Pinealai at SemEval-2024 Task 1: Exploring Semantic Relatedness Prediction using Syntactic, TF-IDF, and Distance-Based Features.

Anvi Alex Eponon[1], **Luis Ramos**[1],
Ildar Batyrshin[1], **Grigori Sidorov**[1], **Olga Kolesnikova**[1], **Hiram Calvo**[1]
[1]Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC),
Mexico City, Mexico
{aeponon2023, lramos2020, ibatyr1, sidorov, kolesnikova, hcalvo}@cic.ipn.mx

Abstract

The central aim of this experiment is to establish a system proficient in predicting semantic relatedness between pairs of English texts. Additionally, the study seeks to delve into diverse features capable of enhancing the ability of models to identify semantic relatedness within given sentences. Several strategies have been used that combine TF-IDF, syntactic features, and similarity measures to train machine learning models to predict semantic relatedness between pairs of sentences. The results obtained were above the baseline with an approximate Spearman score of 0.84.

1 Introduction

The prediction of semantic relatedness between texts is a crucial task with applications in various natural language processing domains. In this study, our focus is on creating a system capable of predicting semantic relatedness across languages while investigating the features that contribute to this prediction. The development of such a system is not only beneficial for understanding semantic relationships within texts but also holds promise for enhancing deep learning models in tasks such as assessing sentence representation methods, question answering, and text summarization (Abdalla et al., 2023).

Despite the advancements in word representation techniques, especially using embeddings, the complexity of human languages presents persistent challenges in accurately capturing semantic relatedness. Our experiment primarily concentrates on the English language, acknowledging its significance as a widely used language in various applications. The inherent difficulty in identifying and quantifying the shared elements between two texts necessitates a thoughtful exploration of diverse features and methodologies.

In the subsequent sections, we describe the dataset used for this experiment, outline the methodology employed for predicting semantic relatedness, discuss the results obtained, and conclude with insights into the implications and potential future directions of this research. The research will delve into exploring features that enhance the prediction of semantic textual relatedness by developing several strategies concerning feature extractions and model training.

2 Literature Review

The complexity of machine-based human language modeling involves a nuanced understanding of various linguistic aspects, notably Pragmatics and Semantics (Abdalla et al., 2021; Miller, 1995). This research specifically emphasizes semantic modeling, with a focus on semantic relatedness, as opposed to the more commonly studied word similarity (Islam et al., 2012; Atoum and Otoom, 2016; Yum et al., 2021).

Traditionally, approaches like Bag of Words have been explored (Islam et al., 2012; Feng F. Jin, 2008), but they often fall short in achieving high performance for semantic relatedness tasks. WordNets models, while prioritized, face limitations in language coverage and comprehensive embedding of semantic relationships (Jordan J. Boyd-Graber, 2005).

A notable contribution by (Gomaa, 2019) introduced a model utilizing multiple similarity features, including cosine similarity and Jaccard. Their multi-layer architecture demonstrated that employing various similarity features collectively yields significantly better results than applying each measure in isolation. However, the approach did not add or consider syntactic features for the enhancement of the semantic prediction on textual data.

Recent advancements in deep learning models exhibit superior semantic similarity and relatedness performance. However, there remains a scarcity of research focusing on the distinctive features between Semantic Textual Similarity (STS) and Se-

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 935–939

mantic Textual Relatedness (STR), and how models can better capture the nuances of semantic relatedness between words and sentences (Kolb, 2005).

3 Task Description

The primary objective of this experiment is to create a system that can predict the semantic relatedness between pairs of texts across various languages but also explore the different features that could help models identify semantic relatedness between given sentences. Although the current experiment is focused on English, the development of such a system holds the potential to enhance deep learning models for various tasks, including assessing sentence representation methods, question answering, and text summarization (Abdalla et al., 2023).

4 Data Description

SemEval 2024 Track 1 utilized data provided by organizers, featuring sentence pairs in training, development, and test sets. Each instance is annotated with a score indicating semantic textual relatedness, which ranges from 0 (unrelated) to 1 (related). Table 1 presents statistics about the dataset, while Figure 1 illustrates the distribution of scores, including the counts for scores of 1.0, 0.0, >0.80 and <0.50, in the training and development set.

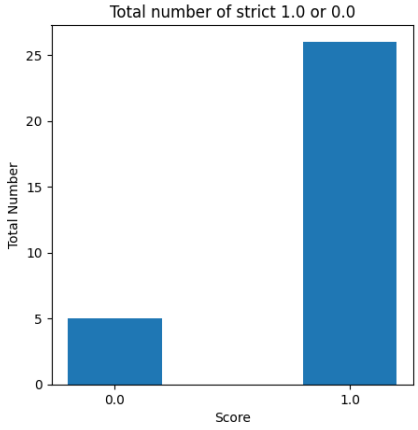
Table 1: Dataset Information

Dataset	Total Pairs	Pairs with Score 1.0	Pairs with Score 0.0
Training	5500	25	5
Development	250	7	123
Test	2600	-	-

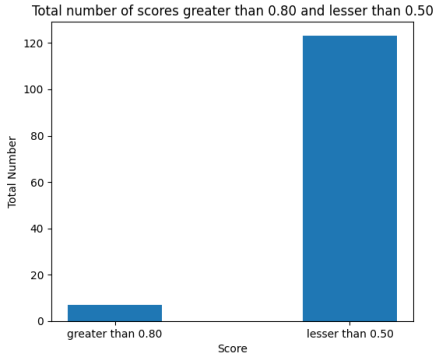
Table 2 displays sample instances, with additional details available in (Abdalla et al., 2021; Ousidhoum et al., 2024b,a).

Instance	Score
It that happens, just pull the plug. if that ever happens, just pull the plug.	1.0
The two little girls jump on the bed. A little girl is jumping down a sandy hill.	0.5
you're taking a sweater in a shop. to taking the life of Conor Greenleaf.	0.03

Table 2: Sample instances of data set



(a) Number of strict 1.0 and 0.0 in the Train set.



(b) Number of scores greater than 0.80 and lesser than 0.5 in the Dev set.

Figure 1: Train and dev scores comparison.

5 Methodology

In this study, our objective is to predict semantic textual relatedness between two texts. We made two key assumptions: firstly, we refrained from preprocessing the corpus to preserve sentence structure, essential for information retrieval and semantic identification (Hirst, 1987).

Secondly, we intentionally excluded Large Language Models (LLMs) from experiments, anticipating challenges in interpreting specific features contributing to semantic identification due to their contextual abilities and complexity (Turton et al., 2020). To extract diverse features, we employed various distance measures, including Jaccard distance, Cosine similarity, Levenshtein distance, and Word Mover's Distance (WMD) (Boubacar, 2014; Kusner et al., 2015; Su et al., 2008). These measures compute the similarity between text pairs based on common words, term vectors, and the minimum distance embedded words need to travel between documents. Feature extraction utilized the

bag-of-words (BoW) technique, specifically Term-Frequency Inverse Document Frequency (TF-IDF) (Hakim et al., 2014), and hidden vectors from a pre-trained SentenceBert model to compute cosine similarity (Reimers and Gurevych, 2019).

Additionally, syntactic features were extracted, parsing each sentence pair to identify words with the same dependency role. This approach resulted in three features: probability of exact word matching, probability of unique words, and probability of related words. The rationale was to explore the impact of words with common dependency roles on predicting semantic relatedness and assess the effect of their absence on the English corpus.

The reason behind using traditional models such as Gradient Boost is that this type of model can easily handle non-linear relationships in texts, which is crucial while capturing semantic relatedness but also can deal with imbalanced datasets (Natekin and Knoll, 2013).

A diagram of our methodology as well as the source code is freely available on GitHub. The link can be found in the Appendix section.

6 Results

The performance metrics of the different models submitted for semantic relatedness are presented. The evaluation metrics include solely the Spearman score. Even though the models submitted were Fasttext and Naive Bayes, during the training phase Linear Regression, XGradient Boost, Random Forest, and an ensemble of two traditional models were trained.

6.1 Train phase

During the training phase, several models were tested using different strategies, and in Table 3 we display the performance for each model. The main strategy that gave the results presented in this document has been explained earlier in the methodology section.

Model	Spearman
Linear Reg.	0.8512
Gradient Boost	**0.8527**
XGB	0.8467
Random Forest	0.8481
Ensemble Model	0.8521

Table 3: Training Model Performances

6.2 Development phase

At this stage of the experiment, the same models were tested on the development dataset which comprised of few number of samples, precisely 250 different instances. In Table 4 we display the performance over the development set.

Model	Spearman
Linear Reg.	0.8512
Gradient Boost	**0.8527**
XGB	0.8467
Random Forest	0.8481
Ensemble Model	0.8521

Table 4: Development Model Performances

6.3 Test phase

At the final stage of the experiments, the Gradient Boost model was chosen for the final tests on the testing dataset which comprised 5000 different instances. In Table 5, we display the results of the Gradient Boost models, compared to the baseline and the highest performance model in the same task.

Team	Spearman	Rank
PALI	0.8595	1
Pinealai	**0.8371**	10
SemRel-Baseline	0.8300	*

Table 5: Final Model Performance Metrics

7 Discussions

In the methodology section, our primary objective is to identify key features that enhance the capability of models in discerning semantic relatedness within textual data. We pursued two distinct approaches. First, we trained various models by employing Tfidf, Jaccard, or extracting specific syntactic features from the texts. This process does not take into account the internal structures of the texts which could give more insights about their meaning.

When exclusively utilizing syntactic features for model training, we achieved a maximum Spearman score of 0.32. Training models solely that used TF-IDF features to compute a cosine similarity and used the metric to predict yielded a separate score of 0.533. Incorporating features extracted from Sbert on top of the syntactic features resulted in a

notable score increase of 2, reaching approximately 0.70. Finally, combining all these strategies during the training phase produced a score of 0.85, with a corresponding score of 0.83 during the testing phase.

8 Conclusion

In conclusion, the study aimed to predict semantic relatedness in English, exploring diverse features and strategies. Limitations included the absence of word sense disambiguation algorithms, the exclusion of Transformer models for explainability, and the decision not to merge training and development sets. Despite these constraints, the models exhibited competitive performance, particularly the Gradient Boost model, which achieved a Spearman score of 0.8371. However, the experiment conducted can not help us derive a conclusion that the model can make a strict difference between semantic relatedness (STR) and semantic textual similarity (STS) in texts. The methodology highlighted the impact of syntactic features, TF-IDF representations, and SentenceBert embeddings. Moving forward, addressing limitations, incorporating advanced algorithms, and leveraging diverse datasets but also developing approaches that help models distinguish between STR and STS will contribute to a deeper understanding of semantic relations in textual data and further improvements in predictive capabilities.

Limitations

The study of semantic relatedness is a vast and tedious endeavor. The research conducted was very limited in many aspects. Firstly, the absence of algorithms specifically targeting direct word sense disambiguation represents a notable limitation. The incorporation of such algorithms could have potentially enhanced the models' effectiveness in this particular task. Also, the research did not explore the preprocessing techniques that could positively impact the semantic relatedness prediction.

Secondly, our study was confined to the training and testing of traditional machine learning models, excluding the exploration of Transformer models. While Transformers might have yielded superior results, their reduced explainability deterred their inclusion in our investigation.

Lastly, the decision not to merge the training and development sets for a final model training phase or add more datasets related to semantics relatedness or even semantic similarity represents another constraint. By solely transitioning to the testing phase with the models having learned solely from the training set given by the organizers, we may have missed opportunities for improvement. Combining both sets or augmenting the datasets through specific techniques in the final training phase could have potentially elevated the models' predictive capabilities, resulting in a more accurate score.

Ethics Statement

We affirm our commitment to ethical research practices and compliance with ACL guidelines in conducting and presenting our study. No ethical concerns or conflicts of interest arose during this research.

Acknowledgements

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico, and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

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.

Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif M Mohammad. 2021. What makes sentences semantically related: A textual relatedness dataset and empirical study. *arXiv preprint arXiv:2110.04845*.

Issa Atoum and Ahmed Fawzi Otoom. 2016. Efficient Hybrid Semantic Text Similarity using Wordnet and a Corpus. *International Journal of Advanced Computer Science and Applications*, 7(9).

Abdoulahi Boubacar. 2014. Valuing semantic relatedness. In *2014 IEEE 7th Joint International Information Technology and Artificial Intelligence Conference*, pages 1–5. IEEE.

Trevor T. Martin Feng F. Jin, Yiming Y. Zhou. 2008. Sentence similarity based on relevance. *Investigation Group of Applied Mathematics in computing*.

WAEL HASSAN Gomaa. 2019. A multi-layer system for semantic relatedness evaluation. *Journal of Theoretical and Applied Information Technology*, 97(23):3536–3544.

Ari Aulia Hakim, Alva Erwin, Kho I Eng, Maulahikmah Galinium, and Wahyu Muliady. 2014. Automated document classification for news article in bahasa indonesia based on term frequency inverse document frequency (tf-idf) approach. In *2014 6th international conference on information technology and electrical engineering (ICITEE)*, pages 1–4. IEEE.

Graeme Hirst. 1987. *Semantic interpretation and the Resolution of Ambiguity*. Cambridge University Press.

Aminul Islam, Evangelos Milios, and Vlado Kešelj. 2012. 29 *Text similarity using Google tri-grams*.

Daniel N. Osherson Robert E. Shapire Jordan J. Boyd-Graber, Christiane C. Fellbaum. 2005. Adding dense, weighted connections to wordnet. *3rd International Global WordNet Conference, Proceedings*.

Peter Kolb. 2005. Experiments on the difference between semantic similarity and relatedness. *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, pages 81–88.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

George A. Miller. 1995. WordNet. *Communications of The ACM*, 38(11):39–41.

Alexey Natekin and Alois Knoll. 2013. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21.

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-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.

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.

Zhan Su, Byung-Ryul Ahn, Ki-Yol Eom, Min-Koo Kang, Jin-Pyung Kim, and Moon-Kyun Kim. 2008. Plagiarism detection using the levenshtein distance and smith-waterman algorithm. In *2008 3rd International Conference on Innovative Computing Information and Control*, pages 569–569. IEEE.

Jacob Turton, David R. Vinson, and Robert E. Smith. 2020. Deriving Contextualised Semantic Features from BERT (and Other Transformer Model) Embeddings. *arXiv (Cornell University)*.

Yunjin Yum, Jeong Moon Lee, Moon Joung Jang, Yoojoong Kim, Jong-Ho Kim, Seong-Tae Kim, Unsub Shin, Sanghoun Song, and Hyung Joon Joo. 2021. A word pair dataset for semantic similarity and relatedness in Korean medical vocabulary: reference development and validation. *JMIR medical informatics*, 9(6):e29667.

Appendix

The diagram of our method and the source code can be found on GitHub at this URL: semEval2024 code