"Judge me by my size (noun), do you?" YodaLib: A Demographic-Aware Humor Generation Framework

Aparna Garimella[♠†], **Carmen Banea**[♠], **Nabil Hossain**[‡] and **Rada Mihalcea**[♠]

[♠]Computer Science and Engineering, University of Michigan
[†]Adobe Research, India
[‡]Computer Science and Engineering, University of Rochester
garimell@adobe.com, nhossain@cs.rochester.edu, {carmennb,mihalcea}@umich.edu

Abstract

The subjective nature of humor makes computerized humor generation a challenging task. We propose an automatic humor generation framework for filling the blanks in Mad Libs® stories, while accounting for the demographic backgrounds of the desired audience. We collect a dataset consisting of such stories, which are filled in and judged by carefully selected workers on Amazon Mechanical Turk. We build upon the BERT platform to predict location-biased word fillings in incomplete sentences, and we fine-tune BERT to classify location-specific humor in a sentence. We leverage these components to produce YODALIB, a fully-automated Mad Libs style humor generation framework, which selects and ranks appropriate candidate words and sentences in order to generate a coherent and funny story tailored to certain demographics. Our experimental results indicate that YODALIB outperforms a previous semi-automated approach proposed for this task, while also surpassing human annotators in both qualitative and quantitative analyses.

1 Introduction

Computer-generated humor is an essential aspect in developing personable human-computer interactions. However, humor is subjective and can be interpreted in different ways by different people. Humor requires creativity, world knowledge, and cognitive mechanisms, which are extremely difficult to model theoretically. Generating humor is considered by some researchers as an AI-complete problem (Stock and Strapparava, 2002), hence humor generation was largely studied in specific settings.

Language preferences vary with user demographics (Tresselt and Mayzner, 1964; Eckert and McConnell-Ginet, 2013; Garimella et al., 2016; Lin et al., 2018; Loveys et al., 2018), and this has led to approaches leveraging the demographic information of users to obtain better language representations and classification performances for various NLP tasks (Volkova et al., 2013; Bamman et al., 2014; Hovy, 2015; Garimella et al., 2017). Humor is a universal phenomenon that is used across all countries, genders, and age groups (Apte, 1985). Likewise, there are variations in how humor is enacted and understood due to demographic differences (Kramarae, 1981; Duncan et al., 1990; Goodman, 1992; Alden et al., 1993; Hay, 2000; Robinson and Smith-Lovin, 2001). However, to the best of our knowledge, the effect of user demographics in computational humor generation has not been studied.

In this paper, we introduce YODALIB, an automatic humor generation framework for Mad Libs®—a story-based fill-in the blank game—which also accounts for the demographic information of the audience and story coherence. We use *location* as the demographic dimension. YODALIB has three stages: (1) a candidate selection stage in which candidate words are selected to fill the sentences in each story, (2) a candidate ranking stage to assess if a filled-in (*transformed*) sentence is a funny version (*transformation*) of the original sentence, and (3) a story completion stage to join individually funny sentences to form complete Mad Lib stories that are humorous. Fig. 1 shows an example filled-in Mad Lib.

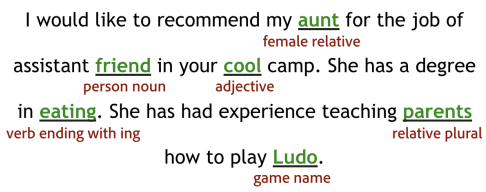

Figure 1: Example of a partial Mad Lib story.

This paper makes four main contributions: **(1)** We collect a novel dataset for location-specific humor generation in Mad Lib stories, which we carefully annotate using Amazon Mechanical Turk (AMT).[1] **(2)** We propose YODALIB, a location-specific humor generation framework that builds on top of BERT-based (Devlin et al., 2019) components while also accounting for location and story coherence. YODALIB typically generates funnier Mad Lib stories than those created by humans and a previously published semi-automatic framework. **(3)** We present qualitative and quantitative analyses to explain what makes the generated stories humorous, and how they differ from the other completions. **(4)** Finally, we outline the similarities and differences in humor preferences between two countries: India (IN) and United States (US), in terms of certain linguistic attributes. To the best of our knowledge, ours is the first computational study to automatically generate humor in a Mad Lib setting while also incorporating the demographic information of the audience, and analyzing its effect in terms of various linguistic dimensions.

2 Related Work

There is a long history of research in general theories of humor (Attardo and Raskin, 1991; Wilkins and Eisenbraun, 2009; Attardo, 2010; Morreall, 2012; O'Shannon, 2012; Weems, 2014). In computational linguistics, a large body of humor research involves humor recognition, and it typically focuses on specific types of jokes (Mihalcea and Strapparava, 2006; Kiddon and Brun, 2011; Bertero and Fung, 2016; Raz, 2012; Zhang and Liu, 2014; Bertero and Fung, 2016; Hossain et al., 2019). Research work on humor generation has been largely limited to specific joke types and short texts, such as riddles (Binsted et al., 1997), acronyms (Stock and Strapparava, 2002), or one-liners (Petrović and Matthews, 2013). In general, it is very difficult to apply humor theories directly to generate humor, as they require a high degree of commonsense understanding of the world.

Owing to the subjective nature of humor, there have been recent efforts in collecting datasets for humor; Blinov et al., (2019) collected a dataset of jokes and funny dialogues in Russian from various online resources, and complemented them carefully with unfunny texts with similar lexical properties. They developed a fine-tuned language model for text classification with a significant gain over baseline methods. Hasan et al., (2019) introduced the first multimodal language (including text, visual and acoustic modalities) dataset of humor detection, and proposed a framework for understanding and modeling humor in a multimodal setting.

In socio-linguistics, the relationship between humor and gender is widely studied. Hay (2000) found that New Zealand women more often shared funny personal stories to create solidarity, while men used other strategies to achieve the same goal. More recently, location has become central in sociolinguistics (Johnstone, 2010). Alden et al., (1993) indicated that humor styles vary with countries, and humorous communications from Korea, Germany, Thailand and US, had variable content for funny advertising, while sharing certain universal cognitive structures.

The effect of demographic background on language use has gained significant attention in computational linguistics, with several efforts focused on understanding the similarities and differences in the language preferences, opinions and behaviors of people (Garimella et al., 2016; Garimella and Mihalcea, 2016; Wilson et al., 2016; Lin et al., 2018; Loveys et al., 2018; Welch et al., 2020). Conversely, there has been work to leverage these demographic differences in language preferences between various groups, to develop better models for NLP tasks, such as sentiment analysis (Volkova et al., 2013), word representations (Bamman et al., 2014), sentiment, topic and author attribute classification (Hovy, 2015), and word associations (Garimella et al., 2017). However, to our knowledge, none of this recent work accounts for demographic information in humor recognition or generation tasks.

We find inspiration in recent work by Hossain et al., (2017), who collected a humor generation dataset with Mad Lib stories, and proposed a semi-automated approach to aid humans in writing funny stories. We go one step further and propose a fully-automated BERT-based demographic-aware humor generation framework. We further study the influence of location on humor preferences via AMT and seek to emulate such preferences in our automatically generated stories. Our work is similar to (Mostafazadeh et al., 2017), as our goal is to engage readers from different demographic groups by generating funnier versions of stories to read.

[1] We release the location-specific Mad Lib humor dataset `http://lit.eecs.umich.edu/downloads.html`.

3 Data Collection

Mad Libs® is a fill-in-the-blank game (Price and Stern, 1974) to create funny stories. A Mad Lib is a textual story template consisting of a title and a short story, with some of the words masked. Players are prompted to provide replacement words for the masked entries based on the provided hints (e.g., part-of-speech (POS) tags, *bodypart*, *food*) without having read the story. The replacement words are then filled in the story; the resulting Mad Lib is usually funny, with the humor aspect coming from the nonsensical filled-in words in an otherwise coherent and sensible story.

We use stories curated by Hossain et al., (2017), namely **Fun Libs**, as (1) Mad Libs are copyrighted and hence it is difficult to release datasets, and (2) experimentation with Fun Libs allows comparison of our approach with that proposed by Hossain et al., (2017).[2] We discard 4 of the 50 Fun Libs, as their themes cater to a US audience,[3] and replace them with 4 new stories we created following the heuristics devised by Hossain et al., (2017).

The data annotation is undertaken by two parties: the **players** who fill-in the blanks to create funny stories, and the **judges** who assess the filled-in stories in terms of their funniness. We assume the three-stage annotation framework devised by Hossain et al., (2017): *judge selection*, *player selection* and *story annotation*, with a few revisions to account for location-specific annotations.

Judge selection. This is done via a linguistic and a demographic survey. Turker judges from each country are given seven pre-filled stories, three of which are taken directly from Wikipedia, with some words underlined as if they were filled-in words, while the remaining four stories were filled-in by English speakers from the corresponding country, who were instructed to create funny stories. We instruct the judges to select for each story a grade from {0: not funny, 1: slightly funny, 2: moderately funny, 3: funny}. To filter spam responses, verification questions are presented that can be answered only after reading the stories. The task ends with a demographic survey prompting the judges to provide their age group, nationality, gender, education, occupation, and income level.

Selected turkers are those who (i) assign 0 to the Wikipedia stories, a grade from $\{1, 2, 3\}$ to at least three of the remaining four stories, (ii) answer the story-based questions correctly, and (iii) spend at least 4 minutes to complete the task. 50 US and 43 IN judges are selected from 60 and 100 candidates respectively, ensuring that they are unbiased in judging the funniness of a filled-in story.

Player selection. Players are expected to be good at writing funny stories. For this, we obtain four stories from Wikipedia, mask some words, and provide hint types next to them. To avoid excess workload to the turkers (leading to possible filling bias), two tasks are created for each country on AMT, each with two stories and a demographic survey, with instructions to fill the stories to make them funny, and answer the demographic survey. Further instructions are provided: the filled-in words must (i) occur in English dictionary, (ii) have exactly one word written in Latin alphabet, (iii) agree with the provided hint types, and (iv) not be slang words or sexual references, as these lead to shallow humor.

In each country, the filled-in stories are graded on the 0-3 humor scale by 5 qualified judges from that country, to reduce the effect of variations in humor preferences, and be representative of an audience rather than an individual on AMT. Players are selected if their **mean funniness grade** (MFG), mean of the 5 judgements, ≥ 1 for at least one story. 30 IN and 26 US players are selected from 80 and 60 turkers respectively.

Annotating Fun Libs. For each story, we ensure that a qualified turker can participate as either a player or a judge but not both. The stories are filled by 3 players from each country following the player selection instructions, and are judged by 5 judges from the same country. The judges rate the overall funniness, coherence, deviation from the story title (on the 0-3 scale), whether incongruity[4] was applied by the player ({Yes, No}), the humor contribution of each filled-in word ({'funny', 'not funny'}), and a verification question. In addition to same country judgements, we also obtain judgements from the opposite country

[2]We limit ourselves to the 50 stories created by Hossain et al., (2017); we believe that ours is the first endeavor in uncovering the demographic-specific idiosyncrasies in humor preferences, and provides motivation for future efforts to collect even larger demographic-specific datasets for humor studies.

[3]Their titles include *Kim Kardashian*, *Baseball*, *Boston Tea Party*, *The Statue of Liberty*.

[4]Incongruity is present in a joke if there is a surprise that defies the expectation of the reader (Weems, 2014), therefore causing the content to be funny.

for the filled-in test stories, to allow for cross-country analyses. We compare coherence to incongruity and deviation, to understand their effect on humor. The Krippendorf's alpha (Krippendorff, 1970) values for IN and US are 0.214 and 0.173 respectively, which indicate positive agreements among AMT judges, and are comparable to those obtained by Hossain et al. (2019), who crowd-sourced a dataset of humorous edited news headlines on the same funniness scale. The total AMT cost is about $1,200.

4 Automatic Humor Generation Framework

We first describe the location-specific training components, namely the location-biased language model and the location-specific humor classifier. Then, we focus on the story generation pipeline that follows three stages: (1) A **candidate selection** stage where possible word replacements for the blanks are generated using a location-biased language model; (2) A **candidate ranking** stage where the selected candidates are ranked by their humor contribution to the sentences they occur in, using a location-specific humor classifier; (3) A **story completion** stage where funny stories are created by selecting the top ranked funny transformations for each sentence, and concatenating them to obtain complete stories.

4.1 Training components

4.1.1 Location-Biased Language Model

In order to be able to generate a Mad Lib-like story, the first step is to automatically fill in a blank with a word that fits the context. Such words can be predicted by a language model, such as the BERT masked language model (MLM), a state-of-the-art deep learning framework (Devlin et al., 2019) based on a multi-layer bidirectional Transformer (Vaswani et al., 2017), trained on the English Wikipedia and Book Corpus (Zhu et al., 2015) datasets, for masked word and next sentence prediction tasks. Since its predictive ability is generic and it does not take the desired demographics into consideration (such as location, in our case), we train it further on location-rich data, thus enabling the model to make word predictions in context biased toward a particular country. To achieve this, we use a large dataset of blog posts (Garimella et al., 2017) authored by users from IN and US (35K blogs, 17M tokens for IN, and 33K, 12M tokens for US). We use the $\text{BERT}_{\text{base}}$ model with default parameters. This allows the language model to incorporate location-based word preferences in its prediction and to provide different replacements for a masked word occurring in a sentence written by an Indian English speaker versus an American English speaker.

4.1.2 Location-Specific Humor Classification

Furthermore, automatically generated word replacements in context need to be assessed for humor given an audience with a particular demographic. To enable us to gauge the funniness level of a word replacement, we train location-specific humor classifiers based on the BERT framework by leveraging the AMT annotations and the country of the turkers. Of the 50 stories, 40 are used for training,[5] and the remaining for evaluation. The stories' sentence splits are retained from (Hossain et al., 2017). The training and validation datasets consist of sentence pairs and their associated humor labels which are derived as follows. First, labels are assigned to each filling using a **majority vote** over the funny judgements in the gold standard. A sentence is considered **funny** if at least 50% of the filled-in words are funny. Sentences that do not contain blanks are not used for training.

Table 1 shows the sizes of train, validation and test sets. As we instructed the players to create funny stories, the datasets are, as expected, skewed toward the funny class. We opt to augment our data with additional non-funny sentence-pairs from Wikipedia. Since candidate sentences are location-agnostic, and therefore have the same form for both India and US, we introduce location-biased completions. We use our location-specific MLM to replace masked words in Wikipedia sentences[6] with the highest probability word given the location model, resulting in non-funny sentence pairs that are different by location.

TYPE	FUNNY		NOT FUNNY	
	IN	US	IN	US
TRAIN	566	574	130	122
VALIDATION	173	193	49	29
TEST	137	210	94	21

Table 1: Statistics on the humor classification datasets.

[5]The 4 newly created stories replace 4 training set stories.

[6]We focus on one of the four POS tags, as other hint types are not trivial to identify.

We use the above augmented dataset to fine-tune FUNNYBERT, a BERT-framework based sentence humor classifier that, accounting for the desired country of the audience, is able to identify whether content is humorous or not. Specifically, we fine-tune BERT for sentence-pair classification by adding a classification layer, with input pair <masked sentence, filled-in sentence>, and output is the prediction $c \in \{\text{funny, not funny}\}$ if the input corresponds to a funny transformation. The final hidden vector h corresponding to [CLS] token represents the sequence. The weights W are learnt during fine-tuning; $p(c|h) = \text{softmax}(Wh)$. We use the following parameters: batch size 32, gelu activation, sequence length 512, vocabulary 30,522, Adam optimizer, learning rate 1e-5, (selected over 5e-5, 5e-6, 1e-6), and 10 epochs (over 1-100 epochs). The limited sizes of annotated datasets make BERT a suitable framework to build on. FunnyBERT ranks the candidate filled-in sentences based on their humor (softmax probability).

Table 2 shows the location-specific validation accuracies of FunnyBERT. The majority vote baseline accuracy is 50%, as the datasets are class-balanced. Metric-wise statistically significant values are marked with *. The funny-class precision and accuracy are lower, and recall is higher for IN than those for US. This may be due to slightly lower quality funny sentence completions from IN players, resulting in FunnyBERT$_{\text{IN}}$ predicting false positives in most cases. We suspect that the familiarity of US turkers with Mad Libs was a factor that led to better quality stories for US. Augmenting Wikipedia sentences without location-specific replacements results in a significantly lower accuracy (80.06%) for IN, suggesting that the improved US scores are caused by the biased nature of the pre-training datasets used by BERT toward US English; IN-specific replacements, however minor, lead to improved performances in this locale.

METRIC	IN	US
PRECISION (FUNNY)	81.48	91.16*
RECALL (FUNNY)	89.02*	85.49
F1 SCORE (FUNNY)	85.08	88.24*
ACCURACY	**84.39**	**88.60***
ACCURACY (Wikipedia sentences without modification)	80.06	**89.12***

Table 2: Classification validation accuracies of FUNNYBERT$_{\text{X}}$, X $\in$ {IN, US} ($p < 0.05$).

4.2 YODALIB for Story Generation

Here we introduce YodaLib, our pipeline for humorous story generation. For a given MadLib-like story, we start out with word candidate selection and ranking, and finalize with story completion.

4.2.1 Candidate Selection

In order to generate candidate words for each story blank, the latter are replaced with the BERT [MASK] token and then input to the location-specific MLM. For each mask, probability scores are obtained for all the words in the vocabulary, and we retain the highest ranking $k = 10,000$ words. These are further filtered to obtain a cleaner candidate list, adhering to hint types and other restrictions imposed in the player selection phase; we further ensure that the number and tense match the hint type for nouns and verbs, respectively.

In sentences with multiple masks, we perform left-to-right selection, at each step pruning our decision space to the top n candidates to avoid assessing an exponential number of combinations.

We use FUNNYBERT to rank the filled-in sentences (Section 4.2.2), and the top $n = 100$ candidates are chosen to fill each mask. We impose left-to-right candidate selection, as this is how stories are read by turkers. It enhances the overall coherence of the resulting funny sentence, as the previously selected candidates are considered in selecting next ones. For a given context, we expect the candidates with high MLM scores to be good fits and hence less humorous (*cats drink milk*), while those with low scores are more likely to be incongruous, and may generate humor in the given context (*cats prepare milk*).

4.2.2 Candidate Ranking

Examining the positions or scores from MLM is not sufficient to predict if candidates are funny substitutes. The second stage involves ranking the candidates for each mask based on their humor contributions to the containing sentences. We leverage the FunnyBERT component by feeding the completed sentences to the model and using the softmax humor probability as a ranking value. The top $n = 100$ sentences are used for candidate selection of the next mask, until all the masks are filled.

METHOD	IN JUDGES		US JUDGES	
	TOP3	TOP10	TOP3	TOP10
FT_{IN}	$1.17^{\ddagger}$	-	$\underline{1.39}^{\ddagger\P}$	-
FT_{US}	$\underline{1.57}^{*\ddagger}$	-	$1.41^{\ddagger}$	-
MLM	0.70	0.91	0.68	0.84
$YODA_{IN}$	$\underline{\mathbf{1.94}}^{*\dagger\ddagger\P}$	$\underline{1.60}^{*\ddagger\P}$	$1.56^{*\ddagger}$	$1.32^{\ddagger}$
$YODA_{US}$	$\underline{\mathbf{2.03}}^{*\dagger\ddagger\P}$	$\underline{\mathbf{1.70}}^{*\dagger\ddagger\S\P}$	$\mathbf{1.77}^{*\dagger\ddagger\S}$	$1.48^{*\ddagger\S}$

Table 3: Average MFG for generated stories. Column-wise significantly higher scores than (i) FT_{IN} are marked with *, (ii) FT_{US} with †, (iii) MLM with ‡, and (iv) $YODA_{IN}$ with §. Row-wise higher values for each country are marked with ¶ ($p < 0.05$). The highest values in columns/ rows are bold/ underlined.

4.2.3 Story Completion

The final stage of the framework involves forming complete stories from the top funny transformations for each of the component masked sentences. Similar to candidate selection, we consider left-to-right story completion. For a sentence to be appended, it must be classified both as funny and as a potential next sentence given the previously selected context. This allows the resulting stories to be both funny and coherent. For (1), we consider the top-ranked funny transformations for each sentence. For (2), we rank the funny transformations based on their semantic similarity to the sentences previously selected in the story. The top $N = 100$ funny *and similar* transformations are selected for each subsequent sentence, resulting in different variations of the completed story. Only the top N as yet completed stories are advanced after processing each sentence (from N^2 sequences), based on the above two scores.

We estimate the similarity between any two transformations as the cosine similarity between their sentence embeddings. We use the average of the word embeddings from location-specific BERT to estimate the sentence embedding. Each word has 12 vectors (from the 12 BERT layers), each of length 768. Different layers of BERT encode different kinds of information, so the appropriate aggregation strategy depends on the task. We consider two variations to obtain the final word vectors: (i) sum of the embeddings from the last 4 hidden layers, and (i) embeddings from only the second to last hidden layer (Devlin et al., 2019). We settle on the second strategy, as it results in better quality stories for the validation set.[7] We then sort the stories in decreasing order of their **story funniness score** (mean of funniness scores of the constituting sentences) , as well as according to their **average word coherence** (mean of the pair-wise similarities of filled-in word embeddings).

5 Evaluation

We evaluate three methods to write funny stories:

1. **FREETEXT (FT)**: Following directions, AMT players complete the funny stories.
2. **MLM**: Word-replacement for each blank is selected based on the original BERT MLM model (without location-specific training).
3. **YODALIB**: The stories are generated using the proposed framework (see Section 4).

For each of these, we obtain 5 judgements on AMT. For each test Fun Lib, grading is done by judges from both locations on the 3 filled-in stories from FT, and the top 10 stories generated from MLM and YODALIB. We grade more stories generated by YODALIB as: (1) unlike the manually crafted FT stories , these are generated automatically, and hence require minimal effort (once the YodaLib framework is in place), and (2) both story funniness and average word coherence scores for the top 10 stories differ only in the fourth decimal place. Hence we treat them as the best diverse humorous variations generated for the given Fun Libs (BEST10). The stories from FT and YODALIB approaches are created in location-specific settings, while MLM does not account for the desired location slant. Stories from MLM are ranked based on their filled-in word probabilities of fitting the corresponding contexts. Hence, we expect the top-ranked MLM stories to be more congruous and less humorous than the low-ranked stories.

[7] The quality of stories is determined by human subjects, as this task is subjective.

Table 3 shows the averages of the mean funniness grades MFG (0-3 scale) of the generated stories, in two settings (for MLM and YODA). (1) TOP3: For MLM, we consider top three stories for each Fun Lib. For YODALIB, we consider three stories from BEST10 which have the highest MFG assigned by judges (as all of them have very similar scores assigned by our framework). (2) TOP10: We consider the top 10 (from filled-in word scores) and BEST10 stories for each Fun Lib for MLM and YODALIB respectively.

Humor generated by both FT and YODALIB is preferred to that by MLM, by both IN and US judges. This is expected, as MLM fills the stories with more plausible words, and hence does not introduce any surprise aspect that is essential for humor. The average MFG for MLM increases in the TOP10 setting, confirming that stories become more humorous as more incongruous words are used to fill them. In the FT stories, both IN and US judges prefer US-written humor (1.57, 1.41) to IN-written humor (1.17, 1.39). On the other hand, IN-written humor is liked more by US judges (1.39), and US-written humor by IN judges (1.57). Hence, an average US turker writes better stories than an average IN turker for Mad Libs, and judges in general enjoy humor written by turkers from a different country more than of their countrymen. This is in contrast to the general expectation that people prefer humor originated from their own group due to a better understanding of the various location-specific subtleties, suggesting that seeing things in a new light possibly contributes to the surprise and creativity factors essential for humor generation.

In the YODALIB approach, US-written humor is preferred to IN-written humor by both IN (2.03, 1.70) and US (1.77, 1.48) judges, in both settings. This may be due to the better performance of FunnyBERT$_{\text{US}}$ in ranking funny transformations to create stories. As seen in the FT approach, US-written stories by turkers are of better quality humor than IN-written ones on average, and this may have led to better quality training datsets for FunnyBERT$_{\text{US}}$. Our YODALIB approach outperforms FT for both IN and US judges (more so in the TOP3 setting for US judges, and both settings for IN judges), indicating that our approach generates better quality humor than AMT players on an average. Our approach also outperforms Libitum (Hossain et al., 2017) in the TOP3 setting. It has an average MFG of 1.51, which is much lower than the 1.77 given by US judges to US-generated humor.[8]

Both the IN- and US-generated stories in the YODALIB approach are liked more by IN judges than by US judges (underlined in Table 3). This suggests that an average IN turker has a more lenient outlook towards humor, whether it is of IN or US origin, and this may have also led to lower quality IN-written stories, whereas an average US turker has a more stricter perspective towards it, possibly due to the familiarity of the game in US, resulting in the more enjoyable humor from US players.

6 Discussion

6.1 Quantitative Analysis

Table 4 shows the correlations of coherence, incongruity and deviation with the corresponding MFG for the test stories generated in each approach. The TOP10 stories are used for MLM and YODALIB approaches. In FT, coherence plays an important role in creating humor for same-location audience on AMT (IN: 0.40, US: 0.77), indicating that humans have a striking ability to apply coherence to create humor, and it is particularly appreciated by audience from the same country, possibly due to a consistent use of location-specific information. This can be seen in the US judgements on IN-written humor, where funniness has no correlation with coherence, and high correlation with incongruity, indicating that US judges find those IN-written stories funny which contain seemingly unexpected words.

METHOD	INDIAN JUDGES			US JUDGES		
	COH	INC	DEV	COH	INC	DEV
FT$_{\text{IN}}$	**0.40**	0.25	**0.59**	0.01	**0.68**	0.19
FT$_{\text{US}}$	0.23	0.24	**0.62**	**0.77**	**0.83**	**-0.59**
MLM	**-0.57**	**0.74**	**0.66**	**-0.64**	**0.36**	**0.72**
YODA$_{\text{IN}}$	**0.22**	**0.41**	**0.52**	0.15	0.12	**0.22**
YODA$_{\text{US}}$	0.07	**0.27**	**0.41**	0.12	**0.27**	0.05

Table 4: Correlations of coherence (Coh), incongruity (Inc) and deviation (Dev) with MFG, with significant values in **bold** ($p < 0.05$).

In the MLM stories, coherence is negatively correlated with funniness, indicating the difficulty in generating humor via coherent words by a language model. In the YODALIB approach, coherence plays very little role; most of the funniness is achieved via incongruity and topic deviations (the only exception

[8]Training and evaluation settings for Libitum are closer to US humor judged by US judges setting (TOP3).

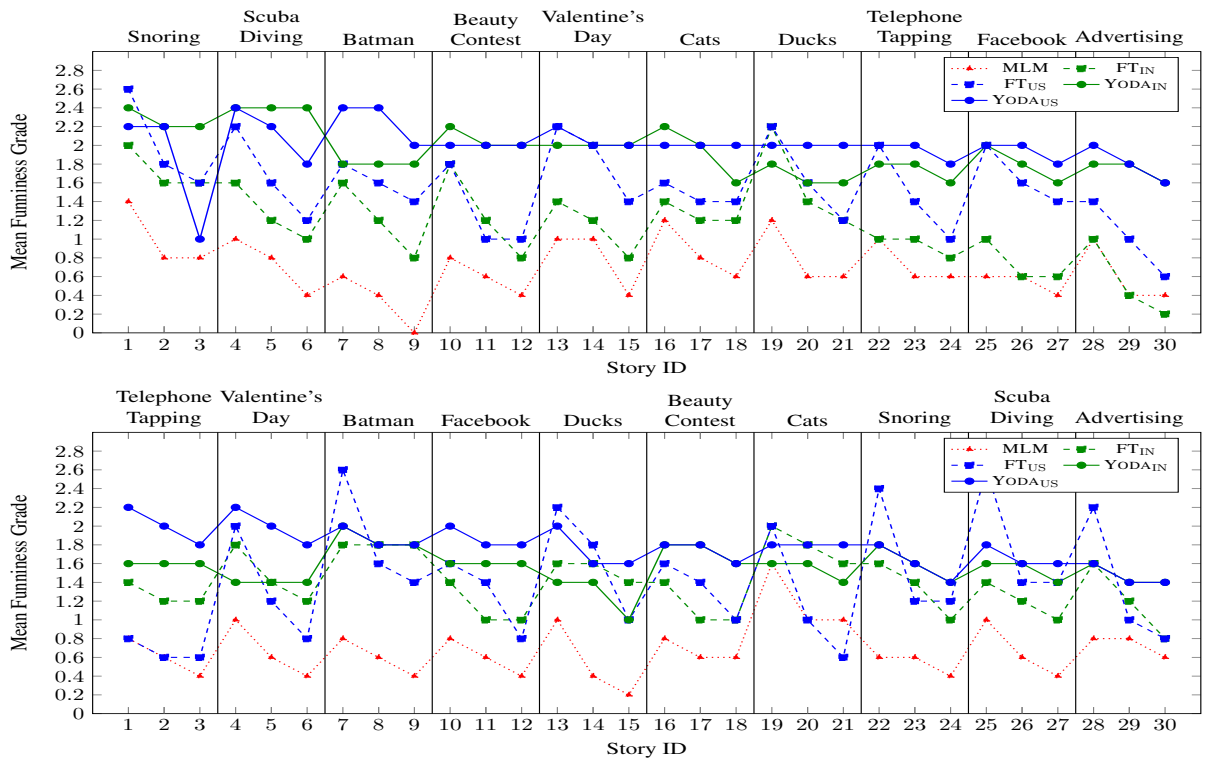

Figure 2: MFG for the 30 test stories (top: IN, bottom: US) using the three approaches graded by US judges.

being US judgements on US-written stories for deviation), though IN judges find IN-generated humor to be somewhat coherent (0.22). Certain skilled turkers are able to write meaningful and coherent stories, something that the YODALIB approach finds very difficult to achieve.

6.2 Qualitative Analysis

Figure 2 shows MFG for stories via the three approaches by IN and US judges respectively. We sort the story titles in decreasing order of MFG. FT consistently outperforms MLM. YODALIB outperforms FT for 28 and 23 stories for IN and US judges respectively. The average gains are 0.79 and 0.46 for IN judges, and 0.17 and 0.37 for US judges, for IN- and US-generated stories respectively. Hence, an average IN judge, though also prefers coherence, finds stories with low coherence and high incongruity and deviation also funny; as these are further incorporated in YODALIB stories, the gains in IN judgements are higher. Contrarily, US judges prefer coherence along with incongruity: the highest MFG by IN judges are for YODALIB stories (e.g., *Snoring*, *Scuba Diving*, *Batman*); those by US judges are for FT stories (e.g., *Batman*, *Snoring*, *Scuba Diving*). It is interesting that the original Mad Libs game introduced humor primarily via incongruous words to fill the stories, while coherence plays an important role in this study, possibly due to the richness of the task in considering the context to fill-in the blanks.

Table 5 shows example story snippets where YodaLib stories receive higher MFG than FT stories. Humor in the YODALIB stories is largely via incongruity (e.g., *advertisement to grill people*, *Valentine's Day is a prank*). Yet, the YODALIB approach is also able to generate small snippets of coherent and funny phrases: Advertising is how a company grills people to buy their products, it can bring sickness and be grim; Valentine's Day is a prank to show mischief, when lovers show their crush to each other; Batman is an abusive superhero, a fiery child, and grew up learning different ways to glare (from Table 6).

We note that similar to the observation in (Hossain et al., 2017), when humans write funnier stories, they do so by incorporating more coherence. Table 6 shows a US-written FT story with high MFG from US judges (IN judges often give higher grades to YODALIB stories). The US turker portrays Batman as a bland and an uninteresting person who is weak, cowardly and moronic, living in a disconcerted city of Gotham. Humor is generated by portraying the title concepts consistently with surprising and

Advertising is how a company *cheats* / grills people to buy their products, services or *dreams* / optics . . . that draws *silly* / hotter attention towards these things. Companies use ads to try to get people to *forget* / fling their products, by showing them the good rather than the bad of their *products* / earrings. For example, to make a *burger* / dynamite look tasty in advertising, it may be painted with brown food colors, sprayed with *oil* / crocodile to prevent it from going *dull* / graceful, and sesame seeds may be super-glued in place. Advertising can bring new *scapegoats* / sickness and more sales for a business. Advertising can be *useless* / grim . . .
Valentine's Day is a *fixture* / prank that happens on February 14 . . . when lovers show their *toe* /crush to each other . . . by giving Valentine's cards or just a *stinky* /handy gift. Some people *kill* /buffet one person and call them their Valentine as a gesture to show *equipment* /mischief and appreciation. Valentine's Day is named for the *gross* /annoying Christian saint . . . who performed *gorillas* /outdoors for couples who were not allowed to get married because their *squids* /circus did not agree with the connection . . . so the marriage was *eaten* /exploded. Valentine gave the married couple flowers from his *casket* /problem. That is why flowers play a very *hungry* /abusive role on Valentine's Day.

Table 5: Examples with YodaLib stories rated as funnier than FT stories (top: IN, down: US. Filled-in word order: *FreeText*/YODALIB).

Batman is one of the most *bland* / abusive superheroes. He was the second *wimp* / pest to be created . . . lives in the *discombobulated* / tool city of Gotham . . . origin story as a *moronic* / fiery child, Bruce Wayne saw a robber *kiss* / spin his parents after the family left a *bakery* / teddy . . . he did not want that kind of *romance* / wayne to happen to anyone else. He dedicated his life to *terrorize* / tolerate Gotham City. Wayne learned many different ways to *grow* / glare as he grew up . . .

Table 6: An example story where US-FT humor is better than US-YodaLib humor Filled-in word order: *FreeText*/YODALIB.

unexpected views. It is interesting that Batman is depicted with two traits consistently: his bland and moronic personality, and his opposition to romance in Gotham due to a robber kissing his parents. This is striking, as it illustrates how good skilled humans of IN and US origin can be at generating humor via multiple coherent and meaningful concepts. However, this is seen in a very few stories, possibly due to the biases humans may have in what they consider humorous. Nevertheless, it provides us with future directions to pursue to generate demographic-aware humor of even higher quality in terms of coherence.

7 Conclusion

In this paper, we studied location-specific humor generation in Mad Libs stories. We first collected a novel location-specific humor dataset on AMT, by selecting players and judges to obtain ground truth data. Next, we proposed an automated location-specific humor generation framework to generate possible candidates to fill-in the blanks, to rank these candidates based on their humor contributions to form funny sentences, and to complete the stories by selecting the best transformations for the constituting sentences. Our approach outperformed a simple language model and human players (in most cases) in generating funny stories.

We also performed a detailed demographic-based analysis of our dataset. We found that humor created with US slant is in general preferred to IN slant humor by both IN and US judges. IN judges seemed to have a more lenient outlook towards humor, while US judges have higher expectations possibly in terms of coherence, which is also reflected in the better quality humor generated in the US-specific setting. When turkers wrote funnier stories, they did so in a coherent manner, indicating the vast potential of coherence in generating humor, contrary to the general incongruent take of Mad Libs. Humor from our approach is generated primarily via incongruity and deviation, despite our preliminary measures to incorporate coherence. We believe that our proposed BERT-based approach is general, and can be used for other affect-based NLP tasks with minor changes. In the future, we aim to extend our research in several directions, using (1) a larger number of and a wider variety of Mad Lib-like stories; and (2) other demographic dimensions (e.g., gender, age group, occupation), and more groups within each dimension (e.g., Singapore, Canada, England for location; Arts, Engineering, Fashion, Publishing for occupation).

We also plan to take steps toward understanding what makes textual humor coherent, and go beyond our word and sentence similarity measures to generate more coherent and funny stories.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation (grant #1815291) and by the John Templeton Foundation (grant #61156). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation or the John Templeton Foundation. The authors would like to thank the reviewers for their comments.

References

Dana L Alden, Wayne D Hoyer, and Chol Lee. 1993. Identifying global and culture-specific dimensions of humor in advertising: A multinational analysis. *The Journal of Marketing*, pages 64–75.

Mahadev L Apte. 1985. *Humor and laughter: An anthropological approach*. Cornell Univ Pr.

Salvatore Attardo and Victor Raskin. 1991. Script theory revis (it) ed: Joke similarity and joke representation model. *Humor-International Journal of Humor Research*, 4(3-4):293–348.

Salvatore Attardo. 2010. *Linguistic theories of humor*, volume 1. Walter de Gruyter.

David Bamman, Chris Dyer, and Noah A. Smith. 2014. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (ACL 2014)*, pages 828–834.

Dario Bertero and Pascale Fung. 2016. A long short-term memory framework for predicting humor in dialogues. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 130–135.

Kim Binsted, Helen Pain, and Graeme D Ritchie. 1997. Children's evaluation of computer-generated punning riddles. *Pragmatics & Cognition*, 5(2):305–354.

Vladislav Blinov, Valeria Bolotova-Baranova, and Pavel Braslavski. 2019. Large dataset and language model fun-tuning for humor recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4027–4032, Florence, Italy, July. 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, June. Association for Computational Linguistics.

W Jack Duncan, Larry R Smeltzer, and Terry L Leap. 1990. Humor and work: Applications of joking behavior to management. *Journal of Management*, 16(2):255–278.

Penelope Eckert and Sally McConnell-Ginet. 2013. *Language and gender*. Cambridge University Press.

Aparna Garimella and Rada Mihalcea. 2016. Zooming in on gender differences in social media. *PEOPLES 2016*, page 1.

Aparna Garimella, Rada Mihalcea, and James Pennebaker. 2016. Identifying cross-cultural differences in word usage. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING 2016)*, pages 674–683, Osaka, Japan.

Aparna Garimella, Carmen Banea, and Rada Mihalcea. 2017. Demographic-aware word associations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2275–2285.

Lizbeth Goodman. 1992. *Gender and humour*. na.

Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, et al. 2019. Ur-funny: A multimodal language dataset for understanding humor. *arXiv preprint arXiv:1904.06618*.

Jennifer Hay. 2000. Functions of humor in the conversations of men and women. *Journal of pragmatics*, 32(6):709–742.

Nabil Hossain, John Krumm, Lucy Vanderwende, Eric Horvitz, and Henry Kautz. 2017. Filling the blanks (hint: plural noun) for mad libs humor. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 638–647.

Nabil Hossain, John Krumm, and Michael Gamon. 2019. "President vows to cut <taxes> hair": Dataset and analysis of creative text editing for humorous headlines. 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 133–142, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL 2015)*, pages 752–762, Beijing, China.

Barbara Johnstone. 2010. Language and place.

Chloe Kiddon and Yuriy Brun. 2011. That's what she said: double entendre identification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 89–94. Association for Computational Linguistics.

Cheris Kramarae. 1981. Women and men speaking: Frameworks for analysis.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Bill Yuchen Lin, Frank F. Xu, Kenny Zhu, and Seung-won Hwang. 2018. Mining cross-cultural differences and similarities in social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 709–719, Melbourne, Australia, July. Association for Computational Linguistics.

Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 78–87, New Orleans, LA, June. Association for Computational Linguistics.

Rada Mihalcea and Carlo Strapparava. 2006. Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence*, 22(2):126–142.

John Morreall. 2012. Philosophy of humor.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51.

Dan O'Shannon. 2012. *What are You Laughing At?: A Comprehensive Guide to the Comedic Event*. A&C Black.

Saša Petrović and David Matthews. 2013. Unsupervised joke generation from big data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 228–232.

Roger Price and Leonard Stern. 1974. *Mad Libs: World's Greatest Party Game: a Do-it-yourself Laugh Kit*. Number 1. Mad Libs.

Yishay Raz. 2012. Automatic humor classification on twitter. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 66–70. Association for Computational Linguistics.

Dawn T Robinson and Lynn Smith-Lovin. 2001. Getting a laugh: Gender, status, and humor in task discussions. *Social Forces*, 80(1):123–158.

Oliviero Stock and Carlo Strapparava. 2002. Hahacronym: Humorous agents for humorous acronyms. *Stock, Oliviero, Carlo Strapparava, and Anton Nijholt. Eds*, pages 125–135.

Margaret E. Tresselt and Mark S. Mayzner. 1964. The Kent-Rosanoff word association: Word association norms as a function of age. *Psychonomic Science*, 1(1-12):65–66.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, number October, pages 1815–1827, Seattle, WA, USA.

Scott Weems. 2014. *Ha!: The science of when we laugh and why*. Basic Books (AZ).

Charles Welch, Jonathan Kummerfeld, Veronica Perez-Rosas, and Rada Mihalcea. 2020. Compositional demographic word embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Julia Wilkins and Amy Janel Eisenbraun. 2009. Humor theories and the physiological benefits of laughter. *Holistic nursing practice*, 23(6):349–354.

Steven Wilson, Rada Mihalcea, Ryan Boyd, and James Pennebaker. 2016. Disentangling topic models: A cross-cultural analysis of personal values through words. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 143–152, Austin, Texas, November. Association for Computational Linguistics.

Renxian Zhang and Naishi Liu. 2014. Recognizing humor on twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 889–898. ACM.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.