Towards analysis on textual inference at ASSIN-2 dataset

Felipe O. da Silva[1], Giovana Meloni Craveiro[1],
Vinícius F. da Silva[1], Vinícius João de Barros Vanzin[1]

[1]Institute of Mathematics and Computer Sciences
University of São Paulo (USP) – São Carlos, SP – Brazil

{felipe.oliveiras2000}@gmail.com,
{giovana.meloni.craveiro, vfsilva, vinicius.vanzin}@usp.br

Abstract. *In this article, we conduct a preliminary analysis of different methods to address the Textual Entailment Recognition (RTE) task in Portuguese. We use the ASSIN-2 dataset as a benchmark to evaluate our models. Our work combines various textual representation approaches, including bag of words and word embeddings, with machine learning models. Additionally, we present a rule-based approach. Our highest performance was achieved by the BERTimbau-large model fine-tuned on ASSIN-2, which attained an $F1$ score of 0.89%, positioning it just 1% below the current state-of-the-art. Our ongoing experiment aims to combine our different approaches to leverage their full potential.*

1. Introduction

Textual Entailment Recognition (RTE), also known as Natural Language Inference (NLI), is the NLP task of determining whether one sentence (premise) entails another (hypothesis). Approaches used for NLI include earlier symbolic and statistical methods to more recent deep learning approaches [Bowman and Zhu 2019]. In the last few years, there has been fast progress on the task [Bowman et al. 2015, Chen et al. 2018] with studies of new model architectures aimed at improving performance on benchmarks as well as at increasing the number of large datasets for evaluating these systems [Williams et al. 2018].

There is a scarcity of datasets on Portuguese for NLI. The ASSIN-2 is a curated dataset proposed at the STIL 2019 conference as an effort to set a new computational semantic benchmark for Portuguese. It contains 10,000 samples of annotated data, divided into balanced portions. The training set contains 6500 sentence pairs, the validation set is composed of 500 pairs, and the test set consists of 2448 pairs [Real et al. 2020]. An example from the data set is shown in Table 1.

Table 1. Example of ASSIN-2 data

Premise	Hypothesis	NLI
Alguns animais estão brincando selvagemente na água	Alguns animais estão brincando na água	Entails
Um avião está voando	Um cachorro está latindo	None

This work uses the ASSIN-2 dataset to analyze and compare the performance of diverse classification approaches for NLI in Portuguese. It combines the representation formats of word embeddings and bag of words with machine learning algorithms. It uses Logistic Regression with L1 and L2 regularization, Random Forests, and eXtreme Gradient Boosting (XGBoost) for the former [Pedregosa et al. 2011] and CatBoost [Prokhorenkova et al. 2018], a Bi-directional recurrent neural network [Schuster and Paliwal 1997] and BERTimbau [Souza et al. 2020], for the latter. Finally, it includes a rule-based symbolic approach.

2. Methodology

2.1. Symbolic approach

The symbolic approach was inspired by the annotation guidelines for ASSIN-2 [Real et al. 2020], which direct towards verifying whether expressions from both sentences could refer to the same extralinguistic elements or whether an expression from one sentence could have a hypernymy relationship with an expression from the other sentence.

Our rule-based method assumes that the sentences must be similar and that if one of them contains a negation, the other should also contain one. Additionally, it assumes that a longer sentence typically carries more specifications than a shorter one and that a more general sentence could be entailed by a more specific one, but not vice versa.

Hence, the classifier declares that sentence A entails sentence B if sentence A is longer than sentence B, their similarity rate is greater than fifty percent, and either no sentence contains the negation term "não" or both sentences do. The similarity rate between them is measured by dividing the number of words that are common in both sentences by the length of sentence B. Preprocessing is based on lowercasing and removing accents.

As this approach does not consider the semantics of different terms and thus cannot identify relationships of synonymy and hypernymy among different words, it cannot be considered a method that appropriately addresses the NLI task. At this stage, it is intended as a baseline for the minimal performance that the other methods should achieve.

2.2. Bag of words

In the approach that uses *Bag of Words* to represent sentences, different techniques were experimented separately and in combination.

The baseline method is the traditional *Bag of Words* [Zhang et al. 2010], which transforms texts into attribute-value tables by calculating the frequency with which words occur in the texts. This technique creates a single matrix for the entire data set, in which each line i contains the number of times that each word j occurred for the i-th sentence in the database. Similarly, the approach based on *Term Frequency-Inverse Document Frequency* (TF-IDF) [Das and Chakraborty 2018], uses attribute-value tables, but also normalizes the frequency of terms in a document, increasing the relevance of rare words. Additionally, an *n-grams* strategy, which aims to include the context of adjacent words instead of the single referred term, is tested as an alternate configuration.

Aiming to reduce the dimensionality of the training set, we also tested adding *Principal Component Analysis* (PCA) [Shlens 2014] with a varying number of components among 3, 4, 5, 10, 100, and 500. As a result, we had 2305 dimensions for the traditional *BOW* method and 33516 dimensions for approaches using 1 to 3 n-grams.

The cited representation formats and techniques are applied to the data set, preprocessed by removing *stopwords*, and fed into machine learning algorithms based on

Logistic Regression, Random Forests, and eXtreme Gradient Boosting (XGBoost), which are offered by scikit-learn [Pedregosa et al. 2011].

For each combination of hyperparameters, the models are fine-tuned with Grid Search and k-fold cross-validation with k = 5 is applied. F1 measure is used to select the best model. It is assumed that no significant weight differences exist between False Positive (FP) and False Negative (FN) errors for this task. It is also the metric used to evaluate RTE systems in the ASSIN-2 benchmark [Real et al. 2020]. A total of 150 different configurations are used, with distinct representation approaches, resulting in 750 predictive models.

2.3. Word embeddings

In this semantic representation format, the NILC pre-trained embeddings [Hartmann et al. 2017] are combined with machine learning classifiers.

1. The first strategy uses the CatBoost algorithm [Prokhorenkova et al. 2018]. Preprocessing consists of normalization to lowercase words and concatenation of premise and hypothesis with a separation token "[SEP]", without removing stopwords. The sentence is represented by the sum of the individual embedding vectors of each word. Tests are conducted with the embeddings *word2vec* CBow of 100 dimensions and *Glove* Skip-gram of 300 dimensions[Mikolov et al. 2013].
2. The second technique employs a bidirectional recurrent neural network (BRNN) [Schuster and Paliwal 1997]. Each sentence was preprocessed with the techniques described in [Hartmann et al. 2017]. The model was trained with the *Adam* optimazation algorithm using at most 25 epochs and a batch of 128 samples. 21 models were trained, varying embeddings (*word2vec skip-gram, word2vec CBoW, wang2vec skip-gram, wang2vec CBoW, FastText skip-gram, FastText CBoW e Glove*) and number of dimensions (50, 300 and 1000).
3. The third method uses BERTimbau [Souza et al. 2020], a Brazilian Portuguese language model, trained on the brWaC corpus [Wagner Filho et al. 2018], fine-tuned [Howard and Ruder 2018] to the RTE task. There are two versions of pre-trained models: one with 12 layers of encoders, 110 million parameters, and 768 dimensions; and one with 24 layers of encoders, 335 million parameters, and 1024 dimensions. The same hyperparameters are used for both versions. The maximum token sequence length is set at 128, the maximum number of epochs is 4, and the batch size is 16 for training and 64 for validation. The remaining hyperparameters were not modified and we use the tokenizer from the pretrained model.

3. Results and Discussion

Among the different strategies used to tackle RTE, several configurations were tested. Table 2 exhibits the results of the ones that obtained the highest scores. For each textual representation method, our code and experiments are openly available at **repository** [Oliveira da Silva et al. 2023], facilitating replication of results.

Our symbolic approach is designed solely with rules that do not attempt to capture semantic relationships among different words. Despite this aspect, it achieves the remarkably high $F1$ score of 0.71%, given its simplicity. This indicates that it is either

Table 2. Result of the best models

Set	Method	Metrics			
		F1	Precision	Recall	Accuracy
Train	BOW	0.94	0.92	0.95	0.93
	BERTimbau-large	0.96	0.97	0.96	0.96
	Symbolic	0.70	0.75	0.67	0.72
Validation	BOW	0.88	0.87	0.89	0.88
	BERTimbau-large	0.96	0.96	0.96	0.96
	Symbolic	0.72	0.75	0.69	0.73
Test	BOW	0.77	0.68	0.88	0.73
	BERTimbau-large	0.89	0.90	0.89	0.89
	Symbolic	0.69	0.74	0.65	0.71

a promising approach or that the data set used to test the experiments is too simplistic to reflect the complexity of the task in real-world examples. The other approaches are expected to outperform this method.

Our fine-tuned model that uses BERTimbau-large indeed reaches an $F1$ score of 0.89%, only one percentage point away from the current state-of-the-art in the RTE task - represented by a BERTimbau-large trained by Neuralmind [Souza et al. 2020] which achieved 0.90%. However, it is important to perform a qualitative analysis of its misclassifications. Table 3 shows an example from the test set in which our BERTimbau model misclassifies.

Table 3. Example of BERTimbau misclassification

Premise	Hypothesis	NLI
um palhaço está cantando no palco e pessoas estão dançando	uma pessoa fantasiada de palhaço está cantando	Non-Entailment

In the given example, the premise says “um palhaço está cantando no palco” and the hypothesis says “uma pessoa fantasiada de palhaço está cantando”. Although the data set classifies this sentence as non-entailment, “a clown” could be considered as equivalent to “a person dressed as a clown”. Therefore, the model seems to be a solid solution, but it is reasonable to further analyze its misclassifications to ensure its robustness and to understand how to enhance it. Nevertheless, its greatest disadvantage is that its computational cost and complexity are significantly greater than those of the other methods.

Our best combination of a BOW method - without PCA and without TF-IDF - achieves an $F1$ measure of 0.73%, which is markedly lower than that of our best BERTimbau model, but its recall is only 0.01% below our BERTimbau's recall, which is notable given its significantly lower computational cost compared to the BERTimbau models.

Given the computational cost and accessibility disadvantages of our BERTimbau model and its performance advantage compared to our other methods, our ongoing work aims to refine and combine our methods, resulting in a Neuro-symbolic approach for Portuguese textual inference that considers all linguistic features necessary to properly address NLI, while remaining accessible and competitive with state-of-the-art models.

References

[Bowman and Zhu 2019] Bowman, S. and Zhu, X. (2019). Deep learning for natural language inference. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 6–8, Minneapolis, Minnesota. Association for Computational Linguistics.

[Bowman et al. 2015] Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

[Chen et al. 2018] Chen, Q., Zhu, X., Ling, Z.-H., Inkpen, D., and Wei, S. (2018). Neural natural language inference models enhanced with external knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2406–2417, Melbourne, Australia. Association for Computational Linguistics.

[Das and Chakraborty 2018] Das, B. and Chakraborty, S. (2018). An improved text sentiment classification model using tf-idf and next word negation. *arXiv preprint arXiv:1806.06407*.

[Hartmann et al. 2017] Hartmann, N. S., Fonseca, E., Shulby, C. D., Treviso, M. V., Rodrigues, J. S., and Aluísio, S. M. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.

[Howard and Ruder 2018] Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

[Mikolov et al. 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Oliveira da Silva et al. 2023] Oliveira da Silva, F., Craveiro, G. M., Siqueira Souza, J. M., Silva, V. F. d., and Vanzin, V. J. d. B. (2023). Natural language inference bow, word embeddings and symbolic experiments at assin-2 dataset. GitHub repository, `https://github.com/jmssouza/nlp_entailment/`.

[Pedregosa et al. 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

[Prokhorenkova et al. 2018] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

[Real et al. 2020] Real, L., Fonseca, E., and Gonçalo Oliveira, H. (2020). *The ASSIN 2 Shared Task: A Quick Overview*, pages 406–412.

[Schuster and Paliwal 1997] Schuster, M. and Paliwal, K. (1997). Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45:2673 – 2681.

[Shlens 2014] Shlens, J. (2014). A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

[Souza et al. 2020] Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

[Wagner Filho et al. 2018] Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brwac corpus: A new open resource for brazilian portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[Williams et al. 2018] Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

[Zhang et al. 2010] Zhang, Y., Jin, R., and Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1):43–52.