OZemi at SemEval-2024 Task 1: A Simplistic Approach to Textual Relatedness Evaluation Using Transformers and Machine Translation

Hidetsune Takahashi and **Xingru Lu**
Sean Ishijima and **Deokgyu Seo** and **Yongju Kim** and **Sehoon Park** and **Min Song** and **Kathylene Marante** and **Keitaro-Luke Iso** and **Hirotaka Tokura** and **Emily Ohman**
Waseda University
ohman@waseda.jp

Abstract

In this system paper for SemEval-2024 Task 1 subtask A, we present our approach to evaluating the semantic relatedness of sentence pairs in nine languages. We use a mix of statistical methods combined with fine-tuned BERT transformer models for English and use the same model and machine-translated data for the other languages. This simplistic approach shows consistently reliable scores and achieves middle-of-the-pack ranks in most languages.

1 Introduction

SemEval 2024 Task 1 (Ousidhoum et al., 2024c) calls for assigning scores indicating semantic textual relatedness (STR) of sentence pairs in 14 different languages. We participate in Track A, which is the supervised subtask for systems that have been trained using the provided labeled datasets (Ousidhoum et al., 2024a). There are data in Algerian Arabic, Amharic, English, Hausa, Kinyarwanda, Marathi, Moroccan Arabic, Spanish, and Telugu for Track A and we provide a solution for all 9 languages. The labeled data has been manually annotated for relatedness using a comprehensive annotation framework (Abdalla et al., 2023).

A large portion of previous work in STR has been conducted for English-language data. This task does include English, but the focus is on lower-resourced languages (Hedderich et al., 2021; Marreddy et al., 2022). STR is a crucial component in information retrieval, summarization, and question answering, as well as in developing Large Language Models (LLMs). The lack of STR or similar NLP resources for low-resource languages means progress is often much slower in related research such as the development of LLMs too making the progress achieved through this task societally highly impactful by providing new tools and datasets for language where NLP resources are lacking (Vulić et al., 2020; Zhang et al., 2020).

Our methodology uses both traditional TF-IDF vectorization and transformer models like BERT (Devlin et al., 2018) fine-tuned for semantic relatedness tasks. We leverage the high availability of resources that exist for English to fine-tune a BERT model that we then use on machine-translated versions of the datasets for the other languages (except for Spanish where a multilingual BERT model yielded better results than with machine translating the data). This approach seems to capture both lexical patterns and deeper semantic relationships, making it effective for linguistically diverse datasets, and cost-effective because there is no need to manually annotate more than one dataset (language). It is therefore an alternative approach to creating language-specific models. Although our approach is simplistic, it has the upside of working reasonably well for any low-resource language that has some machine translation or parallel language data resources.

2 Background

In SemEval-2024 Task 1, the dataset was adapted from the STR-2022 dataset (Abdalla et al., 2023). The STR-2022 dataset contains 5,500 English sentence pairs that were manually annotated using a comparative annotation framework, yielding fine-grained scores ranging from 0 to 1 (maximally unrelated to maximally related). The dataset was constructed by sampling sentences from various sources to capture a wide range of text characteristics such as sentence structure, formality, and grammaticality. The sources include datasets on formality (Rao and Tetreault, 2018), book reviews (Wan and McAuley, 2018), paraphrases (Wieting and Gimpel, 2018), natural language inference (Bowman et al., 2015), semantic textual similarity (Cer et al., 2017), stance (Mohammad et al., 2016), and text simplification (Horn et al., 2014).

The corresponding datasets for the other languages are much smaller and consist of roughly

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 7–12
June 20-21, 2024 ©2024 Association for Computational Linguistics

1000 sentence pairs each with minor variations in size.

Semantic **relatedness** and semantic **similarity** are closely related concepts in natural language processing (NLP), however, the terms are not interchangeable. Semantic similarity is a narrower definition that only takes term similarity into account (e.g. *fork* is similar to *knife*), whereas relatedness in addition to similarity can include terms or concepts that are related beyond hyponymic relationships such as *fork* being related to *eating*) (Asaadi et al., 2019; Batet and Sánchez, 2016). This task focuses on the broader concept of relatedness but utilizes more narrowly defined datasets based on similarity as well in the construction of the datasets.

In recent years the development of NLP resources for low-resource languages has been speeding up, but there are still large discrepancies in what types of tools, models, and resources exist for languages other than English (Hedderich et al., 2021). There are also significant differences in the resources available among low-resource languages and what being a low-resource language entails (Hämäläinen, 2021; Marreddy et al., 2022). For most of the languages in this task, there are at least some models and tools (see e.g Deode et al., 2023) but a handful of research groups working on a language is quite different from nearly all research groups in the world working on producing models and tools for a language (English). When there is a need for more data, often data augmentation methods are used to increase data points. Machine translation is an established method of data augmentation, particularly with low-resource languages where it might not be possible to use language-specific models (Amjad et al., 2020).

3 System overview

Our choice of methodology was shaped by pedagogical considerations as well as technical. As we participated in this task as part of an undergraduate senior research seminar in computational methods, we purposely started with the simplest most readily available tools progressing towards more advanced methods. Along the way, we compared the results and progress at each step in an attempt to better understand how each of the specific NLP tools worked and how accurate their output was when used on real projects such as this dataset.

The main strategy of our system is integrating classic NLP methods, such as the Dice Score and TF-IDF, with advanced deep learning techniques like BERT models, to determine semantic relatedness between sentence pairs. Firstly, our system imports a CSV dataset that contains pairs of English sentences (separated by "\n"), each paired with a relatedness score ranging from 0 to 1. Then, to assess semantic relatedness, the system adopts several basic NLP techniques, including Spacy's Linguistic Features for efficient text processing, TF-IDF for calculating word importance in sentences, Spacy Similarity and Cosine Similarity for measuring sentence similarity, and fine-tuned BERT Models for leveraging contextually rich semantic analysis (Devlin et al., 2018). These techniques collectively contribute to a robust evaluation of semantic relatedness against the given scores. We tried early on to adopt the same approach to the non-English languages with language-specific transformer-based similarity and relatedness models, but the language-specific models yielded much lower evaluation scores than what the English model achieved with machine-translated versions of the non-English datasets. We used the Google Translate API to translate the datasets into English to maintain consistency in analysis. Compared to other translation APIs such as DeepL, for this task, Google Translate seemed to produce better translations, perhaps because of how it favors more common words over context thus being more suited for STR and/or STS tasks (see e.g. Öhman, 2022).

Participating in the semantic relatedness task using the hybrid strategy allows for a comprehensive exploration of the system's performance and methodology. Through a detailed analysis, you can assess the effectiveness of traditional NLP methods, including TF-IDF and Spacy's Linguistic Features, in comparison to more advanced deep learning techniques like BERT. Evaluating the impact of contextual embeddings from fine-tuned BERT models provides insights into how well the model captures nuanced semantic relationships. The inclusion of Google Translate for non-English languages offers an opportunity to examine the system's ability to maintain consistency across languages. Assessing the generalization capability, scalability, and efficiency of the system provides a holistic understanding of its applicability to diverse datasets and real-world scenarios. Through this participation, we can uncover strengths, weaknesses, and potential areas for improvement, guiding future research directions and refining the hybrid strategy for enhanced semantic relatedness evaluation across languages and varied linguistic contexts. In particular, this approach shows that it is possible to achieve reasonable accuracies by leveraging the prevalence of tools and models designed for English with low-resource languages.

Our code is available on GitHub [1].

4 Experimental setup

At the beginning stage of the experiment, we undertook an examination of several readily implementable models on the English baseline dataset and compared the predicted scores with human-labeled scores through Pearson correlation scores.

In the initial English baseline model, we included the SpaCy similarity model[4], cosine vector similarity, and fine-tuned-BERT models[5]. For the SpaCy similarity, we directly applied it to the training dataset, yielding a result of 0.34 (Pearson). In the case of cosine similarity, we tried out two methods of word embedding:

1. **Binary occurrence vectors**: This approach involves creating set-based word vectors using binary occurrence, combining them into a joint space, and comparing them using cosine similarity to quantify the relatedness between the original sets in vectorized forms.

2. **TF-IDF transformer-based vectors**: Using the TF-IDF vectorizer from the sklearn (Pedregosa et al., 2011) library, we obtained TF-IDF weights for each word. The TF-IDF weight is proportional to the word's frequency in the document but is offset by its frequency in the corpus.

Upon comparing these two word-embedding methods, the Pearson correlation results did not reveal a significant difference. Therefore, we selected the Binary occurrence method as the cosine vector similarity, which achieved a score of 0.61 as indicated in Table 1. We use Pearson as opposed to Spearman rank correlation simply because that is what the original task description uses (Ousidhoum et al., 2024a,b).

The final component of the English baseline is the application of the fine-tuned BERT model to compute semantic relatedness with the (unfine-tuned) ClinicalBertSimilarity[5] and WebBertSimilarity[5] models and a batch size of 10 for both. The creators of the model claim that the "project contains an interface to fine-tuned, BERT-based semantic text similarity models. It modifies pytorch-transformers by abstracting away all the research benchmarking code for ease of real-world applicability"[5]. This proved to be the most successful approach with a result of 0.8 for English. Although the task in question is about semantic relatedness, since many of the datasets involved in the creation of the datasets come from similarity data. Additionally, as similarity can be considered a subtype of relatedness, the use of similarity models seemed logical due to their wider availability compared to relatedness models.

After establishing the English baseline, we evaluated several multilingual and language-specific BERT-based similarity models to assess textual relatedness (or similarity) across other language training datasets including the SBERT model for Telugu (Joshi et al., 2022), Sentence-BERT (Reimers and Gurevych, 2019), BioLORD-2023(Remy et al., 2023), etc. However, the results were suboptimal, which is surprising since previous work has shown that sentence transformers show significant improvements to semantic similarity tasks, particularly cross-lingual tasks (Hämmerl et al., 2023). Given the significantly better performance of the English baseline, we decided to translate all language datasets into English before applying the relatedness prediction models. In the case of Spanish we found that using distiluse-base-multilingual-cased-v1(Reimers and Gurevych, 2019) produced higher accuracies than the translation approach, and thus Spanish is the only language we did not translate to English.

When introducing the translation tools, we explored two approaches: utilizing a translation model (Machine Translation) and implementing Google Translate.

1. **Machine Translation**: In the Machine Translation method, we applied M2M100 (Fan et al., 2020) as the translation model. The model can directly translate between the 9,900 directions of 100 languages.

2. **Google Translate**: For the machine translations, we utilized the deep-translator library[6],

[1] https://github.com/esohman/SemEval2024

[3] https://huggingface.co/sentence-transformers/all-mpnet-base-v2 and Reimers and Gurevych (2020)

[4] https://spacy.io/usage/linguistic-features#vectors-similarity

[5] https://github.com/AndriyMulyar/semantic-text-similarity

LANGUAGE	Train Data: English Translation: Spacy Similarity	cos vector	fine-tuned SBERT	Multilingual Model: DBMCv1[2]	all-mpnet-base-v2[3]	Dev Data: Official score	Ranking
Algerian Arabic	0.25	0.44	0.51	0.42	0.39	0.37	18/20
Amharic	0.37	0.61	0.78	0.16	0.12	0.78	11/16
English	*0.34*	*0.61*	*0.80*	*	*	*0.81*	*10/34*
Hausa	0.07	0.43	0.65	0.21	0.34	0.62	12/19
Kinyarwanda	0.18	0.39	0.57	0.3	0.38	0.57	8/14
Marathi	0.45	0.68	0.81	*	*	0.86	13/25
Moroccan Arabic	-0.01	0.45	0.34	0.34	0.16	0.45	18/19
Spanish	0.58	0.7	0.66	*	*	0.62	8/17
Telugu	0.44	0.67	0.78	0.36	0.29	0.78	16/24

Table 1: Task scores for different methods

a versatile tool that facilitates simple language translation using multiple translators.

Despite the relatively high performance claimed by the M2M100 model as described by Fan et al. (2020), the results after the translation process are less than 0.5 for all languages except Spanish, where it achieved a result of 0.67. In contrast, the Google Translate API demonstrated better performance during the training process with the English baseline model (detailed results are listed in Table 1).

Our multilayered approach mirrors that of Jeyaraj and Kasthurirathna (2021) although ours is a much simpler setup.

5 Results

Our rankings show that our approach is nowhere near the state-of-the-art, but it is still a reliable option when more language-specific approaches are unavailable as is often the case with moderately low-resource languages. Out team ranked in the middle of the pack for most languages, but in the top third for English, Marathi, and Spanish, and the bottom for both Arabic dialects, which was expected. The rankings, scores, and models used for each submission can be seen in table 1. We analyze the results in the conclusions section.

6 Conclusions

To sum up, we first focused on English to have a good solution with fine-tuned BERT, and then we applied that solution to other languages by translating the sentences into English using machine translation. Since our English solution is reasonably good (rank 10/34, official score of .81), the application of the solution worked much better than using multilingual models in many languages including Amharic, Marathi, and Telugu for which there exist language-specific semantic similarity models. We speculate that the reason the *MT+English model* worked better than the language-specific relatedness models is due to the higher quality and more diverse training data for the English model(s) as well as machine translation simplifying words to the most commonly used ones, artificially making similar sentences more similar.

The importance of an accurate machine translation can be seen in the failure of our approach with the Arabic dialects in particular. Google Translate does not have specific translators for Moroccan or Algerian Arabic, instead, we had to rely on general Arabic. This likely produced much lower quality translations obfuscating the semantic links between the sentence pairs making it difficult for the English model to accurately judge relatedness. This issue was further exacerbated by the fact that no one on our team speaks any of the languages in the task besides English, which made manual evaluations of the MT output difficult.

Darja and Darija are the names for Algerian and Moroccan Arabic respectively, and they are collectively known as Maghrebi Arabic. Due to its roots in Berber languages, there are notable distinctions between Maghrebi Arabic and Standard Arabic, and using the latter for these two dialects may yield a suboptimal result.

Curiously, a similar issue occurred with Spanish. Spanish is much more closely related to English than the other languages in subtask A, and therefore we expected our approach to get a fairly high score similar to English, especially considering the current state of machine translation between English

[6]https://deep-translator.readthedocs.io/en/latest/README.html#id1

and Spanish. However, it seems that translation of Spanish into English affects the semantic relations of the original sentences, which might be one of the main reasons causing the very low scores and making us choose the multilingual model for Spanish rather than the machine-translated one.

We hypothesize that one of the reasons that Google Translate worked so well on the low-resource languages most dissimilar from English might be because smaller training datasets for MT would force the translation to use less context and instead increase the reliance on individual lexical items leading to sentence pairs with high relatedness becoming more similar via translation. For languages with better MT models, it is conceivable that the better translations work against this approach as it might make the sentence pairs less similar as reflected by the higher scores for Spanish using multilingual models, and the very low scores for both Arabic dialects. In future work, it might be worthwhile to use mixed methods starting with language-specific models and then expanding to incorporate machine translation and larger models developed for, e.g., English.

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.

Maaz Amjad, Grigori Sidorov, and Alisa Zhila. 2020. Data augmentation using machine translation for fake news detection in the urdu language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2537–2542.

Shima Asaadi, Saif Mohammad, and Svetlana Kiritchenko. 2019. Big bird: A large, fine-grained, bigram relatedness dataset for examining semantic composition. 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 505–516.

Montserrat Batet and David Sánchez. 2016. Improving semantic relatedness assessments: Ontologies meet textual corpora. *Procedia Computer Science*, 96:365–374. A paper we can use to show that we understand the difference between semantic relatedness and semantic similarity and looked into potential approaches/methods in order to improve the accuracy of our work.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi, and Raviraj Joshi. 2023. L3cube-indicsbert: A simple approach for learning cross-lingual sentence representations using multilingual bert. *arXiv preprint arXiv:2304.11434*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Mika Hämäläinen. 2021. Endangered languages are not low-resourced! *arXiv preprint arXiv:2103.09567*.

Katharina Hämmerl, Alina Fastowski, Jindřich Libovickỳ, and Alexander Fraser. 2023. Exploring anisotropy and outliers in multilingual language models for cross-lingual semantic sentence similarity. *arXiv preprint arXiv:2306.00458*.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. Learning a lexical simplifier using wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 458–463.

Manuela Nayantara Jeyaraj and Dharshana Kasthurirathna. 2021. Mnet-sim: A multi-layered semantic similarity network to evaluate sentence similarity. *arXiv preprint arXiv:2111.05412*.

Ananya Joshi, Aditi Kajale, Janhavi Gadre, Samruddhi Deode, and Raviraj Joshi. 2022. L3cube-mahasbert and hindsbert: Sentence bert models and benchmarking bert sentence representations for hindi and marathi. *arXiv preprint arXiv:2211.11187*.

Mounika Marreddy, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni, and Radhika Mamidi. 2022. Am i a resource-poor language? data sets, embeddings, models and analysis for four different nlp tasks in telugu language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(1):1–34.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952.

Emily Öhman. 2022. Self & feil: Emotion lexicons for finnish. In *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries conference. CEUR Workshop Proceedings*.

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, 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. 2024b. 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. 2024c. SemEval-2024 task 1: Semantic textual relatedness for african and asian languages. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

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

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

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.

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.

François Remy, Kris Demuynck, and Thomas Demeester. 2023. BioLORD: Semantic textual representations fusing llm and clinical knowledge graph insights.

Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. 2021. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. *arXiv preprint arXiv:2010.05731*.

Mengting Wan and Julian McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, page 86–94, New York, NY, USA. Association for Computing Machinery.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware bert for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9628–9635.