UAlberta at SemEval-2024 Task 1: A Potpourri of Methods for Quantifying Multilingual Semantic Textual Relatedness and Similarity

Ning Shi, Senyu Li, Guoqing Luo, Amirreza Mirzaei, Ali Rafiei, Jai Riley, Hadi Sheikhi
Mahvash Siavashpour, Mohammad Tavakoli, Bradley Hauer, Grzegorz Kondrak
Alberta Machine Intelligence Institute
Department of Computing Science
University of Alberta, Edmonton, Canada
{ning.shi,senyu,gluo,amirzaei,rafiei,jrbuhr,hsheikhi
siavashp,tavakol5,bmhauer,gkondrak}@ualberta.ca

Abstract

We describe our systems for SemEval-2024 Task 1: Semantic Textual Relatedness. We investigate the correlation between semantic relatedness and semantic similarity. Specifically, we test two hypotheses: (1) similarity is a special case of relatedness, and (2) semantic relatedness is preserved under translation. We experiment with a variety of approaches which are based on explicit semantics, downstream applications, contextual embeddings, large language models (LLMs), as well as ensembles of methods. We find empirical support for our theoretical insights. In addition, our best ensemble system yields highly competitive results in a number of diverse categories. Our code and data are available on GitHub.

1 Introduction

In this paper, we describe our submission for SemEval-2024 Task 1: Semantic Textual Relatedness (STR) (Ousidhoum et al., 2024b), which is based on the SemRel2024 dataset (Ousidhoum et al., 2024a). Each instance consists of a pair of sentences in the same language, annotated with a score that quantifies their semantic relatedness. SemRel2024 was annotated by native speakers of the dataset's 14 languages, which span five language families. An example English instance consists of the sentence pair *"the story is gripping and interesting"* and *"it's a brilliant, compelling, and heartfelt story"*, which is annotated with a relatedness score of 0.64. We participated in all three tracks (supervised, unsupervised, and cross-lingual) on all 14 languages.

Semantic relatedness is distinct from semantic similarity. Sentences that express opposite propositions, such as *"it is raining"* and *"it is not raining"*, exhibit low similarity but high relatedness. The impact of relations such as antonymy and meronymy (Budanitsky and Hirst, 2001) make semantic similarity a more specific task: similarity implies relatedness, but not vice versa. Nevertheless, many traditional algorithms make no attempt to distinguish between the two tasks (Jurafsky and Martin, 2009). For example, the word overlap baseline in this shared task could also be applied to measure semantic similarity. The extent to which semantic similarity and relatedness correlate in practice remains an important open question.

In this paper, we test the hypothesis that *similarity is a special case of relatedness* (Pedersen et al., 2007) through implementing an array of methods that are designed to measure similarity, and applying them to the task of measuring relatedness. We experiment with several different approaches: (1) methods that create and compare semantic representations of each input sentence; (2) methods that use the output of systems designed for other semantic tasks, such as paraphrase identification and entailment detection; (3) methods based on prompting large language models using in-context learning; and (4) methods that combine multiple individual methods. We further posit that *semantic relatedness is preserved under translation*. We investigate both hypotheses via supplementary experiments on datasets from the Semantic Textual Similarity (STS) task at SemEval 2017, as well as new cross-lingual datasets that we construct ourselves by translating parts of the SemRel2024 dataset.

Our experimental results provide support for our theoretical insights. The experiments on the supplementary datasets demonstrate a high correlation between the STR and STS tasks. Out of 51 competing teams, we rank among the top three entries in 16 of the language/track settings. In particular, our best-performing supervised ensemble system achieves the highest score in the English Track A among the teams that submitted a system description paper (Ousidhoum et al., 2024b). Taken together, these results support the idea of using similarity as a proxy for relatedness, as predicted by our hypotheses.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1798–1805
June 20-21, 2024 ©2024 Association for Computational Linguistics

2 Methods

We investigate ten different methods, divided into four types. Each method takes as input a pair of sentences, and produces a scalar value, which, possibly after some normalization to place it within the range specified for this shared task, is used as a measure of STR. Thus, each individual method is a complete, functional STR method; our principal innovation is the ensembling of these methods into a single STR system.

2.1 Explicit Semantic Methods

Concept overlap We hypothesize that the number of shared lexical concepts correlates with the relatedness between two sentences. On the basis of this hypothesis, we tag the words in each sentence with WordNet senses (Miller, 1995) using the offline AMuSE-WSD large model Docker image (Orlando et al., 2021). Each such sense corresponds to a unique lexical concept. The concept overlap score is calculated by dividing the number of shared concepts, with a WordNet synset path similarity greater than 0.8, by the total number of unique concepts in both sentences.

AMR similarity We approximate the relatedness of two sentences by measuring the similarity of their abstract meaning representations (AMRs). The AMR of a sentence is a structured labeled graph that represents its meaning (Banarescu et al., 2013). After converting each input sentence into an AMR using the SapienzaNLP API[1], the similarity between the two AMRs is computed as the Smatch F1-score (Cai and Knight, 2013), a metric devised explicitly for analyzing the overlap between graph-based representations. This score is then used as a measure of the relatedness of the sentences.

2.2 Extrinsic Methods

Paraphrase identification (PI) We reduce STR to paraphrase identification, a binary classification task to determine the approximate semantic alignment between two sentences (Bhagat and Hovy, 2013). By utilizing a dedicated PI model, we first compute the probability that one sentence is a paraphrase of the other. The intuition is that a higher probability of a positive classification indicates greater semantic relatedness. While paraphrasing is, in theory, a symmetric relation on sentences, in practice, the order in which sentences are provided to the model impacts its output. We compute the paraphrase identification probability for both orderings of the two sentences, and use their average as the score for STR.

Taking RoBERTa (Liu et al., 2019) as the backbone, we fine-tune a paraphrase classifier on a combined dataset, including six datasets: PIT (Xu et al., 2015), QQP (Iyer et al., 2017), MRPC (Dolan and Brockett, 2005), PAWS QQP (Zhang et al., 2019), PAWS Wiki (Zhang et al., 2019), and PARADE (He et al., 2020). We follow dataset splits and training configurations as established in prior research (He et al., 2020; Peng et al., 2022).

Textual entailment Similar to PI, we use textual entailment as an indicator of sentence relatedness. In particular, we aim to reduce STR to recognizing textual entailment (RTE) or natural language inference (NLI). Both tasks evaluate whether the meaning of one sentence (the hypothesis) can be inferred from another (the premise). RTE frames this as a binary task and NLI expands it into ternary classification with the addition of a neutral label (Dagan and Glickman, 2004). Recognizing that entailment in either direction signifies potential relatedness, we use an off-the-shelf RoBERTa NLI classifier (Nie et al., 2020) to estimate the probability of entailment in both directions. The final STR score is the average of these two probabilities.

2.3 Distributional Methods

Embeddings In this method, we use an LLM to produce dense semantic embeddings representing the meaning of each input sentence. We then compute the cosine similarity between their respective embeddings, and use this as a measure of relatedness. This simple embedding-based approach allows a language model to be used “as-is”, with no need for additional fine-tuning.

We experiment with two variants based on BERT (`Embed-B`) (Devlin et al., 2019) and RoBERTa (`Embed-R`), respectively. For each sentence, hidden states are obtained from the LLM, and an attention mask is applied to ensure the model focuses on meaningful tokens and excludes the other special ones such as the padding token. The resultant hidden states are aggregated into a single vector through average pooling.

2.4 Large Language Models

Prompting We utilize a few-shot prompting strategy to estimate STR between sentence pairs. We

[1] `nlp.uniroma1.it/spring/api/text-to-amr`

use in-context learning (Brown et al., 2020), providing first a small set of examples from the training data, consisting of two sentences and an STR value (i.e., the correct output from the data). For each pair of sentences, To facilitate few-shot prompting, we sample example sentence pairs from the training dataset and query ChatGPT through its API.[2]

Fusion This approach makes use of contextualized embeddings from a variety of open-source LLMs. For each sentence, we extract its sentence embeddings from each LLM, and concatenate them. The result is a "fusion" vector embedding of sentences whose dimensionality is the sum of the dimensionality of the embeddings produced by each LLM. We apply a trainable point-wise linear operation with bias to the fusion embeddings. We train this layer to minimize the distance between the cosine similarity of the fusion embeddings of each sentence pair in the training data and their gold-standard STR scores. In other words, we train this layer to produce the cosine similarity as the STR scores given pairs of fusion embeddings.

We integrate embeddings derived from a range of sentence transformer models (Reimers and Gurevych, 2019). While several of them are multilingual, our training process is exclusively focused on the English dataset. It aims to minimize the mean squared error (MSE) loss between the cosine similarity of the fusion embeddings for sentence pairs and their corresponding gold-standard STR scores. We adopt early stopping to mitigate the risk of overfitting.

Fine-tuning We add a linear regression head to a pre-trained language model, and fine-tune it for STR using the training data. The resulting regression model is therefore optimized for predicting the relatedness score given a pair of sentences. This provides another approach for leveraging the semantic capabilities of modern language models.

We investigate three distinct regression models, with one variant. Each regression model takes an LLM as the backbone with a randomly initialized regression head. We proceed to fine-tune the entire model, both the backbone and the regression head, The AdamW optimizer (Loshchilov and Hutter, 2019) is configured with an initial learning rate of 2e-5 and a batch size of 24. For the backbone, we experiment with T5 (`FT-T5`) (Raffel et al., 2020), GPT-2 (`FT-GPT2`) (Brown et al., 2020), and RoBERTa (`FT-R`). The variant `FT-MPNet` uses MPNet (Song et al., 2020), aligning more with the training process of SBERT. While most models are trained to minimize the MSE loss, MPNet uniquely targets minimizing the cosine similarity loss. This positions the MPNet one as a form of continued pre-training. We categorize this as a variant within our fine-tuning method for better presentation.

2.5 Ensemble Modules

To combine the advantages of the methods above, we assess two ensembling strategies: *unsupervised linear combination* and *supervised regression*.

Linear combination Our first approach is to compute the average of the STR scores produced by the individual methods. We first normalize the scores, based on the observation that some methods tend to produce higher or lower scores (i.e., scores with very different distributions). For instance, one method might typically produce scores between 0.7 and 0.9, while another might tend to produce scores in the range of 0.2 to 0.6. Our normalization is intended to give each method a similar distribution, with the lowest scores being normalized to 0 and the highest scores being normalized to 1. Once normalization is complete, for a given pair of sentences, the final ensemble STR score is obtained by computing the average score across all methods.

Our official submission for Track B is a linear ensemble system `Linear-2Ms` applied to synthesize the results of `Embed-B` and `Embed-R`. It operates entirely unsupervised, meaning it does not require exposure to any samples from the training set.

Regression One limitation of the linear combination is that it makes no distinction between methods; each method contributes equally to the average, regardless of how reliable it is in practice. Our second method combines the scores from the individual methods by treating each score as a feature in a linear regressor. Once trained, the method is applied by first computing the outputs of each method in the ensemble, and then applying the regression model to obtain the final score.

Our official submission for non-English languages in Track C, as well as English in Track A, is a regression ensemble system `XGB-4Ms` designed to synthesize the outputs from fine-tuning T5, GPT2, RoBERTa, and MPNet. At the heart of this ensemble system, we deploy an XGBoost regressor (Chen and Guestrin, 2016) as the central

[2] GPT-3.5-turbo-1106. Our experiments with LLaMA-2 (Touvron et al., 2023) were unsuccessful.

Method	afr	amh	arb	arq	ary	eng	esp	hau	hin	ind	kin	mar	tel	Avg.
Overlap	71.0	63.0	32.0	40.0	63.0	67.0	67.0	31.0	53.0	55.0	33.0	62.0	70.0	54.4
LaBSE-Cross	79.0	84.0	61.0	46.0	40.0	80.0	62.0	62.0	76.0	47.0	57.0	84.0	82.0	66.2
LaBSE-Sup	-	85.0	-	60.0	77.0	83.0	70.0	69.0	-	-	72.0	88.0	82.0	-
WordOvlap	73.2	64.3	31.4	40.2	57.7	73.9	63.5	38.7	57.1	53.2	31.5	68.9	64.3	55.2
EngWordOvlap	74.5	65.9	34.8	42.4	40.9	73.9	61.3	39.6	59.3	49.6	27.5	69.3	71.7	54.7
ConceptOvlap	69.7	60.4	42.6	38.2	40.7	68.8	61.1	34.9	59.0	38.5	30.7	64.7	70.3	52.3
AMR	70.1	62.7	30.7	35.9	33.7	71.4	60.6	36.0	59.5	45.7	33.2	67.8	66.4	51.8
PI	48.2	66.7	32.6	30.6	32.0	73.6	42.6	49.4	63.3	32.4	46.1	71.0	74.3	51.0
NLI	25.9	33.2	21.5	1.5	8.5	64.5	24.4	39.6	57.5	18.3	41.6	64.4	67.3	36.0
Embed-B	77.9	71.3	44.1	36.5	6.0	77.4	67.7	41.9	69.1	46.1	36.4	77.9	71.5	55.7
Embed-R	79.3	71.5	45.9	35.0	11.4	75.2	67.7	32.0	67.0	49.3	38.9	75.5	65.7	55.0
Prompt	80.4	78.2	64.0	40.2	38.1	82.0	62.2	57.6	80.3	47.1	53.0	85.5	82.9	65.5
Fusion	82.5	80.6	70.2	42.9	30.8	84.6	64.1	64.7	80.0	48.8	56.0	84.7	84.1	67.2
FT-T5	78.8	80.5	58.9	35.0	56.1	82.3	53.8	63.3	78.8	39.9	62.2	82.0	84.8	65.9
FT-GPT2	79.4	78.4	54.8	44.4	51.1	82.9	58.7	60.4	75.5	46.0	57.8	80.7	82.3	65.6
FT-R	81.1	80.8	65.7	43.2	54.4	83.6	55.8	67.7	82.3	39.4	64.1	86.6	84.1	68.4
FT-MPNet	81.7	80.3	67.9	44.7	25.3	84.9	64.0	61.4	80.2	52.2	53.5	84.7	82.4	66.4
Linear-2Ms	78.9	72.3	46.7	36.8	8.1	77.5	68.0	38.0	69.1	48.4	37.8	78.0	69.3	56.1
XGB-3Ms	80.6	81.5	67.3	45.1	60.4	84.6	57.0	67.4	82.2	45.6	63.7	85.5	85.2	69.7
XGB-4Ms	81.8	82.1	70.2	47.0	48.0	85.6	60.4	67.4	82.8	49.4	62.1	86.5	86.1	70.0
Target-XGB	-	85.4	-	57.5	80.6	85.6	70.5	73.5	-	-	77.4	89.0	85.7	-

Table 1: The results on the test sets of SemRel2024 in terms of the Spearman correlation (%).

model, tasked with integrating the results of the four individual systems as input features to predict the STR score. The XGBoost regressor is configured with a squared error regression objective and tailored configurations to optimize performance: a column sample by tree of 0.1, a learning rate of 0.1, a maximum depth of 8, an alpha value of 0.1, and 128 estimators. Training will stop if there is no improvement during validation for 32 consecutive rounds. This set of hyper-parameters is optimized on the English development set and remains constant throughout experiments.

Our official submission for non-English languages in Track A is Target-XGB, a tailored variant of the XGB-4Ms system, which is specifically engineered to navigate the linguistic distribution shifts inherent across languages. Underpinning this approach is the assumption that STR between sentence pairs remains consistent across languages. To this end, we fine-tune each individual system and the XGB regressor on English translations within the target language. Recognizing the potential introduction of noise from imperfect machine translation systems, we implement a data augmentation technique. To be specific, we merge the English training and development sets with the translated training set of the target language. The translated development set of the target language is kept as it is out of the training. By that, we intend to treat the English dataset as a stabilizing anchor, mitigating translation noise and ensuring that our system still remains sensitive to the target language.

3 Semantic Textual Relatedness

Our principal evaluation is on Tracks A, B, and C of the shared task datasets. The evaluation results are reported in Table 1, excluding Punjabi (pan), where most results are negative without any observable pattern. Our results may differ from those submitted due to adjustments in methods. We report the Spearman correlation (%) between the prediction and the golden STR scores.

We employ several baseline methods. Overlap, LaBSE-Cross and LaBSE-Sup are the official baselines reported by Ousidhoum et al. (2024a); the key distinction between the last two lies in whether the backbone LaBSE (Feng et al., 2022) is fine-tuned or not. WordOvlap is our re-implementation of Overlap; EngWordOvlap is its variant which requires translating sentences into English first. Most of our systems are English-specific; we translate sentences in other languages into English via the Google Translate API.

Explicit methods ConceptOvlap performs comparably to WordOvlap, indicating a similar level of efficacy in capturing semantic relatedness. However, AMR lags behind, suggesting that representing sentences as semantic graphs may introduce information that does not contribute to determining semantic relatedness, or that the quality of the AMR representations is insufficient.

Extrinsic methods Reducing the STR task to either PI or NLI yields markedly distinct outcomes. While PI approaches the performance level of the WordOvlap baseline, NLI generally underperforms. This discrepancy may be attributed to the inherent unidirectional nature of entailment. Our implementation takes the average of two entailment probabilities, and thus imposes strict constraints on the relatedness of sentences.

Distributional methods We can see that both Embed-B and Embed-R secures commendable results, matching the overall performance of the WordOvlap baseline. The observed superiority of BERT over RoBERTa could stem from differences in their score distributions. The predictions of Embed-R tend to be more clustered (e.g., ranging from 0.84 to 0.99 on eng). We found that simply rounding its results to two decimal places could reduce its performance from 75.2 to 72.4 on eng. In contrast, the predicted score distribution of Embed-B is relatively more dispersed.

LLM methods Prompt is competitive with LaBSE-Cross, but well below other LLM methods, such as Fusion and FT. This shows that, despite its strong performance on many other tasks, ChatGPT's STR capabilities are still limited. Furthermore, training on the provided dataset is observed to be pivotal in enhancing performance. Overall, our findings underline that there remains considerable scope for exploring and enhancing the application of LLMs in this field.

Ensemble modules Target-XGB obtains the best results across most languages. It surpasses LaBSE-Sup and consistently exceeds XGB-4Ms across all evaluated languages by a significant margin. These results show the importance of additional fine-tuning using the translations of the target language. Furthermore, incorporating the English dataset alongside the translated dataset proves to be advantageous. Notably, our ensemble systems, using either linear or regression modules, demonstrate superior performance over the individual systems they comprise, supporting the efficacy of our proposed ensemble approach.

4 Cross-Lingual Textual Relatedness

In this section, we discuss our experiments on new cross-lingual datasets which we created from the shared task data. The purpose of these experiments is to test our hypothesis that *semantic relatedness is preserved under translation.* Our bilingual dataset contains pairs of sentences from English and Spanish, respectively. The Spanish sentences are obtained by alternately translating one of the two English sentences from each instance of the SemRel2024 development set. The task is to determine the cross-lingual STR score, which is assumed to be the same as that for the original monolingual English sentence pair.

Method	eng	esp*	eng-esp	eng-esp*
WordOvlap	62.7	57.8	33.1	62.5
ConceptOvlap	63.4	62.1	51.5	64.1
AMR	66.1	61.4	-	64.2
PI	71.6	40.3	-	71.5
NLI	62.0	38.2	-	61.6
Embed-B	72.4	71.0	-	70.9
Embed-R	72.1	**72.0**	-	67.0
Prompt	79.0	67.5	77.1	78.7
Fusion	82.5	68.2	**80.4**	81.7
XGB-4Ms	**85.6**	68.4	-	**84.9**

Table 2: Results of primary methods evaluated using Spearman correlation (%) in our cross-lingual setting. Translating inputs into English is denoted by "*".

Table 2 shows the experimental results on our cross-lingual STR dataset. The eng-esp column shows the results of those methods that can be directly applied to languages other than English. The eng-esp* column shows the results of a larger subset of methods obtained after translating the Spanish sentence in each instance back into English. For reference, we also include the results on the official English (eng) and Spanish (esp*) development sets, of which the latter is translated into English.

The WordOvlap baseline performs poorly when applied to the eng-esp dataset because orthographic forms rarely match across languages even if they have the same meaning. In contrast, ConceptOvlap performs much better, as it is entirely multi-lingual and independent of orthography and script. However, both methods obtain similar results on eng-esp*, where the Spanish text is translated into English. Our AMR, NLI, and XGB-4Ms systems cannot be applied to cross-lingual pairs, but when Spanish is translated into English, their performance on eng-esp* is comparable to what is observed on the English test set.

The most interesting findings emerge from the results of Prompt and Fusion. Both are applicable directly to cross-lingual data, without translating the Spanish sentences into English. Surprisingly, for both methods, we observe only small differ-

Method	eng-eng	eng-esp	eng-esp*
ECNU	85.2	81.3	-
WordOvlap	72.8	13.5	64.4
ConceptOvlap	74.8	50.3	69.0
AMR	71.5	-	59.2
PI	76.9	-	72.2
NLI	68.4	-	69.7
Embed-B	73.5	-	63.4
Embed-R	71.5	-	59.2
Prompt	89.2	**87.9**	**88.4**
Fusion	90.3	84.4	87.8
XGB-4Ms	**91.0**	-	87.6

Table 3: Evaluation results of our primary methods using Spearman correlation (%) for the STS task. Translating inputs into English is denoted by "*". ECNU (Tian et al., 2017) ranked first in the SemEval 2017 Task 1.

ences between the relatively high numbers in the three columns. This finding supports our hypothesis that translation does not affect the degree of relatedness between a pair of sentences.

5 Semantic Textual Similarity

Another hypothesis that we investigate is that *similarity is a special case of relatedness*. Therefore, we expect a strong correlation between the two concepts: sentences that are highly similar should also be considered highly related. In this section, we test this hypothesis by applying our methods to STS datasets (i.e., track 5 and 4a) from SemEval 2017 Task 1 (Cer et al., 2017).

Since both STR and STS tasks output numerical scores on pairs of sentences, our methods can be directly applied to STS without modification. For supervised methods, we apply the models to STS in the same way as to STR, without any additional training or fine-tuning. This approach can be viewed as transfer learning: models trained on STR datasets are tested on the STS task. For cross-lingual datasets, we again experiment with both direct application to different languages (eng-esp) and pre-translation into English (eng-esp*).

Table 3 shows the results of our STS experiments. While these STS results are not directly comparable to any STR results, we observe that the best three methods are the same for both monolingual and bilingual STS and STR. As shown in Figure 1, the progressive improvement from left to right across methods indicates a strong correlation between STR and STS. Furthermore, the overall alignment between the blue and green lines, as well as between the red and yellow lines, support our hypothesis that both STR and STS are preserved under translation.

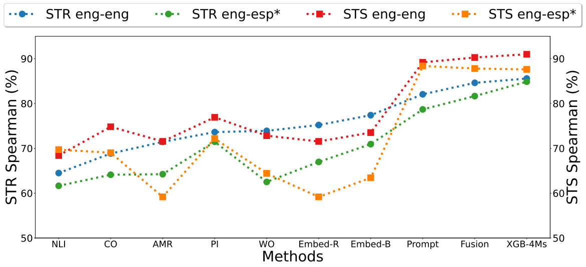

Figure 1: Summary of evaluation results for our primary methods in both STR and STS, Methods are ordered by their performance on STR eng-eng.

A more detailed analysis of the individual methods reveals several additional insights. Among the explicit methods, WordOvlap and AMR are both outperformed by ConceptOvlap on STS, which is likely due to their lack of robustness in cross-lingual settings. Among the extrinsic methods, PI works better than NLI, exhibiting remarkable stability across tasks, likely due to the training of our PI model on diverse benchmark datasets, including adversarial examples from PAWS QQP and PAWS Wiki. Among the distributional methods, while Embed-B consistently outperforms Embed-R, both experience a decline in cross-lingual performance, revealing sensitivity of sentence embeddings to translation noise. Among LLMs-based methods, Prompt excels on cross-lingual STS benchmarks, possibly because of data leakage in training ChatGPT, as well as its multilingual design. Our ensemble system XGB-4Ms generally delivers the best results.

6 Conclusion

We have investigated a wide array of methods on two of sentence-level semantic tasks in both mono-lingual and cross-lingual settings. In the process, we assembled a comprehensive benchmark of datasets for future explorations in this domain. The experiments furnish evidence for two hypotheses: (1) semantic similarity is a special case of semantic relatedness, and (2) both similarity and relatedness are preserved under translation. In practical terms, the evaluation results indicate that ensembling LLMs with diverse architectural designs yields the most robust and effective performance across languages and tasks. Notably, our strongest system is at the top of the ranked teams in English Track A, the setting with the highest number of participants.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Alberta Machine Intelligence Institute (Amii).

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Alexander Budanitsky and Graeme Hirst. 2001. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and other lexical resources*, volume 2, pages 2–2.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

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.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Ido Dagan and Oren Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. *Learning Methods for Text Understanding and Mining*, 2004(26-29):2–5.

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.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

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.

Yun He, Zhuoer Wang, Yin Zhang, Ruihong Huang, and James Caverlee. 2020. PARADE: A New Dataset for Paraphrase Identification Requiring Computer Science Domain Knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7572–7582, Online. Association for Computational Linguistics.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. Quora question pairs. *First Quora Dataset Release: Question Pairs*.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing*, 2nd edition. Prentice Hall.

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*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

George A Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Riccardo Orlando, Simone Conia, Fabrizio Brignone, Francesco Cecconi, and Roberto Navigli. 2021. AMuSE-WSD: An all-in-one multilingual system for easy Word Sense Disambiguation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 298–307, Online and Punta Cana, Dominican Republic. 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, Meriem Beloucif, Chris Biemann, Sofia Bourhim, Christine De Kock, Genet Shanko Dekebo, Oumaima Hourrane, Gopichand Kanumolu, Lokesh Madasu, Samuel Rutunda, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Hailegnaw Getaneh Tilaye, Krishnapriya Vishnubhotla, Genta Winata, Seid Muhie Yimam, and Saif M. Mohammad. 2024a. Semrel2024: A collection of semantic textual relatedness datasets for 14 languages.

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 task 1: Semantic textual relatedness for african and asian languages.

Ted Pedersen, Serguei V.S. Pakhomov, Siddharth Patwardhan, and Christopher G. Chute. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299.

Qiwei Peng, David Weir, Julie Weeds, and Yekun Chai. 2022. Predicate-argument based bi-encoder for paraphrase identification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5579–5589, Dublin, Ireland. 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. *J. Mach. Learn. Res.*, 21(1).

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*. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Junfeng Tian, Zhiheng Zhou, Man Lan, and Yuanbin Wu. 2017. ECNU at SemEval-2017 task 1: Leverage kernel-based traditional NLP features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 191–197, Vancouver, Canada. Association for Computational Linguistics.

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

Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015. SemEval-2015 task 1: Paraphrase and semantic similarity in Twitter (PIT). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 1–11, Denver, Colorado. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.