GIL-IIMAS UNAM at SemEval-2024 Task 1: SAND: An In Depth Analysis of Semantic Relatedness Using Regression and Similarity Characteristics.

F. López-Ponce[1], **Ángel Cadena**[1,2], **K. Salas-Jimenez**[1,2],
D. Preciado Márquez[1], **G. Bel-Enguix**[1,3]
[1]Grupo de Ingeniería Lingüística - UNAM
[2]Posgrado en Ciencias e Ingeniería de la Computación - UNAM
[3]Departament de Filologia Catalana i Lingüística General - Universitat de Barcelona
{francisco.lopez.ponce,karla_dsj}@ciencias.unam.mx
{angelcaden,davidpreciado115}@gmail.com
gbele@iingen.unam.mx

Abstract

The STR shared task aims at detecting the degree of semantic relatedness between sentence pairs in multiple languages. Semantic relatedness relies on elements such as topic similarity, point of view agreement, entailment, and even human intuition, making it a broader field than sentence similarity. The **GIL-IIMAS UNAM** team proposes a model based in the SAND characteristics composition (Sentence Transformers, AnglE Embeddings, N-grams, Sentence Length Difference coefficient) and classical regression algorithms. This model achieves a 0.83 Spearman Correlation score in the English test, and a 0.73 in the Spanish counterpart, finishing just above the SemEval baseline in English, and second place in Spanish.

1 Introduction

The Semantic Textual Relatedness (STR) task (Ousidhoum et al., 2024b) aims at creating systems that measure STR on pairs of sentences based on their closeness in meaning (Abdalla et al., 2023). This task is comprised of three tracks. Tracks A and B focus on monolingual models. Track A only accepts supervised models trained with the available tagged datasets, whereas track B focuses on the unsupervised approach relying on the same datasets but without the tagged score. Track C is the cross-lingual case where the target language has to follow an unsupervised approach. The datasets provided consist of sentence pairs that were sampled from various semantic similarity datasets.

This task expands upon classic sentence similarity comparisons, encouraging the use models and algorithms capable of analyzing more than meaning of a pair of sentences, focusing deeper on characteristics such as the syntactic structure of the sentences, lexicon relationships, as well as meaning and emotion. The GIL-IIMAS UNAM team participated in Track A. Although it included nine languages we have only worked with the Spanish and English dataset.

Track A is a regression problem since each dataset contains the sentence pairs as well as a corresponding sentence relatedness score that ranges from 0 to 1. The evaluation compares the ground values in the test set with the proposed model's prediction, meaning track A is evaluated using the Spearman Correlation.

This paper makes use of the SAND composition, a set of STR relevant characteristics, as well as regression algorithms trained with these characteristics in order to predict the STR score of other sentence pairs. In this paper the precise characteristics, algorithms, and parameters are presented as well as language based analysis. The final scores correspond to the best performing regression algorithm.

Our work also compares different metrics and their influence on the STR task compared to classic semantic similarity, as well as the model's varying behavior over the different languages used.

The paper is structured as follows: Section 2 explains the theories and models that are the background of our proposal. Section 3 explains the set of characteristics that have been chosen for our experiments. Section 4 explains the configuration of the data and the experimental methodology. In Section 5 we discuss the results obtained in the experiments, and compare them to the scores of other participants in the track. We close in Sec-

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1288–1292

tion 6 with some conclusions and ideas for further experiments.

2 Related Work

In the field of STR, various methodologies have been proposed. One such approach, outlined by Asaadi et al. (2019), involves analyzing the relatedness between word bi-grams. They describe the construction of a dataset tailored for this purpose. To compute the STR between bi-grams, or between bi-grams and unigrams, they utilize word embeddings represented as vectors generated by GloVe, fastText, and models employing matrix factorization of word-context co-occurrence matrices. They explore various methods for composing bi-gram vectors, such as addition, multiplication, tensor product with convolution, and dilation. The relatedness between two vectors is determined by computing the cosine similarity between them.

In a study by Abdalla et al. (2023), the concept of Semantic Textual Relatedness (STR) is extended to encompass the comparison between entire sentences. The authors delineate the construction of a specialized dataset tailored for this task and show the annotation process applied to this dataset. Their investigation delves into the influence of various linguistic factors, including lexical overlap, related words, related words belonging to the same part of speech, and the relatedness of subjects or objects, on the semantic relatedness between pairs of sentences. They represent each sentence as a vector and employ cosine similarity between these vectors as a metric for predicting semantic relatedness. To facilitate this analysis, they use static word embeddings such as Word2Vec, GloVe, and fastText, as well as contextual word embeddings like BERT and RoBERTa.

3 SAND Composition

This section describes the SAND (named based on the used characteristics: **S**entence Transformers, **A**nglE Embeddings, **N**-grams, Sentence Length **D**ifference coefficient) regression system used for the task. The STR datasets are comprised of senbtence pairs and a target score, in order to train the model with task relevant information certain similarity metrics were chosen in order to create a vector of characteristics that represent each sentence pair, such vectors were used as training data for the regression algorithms. An initial approach relied on similarity metrics such as Jaccard, and Dice coefficients as well as Jaro-Winkler and Levenshtein distance, nonetheless these metrics proved to be inefficient at training the model adequately, returning poor results when evaluated with partitioned training data. After revising the dataset and the nature of the sentences in question, the initial chosen characteristics were a coefficient analyzing the length of the sentences with and without stoptwords. Consider x, y two sentences, then both coefficients are obtained from the following:

$$\text{LenCoef}(x, y) = |\frac{\text{length}(x) - \text{length}(y)}{\text{length}(x) + \text{length}(y)}| \quad (1)$$

An observation on lexical overlapping in highly related sentences led to the choice of an n-gram based coefficient. For $n \in \{1, 2, 3\}$, the n-gram coefficient is defined as:

$$\text{n-gramCoef}(x, y) = \frac{\text{n-grams}(x) \cap \text{n-grams}(y)}{\text{n-grams}(x)} \quad (2)$$

Sentence pairs in the dataset often rely on contextual similarity apart from lexical overlap when it comes to measuring relatedness, meaning the use of a pretrained model is in order. The first approach relied on the use of Sentence Transformers (ST) (Reimers and Gurevych, 2019), a siamese neural net that uses pretrained encoders in order to generate contextualized sentence embeddings. Each pair of sentences was embedded using the ST architecture and the cosine similarity of the resulting vectors was obtained as the initial pretrained characteristic. Formally speaking:

$$\text{ST}_{sim}(x, y) = \cos(\theta) = \frac{x \cdot y}{\|x\| \, \|y\|} \quad (3)$$

Nonetheless standard encoder generated embeddings have shown to be improvable, for that the final characteristic relies on similarity based on AnglE-optimized embeddings (Li and Li, 2023). These type of embeddings optimize the cosine similarity saturation zones during training using complex space embeddings so that resulting vectors achieve a higher level of similarity. Nonetheless the final comparison between AnglE vectors is done in the same manner as equation 3.

Once these characteristics are extracted from each sentence pair they are passed to various regression algorithms for training, validation and testing. For this task four regression algorithms

were used. The reported scores correspond to the best model for each language. The precise details of the implementation are presented next.

4 Experimental Setup

4.1 Data and Evaluation Methodology

We use the official dataset (Ousidhoum et al., 2024a) in English and Spanish for Task 1, track A (supervised), which is structured as follows: PairID, Text and Score. PairID is an identifier of the pair, the Text column is the sentence pair separated by a line break, and the Score column is a float number between 0 (completely unrelated) and 1 (maximally related) which indicates the degree of semantic textual relatedness between the two sentences.

The English training corpus is composed of 5500 sentence pairs meanwhile the Spanish counterpart has 1561 pairs. The score distribution comparison in figures 1 and 2 indicates that the English scores have a wider variance than the Spanish ones, nonetheless the most represented scores (scores corresponding to over 50 sentence pairs) follow roughly a Gaussian distribution. In contrast the Spanish score distribution doesn't behave as cleanly.

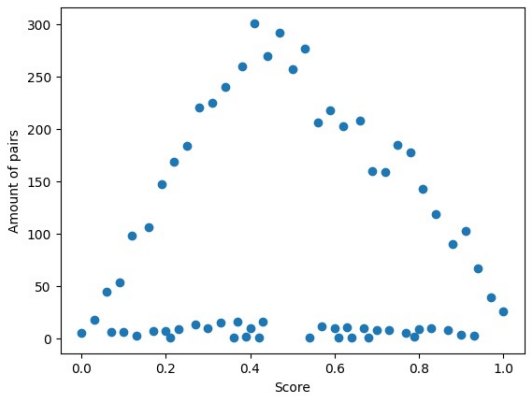

Figure 1: English score distributions.

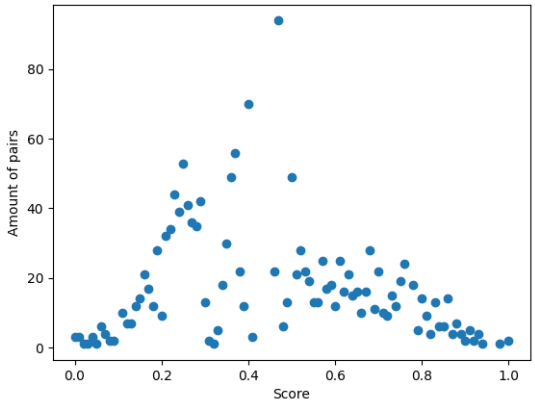

Figure 2: Spanish score distributions.

The results of the shared task are evaluated with the Spearman rank correlation coefficient which is used to discover the strength of a relationship between two sets of data, in this case two sentences.

4.2 Algorithms, parameters and pretrained models

The following regression algorithms, along with the particular parameters, were trained and validated using 7-dimensional vectors corresponding to each characteristic in the SAND Composition:

- SVM and SVM with epsilon: Both algorithms' used the default regularization parameter of 1, meanwhile the SVM ϵ variation used $\epsilon = 0.3$.
- RandomForest: Default parameters were used except for maximum depth which was set to 5, and randomness of the bootstrapping of the samples used when building trees was set at 0.
- Ridge regression: Default parameters except the constant that multiplies the L2 term α. For the English corpus α was set at 0.8, while for the Spanish corpus it was set at 0.9.

Regarding ST and AnglE Embeddings the use of a pretrained model was necessary in order to compute the embeddings and eventually the similarity score. For the English version of ST the ALL-MPNET-BASE-V2 checkpoint was used, meanwhile for the Spanish version the PARAPHRASE-MULTILINGUAL-MPNET-BASE-V2 checkpoint was chosen, both developed particularly for the ST architecture (Reimers and Gurevych, 2020). Meanwhile AnglE embeddings

used the ANGLE-BERT-BASE-UNCASEDNLI-EN-V1 checkpoint, prioritizing the BERT (Devlin et al., 2019) based model over the LLaMA (Touvron et al., 2023) based one due to computational power needed for each model.

5 Results and Analysis

The final SAND composition was the result of an ablation test performed using various combinations of different characteristics. The previously mentioned regression algorithms were trained and tested using each individual characteristic as well as different combinations of each. The SAND composition was the best performing combination for both English and Spanish, each of the four regression algorithms achieved the best result in the ablation test with SAND than with each simpler combination.

Tables 1 and 2 show the best Spearman Correlation for each characteristic combination as well as the model that achieved said score in each language when evaluating in the validation dataset.

Char	SVM	RF	SVM_ϵ	Ridge
ST	0.7891	0.7847	0.7865	0.7891
AnglE	0.7789	0.7737	0.7772	0.7789
N-grams	0.6634	0.6496	0.6620	0.6584
Distance	0.2343	0.2090	0.2839	0.2888
SAND	**0.7986**	**0.8197**	**0.7992**	**0.7921**

Table 1: English Spearman Correlation for different characteristics.

Char	SVM	RF	SVM_ϵ	Ridge
ST	0.6397	0.6058	0.6310	0.6397
AnglE	0.6140	0.5976	0.6038	0.6212
N-grams	0.6425	0.6232	0.6419	0.6431
Distance	0.5595	0.5584	0.5586	0.5594
SAND	**0.688**	**00.6783**	**0.6937**	**0.7029**

Table 2: Spanish Spearman Correlation for different characteristics.

Finally table 3 shows the results of each model with the SAND Composition. The reported results correspond to the predicted scores made by the highlighted models: Random Forest for English, and Ridge Regression for Spanish.

It is important to note that the embeddings created with pretrained models were the feature with the greatest impact on our model. Even as an isolated measure they both prove to be better metrics than their n-grams and distance counterparts. Similarly, considering that both BERT and ALL-MPNET-BASE-V2 are trained in English primarily, it is logical that the regression algorithms performed better in said language.

Model	Spanish	English
RandomForest	0.6968	**0.8197**
SVM	0.6881	0.8133
Ridge	**0.7029**	0.8117
SVM Epsilon	0.6997	0.7945

Table 3: Spearman coefficient for SAND composition.

Nonetheless the SAND Composition proves that these characteristics can be complemented and improved using relevant information such as n-grams coefficients. Since they don't rely on pretrained models and rather focus on lexical overlapping, this coefficient was able to discern certain relatedness measures.

SAND Composition was able to achieve the best results of the ablation test meaning that regression models do benefit from the mix of characteristics and still be relevant in a competition setting.

6 Conclusion

In this paper, we describe the SAND composition for the STR shared task, which is based on both semantic and lexical features, because we observe that: two sentences can share most of the words and apparently have no semantic relation but a high value of Spearman coefficient and vice versa, they can share semantics without matching words. With this in mind the SAND Composition contains half semantic features, and half lexical ones. This approach allowed achieved the 18th place in English and 2nd place in Spanish, with 0.83 and a 0.73 Spearman Correlation score respectively. In both cases the results are over the baseline and only 0.05 of the first place in English and 0.01 in Spanish, meaning SAND proves to have relevant characteristics.

For future experiments added features that consider both semantic, lexical and contextual parts simultaneously might prove to be more efficient than various unrelated metrics. Mixing word embeddings and PoST tagging might generate a relatedness score that proves to be more useful than separate similarity metrics.

Acknowledgments

This research was funded by CONAHCYT (CF-2023-G-64) and PAPIIT project IT100822. G.B.E. is supported by a grant for the requalification of the Spanish university system from the Ministry of Universities of the Government of Spain, financed by the European Union, NextGeneration EU (María Zambrano program, Universitat de Barcelona).

K. Salas-Jimenez thanks CONAHCYT scholarship program (CVU: 1291359).

Ángel Cadena thanks CONAHCYT scholarship program (CVU: 1227093).

References

Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif M. Mohammad. 2023. What makes sentences semantically related: A textual relatedness dataset and In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Dubrovnik, Croatia.

Shima Asaadi, Saif Mohammad, and Svetlana Kiritchenko. 2019. Big BiRD: A large, fine-grained, bigram relatedness dataset for examining semantic composition. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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)*. Association for Computational Linguistics, Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-1050.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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)*. Association for Computational Linguistics, Minneapolis, Minnesota, pages 4171–4186. https://doi.org/10.18653/v1/N19-1423.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871* .

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. http://arxiv.org/abs/1908.10084.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *CoRR* abs/2004.09813. https://arxiv.org/abs/2004.09813.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.