IPR: The Semantic Textual Similarity and Recognizing Textual Entailment systems

Rui Rodrigues[1], Paula Couto[1], and Irene Rodrigues[2]

[1] Centro de Matemática e Aplicações (CMA), FCT, UNL
Departamento de Matemática, FCT, UNL
Portugal
[2] Laboratório de Informática, Sistemas e Paralelismo (LISP)
Departamento de Informática, Universidade de Évora
Portugal

Abstract. We describe IPR's systems developed for ASSIN2 (Evaluating Semantic Similarity and Textual Entailment). Our best submission ranked first in the Semantic Textual Similarity task and second in the Recognizing Textual Entailment task. These systems were developed using BERT, for each task we added one layer to a pre-trained Bert model and fine-tuned the whole task network.

1 Introduction

In this paper, we describe the IPR team participation in the ASSIN2[11] (Evaluating Semantic Similarity and Textual Entailment) tasks, Semantic Textual Similarity (STS) and Recognizing Textual Entailment (RTE).

STS and RTE are semantic tasks that infer two semantic relations between sentences, similarity and entailment. The STS task classifies, on a scale from 1 to 5, the level of semantic equivalence between sentences. The RTE task classifies the sentences entailment, Yes/No. In the ASSIN2, challenge the metrics used to evaluate the the predictions of the systems on the two tasks were: F1-measure (primary metric) and Accuracy (secondary metric), for RTE, and Pearson Correlation (primary metric) and Root Mean Square Error (MSE, secondary metric), for STS.

Our systems were developed using BERT (Bidirectional Encoder Representations from Transformers) [3]. We took a pre-trained BERT model, add, for each task, one untrained layer of neurons on the end, and then train the new models for our classification tasks.

The train data used include: the ASSIN2 train data; the previous ASSIN Brazilian Portuguese and European Portuguese train and test data; and a Portuguese corpus build with the Portuguese Wikipedia and the journal extracts of Público and Folha de São Paulo included in corpus CHAVE[12].

In both tasks, in our best submission, we departed from the Multilingual BERT model. In the STE, task we fine-tuned the model to Portuguese using the Portuguese corpus. Then, for each task, we added a new layer and we used the

ASSIN2 train data and the ASSIN train and test data to fine-tune the resulting network, giving rise to our systems.

IPR's best submission in the STS task ranked first (Pearson correlation). And in the RTE task ranked second (F1 score).

In these semantic tasks, the strategy of using a BERT language model fine-tuned with the classification task data has obtained very good results, in section 2 it is provided an explanation of the method we followed. Section 3 describes our approach to the specific tasks and presents the results we achieved on these tasks along with instances where the systems did not perform so well. Section 4 discusses our performance in ASSIN2 and includes also future plans for improving the systems.

2 Using BERT for NLP semantic tasks

BERT models use what is known as Word Embeddings, models where words are represented as real number vectors in a predefined vector space. The use of real number vectors to represent words dates back to 1986 [5]. More recently, in 2003, in [1] a vector representation of words is obtained by using a neural network and the vectors are elements of a probabilistic language model.

In 2008, [2], the network that produced such a vector representation was trained to create a language model together with several NLP tasks: part-of-speech tags, chunks, named entity tags, semantic roles, semantic similar words.

In 2013 a significant advance was achieved with Word2vec [8].

Word2vec is a shallow model, log-bilinear, without non-linearities that enables the use of higher dimension vector representation and can be training on larger corpus. It achieved important results in several NLP tasks involving word semantic and syntactic similarity.

GloVe [9], is also a log-bilinear model but it's train differs from the word2vec train since it uses global occurences of matrices. Glove achieved important results on NLP tasks like word analogy and Named Entities Recognition.

In the these models, word representations do not distinguish the different meanings of some words: the vector representation is always the same for each word independently of context [3].

More recently, language models ELMo [10] and ULMFit [6] used recurrent bidirectional neural networks (LSTMs) to generate word vector representation of words based on the context. This contextualized word representations allowed improvements that brought these models to state-of-art in many NLP tasks.

BERT [3] is a Transformer neural network [14] which has a better integration of bidirectional context. The use of feedforward neural networks instead of recurrent allows for a much bigger model. BERT achieves on most NLP tasks better results than any previous models. The version we used, BERT-Base has 110 million parameters.

Each BERT's input is one sentence or a pair of sentences. Each sentence is previously converted to a sequence of tokens using 'WordPiece' tokenizer [13].

[3] see [10] for more details on the Word embeddings

More concretely each word is converted in one or more tokens: tokens are more meaningful when they correspond to frequent words, suffixes or prefixes. In our case, the output of BERT is a vector of 768 floats.

BERT is pre-trained simultaneously on two tasks:

- 10% of the words in a sentence are masked and BERT tries to predict them.
- Two sentences, A and B, are given and BERT must decide if B is the sentence that follows A.

Two network layers (in parallel) are added to BERT in order to train it on these tasks.

3 Our systems

In this section we describe our systems approach using BERT, a Portuguese Corpus and the ASSIN1 and ASSIN2 datasets.

3.1 BERT versions

The authors of BERT made available a pre-trained multilingual version. They used 104 languages, including Portuguese, Arabic, Russian, Chinese and Japanese. The resulting vocabulary consists of 119547 tokens. To compare with the English BERT version which vocabulary contains 30000 tokens. Therefore, the set of tokens, in the multilingual version, can not be well adapted to each language .

The example below presents the tokenization of an ASSIN2 sentence where tokens are separated by a space:

> O meni ##no e a meni ##na estão br ##in ##cando na academia ao ar livre[4]

We can see that words as "menina" and the verb "brincar" do not have a natural decomposition in tokens.

The Multilingual Portuguese tokenization leads us to suspect that the use of this BERT version can not be optimal for Portuguese and possibly for other languages. So to adjust the network resources (weights) to Portuguese we decided to build a new version by using this Multilingual version fine-tuned in a Portuguese corpus[5]. We used Portuguese Wikipedia and the journal extracts of Publico and Folha de São Paulo included in corpus CHAVE [12]. The set of tokens in this new BERT version, Multilingual fine-tuned in Portuguese, is the same set of the Multilingual version.

To try to improve the tokenization we trained BERT from scratch on the same Portuguese Corpus used to fine-tune the Multilingual version. We used a set of 32000 tokens constructed only from our Portuguese Corpus. The tokenization we obtain for the previous sentence is now:

[4] Sentence translation: The boy and the girl are playing at the gym outdoor

[5] Note that this fine-tuning uses the original Multilingual tokenization.

O menino e a menina estão brinca ##ndo na academia ao ar livre[6]

In this example, only the word "brincando" ("playing") is represented by two or more tokens, it is divided into a verb lemma and a common termination. This is one of the advantages of creating a tokens vocabulary based on the Portuguese language.

This Portuguese BERT version was one of those used in our ASSIN2 tasks submissions.

3.2 Training datasets

Since the previous ASSIN challenge data (ASSIN1) is available, we used it's train and test datasets to fine-tune the network in each task.

The ASSIN dataset for the RTE task is annotated with three-labels: entailment, paraphrase and neutral[4], for the STS task is annotated with a value between 1 and 5 as in ASSIN2. The ASSIN1 data has a subset for European Portuguese and another for Brazilian Portuguese. It is based on news with some linguistic complexity phenomena like temporal expressions.

The ASSIN2 dataset has about 10,000 sentence pairs with no linguist challenges: 6,500 used for training, 500 for validation, and 2,448 for test. It is available at https://sites.google.com/view/assin2/.

In figure 1 we present three ASSIN2 dataset examples. The tag *entailment* can have the values "Entailment"/"None" and *similarity* a value between 1 and 5.

- *entailment="None" id="12" similarity="2.4"*
  Um homem está tocando teclado[7]
  Um homem está tocando um violão elétrico[8]
- *entailment="Entailment" id="451" similarity="1.5"*
  Um cara está brincando animadamente com uma bola de meia[9]
  O homem não está tocando piano[10]
- *entailment="Entailment" id="459" similarity="4.7"*
  Um homem está andando de cavalo na praia[11]
  Um cara está montando um cavalo[12]

Fig. 1. Example of ASSIN2 data.

3.3 Neural Network

For each ASSIN2 task, systems were built by adding one layer to each pre-trained BERT version (Multilingual, Multilingual-Portuguese and Portuguese) and fine-tuned the whole network on the task.

[6] Sentence translation: The boy and the girl are playing at the gym outdoor

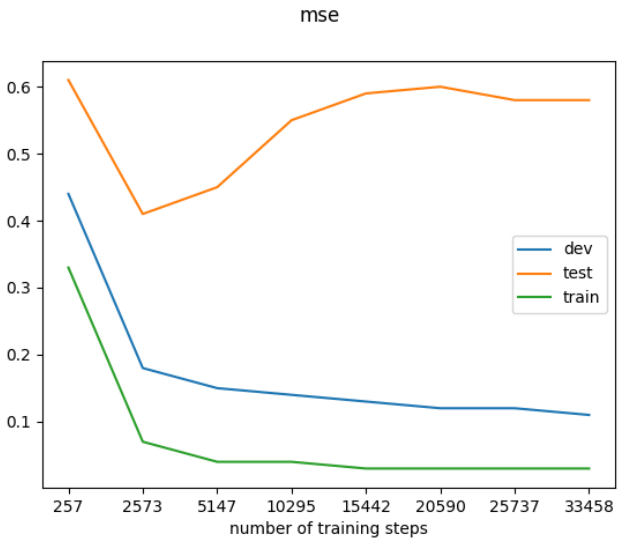

Fig. 2. Train run for the Similarity task with MSE measure.

To train our systems in each task, several epochs of mini-batch gradient descent were run until the results on the dev set started to decline. In figure 2 we present the MSE values for a typical training run of the Similarity task. In this example, we used 33458 steps to train the model. Each epoch corresponds to 257 steps. In Figure 3 we present Pearson correlation values for the same training run.

In Recognizing Textual Entailment task, the loss used for training was the binary cross-entropy, while in the Semantic Textual Similarity task, the loss used for training was Mean Square Error although the main metric for the task was Pearson Correlation.

3.4 Recognizing Textual Entailment

In this task, our starting point was always the Multilingual-Portuguese fine-tuned BERT.

Table 3.4 presents our results for 3 systems that were built with three sets of training data:

1. ASSIN2 training data (ASSIN2)
2. ASSIN1 Brazilian Portuguese training and test data plus ASSIN2 training data (ASSIN2+ASSIN1:ptbr)
3. ASSIN1 Brazilian and European Portuguese training and test data plus ASSIN2 training data (ASSIN2+ASSIN1:ptbr+pteu)

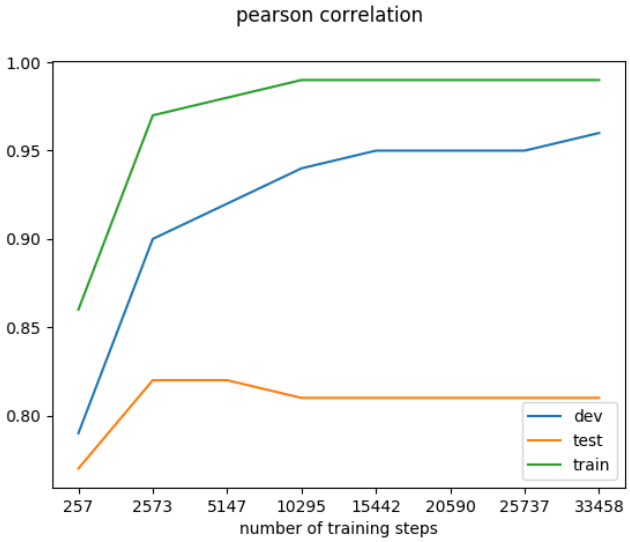

Fig. 3. Train run for the Similarity task with Pearson correlation.

The use of ASSIN2+ASSIN1:ptbr training data had slightly better results than the others as it can be seen in Table 3.4 in bold. We used 25 epochs to train the system.

Our best results are:

- When the system is evaluated on *dev*, the ASSIN2 dataset used for testing/improving our systems, F1 - 0.956, Accuracy - 95.60.
- When the system is evaluated on *test*, the ASSIN2 final competition dataset, F1 - 0.876, Accuracy - 87.58.

Surprisingly, when we use ASSIN2+ASSIN1:ptbr+pteu training data the results get worse in both sets: *dev* and *test*. This can be due to the fact that ASSIN2 dataset was built with Brazilian Portuguese.

When only ASSIN2 is used as training data, the results get even worse in both sets, this confirms that the use of more data in the training can improve our systems.

3.5 Semantic Textual Similarity

In this task we always used ASSIN1+ptbr+pteu data and the ASSIN2 training data to fine-tune some BERT version.

We tried the three BERT versions described above.

As Table 2 reports, the best results were achieved with the Multilingual version without fine-tuning to Portuguese and we used 235 epochs for training in the best submission.

Multilingual best results were:

training dataset	dev/test	F1	accuracy
ASSIN2 + ASSIN1:ptbr	test	**0.876**	**87.58**%
	dev	**0.956**	**95.60**%
ASSIN2 + ASSIN1:ptbr+ptpt	test	0.873	87.38%
	dev	0.952	95.20%
only ASSIN2	test	0.870	87.01%
	dev	0.950	95.0%

Table 1. Results of the RTE task

- When the system is evaluated on *dev*, the ASSIN2 dataset used for testing/improving our systems, Pearson Correlation - 0.968, MSE - 0.078.
- When the system is evaluated on *test*, the ASSIN2 final competition dataset, Pearson Correlation - 0.826, MSE - 0.523.

The Multilingual BERT fine-tuned Portuguese version that was submitted to ASSIN2 contained an error, so in Table 2 we present the results for the non official version. As you can see, in the Table, this version has a lower performance than the Multilingual version. The Portuguese version has the worst results, but encourage us to improve it by using more Portuguese data in the training of BERT.

BERT version	dev/test	Pearson Correlation	MSE
Multilingual	test	**0.826**	**0.523**
	dev	**0.968**	**0.078**
Multilingual fine-tuned Portuguese (non official)	test	0.821	0.552
	dev	0.965	0.080
Portuguese	test	0.809	0.625
	dev	0.938	0.15

Table 2. Results of the STS task

4 Discussion

Our results in the ASSIN2 challenge, see Table 3, first place in the Similarity task and second place in Entailment task, show that fine-tuning BERT is at the moment one of the best approaches on Portuguese semantic NLP tasks. We expect to improve the results by properly training BERT from scratch on a big and adapted Portuguese Corpus that has still to be assembled. Different versions of BERT need to be considered. We used BERT-Base but a larger version, BERT-Large (340 million parameters), achieved the better results on English NLP tasks. Given that the performance of a model depends also on the available training data and for the Portuguese language the available data is not so large as for English, we plan to experiment with ALBERT [7], a more light version of BERT.

Team	Entailment		Similarity	
	F1*	Accuracy	Pearson*	MSE
ASAPPj	0.606	62.05	0.652	0.61
ASAPPpy	0.656	66.67	0.740	0.60
IPR	0.876	87.58	**0.826**	0.52
LIACC	0.770	77.41	0.493	1.08
NILC	0.871	87.17	0.729	0.64
PUCPR	-	-	0.678	0.85
L2F/INESC	0.784	78.47	0.778	0.52
Deep Learning Brasil	**0.883**	88.32	0.785	0.59
Stilingue	0.866	86.64	0.817	0.47

* : primary metric

Table 3. Comparison of our results, IPR team, with the other teams' results

Acknowledgements

This work was partially supported by the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the project UID/MAT/00297/2019 (Centro de Matemática e Aplicações) and the grant UID/CEC/4668/2016 (Laboratório de Informática, Sistemas e Paralelismo).

References

1. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**, 1137–1155 (Mar 2003)
2. Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning. pp. 160–167. ICML '08, ACM, New York, NY, USA (2008)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)
4. Fonseca, E., Borges dos Santos, L., Criscuolo, M., Aluísio, S.: Visão geral da avaliação de similaridade semântica e inferência textual. Linguamática **8**(2), 3–13 (12 2016)
5. Hinton, G.E., McClelland, J.L., Rumelhart, D.E.: Distributed representations. In: Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations, p. 77–109. MIT Press, Cambridge, MA, USA (1986)
6. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 328–339. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)
7. Lan, Z.Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. ArXiv **abs/1909.11942** (2019)

8. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
9. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014)
10. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. 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. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018)
11. 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)
12. Santos, D., Rocha, P.: The key to the first clef with portuguese: Topics, questions and answers in chave. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) Multilingual Information Access for Text, Speech and Images. pp. 821–832. Springer Berlin Heidelberg, Berlin, Heidelberg (2005)
13. Schuster, M., Nakajima, K.: Japanese and korean voice search. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 5149–5152 (2012)
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 6000–6010. NIPS'17, Curran Associates Inc., USA (2017)