# Incorporating multiple feature groups to a Siamese Neural Network for Semantic Textual Similarity task in Portuguese texts

João Vitor Andrioli de Souza[1], Lucas Emanuel Silva e Oliveira[1], Yohan Bonescki Gumiel[1], Deborah Ribeiro Carvalho[1] and Claudia Maria Cabral Moro[1]

Graduate Program on Health Technology (PPGTS), Pontifical Catholic University of Paraná (PUCPR). Curitiba, Brazil
joao.vitor.andrioli@gmail.com,{lucas.oliveira,yohan.gumiel,ribeiro.carvalho,c.moro}@pucpr.br

**Abstract.** The Semantic Textual Similarity (STS) algorithms have a key role in Natural Language Processing (NLP) studies since it can support various NLP tasks such as Text Summarization and Information Retrieval. Although we found several STS initiatives in the literature, just a few authors explored Siamese Neural Networks (SNN) to solve this problem, especially for the Portuguese language, even considering their lower need for training data and an architecture built for similarity tasks. We defined a set of lexical, semantic, distributional and graph-based feature groups to capture distinct aspects of the text and incorporated to a SNN architecture to perform STS in ASSIN 1 and ASSIN 2 datasets. The experiments indicate positive results since we improved the results of previous attempts of STS using SNNs in Portuguese texts.

**Keywords.** Semantic Textual Similarity, Siamese Neural Networks, Shared Tasks.

## 1 Introduction and Background

The Semantic Textual Similarity (STS) task consists of quantifying the degree of semantic equivalence of one text to another. It is essential for many Natural Language Processing (NLP) research and applications, supporting tasks such as Plagiarism Detection, Text Deduplication, Text Summarization, Information Retrieval and Text Clustering [1].

The Neural Network (NN) architectures have been outperforming traditional Machine Learning (ML) models in several fields of study including NLP. One example of successful NN is the Siamese Neural Network (SNN), which is a type of NN used to calculate similarity in studies like [2–5], it has been successful in various tasks focused on both image, and more recently on text context, achieving good results using less data than other approaches. Additionally, it has shown less susceptibility to overfitting [5].

The SNNs layers are configurable to any layer type that suits the problem resolution, such as Convolutional Neural Networks or Long-Short Term Memory, the architecture is composed of two or more equal subnetworks that share the same configurations, parameters, and weights, which has the values updated simultaneously during the learning process [6].

Many features have been tested to tackle the STS problem, such as lexical with string-based approaches, semantical with corpus-based and knowledge-based approaches [7], structural syntactic or morphological information, and recently, several studies have used distributional and contextual Word Embeddings (WE).

The shared tasks have an important role in the STS task, for the English language, the SemEval initiatives released various STS tasks over the years (e.g., [1]). The Portuguese language is represented by the ASSIN 1 [8] and ASSIN 2 [9] shared tasks, which focused on STS texts from the journalistic domain. The winning team [10] of ASSIN 1 used an approach in which they combined TF-IDF calculation with WE. The SNN architecture was not fully explored, with just one group of ASSIN 1 exploring it [11] and just a few other studies applying to other languages [e.g., [4–6, 12].

We hypothesize that associating the SNN's efficacy to train with data limitations (which is often the case in shared-tasks), and the use of a set of lexical, semantic, graph representation and distributional features, it could be able to capture different aspects of the text, establishing a consistent model.

In this work, we present a SNN architecture inputted with lexical, semantic, distributional and graph-based features, aiming to perform STS in ASSIN 1 & 2 datasets.

## 2 Materials and Methods

### 2.1 Dataset

The ASSIN 1 & 2 datasets are composed of manually annotated Portuguese sentence pairs with their respective similarity/relatedness scores ranging from 1 to 5, where 1 depicts no similarity and 5 depicts equivalence. The ASSIN 1 dataset is divided into two parts, the Brazilian Portuguese (BR) and the European Portuguese (PT), while the ASSIN 2 contains BR sentences only. Table 1 presents the sizes of the datasets while Table 2 depicts some sentence pairs and their respective similarities.

Aiming to conduct exploratory data analysis and compare the data sparsity and density of both datasets, we applied a graph representation algorithm to the data and verified interesting aspects such as low vocabulary volume on ASSIN 2 dataset compared to the ASSIN 1, that despite having more sentence pairs, have less unique tokens (shown in Table 3). When we compare the number of edges of each dataset (see Table 4) it implies the high and low data sparsity of ASSIN 1 and ASSIN 2 respectively.

**Table 1**. The number of text pairs in each dataset.

| | ASSIN 2 | ASSIN 1 | | ASSIN 1 & 2 |
|---|---|---|---|---|
| | | PT | BR | |
| **Train** | 7000 | 3000 | 3000 | 13000 |
| **Test** | 2448 | 2000 | 2000 | 6448 |
| **Total** | 9448 | 5000 | 5000 | 19448 |

**Table 2.** ASSIN 1 & 2 text pairs with their corresponding similarity score and description.

| **ASSIN 1 & ASSIN 2** | | |
|---|---|---|
| **1** | **Description** | The two sentences are totally unrelated |
| | **S1** | *Um cachorro branco de coleira está andando na água* |
| | **S2** | *Um homem sem camisa está jogando futebol em um gramado* |
| | **Score** | 1.0 |
| **2** | **Description** | The two sentences have similar actions or objects. |
| | **S1** | *Um cachorro aparentemente desnutrido está em pé nas patas de trás e se preparando para pular* |
| | **S2** | *Um cachorro de aparência saudável está deitado no chão* |
| | **Score** | 2.0 |
| **3** | **Description** | The two sentences share details. |
| | **S1** | *Um homem e uma criança estão andando de caiaque pelas águas calmas* |
| | **S2** | *Um caiaque amarelo está sendo navegado por um homem e um menino jovem* |
| | **Score** | 3.0 |
| **4** | **Description** | The two sentences are closely related. |
| | **S1** | *O cara está montando um cavalo perto de um riacho* |
| | **S2** | *O cara está montando um cavalo perto de uma correnteza* |
| | **Score** | 4.0 |
| **5** | **Description** | The two sentences are equivalent. |
| | **S1** | *Um cara está brincando com uma bola de meia* |
| | **S2** | *Tem um cara brincando com uma bola de meia* |
| | **Score** | *5.0* |

**Table 3.** The number of Unique Tokens (Nodes) in the ASSIN 1 & 2 datasets on train, test and concatenated train+test portions.

| | **ASSIN 2** | **ASSIN 1** | |
|---|---|---|---|
| | | **PT** | **BR** |
| **Train** | 2342 | 11075 | 10058 |
| **Test** | 1967 | 9282 | 8675 |
| **Train+Test** | 2542 | 15249 | 13757 |
| | | 22389 | |
| | 23673 | | |

**Table 4.** The number of Unique Token Bigram (Edges) in the ASSIN 1 & 2 datasets on train, test and concatenated train+test portions.

| | **ASSIN 2** | **ASSIN 1** | |
|---|---|---|---|
| | | **PT** | **BR** |
| **Train** | 8787 | 45218 | 40039 |
| **Test** | 7090 | 35075 | 33441 |
| **Train+Test** | 10327 | 70965 | 63710 |
| | | 120453 | |
| | 129143 | | |

### 2.2 Machine Learning classifier and Features

Our STS algorithm was based on formerly proposed SNN in [6], but the Manhattan's distance was replaced with a 50-units dense layer since the addition of new features would not work well with the Manhattan's distance. We used two 300-sized BiLSTM subnetworks and trained for 70 epochs with the mean squared error (*mse*) loss function.

We decided to use the pre-trained Word2Vec CBOW 300-sized vector from NILC [13] since it was utilized in previous STS works for Portuguese and achieved good results [10, 14] (from now on named *NILC Word2vec*) . The 100-sized Word2Vec skip-gram vector ID 63 from NLPL WE Repository[1] was used as well, hereafter *NLPL ID 63 Word2vec*. A 100-sized vector was trained with each of the dataset groups texts using the Word2Vec algorithm [15], to work as our baseline, henceforth *ASSIN Word2Vec*.

[1] http://vectors.nlpl.eu/repository/

A new parallel dense layer with 100-units was incorporated in the SNN to accommodate five new feature groups: (i) lexical-based similarities, (ii) knowledge-based wordnet tokens distances, (iii) distributional-based WE cosine distance between the sentences, (iv) the sum of the degree centrality of the tokens using a graph created from the dataset, and (v) overlap of common words around the sentences.

The lexical-based feature group is composed of three different equations, computed with the Jaccard index, Dice coefficient and Cosine distance [7]. They represent the overlap of common tokens of both sentences and have shown very similar results to the word overlap metric used as the baseline of the ASSIN.

We used the Open Multilingual Wordnet (OMW)[2] from the Natural Language Toolkit (NLTK) to build our knowledge-based feature group. Three different similarity calculations were used as features, such as the Wu-Palmer similarity, Leacock-Chodorow similarity and shortest path distance that connects the hypernym/hyponym taxonomy.

The third feature group is the cosine distance of both inputted sentences using the WE model selected for each experiment (more details on the different models later in this chapter). Each sentence position is the average position of their words in the WE model, as presented in Figure 1.

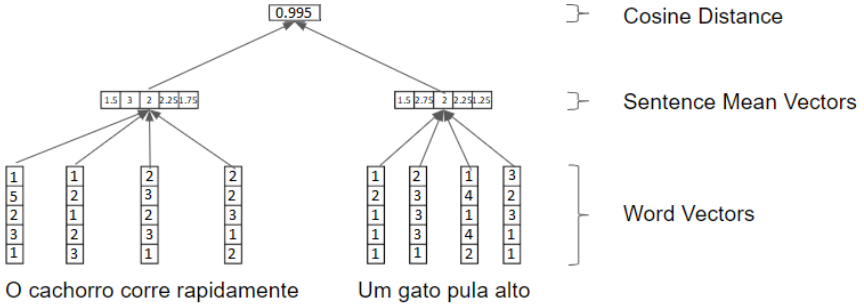


**Figure 1.** Process for calculating the Cosine distance of two sentences Word Vectors.

Intending to use the information found in the dataset itself as features, a directed graph of each dataset was generated to calculate both the degree centrality of a token and the overlap of common words around the sentences, each node is a token, and each edge is the pair of token (current token, next token), the edges possess some syntactic information.

The degree centrality of a token is the link incident over the token, for instance in Figure 2 the degree centrality of "*jovem*" is 8, because there are 8 links around it, this metric was selected as a way to measure the significance of the words. We have chosen the degree centrality equation due to the low computational cost and the high importance of words closely related, other centrality equations can be used and need to be tested.

[2] https://www.nltk.org/howto/wordnet.html

The overlap of common words around the sentences is the number of common words directly around both sentences; the equation is normalized between 0 and 1 by dividing the overlap with the sum of tokens around the sentences.

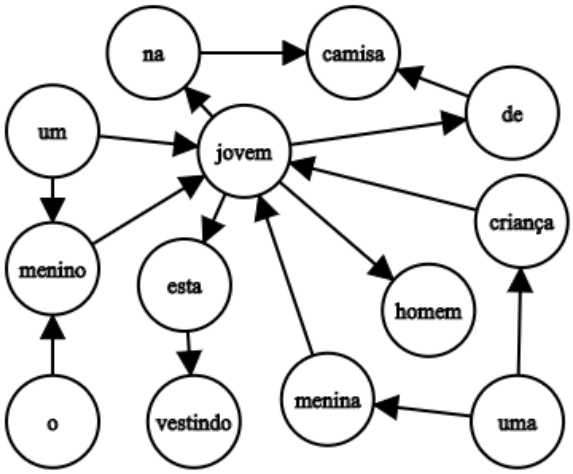


**Figure 2.** Example of the generated Graph.

Our experiments were executed in each of the following dataset configurations: (i) a concatenation of ASSIN 1+2, (ii) ASSIN 2, (iii) ASSIN 1 BR, (iv) ASSIN 1 PT and (v) a concatenation of ASSIN 1 BR and ASSIN 1 PT. Our experimental setup involved the algorithm evaluation using only the WE as features to our SNN, and the WE plus additional hand-crafted features. The performance was evaluated using the same metrics used in the ASSIN, the Pearson correlation (*p*) and the mean squared error (*mse*).

## 3 Results

The results of our SNN algorithm are displayed in Table 5, which distinguishes the scores for each dataset configuration (i.e., ASSIN1, ASSIN2, ASSIN1-PT, ASSIN1-BR, ASSIN1&2), WE model (i.e., ASSIN Word2Vec, NLPL ID 63 Word2vec and NILC Word2vec) and features used (i.e., WE plus five features and WE as only feature). The best scores for each dataset configuration are in bold (*p* value the larger the better, *mse* the smaller the better).

The five proposed features seem to improve most of the results if compared to the "WE as only feature" runs. The best *p* score was achieved by the five features model with the NLPL ID 63 Word2vec pre-trained model for all datasets, for ASSIN 2 it scored 0.72 *p* and 0.65 *mse*. Our feature selection has improved the scores mostly for the pre-trained WE.

In Table 6, we compare the performance of our current method with the language-independent approach developed in [16]. The improvement is evident in all dataset configurations.

Table 7 shows a comparison with the five feature groups in isolation and pairs, that were evaluated exclusively with the NLPL ID 63 WE in the concatenation of ASSIN 1 & 2 datasets, giving that the NLPL ID 63 WE achieved the best results in most runs. Differently from previous experiments, which were trained with 70 epochs, this table was trained with only 50 epochs, due to time constraints and that the focus was only to compare the features results.

**Table 5.** Pearson correlation and Mean squared error scores for each SNN algorithm execution

| | **ASSIN Word2Vec** | | | | **NLPL ID 63 Word2vec** | | | | **NILC Word2vec** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *p* | | *mse* | | *p* | | *mse* | | *p* | | *mse* | |
| | basic | feat | basic | feat | basic | feat | basic | feat | basic | feat | basic | feat |
| **ASSIN 2** | 0.67 | 0.70 | 0.64 | 0.65 | 0.69 | **0.72** | 0.70 | 0.65 | 0.68 | 0.68 | 0.73 | **0.60** |
| **ASSIN 1** | 0.62 | 0.62 | 0.59 | 0.61 | 0.62 | **0.66** | 0.58 | **0.56** | 0.61 | 0.64 | 0.62 | 0.57 |
| **ASSIN 1 PT** | 0.64 | 0.64 | 0.75 | 0.76 | 0.62 | **0.66** | 0.72 | **0.64** | 0.62 | 0.65 | 0.75 | 0.66 |
| **ASSIN 1 BR** | 0.61 | 0.61 | 0.49 | 0.50 | 0.63 | **0.64** | 0.47 | 0.46 | 0.62 | **0.64** | 0.51 | **0.45** |
| **ASSIN 1+2** | 0.66 | 0.66 | 0.64 | 0.64 | 0.68 | **0.70** | 0.62 | **0.60** | 0.65 | 0.67 | 0.70 | 0.63 |

***basic** denotes for the WE as the only feature approach.

***feat** denotes for the WE and additional five features approach.

**Table 6.** Best scores of our current method compared to a language-independent method

| | ***p*** | | ***mse*** | |
|---|---|---|---|---|
| | *lind* | *new* | *lind* | *new* |
| **ASSIN 2** | 0.69 | **0.72** | 0.61 | **0.60** |
| **ASSIN 1** | 0.63 | **0.66** | 0.60 | **0.56** |
| **ASSIN 1 PT** | 0.64 | **0.66** | 0.75 | **0.64** |
| **ASSIN 1 BR** | 0.63 | **0.64** | 0.46 | **0.45** |
| **ASSIN 1+2** | 0.64 | **0.70** | 0.71 | **0.60** |

***lind** denotes for the language-independent approach by [16].

***new** denotes the best result of our current approach.

## 4 Discussion

The degree centrality and the overlap of common words around are dependent on the generated graph, and the graph for ASSIN 1 and ASSIN 2 have very different aspects, for instance, the low data sparsity of ASSIN 2, shown in Table 3 and Table 4, may have led to the different performances for each WE on Table 5, contrasting with each other distinct dataset performance.

**Table 7.** Comparison of the features in pairs and isolated.

| | ***(isolated)*** | | **Lexical** | | **WordNet** | | **WE Cosine Distance** | | **Degree Centrality** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *p* | *mse* | *p* | *mse* | *p* | *mse* | *p* | *mse* | *p* | *mse* |
| **Lexical** | 0.68 | 0.64 | - | - | - | - | - | - | - | - |
| **WordNet** | 0.66 | 0.71 | 0.68 | 0.63 | - | - | - | - | - | - |
| **WE Cosine Distance** | 0.67 | 0.63 | 0.70 | **0.58** | 0.67 | 0.61 | - | - | - | - |
| **Degree Centrality** | 0.60 | 0.74 | 0.68 | 0.67 | 0.61 | 0.72 | 0.68 | 0.72 | - | - |
| **Common Words Around** | 0.64 | 0.65 | 0.69 | 0.60 | 0.69 | 0.65 | **0.71** | 0.60 | 0.64 | 0.65 |

The lexical and the WE cosine distance were the less costly features to calculate, while the overlap of common words around was the slower to calculate, despite that they have shown good results, while the WordNet have shown smaller score increase and the degree centrality have not shown good results.

The degree centrality got the worst results, this might be an indication of low representativeness of the significance of words for similarity tasks, nonetheless, other centralities and node influence metrics need to be tested and the graph representation features can be improved with contextual weighting.

One of the difficulties relative to the method selection was due to the lack of annotation guidelines released, therefore some aspects of the text were not fully presented, for instance, if the scores are about sentence relatedness or similarity. For example, in the sentences "*O menino está tocando o piano*" and "*O menino não está tocando o piano*" with score "3.0" the sentence is related, but due to the negation the meaning is opposite to each other, this score does not corroborate with the sentences "*Não tem nenhum homem executando um truque em uma bicicleta verde*" and "*Um homem está realizando um truque em uma bicicleta verde*" with score "4.8".

As future work we intend to use at least the three best features of Table 7 with a contextual WE, such as BERT and ELMO, this could improve the scores by taking advantage of the contextual aspects stored in this kind of embeddings. We hypothesize that a contextual WE would improve the results mainly of ASSIN 2, due

to the low data sparsity of the dataset, while a distributional WE could fail when dealing with words used in multiple contexts and with more similarity variation.

## 5 Conclusion

We trained a SNN architecture in association with pre-trained Word Embeddings and a set of features trying to cover different aspects of the text (e.g., lexical, semantic), aiming to perform the Semantic Textual Similarity task in Portuguese texts. The experiments showed promising results since we improved all the last attempts to use SNN for Portuguese STS, and checked each feature contribution by isolating each one, and training separated models.

## References


1. Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A.: SemEval-2012 Task 6: A pilot on semantic textual similarity. In: *SEM 2012 - 1st Joint Conference on Lexical and Computational Semantics. pp. 385–393. Association for Computational Linguistics, Montréal, Canada (2012).
2. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a "Siamese" time delay neural network. In: NIPS'93 Proceedings of the 6th International Conference on Neural Information Processing Systems. pp. 737–744. Morgan Kaufmann Publishers Inc., Denver, Colorado (1993).
3. Chopra, S., Hadsell, R., LeCun, Y.: Learning a Similarity Metric Discriminatively, with Application to Face Verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). pp. 539–546. IEEE, San Diego, CA, USA (2005).
4. Neculoiu, P., Versteegh, M., Rotaru, M.: Learning Text Similarity with Siamese Recurrent Networks. In: Proceedings of the 1st Workshop on Representation Learning for NLP. pp. 148–157. Association for Computational Linguistics, Stroudsburg, PA, USA (2016).
5. Ranasinghe, T., Orasan, C., Mitkov, R.: Semantic Textual Similarity with Siamese Neural Networks. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2019. , Varna, Bulgaria (2019).
6. Mueller, J., Thyagarajan, A.: Siamese recurrent architectures for learning sentence similarity. In: 30th AAAI Conference on Artificial Intelligence, AAAI 2016. pp. 2786–2792. AAAI Press, Phoenix, Arizona (2016).
7. H.Gomaa, W., A. Fahmy, A.: A Survey of Text Similarity Approaches. Int. J. Comput. Appl. 68, 13–18 (2013).
8. Fonseca, E.R., Santos, L.B. dos, Criscuolo, M.: Visão Geral da Avaliação de Similaridade Semântica e Inferência Textual. In: Linguamática. pp. 3–13 (2016).
9. Real, L., Fonseca, E., Gonçalo Oliveira, H.: The ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese. In: Proceedings of the ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese. p. In this volume. CEUR-WS.org (2020).

10. Hartmann, N.S.: Solo queue at ASSIN: Combinando abordagens tradicionais e emergentes. Linguamatica. 8, 59–64 (2016).
11. Barbosa, L., Cavalin, P., Guimarães, V., Kormaksson, M.: Blue Man Group at ASSIN : Using Distributed Representations for Semantic Similarity and Entailment Recognition. Linguamática. 8, 15–22 (2016).
12. Barrow, J., Peskov, D.: UMDeep at SemEval-2017 Task 1: End-to-End Shared Weight LSTM Model for Semantic Textual Similarity. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 180–184. Association for Computational Linguistics, Stroudsburg, PA, USA (2017).
13. Hartmann, N.S., Fonseca, E., Shulby, C.D., Treviso, M. V, Rodrigues, J.S., Aluísio, S.M.: Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. In: Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology. pp. 122–131. Sociedade Brasileira de Computação, Uberlândia, MG, Brazil (2017).
14. Alves, A., Gonçalo Oliveira, H., Rodrigues, R., Encarnação, R.: ASAPP 2.0: Advancing the state-of-the-art of semantic textual similarity for Portuguese. In: Informatik, S.D.-L.-Z. fuer (ed.) Proceedings of 7th Symposium on Languages, Applications and Technologies (SLATE 2018). pp. 1–12 (2018).
15. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. arXiv Prepr. arXiv1301.3781. (2013).
16. de Souza, J.V.A., Oliveira, L.E.S., Gumiel, Y.B., Carvalho, D.R., Moro, C.M.C.: Exploiting Siamese Neural Networks on Short Text Similarity tasks for multiple domains and languages. In: Computational Processing of the Portuguese Language - 13th International Conference, PROPOR 2020, Évora, Portugal, March 2-4, 2020, Proceedings (2020).