NILC at ASSIN 2: Exploring Multilingual Approaches

Marco A. Sobrevilla Cabezudo, Marcio Inácio, Ana Carolina Rodrigues, Edresson Casanova, and Rogério Figueredo de Sousa

NILC - Interinstitutional Center for Computational Linguistics
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos SP 13566-590, Brazil
{msobrevillac, marciolimainacio, ana2.rodrigues, edresson, rogerfig}@usp.br

Abstract. Recognizing Textual Entailment, also known as Natural Language Inference recognition, aims to identify if it is possible to infer the meaning of a text from another fragment of text. In this work, we investigate the use of multilingual models, through BERT, for recognizing inference and similarity in the ASSIN 2 dataset, an entailment recognition and sentence similarity corpus for Portuguese. We also investigate possible features that could enhance the results, such as similarity scores or WordNet relations. Our results show that a multilingual pre-trained BERT model may be sufficient to outperform the current state-of-the-art in this task for the Portuguese Language. We also show that using other features did not necessarily improve the performance of the model, however deeper studies are needed to investigate the causes for this.

Keywords: Natural Language Inference · BERT · Multilingual Training · Cross-lingual Training

1 Introduction

Recognizing meaning connections such as entailment relations and content similarity among different statements is part of daily communication and usually done effortless by humans. However, automatizing such communication component has been a challenge. Overcome it can help many Natural Language Processing (NLP) applications such as Machine Translation, Question Answering, Semantic Search and Information Extraction.

Particularly, the task of recognizing textual entailment (RTE) has been widely explored in natural language processing field. Inference recognition in NLP, also called text entailment recognition, consist recognizing a directional relationship between pairs of text expressions, in which a human reading the first text would infer the second one is likely true.[7].

Initially spread by the Pascal Challenge [6], several inference-annotated corpus for English have been released in the last decade, such as MultiNLI [19], XNLI [4], and SICK [13]. Specifically for Portuguese, multiple efforts have been

made to develop an inference-annotated corpus[11][16]. In 2016 the first shared task for inference recognition for Portuguese, ASSIN [11], took place, followed by the second edition in 2019 (ASSIN 2) [15].

In addition to the growth of available corpora, several techniques have been tested to improve inference recognition in NLP, including probabilistic models and rule-based approaches [5]. Recently, with the expansion of machine learning applications (and neural networks in particular), it has been tested on the lights of pre-trained language representations [8][18].

A broadly known one is the Bidirectional Encoder Representations from Transformers (BERT). BERT has been used effectively in multiple tasks like Semantic Text Similarity, Paraphrase detection, among others[8][1]. BERT has improved fine-tuning approaches using a masked language model (MLM), in which parts of the input are randomly hidden to be predicted based only on their context.

In addition to MLM, the authors also use next sentence prediction task that jointly pre-trains text-pair representations. BERT was the first fine-tuned representation model to achieve state-of-the-art performance for a large number of token and phrase-level tasks, outperforming models developed specifically for these tasks. As far as we know, there are two monolingual pre-trained models publicly available: one trained for English and one for Chinese, along with one multilingual model. The multilingual model was trained for the 100 languages with most articles on Wikipedia[2].

This work presents the results achieved by the NILC group for the ASSIN 2 shared task. We firstly analyse the corpus in order to find correlation between features and classification labels. Then we fine-tune multilingual BERT on sentence-pairs from ASSIN 2 corpus for the RTE task and finally, use the generated embeddings and incorporate some linguistic features for the Semantic Textual similarity (STS) evaluation. In general, we rank 3rd place for the RTE task and 5th place for STS task.

This paper is organized as follows. Firstly, we discuss previous related work in section 2. Afterwards, we describe the ASSIN dataset in section 3, followed by our experiments and results in section 5. Finally, some conclusions are presented in section 6.

2 Related Work

There are several works on textual inference for multiple languages. However, due to differences in corpora for other languages, we report mainly works done in Brazilian Portuguese. This way, the works reported here use the corpus ASSIN, thus making a closer comparison to our work.

Rocha and Lopes Cardoso [17] reported only the result for European Portuguese (PT-PT) and obtained F1 of 0.73. they explored the use of named en-

[1] Available at https://github.com/google-research/bert.

[2] Available at https://github.com/google-research/bert/blob/master/multilingual.md.

tities as a feature along with word similarity, number of semantically related tokens, and whether both sentences have the same verb tense and voice.

Fialho et al. [9] (from the INESC-ID group) obtained an F1 score of 0.71 in Brazilian Portuguese (PT-BR) and 0.66 in PT-PT for textual inference in their best experiment. The authors trained a Support Vector Machine (SVM) model using 96 lexical features, including editing distance, BLEU score, word overlap, ROUGE score, among others.

Barbosa et al. [2] (from the Blue Man Group) obtained, in their best experiment, an F1 score of 0.52 for PT-PB and 0.61 for PT-PT exploring the use of word embeddings similarity. For classification, the authors used SVM and Siamese Networks [3].

Reciclagem and ASAPP were proposed by Alves et al. [1] (from the ASAPP group). Reciclagem is based only on heuristics in semantic networks. While ASAPP explores the use of lexical, syntactic, and semantic features extracted from texts. Their best results were an F1 score of approximately 0.5 for PT-PB and 0.59 for PT-PT.

Finally, Fonseca and Aluísio proposed the Infernal system [10]. The authors explored some features such as syntactic knowledge, embedding-based similarity, and well-established features that deal with word alignments, totalizing 28 features. Their best experiment for PT-BR achieved F1 score of 0.71, similarly to the previously reported INESC-ID system. On the other hand, for PT-PT, F1 score of 0.72 has been reported, lower than the one obtained by Rocha and Lopes Cardoso [17]. When considering the entire dataset, i.e. both PT-BR and PT-PT, the Infernal system reached F1 score of 0.72, currently the best result reported for this dataset.

3 ASSIN 2 Dataset

The ASSIN 2 corpus consists of 10,000 pairs of sentences tagged with similarity grading and entailment classification. Tags for the Recognizing Textual Entailment (RTE) task are *None*, when both sentences are not related in any way and *Entailment* when the second sentence is a direct inference of the first one.

ASSIN 2 was also manually annotated for semantic textual similarity in a range from 0 to 5, for which the pair was considered more similar by the annotators as higher is the number.

The corpus was split in two parts for the shared task, 7000 pairs were provided in advance as a training set and the other 3000 pairs later. The dataset provided for training was balanced, with 3500 pairs labeled as *None* and 3500 as *Entailment*.

4 Corpus Analysis

As the dataset provided for training was balanced for the two entailment classes, we verified if it was equally balanced for other features. We use OpenWordnet-PT to extract the number of synonyms and hypernyms in each pair per class.

Hypernyms were counted when the second phrase contained any hypernym of any word in the first one. Synonym values were calculated in a similar way. Results are shown in figures 1 and 2.

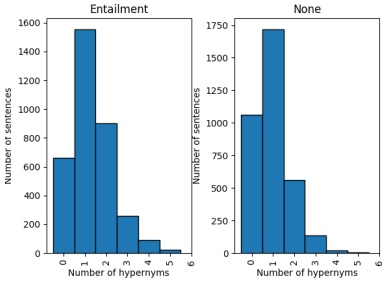

Fig. 1. Hypernym counts

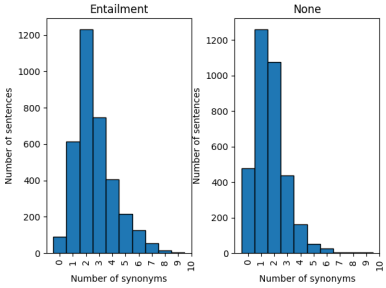

Fig. 2. Synonym counts

As can be seen, sentence pairs with entailment tend to have more hypernyms (as can be seen by the bars representing counts of 0 and 2). The same can be observed for synonyms: there are more sentence pairs without entailment with no synonyms than those with an entailment relation.

Since the corpus was also annotated for semantic similarity in a continuous range from 0 to 5, we investigate the relation between similarity index and entailment classes. As a result, we find that *entailment* pairs have lower dispersion than *none* ones, and, differently from *none*, its range is mostly concentrated in higher similarity values, as can be seen in Figure 3.

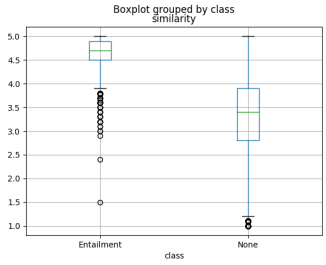

Fig. 3. Similarity per class

Although similarity values shows notable correlation to entailment classes, it was not possible to considered them for the entailment recognition task, once it was part of the expected predicted result. In other words, textual similarity

annotation were hidden in the test set. Therefore, in order to incorporate this kind of knowledge into the model, some metrics have been explored, namely BLEU [14] and Levenshtein's Edit Distance [12].

Both metrics were calculated for each pair of sentences resulting in the distributions presented in figures 4 and 5. The figures show the distributions of the metrics according to the classes in the corpus. The results for the BLEU metric, as it is based on the calculation of string overlaps between texts, show greater values for *entailment* (higher similarity). Accordingly, Edit Distance values are lower for this class, as it computes how different a pair of sentences is, i.e. their dissimilarity.

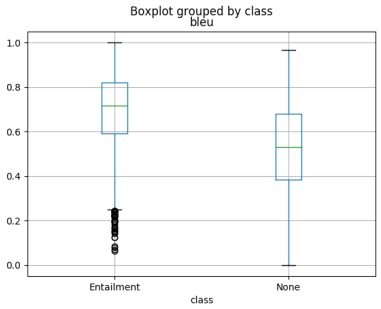

Fig. 4. BLEU Metric

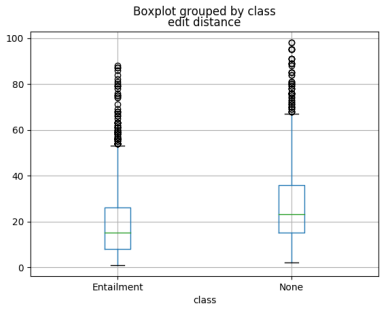

Fig. 5. Edit distance

From these analyses, these four features (synonyms and hypernyms counts, BLEU and Edit Distance scores) seem applicable to text inference prediction. Thus, we try to combine these features with the model selected for this task: the BERT language model, as will be discussed later.

5 Experiments and Results

We analyze how BERT performs on Recognizing Textual Entailment (RTE) and Semantic Textual Similarity (STS) for Brazilian Portuguese in ASSIN 2 corpus.

It is worth noting that our methods were tested on three variations of the original dataset for RTE, which are the runs submitted to the shared task. On the other hand, our submission to the Semantic Textual Similarity task only was tested on the original dataset.

5.1 Recognizing Textual Entailment (RTE)

We use BERT to train the RTE classifier. Specifically, we use a pre-trained BERT model, add an untrained layer of neurons at the end, and train the new model

for the RTE task. For this purpose, we use the pre-trained BERT multilingual model that includes Brazilian Portuguese[3] along with other 103 languages. The model was trained for 7 epochs with a learning rate of 0.00002, a batch size of 22 and a maximum sequence length of 128 tokens[4].

As mentioned previously, linguistic features like BLEU, Edit-distance, number of hypernymns and synonymns between the sentence in a sentence-pair were also used in the model, however, the introduction of these features did not contribute positively to the final result.

5.2 Semantic Textual Similarity

Due to the high correlation between the Entailment class and the similarity values, we used the previous trained model for RTE to obtain embeddings for each sentence-pair. Initially, we experiment using separate embeddings (an embedding for each sentence, obtained through BERT), but the results were poor. Thus, we use joint embeddings (size of 768) as input, obtained by providing both sentences in the pair as an input to the BERT model, which creates a single representation of the whole pair. Additionally, we incorporate the features BLEU, number of synonyms and hypernyms in common for each sentence-pair.

For experiments, we use a multilayer perceptron with a hidden layer (64 neurons), the logistic function as the activation function, the adam optimizer, a learning rate of 0.001 and a maximum number of iterations of 1,000.

5.3 Results

As mentioned, our methods were trained and fine-tuned on three variations of the original dataset. The first variation ("own" in Table 1), consists in splitting the original training dataset (7,000 sentence-pairs) into 6,300 for training and 700 for development. This split has been done by a stratified sampling according to both entailment and similarity values, to guarantee that their distribution is also represented in the development set. The second one was provided by the organization and it contains 6,500 sentence-pairs for training and 500 for development ("assin-2" in Table 1). The last variation comprised all 7000 sentences ("all" in Table 1).

Table 1 shows the results of the best three teams (excluding our team), our results and the baseline results for RTE. In general, our proposal obtained the third place in the RTE task (being only surpassed by the Deep Learning Brasil and IPR teams) and the difference between our proposal and their proposals is small.

Concerning our proposal, it is important to highlight three regards. Firstly, the split performed by us shows the best results even containing fewer instances

[3] The model is available at https://storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip

[4] It is worth noting that other hyperparameters were tested. However, the results were not shown improvements and they are no reported in this paper.

in the training set than the original split, which could note the relevance of the splitting strategy. However, we cannot affirm this due to the small improvements.

Secondly, the introduction of linguistic features (BLEU, edit-distance, among others) did not contribute positively to the final result. Thus, we only fine-tuned the multilingual BERT on our RTE task. In principle, it could be thought that multilingual BERT learned all these features, this way by including them, the results did not improve. Another explanation could be that it is necessary to explore other ways to integrate this kind of information. However, a deeper study must be performed.

Finally, it is worth noting the potential of multilingual BERT. This model made our proposal easier in comparison to other approaches as it only needs the pre-trained model and adding an untrained layer to perform the fine-tuning process on the RTE task.

Table 1. Best teams entailment results

Team	Run	Results	
		F1*	Acc.
Deep Learning Brasil	Ensemble	0.883	88.32%
IPR	1	0.876	87.58%
Stilingue	2	0.866	86.64%
NILC	own	0.871	87.17%
NILC	assin-2	0.868	86.85%
NILC	all	0.865	86.56%
Baseline	BoW sentence 2	0.557	56.74%
Baseline	Word Overlap	0.667	66.71%
Baseline	Infernal	0.742	74.18%

Concerning the Semantic Textual Similarity (STS) task, our proposal obtained smaller results than the other proposals. However, our results outperformed all baselines. It is interesting to note that fine-tuning multilingual BERT on the RTE task contributes positively to the STS task. However, some linguistic features had to be incorporated to obtain better results, showing that BERT could not learn this kind of information. In that sense, an interesting direction could be experimenting fine-tuning on STS task and then using this information to apply to RTE task or trying to learn both tasks in a multi-task learning approach.

Table 2. Best teams textual similarity results

Team	Run	Results	
		Pearson*	MSE
IPR	1	0.826	0.52
Stilingue	3	0.817	0.47
Deep Learning Brasil	Ensemble	0.785	0.59
L2F/INESC	BL	0.778	0.52
ASAPPpy	2	0.740	0.60
NILC	-	0.729	0.64
Baseline	Word Overlap	0.577	0.75
Baseline	BoW sentence 2	0.175	1.15

6 Conclusions and Future Works

This work presents the results obtained by the NILC group for the shared task ASSIN 2 on entailment and textual similarity recognition. We analyzed characteristics of the ASSIN 2 dataset according to its entailment classes and tested two approaches of classification using BERT.

Fine-tuning BERT, on the ASSIN 2 corpus without any extra feature presented the best results, largely outperforming the baselines. Therefore, we show that using a simple BERT model can provide satisfactory results in these tasks.

As shown in the corpus analyses, similarity, BLEU and Edit Distance metrics seem to be suitable for discriminating the entailment classes of the ASSIN 2 corpus. In particular, *entailment* pairs have higher similarity values and higher BLEU scores, as well as lower Edit Distance values, than *none* class.

The number of hypernyms and synonyms calculated consulting OpenWordNet-PT for each pair also indicates some level of distinction between the two classes, there were considerable more *entailment* pairs containing these relations. However, incorporating these features as values concatenated to BERT embedding vectors achieved poorer results.

We considered some possibilities for these negative results, such as the way features were incorporated, as a concatenation to BERT embeddings. Another reason may be that BERT embeddings already incorporate such knowledge (similarity, synonym and hypernym relations) within their representation. A future deeper analysis about the incorporation of these features may lead to further conclusions about these hypothesis.

References

1. Alves, A.O., Rodrigues, R., Oliveira, H.G.: ASAPP: alinhamento semântico automático de palavras aplicado ao português. Linguamática **8**(2), 43–58 (2016)
2. Barbosa, L., Cavalin, P., Guimaraes, V., Kormaksson, M.: Blue Man Group no ASSIN: Usando representações distribuídas para similaridade semântica e inferência textual. Linguamática **8**(2), 15–22 (2016)

3. Chopra, S., Hadsell, R., LeCun, Y., et al.: Learning a similarity metric discriminatively, with application to face verification. In: CVPR (1). pp. 539–546 (2005)
4. Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S.R., Schwenk, H., Stoyanov, V.: Xnli: Evaluating cross-lingual sentence representations. arXiv preprint arXiv:1809.05053 (2018)
5. Dagan, I., Glickman, O.: Probabilistic textual entailment: Generic applied modeling of language variability. Learning Methods for Text Understanding and Mining **2004**, 26–29 (2004)
6. Dagan, I., Glickman, O., Magnini, B.: The pascal recognising textual entailment challenge. In: Machine Learning Challenges Workshop. pp. 177–190. Springer (2005)
7. Dagan, I., Roth, D., Sammons, M., Zanzotto, F.M.: Recognizing textual entailment: Models and applications. Synthesis Lectures on Human Language Technologies **6**(4), 1–220 (2013)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Fialho, P., Marques, R., Martins, B., Coheur, L., Quaresma, P.: INESCID at ASSIN: Measuring semantic similarity and recognizing textual entailment. Linguamática **8**(2), 33–42 (2016)
10. Fonseca, E., Aluísio, S.M.: Syntactic knowledge for natural language inference in portuguese. In: International Conference on Computational Processing of the Portuguese Language. pp. 242–252. Springer (2018)
11. Fonseca, E., Santos, L., Criscuolo, M., Aluísio, S.: Visão geral da avaliação de similaridade semântica e inferência textual. Linguamática **8**(2), 3–13 (2016)
12. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady. vol. 10, pp. 707–710 (1966)
13. Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., Zamparelli, R.: Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014). pp. 1–8 (2014)
14. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
15. Real, L., Fonseca, E., Gonçalo Oliveira, H.: The ASSIN 2 shared task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese. In: Proceedings of the ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese. p. [In this volume]. CEUR Workshop Proceedings, CEUR-WS.org (2020)
16. Real, L., Rodrigues, A., e Silva, A.V., Albiero, B., Thalenberg, B., Guide, B., Silva, C., de Oliveira Lima, G., Câmara, I.C., Stanojević, M., et al.: Sick-br: a portuguese corpus for inference. In: International Conference on Computational Processing of the Portuguese Language. pp. 303–312. Springer (2018)
17. Rocha, G., Lopes Cardoso, H.: Recognizing textual entailment: challenges in the portuguese language. Information **9**(4), 76 (2018)
18. Singh, J., McCann, B., Keskar, N.S., Xiong, C., Socher, R.: Xlda: Cross-lingual data augmentation for natural language inference and question answering. arXiv preprint arXiv:1905.11471 (2019)

19. Williams, A., Nangia, N., Bowman, S.: 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). pp. 1112–1122. Association for Computational Linguistics (2018)