# ИНФОРМАТИКА

UDC 004.912
MSC 68T50

## Semantic Textual Similarity on Brazilian Portuguese: An approach based on language-mixture models

*A. Silva*[1], *A. Lozkins*[2], *L. R. Bertoldi*[1], *S. Rigo*[1], *V. M. Bure*[2]

[1] University of Vale do Rio dos Sinos, 950, Av. Unisinos, São Leopoldo, RS, 93020-190, Brazil
[2] St. Petersburg State University, 7–9, Universitetskaya nab., St. Petersburg, 199034, Russian Federation

**For citation:** Silva A., Lozkins A., Bertoldi L. R., Rigo S., Bure V. M. Semantic Textual Similarity on Brazilian Portuguese: An approach based on language-mixture models. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2019, vol. 15, iss. 2, pp. 235–244. https://doi.org/10.21638/11702/spbu10.2019.207

The literature describes the Semantic Textual Similarity (STS) area as a fundamental part of many Natural Language Processing (NLP) tasks. The STS approaches are dependent on the availability of lexical-semantic resources. There are several efforts to improve the lexical-semantics resources for the English language, and the state-of-art report a large amount of application for this language. Brazilian Portuguese linguistics resources, when compared with English ones, do not have the same availability regarding relation and contents, generation a loss of precision in STS tasks. Therefore, the current work presents an approach that combines Brazilian Portuguese and English lexical-semantics ontology resources to reach all potential of both language linguistic relations, to generate a language-mixture model to measure STS. We evaluated the proposed approach with a well-known and respected Brazilian Portuguese STS dataset, which brought to light some considerations about mixture models and their relations with ontology language semantics.

*Keywords*: Semantic Textual Similarity, natural language processing, computational linguistics, ontologies.

**1. Introduction.** One of the areas of Natural language processing (NLP), the task of assessing the Semantic Textual Similarity (STS) is one of the challenges in NLP, which plays an increasingly important role in related applications. The STS is a fundamental part of techniques and approaches in several areas, such as information retrieval, text classification, document clustering, applications in the areas of translation, among others [1, 2]. NLP is a very mature discipline that uses shared tasks to improve the state-of-

the-art of well defined tasks. In Semantic Evaluation (SemEval), STS is one of the tasks that has received a lot of attention [2–6]. Also, events such as International Workshop on SemEval and the International conference on the Computational Processing of Portuguese (PROPOR), which have specific tasks to measure semantic similarity between sentences, are gaining in popularity and promoting the development of a host of other applications.

Meanwhile, as appointed by works [7, 8], the literature on textual entailment presents a considerable amount of research on assessing similarity in the English language. Works dealing with the Portuguese language represent still a few sets of initiatives. Some works include a broad set of elements, representing lexical, syntactic and semantic dimensions [1, 9–12]. Some other approaches include natural deduction proofs to identify bidirectional entailment relations between sentence pairs [13], and evaluating improvements in the sentence similarity identification by applying constraints in iterative process [14].

Although a crescent number of work in English STS literature make use of resources such as WordNet*, FrameNet** and VerbNet*** for integrating some linguistic relationships to the STS process [4, 12, 15, 16]. Is already known that the available Portuguese linguistics resources do not have all relations and contents than the sibling in the English language. To workaround this issue, the proposed approach makes use of machine translate resources, and lexical-semantics resources to use all potential of Portuguese and English linguistic relations on sentences.

We assessed the proposed approach with a dataset made available in the PROPOR, which is a well-known and respected event in the Brazilian Portuguese STS research community. The achieved results appear among the best in the literature of STS for Brazilian Portuguese. One important aspect to highlight is that, although this approach does not overcome current state-of-art results for Portuguese STS, our experiments showed that combine English resources to deal with limited language resources insert more noise than help classifiers to estimate the similarities. This is considered as an indication of the disadvantages of using language mixture resources to obtain linguistic aspects from sentences. Moreover, our experiments show that the use of linguistic relations combined with Vector Space Models (VSM) techniques scored worst results than state-of-art for only one of the languages.

The structure of this paper is the following. Section 2 describes related work. In section 3 the adopted approach is presented. Section 4 describes the simulation study. In section 5 we present the obtained results. Finally, the conclusions are presented in section 6.

**2. Related works.** The literature on STS presents few works on assessing similarity in the Brazilian Portuguese language. A lot of work has been done, some of they used only linguistic features [12, 17] when others used probabilistic techniques [5, 7, 18].

Some linguistic-based approaches like [12], represent each pair of the sentence as a combination of different similarity measures. According to the author, the similarity measures used were defined considering lexical, syntactical and semantic layers. Instead of him, [17] proposed an heuristics-based approach under semantic lexical networks for the Portuguese language and another that uses supervised automatic learning resources. According to the author, he counted the nominal, verbal and prepositional groups were in each one of the sentences of each pair besides calculating both the Named Entity Recognition (NER), and the absolute value of the difference for each type of group. In

* Available at: https://wordnet.princeton.edu (accessed: 08.01.2019).

** Available at: https://framenet.icsi.berkeley.edu/fndrupal (accessed: 08.01.2019).

*** Available at: https://verbs.colorado.edu/ mpalmer/projects/verbnet.html (accessed: 08.01.2019).

addition, [17] also used nine lexical-semantic networks in order to obtain five types of relations: antonyms, hypernym, hyponymy, synonymy and the group of all other existing relationships.

Contrary to other approaches, [5] describes the problem of spreading of data caused by techniques exclusively mathematics or lexicon based. The author used a Word2Vec-based technique to get word embeddings, and [19] technique's to measure the similarity of the sentences through cosine distance between the sum of the pairs of sentence vectors as input to Support Vector Machine (SVM) algorithm. Another probabilistic approach was done by [20] that has used as language resources the polarity and negation of the sentence, which are linguistic resources related to textual similarity.

In his more recent STS approach, [7] presents the training of word embeddings using different windows size (50, 100, 300, 600 and 100) through four techniques (FastText, GloVe, Wang2Vec, Word2Vec). Their results through intrinsic and extrinsic evaluations were not aligned with each other, contrary to what was expected by the author. GloVe produced the best results for syntactic and semantic analogies, and worst, along with FastText, for both POS marking and sentence similarity. The results were aligned with those of [21], who suggested that word analogies are not appropriate for evaluating word embedding. The vectors of Wang2Vec have performed very well in our assessments, indicating that they can be useful for a variety of NLP tasks.

The current owner of Brazilian Portuguese STS state-of-art result in the ASSIN dataset, [18] proposed a hybrid approach to measure the semantic similarity between sentences. According to the author, his technique overcomes the problem of the meaning of sentences by combining TF-IDF techniques, sentence size, and similarity of word embeddings obtained using both matrix similarity and binary matrix techniques. The [18] calculated TF-IDF in the same way to that was done by [5], where the author makes use of the technique in conjunction with both stemming and expansion of the synonymous relations. According to the word embeddings matrix similarity technique, which is quite similar to that proposed by [4], the author calculates the similarity value between each word of the two sentences through Word2Vec and then removes the terms that have the highest values. Since no more words are left, the mean of the highest similarity values obtained between sentences is then averaged.

**3. Proposed approach.** The general steps of the method is presented in the current section. The proposed approach assumes that there is the basic language for which the initially textual units' contents are compared (similarity estimation) and the same textual units' contents in a secondary language. The introduction of the additional languages aims at using the advantages of semantic connections, word meaning, lexicon, and aspects of different languages to improve the text similarities estimation model on the basic language. The relationship between languages occurs through the use of machine translation algorithms, excluding cases then it is error prone or even impossible sometimes due to intrinsic differences in languages.

Current work addresses the following problem: there are similarities between levels for the same sentence pair calculated by using the STS models of different languages, how the final similarity level should be estimated? We propose the regression model as a way to combine similarity scores obtained for each language, i. e. to produce the goal similarities of sentence pairs on the basic language.

Let denote the indices set $\{0, \ldots, N\}$ of languages to be used. The similarities for language $i$ is the row vector $sim_i$, where the column $j$ of the $sim_i$ corresponds to the pair of sentences in the data set. The matrix $sim = [sim_0; \ldots; sim_N]$ represents all similarities

for sentence pairs (columns) for all languages (rows), the sentences pairs for each language are denoted by $pair_{ij}$ for $j \in 1, \ldots, M$, where the $M$ is a number of sentence pairs.

**Algorithm:** ***Textual similarities consolidation model***

**Step 1.** Sentence pairs translation $T_i(pair_{0j}) = pair_{ij}$, where the $T_i(x)$ is machine translation model from language 0 to language $i$ (there are no restrictions for translation approaches, the human translation is possible to use as well).

**Step 2.** Similarities estimation for each language $STS_i(pair_{ij}) = sim_{ij}$, where $STS_i(x)$ is STS model for language $i$.

**Step 3.** Similarities consolidation model: $Reg(sim) = finalSim$, there $Reg(x)$ is the regression model, $finalSim$ is vector of goal similarities.

The different languages make it own influence for final similarities which are characterized by regression coefficients. If in the training process of $STS_i$ model phase the estimated similarities $sim_i$ is close to real similarities then language $i$ would have high weight in the regression model (excluding situation in which $\exists i, j \in \{k\}_0^N, i \neq j : sim_i - sim_j \approx 0$, because of the high correlation level).

The regression model is chosen as consolidation model for a few reasons. The regression model represents similarities of different languages as independent values. This property helps to identify each languages utility. Therefore, the precise STS model's (the $sim_i$ value is closer to real similarities then $sim_0$) contribution in the final similarities will be higher. The regression model does not need the information about the lexical and semantic connections between all languages, but instead, all information is encapsulated into the STS model, and the training process of STS and regression model could be carried out separately. The separated training process has a negative side, the textual connections between different languages are lost. This information could be important for the STS model, but the training process becomes significantly more complicated. Instead of training $N + 1$ separate models, the single model with all sentences and corpus of $N + 1$ languages should be treated.

**4. Simulation study.** The ASSIN dataset contains 5.000 pairs of sentences on Brazilian Portuguese language with corresponding human estimated similarity levels.

The considered dataset* contains 10.000 pairs of sentences collected through Google News (divided equally into Brazilian Portuguese and European Portuguese). Within these, 5.000 records are for training and the others for testing. Once the proposed work is aimed for Brazilian Portuguese, only respective entries on dataset were used. With this in mind, the two languages model IS presented in this section. As previously stated, the primary language is Brazilian Portuguese, the secondary language is English and $N = 1$.

The yandex.translate service API** is used for machine translation Portuguese—English for each pair of sentences. The yandex.Catboost and scikit-learn Python libraries are used for regression models. The data translation scheme is shown in Figure 1.

The indices set of languages is $\{0, 1\}$, where 0 corresponds to Brazilian Portuguese language and 1 to English language. The $pair^1$, $pair^2$ and $pair^3$ is the sentence pairs matrices $(pair_0^k, T_1(pair_0^k)),\ \forall k = 1, \ldots, 3$, correspondingly for each data subsets, $pair_1^k = T_1(pair_0^k), \forall k = 1, \ldots, 3$. The target values of similarities lets denote by $tar^1$, $tar^2$ and $tar^3$. The $tar^k$, $\forall k = 1, \ldots, 3$, are assumed to be the same for Portuguese and English languages.

* Available at: http://nilc.icmc.usp.br/assin/ (accessed: 08.01.2019).

** Available at: https://yandex.ru/legal/translate_api/ (accessed: 08.01.2019).

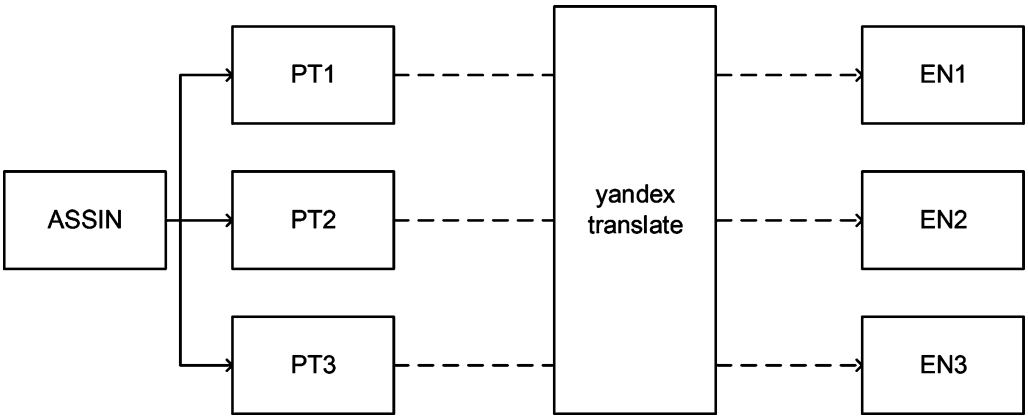


*Figure 1.* Translation step

Following the notation introduced in the previous section the estimated similarity row vectors are $STS_0(pair_0^k) = sim_0^k$ and $STS_1(pair_1^k) = sim_1^k \ \ \forall k = 1, \ldots, 3$.

The authors use the following algorithm of training and testing models:

**Step 1.** The $STS_i(pair_i^1) \ \ \forall i = 0, \ldots, 1$ models training on PT1, EN1 on Figure 1 and $tar^1$ date correspondingly.

**Step 2.** The similarities estimation for the second (PT2 and EN2 on Figure 1) and the third (PT3 and EN3) subsets of data: $STS_i(pair_i^k) = sim_i^k \ \ \forall i = 0, \ldots, 1, \ k = 2, \ldots, 3$.

**Step 3.** Regression model training on the second (PT2 and EN2 on Figure 1) subset of data and $tar^2$: $Reg(sim^2) = finalSim^2$, where $sim^2 = (sim_0^2, sim_1^2)$.

**Step 4.** The prediction of the final similarities: $Reg(sim^3) = finalSim^3$. The statistics calculation using $finalSim^3$ and $tar^3$: Mean Squared Error, Pearson's Correlation (PC) and Spearman's Correlation (SC).

The $sim_0^2$ and $sim_1^2$ correspond to AP1 and AE1 on Figure 2, the $sim_0^3$ and $sim_1^3$ correspond to AP2 and AE2 on Figure 2, the $finalSim^3$ corresponds to A3 on Figure 2. The results of the algorithm application are presented in the section 5.

**5. Results.** For this study, we did experiment with all the attributes proposed by [5], [7] and [8]. As stated by authors, were used an attribute through the similarity of the cosine between the sum of the word vectors of each sentence, and another obtained with Principal Component Analysis (PCA) technique with the calculation of the Euclidean distance between the first component of each sentence, which contains the items with greater variation in the embeddings matrix. The next attributes were obtained using OpenWordNet-pT [22] and Princeton WordNet [23], in order to obtain lexical and semantic aspects of the sentences. These resources were used to antonym relation count besides hypernyms, and synonyms generalization in the sentences [8]. As [8], we also the normalized word-overlap, and inverse word-overlap between sentences (we also consider $n$-gram technique). Contrary to the author, we do not use the penalization coefficient due to not impact in similarity estimation. Following, we also obtained an attribute from the TF-IDF.

The best results obtained, along with some specific combinations of interest for the overall analysis of the classification process, are described along this section. In our results, we tried a large set of experiments with attribute normalization through MaxMin,

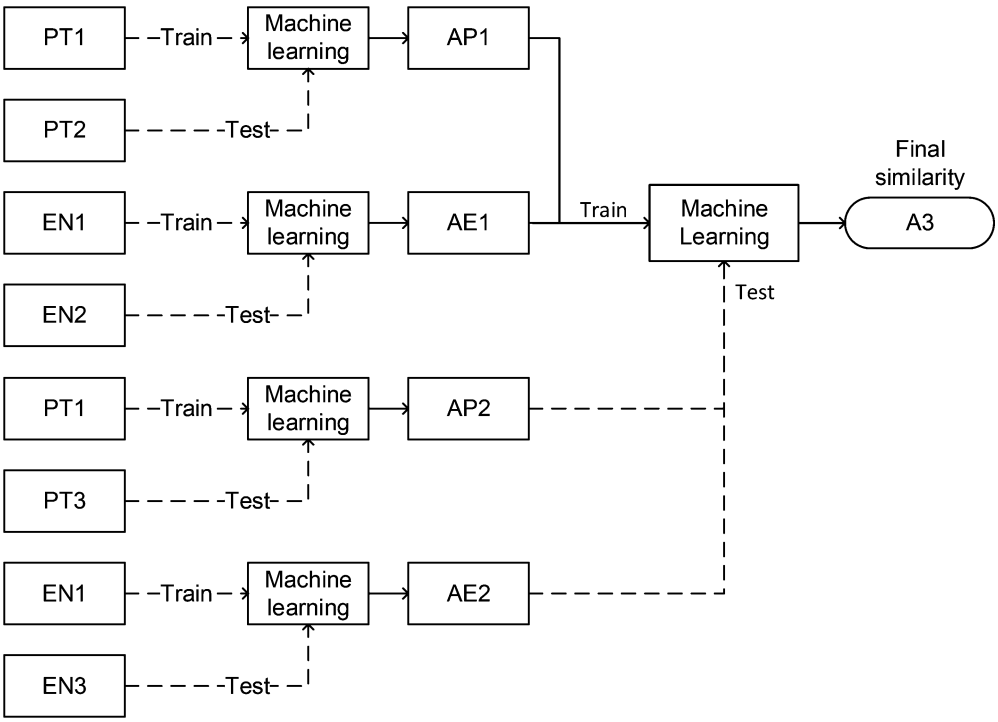


*Figure 2.* Proposed approach overview

Z-score, and L2 besides applied correlation, and variance inflation factors to find the more relevant attributes. However, the results were satisfied, and due to this, we performed chi-squared attribute selection to retain the best four variables. After doing all pre-processing, transformations, and feature selection, the present study used as an attributes the word embeddings difference between sentences [7], cosine distance between TF-IDF vectors [5], shared $n$-gram proportion between sentences [8] and word overlap measure [8].

Throughout our experiments, we tried measure similarity with the following common use machine learning models: Support Vector Machines (kernels: linear, radial, and polynomial), Random forest, linear regression, ridge regression, bayesian ridge regression, elastic net, catboost regressor* and cluster regression (using stable by [24] number of clusters). However, almost all of they showed similar results, and a statical $t$-paired test did not reject the null hypothesis' ($p > 0.05$). Therefore, there is no significant difference between all classifiers inner each group (AP1, AP2, AE1, AE2, and A3) experiments. Although this, simple linear regression models archived the best results shown in table 1.

To allow a comparison with Brazilian Portuguese, and others work on semantic textual similarity literature, the current article measure obtained results with Pearson's Correlation ($r$), Spearman's Correlation ($\rho$), and Mean Squared Error (MSE) metrics. Analyzing the results of table 1, we can observe that the higher $r$ and the smaller MSE were achieved through Portuguese experiments. Enforcing our expectations, AP1 and AP2 performed better than English translated sentences, which we believe be despite the noise inserted by automatic translation process. Therefore, this explain why A3 process did not overcome P1 & P2 $\rightarrow$ P3.

* Available at: https://catboost.yandex/ (accessed: 08.01.2019).

*Table 1.* **Results of experiments with proposed approach**

| Experiment | Artefact | Algorithm | $r$ | $\rho$ | MSE |
|---|---|---|---|---|---|
| P1 → P2 | AP1 | Linear regression | 0.70 | 0.69 | 0.39 |
| P1 → P3 | AP2 | Linear regression | 0.68 | 0.66 | 0.41 |
| P1 & P2 → P3 | – | Linear regression | 0.68 | 0.66 | 0.41 |
| P1 & P3 → P2 | – | Linear regression | 0.70 | 0.69 | 0.39 |
| E1 → E2 | AE1 | Linear regression | 0.65 | 0.63 | 0.45 |
| E1 → E3 | AE2 | Linear regression | 0.64 | 0.62 | 0.45 |
| E1 & E2 → E3 | – | Linear regression | 0.64 | 0.62 | 0.45 |
| E1 & E3 → E2 | – | SVM radial kernel | 0.65 | 0.63 | 0.45 |
| AP1 & AE1 → AP2 & AE2 | A3 | SVM linear kernel | 0.68 | 0.67 | 0.45 |
| AP2 & AE2 → AP1 & AE1 | – | Linear regression | 0.24 | 0.22 | 0.77 |

To improve our approach, we start an exploration of linguistic relations to better measure the similarities over sentences. As stated by [8], the antonym, synonym, and hypernym could contribute to more representative sentences. Therefore, we applied the synonym and hypernym generalization suggested by the author. The results were summarized in table 2. As shown in this table, the archived results with synonym, and hypernym generalization did not overcome the past obtained without any word replacement in the sentences. A more deep analysis on A3 showed that all methods failed to estimate border lower ($< 2$) and upper ($> 4$) similarities. In addition, the classifiers result mainly concentrated around on target mean but do not respect data variance. Therefore, the proposed approach seems to perform better on the measure the semantic similarity of closed sentences.

*Table 2.* **Comparison through the use of linguistics resources**

| Artefact | Linguistic | Algorithm | $r$ | $\rho$ | MSE |
|---|---|---|---|---|---|
| | | Linear regression | **0.70** | **0.69** | **0.39** |
| AP1 | Synonym | SVM polynomial kernel | 0.67 | 0.66 | 0.42 |
| | Synonym and hypernym | SVM polynomial kernel | 0.67 | 0.66 | 0.42 |
| | | Linear regression | **0.68** | **0.66** | **0.41** |
| AP2 | Synonym | CatBoostRegressor | 0.66 | 0.65 | 0.43 |
| | Synonym and hypernym | Linear regression | 0.66 | 0.65 | 0.43 |
| | | Linear regression | **0.65** | **0.63** | **0.45** |
| AE1 | Synonym | Linear regression | 0.58 | 0.56 | 0.51 |
| | Synonym and hypernym | CatBoostRegressor | 0.59 | 0.56 | 0.51 |
| | | Linear regression | **0.64** | **0.62** | **0.45** |
| AE2 | Synonym | Linear regression | 0.60 | 0.58 | 0.48 |
| | Synonym and hypernym | SVM polynomial kernel | 0.60 | 0.57 | 0.49 |
| | | SVM linear kernel | **0.68** | **0.67** | **0.45** |
| A3 | Synonym | Linear regression | 0.59 | 0.56 | 0.49 |
| | Synonym and hypernym | Linear regression | 0.59 | 0.56 | 0.50 |

We can see the best results for the assessment of Portuguese STS besides a comparison of the proposed approach with state-of-art results on PROPOR dataset through Pearson's Correlation ($r$), and Mean Squared Error (MSE) on table 3.

As it is possible to observe in table 3, the results obtained in this work figure on the top four best results for Pearson's Correlation or MSE, when compared to the related work that used the same dataset. The table indicates the set of attributes and the metric values obtained. Although we do not overcome current state-of-art results for Portuguese STS, our experiments showed that combining English resources to deal with limited language resources insert more noise than help classifiers to estimate the similarities. As previously stated, the automatic translation service could insert some noise due to the

*Table 3.* **Results comparison with state of the art**

| Method | | $r$ | MSE |
|---|---|---|---|
| Proposed approach | AP1 | 0.70 | 0.39 |
| | AP2 | 0.68 | 0.41 |
| | AE1 | 0.65 | 0.45 |
| | AE2 | 0.64 | 0.45 |
| | A3 | 0.68 | 0.45 |
| State-of-art | [5] | 0.70 | 0.38 |
| | [18] | 0.71 | **0.37** |
| | [20] | **0.73** | 0.63 |
| | [17] | 0.65 | 0.44 |
| | [7] | 0.60 | 0.49 |
| | [8] | 0.64 | 0.44 |

loss of sentence semantics. Therefore, a more deep experiment with another translation services is recommended to appropriately discard mixture languages resources approach.

**6. Conclusions.** In this study, we presented an approach that makes use of machine-translate and linguistic resources, to bring all potential of Portuguese and English linguistic relations on sentences, in order to measure the STS between short sentences. To do so, we have applied word embeddings, TF-IDF, PCA, and the linguistic relations of antonyms, hypernym, and synonymy. This approach allowed us to obtain a set of different attributes combination, used then in experiments with a large set of classifiers. The results achieved show that combine English resources to deal with Brazilian Portuguese ones insert more noise than help classifiers to estimate the similarities. This is considered as an indication of the disadvantages of using language mixture resources to obtain linguistic aspects from sentences. Moreover, our bests results were obtained with the combination of attributes which not incorporate linguistic and probabilistic aspects. This was observed with all the different classifiers used.

The results achieved show that generalization of synonym, and hypernym did not increase information for a better identification of similarity in language mixture scenario. Moreover, our results showed that the use of linguistic relations combined with probabilistic techniques scored worst than using only Portuguese or English languages. Although this approach does not overcome current state-of-art results for Brazilian Portuguese STS, our achieved results appear among the best in the literature of STS for the language.

Future works may consider not using mixture language models to measure the STS for only one language, in order to avoid in the current results.

**References**

1. Gomaa W. H., Fahmy A. A. A survey of text similarity approaches. *Intern. Journal of Computer Applications*, 2013, vol. 68, no. 13, pp. 13–18.

2. Freire J., Pinheiro V., Feitosa D. LEC UNIFOR no ASSIN: FlexSTS-Um framework para Similaridade Semantica Textual. *PROPOR-Intern. conference on the Computational Processing of Portuguese*. Tomar, Portugal, 2016. Available at: http://proper206.di.fc.ul.pt/ (accessed: 03.08.2018).

3. Barbosa L., Cavalin P., Kormaksson M., Guimaraes V. Blue man mroup at ASSIN: Using distributed representations for semantic similarity and entailment recognition. *PROPOR-Intern. conference on the Computational Processing of Portuguese*. Tomar, Portugal, 2016. Available at: http://proper206.di.fc.ul.pt/ (accessed: 03.08.2018).

4. Ferreira R., Lins R. D., Simske S. J., Freitas F., Riss M. Assessing sentence similarity through lexical, syntactic and semantic analysis. *Computer Speech & Language*, 2016, vol. 39, pp. 1–28.

5. Hartmann N. S. Solo queue at ASSIN: Combinando abordagens tradicionais e emergentes. *Linguamática*, 2016, vol. 8, no. 2, pp. 59–64.

6. Cer D., Diab M., Agirre E., Lopez-Gazpio I., Specia L. *Semeval-2017 Task 1: Semantic Textual Similarity-multilingual and cross-lingual focused evaluation.* arXiv preprint arXiv:1708.00055, 2017. doi: 10.18653/v1/S17-2001

7. Hartmann N., Fonseca E., Shulby C., Treviso M., Rodrigues J., Aluisio S. *Portuguese word embeddings: Evaluating on word analogies and natural language tasks.* arXiv preprint arXiv:1708.06025, 2017.

8. Silva A., Rigo S., Alves I. M., Barbosa J. Avaliando a similaridade semântica entre frases curtas através de uma abordagem híbrida. *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, Uberlândia, 2017, pp. 93–102.

9. Pradhan N., Gyanchandani M., Wadhvani R. A review on text Similarity Technique used in IR and its application. *Intern. Journal of Computer Applications*, 2015, vol. 120, no. 9, pp. 29–34.

10. Chen F., Lu C., Wu H., Li M. A semantic similarity measure integrating multiple conceptual relationships for web service discovery. *Expert Systems with Applications*, 2017, vol. 67, pp. 19–31.

11. Berrahou S. L., Buche P., Dibie J., Roche M. Xart: Discovery of correlated arguments of $n$-ary relations in text. *Expert Systems with Applications*, 2017, vol. 73, pp. 115–124.

12. Ferreira R., Cavalcanti G. D., Freitas F., Lins R. D., Simske S. J., Riss M. Combining sentence similarities measures to identify paraphrases. *Computer Speech & Language*, 2018, vol. 47, pp. 59–73.

13. Yanaka H., Mineshima K., Martinez-Gomez P., Bekki D. *Determining Semantic Textual Similarity using natural deduction proofs.* arXiv preprint arXiv:1707.08713, 2017.

14. Kajiwara T., Bollegala D., Yoshida Y., Kawarabayashi K. I. An iterative approach for the global estimation of sentence similarity. *PloS one*, 2017, vol. 12, no. 9, pp. e0180885.

15. Brychcín T., Svoboda L. UWB at Semeval-2016 Task 1: Semantic Textual Similarity using lexical, syntactic, and semantic information. *Proceedings of the 10th Intern. Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, 2016, pp. 588–594.

16. Kashyap A., Han L., Yus R., Sleeman J., Satyapanich T., Gandhi S., Finin T. Robust semantic text similarity using LSA, machine learning, and linguistic resources. *Language Resources and Evaluation*, 2016, vol. 50, no. 1, pp. 125–161.

17. Oliveira Alves A., Rodrigues R., Gonçalo Oliveira H. ASAPP: Alinhamento Semântico Automático de Palavras aplicado ao Português (eng. ASAPP: Automatic semantic alignment for phrases applied to portuguese). *Linguamática*, 2016, vol. 8, no. 2, pp. 43–58.

18. Cavalcanti A. P., de Mello R. F. L., Ferreira M. A. D., Rolim V. B., Tenório J. V. S. Statistical and semantic features to measure sentence similarity in Portuguese. *Intelligent Systems (BRACIS), 2017 Brazilian conference on.* IEEE, 2017, pp. 342–347.

19. Mikolov T., Chen K., Corrado G., Dean J. *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:1301.3781, 2013.

20. Fialho P., Marques R., Martins B., Coheur L., Quaresma P. *Medição de Similaridade Semântica e Reconhecimento de Inferência Textual (eng. Measuring Semantic Similarity and Recognizing Textual Entailment).* INESC-ID@ASSIN, 2016.

21. Faruqui M., Tsvetkov Y., Rastogi P., Dyer C. *Problems with evaluation of word embeddings using word similarity tasks.* arXiv preprint arXiv:1605.02276, 2016.

22. Paiva V., Rademaker A., Melo G. Openwordnet-pt: An open brazilian wordnet for reasoning. *COLING 2012*. Mumbai, 2012.

23. Miller G. A. WordNet: a lexical database for English. *Communications of the ACM*, 1995, vol. 38, no. 11, pp. 39–41.

24. Lozkins A., Bure V. M. The probabilistic method of finding the local-optimum of clustering. *Vestnik of Saint Petersburg University. Series 10. Applied Mathematics. Computer science. Control Processes*, 2016, iss. 1, pp. 28–37.

Received: November 18, 2018.
Accepted: March 15, 2019.

Author's information:

*Allan Silva* — Master; allanbs@unisinos.br

*Aleksejs Lozkins* — Postgraduate Student; aleksejs.lozkin@gmail.com

*Luiz Ricardo Bertoldi* — Master; luizbertoldi@unisinos.br

*Sandro Rigo* — Dr. Sci. in Computers; rigo@unisinos.br

*Vladimir M. Bure* — Dr. Sci. in Technics, Professor; vlb310154@gmail.com

# Семантическое сходство текстов на бразильском португальском языке: Подход, основанный на комбинировании нескольких языков

*А. Сильва*[1], *А. Ложкинс*[2], *Л. Р. Бертольди*[1], *С. Риго*[1], *В. М. Буре*[2]

[1] Университет Вале до Рио дос Синос, Бразилия, 93020-190, Сао Леопольдо, РС, пр. Унисинос, 950

[2] Санкт-Петербургский государственный университет, Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9

**Для цитирования:** *Silva A., Lozkins A., Bertoldi L. R., Rigo S., Bure V. M.* Semantic Textual Similarity on Brazilian Portuguese: An approach based on language-mixture models // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2019. Т. 15. Вып. 2. С. 235–244. https://doi.org/10.21638/11702/spbu10.2019.207 (In English)

В литературе исследование семантического текстового сходства (СТС) описывается как фундаментальная часть многих задач обработки естественного языка. Подходы СТС зависят от наличия и объема лексико-семантической базы. Существуют несколько попыток по улучшению лексико-семантической базы, и представлено большое количество приложений для английского языка. Лингвистическая база бразильского португальского, по сравнению с английской, не имеет одинаковой доступности в отношении семантических связей и содержания, что приводит к потере точности в задачах СТС. В настоящей работе описан подход, сочетающий лексико-семантические онтологические базы бразильского португальского и английского языков, для использования всех возможностей языковых отношений и создания комбинированной модели для измерения семантического текстового сходства. Предложенный подход проанализирован на известном и признанном наборе данных бразильского португальского языка СТС, который позволил выявить преимущества и недостатки комбинированной модели.

*Ключевые слова*: семантическое сходство текстов, обработка естественного языка, компьютерная лингвистика, онтологии.

Контактная информация:

*Сильва Аллан* — магистр; allanbs@unisinos.br

*Ложкинс Алексейс* — аспирант; aleksejs.lozkin@gmail.com

*Бертольди Луиз Рикардо* — магистр; luizbertoldi@unisinos.br

*Риго Сандро* — д-р информ. наук, проф.; rigo@unisinos.br

*Буре Владимир Мансурович* — д-р техн. наук, проф.; vlb310154@gmail.com