# Multilingual Transformer Ensembles for Portuguese Natural Language Tasks

Ruan Chaves Rodrigues[1], Jéssica Rodrigues da Silva[2], Pedro Vitor Quinta de Castro[1], Nádia Félix Felipe da Silva[1], and Anderson da Silva Soares[1]

[1] Institute of Informatics, Federal University of Goiás, Brazil
ruanchaves93@gmail.com
{pedrovitorquinta, nadia, anderson}@inf.ufg.br
[2] Department of Computer Science, Federal University of São Carlos, Brazil
jsc.rodrigues@gmail.com

**Abstract.** Due to the technical gap between the language models available for low-resource languages and the state-of-the-art models available in English and Chinese, a simple approach that deploys automatic translation and ensembles predictions from Portuguese and English models is competitive with monolingual Portuguese approaches that may demand task-specific preprocessing and hand-crafted features. We performed our experiments on ASSIN 2 – the second edition of the *Avaliação de Similaridade Semântica e Inferência Textual* (Evaluating Semantic Similarity and Textual Entailment). On the semantic textual similarity task, we performed multilingual ensemble techniques to achieve results with higher Pearson correlation and lower mean squared error than BERT-multilingual, and on the textual entailment task, BERT-multilingual could be surpassed by automatically translating the corpus into English and then fine-tuning a large RoBERTa model over the translated texts.

**Keywords:** Semantic textual similarity · Textual entailment · Transformer architecture

## 1 Introduction

Although recent advances in Transformer architectures [8] have significantly improved the state-of-the-art for several downstream natural language processing tasks, such models usually require training to be performed with billions of parameters on massive datasets. And as only major companies and research centers can afford this process, a linguistic bias has arisen in the field: most state-of-the-art models are available only for the languages that predominate on the areas where these entities are based, namely English and Chinese [14].

We demonstrate that, due to the current gap between the models available for Portuguese and these two languages, a simple approach that deploys automatic translation and ensembles predictions from Portuguese and English models is competitive with monolingual approaches that may demand task-specific preprocessing and hand-crafted features to achieve the same accuracy.

For this purpose, we examine the effectiveness of multiple ensemble techniques on the Semantic Textual Similarity (STS) and Recognizing Textual Entailment (RTE) tasks proposed on the ASSIN 2 dataset[3] [13], while fine-tuning BERT-multilingual [8] over the original Portuguese datasets and RoBERTa [11] over the automatic translation of the original datasets into English. In order to ensure the reproducibility of our results, we have open-sourced our models and experiments[4].

In section 2 we describe some of the related work for Transformer architectures in low-resource languages. The approaches to the Transformer architecture investigated in this paper are described in section 3. Section 4 shows the approach considered for ensemble modeling. The experiments carried out for evaluating our model for the ASSIN2 tasks are described in section 5. Section 6 finishes this paper with its conclusions and proposals for future work.

## 2 Related Work

As Transformer architectures are reasonably recent, there are few works which have attempted to apply large pretrained English models to other linguistic domains. Fonseca fine-tuned the GPT-2 model in Portuguese, and his results[5] indicate that this can become a viable approach for tackling text generation problems in the Portuguese language.

However, the strategy of automatically translating a dataset into a foreign language and then ensembling models from distinct linguistic domains to address a natural language processing task has already been throughly researched in the past. In the field of sentiment analysis, Wan et al. [20] successfully improved the accuracy of sentiment analysis on Chinese customer reviews by ensembling the predictions produced from both the original Chinese dataset and its translation into English, achieving combined results which were better than either approach considered in isolation.

Under different circumstances, Araujo et al. [3] has shown that, for some particular languages, simply translating the dataset into English and applying the state-of-the-art sentiment analysis methods available for the English language yielded better results than the existing language-specific approaches evaluated during their experiments.

Tian et al. [19] applied a similar approach for the Spanish STS task proposed at SemEval-2017. After automatically translating the entire dataset, they applied an ensemble of the state-of-the-art techniques available for the task. However, they also included in his ensemble nine features that measured the quality of the translated text, namely *BLEU*, *GTM-3*, *NIST*, *-WER*, *-PER*, *Ol*, *-TERbase*, *METEOR-ex* and *ROUGE-L*.

Belinkov et al. [4] have demonstrated that current Neural Machine Translation (NMT) models, which include the external translation service utilized in

[3] `https://sites.google.com/view/assin2/`
[4] `https://github.com/ruanchaves/assin/`
[5] `https://medium.com/ensina-ai/ensinando-portugues-ao-gpt-2-d4aa4aa29e1d`

our experiments, are highly vulnerable to adversarial noise, as well as natural noise produced by native speakers of the language. Such noise cannot be easily addressed by existing tools, such as spell checkers. Therefore, having features that account for the quality of the translation are certainly helpful while building ensembles with features extracted from translated text either through deep learning methods or traditional natural language processing techniques.

Although machine translation may be effective, it may not be suitable for certain industrial settings, specially when one cannot afford to call an external translation service several times. Tang et al. [18] trained a shared multilingual encoder that successfully leveraged knowledge from English labeled data in Spanish STS tasks, without requiring the annotation of hand-crafted features, or having to resort to machine translation during inference. His results were consistently better than those achieved by monolingual approaches, and they were able to reach the same performance as the machine translation methods that were evaluated during his experiments. Shared multilingual encoders for STS tasks have also been trained by Chidambaram et al. [6].

Compared to the Portuguese language, a relatively large amount of Transformer models have been trained for Spanish and other major Romance languages. To the best of our knowledge, no large Transformer model trained exclusively for Portuguese has ever been made publicly available, and no other paper has ever investigated how Transformer architectures and machine translation can be leveraged for natural language processing tasks in the Portuguese language.

## 3 Transformer architectures

Only two transformer architectures were considered for our experiments: BERT-multilingual and RoBERTa. We started from pretrained models, which were fine-tuned to a STS task through the Transformers library developed by Hugging Face [22].

### 3.1 BERT-multilingual

BERT stands for Bidirectional Encoder Representations from Transformers [8]. Devised as an alternative to unidirectional language models, BERT pretrains deep bidirectional representations that are simultaneously learned both for left and right contexts. As a result, the last layers of a pretrained BERT model can be fine-tuned to specific natural language tasks without requiring any substantial modifications on its architecture.

In order to define its prediction goal, BERT utilizes two training strategies. One of them is called Masked LM (MLM). Before being fed into the model, 15% of all tokens are replaced by either a random token or a fixed `[MASK]` token. In this way, the model has the training goal to improve its ability to predict the original token that occupied these positions.

For its second strategy, the model is pretrained for a binarized Next Sentence Prediction (NSP) task with sentences `A` and `B`, where `B` has an equal chance of being either a randomized sentence or the actual next sentence after `A`.

In our experiments, we utilized the multilingual, cased BERT model, trained with 110 million parameters on 104 languages, including Portuguese. These languages were chosen because they are the ones with the largest Wikipedias, and the entire Wikipedia dump for each language was taken as the training data. Although overrepresented languages were downsampled during the training stage, no downsampling was performed among dialects within the same language. It should also be noticed that Wikipedia does not impose any Portuguese dialect for any of its articles, and therefore, it is not possible to differentiate between them on its dumps.

The flexibility of BERT has contributed to establish it as one of the main paradigms for advancing the state-of-the-art in natural language processing tasks. As of today, most state-of-the-art models are either directly or indirectly based on the original BERT architecture [17].

### 3.2 RoBERTa

After the release of BERT, several authors studied the model and pointed out architectural choices that should be reconsidered. The main issues lied on its training strategies, the Masked LM and Next Sentence Prediction tasks. In fact, some of the limitations found within these strategies had already been acknowledged on the paper that first presented the BERT architecture [8].

In our experiments, we have used one of the models that improved on BERT while trying to reach a proper training strategy: RoBERTa, which stands for Robustly Optimized BERT Pretraining Approach [11]. While maintaining the same base architecture as BERT, RoBERTa improves on its results by removing the next sentence prediction task, dynamically changing the masking pattern applied to the training data, and increasing the training time, the size of the batches, the volume of data and the input sequence length.

As demonstrated by [2], the Next Sentence Prediction (NSP) objective hurts the performance of multilingual versions of BERT even more than it does when it is trained on a single language. Therefore, in situations where both models receive an input of the same quality, we can only expect that RoBERTa will consistently perform better than BERT-multilingual.

We can also expect better performance even when comparing multilingual versions of RoBERTa with multilingual versions of BERT. In fact, Conneau et al. [7] recently released the a model called XLM-RoBERTa, which not only achieves results better than BERT-multilingual, but also obtains results competitive with monolingual versions of RoBERTa and other state-of-the-art monolingual models after monolingual fine-tuning on the GLUE and XNLI benchmarks.

Although there are versions of RoBERTa available which have been previously fine-tuned on MNLI and the output of GPT-2, we performed our experiments only on a pretrained version of RoBERTa which was based on the architecture of BERT-large.

## 4 Ensemble Techniques

For both STS and RTE tasks, BERT-multilingual was fine-tuned on the original, Portuguese dataset, and RoBERTa was fine-tuned on the same dataset after it was automatically translated into English. In this way, it can be said that a stronger model (RoBERTa) is receiving an input of lower quality (an automatically translated dataset) while the weaker model (BERT-multilingual) is receiving an input of better quality (the original dataset).

After fine-tuning, we tested two distinct ensemble techniques to combine the predictions generated by both Transformers. The first one was averaging: the fine-tuned models generated predictions for the test set, and then these scores were averaged through an arithmetic mean to produce the final submission.

The second technique utilized is most commonly known as stacking. A careful and diagrammatic description of the stacking ensemble technique has been made by Kyriakides et al. [10]. In this approach, BERT and RoBERTa are called base learners (or level 0 learners), which will provide metadata to train a meta-learner (or level 1 learner), which in our particular setup happens to be a Multilayer Perceptron (MLP).

The metadata was generated through K-fold cross validation over the entire training set. Both BERT and RoBERTa were fine-tuned on their corresponding training sets for each possible combination of $K - 1$ folds. Therefore, an ensemble made through K-fold stacking will generate K distinct BERT models fine-tuned on the original dataset, and K distinct RoBERTa models fine-tuned on the translated dataset.

After generating $2K$ fine-tuned models, each one of them will generate predictions for the fold from the training set which was missing during its fine-tuning process. Although all labels from the training set are known to us, they are temporarily ignored for the sake of generating the metadata for the MLP, and we concatenate the predictions from all fine-tuned models. As a result, we will have produced metadata for the entire training set, which will be used to train the MLP.

The MLP, our meta-learner, is trained while taking the metadata generated by each transformer as its input, and the gold score for the sentence pairs of the training set as the training goal. During this process, the MLP will try to learn how to properly weight the contributions of each transformer on each score range.

After training the MLP, we fine-tune one BERT and one RoBERTa model over their entire corresponding training sets, without considering any division into folds. These models produce predictions, which are combined by the trained MLP to produce the final predicted scores. If successful, this trained meta-learner will have produced scores that are more accurate than the predictions generated by either base learner considered in isolation.

We did not test any averaging techniques other than a simple arithmetic mean, and our MLP was made only of two Dense layers of 64 neurons with a ReLU activation function followed by a single Dense layer of one neuron with a linear activation function.

It should be noticed that there is a trade-off between improving the overall performance of the ensemble and the amount of computational resources required to fine-tune transformer models for a higher amount of folds. There are also diminishing returns for increasing the amount of folds: although the accuracy will initially increase, it is expected to stabilize after the folds become small enough for the MLP to learn the general patterns present on the metadata.

## 5 Experimental Evaluation

We evaluated the performance of our models on the ASSIN 2 dataset. Part of our experiments were submitted during the ASSIN 2 workshop, and then their accuracy was compared with the results obtained by the other participants. In this section, we examine some particular features of the dataset and how they influenced the obtained results. These results are described both in their entirety and, in the case of the STS task, also separately for five distinct gold score intervals.

### 5.1 Dataset

The dataset consists of sentence pairs in the Portuguese language, with human-annotated scores for Semantic Textual Similarity and Textual Entailment. The Semantic Textual Similarity task is available in the GLUE benchmark [21] as the Semantic Textual Similarity Benchmark (STS-B) [1], and the Textual Entailment task is available as Recognizing Textual Entailment (RTE) [5]. Both tasks are present on every ASSIN dataset.

During our experiments, RTE was treated as being a regression task, exactly like the STS task, rather than being a classification task. In other words, while STS was treated as the task of predicting scores ranging from 1 to 5, RTE was similarly treated as the task of predicting scores ranging from 0 to 1, taking "None" as 0, "Entailment" as 1. And then, before submitting our files to the official evaluation script for the ASSIN 2 dataset, the entailment scores were rounded up and converted back to their corresponding labels.

Datasets for Brazilian and European Portuguese were released during the first edition of ASSIN [9]. Their sentence pairs were collected from Google News, and had the linguistic complexity that can be expected from real-world sources. However, during ASSIN 2, a single dataset was released which was purposefully simple, without any named entities, instances of indirect speech or verbs not conjugated in the present. Furthermore, a certain portion of the sentence pairs in ASSIN 2 was translated from existing English datasets into Portuguese. In this way, we were not able to distinguish such pairs from those which were created from scratch, as there are not any labels on the dataset itself that indicate this feature. We translated the ASSIN 2 dataset into English through Google Cloud Translation API[6] under its default settings.

[6] `https://cloud.google.com/translate/docs/`

It should be noticed that we did not apply any preprocessing steps to the datasets other than those required by the Transformer models themselves. We also did not take any measures to modify or increase the quality of the translated sentence pairs after they had been retrieved from the translation API.

### 5.2 Results

The Transformers (BERT and RoBERTa) were fine-tuned both on the standard training and validation sets for ASSIN 2. After fine-tuning, they produced their own predictions for each one of the available test sets. These predictions were submitted to the standard evaluation script for the ASSIN datasets, both in isolation and combined through two different ensemble techniques: averaging and 5-fold stacking.

It can be seen on Table 1 that the best Pearson correlation for the STS task has been achieved by combining the predictions from both Transformers through an aritmetic mean, although 5-fold stacking achieved a mean squared error considerably lower than any other of the evaluated approaches. For the entailment task, RoBERTa singlehandedly performed better than either BERT or any of the ensembles. However, 5-fold stacking achieved results pratically equivalent to RoBERTa.

Table 1: Results for evaluation on the tasks provided by ASSIN 2. For the STS task, the predictions were ranked by their Pearson correlation coefficient ($\rho$) and mean squared error (MSE) relative to the gold scores. For evaluation on the RTE task, the predictions were ranked by their accuracy (Acc), Macro-F1 score (F1) and Matthews correlation coefficient (MCC). Arrows indicate whether lower (↓) or higher (↑) is better.

| **Model** | **Similarity** | | **Entailment** | | |
|---|---|---|---|---|---|
| | $\rho$ (↑) | **MSE** (↓) | **Acc** (↑) | **F1** (↑) | **MCC** (↑) |
| BERT-multilingual | 0.75 | 1.20 | 0.8190 | 0.82 | 0.659 |
| RoBERTa | 0.81 | 0.77 | **0.8840** | **0.88** | **0.772** |
| Ensemble ( averaging ) | **0.83** | 0.91 | 0.8476 | 0.84 | 0.720 |
| Ensemble ( stacking, 5-fold ) | 0.78 | **0.59** | 0.8832 | 0.88 | 0.771 |

### 5.3 Discussion

Reimers et al. [15] demonstrates that the Pearson correlation coefficient can be misleading and especially ill-suited to predict the best STS system for natural

language processing tasks. In plagiarism detection, for instance, the STS system may receive documents that have been previously pre-filtered. In this way, it will only analyze documents that score above a certain threshold. And for semantic search, Reimers [7] mentions that, as Okapi BM25 has a lower false positive probability than BERT and other Transformer architectures, query results should be pre-filtered through Okapi BM25 to build a clean candidate set, and then BERT can be used to perform reranking among the selected candidates.

With these concerns in mind, we have measured both Pearson correlation and the mean squared error individually for five distinct gold score intervals.

Regarding the mean squared error, in Figure 1 we can see that, while the performance of BERT-multilingual linearly improves for higher gold scores, RoBERTa performs exceptionally better than all other models for the lowest score range between 1.0 and 1.4 . After this range, its performance suddenly decreases. And after decreasing, it also starts to linearly improve.

As sentence pairs on the lowest gold score intervals are very different from each other, one possible explanation is that they were able to maintain their high degree of dissimilarity even after automatic translation. In this way, RoBERTa was able to provide results closer to its performance for sentence pairs which were originally written in English; and as expected, such results are consistently better than what can be achieved through BERT-multilingual.

Although averaging performed slightly better than 5-fold stacking on the lowest score range, 5-fold stacking was able to provide a mean squared error lower than any other model on all other score ranges.

Regarding the Pearson correlation, it is interesting to notice that RoBERTa provided results which were superior or otherwise competitive with all other models except on the higher score interval between 4.5 and 5.0, and BERT-multilingual has a performance inferior to all other models except on the same higher score interval. Although 5-fold stacking failed to notice such complementary behavior, it was possible to achieve results competitive with both models for all score ranges by simply averaging their predictions.

As we treated the RTE task as if it were a STS task with only two scores, the results obtained by RoBERTa on this task can be explained by its ability to accurately determine when two automatically translated sentence pairs are totally dissimilar to each other, that is, when their entailment relationship is equal to `None`. In other words, RoBERTa has a lower false positive probability than BERT-multilingual or our evaluated ensemble approaches. As a result, RoBERTa achieved not only high accuracy and Macro-F1 scores, but also a high Matthews correlation coefficient.

Therefore, for the RTE task, an ensemble will not be necessary, and we can simply take the predictions produced by RoBERTa as our final result. However, an ensemble made through 5-fold stacking is able to achieve equivalent results, as it can recognize that BERT-multilingual performs consistently worse than RoBERTa on this task, and thus it minimizes the weight given to the predictions made by BERT-multilingual as much as possible.

[7] `https://github.com/huggingface/transformers/issues/876#issuecomment-536792338`

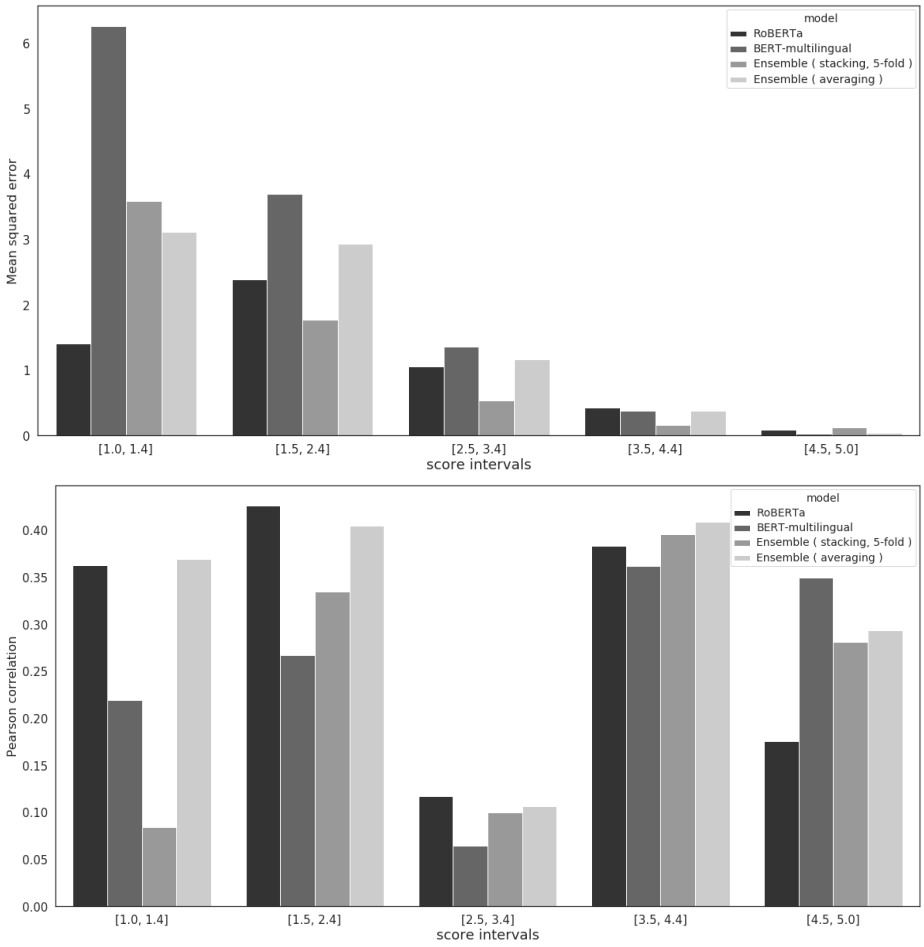


Fig. 1: We break down the dataset into five parts determined by the intervals in which their gold scores are located. Then we calculate the Pearson correlation and the mean squared error of all experiments listed on Table 1 separately for each one of these intervals.

## 6 Conclusions

We have discovered that, for the evaluated Portuguese RTE task, automatically translating the dataset into English and then fine-tuning a large RoBERTa model over the translated dataset can produce better results than BERT-multilingual or monolingual approaches that rely only on the resources natively available for the Portuguese language.

And although the same results could not be achieved in isolation by RoBERTa on the STS task, state-of-the-art results in the Portuguese language for this task can be achieved by combining the predictions made by BERT-multilingual and RoBERTa through an adequate ensemble technique, which in some cases may be as simple as taking the arithmetic mean of both predictions.

In spite of the simplicity of our approach, we have achieved the best results for the RTE task on the ASSIN 2 workshop. Later on, we also found out that simply taking the arithmetic mean of the predictions generated by BERT and RoBERTa was able to surpass all the Pearson correlation scores which had been achieved by the participants to the STS task. We did not perform any preprocessing on the training data, either before or after automatic translation. We also did not take into account any features other than the predictions generated by the Transformer architectures as they have been implemented in standard libraries.

Therefore, we believe that there is room for significant improvements to our results. For both tasks, it may be convenient to test more sophisticated techniques for ensemble building [16] [19]. And although our research was limited to the ASSIN 2 dataset, it may be interesting to extend the same experiments to its first edition, ASSIN 1. We should also investigate which linguistic features were easily learned by our models, and which ones were not.

For future experiments, new English models that reach the top positions on the GLUE Benchmark Leaderboard[8] for the STS-B and RTE tasks should be considered, and these models may be combined or compared with Transformers trained in Romance languages. For instance, experiments may be performed on CamemBERT [12], with datasets automatically translated from Portuguese to French.

However, as a long-term strategy, we should also consider solutions which entirely avoid the inevitable limitations of machine translation, such as training our own Transformer models exclusively for the Portuguese language.

## Acknowledgments

The authors would like to acknowledge the support of the Institute of Informatics at the Federal University of Goiás, Americas Health Co., B2W Digital, Datalawyer and Data-H Data Science and Artificial Intelligence.

[8] `https://gluebenchmark.com/leaderboard/`

# References


[1] Agirre, E., M'arquez, L., Wicentowski, R.: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007). Association for Computational Linguistics, Prague, Czech Republic (June 2007)

[2] Anonymous: Cross-lingual ability of multilingual {bert}: An empirical study. In: Submitted to International Conference on Learning Representations (2020), https://openreview.net/forum?id=HJeT3yrtDr, under review

[3] Araújo, M., Pereira, A., Benevenuto, F.: A comparative study of machine translation for multilingual sentence-level sentiment analysis. Information Sciences (2019), http://www.sciencedirect.com/science/article/pii/S0020025519309879

[4] Belinkov, Y., Bisk, Y.: Synthetic and natural noise both break neural machine translation (2017)

[5] Bentivogli, L., Dagan, I., Dang, H.T., Giampiccolo, D., Magnini, B.: The fifth PASCAL recognizing textual entailment challenge (2009)

[6] Chidambaram, M., Yang, Y., Cer, D., Yuan, S., Sung, Y.H., Strope, B., Kurzweil, R.: Learning cross-lingual sentence representations via a multi-task dual-encoder model (2018)

[7] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale (2019)

[8] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018)

[9] Fonseca, E., Santos, L., Criscuolo, M., Aluísio, S.: Visão geral da avaliação de similaridade semântica e inferência textual. Linguamática **8**(2), 3–13 (2016)

[10] Kyriakides, G., Margaritis, K.G.: Hands-On Ensemble Learning with Python (07 2019)

[11] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019)

[12] Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., Villemonte de la Clergerie, É., Seddah, D., Sagot, B.: CamemBERT: a Tasty French Language Model. arXiv e-prints arXiv:1911.03894 (Nov 2019)

[13] 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)

[14] Rei, M.: The geographic diversity of nlp conferences. https://web.archive.org/web/20191009171059/http://www.marekrei.com/

blog/geographic-diversity-of-nlp-conferences/ (2019), accessed: 2019-10-09
[15] Reimers, N., Beyer, P., Gurevych, I.: Task-oriented intrinsic evaluation of semantic textual similarity. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 87–96. The COLING 2016 Organizing Committee, Osaka, Japan (Dec 2016), https://www.aclweb.org/anthology/C16-1009
[16] Sagi, O., Rokach, L.: Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **8**(4), e1249 (2018)
[17] Storks, S., Gao, Q., Chai, J.Y.: Recent advances in natural language inference: A survey of benchmarks, resources, and approaches (2019)
[18] Tang, X., Cheng, S., Do, L., Min, Z., Ji, F., Yu, H., Zhang, J., Chen, H.: Improving multilingual semantic textual similarity with shared sentence encoder for low-resource languages (2018)
[19] Tian, J., Zhou, Z., Lan, M., Wu, Y.: ECNU at SemEval-2017 task 1: Leverage kernel-based traditional NLP features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 191–197. Association for Computational Linguistics, Vancouver, Canada (Aug 2017). https://doi.org/10.18653/v1/S17-2028, https://www.aclweb.org/anthology/S17-2028
[20] Wan, X.: Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 553–561. EMNLP '08, Association for Computational Linguistics, Stroudsburg, PA, USA (2008), http://dl.acm.org/citation.cfm?id=1613715.1613783
[21] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: A multi-task benchmark and analysis platform for natural language understanding (2019), in the Proceedings of ICLR.
[22] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J.: Huggingface's transformers: State-of-the-art natural language processing (2019)