SATLab at SemEval-2024 Task 1: A Fully Instance-Specific Approach for Semantic Textual Relatedness Prediction

Yves Bestgen
Statistical Analysis of Text Laboratory (SATLab)
Université catholique de Louvain
Place Cardinal Mercier, 10 1348 Louvain-la-Neuve, Belgium
yves.bestgen@uclouvain.be

Abstract

This paper presents the SATLab participation in SemEval 2024 Task 1 on Semantic Textual Relatedness. The proposed system predicts semantic relatedness by means of the Euclidean distance between the character ngram frequencies in the two sentences to evaluate. It employs no external resources, nor information from other instances present in the material. The system performs well, coming first in five of the twelve languages. However, there is little difference between the best systems.

1 Introduction

Semantic similarity between words, phrases and texts has long attracted the attention of NLP researchers. It is obviously a useful source of information in tasks such as information retrieval, text summarization, question answering or machine translation (Agirre et al., 2012). It has been the subject of several shared tasks within SemEval since 2012 (Agirre et al., 2012; Marelli et al., 2014; Cer et al., 2017). More recently, interest has also focused on Semantic Textual Relatedness (STR), which is supposed to be a more general concept. As Abdalla et al. (2023) point out, two sentences must be paraphrases or present an entailment relation to be semantically similar, whereas to be related, it is sufficient that they deal with similar themes or express similar points of view on a given issue. Work on STR is less advanced due to the lack of annotated datasets on this dimension (Abdalla et al., 2023). It should be noted, however, that the human annotators who evaluated semantic textual similarity for the SILK dataset (Marelli et al., 2014) clearly evaluated relatedness, since they considered pairs of sentences that contradict each other as semantically very similar (96% similarity), such as in SILK Instance 466:
- *A man is performing a trick on a green bicycle.*
- *There is no man performing a trick on a green bicycle.*

The SILK dataset contains many other examples of this kind of judgement. This observation suggests that the term "relatedness" is more appropriate to describe this field of research, at least when dealing with the intuition of native speakers. It also suggests that techniques that are effective in automatically estimating semantic similarity should also be effective in estimating relatedness. These are mainly state-of-the-art deep learning algorithms (Cer et al., 2017).

In this context, Ousidhoum et al. (2024b) have proposed the SemEval 2024 Task 1, which has a number of specific features compared with previous work. Firstly, the task focuses on relatedness, and is based on material consisting of sentence pairs that have been annotated on this dimension by native speakers. Secondly, the task is highly multilingual, covering more than ten languages, some of which are very poorly resourced. Finally, it includes three subtasks: supervised, unsupervised and crosslingual. In the supervised subtask, the systems were to be trained using training datasets provided by the task organizers. In the unsupervised subtask, no datasets labeled according to semantic relatedness or semantic similarity could be used. In the crosslingual subtask, the system had to be trained on a language other than the target language.

2 The Proposed Approach

Due to its highly multilingual nature (twelve languages), the unsupervised subtask seemed a priori to be particularly interesting for the development of a generic approach, as language-independent as possible. This would be the case of a system that estimates the semantic relatedness of a pair of sentences without recourse to any resources external to the material and even without taking into account the other instances present in the material. A system takes other instances into account when, for example, it weights an instance features according

to their frequency in the complete material, using the classic TF-IDF. A system is completely independent of other instances when the processing of one instance is not affected in any way by the other instances it has to predict. The system proposed by the SATLab fulfills this requirement by using the Euclidean distance between the two sentences, calculated on the basis of the frequency of the ngrams of characters that make them up. If such a system proves successful to predict semantic relatedness, it could become a potential candidate for the analysis of any language.

Admittedly, such a system is more akin to a baseline than a state-of-the-art system. However, it should also be noted that systems based on character ngrams have for many years been considered particularly effective for NLP tasks such as language identification, error correction, information retrieval and even for hate speech and offensive content identification (Damashek, 1995; Bestgen, 2021b). Character ngrams have the advantage of not requiring material to be tokenized, which can be problematic in some Asian languages, and of being able to extract morphological information at very low cost (Peng et al., 2003).

This paper presents SATLab's participation in SemEval 2024 Task 1 with this fully instance-specific system. The following section introduces the task and describes the proposed system. The results obtained are then reported.

3 The Unsupervised Task

Subtask 1B of SemEval 2023 (Ousidhoum et al., 2024b) asked participating teams to estimate the semantic relatedness between pairs of sentences in twelve languages: five Afro-Asiatic (Algerian Arabic [arq], Amharic [amh], Hausa [hau], Modern Standard Arabic [arb], Moroccan Arabic [ary]), five Indo-European (Afrikaans [afr], English [eng], Hindi [hin], Punjabi [pan] and Spanish [spa]), one Austronesian (Indonesian [ind]) and one from the Niger-Congo family (Kinyarwanda [kin]). The material, collected by Ousidhoum et al. (2024a), was selected from various resources such as semantic similarity datasets, news articles and Wikipedia texts. After this material had been carefully checked, it was submitted to native speakers whose task was to assess the semantic relatedness between pairs of sentences using the Best-Worst Scaling procedure. Ousidhoum et al. (2024a) reported high to near-perfect inter-rater reliabilities (split-half correlations: Min = 0.64, Max = 0.96).

In this Task 1B, the systems had to be unsupervised, since no dataset including evaluations of semantic relatedness between sentence pairs or texts could be employed. It should be noted, however, that the organizers provided participants with development data similar to that provided later for the testing phase, and that a team's predictions for these data could be evaluated by submitting them to CodaLab. The few tests I carried out showed that performance varied greatly depending on the language. It therefore didn't seem advisable to rely on this development material to make general decisions about the system to be developed. In the testing phase, only one prediction for each language could be submitted, and the performance measure was Spearman's rank correlation coefficient.

4 The SATLab System

A single system was used for all twelve languages. It is adapted from the one developed for the authorship identification of source code (Bestgen, 2020). This system takes as input each pair of utterances and outputs a distance between them without any other information, either from the rest of the material or external to it. Each pair of utterances is therefore processed in a way that is completely independent of the other pairs present in the material.

The only pre-processing is the lower-casing of all texts as included in SAS. I have to admit that it's not obvious to me what impact this has on languages as unknown to me as Kinyarwanda or Amharic. No tokenization or lemmatization has been applied. The system uses character ngrams made up of 1 to 5 characters. All characters are taken into account, including spaces, punctuation marks, symbols, characters from other writing systems, etc. The ngrams at the beginning and end of each statement are distinguished from the others. All ngrams in a statement are retained, so there is no frequency threshold. The frequency of each feature is weighted by a logarithmic function using the formula: $1 + \log(Freq)$. Finally, the features of each statement are weighted by the L2 norm (thus instance-wise). Most of these system components have be taken from the one developed for a difficult language identification problem (Bestgen, 2021a).

The Euclidean distance between the sets of ngrams of each utterance in a pair is used to estimate the semantic dissimilarity between these utterances. Before submission, these distances are

transformed into similarity by ranking them from largest to smallest. No information is lost through such ranking, since the organizers have chosen a rank correlation as the efficiency criterion.

5 Analysis and Results

5.1 Official Results

Twelve teams took part in the test phase of Task 1B, but only five proposed solutions for all twelve languages. One team proposed a solution for all languages except Spanish. The organizers provided a baseline based on the number of shared words between the two sentences of a pair (SemRel Lexical Overlap Baseline, see Ousidhoum et al. (2024b) for details).

Figure 1 shows the performance of all the systems for the twelve languages, highlighting the baseline and the system proposed by the SATLab. Marks not connected by a line are from systems that did not submit a solution for all languages. I don't know whether the systems proposed by the other teams are identical for all twelve languages, as is the case for the baseline and the SATLab.

This figure merits several comments. Firstly, when we analyze the overall results, we observe that the profiles of the teams[1] who submitted for all languages are similar. This observation is confirmed by an analysis of the Pearson correlations between these profiles. The lowest correlation is 0.54, only two are below 0.63 and half of them are above 0.73. These profiles highlight strong variations in performance according to language. While almost all the teams performed well to very well for Afrikaans (afr), Amharic (amh), English (eng) and Spanish (spa), they performed poorly for Punjabi (pan), with the SATLab system even achieving a negative correlation. It therefore appears that the material for some languages is considerably more complicated than for others. A detailed analysis of the differences between these materials would therefore be very useful.

Figure 1 also shows that the SATLab's performance is as good as or better than that of other teams in the vast majority of languages, but there is little difference between the best teams. This second observation would certainly be confirmed if confidence intervals, obtained by bootstrapping (Bestgen, 2022), were presented, but their calculation requires access to the predictions of all systems. In any case, when performances are so close, it is essential to take into account other factors such as computational complexity, which will be possible when reading the system descriptions of the other teams.

Finally, Figure 1 also shows that the organizers' baseline is superior to all other systems for two languages: Hindi (hin) and Moroccan Arabic [ary]. Clearly, this is an underperformance by all participants.

5.2 System Component Analysis

To assess the contribution of each component to the system overall performance, all of them were modified, one at a time, and the system was re-evaluated using the gold standard provided by the task organizers for eleven languages. The results are shown in Table 1 using the difference between each modified system and the official SATLab system, whose performance is shown in the first row.

The only pre-processing of the material carried out, the lower casing, brings benefits in only two languages. Presumably, it doesn't affect the many languages that don't use Latin characters. Using ngrams whose maximum length is one character shorter or one character longer has very little impact. On the other hand, feature weighting by TF-IDF is beneficial in ten out of eleven languages. Not using L2 normalization profoundly alters performance. While it brings significant benefit in one language, the impact is negative in nine languages, and can reach -0.574. As far as distance is concerned, Dice is more efficient than the Euclidean distance, but the gain is significantly lower than that obtained by applying the Euclidean distance to the weights transformed by TF-IDF.

The last line gives the correlations obtained by the system when TF-IDF is used instead of the logarithmic weighting. The gains over the official SATLab submission are sufficiently large to conclude that a fully instance-specific approach is significantly less effective at predicting STR than an approach that takes into account the other instances of the test material (which TF-IDF does, as explained in the introduction). There is no point in comparing these correlations with those of the other participants, since they would certainly have submitted a different system if they had been able to optimize it as just done.

[1] In this discussion of results, the baseline is considered a "team".

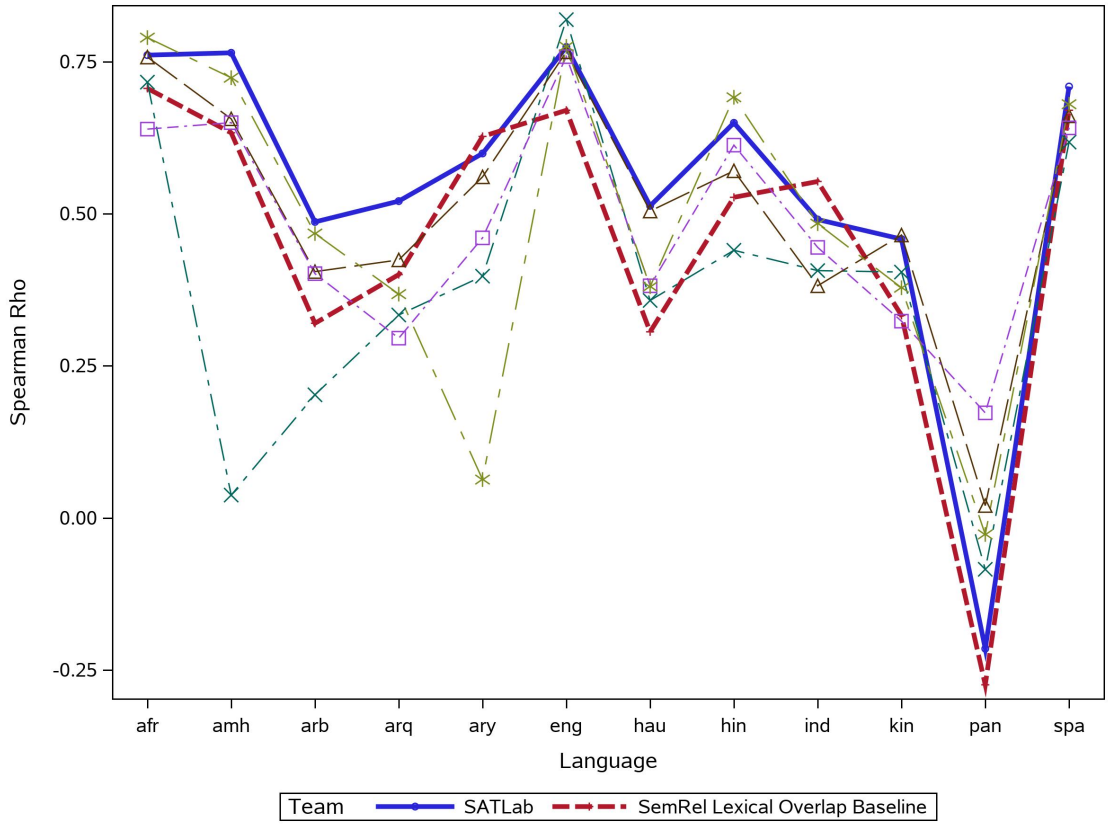

Figure 1: Performances of all systems for the twelve languages

Expe	afr	amh	arb	arq	ary	eng	hau	hin	ind	kin	pan
Submitted	0.761	0.764	0.487	0.521	0.599	0.774	0.513	0.649	0.491	0.458	-0.215
No Lowercase	0.005	0.000	0.000	0.000	0.000	-0.02	-0.028	0.000	0.005	0.005	0.000
4-grams	-0.003	-0.001	-0.016	0.005	-0.012	-0.002	-0.015	-0.007	0.007	0.012	0.018
6-grams	0.001	0.000	0.004	-0.009	-0.002	-0.001	0.003	0.000	-0.012	-0.004	-0.009
TF-IDF	0.021	0.001	0.061	0.052	0.024	0.024	0.057	0.046	-0.052	0.069	0.002
BM25	0.011	-0.008	0.043	-0.085	0.005	0.014	0.026	-0.091	-0.077	0.083	0.032
No L2	-0.144	-0.247	-0.421	-0.574	0.211	-0.245	-0.059	-0.432	0.003	-0.157	-0.135
Cosinus	-0.008	0.013	-0.013	-0.022	-0.003	-0.005	-0.012	-0.020	0.001	-0.035	0.005
Dice	-0.001	0.003	0.032	0.036	0.022	0.012	0.029	0.002	0.005	-0.021	0.003
Best	0.782	0.765	0.548	0.573	0.623	0.798	0.570	0.695	0.439	0.527	-0.213

Table 1: Analysis of the impact of the system components

6 Conclusion

This paper presents the SATLab participation in SemEval 2024 Task 1: Semantic Textual Relatedness (STR). The proposed system predicts semantic relatedness by means of the Euclidean distance between two sentences, calculated on the basis of the frequency of the ngrams of characters that make them up. It employs no resources external to the material and extracts no information from other instances present in the material. The system performs well, coming first in five of the twelve languages. However, there is little difference between the best systems. What's more, the baseline proposed by the organizers was better than all the systems proposed by the participants in two languages.

Analysis of the system's components shows that the decision to develop a fully instance-specific approach was clearly the wrong one. Simply taking into account the frequencies of features in the material as a whole, as the TF-IDF weighting system does, provides a significant benefit, as Damashek (1995) has already pointed out when character ngrams are used in other NLP tasks.

The performance of all teams varies considerably according to language. It would be very interesting to carry out further research to try and understand the origin of these fluctuations. Otherwise, this type of unsupervised approach cannot be recommended, since negative correlations are observed for one of the languages. It is possible that this is linked to the way in which the material has been designed, which varies greatly depending on the language for obvious reasons of unavailability of certain resources (Ousidhoum et al., 2024a).

7 Ethical Considerations

The ethical issues raised by this research are identical to those described by the researchers who collected the data (Ousidhoum et al., 2024a) and by the researchers who organized this task (Ousidhoum et al., 2024b).

Acknowledgements

The author wishes to thank the organizers of this shared task for putting together this valuable event and for their availability throughout the task. He is a Research Associate of the Fonds de la Recherche Scientifique - FNRS (Fédération Wallonie Bruxelles de Belgique).

References

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

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

Yves Bestgen. 2020. Boosting a KNN classifier by improving feature extraction for authorship identification of source code. In *Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation*, CEUR Workshop Proceedings, pages 705–712. CEUR-WS.org.

Yves Bestgen. 2021a. Optimizing a supervised classifier for a difficult language identification problem. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 96–101, Kiyv, Ukraine. Association for Computational Linguistics.

Yves Bestgen. 2021b. A simple language-agnostic yet strong baseline system for hate speech and offensive content identification. In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation*, CEUR Workshop Proceedings, pages 1–10. CEUR-WS.org.

Yves Bestgen. 2022. Please, don't forget the difference and the confidence interval when seeking for the state-of-the-art status. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5956–5962, Marseille, France. European Language Resources Association.

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.

Marc Damashek. 1995. Gauging similarity with n-grams: Language-independent categorization of text. *Science*, 267(5199):843–848.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full

sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8, Dublin, Ireland. 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-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.

Fuchun Peng, Dale Schuurmans, and Shaojun Wang. 2003. Language and task independent text categorization with simple language models. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 189–196.