Tübingen-CL at SemEval-2024 Task 1: Ensemble Learning for Semantic Relatedness Estimation

Leixin Zhang
University of Tübingen, Germany
leixin.zh@gmail.com

Çağrı Çöltekin
University of Tübingen, Germany
cagri.coeltekin@uni-tuebingen.de

Abstract

The paper introduces our system for SemEval-2024 Task 1, which aims to predict the relatedness of sentence pairs. Operating under the hypothesis that semantic relatedness is a broader concept that extends beyond mere similarity of sentences, our approach seeks to identify useful features for relatedness estimation. We employ an ensemble approach integrating various systems, including statistical textual features and outputs of deep learning models to predict relatedness scores. The findings suggest that semantic relatedness can be inferred from various sources and ensemble models outperform many individual systems in estimating semantic relatedness.

1 Introduction

Identifying semantic relatedness is a 'related' task to many well-studied tasks of semantic similarity. According to Abdalla et al. (2023), two sentences are considered similar if they are paraphrases or share a relation of entailment. Semantic relatedness, however, is a broader concept than semantic similarity. Two expressions are considered related if they share any semantic association. For instance, 'teacher' and 'student' are related because they frequently occur within the same context or domain. Similarly, 'tasty' and 'unpalatable' are related, as both terms are used to describe food, albeit with opposite meanings.

SemEval-2024 Task 1 (Ousidhoum et al., 2024b) is designed to estimate the relatedness of sentence pairs. The task is based on a multilingual dataset of 14 languages and offers supervised, unsupervised and cross-lingual tracks. Our team participated in two tracks, and a subset of available languages: Track A (supervised learning) for English, and Track B (unsupervised learning) for English, Spanish, and Hindi.

We posit that semantic relatedness can be inferred from a multitude of sources and therefore propose an ensemble approach that integrates outcomes from diverse systems to estimate semantic relatedness. Our study explores features from textual statistical analysis, general large language models, word embedding models, and models trained on semantic labeled datasets, question-answering pairs, or title-passage pairs in estimating semantic relatedness, and we conducted ensemble experiments with these features.

2 Related Work

SemEval in previous years has introduced tasks focusing on semantic textual similarity to evaluate the degree of similarity between sentence pairs (Agirre et al., 2012; Manandhar and Yuret, 2013; Agirre et al., 2014; Cer et al., 2017). There tasks provided datasets with human labeled similarity scores, which have been extensively utilized for training sentence embedding models and conducting semantic evaluations (Wieting et al., 2015; Cer et al., 2018; Reimers and Gurevych, 2020; Feng et al., 2022).

2.1 Sentence Embeddings

Word embedding models such as BERT (Devlin et al., 2019), GloVe (Pennington et al., 2014), RoBERTa (Liu et al., 2019), and Word2Vec (Mikolov et al., 2013) are frequently employed to assess the semantic distance between words. Sentence embeddings with a fixed length are often generated via mean/max pooling of word embeddings or employing `CLS` embedding in BERT. The semantic distances are commonly measured using the cosine similarity of embeddings of two expressions.

Siamese or triplet network architectures are frequently employed in sentence embedding training. For example, models such as Sentence-BERT (Reimers and Gurevych, 2019, 2020) utilize a dual-encoder architecture with shared weights for

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1019–1025
June 20-21, 2024 ©2024 Association for Computational Linguistics

predicting sentence relationships (e.g., semantic contradiction, entailment, or neutral labeling) or for similarity score prediction using regression objectives, e.g., the difference between human annotated similarity score (sim) of two sentences and the cosine of two sentence embeddings (v and u), illustrated in Equation (1).

$$\mathcal{L} = |\cos(v, u) - \text{sim}| \tag{1}$$

In triplet neural networks, an anchor sentence (u) can be trained along with a positive sample (a sentence with a similar meaning) and a negative sample (a sentence with a dissimilar meaning), with contrastive loss. InfoNCE (Noise-Contrastive Estimation) can be utilized as the objective function. A larger number of negative samples can also be integrated into neural networks through the application of InfoNCE, as demonstrated in Equation (2). Here, v^+ denotes positive samples. The negative sample size is denoted as K, and the total sample size (including one positive sample) as $K + 1$. This approach is adopted by the Jina embedding model (Günther et al., 2023), which is used in our ensemble system.

$$\mathcal{L} = -\mathbb{E}\left[\log \frac{f(v^+, u)}{\sum_{i=1}^{K+1} f(v_i, u)}\right] \tag{2}$$

2.2 Ensemble Learning

In previous studies, ensemble learning presents several advantages. The ensemble approach can reduce the errors from individual models by amalgamating results from multiple sources or can make the system more robust. In our study, using multiple pre-trained models can also save a substantial amount of computation while making use of information from the large data during pre-training. Previous research has demonstrated that ensemble learning can achieve remarkable success (Huang et al., 2023; Osika et al., 2018).

In our study, we aim to integrate multiple deep learning models to assess semantic relatedness. When models are trained on diverse datasets with different architectures, they may produce varied predictions on semantic relatedness, and combining them may improve overall performance.

We use sentence embeddings mainly from the following models. Sentence-BERT (Reimers and Gurevych, 2019) is trained on datasets involving SNLI (a collection of 570,000 sentence pairs) and MultiNLI (comprising 430,000 sentence pairs). The Jina Embedding model (Günther et al., 2023) utilizes 385 million sentence pairs and 927,000 triplets (comprising positive and negative samples of semantic similarity) after a filtering process. The T5 model is trained on approximately 7 TB of text data derived from Common Crawl, serving various text-to-text purposes (Raffel et al., 2020; Ni et al., 2021).

3 Methodology

In this study, we hypothesize that semantic relatedness covers a broader spectrum than semantic similarity in theory. Consequently, the integration of various systems and features should achieve superior results compared to individual systems.

3.1 Supervised Learning

For the supervised track[1], we first evaluated subsystems in an unsupervised manner and selected those with a higher Spearman's correlation with human annotations for ensemble learning. The selected results were then further fine-tuned using the training data (5,500 English sentence pairs labeled with relatedness scores provided by the shared task, Ousidhoum et al., 2024a) to achieve closer alignment with human annotations.

In the following subsections, we present the features and systems utilized for ensemble learning. The features can be classified into three categories: textual statistical features (Section 3.1.1), word embedding models (Section 3.1.2), and sentence embedding models (Section 3.1.3).

3.1.1 Textual Statistical Features

Our analysis began with surface-level textual statistical features, including word overlap and the Levenshtein distance measurement at the character level. These scores were then normalized into ratios to estimate their correlation with human-annotated relatedness. Specifically, we considered the following features:

- Character Distance Ratio: normalization of Levenshtein distance. Levenshtein distance (represented as $Dist$ in Equation (3)) or edit distance is a string metric for measuring the

[1] In the supervised track, we only participated English subtask, in which relatively more training data was provided. For this reason, our analysis of supervised learning is specific to English.

Statistic Features	Spearman *r*
Char Distance Ratio	0.513
Word Overlap Ratio	0.593
Content Words Overlap Ratio	0.604

Table 1: Correlation between human-annotated relatedness scores with ratios of textual statistical features.

difference or distance between two sequences at the character level. The character ratio we use in this study is defined as:

$$\frac{\text{len}(\text{Sent}_1) + \text{len}(\text{Sent}_2) - \text{Dist}}{\text{len}(\text{Sent}_1) + \text{len}(\text{Sent}_2)} \quad (3)$$

- Word Overlap Ratio: the count of overlapped words over the total word count in sentence pairs, expressed as:

$$\text{Ratio} = \frac{|\text{Words(A)} \cap \text{Words(B)}|}{|\text{Words(A)} \cup \text{Words(B)}|} \quad (4)$$

- Content Word Overlap Ratio: the overlap ratio with content word considered only. Content words and functional words are distinguished by analyzing their part-of-speech (POS) using `SpaCy` python package.

We found that the overlap ratio computed solely on content words shows a better correlation with the human judgment of relatedness (Table 1). Furthermore, we tested the correlation of the word overlap ratio with the other two scores: Spearman's *r* with content word overlap ratio is 0.77, and Spearman's *r* with character distance ratio is 0.78. This suggests that the combination of two or more results may improve the relatedness estimation.

3.1.2 Word Embedding Models

In this subsection, we evaluate the performance of word embedding models' potential to estimate semantic relatedness. Sentence embeddings are represented as the mean of the word embeddings of all words in the sentence. We explored static word embeddings (GloVe and first layer BERT embeddings) and contextual word embeddings (the last layer of BERT embeddings) in relatedness estimation. The performance of the following variations is presented in Table 2:

- PCA transformation of embeddings. By using the PCA technique, we do not intend to reduce the dimension of the sentence embeddings, but transform sentence embeddings onto a new coordinate system such that the principal components capture the largest variation in the data. In practice, the maximum dimension that fits the dataset is adopted: *min* (*embedding_length, sample_size*).

- Content word embeddings: the average of word embeddings of content words only.

- Noun embeddings: the average of word embeddings for nouns only.

- Tree-Based word embeddings: the mean of embeddings of words that are at the top three levels of dependency trees,[2] namely the root (main predicate), direct dependents of the root, and dependents with the dependency distance of 2 from the root.

Our preliminary analysis offers the following insights for further ensemble learning:

1. Excluding functional words (using content words only) can enhance the effectiveness of GloVe embedding.

2. Focusing on words closer to the sentence's 'root' in terms of dependency distance did not yield better results.

3. Contextualized BERT embeddings do not necessarily outperform uncontextualized embeddings in semantic relatedness estimation.

4. PCA-transformed embeddings show improved correlation with human annotation of relatedness.[3]

3.1.3 Models for Sentence Representations

For supervised learning, we also incorporate sentence representations from pre-trained language models into our ensemble system. This includes models known for their strong performance in sentence similarity tasks, involving Sentence-BERT (`mpnet-base`, Reimers and Gurevych, 2019) and Jina Embedding (`jina-v1`, Günther et al., 2023), as well as the general large language model, T5

[2]We use SpaCy to parse sentences and select the root and dependents

[3]Despite the better performance of PCA-transformed embeddings in Spearman's correlation when word embedding models are tested individually, it was not beneficial in later supervised training. Ultimately, $\text{GloVe}_{\text{Content}}$ word embedding was utilized in supervised and unsupervised ensemble learning for English.

Model	Spearman *r*
GloVe	0.460
$\text{GloVe}_{\text{PCA}}$	0.533
$\text{GloVe}_{\text{Content-words}}$	**0.554**
$\text{GloVe}_{\text{Tree-Based}}$	0.249
$\text{GloVe}_{\text{Noun}}$	0.430
$\text{BERT}_{\text{LastLayer}}$	0.399
$\text{BERT}_{\text{LastLayer/PCA}}$	0.446
$\text{BERT}_{\text{FirstLayer}}$	0.570
$\text{BERT}_{\text{FirstLayer/PCA}}$	**0.593**

Table 2: Spearman's correlation between human-annotated relatedness scores with the cosine similarity of average embeddings of all words, content words, all nouns or tree-based word selections within a sentence. PCA-transformed average embeddings of all words in a sentence are also presented.

encoder (Raffel and Chen, 2023; Ni et al., 2021). Among all models tested in this study for English (refer to Table 3), T5 demonstrates the highest performance, achieving a Spearman's correlation of approximately 0.82 with human annotation.

3.1.4 Ensemble Learning

We explored two approaches for ensemble learning. The first approach operated directly on sentence representations from multiple models. This included concatenating sentence embeddings from various models and applying transformation (e.g., PCA transformation) in the embedding space to achieve a better correlation with human judgment. Our analysis indicates that while concatenation and transformation operations can slightly improve Spearman's correlation, they are not as effective as incorporating more statistical features into supervised fine-tuning.

In the final system, we directly used the cosine similarity values from sentence embedding and word average embeddings as features (from models `mpnet-base`, `jina embedding`, `T5-base` and mean of content word embeddings from GloVe), along with textual statistic features (content word overlap ratio and character distance ratio) to estimate the relatedness of sentence pairs. These features are fed into Support Vector Machine (SVM) regression models (with RBF kernel) to predict human annotated relatedness.

3.2 Unsupervised Ensemble

In the unsupervised track, without utilizing labeled datasets for sentence similarity or relatedness and without employing models pre-trained on labeled datasets, we aim to evaluate whether models trained on other types of datasets intended for different purposes could generate representations suitable for estimating semantic relatedness.

In addition, we investigated whether integrating additional features, such as the cosine distance of average word embeddings and word overlap ratios, could enhance performance. We calculated the arithmetic mean of the cosine distances and ratios from textual statistics as the relatedness prediction of sentence pairs. Various feature combinations are tested with the provided validation dataset.

For the unsupervised task of English, we utilized two models to generate sentence representations: a model designed for semantic search (`multi-qa-MiniLM-L6-cos-v1`, Reimers and Gurevych, 2019), trained on 215 million question-answer pairs; and e5 (`e5-base-unsupervised`, Wang et al., 2022),[4] trained on question-answer pairs, post-comment pairs, and title-passage pairs. These models were further refined with an unsupervised transformation (PCA). Additionally, we incorporated two other features: PCA-transformed GloVe embeddings (average of content word embeddings within a sentence) and content word overlap ratios into the unsupervised ensemble system.

For the unsupervised tasks in Spanish and Hindi, we used a similar method for predicting relatedness, combining features involving the cosine distance of `multi-qa-MiniLM` model representations, word embedding model and word overlap ratios. For word embeddings, we employed multilingual BERT (`bert-base-multilingual-uncased`), utilizing both the first-layer (uncontextualized) and last-layer (contextualized) embeddings for relatedness estimation.

4 Results and Analysis

The shared task evaluates the participating systems based on Spearman's correlation (r) between the human-annotated scores, which ranges from 0 to 1. In Table 3, we compare the correlation scores for our systems and other popular models on the official test set.

[4]The e5 monolingual model is exclusively used for English, not for the other two languages: Spanish and Hindi

Models	English	Spanish	Hindi
Lexical Overlap	0.741	0.661	0.587
$\text{mBERT}_{\text{\_Ave}}$	0.640	0.655	0.566
mpnet-base[5]	0.809	0.590	**0.746**
T5 (base)	0.825	-	-
LaBSE[6]	0.818	0.651	0.709
multi-qa-Mini	0.793	0.638	0.466
$\text{Ensemble}_{\text{Sup}}$	**0.850**	-	-
$\text{Ensemble}_{\text{Unsup}}$	**0.837**	**0.705**	0.649

Table 3: Spearman correlation between human-annotated relatedness scores and system predicted scores on the test dataset.

Results presented in Table 3 suggest that the ensemble approach generally outperforms single models. Specifically, the ensemble system trained with true labels, for the supervised English task, achieved the best result among all listed systems, with an improvement in Spearman's correlation of 0.025 compared to the T5 base model.

The ensemble approach for English and Spanish unsupervised tasks also achieved relatively high scores, despite the absence of similarity or relatedness scores in learning. It suggests that semantic relatedness can be estimated without necessarily relying on human-annotated scores of semantic similarity or semantic relatedness. Other sources like question-answering pairs or statistical features of texts also play a role in relatedness estimation. Thus, the ensemble of statistical text features, word embedding models, and models trained on question-answer pairs can achieve good results.

Although the results for Hindi did not match the superior outcomes of other supervised models, such as `mpnet-base` and `LaBSE`, which were trained with semantic labels or similarity scores, the ensemble system's performance still surpasses that of the multilingual BERT embedding model and the `multi-qa` model, both of which were utilized for ensemble learning as base models.

4.1 Biased Performance

We also observe that the unsupervised results for Hindi are not comparable with those from Spanish and English though with the same ensemble approach. This discrepancy stems from the suboptimal performance of the sub-models used in the unsupervised ensemble. For example, the multi-qa-MiniLM model utilized for Hindi only achieves a correlation of 0.466, and the multilingual BERT for Hindi is also less effective compared to the other two languages.

Apart from Hindi, we also applied the same ensemble method to other non-Indo-European languages in the unsupervised track, yet the results scarcely surpassed 0.60 for the validation dataset, so results of other languages were ultimately not submitted.

The results indicate that some multilingual models are biased towards English and Indo-European languages, and perform less effectively for other languages. This bias may be attributed to imbalanced data during the models' pre-training phase.

5 Conclusion

Our system employs an ensemble approach to estimate semantic relatedness, integrating results from multiple systems: textual statistical features, word embedding models, and sentence representation models. Our findings suggest that semantic relatedness can be deduced from a variety of sources. Although some features (e.g., lexical overlap ratio) may not perform as strongly as models specifically designed to obtain sentence representations, the results demonstrate that these features, when used in a combined manner, can outperform many individual systems and collaboratively achieve a better correlation with human judgment on semantic relatedness.

6 Limitation and Future Work

Constrained by the size of the training data and the availability of pre-trained language models, it is regrettable that we did not offer insights into other Asian and African languages. In future research, studies on low-resource languages will be valuable, including tasks such as data collection, annotation, and pre-training models tailored to these languages.

Acknowledgements

We are very grateful for the assistance and discussions provided by Leander Girrbach and Milan Straka.

[5]Table 3 shows `all-mpnet-base-v2` result for English and `paraphrase-multilingual-mpnet-base-v2` model results for Spanish and Hindi, model details: https://www.sbert.net/docs/pretrained_models.html

[6]Feng et al., 2022

References

Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif Mohammad. 2023. What makes sentences semantically related? a textual relatedness dataset and empirical study. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 782–796, Dubrovnik, Croatia. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Eneko Agirre, Johan Bos, Mona Diab, Suresh Manandhar, Yuval Marton, and Deniz Yuret, editors. 2012. **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Association for Computational Linguistics, Montréal, Canada.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Michael Günther, Georgios Mastrapas, Bo Wang, Han Xiao, and Jonathan Geuter. 2023. Jina embeddings: A novel set of high-performance sentence embedding models. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 8–18, Singapore. Association for Computational Linguistics.

Xin Huang, Kye Min Tan, Richeng Duan, and Bowei Zou. 2023. Ensemble method via ranking model for conversational modeling with subjective knowledge. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 177–184, Prague, Czech Republic. 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. *arXiv preprint arXiv:1907.11692*.

Suresh Manandhar and Deniz Yuret, editors. 2013. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, Atlanta, Georgia, USA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*.

Anton Osika, Susanna Nilsson, Andrii Sydorchuk, Faruk Sahin, and Anders Huss. 2018. Second language acquisition modeling: An ensemble approach. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 217–222, New Orleans, Louisiana. Association for Computational Linguistics.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Abinew Ali Ayele, Pavan Baswani, et al. 2024a. Semrel2024: A collection of semantic textual relatedness datasets for 14 languages. *arXiv preprint arXiv:2402.08638*.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Meriem Beloucif, Christine De Kock, Oumaima Hourrane, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Krishnapriya Vishnubhotla, Seid Muhie Yimam, and Saif M. Mohammad. 2024b. Semeval-2024 task 1: Semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word

representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Matthew Raffel and Lizhong Chen. 2023. Implicit memory transformer for computationally efficient simultaneous speech translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12900–12907, Toronto, Canada. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*.