LexDivPara: A measure of paraphrase quality with integrated sentential lexical complexity

Thanh Thieu[1], Ha Do[2], Thanh Duong[1], Shi Pu, Sathyanarayanan Aakur[1], and Saad Khan[3]

[1] Oklahoma State University, Stillwater OK 74078, USA
tthieu@okstate.edu
WWW home page: http://languageandintelligence.cs.okstate.edu
[2] University of Louisville, Louisville, KY 40292, USA
[3] FineTune Learning, Boston, MA, USA

Abstract. We present a novel method that automatically measures *quality* of sentential paraphrasing. Our method balances two conflicting criteria: semantic similarity and lexical diversity. Using a diverse annotated corpus, we built learning to rank models on edit distance, BLEU, ROUGE, and cosine similarity features. Extrinsic evaluation on STS Benchmark and ParaBank Evaluation datasets resulted in a model ensemble with moderate to high quality. We applied our method on both small benchmarking and large-scale datasets as resources for the community.

Keywords: monolingual rewriting, lexical diversity, paraphrasing quality

1 Introduction

In linguistics, lexical complexity is a multidimensional measure encompassing lexical diversity, lexical density, and lexical sophistication [13, 16, 22]. Modern natural language processing (NLP) adopted a bag-of-features approach on lexical complexity for paraphrase simplification. The general strategy is to perform complex word identification (CWI) [1, 14, 17, 25] and then substitute those with simpler words. Four categories of features used in CWI were: (1) word-level features such as word length, syllable counts, (2) morphological features such as part-of-speech, suffix length, noun gender, (3) semantic features derived from WordNet or cosine similarity between word embedding vectors, and (4) corpus-based features such as word frequencies, n-gram frequencies, or topic distribution in some reference corpora. These strategies measured single word and short phrase complexity, thus rendering them unsuitable for measuring complexity of complete sentences.

In sentential monolingual rewriting, most modern NLP methods focused on semantic similarity between a reference sentence and its paraphrases [5, 9, 24]. Recent work sought to improve the lexical diversity of paraphrases by adding

heuristic lexical constraints to the decoder [8, 10]. However, these works resulted in most highly ranked paraphrases that were almost lexically identical to the references. Thus, paraphrase generation became a trivial task unusable for practical purposes such as: content generation in education, data augmentation in language modeling, question answering, textual entailment, etc. Table 1 shows examples of top ranking paraphrases from two human annotated datasets: STS Benchmark[4] and ParaBank Evaluation[5].

Table 1. Example top ranking **R**eference/**P**araphrase pairs in semantic similarity by humans in **STS** Benchmark and **ParaB**ank Evaluation datasets.

Datasets	**Top Examples**
STS	**R**: A man with a hard hat is dancing.
	P: A man wearing a hard hat is dancing.
	R: A man is feeding a mouse to a snake.
	P: The man is feeding a mouse to the snake.
ParaB	**R**: You weigh a million pounds.
	P: You weigh one million pounds.
	R: Ladies and gentlemen, young people.
	P: Ladies and gents, young people.

In this study, we present a learnt quality measure of paraphrases that addresses the low lexical diversity issue in sentential paraphrasing. Our method not only aligns with semantic similarity, but also significantly enhances the difference in lexical use between a paraphrase and its reference. Such desideratum was referred to as *quality* or *fluency* of paraphrases [8]. We also adopted a bag-of-features approach but did not use the four feature categories of CWI since these features were developed for single-word complexity while our task aims to measure sentential complexity. We modeled paraphrase quality as a learning-to-rank problem on a controlled corpus generated by educational specialists and annotated by Amazon Mechanical Turk workers. We then used the trained model to re-rank paraphrases in STS Benchmark and ParaBank Evaluation datasets and showed that our model picked paraphrases with superior quality.

Hereinafter, we detail data collection, feature engineering, and measure modeling in the Method section. Extrinsic evaluation is presented in the Results section, and contribution together with model characteristics are explained in the Discussion section. Lastly, we highlight the novelty and impact of this study in the Conclusion section.

[4]https://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark

[5]https://github.com/decompositional-semantics-initiative/ParaBank-Eval-Data

2 Method

We define *quality* as the holistic fusion between semantic similarity and lexical variation of a paraphrase compared to its reference sentence.

2.1 Data collection

Our dataset comprises of 5 documents, totaling 92 English sentences, from ACT Test Preparation textbooks. Topics includes sport, history, biological science, tourism, and geography.

To generate paraphrases, we used Google Cloud Translation to translate each reference English sentence into 10 foreign languages, then back-translate these 10 foreign sentences into English. The 10 foreign languages included Japanese, Korean, Chinese, German, Spanish, Portuguese, Greek, Arabic, Slovenian, and Turkish. This process generated 10 paraphrases per reference, resulting in a dataset of 920 paraphrase/reference pairs.

We hired English speakers on Amazon Mechanical Turk to annotate quality scores for the paraphrases. We restricted annotators to at least had graduated from a U.S. high school and possessed excellent reviews by previous requesters. Given the ACT documents were for college entrance exam, the selected annotator population qualified to perform our task. We adopted the EASL framework [23] to increase annotation efficiency by presenting all ten paraphrases (of the same reference sentence) simultaneously in one page so that annotators could compare between items while giving scores. We re-used the HTML template from EASL to generate task pages on Amazon MTurk. The annotators were asked to give a score in a range [0, 100] to each paraphrase/reference pair. Additionally, each pair was annotated by 10 different annotators. Score 100 corresponds to same meaning and different wording, while the contrasted score 0 is hypothetical and corresponds to different meaning and same wording. Thus, the score measures *quality* of the paraphrase. In total we obtained 9,200 paraphrase/reference rankings.

2.2 Feature engineering

Table 2. Summary of semantic and lexical features

Type	Category	Metric
Semantic	Sentence embedding	Cosine similarity
Lexical	Edit distance	Constituent tree, word sequence, character sequence
	BLEU	1-gram, 2-gram, 3-gram, 4-gram
	ROUGE	1-gram, 2-gram, 3-gram, 4-gram, longest common subsequence (LCS), weighted LCS

To model the fusion between semantic similarity and lexical difference, we combined cosine similarity with edit distance and machine translation scores. In total there were one semantic feature and thirteen lexical features. Table 2 gives a summary of the features. In following description, we refer to *two sentences* as a reference/paraphrase pair.

Cosine similarity: We invoked the universal sentence encoder [6] on deep averaging network (DAN) [12] to generate embeddings of the two sentences, then calculated cosine distance between the two embedding vectors to represent semantic similarity.

Tree edit distance: We invoked the Stanford CoreNLP toolkit [18] to parse constituent trees of the two sentences, then used Zhang-Shasha algorithm [27] to compute the edit distance between the two trees. In addition, we normalized the distance by the total number of nodes in both trees. Tree edit distance represents the difference in grammatical structure between the two sentences.

Word and character edit distances: We used NLTK [3] implementation of Levenshtein edit-distance [19] with substitution cost set to 2 to compute the transformation cost between the two sequences of words (or characters) of the two sentences. We also normalized the cost by the total number of words (or characters) of both sentences. This normalization engulfs the substitution cost of 2, which represent the removal of one word (or character) from a sentence while adding another word (or character) into the other sentence. Sequence edit distances represent the ordinal difference in vocabulary use between the two sentences.

BLEU scores: We used NLTK [3] implementation of bilingual evaluation understudy [20] to compute modified precision of overlapping n-grams for individual orders of n-gram. BLEU scores represent the precision of paraphrase n-grams that match the reference sentence.

ROUGE scores: We used PyPI's py-rouge package [2] to compute recall-oriented understudy [15] of overlapping n-grams for individual orders of n-grams, longest common subsequence (LCS), and weighted LCS. ROUGE scores represent the recall of reference sentence's subsequences that match the paraphrase.

2.3 Learning to rank paraphrase quality

We formulated paraphrase quality as a learning-to-rank problem in information retrieval. The reference sentence serves as a query; paraphrases serves as retrieved documents; and paraphrase quality score serves as the relevance score. The learning-to-rank formulation optimizes relative orders of paraphrases; thus, it is robust to both the inconsistency in score ranges and distances.

We utilized XGBoost gradient boosted trees [7] to train our ranking model. XGBoost was successfully used by multiple winners in machine learning challenges and Kaggle competitions. We parametrized XGBoost to use LambdaMART [4] to perform list-wise ranking with mean average precision (MAP) objective function. Learning rate was set to 0.1; minimum loss reduction was set to 1.0; minimum sum of instance weights in a child was set to 0.1; maximum depth of a tree was set to 6; number of trees was set to 10. To evaluate model performance,

we implemented five-fold cross-validation with 80% data for training and 20% data for validation, then used Scikit-learn's implementation [21] of normalized discounted cumulative gain (NDCG) for evaluation. We experimented with some variation of above parameter settings and found insignificant NDCG gain/loss.

2.4 Label smoothing regularization

We observed that annotators diverted to different score ranges and scales. For example, ten paraphrases of a sentence "The Fulton fish market" were given scores [60, 61, 60, 77, 50, 50, 50, 60, 70, 60] by one annotator and scores [42, 45, 51, 57, 49, 55, 55, 53, 45, 51] by another annotator. Thus, the first annotator preferred scores in range [50, 80] while the second annotator preferred scores in range [40, 60]. Calculating Spearman's rank correlation coefficient between annotation scores of paraphrases from the same reference sentence resulted in mean = 0.24 and standard deviation = 0.40. Agreement between annotators were low and spreading. We hypothesizes that quantification of paraphrase quality was affected by annotators' individual bias. Since we pioneered the study of sentential paraphrase's holistic quality in this work, we were not aware of any prior formal composition of paraphrase quality, nor a proven scale to reduce annotator bias.

To smooth annotator bias, we first experimented with z-score normalization in various scopes (e.g. per annotator, per reference sentence, and per paraphrase), but they all resulted in the same rank correlation coefficient. By trials and errors, we discovered that the scores could be smoothed using their sorted indices. Specifically, we sorted the original scores and then substituted them by their indices in the sorted list. Thus, scores [80, 89, 60, 78, 76, 74, 63, 32, 72, 70] becomes [1, 0, 8, 2, 3, 4, 7, 9, 5, 6]. When ties occured, the earlier item in the list was aribitrarily assigned a smaller index score. Hereinafter, we denote models using the smoothed labels as Index models, and models using the original annotator scores as Raw models. Spearman's rank correlation coefficient on Index scores agreement reached higher means = 0.28 and smaller standard deviation = 0.38 compared to Raw scores agreement.

2.5 Augmenting semantic similarity

Measuring semantic similarity based on sentence embedding is a challenging task [5, 26], and we expected our semantic feature to be a weak one. To offset this weakness, we experimented augmentation of our paraphrase quality score (Q) with the human annotated semantic similarity score (S) in benchmarking datasets (STS Benchmark and ParaBank Evaluation). We experimented with linear combinations of Q and S and found the best linear coefficient performed equally well as a harmonic mean F1 combination. We picked the balanced F1 as the combined score to simplify hyper-parameter tuning.

$$F_1(Q, S) = 2 \times Q \times S/(Q + S)$$

Table 3. Top ranking examples from the STS-Benchmark and ParaBank datasets. Raw and Index (Idx) models were ranked by Q. Augmented models (prefix A-) were ranked by F_1. S is ground truth for semantic similarity. Q reflects lexical diversity. STS score range is [0, 5]; ParaBank score range is [0, 100].

Model	Reference(**R**)-Paraphrase(**P**) Sentence Pair	S	Q
	STS Benchmark		
Raw	**R**: A man is cutting a potato. **P**: A man is slicing some potato.	4.4	3.16
ARaw	**R**: A man is playing the drums. **P**: A man plays the drum.	5.0	1.47
Idx	**R**: A man plays an acoustic guitar. **P**: A woman and dog are walking together.	0.0	4.51
AIdx	**R**: I realized there is already an accepted answer but I figure I would add my 2 cents. **P**: I know this is an old question but I feel I should add my 2 cents.	5.0	4.01
	R:You may have to experiment and find what you like. **P**:You have to find out what works for you.	5.0	4.49
	ParaBank Evaluation		
Raw	**R**: I've known Miguel since childhood. **P**: I knew Miguel from childhood.	87	36
ARaw	**R**: You're confusing humility, with humiliation. **P**: I think you mistake humility with humiliation.	100	24
Idx	**R**: I am at your service. **P**: Dyce's here to see you.	20	90
AIdx	**R**: One doesn't detect the tiniest trace of jealousy, does one? **P**: I don't hear a tiny undertone of jealousy in your voice?	99	90
	R: Let me check once again. **P**: I'll look again.	100	87

3 Results

Our models were evaluated on two extrinsic benchmarking datasets: STS Benchmark and ParaBank Evaluation. In each dataset, we applied both Raw and Index models together with their augmneted versions (Section 2.5). We then sorted the datasets in descending order of computed scores and compared top ranking paraphrases. In Table 3, each pair of reference/paraphrase sentences was accompanied by a semantic score S, and a lexical diversity score Q. We chose score Q of the Index model to represent lexical diversity because it best reflected the lexico-grammatical difference between the two sentences.

Results showed that the Raw model preserved the reference meaning well but only performed moderately on lexical diversity. Paraphrases of the Raw model repeated key phrases from the reference sentences. Augmented Raw model gave higher semantic similarity but at the cost of reduced lexical diversity. Index model failed at preserving reference meaning, but prevailed at promoting lexical diversity. Paraphrases found by the vanilla Index model showed large lexico-

grammatical difference, but also carried almost different meaning from the reference sentences. The Augmented Index model was the most interesting one as it not only throve at semantic similarity, but also showed high lexical diversity. Paraphrases found by the Augmented Index model expressed significant lexico-grammatical difference while preserving the original meaning of the reference sentences. In addition to Table 3, we made the full ranking of STS Benchmark and ParaBank Evaluation datasets publicly available for community investigation.

To gain insights and intuition about model behaviors, we analyzed models' feature importance based on decision trees' node split gain (Figure 1). The Raw model prioritized cosine similarity for sentence meaning and ROUGE-1 for single-word lexical difference. This explained its tendency to keep almost identical meaning and picked paraphrases with few single-word difference. In contrast, the Index model prioritized tree edit distance for difference in grammatical structure and ROUGE-L for variation of long sub-phrases. The Index model was the opposite of the Raw model. The Index model's features favored lexico-grammatical difference at the expense of reference meaning. When augmented with a strong semantic similarity signal S, the Augmented Raw model inclined even more toward preserving reference meaning, while the Augmented Index model achieved a rare equilibrium that produced both significant lexico-grammatical difference and strong similarity with reference meaning. In our study, Augmented Index was the highest *quality* ranking model for monolingual paraphrasing.

We call our method LexDivPara for lexical diversity in paraphrasing. Our experiment and evaluation suggested that the Augmented Index model should be used when a strong feature for semantic similarity is available. Otherwise, the Raw model should be used to deliver a moderate quality device for paraphrase ranking. In addition to scoring paraphrase quality of STS Benchmark and ParaBank Evaluation datasets, we have scored and sorted the ParaBank 2.0 dataset [11] comprising of 19 million reference sentences and made it publicly available as a large-scale resource for researchers interested in training good quality paraphrase generative models.

4 Discussion

Measuring quality of monolingual paraphrasing is a challenging task, as it struggles to balance between two conflicting desiderata: semantic preservation and maximal lexical variation. In this work, we projected a holistic quality score for paraphrases and factorized it into two contradicting components: semantic similarity, and lexical diversity. While semantic similarity had been studied well in the computational linguistics literature [5, 9, 24], sentential lexical diversity was mostly un-explored. A recent work close to ours is the development of ParaBank [8, 10] that used heuristic lexical constraints to encourage diverse use of words in the decoding sequence. However, this work only used one single measure, BLEU without length penalty, to evaluate how different the paraphrases are to the

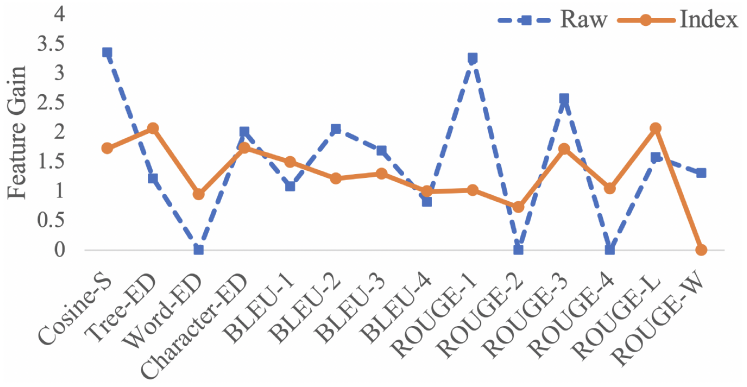

Fig. 1. Feature importance of XGBoost ranking models. Suffixes: -S:similarity, -ED:edit distance.

reference sentences. Furthermore, their use of an evaluation data set that contained many of mostly identical paraphrase/reference pairs made the treatment of lexical diversity subjective and incomplete. To comprehensively assess lexical diversity at sentential level, our work expanded the set of features to encompass multiple machine translation measures that reflected precision, recall, and edit distance statistics.

Our contribution involves three folds: (1) quantify lexical diversity measure between two sentences, (2) annotate a sentential paraphrase quality corpus, and (3) evaluate learnt models on extrinsic datasets. Nevertheless, our work is restricted to a limited training dataset and our top-performing model relied on the availability of high-quality semantic similarity scores. Interested researchers could overcome this challenge by either leveraging STS Benchmark and ParaBank Eval in semi-supervision learning, or utilizing multiple embedding methods to improve semantic similarity representation. To expedite research on sentential paraphrasing, we release the scored and sorted, large scale ParaBank 2.0 dataset that contains millions of sentences.

5 Conclusion

We presented a novel, accurate method to measure *quality* of paraphrases. Our work incorporated lexical diversity at the sentential level, in contrast to existing work in computational linguistics that was constrained to single-word and phrasal levels. We established the first learnt measure for paraphrase quality using supervised learning. Our machine learning models were free from heuristic rule construction and lexical choice guess work. We expect this study to provide resource and methodology to the under-active lexical diversity aspect of language generation. Potential future work includes investigating alternative high-quality semantic similarity scores, filtering high quality bitext corpora for machine translation, or embedding quality measure into end-to-end language generation mod-

els. Our source code, feature set, annotated data, and ranked datasets are freely available at: http://languageandintelligence.cs.okstate.edu/tools.

Acknowledgment

The authors would like to thank ACT for assisting with collection of the original text and annotation on Amazon Mechanical Turk. This work is partly supported by the first author's start-up fund, the first author's OSU ASR FY22 summer program, NSF CISE/IIS 1838808 grant, and NSF OIA 1849213 grant.

References

[1] Alfter D, Volodina E (2018) Towards Single Word Lexical Complexity Prediction. In: Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, New Orleans, Louisiana, pp 79–88, URL https://www.aclweb.org/anthology/W18-0508http://dx.doi.org/10.18653/v1/W18-0508

[2] Antognini D (2018) Py-rouge. URL https://github.com/Diego999/py-rouge

[3] Bird S, Klein E, Loper E (2009) Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc."

[4] Burges CJC, Svore KM, Wu Q, Gao J (2008) Ranking, boosting, and model adaptation

[5] Cer D, Diab M, Agirre E, Lopez-Gazpio I, Specia L (2017) SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. Association for Computational Linguistics, Vancouver, Canada, pp 1–14, DOI 10.18653/v1/S17-2001, URL https://www.aclweb.org/anthology/S17-2001https://doi.org/10.18653/v1/S17-2001

[6] Cer D, Yang Y, Kong Sy, Hua N, Limtiaco N, St John R, Constant N, Guajardo-Cespedes M, Yuan S, Tar C, Strope B, Kurzweil R (2018) Universal Sentence Encoder for English. In: 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Brussels, Belgium, pp 169–174, URL https://www.aclweb.org/anthology/D18-2029

[7] Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System. DOI 10.1145/2939672.2939785, URL https://doi.org/10.1145/2939672.2939785

[8] Edward Hu J, Rudinger R, Post M, Van Durme B (2019) ParaBank: Monolingual Bitext Generation and Sentential Paraphrasing via Lexically-constrained Neural Machine Translation. In: AAAI 2019, Honolulu, Hawaii

[9] Ganitkevitch J, Van Durme B, Callison-Burch C (2013) PPDB: The Paraphrase Database. Association for Computational Linguistics, pp 758–764, URL http://aclweb.org/anthology/N13-1092

[10] Hu JE, Khayrallah H, Culkin R, Xia P, Chen T, Post M, Durme BV (2019) Improved Lexically Constrained Decoding for Translation and Monolingual Rewriting. In: NAACL 2019, Minneapolis, Minnesota

[11] Hu JE, Singh A, Holzenberger N, Post M, Van Durme B (2019) Large-Scale, Diverse, Paraphrastic Bitexts via Sampling and Clustering. In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, Hong Kong, China, pp 44–54, URL https://www.aclweb.org/anthology/K19-1005http://dx.doi.org/10.18653/v1/K19-1005

[12] Iyyer M, Manjunatha V, Boyd-Graber J, DaumÈ III H (2015) Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Beijing, China, pp 1681–1691, URL https://www.aclweb.org/anthology/P15-1162http://dx.doi.org/10.3115/v1/P15-1162

[13] Johansson V (2009) Lexical diversity and lexical density in speech and writing: a developmental perspective. Lund Working Papers in Linguistics 53:61–79

[14] Kriz R, Miltsakaki E, Apidianaki M, Callison-Burch C (2018) Simplification Using Paraphrases and Context-Based Lexical Substitution. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, pp 207–217, URL https://www.aclweb.org/anthology/N18-1019http://dx.doi.org/10.18653/v1/N18-1019

[15] Lin CY (2004) ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, pp 74–81, URL https://www.aclweb.org/anthology/W04-1013

[16] Lu X (2012) The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. The Modern Language Journal 96(2):190–208, DOI 10.1111/j.1540-4781.2011.01232_1.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-4781.2011.01232{_}1.x

[17] Maddela M, Xu W (2018) A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, pp 3749–3760, URL https://www.aclweb.org/anthology/D18-1410http://dx.doi.org/10.18653/v1/D18-1410

[18] Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D (2014) The Stanford CoreNLP Natural Language Processing Toolkit. Association for Computational Linguistics, pp 55–60, DOI 10.3115/v1/

P14-5010, URL http://aclweb.org/anthology/P14-5010http://dx.doi.org/10.3115/v1/P14-5010

[19] Miller FP, Vandome AF, McBrewster J (2009) Levenshtein Distance: Information theory, Computer science, String (computer science), String metric, Damerau?Levenshtein distance, Spell checker, Hamming distance. Alpha Press

[20] Papineni K, Roukos S, Ward T, Zhu WJ (2002) Bleu: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp 311–318, URL https://www.aclweb.org/anthology/P02-1040http://dx.doi.org/10.3115/1073083.1073135

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

[22] Read J (2000) Assessing Vocabulary. Cambridge University Press, Cambridge, DOI DOI:10.1017/CBO9780511732942, URL https://www.cambridge.org/core/books/assessing-vocabulary/C38BC270DC287C6F6107CDE34BF73E88

[23] Sakaguchi K, Van Durme B (2018) Efficient Online Scalar Annotation with Bounded Support. Association for Computational Linguistics, Melbourne, Australia, pp 208–218, URL https://www.aclweb.org/anthology/P18-1020

[24] Wieting J, Gimpel K (2018) ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations. Association for Computational Linguistics, pp 451–462, URL http://aclweb.org/anthology/P18-1042

[25] Wilkens R, Vecchia AD, Boito MZ, Padró M, Villavicencio A, Bazzan ALC, Pichara K (2014) Size Does Not Matter. Frequency Does. A Study of Features for Measuring Lexical Complexity. Springer International Publishing, Cham, pp 129–140

[26] Yang Y, Cer D, Ahmad A, Guo M, Law J, Constant N, Abrego GH, Yuan S, Tar C, Sung YH, Strope B, Kurzweil R (2019) Multilingual Universal Sentence Encoder for Semantic Retrieval pp 87–94, URL http://arxiv.org/abs/1907.04307, 1907.04307

[27] Zhang K, Shasha D (1989) Simple fast algorithms for the editing distance between trees and related problems. SIAM J Comput 18(6):1245–1262, DOI 10.1137/0218082