Journal of Computer Science

Original Research Paper

Enhancing Brazilian Portuguese Textual Entailment Recognition with a Hybrid Approach

Allan de Barcelos Silva and Sandro José Rigo

Applied Computing Graduate Program Program,
Universidade do Vale do Rio dos Sinos - UNISINOS, São Leopoldo, Brazil

Article history
Received: 05-05-2018
Revised: 15-06-2018
Accepted: 06-07-2018

Corresponding Author:
Allan de Barcelos Silva
Applied Computing Graduate Program Program,
Universidade do Vale do Rio dos Sinos - UNISINOS, São Leopoldo, Brazil
Email: albarsil@gmail.com

Abstract: Previous work on textual entailment has not fully exploited aspects of deep linguistic relations, which have been shown as containing important information for entailment identification. In this study, we present a new method to compute semantic textual similarity between two sentences. Our proposal relies on the integration of a set of deep linguistic relations, lexical aspects and distributed representational resources. We used our method with a large set of annotated data available from the ASSIN Workshop in the PROPOR 2016 event. The achieved results score among the best-known results in the literature. A perceived advantage of our approach is the ability to generate good results even with a small corpus on training tasks.

Keywords: Semantic Textual Similarity, Computational Linguistics, Textual Entailment, Word Embeddings, Machine Learning

Introduction

Semantic Textual Similarity (STS) analysis performs an increasingly important role in research and applications related to the Natural Language Processing (NLP) field. The ability to identify the degree of similarity between sentences is crucial to many of the NLP tasks, such as information retrieval, text classification, document clustering, topic detection, among others (Gomaa and Fahmy, 2013; Freire *et al*., 2016). We can consider STS process as composed of three main steps (Ferreira *et al*., 2016). The first one deals with the representation of the sentences, typically using the words and corresponding syntactic information. A second step implements a set of similarity measures to be applied between sentences. These measures are directly related to the kind of information used in the first step. In the last step, the initial representations and the similarity measures results are applied as an input to classification algorithms.

STS methods relying on the similarity identification of shared words between sentences restricts the analysis to the syntactic information only, causing aspects akin to the word order and the sentences meaning to be bypassed (Ferreira *et al*., 2016; 2018). Approaches to reduce this restriction include the use of a broad set of elements, representing lexical, syntactic and semantic dimensions (Gomaa and Fahmy, 2013; Pradhan *et al*., 2015; Ferreira *et al*., 2018; Chen *et al*., 2017; Berrahou *et al*., 2017). Another line of investigation is dedicated to evaluating improvements in the sentence similarity identification by applying constraints in iterative process (Kajiwara *et al*., 2017), while some other approaches include natural deduction proofs to identify bidirectional entailment relations between sentence pairs (Yanaka *et al*., 2017).

A crescent number of works in STS literature rely on the use of resources such as WordNet, FrameNet and VerbNet for integrating some linguistic relationships to the STS process (Al-Alwani, 2015; Yousif *et al*., 2015; Brychcín and Svoboda, 2016; Ferreira *et al*., 2016; Kashyap *et al*., 2016; Ferreira *et al*., 2018). As a complement aspect, probabilistic-based techniques, as we can see in the Vector Space Models (VSM) has been motivating studies about its advantages, such as domain independence and the ability to automatically obtain some of the semantic relations between sentences considering a space of contexts (Hartmann, 2016; Barbosa *et al*., 2016; Freire *et al*., 2016).

Although the number of works integrating linguistic and probabilistic aspects is growing, extensive studies experimenting with details these sets of attributes are still a necessity. In the present work, we present an experiment with a new method to compute the similarity

between two sentences in Brazilian Portuguese. We integrate a set of linguistic resources and probabilistic techniques to better represent the phrases as means to maximize similarity classification assertiveness between sentence pairs. The linguistic relations antonymy, hyperonym, hyponymy and synonymy were explored under the aegis of the TeP synonymous database (Maziero *et al*., 2008) and the Portuguese Unified Lexical Ontology (PULO) (Simões and Guinovart, 2014). Resources achieved with the use of Vector Space Model, a metric of Term Frequency- Inverse Document Frequency (TF-IDF) and Principal Component Analysis (PCA) were applied in the method.

The primary investigation interest in this study is related to the advantages observed in the use of deep linguistic relations as part of the attributes representing the sentences. These relations can explain the sentence similarity when used by linguist experts in text analysis tasks (Evans and Green, 2006), but the studied literature in STS area did not describe it as a resource applied in the models. We show that the contribution of these main linguistic relations regarding the sentences elements can be proven a promising complement to identify similarity in sentences.

We assessed the proposed approach with a dataset made available in the International Conference on the Computational Processing of Portuguese (PROPOR), which is a well-known and respected event in the Brazilian Portuguese STS research community. The results achieved appear among the best results in the literature of semantic textual similarity for Brazilian Portuguese. One important aspect to highlight is the fact that we used a small corpus for training tasks when compared with the other works. Our training corpus is about 0.53% of the compared corpora. This is considered as an indication of the advantages one can obtain incorporating linguistic aspect in the process. Moreover, our experiments show that the use of linguistic relations combined with probabilistic techniques scored better results than using only one of the approaches.

The structure of this paper is the following. Section two presents linguistic background. In section three are described related works. Section four presents the adopted approach. In section five, we present the obtained results. Finally, the conclusions are presented in section six.

Linguistic Aspects of Similarity in Texts

In this section, we present some aspects of Linguistic studies regarding the similarity phenomenon. Our approach was developed based on the assumption that a strong support of these Linguistic theories can increase the quality of our work, in the sense that allows representing computationally the issues of importance to the similarity identification.

Similarity can be taken as a criterion for the identification of different semantic properties. Regarding paradigmatic relations, under the onomasiological point of view, the typical phenomenon evidencing similarity is the synonymy, which reflects the construction of maximum semantic identity between two distinct lexical items. The relationship of hyponymy is also commonly seen as a factor that evidences similarity of some type. In the semasiological perspective, polysemy is the phenomenon that is directly related to similarity. The identification of similarity between meanings associated with the same lexical item or to the same lexical category is considered as the primary criterion to characterize the polysemy. In both cases, the explanation for similarity, in its almost totality, revolves around the notion of metaphor (Cruse, 1995).

More elements for the characterization of similarity are found when we look at the phenomenon as a cognitive principle responsible for the construction of approximations between different entities. In this context, we can explain the similarity regarding the Gestalt principles, which define the unconscious perceptual mechanisms responsible for the construction of ’all’ or ’gestalts’ from the processing of inputs that are incomplete (Evans and Green, 2006). According to this criterion, we can find (if any) a unifying element relevant to the interpretation of the senses in comparison. This association may occur in different terms: Based on objective factors and/or subjective factors. Objective factors assume that entities perceived as similar in a scene share physical characteristics, such as size, shape or color and that they are perceived as belonging to a group (Evans and Green, 2006). Looking at the senses, this type of association requires that words denote similar entities objectively (Hirsch, 1997).

In addition to varying in terms of objectivity/subjectivity and intensity, similarity can be constructed with a lesser or greater degree of linearity or regularity. From these observations, we understand that looking at the notion of similarity as a cognitive principle responsible for the construction of sense relations of different types helps to figure out less intuitively what is intended to communicate when recognizing the existence of the similarity between two senses.

Related Works

The literature on textual entailment presents a high number of works on assessing similarity in the English language. Works dealing with the Portuguese language represent still a few sets of initiatives. We studied works in both languages and focused on the construction of contributions to the Portuguese language domain. This section, therefore, presents an overview of methods and approaches to semantic similarity detection. We selected examples of recent works that discusses in depth the classification attributes.

Among the English language dedicated works, we can highlight the experiments of Hänig *et al*. (2015) and Kashyap *et al*. (2016), in which the authors obtained good results through hybrid approaches. In Kashyap *et al*. (2016) is proposed a technique that seeks to calculate the similarity between words and sentences through the combination of Hyperspace Analog to Language (Burgess *et al*., 1998) together with similarity measures extracted from WordNet. In Hänig *et al*. (2015), named entities and temporal expressions are used, as well as a series of distance measures and manipulation of negation to assess the similarity between sentences. Additionally, in agreement with the previous resources, the author uses antonymy, hypernymy, hyponymy and synonymy contained in WordNet to compose the attributes used in a Support Vector Machine (SVM) classifier.

Xie and Hu (2017) presents an approach based on max-cosine matching for natural language inference in short sentences. In this approach, the first step involves word similarity evaluation and the next step is to represent this word pairs in order to apply a LSTM Artificial Neural Network architecture to identify the sentences similarity.

Ferreira *et al*. (2018) proposes a paraphrase identification system that represents each pair of the sentence as a combination of different similarity measures. In this case, the similarity measures used were defined considering lexical, syntactical and semantic layers. The similarity representation between the two sentences being analyzed is given as input to a machine learning algorithm that classifies these two sentences as similar or not.

Another recent line of investigation is dedicated to evaluating improvements in the sentence similarity by applying constraints in iterative process (Kajiwara *et al*., 2017). The work of Yanaka *et al*. (2017), has the objective of actuating in the aspects related to the capture of the semantics in the sentences. Therefore, the authors propose a method for determining semantic textual similarity by combining shallow features with features extracted from natural deduction proofs of bidirectional entailment relations between sentence pairs. Besides this, they applied logical semantic representations to capture deeper levels of sentence semantics.

Regarding works dedicated to the Portuguese Language, a similar set of these aspects can be identified. Barbosa *et al*. (2016) creates metrics using the Word Embeddings, Inverse Document Frequency (IDF) and use a Siamese network (Chopra *et al*., 2005) to sentence similarity classification. In Freire *et al*. (2016) is proposed a framework composed of three systems: MachineLearning, HAL e WORDNET_HAL. Respectively, the first uses word similarity through the Dice coefficient and WordNet, while the other two use a symbolic approach to calculate word similarity through Latent Semantic Analysis (LSA).

The best results for Brazilian Portuguese STS in PROPOR 2016, Hartmann (2016), describes the problem of spreading of data caused by techniques exclusively mathematics or lexicon based. It points to domain dependency caused by tools such as WordNet, which he claims to restrict the application of the method only to a given language, due to the unique language characteristics found in the resources. The author uses the word2vec technique to get word embeddings, through CBOW and SG using a 600-dimensional window and a corpus containing three million of tokens in Brazilian Portuguese, composed of texts from the Wikipedia and PLN-Br corpus of Bruckschen *et al*. (2008). The author uses the technique of Mikolov *et al*. (2013) and the similarity of the cosine between the sum of the pairs of sentence vectors for the linear regression with SVM.

Another author, Alves *et al*. (2016) brings two approaches. The first based on heuristics under semantic lexical networks for the Portuguese language. The second uses supervised automatic learning resources. Initially, the nominal, verbal and prepositional groups were counted in each one of the sentences of each pair, as well as calculating the absolute value of the difference for each type of group. With the identification of Entities mentioned and each entity type found was calculated the absolute value of the difference of the count in both sentences. The heuristic approach applies atomization and labeling semantics as pre-processing of sentences. Subsequently, the author makes use of the lemmatization through LemPort as well as REM through Apache OpenNLP. Once the sentence characteristics are obtained, nine lexical-semantic networks are used for the calculation of similarity, through the highest similarity between neighboring words of each verdict. For this, the networks are used in order to obtain five types of relations: antonyms, hypernymy, hyponymy, synonymy and the group of all other existing relationships.

Has been seen authors (Fialho *et al*., 2016) using as language resources the sentence polarity and negation. These linguistic resources are related to measures of similarity. The Probabilistic resources are restricted to TF-IDF method and the classification algorithm used was the Support Vector Machine.

Table 1 presents a summary of some aspects present in the proposed method and in the studied works. The following aspects are compared: (a) The use of resources and substitution of antonyms, hypernyms, hyponyms and synonyms; (b) relations features treated in the work, where C is the counting of relation occurrences, M is a similarity measure between the terms under relation, S is the substitution of the words found on relation and E is check if the linguistic relation exists; (c) The use of probabilistic complementary resources; (d) the language that is treated by the work. The main reason to choose these aspects is to identify and compare the use of several linguistic relations as well as probabilistic resources.

Table 1: Related works Comparison

	Antonym	Hypernym and Hyponym	Synonym	Features	Probabilistic recources	Language
Ferreira *et al*. (2016)				S/M		English
Kashyap *et al*. (2016)		X	X	E	X	English
Hänig *et al*. (2015)	X	X	X	M	X	English
Xie and Hu (2017)			X	M	X	English
Yanaka *et al*. (2017)			X	M		English
Hartmann (2016)			X	S	X	Portuguese
Alves *et al*. (2016)	X	X	X	C/M		Portuguese
Fialho *et al*. (2016)				M	X	Portuguese
Barbosa *et al*. (2016)			X	S		Portuguese
Proposed work	X	X	X	C/S	X	Portuguese

Observing the Table 1 is possible to identify that although most of the works cited use linguistic resources, the set of aspects used in this regard are frequently very small and do not fully exploit the potential of these resources in the STS task. For example, some studies present the use of probabilistic, or heuristic resources solely, to assess the similarity between sentences (Brychcín and Svoboda, 2016; Ferreira *et al*., 2016; Barbosa *et al*., 2016). It is also observed that the chosen probabilistic approach differs in the works previously mentioned and although most of them use word embeddings together with other resources, the linguistic resources are not applied largely. Therefore, in this study, we investigate the specific advantages in the use of deep linguistic relations combined with probabilistic techniques, which is an approach that is not observed in the literature.

Material and Methods

This section provides a comprehensive framework of our approach and the resources and corpora involved. Linguistic resources and probabilistic techniques are used in the proposed approach to better represent the sentences and to pursue better opportunities for similarity identification. Regarding the linguistic aspects, the concepts of antonymy, hypernymy, hyponymy and synonymy were used through the TeP (Maziero *et al*., 2008) and PULO (Simões and Guinovart, 2014) resources, which are Brazilian Portuguese versions of WordNet. Considering the probabilistic scope, the concepts of VSM, TF-IDF and PCA were explored. These techniques have an important contribution to semantic textual similarity, once they can provide more information about sentence contexts.

The VSM models were used to obtain a distributed representation for words in the sentences, where each of these present in a corpus is mapped to an attribute vector, which represents different contexts of the word and considers each of them as a point in the space of vectors. Although the high dimensionality of VSM representation could be useful in some cases, there are others where the use of PCA to promote a dimensionality reduction could improve the results and preserve the context of sentences.

Proposed Approach

The proposed approach is divided into four main tasks, which are corpus acquisition, sentence representation, similarity analysis and classification. The corpus acquisition task is designed to collect and pre-process the corpus necessary to the experiments. Sentence representation task aims to express the required elements and relations from the sentences, according to linguistic resources used. In the similarity analysis task several possible similarity measures are explored, to generate the necessary material for the sentences classification. The last task is the classification, which applies machine learning and regression models to classify the sentences pairs.

Figure 1 describes the elements and processes of our approach. Each one of the tasks can be implemented with independence regarding the resources applied. Step 1 captures texts in news websites through a Web Crawler. Subsequently, as indicated in step 2, the collected corpus is stored to be used to generate word vector representations. During step 3, operations are applied to prepare the corpus as an input to the GloVe Algorithm. At step 4, the (Pennington *et al*., 2014) algorithm generates word embeddings and stores them. Step 5 is dedicated to the pre-processing of ASSIN dataset (Fonseca *et al*., 2016), as the last step of the corpus acquisition task, which contains 10,000 pairs of sentences collected through Google News (divided equally into Brazilian Portuguese and European Portuguese). Within these, 6000 records are data for training and the others for testing. Both sets contain the similarity value between sentence pairs in a numeric interval (in this case, the interval [1, 5]).

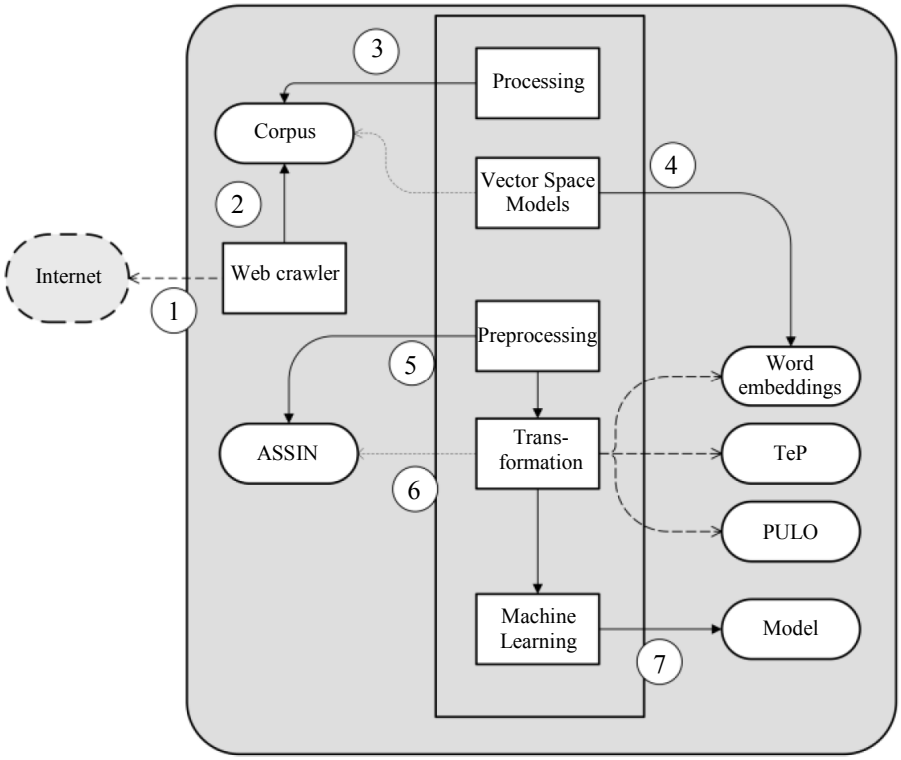

Fig. 1: Proposed methodology detail

The sentence representation task involves both steps 5 and 6. In the pre-processing, the following techniques were used: Punctuation removal, the conversion of the text into lowercase happens and likewise the removal of numerical data. In the transformation, lexical-semantic resources (TeP and PULO) are used to treat the hypernyms, hyponyms and synonymy relations. A brief example of transformation considering original sentences using the resources in the methodology is described below, considering the sentences numbered as sentence 1.1 and sentence 1.2, presented in Portuguese and English versions:

1 Portuguese

1.1 "A comissão apura denúncias de abuso e exploração sexual em meninas da comunidade quilombola"

1.2 "O grupo apura denúncias de abusos e exploração sexual de crianças da Comunidade Quilombola"

2 English

2.1 "The commission investigates allegations of sexual abuse and exploitation of girls of quilombola community"

2.2 "The group investigates allegations of abuse and sexual exploitation of children of quilombola community"

After the processing with sentence representation resources, the following sentences (sentence 2.1 and sentence 2.2) are obtained:

1. Portuguese

   1.1. "Comissao apurar denunciar abusar exploracao sexual criancas comunidade quilombola"

   1.2. "Comissao apurar denunciar abusos exploracao sexual criancas comunidade quilombola"

2. English

   2.1. "Commission to investigate denouncing abuse sexual exploitation children community quilombola"

   2.2. "Commission to report denouncing abuse sexual exploitation children community quilombola"

It stands out in the resulting phrases the influence of hypernym relations, with the words "children" and "girls". In addition, the example shows the substitution of "group" for "commission", in the case of synonymy relation.

The tasks of similarity analysis and classification are represented in step 7, in which the data resulting from the previous stps are used as an input for training and testing the machine learning algorithms where the similarity classifier models will be generated.

Corpus Acquisition Considerations

We considered two steps involving textual data used as a corpus in our approach. The first one is dedicated to obtaining and processing a large corpus of Brazilian Portuguese sentences to generate word embeddings. This word embeddings are applied in the classification task of our approach. The second step is the use of the ASSIN annotated corpus to validate our results.

Web Crawler was developed to capture Brazilian Portuguese texts in news websites, such as Google News and Wikipedia. During the collection process, to each page visited the software extract textual elements, removes HTML markups and saves the text into a file containing one paragraph per line. At the end of this process, we obtain the corpus that was used by the GloVe algorithm to produce word embeddings. The number of words collected was similar between different domains. Notwithstanding, when manually inspecting the corpus, we verified the existence of sentences containing a few words, or special strings. This type of occurrence might cause a change in the score performed by GloVe since the VSM models uses a word occurrences theory (Harris, 1954) as a premise. As these sentences frequently lack texts and present markers, or indicators of web pages solely, consequently they would not add useful information to the training of word embeddings. Hence, we performed the corpus processing for the removal of sentences composed only of numbers, or those containing less than five words.

This research used the dataset made available by the ASSIN task as the basis for the comparison of results, which belongs to the PROPOR 2016 event. We aim at identifying the semantic textual similarity and classification between pairs of short sentences made available in that dataset. According to Fonseca *et al*. (2016), the dataset was annotated by a total of 36 people, each sentence being assessed by four people. The dataset contains 10,000 pairs of sentences collected through Google News (divided equally into Brazilian Portuguese and European Portuguese). Within these, 6000 records are data for training and the others for testing. Both sets contain the similarity value between sentence pairs in a numeric interval (in this case, the interval [1, 5]). According to Fonseca *et al*. (2016), the assessment of the work submitted to the task was done through the Pearson's Correlation (PC) and the Mean Squared Error (MSE). The first one measures how linearly the result and the expected value are related, whereas the second estimates the error when classifying the correlation.

Sentence Representation Attributes

For this study, the GloVe algorithm (Pennington *et al*., 2014) was used to obtain the proposed corpus word embeddings. The GloVe was trained during 10 epochs with 6 elements in the context window, 100 co-occurrences and a learning rate of 0.15. Moreover, the size of the vectors was set at 600 positions because previously works (Pennington *et al*., 2014) showed an increased accuracy in capturing semantic textual similarity. Initially, the composition of each sentence was performed through the word embeddings correspondent to each word and, in this fashion, the matrix of contexts with the relevant words and 600 dimensions was obtained. At this point, as shown in Hartmann (2016) and Mikolov *et al*. (2013), an attribute was created through the similarity of the cosine between the sum of the matrix of contexts of each sentence. However, Hartmann (2016) states that the sum of the word embeddings matrix generates the generic representation of the sentence and does not reflect its contexts. In this sense, the PCA technique for dimensionality reduction was applied and later the calculation of the Euclidean distance between the first component of each sentence, which contains the items with greater variation in the context matrix.

Besides the attributes that make use of word embeddings (9 and 10), others eight were elaborated, hence, computing ten attributes presented in Table 2.

The first three attributes (1, 2 and 3) use lexical and semantic aspects of the sentences, obtained from PULO and TeP databases and applied for the substitution of hypernyms, hyponyms and synonyms in the sentences. The next three attributes (4,5 and 6) were obtained using word counting and the search for uni-grams, bi-grams or tri-grams in both sentences, using the WEKA tool libraries (Witten *et al*., 2016) to find compound and common terms with one occurrence at most; in an empirical analysis, we noticed that many of the named entities were scored in this attribute.

Table 2: List of attributes used in the experiments

Index	Attribute
1	Synonyms substitution in sentences
2	Hyponyms and hypernyms substitution in sentences
3	Antonyms score evaluation in sentences
4	Different words proportion in sentences
5	Common n-grams proportion in sentences
6	Words in common proportion between sentences
7	Sentence size penalization coefficient
8	Cosine similarity between the sum of word embeddings
9	Euclidean distance between the first main component of each sentence
10	Cosine similarity between the TF-IDF vectors of each sentence

The equation indicated by Ferreira *et al*. (2016) for calculating the penalization of sentences with different sizes was used as the attribute 7; however, the similarity value used in the author's equation was replaced by the arithmetic mean of the word embeddings and TF-IDF similarities. Attributes 8, 9 and 10 were obtained using GloVe and TF-IDF technique with the use of both original sentences and the variation obtained through substitutions. The option to perform the modifications in the sentences is an attempt to reduce the sparsity of the data since both of word embeddings and TF-IDF approaches use in its calculation the score of words or contexts shared between the sentences. Therefore, the more elements and contexts shared between texts, the greater the correlation between them.

Experimental Results

We obtained the experimental results using the same corpus which related works applied, the ASSIN dataset, therefore enabling an adequate comparison. A series of 49.726 experiments was initially accomplished. In these experiments were generated several combinations of the attributes shown in Table 3 and the same proceeding was repeated, allowing that each combination outcome could be compared with any other. We used the SVM, Artificial Neural Networks (ANN) and Generalized Linear Models (GLM) to generate the linear regression models, which are a well known and respected machine learning algorithms on STS area. The use of selected algorithms allow a state-of-art reproducibility and allow a comparison through their algorithms. In addition, we also made experiments with normalization techniques such as Max-Min and Z-score.

The best results obtained, along with some specific combinations of interest for the overall analysis of the classification process, are described in Table 3. As we can see in Table 3, the results achieved with only word embeddings feature were not sufficient for a good SVM, ANN or GLM performance, which is also observed in (Hartmann, 2016). We understand that the use of PCA instead of a sum to obtain similarity from word embeddings maintains the unsatisfactory performance because the reduction of dimensionality might lead to the loss of sentences peculiarities and context. Therefore, as depicted in Table 3, the results obtained using exclusively probabilistic resources are not satisfactory and are very far from the best. These results are exhibited in the first three lines, with the attributes indexes 8, 9 and 10. When using only the antonymy relation, the classification metrics are even worsting that with the probabilistic resources, as described in the fourth line of Table 3, with the index attribute 3. Combining few probabilistic attributes and metrics that actuate only in the syntax scope, as line five, which contains the results of the combination of the attributes indexes 7 and 9, reveal poor results as well. These first five lines of Table 3 are indicated here to allow comparison with the improvements in results when incorporating more attributes to the classification process.

When the set of attributes includes a combination of linguistic relations and probabilistic elements, the level of the obtained results is improved significantly. In the Table 3 this can be seen comparing the level of results obtained in the first five lines, in which the best value to the Pearson's Correlation (PC) is 0.4448 and the lowest Mean Squared Error (MSE) is 0.6847 in front of the last lines where the results are improved to the better PC equal to 0.6626 and the better MSE equal to 0.4302, which correspond to a better level of results. The mean of the results in the more complete sets of attributes, comprising the combination of linguistic and probabilistic elements, is equal to 0.6364 for the PC, while the mean of the results with the exclusive sets (only linguistic or only probabilistic resources) is equal to 0.2672 for PC.

Therefore, the experiments indicated in Table 3 present better results when a great set of attributes is used. The final four lines of Table 3 presents the best results and also represents the biggest sets of attributes. The differences observed in these four sets are associated with the representativity in the attributes used, indicating that the linguistic relations such as synonymy and hyponymy play an important role in the results.

Table 3: Results of the proposed approach using different sets of features

Index	Attributes (Index in Table 2)*	Pearson's Correlation	Mean Squared Error
1	8	0.3165	0.6847
2	9	0.2641	0.7226
3	10	0.4448	0.6174
4	3	0.0355	0.7754
5	7,9	0.2672	0.7087
6	5,6,8,10	0.6364	0.4535
7	4,5,6,8,10	0.5782	0.5102
8	5,6,9,10	0.6357	0.4543
9	4,5,6,9,10	0.6343	0.4622
10	5,6,10	0.6160	0.4790
11	1,2,3,4,5,6,8,10	0.6394	0.4499
12	1,2,5,6,8,10	0.6625	0.4302
13	1,2,4,6,7,10	0.6625	0.4303
14	1,2,4,5,6,8,10	**0.6626**	**0.4304**

Table 4: Results comparison with state of the art

	Attributes	Pearson's Correlation	Mean Squared Error
Proposed approach	Word embeddings with PCA	0.30	0.69
	Sum of word embeddings	0.30	0.68
	TF-IDF	0.44	0.61
	Word embeddings with PCA and TF-IDF	0.46	0.59
	Sum of word embeddings and TF-IDF	0.55	0.52
	Combination with the best results of Table 3*	0.66	0.43
Hartmann (2016)	Sum of word embeddings	0.58	0.50
	TF-IDF	0.68	0.41
	Sum of word embeddings and TF-IDF	0.70	**0.38**
	Soft TF-IDF		
Fialho *et al*. (2016)	Similarity between words	**0.73**	0.63
	Ngram overlap		
Alves *et al.* (2016)	ASAPP	0.65	0.44
	Reciclagem	0.59	1.31

This can be observed when comparing the improvements obtained from the experiment in line 6 to the experiment in line 14, once in this pair of attributes the main change is the inclusion of synonym and hypernym and hyponym information.

In the experiments, due to the required processing capacity, a server with two 1GHz version 4 E5-2620 processors, 128 gigabyte RDIMM (2400 MT/s) and 16 megabyte Matrox G200eR2 video card was used. Furthermore, R Studio Server environment had as main tasks the generation of machine learning models, preprocessing and data transformation was also configured.

Discussion

Analyzing the results of Table 3, we can observe that the higher PC and the smaller MSE were achieved through experiments using linguistic properties, such as antonymy and hyponymy relations. However, when analyzing the number of antonyms by pairs in the dataset used, we can notice that rarely one or more antonyms in the same sentence were identified. This might be justified by the low volume of records in the data. These results made it difficult to use linguistic relations and showed some repercussions on the performance of the technique to use the attributes of antonymy and hyponymy. Moreover, we can observe the low performance of the size penalization attribute due to the difference in size between sentences.

The Table 4 shows the best results in state of the art for the assessment of semantic textual similarity, which may be compared with the current work through the ASSIN dataset. As it is possible to observe in the above table, the results obtained in this study score in the same level of the best results for PC or MSE, when compared to the related work that used the same dataset. We emphasize that the number of tokens in the corpus used to obtain word embeddings in our experiment was extremely low, representing a very small percentage of the word embeddings corpus size related to other works. In Hartmann (2016), the author used a corpus containing about 300 million tokens collected from the G1 and Wikipedia websites, besides the additional use of the Bruckschen *et al*. (2008) corpus to VSM algorithm training and word embeddings generation. In contrast, within our study, only 1,584,492 tokens were used for word embeddings training, which corresponds to about 0.05% of what Hartmann (2016) used.

The best results achieved in this research it was obtained with ANN model using a set of 15 neurons, 0.12 of learning rate, sigmoid activation function and 1000 epochs of training time besides the use of the substitutions of synonymy and hyponymy relations in the sentences. This feature does not affect the meaning of the sentence and it allows direct comparison between occurrences of common words in both. The approach described maximized the results of the TF-IDF technique, adding to it a crucial role in achieving similarity. Notwithstanding the preceding, it can be observed that although the antonyms metrics does not correlate with the expected value of similarity (p>0.05), it showed a good performance when used along with other attributes (as we can see in Table 3).

Despite the attempt to use PCA to reduce the dimensionality of word embeddings and preserve their linearity, overall results obtained with the use of this technique were worst than the simple sum of word embeddings suggest by Mikolov *et al*. (2013). Therefore, we figure that PCA suppresses some of the contexts and cannot capture all the semantic information hidden in embeddings space.

All our experiments used a set of many combinations of the attributes shown in Table 2, including normalization techniques. In our experience, the normalization of attributes got their best results with ANN but when we consider only SVM and GLM algorithms, the normalized attributes performed closely to the standard.

Table 5: Related works comparison

	Pearson's Correlation		Mean Squared Error	
	Mean	Standard deviation	Mean	Standard deviation
Standard dataset	0.5864	0.0649	0.5020	0.0587
Synonym generalization	0.5939	0.0674	0.5197	0.0837
Synonym, hyponym and hypernym generalization	0.5969	0.0676	0.5163	0.0828

As we can see in Table 5, the generalization of both sentences words by synonym relation improved the results in PC, showing that results could be close to the expected, but the MSE raise rates shows that there were more errors in measuring similarity.

The comparison of results obtained by machine learning algorithms shown in Fig. 2 demonstrate the ANN better performance over SVM and GLM. Although the ANN obtained the best results in both of CP and MSE, the SVM had a close median with other algorithms and shown less variation in his results. Therefore, we highlight that the use of ANN appears to be promising, but it needs more experiments with larger set of parameters once our analysis used only a set of three epochs (500, 700 and 1000), seven hidden neurons (5, 7, 10, 15, 19, 22 and 30) and three learning rates (0.012, 0.01 and 0.12).

We can highlight some aspects that represent limitations in this research and will be addressed in the future works. At first, there are some resources to support Brazilian Portuguese Natural Language Processing that were not applied in these experiments but could be important to improve the quality of linguistic resources. One of such resources is the Brazilian WordNet OpenWordnet (Paiva *et al*., 2012), that is currently being studied to be part of next experiments. Another limitation highlighted is the size of the corpus used to treat the word embeddings. We believe that the probabilistic resource can represent a good possibility to achieve better results, but so far the size of the corpus is extremely small when compared with other works. As a final point, we deal only with a limited set of linguistic relations due that these choose relations are the ones described by the literature as the most promising to be used in this task, we should improve the set of linguistic relations in order to implement new experiments with a broader set of options.

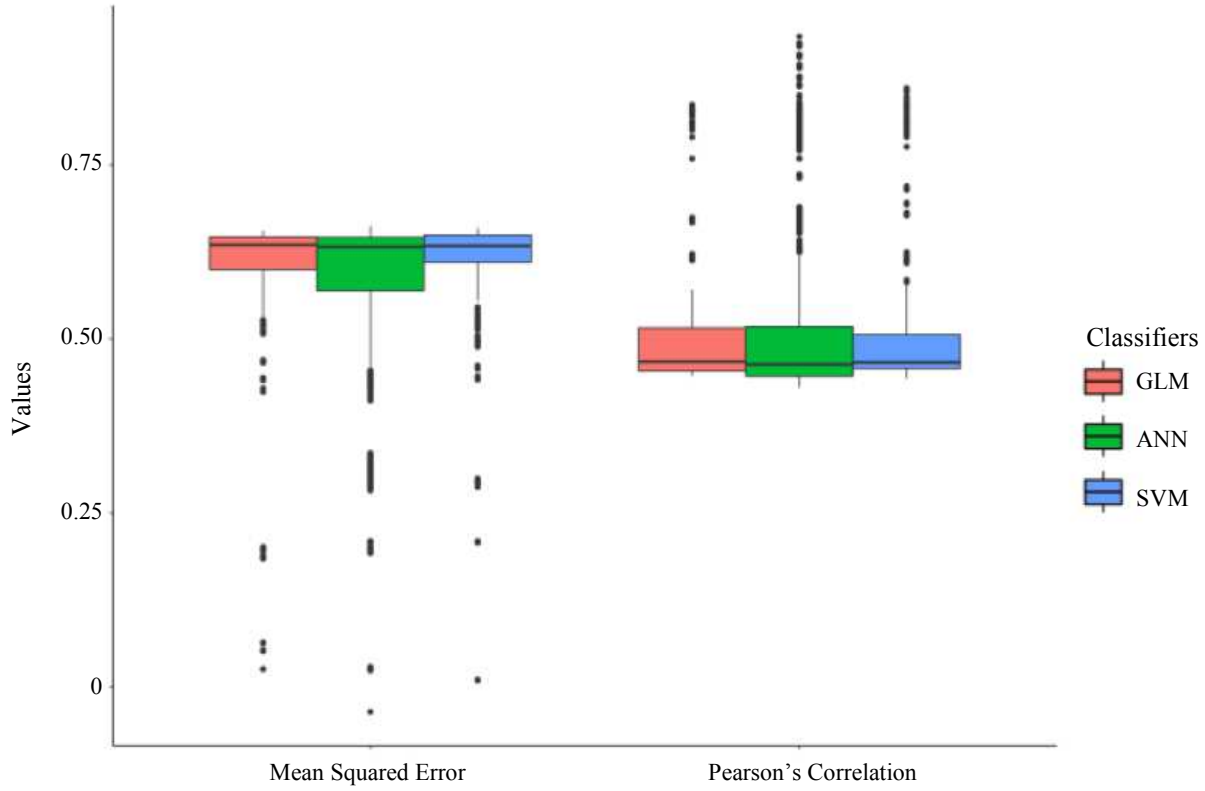

Fig. 2: Machine learning algorithms comparison

Conclusion

In this study, we presented a hybrid approach to semantic textual similarity identification between short sentences. To do so, we have applied the concepts of VSM, TF-IDF and PCA and likewise the linguistic relations of antonyms, hypernymy, hyponymy and synonymy. This approach allows us to obtain a set of different attributes combination, used then in experiments with SVM, ANN and GLM classifiers. Our bests results were obtained with the combination of attributes which incorporate linguistic and probabilistic aspects. This was observed with all the different classifiers used. As for the classifiers, the best results were obtained with the ANN experiments.

The number of tokens in the corpus used to train the Vector Space Models with GloVe algorithm may have directly influenced the scale of the proximity of words and therefore, the similarity of sentences. As mentioned, even with a limited word embeddings corpus for training the classifiers, results equivalent to state of the art were possible to achieve through the use of linguistic properties.

The results achieved show that the use of hypernymy and hyponym relations alone did not present sufficient information for a better identification of similarity. However, the use of them as attributes was supportive in the generalization of sentence terms. Hence, they brought better results when used together with techniques such as TF-IDF and word embeddings, which are dependent on the occurrence of similar words or similar contexts between sentences.

To promote the reproducibility of this research, all the tools used were open-source and all the resources created from the development of this work were made available on the web, as well as the source code containing the procedures performed to obtain the results achieved in this study. Consequently, the present work has contributed to the NLP field through the publication of a non-annotated corpus word embeddings in Brazilian Portuguese containing 1,584,492 tokens, obtained through GloVe and a web service for querying lexical-semantic information.

We highlight as main contributions of this work: (i) An applicable approach for measuring the semantic textual similarity between short sentences in Brazilian Portuguese; (ii) results of experiments with reduction of dimensionality; and (iii) a methodology for the availability of resources on the web. In addition, we can highlight the application of size penalization between sentences, as well as the use of hypernym, hyponymy and synonymy relations in support of vector space models' representation for similarity analysis.

In future works, despite the high hardware requirements for solutions involving deep learning, we intend to assess the performance of the SVM, ANN and GLM algorithms compared to LSTM networks. They are already seen in Mueller and Thyagarajan (2016), who show their ability to handle complex semantic representations and modeling to calculate sentence similarity. We also intend to apply the same approach in the English language. Finally, we intend to explore resources such as OpenWordnet and a broader set of linguistic relations.

Acknowledgment

This research paper was made possible through the help and support from my guide, Dr. Sandro José Rigo. First and foremost, thanks for his most support and encouragement.

Author's Contributions

Allan de Barcelos Silva: Participated in all experiments, coordinated the data-analysis and contributed to the writing of the manuscript.

Sandro José Rigo: Designed the research plan and organized the study. He also contributed in drafting the article and reviewing it critically for significant intellectual content. In addition, he gave the final approval of the version to be submitted.

Ethics

We testify that this research paper submitted to the Journal of Computer Science has not been published elsewhere and that has no ethical issues. All authors have been personally and actively involved in substantive work leading to the manuscript and will hold themselves jointly and individually responsible for its content.

References

Al-Alwani, A., 2015. Improving email response in an email management system using natural language processing based probabilistic methods. J. Comput. Sci., 11: 109-119. DOI: 10.3844/jcssp.2015.109.119

Alves, A.O., R. Rodrigues and H.G. Oliveira, 2016. ASAPP: Alinhamento Semântico Automático de Palavras aplicado ao Português. Linguamática.

Barbosa, L., P. Cavalin, V. Guimarães and M. Kormaksson, 2016. Blue man group at ASSIN: Using distributed representations for semantic similarity and entailment recognition. Linguamática, 82: 15-22.

Berrahou, S.L., P. Buche, J. Dibie and M. Roche, 2017. Xart: Discovery of correlated arguments of n-ary relations in text. Expert Syst. Applic., 73: 115-124. DOI: 10.1016/j.eswa.2016.12.028

Bruckschen, M., F. Muniz, J. Guilherme, C. De Souza and J.T. Fuchs *et al*., 2008. Anotação linguística em XML do Corpus PLN-BR Tech. Rep. Universidade de São Paulo, São Carlos, Brasil.

Brychcín, T. and L. Svoboda, 2016. UWB at SemEval-2016 task 1: Semantic textual similarity using lexical, syntactic and semantic information. Proceedings of the International Workshop on Semantic Evaluation, Jun. 16-17, Association for Computational Linguistics, San Diego, California, pp: 588-594.

Burgess, C., K. Livesay and K. Lund, 1998. Explorations in context space: Words, sentences, discourse. Discourse Processes, 252: 211-257. DOI: 10.1080/01638539809545027.

Chen, F., C. Lu, H. Wu and M. Li, 2017. A semantic similarity measure integrating multiple conceptual relationships for web service discovery. Expert Syst. Applic., 67: 19-31. DOI: 10.1016/j.eswa.2016.09.028

Chopra, S., R. Hadsell and Y. LeCun, 2005. Learning a similiarty metric discriminatively, with application to face verification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 20-25, IEEE Xplore Press, San Diego, CA, USA, pp: 539-546. DOI: 10.1109/CVPR.2005.202

Cruse, D.A., 1995. Polysemy and Related Phenomena from a Cognitive Linguistic Viewpoint. In: Computational Lexical Semantics, Saint-Dizier, P. and E. Viegas (Eds.), Cambridge University Press, pp: 33-49.

Evans, V. and M. Green, 2006. Cognitive Linguistics: An Introduction. 1st Edn., L. Erlbaum, Mahwah, ISBN-10: 0805860142, pp: 830.

Ferreira, R., G.D.C. Cavalcanti, F. Freitas, R.D. Lins and S.J. Simske and M. Riss, 2018. Combining sentence similarities measures to identify paraphrases. Comput. Speech Lang., 47: 59-73. DOI: 10.1016/j.csl.2017.07.002

Ferreira, R., R.D. Lins, S.J. Simske, F. Freitas and M. Riss, 2016. Assessing sentence similarity through lexical, syntactic and semantic analysis. Comput. Speech Lang., 39: 1-28. DOI: 10.1016/j.csl.2016.01.003

Fialho, P., R. Marques, B. Martins, L. Coheur and P. Quaresma, 2016. INESC-ID@ASSIN: Medição de similaridade semântica e reconhecimento de inferência textual. Linguamática, 82: 33-42.

Fonseca, E.R., L. Borges, D. Santos, M. Criscuolo and S.M. Aluísio, 2016. Visão geral da avaliação de similaridade semântica e inferência textual. Linguamática, 8: 15-22.

Freire, J., V. Pinheiro and D. Feitosa, 2016. LEC_UNIFOR no ASSIN: FlexSTS um framework para similaridade semântica textual. Linguamática, 8: 23-31.

Gomaa, W. and A. Fahmy, 2013. A survey of text similarity approaches. Int. J. Comput. Applic., 68: 13-18. DOI: 10.5120/11638-7118

Hänig, C., R. Remus, X. De and L. Puente, 2015. ExB themis: Extensive feature extraction from word alignments for semantic textual similarity. Proceedings of the International Workshop on Semantic Evaluation Valuation, Jun. 4-5, Association for Computational Linguistics, Denver, USA, pp: 264-268.

Harris, Z.S., 1954. Distributional structure. WORD, 10: 146-162. DOI: 10.1080/00437956.1954.11659520

Hartmann, N.S., 2016. Solo queue at ASSIN: Combinando abordagens tradicionais e emergentes. Linguamática, 82: 59-64.

Hirsch, E., 1997. Dividing Reality. 1st Edn., Oxford University Press, New York, ISBN-10: 0195111427, pp: 247.

Kajiwara, T., D. Bollegala, Y. Yoshida and K.I. Kawarabayashi, 2017. An iterative approach for the global estimation of sentence similarity. PloS ONE, 12: e0180885-e0180885. DOI: 10.1371/journal.pone.0180885

Kashyap, A., L. Han, R. Yus, J. Sleeman and T. Satyapanich *et al*., 2016. Robust semantic text similarity using LSA, machine learning and linguistic resources. Language Resources Evaluat., 50: 125-161. DOI: 10.1007/s10579- 015-9319-2

Maziero, E.G., T.A.S. Pardo, A. Di Felippo and B.C. Dias-da Silva, 2008. A base de dados lexical e a interface web do TeP 2.0: Thesaurus eletrônico para o Português do Brasil. Proceedings of the 14th Brazilian Symposium on Multimedia and the Web, Oct. 26-29, ACM Press, Vila Velha, Espírito Santo, Brazil, pp: 390-392. DOI: 10.1145/1809980.1810076

Mikolov, T., I. Sutskever, K. Chen, G. Corrado and J. Dean, 2013. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems, Dec. 05-10, Curran Associates Inc., Lake Tahoe, Nevada, pp: 3111-3119.

Mueller, J. and A. Thyagarajan, 2016. Siamese recurrent architectures for learning sentence similarity. Proceedings of the 13th AAAI Conference on Artificial Intelligence, Feb. 12-17, AAAI Press, Phoenix, USA, pp: 2786-2792.

Paiva, V., A. Rademaker and G. Melo, 2012. OpenWordNet-PT: An open Brazilian wordnet for reasoning. Proceedings of the International Conference on Computational Linguistics, (CCl' 12), Organizing Committee, Mumbai, India, pp: 353-360.

Pennington, J., R. Socher and C. Manning, 2014. Glove: Global vectors for word representation. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Oct. 25-29, Association for Computational Linguistics, Stroudsburg, USA, pp: 1532-1543. DOI: 10.3115/v1/D14-1162

Pradhan, N., M. Gyanchandani and R. Wadhvani, 2015. A review on text similarity technique used in IR and its application. Int. J. Comput. Applic., 1209: 29-34. DOI: 10.5120/21257-4109

Simões, A. and X.G. Guinovart, 2014. Bootstrapping a Portuguese WordNet from Galician, Spanish and English Wordnets. Proceedings of the International Conference of Advances in Speech and Language Technologies for Iberian Languages, Nov. 19-21, Springer International Publishing, Cham, Germany, pp: 239-248. DOI: 10.1007/978-3-319-13623-3_25

Witten, I.H., E. Frank, M.A. Hall and C.J. Pal, 2016. Data Mining: Practical Machine Learning Tools and Techniques. 4th Edn., Elsevier Science & Technology, ISBN-10: 0128042915, pp: 621.

Xie, Z. and J. Hu, 2017. Max-cosine matching based neural models for recognizing textual entailment. Proceedings of the International Conference on Database Systems for Advanced Applications, (SAA' 17), pp: 295-308.

Yanaka, H., K. Mineshima, P. Martinez-Gomez and D. Bekki, 2017. Determining semantic textual similarity using natural deduction proofs. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Sept. 7-11, Association for Computational Linguistics, Copenhagen, Denmark, pp: 681-691.

Yousif, S., V. Samawi, I. Elkabani and R. Zantout, 2015. Enhancement of Arabic text classification using semantic relations of Arabic WordNet. J. Comput. Sci., 11: 498-509. DOI: 10.3844/jcssp.2015.498.509