FigLang 2024

4th Workshop on Figurative Language Processing

Proceedings of the Workshop

June 21, 2024

©2024 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-110-0

Introduction

Welcome to the 4th Workshop on Figurative Language Processing (FigLang 2024), to be held on June 21, 2024 as part of NAACL in Mexico City, Mexico.

The use of figurative language enriches human communication by allowing us to express complex ideas and emotions. Consequently, it is not surprising that figurative language processing has become a rapidly growing area in Natural Language Processing (NLP), including metaphors, idioms, puns, irony, sarcasm, among others. Characteristic to all areas of human activity (from poetic to ordinary to scientific) and, thus, to all types of discourse, figurative language becomes an important problem for NLP systems. Its ubiquity in language has been established in several corpus studies, and the role it plays in human reasoning has been confirmed in psychological experiments. This makes figurative language an important research area for computational and cognitive linguistics, and its automatic identification and interpretation indispensable for any semantics-oriented NLP application. Recent advent of large language model-based NLP has led to novel techniques for understanding, interpreting, and creating figurative language.

This workshop is the fourth in a series of biannual workshops on Figurative Language Processing (following ACL 2018, ACL 2020, and EMNLP 2022 installments). This new workshop series builds upon the successful start of the Metaphor in NLP workshop series (at NAACL– HLT 2013, ACL 2014, NAACL–HLT 2015, NAACL–HLT 2016), expanding its scope to incorporate the rapidly growing body of research on various types of figurative language such as sarcasm, irony and puns, with the aim of maintaining and nourishing a community of NLP researchers interested in this topic. The workshop features both regular research papers and two shared tasks on Multilingual Euphemism Detection and Multimodal Figurative Language. The workshop is privileged to present one invited talk this year. Dr. Vered Shwartz will be presenting talks at this year's workshop on whether LLMs have solved figurative language.

In the regular research track, we received twenty two research paper submissions and accepted nine. The featured papers cover a range of aspects of figurative language processing such as disagreement in sarcasm detection (Jang et al.), multimodal generation such as images (Khaliq et al.), metaphor detection in cross-lingual setting (Hulsing et al.) annotation guidelines for identifying metaphors (Dippet et al.), metaphor annotation in Mexican Spanish popular science tweets (Montero et al.), expectation-realization model for metaphor detection (Uduehi and Bunescu), idiom detection (Fornaciari et al.), distribution of personification in Hungarian (Simon), and a summary paper on challenges of rhetorical figures detection (Kuhn and Mitrović).

The two shared tasks on Multilingual Euphemism Detection and Multimodal Figurative Language serve to benchmark various computational approaches to euphemism and different types of figurative language, clarifying the state of this steadily growing field and facilitating further research.

In the Multilingual Euphemism Detection Shared Task, participants were invited to develop models to classify texts in various languages as either euphemistic or not. The previous iteration used only an English dataset. This time, we included data in American English (EN), Spanish (ES), Yorùbá (YO), and Mandarin Chinese (ZH) to broaden the insights across languages and facilitate transfer learning for identifying cross-lingual patterns. The datasets consisted of texts from diverse sources including online articles, webpages, transcribed texts, and social media posts. Each text, containing up to three sentences with a potentially euphemistic term (PET), was annotated by humans to indicate euphemistic (1) or non-euphemistic (0) usage. During the development phase, participants were provided with datasets in all four languages. During the test phase, participants were provided a test set for each language and had the option of submitting predictions for one to four of them for scoring. However, all teams ultimately chose to submit predictions for all four. Submissions were evaluated based on the Macro-F1 score,

with equal weighting across languages. Three participating teams submitted system descriptions and achieved scores significantly above baselines but below their reported validation metrics. The different approaches are described in the shared task's summary paper, and the outcomes not only demonstrate the effectiveness of current approaches but also underscore the need for further research into large language models, ensemble techniques, and task-related strategies. Future studies should also explore the broader impact of PETs on model behavior and the potential connections to other linguistic tasks.
The second shared task on understanding figurative language is designed to challenge the participants to build models to not only identify the type of figurative language but also to explain the decision via natural language. The task is based on the recently developed FLUTE dataset, which is based on four types of figurative language – idiom, sarcasm, metaphor, and simile. Out of all the models submitted, four system papers were submitted to the shared task. Although all the submitted models were based on the transformer architecture, participants did attempt different approaches – such as using elaboration of the situation first as additional contexts, sequential training on a variety of NLI datasets, and conducting multi sequence2sequence tasks. Two participants attained the highest accuracy (accuracy@60) scores of 63.33.

Finally, we acknowledge NSF for their generous grant (grant #2226006) with which we are able to support registrations as well as travel and accommodation of a few individual.

We wish to thank everyone who showed interest and submitted a paper, all of the authors for their contributions, the members of the Program Committee for their thoughtful reviews, the invited speakers for sharing their perspective on the topic, and all the attendees of the workshop. All of these factors contribute to a truly enriching event!

Debanjan Ghosh, Smaranda Muresan, Anna Feldman, Tuhin Chakrabarty, Emmy Liu, Workshop Co-Chairs

Organizing Committee

Workshop Organizers

Debanjan Ghosh, Educational Testing Service, USA
Smaranda Muresan, Columbia University, USA
Anna Feldman, Montclair State University, USA
Tuhin Chakrabarty, Columbia University, USA
Emmy Liu, Carnegie Mellon University, USA

Euphemisms Shared Task Organizers

Patrick Lee, Amazon
Anna Feldman, Montclair State University, USA

Multimodal Shared Task Organizers

Arkadiy Saakyan, Columbia University, USA
Shreyas Kulkarni, Columbia University, USA
Tuhin Chakrabarty, Columbia University, USA
Smaranda Muresan, Columbia University, USA

Program Committee

Program Chairs

Tuhin Chakrabarty, SalesForce Research
Anna Feldman, Montclair State University and Montclair State University
Debanjan Ghosh, Educational Testing Service and Massachusetts Institute of Technology
Emmy Liu, School of Computer Science, Carnegie Mellon University
Smaranda Muresan, Columbia University

Area Chairs

Tuhin Chakrabarty, SalesForce Research
Anna Feldman, Montclair State University and Montclair State University
Debanjan Ghosh, Educational Testing Service and Massachusetts Institute of Technology
Emmy Liu, School of Computer Science, Carnegie Mellon University
Smaranda Muresan, Amazon and Columbia University

Reviewers

Arav Kapish Agarwal, Khalid Alnajjar

Yulia Badryzlova, Yuri Bizzoni

Tuhin Chakrabarty, Jiale Chen

Verna Dankers, Rahul Divekar, Jonathan Dunn

Anna Feldman, Michael Flor

Lingyu Gao, Debanjan Ghosh

Nicholas Hankins, Mika Hämäläinen

Loukas Ilias

Hyeju Jang

Raghav Kapoor, Sedrick Keh, Ilker Kesen, Guneet Singh Kohli, Valia Kordoni, Shreyas Kulkarni

Yash Kumar Lal, Mark G. Lee, Patrick Lee, Els Lefever, Emmy Liu, Ziqian Luo

Smaranda Muresan, Elena Musi

Geetanjali Rakshit, Ellen Riloff, Anthony Rios, Andrés Torres Rivera

Arkadiy Saakyan, Farig Sadeque, Eyal Sagi, Sabine Schulte Im Walde, Meghdut Sengupta, Egon Stemle, Kevin Stowe, Carlo Strapparava, Tomek Strzalkowski, Stan Szpakowicz

Yufei Tian, Xiaoyu Tong

Tony Veale, Fedor Vitiugin

Peratham Wiriyathammabhum

Qihao Yang

Keynote Talk

Did language models "solve" figurative language?

Vered Shwartz
University of British Columbia
2024-06-21 –

Abstract: Figurative expressions, such as idioms, similes, and metaphors, are ubiquitous in English. For many years, they have been considered a pain in the neckfor NLP applications, due to their non-compositional nature. With LLMs excelling at understanding and generating English texts, it's time to ask: did LLMs solvefigurative language? Is it possible that the sheer amount of exposure to figurative language in their training data equipped them with the ability to understand and use figurative language? I will discuss the state of LLMs in recognizing figurative usage, interpreting figurative expressions in context, and usage of figurative language in generated text.

Bio: Vered Shwartz is an Assistant Professor of Computer Science at the University of British Columbia. Her research interests include commonsense reasoning, computational semantics and pragmatics, and multiword expressions. Previously, Vered was a postdoctoral researcher at the Allen Institute for AI (AI2) and the University of Washington, and received her PhD in Computer Science from Bar-Ilan University. Vered's work has been recognized with several awards, including The Eric and Wendy Schmidt Postdoctoral Award for Women in Mathematical and Computing Sciences, the Clore Foundation Scholarship, and an ACL 2016 outstanding paper award.

Table of Contents

Program

Friday, June 21, 2024

08:50 - 09:00 *Opening Remarks*

09:00 - 10:30 *Research Track*

Context vs. Human Disagreement in Sarcasm Detection
Hyewon Jang, Moritz Jakob and Diego Frassinelli

The Register-specific Distribution of Personification in Hungarian: A Corpus-driven Analysis
Gabor Simon

Comparison of Image Generation Models for Abstract and Concrete Event Descriptions
Mohammed Abdul Khaliq, Diego Frassinelli and Sabine Schulte Im Walde

Cross-Lingual Metaphor Detection for Low-Resource Languages
Anna Hülsing and Sabine Schulte Im Walde

A Hard Nut to Crack: Idiom Detection with Conversational Large Language Models
Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios and Maite Melero

The Elephant in the Room: Ten Challenges of Computational Detection of Rhetorical Figures
Ramona Kühn and Jelena Mitrović

10:30 - 11:00 *Coffee Break*

11:00 - 12:30 *Research Track + Shared Tasks*

Report on the Multilingual Euphemism Detection Task
Patrick Lee and Anna Feldman

A Report on the FigLang 2024 Shared Task on Multimodal Figurative Language
Shreyas Kulkarni, Arkadiy Saakyan, Tuhin Chakrabarty and Smaranda Muresan

Friday, June 21, 2024 (continued)

Optimizing Multilingual Euphemism Detection using Low-Rank Adaption Within and Across Languages
Nicholas Hankins

Guidelines for the Annotation of Intentional Linguistic Metaphor
Stefanie Dipper, Adam Roussel, Alexandra Wiemann, Won Kim and Tra-my Nguyen

Evaluating the Development of Linguistic Metaphor Annotation in Mexican Spanish Popular Science Tweets
Alec Misael Sánchez Montero, Gemma Bel-Enguix, Sergio-Luis Ojeda-Trueba and Marisela Colín Rodea

Can GPT4 Detect Euphemisms across Multiple Languages?
Todd E Firsich and Anthony Rios

12:30 - 14:00 *Lunch Break*

14:00 - 15:00 *Keynote Talk 1: Vered Shwartz: Did language models "solve" figurative language?*

15:00 - 15:30 *Research Track + Shared Tasks*

A Textual Modal Supplement Framework for Understanding Multi-Modal Figurative Language
Jiale Chen, Qihao Yang, Xuelian Dong, Xiaoling Mao and Tianyong Hao

An Expectation-Realization Model for Metaphor Detection: Within Distribution, Out of Distribution, and Out of Pretraining
Oseremen Oscar Uduehi and Razvan Bunescu

15:30 - 16:00 *Coffee Break*

16:00 - 16:30 *Shared Tasks*

Ensemble-based Multilingual Euphemism Detection: a Behavior-Guided Approach
Fedor Vitiugin and Henna Paakki

FigCLIP: A Generative Multimodal Model with Bidirectional Cross-attention for Understanding Figurative Language via Visual Entailment
Qihao Yang and Xuelin Wang

Context vs. Human Disagreement in Sarcasm Detection

Hyewon Jang[1], **Moritz Jakob**[1], **Diego Frassinelli**[1,2]
[1]Department of Linguistics, University of Konstanz, Germany
[2]Center for Information and Language Processing, LMU Munich, Germany
{hye-won.jang, moritz.jakob, diego.frassinelli}@uni-konstanz.de

Abstract

Prior work has highlighted the importance of context in the identification of sarcasm by humans and language models. This work examines *how much* context is required for a better identification of sarcasm by both parties. We collect textual responses to dialogical prompts and sarcasm judgment to the responses placed after long contexts, short contexts, and no contexts. We find that both for humans and language models, the presence of context is generally important in identifying sarcasm in the response. But increasing the amount of context provides no added benefit to humans (long = short > none). This is the same for language models, but only on easily agreed-upon sentences; for sentences with disagreement among human evaluators, different models show different behavior. Also, we show how, despite the low agreement in human evaluation, the sarcasm detection patterns by the manipulation of context amount stay consistent.

1 Introduction and related work

This work examines the role of the presence and amount of contextual information in detecting sarcasm. Previous work in cognitive science has shown the importance of context in sarcasm comprehension (Woodland and Voyer, 2011) and production (Jang et al., 2023) for humans. In computational linguistics, similar observations were made: supplying context to the target utterance boosts sarcasm detection performance of language models, though with more conflicting results: some studies report that supplying context leads to a performance boost in sarcasm detection by neural models (Jaiswal, 2020; Ghosh et al., 2018), whereas other studies report no such benefit (Castro et al., 2019) or marginal benefit (Jang and Frassinelli, 2024) in using context for the same task. However, there has not been much effort in exploring the benefit of varying amounts of contextual information, or in addressing what counts as context. The term 'context' varies a lot work by work; it can mean any number of preceding strings such as previous posts on social media (Jaiswal, 2020; Joshi et al., 2016) or previous utterances in a dialogue (Castro et al., 2019), or any additional information that can help detect sarcasm, such as eye-tracking data (Mishra et al., 2016) or images (Schifanella et al., 2016).

In this work, we define context as the preceding textual utterances that can trigger sarcasm in people (Section 2), and then examine what is a good amount of contextual information that facilitates sarcasm identification for humans (Section 3) and language models (Section 4). We further show how context interacts with the level of disagreement among human evaluators (Section 4.3).

2 Data creation

We created a new dataset based on the Multimodal Sarcasm Detection Dataset (MUStARD; Castro et al., 2019). The MUStARD dataset contains written transcriptions of "contexts" (preceding utterances) and the following "response"[1] from multiple TV series, and binary labels of sarcasm for the responses (*sarcastic* or *not sarcastic*). We selected 24 contexts that are generalizable enough, all of which were from the TV series 'Friends' and situations happening between two conversation partners. The names of all conversation partners were modified to detach the stimuli from the TV show as much as possible. For all the selected contexts, we collected new responses in an online data collection.

Here, we manipulated the amount of context. Additional to the original contexts available in a short utterance form, we described each context in a narrative form by manually referring to the scenes and episodes of the TV show to restore the relevant information that would allow the following utterance to be correctly judged as sarcastic or not. This information in the original dataset often came

[1]The term used in the MUStARD dataset is 'utterance'.

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 1–7
June 21, 2024 ©2024 Association for Computational Linguistics

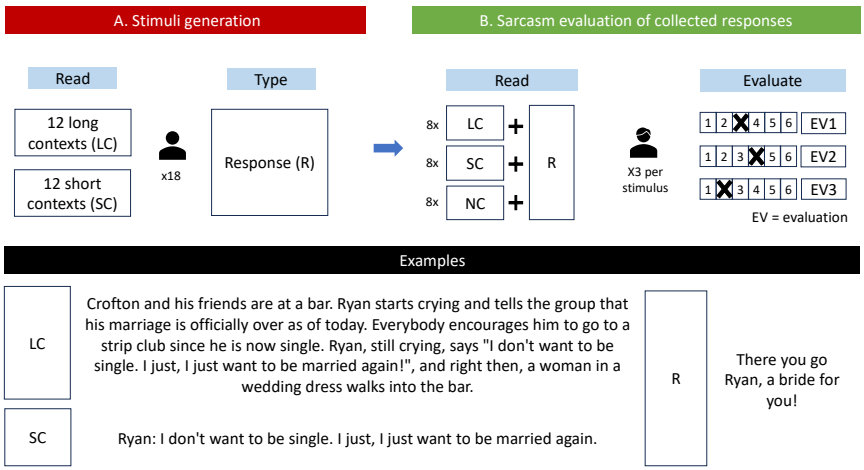


Figure 1: Data collection (A), data evaluation (B), and example stimuli for long (LC) and short (SC) contexts, and an example response (R) collected from participants.

from multimodal, episode-level, or series-level information not reflected in the transcripts.

Therefore, each context was represented twice both as *short context* (SC) in its original utterance form and as *long context* (LC) in a descriptive/narrative form. The average number of words was 26 for SC and 66 for LC. For each LC and SC, we collected new responses to make the stimuli comparable, given that the original dataset had responses only to short contexts. This also allowed us to collect spontaneous responses from multiple lay people as opposed to responses generated by professional screenwriters.

We recruited 32 native English-speaking participants based in the UK, USA, Canada, Australia, New Zealand or Ireland[2]. They read 24 contexts and freely responded to each (they were not instructed to be sarcastic). Half of the contexts (N = 12) were presented as SC and the other half as LC (See A in Figure 1). At the end of the collection, participants reported their familiarity to the TV show Friends and how many of the situations they recognized as being from the show.

To control for the expectation of sarcasm arising from the familiarity to the TV show, we discarded data from the participants who were *quite familiar*, *very familiar*, or *extremely familiar* to the show or who recognized at least 3 scenes from the show. After removing data from 14 such participants, data by 18 respondents remained.[3]

[2]We used FindingFive (https://www.findingfive.com) for experiment building and Prolific (https://www.prolific.co) for participant recruitment.

[3]The new data consisting of responses and evaluation ratings are available at https://github.com/copsyn.

3 Influence of context for sarcasm judgment by humans

Here we identify what amount of context affects human judgment of sarcasm on the following response.

3.1 Experiment

In an online experiment, new participants evaluated the level of sarcasm of the responses in isolation (NC) or placed after long context (LC) or short context (SC) as shown in Table 1.

Table 1: Number of items for different combinations of context (**C**) and response (**R**).

	Condition	N
i	SC (24) + R (18)	432
ii	LC (24) + R (18)	432
iii	NC (R-only)	432
Total		1,296

In conditions **i** and **ii**, each context is paired with the generated responses and condition **iii** consists of the responses only (See Section 2).

Each stimulus was evaluated by 3 participants recruited with the same criteria as before. Each participant was presented with 24 stimuli, distributed evenly across the 3 conditions (See B in Figure 1). Participants rated the sarcasm level of the responses on a six-point Likert scale (*not at all*, *mostly not*, *not so much*, *somewhat*, *mostly*, and *completely*). Participants who failed attention check questions

or were familiar with the TV show were replaced with new ones.

Context length and disagreement Table 2 shows the proportions of sarcasm (binary-coded from the six-point scale; *completely*, *mostly*, *somewhat* into *sarcastic*) in each contextual condition by three evaluators and by their average per stimulus. The probability of judging a response as sarcastic increases when contextual information is present. Around 38% of instances that were judged as 'not sarcastic' in the NC condition were judged as 'sarcastic' when more context became available (LC or SC condition). However, adding context also increases disagreement among evaluators (lower Kappa).

Table 2: Proportions of sarcastic responses (binary-coded) by context amount according to three distinct evaluations per stimulus (EVs) and inter-rater agreement (Fleiss' Kappa) by context amount.

	AVG	EV1	EV2	EV3	Kappa
LC	0.46	0.36	0.44	0.49	0.10
SC	0.42	0.36	0.43	0.41	0.13
NC	0.23	0.25	0.25	0.28	0.18

3.2 Analysis and results

We tested whether the presence and amount of contextual information are important factors for humans to identify sarcasm in the following response. To easily compare the behavior of humans and LMs, we binarized the sarcasm ratings. The overall inter-rater agreement across all stimuli measured by Fleiss' Kappa was 0.17 (See Appendix B for Spearman correlations).[4]

We fit a generalized linear mixed-effects model for each evaluation (See Appendix C for details)[5]. Random intercepts for participants and items were included in the statistical model. We used R (R Core Team, 2021) and the *lme4*-package (Bates et al., 2015) for the main models and the *emmeans*-package for post-hoc pairwise comparisons (Lenth, 2023).

For all evaluations, the presence of context, either long or short, triggered significantly higher probability of perceiving sarcasm in the following response. Long contexts caused more frequent sarcasm judgment compared to short contexts only in EV3 ($p < 0.005$), but not in EV1 ($p = 0.98$), EV2 ($p = 0.97$), or AVG ($p = 0.27$). The results indicate that the presence of context is important for human evaluators to identify sarcasm, but a greater amount of context does not necessarily lead to any added benefit.

[4]For comparison, the Kappa score reported in the original MUStARD paper is 0.23 (Castro et al., 2019).

[5]In this work, unless otherwise specified, statistically significant scores correspond to a p-value smaller than 0.001.

4 Influence of context on sarcasm detection by large language models

Here we test if manipulating the amount of context directly affects the performance of three language models in the detection of sarcasm on the following response. As gold standard we use the human-evaluated scores described in Section 3.

4.1 Data and model

We performed sarcasm detection using three pretrained LMs: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DistilBERT (Sanh et al., 2019). We fine-tuned these models on the contexts and responses from the MUStARD dataset excluding the 24 contexts we used in our experiments. We then used our data as test data to classify the responses in the three conditions (LC, SC, NC) as either *sarcastic* or *not sarcastic*. Given the high subjectivity in identifying sarcasm indicated by the low inter-rater agreement (Kappa 0.17), we predicted the binary-coded human ratings by the three evaluations (EVs) independently and combined. We conducted an error analysis comparing the results from the three EVs. We used four different seeds and five folds for validation. All the reported results in this paper are an average of all the models (4 seeds $\times$ 5 folds) trained for 10 epochs, which yielded the best prediction results. See Appendix A for the full model parameters.

Table 3: Macro F-scores of sarcasm detection on the new dataset described in Section 2 by three LMs trained on MUStARD for 10 epochs. Labels provided by each evaluation (EV) or combined (C) across three EVs.

		EV1	EV2	EV3	C
BERT	LC	0.49	0.52	0.57	0.55
	SC	0.53	0.51	0.53	0.54
	NC	0.47	0.39	0.41	0.36
RoBERTa	LC	0.46	0.54	0.52	0.53
	SC	0.54	0.50	0.52	0.50
	NC	0.36	0.34	0.38	0.29
DistilBERT	LC	0.53	0.52	0.51	0.53
	SC	0.53	0.51	0.52	0.53
	NC	0.44	0.38	0.40	0.32

4.2 Results

Overall, the three LMs achieve comparable classification results. Supplying context, either short or long, always improves the performance of all LMs. The performance results in Table 3 suggest that there are no strong differences between supplying long context and short context. A noteworthy aspect of these results is that despite low agreement among three evaluations, the prediction results by context amount show similar patterns for all EVs (LC and SC lead to a higher number of correct predictions than NC).

4.3 Error analysis

Disagreement among human evaluators To identify the reasons behind the similar patterns in model performance despite low agreement, we divided the data into *agreed-upon* (all evaluators agreed on a label) and *disagreed-upon* (evaluators disagreed on the label: 2 vs. 1) instances of sarcasm based on the binarized labels. From the *disagreed-upon* category, we extracted the number of instances for which LMs chose the majority label (better choice) or the minority label (worse choice), neither of which is completely correct or incorrect. Table 4 shows that LMs choose the labels given by each evaluation at a similar rate. This pattern suggests that LMs misclassify some sentences when tested with labels from one evaluation, but misclassify other sentences when tested with labels from another evaluation, thus holding the general classification patterns stable.

Table 4: Proportions (Prop.) of predictions by BERT. Correct & incorrect predictions apply to *agreed-upon* (A) instances. Majority (better choice) & minority (worse choice) predictions apply to *disagreed-upon* (D) instances. The other models show the same pattern (See Appendix D).

Type	Prediction	Evaluations that predictions match			Prop.
A	Correct	All			0.58
	Incorrect	None			0.42
		Match_EV1	**Match_EV2**	**Match_EV3**	
D	Majority	0	1	1	0.18
		1	0	1	0.17
		1	1	0	0.17
	Minority	0	0	1	0.15
		0	1	0	0.16
		1	0	0	0.17

The interaction between context amount and degree of disagreement To analyze the interaction between the amount of context (LC, SC, NC) and disagreement levels (agreed vs. disagreed), we categorized the predicted labels according to these factors. Table 5 shows that for *agreed-upon* instances, providing context helps LMs predict (more) correct labels than when no contexts are available (LC/SC >NC for correct & majority). For *disagreed-upon* instances, more variability is shown: For BERT, only long context significantly improves the detection of sarcasm (LC >SC = NC), whereas for RoBERTa and DistilBERT, no amount of context is beneficial (LC = SC = NC).

Table 5: Proportions of classification choice of BERT (average across all seeds and folds) by context length × disagreement level.

		Agreed-upon			**Disagreed-upon**		
		Correct	Incorrect	Std.	Majority	Minority	Std.
BERT	LC	0.60	0.40	0.07	0.54	0.46	0.04
	SC	0.60	0.40	0.07	0.51	0.49	0.05
	NC	0.55	0.45	0.16	0.50	0.50	0.06
RoBERTa	LC	0.61	0.39	0.09	0.50	0.50	0.04
	SC	0.57	0.43	0.08	0.48	0.52	0.04
	NC	0.52	0.48	0.14	0.49	0.51	0.06
DistilBERT	LC	0.57	0.43	0.08	0.52	0.48	0.05
	SC	0.59	0.41	0.10	0.51	0.49	0.05
	NC	0.54	0.46	0.19	0.50	0.50	0.08

In summary, the presence of context is important for LMs to significantly improve their performance of sarcasm detection for sentences with a high agreement, but adding more context does not present clear benefit compared to a lower amount of context. For sentences with disagreement, the contribution of contextual information heavily depends on each model. Only BERT uses the extra contextual information provided by a longer context to detect sarcasm significantly better.

5 Conclusion

This work systematically tested the amount of contextual information required for humans and language models to evaluate the following utterance in terms of sarcasm. We showed that in general, the presence of context leads to better detection of sarcasm both by humans and by three LMs. But, providing a higher amount of information in the context did not present clear additional benefit for humans, which was also true for LMs for sentences for which human evaluators agreed on a label. When humans disagreed, the presence of context stopped playing any role in facilitating the detection of sarcasm in RoBERTa and DistilBERT, whereas the performance of BERT improved when a longer context was provided. We lastly showed

that low inter-rater agreement did not affect the overall classification patterns, due to a high variability in the sentences that the models misclassify each time they are tested against labels from different human evaluators. This is a relevant finding for many NLP tasks prone to disagreement and susceptible to subjectivity, which must continue to be addressed in future research.

Limitations

This work investigated the influence of the amount of information embedded in the context. However, we did not systematically calculate the amount of information available in the different contextual conditions (SC vs. LC). Future work should address how to draw a line between sufficient and redundant contextual information by investigating a gradient change in the amount of context.

The data collected in this work is small because we had to go through rigorous filtering of an existing dataset to obtain sufficiently generalizable contexts for further experiments. Future work should test the same effect with a bigger sample size.

In the data collection (Section 2), we only recruited male participants because some of the selected situations were much more suitable for male speakers than female speakers and the already small number of generalizable contexts could not be further reduced. A follow-up study should include gender as a variable for a more comprehensive evaluation of the use of sarcasm by humans.

Ethics Statement

We see little ethical issue related to this work. All our experiments involving human participants were conducted anonymously, on a voluntary basis, and with a fair compensation suggested by the recruitment platform Prolific (9 GBP per hour) and are in line with the ethical regulations of the University of Konstanz (IRB number 05/2021). All our modeling experiments were conducted with open-source libraries, which received due citations. However, we acknowledge that some of the stimuli extracted from the original MUStARD dataset contain sensitive language that could potentially be insulting for the reader.

Acknowledgements

We thank Matteo Guida and Hsun-Hui Lin for their work in initial data selection and preparation.

References

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Debanjan Ghosh, Alexander R. Fabbri, and Smaranda Muresan. 2018. Sarcasm Analysis Using Conversation Context. *Computational Linguistics*, 44(4):755–792.

Nikhil Jaiswal. 2020. Neural sarcasm detection using conversation context. In *Proceedings of the second workshop on figurative language processing*, pages 77–82.

Hyewon Jang, Bettina Braun, and Diego Frassinelli. 2023. Intended and perceived sarcasm between close friends: What triggers sarcasm and what gets conveyed? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.

Hyewon Jang and Diego Frassinelli. 2024. Generalizable sarcasm detection is just around the corner, of course! *arXiv preprint arXiv:2404.06357*.

Aditya Joshi, Vaibhav Tripathi, Pushpak Bhattacharyya, and Mark J. Carman. 2016. Harnessing sequence labeling for sarcasm detection in dialogue from TV series 'Friends'. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 146–155, Berlin, Germany. Association for Computational Linguistics.

Russell V. Lenth. 2023. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.8.6.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2016. Harnessing

cognitive features for sarcasm detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104, Berlin, Germany. Association for Computational Linguistics.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Rossano Schifanella, Paloma De Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1136–1145.

Jennifer Woodland and Daniel Voyer. 2011. Context and intonation in the perception of sarcasm. *Metaphor and Symbol*, 26(3):227–239.

A Fine-tuning implementation details

We used `bert-base-uncased`, `roberta-base`, and `distilbert-base-uncased`. Each language model was fine-tuned for 2, 5, and 10 epochs with a batch size of 64, a learning rate of 5e-5, and a weight decay of 1e-2. The fine-tuning was implemented using the Trainer class from the Hugging Face library, and conducted on an NVIDIA A100 GPU with a total memory of 40GB.

B Inter-rater agreement

Table 6 reports the Spearman's correlation coefficients (r) calculated between the original ratings (1-6 Likert scale) that each evaluation group (EV) assigned to responses alone (NC) and responses following long contexts (LC) or short contexts (SC). The trends observed here are consistent with the results on the binarized sarcasm scores reported in Table 2 in the main text.

Table 6: Inter-rater agreement of the original ratings (1-6) measured by Spearman's correlations between each pair of evaluation (EV), $p < 0.005$.

	EV1-EV2	EV1-EV3	EV2-EV3
LC	0.26	0.17	0.15
SC	0.26	0.17	0.19
NC	0.18	0.24	0.20

C Details of statistical tests

The formula used for the GLMER models is as follows:

sarcasm_binary_labels ∼
context_amount
+ (1 | item) + (1 | participant)

The model indicates if there are differences in the sarcasm label (yes/no) distribution given contextual manipulation. The random intercepts account for the variability between participants and items that cannot be explained by the fixed effects alone.

The *emmeans* library conducts a pairwise comparison of the three context conditions (LC vs. SC, LC vs. NC, and SC vs. NC) by performing automatic alpha correction.

D Error analysis for the other models

Proportions of predictions by RoBERTa (see Table 7) and DistilBERT (see Table 8).

Table 7: Proportions (Prop.) of predictions by RoBERTa. Correct & incorrect predictions apply to *agreed-upon* (A) instances. Majority (better choice) & minority (worse choice) predictions apply to *disagreed-upon* (D) instances.

Type	Prediction	Annotator groups that predictions match			Prop.
A	Correct	All			0.56
	Incorrect	None			0.44
		Match_EV1	**Match_EV2**	**Match_EV3**	
D	Majority	0	1	1	0.16
		1	0	1	0.16
		1	1	0	0.17
	Minority	0	0	1	0.18
		0	1	0	0.16
		1	0	0	0.17

Table 8: Proportions (Prop.) of predictions by DistilBERT. Correct & incorrect predictions apply to *agreed-upon* (A) instances. Majority (better choice) & minority (worse choice) predictions apply to *disagreed-upon* (D) instances.

Type	Prediction	Annotator groups that predictions match			Prop.
A	Correct	All			0.56
	Incorrect	None			0.44
		Match_EV1	**Match_EV2**	**Match_EV3**	
D	Majority	0	1	1	0.17
		1	0	1	0.17
		1	1	0	0.18
	Minority	0	0	1	0.17
		0	1	0	0.16
		1	0	0	0.16

Optimizing Multilingual Euphemism Detection using Low-Rank Adaptation Within and Across Languages

Nicholas Hankins

nicholasjhankins@gmail.com,
hankinsn1@montclair.edu

Abstract

This short paper presents an investigation into the effectiveness of various classification methods as a submission in the Multilingual Euphemism Detection Shared Task for the Fourth Workshop on Figurative Language Processing co-located with NAACL 2024. The process utilizes pre-trained large language models combined with parameter efficient fine-tuning methods, specifically Low-Rank Adaptation (LoRA), in classifying euphemisms across four different languages - Mandarin Chinese, American English, Spanish, and Yorùbá. The study is comprised of three main components that aim to explore heuristic methods to navigate how base models can most efficiently be fine-tuned into classifiers to learn figurative language. Multilingual labeled training data was utilized to fine-tune classifiers for each language, and later combined for one large classifier, while unseen test data was finally used to evaluate the accuracy of the best performing classifiers. In addition, cross-lingual tests were conducted by applying each language's data on each of the other language's classifiers. All of the results provide insights into the potential of pre-trained base models combined with LoRA fine-tuning methods in accurately classifying euphemisms across and within different languages.

1 Introduction

In order to best understand this task, it is important to define what a euphemism is. Euphemisms are a linguistic device used to soften statements, or to make statements more polite. Some examples of a euphemism might be using the terms "between jobs" or "late" instead of "unemployed" or "dead," respectively (Lee et al. 2024). Research proves that euphemisms are a multilingual feature that exists in numerous languages (Gavidia et al. 2022). By collecting more training data and testing on unseen data, we are further able to see the extent of how state-of-the-art language modeling captures these universally figurative traits.

The ability to observe whether these elements of figurative language are taken into consideration during tasks like classification by large language models (LLM) can be speculated as a topic of increasing interest in natural language processing communities. The growing number of base models, such as XLM-RoBERTa, that can be utilized for downstream tasks like text classification, reasoning, and sequence generation is staggering and leads to further questions of how the existing methods can be tested and improved (Conneau et al. 2019). By addressing the numerous kinds of euphemistic categories, and how they can be represented multilingually, this kind of research enables a greater level of natural language understanding by embodying an ambiguous and subjective aspect of languages (Lee et al. 2024). Furthermore, by aiming to solve the problem of accurate classification of figurative language using machine learning, this task importantly measures how well a human language characteristic can be interpreted by LLMs.

2 Related Work

Prior research determined that semantic category might influence cross-lingual transfer of information (Lee et al. 2024). This insight drives the intuition for this experiment. Once the ostensibly optimal classification method is discovered, then we can perform a cross-lingual comparison to see how all other languages performed on classifiers fine-tuned for other languages. Previous work is helpful in this regard, as it enables us to have a starting point to compare and contrast base models, which were chosen heuristically. The Multilingual

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 8–14
June 21, 2024 ©2024 Association for Computational Linguistics

Classifier predictions will be included in the results as a comparison. It is important to note that the base model for the cross-lingual experiment remained the same. In other words, all languages had context for each inference, yet the fine-tuning the classifier certainly made a difference in the results. The complete visualization of this can be seen in Figure 1.

By freezing all the parameters in the base model with the Parameter-Efficient Fine-Tuning (PEFT) method, we are able to explicitly train the new classifier on our input language datasets. This will ideally be able to increase the speed with which we fine-tune and keep majority of the base model parameters frozen (Hu et al. 2021). Using parameter efficient fine-tuning essentially allows our model to be trained on a small set of new parameters, which is why PEFT was used for this experiment. The goal is to see how well we can utilize these methods for our classification purposes. Multilingual word embeddings have been shown to also have produced positive results in text classification tasks (Plank 2017).

Through examining the problem posed in the introduction of best classifying figurative language through fine-tuning LLMs, and incorporating adjacent work that has proven successful on tasks with similar goals, we can begin to formulate an overarching methodology. Starting with different base models to explore a variety of new options, keeping in mind the limited compute resources available, we can focus our efforts in changing as little as possible from the base model in an effort to highlight the impact of the task's training and test data throughout the experiment. LoRA, specifically PEFT, allows us to do this by inspecting the given data, specifically how semantic information is transferred accordingly, in the model predictions. The aim is to emphasize how the arrangement of the model's data can affect classification predictions.

3 Experiment 1 - Choosing the Base Model

3.1 Methodology

The training data included examples with a column for the text containing the Potentially Euphemistic Term (PET), the assigned label of 1 signifying that the term is euphemistic, or 0 if not, and the PET category. After obtaining the training data, one of the primary characteristics observed was the imbalance of the data. More specifically, each language had more positive, or euphemistic, labels which indicates that the training datasets are imbalanced (See Table 1). This imbalance problem was addressed with adjusting the learning rate during hyperparameter weight setting.

	En.	Sp.	Ch.	Yo.
Euph.	1383	1143	1484	1281
Non-Euph.	569	718	521	660

Table 1: Training Data Split between Euphemistic and Non-Euphemistic Examples (English, Spanish, Chinese, Yorùbá)

It is important to note the unique counts of PET categories present in each dataset considering the impact that they might have, despite the PET category not being explicitly included in the fine-tuning process. Only the text and label columns were input into the trainer function. Nonetheless, this way, we can intuitively observe the classifier results and see the PET characteristics, such as quantity, uniqueness, and frequency in the data. The total counts by themselves do not provide much insight given that each example has one by design, however, it is helpful to know the count of unique PET categories that may be prompting the fine-tuning step with semantically relevant information embedded within the associated input text. English has 163 unique PET categories, Spanish has 147, Chinese has 110, and Yorùbá has 133.

All the classifiers were trained on T4 GPU and incorporated the models, tokenizers, and LoRA adapters via HuggingFace's platform (Wolf et al. 2020). They were all preprocessed the same way by utilizing an Autotokenizer and were set to initiate data collation for training (*YouTube-Blog* 2024). While tuning the hyperparameters, it was discovered that having all the linear target modules instantiated maximized the number of trainable parameters, as supported by prior studies into LoRA techniques (Dettmers et al. 2023). This meant that the most crucial hyperparameter was the number of LoRA adapters, thus ensuring full capability of fine-tuning performance (Dettmers et al. 2023). The number of LoRA linear layers included in the

Languages	First Test	Second Test
English	0.85045	0.84158
Spanish	0.77704	0.74321
Chinese	0.84438	0.85454
Yorùbá	0.81308	0.80423

Table 2: Maximum Validation F1 scores of the 10 epochs for both experiments. The final epoch results may be lower during inference on the test data.

PEFT model instantiation was 6 in total [1].

Moreover, all the classifiers ultimately were set to having a learning rate of 1e-5 and trained on 10 epochs in order to create consistency. It should be noted that the original metric scores of the first base model experiments had varying learning rates, which may have had an impact on the training process due to the inherent data imbalance.

The first iteration of fine-tuning used the uncased DistilBERT base model for English (Sanh et al. 2019), main branch of XLM-Roberta for Spanish and Chinese (Conneau et al. 2019), and a fine-tuned version of XLM-Roberta for Yorùbá (Adelani 2021). After that, in the second iteration, the cased multilingual DistilBERT base model was utilized for each language classifier's training due to its ability to train more quickly without forfeiting performance on predicting output labels (Sanh et al. 2019). The process included splitting the data into a training set and a test set of 80/20, respectively. This split was the same for both experiments. The metric that would be used in the shared task competition was Macro F1, so the efforts of enhancing the training process made sure to especially track those results in the trainer outputs.

At this point, the decision for which base models to incorporate in training the classifiers was made after observing changes in model performance after each epoch output. Some undesirable trends were noticed, such as overfitting in one case as suggested by an increasing validation loss in the uncased DistilBERT base model for English. This trial and error process facilitated the choice for which base model would be used later by ruling out the options that do not perform well.

3.2 Results

The cased multilingual DistilBERT base model proved to be the better option moving forward, since the difference in maximum F1 validation scores were marginal, and keeping this model as the base one allowed for consistency in creating one large multilingual classifier. The reason for this is because all four languages in this experiment were all included in that particular base model's training data. Given that the cased multilingual DistilBERT model was originally chosen as simply a new option to explore, combined with its lightweight characteristics, the decision was then confirmed to move forward with a uniform base model due to its ability to include all of the languages, an increased training time efficiency, sufficient F1 metric performance, and a confidence in the prediction labels (See Table 2 for more details). The prediction labels held great importance in seeing how the configuration of the models and the training data impacted the final results.

This importance of prediction label analysis was another significant contributing factor to abandoning the implementation of different base models for a consistent one in how some languages appeared to have exceptional F1 scores during training, yet when tested on the data, the prediction labels were incredibly wrong. For example, training Mandarin Chinese on XLM-RoBERTa proved to have high F1 and accuracy scores (using glue and mrpc), yet when the training data was tested as an inference, everything was labeled as euphemistic.

4 Experiment 2 - Multilingual Classifier and Cross-Lingual Comparison

4.1 Methodology

Since there have been positive results making a large multilingual classifier for text classification, the next step of this paper will detail how that process was completed for this shared task (Plank 2017). In an effort to maximize the F1 validation scores, the first step was concatenating the data so it would all be trained at the same time. Once it was prepared, the training pipeline remained the same. That is, LoRA was used again for its ability to keep

[1] https://github.com/nhankins/multilingual-euphs-figlang2024

most parameters the same as the base model, drawing attention to the training data in particular. The results can be found in Figure 1, or numerically in Table 4 and Figures 2-5.

Another aspect of this paper focuses on how information is transferred across languages. This portion ran concurrently with the multilingual portion to see if there was a major difference in the results when compared side-by-side. In each of these experiments, it is important to emphasize the motivations in choosing what constitutes a classifier as being better is directly related to its ability to both satisfy higher Macro F1 validation scores, but give confident label predictions on completely unseen data. This was done as a way to succeed in meeting the shared task requirements, and likewise further improve figurative language text classification.

At this point, the study requires verification that there is indeed an effect in using one language classifier with a specific dataset over another. The expectation is that the languages will output more accurate predictions on the classifier which has been fine-tuned with its own language. Therefore, the cross-lingual exercise demonstrates the results of this expectation.

4.2 Results

Analyzing the euphemistic and non-euphemistic splits from all 4 individual classifiers did not appear to yield any glaringly significant observations, yet when visualized it became easier to see overarching correlations (See Figure 1). Languages did not always appear to align more closely with the multilingual classifier predictions even on their own languages, which suggests that the greater quantity of training data plays an important role in favorable predictions. The Multilingual results were added to show contrast between the splits, noting that the multilingual classifier performed better than the individual ones. The detailed shared task final results on the test data of both methods can be found in the appendix, yet the F1 scores are as follows in Table 3.

Languages	**Individual**	**Multilingual**
English	0.57	0.64
Spanish	0.59	0.60
Chinese	0.59	0.68
Yorùbá	0.60	0.65

Table 3: Final F1 Scores for the shared task after submitting predictions. The First experiment (individual classifier) showed consistently lower values for all languages compared with the Second experiment (multilingual classifier).

	En.	Sp.	Ch.	Yo.
Euph.	687	714	794	486
Non-Euph.	509	377	432	183

Table 4: Predicted labels on Test Data using Multilingual Classifier (English, Spanish, Chinese, Yorùbá)

5 Conclusion

In conclusion, we learned that the cased multilingual DistilBERT base model proved to have a faster performance and learned more about the training data during fine-tuning. Despite not much change in the metrics output, the adherence of predicted labels to the ground truth gold standard labels was much closer.

Some concluding speculations could be that English and Spanish potentially have lower success rates in predicting whether a term is being used euphemistically or not due to less ambiguity in the instances for which they are being used. This is a curious assertion to prove in future work as the definition of what is ambiguous varies between speakers of a language. A major consideration that should also be noted, and potentially the subject of future work, is the impact that the unique number of PET categories has on the training process. English, for example, as mentioned before has the highest number, whereas Chinese has the lowest number. As mentioned previously, the data imbalance problem was addressed with learning rate adjustment, due to concerns that alternative methods, such as undersampling, might eliminate crucial semantic information. Another factor that should be noted is that Chinese and Yorùbá both needed to have truncation at inference time encoding, most likely due to BERT models using word-piece tokenization (Devlin et al. 2018). In other words, they saw more unknown words in their vocabularies, thus needing them to create more tokens and increasing the total length of the sequence for each example. Future work could explore if token length, language family, more balanced training data, or different

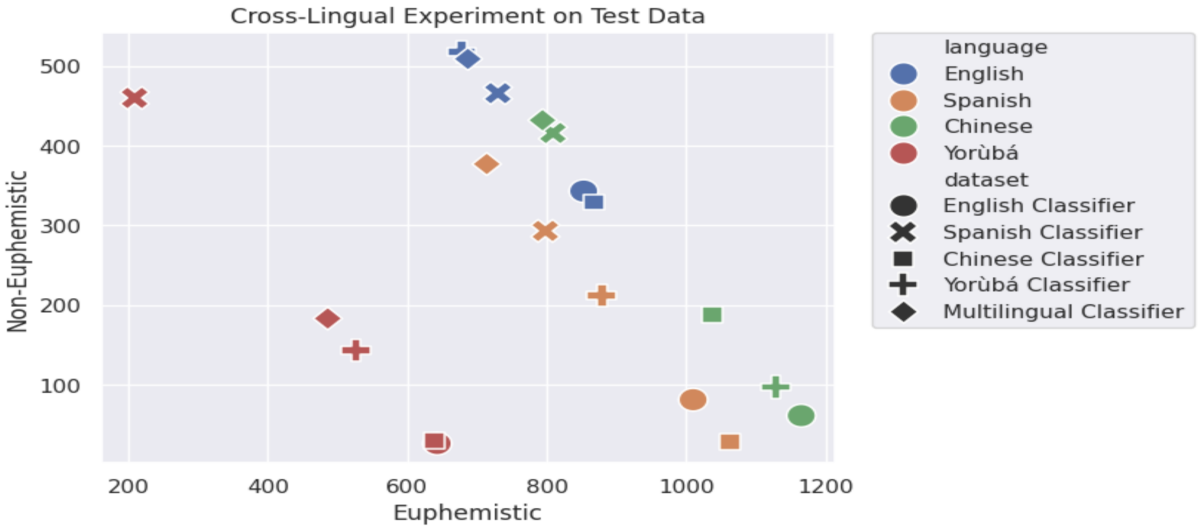


Figure 1: This chart portrays the inference results of the cross-lingual split between which sequences were labeled as euphemistic, and which were labeled as non-euphemistic. The individual language classifiers (fine-tuned only on their respective language data) are included along with the multilingual classifier to show contrast. Gold Standard labels are unknown and were not available to include in this Figure. Values can be found in Figures 2-5.

dialects play a role in greater euphemistic language understanding. The overall implications can suggest which kinds of base models could be optimal for assessing complicated linguistic devices in downstream language tasks, as well as how semantic correlation impacts deep learning throughout different languages.

6 Limitations

Please note that this paper does not account for varying dialects of all the presented languages. The only dialect of Chinese in the data is Mandarin Chinese, and the only dialect of English is American English. The Spanish and Yorùbá language data sets do, however, contain examples from different dialects. The selection of DistilBERT was partially due to the limited computatational resources of the author.

7 Ethics Statement

The author does not foresee any ethical concerns with the findings presented in this paper.

8 Acknowledgments

Special thanks goes to the Montclair State University NLP Lab for their work in organizing this shared task, answering questions concerning the limitations of this paper, as well as assembling the training and test data sets.

References

Adelani, David (2021). *Davlan/xlm-roberta-base-finetuned-yoruba · Hugging Face — huggingface.co*. https://huggingface.co/Davlan/xlm-roberta-base-finetuned-yoruba. [Accessed 03-03-2024].

Conneau, Alexis et al. (2019). "Unsupervised Cross-lingual Representation Learning at Scale". In: *CoRR* abs/1911.02116. arXiv: 1911.02116. URL: http://arxiv.org/abs/1911.02116.

Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer (2023). *QLoRA: Efficient Finetuning of Quantized LLMs*. arXiv: 2305.14314 [cs.LG].

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805. arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

Gavidia, Martha, Patrick Lee, Anna Feldman, and Jing Peng (2022). *CATs are Fuzzy PETs: A Corpus and Analysis of Potentially Euphemistic Terms*. arXiv: 2205.02728 [cs.CL].

Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv: 2106.09685 [cs.CL].

Lee, Patrick, Alain Chirino Trujillo, Diana Cuevas Plancarte, Olumide Ebenezer Ojo, Xinyi Liu, Iyanuoluwa Shode, Yuan Zhao, Jing Peng, and Anna Feldman (2024). *MEDs for PETs: Multilingual Euphemism Disambiguation for Potentially Euphemistic Terms*. arXiv: 2401.14526 [cs.CL].

Plank, Barbara (2017). *ALL-IN-1: Short Text Classification with One Model for All Languages*. arXiv: 1710.09589 [cs.CL].

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *ArXiv* abs/1910.01108.

Wolf, Thomas et al. (2020). *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. arXiv: 1910.03771 [cs.CL].

YouTube-Blog (2024). https://github.com/ShawhinT/YouTube-Blog.

9 Appendix

A Cross-Lingual Experiment Data

Figures 2-5 are the predicted label splits from the Cross-Lingual Experiment.

English Classifier

	En.	Sp.	Ch.	Yo.
Euph.	853	1010	1165	643
Non-Euph.	343	81	61	26

Figure 2: English Classifier with Cross-Lingual Experiment

Spanish Classifier

	En.	Sp.	Ch.	Yo.
Euph.	730	798	809	209
Non-Euph.	466	293	416	460

Figure 3: Spanish Classifier with Cross-Lingual Experiment

Chinese Classifier

	En.	Sp.	Ch.	Yo.
Euph.	867	1063	1038	639
Non-Euph.	329	28	188	30

Figure 4: Chinese Classifier with Cross-Lingual Experiment

Yorùbá Classifier

	En.	Sp.	Ch.	Yo.
Euph.	678	879	1129	526
Non-Euph.	518	212	97	143

Figure 5: Yorùbá Classifier with Cross-Lingual Experiment

B Full Results from Shared Tasks

These are the detailed results output after the first and second submissions to the shared task. They include the individual classifier results, and the multilingual classifier results. As mentioned, the F1 scores were most important for this task, yet the precision and recall were included for transparency.

	En.	Sp.	Ch.	Yo.
F1	0.5736	0.5997	0.5995	0.6091
Precis.	0.6410	0.5986	0.7076	0.6537
Recall	0.6184	0.6011	0.6130	0.6104

Table 5: Detailed Results of Individual Classifiers

	En.	Sp.	Ch.	Yo.
F1	0.6446	0.6054	0.6808	0.6500
Precis.	0.6601	0.6024	0.6861	0.6716
Recall	0.6607	0.6209	0.6780	0.6457

Table 6: Detailed Results of Multilingual Classifier

Comparison of Image Generation Models for Abstract and Concrete Event Descriptions

Mohammed Abdul Khaliq[1] and **Diego Frassinelli**[2,3] and **Sabine Schulte im Walde**[1]
[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]Department of Linguistics, University of Konstanz, Germany
[3]Center for Information and Language Processing, LMU Munich, Germany
{mohammed.abdul-khaliq,schulte}@ims.uni-stuttgart.de,
diego.frassinelli@uni-konstanz.de

Abstract

With the advent of diffusion-based image generation models such as DALL-E, Midjourney and Stable Diffusion, high-quality images can be easily generated using textual inputs. It is unclear, however, to what extent the generated images resemble human mental representations, especially regarding abstract event knowledge, in contrast to concrete event knowledge. We analyse the capabilities of four state-of-the-art models in generating images of verb-object event pairs when we systematically manipulate the degrees of abstractness of both the verbs and the object nouns. Human judgements assess the generated images and indicate that DALL-E is strongest for event pairs with concrete nouns (e.g., *pour water*; *believe person*), while Midjourney is preferred for event pairs with abstract nouns (e.g., *remain mystery*; *raise awareness*), in both cases irrespective of the concreteness of the verb. Across models, humans were most unsatisfied with images of events pairs that combined concrete verbs with abstract direct-object nouns (e.g., *speak truth*; *steal idea*). We hypothesised that this is due to the tendency of these combinations to express figurative language, which was confirmed by post-hoc collected human judgements.

1 Introduction

Nowadays tools for automatic image generation are accessible to laypeople as much as to experts. But do the generated images capture human mental representations? And which images are generated for abstract concepts and events that are not easily depictable, such as the concept *patience* and the event *speak the truth*, given that what we really **see** in the images depicting abstract knowledge are concrete objects?

The current study assesses four image generation models on how well they depict abstract vs. concrete event descriptions: we compare DALL-E 2 (Ramesh et al., 2022), Stable Diffusion (Rombach et al., 2022), Stable Diffusion XL (Podell et al., 2023) and Midjourney[1], as well as images retrieved by the search engine Bing[2]. Following Frassinelli and Schulte im Walde (2019), the prompts for the models are represented by 40 phrase-level events consisting of a verb and a direct object noun, where we systematically vary the words' degrees of abstractness by relying on the ratings in Brysbaert et al. (2014), cf. *build a perspective* vs. *carry a box*. We evaluate the generated images through human ratings (i) in a standard large-scale crowd-sourcing task, and (ii) in a two-step small-scale setup where we prime our participants on their expectations by asking them to first describe what they would expect to see in an image of a specific event, before asking them to judge the quality of the automatically generated images. Our hypothesis is that humans will be less satisfied with the depiction of abstract in comparison to concrete event knowledge, while it is unclear how and to what extent the abstractness of verbs vs. nouns influences the human judgements with regard to the four-way combinations of abstract/concrete verb-noun events.

We thus propose an exploration of the capabilities of image generation models regarding abstract vs. concrete event descriptions, while previous work primarily focused on concrete events such as scenes with concrete objects and relations (Johnson et al., 2018), person appearance and shape (Tang et al., 2020), and transformer-based text-to-image generation across different styles (Ding et al., 2021), or on investigating prompts variants for optimising the generation of abstract and figurative concepts (Chakrabarty et al., 2023; Liao et al., 2023). Examples of research that not only targeted concrete but also abstract knowledge in images, are studies by McRae et al. (2018) who performed priming experiments for abstract words in images,

[1]https://www.midjourney.com
[2]https://www.bing.com/

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 15–21
June 21, 2024 ©2024 Association for Computational Linguistics

Akula et al. (2023) who proposed standard vision detection and retrieval tasks to distinguish between concrete and abstract concepts in visual metaphors, and Shahmohammadi et al. (2023) who trained image generation models to illustrate any kind of textual input, including figurative language.

2 Target and Data Collections

As the basis for our experiments we create verb-noun event pairs of varying degrees of concreteness (Section 2.1). These event pairs are used as prompts for the image generation models (Section 2.2).

Verb	Score	Noun	Score	Category V + N
eat	4.44	meal	4.66	C + C
know	1.68	man	4.79	A + C
raise	3.80	awareness	1.84	C + A
assume	1.75	responsibility	1.40	A + A

Table 1: Examples of verb-noun event pairs, together with the individual verb/noun mean concreteness rating scores from Brysbaert et al. (2014) on a scale from 1 (abstract) to 5 (concrete), and the event category type.

2.1 Verb-Noun Event Pairs

We rely on the concreteness ratings by Brysbaert et al. (2014) to systematically create a total of 40 pairs combining 10 strongly concrete verbs and strongly concrete nouns (`ConcV+ConcN`), 10 strongly abstract verbs and strongly concrete nouns (`AbstV+ConcN`), 10 strongly concrete verbs and strongly abstract nouns (`ConcV+AbstN`), and 10 strongly abstract verbs and strongly abstract nouns (`AbstV+AbstN`). Table 1 presents one example per verb-noun event category and the corresponding individual word concreteness ratings. The full table is provided in Appendix A.

2.2 Image Generation

We employ four image generation models. In addition to these models we also use Bing images.

DALL-E 2 is a text-to-image image generation model from OpenAI released in April, 2022. DALL-E 2 can be accessed through the OpenAI's API at a fixed cost per image basis. It is able to create an image in 1:1 aspect ratio with a maximum resolution of 1024x1024, which is what we use.

Midjourney (MJ) v5.1 is a text-to-image model developed by Midjourney Inc. Unlike the other models, Midjourney is not accessible through an API, and it requires manual prompting in a Discord interface. It also has a fixed subscription-based payment to generate images. Midjourney v5.1 generates images at 1024x1024 resolution which can be altered for different aspect ratios. We use the default 1024x1024 resolution of v5.1.

Stable Diffusion (SD) v2.1 is a text-to-image model developed by Stability AI which makes use of the latent diffusion model architecture to generate images. It is open-source and can be run locally or accessed via API through DreamStudio[3]. It is able to create images of varying aspect ratios and resolutions at the cost of degrading quality the further you go away from the 768x768 native resolution. We use the 768x768 resolution for all our generations setting the inference steps to 75.

Stable Diffusion XL (SDXL) v1.0 is the latest Stable Diffusion model from Stability AI. It improves over Stable Diffusion v2.1 by requiring shorter and less detailed prompts and being able to generate text within the images. Additionally, its three times larger UNet Backbone (used for image segmentation) and architectural improvements enable it to create more prompt-consistent and high-quality images with a native resolution of 1024x1024. It is open-source and can be run locally or accessed via API through DreamStudio. We set the resolution to 1024x1024 with the number of inference steps set to 50 (default).

Bing is a search engine that we use for image search as an upper bound to evaluate the image generation models. We feed our prompts via the Bing API to retrieve images that are not restricted by resolution or aspect ratio.

For all four models as well as Bing, we use as prompts the verb-noun event pairs introduced above. The image generation models were prompted using their default parameters. We collect four images from each of the four models' outputs as well as from Bing, for each of the 40 verb-noun pairs, a total of 800 images. Figure 1 presents one example image for each model and for two event pairs, *serve food* (`ConcV+ConcN`) and *remain mystery* (`AbstV+AbstN`).

3 Model Evaluation

We evaluate the generated images through human ratings in two studies. The images, the full annotation instructions and all collections are publicly available at `https://www.ims.uni-stuttgart.de/data/image-generation`.

[3] `https://dreamstudio.ai/generate`

(a) DALL-E 2

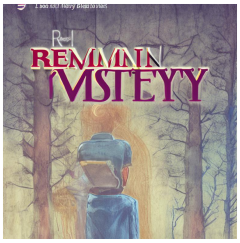
(b) Midjourney v5.1

(c) Stable Diffusion v2.1

(d) Stable Diffusion XL v1.0

(e) DALL-E 2

(f) Midjourney v5.1



(g) Stable Diffusion v2.1

(h) Stable Diffusion XL v1.0

Figure 1: Example images for the event pairs *serve food* and *remain mystery*, as generated by the four models.

serve food	**reduce noise**	**steal idea**	**remain mystery**
a waiter bringing a platter filled with food to a table at a dimly lit diner, three people sitting at the table; a man stands behind a counter and dishes up a variety of foods to a customer	a slider with a speaker symbol next to it and an arrow over the slider pointing away from the speaker symbol; a grainy picture followed by an arrow and a very soft looking version of the picture	a person in a lab coat leafing through a notebook, the body language shows unease; person with thought bubble above their head, the thought bubble is being snatched away by another person	a woman burns a letter from an ex without reading it; an archaeologist tries to decipher a text from an unknown language

Table 2: Examples of human descriptions for four verb-noun event pairs in Task 1 of the expectation-based study.

Study 1: Crowdsourcing Ratings We gather ratings of the generated images for our verb-noun events from Amazon Mechanical Turk (AMT)[4] workers based in either USA or UK, and with more than 10,000 prior submissions and a ≥99% approval rate. The workers are asked to rate on a scale from 1 to 6 how well each of the 800 generated images depicts the associated verb-noun pair event. We also add 80 images as sanity checks; these include an obviously wrong image for additional verb-noun pairs, e.g., an image of a car for *play football*.

Study 2: Expectation-based Ratings This evaluation is conducted in two consecutive tasks.

In Task 1 we aim at collecting precise descriptions of what our participants expect to see in an image of a particular verb-noun event, by asking them to provide one or more phrases describing the mental image they created of the given event. In this way, participants can reflect on the given event and the mental representations they are generating.

In Task 2, the same participants are presented with the same verb-noun pairs, their own descriptions for the pairs, and four images from each of the four models and Bing. They are asked to select all images that depict the event well, without providing any ranking. The annotators can also select images that do not directly match their own descriptions, as long as they judge the image good.

The annotators are university students highly proficient in English (B2 level or higher). We collect 19 responses from our annotators describing their image expectations for the verb-noun event in Task 1 (see examples in Table 2). 12 out of the 19 annotators also completed Task 2.

4 Results

Study 1: Crowdsourcing Ratings We collected a total of 7,200 ratings for our 800 images, with nine unique annotators rating each image on a scale from 1 to 6. After removing all ratings by annotators that failed the sanity check, and using only those images that received ≥4 approved ratings, our final set contains 4,212 ratings.

[4] https://www.mturk.com/

These ratings distribute over our event categories as follows.

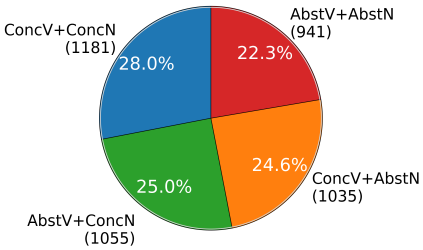


Figure 2: Final set of ratings across categories.

Figure 3 presents the proportions of how often a model received an extremely low (bad) rating of 1 or 2 (left plot) or an extremely high (good) rating of 5 or 6 (right plot), out of the total number of ratings for that model and a specific verb-noun event category. For example, SD received a very low rating for 62 (48%) of the generated images in the `AbstV+AbstN` category and a very high rating for only 13 (10%) generated images in this category.

Overall, we can clearly see that SD (orange bars) received most low ratings and fewest high ratings across event categories; Bing (blue bars) serves as an upper bound (i.e., receiving few low and many high ratings across most event categories); and DALL-E, SDXL and MJ show more variable results across event categories. More specifically, the right plot in Figure 3 displays closer competitions across the image generation models: Our best performing model for `AbstV+ConcN` and `ConcV+ConcN` is DALL-E, while MJ is best regarding the other two categories. Therefore, DALL-E performs best when the direct-object noun is concrete, while MJ performs best when the direct-object noun is abstract, irrespective of the concreteness of the verb. MJ also exhibits a rather uniform success rate across categories. SDXL (green bars) is the second best generation model in three out of four categories.

Study 2: Expectation-based Ratings Figure 4 shows how many images from each model were selected by the annotators across verb-noun categories in Task 2, after they had previously described their expectations (see examples in Table 2). Similar to our large-scale experiment, we notice the consistently poor performance of SD, while DALL-E, SDXL and MJ are more favoured, and Bing serves as the upper bound. The plot confirms that DALL-E performs best when the direct-object noun is concrete, while MJ performs best when the direct-object noun is abstract. Finally, the annotators were much less satisfied across models with images for the `ConcV+AbstN` event category than with images for any of the other event categories.

Table 3 once more confirms the general trends by showing the total number of images for each model that were selected in Task 2. Again we notice that MJ, DALL-E and also SDXL are more favoured than SD, and that Bing serves as the upper bound. Table 3 also shows the mean and standard deviation scores across our four event categories, pointing out that especially DALL-E varies strongly.

	#selected	*mean*	*stdev*
Bing	760	190.00	47.10
DALL-E	513	128.25	60.75
MJ	530	132.50	33.27
SD	181	45.25	19.76
SDXL	429	107.25	35.91

Table 3: Overall selected images per model/Bing.

Overall, our human expectations evaluation confirms the general trends from the crowdsourced evaluation regarding (dis)preferences that annotators perceived when judging the generated images. In fact, Figure 4 presents a similar yet sharper picture of the human evaluation preferences in comparison to Figure 3.

Abstract Events and Figurative Language Our initial hypothesis was that humans would be less satisfied with the depiction of abstract in comparison to concrete event knowledge. Looking into our best model results, this hypothesis has been confirmed but in an unexpected way. We found that DALL-E performs best when the direct-object noun is concrete (however with a rather large standard deviation), while MJ performs best when the direct-object noun is abstract, irrespective of the concreteness of the verb. In particular, annotators were much less satisfied across models with images for the `ConcV+AbstN` event category. So overall it seems as if the abstractness of the noun plays a core role in how well the generated images depict verb-noun events.

We suspected that this is the case because `ConcV+AbstN` events predominantly express figurative language usage, as suggested by Frassinelli and Schulte im Walde (2019), which is inherently difficult to depict. In order to look into this follow-up hypothesis, we ran an additional annotation study by asking 12 annotators for their binary judgements

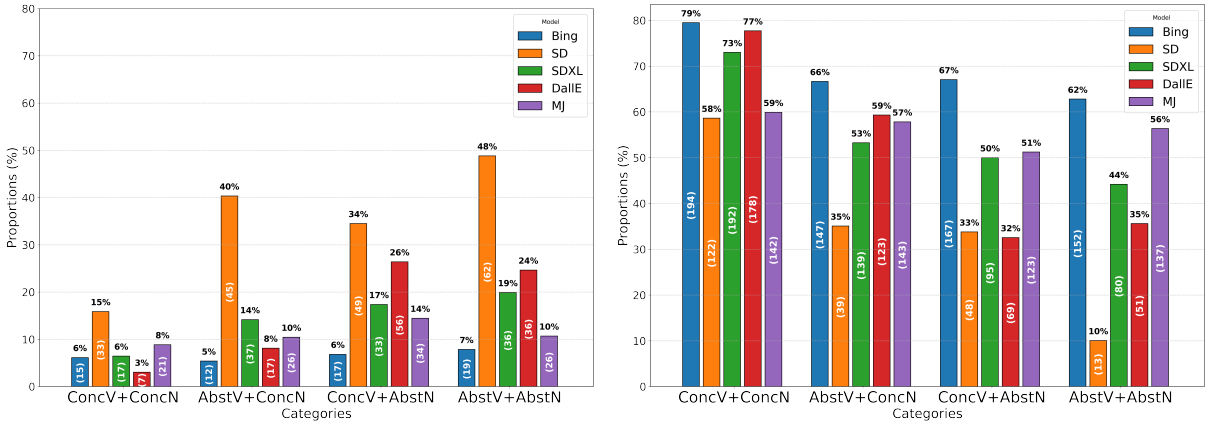


Figure 3: Proportions of how often the four models or Bing received an extremely low rating (1 or 2, *left plot*) or an extremely high rating (5 or 6, *right plot*) in the crowdsourcing evaluation, out of the total number of ratings for that model and a specific event category.

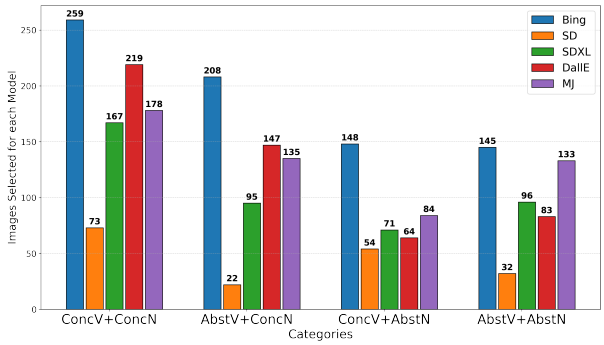


Figure 4: Number of images selected for each model in Task 2 of the human expectations setup, i.e., where the annotators judged the images as well-depicting the respective events.

on figurative vs. literal language of our 40 event pairs.[5] Figure 5 shows that indeed `ConcV+AbstN` (and to a lesser degree also the most abstract combination `AbstV+AbstN`) are strongly perceived as figurative language.

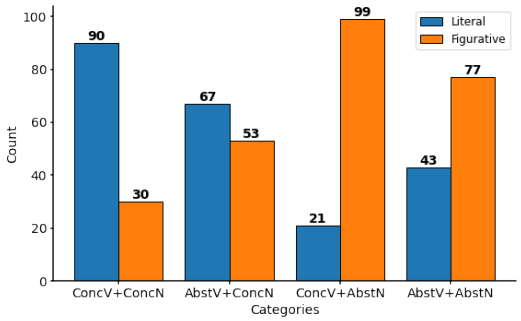


Figure 5: Number of literal vs. figurative language judgements of event pairs across event categories.

[5]We also asked the annotators to provide examples sentences, so that we could check that they understood the task, and to obtain textual event information. The data are publicly available from the same URL as above.

5 Conclusion

This paper systematically assessed image generation models on their capacity to generate images for abstract vs. concrete event descriptions. We demonstrated through human evaluations that DALL-E is strongest for event pairs with concrete nouns, while MJ is strongest for event pairs with abstract nouns. Regarding images for events with a concrete verb and an abstract direct-object noun, humans were generally not satisfied with any model, which an additional annotation attributed to a strong tendency for representing figurative language.

We cannot conclusively say why some models perform better than others, but we suspect that this is due to reasons such as MJ's tendency to produce more creative images in contrast to DALL-E producing simplistic and to-the-point images (which humans seem to like for concrete nouns). Overall, all models were outperformed by Bing images, which we attribute to less artifacting, randomness and consistency issues in those images.

Acknowledgements

This research was supported by the DFG Research Grant SCHU 2580/4-1 (*MUDCAT – Multimodal Dimensions and Computational Applications of Abstractness*). We also thank the reviewers and the SemRel group for useful feedback and suggestions, and our friends and colleagues who supported us in our collections of image expectations and expectation-based ratings.

Ethics Statement

We are aware that image generation models – as all models trained on some selection of natural language data – are likely to capture biases regarding societal inequalities. With respect to our experiments involving human participants, we do not see any ethical issues related to this work: All collections were conducted on a voluntary basis with a fair compensation (12 Euros per hour), and we kept the data collection anonymous.

References

Arjun R. Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T. Freeman, Yuanzhen Li, and Varun Jampani. 2023. MetaCLUE: Towards comprehensive visual metaphors research. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 23201–23211, Vancouver, Canada.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. *arXiv preprint arXiv:2305.14724*.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. 2021. CogView: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835.

Diego Frassinelli and Sabine Schulte im Walde. 2019. Distributional interaction of concreteness and abstractness in verb–noun subcategorisation. In *Proceedings of the 13th International Conference on Computational Semantics*, pages 38–43, Gothenburg, Sweden.

Justin Johnson, Agrim Gupta1, and Li Fei-Fei. 2018. Image generation from scene graphs. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1219–1228, Salt Lake City, Utah.

Jiayi Liao, Xu Chen, Qiang Fu, Lun Du, Xiangnan He, Xiang Wang, Shi Han, and Dongmei Zhang. 2023. Text-to-image generation for abstract concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 38(4), 3360–3368, Vancouver, Canada.

Ken McRae, Daniel Nedjadrasul, Raymond Pau, Bethany Pui-Hei Lo, and Lisa King. 2018. Abstract concepts and pictures of real-world situations activate one another. *Topics in Cognitive Science*, 10:518–532.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Anirudh Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Hassan Shahmohammadi, Adhiraj Ghosh, and Hendrik Lensch. 2023. ViPE: Visualise pretty-much everything. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5477–5494, Singapore.

Hao Tang, Song Bai, Li Zhang, Philip H. S. Torr, and Nicu Sebe. 2020. XingGAN for person image generation. In *Proceedings of the European Conference on Computer Vision*, pages 717–734. Springer.

A All 40 Verb-Noun Pairs and their Event Categories

Verb	Score	Noun	Score	Category V + N
eat	4.44	meal	4.66	
write	4.22	song	4.66	
pour	4.14	water	5.00	
throw	4.04	money	4.54	
carry	4.04	weight	3.94	ConcV+ConcN
raise	3.80	family	4.23	
serve	3.78	food	4.80	
build	3.71	company	4.11	
hold	3.68	pillow	5.00	
read	3.56	paper	4.93	
put	2.50	weight	3.94	
keep	2.37	money	4.54	
investigate	2.27	case	3.93	
generate	2.23	electricity	3.90	
sustain	2.17	injury	4.00	AbstV+ConcN
educate	2.12	child	4.78	
reduce	2.00	noise	3.52	
develop	1.87	company	4.11	
know	1.68	man	4.79	
believe	1.55	person	4.72	
pave	4.03	way	2.34	
seize	3.97	moment	1.61	
steal	3.84	identity	2.00	
steal	3.84	idea	1.61	
raise	3.80	awareness	1.84	ConcV+AbstN
raise	3.80	expectation	1.62	
build	3.71	perspective	2.38	
speak	3.70	truth	1.96	
hold	3.68	responsibility	1.40	
unfold	3.55	drama	2.34	
understand	2.28	reason	1.93	
understand	2.28	meaning	1.85	
learn	2.20	language	2.35	
reduce	2.00	loss	2.19	
remain	1.96	mystery	2.33	AbstV+AbstN
develop	1.87	idea	1.61	
improve	1.82	safety	2.37	
improve	1.82	health	2.28	
fulfill	1.78	obligation	2.04	
assume	1.75	responsibility	1.40	

Table 4: All our 40 verb-noun event pairs, together with the individual verb/noun mean concreteness rating scores from Brysbaert et al. (2014) on a scale from 1 (abstract) to 5 (concrete), and the event category type.

Cross-Lingual Metaphor Detection for Low- to High-Resource Languages

Anna Hülsing
University of Hildesheim
anna.huelsing@uni-hildesheim.de

Sabine Schulte im Walde
University of Stuttgart
schulte@ims.uni-stuttgart.de

Abstract

Research on metaphor detection (MD) in a multilingual setup has recently gained momentum. As for many tasks, it is however unclear how the amount of data used to pretrain large language models affects the performance, and whether non-neural models might provide a reasonable alternative, especially for MD in low-resource languages. This paper compares neural and non-neural cross-lingual models for English as the source language and Russian, German and Latin as target languages. In a series of experiments we show that the neural cross-lingual adapter architecture MAD-X performs best across target languages. Zero-shot classification with mBERT achieves decent results above the majority baseline, while few-shot classification with mBERT heavily depends on shot-selection, which is inconvenient in a cross-lingual setup where no validation data for the target language exists. The non-neural model, a random forest classifier with conceptual features, is outperformed by the neural models. Overall, we recommend MAD-X for metaphor detection not only in high-resource but also in low-resource scenarios regarding the amounts of pretraining data for mBERT.

1 Introduction

Song titles such as *Life is a Highway* are prominent examples of how we use metaphors in our everyday life. But songs are by far not their only habitats: on average and across domains, metaphors can be found in every third sentence (Shutova and Teufel, 2010). Lakoff and Johnson (1980) define a conceptual metaphor as "understanding one conceptual domain [A] in terms of another conceptual domain [B]" (Kövecses, 2010). In the above example, the domain *Life* (A) is understood in terms of the domain *Journey* (B). Detecting whether or not a word or expression is a metaphorical linguistic expression (i.e. whether or not it is used metaphorically) is vital for many NLP applications, such as sentiment analysis, machine translation, information extraction, and dialog systems, cf. Tsvetkov et al. (2014). Metaphor detection (MD) can further support automatic essay scoring (Beigman Klebanov et al., 2018), schizophrenia detection (Gutiérrez et al., 2017), and propaganda identification (Baleato Rodríguez et al., 2023).

Many efforts have been made to tackle the task of metaphor detection (MD),[1] and successfully so: close-to-human performance was reached by systems using large pretrained language models like BERT (Devlin et al., 2019) for English datasets containing single sentences with a metaphorical expression (Ma et al., 2021). For a long time, Tsvetkov et al. (2014) were the only ones to perform MD cross-lingually, namely for Spanish, Russian and Farsi. Only recently, Aghazadeh et al. (2022) and Lai et al. (2023) addressed metaphor detection in a multilingual setup with the same languages as Tsvetkov et al. (2014). Whereas Aghazadeh et al. (2022) focused on probing metaphoricity within the transformer layers, Lai et al. (2023) used a template-based prompt learning approach to MD. These multilingual MD approaches focus on languages where large amounts of data are available for pretraining. Insights are missing, however, on whether or not large language models are also suitable for MD in languages with small amounts of pretraining data.

The current study addresses this bottleneck and compares neural and non-neural cross-lingual models for detecting metaphors in languages with varying degrees of pretraining data, including the low-resource language Latin.

Our metaphor detection focuses on word-based classification, as in the following example from the metaphor dataset by Tsvetkov et al. (2014):

(1) Actions talk even louder than phrases.

[1] See Shutova (2015), and Tong et al. (2021) for two prominent surveys.

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 22–34
June 21, 2024 ©2024 Association for Computational Linguistics

Language	# Wikipedia articles
English (source)	$\approx 6.7m$
German	$\approx 2.8m$
Russian	$\approx 1.9m$
Latin	$\approx 0.1m$

Table 1: Amount of articles in millions (m) regarding the four languages used in the current study. The numbers are taken from https://meta.wikimedia.org/wiki/List_of_Wikipedias, accessed 25 Sep. 2023. Altogether, mBERT was pretrained on Wikipedia articles from 104 languages.

We define a binary classification task to detect whether or not the underlined target word is used metaphorically in the given context. For zero- and few-shot classification we apply multilingual BERT (mBERT) (Devlin et al., 2019) and the adaptation method MAD-X (Pfeiffer et al., 2020b), which have shown state-of-the-art results for e.g. named entity recognition and question answering. As our non-neural model, we apply a random forest classifier (Breiman, 2001), as random forest classifiers generally perform well in low-resource scenarios (Tsvetkov et al., 2014). Our model utilizes a vector space model and conceptual features (abstractness and supersenses) – similarly to the model introduced by Tsvetkov et al. (2014).

As for target languages, we investigate modelling performances for German, Russian and Latin, because the amount of data used to pretrain mBERT varies greatly across these three languages (see Table 1). Whereas German and Russian are not considered low-resource languages in terms of pretraining data, we simulate low-resource conditions and explore the influence of different amounts of pretraining data by using as little as 20 instances or no labelled data at all from the target languages for training, and no data at all for validation. Latin, on the other hand, is a low-resource language in terms of pretraining data and in terms of labelled training data. English as a high-resource language was used as the source language for cross-lingual transfer.

Contributions. The main contribution of this paper is a comparison and a series of insights regarding cross-lingual neural and non-neural models for MD in languages with high-to-low degrees of pretraining data, i.e., German, Russian and Latin. More specifically, **1)** we find that with default hyperparameters, zero-shot mBERT performs best: results are above a majority vote baseline for all three target languages. **2)** MAD-X performs best when hyperparameter-tuning is carried out or large amounts of source language training data are used. **3)** We show that few-shot mBERT depends largely on shot-selection, which cannot be carried out in a low-resource environment where no validation data exists. **4)** Overall, the non-neural model is outperformed by the neural classifiers, and we recommend using MAD-X with suitable hyperparameters for MD in languages with both large and little amounts of data used for pretraining mBERT.[2]

2 Related Work

Metaphor Detection. Turney et al. (2011) were among the first to apply insights from cognitive linguistics to their MD model, i.e., exploiting that metaphors transfer knowledge from a concrete domain to an abstract domain (Lakoff and Johnson, 1980). Since metaphoricity is correlated with the degree of contextual abstractness, the authors used abstractness scores of context words as features in a logistic regression model.

The idea of "conceptual features" also inspired Tsvetkov et al. (2014), who used abstractness scores, imageability scores and semantic supersenses as classification features. Whereas Turney et al. (2011) focused on English data only, Tsvetkov et al. (2014) trained on English data and then evaluated the model cross-lingually on Spanish, Farsi and Russian. Their model represents the basis for the random forest classifier used in our experiments. Köper and Schulte im Walde (2016) focused on MD for German particle verbs. They also used 1) abstractness and imageability ratings as well as 2) scores indicating the distributional fit of particle verbs with regard to base verb contexts. In addition, they used 3) unigram context words and 4) noun clusters as features.

Do Dinh and Gurevych (2016) were the first to use a neural model architecture for MD, namely a multilayer perceptron with word embeddings. Their approach performed comparable to existing models without requiring feature engineering. Dankers et al. (2019) explored the relationship between metaphors and emotions by building several multi-task learning models. The best performing architecture made use of BERT embeddings used as input to a multilayer perceptron or to additional attention layers. They reached state-of-the-art results in 2019 for both metaphor and VAD prediction.

[2]The code can be accessed here: https://github.com/AnHu2410/MD_crosslingual.

Su et al. (2020) transformed word-based metaphor detection into a reading comprehension problem; their approach, DeepMet, was the most successful model in the 2020 metaphor detection shared task (Leong et al., 2020). Ma et al. (2021) fine-tuned BERT for MD. To perform word-based binary metaphor classification, they copied the input sentence and masked the target word. The original sentence and the masked copy were used as input for a sequence classification task. The BERT model then predicted whether the two sentences appeared in the same context; if yes, they predicted a literal usage of the masked word; otherwise they predicted a metaphorical usage. They also performed sentence-level classification and sequential labelling of metaphorical expressions. Their results showed an increase over previous state-of-the-art models. We use their word-based classification approach for the mBERT-based classifiers in our experiments. While their focus was on English, we use it in a multilingual setup.

Li et al. (2023) exploited the fact that many datasets are based on the Metaphor Identification Process (MIP; Pragglejaz Group, 2007), where a word is annotated as metaphorical if its contextual meaning is dissimilar to its "more basic meaning" (among further criteria). While prior models (such as MelBERT by Choi et al. 2021) grounded on MIP use decontextualized representations of the target word, Li et al. (2023) successfully gathered the representation of the target word from sentences where it was used literally.

Cross-Lingual Representations. Vulić and Moens (2013) proposed a bootstrapping method to create bilingual vector spaces from non-parallel data. Usually, a high-dimensional vector in a feature vector space uses context features as dimensions. For the proposed bilingual vector space, these features consisted of translation pairs. This method can be applied to any language pair.

Multilingual BERT (Devlin et al., 2019) was pretrained on data from 104 languages. Lauscher et al. (2020) pointed out limitations of large multilingual pretrained language models by demonstrating that these models do not transfer knowledge well for low-resource target languages (i.e. languages with small pretraining corpora) and for distant language pairs. They showed that first fine-tuning on large amounts of data and then continuing fine-tuning with very few examples from the target language considerably improves results across all languages and tasks. The current paper investigates whether these findings also apply to MD. Pfeiffer et al. (2020b) tried to mitigate problems of multilingual language models targeting low-resource languages by using an adaptation method, i.e. by inserting small amounts of trainable weights into an existing pretrained model (see Section 4). We also apply these Multiple ADapters for Cross-lingual transfer (MAD-X) to MD in our experiments.

3 Datasets and Preprocessing

Source Language. We used the dataset from Tsvetkov et al. (2014) as our basic English training dataset. It is based on the TenTen[3] Web Corpus, contains 222 instances, and is balanced. This basic training dataset was previously used by Tsvetkov et al. (2014) for evaluation. In the course of our experiments we augmented the amount of training data by adding the imbalanced dataset by Mohammad et al. (2016), which consists of 1,639 instances. The augmented version comprises 1,861 instances.[4]

Target Languages. Tsvetkov et al. (2014) also provide the Russian dataset that we used for evaluation, which is balanced, consists of 240 instances, and is based on the TenTen Web Corpus. For evaluation in German, we used the MD dataset provided by Köper and Schulte im Walde (2016), which is based on the web corpus DECOW14AX (Schäfer and Bildhauer, 2012) and where the target words are particle verbs. To balance the dataset, we reduced the original dataset from Köper and Schulte im Walde (2016) to 896 metaphorical and 896 literal instances.

For our Latin dataset we used the Lexham Figurative Language of the New Testament Dataset (Westbury et al., 2016), which is published in the Logos[5] Bible Software. It shows passages from the New Testament (we used the American Standard Version of the Bible), and highlights the metaphors in each verse. We extracted 100 sentences, of which 50 were annotated as metaphorical and 50 were annotated as literal. As the metaphors were annotated in the English Bible text, we then manually searched for the Latin translations in the Vulgate[6].

[3] https://www.sketchengine.eu/

[4] For the random forest classifier, only a subset of the dataset by Mohammad et al. (2016) was used for augmentation, because lemmatized subjects, verbs and objects had to be annotated, but this annotation was available only for 100 instances.

[5] https://www.logos.com

[6] https://vulgata.info/index.php?title=Kategorie:BIBLIA_SACRA.

The first author of this paper, a classical philologist, ensured that the metaphors found in the English texts correspondingly occurred in the Latin texts, i.e. that the American Standard Version did not introduce metaphors that were not present in the Vulgate.[7]
Below we provide two example sentences for each dataset, together with the respective categorization into metaphorical vs. literal.

- English (Tsvetkov et al., 2014, source):

  (2) The twentieth century saw intensive development of new technologies. → *metaphorical*

  (3) The young man shook his head. → *literal*

- English (Mohammad et al., 2016, source):

  (4) This young man knows how to climb the social ladder. → *metaphorical*

  (5) Did you ever climb up the hill behind your house? → *literal*

- Russian (Tsvetkov et al., 2014, target):

  (6) Бедность давит на людей.[8] (translation: "Poverty weighs on people.") → *metaphorical*

  (7) Повар варит суп на кухне.[9] (translation: "The cook cooks soup in the kitchen.") → *literal*

- German (Köper and Schulte im Walde, 2016, target):

  (8) Dort wird das Wasser aufgestaut und an Nimroz verkauft. (translation: "There, the water is dammed up and sold to Nimroz.") → *literal*

  (9) Über die Zeit hatte sich in ihnen Sehnsucht und Verlangen aufgestaut. (translation: "Over time, longing and desire had dammed up inside them.") → *metaphorical*

- Latin (Westbury et al., 2016, target):

  (10) Et venerunt, et impleverunt ambas naviculas, ita ut pene mergerentur. ("And they came, and filled both the boats, so that they began to sink.") → *literal*

  (11) Et dixerunt ei: Quia heri hora septima reliquit eum febris.("They said therefore unto him, Yesterday at the seventh hour the fever left him.") → *metaphorical*

We preprocessed all datasets such that the original sentence was available, as well as a copy of the original sentence, where we replaced the target word by the [MASK]-token. These two sentences were then further preprocessed by the HuggingFace[10] tokenizer pipeline. In addition, the random forest classifier required the target word (a verb) and its dependent subject and object as lemmas, which we annotated in cases where the information was missing. Figure 1 illustrates an example of input and output across models.

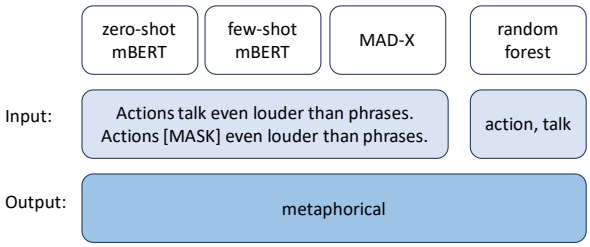


Figure 1: Example input and output of our models.

4 Models

For zero-shot and few-shot classification, we used mBERT (Devlin et al., 2019). For zero-shot classification, we fine-tuned the pretrained language model for MD on the source language data and used this model for predictions in all three target languages. For few-shot classification, we first fine-tuned mBERT on source-language training data, and then fine-tuned it again on a small amount of target language data (see Lauscher et al., 2020). Additionally, we applied MAD-X (Pfeiffer et al., 2020b), which consists of three types of adapters: language adapters, task adapters and invertible adapters. For this method, the pretrained model was frozen and two language adapters were trained on a masked language modelling task: one adapter was trained on unlabelled data from the source language, and one on unlabelled data from

[7] The German and Latin dataset are available here: https://github.com/AnHu2410/MD_crosslingual

[8] Transliteration: *Bednost' davit na lyudey.*

[9] Transliteration: *Povar varit sup na kukhne.*

[10] https://huggingface.co/docs/transformers/main_classes/tokenizer

the target language. Then, the source language adapter was inserted in addition to the task adapter, and the latter was trained on labelled data from the source language. Finally, inference was performed by plugging in the target language adapter and the (language-agnostic) task adapter. The invertible adapters were plugged in simultaneously with the language adapters, but come with a slightly different architecture because they adapt the embeddings, while the language and task adapters were inserted into each transformer layer. For all neural models we utilized the word-based MD method by Ma et al. (2021), where the original sentence and a copy of that sentence with the masked target word were used as input for sequence classification.

As our non-neural model we replicated the random forest classifier by Tsvetkov et al. (2014). This model contains three feature types: 1) abstractness and imageability scores, which Tsvetkov et al. (2014) generated on the basis of the MRC ratings by Wilson (1997), 2) supersenses, i.e., "coarse semantic categories", where a word can belong to several synsets in WordNet (Fellbaum, 1998), each of which is associated with several supersenses. We created a feature vector with these supersenses as dimensions, e.g., the noun "head" occurs in 33 synsets, 3 of which are related to the supersense *noun.body*. The dimension corresponding to the supersense *noun.body* then receives 3/33 (example taken from Tsvetkov et al., 2014). 3) Further features were produced with the vector space model by Faruqui and Dyer (2014). This model utilizes multilingual information in order to generate similar vectors for synonymous words. All these features were extracted from the target word – in our case, a verb – and from its dependent subject and object. For cross-lingual inference, the model relies on one-to-many translations: all translations were given for a target language word, and the scores obtained for the translations were averaged (see Tsvetkov et al., 2014). For translation, we used Word2Word by Choe et al. (2020).

5 Experiments and Results

5.1 Experimental Setup with Basic and Augmented Training Data

We used the basic English dataset by Tsvetkov et al. (2014) for training, and the target language datasets for Russian, German and Latin for evaluation (see Section 3). Then we explored how each of the following cross-lingual classifiers performed on each of the target languages: zero-shot mBERT (mB0); few-shot mBERT with a second fine-tuning on 20 instances of target language data (mB20)[11]; MAD-X; and the random forest classifier (RF).

As hyperparameters for zero- and few-shot mBERT in this basic experimental setup we used the default hyperparameters from Huggingface (Wolf et al., 2020), namely a batch-size of 8, a learning rate of 5e-5, and 3 training epochs. As hyperparameters for MAD-X we used those mentioned by Pfeiffer et al. (2020b): a learning rate of 1e-4, a batch-size of 8 and 100 training epochs. As hyperparameters for the random forest classifier we used those from scikit-learn 1.2 (Pedregosa et al., 2011), namely 100 estimators, no max-depth limit, and Gini as split criterion. We repeated the runs for three different seeds in order to simulate the variance of results achieved on different GPU machines, and report the mean F1-scores as well as the standard deviation (SD). We ran the experiments on an AMD EPYC 7282 16-Core Processor with 32 threads and NVIDIA RTX A6000 GPUs[12]. Our baseline predicts all instances to be metaphorical.

The results for the basic training dataset are presented in the left panel of Table 2. Zero-shot mBERT (mB0) outperformed the baseline for all three languages, while the results for the other three models were all similar or lower (with the exception of mB20 for Russian), and the results for Latin even dropped below the baseline. The random forest classifier produced results lower than mB0.

In order to investigate whether or not the small amount of training data (222 instances) could be responsible for the partly low results, we augmented the basic training data with data from Mohammad et al. (2016) to 1861 training instances, and repeated the experiments. The results are presented in the right panel of Table 2. For mB0, Russian showed slightly higher F1-scores, while the other two languages showed lower F1-scores compared to the basic training dataset. mB20 only achieved a performance comparable to the baseline (except for German). For the random forest classifier the results improved for Russian but remained the same for German and Latin. MAD-X clearly profited from the augmented training data.

[11] The 20 instances are taken from the test datasets for mB20, so here the test datasets are slightly reduced in comparison to the test datasets used for the other experiments.

[12] Training times were for the most part shorter than 10 minutes. The only exception was the training with augmented training dataset for MAD-X with 100 epochs (<30 minutes).

	basic training dataset			augmented training dataset		
	ru	**ge**	**la**	**ru**	**ge**	**la**
baseline	66.7	66.7	66.7	66.7	66.7	**66.7**
mB0	81.1 ±6.9	**77.1** ±1.6	**69.6** ±1.9	82.8 ±14.0	72.5 ±5.2	66.1 ±1.0
mB20	**82.0** ±2.3	67.3 ±1.2	62.1 ±0.0	66.9 ±37.3	70.9 ±3.9	62.2 ±0.8
MAD-X	68.3 ±10.5	64.2 ±10.7	42.0 ±21.3	**87.6** ±2.1	**75.2** ±0.3	63.3 ±3.6
RF	78.6 ±0.7	71.2 ±0.7	66.7 ±1.5	86.2 ±0.7	71.3 ±0.5	66.5 ±0.3

Table 2: Mean F1-scores for verbal MD across three runs with different seeds (± SD) for hyperparameters with the basic and the augmented training dataset and across our target languages Russian (ru), German (ge) and Latin (la).

5.2 Few-Shot Classifier: Shot-Selection

Even though Lauscher et al. (2020) showed that few-shot fine-tuning improves the performance of using zero-shot mBERT, the results obtained in our experiments did not improve with a second round of fine-tuning with 20 target language instances (except for Russian when using the basic training dataset). We therefore investigated shot-selection by selecting five different randomly selected shots instead of one randomly selected shot as in the previous experiments. The results for using default hyperparameters[13] and the basic training dataset are shown in Table 3. While the mean scores are lower than for the best-performing other models, the maximum scores were competitive; SD was rather high across all languages. We manually checked whether the successful shots exhibit specific features in comparison to the non-successful shots, but no pattern could be identified.

	max.	**mean**	**SD**
ru	87.3	76.3	15.1
ge	80.9	75.2	6.0
la	66.7	51.8	29.0

Table 3: Maximum and mean F1-scores as well as SD for using five different shots of the target language datasets for the second fine-tuning of mBERT (default hyperparameters, basic training dataset).

5.3 MAD-X: Hyperparameter-Tuning

As preliminary experiments have shown that MAD-X heavily relies on suitable hyperparameters, as a next step hyperparameter-tuning[14] was carried out. Given that in the cross-lingual setup no validation data for the target language exists, we explored whether using a dataset from the source language English for validation is a valid option. To do so, we performed a grid search, where we fine-tuned the task adapter on the basic English dataset for different hyperparameter sets (see Table 4).

learning rates	**epochs**	**batch size**
1e-3, 1e-4, 1e-5	10, 50, 100	8, 16, 32

Table 4: Hyperparameter values used for the grid search for MAD-X. We ran each combination, with a total of 27 hyperparameter sets.

We then used the English dataset by Mohammad et al. (2016) as our validation dataset, and pretended that the datasets for German, Russian and Latin were also validation datasets. We obtained the F1-scores for each hyperparameter set across all four validation datasets (see Appendix A). We then calculated Spearman's rank-order correlation coefficient ρ between the F1-scores for the English validation dataset and the target-language validation datasets. I.e., we examined whether we find a correlation between the hyperparameter sets that lead to high (low) results for the English validation set and the hyperparameter sets that lead to high (low) results for each of the target-language datasets. If the same sets lead to high (low) F1-scores for English and some target language, then we could infer that fine-tuning the hyperparameters on a source-language dataset is sufficient and no target-language material is necessary for the validation. We however found no strong correlation between English and any target language, see top row in Figure 2.

What we did observe, though, was a strong correlation between the target language datasets, which indicates that a dataset from a language other than the source or target language, i.e. from a third language, can be used for validation. Accordingly, we

[13] We only used one seed (42) to produce the results, because our aim is to show variance across shots, not seeds.

[14] Hyperparameter-tuning was carried out for the task adapter, the language adapter was taken off-the-shelf from AdapterHub, see Pfeiffer et al., 2020a.

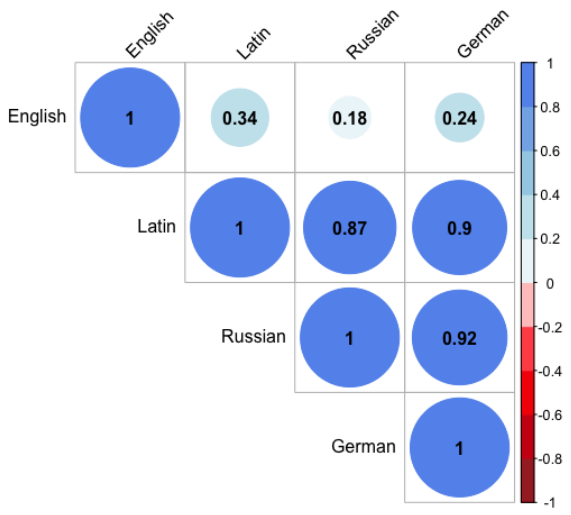


Figure 2: Spearman's rank-order correlations ρ between the hyperparameter sets of the three target languages with regard to the achieved F1-scores for MAD-X.

used the Russian dataset as a validation dataset for the target languages German and Latin (batch-size: 32, learning rate: 1e-3, 50 training epochs), and the German dataset as a validation dataset for Russian (batch-size: 32, learning rate: 1e-3, 100 training epochs).[15] The results are presented in Table 5. Russian shows a result that is comparable to the best results of the other classifiers (except MAD-X with default hyperparameters and the augmented training dataset). The results for German and Latin, in contrast, are the highest across all experiments, and SD is rather low ($<$ 2.5 F1-points).

	ru	ge	la
MAD-X	82.7±2.5	77.3±0.4	73.8±0.9

Table 5: Mean F1-scores (± SD) for using the best performing hyperparameter set from Russian validation data for Latin and German, and the best performing hyperparameter set from German validation data for Russian with MAD-X across three different seeds.

5.4 Summary of Results

MAD-X showed the best performance. For Russian, using default hyperparameters and an augmented training dataset led to the best performance across all models, whereas for German and for Latin hyperparameter-tuning with the basic training dataset led to the best results across all models. These two scenarios (i.e. augmented training dataset, hyperparameter-tuning) also show a small SD across different seeds, which means that the results are robust in terms of different hardware. The results that we obtained with hyperparameter-tuning were generated by using data from a third language (i.e. neither from the source nor from the target language) as validation data. The use of a third language dataset for validation should be confirmed by more experiments for other high- and low-level tasks, as well as for other languages.

When using the basic training dataset (which covers very few training instances) and default hyperparameters, mB0 performed best (only for Russian mB20 showed slightly higher results). mB0 was even able to produce significantly[16] better results than the baseline for Latin, which no other model achieved besides MAD-X. Even though mB20 achieved high results for Russian when using the basic training dataset, all other results are worse or only slightly better than the baseline. As the SD across different shots is very high (see Table 3), it is important to select an appropriate shot. This is inconvenient in a cross-lingual setup, since no validation data in the target language is available. Finding a solution for this problem would be beneficial, since the best shot for German led to results even higher than the results from MAD-X.

Overall, we reach an F1-score of 86.2 for Russian, comparable to Tsvetkov et al. (2014) with an F1-score of 86.0, but the random forest classifier was not able to outperform the neural models.

6 Qualitative Analysis

It was expected that the models perform better on German than on Russian. Afterall, more German than Russian data was used to pretrain mBERT, and German is typologically closer to the source language English than Russian. This expectation was not confirmed. Therefore, we carried out a qualitative analysis. Here, possible sources of errors were identified for German by looking at the predictions of zero-shot mBERT with default hyperparameters and basic training dataset. One hypothesis as to why the models performed worse for German than for Russian is that the target words consist of "computationally challenging" particle verbs (Köper and Schulte im Walde, 2016), i.e. combinations of a base verb (e.g. "schminken") with a prefix parti-

[15] We also applied this hyperparameter-tuning to the other neural and non-neural models, but observed no improvement.

[16] According to χ^2-testing for the model with seed 42 and p<0.05.

cle (e.g. “ab-”)[17]. They are highly productive and notoriously ambiguous. Also, the particle may be separated from the base verb. In contrast, the target words in the Russian dataset are frequent verbs.

Another hypothesis as to why the models performed worse on the German dataset than on the Russian dataset is that the German dataset contains many idioms. For example:

(12) Da wird der Teufel mit dem Beelzebub ausgetrieben. (translation: “One evil is replaced by another.”)

Interestingly, similar variants of this idiom were classified inconsistently. While the target word in (12) was misclassified as literal, it was correctly classified as metaphorical in (13):

(13) Denn die Elite und die USA werden den Teufel nicht mit einem Beelzebub austreiben. (translation: “For the elites and the U.S. will not replace one evil with another.”)

In total, three out of seven sentences that contain the idiom “den Teufel mit dem Beelzebub austreiben” were classified incorrectly. Similar behaviour was also observed for other highly conventionalized expressions, such as “Dampf ablassen” (translation: "let off steam"). In order to test whether the classifier indeed struggles with idioms, the dataset from Ehren et al. (2020) was used. This dataset consists of sentences from 34 preselected verbal idioms. For each sentence the information is given whether it contains a figuratively used idiom or not. In order to make it comparable to our version of the dataset by Köper and Schulte im Walde (2016), it was balanced and reduced to 2000 instances.

All neural models were applied to this dataset. As can be seen in Table 6, the results for the dataset by Ehren et al. (2020) were lower than the results for the dataset by Köper and Schulte im Walde (2016) across all models. This suggests that the neural methods for word-based MD do not work as well on idioms as they do on less conventionalized metaphors, especially since the target words (non-complex German verbs) are less computationally challenging in this dataset than the particle verbs in Köper and Schulte im Walde (2016).

A third hypothesis attributes classifier weakness to instances where the target verb is part of an extended metaphor:

(14) In der Gerüchteküche wurde tagelang deftig aufgekocht. (translation: “For days the gossip factory was working overtime.”)

Here, not only the target word is used metaphorically, but also most context words. This and comparable sentences were misclassified; apparently, too little evidence hinted at the metaphoricity. From 1792 sentences in the balanced dataset (Köper and Schulte im Walde, 2016) that we used for our experiments, 398 were misclassified. We analysed all 398 misclassifications. Our possible explanations regarding idiomatic rather than metaphorical expressions, and regarding larger metaphorical contexts, however, only account for roughly 26 misclassifications. We conclude that the vast majority of instances were misclassified either due to the structural difficulty of particle verbs, or that further reasons for the misclassifications still have to be identified. Additionally, the sentences in the Russian dataset are shorter, which makes it easier for the neural models to make correct predictions: the sentences in the Russian dataset contain an average of nine tokens, while the average sentence length for the German dataset is 13 tokens.

	Ehren	**Köper**
baseline	0.67	0.67
mB0	69.7±3.3	72.5±5.2
mB20	66.9±0.7	70.9±3.9
MAD-X	67.9±2.2	75.2±0.3

Table 6: Mean F1-scores (± SD) for detecting metaphorical usage in the dataset by Ehren et al. (2020) using three seeds (default hyperparameters, augmented training set); results for dataset by Köper and Schulte im Walde (2016) shown in gray.

7 Conclusion

While research on MD has focused on languages with comparably large amounts of data used for pretraining large language models, our experiments have shown that neural cross-lingual methods are suitable for languages with relatively large (Russian and German) and small amounts of pretraining data (Latin). Especially MAD-X performed very well, with the highest results across all experiments for German and Latin using a small training dataset

[17]The literal translation of the particle verb “abschminken” is “to remove makeup”.

and tuned hyperparameters, and for Russian using a large training dataset and default hyperparameters.

Zero-shot classification with mBERT performed decently on a small training dataset and default hyperparameters across all three languages. Few-shot classification with mBERT as applied in our experiments was not successful, as it relies on validation data for shot-selection, which is not possible in the cross-lingual setup. The non-neural random forest classifier, even though it yielded competitive results for Russian and German, was generally outperformed by the neural models – even for Latin, where small amounts of data were used to pretrain the neural models. It is unclear, however, why performance was better for Russian than for German across experiments. A qualitative analysis revealed a range of possible explanations, namely the inherent difficulty of particle verbs, idioms, and rich metaphorical contexts in the German dataset.

Whereas for the few-shot experiments we conducted sequential fine-tuning on source and target language data, Schmidt et al. (2022) showed that joint (instead of sequential) fine-tuning leads to few-shot models that yield higher results and are more robust in terms of hyperparameters (e.g. number of training epochs). We plan to employ this method for MD in future work, because few-shot fine-tuning showed promising results but still depends on target-language validation data. Another next step will be to compare our models' performance for Latin to their performance for Romance languages, in order to minimize the typological differences between the target languages. We will also investigate how the models presented in this paper perform in contrast to newer multilingual large language models such as mT5[18] (Xue et al., 2021).

8 Limitations

The MD methods described in this paper were investigated only for individual, curated sentences. Optimally, however, MD should be carried out on the basis of longer sequences from authentic data; here, also sequence-based metaphor detection should be applied to detect entire metaphorical phrases.

The target languages chosen for the experiments only cover a small subset of languages that were used to pretrain large language models; they should be repeated for other target languages with low amounts of pretraining data, especially those that do not belong to the Indo-European language family. Finally, English is studied as the only source language for the cross-lingual transfer, but it is possible that other languages with rather large amounts of pretraining data might be better suited as source languages.

[18] mT5 was pretrained on modern languages as well as on Latin data (Xue et al., 2021).

9 Acknowledgements

We thank the reviewers for their valuable feedback. We are also very grateful for the support of Filip Miletić, Prisca Piccirilli, Annerose Eichel and Andrea Horbach in the various stages of this study. This research was supported by the DFG Research Grant SCHU 2580/4-1 *(MUDCAT -– Multimodal Dimensions and Computational Applications of Abstractness).*

References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland. Association for Computational Linguistics.

Daniel Baleato Rodríguez, Verna Dankers, Preslav Nakov, and Ekaterina Shutova. 2023. Paper bullets: Modeling propaganda with the help of metaphor. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 472–489, Dubrovnik, Croatia. Association for Computational Linguistics.

Beata Beigman Klebanov, Chee Wee (Ben) Leong, and Michael Flor. 2018. A corpus of non-native written English annotated for metaphor. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 86–91, New Orleans, Louisiana. Association for Computational Linguistics.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Yo Joong Choe, Kyubyong Park, and Dongwoo Kim. 2020. word2word: A collection of bilingual lexicons for 3,564 language pairs. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3036–3045, Marseille, France. European Language Resources Association.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification

theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2218–2229, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Erik-Lân Do Dinh and Iryna Gurevych. 2016. Token-level metaphor detection using neural networks. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 28–33, San Diego, California. Association for Computational Linguistics.

Rafael Ehren, Timm Lichte, Laura Kallmeyer, and Jakub Waszczuk. 2020. Supervised disambiguation of German verbal idioms with a BiLSTM architecture. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 211–220, Online. Association for Computational Linguistics.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet – An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA, USA.

E. Darío Gutiérrez, Guillermo Cecchi, Cheryl Corcoran, and Philip Corlett. 2017. Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2923–2930, Copenhagen, Denmark. Association for Computational Linguistics.

Maximilian Köper and Sabine Schulte im Walde. 2016. Distinguishing literal and non-literal usage of German particle verbs. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 353–362, San Diego, California. Association for Computational Linguistics.

Zoltan Kövecses. 2010. *Metaphor: A Practical Introduction*. Oxford University Press.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multilingual multi-figurative language detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9254–9267, Toronto, Canada. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 1980. *Metaphors we Live by*. University of Chicago Press, Chicago.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29, Online. Association for Computational Linguistics.

Yucheng Li, Shun Wang, Chenghua Lin, and Frank Guerin. 2023. Metaphor detection via explicit basic meanings modelling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 91–100, Toronto, Canada. Association for Computational Linguistics.

Weicheng Ma, Ruibo Liu, Lili Wang, and Soroush Vosoughi. 2021. Improvements and extensions on metaphor detection. In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, pages 33–42, Online. Association for Computational Linguistics.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

Roland Schäfer and Felix Bildhauer. 2012. Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 486–493, Istanbul, Turkey.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2022. Don't stop fine-tuning: On training regimes for few-shot cross-lingual transfer with multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10725–10742, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ekaterina Shutova. 2015. Design and evaluation of metaphor processing systems. *Computational Linguistics*, 41(4):579–623.

Ekaterina Shutova and Simone Teufel. 2010. Metaphor corpus annotated for source-target domain mappings. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. DeepMet: A reading comprehension paradigm for token-level metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39, Online. Association for Computational Linguistics.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686, Online. Association for Computational Linguistics.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Ivan Vulić and Marie-Francine Moens. 2013. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1613–1624, Seattle, Washington, USA. Association for Computational Linguistics.

Joshua R. Westbury, Kris Lyle, Jimmy Parks, and Jeremy Thompson. 2016. *The Lexham Figurative Language of the Bible Glossary*. Lexham Press.

Michael Wilson. 1997. MRC psycholinguistic database: Machine usable dictionary, version 2.00. *Behaviour Research Methods*, 20.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Hyperparamter-Tuning for MAD-X: Additional material

Table 7 reports the sets of hyperparameters that were used during hyperparameter search for the MAD-X classifier. Figure 3 shows which hyperparameter set led to which F1-score for each of the four languages. This figure hints at the fact that the correlations between English and each of the three languages Russian, German and Latin are low, while the correlation for language pairs not including English are high. We quantified this assumption by calculating Spearman's rank-order correlations presented in Figure 2 (see Section 5.3).

index	learning rate	epochs	train batch size
1	1e-3	10	8
2	1e-3	10	16
3	1e-3	10	32
4	1e-3	50	8
5	1e-3	50	16
6	1e-3	50	32
7	1e-3	100	8
8	1e-3	100	16
9	1e-3	100	32
10	1e-4	10	8
11	1e-4	10	16
12	1e-4	10	32
13	1e-4	50	8
14	1e-4	50	16
15	1e-4	50	32
16	1e-4	100	8
17	1e-4	100	16
18	1e-4	100	32
19	1e-5	10	8
20	1e-5	10	16
21	1e-5	10	32
22	1e-5	50	8
23	1e-5	50	16
24	1e-5	50	32
25	1e-5	100	8
26	1e-5	100	16
27	1e-5	100	32

Table 7: Index to hyperparameter sets for MAD-X.

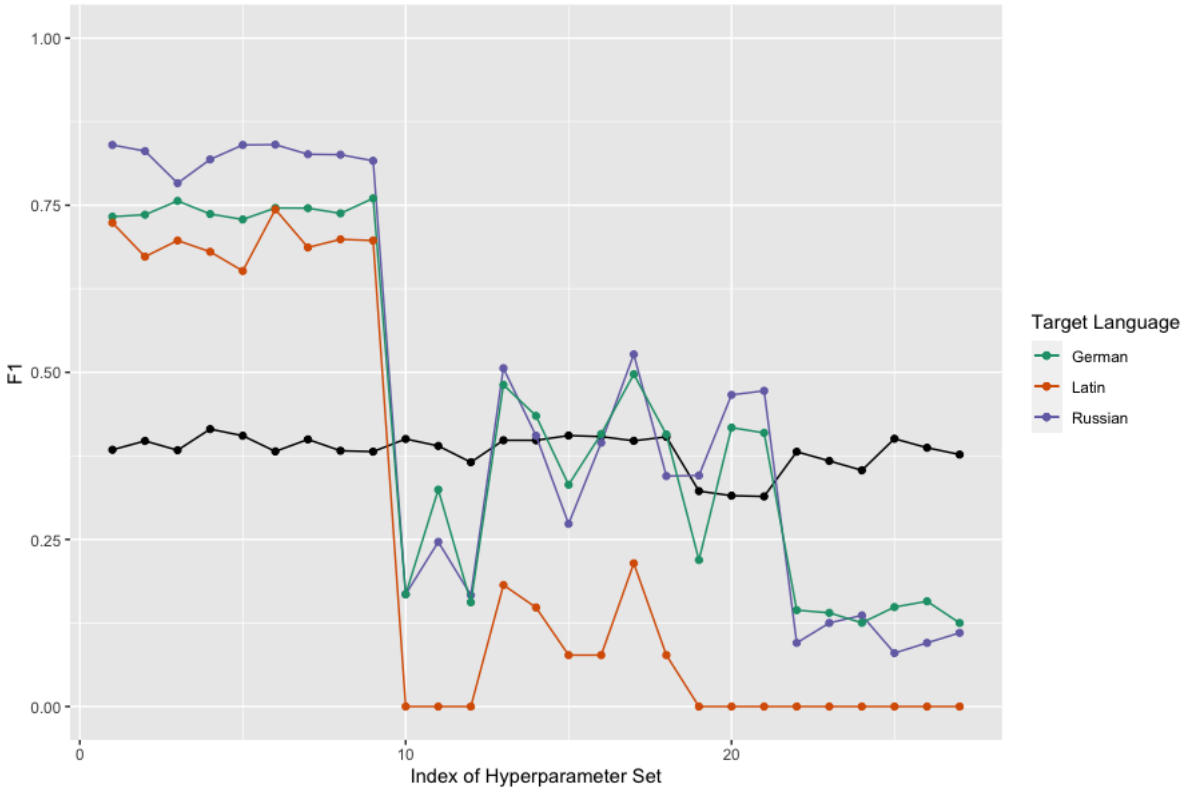


Figure 3: Result for using both the data from Mohammad et al. (2016) (black line) and the different test sets for target languages Russian, German and Latin as dev sets for the grid search on zero-shot classification with MAD-X.

A Hard Nut to Crack: Idiom Detection with Conversational Large Language Models

Francesca De Luca Fornaciari[1], Begoña Altuna[2], Itziar Gonzalez-Dios[2], Maite Melero[1]
[1]Barcelona Supercomputing Center (BSC)
[2]HiTZ Center - Ixa, University of the Basque Country UPV/EHU
[fdelucaf,maite.melero]@bsc.es, [begona.altuna,itziar.gonzalezd]@ehu.eus

Abstract

In this work, we explore idiomatic language processing with Large Language Models (LLMs). We introduce the Idiomatic language Test Suite IdioTS, a new dataset of difficult examples specifically designed by language experts to assess the capabilities of LLMs to process figurative language at sentence level. We propose a comprehensive evaluation methodology based on an idiom detection task, where LLMs are prompted with detecting an idiomatic expression in a given English sentence. We present a thorough automatic and manual evaluation of the results and an extensive error analysis.

1 Introduction

The continuous improvements in LLM performance raise the hypothesis that their exposure to vast amounts of pre-training data may give them the capability to accurately process the meaning of natural language utterances. We conducted a thorough analysis of the behaviour of three small-sized, instruction-tuned LLMs, tasked with figurative uses of language. The goal of this work is to provide a comprehensive evaluation methodology centred around a new test suite, IdioTS,[1] designed to assess the capabilities of LLMs to distinguish between figurative and literal meanings of Potentially Idiomatic Expressions (PIEs). The adopted definition of PIE is the one provided by Haagsma et al. (2020): expressions that can have an idiomatic meaning, regardless of whether they actually have that meaning in a given context.

2 Related work

The question about to what extent LLMs can interpret non-literal phrases remains open (Jhamtani et al., 2021). The creation of numerous figurative language datasets as a fundamental resource for evaluation underscores the importance of this issue in Natural Language Processing (NLP). To the best of our knowledge, these are some of the most significant existing datasets on figurative language.

The MAGPIE corpus, created by Haagsma et al. (2020), is a large sense-annotated corpus of PIEs created from a highly curated list of idioms. This dataset has been employed in numerous studies (Tan and Jiang, 2021; Madabushi et al., 2021; Dankers et al., 2022), exploring figurative language processing from the most diverse perspectives.

The Fig-QA dataset was developed by Liu et al. (2022) to test the ability of LLMs to reason about figurative language. The findings of the conducted experiments underscored that LLMs still fall short of human performance, particularly in zero- or few-shot settings.

The IMPLI dataset (Stowe et al., 2022) is a human annotated dataset consisting of paired sentences spanning idioms and metaphors, designed for natural language inference (NLI). The task consists in predicting whether the meaning of one text fragment (premise) entails another (hypothesis). Experiment findings indicate that, even when pre-training data includes figurative sentences, idiomatic language remains a challenge for pre-trained language models.

The ID10M multilingual dataset developed by Tedeschi et al. (2022) was proposed as part of a complete framework for idiom identification in several languages. The conducted experiments demonstrate that a model fine-tuned on this dataset is able to correctly predict the majority of idiomatic PIEs, but struggles with literal PIEs, tending to attribute them an idiomatic meaning.

The FLUTE dataset, introduced by Chakrabarty et al. (2022), is a dataset of textual explanations of figurative expressions. The results of the experiments conducted with models fine-tuned on

[1]The resource is published under an open licence (CC BY-SA-NC 4.0) and can be accessed at this URL: https://ixa.si.ehu.es/node/14017

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 35–44
June 21, 2024 ©2024 Association for Computational Linguistics

FLUTE showed how such dataset can contribute to developing models that understand figurative language through textual explanations.

3 Dataset creation process

We introduce a new evaluation dataset specifically crafted for idiom detection in English. The rationale behind the creation of a new resource from scratch, rather than building on a pre-existing dataset, is grounded in the need to avoid data contamination by providing data that we can guarantee the assessed models have not seen before. In this section, we describe all the steps involved in the creation process of the Idiomatic language Test Suite IdioTS.

3.1 Idioms list creation

Drawing inspiration from Haagsma et al. (2020), we manually built a highly curated list of idioms extracted from diverse online platforms, such as *Amigos Ingleses*,[2] *The idioms*,[3] and *EF English idioms*.[4] We selected the idioms with a view to producing a sufficiently comprehensive list in terms of diversity of syntactic structures. We included not only phrases with a completely fixed morpho-syntactic structure ("Nothing to write home about"), but also constructions with a high morpho-syntactic variability ("To blow your own trumpet"). The idioms within the resource encompass verb-object constructions ("Hold your horses"), a wide range of structures with the verb "to be" followed by a prepositional phrase ("To be on the ball", "To be up your street"), adjective-noun combinations ("Cold turkey"), more or less complex prepositional phrases ("By the skin of your teeth", "Out of the blue"), binomial pairs consisting of two nouns linked by a conjunction ("Bits and bobs"), and appositional compounds ("Easy-peasy"), among others. The idioms included in our list pertain to a colloquial text style and are frequent in spoken everyday language.

As a following step, we meticulously reviewed the idioms list to ensure a high degree of homogeneity. For syntactically flexible and semi-fixed expressions, adjustments were made by placing the main verb in the infinitive tense and in the active form. Personal pronouns and determiners were replaced with indefinite pronouns (e.g. "It serves you right" became "To serve someone right"). Idioms with a fixed morpho-syntactic structure were preserved in their original form (e.g. "Don't quote me", "Hold your horses"), as this is the sole form in which they appear in authentic usages. The resulting database consists of 93 idioms, each associated with a unique alphanumeric identifier and the original source from which it was extracted.

3.2 Idiomatic sentence crafting

Even though for the majority of the idioms an example sentence was provided in the original source, we decided to craft entirely new sentences in order to minimise the risk of data contamination.

As crowdsourcing has become increasingly popular for language resource development in NLP applications (Drutsa et al., 2021), and is considered a valid method to outsource data generation by mitigating potential researcher bias, we organised a small-scale crowdsourcing on a voluntary basis. To ensure the quality of the generated sentences, we established the essential requirements collaborators had to fulfil: native English speakers, predominantly of British origins, with a demonstrated high linguistic proficiency attaining at least a C1 level.

Collaborators were eight language professionals with a linguistic background (English teachers, linguists, translators, and NLP experts). They were provided with a spreadsheet containing just the idioms and an empty cell to fill with a sentence, without any additional context. They were instructed to select a few idioms of their choice and to craft a sentence per chosen idiom. They were asked to produce sentences representative of natural, spontaneous language use, provided it resonated authentically with their native speaker experience. An idiom with its corresponding sentence was included as an example. Through this initiative we obtained the 164 idiomatic sentences corresponding to the positive class of our dataset.

3.3 Distractor sentence crafting

At this point, the dataset needed to be augmented with instances of the negative class, i.e. plausible, grammatically and syntactically correct sentences containing a set of words that might belong to an idiomatic expression, but in fact are employed in a less common, literal way. These are meant to be the most challenging portion of our dataset. Whereas the interpretation of the meaning of distractor sentences would pose minimal difficulty for a human

[2] https://www.amigosingleses.com/

[3] https://www.theidioms.com/

[4] https://www.ef.com/wwen/english-resources/english-idioms/

reader, our intuition was that a LLM would encounter issues with this particular type of sentences. The complex task of generating this kind of sentences was undertaken internally to ensure both their quality and correctness, while also providing a subtle suggestion of idiomaticity. We employed various approaches. Whenever possible, a new sentence was crafted by selecting the complete set of words composing the idiom and placing it unchanged in a different semantic context that, from a human perspective, unequivocally determined its literal meaning. This happened, for instance, with the idiom idi07 "Bob's your uncle" (1):

(1) What a surprise! I didn't know Bob's your uncle.

For other idioms, like idi25 "It's like talking to a brick wall", not the entire expression but only certain elements — the verb "to talk" and the noun phrase "brick wall" — were extracted and placed in a different context that changed their meaning to literal (2):

(2) Let's talk about how a brick wall can add charm and character to any space.

In other cases, the applied strategy was to use some of the words composing the idiom with a different syntactic or even morphological role, like it happened for the idiom idi82 "To make a living" (3):

(3) I bought a new lamp and lots of plants to make our living room warmer and more cosy.

One final employed method involved proposing an expression with a certain character overlapping and assonance with the idiom, for example "speed and span" and "spick and span" (4):

(4) It is difficult to measure the speed and span of the dissemination of the virus.

3.4 Sentence proofreading and final layout

As far as possible, efforts were made to avoid having more than one PIE in a single sentence. This strategy aimed to simplify the comprehension and execution of the task for the models as well as the collection and analysis of the model's responses for the researchers.

Additionally, a concerted effort was made to mitigate gender bias within our newly developed resource. Whenever possible, gender-specific terms were either eliminated or neutralised, a large number of sentences were reformulated adopting a gender neutral first person plural ("we"/"us"), second person singular or plural ("you"), or third person plural ("they"). Since the gender neutralisation is not always possible due to grammatical or syntactical constraints, meticulous attention was devoted to ensuring a representation of feminine and masculine gender terms as balanced as possible throughout the dataset.

Finally, each sentence was assigned a unique alphanumeric identifier containing information about the related idiom and a suffix indicating whether it is an idiomatic or a distractor sentence.

The final Idiomatic language Test Suite IdioTS is composed by a total of 250 sentences, 164 of which are idiomatic and 86 distractor sentences.

4 Experiment definition

Our experimental focus was pointed at evaluating the ability of the selected LLMs to detect an idiomatic expression in a given sentence. This experiment falls within the context of "idiom detection" and involved a binary sentence classification task, being the two classes to predict "idiomatic" (positive class) and "non-idiomatic" (negative class). The goal was to assess whether LLMs are able to accurately capture the meaning of a PIE, distinguishing between figurative and literal meaning based on the formulation of the sentence.

Assuming that the pre-training data for these models contained the specific PIEs far more frequently with idiomatic than with literal meaning, the models may be inclined to attribute a figurative meaning to the expression based on probability distribution.

4.1 Assessed LLMs

Ensuring a fair comparability among models is an unresolved challenge, due to the many internal aspects of a model that remain undisclosed. Nevertheless, for the scope of this study, we attempted to minimise differences, focusing on three LLMs that have the following characteristics in common: they have a transformer-based architecture and approximately 7 billion parameters in size, they are open source and fine-tuned for dialogue. The preference for open-source over proprietary models was motivated by transparency and reproducibility reasons, along with cost implications. The choice of the smallest model within a specific model family

was motivated by the possibility to conduct experiments in a resource-efficient way, by using a local machine without a GPU. The choice of instruction fine-tuned, conversational models was based on the idea of simulating a real-world scenario where a user employs a chatbot application to solve a task or find an answer to a question.

In accordance with these considerations, we included the following models in our assessment:

- Llama-2-7b-chat (Touvron et al., 2023).
- Mistral-7b-Instruct (Jiang et al., 2024).
- Vicuna-7b (Zheng et al., 2023).

Regarding configuration, we maintained default values for most hyper-parameters, such as top-k: 40 and top-p: 0.95, as we observed that altering these values in the development phase did not significantly impact the output. However, we had to extend the default token limit to 800 to accommodate the long prompt and the verbose model responses, and prevent errors related to exceeding the maximum token length. We also set the temperature to 0 in order to make the model output deterministic and the experiment reproducible.

4.2 Prompt engineering

At a broad level, the key of successful prompts lies in incorporating all necessary information while avoiding excessively complex instructions. For our experiment, we employed the following question as the central component of the prompt: "Is there an idiom in the sentence?", followed by the sentence to analyse.

Conversational LLMs typically accept prompts structured in two parts: the system prompt, a generic instruction about the models behaviour in interactions, and the user prompt, containing the specific question or request. In development, we accurately chose the optimal prompt structure for our experiment, which is exemplified in Appendix A, Figure 1 and contains all the elements listed in the following lines.

Defining the persona. This technique consists in assigning the model a specific role by including a short description in the prompt. In our case, we adopted this formulation: "You are a professional linguist specialising in figurative language". Introducing the concept of "figurative language" we intended to guide the model to focus on this specific linguistic phenomenon. However, we acknowledge the potential risk of introducing some level of researcher bias.

Describing the task. This was expressed through this wording: "Your task is to analyse English sentences that may contain an idiom, also known as an idiomatic expression". To ensure accurate language identification, we specified the language name. Additionally, we employed two distinct forms to refer to idiomatic expressions, aiming to provide the most precise task description.

Zero-shot prompting. We added no examples to the prompt. Through this approach we intended to test the model's ability to perform the task based on the task description alone.

Including a definition of "idiom". Due to the lack of an unique agreed-upon definition of idiom, we saw the need to include a concise definition, in an effort to narrow down the potential variations in model outputs: "A phrase, expression, or group of words that has a meaning different from the individual meanings of the words themselves, and employed to convey ideas in a non-literal or metaphorical manner".

Requiring an answer in JSON format. In order to mitigate the issue of overgeneration related to conversational LLMs, an explicit instruction was added to guide the model to provide an answer in a JSON format, specifying the fields and the information to include in each field of the JSON file. This approach forced the model to provide all and only the required information, structured in a way that facilitated the collection and analysis of the output. The wording for this instruction was the following: "The response should be in strict JSON format including four fields", where we specified header and content for each field as follows:

- 'hasIdiom': Is there an idiom in the sentence? Give a true/false answer.
- 'idiom': Should include which is the idiom contained in the sentence.
- 'meaning': Should explain the meaning of the identified idiom.
- 'explanation': Should include a concise elaboration.

From a technical point of view, we used the llama-cpp-python binding[5] that supports inference for many LLMs models and played a crucial role in

[5]https://github.com/abetlen/llama-cpp-python

converting the standardised prompt into a specific input format compatible with each of the models during the inference process. As an example, in Appendix A, Figure 2 we show the input format generated for Llama2, where the tags delimiting system and user prompt were replaced with the standard ones accepted by this particular model.

5 Findings

We established two different levels of evaluation for the experiment. The first level consists of a completely automatic evaluation, whereas the second level is complemented with a thorough manual evaluation and error analysis.

5.1 First level of evaluation

At the first level, we employed the following automatic metrics to assess the capability of a model to detect an idiom in a given English sentence: Accuracy, Misclassification Rate (MR), Recall, Specificity, Precision, and Balanced Accuracy.

These metrics offer a general overview of the behaviour of the models and facilitate comparisons.

In Table 1 we present the aggregated results. At a broad level, all three models fall within the same range of results, as they show close scores in terms of Accuracy and Misclassification Rate.

	Llama2	Mistral	Vicuna
Accuracy ↑	0.656	0.660	**0.676**
MR ↓	0.344	0.340	**0.324**
Recall ↑	**1.0**	0.896	0.988
Specificity ↑	0.0	**0.209**	0.081
Precision ↑	0.656	**0.680**	0.672
Balanced Accuracy ↑	0.5	**0.553**	0.535

Table 1: Automatic metrics calculated for the three models. Numbers in bold indicate which model achieved the best result for each metric. For Misclassification Rate, lower values are indicative of better performance, as denoted by the downward arrow.

When we observe further metrics, such as Recall and Specificity, we immediately notice a particular behaviour for Llama2. The model shows a Recall of a hundred percent, meaning that it correctly classified all the idiomatic sentences, and a Specificity of 0.0, meaning that it did not correctly classify any of the distractor sentences. In fact, Llama2 only provided positive answers. This behaviour is known as *acquiescence* or *agreement bias* and consists in the model trying to always provide an answer that is compliant or satisfies the user request. As demonstrated by our experiment, this can have counterproductive effects, leading the model to provide inaccurate responses.

Specificity, also known as True Negative Rate, is especially significant in our study, since it expresses the number of distractor sentences that were correctly classified. Given our initial assumption about distractor sentences being especially challenging, a high score for this metric reflects a good performance within the scope of the proposed task. Mistral not only exhibits the best Specificity score, but also a considerable lead over the other models, clearly demonstrating its superiority in this specific aspect. Furthermore, it achieves the best score for Precision, even though the difference compared to the other models is less pronounced.

Even though Vicuna obtained slightly better scores than Mistral and Llama2 in terms of Accuracy and MR, we can observe that Mistral strikes the best score in terms of **Balanced Accuracy**. In our scenario, where the positive class in the dataset is double the size of the negative class, Balanced Accuracy is a more robust metric, and it provides a more reliable measure of classification performance in the face of imbalanced data.

Analysis of misclassifications All incorrect classifications for Llama2 are of the type false positive. Regarding Mistral and Vicuna, the two models share a similar distribution of misclassifications, being false positive the predominant type for both. This indicates that the most common behaviour pattern across models was incorrectly attributing idiomaticity to a sentence that is not idiomatic. Conversely, both models exhibit fewer misclassifications of the false negative type, suggesting that they were generally effective in identifying the presence of an idiomatic expression in a sentence.

These observations align with findings from Tedeschi et al. (2022), and with our initial intuition that, given the pre-training data likely contains the given PIEs with idiomatic meaning more frequently than with a literal meaning, the models tend to classify these expressions as idiomatic rather than literal based on probability distribution.

5.2 Second level of evaluation

In our study, for each sentence classified as idiomatic, the models were asked to additionally specify the detected idiomatic expression. We observed that in a certain number of cases the models, despite correctly classifying a sentence as id-

iomatic, did not detect the correct idiom and rather identified some other part of the sentence as idiomatic, such as a phrasal verb, a collocation, or a single word. This observation underscores that general metrics are insufficient to conclusively demonstrate the capability of a LLM to detect an idiom in a sentence and motivated us to perform an additional verification step to validate the accuracy of true positive classifications. We calculated **True Positive Consistency** as the proportion of true positive predictions where the correct idiomatic expression was accurately identified as well. This additional score allowed us to validate whether the models response was grounded in the correct reason.

Table 2 displays True Positive Consistency values for the three analysed models. Mistral exhibits the best score, achieving a True Positive Consistency of 0.905, followed by Vicuna, and lastly Llama2.

	Idiomatic sentences (positive class)	True positives	True positives with correct reason	True Positive Consistency↑
Llama2	164	164	138	0.841
Mistral	164	147	133	**0.905**
Vicuna	164	162	144	0.889

Table 2: True Positive Consistency values per model.

5.2.1 Error analysis

By carefully examining the responses in the 'idiom' and 'meaning' fields, we identified the elements that the models incorrectly detected as an idiom and upon which they based their classification of the sentence as idiomatic. We identified recurring error patterns across the three models.

True positive wrong reason error types Regarding true positive with wrong reason, all of the three models in some cases identified a phrasal verb as an idiom ("to run off", "to look up").

In other cases, the models detected an element in the sentence and reported an existing idiom — containing this element — which meaning is unrelated with the analysed sentence, like in the following example (5) from Llama2:

(5) Just cut to the chase! What did she say about us using her *ironing* board?

Idiom: *Ironing* out differences.

Meaning: To resolve conflicts or disagreements.

In some other cases, the models reported an existing idiom, that seemed completely unrelated with the given sentence. Regarding Vicuna, it seems that the model engaged in an inverse reasoning process by firstly extracting the meaning of the idiomatic expression from the sentence and subsequently identifying a different idiomatic expression that conveys a similar meaning.

(6) That's the last straw! I won't let you push me around any further!

Idiom: To break the camel's back.

Meaning: To reach a point where one can no longer tolerate or endure something.

As we can observe in (6), the explanation provided by the model is consistent with the meaning of the original idiom "To be the last straw", as well as the identified idiom "To break the camel's back".

A last detected pattern for true positive with wrong reason was the models inventing a non-existing idiom by using a single word or a chunk of the examined sentence as in (7). With 23 occurrences, Llama2 was the model that showed this error type with the highest frequency:

(7) I'm going to go out on a limb and say they *used the old model* for this task.

Idiom: *Used the old model*.

Meaning: To use something that is no longer current or fashionable.

We observed that the meaning explanations offered by the models appear plausible and coherent. This represents a potential concern, as it may lead to an over-reliance on model outputs, especially in situations where users lack the necessary language knowledge to verify the semantic content of the response.

False negative error types In our study, false negatives are idiomatic sentences that were incorrectly classified as non-idiomatic. This type of classifications were entirely absent for Llama2. Regarding Vicuna, this happened in two instances, where the model failed to detect the idioms "To make up your mind" and "To be crystal clear". However, the model successfully detected these same idioms in the context of a different sentence. Mistral classified 17 instances as false negatives. Despite the higher number of cases, the model did not exhibit a consistent behaviour. In some cases, it demonstrated a lack of knowledge of a specific idiom, failing to recognise it in different contexts.

In other cases, its ability to detect a specific idiom varied as the sentence context changed. Due to the lack of consistency, we were unable to identify any discernible patterns or offer an explanation for the behaviour of the models for this specific error type.

False positive error types Regarding instances where models wrongly attributed an idiomatic meaning to a literal sentence, we identified two main error types. By the first type, the models wrongly classified a sentence as idiomatic and erroneously attributed an idiomatic meaning to some random element of the sentence, like in the following example (8) from Mistral:

(8) We need to cut and tighten this rope: it is too slack!

Idiom: Cut and tighten.

Meaning: To shorten or make something more concise.

We consider that this type of error might be, to some extent, a consequence of researcher bias: by explicitly asking the model whether a sentence contains an idiom may force the model to actively search for an idiom, altering the neutrality of its response. Therefore, we defer a detailed analysis of this type of errors to future work.

Conversely, we consider the second type, where models wrongly classified a sentence as idiomatic and erroneously attributed an idiomatic meaning to the PIE associated with the sentence, of significant interest for our study. In these cases, it seems plausible to assume that the models might have fallen into the intentional "traps" we set by incorporating distractor sentences into our dataset.

	Distractor sentences (negative class)	**False positives associated PIE: total↓**	**False positives associated PIE: ratio↓**
Llama2	86	55	0.640
Mistral	86	53	0.616
Vicuna	86	**47**	**0.546**

Table 3: Number of false positives with idiomatic meaning attributed to the associated PIE over total distractor sentences per model.

Table 3 presents, for each model, the ratio of distractor sentences where the model attributed an idiomatic meaning to the associated PIE over the total number of distractor sentences (86) in the dataset. As we can observe, the three examined models exhibit a comparable behaviour, with Vicuna showing the smallest number of errors of this type.

6 Conclusions and future work

The use of figurative language is a complex linguistic phenomenon that poses hard challenges for LLMs. Despite its critical role within numerous NLP tasks, it still remains a relatively underexplored area of investigation.

In this work we addressed the specific domain of idiomatic expressions in English as a special case of figurative language use. As a part of our contribution:

- We introduced the new Idiomatic language Test Suite IdioTS, manually curated by language experts, and covering especially challenging idiomatic and literal uses of language.
- We proposed a comprehensive methodology for the assessment of the linguistic capabilities of LLMs in relation to idiomatic language.
- We conducted an idiom detection experiment focused on the assessment of the capabilities of small conversational LLMs to detect idioms within ambiguous English sentences.
- We conducted a thorough manual evaluation and error analysis and observed the main behaviour patterns of LLMs within this task.

The findings from our study indicate that when it comes to capturing the meaning of an ambiguous sentence, LLMs struggle to distinguish between literal and idiomatic uses of language. In line with the observations in the literature, a high acquiescence or agreement bias was observed: LLMs tend to force the identification of an idiom by assigning idiomatic meaning to an aleatory element in the sentence. Additionally, they offer coherent explanations to reinforce their inaccurate answers, which can be a cause for concern.

As future research directions, we intend to broaden our experiments by extending them to one- and few-shot scenarios, by exploring other prompting techniques focused on mitigating researcher bias and incorporating the possibility to interact with conversational models in multi-turn conversations.

Regarding the proposed IdioTS, we plan to explore several data augmentation techniques to generate additional idiomatic and distractor sentences. Additionally, a categorisation of distractor types could be incorporated to gain an understanding of which constructions are the most challenging for

the models. Moreover, we intend to translate the sentences into other languages to create a multilingual dataset and open a path for MT experiments aimed to investigate possible correlations between idiom detection and translation.

At a broad level, exploring models with different architectures, sizes, and hyper-parameter configurations could provide valuable insights into how these models characteristics relate to the capabilities of LLMs to process natural language and could open avenues for targeted experimentation, such as specific fine-tuning strategies, aimed at enhancing the performance of LLMs across various natural language tasks.

Acknowledgements

This work has been partially funded by i) DeepR3 (TED2021-130295B-C31) funded by MCIN/AEI/10.13039/501100011033 and European Union NextGeneration EU/PRTR. ii) Ixa group A type research group (IT-1805-22) funded by the Basque Government, iii) DeepKnowledge (PID2021-127777OB-C21) project funded by MCIN/AEI/ 10.13039/501100011033 and by ERDF A way of making Europe and iv) AWARE Commonsense for a new generation of natural language understanding applications (TED2021-131617B-I00) funded by MCIN/AEI /10.13039/501100011033 by the European Union NextGenerationEU/ PRTR.

References

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative Language Understanding through Textual Explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Alexey Drutsa, Dmitry Ustalov, Valentina Fedorova, Olga Megorskaya, and Daria Baidakova. 2021. Crowdsourcing Natural Language Data at Scale: A Hands-On Tutorial. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 25–30, Online. Association for Computational Linguistics.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A Large Corpus of Potentially Idiomatic Expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Taylor Berg-Kirkpatrick. 2021. Investigating robustness of dialog models to popular figurative language constructs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L'elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts. *ArXiv*, abs/2401.04088.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the Ability of Language Models to Interpret Figurative Language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. Astitchinlanguagemodels: Dataset and methods for the exploration of idiomaticity in pre-trained language models.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI Models' Performance on Figurative Language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Minghuan Tan and Jing Jiang. 2021. Does BERT understand idioms? a probing-based empirical study of BERT encodings of idioms. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1397–1407, Held Online. INCOMA Ltd.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. ID10M: Idiom Identification in 10 Languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *ArXiv*, abs/2307.09288.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Prompt formats used for the idiom detection experiment

In Figure 1 we present the content of the prompt passed to all the assessed models.

“role”: “system”,
“content":
You are a professional linguist specialising in figurative language and your task is to analyse English sentences that may contain an idiom, also known as an idiomatic expression. This is a definition of idiom: ‘A phrase, expression, or group of words that has a meaning different from the individual meanings of the words themselves, and employed to convey ideas in a non-literal or metaphorical manner’.
The response should be in strict JSON format including four fields:
‘hasIdiom’: Is there an idiom in the sentence? Give a true/false answer.
‘idiom’: Should include which is the idiom contained in the sentence.
‘meaning’: Should explain the meaning of the identified idiom.
‘explanation’: Should include a concise elaboration.
“role”: “user”,
“content”: sentence

Figure 1: Prompt passed to all the assessed models.

In Figure 2 we present the specific layout of the prompt generated by the llama-cpp-python binding for Llama2.

```
<s>[INST] <<SYS>>

You are a professional linguist specialising in figurative language and your task
is to analyse English sentences that may contain an idiom,
also known as an idiomatic expression. This is a definition of idiom: 'A phrase, expression,
or group of words that has a meaning different from the individual meanings of the words
themselves, and employed to convey ideas in a non-literal or metaphorical manner'.
The response should be in strict JSON format including four fields:
 'hasIdiom': Is there an idiom in the sentence? Give a true/false answer.
 'idiom': Should include which is the idiom contained in the sentence.
 'meaning': Should explain the meaning of the identified idiom.
 'explanation': Should include a concise elaboration
<</SYS>> She didn't know the chords of that song, she was playing it by ear. [/INST]
```

Figure 2: Specific layout of the prompt generated by the llama-cpp-python binding for Llama2.

The Elephant in the Room: Ten Challenges of Computational Detection of Rhetorical Figures

Ramona Kühn
Innstraße 43
94032 Passau
University of Passau
ramona.kuehn@uni-passau.de

Jelena Mitrović
Innstraße 43
94032 Passau
University of Passau/
Institute for AI R&D of Serbia
jelena.mitrovic@uni-passau.de

Abstract

Computational detection of rhetorical figures focuses mostly on figures such as metaphor, irony, or sarcasm. However, there exist many more figures that are neither less important nor less prevalent. We want to pinpoint the reasons why researchers often avoid other figures and shed light on the challenges they struggle with when investigating those figures. In this comprehensive survey, we analyzed over 40 papers dealing with the computational detection of rhetorical figures other than metaphor, simile, analogy, sarcasm, and irony. We encountered recurrent challenges from which we compiled a ten point list. Furthermore, we suggest solutions for each challenge to encourage researchers to investigate a greater variety of rhetorical figures.

1 Introduction

Rhetorical figures such as metaphor, alliteration, or irony are present in our daily lives. They make language vivid, more emotional, or more persuasive. Each figure has a special function, e.g., figures with repetition create more emphasis (Fahnestock, 2002), while sarcasm and irony are often used in the context of hate speech (Frenda et al., 2023). To understand the often non-literal meaning and subtle nuances of a text containing rhetorical figures, it is important to reliably detect those figures computationally. Furthermore, the performance of classical NLP tasks improves when taking features of rhetorical figures into account. This was demonstrated for sentiment analysis (Nguyen et al., 2015), argumentation mining (Mitrović et al., 2017), text summarization (Alliheedi and Di Marco, 2014), and hate speech and abusive language detection (Lemmens et al., 2021).

Most detection approaches only consider the rhetorical figures metaphor (Shutova et al., 2013; Ghosh et al., 2015; Bizzoni et al., 2017; Bizzoni and Ghanimifard, 2018; Chakrabarty et al., 2022; Rai and Chakraverty, 2020; Tong et al., 2021; Ge et al., 2023), or irony and sarcasm (Ghosh et al., 2015; Wallace, 2015; Joshi et al., 2017; Yaghoobian et al., 2021). However, the Silva Rhetoricae[1] (Burton, 2007), an online resource for rhetorical figures and their descriptions, lists 435 different rhetorical figures. Most of those figures are neither less present nor less important than metaphor. For example, the figure antithesis is important in environmental (Green, 2021) or populist communication (Kühn et al., 2024), litotes is important in sentiment analysis (Karp et al., 2021), and polyptoton can highlight similarities while showing a distinction (Fahnestock, 2002).

We believe that it is essential to pay attention to the other figures, too. In this survey, we investigate the main challenges and problems researchers struggle with when computationally dealing with those figures. We examined over 40 papers describing computational detection approaches for rhetorical figures other than metaphor, simile, analogy, sarcasm, and irony. The figures range from A like alliteration to Z like zeugma. Table 3 in Appendix A illustrates the distribution of figures across the papers we examined, showing the frequency of appearance for each figure. The investigated papers were published between 2006 and 2024.

We focus on papers that consider the detection of rhetorical figures in written text, as speech or multimodal approaches further increase both the complexity and challenges. We were looking on Google Scholar[2] for relevant work by searching for figure names along with "detection", and including relevant related work. We explicitly did not look only into libraries such as the ACL anthology[3], as the field of rhetorical figure detection is not that represented at big conferences. From these results, we compiled a comprehensive list of ten key chal-

[1] http://rhetoric.byu.edu/
[2] https://scholar.google.com/
[3] https://aclanthology.org/

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 45–52
June 21, 2024 ©2024 Association for Computational Linguistics

lenges and problems that show a recurrent pattern. We also provide suggestions for overcoming those challenges in order to further strengthen the field of computational detection of rhetorical figures in the future.

2 Rhetorical Figure Detection: Ten Challenges

We present ten challenges that most researchers face when trying to computationally identify rhetorical figures. We also suggest solutions for each of the challenges.

2.1 Inconsistent Definitions and Binary Classification

Although rhetorical figures have been studied for hundreds of years from a linguistic perspective, their spellings and definitions are often inconsistent (Harris et al., 2018; Gavidia et al., 2022; Kühn and Mitrović, 2024). This leads to different interpretations of what a rhetorical figure consists of. Consider, for example, the figure antithesis ("working all day, sleeping all night"). Most definitions agree on the antonymous relation (working vs. sleeping), but not every definition requires syntactic parallelism. Another example comes from the work of Dubremetz and Nivre (2017), in which the figure chiasmus is described, but the authors actually refer to a more specific form of chiasmus called antimetabole (Schneider et al., 2021). A further problem is that some figures are language dependent, i.e., a rhetorical figure in English does not have a matching counterpart in another language (Kühn et al., 2022; Zhu et al., 2022). For example, metaphor and simile are considered one figure in Chinese (Zhu et al., 2022), or figures with the same name have deviating definitions in different languages (Wang et al., 2022). We think these inconsistent definitions cause problems when figures are binary annotated, e.g., as *present* or *not present*, because figures deviate in their salience.

Suggested solution: Consulting different sources before approaching the figure detection task is a good way to start. More importantly, we think that the detection of rhetorical figures should not be considered a binary classification task. We suggest a ranking scheme (e.g., continuous values) tailored to every figure based on its salience and conspicuousness or how many properties from the textual definitions are fulfilled. Rankings for rhetorical figures have already proven to be useful (Dubremetz and Nivre, 2015; Troiano et al., 2018; Zhang and Wan, 2021). For example, in the case of antithesis, sentences that contain both parallelism and antonyms can be ranked higher than sentences with antonyms and no parallelism. Nevertheless, it is necessary to remember that annotations with continuous values are often more unreliable than binary annotations (Bagdon et al., 2024). To avoid this problem, we suggest a comparison-based annotation, e.g., best-worst scaling. This method already performed well in emotion intensity annotation with language models (Bagdon et al., 2024), which we consider related to rhetorical figure annotation.

2.2 Defining Boundaries and Intentional Usage

Another problem that most researchers encountered is the definition of the boundary in which to look for figures. As figures can span over multiple sentences, paragraphs, or the whole text, it is important to define where to start and where to end. If a repetition of two words is too far apart, it is not recognized as salient anymore by humans, while automatic parsers detect the repetition (Strommer, 2011). Properly defining boundaries determines the success of rhetorical figure detection (Strommer, 2011). An additional challenging aspect is to decide whether the figure is accidentally or intentionally present. Especially repetitions can occur without a rhetorical purpose (Strommer, 2011; Dubremetz and Nivre, 2015). This leads to the problem that annotators often cannot agree if it is actually a figure and which figure it is, decreasing the agreement between annotators and the reliability of the annotation itself. Strommer (2011) describes that in the case of his 156 instances, the annotators agreed only on two of them to be an intentional anaphora. Troiano et al. (2018) also mention that they had diverse annotations in their hyperbole dataset.

Suggested solution: It is important for future dataset construction to not only include one or two sentences containing the figure itself but also to consider larger text chunks. A ranking scheme mentioned in Section 2.1 can also help with expressing the salience of figures and deciphering between a rather accidental or intentional use.

2.3 Lack of Data/Datasets

When considering popular figures such as metaphor, irony, or sarcasm, researchers can profit

from users that tag their posts in social media, e.g., `#sarcasm` (Ranganath et al., 2018). This makes it easier to compile larger annotated datasets. Other figures that also play an important role in persuasive communication, but are not that present in the minds of the average social media users are often neglected. It can be more difficult to find instances of those figures (Dubremetz and Nivre, 2015). Another problem when creating datasets could be an inherent bias, as only sentences with salient rhetorical figures are chosen. This means that edge cases, where it is arguable whether it is a rhetorical figure or not (see Section 2.2,) are not included in the dataset.

Suggested solution: Generative large language models (LLMs) can help create sentences containing rhetorical figures. A downside is, however, that the LLM was probably pre-trained on data in which rhetorical figures other than metaphor are not explicitly annotated, making the generation more difficult. Furthermore, one must be aware of the vicious cycle that LLMs can only generate sentences with rhetorical figures they already know. If the LLM does not know the construction rules of a rhetorical figure, it cannot reliably generate sentences containing the figure. It is still necessary that human annotators oversee the process, as in Chakrabarty et al. (2022) where three annotators verified the texts generated by GPT-3. Another solution to collect more annotated data is to develop platforms where users can submit instances of rhetorical figures in a game-like scenario (Kühn and Mitrović, 2023).

2.4 Imbalanced Datasets and Deceptive Performance Metrics

If datasets for rhetorical figures are constructed, researchers like Bhattasali et al. (2015); Dubremetz and Nivre (2017); Ranganath et al. (2018); Adewumi et al. (2021); Kühn et al. (2023) face highly imbalanced datasets, i.e., the majority of data points are not a rhetorical figure. Using then accuracy as a performance metric can be highly deceptive. In a dataset where 99 % of instances are not a rhetorical figure, a model that consistently predicts a particular class will achieve a classification accuracy of 99 %. Also, other metrics such as precision and recall have to be considered carefully as their problems became obvious in the work of Gawryjolek (2009) and Java (2015). Furthermore, with only a few datasets with positive examples of rhetorical figures, it is more difficult to train machine models on (Dubremetz and Nivre, 2017; Zhang and Wan, 2021) or fine-tune language models to achieve better performance.

Suggested solution: Augmentation techniques or over- or undersampling can help decrease the imbalance. LLMs can also help create more sentences containing rhetorical figures. Evaluation metrics have to be chosen wisely.

2.5 Not Including Ontologies

Formal domain ontologies of rhetorical figures have the goal of overcoming the problem of inconsistent definitions and spellings (see Section 2.1). There exist ontologies such as the English RhetFig ontology (Harris et al., 2017), the Ploke (Wang et al., 2021), the Serbian Retfig (Mladenović and Mitrović, 2013), the German GRhOOT (Kühn et al., 2022), and a multilingual ontology (Wang et al., 2021). They all represent rhetorical figures in the form of classes and relations, describing how they are constructed, where they appear, and which cognitive effects they have. However, we realized that none of the investigated approaches use those ontologies.

Suggested solution: We suggest including those ontologies in the process of detecting rhetorical figures. We are confident that those ontologies can help improve detection rules or help annotators achieve higher agreement. Further applications are also possible when the ontologies are combined with LLMs, especially in a retrieval augmented generation (RAG) system (Lewis et al., 2020), where the context of an LLM is enhanced with rhetorical knowledge from the ontologies. In addition, it is possible that the data generation and annotation capabilities of LLMs are improved, too.

2.6 Missing Context

Rhetorical figures are often implicit, subtle, and can only be understood with context knowledge (Lawrence et al., 2017; Ranganath et al., 2018; Troiano et al., 2018). Some figures can even be used both in a figurative and literal meaning, e.g., rhetorical questions, which are syntactically not different from regular questions (Ranganath et al., 2018), or hyperboles that can also have both a literal and a figurative meaning, depending on context: Troiano et al. (2018) give the example of *"It took ages to build the castle"* vs. *"It took ages to build the castle. After a few minutes, my little*

brother had already destroyed it!" For an efficient detection of rhetorical figures, it is important to understand the semantics, syntax, and pragmatics (Medková, 2020).

Suggested solution: For the detection of most figures, it is necessary to include sentences/paragraphs pre- and succeeding the sentence of interest for context knowledge. In addition, LLMs can help to resolve contextual ambiguities and syntactic knowledge about figure formation can be extracted from ontologies.

2.7 Focus on Rule-based Methods

While deep-learning methods are already implemented successfully for the detection of metaphors (Bizzoni et al., 2017; Bizzoni and Ghanimifard, 2018), we observe a focus on rule-based approaches for lesser-known figures. We are certain that approaches based on LLMs will massively increase in the future and may overcome the performance of current state-of-the-art rule-based approaches. Zhu et al. (2022) experience lower performance with rule-based approaches for various rhetorical figures. They note that a complex task such as the detection of rhetorical figures cannot be solved by identifying "shallow and obvious patterns." Similar to the field of mail spam detection, there is no use in creating lists with known rhetorical figures, as humans are creative and come up with new metaphors or analogies. From the over 40 papers we investigated, the authors implemented 87 different detection techniques for various figures (see Table 1). 68.97 % are rule-based approaches, whereas only 27.59 % are model-based or deep learning approaches. Only one approach from Kühn et al. (2024) combines a rule-based with a model-based approach to detect the figure antithesis.

Suggested solution: We suggest using LLMs. However, as even powerful language models show a decreased performance in the understanding of rhetorical figures compared to humans (Liu et al., 2022), we believe that the combination of LLMs and rule-based approaches can be fruitful. For example, the presence of figures with perfect lexical repetition can be better verified by rules.

2.8 Focus on English

Existing datasets of rhetorical figures mainly contain sentences in English. This makes it even more challenging to investigate rhetorical figures in other languages. A direct translation from English into another language is often not possible without losing the original form of the rhetorical figure, especially if it contains syntactical aspects (Kühn et al., 2023). Another problem is that English is uncased and has neither a grammatical gender nor inflection. Some figures based on a change in inflection (such as polyptoton) appear less frequently than in languages with strong inflection, e.g., German (Fahnestock, 2002). Furthermore, English does not have separable verbs. These are verbs where the prefix is split from the main verb. This can create repetitions without a rhetorical purpose: "Wir fingen an, an danach zu denken" (*"We began to think about what comes after."*), where "an" is repeated while referring to different concepts. This highlights once again why rule-based approaches can fail (see Section 2.7). Table 2 shows that 66.81 % of the investigated approaches focus on rhetorical figures in English. When authors consider figures in multiple languages (e.g., Hromada (2011) investigates English, Latin, French and German, or Lagutina et al. (2019) in Russian and English), we counted them individually for every language.

Approach category	#Approaches	In Percent
Rule-based	**60**	**68.97 %**
Model-based	24	27.59 %
Rule-& Model-based	1	1.15 %
Unknown	2	2.30 %

Table 1: Distribution of the approach categories over the 86 approaches.

Language	#Approaches	In Percent
English	**74**	**69.81 %**
German	10	9.43 %
Russian	8	7.55 %
French	4	3.77 %
Latin	4	3.77 %
Chinese	3	2.83 %
Czech	2	1.89 %
Japanese	1	0.94 %

Table 2: Distribution of languages.

The focus on English leads to another problem. Most NLP tools are developed for English. According to the #BenderRule (Bender, 2019), it is "undesirable" that language technologies are only developed for one or two popular languages. This

leads to a vicious cycle: The more tools are tailored to the English language, the more researchers only focus on the detection of rhetorical figures in English. Because appropriate tools are lacking for other languages, identifying rhetorical figures is more challenging and might be neglected. As we mentioned previously, translating the data into English to be able to use existing tools is not an option.

The focus on English already created inequalities regarding model creation, leading to a lower acceptance rate at NLP conferences for papers not dealing with English (Søgaard, 2022).

Suggested solution: This is not an easy challenge to overcome as it affects the entire discipline of NLP. Nevertheless, we would like to encourage researchers to perform their work in languages other than English. Also, we think that it is necessary to reward research that focuses on other languages. Another solution can be the creation of adequate tools in multiple languages.

2.9 Neglecting Cognitive Effects

Another point of critique is that research about rhetorical figures focuses on detection but often forgets about the cognitive effects of the figures (Mitrović et al., 2020). This seems to be especially the case when approaching rhetorical figures from a computational perspective, as it is already challenging to implement detection algorithms. Often, the interpretation of the figure in the given context is then neglected. However, as every form of a figure has a certain function (Givón, 1995), it is important to not only identify figures but also interpret their usage.

Suggested solution: It is important to have a holistic look at the task of rhetorical figure detection. We suggest including explanations of what the usage of a certain figure in a given context actually means and analyzing which emotions are created for readers and listeners.

2.10 Lack of Interdisciplinary Efforts

Dealing with rhetorical figures is a highly interdisciplinary task that includes all obstacles from other disciplines. From an NLP perspective, rhetorical figures are not only syntactic constructions. They also include semantic features, have a transferred meaning, or depend on sound. For certain figures, it is necessary to identify negation, which is still a hard task in NLP. As rhetorical figures appear in all areas of our daily lives, we encounter them in the domain of advertising, politics, sentiment analysis, hate speech, machine translation, and many more. Rhetorical figures are also interesting for neuroscience in terms of their effect on the human brain. Green (2021) showed how rhetorical figures are applied in environmental arguments. Fahnestock (2002) highlights the importance of rhetorical figures in disciplines such as biology or chemistry, among others. If those fields understand rhetorical figures better, they can communicate more effectively with convincing arguments. In the field of law, there is a growing body of work devoted to argumentation and deciphering the effects of figures on persuasiveness (Al Zubaer et al., 2023).

Suggested solution: Researchers coming from different disciplines should join forces to build a holistic view of rhetorical figures, their purpose, function, and effect. Computer scientists and linguists can benefit from one another especially. Other disciplines can also profit from collaboration and open up new areas of research.

3 Conclusion

Our comprehensive review of over 40 papers highlights the prevalent challenges in computationally detecting rhetorical figures. As each rhetorical figure plays a crucial role in our daily communication, we urge researchers to tackle the presented challenges. When we can understand the non-literal and subtle meaning of rhetorical figures, we can improve existing systems and better understand language. In the future, we would like to see some of the suggestions implemented. Furthermore, we aim to inspire researchers to also focus on the detection of lesser-known figures.

Acknowledgements

The project on which this report is based was funded by the German Federal Ministry of Education and Research (BMBF) under the funding code 01|S20049. The authors are responsible for the content of this publication. Furthermore, we thank the anonymous reviewers for their valuable feedback, insightful comments, and new perspectives that enhanced the quality of our work.

References

Tosin P Adewumi, Roshanak Vadoodi, Aparajita Tripathy, Konstantina Nikolaidou, Foteini Liwicki, and Marcus Liwicki. 2021. Potential idiomatic expression (pie)-english: Corpus for classes of idioms. *arXiv preprint arXiv:2105.03280*.

Abdullah Al Zubaer, Michael Granitzer, and Jelena Mitrović. 2023. Performance analysis of large language models in the domain of legal argument mining. *Frontiers in Artificial Intelligence*, 6.

Mohammed Alliheedi and Chrysanne Di Marco. 2014. Rhetorical figuration as a metric in text summarization. In *Advances in Artificial Intelligence: 27th Canadian Conference on Artificial Intelligence, Canadian AI 2014, Montréal, QC, Canada, May 6-9, 2014. Proceedings 27*, pages 13–22. Springer.

Christopher Bagdon, Prathamesh Karmalker, Harsha Gurulingappa, and Roman Klinger. 2024. " you are an expert annotator": Automatic best-worst-scaling annotations for emotion intensity modeling. *arXiv preprint arXiv:2403.17612*.

Emily Bender. 2019. The #benderrule: On naming the languages we study and why it matters. *The Gradient*, 14.

Shohini Bhattasali, Jeremy Cytryn, Elana Feldman, and Joonsuk Park. 2015. Automatic identification of rhetorical questions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 743–749.

Yuri Bizzoni, Stergios Chatzikyriakidis, and Mehdi Ghanimifard. 2017. "deep" learning : Detecting metaphoricity in adjective-noun pairs. In *Proceedings of the Workshop on Stylistic Variation*, pages 43–52, Copenhagen, Denmark. Association for Computational Linguistics.

Yuri Bizzoni and Mehdi Ghanimifard. 2018. Bigrams and bilstms two neural networks for sequential metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 91–101.

Gideon O Burton. 2007. The forest of rhetoric. *Silva Rhetoricae*, 6.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. Flute: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159.

Marie Dubremetz and Joakim Nivre. 2015. Rhetorical figure detection: The case of chiasmus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 23–31.

Marie Dubremetz and Joakim Nivre. 2017. Machine learning for rhetorical figure detection: More chiasmus with less annotation. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 37–45.

Jeanne Fahnestock. 2002. *Rhetorical figures in science*. Oxford University Press on Demand.

Simona Frenda, Viviana Patti, and Paolo Rosso. 2023. When sarcasm hurts: Irony-aware models for abusive language detection. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 34–47. Springer.

Martha Gavidia, Patrick Lee, Anna Feldman, and Jing Peng. 2022. Cats are fuzzy pets: A corpus and analysis of potentially euphemistic terms. *arXiv preprint arXiv:2205.02728*.

Jakub Jan Gawryjolek. 2009. Automated annotation and visualization of rhetorical figures. Master's thesis, University of Waterloo.

Mengshi Ge, Rui Mao, and Erik Cambria. 2023. A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. *Artificial Intelligence Review*, pages 1–67.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 470–478.

Talmy Givón. 1995. Isomorphism in the grammatical code. *Iconicity in language*, pages 47–76.

Nancy L Green. 2021. Some argumentative uses of the rhetorical figure of antithesis in environmental science policy articles. In *Proceedings of the Workshop on Computational Models of Natural Argument*, pages 85–90.

Randy Allen Harris, Chrysanne Di Marco, Ashley Rose Mehlenbacher, Robert Clapperton, Insun Choi, Isabel Li, Sebastian Ruan, and Cliff O'Reilly. 2017. A cognitive ontology of rhetorical figures. *Cognition and Ontologies*, pages 18–21.

Randy Allen Harris, Chrysanne Di Marco, Sebastian Ruan, and Cliff O'Reilly. 2018. An annotation scheme for rhetorical figures. *Argument & Computation*, 9(2):155–175.

Daniel Hromada. 2011. Initial experiments with multilingual extraction of rhetoric figures by means of perl-compatible regular expressions. In *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, pages 85–90.

James Java. 2015. *Characterization of prose by rhetorical structure for machine learning classification*. Ph.D. thesis, Nova Southeastern University.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22.

M Karp, N Kunanets, and Y Kucher. 2021. Meiosis and litotes in the catcher in the rye by jerome david salinger: text mining. In *CEUR Workshop Proceedings*, volume 2870, pages 166–178.

Ramona Kühn, Jelena Mitrović, and Michael Granitzer. 2022. Grhoot: Ontology of rhetorical figures in german. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4001–4010.

Ramona Kühn and Jelena Mitrović. 2023. Multilingual domain ontologies of rhetorical figures and their applications. In *UniDive 1st General Meeting*.

Ramona Kühn and Jelena Mitrović. 2024. Status Quo der Entwicklungen von Ontologien Rhetorischer Figuren in Englisch, Deutsch und Serbisch. In *Book of Abstracts - DHd2024*. Zenodo.

Ramona Kühn, Jelena Mitrović, and Michael Granitzer. 2023. Hidden in plain sight: Can german wiktionary and wordnets facilitate the detection of antithesis? In *Proceedings of the 12th Global Wordnet Conference*, pages 106–116.

Ramona Kühn, Khouloud Saadi, Jelena Mitrović, and Michael Granitzer. 2024. Using pre-trained language models in an end-to-end pipeline for antithesis detection. In *Proceedings of the 14th Language Resources and Evaluation Conference*. European Language Resources Association.

Nadezhda Stanislavovna Lagutina, Kseniya Vladimirovna Lagutina, Elena Igorevna Boychuk, Inna Alekseevna Vorontsova, and Il'ya Vyacheslavovich Paramonov. 2019. Automated search of rhythm figures in a literary text for comparative analysis of originals and translations based on the material of the english and russian languages. *Modelirovanie i Analiz Informatsionnykh Sistem*, 26(3):420–440.

John Lawrence, Jacky Visser, and Chris Reed. 2017. Harnessing rhetorical figures for argument mining. *Argument & Computation*, 8(3):289–310.

Jens Lemmens, Ilia Markov, and Walter Daelemans. 2021. Improving hate speech type and target detection with hateful metaphor features. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 7–16.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Emmy Liu, Chen Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. *arXiv preprint arXiv:2204.12632*.

Helena Medková. 2020. Automatic detection of zeugma. In *RASLAN*, pages 79–86.

Jelena Mitrović, Cliff O'Reilly, Miljana Mladenović, and Siegfried Handschuh. 2017. Ontological representations of rhetorical figures for argument mining. *Argument & Computation*, 8(3):267–287.

Jelena Mitrović, Cliff O'Reilly, Randy Allen Harris, and Michael Granitzer. 2020. Cognitive modeling in computational rhetoric: Litotes, containment and the unexcluded middle. In *ICAART (2)*, pages 806–813.

Miljana Mladenović and Jelena Mitrović. 2013. Ontology of rhetorical figures for serbian. In *Text, Speech, and Dialogue*, pages 386–393, Berlin, Heidelberg. Springer.

Hoang Long Nguyen, Trung Duc Nguyen, Dosam Hwang, and Jason J Jung. 2015. Kelabteam: A statistical approach on figurative language sentiment analysis in twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 679–683.

Sunny Rai and Shampa Chakraverty. 2020. A survey on computational metaphor processing. *ACM Computing Surveys (CSUR)*, 53(2):1–37.

Suhas Ranganath, Xia Hu, Jiliang Tang, Suhang Wang, and Huan Liu. 2018. Understanding and identifying rhetorical questions in social media. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(2):1–22.

Felix Schneider, Björn Barz, Phillip Brandes, Sophie Marshall, and Joachim Denzler. 2021. Data-driven detection of general chiasmi using lexical and semantic features. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 96–100.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

Anders Søgaard. 2022. Should we ban english nlp for a year? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260.

Claus Walter Strommer. 2011. Using rhetorical figures and shallow attributes as a metric of intent in text.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686.

Enrica Troiano, Carlo Strapparava, Gözde Özbal, and Serra Sinem Tekiroğlu. 2018. A computational exploration of exaggeration. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3296–3304.

Byron C Wallace. 2015. Computational irony: A survey and new perspectives. *Artificial intelligence review*, 43:467–483.

Yetian Wang, Randy Allen Harris, and Daniel M Berry. 2021. An ontology for ploke: Rhetorical figures of lexical repetitions. In *JOWO*.

Yetian Wang, Ramona Kühn, Randy Allen Harris, Jelena Mitrović, and Michael Granitzer. 2022. Towards a unified multilingual ontology for rhetorical figures. In *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. Valletta, Malta: SCITEPRESS-Science and Technology Publications*, pages 117–127.

Hamed Yaghoobian, Hamid R Arabnia, and Khaled Rasheed. 2021. Sarcasm detection: A comparative study. *arXiv preprint arXiv:2107.02276*.

Yunxiang Zhang and Xiaojun Wan. 2021. Mover: Mask, over-generate and rank for hyperbole generation. *arXiv preprint arXiv:2109.07726*.

Dawei Zhu, Qiusi Zhan, Zhejian Zhou, Yifan Song, Jiebin Zhang, and Sujian Li. 2022. Configure: Exploring discourse-level chinese figures of speech. *arXiv preprint arXiv:2209.07678*.

A Appendix

Following Table 3 shows which figures were considered in the papers and how often they are investigated. If multiple figures are investigated in one paper, we counted them multiple times.

Figure	# Approaches
Alliteration	2
Anadiplosis	5
Anaphora/Epanaphora	7
Antimetabole	**10**
Antithesis	5
Assonance	1
Chiasmus	8
Conduplicatio	1
Diacope	1
Dirimens copulatio	1
Duality	1
Dysphemism	1
Epanalepsis	3
Epanaphora	1
Epiphora/Epistrophe	6
Epizeuxis	4
Euphemism	3
Eutrepismus	1
Hyperbole	4
Isocolon	2
Litotes	3
Meiosis	1
Metonymy	6
Oxymoron	3
Parallelism	3
Personification	1
Ploke/Ploce	2
Polyptoton	4
Polysyndeton	3
Quote	1
Repetition	1
Rhetorical question	3
Symploke	2
Synaesthesia	1
Zeugma	1

Table 3: Frequency of appearance for each figure in the investigated papers.

Guidelines for the Annotation of Deliberate Linguistic Metaphor

Stefanie Dipper and **Adam Roussel** and **Alexandra Wiemann**
and **Won Kim** and **Tra-My Nguyen**
Department of Linguistics
Fakultät für Philologie
Ruhr-Universität Bochum
firstname.lastname@ruhr-uni-bochum.de

Abstract

This paper presents guidelines for the annotation of deliberate linguistic metaphor. Expressions that contribute to the same metaphorical image are annotated as a chain along with a semantically contrasting expression of the target domain, which helps to make the domain contrast inherent to metaphor more explicit. So far, a corpus of ten TEDx talks with a total of ca. 20k tokens has been annotated according to these guidelines. 1.35% of the tokens are deliberate metaphorical expressions according to our guidelines, which shows that our guidelines successfully identify a significantly higher proportion of deliberate metaphorical expressions than previous studies.

1 Introduction

In conceptual metaphor theory (Lakoff and Johnson, 1980, CMT), the idea of a conceptual metaphors refers to the understanding of one conceptual domain in terms of another. This involves taking an expression from a literal, usually more concrete, source domain and transferring it onto a target domain in order to shape our understanding of this target domain concept in some way. This cross-domain mapping effects a transfer of properties of the source domain to the target domain, as the source domain is reinterpreted.

Such conceptual metaphors can be implemented in any of a number of ways, but one common medium for conceptual metaphors is language. Linguistic metaphor is often associated with certain properties: there is usually some kind of semantic mismatch between certain words in a sentence, which triggers the reinterpretation of the metaphorically-used words. According to Hanks (2013), such mismatches, which he calls 'exploitations', stem from a deliberate departure from an established pattern of normal word use. For instance, in example (1), the subject *Bodenschätze* 'natural resources' (lit. 'ground-treasure'), is normally used with container expressions referring to soil or huge shipping containers, so referring to people's minds as containers deviates from the norm. As a consequence, *Bodenschätze* is reinterpreted as the valuable content of minds, such as intelligence or creativity.

(1) Das kann sich ein Land, dessen Bodenschätze in den Köpfen unserer Bevölkerung stecken, nicht leisten.
'A country whose natural resources are in the minds of our population cannot afford this.'

There is a related notion that metaphoric expressions can be observed to stand out in their immediate context, that it will be surprising to find language pertaining to product packaging in the context of a poetry slam for instance, as in example (2), and this element of surprise can also trigger the reinterpretation of expressions that are intended metaphorically.

(2) Du bist so vakuumverpackt, so in deiner Komfortzone versackt.
'You are so vacuum-packed, so stuck in your comfort zone.'

In order to learn more about the linguistic dimensions of metaphor and the relationship between linguistic metaphors and their context, we annotate whole texts and will eventually expand our corpus to encompass a variety of text genres.

Previous annotation efforts that have covered the annotation of complete texts, most notably the VUA Metaphor Corpus (Steen et al., 2010), often used guidelines oriented broadly towards the annotation of all kinds of metaphor, and accordingly their datasets consist mostly of conventionalized metaphors, of which speakers are mostly unaware and which don't serve a particular discourse-communicative purpose. In contrast, our guidelines

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 53–58
June 21, 2024 ©2024 Association for Computational Linguistics

are focussed more squarely on *deliberate metaphors* in the sense of Steen (2008), which play an important role in a discourse and of which speakers and listeners are likely aware.

The contributions of this paper are: (i) annotation guidelines for identifying deliberate metaphor; (ii) an annotated corpus of TEDx talks with 20k tokens, which is made freely available.[1]

2 Related work

The first work on the annotation of metaphors in texts comes from an interdisciplinary group of researchers who define a Metaphor Identification Procedure (MIP) to recognize metaphorically used expressions in texts (Pragglejaz Group, 2007). The MIPVU guidelines went beyond MIP by also taking into account explicit comparisons or similes (Steen et al., 2010). In both approaches, the annotator must first determine the contextual meaning of a word, i.e. the current meaning in the text, and then use a reference lexicon to check whether there is a 'more basic' literal meaning (e.g. a more concrete meaning). If the contextual meaning is in contrast to the literal meaning, but is at the same time in some way similar and can be understood in comparison to it, the word is labeled as 'MRW' (metaphor-related word). The guidelines are designed as to identify all metaphors, including conventionalized ones.

Steen et al. (2010) annotated the VUAMC (VU Amsterdam Metaphor Corpus) according to MIPVU. The corpus contains 190k words and consists of fragments from four registers of the BNC-Baby corpus (academic texts, conversation, fiction, and news texts). 86% of the words are clearly non-metaphorical and 13% are clear MRWs, and 1% are borderline cases. The highest proportion of MRWs is found among prepositions. In different studies on inter-annotator agreement (IAA), Steen et al. (2010) achieved Fleiss' κ between 0.70 and 0.96 (with texts in English and Dutch).

Deliberate metaphor The DMIP guidelines (Deliberate Metaphor Identification Procedure) aim at excluding dead and conventionalized metaphor (Reijnierse et al., 2018). Deliberate metaphors are those that are intentionally used as metaphor and draw attention to the cross-domain mapping, as opposed to conventionalized metaphors where no such processes take place. According to the DMIP guidelines, only *potentially* deliberate metaphors can be identified sensibly. Rather than providing detailed and specific criteria for the identification of deliberate metaphor, Reijnierse et al. (2018, p. 137) give the following instruction: "Determine whether the source domain of the MRW is part of the referential meaning of the utterance in which the MRW is used." However, they mention some typical indicators of deliberate metaphor, including novel metaphor and extended metaphor, consisting of multiple words that relate to the same metaphor, as well as direct metaphor, signaled by lexical cues such as *as* or *like*, or topic-triggered metaphor, where lexis related to the overall topic of the text is used metaphorically.

The DMIP guidelines have been tested on pre-marked MRWs of a set of selected VUAMC sentences, resulting in Cohen's κ between 0.70 and 0.73 (with 129 and 130 pre-marked MRWs from VUAMC, respectively). In the two datasets, 11.6% and 9.2% of the MRWs are annotated as deliberate.[2] The size of the data sets is not specified in the paper, though. Since around 11.1% of all tokens in VUAMC are MRWs, it can be estimated that deliberate metaphor accounts for approximately 1.2% of all tokens.

Beigman Klebanov and Flor (2013) present an annotation protocol for the identification of "metaphorical expressions that are noticeable and support the author's argumentative moves" (p. 15). The guidelines do not specify detailed criteria for identification, but rather describe metaphors in general terms: "Generally speaking, a metaphor is a linguistic expression whereby something is compared to something else that it is clearly literally not, in order to make a point." (p. 14). A total of 116 test-taker essays, discussing the role of electronic media for communication, are annotated with 55k tokens (κ = .575). On average, the two student annotators marked 4.86% of all tokens as metaphorical according to the guidelines; the union set, which serves to account for the fact that disagreement is often due to attention slips (Beigman Klebanov et al., 2008), comprises 6.83% of all tokens. The evaluation shows that verbs in particular are used metaphorically disproportionately often.

Novel metaphor Do Dinh et al. (2018) investigate novel metaphors (which constitutes a subset of deliberate metaphor). Their work is based on the VUAMC. For all content-word MRWs (i.e. excluding auxiliaries and prepositions), they an-

[1] https://gitlab.rub.de/comphist/figlang2024

[2] The annotated MRWs are freely available at https://osf.io/c8bxs.

notate whether the metaphor is novel, i.e. non-conventionalized. Crowd workers receive random samples with four MRWs each and annotate which of these is the most novel and which is the most conventionalized (no IAA calculable). The proportion of novel metaphors (353) of all tokens (240k) ranges from 0.04–0.26% across the four registers.

Parde and Nielsen (2018) also investigate novel metaphor and annotate MRWs from the VUAMC, similar to Do Dinh et al. (2018). However, the crowd workers only annotate selected word pairs that consist of content words (or a personal pronoun), at least one of which is an MRW and which are syntactically linked. The annotations consist of gradual scores, from 0 'not metaphoric' to 1 for 'low metaphor novelty' up to 3 for 'high metaphor novelty'. IAA was calculated between trained annotators with κ scores of 0.435, and, with relaxed constraints, 0.897 (on 3k instances). In total, the corpus contains more than 18k annotated word pairs, however, the exact proportion of novel metaphor (with scores 2 or 3) is not specified in the paper.[3]

Alnajjar et al. (2022) annotate metaphors in 27 YouTube videos of the start-up domain. The criteria for annotation are kept very simple: A word is considered a metaphor if its meaning is not literal, if the meaning is not listed in the lexicon (i.e. it is not a conventionalized metaphor), or if it is not meant sincerely but sarcastically. However, if the metaphor includes several words, it is considered an idiom and annotated, even if it is conventionalized (e.g. *give it a shot*). The two expert annotators annotate both vehicle (the metaphorical expression from the source domain) and tenor (the expression from the target domain) – the criteria for tenor, however, remain unclear, as these are typically interpreted literally. No IAA is reported. In total, 672 metaphorical tokens have been annotated, among them 45% novel metaphors, which roughly seem to correspond to 0.23% of all tokens.

Resources for German To the best of our knowledge, there are no annotated texts for German available. Herrmann et al. (2019) adapt MIPVU to German. They calculate IAA for set of 559 sentences, obtaining Fleiss' $\kappa = 0.71$. The analyzed corpus of 20k sentences is not available.

Egg and Kordoni (2022, 2023) also adopt the MIPVU guidelines, but extend them to include the annotation of elements in the context of the metaphorical expression that trigger the metaphorical meaning, which they call 'background'. They also determine the conventionality of an MRW: An MRW is conventionalized if its meaning is listed in the lexicon. Using INCEpTION (Klie et al., 2018), they annotate a corpus with five different registers, which should ultimately contain 150k words. In Egg and Kordoni (2023) an IAA of Krippendorff's $\alpha = 0.89$ is reported, but it is unclear on which data this was calculated. In their data, the conventionalized MRWs have a proportion of 4–15% and the non-conventionalized MRWs of 0.01–0.29% (again, the size of the underlying data is unclear). The guidelines and the corpus are not yet available.

3 Guidelines

We are interested in deliberate metaphor in German-language data. In most studies, deliberate MRWs represent a very small proportion of all tokens, less than 0.3%. The study by Beigman Klebanov and Flor (2013) clearly deviates from this with proportions of 4.86 and 6.83%, but it is unclear whether this is due, for example, to the open guidelines or to the text type or to the fact that the texts come from learners.

Our aim is to produce guidelines with specific criteria, offering supportive guidance for the annotators, so that the proportion of overlooked cases due to attention slips is minimized and we are able to identify more instances of deliberate MRWs than has been the case in previous studies. Our criteria, detailed in the following, are based on those for deliberate MRWs in Reijnierse et al. (2018).

Deliberate An MRW is considered deliberate if the metaphorical image is new or if the MRW used for an known metaphorical image is unusual and innovative. Alternatively, the MRW can be deliberate because it is marked in some way, e.g. if it occurs in a construction that is normally used in the active voice but now occurs in the passive voice, if the MRW is typographically emphasized, e.g., by italics or quotation marks, or if it stands out because it also appears in the title of the text.

For instance, example (3) contains a well-known metaphor, *ein Strauß an Forderungen* 'a bouquet of demands'. However, this established metaphor is expanded and modified by the adjective *bunt* 'colorful' and the verb *binden* 'to bind', so we consider it a deliberate metaphor.

[3]The data are available at https://computerscience.engineering.unt.edu/labs/hilt/resources.

(3) einen Strauß bunter Forderungen binden
'tying a bouquet of colorful demands'

The label **grey area** is used when an MRW shows characteristics of both deliberate and conventionalized metaphors.

Revitalized A subset of conventionalized expressions is also relevant here, namely revitalized usages: A conventionalized MRW can appear in a new light in a particular context, e.g. when a deliberate MRW that refers to the same image occurs in the immediate vicinity, so that the conventionalized expression could plausibly have been chosen deliberately rather than arbitrarily or a listener might plausibly perceive it in this way. The otherwise conventionalized expression is thus considered 'revived' or revitalized.

Anchor Usually, the annotation process begins when an annotator, in the course of reading through a text, notices some unusual or conspicuous combination of words, which impression is often the result of a domain clash or a kind of semantic incompatibility between them. One of the words, corresponding to the source domain, then needs to be re-interpreted metaphorically, while the other, corresponding to the target domain, is taken literally. We label this second expression the 'anchor', as this is the expression that 'anchors' the metaphorical image in reality. In example (3) above, the anchor is *Forderungen* 'demands', because this is the expression that is intended literally – the statement is ultimately really about 'demands' of some kind and not flowers.

In addition, we mark **flags** (Steen et al., 2010) indicating a comparison, e.g. expressions such as *wie* 'like' or *sozusagen* 'so to speak'.

MRW chains A metaphorical image is often verbalized by several MRWs and enriched with details. All MRWs that contribute to the same metaphorical image are annotated together and linked as a chain annotation, that is, an unordered set of token spans.

Of these MRW expressions, one can often be considered central, insofar as it best characterizes or names the metaphorical image. In example (3) above, *Strauß* 'bouquet' is the central expression, and *binden* 'tie' and *bunt* 'colorful' also contribute to the metaphorical image.

This central expression is the one that is given a specific label in the annotation that characterizes the whole metaphorical instance, while all of the other MRW expressions in the chain are only marked with the general label 'MRW'. Such specific labels are 'deliberate', 'grey area', 'revitalized', and 'extended'.

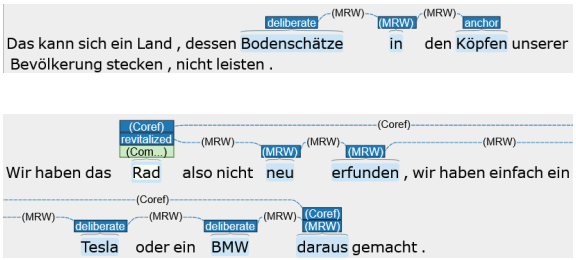


Figure 1: Metaphor annotations in INCEpTION for examples (1) and (4).

Locality principle As a general rule, though not a strict requirement, the anchor should be determined in such a way that there is a direct syntactic dependency relation between the anchor and the central MRW, e.g. an MRW verb with its subject as the anchor, or an MRW noun with its modifier as the anchor. Very often a suitable anchor is easily found among the syntactically close expressions, since this direct relation is what allows the two expressions to better clash semantically.

Due to this close syntactic relationship between the MRWs and the anchor, an MRW chain usually only involves one clause or at most one sentence.[4]

Extended If a metaphorical image extends over several sentences, e.g. because it is introduced and then elaborated in subsequent sentences, we annotate the 'local' chains in each sentence individually. This can lead to there being no clear anchor in these subsequent sentences, therefore, in such cases, the MRWs may be annotated without an anchor. The otherwise deliberate MRW is then labeled 'extended'.

Examples (1) from above and (4) are two examples from our corpus. Figure 1 shows the annotation of these examples in INCEpTION.

(4) Wir haben das Rad also nicht neu erfunden, wir haben einfach ein Tesla oder ein BMW daraus gemacht.
'So we haven't reinvented the wheel, we've simply made a Tesla or a BMW out of it.'

[4]If a chain contains a pronoun, the pronoun is additionally linked to its antecedent via a coreference link. Such a chain is not extended to multiple sentences.

In example (1) there is a clear semantic clash between *Bodenschätze* 'natural resources' (= metaphorical, meaning 'intelligence', 'creativity', etc.) and *Köpfen* 'heads, minds' (= literal). Normally we only annotate nouns, verbs, and adjectives for metaphoricity. In this case, however, the preposition *in* 'in' plays an important role, so it is also annotated as MRW and included in the chain.

Example (4) contains what would ordinarily be considered a conventionalized metaphor: *das Rad neu erfinden* 'reinvent the wheel'. The second clause takes up part of the conventionalized image through the pronoun *daraus* 'out of it', which refers to *Rad* 'wheel' (see the coreference link in Fig. 1), and then elaborates upon this image, thereby revitalizing it. There is no clear clash in either clause and thus no anchor. However, the wider context makes it clear that *wir* 'we', the speakers, do not work in the automotive industry and are not talking about actually producing vehicles of any kind.

4 Data and results

Corpus The current corpus consists of the transcriptions of a total of ten TEDx Talks which were given in German on a range of different topics. Four of the texts have been doubly annotated and curated (see below). The texts are subject to licenses that permit free redistribution.[5] The corpus contains 20k tokens (averaging 1979.4 ±481.7 tokens per document). 1.35% of the tokens are deliberate metaphorical expressions, which shows that our guidelines successfully identify a significantly higher proportion of deliberate MRWs than previous studies. Of course, we cannot say what part the text type – TEDx Talks – has in this. Future work with annotations of other text types will have to show this.

Table 1 shows the distribution of the different types of MRWs. The numbers indicate the total number of chains per label, where a chain is categorized according to the label of its 'central MRW', such as 'deliberate', as well as the total number of tokens (including anchors) in each kind of chain.

Inter-annotator agreement Our validation corpus consists of four talks from the TEDx series. These texts were doubly annotated in their entirety according to our guidelines by two of the authors.

[5]The TEDx Talks are part of this playlist: https://www.youtube.com/playlist?list=PLzPiBVgAHXijVDasy92X6lZkl0DvFgSEg, accessed 2024-02-26. Our annotations are based on the subtitles extracted from these videos.

Type	# Chains	# Tokens
deliberate	85	264
extended	25	52
grey area	15	30
revitalized	20	46

Table 1: Distribution of different types of MRWs.

Our annotation scheme aims to capture more of the complexity of linguistic metaphor than previous annotation efforts, but the increased complexity of the annotation scheme brings with it both benefits and drawbacks. The information that is made available in the annotations is accordingly rich, but evaluating the reliability of the annotation effort becomes more difficult – in addition to the increased difficulty of the task itself.

To evaluate the reliability of the annotations, we employ the γ agreement measure (Mathet et al., 2015), specifically the implementation of Titeux and Riad (2021). This is a holistic agreement measure that determines the alignment between annotated units jointly with the measurement of disagreements in categorization.

We use a dissimilarity measure that takes into account the conceptual similarity between the category labels. For instance, metaphors that are labeled 'deliberate' can be considered more similar to those labeled 'grey area' than 'anchor'. As such, our dissimilarity measure will consider disagreement between 'deliberate' and 'grey area' to be less than between 'deliberate' and 'anchor'.

The γ statistic, calculated on these data with the parameters described above is 0.35, 0.43, 0.49 and 0.56 for each of the four evaluation texts, respectively. Especially considering the complexity of the phenomenon itself and the annotation scheme, these are promising results, which we expect could be improved in the future with further refinement of the annotation guidelines.

Acknowledgements

We are very grateful to the anonymous reviewers for their helpful and valuable comments.

Funded by/gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – SFB 1475 – Projektnummer 441126958.

References

Khalid Alnajjar, Mika Hämäläinen, and Shuo Zhang. 2022. Ring that bell: A corpus and method for multimodal metaphor detection in videos. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 24–33, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Beata Beigman Klebanov, Eyal Beigman, and Daniel Diermeier. 2008. Analyzing disagreements. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 2–7, Manchester, UK. Coling 2008 Organizing Committee.

Beata Beigman Klebanov and Michael Flor. 2013. Argumentation-relevant metaphors in test-taker essays. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 11–20, Atlanta, Georgia. Association for Computational Linguistics.

Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. 2018. Weeding out conventionalized metaphors: A corpus of novel metaphor annotations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1424, Brussels, Belgium. Association for Computational Linguistics.

Markus Egg and Valia Kordoni. 2022. Metaphor annotation for German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2556–2562, Marseille, France. European Language Resources Association.

Markus Egg and Valia Kordoni. 2023. A corpus of metaphors as register markers. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 220–226, Dubrovnik, Croatia. Association for Computational Linguistics.

Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations*. MIT Press.

Berenike Herrmann, Karola Woll, and Aletta Dorst. 2019. Linguistic metaphor identification in German. In Susan Nacey, Aletta Dorst, Tina Krennmayr, and Gudrun Reijnierse, editors, *Metaphor identification in multiple languages. MIPVU around the world*, pages 113–135. Benjamins, Amsterdam.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Natalie Parde and Rodney Nielsen. 2018. A corpus of metaphor novelty scores for syntactically-related word pairs. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22:1–39.

W. Gudrun Reijnierse, Christian Burgers, Tina Krennmayr, and Gerard J. Steen. 2018. DMIP: A method for identifying potentially deliberate metaphor in language use. *Corpus Pragmatics*, 2:129–147.

Gerard Steen. 2008. The paradox of metaphor: Why we need a three-dimensional model of metaphor. *Metaphor & Symbol*, 23(4):213–241.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification. From MIP to MIPVU*. Number 14 in Converging Evidence in Language and Communication Research. John Benjamins, Amsterdam.

Hadrien Titeux and Rachid Riad. 2021. pygamma-agreement: Gamma γ measure for inter/intra-annotator agreement in Python. *Journal of Open Source Software*, 6(62):2989.

A Sources of example sentences

The example sentences (1), (2) and (4) are taken from the following talks in the TEDx series:

- Example (1): *Schüler, Zukunft & Motivation*
- Example (2): *Vacuum-packed*
- Example (4): *Der Supermarkt der Zukunft*

Evaluating the Development of Linguistic Metaphor Annotation in Mexican Spanish Popular Science Tweets

Alec Sánchez-Montero alecm@comunidad.unam.mx
Gemma Bel-Enguix gbele@iingen.unam.mx
Sergio-Luis Ojeda-Trueba sojedat@iingen.unam.mx
Universidad Nacional Autónoma de México

Abstract

Following previous work on metaphor annotation and automatic metaphor processing, this study presents the evaluation of an initial phase in the novel area of linguistic metaphor detection in Mexican Spanish popular science tweets. Specifically, we examine the challenges posed by the annotation process stemming from disagreement among annotators. During this phase of our work, we conducted the annotation of a corpus comprising 3733 Mexican Spanish popular science tweets. This corpus was divided into two halves and each half was then assigned to two different pairs of native Mexican Spanish-speaking annotators. Despite rigorous methodology and continuous training, inter-annotator agreement as measured by Cohen's kappa was found to be low, slightly above chance levels, although the concordance percentage exceeded 60%. By elucidating the inherent complexity of metaphor annotation tasks, our evaluation emphasizes the implications of these findings and offers insights for future research in this field, with the aim of creating a robust dataset for machine learning in the future.

1 Introduction

Computational approaches to metaphor date back at least to the 1980s, when Artificial Intelligence (AI) and Natural Language Processing (NLP) became interested in the structure and mechanisms of the phenomenon (Shutova et al., 2013, Introduction). Since then, there has been growing interest among researchers in understanding how computers can effectively process both linguistic and non-linguistic metaphors. An instance of this progressive work has been the various workshops developed within the ACL, the NAACL and the EMNLP, since 2007, on metaphor, in particular, and on figurative language, in general.

Broadly speaking, automatic metaphor processing has branched into three fundamental areas: metaphor identification or detection, metaphor interpretation, and metaphor generation (Sánchez-Bayona, 2021). Usually regarded as the 'first step', metaphor identification aims to automatically recognize linguistic expressions that convey metaphorical meaning within a text. For this task, supervised machine learning techniques trained on annotated datasets are often used to distinguish linguistic patterns indicative of metaphor.

However, despite recent advances in Figurative Language Processing (FLP) focused on metaphor processing for English, the situation for the Spanish language is quite different. Although there are tools and models developed for automatic metaphor processing tasks in English, the same level of development and availability has not been reached for Spanish. More precisely, our literature review has revealed a substantial gap regarding NLP approaches to metaphor in Mexican Spanish tweets within the realm of science communication. This represents a novel and unexplored area of research, where the intersection of metaphorical language and science popularization discourse in the context of Mexican Spanish on X (previously Twitter) remains a largely unexplored territory. This study has the objective of analyzing the usage of linguistic metaphors through NLP techniques to provide an overview of metaphor identification and classification within short scientific communication posts on X in Mexico.

2 Preliminaries

2.1 Conceptual Metaphor Theory

According to conceptual metaphor theory (CMT), the fundamental feature of metaphor, as a cognitive phenomenon, lies in the conceptual mapping between source and target domains, i.e. a process whereby our understanding of concrete experiences is projected onto more abstract domains, facilitating comprehension and communication of complex

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 59–64
June 21, 2024 ©2024 Association for Computational Linguistics

ideas (Lakoff and Johnson, 1980). With this theoretical background in mind, it is vital to understand that linguistic metaphors are the linguistic expressions that manifest conceptual metaphors. In that regard, linguistic metaphors are made of language units, and they permeate various aspects of communication, from everyday conversation to specialized fields such as scientific communication, where they play a crucial role in shaping the way scientific concepts are articulated and understood by the public.

Furthermore, subsequent approaches within cognitive linguistics, such as conceptual blending challenged the notion of mapping as the sole foundation of the cognitive operation underlying metaphor. Instead, authors like Fauconnier and Turner (2008) hypothesize that metaphors are part of a continuum of mental operations (including metonymy and framing) where different domains are integrated into several networks within a mental space. In this integrated networks, specific features are selected for contrast, resulting in conceptual blending. Thus, conceptual metaphors are mental constructions resulting from the integration of multiple spaces and multiple mappings.

2.2 Metaphor Identification Procedure Vrije Universiteit

The Pragglejaz Group (2007) published the Metaphor Identification Procedure (MIP) to detect metaphorically used words in discourse. This method was later extended by Steen et al. (2010) in the Metaphor Identification Procedure Vrije Universiteit (MIPVU), which has served as a consistent methodology for detecting linguistic metaphor in authentic written texts through the annotation of metaphor related words (MRWs). According to MIPVU, MRWs encompass indirect, direct, and implicit types of metaphorical expressions. Additionally, MRW also include signals, which explicitly indicate the use of metaphor within the text and are characteristic of direct metaphor. Finally, within this framework, personification is recognized as a form of conceptual mapping that leads to metaphor.

MIPVU has proven particularly useful for annotating textual corpora across multiple languages (Nacey et al., 2019), as it allows for the integration of both semantic and contextual meaning in linguistic metaphor identification. Due to these properties, annotated datasets resulting from MIPVU (such as the VUAM corpus) (Steen et al., 2010) have been extensively used for training and evaluating machine learning models for automatic metaphor processing in FLP studies.

3 Related Work

Research and advances in automatic metaphor processing in Spanish remain scarce to this day. Specifically, if we focus solely on annotated corpora approaches for training supervised machine learning models, we have limited resources available. So far, the only publicly available annotated dataset on Spanish linguistic metaphors is the Corpus for Metaphor Detection in Spanish (CoMeta) (Sánchez-Bayona, 2021; Sánchez-Bayona and Agerri, 2022). This linguistic dataset represents the first documented effort to compile a collection of general domain texts for everyday metaphor detection in Spanish. CoMeta also marks the first adaptation of the MIPVU guidelines to this romance language, although during our literature review, it has not been possible to find the annotation guidelines used for the CoMeta.

For English, besides specifically trained models for metaphor processing such as MelBERT (Choi et al., 2021) and MIss RoBERTa WiLDe (Babieno et al., 2017), the work of Kim and Cho (2023) is remarkable, since it focuses on the generation of scientific metaphors. Using GPT-3 as a base model, these authors developed Metaphorian, a system that assists science writers in the creation of scientific metaphors. The Metaphorian system allows users to search for, add and modify scientific metaphors, which is a valuable creative assistance tool for formulating difficult-to-explain scientific concepts in terms of more familiar concepts.

4 Corpus Annotation

It is important to clarify that the primary subject of this research is linguistic metaphor annotation, according to the theoretical-methodological foundation of MIPVU, rather than conceptual metaphor analysis. Nonetheless, we have resorted to some CMT notions in the annotation guide, similar to the approach used by Zayed (2021), for didactic purposes in explaining metaphors to the annotators. Moreover, given our selection of popular science as the genre of interest, our annotation focuses on identifying both scientific metaphors and everyday or colloquial metaphors in the corpus, which is appropriate as these texts bridge the specialized realm of science and the colloquial domain of language.

In our annotation protocols, we center on identifying three types of linguistic metaphor across

popular science tweets: direct (DM), indirect (IM) and personification (PM). We define DM as an explicit comparison between the source domain and the target domain, characterized by three units: the source unit (label: 'md_fuente'), the target unit (label: 'md_meta) and the signal or cue (label: 'md_señal'). IM is understood as an implicit comparison between the source domain and the target domain, consisting of only one unit - the source unit (label: 'md_indirecta') - since the target unit is elided. Finally, we explain PM as the attribution of human or animate semantic features (label: 'personificador') to an inanimate or abstract object (label: 'pers_obj'). As far as we know, this is the only public effort to annotate linguistic metaphors specifically in Mexican Spanish. Both the original guidelines in Spanish and the English translation are accessible in our GitHub repository.

Following this, we have annotated a corpus of 3733 popular science communication tweets. This dataset comprises Mexican Spanish tweets from 19 science communicators on X based in Mexico, which were published from January 2020 to May 2023 and extracted with the X API.[1] It should be emphasized that the information on these user accounts was collected without specific preferences for a particular scientific domain, which led to a wide topic range in the corpus, from astronomy and general physics to genetics and history of science, among other areas.

We gathered a group of 4 native Mexican Spanish-speaking annotators to conduct an initial annotation of the entire corpus. These annotators are all undergraduate linguistics students, aged between 18 and 25, 1 female and 3 male. To enhance annotation, we opted for the Argilla platform as it supports token classification tasks on loaded datasets in Spanish. Subsequently, we divided the corpus into two halves, and assigned each half to a pair of annotators (1866 and 1867 tweets respectively), ensuring balanced coverage and consistency in the annotation process. This approach allowed us to efficiently distribute the workload while maintaining a rigorous and systematic approach to linguistic metaphor annotation.

We trained this group of annotators to apply 6 labels corresponding to the 3 metaphor types. Of these 6 labels, 3 belong to DM, 1 to IM and 2 to PM. Table 1 displays the distribution of such labels and their respective meanings in the context of the annotation, while Table 2 provides some examples of target annotations included in the guide for each metaphor type. For non-metaphorical tweets, annotators were instructed to save records without annotations, facilitating data collection and interpretation.

Metaphor Type	Label	Refers to
Direct	(1) `md_fuente` (2) `md_meta` (3) `md_señal`	Source domain unit Target domain unit Metaphor signal/cue
Indirect	(4) `m_indirecta`	Source domain unit, full scope of IM
Personification	(5) `pers_obj` (6) `personificador`	Personified object Linguistic unit giving human features to (5)

Table 1: Label classification by type of metaphor

During the annotation process, communication channels with annotators remained open for ongoing support. In addition to virtual meetings for annotation training, where we included both, examples of correct and incorrect annotations, all their questions were continuously answered and feedback on their work was provided. Naturally, annotators had access to the guidelines for annotation at all times.

5 Evaluation of the Annotation Task

5.1 Binary Classification

After completion of corpus annotation, we collected the data of the labels assigned to each tweet by the different annotators, using the Argilla library for Python. Next, we analyzed the annotated data to assess the level of agreement among annotators in a binary classification task, i.e. the distinction between metaphorical and non-metaphorical tweets. For this purpose, tabular data structures were created, in which we assigned the label '0' to records without annotations (representing non-metaphorical tweets) and '1' to tweets annotated with either DMs, IMs, PMs, or a combination of them. Using this methodology, we were able to calculate the percentage of inter-tag matches, indicating whether both annotators classified the tweet

[1] After its acquisition by Elon Musk, Twitter was renamed 'X' and the texts published on it became known as 'posts'. However, since at the time of data collection, this platform was called Twitter and its texts 'tweets', we have decided to preserve said term for this paper.

Metaphor Annotation Example	Observations
Además tienen una capa de tejido que refleja la luz, como un espejo detrás de la retina, llamada tapetum lucidum, que mejora su visión nocturna considerablemente.	Direct Metaphor: A "layer of tissue" (*capa de tejido*) is explicitly compared to a "mirror" (*espejo*) through the expression "like a" (*como un*)
Nuevas simulaciones numéricas sobre la distribución de materia en la telaraña cósmica	Indirect Metaphor: The structure of the universe is expressed in terms of a "cosmic web" (*telaraña cósmica*)
En 1986 surgió en Reino Unido una enfermedad que atacaba el sistema nervioso de las vacas.	Personification Metaphor: A "disease" (*enfermedad*) is described as an entity which can "attack" (*atacaba*) other things, as a human would

Table 2: Examples of metaphor annotation in the guidelines

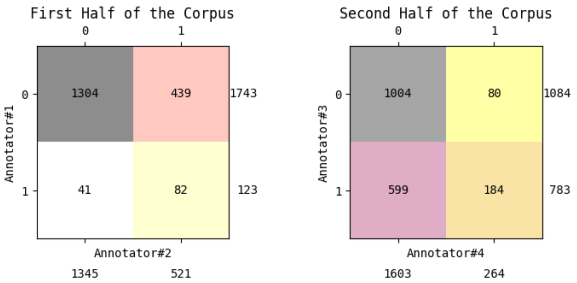


Figure 1: Binary classification of tweets in the corpus by halves

Corpus Half	Agreement (%)	Cohen's Kappa
First Half	74.27	0.16
Second Half	63.63	0.17

Table 3: Agreement Percentage and Cohen's Kappa Score by section of the corpus

as metaphorical or non-metaphorical, as well as the kappa coefficient of inter-annotator reliability.

For this study, we used Cohen's kappa (Cohen, 1960), since we evaluated the annotation of only 2 raters at the same time. The equation for this coefficient is: *K=(P0-Pe)/(1-Pe)*, where *P0* represents the observed agreement between annotators and *Pe* represents the agreement expected only by chance (Cohen, 1960). The use of this coefficient made it possible to calculate the possibility that the match occurred by chance and, as we will discuss later, contrasted with the percentage of matches between labels in this annotation. Afterwards, we extracted the labels in tuples to identify matches between labels and calculate the percentage of agreement.

As depicted by Figure 1, concerning results of the binary classification of the corpus by halves, the highest rate of inter-annotator agreement was in the non-metaphorical tweets, as both pairs of annotators agreed on 1304 tweets for the first half of the corpus and 1004 for the second half. In terms of tweets labeled as metaphorical by both annotators, the first pair identified 82 common tweets as metaphorical, while the second pair annotated 184 metaphorical tweets in common. Based on this information, we calculated that the percentage of agreement for the first half of the corpus was **74.27%**, while for the second half it was **63.63%**.

However, in terms of Cohen's kappa, the results were **0.16** for the first half and **0.17** for the second half. Both scores are considered as a "slight" agreement (Landis and Koch, 1977). While percentages of agreement are high, kappa scores remain low, in part, by the difference in the number of tweets that were identified as metaphorical in each half of the corpus. In both pairs of annotators, there was one annotator who recorded fewer metaphorical tweets compared to the other annotator. In the first pair of annotators, annotator 1 labeled only 123 tweets as metaphorical, while annotator 2 labeled a total of 521. In the second half of the corpus, annotator 4 labeled 264 tweets as metaphorical compared to the 783 by annotator 3. Annotators who identified fewer metaphorical tweets may have influenced the overall agreement score, as their annotations would have less impact on the kappa calculation. Furthermore, it is noteworthy that not all of the tweets metaphorically labeled by these annotators with fewer metaphorically labeled tweets were comprised of the other annotator's analogous tweets in each pair. Table 3 presents a synthesis of the data relating to the percentage of agreement and Cohen's kappa score for each half of the corpus.

Upon analysis of the low inter-annotator agreement rates, we have formulated some hypotheses. First, we believe that the annotators' lack of experience in explicitly identifying metaphors may have contributed to divergent interpretations and annotation errors. Second, despite the specific linguistic criteria in our guide for identifying metaphors, the interpretation of metaphorical expressions by human annotators is largely a subjective task. This

Otro estudio más reciente indica que el bostezo ayuda a enfriar el cerebro que , como las computadoras , puede sobrecalentarse.🥵🥵

A la fecha existen cinco redes sismológicas permanentes e independientes operando en la capital del país que recientemente conformaron la " Red Sísmica de la Ciudad de México "

Figure 2: Examples of commonly annotated metaphors with exact matches

means that we will have to rework a new version of the annotation guide with even clearer and more defined parameters that do not give rise to ambiguity in the reading.

5.2 Metaphor Annotation Matches

Despite the overall lower agreement rates observed in both kappa scores, there were some instances where both annotators identified the same tweets as metaphorical and even placed the same label on the same text sections. In the first half of the corpus, only 12 of the 82 common metaphorical tweets matched exactly. Similarly, in the second half of the corpus, of the 184 metaphorical tweets identified in common, only 27 showed complete agreement between annotators. In percentage terms, exact matches constitute **14.6%** of the total number of metaphorical tweets for both corpus halves. Figure 2 shows an exact match in metaphor annotation for the first corpus half (direct metaphor on top) and for the second corpus half (indirect metaphors at the bottom).

From this total of 39 tweets exhibiting exact annotation agreement, we proceeded to analyze the identified metaphors to determine whether there was a prevailing metaphor type in annotation agreement. As shown in Figure 3, our findings revealed the distribution of metaphor types as follows: 6 DMs (4 in the first half and 2 in the second half), 29 IMs (5 in the first half and 24 in the second half), and 5 PMs (3 in the first half and 2 in the second half). Although annotators were told that there could be more than one metaphor in each tweet, only one of the exact matches contemplates 2 IMs in the same tweet, so the total number of matching metaphors is 40. Figure 3 indicates a notable predominance of IM (72.5% of the exact matches), which corresponds to the general structure of the corpus, since it is the most frequent type of annotated metaphor. On the other hand, this can also be explained by the fact that every IM requires only one label per metaphor, while a DM requires three and a PM requires two.

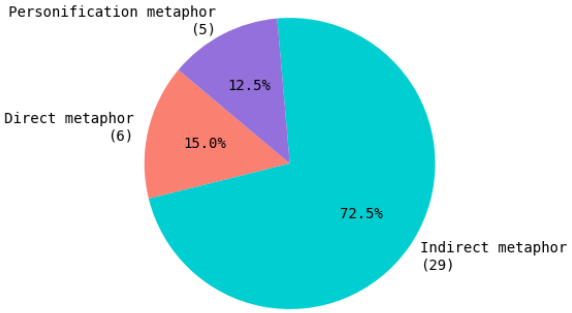


Figure 3: Distribution of Metaphor Types with Exact Agreement

6 Conclusions and Future Work

Metaphor detection is a complex task for human annotators. As we have found in this study, although native speakers of Spanish have an intuition about metaphorical language, when following annotation guidelines the exact correspondence between identified metaphors may be very low. Our research provides insights into the challenges of developing a manually annotated corpus for automatic metaphor detection in Mexican Spanish.

As Pustejovsky and Stubbs (2013) point out, the annotation of a linguistic corpus is an iterative process that involves multiple cycles of modeling and annotation, a situation that is emphasized when the goal is to annotate forms of figurative language. Moving forward in our research, efforts must be made towards refining metaphor annotation guidelines, with the follow-up goal of establishing a Gold Standard dataset of metaphorical tweets in the corpus, so that human annotators can place the corresponding labels for each particular type of metaphor in the texts. This new phase would involve another round of annotation using an updated version of the annotation guide, incorporating lessons learned from previous iterations. Through these iterative cycles of modeling and annotation, we can progressively enhance the quality and reliability of our annotated dataset, ensuring that it can be used effectively for the automatic detection of linguistic metaphors in Mexican Spanish.

References

Mateusz Babieno, Masashi Takeshita, Dusan Radisavljevic, Rafal Rzepka, and Kenji Araki. 2017. MIss RoBERTa WiLDe: Metaphor identification using masked language model with wiktionary lexical definitions. *Applied Sciences*, 12(4):2081.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Gilles Fauconnier and Mark Turner. 2008. *The Cambridge handbook of metaphor and thought*, chapter 3. Rethinking Metaphor. Cambridge University Press.

Kang Kim and Hankyu Cho. 2023. Enhanced simultaneous machine translation with word-level policies. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15616–15634, Singapore. acl.

George Lakoff and Mark Johnson. 1980. *Metafors We Live By*. The University of Chicago Press.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159.

Susan Nacey, W. Gudrun Reijnierse, Tina Krennmayr, and Aletta G. Dorst. 2019. *Metaphor Identification in Multiple Languages*. Converging Evidence in Language and Communication Research. John Benjamins Publishing Company.

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

James Pustejovsky and Amber Stubbs. 2013. *Natural language annotation for machine learning*. O'Reilly Media.

Ekaterina Shutova, Beata Beigman Klebanov, Joel Tetreault, and Zornitsa Kozareva, editors. 2013. *Proceedings of the First Workshop on Metaphor in NLP*. Association for Computational Linguistics.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*, volume 14 of *Converging Evidence in Language and Communication Research*. John Benjamins Publishing Company.

Elisa Sánchez-Bayona. 2021. Detection of everyday metaphor in spanish: annotation and evaluation.

Elisa Sánchez-Bayona and Rodrigo Agerri. 2022. Leveraging a new spanish corpus for multilingual and crosslingual metaphor detection. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*.

Omnia Zayed. 2021. Metaphor processing in tweets.

Can GPT-4 Detect Euphemisms across Multiple Languages?

Todd Firsich and **Anthony Rios**
Department of Information Systems and Cyber Security
The University of Texas at San Antonio
todd.firsich@utsa.edu, anthony.rios@utsa.edu

Abstract

Euphemisms are words or phrases used instead of another word or phrase that might be considered harsh, blunt, unpleasant, or offensive. Euphemisms generally soften the impact of what is being said, making it more palatable or appropriate for the context or audience. Euphemisms can vary significantly between languages, reflecting cultural sensitivities and taboos, and what might be a mild expression in one language could carry a stronger connotation in another. This paper uses prompting techniques to evaluate GPT-4 for detecting euphemisms across multiple languages as part of the 2024 FigLang shared task. We evaluate both zero-shot and few-shot approaches. Our method achieved an average macro F1 of .732, ranking first in the competition. Moreover, we found that GPT-4 does not perform uniformly across all languages, with a difference of .233 between the best (English .831) and the worst (Spanish .598) languages.

1 Introduction

A euphemism is a term or expression substituted for another that may be deemed too direct, harsh, or offensive. Euphemisms play a nuanced role in linguistic expression, serving as a polite or softer alternative to potentially sensitive or direct language (Danescu-Niculescu-Mizil et al.; Magu and Luo). However, their inherent ambiguity challenges Natural Language Processing (NLP) systems in comprehending meaning because they must pick up on subtle contextual cues (Bisk et al.; Carbonell and Minton). This difficulty is magnified in multilingual contexts, where the same euphemism could have different meanings across cultures. Hence, this paper describes an approach for the 2024 FigLan shared task for multilingual euphemism detection.

Much of the recent research on euphemism detection has focused on fine-tuning transformer-based models (Zhu and Bhat, 2021; Maimaitituoheti et al., 2022; Wang et al., 2022). For instance, Wang et al. (2022) combined a BERT-based transformer with a relational graph attention network and fine-tuned it for euphemism detection. However, recent advancements in the development of large language models (LLMs) like GPT-4 have been shown to be successful in similar tasks such as offensive and abusive language detection (OpenAI et al.; Wu et al.; Matter et al., 2024; Li et al., 2023). GPT-4 is supposedly trained on extensive datasets of multilingual text containing wide variations of linguistic styles, which would be very helpful in understanding and interpreting euphemistic language. The tool's ability to generate human-like dialogue and adapt itself to nuanced language suggests that it could be used to distinguish between literal and euphemistic language use.

Recent research has shown limitations of GPT-4 and related models in multi-lingual settings (Zhang et al., 2024; Ahuja et al., 2023). For example, Qiu et al. (2024) report substantial differences in medical applications performance of GPT-4 across different languages. Hence, understanding how GPT-4 performs for multilingual classification, particularly for tasks that involve figurative language, can provide unique insights into its limitations.

In this paper, we explore the application of prompting techniques (Ouyang et al.; Lester et al.; Liu et al.) to detect euphemisms using GPT-4. We note that recent work has explored prompting-based euphemism detection (Maimaitituoheti et al., 2022). However, the system still required fine-tuning model parameters. Here, we explore zero-shot and few-shot prompting strategies without any fine-tuning. We analyze a various number of in-context examples. Moreover, we performed a small error analysis to understand the limitations of GPT-4 for euphemism detection and to understand when GPT-4 fails for multilingual euphemism detection.

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 65–72
June 21, 2024 ©2024 Association for Computational Linguistics

2 Related Work

Despite the general advancements in NLP, the automated detection of euphemisms remains a relatively under-explored area. Early approaches to identify euphemistic speech focused on rule-based systems and statistical methods (Felt and Riloff). Keh et al. (2022) explored kNN and data augmentation for euphemism detection. Likewise, fine-tuning pretrained transformer models is a popular approach. For instance, Wiriyathammabhum (2022) fine-tune RoBERTa (Liu et al., 2019) models for euphemisim detection. Trust et al. (2022) combined RoBERTa models with cost-sensitive learning to handle class imbalance issues. Wang et al. (2022) combined a BERT-based transformer with a relational graph attention network and fine-tuned it for euphemism detection. However, these approaches cannot capture euphemisms' nuanced nature or how euphemisms change over time. With the advent of models such as BERT and its successors, researchers have been able to show the potential for neural network models to understand complex language phenomena like metaphors, sarcasm, and idioms (Magu and Luo; Wang et al.; Zhu and Bhat; Gavidia et al.).

While the LLMs have shown to be more capable, researchers identified that not only the size of the model and the training data used are important, but how a task is presented to the LLM is equally important (Wei et al.; Li et al.). Prompting offers a few benefits over fine-tuning a LLM. Prompting does not require a model to undergo an additional round of training, making it more resource-efficient and accessible. Also, prompting leverages the model's pre-trained knowledge, enabling quick adaptation to new tasks without the risk of overfitting. Prompting is particularly appealing for subtle language tasks like euphemism disambiguation, allowing the LLM to focus on the subtleties of euphemistic language without extensive training.

A few researchers have used prompting in previous euphemism studies (Keh; Maimaitituoheti et al.). Maimaitituoheti et al. used a RoBERTa model and fine-tuned the model to improve its performance using prompts. The most similar work to this paper is by Keh (2022), which used an older GPT-3 model and post-processing rules to classify the evaluation as euphemistic or literal. Their work found that fine-tuned models (e.g., RoBERTa) outperformed zero-shot and few-shot methods using GPT-3. In this work, we extend the idea of using prompting in two ways. First, we use GPT-4, which is more capable than GPT3-3. Second, this model is evaluated on the new multilingual euphemism dataset.

3 Methodology

In this section, we discuss the general task, dataset, and our prompting strategy. Overall, we use a few-shot prompting framework for our submission.

Task. The Multilingual Euphemism Detection Shared Task for the Fourth Workshop on Figurative Language Processing involves predicting whether a substring within a sentence is a euphemism. Specifically, given a string, "*This summer, the budding talent agent was <PET>between jobs</PET> and free to babysit pretty much any time*," participants need to detect whether the embedded Potential Euphemistic Terms (PET) is a euphemism or not for this specific context. This means that each PET can be a literal (not a euphemism or a euphemism). The participants' results are collected and evaluated on the shared task site at Codabench.[1]

Dataset. For this shared task, two sets of data are provided, each consisting of samples in Chinese, English, Spanish, and Yorùbá. The first sets are the training datasets to help refine the participants' methodology, consisting of rows of sentences, the embedded PET, and a classification label (euphemism or not). The composition of the datasets by language is provided in Table 1. The second set is the test dataset, which consists of only sentences and the embedded PET without ground truth labels. The composition of these datasets by language is also provided in Table 1. It was observed that the PETs in the training and test datasets match relatively often. For instance, we may find both "passed away" in the test and training data. Only 47 of the 67 PETs from the test dataset are in the training dataset for English. Each English PET in the test data matched an average of 1.83 euphemisms and 1.54 literal PETS. For Spanish, there are no PETs in the test dataset that are also in the training dataset. The Chinese dataset has 7 of the 48 PETs in both datasets (.38 euphemisms and .29 literal PETs on average), and Yorùbá has 14 of the 28 PETs in both datasets (0.41 euphemisms and .30 literal PETs on average). We split the training datasets into both a training and validation dataset, with 20% used for validation and 80% used as train-

[1] https://www.codabench.org/competitions/1959

Language-Set	PETs	Num Sent.	Euph.
Chinese-Train	111	2005	1484
Chinese-Test	48	1226	—
English-Train	163	1952	1383
English-Test	67	1196	—
Spanish-Train	147	1861	1143
Spanish-Test	85	1091	—
Yorùbá-Train	133	1941	1281
Yorùbá-Test	28	669	—

Table 1: Dataset Composition for Training and Testing

ing examples (i.e., to find matching PETs).

Prompt Development. We use a few-shot prompting framework for our approach. Specifically, we prompt GPT-4 using the OpenAI API to predict whether a given PET is either a euphemism (True), or not (False). We provide the prompt template below:

Given the context, determine if the phrase 'PET' is used as a Euphemism. Reply with the word 'True' if it is used as a Euphemism in this context else 'False'.

«context»

A euphemism is a mild or indirect word or expression substituted for one considered to be too harsh, blunt, or offensive. Euphemisms are used to avoid directly mentioning unpleasant or taboo topics, and they are often employed to soften the impact of the information being conveyed

«Euphemism examples»

Example - Is the phrase '{PET}' a Euphemism in the following text. {text} — Answer - 'True'

Example - Is the phrase '{PET}' a Euphemism in the following text. {text} — Answer - 'True'

«Literal examples»

Example - Is the phrase '{PET}' a Euphemism in the following text. {text} — Answer - 'False'

Example - Is the phrase '{PET}' a Euphemism in the following text. {text} — Answer - 'False'

«task»

Given the context, is the phrase '{PET}' used as a Euphemism in the following text? Context: {Text}

The prompt has five main components: instruction, context, examples of euphemism, and literal examples. The instruction provides the high-level task (e.g., return True or False). The context defines euphemisms. The euphemism and literal examples are instances directly from the training dataset. Each example is formatted in the form of "*Is the phrase [PET] a Euphemism in the following text [text].*" The PET is the substring of interest, e.g., 'between jobs." The text is the actual context that the PET appears in, e.g., "*the budding talent agent was <PET>between jobs</PET> and free to babysit pretty much any time*." Each example is followed by a "*Label*" token and either a "*True*" or "*False*" value. Finally, the task is a single test instance that we wish to classify as either the PET being a euphemism or not.

For the study, five different styles of prompting were examined. The first style is "Zero-Shot," which only uses the instruction and the task. "Zero-Shot with context" adds the context information. Next is the "Few-Shot with Random Examples" method, which uses only one random euphemism and one literal example. Research suggests that better prompt performance is achieved when similar examples are provided to the LLM in the prompt (Wei et al.; Brown et al.). Hence, we also experiment with variations called "Few-Shot with Targeted Examples," where we use k euphemism and k literal examples with the same PET as the text instance. Specifically, if the text instance's PET is "between jobs," then we will find both up to k euphemism and k literal examples that also have the "between jobs" PET. If there are no other matching examples with the same PET, or there are fewer than k matching examples, we choose the remaining examples at random.

Experimental Details. The process to evaluate the PETs used the GPT-4 APIs provided by OpenAI (OpenAI, 2023). The GPT-4 model used in our experiments is the "gpt-4-0125-preview" version and the processing occurred between 2024-02-06 and 2024-03-07. The model temperature was set at "0" to make the model less random. All other model parameters were accepted at their default values. The software developed to process each sample using the APIs was written in Python based on examples provided on the OpenAI developer website.[2]

4 Results

In this section, we report the results on both the validation and test datasets.

Validation Dataset Results. The validation dataset results are shown in Table 2. In total, we executed

[2] https://platform.openai.com/docs/guides/text-generation

Technique	Language	F1	Precision	Recall
Zero-Shot	Chinese	.650	.581	**.962**
Zero-Shot w context	Chinese	.748	.916	.795
Few Shot - Ran. Examples	Chinese	.760	.906	.832
Few Shot - Targ. Examples (2)	Chinese	.801	.941	.838
Few Shot - Targ Examples (8)	Chinese	**.858**	**.957**	.891
Zero-Shot	English	.707	**.912**	.675
Zero-Shot \w context	English	.732	.861	.805
Few Shot - Ran. Examples	English	.715	.841	.819
Few Shot - Targ. Examples (2)	English	.747	.877	.801
Few Shot - Targ. Examples (8)	English	**.820**	.907	**.877**
Zero-Shot	Spanish	.545	.794	.345
Zero-Shot + context	Spanish	.666	.800	.592
Few Shot - Ran. Examples	Spanish	.662	.772	.623
Few Shot - Targ. Examples (2)	Spanish	.698	.825	.632
Few Shot - Targ. Examples (8)	Spanish	**.761**	**.911**	**.776**
Zero-Shot	Yorùbá	.400	1.000	.181
Zero-Shot with context	Yorùbá	.610	.926	.498
Few Shot - Ran. Examples	Yorùbá	.674	.923	.61
Few Shot - Targ. Examples (2)	Yorùbá	.761	.911	.776
Few Shot - Targ. Examples (8)	Yorùbá	**.872**	**.951**	**.916**

Table 2: F1, Precision, and Recall for each prompting technique for each language dataset from the Training dataset.

20 experiments across each model and language combination (i.e., five model comparisons for each language). Overall, we make several findings. First, we find that the Zero-Shot prompting style underperforms all other methods. Interestingly, adding the context information in the "Zero-Shot with Context" method improves the results. This suggests that including more information about the task (e.g., the definition of a euphemism) can improve performance.

Next, we can find that adding in-context examples in the "Few-Shot - Random Examples" and Few -Shot - Targeted Example" methods improves the "Zero-Shot with context" methods. Furthermore, we find that using Targeted examples universally improves performance over random examples. When we add more in-context examples, the performance continues to improve. For instance, "Few-Shot - Targeted Examples" improves from .801 with four in-context examples to .859 with eight examples. From a language-to-language perspective, we obtained the worst in Spanish, which is about 5% lower than the English results.

Test Dataset Results. The final competition results for our best system (i.e., Few Shot - Targeted Examples (8)) on the test dataset are shown in Table 3. The results indicate that the prompting with the English test cases performed substantially better than the prompting with the Spanish test cases, while the Chinese and Yorùbá test cases fell in between these two extremes. For the test experiments, the source of the sample cases to be included as random or

Language	F1	Precision	Recall
Chinese	.776	.774	.780
English	.831	.829	.834
Spanish	.598	.622	.659
Yorùbá	.723	.721	.733

Table 3: F1, Precision, and Recall for each prompting technique for each language dataset from the Test dataset

targeted examples were pulled from the training datasets. The prompting proved most effective for the English dataset, and the results (F1=.831) were slightly higher than those measured during training. The results for both the Chinese (F1=.776) and the Yorùbá (F1=.723) datasets ended up falling between the "few shot random" and "few shot targeted (2)" prompt results for the training results for each language. The performance for the Spanish dataset fell (F1=.598) to only slightly better than the original "zero-shot" results.

When we look at the potential number of example cases to include with the targeted prompt, we find that with the English test cases, there was nearly 75% coverage. This means that 75% of the test PETs were also included in the training dataset. However, with the Spanish test cases, there was no overlap between the training data set and the test data set. The Chinese and Yorùbá data had test coverage between these two extremes. This may explain why the results with the Spanish dataset were so poor (0% coverage) and why the Chinese and Yorùbá datasets fell between random and targeted (some coverage).

Error Analysis. We analyzed a few of the errors to better understand how the model performed. For this analysis, we select one PET from the English dataset and one PET from the Chinese dataset.

In the English training dataset, the PET "disabled" showed good improvement by using the prompts. With the simple zero-shot prompt, all 16 examples were evaluated as being classified as a euphemism; however, seven of these examples were labeled as being literal in the ground-truth annotations. Adding context to the zero-prompt resulted in no improvement. Only slight improvement was realized when the few-shot prompt was used. However, with the few-shot prompt and eight examples, the evaluation matched 100%. The additional examples appeared to have given the model good context to discern between the nine euphemisms and seven literal cases. Overall, one potential cause

for these findings is that certain terms, such as disabled, can appear in many contexts (euphemistic and not). The model is unable to understand which applies in a given context without strong examples. Other terms mostly used in euphemistic settings are easier for the system to detect.

In the Chinese training dataset, one of the PETs that showed improvement with each new prompt technique was the PET "环卫工人," which translates to "sanitation worker." GPT-4 sometimes translates this to "city beautician," which would be a euphemism. There are 30 examples in the training dataset, and each one is classified as a euphemism.

Only 5 of the 30 examples were included in the evaluation. With zero-shot prompting, all five failed to be classified as euphemisms. With each subsequent prompt technique, the performance improved to the last prompt, where four cases were identified correctly based on the label. This would indicate that the prompting added contextual data that influenced GPT-4. We believe that the term sanitation worker may not be a strong euphemism and needs substantial evidence from examples to change the prior of the model.

5 Future Work

While demonstrating the viability of our approach in identifying euphemisms, we also uncovered several research directions to pursue that could further enhance our understanding of the euphemistic speech capabilities of LLMs.

OpenAI's Chat GPT-4 model is a high-performing LLM trained on multi-lingual data. The LLM demonstrated its capability of translating the training datasets from the original language into English without additional fine-tuning. Limited testing during the development phase was performed using Mistral (Jiang et al.) and Llama-2 LLMs (Touvron et al.) but both exhibited zero-shot performance below Chat GPT-4. The main focus of the study was on improving performance using prompting strategies, so the team directed its efforts to refine the prompts. As highly capable LLM models are being released frequently, evaluating a variety of these models is an area of focus for future studies.

Our approach utilized only the model's inherent knowledge and a subset of the training data as additional knowledge to identify euphemisms. This additional knowledge was shown to significantly improve performance during the training phase. For the cases in which there were multiple samples to choose from, the current approach randomly selected the samples to include and the order they were listed. A future research direction is to determine if the selection of examples using those that are more closely related to the test case improves the performance. Also, does the order the samples are listed in the prompt affect the results?

When reviewing the test performance (Table 3), we noticed that not all languages performed comparably between training (Table 2) and test. When investigating the results for the lowest-performing dataset during the test phase (Spanish), we identified that no samples from the training dataset matched the PET in the test dataset. As noted, this additional knowledge was shown to be beneficial.

There are two approaches we could pursue to address this. One would be to locate additional datasets online or create datasets from open-source language repositories. A second approach would be to use a language model to generate the additional samples. The attraction to this approach is that we could generate samples of a new PET being used in a previously unseen manner and assist the model in recognizing the new usage of a phrase.

6 Conclusion

In this paper, we presented our approach for the 2024 FigLang Shared Task for multilingual Euphemism detection. We introduced a method using GPT-4 and in-context learning. This adjustment would be beneficial in a scenario in which the usage of a euphemism has changed over time, but the model has not yet been learned, or the model does not have a strong indication of being a euphemism without strong evidence. Future areas to research include 1) using the LLM to generate samples to include as examples to include in the multi-targeted prompt 2) improving the selection of targeted examples to identify those examples that are more closely related to the test case. 3) using the LLM to identify potential euphemisms from the text in question without being supplied with this information.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 2145357.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, et al. 2023. Mega: Multilingual evaluation of generative ai. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, and Tom Henighan. Language models are few-shot learners.

Jaime G. Carbonell and Steven Minton. Metaphor and common-sense reasoning:.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. A computational approach to politeness with application to social factors.

Christian Felt and Ellen Riloff. Recognizing euphemisms and dysphemisms using sentiment analysis. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 136–145. Association for Computational Linguistics.

Martha Gavidia, Patrick Lee, Anna Feldman, and Jing Peng. CATs are fuzzy PETs: A corpus and analysis of potentially euphemistic terms.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b.

Sedrick Scott Keh. Exploring euphemism detection in few-shot and zero-shot settings.

Sedrick Scott Keh. 2022. Exploring euphemism detection in few-shot and zero-shot settings. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 167–172.

Sedrick Scott Keh, Rohit Bharadwaj, Emmy Liu, Simone Tedeschi, Varun Gangal, and Roberto Navigli. 2022. EUREKA: EUphemism recognition enhanced through knn-based methods and augmentation. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 111–117, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059. Association for Computational Linguistics.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2023. "hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*.

Moxin Li, Wenjie Wang, Fuli Feng, Yixin Cao, Jizhi Zhang, and Tat-Seng Chua. Robust prompt optimization for large language models against distribution shifts.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. 55(9):1–35.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rijul Magu and Jiebo Luo. Determining code words in euphemistic hate speech using word embedding networks. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 93–100. Association for Computational Linguistics.

Abulimiti Maimaitituoheti, Yang Yong, and Fan Xiaochao. A prompt based approach for euphemism detection.

Abulimiti Maimaitituoheti, Yang Yong, and Fan Xiaochao. 2022. A prompt based approach for euphemism detection. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 8–12.

Daniel Matter, Miriam Schirmer, Nir Grinberg, and Jürgen Pfeffer. 2024. Close to human-level agreement: Tracing journeys of violent speech in incel posts with gpt-4-enhanced annotations. *arXiv preprint arXiv:2401.02001*.

OpenAI. 2023. Chatgpt turbo 4 preview model 0125. `https://openai.com`. Accessed: 2024-03-02.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko,

Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *arXiv preprint arXiv:2402.13963*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models.

Paul Trust, Kadusabe Provia, and Kizito Omala. 2022. Bayes at FigLang 2022 euphemism detection shared task: Cost-sensitive Bayesian fine-tuning and Venn-abers predictors for robust training under class skewed distributions. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 94–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, and Michael Witbrock. Improving natural language inference using external knowledge in the science questions domain. 33(1):7208–7215.

Yuting Wang, Yiyi Liu, Ruqing Zhang, Yixing Fan, and Jiafeng Guo. 2022. Euphemism detection by transformers and relational graph attention network. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 79–83.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models.

Peratham Wiriyathammabhum. 2022. Tedb system description to a shared task on euphemism detection 2022. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 1–7.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. A brief overview of ChatGPT: The history, status quo and potential future development. 10(5):1122–1136.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2024. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36.

Wanzheng Zhu and Suma Bhat. Euphemistic phrase detection by masked language model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 163–168. Association for Computational Linguistics.

Wanzheng Zhu and Suma Bhat. 2021. Euphemistic phrase detection by masked language model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 163–168.

Ensemble-based Multilingual Euphemism Detection: a Behavior-Guided Approach

Fedor Vitiugin and **Henna Paakki**
Department of Computer Science, Aalto University, Finland
{fedor.vitiugin, henna.paakki}@aalto.fi

Abstract

This paper describes the system submitted by our team to the Multilingual Euphemism Detection Shared Task for the Fourth Workshop on Figurative Language Processing (FigLang 2024). We propose a novel model for multilingual euphemism detection, combining contextual and behavior-related features. The system classifies texts that potentially contain euphemistic terms with an ensemble classifier based on outputs from behavior-related fine-tuned models. Our results show that, for this kind of task, our model outperforms baselines and state-of-the-art euphemism detection methods. As for the leader-board, our classification model achieved a macro averaged F1 score of 69%, reaching the third place.

1 Introduction

Euphemism, as defined by the Oxford English Dictionary, is the substitution of mild or indirect expressions for harsh or blunt ones when referring to unpleasant topics. The American Heritage Dictionary of the English Language similarly defines euphemism as replacing harsh or offensive terms with milder, indirect ones.

This paper explores the task of detecting euphemisms across multiple languages. Euphemism is a linguistic strategy employed to soften the impact of direct or uncomfortable language, such as using 'collateral damage" instead of "war-related civilian deaths". Euphemisms are commonly employed to maintain politeness, ease discomfort, or veil harsh realities in everyday communication. Despite cultural differences in their usage, the universal need to discuss sensitive topics without causing offense suggests commonalities in how euphemisms are applied across languages and cultures. This study investigates how multilingual models can leverage these similarities in processing euphemisms.

Our work is part of a Shared Task for the Fourth Workshop on Figurative Language Processing (FigLang 2024) and focuses on the euphemism disambiguation task, in which potentially euphemistic termss (PETs) are classified as euphemistic or not in a given context in four languages (Chinese, English, Spanish, and Yorùbá). This set of languages helps to encompass a diverse range of linguistic and cultural backgrounds (Lee et al.).

Our approach achieved the third-best score in the multilingual euphemism detection shared task. This paper describes our model [1] participating in the task.

2 Related Work

In this section, we explore related work about figurative language detection and euphemism detection in particular, utilization of behavior-related models for detecting specific types of content, and use of ensemble learning for combining different approaches for text classification.

2.1 Euphemism Detection

Euphemism allows writers to address taboo topics indirectly, facilitating better cross-cultural communication. Consequently, there's a growing interest in computational methods for detecting euphemisms within Natural Language Processing (NLP) (Lee et al., 2022; Gavidia et al., 2022; Lee et al., 2023).

Recent work demonstrates semantic lexicon induction and the development of sentiment analysis methods could help to detect of euphemisms by investigating their connection with sentiment analysis. The study suggests analyzing affective polarity and connotation within sentence contexts yields better results than directly labeling phrases (Felt and Riloff, 2020).

[1] Our code is available at https://github.com/vitiugin/med

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 73–78
June 21, 2024 ©2024 Association for Computational Linguistics

Pre-trained transformer models are extensively employed in various NLP-related tasks including euphemism detection through task-specific fine-tuning (Tiwari and Parde, 2022), in combination with relational graph attention network (Wang et al., 2022), with adversarial augmentation technique (Kohli et al., 2022). Additionally, the utilization of clustering algorithms to provide additional signals of PETs similarity improves performance of pre-trained model in ensemble methods (Keh et al., 2022).

Leveraging of prompt tuning pre-trained language models is another direction in euphemism detection. Use of RoBERTa as the pre-trained language model and creation of suitable templates and verbalizers could be effectively used (Maimaitituoheti et al., 2022).

Large Language Modelss (LLMs) have been the subject of exploration regarding their multilingual and cross-lingual transfer capabilities in prior studies (Lee et al.). Multilingual LLMs extensively leverage data from multiple languages, acquiring both complementary and reinforcing information (Choenni et al., 2023). Transfer learning from out-of-language data within a particular domain yielded superior results compared to utilizing same-language data from a different domain (Shode et al., 2023).

2.2 Behavior-Related Fine-Tuning for Euphemism Detection

Since euphemisms are established social speaking and behaving norms, ways of thinking as well as outlook of value, it is essential to study their application. Euphemism exists in all aspects of English in great numbers and is categorized into eight types (Li-Na, 2015): *death, aging and disease* ("passed away", "passed", "departed"), *disability and handicap* ("mentally challenged", "special needs", "full-figured"), *education* ("slow student", "peer homework"), *marriage and pregnancy* ("renovate", "unwedding", "tie the knot"), *military* ("collateral damage", "neutralizing", "involvement"), *profession* ("sanitation engineer", "comfort woman"), *politics* ("the deprived", "economic downturn"), *profanity* ("private parts", "choke the chicken").

Utilizing models to detect sociopolitical threads can enhance euphemism detection performance according to the provided classification. Behavior-related fine-tuning (Ruder, 2021) involves teaching models relevant capabilities for excelling in a target task, necessitating an understanding of diverse human behavioral patterns in language (Founta et al., 2019; Zhang et al., 2023). This process involves fine-tuning the model on related tasks to acquire practical behaviors (Vitiugin and Purohit, 2024), contrasting with adaptive fine-tuning. Behavioral fine-tuning, particularly with labeled data, has proven effective in teaching models various linguistic features such as named entities (Broscheit, 2020), paraphrasing (Arase and Tsujii, 2019), syntax (Glavaš and Vulić, 2021), answer sentence selection (Garg et al., 2020), and question answering (Khashabi et al., 2020). A recent study emphasized the importance of a diverse task selection for optimal transfer performance, based on fine-tuning a model on nearly 50 labeled datasets in a massively multitask environment (Aghajanyan et al., 2021).

2.3 Ensemble Learning

Ensemble multifeatured deep learning is a powerful method to improve model generalization and performance, which has been used effectively in figurative language detection. Combining ensemble outputs can boost metaphor detection performance (Brooks and Youssef, 2020). Additionally, utilizing an Adaptive Boosting classifier with Decision Tree as a base estimator shows promise in predicting sarcasm probabilities (Lemmens et al., 2020).

By combining the strengths of multiple models and features, ensemble multifeatured deep learning models have demonstrated improved performance and adaptability in diverse problem settings. While these models have such challenges as model interpretability, computational complexity, ensemble model selection, adversarial robustness, and personalized and federated learning (Abimannan et al., 2023).

3 Model Architecture

The model's architecture is presented in Figure 1 and includes two main steps: fine-tuning for behavior-related downstream tasks and ensemble method for classification.

First, we fine-tuned the multilingual transformer-based model (XLM-RoBERTa (Conneau et al., 2019)) for classifying contextual texts (without PETs) and classifying PETs separately. Based on review of related work, we fine-tuned the same pre-trained language model for the several behav-

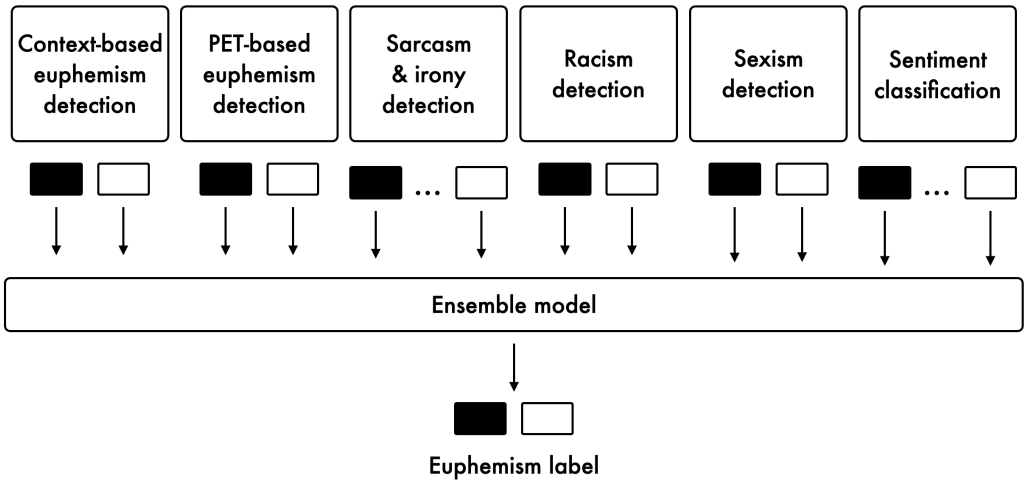


Figure 1: Model architecture

ioral tasks: detection of sarcasm and irony (Ling and Klinger, 2016), sexism, racism (Albright, 2021), and sentiment classification (Passionate-NLP, 2021). After fine-tuning, we had 6 fine-tuned models with the same architecture, and tokenizers.

Second, our final model used the ensemble learning method for classification, which received logits from described models as features. During the developing step, we tested several ensemble models including: Adaptive Boosting, Extra Trees, Gradient Boosting, and Random Forest.

Finally, we used the best performing ensemble learning method to train model for detection euphemisms in four languages.

4 Experiment

For the shared task, we made only multilingual experiments, i.e. training and developing datasets contain entities in all four presented languages.

4.1 Dataset

The dataset for the experiment includes texts in four languages: Mandarin Chinese (ZH), American English (EN), Spanish (ES), and Yorùbá (YO) (Lee et al., 2023). The dataset for each language contains texts, PETs, and labels (euphemistic or non-euphemistic). Dataset statics is presented in Table 1. For each test run, we use 80-10-10 split to create training, validation, and test sets.

4.2 Implementation Details

We maintain the same number of layers in each model – 24 layers for XML-RoBERTa (Conneau et al., 2019). During fine-tuning, we used the same hyperparameters and number of frozen layers (detected for task-related fine-tuning by grid search.) For LLMs' fine-tuning, we used $0.5 * 10^{-5}$ learning rate, 10 epochs. The number of frozen layers for each model were detected by grid search. The models were trained on NVIDIA A100-SXM4 with 40Gb GPU RAM.

Table 1: Experiment dataset statistics

language	euphemistic	non-euphemistic	total
Chinese (ZH)	1484	521	2005
English (EN)	1383	569	1952
Spanish (ES)	1143	718	1861
Yorùbá (YO)	1281	660	1941

Table 2: Comparison of ensemble learning methods for classification. 10-fold CV for multilingual data.

scheme	ACC	AUC	F1
Adaptive Boosting	96.06	95.38	95.13
Extra Trees	96.01	95.32	95.06
Gradient Boosting	**96.10**	95.39	94.75
Random Forest	**96.10**	**95.42**	**95.27**

4.3 Baselines and Compared Methods

To compare our proposed method for multilingual euphemism detection problem, we construct baseline scheme using deep learning model that use LASER embeddings (Artetxe and Schwenk, 2019) as input features. Additionally, we also compare our method in combination with varied sets of behavior-related models. The full list of schemes includes:

- **[LSTM_text&PET]** – method uses combines pre-trained LASER embeddings of text and PET, which are passed as input to a

Table 3: Comparison of baseline schemes and proposed approach. 10-fold CV for multilingual data.

scheme	ACC	AUC	F1
LSTM_text&PET	79.52 ± 0.5	79.66 ± 0.4	88.30 ± 0.9
RoBERTa_text&PET	91.29 ± 0.7	90.42 ± 0.9	90.25 ± 1.1
RoBERTa_text&PET&sexism	95.84 ± 0.8	95.13 ± 1.0	94.92 ± 0.9
RoBERTa_text&PET&racism	95.79 ± 0.7	95.07 ± 0.9	94.90 ± 1.1
RoBERTa_text&PET&social	95.82 ± 0.7	95.11 ± 0.9	94.87 ± 1.1
RoBERTa_text&PET&social&sarcasm	96.02 ± 0.7	95.23 ± 0.9	94.94 ± 1.1
RoBERTa_text&PET&social&sentiment	96.03 ± 0.7	95.35 ± 0.8	95.09 ± 1.1
RoBERTa_text&PET&all	**96.10 ± 0.7**	**95.42 ± 0.9**	**95.27 ± 1.1**

Long Short-Term Memory (LSTM) Network model (Vitiugin and Barnabo, 2021);

- **[RoBERTa_text&PET]** – method uses logits of fine-tuned RoBERTa for euphemism detection in text and PET;
- **[RoBERTa_text&PET&sexism]** – method uses logits of fine-tuned RoBERTa for euphemism detection in text and PET, as well as logits of the model for sexism detection;
- **[RoBERTa_text&PET&racism]** – method uses logits of fine-tuned RoBERTa for euphemism detection in text and PET, as well as logits of the model for racism detection;
- **[RoBERTa_text&PET&social]** – method uses logits of fine-tuned RoBERTa for euphemism detection in text and PET, as well as logits from models for sexism and racism detection;
- **[RoBERTa_text&PET&social&sarcasm]** – method uses logits of fine-tuned RoBERTa for euphemism detection in text and PET, as well as logits from models for sexism, racism, and sarcasm detection;
- **[RoBERTa_text&PET&social&sentiment]** – method uses logits of fine-tuned RoBERTa for euphemism detection in text and PET, as well as logits from models for sexism and racism detection and from sentiment classification model;
- **[RoBERTa_text&PET&all]** – method uses logits of fine-tuned RoBERTa for euphemism detection in text and PET, as well as logits from all behaviour-related models.

4.4 Results

First, we compare several ensemble methods applying for the euphemism detection task. In this experiment we use outputs from all fine-tuned models and all ensemle methods' parameters were optimized by applying Greed Search. Table 2 demonstrates that the Random Forest classifier reaches the highest results. While Adaptive Boosting, Extra Trees, and Gradient Boosting perform less effective, 10-fold cross-validation demonstrates that the difference between the performance of different models is insignificant (p-value ≥ 0.05). As a result of this experiment, we chose the Random Forest model for combining outputs of fine-tuned models.

Comparison of baseline and proposed models on training data provided by organizers of the shared task demonstrates high performance of ensemble learning method with behavior-related models. Use of logits from all fine-tuned models shows the best performance. Even use of logits from the only one behaviour-related model significantly improves results (p-value ≤ 0.05) comparing to combination of logits provided only by contextual and PET models. While our experiments didn't show significant improvement of performance between models used outputs from one behaviour-related model and outputs from all behaviour related models (p-value ≈ 0.4). The full results of schemes comparison are presented in Table 3.

4.5 Shared Task Results

During the test phase of the shared task, we employed our most effective model, *RoBERTa_text&PET&all*. However, its performance significantly declined compared to the development phase, achieving a macro-averaged F1 score of 69%. This highlights the model's reliance on contextual familiarity, particularly as the test data incorporates numerous new PETs. Notably, English and Chinese languages exhibited better performance overall, aligning with trends observed in similar methods. Noteworthy, our model excelled with the Spanish dataset. For detailed results, please refer to Table 4.

Table 4: Shared task results for test dataset provided by organizers.

Language	P	R	F1
English	75.29	75.57	73.90
Spanish	68.78	66.56	67.43
Yorùbá	65.53	62.77	63.06
Chinese	71.10	82.00	70.44

5 Conclusion

We have described a method for multilingual euphemism detection. This method is based on behaviour-related fine-tuning of transformer model for combining their logits in ensemble learning. Experiments with four different languages demonstrate that our approach could reach high performance in the task.

5.1 Limitations

In the work, we used only English datasets for behavior-related fine-tuning. The use of datasets in other languages could show different results.

5.2 Future Work

One of the directions of future research is exploration of grammatical features of euphemisms. Grammatical methods, such as past tense and passive voice, create psychological distance and politeness. Extracting these types of features from the text could enhance multilingual euphemism detection.

6 Acknowledgements

This work is supported by the Trust-M research project, a partnership between Aalto University, University of Helsinki, Tampere University, and the City of Espoo, funded in-part by a grant from the Strategic Research Council (SRC) in Finland. The research is also supported in-part by a grant from the Helsingin Sanomat Foundation for the project AI-infused Disinformation in Media Communications (DiME).

References

Satheesh Abimannan, El-Sayed M El-Alfy, Yue-Shan Chang, Shahid Hussain, Saurabh Shukla, and Dhivyadharsini Satheesh. 2023. Ensemble multi-featured deep learning models and applications: A survey. *IEEE Access*.

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811.

Munki Albright. 2021. Suspicious tweets dataset. https://www.kaggle.com/datasets/munkialbright/classified-tweets.

Yuki Arase and Jun'ichi Tsujii. 2019. Transfer fine-tuning: A BERT case study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5393–5404, Hong Kong, China. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Jennifer Brooks and Abdou Youssef. 2020. Metaphor detection using ensembles of bidirectional recurrent neural networks. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 244–249.

Samuel Broscheit. 2020. Investigating entity knowledge in bert with simple neural end-to-end entity linking. *arXiv preprint arXiv:2003.05473*.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023. How do languages influence each other? studying cross-lingual data sharing during llm fine-tuning. *arXiv preprint arXiv:2305.13286*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Christian Felt and Ellen Riloff. 2020. Recognizing euphemisms and dysphemisms using sentiment analysis. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 136–145.

Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM conference on web science*, pages 105–114.

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2020. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7780–7788.

Martha Gavidia, Patrick Lee, Anna Feldman, and Jing Peng. 2022. Cats are fuzzy pets: A corpus and analysis of potentially euphemistic terms. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2658–2671.

Goran Glavaš and Ivan Vulić. 2021. Is supervised syntactic parsing beneficial for language understanding? an empirical investigation.

Sedrick Scott Keh, Rohit K Bharadwaj, Emmy Liu, Simone Tedeschi, Varun Gangal, and Roberto Navigli. 2022. Eureka: Euphemism recognition enhanced through knn-based methods and augmentation. *arXiv preprint arXiv:2210.12846*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Guneet Kohli, Prabsimran Kaur, and Jatin Bedi. 2022. Adversarial perturbations augmented language models for euphemism identification. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 154–159.

Patrick Lee, Anna Feldman, and Jing Peng. 2022. A report on the euphemisms detection shared task. *arXiv preprint arXiv:2211.13327*.

Patrick Lee, Iyanuoluwa Shode, Alain Chirino Trujillo, Yuan Zhao, Olumide Ebenezer Ojo, Diana Cuervas Plancarte, Anna Feldman, and Jing Peng. 2023. Feed pets: Further experimentation and expansion on the disambiguation of potentially euphemistic terms. *arXiv preprint arXiv:2306.00217*.

Patrick Lee, Alain Chirino Trujillo, Diana Cuevas Plancarte, Olumide Ebenezer Ojo, Xinyi Liu, Iyanuoluwa Shode, Yuan Zhao, Jing Peng, and Anna Feldman. Meds for pets: Multilingual euphemism disambiguation for potentially euphemistic terms.

Jens Lemmens, Ben Burtenshaw, Ehsan Lotfi, Ilia Markov, and Walter Daelemans. 2020. Sarcasm detection using an ensemble approach. In *proceedings of the second workshop on figurative language processing*, pages 264–269.

Zhou Li-Na. 2015. Euphemism in modern american english. *Sino-US English Teaching*, 12(4):265–270.

Jennifer Ling and Roman Klinger. 2016. An empirical, quantitative analysis of the differences between sarcasm and irony. In *The Semantic Web: ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29–June 2, 2016, Revised Selected Papers 13*, pages 203–216. Springer.

Abulimiti Maimaitituoheti, Yang Yong, and Fan Xiaochao. 2022. A prompt based approach for euphemism detection. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 8–12.

Passionate-NLP. 2021. Twitter sentiment analysis dataset. https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis.

Sebastian Ruder. 2021. Recent advances in language model fine-tuning. http://ruder.io/recent-advances-lm-fine-tuning,.

Iyanuoluwa Shode, David Ifeoluwa Adelani, Jing Peng, and Anna Feldman. 2023. Nollysenti: Leveraging transfer learning and machine translation for nigerian movie sentiment classification. *arXiv preprint arXiv:2305.10971*.

Devika Tiwari and Natalie Parde. 2022. An exploration of linguistically-driven and transfer learning methods for euphemism detection. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 131–136.

Fedor Vitiugin and Giorgio Barnabo. 2021. Emotion detection for spanish by combining laser embeddings, topic information, and offense features.

Fedor Vitiugin and Hemant Purohit. 2024. Multilingual serviceability model for detecting and ranking help requests on social media during disasters. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18.

Yuting Wang, Yiyi Liu, Ruqing Zhang, Yixing Fan, and Jiafeng Guo. 2022. Euphemism detection by transformers and relational graph attention network. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 79–83.

Dong Zhang, Wenwen Li, Baozhuang Niu, and Chong Wu. 2023. A deep learning approach for detecting fake reviewers: Exploiting reviewing behavior and textual information. *Decision Support Systems*, 166:113911.

An Expectation-Realization Model for Metaphor Detection

Oseremen O. Uduehi
School of EECS
Ohio University
Athens, OH 45701
ou380517@ohio.edu

Razvan C. Bunescu
Department of Computer Science
University of North Carolina at Charlotte
Charlotte, NC 28223
razvan.bunescu@uncc.edu

Abstract

We propose a new model for metaphor detection in which an expectation component estimates representations of expected word meanings in a given context, whereas a realization component computes representations of target word meanings in context. We also introduce a systematic evaluation methodology that estimates generalization performance in three settings: within distribution, a new strong out of distribution setting, and a novel out-of-pretraining setting. Across all settings, the expectation-realization model obtains results that are competitive with or better than previous metaphor detection models.

1 Introduction and Motivation

Metaphors enhance the communicative aspects of language by connecting concepts from new domains, often abstract, with more familiar ones, usually concrete (Lakoff and Johnson, 1980). Metaphorical expressions have many uses, from helping frame an issue in order to emphasize some aspects of reality (Boeynaems et al., 2017), to creating a strong emotional effect (Blanchette and Dunbar, 2001; Citron and Goldberg, 2014). The ubiquity of metaphors means their computational treatment (Veale et al., 2016) has received significant attention in the NLP community, as surveyed by Shutova (2015) and more recently Tong et al. (2021). Owing to its important communicative function, metaphorical expression detection has been approached over the years using a wide variety of NLP techniques, ranging from models employing hand-engineered features (Shutova et al., 2010; Bulat et al., 2017), to RNNs (Gao et al., 2018; Mao et al., 2019), to more recently pre-trained language models (Choi et al., 2021; Ghosh et al., 2022; Li et al., 2023), to mention just a few.

Recent state of the art models for metaphor detection rely on the Metaphor Identification Procedure (MIP) (Group, 2007), according to which metaphors happen whenever the contextual meaning of a word is different from its basic, literal meaning. Implementations of MIP vary mainly in how they estimate representations of the basic meaning of a word: MelBert (Choi et al., 2021) uses simply the BERT embedding of the word without any context, whereas BasicBERT and BasicMIP (Li et al., 2023) use an average of all literal uses of the word as marked in the training data.

In this paper we propose a new theory of metaphor identification, the *Expectation-Realization* model, that is motivated by the observation that the metaphorical use of a word, i.e. its *realization* in context, leads to surprise due to a violation of a literal word *expectation* engendered by the same context. Surprise offers a general mechanism through which stories and music trigger emotion (Meyer, 1961), and correlates with creative uses of language, such as humor and metaphor (Bunescu and Uduehi, 2022). Correspondingly, we propose an architecture that is structured around two modules: one module aims to estimate the literal meaning expectation through the use of a context where the target word is masked, whereas the other module aims to estimate the realized meaning of the target word as used in context. The new model is competitive with previous SoA in terms of within distribution (WiD) generalization. We further propose two new evaluation scenarios: a *strong out-of-distribution* (OoD) setting that ensures target lexemes do not appear during training, and a novel *out-of-pretraining* (OoP) setting that aims to ensure that the metaphorical phrase was not seen during pretraining. The large gap between OoP and WiD results elucidates why pretrained LMs struggle with metaphor identification.

2 The Expectation-Realization Model

The architecture of the Expectation-Realization (ER) model for metaphor detection is shown in

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 79–84
June 21, 2024 ©2024 Association for Computational Linguistics

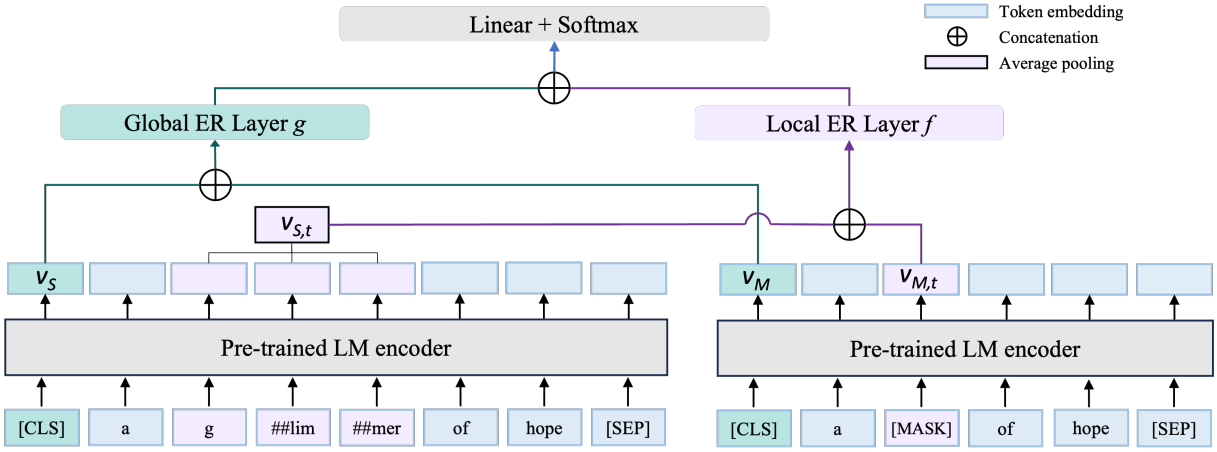


Figure 1: ER architecture: left branch for Realization embeddings, right branch for Expectation embeddings. The high expectation for the literal meaning "a [bit] of hope" is confounded by the word "glimmer", causing surprise.

Figure 1. To compute the realized (R) meaning $v_{S,t}$ of the target word in context, a copy of the Transformer encoder of a pre-trained language model (shown on the left) processes the input text S where the target word at position t is marked with a special token. To compute the expectation (E) of the literal meaning $v_{M,t}$ induced by the context, the same pre-trained language model (shown on the right) process the same input text M where the target word is masked. Additionally, global expectation v_M and realization v_S representations are also computed at the sentence level using the embeddings for the special [CLS] token. The concatenation of the local target word ER embeddings and the sentence-level ER embeddings are passed through non-linear layers f and g, respectively, to capture interactions between expectation and realization embeddings at word-level as $h_{local} = f[v_{M,t}; v_{S,t}]$, and at sentence level as $h_{global} = g[v_M; v_S]$. To enable a fair comparison with previous models, we instantiate the pre-trained Transformer encoder using RoBERTa base (Liu et al., 2019). The concatenated local and global ER representations are then used as input features to a logistic regression model that estimates the probability $\hat{y}$ that the target word is used metaphorically.

$$\hat{y} = \sigma(w^{\mathsf{T}}[h_{local}; h_{global}] + b)$$

The ER model parameters together with the pre-trained LM parameters are trained and fine-tuned, respectively, in order to minimize a loss function $L^i = L^i_{CE} - L^i_{Sim}$ that contains a *cross-entropy loss* L^i_{CE} and a *similarity loss* L^i_{Sim} computed as:

$$\begin{aligned} L^i_{CE} &= y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \\ L^i_{Sim} &= \alpha_1 \cos(u_{M,t}, v_{M,t}) + \alpha_2 \cos(u_M, v_M) \end{aligned}$$

where y_i and $\hat{y}_i$ are the ground truth and predicted labels, respectively, for training sample i. The embeddings u are obtained from the original pre-trained LM with fixed parameters, whereas the embeddigns v are obtained from the fine-tuned LM. Importantly, the similarity loss encourages the fined-tuned LM to learn expectation embeddings v that do not deviate much from the original embeddings produced by the pre-trained LM. The hyper-parameters α_1 and α_2 trade-off the global and local components of the similarity term within the overall loss. Given that most words in the vocabulary are used with their literal meaning most of the time, the similarity loss has the effect of anchoring the fine-tuned LM such that its expectation embeddings v reflect a literal meaning of words.

3 Experimental Evaluation

We run evaluations on three English metaphor datasets: the VUA-18 Amsterdam Metaphor Corpus (Chen et al., 2020), TroFi (Birke and Sarkar, 2006) and LCC (Mohler et al., 2016). Table 1 summarizes the statistics of the datasets used in our evaluations. The VUA-18 dataset is split into training, validation and test datasets denoted by VUA-18$_{tr}$, VUA-18$_{dev}$ and VUA-18$_{te}$ respectively. The examples in the VUA-18 dataset are sentences where selected words of the sentence are annotated as metaphorical or not. The LCC Metaphor dataset is a large, multilingual dataset of metaphor annotations created by a team of researchers at the Language Computer Corporation (LCC). Each target word is annotated with a metaphoricity rating on a four-point scale [0, 3]. In our experiments we use a subset of the English dataset where examples with

Dataset	#words	%M	#Sent	Len
VUA-18$_{tr}$	116,622	11.2	6,323	18.4
VUA-18$_{dev}$	38,628	11.6	1,550	24.9
VUA-18$_{te}$	50,175	12.4	2,694	18.6
LCC	5,646	28.9	5,390	28.9
TroFi	3,737	43.5	3,737	28.3

Table 1: Detailed statistics of datasets. #words is the number of target words to be classified, %M is the percentage of metaphorical words, #Sent is the number of sentences, and Len is the average sentence length.

metaphoricity score of 3 are considered as positive and examples with metaphoricity score of 0 as negatives. The TroFi dataset consists of a collection of literal and nonliteral usage of 50 verbs which occur in 3,737 sentences selected from the WSJ corpus.

For the evaluations on VUA-18 dataset, we use the same hyperparameter settings from (Choi et al., 2021) for training all models. For the LCC and TroFi experiments, the development dataset was used for determining the best hyperparameter settings. We use the same hyperparameter settings for all the models. The batch size and max sequence length were set at 32 and 150, respectively. We train for 12 epochs without dropout, and linearly increase the learning rate from 0 to 5e-5 in the first two epochs, after which we decreased it linearly to 0 during the remaining 10 epochs. The tuned similarity weights α_1 and α_2 were 1.0. for the within-distribution experiments and 0.0 for out-of-distribution experiments. Results are averaged over 5 runs with different random seeds. The detailed ranges used for hyperparameters tuning are presented in Appendix A.

Given that VUA-18 is the only dataset on which all 3 metaphor-detection baselines were previously evaluated, we use it to compare their performance against ER. As shown in Table 2, the ER model outperforms both MDGI-Joint-S (Wan et al., 2021) and MelBERT (Choi et al., 2021), and is competitive with the more complex BasicBERT (Li et al., 2023) that requires annotation of literal tokens.

3.1 Three Generalization Scenarios

The generalization performance of each of the 3 models is evaluated in three settings: *within distribution (WiD)*, strong *out of distribution (OoD)*, and *out-of-pretraining (OoP)* metaphor generalization. For the WiD generalization, we randomly split the dataset into 10 folds and run 10-fold evaluation, where 9 folds are used for training and development, and 1 fold is used for testing, with the procedure repeated 10 times so that each folds gets to be used as a test fold. For strong OoD generalization, the 10 folds are created such that the lemmas of target words are disjoint across the folds. For the OoP generalization setting, we identify a subset of 237 positive examples within the LCC dataset that are novel or unconventional metaphors. The criteria for creating this subset were example with the highest metaphoricity score of 3.0 that were also rare according to a Google search, i.e. returning fewer that 25 search results. To complete the novel version of the dataset, negative examples are randomly sampled from the LCC dataset such that the ratio of positive to negatives for this novel dataset is similar to that of the original LCC dataset. Note that the OoP examples, which are *novel to the pretrained LM*, are different from the crowdsourced novel metaphors from (Do Dinh et al., 2018), which are *novel to the average human* annotator. For the OoP evaluation we only compute the test performance on the OoP subset of examples using the models already trained on data from the within-distribution setting, ensuring that no OoP test example has been used during training.

Due to the imbalanced distribution of positive and negative examples in the datasets, we report only precision, recall and F1-score metrics. For 10-fold evaluation we report their micro-averages.

3.2 Generalization Results

Tables 3 and 4 show the results of comparison of the ER Model against MelBERT and R-SPV on the LCC and TroFi datasets. The R-SPV model implements only the realization component of the ER model, using as input the sentence with the target word marked, as shown on the left of Figure 1. Note that even though this is equivalent with the SPV component of the MelBERT model, it is found to perform as well as MelBERT. Additionally, for the LCC and TroFi datasets in the WiD setting we also report the performance of a logistic regression model that trained on binary responses from GPT-4 on 13 questions that are aimed at identifying metaphors and also distinguishing metaphors from other types of figurative language (Appendix B).

For the the within distribution (WiD) setting of VUA-18, LCC and TroFi, the ER model statistically significantly outperforms R-SPV and MelBERT, as determined through a one-tailed, paired

Dataset	Model	Prec	Rec	F1
VUA-18 (WiD)	MDGI-Joint-S	81.3	73.2	77.0
	MelBERT	80.1	76.9	78.5
	BasicBERT	79.5	78.5	79.0
	ER	80.2	77.5	78.8

Table 2: Performance comparison of ER model with baselines on the VUA-18 dataset.

Dataset	Model	Prec	Rec	F1
LCC (WiD)	R-SPV	86.2	83.9	85.0
	MelBERT	86.1	83.8	84.9
	GPT-4	82.1	77.5	79.7
	ER	**86.9**	**84.3**	**85.5***†
	ER-Ens	87.7	85.3	86.5
LCC (OoD)	R-SPV	83.6	79.8	81.6
	MelBERT	83.4	79.8	81.5
	ER	**84.0**	**80.6**	**82.2***†
	ER-Ens	85.9	81.9	83.9
LCC (OoP)	R-SPV	88.0	94.3	91.1
	MelBERT	87.6	94.5	90.9
	ER	**88.8**	**95.1**	**91.8***†
	ER-Ens	89.3	95.7	92.4

Table 3: Performance comparison of ER model with baselines on LCC dataset. * and †indicate significantly better F1 than R-SPV and MelBERT, respectively.

t-test of significance at $p < 0.05$ level. The VUA-18 results are notably lower than the LLC results for all methods. Error analysis revealed that almost any non-literal use of a word is annotated as a positive example in VUA, including idioms. Therefore, the patterns are more complicated. Idioms, in particular, lack any clear pattern, hence they require memorization, which may explain the much lower VUA performance. The logistic regression model on top of features from GPT-4 had the lowest WiD F1 on LCC and TroFI, indicating that, despite its language understanding capabilities, it still struggles to accurately identify metaphors, a result that can also be understood in light of insights drawn from the OoP scenario below. The GPT-4 results were obtained using binary answers to questions in a zero-shot setting; it is expected that in-context learning with few-shot examples or fine-tuning of GPT models, while more computationally demanding than using BERT-like models, will lead to better results. We leave such experiments for future work.

Dataset	Model	Prec	Rec	F1
TroFi (WiD)	R-SPV	70.2	71.8	71.0
	MelBERT	69.5	73.3	71.3
	GPT-4	63.5	60.9	62.1
	ER	**70.2**	**73.7**	**71.9***†
	ER-Ens	72.2	73.5	72.8
TroFi (OoD)	R-SPV	57.4	69.6	62.8
	MelBERT	57.1	69.8	62.7
	ER	57.0	70.5	63.0
	ER-Ens	58.1	71.8	64.2

Table 4: Performance comparison of ER model with baselines on TroFi dataset. * and †indicate significantly better F1 than R-SPV and MelBERT, respectively.

For the strong out-of-distribution (OoD) evaluation on the LCC and TroFi datasets, the ER model on average performs better than both R-SPV and MelBERT, with the comparison on LCC being statistically significant. The results from the OoD settings show a significant drop compared to the within distribution setup with the result being less worse for LCC than TroFi because of the more diverse nature of the target words in the LCC dataset. This drop in performance in the OoD scenario suggests that the models rely on some form of memorization, which is detrimental to identifying metaphors that use unseen words. The nature of the TroFi dataset makes the OoD generalization even worse, as the dataset contains only 50 words and thus the model has limited diversity in terms of target metaphorical words.

In the out-of-pretraining (OoP) evaluation setting conducted for the LCC dataset, the ER model again outperforms both baselines, obtaining a 9.8% relative error reduction over MelBERT. Note that the OoP results are much higher than the WiD results for all methods, which seems to indicate that the difficulty of metaphor detection comes from the large number of conventional metaphors that appear often in the pretraining data; that in turn makes it hard for pretrained models such as BERT or GPT to create embeddings that can discriminate conventional metaphors from literal language.

Lastly, ensembles ER-Ens of 5 ER models further improve metaphor detection in all settings.

4 Conclusion and Future Work

We introduced a new model for metaphor detection rooted in the hypothesis that non-literal uses of

words trigger surprise, or violation of expectations given by the context. We further proposed two new evaluation scenarios: strong out-of-distribution and out-of-pretraining. Extensive experiments show that the simple ER model is competitive with, and often outperforms, state-of-the-art models.

In this work, expectations of literal meaning were computed based on context words. In future work, we plan to also compute expectations of literal meanings of words by leveraging large amounts of text where words are known to be used literally, such as descriptions of physical, concrete concepts in Wikipedia. Furthermore, we plan to generalize the ER approach from word-level metaphors to phrase-level constructions, such as idioms, which too violate expectations of literal language use.

Acknowledgements

We would like to thank the anonymous reviewers for their suggestions and constructive feedback. This research was partly supported by the United States Air Force (USAF) under Contract No. FA8750-21-C-0075. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the USAF.

References

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 329–336, Trento, Italy. Association for Computational Linguistics.

Isabelle Blanchette and Kevin Dunbar. 2001. Analogy use in naturalistic settings: The influence of audience, emotion, and goals. *Memory & Cognition*, 29(5):730–735.

Amber Boeynaems, Christian Burgers, Elly Konijn, and Gerard Steen. 2017. The impact of conventional and novel metaphors in news on issue viewpoint. *International Journal of Communication*, 11(0).

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. Modelling metaphor with attribute-based semantics. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 523–528, Valencia, Spain. Association for Computational Linguistics.

Razvan C. Bunescu and Oseremen O. Uduehi. 2022. Distribution-based measures of surprise for creative language: Experiments with humor and metaphor. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 68–78, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Xianyang Chen, Chee Wee (Ben) Leong, Michael Flor, and Beata Beigman Klebanov. 2020. Go figure! multi-task transformer-based architecture for metaphor detection using idioms: ETS team in 2020 metaphor shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 235–243, Online. Association for Computational Linguistics.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor Detection via Contextualized Late Interaction using Metaphorical Identification Theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

Francesca M. M. Citron and Adele E. Goldberg. 2014. Metaphorical sentences are more emotionally engaging than their literal counterparts. *Journal of Cognitive Neuroscience*, 26(11):2585–2595.

Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. 2018. Weeding out Conventionalized Metaphors: A Corpus of Novel Metaphor Annotations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1424, Brussels, Belgium. Association for Computational Linguistics.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613, Brussels, Belgium. Association for Computational Linguistics.

Debanjan Ghosh, Beata Beigman Klebanov, Smaranda Muresan, Anna Feldman, Soujanya Poria, and Tuhin Chakrabarty, editors. 2022. *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid).

Pragglejaz Group. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

George Lakoff and Mark Johnson. 1980. *Metaphors we Live by*. University of Chicago Press, Chicago.

Yucheng Li, Shun Wang, Chenghua Lin, and Frank Guerin. 2023. Metaphor Detection via Explicit Basic Meanings Modelling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 91–100, Toronto, Canada. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3898, Florence, Italy. Association for Computational Linguistics.

Leonard Meyer. 1961. *Emotion and Meaning in Music*. University of Chicago.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).

Ekaterina Shutova. 2015. Design and Evaluation of Metaphor Processing Systems. *Computational Linguistics*, 41(4):579–623. _eprint: https://direct.mit.edu/coli/article-pdf/41/4/579/1807226/coli_a_00233.pdf.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1002–1010, Beijing, China. Coling 2010 Organizing Committee.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686, Online. Association for Computational Linguistics.

Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. Metaphor: A Computational Perspective. *Synthesis Lectures on Human Language Technologies*, 9(1):1–160. Publisher: Morgan & Claypool Publishers.

Hai Wan, Jinxia Lin, Jianfeng Du, Dawei Shen, and Manrong Zhang. 2021. Enhancing Metaphor Detection by Gloss-based Interpretations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1971–1981, Online. Association for Computational Linguistics.

A Hyperparameter Tuning

Details for the hyperparameter tuning for the models and dataset are presented in Table 5.

Hyperparameter	Tuning values
learning rate	[1e-5, 2e-5, 3e-5, 4e-5, 5e-5]
dropout ratio	[0.0, 0.1, 0.2, 0.25, 0.4, 0.5]
similarity weight α	[0, 0.5, 1, 2, 4]
hidden dims	[[768], [768,768], [768,768,1]]
hidden activation	[None, relu]
optimizer	[Adam]
train batch size	[32]

Table 5: Hyperparameters tuning range used in experiments. For the similarity weight, $\alpha = \alpha_1 = \alpha_2$.

B GPT-4 prompt template

The sample prompt we used to query GPT-4 is shown below:

You are a professional linguist. For the text below, answer precisely the following questions. Only print out a Python list containing your answers.

text: The sun *walked* between the clouds.

1. What word is emphasized?
2. Is the emphasized word "walked" used literally in the text? Yes or No?
3. Is the emphasized word "walked" used figuratively in the text? Yes or No?
4. Is the emphasized word "walked" used metaphorically in this text? Yes or No?
5. Is the emphasized word "walked" used with its literal meaning in the text? Yes or No?
6. Is the emphasized word "walked" used with its most common literal meaning in this text? Yes or No?
7. Is the emphasized word "walked" used with a concrete meaning in the text? Yes or No?
8. Is the emphasized word "walked" used with a physical meaning in the text? Yes or No?
9. Is the emphasized word "walked" used with its conventional meaning in the text? Yes or No?
10. Is the emphasized word "walked" used with its most common meaning in this text? Yes or No?
11. Is the emphasized word "walked" used with its original (oldest) meaning in this text? Yes or No?
12. Is the emphasized word "walked" part of a metaphorical expression in the text? Yes or No?
13. Is the emphasized word "walked" part of an idiomatic expression in the text? Yes or No?
14. Is the emphasized word "walked" part of a multiword expression in the text? Yes or No?

A Textual Modal Supplement Framework for Understanding Multi-Modal Figurative Language

Jiale Chen[1], Qihao Yang[1], Xuelian Dong[1], Xiaoling Mao[2] and **Tianyong Hao[1]***
[1] School of Computer Science, South China Normal University
[2] School of Languages and Literature, University of South China
{jlchen, charlesyeung}@m.scnu.edu.cn,
{xueliandong01, BriannaMxl}@163.com,
haoty@m.scnu.edu.cn

Abstract

Figurative language in media such as memes, art, or comics has gained dramatic interest recently. However, the challenge remains in accurately justifying and explaining whether an image caption complements or contradicts the image it accompanies. To tackle this problem, we design a modal-supplement framework MAPPER consisting of a describer and a thinker. The describer based on a frozen large vision model is designed to describe an image in detail to capture entailed semantic information. The thinker based on a finetuned large multi-modal model is designed to utilize description, claim and image to make prediction and explanation. Experiment results on a publicly available benchmark dataset from FigLang2024 Task 2 show that our method ranks at top 1 in overall evaluation, the performance exceeds the second place by 28.57%. This indicates that MAPPER is highly effective in understanding, judging and explaining of the figurative language. The source code is available at https://github.com/Libv-Team/figlang2024.

1 Introduction

Figurative language in media has gained much interests recently. By understanding similes and metaphors in the figurative language, it is possible to deepen the understanding of specific cultural contexts and social phenomena (Hwang and Shwartz, 2023). This task is challenging because it involves abstract reasoning about images, as well as it involves understanding social common sense and cultural contexts.

Traditional solutions extract features from images using CNNs and encode textual descriptions with RNNs (Mo et al., 2023; Chen et al., 2024), employ multi-modal fusion for inference (Karpathy and Fei-Fei, 2015; Vinyals et al., 2015), and determine and elucidate their interrelations through

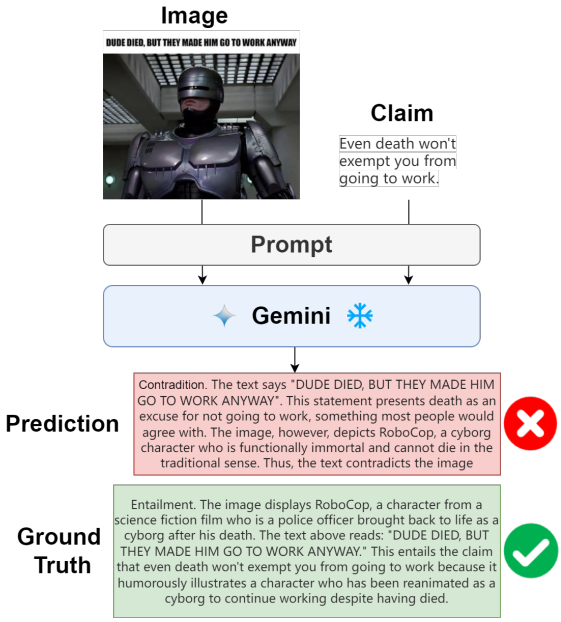


Figure 1: A typical method uses zero shot prompts to induce responses from a multi-modal large language model.

classification and explanation generation. With the development of Multi-modal Large Language Models (MLLM) in image captioning and Visual Question Answering (VQA), it turns to be a visual entailment task. The task first predicts whether an image caption entails the image or not and provide a text explanation for labeling prediction. New ideas also involve using formulated prompts according to heuristic rules to guide a large model in producing a relevant answer. The main framework of a typic method utilizing large model and prompt is shown in Figure 1.

Despite the progress made by these methods in dealing with visual entailment tasks, when faced with specific cultural and social contexts, the model ability to explain and reason is limited due to the lack of relevant context. Subsequently, the inconsistency between images and texts may make the models more challengeable to determine the entail-

*Corresponding author

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 85–91
June 21, 2024 ©2024 Association for Computational Linguistics

ment relationships. Thereby, their performance is much worse than that of human beings.

To that end, we propose MAPPER (textual ModAl suPPlement framEwoRk), a figurative language understanding model. It consists of a describer and a thinker. The describer provides a textual model of the image as a modality supplement for further prediction and explanation by the thinker. Experiment result indicated that MAPPER is effective in understanding, judging and explaining of the figurative language. It shows that a simple fine-tuning method can significantly enhance the model performance in figurative language understanding with just minor prompt adjustments.

2 Related work

In recent years, advances in language modeling notably improved model comprehension of metaphorical language. Chakrabarty et al. proposed a model that fine-tuned T5 to understand metaphorical language through textual interpretation. Chakrabarty et al. introduced a knowledge augmentation model employing human strategies for explaining types of figurative language: inferring meaning from context and drawing on the literal meanings of constituent words. This knowledge augmentation model enhanced performance on discriminative and generative tasks, further narrowing the gap with human performance. Liu et al. created a Fig-QA benchmark through crowdsourcing for a broader study of metaphorical language. Their findings indicated that although pre-trained language models could achieve commendable performance after fine-tuning, their performance on a limited number of samples still fell significantly short of human capabilities.

In addition, with the development of multimedia, there had been an increased focus on generative understanding of multimodal metaphorical language. Hessel et al. investigated visual language models and language-only models for understanding multimodal metaphorical language and found that both types of models had difficulties in all three tasks. Desai et al. introduced an architecture based on a multimodal Transformer, which included a cross-modal attention mechanism focusing on the distinctive features between images and captions. This model obtained relatively high consistency scores in human evaluations. Yosef et al. utilized the state-of-the-art vision and language model CLIP (Radford et al., 2021) to perform on a multimodal metaphorical language comprehension task and found that it performed relatively poorly. The experimental results showed that the best model was only 22% accurate in the detection task, much lower than the 97% accuracy achieved by humans. This discrepancy was mainly due to the poor performance of model in understanding the connection between metaphorical language and images, with a tendency to prefer partially literal images over metaphorical ones.

These studies have primarily improved performance through methods such as model fine-tuning and knowledge enhancement. However, they still face challenges in understanding multimodal metaphorical language. To enhance the capability of visual language models to comprehend metaphorical language, we design prompts to clarify task requirements and employ modal supplement methods to boost the integration of multimodal data, aiming to narrow the gap between models and humans in multimodal metaphor comprehension.

3 The Method

The task of multimodal figurative language is defined as follows: Given an image claim C and an image I, a label L that indicate the caption entails or contracts to the image need to be predict. A corresponding explanation E of the predicted label is needed to be generated.

This paper proposes a textual modal supplement framework MAPPER, which is consisted of a describer and a thinker. The describer read the i-th image I_i, and used self-knowledge to describe the image as inherent thinking D_i. The thinker uses the inherent thinking D_i, image I_i and claim C_i to generate final predict L_i iand explanation E_i. The overview of the model is shown in Figure 2.

The Describer. To better understanding the image content, a MLLM-based describer is designed according the prompt instruction P_1 from the prompt template PTR_1 and the i-th image I_i to generate the image description D_i, It is worth noting that the parameter weights in MLLM are frozen. Formally:

$$D_i = MLLM_{frozen}\left(Prompt\left(I_i, C_i\right)\right) \quad (1)$$

Here, P_1 is designed as follows:

```
<Image> Please describe in detail what
you see in the provided image.
```

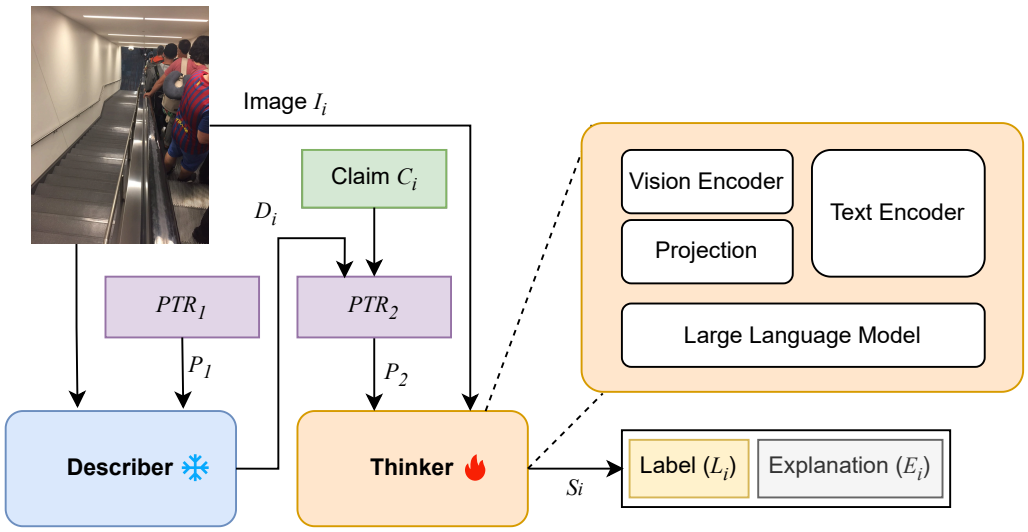


Figure 2: The overall architecture of our MAPPER framework with a describer and a thinker.

Data	Train/Valid		Test	
	Absolute	Proportion	Absolute	Proportion
Nycartoons	520	11.7%	87	12.6%
IRFL	1322	29.9%	198	28.7%
Muse	1000	22.6%	150	21.8%
Mamecap	853	19.3%	128	18.6%
Vismet	731	16.5%	126	18.3%

Table 1: Dataset statistics.

The Thinker. A prompt template PTR_2 is used firstly to generate a prompt based on the image description D_i and claim C_i. In this way, we unify the original classification and the generation tasks into one generation task. In this way, the thinker generates the responses S_i consisted of the concatenation of the label L_i and explanation E_i. The design of prompt template is shown as follows:

```
The description of this picture is
<Description>. The claim of this picture
is <Claim>. You need to predict the claim
of this picture is 'entailment' or
'contradiction' firstly according to the
picture and its description. Then you need
to give an explanation for the prediction.
The prediction and the explanation are
related to the meaning of the figurative
language expression. Your response must
follow the format shown as below:
"Prediction. Explanation".
```

Next, a vision encoder Enc_p is designed to encode the image I_i, and a text encoder Enc_e is used to encode the prompt p_2. $f(\bullet)$ is the projection function. The process is as shown in Equation 2–4.

$$H_v = f(Enc_p(I_i)) \tag{2}$$

$$H_l = Enc_e\left(Prompt\left(C_i, D_i\right)\right) \tag{3}$$

$$[L_i; E_i] = s_i = LLM\left(H_v, H_l\right) \tag{4}$$

Training. During training epoch, the model is trained as a minimized negative log likelihood as Equation 5.

$$\mathcal{L} = \sum_{j}^{n} -\log p\left(s_i^j \middle| s_i^{<j}, I_i, C_i, D_i\right) \tag{5}$$

s_i^j is the generated word output in the j-th time step that generated by the system. n is the maximum response length.

4 Experiments

4.1 Datasets

The V-FLUTE (Saakyan et al., 2024) was used in the experiments. It consisted of five small datasets, with data compiled from a series of prior work on visual metaphor and multimodal understanding, supplemented with annotated explanations detailing the implicit relationships (Yosef et al., 2023; Chakrabarty et al., 2023; Hwang and Shwartz, 2023; Hessel et al., 2023b; Jain et al., 2020; Shahaf et al., 2015). The statistical details of these datasets weres presented in Table 1. We followed datasets splits from the competition "UNDERSTANDING

Type	Method	Metrics		
		F1@0	F1@50	F1@60
Zero Shot	LLava-7B-v1.6 (offical baseline)	44.82	37.38	19.99
	LLava-7B-v1.5	43.40	40.42	20.30
	Gemini-Pro-Vision	59.57	58.61	42.36
	Gemini (Text only)	57.24	56.30	36.09
	GPT-4V	69.56	63.78	48.89
Fine-tune	TinyLLava-1.5B	72.56	71.39	59.24
	TinyLLava-3.1B	<u>86.12</u>	<u>85.40</u>	<u>71.56</u>
	MAPPER (Ours)	**89.67**	**89.09**	**74.15**

Table 2: Performance comparison of the models on the V- FLUTE datasets. The best performance is bolded and the second is underlined.

OF FIGURATIVE LANGUAGE THROUGH VISUAL ENTAILMENT" [1] for training, validation, and testing.

4.2 Evaluation Metrics

The evaluation metrics were primarily F1 scores for the label prediction. In addition, we used BERT-score (Yuan et al., 2021) to assess the quality of the explanation. Thus, the evaluation metrics were F1@0 (only F1 scores), F1@50 (computed F1 scores where only instances whose interpretations matched a reference with BERT-score higher than 50 were treated as correct), and F1@60 (computed F1 scores where only instances whose interpretations matched a reference with BERT-score higher than 60 were treated as correct). These metrics were based on previous work in FigLang2022.

4.3 Baseline

Three categories of baseline models were evaluation in this experiment. 1) The origin multi-modals models: Gemini-Pro-Vision (Team et al., 2023), and GPT-4V. 2) The models consisted of an image encoder and a large language model: LLava-7B (Liu et al., 2023), TinyLLava (Zhou et al., 2024). 3) The large language model: Gemini-Pro.

4.4 Hyperparameters

A frozen parameter LLava-7B-v1.5 model was used for the describer, while a finetuned LLava-7B-v1.5 model with Lora (Hu et al., 2021) for the thinker. The training epoch was set to 3. The batch size was set to 4 and the learning rate was set to $2.5e^{-5}$. The rank of the Lora model was set to 128. The learning rate scheduler type was used "cosine", and the max length of model was constraint to 2048. The vision tower of MAPPER used CLIP. The warm up ration was set to 0.03. All experiments were conducted in a NVIDIA 4090 GPU with 24GB memory.

5 Results and Analysis

5.1 Main Result

The results of comparing with the baseline models were shown in Table 2. It could be seen that our MAPPER achieved the highest scores on all three metrics through the supervised fine-tuning method. Specifically, F1@0 reached 90.06, F1@50 reached 89.49, and F1@60 reached 76.33. Compared with the Prompt-based Zero Shot method, our method improved the performance of these three indicators by 28.57%, 34.33%, and 50% respectively after supervised fine-tuning compared with the method without fine-tuning. These high F1 scores indicated that MAPPER could effectively understand the information in images and give accurate predictions and explanations. Moreover, some interesting phenomena were identified. The Gemini-Pro-Vision model performed better than the Gemini (Text only) model in zero-shot conditions. This might indicate that replacing images with their descriptions could cause information losing in predication, resulting in worse performance.

5.2 Ablation Study

To explore the effectiveness of each component, an ablation study was conducted. We removed describer and PTR_2 of the MAPPER one by one and then analyzed the performance changes. The result was shown in Table 3. From the table result,

[1] https://www.codabench.org/competitions/1970/#/pages-tab

Method	F1@0	F1@50	F1@60
Ours	**89.67**	**89.09**	**74.15**
- Describer	89.20 (0.47 ↓)	89.04 (0.05↓)	75.58 (1.43 ↑)
- PTR_2	51.16 (38.51 ↓)	50.46 (38.63 ↓)	41.45 (32.70 ↓)

Table 3: The ablation experiment with/without two main modules Describer and PTR_2.

Image:  Claim: The cat is busy as a bee

Method	Prediction + Explanation	F1	BERT-score
MAPPER	Entailment. The image depicts a cat lying on a bed with glasses on its face, appearing to use a laptop with its paws, as if it were working or using the computer. The simile "The cat is as busy as a bee" is entailed by the image because the cat is engaged in an activity that resembles work or productivity, similar to how a bee would be busy with its tasks.	1	81.85
TinnyLLava-3.1B	Entailment. The image shows a cat sitting on a bed with a laptop, wearing glasses and a bowtie, as if it were a human working on the computer. The simile "The cat is as busy as a bee" is entailed by the image because the cat's posture and attire suggest that it is engaged in a task, similar to how a bee is industrious and focused on its work.	1	80.98
TinyLLava-1.5B	Entailment. The image depicts a cat wearing glasses and a tie, sitting on a bed with a laptop in front of it. The simile ""The cat is as busy as a bee"" is entailed by the image because the cat's attire and the context suggest that it is engaged in a task that requires focus and attention, similar to the industrious nature of a bee."	1	79.44
Ground Truth	Entailment. The image shows a cat with glasses on, looking at a laptop as if it's working or engaged in an activity. The simile "The cat is as busy as a bee" is entailed by the image because the cat appears to be actively involved with the laptop, mimicking the busy behavior of a bee with human-like working characteristics.	-	-

Figure 3: MAPPER and two baseline generated explanations in the V-FLUTE dataset.

we could draw the following observations:

The describer had relatively little impact on model performance. Without "Describer", the model performance scores on F1@0 and F1@50 dropped by 0.47 and 0.05 respectively, which were a relatively small change. However, it was worth noting that the F1@60 score increased by 1.43, which might indicate that the describer might had limitations when dealing with complex or difficult-to-classify cases.

The PTR_2 had a large impact on model performance. When we removed PTR_2, the model performance scores on all three metrics dropped significantly, especially on F1@0 and F1@50, where the scores dropped by 38.51 and 38.63 respectively. This indidated that the PTR_2 component played a key role in the model and had a significant impact on the model performance.

Overall, these results indicated that the performance of our MAPPER relied heavily on the PTR_2, while the describer component had a relatively less impact. This provided us with important guidance when improving the model and optimizing performance in the future.

5.3 Case Study

Figure 3 presents a case study demonstrating the comparesion between MAPPER and two baselines. The data of this case study was sourced from the V-FLUTE dataset. The label predictions and explanations were generated by MAPPER and two

baseline methods, TinyLLava-1.5B and TinyLLava-3.1B. The input consisted of an image and a claim. The image showed a cat working in front of a computer accompanied by the claim "That cat is busy as a bee". All three methods accurately predicted the labels. However, the explanation generated by MAPPER achieved the highest BERT-score compared to the baseline methods. There were some biased words using in the baseline explanations in contrast to the ground truth. Our MAPPER explicitly indicated the "busy" and generated explanation more closely resembling the ground truth, resulting in the highest BERT-score. This case exemplifies capacity of MAPPER to generate explanations that closely align with the ground truth.

6 Conclusion

This paper proposed a textual modal supplement method MAPPER for figurative language understanding. The MAPPER used a frozen LLava as the describer to generate a description of the image and a finetuned MLLM as the thinker to make predictions and explanations for the figurative language within image and claim. Experiment results on the public datasets indicated that our MAPPER achieved the state-of-the-art performance. The results illustrated that a finetune in small dataset about understanding of figurative language could highly improve MLLM model performance.

Limitation

Due to competition time constraints, we did not explore clearly in this experiment why the textual modal supplement generated by describer can have a negative impact on F1@60. Although our method ranked first in the competition, this paper did not design different prompts to test the robustness of our method. In addition, we did not further explore whether a MLLM with a larger number of parameters can learn more accurate judgment and understanding of figurative language in pictures.

Acknowledgments

The work is supported by grants from National Natural Science Foundation of China (No. 62372189) and the Research Grants Council of the Hong Kong Special Administrative Region, China (UGC/FDS16/E09/22).

References

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022a. It's not Rocket Science: Interpreting Figurative Language in Narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022b. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7370–7388, Toronto, Canada. Association for Computational Linguistics.

Guanhua Chen, Qiqi Xu, Choujun Zhan, Fu Lee Wang, Kai Liu, Hai Liu, and Tianyong Hao. 2024. Improving open intent detection via triplet-contrastive learning and adaptive boundary. *IEEE Transactions on Consumer Electronics*.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023a. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023b. Do androids laugh at electric sheep? Humor "understanding" benchmarks from The New Yorker Caption Contest. In *Proceedings of the ACL*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A dataset for captioning and interpreting memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445, Singapore. Association for Computational Linguistics.

Lalit Jain, Kevin Jamieson, Robert Mankoff, Robert Nowak, and Scott Sievert. 2020. The New Yorker cartoon caption contest dataset.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Di Mo, Bangrui Huang, Haitao Wang, Xinyu Cao, Keqin Gan, Jie Wei, Heng Weng, and Tianyong Hao. 2023. Sclert: A span-based joint model for measurable quantitative information extraction from chinese texts. *IEEE Transactions on Consumer Electronics*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2024. V-FLUTE: Visual Figurative Language Understanding with Textual Explanations Dataset. https://huggingface.co/datasets/ColumbiaNLP/V-FLUTE. Dataset associated with the paper "V-FLUTE: Visual Figurative Language Understanding with Textual Explanations".

Dafna Shahaf, Eric Horvitz, and Robert Mankoff. 2015. Inside jokes: Identifying humorous cartoon captions. In *KDD*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. IRFL: Image recognition of figurative language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1044–1058, Singapore. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. 2024. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*.

FigCLIP: A Generative Multimodal Model with Bidirectional Cross-attention for Understanding Figurative Language via Visual Entailment

Qihao Yang
School of Computer Science
South China Normal University
Guangzhou, China
charlesyeung@m.scnu.edu.cn

Xuelin Wang✉
College of Chinese Language and Culture
Jinan University
Guangzhou, China
wangxuelin@stu2022.jnu.edu.cn

Abstract

This is a system paper for the FigLang-2024 Multimodal Figurative Language Shared Task. Figurative language is generally represented through multiple modalities, facilitating the expression of complex and abstract ideas. With the popularity of various text-to-image tools, a large number of images containing metaphors or ironies are created. Traditional recognizing textual entailment has been extended to the task of understanding figurative language via visual entailment. However, existing pre-trained multimodal models in open domains often struggle with this task due to the intertwining of counterfactuals, human culture, and imagination. To bridge this gap, we propose FigCLIP, an end-to-end model based on CLIP and GPT-2, to identify multimodal figurative semantics and generate explanations. It employs a bidirectional fusion module with cross-attention and leverages explanations to promote the alignment of figurative image-text representations. Experimental results on the benchmark demonstrate the effectiveness of our method, achieving 70% F1-score, 67% F1@50-score and 50% F1@60-score. It outperforms GPT-4V, which has robust visual reasoning capabilities.

1 Introduction

Figurative language is typically divided into metaphor, simile, and sarcasm (Saakyan et al., 2022). It serves as an implicit way for us to convey complex and imaginative expressions. In recent years, researchers have focused on developing neural networks through mining contextual information. They also aim to construct large-scale figurative datasets to facilitate in-depth research on recognizing textual entailment (Gu et al., 2022; Bigoulaeva et al., 2022; Phan et al., 2022). Despite increasing in parameter size, pre-trained language models (Devlin et al., 2018; Liu et al., 2019) are still unable to fully comprehend cultural knowledge and the social context within figurative language.

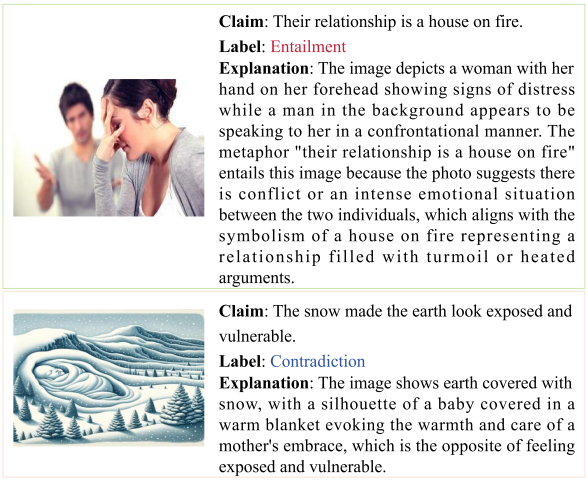


Figure 1: Illustration of the Multimodal Figurative Language Shared Task.

With the prevalence of social media, individuals sometime use images with visual metaphors (i.e., figurative images) to convey counterfactual or humorous meanings, particularly in the advertising industry (Yosef et al., 2023). Various text-to-image AI tools can also be used to create a vast number of figurative images (Chakrabarty et al., 2023). To promote the research on figurative language, the Multimodal Figurative Language Shared Task[1] (named Understanding of Figurative Language Through Visual Entailment) is first introduced by FigLang-2024[2]. Given an <image, text> pair, the goal of this task is to 1) predict whether the image entails or contradicts the text, where the text is referred to as "claims"; 2) generate an explanation for the entailment or contradiction. The illustration of this task is shown as Figure 1.

Different from previous research that focused

✉ Corresponding author.

[1] https://www.codabench.org/competitions/1970

[2] https://sites.google.com/view/figlang2024/home

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 92–98
June 21, 2024 ©2024 Association for Computational Linguistics

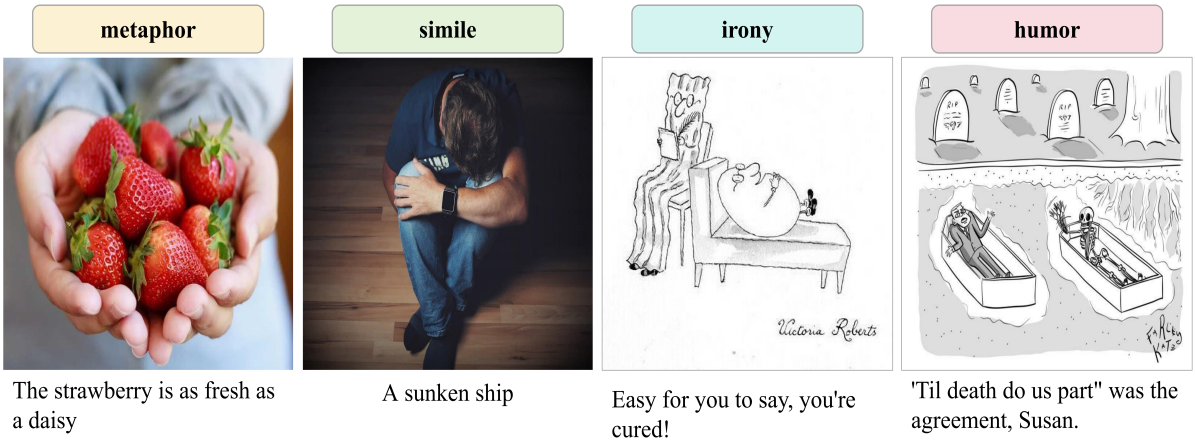


Figure 2: Examples of visual entailment between images and claims.

on recognizing textual entailment (Chakrabarty et al., 2022b), the Multimodal Figurative Language Shared Task introduces an image modality to interpret figurative language. Empirically, images can carry richer contextual information than words. Awareness of abstract implications beyond literal and intuitive meanings is the most significant challenge for this task. Even CLIP (Radford et al., 2021), a state-of-the-art architecture in image-text understanding, achieves only 62% accuracy in multimodal entailment test settings, which is far less than the human accuracy of 94% (Hessel et al., 2023). Moreover, existing vision-language models (Radford et al., 2021; Li et al., 2023, 2022) and generative language models (Raffel et al., 2020; Radford et al.) are utilized separately to predict image-text labels and generate explanations. This results in a decoupling of the task, which is inconsistent with the widely accepted paradigm of end-to-end training. Although many large multimodal models (Liu et al., 2024; Jin et al., 2023) perform well on diverse downstream tasks, the availability of large-scale figurative image-text datasets and the requirement for high computational resources are prerequisites for fine-tuning them. Therefore, developing a generic, low-cost, end-to-end multimodal model for multimodal figurative language can potentially further advance the future associated research.

In this paper, we propose FigCLIP. It is built upon CLIP and GPT-2 (Radford et al.) and can jointly achieve the two requirements of label prediction and explanation generation. The main contributions of this work can be summarized as follows:

- A low-cost and end-to-end model is proposed, which is competitive in multimodal figurative language task.
- A bidirectional fusion module with cross-attention is introduced, which enhances the alignment of figurative image-text representations within the mapping space defined by CLIP and GPT-2.
- We compare the model performance for understanding multimodal figurative language at different resolutions.

2 Related Work

Understanding figurative language has been framed as a recognizing textual entailment (RTE) task (Chakrabarty et al., 2022b). Given a <premise, hypothesis> pair, a RTE model is required to determine whether the texts entail or contradict each other. Pre-trained language models like BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) are used to encode both premise and hypothesis texts. The deep representations of premise and hypothesis texts are concatenated and then input into a linear-layer classifier to output an entailment or contradiction label (Chakrabarty et al., 2021, 2022a; Hu et al., 2023). However, these methods cannot enable us to probe whether language models are right for the right reasons. Thus, researchers are committed to construct refined RTE datasets to avoid spurious correlations and annotation artifacts and provide profound figurative knowledge. Explanation-based RTE datasets such as e-SNLI (Camburu et al., 2018) and FLUTE (Chakrabarty et al., 2022b) are increasingly favored. Employing large language models (LLMs) has become

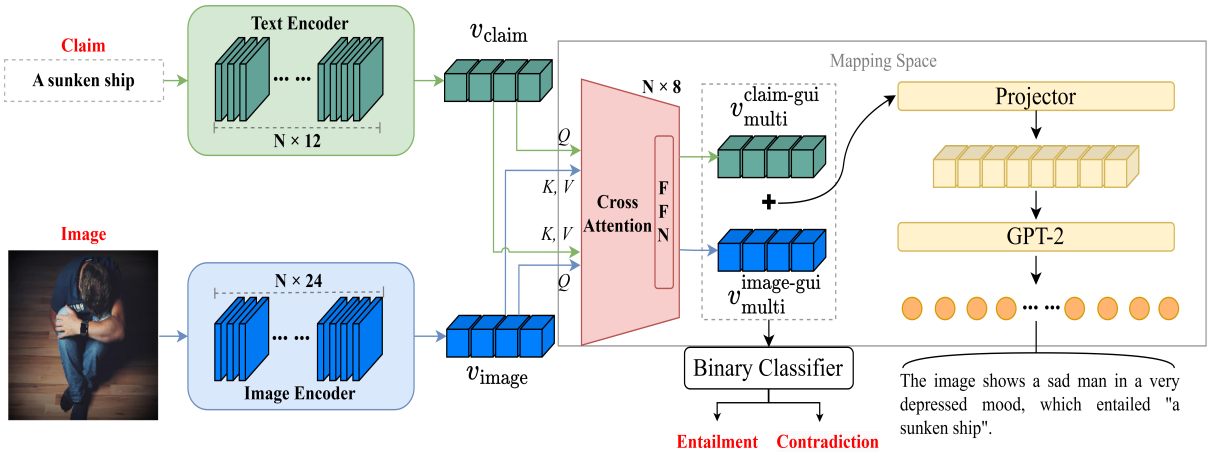


Figure 3: Overview framework of the proposed FigCLIP model.

the mainstream approach to address the RTE task (Kim et al., 2023). Premise and hypothesis texts are combined into prompts to guide the LLMs for generating answers. This implies that the RTE task is simplified into a question-answering problem, allowing the full utilization of the LLMs' capabilities in natural language inference.

Figurative language in images has recently received increasing attention (Yosef et al., 2023; Hessel et al., 2023). As shown in Figure 2, images with claim texts can present metaphors, similes, irony and humor. With the help of diffusion-based text-to-image models such as DALL-E (Ramesh et al., 2021), a number of comic-like figurative images is created based on figurative texts. A high-quality dataset is constructed by (Chakrabarty et al., 2023), containing 6,476 visual metaphors for 1,540 linguistic metaphors and their associated visual elaborations. The Image Recognition of Figurative Language (IRFL) dataset is developed by (Yosef et al., 2023), with human annotation and an automatic pipeline. Although the size of figurative multimodal datasets is increasing, it is still not enough for training a model with strong generalization ability. Thus, pre-trained multimodal models can serve as the backbone and are used to learn the fine-grained figurative image-text representations by fine-tuning on limited figurative multimodal datasets. They only perform the label prediction. For generating explanation, captions generated from images are concatenated with claim texts into pure textual questions. The questions are fed into language models such as GPT-2 and T5 (Raffel et al., 2020), then an explanations are output. To meet the two needs of prediction and explanation at the same time, several large multimodal models, such as GPT-4V (Achiam et al., 2023), MiniGPT4 (Zhu et al., 2023), Flamingo (Alayrac et al., 2022), LlaVA (Liu et al., 2024), are used to accept image and text input and then generate labels and explanations. However, they are commonly evaluated by zero-shot or few-shot due to the high training cost. Research on fine-tuning them on figurative multimodal datasets is still scarce.

3 Method

3.1 Task formulation

The Multimodal Figurative Language (MFL) Shared Task can be treated as a classification and generation problem. Given an <image, claim> pair, a MFL model is required to align image-claim representations, learn a binary classification function F_c to predict entailment or contradiction labels by following Eq. 1, and learn a generation function F_g to generate explanations by following Eq. 2.

$$label = \arg\max F_c\left(image, claim\right) \quad (1)$$

$$explanation = \arg\max F_g\left(image, claim\right) \quad (2)$$

3.2 The FigCLIP model

Architecture. The proposed FigCLIP model employs 12-layer transformers as the text encoder and 24-layer vision transformers as the image encoder. The text encoder and the image encoder are both initialized by CLIP. A GPT-2 model is utilized to generate explanations. The framework of the FigCLIP model is shown in Figure 3.

Specifically, a given claim is input to the text encoder and a claim vector v_{claim} is output. A given image is fed into the image encoder and an image

vector v_{image} is output. For label prediction, the FigCLIP model needs to consider whether the claim is semantically entailed by the image. To fuse the deep representations of the claim and the image, a bidirectional fusion module with 8-layer cross-attention is designed. The fusion process is divided into two steps. The claim vector v_{claim} serves as Q, and the image vector v_{image} serves as K and V. They are fed into the fusion module and then a claim-guided multimodal vector $v_{\text{multi}}^{\text{claim-gui}}$ is calculated by $\text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_k}}V\right)$, where d_k denotes the dimension of 768. This claim-guided multimodal vector achieves an effective interaction of observing details in images based on text. Similarly, the image vector v_{image} serves as Q, and the claim vector v_{claim} serves as K and V. They are fed into the fusion module and then a image-guided multimodal vector $v_{\text{multi}}^{\text{image-gui}}$ is calculated by the same cross-attention calculation process. This image-guided multimodal vector achieves an effective interaction of observing details in text based on images. These two mentioned-above steps share parameters, enhancing the alignment of figurative image-text representations. After that, the $v_{\text{multi}}^{\text{claim-gui}}$ and the $v_{\text{multi}}^{\text{image-gui}}$ are concatenated and input to a binary linear-layer classifier to predict a label of entailment or contradiction.

The original representation space of CLIP is inconsistent with that of GPT-2. GPT-2 relies on a 50257-dimensional vocabulary to generate text, while the CLIP multimodal space is 768-dimensional. For generating explanation, the FigCLIP model needs to match the low-dimensional multimodal representations to 50257 dimensions in a mapping space. Inspired by ClipCap (Mokady et al., 2021), we stack multiple linear layers of different dimensions as a projector. This projector is composed of three sets of linear layers of (768*2→2048), (2048→4096), (4096→50257). In order to further compress the size of parameters to reduce training costs, the parameters of this (4096→50257) linear layer are frozen and treated as a fixed matrix. This is the reason why FigCLIP is more lightweight than ClipCap, despite their similar model architectures. The multimodal representations after projector mapping is fed into GPT-2 to generate an explanation about why the image and claim are semantically entailed or contradicted.

Loss. Two cross-entropy losses are defined to optimize the FigCLIP model jointly, comprising a classification loss ($\mathcal{L}_{\text{cls}}$) and a generation loss ($\mathcal{L}_{\text{gen}}$). The predicted labels and the ground-truth labels are used to calculate the classification loss, which can promote semantic alignment between images and claims to learn more fine-grained details of entailment or contradiction. The generated explanations and the ground-truth explanations are used to calculate the generation loss, which also can facilitate the mapping of multimodal deep representations to establish a reliable mapping space. Finally, the sum of the $\mathcal{L}_{\text{cls}}$ and the $\mathcal{L}_{\text{gen}}$ is regarded as the complete training objective.

Algorithm 1: Pseudocode of Training FigCLIP

data : a claim c, an image i;
a ground-truth label l_{gt}, a ground-truth explanation e_{gt};

while c, i, l_{gt}, e_{gt} **do**
  # the claim vector
  $v_{\text{claim}} \leftarrow$ Text-Encoder(c);
  # the image vector
  $v_{\text{image}} \leftarrow$ Image-Encoder(i);
  # the claim-guided multimodal vector
  # the parameter order is Q, K, V
  $v_{\text{multi}}^{\text{claim-gui}} \leftarrow$ Fusion(v_{claim}, v_{image}, v_{image});
  # the image-guided multimodal vector
  # the parameter order is Q, K, V
  $v_{\text{multi}}^{\text{image-gui}} \leftarrow$ Fusion(v_{image}, v_{claim}, v_{claim});
  # the concatenated multimodal vector
  $v_{\text{multi}} \leftarrow v_{\text{multi}}^{\text{claim-gui}} + v_{\text{multi}}^{\text{image-gui}}$

  # the classification loss
  $label \leftarrow$ Classifier(v_{multi});
  $\mathcal{L}_{cls} \leftarrow$ CrossEntropyLoss($label$, l_{gt});

  # the generation loss
  $v_{\text{multi}}^{\text{mapping}} \leftarrow$ Projector(v_{multi});
  $explanation \leftarrow$ GPT-2($v_{\text{multi}}^{\text{mapping}}$);
  $\mathcal{L}_{gen} \leftarrow$ CrossEntropyLoss($explanation$, e_{gt});

  # the complete training objective
  $\mathcal{L} \leftarrow \mathcal{L}_{cls} + \mathcal{L}_{gen}$;
end

The FigCLIP model enables end-to-end training because it can jointly address the problems of label prediction and explanation generation. The whole training procedure of the PigCLIP model can be abstracted in Algorithm 1.

4 Experiments and Results

4.1 Datasets

According to the official data description, the training data is compiled from the following five datasets about visual metaphors and multimodal understanding:

(1) a subset of 731 Visual Metaphors dataset (Chakrabarty et al., 2023);

(2) a subset of 1,322 textual metaphors with images (Yosef et al., 2023);

(3) a susbet of 853 memes with annotated claims and explanations (Hwang and Shwartz, 2023);

Data Source	Train/Valid		Test	
	absolute	proportion	absolute	proportion
nycartoons (Hessel et al., 2023)	520	11.7%	87	12.6%
irfl (Yosef et al., 2023)	1322	29.9%	198	28.7%
muse (Desai et al., 2022)	1000	22.6%	150	21.8%
memecap (Hwang and Shwartz, 2023)	853	19.3%	128	18.6%
vismet (Chakrabarty et al., 2023)	731	16.5%	126	18.3%
total	4426	100%	689	100%

Table 1: The statistical details of the datasets for the MFL task.

Model	V-FLUTE test set (%)		
	F1	F1@50	F1@60
jalor	90	89	75
FigCLIP$_{336\times336}$	70	67	50
FigCLIP$_{224\times224}$	68 (-2)	65 (-2)	49 (-1)
GPT-4V (zero-shot)	70	64	49
mrshu	63	62	43
yangst	51	48	31
LlaVA (baseline)	45	38	21

Table 2: Evaluation results on the V-FLUTE test set.

(4) a subset of 1,000 sarcastic captions with images (Desai et al., 2022);

(5) a subset of 520 unique images with captions accompanied with textual explanations (Hessel et al., 2023).

The test data is available at huggingface[3]. Table 1 displays the statistical details of the datasets (named V-FLUTE (Saakyan et al., 2024)) for the MFL task.

4.2 Settings

Our model is implemented on Pytorch 2.0.1 and only one RTX 4090 GPU. Both the text encoder and image encoder are initialized by CLIP-ViT-L/14 or CLIP-ViT-L/14@336px (Radford et al., 2021). All parameters of the text encoder and GPT-2 are optimized, while the image encoder is completely frozen for reducing the training costs. The batch size is set to 32, and the epoch is set to 20. AdamW is applied to optimize model parameters with a learning rate of 1e-04 and weight decay of 0.05. The image resolution is specified as 224×224 or 336×336, and the maximum text length is set to 77. Following previous work (Saakyan et al., 2022), three metrics are used to evaluate the model performance, including F1@0 (pure F1 score), F1@50 (F1 score computed where only instances which had their explanation match the reference with BERTscore (Zhang et al., 2019) above 50 are counted as correct), and similarly F1@60.

4.3 Results

The official evaluation results are reported in Table 2. Our submission ranked second on the leaderboard, where the FigCLIP model was initialized by CLIP-ViT-L/14@336px. The FigCLIP$_{336\times336}$ model achieved 70% F1-score, 67% F1@50-score and 50% F1@60-score on the benchmark test set. LlaVA, the official baseline, only obtained 45% F1-score, 38% F1@50-score and 21% F1@60-score by zero-shot. This means that LlaVA can be applied to this task but it is not proficient in multimodal figurative language understanding. Nevertheless, the FigCLIP$_{336\times336}$ model outperformed LlaVA by 25% F1-score, 29% F1@50-score and 29% F1@60-score respectively. Compared with GPT-4V (a state-of-the-art model in image-text understanding), the FigCLIP$_{336\times336}$ model leaded by 3% and 1% in F1@50-score and F1@60-score respectively, even though their F1-scores ware the same. It is worth noting that calling GPT-4V's API for zero-shot on the test set took approximately $19 and 2 hours, while training an epoch of the FigCLIP model only took less than 1 minute on one 24GB GPU. This demonstrates the low cost and effectiveness of our method. Moreover, we initialized FigCLIP using CLIP-ViT-L/14 to explore the impact of low resolution (224×224). We found that all three metrics dropped slightly when understanding images at low resolution. This shows that the FigCLIP$_{336\times336}$ model can capture more subtle image semantics and facilitate the identification of fine-grained implication relationships with claims.

5 Conclusion

This paper propose an end-to-end model FigCLIP for the FigLang-2024 Multimodal Figurative Language shared task. We introduce a shared bidirectional fusion module with cross-attention to advance the alignment of figurative image-text pairs. In the mapping space defined by CLIP and GPT-2, we utilize a projector to bridge multimodal representations and explanation representations and make FigCLIP lightweight. Experimental results on the benchmark test set demonstrates the effectiveness of our method, which achieves competitive performance and outperforms GPT-4V. Moreover,

[3] https://huggingface.co/datasets/ColumbiaNLP/V-FLUTE-test

understanding images at high resolution has been proven to be beneficial for capturing more fine-grained details of figurative language.

Limitations

To alleviate the training burden and reduce training costs, the image encoder was completely frozen. This may prevent the model from learning richer and more accurate knowledge of multimodal figurative language. Limited by the short duration of this task, we did not explore the impact of different generative models on model performance. In future work, we will optimize the different layers of the image encoder to find the optimal trade-off between performance and cost. Furthermore, we will replace the current generative model with several large language models such as Llama and Vicuna to enhance FigCLIP's generalization ability in understanding and explaining multimodal figurative language.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: A visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Irina Bigoulaeva, Rachneet Singh Sachdeva, Harish Tayyar Madabushi, Aline Villavicencio, and Iryna Gurevych. 2022. Effective cross-task transfer learning for explainable natural language inference with t5. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 54–60, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022a. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021. Figurative language in recognizing textual entailment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022b. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7370–7388, Toronto, Canada. Association for Computational Linguistics.

Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10563–10571.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yuling Gu, Yao Fu, Valentina Pyatkin, Ian Magnusson, Bhavana Dalvi Mishra, and Peter Clark. 2022. Just-DREAM-about-it: Figurative language understanding with DREAM-FLUTE. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 84–93, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

EunJeong Hwang and Vered Shwartz. 2023. Memecap: A dataset for captioning and interpreting memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445.

Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, CHEN Bin, Chengru Song, Di ZHANG, Wenwu Ou, et al. 2023. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. In *The Twelfth International Conference on Learning Representations*.

Jeongyeon Kim, Sangho Suh, Lydia B Chilton, and Haijun Xia. 2023. Metaphorian: Leveraging large language models to support extended metaphor creation for science writing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, pages 115–135.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Khoa Thi-Kim Phan, Duc-Vu Nguyen, and Ngan Luu-Thuy Nguyen. 2022. NLP@UIT at FigLang-EMNLP 2022: A divide-and-conquer system for shared task on understanding figurative language. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 150–153, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.

Arkadiy Saakyan, Tuhin Chakrabarty, Debanjan Ghosh, and Smaranda Muresan. 2022. A report on the FigLang 2022 shared task on understanding figurative language. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 178–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2024. V-FLUTE: Visual Figurative Language Understanding with Textual Explanations Dataset. `https://huggingface.co/datasets/ColumbiaNLP/V-FLUTE`. Dataset associated with the paper "V-FLUTE: Visual Figurative Language Understanding with Textual Explanations".

Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. IRFL: Image recognition of figurative language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1044–1058, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.

The Register-specific Distribution of Personification in Hungarian: A Corpus-driven Analysis

Gábor Simon

Abstract

Although several promising initiations have been proposed recently about how to identify personifications, a comprehensive corpus linguistic analysis of personifying meaning generation still have to be carried out. The aim of the paper is twofold: (i) to present an extended version of the PerSE corpus, the language resource for investigating personification in Hungarian; (ii) to explore the semantic and lexicogrammatical patterns of Hungarian personification in a corpus-driven analysis, based on the current version of the research corpus. PerSE corpus is compiled from online available Hungarian texts in different registers including journalistic (car reviews and reports on interstate relations) and academic discourse (original research papers from different fields). The paper provides the reader with the infrastructure and the protocol of the semi-automatic and manual annotation in the corpus. Then it gives an overview of the register-specific distribution of personifications and focuses on some of its lexicogrammatical patterns.

1 Introduction

Despite its apparent clarity, the category of personification is far from being simple and homogeneous. In the last decades, at least four different conceptual models of personifying meaning-making have been proposed in cognitive linguistics. Beyond the general metaphorical explanation (personification is an ontological conceptual metaphor with a human being as its source domain, see Kövecses, 2010) there is an alternative model within the framework of conceptual metaphor theory (based on the EVENTS ARE ACTIONS generic-level metaphor, see Lakoff, 2006), but a metonymic (Low, 1999) and a conceptual integration model (Long, 2018) are also available in the literature. Moreover, a solid methodological framework for identifying personifications in texts has been proposed by Dorst et al. (2011). However, systematic and extended research on the linguistic variability of personification has not been carried out yet. Although the protocol for identification may serve as a promising vantage point for a comprehensive corpus study, there is not any available language resource in terms of personification annotation. The present paper aims to fill this gap by proposing an extended version of the PerSE[1] corpus, a new language resource for studying personifying language use in Hungarian. Beyond merely demonstrating the corpus, some initial analyses of the register-specific patterns of personification in Hungarian are provided here, too.

The study is based on the extended and improved version of the PerSE corpus introduced previously (Simon, 2022). The former study provided the reader with the annotation protocol, the basic infrastructure of the corpus and some preliminary results of personification identification in a pilot sample of only one register. Compared to it, this paper demonstrates the annotation of personifications on a relatively larger scale (analyzing three different registers) and with advanced infrastructure. This modest expansion of the corpus made it possible to consider the register-specificity of personifying language use. Consequently, the scope of the study also encompasses the quantitative analysis of

[1] The name of the corpus is the abbreviation of the phrase „Personifying Structures Encoded".

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 99–109
June 21, 2024 ©2024 Association for Computational Linguistics

personifying language use within and between registers. For the latter, both a whole corpus design and a linguistic feature design (Brezina, 2018) have been implemented.

The paper is structured as follows. After the introduction, the basic notions and principles of the analysis are discussed (2). Then the material and the methodology of the study are detailed, including corpus building, annotation and quantitative analysis (3). The fourth section deals with the results of the analysis, and the paper ends with some concluding remarks (5).

2 Theoretical Background

According to the glossary of Kövecses's volume on conceptual metaphor theory, personifications "involve understanding nonhuman entities, or things, in terms of human beings. They thus impute human characteristics to things" (Kövecses, 2010). This very basic definition needs to be detailed with the further aspects of personifying meaning-making: it attributes agency to non-human entities (Dorst, 2011), it can rely on the metonymic link between human and non-human entities, and it can be conventionalized in different degrees ("dead personifications", see Dorst, 2011). Therefore, personification as a semantic phenomenon is much more complex than it is implied in its definition.

Dorst et al. (2011) operationalize the notion in the following way: if the basic meaning of a lexical unit is human-oriented (i.e., the primary figure of the meaning is typically a human being), and the contextual (or actual) meaning of the unit refers to a non-human entity, it can be labelled as personification. By way of explanation, identifying personification in discourse is a specific process of word sense disambiguation, rendering it possible not only to highlight personifications in a text but also to categorize them in terms of conventionalization. If both the human basic meaning and the non-human contextual meaning are offered by the dictionary, the personification can be considered a conventionalized one. In (1) invasion is a military process in its basic meaning, but it has a more general meaning in the dictionary as well ('Someone or something appears somewhere en masse'), therefore the personifying usage of the noun is conventional. However, if only the human basic meaning can be found in the dictionary, the expression is rather a novel personification. In (2) (referring to an engine of a car) greedy has only a human-related meaning in the dictionary ('[someone] trying to satisfy their desire ardently'), thus the personifying usage of the adjective has not been lexicalized yet. In the case of not referring directly to human beings in the description of the basic meaning by the dictionary, but the prototypical or default figure is human, the personification belongs to the default type. The basic meaning of develop (3) is '<living being> grows, their features evolve gradually', in which the primary figure is not explicitly a human being, but in its default interpretation, it refers typically to people, therefore it has a default personifying usage. Finally, (4) illustrates metonymic personification with the reference of Russia to the leaders or the members of the Russian army.

(1) biológia-i inváz**ió**
biology-ADJ invasion
'biological invasion'

(2) nem egy mohó szerkezet
not a.DET greedy gear
'not a greedy gear'

(3) harc-művészet-i hagyomány-ok fejlőd-t-ek
fight-art-ADJ tradition-PL develop-PST-3PL
'[the] traditions of martial arts developed'

(4) Oroszország helikopter-t veszít-ett
Russia helicopter-ACC loose-PST.3SG
'Russia lost [a significant amount of] helicopters'

This lexical semantic approach to personification provides a solid theoretical foundation for a corpus-driven analysis of the linguistic patterns of personifying language use, ensuring such a scope that is broader than in previous research. Low (1999), for instance, considers metonymic and personifying readings as alternatives in meaning-making. In the cognitive linguistic research of the discourse on interstate relations (Twardzisz, 2013) metonymy is completely excluded from the realm of personification. As a consequence of the latter decision, the personifying use of state names proves to be rather infrequent in the journalistic corpus of the previous analysis. However, the present study adopts such an operationalization of the notion of personification that results in a better recall of the corpus analysis without decreasing the level of precision.

The chosen theoretical and methodological orientation is based on the MIPVU protocol for

metaphor identification (Steen et al., 2010), and since a systematic and comprehensive analysis sheds light on the register-specific patterns of metaphorization in English, we can assume that personifying language use will also show different realizations in terms of discourse types in Hungarian. As Steen et al. (2010) observe, the academic register has the highest percentage of metaphor-related words, while fiction is only the third on the list regarding the frequency of metaphors. News texts are almost as metaphorical as academic papers, and conversation is the least metaphorical. Compared to the previous research, the PerSE corpus represents two broad fields of discourse on a higher level of granularity than the Amsterdam Metaphor Corpus: from journalism, I analyzed two specific registers (car reviews and reports on interstate relations), while from academic discourse I sampled original research papers from a wide range of scientific fields, including natural and social sciences.

Considering the linguistic structure of personification, we can claim that it is not limited to only one word in the discourse. In a previous experimental study, Dorst et al. (2011) found that 61.90% of the personifications identified by the informants were word combinations. From a rather theoretical point of view, Long (2018) defines personification as an "extended unit of meaning" (relying on Sinclair's notion), encompassing a node word and its collocations. In my previous analysis (Simon, 2022), personification-related arguments were slightly more frequent than words related directly to personifying meaning, which means that on average every personification had at least one argument in the corpus. Consequently, the present study also focuses on personifications as potentially multi-word expressions and provides data not only on the raw frequencies of personifications in different texts but also on their size and distribution in terms of node and argument structure.

As a result of the overview of the theoretical background of the present study, the central research question is as follows: what are the differences between the patterns of linguistic personification in different registers in Hungarian? This general question can be answered from more than one perspective, regarding the distribution and frequency of personifications on the one hand (whole corpus design, see Brezina, 2018), and on the other hand taking the lexicogrammatical features of personification in different registers into consideration (linguistic feature design, see Brezina, 2018). In this paper, I apply both points of view.

3 Material and Methods

Before turning to the results of the annotation process and the corpus analysis, I introduce the language resource that served as the basis of the research: the PerSE corpus. The process of corpus building, the infrastructure of the research, the annotation protocol and the methods of the quantitative analysis are outlined in this section.

3.1 Sampling and Research Infrastructure

The overall aim of the research is to explore systematically, in a corpus-driven way how personifying meanings are symbolized in Hungarian, i.e. what are the central linguistic patterns of personification in this language. To discover these patterns, it is essential to sample texts from as wide a repertoire as possible. The pilot version of the corpus was compiled from online car reviews written in Hungarian (see Simon, 2022 for a detailed description of this version), representing a variety of personifications in Hungarian without any reference to their register-specific character. The extended version of the corpus includes Hungarian reports on interstate relations published in an online daily news site, which makes a comparison of personifications across journalistic registers possible.[2]

There was only one aspect for sampling in the interstate relations subcorpus: the article needed to describe a prominent geopolitical event or scenario in the time frame of the sampling period (from 2022 October to 2023 June). 6 reports were chosen with the topics of Italian politics (the political agenda of the Meloni government), the French-German relationship, British politics (the political agenda of the Sunak government), the war in Ukraine and the legal investigation against Donald Trump.

Moreover, the PerSE corpus contains online available Hungarian academic texts as well, namely original research papers from online journals in the following fields of research: health

2 It is worth noting that the complete version of the corpus will consist of literary fiction and conversations, too.

sciences, dentistry, hydrology, orientalism and conservation biology. Here, the criteria of sampling were more complex: (i) the journal needs to have an open access declaration; (ii) the paper needs to be published under a CC BY licence; (iii) the size of the paper needs to be short or medium; (iv) the paper needs to be published recently (from 2020 to 2023). The papers fulfilling these criteria were sampled and converted to .txt format. Diagrams, tables, figures, original quotations not written in Hungarian and references were omitted from the samples. Although the abstracts frequently repeat some sentences from the main text, they constitute an essential component of the genre, therefore the papers were sampled to the corpus with abstracts (and keywords, if there were any). Appendix A presents the size of the corpus and its subcorpora.

The plain texts sampled into the corpus were automatically processed with the e-magyar digital language processing system[3] (Indig et al., 2019) with the preset "Raw text to dependency parsing in CoNLL-U format using Stanza Dependency parser". The automatic preprocessing thus included tokenization, lemmatization, PoS tagging and morphosyntactic analysis. The results of the preprocessing were exported in CoNLL-U format. For the manual annotation of personifications, the INCEpTION platform was used (Klie et al., 2018). (The reliability test of the procedure was carried out in WebAnno (Eckart de Castillho et al., 2016)). According to the annotation protocol, I used the Concise Dictionary of Hungarian (Pusztai ed. in chief, 2003) for word sense disambiguation. To estimate the idiomaticity of linguistic personifications, the Hungarian National Corpus[4] (v2.0.5, Oravecz, Váradi and Sass, 2014) was used as a reference corpus.

The integrated result of the automatic preprocessing and the manual annotation was exported in WebAnno TSV v3.3 file format and further analyzed in MS Excel. Statistical analysis was carried out in R (v4.1.0, R Core Team, 2021).

3.2 Annotation Procedure

In this subsection, I give a brief overview of the annotation process, for a detailed description see Simon (2022). The procedure, which is based on the identification protocol proposed by Dorst et al. (2011), and borrows some elements from a MIPVU-inspired, language-specific metaphor annotation protocol (the MetaID protocol, Simon et al., 2023) as well, has the token as its basic unit, and relies on the previously described operationalization of the notion of personification.

The annotation is carried out on three layers (summarized in Appendix A). First, the components of linguistic personification are labelled (ptags). Then the annotator analyses the semantic relations (prel) between the components (if the actual target is a multi-word expression), using the basic semantic categories of cognitive grammar (Langacker, 2013): the trajector for the primary figure of a process or relation (basically, it is the agent) and the landmark for the secondary figure (the patient, experient, recipient of the process, or the element of the setting in the represented scenario). Since possessive relation is also frequent in personification (mainly in body-part constructions), it receives a distinct label. (A technical label for separated elements of a construction (e.g., preverbs) was also used in the procedure, but it is of peripheral importance, thus, I omitted it from the analysis.) Lastly, the semantic quality of personifications is classified by the aforementioned four categories (conventional, novel, default and metonymic personifications, pqual).

Since the semantic categories of personification have been discussed in section 2, here I focus rather on components and their labels. PRW refers to personification-related words, all tokens that initiate personifying meaning-making. For example, in (2) the adjective *mohó* ('greedy') personifies the concept of the engine.

PRA is for all the tokens semantically and grammatically linked to an initiator of personification (labelled as PRW), and contribute to the elaboration of a personifying meaning. As an example, the *hagyományok* ('traditions') in (3) constitutes an argument of a multi-word personification.

PRWid stands for idiomatic personification, i.e., for those units that are considered an element in a prefabricated structure in Hungarian. Prefabricatedness is measured by exploring the collocational behavior of the candidate expressions in the reference corpus (see section 3.1). Collocations are identified by the logDice score

[3] The pipeline is available here: http://emtsv.elte-dh.hu:5000/ (last access: 01/03/2024).

[4] The HNC corpus is available here: https://clara.nytud.hu/mnsz2-dev/ (last access: 01/03/2024).

(Rychlý, 2008) with a threshold of 6. PRAid refers to the idiomatic counterpart of the PRA category.

PRWimp stands for implicit personification: in this case, the token refers to some other labelled token via coreference, therefore it implicitly conveys personifying meaning. For instance, in (5) the verb *szankcionálták* ('sanctioned') refers back to the noun *törvények* ('laws') in the context, therefore it has an implicit personification in its semantic structure.

(5) szankcionál-t-ák […] bűn-cselekmény-ek-et
sanction-PST-3PL […] crime-action-PL-ACC
'[they] sanctioned criminal acts'

In the Hungarian construction, *Németország befelé fordul* ('Germany turns inward') the word *befelé* ('inward') collocates with the verb *fordul* ('turn') with a logDice score of 7.716, which means that in the expression the verb can be labelled as PRWid, the adverb *befelé* can be identified as PRAid, and the noun *Németország* ('Germany') is a simple argument of the personification. Additionally, within the idiomatic expression, there is a landmark relationship (*befelé* ('inward') symbolizes the orientation of the action as a container), whereas the nominal component explicates the trajector of the process.

The reliability of the annotation process was tested with two annotators, one of them was the author of the present paper, and the other was a university student who learned the procedure during a workshop and practiced it alone. The test was performed in one text sampled to the corpus, the annotators worked independently. The inter-annotator agreement was automatically calculated in Cohen's Kappa by the WebAnno platform. At two layers out of three, the annotation has good reliability (above or very close to the threshold of 0.8: at the layer of the components, it was 0.79, at the layer of the relations it was 1). Regarding the decision about the semantic quality of personifications, the test demonstrated a more modest agreement (with a Kappa measure of 0.68), but even in this case, the procedure seems to be tentatively reliable (Artstein and Poesio, 2008). Further improvement of the annotation protocol needs to be done in the future to improve the reliability of the process in the latter case, especially in terms of the sematic categories of personification.

3.3 Methods of Quantitative Analysis

Manual annotation is a time-consuming process, and the identification of personifications requires a lot of effort from the analyst. In the PerSE corpus, a relatively small-scale language resource, the sample sizes are low, and the distributions of the data are not normal in every case. As a consequence, only non-parametric statistical tests can be taken into consideration, if we are interested in register-specific tendencies of personifying language use in Hungarian.

As a baseline, two-tailed Wilcoxon tests were performed in pairs of the subcorpora. Then, a non-parametric one-way ANOVA was carried out (with a non-parametric post hoc test) to shed light on the between-group variability of personifications in the whole corpus. Note that only the pqual data have been tested in this way since the preliminary visualizations suggested significant differences only at this layer.

Considering the lexicogrammatical patterns of personification across registers, I focused on the four most frequent personified verbs in all three subcorpora and analyzed their register-specific personified use with Pearson's Chi-squared tests (see Brezina, 2018).

4 Results and Discussion

After introducing an extended version of the PerSE corpus, this section demonstrates why such a language resource is useful in researching figurative language use (especially in cognitive linguistics). First of all, I explore the distributions of the allocated labels in the entire corpus. The statistical testing of register-specific differences is the next focus of the analysis. Finally, I zoom in on some lexicogrammatical patterns of personification in the corpus.

4.1 The Frequency of Personifications

If we are interested in the overall tendencies of personifying language use in the corpus, we can observe that reports on interstate relations use personifications most frequently (34,759.33 per million words), then come the car reviews (with a relative frequency of 33,985.32 pmw), and the least personifying language use is characteristic of research papers (28,267.87 personifications pmw). Thus, the distribution of personifications across registers appears to be the reverse of the tendencies

of metaphorization observed by Steen et al. (2010) in a previous study: although they didn't find great differences between academic discourse and news texts in terms of metaphor use, the former proved to be the most metaphorical in their research, followed by the latter register. In the PerSE corpus, however, academic texts are relatively deficient in personifying language use compared to the two other journalistic registers. However, these numbers hide important within-group differences: one of the three texts being the richest in personifications comes from the academic subcorpus (after a car review and a report). Since natural and health sciences are overrepresented in the research paper subcorpus, and since the discussion of these topics seems to be weak in personifying language use, one possible explanation behind the overall frequency distribution is that humanities prefer personification more than sciences. Meanwhile, however, the very technical language of car reviews also gives a lot of personifications, thus, any absolute distinction would be hard to make.

The distribution of the components of personifications indicates slight differences. Regarding the relative frequencies of the allocated component labels in the corpus, it is a general tendency that the arguments outnumber the node words of personifications. But this ratio is the greatest in reports on interstate relations (2.60% : 5.24%, which means that almost 2 arguments are linked to a node on average in this subcorpus), and the lowest in car reviews (3.00% : 3.99%, the

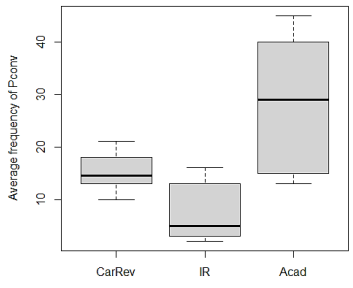


Figure 1: Box plots of the frequencies of pconv label

average argument of a node is close to 1), and the research papers are in between these two extremes (2.30% : 3.85%, the average number of arguments per node is slightly above 1.5). In other words, interstate reports provide the most extended (and elaborated) personifications.

The other issue of the distribution of the components is idiomaticity: again, reports contain the highest number of idiomatic personifications (with a relative frequency of above 0.70%); car

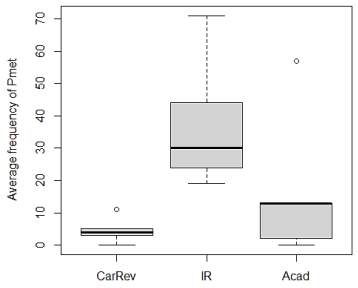


Figure 2: Box plots of the frequencies of pmet label

reviews almost lack idiomatic personifications (the average relative frequency of them is 0.22%); while research papers come close to reports in this respect (with 0.675% average relative frequency). We can claim, thus, that political journalism prefers prefabricated expressions in personifying language use the most.

The difference in allocated semantic relations between the subcorpora is not remarkable. In general, trajector labels are more frequent than landmarks in the entire corpus, which demonstrates that the personified entity (symbolized as the trajector of a process or relationship) receives linguistic elaboration in a higher percentage of the cases than other participants of the scenario. The possessive relationship reaches its maximum in car reviews, due to the preference of body-part personifications in this register.

4.2 The Semantic Quality of Personifications

Based on the observed diversity in idiomaticity in the three subcorpora, I tested first the hypothesis of whether there are statistically significant differences between registers in terms of idiomatic personifications. According to a two-tailed Wilcoxon test, interstate reports use a significantly higher number of idiomatic arguments in personifications ($W=3.5$, $p<0.05$) than car reviews, but no further significant differences could be observed either in other labels or between other registers. This means that there are no other

remarkable structural register-specific patterns of personification.

However, moving on to the semantic quality of the identified expressions, we can assume that the investigated registers significantly differ in more than one aspect.

As can be observed, default personifications are distributed evenly in the corpus, while the other three categories seem to be preferred by different subcorpora. Figures 1-3 show the frequency of the conventional, the metonymic and the novel personifications. It is clear that conventional personifications dominate in the research papers, metonymic expressions are overrepresented in interstate reports, while novel personifications belong to the register of car review.

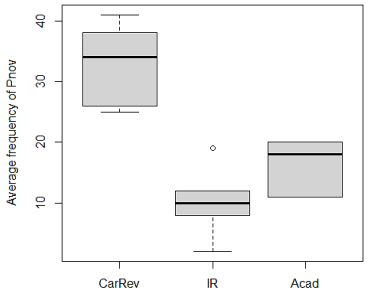


Figure 3: Box plots of the frequencies of pnov label

In the case of conventional personifications, the non-parametric one-way ANOVA did not result in any significant difference ($F(2, 6.7999)=4.7349$, $p>0.05$), thus, the register does not have a significant effect on the conventionality of personifications. However, the register affects the distribution of metonymic personifications ($F(2, 5.5681)=6.0275$, $p<0.05$), and the frequency pattern of novel personifications is affected, too, by it ($F(2, 8.5015)=18.871$, $p<0.001$). Considering the post hoc tests reports on interstate relations contain significantly more metonymic personifications than car reviews ($p<0.001$), while the latter register uses significantly more novel personification than the other two ($p<0.001$) according to a non-parametric version of the Tukey HSD test.

4.3 Verbal Personifications in the Corpus

As a language resource, the PerSE corpus provides data not only about the general distributional patterns of personification in different registers but also about the register-related personifying behavior of specific linguistic structures. While studying the frequencies of the allocated labels can be considered a top-down analysis, personifying language use can be explored from a bottom-up perspective as well, in which the frequency of personification is observed by concrete words. The latter orientation makes it possible to characterize the lexicogrammatical features of personification in Hungarian, e.g., the part-of-speech categories associated with personifying meaning, or the complexity of personifications as constructions. Due to the limitations of the present paper, I provide the reader only with a brief, rather illustrative analysis of the personifying and non-personifying use of some basic Hungarian verbs in the corpus. This analysis can be considered neither exhaustive nor comprehensive, but it may shed light on the perspectives the corpus can open for cognitive corpus linguistics.

First, relying on the verb frequency lists of the corpus, four verbs were selected for further analysis, because they belong to the most frequent verbs in all three subcorpora. The verbs are the following: *tud* (‘know/can’), *ad* (‘give’), *tesz* (‘put/do’) and *vesz* (‘take’). Then I counted all the occurrences of these verbs in the corpus, considering both their personifying and non-personifying use.

The basic tendency of all four verbs is that the non-personifying use is more frequent than personification. There are only two exceptions: *tud* (‘know/can’) is more associated with personification in car reviews, and *tesz* (‘put/do’) is rather personified in research papers. The effect size (Cramer’s V) was moderate in both cases (0.467 and 0.326, respectively). I have found a significant association only between the personifying use of the verb *tud* and register:

$X^2(2)=14.841$, $p<0.001$. Figure 4 shows the register preferences in the usage of the verb *tud*.

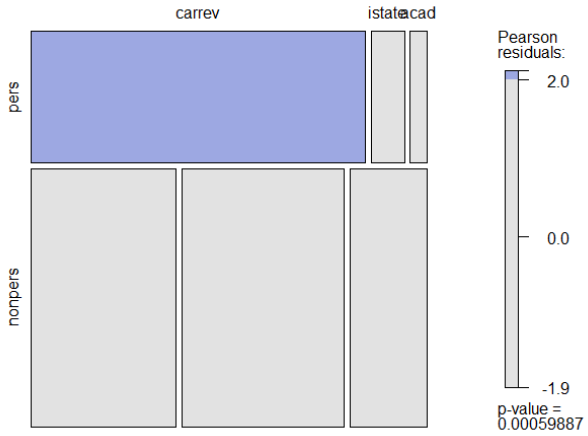


Figure 4: Mosaic plot of personifying and non-personifying use of *tud* ('know/can') in the corpus

This analysis of the linguistic features of personification across registers opens only a small window to the lexicogrammatical patterns of personifying language use in Hungarian. However, it may illustrate the potential of having a manually annotated research corpus of personifications. Moreover, it demonstrates one interesting aspect of register-specificity in personifying language use: cars (and their components) are described as human beings with physical and mental capacities in car reviews, and it is a significant linguistic pattern that cannot be found in other registers.

5 Limitations of the study

The PerSE corpus can be considered a new language resource in Hungarian that makes it possible to analyze the expressions of personification within and across different registers in a systematic way. It has, however, three major limitations that need to be addressed here. First, it is the manifestation of in-progress research, which means that other texts will be sampled into it on the one hand, and on the other, it needs to be made available for the broader research community. In its present version, it is rather a modest-scale research corpus, thus, the long-term goal of corpus compilation is to provide an open-source database designed to support cognitive corpus linguistic investigations.

Secondly, further refinement of the annotation procedure needs to be carried out to increase the reliability of the identification process and create a benchmark for cross-linguistic exploration of personification. This also means that not only do additional annotators have to be involved in the curation phase of the corpus but also that the reliability of personification identification needs to be tested via alternative empirical (experimental and/or questionnaire-based) psycholinguistic methods as well. The dictionary-based analysis demonstrates currently that the precise identification of potential personifications is feasible, but whether the annotated expressions are true personifications in actual discourse comprehension has remained an open question.

Finally, the corpus paves the way for automatic personification detection providing a precise and comprehensive data set for training large language models to this task. However, it is not clear whether these models would gain enough information from the corpus to produce good results, and if so, how this development could contribute to the current NLP field. Nevertheless, personification is a pervasive phenomenon in discourse, which makes its identification a good start for improving text classification or observing the patterns of how (mental) health issues or other negative factors are construed figuratively in everyday life.

6 Conclusion

Personification is a complex phenomenon in terms of both conceptualization and linguistic organization. Thus, cognitive and corpus linguistics need to cooperate in exploring the functioning of personifying meaning-making. The PerSE corpus is a unique language resource for analyzing personification in Hungarian from a

corpus-driven perspective. With its extended size (exceeding 30,000 tokens), genre and register variability (including technical, political and scientific language use, but also formal and informal styles in three different registers) and hybrid annotation design (extending to automatic preprocessing of grammatical features and manual identification of personifications as well) the PerSE corpus provides the analyst with a vast amount of information on personification, making it possible to approach it from different theoretical perspectives and with a wide range of methods. The present paper introduced the new, extended version of the corpus, outlining its methodological framework, and demonstrated how to exploit this language resource in a corpus-driven analysis.

The corpus annotation relies partly on automatic language processing, and hence the texts in the corpus can be analyzed on different levels of granularity (from lexical density based on tokenization and lemmatization to word class categories, and morphological and syntactic analyses). The identification of personification is based on the operationalization of the notion proposed in the literature. The dictionary-based word sense disambiguation maximizes the transparency of the annotation while minimizing its intuitive nature. Moreover, the protocol extends the task of identification to measuring the idiomatic character of personifying expressions and allocating semantic relation labels in the corpus. Thus, not only the lexicogrammatical patterns of personification can be observed but also their internal semantic organization and their prefabricatedness. This line of analysis can lead us toward the exploration of the construction-like behavior of personification in Hungarian in the future.

Compared to its pilot version, the extended PerSE corpus sheds new light on the language-internal variability of personification as well. The most important findings in this regard are as follows. (i) Journalistic registers use more personifications than academic discourse (although register-specificity is assumable based on the observations). (ii) Personifications in academic texts and interstate reports appear to be more complex in their linguistic structure with a stronger tendency to use idiomatic patterns of Hungarian. (iii) Register has a significant effect on the semantic quality of personification: interstate reports prefer metonymic personifications whereas car reviews exploit the potential of novel personifications. (iv) Some frequent Hungarian verbs are associated more with personifying language use in particular registers (e.g., the verb *tud* ('can/know') in car reviews).

The PerSE corpus also provides a solid methodological grounding for an even more extended analysis of Hungarian personifications in the future. The closest aim of the author of the present paper is to sample literary texts into the corpus and test the well-known assumption that literature would be the richest source of figurative language use. Additionally, the corpus may serve as an input data set for improving large language models in the direction of detecting and automatically identifying personification in language. In other words, the PerSE corpus as a reliable language resource with precise and multi-faceted processing of lexicogrammatical features can be used as a corpus for training existing NLP resources in Hungarian toward automatic personification annotation. Finally, the design of the corpus and the identification protocol may serve as a vantage point for creating other similar language-specific resources in various languages, bringing personification onto the top of the agenda of cross-linguistic cognitive corpus analyses. This would motivate the reinterpretation of personification as figurative language use evaluating it not as a subtype of metaphor but rather as a complex and colorful phenomenon, which is worth investigating in its own right.

Acknowledgments

The research was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (BO/00382/21).

References

Ron Artstein, and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34(4):555–596.

Vaclav Brezina. 2018. *Statistics in Corpus Linguistics. A Practical Guide.* Cambridge University Press, Cambridge, UK.

Aletta G. Dorst. 2011. Personification in discourse: Linguistic forms, conceptual structures and communicative functions. *Language and Literature* 20(2):113–135. https://doi.org/10.1177/0963947010395522

Aletta G. Dorts, Gerben Mulder, and Gerard J. Steen. 2011. Recognition of personifications in fiction by

non-expert readers. *Metaphor and the Social World* 1(2):174–201. https://doi.org/10.1075/msw.1.2.04dor

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chirs Biemann. 2016. A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In *Proceedings of the LT4DH workshop at COLING 2016.* Osaka, Japan, pp. 76–84.

Gábor Simon. 2022. Identification and Analysis of Personification in Hungarian: The PerSECorp project. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022).* Marseille, France: European Language Resources Association, pp. 2730–2738.

Gábor Simon, Tímea Bajzát, Júlia Ballagó, Zsuzsanna Havasi, Emese K. Molnár, and Eszter Szlávic. 2023. When MIPVU goes to no man's land: a new language resource for hybrid, morpheme-based metaphor identification in Hungarian. *Lang Resources and Evaluation* (2023). https://doi.org/10.1007/s10579-023-09705-9

Balázs Indig, Bálint Sass, Eszter Simon, Iván Mittelholcz, Noémi Vadász, and Márton Makrai. 2019. One format to rule them all. The emtsv pipeline for Hungarian. In *Proceedings of the 13th Linguistic Annotation Workshop.* Florence, Italy: Association for Computational Linguistics, pp. 155–165.

Jan-Christoph Klie, Michael Burgert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics (COLING 2018).* Santa Fe, New Mexico, USA, pp 5–9.

Zoltán Kövecses. 2010. *Metaphor. A Practical Introduction.* Oxford University Press, Oxford, UK.

George Lakoff. 2006. The contemporary theory of metaphor. In Dirk Geeraerts (Ed.), *Cognitive Linguistics: Basic Readings.* Mouton de Gruyter, Berlin, GE and New York, NY, pp. 185–238. https://doi.org/10.1515/9783110199901

Ronald W. Langacker 2013. *Essentials of Cognitive Grammar.* Oxford University Press, New York, NY.

Deyin Long. 2018. Meaning construction of personification in discourse based on conceptual integration theory. *Studies in Literature and Language* 17(1):21–28. http://dx.doi.org/10.3968/10361

Graham Low. 1999. "This paper thinks…": Investigating the acceptability of the metaphor AN ESSAY IS A PERSON. In Lynne Cameron, and Graham Low (Eds.), *Researching and Applying Metaphor.* Cambridge University Press, Cambridge, UK. pp. 221–248.

Csaba Oravecz, Tamás Váradi, and Bálint Sass. 2014. The Hungarian Gigaword Corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014).* Reykjavik, Iceland: European Language Resources Association, pp. 1719–1723.

Ferenc Pusztai (Ed. in chief). 2003. *The Concise Dictionary of Hungarian*. Akadémiai Kiadó, Budapest, HU.

Pavel Rychlý. 2008. A lexicographer-friendly association score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing RASLAN.* Masaryk University, Brno, CZ. pp. 6–9.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, AU. URL https://www.R-project.org/.

Gerard J. Steen, Aletta G. Dorst, Berenike J. Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijnjte Pasma. 2010. *A Method for Linguistic Metaphor Identification. From MIP to MIPVU.* John Benjamins, Amsterdam, NL and Philadelphia, PA. https://doi.org/10.1075/celcr.14

Piotr Twardzisz. 2013. *The Language of Interstate Relations. In Search of Personifications.* Palgrave Macmillan, Houndmills, UK and New York, NY. https://doi.org/10.1057/9781137332707

A The PerSE Corpus: numbers and labels

Register	No. of texts	Size in tokens
Car reviews	6	10,468
Reports on interstate relations	5	7,938
Research papers	5	11,500

Table 1: The structure of the PerSE corpus

Ptags (components)	PRW	PRA	PR Wid	PR Aid	PRW imp
Prel (relations)	tr	lm	poss	r	
Pqual (qualities)	pconv	pnov	pdef	pmet	

Table 2: The layers of the manual annotation

B Personifying Usage of the Most Frequent Verbs in the Corpus

Verb	Car reviews	Interstate reports	Research papers
tud ('know/can')			
personifying	20	2	1
non-personifying	17	19	9
ad ('give')			
personifying	3	2	5
non-personifying	12	3	8
tesz ('put/do')			
personifying	5	4	8
non-personifying	9	10	4
vesz ('take')			
personifying	2	2	5
non-personifying	10	3	6

Table 3: Contingency table
of (non-)personifying use of verbs in the corpus

Report on the Multilingual Euphemism Detection Task

Patrick Lee and **Anna Feldman**
Montclair State University
New Jersey, USA
{leep,feldmana}@montclair.edu

Abstract

This paper presents the Multilingual Euphemism Detection Shared Task for the Fourth Workshop on Figurative Language Processing (FigLang 2024) held in conjunction with NAACL 2024. Participants were invited to attempt the euphemism detection task on four different languages (American English, global Spanish, Yorùbá, and Mandarin Chinese): given input text containing a potentially euphemistic term (PET), determine if its use is euphemistic or not. We present the expanded datasets used for the shared task, summarize each team's methods and findings, and analyze potential implications for future research.

1 Introduction

Euphemisms are a linguistic device used to soften or neutralize language that may otherwise be harsh or awkward to state directly (e.g., "between jobs" instead of "unemployed","late" instead of "dead", "collateral damage" instead of "war-related civilian deaths"). By acting as alternative words or phrases, euphemisms are used in everyday language to maintain politeness, mitigate discomfort, or conceal the truth. While they are culturally-dependent, the need to discuss sensitive topics in a non-offensive way is universal, suggesting similarities in the way euphemisms are used across languages and cultures.

Terms which may be used euphemistically sometimes require context to determine a euphemistic usage:

Asked to choose between jobs and the environment, a majority – at least in our warped, first-past-the-post system – will pick jobs. (non-euphemistic)

This summer, the budding talent agent was between jobs and free to babysit pretty much any time. (euphemistic)

In this shared task, participants were invited to develop approaches and models to disambiguate texts (in multiple languages) as either euphemistic or not. The previous iteration of this task resulted in numerous insights from participating teams, but featured only an English dataset (Lee et al., 2022a). By providing a multilingual iteration, we hoped to extend these findings to other languages and employ transfer learning to uncover possible cross-lingual patterns (Shode et al., 2023). This paper is structured as follows: Section 2 describes related work, Section 3 describes the additional data collected for the competition[1] and the task setting, Section 4 summarizes the participants' methods and results, and Section 5 analyzes common findings and the future directions they suggest.

2 Related Work

Magu and Luo (2018) and Felt and Riloff (2020) explored word embeddings and sentiment analysis, respectively, for detecting euphemisms. Zhu and Bhat (2021) and subsequent works such as (Lee et al., 2022a) and Lee et al. (2023) advanced this research using BERT and other transformers for euphemism detection and disambiguation. Keh (2022) focused on classifying previously unseen euphemistic phrases. Gavidia et al. (2022) built a corpus of potentially euphemistic terms (PETs) influencing further studies (Lee et al., 2022b,a, 2023). Most recently, (Lee et al., 2024) demonstrated the effectiveness of XLM-RoBERTa in multilingual euphemism disambiguation, showing superior performance of multilingual over monolingual models and enabling zero-shot learning across languages (refer to Table 1 for the average macro-F1 scores from multilingual and cross-lingual experiments).

[1]The final datasets, as well as the specific train-test split used for the competition, are available at https://github.com/pl464/euph-detection-datasets/tree/main/EACL_2024

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 110–114
June 21, 2024 ©2024 Association for Computational Linguistics

Table 1: Average Macro-F1s for multi- and cross-lingual experiments. ZH=Mandarin Chinese, EN=American English, ES=Global Spanish, and YO=Yorùbá

TrainTest	ZH	EN	ES	YO
Baseline	0.426	0.416	0.381	0.394
ZH	**0.879**	0.653	0.535	0.300
EN	0.607	**0.765**	0.567	0.381
ES	0.613	0.639	**0.752**	0.384
YO	0.417	0.407	0.383	**0.790**
ZH+EN	0.897	0.804	0.508	0.397
EN+ES	0.650	0.781	0.764	0.416
ES+YO	0.605	0.630	0.758	0.794
ZH+ES	0.884	0.670	0.764	0.377
EN+YO	0.616	0.772	0.602	0.802
ZH+YO	0.881	0.646	0.585	0.795
ZH+EN+ES	0.898	0.805	0.775	0.389
EN+ES+YO	0.647	0.783	0.772	0.791
ZH+EN+YO	0.899	0.801	0.555	0.794
ZH+ES+YO	0.885	0.664	0.778	0.778
All	0.895	0.792	0.776	0.793

3 Task Setting

3.1 Multilingual Datasets

The training data used in this competition were the labelled datasets in American English (EN), Spanish (ES), Yorùbá (YO), and Mandarin Chinese (ZH) constructed and described by Lee et al. (2023). Source texts were collected from a variety of sources that comprised primarily of online articles and webpages (though the Spanish and Yorùbá datasets included other sources, such as transcribed texts and social media posts). Each instance contained up to 3 sentences and contained a potentially euphemistic term (PET). These texts were also human-annotated with labels indicating either a euphemistic (1) or non-euphemistic (0) usage of the PET. Special tokens were placed before and after the PET in each instance, which we standardize for the shared task as "**[PET_BOUNDARY]**". Additionally, as euphemisms can be language-specific, data for each language were collected separately (i.e. are not translations of each other) and differed in PET and label distributions.

Since these datasets were already publicly available, we collected additional data in each of the four languages to comprise the test sets. The data were from the same source corpora as the training data and were annotated by 2-3 native speakers in each language. The final distribution of examples in the training and test set can be found in Table 2. Note that the goal was not only to provide unseen examples for the shared task, but also to contribute additional data for multilingual euphemism detection in general; therefore, test sets sometimes contained entirely new PETs, but to varying extents across languages as shown in Table 3. Prior work has shown that when new PETs are introduced at test time, models have a more difficult time correctly classifying them (Keh, 2022). As a result of this and other differences between the datasets, classification metrics among the participants should not be compared across languages, but "within" languages.

Lang	Train		Test	
	1s	0s	1s	0s
EN	1339	563	502	694
ES	1146	715	809	282
YO	1270	659	419	250
ZH	1469	516	744	482

Table 2: Number of examples per label in train and test. "1s" refers to euphemistic examples, and "0s" refers to non-euphemistic examples.

Lang	Number of PETs		
	Train	Test	Overlap
EN	121	67	44
ES	148	85	0
YO	133	28	4
ZH	110	48	7

Table 3: Number of PETs/overlap between train and test

3.2 Task Description

The shared task was hosted as a competition on Codabench[2]. During the development phase, participants were provided with datasets in all four languages. During the test phase, participants were provided a test set for each language and had the option of submitting predictions for one to four of them for scoring. However, all teams ultimately chose to submit predictions for all four. The metric for comparison was Macro-F1, and the submissions were ranked using the average Macro-F1 across all four languages, weighted equally.

#	User	EN	ES	YO	ZH	AVG	Title of Paper
1	amri228	0.83	0.60	0.72	0.78	0.73	Can GPT4 Detect Euphemisms across Multiple Languages? (Firsich and Rios, 2024)
2	vitiugin	0.74	0.67	0.63	0.71	0.69	Ensemble-based Multilingual Euphemism Detection: a Behavior-Guided Approach (Vitiugin and Paaki, 2024)
3	nhankins	0.65	0.61	0.65	0.68	0.65	Optimizing Multilingual Euphemism Detection using Low-Rank Adaptation Within and Across Languages (Hankins, 2024)
4	Baseline	0.30	0.43	0.39	0.38	0.37	–

Table 4: Results of submitted systems to the Multilingual Euphemism Detection Task

4 Participants and Results

In all, there were 3 teams that participated in the task and also submitted descriptions of their systems. A summary of their performances are in Table 4, along with a majority class baseline. In this section, we briefly describe each team's approach and results.

4.1 GPT-4 in Zero-Shot and Few-Shot Settings

Firsich and Rios (2024) submitted the highest-scoring approach (based on averaged F1 across all four languages), which explored zero-shot and few-shot prompts with GPT4 for the task. Their zero-shot setting consisted of instructions and then the task prompt, optionally accompanied by "Context", or a description of what euphemisms are. Their few-shot setting consisted of the above, plus k examples of euphemistic and non-euphemistic instances with labels. On the development set, they confirm that the highest setting of k=8 yields the highest scores by a significant margin over k=2, which is also significantly better than k=0. Moreover, providing few-shot examples that contained the same PET as in the task prompt was always better. This is an intuitive result, and it seems the model is able to better leverage more directly related examples to do a better job of disambiguating PET usages. Additionally, providing the "context" of what euphemisms are boosted performance significantly for the zero-shot setting (e.g. for Yorùbá, 0.400 $\rightarrow$ 0.610).

On the shared task's test set, they scored the highest in all categories except Spanish. Performances on all languages except English dropped significantly from the best setting in the development set (ES: 0.761 $\rightarrow$ 0.598, YO: 0.872 $\rightarrow$ 0.723, ZH: 0.858 $\rightarrow$ 0.776). This likely correlates with the degree of "PET overlap" (see Table 3) for which English is very high, Spanish and Yorùbá very low, and Chinese in-between.

4.2 Behavior-Guided, Ensemble-based Approach

Vitiugin and Paaki (2024) develop an approach using an ensemble of multilingual transformers (XLM-RoBERTa-large, or XLM-R), each fine-tuned on either the euphemism detection task or one of several "behavior-related" tasks (sarcasm and irony detection, sexism detection, racism detection, and sentiment classification) that are potentially related to general euphemism understanding. The authors cite multiple works in which training on such tasks, as well as ensembling, have been shown to improve performance on figurative language tasks. Unlike the previous system, they train and test on data from all four languages at once.

They found the best approach on the development set to be a Random Forest ensemble of 6 models: all 4 behavior-related fine-tuned models, and 2 trained on the euphemism detection task, one of which with PETs removed from the text, and the other as normal. This decision may have stemmed from the observation that PETs are unevenly distributed in the dataset, and the model should learn to classify based on context. While their reported performance on the development set was very high (F1 = 0.95), it was much lower on the test set (average F1 across the four languages = 0.69), though they yielded the highest score on Spanish in the competition (F1 = 0.67). This suggests some kind of significant overfitting, perhaps in regards to PETs, though no connection to "PET overlap" can be made, as their validation perfor-

[2] https://www.codabench.org/competitions/1959/

mance was not reported for each of the languages separately.

4.3 Optimizing and Low-Rank Adaptation Approach

Hankins (2024) experiment with multiple multilingual transformer models with a focus on efficient methods. On the development data, they find that fine-tuning multilingual DistilBERT (base, cased) with Low-Rank Adaptation (LoRA) yields comparable performances to using XLM-R (F1 ∼0.74-0.85), while being much lighter and faster to train. However, as with the other teams' approaches, the performance on the test set was much lower all around (∼0.61-0.68). This may suggest that, while more parameter-efficient approaches work well when tested on PETs seen during training, a larger number of parameters may be needed for capturing the nuances associated with unseen PETs.

5 Discussion and Future Work

Here, we discuss some common themes among the participants' approaches and suggest related directions for future work.

5.1 PETs Matter

There are many indications that the distribution of PETs in the data seems to matter to a large extent. Not only are the test score degradations correlated with PET overlaps in each language, but each language's relative score also seems correlated with the overall number of PETs in the dataset (e.g., Spanish had the most unique PETs total, 233, and generally performed the worst; English had the least, 144, and performed the best). Furthermore, the degree of difficulty may vary by PET, as well. In addition to the varying label distributions per PET (e.g. ten 1's and ten 0's for one PET, but thirty 1's and no 0's of another), the complexity of some PETs may also differ. Firsich and Rios (2024) noted the examples of the English PET "disabled" and Chinese PET "环卫工人", which intuitively seemed difficult to classify and in fact required relatively many examples of the same PET to improve performance.

All in all, it seems that this task is inherently tied to the varying types of PETs present. It is suggested that future work should pay special attention to this aspect, perhaps experimenting with different ranges, amounts, or linguistic qualities of PETs.

5.2 Analyzing Model Predictions

As mentioned in the previous section, Firsich and Rios (2024) observed that PETs may have different "classification difficulties" by looking past the classification metrics and at actual predictions. Hankins (2024) additionally report the distributions of predictions made by models trained on different languages. While they found, somewhat unsurprisingly, that test performance on language X is highest with a model trained on data from all four languages (i.e. is trained four times as much data), it makes significantly different predictions than a model only trained on language X, particularly for Chinese and English. This suggests that training on multiple languages results in significantly different learned representations of languages for this task. Overall, it is suggested to analyze prediction distributions and error analyses to further understand model behavior.

5.3 Linguistically Related Knowledge

Euphemism detection may involve many different forms of pragmatic knowledge - politeness, offensiveness, directness, conciseness, sentiment, sensitivity, etc. One way to leverage this intuition computationally is to explicitly teach models these tasks, as explored by Vitiugin and Paaki (2024), or include them as part of model inputs. The validation scores from Firsich and Rios (2024) show that including a definition of euphemisms in prompts benefits GPT4 in the zero-shot setting almost as much as providing randomized (i.e. not having the same PET) few-shot examples. Additionally, models trained on euphemism detection may also implicitly encode this knowledge, and perhaps differently across languages. These are all potential findings for future computational work to uncover.

6 Conclusion

We present the results of the Multilingual Euphemism Detection Shared Task. Participants' systems scored well above the baselines, but well below their reported validation metrics. Taken together, these results invite further work into using LLMs, ensembling/related tasks, and efficient models, which showed proficiency across languages, but leave much room for improvement. From a synthesis of the teams' findings, we also suggest that future work explore the impact of PETs, model behavior beyond performance metrics, and connections with related linguistic tasks.

Limitations

The primary limitations of the work include inconsistent performance across languages, particularly in non-English languages due to varying degrees of potentially euphemistic term overlap and limited model robustness in handling diverse linguistic data.

Ethics Statement

The authors foresee no ethical concerns with the work presented in this paper.

Acknowledgements

We would like to thank Alain Chirino Trujillo, Diana Cuevas Plancarte, Iyanuoluwa Shode, Julia Sammartino, Olumide Ebenezer Ojo, Thomas Hicks, Xinyi Liu, and Yuan Zhao for for all the help with collecting and annotating datasets and preparing them for the shared task.

This material is based upon work supported by the National Science Foundation under the Grant number 2226006.

References

Christian Felt and Ellen Riloff. 2020. Recognizing euphemisms and dysphemisms using sentiment analysis. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 136–145.

Todd Firsich and Anthony Rios. 2024. Can gpt4 detect euphemisms across multiple languages? In *In Proceedings of the 4th Workshop on Figurative Language Processing co-located with NAACL 2024, Mexico City, June, 2024. To appear.*

Martha Gavidia, Patrick Lee, Anna Feldman, and JIng Peng. 2022. CATs are fuzzy PETs: A corpus and analysis of potentially euphemistic terms. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2658–2671, Marseille, France. European Language Resources Association.

Nicholas Hankins. 2024. Optimizing multilingual euphemism detection using low-rank adaptation within and across languages. In *In Proceedings of the 4th Workshop on Figurative Language Processing co-located with NAACL 2024, Mexico City, June, 2024. To appear.*

Sedrick Scott Keh. 2022. Exploring euphemism detection in few-shot and zero-shot settings. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 167–172, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Patrick Lee, Alain Chirino Trujillo, Diana Cuevas Plancarte, Olumide Ojo, Xinyi Liu, Iyanuoluwa Shode, Yuan Zhao, Anna Feldman, and Jing Peng. 2024. MEDs for PETs: Multilingual euphemism disambiguation for potentially euphemistic terms. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 875–881, St. Julian's, Malta. Association for Computational Linguistics.

Patrick Lee, Anna Feldman, and Jing Peng. 2022a. A report on the euphemisms detection shared task. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 184–190, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Patrick Lee, Martha Gavidia, Anna Feldman, and Jing Peng. 2022b. Searching for PETs: Using distributional and sentiment-based methods to find potentially euphemistic terms. In *Proceedings of the Second Workshop on Understanding Implicit and Underspecified Language*, pages 22–32, Seattle, USA. Association for Computational Linguistics.

Patrick Lee, Iyanuoluwa Shode, Alain Trujillo, Yuan Zhao, Olumide Ojo, Diana Plancarte, Anna Feldman, and Jing Peng. 2023. FEED PETs: Further experimentation and expansion on the disambiguation of potentially euphemistic terms. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 437–448, Toronto, Canada. Association for Computational Linguistics.

Rijul Magu and Jiebo Luo. 2018. Determining code words in euphemistic hate speech using word embedding networks. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 93–100.

Iyanuoluwa Shode, David Ifeoluwa Adelani, Jing Peng, and Anna Feldman. 2023. Nollysenti: Leveraging transfer learning and machine translation for nigerian movie sentiment classification. pages 986–998.

Fedor Vitiugin and Henna Paaki. 2024. Ensemble-based multilingual euphemism detection: a behavior-guided approach. In *In Proceedings of the 4th Workshop on Figurative Language Processing co-located with NAACL 2024, Mexico City, June, 2024. To appear.*

Wanzheng Zhu and Suma Bhat. 2021. Euphemistic phrase detection by masked language model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 163–168, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Report on the FigLang 2024 Shared Task on Multimodal Figurative Language

Shreyas Kulkarni[1], Arkadiy Saakyan[1], Tuhin Chakrabarty[1], Smaranda Muresan[1]

[1]Department of Computer Science, Columbia University,
shreyas.kulkarni@columbia.edu, a.saakyan@columbia.edu, tuhin.chakr@cs.columbia.edu, smara@cs.columbia.edu

Abstract

We present the outcomes of the Multimodal Figurative Language Shared Task held at the 4th Workshop on Figurative Language Processing (FigLang 2024) co-located at NAACL 2024. The task utilized the V-FLUTE dataset (Saakyan et al., 2024) which is comprised of <image, text> pairs that use figurative language and includes detailed textual explanations for the entailment or contradiction relationship of each pair. The challenge for participants was to develop models capable of accurately identifying the visual entailment relationship in these multimodal instances and generating persuasive free-text explanations. The results showed that the participants' models significantly outperformed the initial baselines in both automated and human evaluations. We also provide an overview of the systems submitted and analyze the results of the evaluations. All participating systems outperformed the LLaVA-ZS baseline, provided by us in F1-score.

1 Introduction

Figurative language, which demands an understanding of the implied meanings behind expressions, has been extensively studied, as demonstrated in prior research (Chakrabarty et al., 2022; Saakyan et al., 2022). Similar complexities exist in visual domains, notably in visual metaphors (Chakrabarty et al., 2023; Akula et al., 2023), though most research on large multimodal models (LVMs) has primarily addressed the interpretation of literal meanings in images, as seen in benchmarks like e-ViL (Kayser et al., 2021), ScienceQA (Lu et al., 2022), and MMMU (Yue et al., 2024).

In this shared task, we aim to explore how LVMs handle figurative content in multimodal inputs. Our task, explainable figurative visual entailment, challenges a model to determine whether an image (the premise) supports or contradicts a given claim (the hypothesis) and to provide a reasoned explanation for its decision. Examples from our dataset are shown in Table 2.

The dataset leverages extensive prior research on both figurative language and images (Chakrabarty et al., 2023; Yosef et al., 2023; Hessel et al., 2023; Hwang and Shwartz, 2023; Desai et al., 2022). It is designed specifically for the visual entailment task and is enhanced with high-quality annotations that include explanations.

This paper reports the results of the shared task that is part of the 4th Workshop on Figurative Language Processing (FigLang 2024) at NAACL 2024. Details of the task, datasets, and evaluation methods are discussed in Section 2. Summaries of each participating system are provided in Section 4, and Section 4.4 offers a comparative analysis of these systems.

2 Datasets and Task Description

Subset	Fig. Lang. Type	Fig. Part
IRFL	Metaphor, Idiom, Simile	Caption
VisMet	Metaphor, Simile	Image
MemeCap	Humor	Image
MuSE	Sarcasm	Caption
NYCC	Humor	Both
V-FLUTE	Metaphor, Idiom, Simile, Sarcasm, Humor	Image, Caption, Both

Table 1: Overview of subsets for visual entailment and multimodal figurative language understanding.

The shared task utilizes an early version of the V-FLUTE dataset, introduced by Saakyan et al. (2024). The dataset is comprised of <image, text> pairs, each annotated with labels indicating either entailment or contradiction, along with explanations for each pair (see Table 2). Originating from five previous studies (Chakrabarty et al., 2023; Yosef et al., 2023; Hessel et al., 2023; Hwang and Shwartz, 2023; Desai et al., 2022), V-FLUTE includes figurative language elements such as metaphors, idioms,

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 115–119
June 21, 2024 ©2024 Association for Computational Linguistics

Subset	Image (Premise)	Claim (Hypothesis)	Label and Explanation
VisMet		The faculty meeting was peaceful.	**Label:** Contradiction *Explanation:* The image shows a faculty meeting transformed into a dramatic battlefield scene, with members dressed as knights discussing academic content on boards behind them as if they were battle tactics. This visual metaphor suggests the faculty meeting was like a war, and not peaceful.
IRFL		Their relationship is a house on fire.	**Label:** Entailment *Explanation:* [...] the photo suggests there is conflict or an intense emotional situation between the two individuals, which aligns with the symbolism of a house on fire representing a relationship filled with turmoil or heated arguments.
MuSE		Oh I just #love having to stare at this while I #work.	**Label:** Contradiction *Explanation:* the author wants to go to the disneyland and not just stare at it while working.
MemeCap		Even death won't exempt you from going to work.	**Label:** Entailment *Explanation:* The image displays RoboCop [...] This entails the claim that even death won't exempt you from going to work because it humorously illustrates a character who has been reanimated as a cyborg to continue working despite having died.
NYCC		Easy for you to say, you're cured!	**Label:** Entailment *Explanation:* A play on the word "cured". People go to therapy to have their mental problems remedied or cured. But "cured" can also refer to a meat preparation technique — here the therapist is cured bacon, and the patient is an egg (which is not cured). The egg is saying that the therapist doesn't understand his problems because he's "cured" in both senses.

Table 2: Sample dataset instances form V-FLUTE corresponding to the source datasets.

Subset	Train		Test	
	#	%	#	%
VisMet	731	16.5	126	18.3
IRFL	1322	29.9	198	28.7
MuSE	1000	22.6	150	21.8
MemeCap	853	19.3	128	18.6
NYCC	520	11.7	87	12.6
Total	4426	100.0	689	100.0

Table 3: Summary of subset distribution statistics involved in V-FLUTE.

similes, humor, and sarcasm. It consists of 5,115 multimodal pairs of high-quality images and texts, complete with labels and explanations. For statistics on the dataset, please see Table 3.

3 Evaluation Setup

To evaluate the participant models, we developed a test set by randomly selecting 689 instances, each comprising an <image,text> pair with corresponding explanations, from our dataset. We describe below the automatic metrics used to evaluate the models' capability in interpreting figurative language.

Automatic Metrics We used BERTScore (using `microsoft/deberta-xlarge-mnli`), termed here as the *explanation score*, which ranges from 0 to 100, to evaluate the quality of the explanations. Rather than just reporting label accuracy, we report the label F1 score at three explanation score thresholds: 0, 50, and 60. An F1@0 score corresponds to basic label F1, while an F1@50 score includes only those correct label predictions with an explanation score above 50.

4 Participants and Results

4.1 Training Phase

The competition began on January 25, 2024, with the release of training data and auxiliary scripts to all registered participants. Participants had the option to further divide the training data into a validation set for tuning hyperparameters or to use the data for cross-validation.

4.2 Evaluation Phase

The test instances were made available on February 15, 2024, for evaluation. The deadline for submissions was March 25, 2024. From the submissions, two system papers were accepted for presentation at the Workshop. Submissions were made through the Codalab site and evaluated against the test instances' gold labels. We utilized Codabench (Xu et al., 2022) for the competition due to its user-friendly interface, its ability to facilitate communication (such as mass emailing) with participants, and its real-time leader-board updates. Additionally, we established our own GPU-based evaluation system using custom Docker architecture. The leader-board showcased the F1@60 scores in descending order.

4.3 Participants

Overall, five teams participated in the competition, excluding the organizing team. The following section details the two systems that were accepted.

Baselines We report a fine-tuned baseline and a zero-shot baseline for the task. These baselines utilize a Zero-shot `LLaVA-v1.6-mistral-7B` model and a fine tuned `LLaVA-v1.6-mistral-7B` model on the V-FLUTE dataset.

MAPPER (map, 2024) is a modal-supplement framework, consisting of a describer and a thinker. The describer uses a frozen large vision model (`LLava-7B-v1.5`) to detail images capturing essential semantic information. The thinker, enhanced with LoRA (Hu et al., 2021) on a fine-tuned large multi-modal model (`LLava-7B-v1.5`), leverages these descriptions along with claims and images to form predictions and explanations. MAPPER's vision component uses CLIP (Radford et al., 2021) for image understanding.

FigCLIP (fig, 2024) merges CLIP and GPT-2 to identify and elucidate multimodal figurative semantics. It features separate text and image encoders initialized by CLIP (`CLIP-ViT-L/14`), connected via a bidirectional fusion module with cross-attention mechanisms. A GPT-2 model generates explanations, and a special projector aligns multimodal embeddings with explanation representations, enhancing the model's efficiency in handling figurative image-text alignment. The projector involved makes FigCLIP lightweight.

4.4 Analysis

The best performing method according to (Table 4) is MAPPER. The system outperforms others on both F1@0 and F1@60 metrics. We note that the system improvement is quite high compared to the zero-shot system. Interestingly, the FigCLIP system performs very well and only slightly lower

#	Participant	F1@0	F1@50	F1@60
1	MAPPER	0.90	0.89	0.75
2	LLaVA-FT	0.73	0.72	0.59
3	FigCLIP	0.70	0.67	0.50
4	GPT-4V	0.70	0.64	0.49
5	mrshu	0.63	0.62	0.43
6	yangst	0.51	0.48	0.31
7	LLaVA-ZS	0.45	0.38	0.21

Table 4: Automatic evaluation results by team with rank. FT refers to fine-tuned model and ZS represents the Zero-Shot model. The GPT-4V model submitted is not our baseline but a participants submission.

than the fine-tuned LLaVA model that utilizes a much stronger language model backbone.

5 Conclusion

This paper presents the outcomes of the shared task on multimodal figurative language, conducted at the 4th Workshop on Figurative Language Processing at NAACL 2024 (FigLang 2024). The goal of this shared task was to accurately classify figurative <image, text> instances and provide a persuasive explanation for the classification. We included a brief overview of each system that participants submitted to the shared task. All systems submitted by participants surpassed the LLaVA-ZS baseline in terms of F1-score. In conclusion, we anticipate that this shared task will encourage continued research into the understanding of figurative language.

References

2024. Figclip: A generative multimodal model with bidirectional cross-attention for understanding figurative language via visual entailment.

2024. A textual modal supplement framework for understanding multi-modal figurative language.

Arjun R Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T Freeman, et al. 2023. Metaclue: Towards comprehensive visual metaphors research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23201–23211.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7370–7388, Toronto, Canada. Association for Computational Linguistics.

Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10563–10571.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A dataset for captioning and interpreting memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445, Singapore. Association for Computational Linguistics.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1244–1254.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Arkadiy Saakyan, Tuhin Chakrabarty, Debanjan Ghosh, and Smaranda Muresan. 2022. A report on the FigLang 2022 shared task on understanding figurative language. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 178–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2024. V-flute: Visual figurative language understanding with textual explanations.

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543.

Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. IRFL: Image recognition of figurative language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1044–1058, Singapore. Association for Computational Linguistics.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Author Index