Organizing the ASSIN 2 Shared Task

Livy Real[1], Erick Fonseca[2], and Hugo Gonçalo Oliveira[3][0000−0002−5779−8645]

[1] B2W Digital/Grupo de Linguística Computacional – University of São Paulo, livyreal@gmail.com
[2] Instituto de Telecomunicações, Lisboa, Portugal, erick.fonseca@lx.it.pt
[3] CISUC, University of Coimbra, Portugal, hroliv@dei.uc.pt

Abstract. We describe ASSIN 2, the second edition of a task on the evaluation of Semantic Textual Similarity (STS) and Textual Entailment (RTE) in Portuguese. The ASSIN 2 task uses as dataset a collection of pairs of sentences annotated with human judgments for textual entailment and semantic similarity. Interested teams could participate in either of the tasks (STS or RTE) or in both. Nine teams participated in STS and eight in the RTE. A workshop on this task was collocated with STIL 2019, in Salvador, Brazil. This paper describes the ASSIN 2 task and gives an overview of the participating systems.

Keywords: Shared Task · Semantic Textual Similarity · Recognizing Textual Entailment · Natural Language Inference · Portuguese

1 Introduction

ASSIN[4] stands for *Avaliação de Similaridade Semântica e INferência Textual* (Evaluating Semantic Similarity and Textual Entailment) and is an evaluation shared task in the scope of the computational processing of Portuguese. In fact, as in the first ASSIN (hereafter, ASSIN 1), ASSIN 2, the second edition of this task, consisted of two different tasks: Recognizing Textual Entailment (RTE), also known known as Natural Language Inference (NLI), and Semantic Textual Similarity (STS).

Following ASSIN 1 [11], ASSIN 2 offered the interested community a new benchmark for computational semantic tasks in Portuguese, thus advancing the state-of-the-art. The shared task was collocated with the Symposium in Information and Human Language Technology (STIL) in Salvador, BA, Brazil, with a workshop held on October, 15^{th}, 2019. A short paper on ASSIN 2 was published in the Proceedings of the 14^{th} International Conference on the Computational Processing of Portuguese [24].

Briefly, as defined in a SemEval 2012 task [1] on the topic, Semantic Textual Similarity (STS) 'measures the degree of semantic equivalence between two sentences'. Given a set of pairs of textual fragments, often sentences, this kind of task asks for the assignment of a score for their similarity. The similarity scale

[4] *Assim* in Portuguese means 'in the same way', so arguably an adequate name for a similarity task.

adopted in ASSIN 2 ranges between 1 and 5, with 1 meaning that the sentences are totally different and 5 that they have virtually the same meaning. The pair *O cachorro está pegando uma bola azul/Uma bola azul está sendo pega pelo cachorro*[5] is an example of a pair scored with 5, while *A menina está andando de cavalo/O menino está borrifando as plantas com água* [6] is scored 1.

Recognizing Textual Entailment (RTE), or Natural Language Inference (NLI), is the task of predicting whether a given text entails another (i.e., a premise implies a hypothesis). The **entailment** relation happens when, from the premise [A], we can infer that another sentence [B] is also true. That is, from [A] we can conclude [B]. For the pair [A] *Um macaco está provocando um cachorro no zoológico/* [B] *Um cachorro está sendo provocado por um macaco no zoológico*[7], we say A `entails` B. While for the pair [A]*Um grupo de meninos em um quintal está brincando e um homem está de pé ao fundo/* [B]*Os meninos jovens estão brincando ao ar livre e o homem está sorrindo por perto* [8], there is no entailment relation from A to B[9].

We follow the tradition of shared tasks for RTE that can be traced back to 2005 with the first Pascal Challenge [7], targeting RTE in a corpus of 1,367 pairs annotated for `entailment` and `non-entailment` relations. Back then, the best teams (MITRE and Bar Ilan teams) achieved an accuracy of 0.586. In the next Pascal Challenges, different corpora and task designs were tried: paragraphs were used instead of short sentences (Challenge 3 [12]); contradictions were added to the data (Extended Challenge 3[27]); non-aligned texts were given to the participants (Challenges 6 and 7) and, more recently, the task was presented as multilingual [22,23].

Regarding STS, shared tasks for English go back to SemEval 2012 [1]. Recently, in 2017 [5], Arabic and Spanish were also included. SemEval 2014 also included a related task on Compositionality, that put together both Semantic Relatedness and Textual Entailment [19], which we modeled our dataset after. For both STS and RTE, this task used the SICK corpus ('Sentences Involving Compositional Knowledge') as its data source, the first corpus in the order of 10,000 sentence pairs annotated for inference.

In 2015, SNLI, a corpus with more than 500,000 human-written English sentences annotated for NLI was released[3] and, in 2017, RepEval [21] included the MultiNLI corpus, with more than 430,000 pairs annotated for NLI, covering different textual genres.

---

[5] The dog is catching a blue ball/A blue ball is being caught by the dog.

[6] The girl is riding the horse/The boy is spraying the plants with water.

[7] A monkey is teasing a dog at the zoo/A dog is being teased by a monkey at the zoo.

[8] A group of boys in a backyard are playing and a man is standing in the background/ Young boys are playing outdoors and the man is smiling nearby.

[9] One could possibly think there is an entailment relation among these sentences, since 'meninos' (boys) are always 'meninos jovens' (young boys) and that probably the man standing would be also smiling. Since it is also equally possible that the man nearby is not smiling, this pair is considered a non-entailment, that is, it is possible that the two scenes happens at the same time, but it is not necessary.

When it comes to Portuguese processing, data availability and shared tasks for semantic processing are still starting to become popular. In 2016, ASSIN [11] was the first shared task for Portuguese STS and RTE. Its dataset included 10,000 pairs of annotated sentences, half in European Portuguese and half in Brazilian Portuguese. ASSIN 2 follows the goal of ASSIN by offering a new computational semantic benchmark to the community interested in computational processing of Portuguese.

2 Task and Data Design

When designing ASSIN 2, we considered the previous experience of ASSIN and made some changes towards an improved task. This section describes the data used in the ASSIN 2 collection, its annotation process, decisions taken and the main differences to ASSIN 1. It ends with a brief schedule of ASSIN 2.

2.1 Data Source

The ASSIN 1 dataset is based on news and imposes several linguistic challenges, such as temporal expressions and reported speech. Following thoughts of the ASSIN 1 organization [11], we opted to have a corpus specifically created for the tasks, as SNLI and MultiNLI, and containing only simple facts, as SICK. Therefore, the ASSIN 2 data was based on SICK-BR [25], a translation and manual adaptation of SICK [19], the corpus used in SemEval 2014, Task 1. SICK is known to be based on captions of pictures and to have fewer complex linguistic phenomena, which perfectly suits our purposes. Since ASSIN 2 collection is made upon SICK-BR, it only contains the Brazilian variant of Portuguese.

2.2 Data Balancing

Another goal considered in data design was to have a balanced corpus in terms of RTE labels. Both ASSIN and SICK-BR data are unbalanced, offering many more neutral pairs than entailments. Even if this is more representative of the reality of language usage by people, this is undesirable for machine learning techniques. Since SICK-BR has less than 25% of entailment pairs, we had to create and annotate more of them. To create such pairs we followed a semi-automated strategy, starting from entailment SICK-BR pairs and changing synonyms or removing adverbial or adjectival phrases in those. All generated pairs were manually revised. We also manually created pairs hoping they would be annotated as entailments, but trying as much as possible not to introduce artifact bias [14].

2.3 Data Annotation

All the annotators involved in the ASSIN 2 annotation task have linguistic training, being them professors, linguistics students or computational linguists. We

would like to express the deepest appreciation to Alissa Canesim (Universidade Estadual de Ponta Grossa), Amanda Oliveira (Stilingue), Ana Cláudia Zandavalle (Stilingue), Beatriz Mastrantonio (Stilingue - Universidade Federal de Ouro Preto), Carolina Gadelha (Stilingue), Denis Owa (Pontifícia Universidade Católica de São Paulo), Evandro Fonseca (Stilingue), Marcos Zandonai (Universidade do Vale do Rio dos Sinos), Renata Ramisch (Núcleo Interinstitucional de Linguística Computacional - Universidade Federal de São Carlos), Talia Machado (Universidade Estadual de Ponta Grossa) for taking part of this task with the single purpose of producing open resources for the community interested on the computational processing Portuguese. We also thank the Group of Computational Linguistics from University of São Paulo for making available SICK-BR, which served as the base for the ASSIN corpus.

All the pairs created for ASSIN were annotated by at least four native speakers of Brazilian Portuguese. The annotation task was conducted using an online tool prepared for the RTE and STS tasks, the same as in ASSIN 1.

For the RTE task, only pairs annotated the same way by the majority of the annotators were actually used in the dataset. It means that at least three of four annotators agreed on the RTE labels present in ASSIN 2 collection. For the STS task, the label is the average of the score given by all the annotators. The final result was a dataset with about 10,000 sentence pairs: 6,500 used for training, 500 for validation, and 2,448 for test, now available at `https://sites.google.com/view/assin2/`.

Since we wanted to have a balanced corpus and a sound annotation strategy, we opted for having only two RTE labels; `entailment` and `non-entailment`. Differently from ASSIN 1, we did not use the `paraphrase` label because a paraphrase happens when there is a double-entailment, being, somehow, unnecessary to annotate a double-entailment with a third label. This was further motivated by the results of ASSIN 1, where systems showed much difficulty to outperform the proposed baselines, which were the same as in ASSIN 2. For example, no participant run did better than the RTE baseline in Brazilian Portuguese. Thus, we decided to pursue a new task design having in mind its utility to the community.

In fact, our original intent was to follow a tradition in inference that pays attention to contradictions as much as to entailments, as Zaenen et al. [28] and de Marneffe et al. [20], as well as most recent datasets. However, having a soundly annotated corpus for contradictions is not a trivial task. Firstly, defining contradictions and having functional guidelines for the phenomenon is a task on its own. While recent datasets aim to have a "human" perspective of the phenomenon [4], semanticists and logicians have already pointed out that this lay perspective on contradictions can lead to much noise on inference annotation[10], especially when considering contradictions' annotation. For example, the work of Kalouli et al. [16] shows that almost 50% of the contradictions in the SICK dataset, around 15% of all the pairs, do not follow the basic 'logical' assumption that, if the premise (sentence A) contradicts the hypothesis (sentence B), the hypothesis (B) must also contradict the premise (A). After all, contradic-

[10] See Crouch (et al.)-Manning controversy for details on this point [28,18,6].

tions should be symmetric. Secondly, considering that we used SICK-BR as the base of our dataset, we would have needed to correct all the contradictions that were already in SICK, following Kalouli et al. [16], that finds many inconsistencies on contradictions annotation. Another point for excluding contradictions in ASSIN 2 is that we would also not have a balanced corpus among the labels, since SICK (and SICK-BR) has less than 1,500 contradictions in a corpus of 10,000 pairs.

Table 1. Examples of ASSIN 2 data

Premise	**Hypothesis**	**RTE**	**STS**
O cachorro castanho está correndo na neve	*Um cachorro castanho está correndo na neve*	Entails	5
Alguns animais estão brincando selvagemente na água	*Alguns animais estão brincando na água*	Entails	4.4
Dois meninos jovens estão olhando para a câmera e um está pondo sua língua para fora	*Duas jovens meninas estão olhando uma câmera e uma está com a língua para fora*	None	3.7
A menina jovem está soprando uma bolha que é grande	*Não tem nenhuma menina de rosa girando uma fita*	None	2.1
Um avião está voando	*Um cachorro está latindo*	None	1

Table 1 illustrates the dataset with five annotated pairs. The collection of ASSIN 2 is distributed in the same XML format adopted in ASSIN 1. Sentence pairs are marked by the `<pair>` element that includes elements `<t>` and `<h>`, respectively for the first and second sentence. For illustrative purposes, Figure 1 represents the first example in Table 1.

```
<pair entailment="Entailment" id="681" similarity="5">
  <t>O cachorro castanho está correndo na neve</t>
  <h>Um cachorro castanho está correndo na neve</h>
</pair>
```

Fig. 1. Data format of the ASSIN 2 collection.

2.4 Schedule

ASSIN 2 was announced on May 2019, in several NLP mailing lists. On June 2019, a Google Group was created for communication between the organization and participants or other interested people (`https://groups.google.com/forum/#!forum/assin2`). Training and validation data were released on 16^{th} June and testing data on 16^{th} September, which also marked the beginning of

the evaluation period. The deadline for result submission was 10 days later, on 26^{th} September, and the official results were announced a few days after this.

A physical workshop where ASSIN 2 was presented, as well as some participations, was held on 15^{th} October 2019, in Salvador, Brazil, collocated with the STIL 2019 symposium[11]. On 2^{nd} March 2020, a second opportunity was given to participants to present their work in the POP2 workshop, in Évora, Portugal, collocated with the PROPOR 2020 conference[12].

2.5 Metrics

As it happened in ASSIN 1 and in other shared tasks with the same goal, systems' performance on the RTE task was measured with the macro F1 of precision and recall as the main metric. For STS, performance was measured with the Pearson correlation index (ρ) between the gold and the submitted scores, with Mean Squared Error (MSE) computed as a secondary metric. The evaluation scripts can be found at `https://github.com/erickrf/assin`.

3 Participants and Results

ASSIN 2 had a total of nine participating teams, five from Portugal and four from Brazil, namely:

- CISUC-ASAPPj (Portugal)
- CISUC-ASAPPy (Portugal)
- Deep Learning Brasil (Brazil)
- IPR (Portugal)
- L2F/INESC (Portugal)
- LIACC (Portugal)
- NILC (Brazil)
- PUCPR (Brazil)
- Stilingue (Brazil)

Each team could submit up to three runs and participate in both STS and RTE, or in only one of them. Moreover, each team could participate without attending the workshop venue, held in Salvador. We believe this was an important point for increasing participation, because travelling expenses can be high, especially for those that were coming from Europe. The main drawback was that only four teams actually presented their approaches in the ASSIN 2 workshop, namely CISUC-ASAPP, CISUC-ASAPPy, Deep Learning Brasil and Stilingue. On the other hand, a total of six teams submitted a paper describing their participation, to be included in this volume.

Team	Run	STS		RTE	
		ρ	MSE	F1	Accuracy
CISUC-ASAPPj	1	0.642	0.63	0.560	58.91
	2	0.652	0.61	0.606	62.05
	3	0.616	0.82	0.576	59.76
CISUC-ASAPPpy	1	0.726	0.58	0.401	53.10
	2	0.730	0.58	0.656	66.67
	3	0.740	0.60	0.649	65.52
Deep Learning Brasil	1	0.751	1.20	0.816	81.90
	2	0.785	0.59	**0.883**	**88.32**
	3	0.657	0.65	0.333	50.00
IPR	1	**0.826**	0.52	0.876	87.58
	2	-0.037	15.48	0.873	87.38
	3	0.809	0.62	0.87	87.01
L2F/INESC	1	0.771	0.54	0.775	77.66
	2	0.778	0.52	0.784	78.47
	3	0.751	1.20	0.816	81.90
LIACC	1	0.493	1.08	0.77	77.41
	2	0.459	1.02	0.72	73.20
	3	0.458	1.04	0.733	74.31
NILC	1	0.729	0.64	0.871	87.17
	2	0.729	0.64	0.868	86.85
	3	0.729	0.64	0.865	86.56
PUCPR	1	0.643	0.90	N/A	N/A
	2	0.678	0.85	N/A	N/A
	3	0.646	0.92	N/A	N/A
Stilingue	1	0.748	0.53	0.788	78.84
	2	0.800	**0.39**	0.866	86.64
	3	0.817	0.47	0.866	86.64
WordOverlap (baseline)	–	0.577	0.75	0.667	66.71
BoW sentence 2 (baseline)	–	0.175	1.15	0.557	56.74
Infernal (baseline)	–	N/A	N/A	0.742	74.18

Table 2. Results of each submitted run and baselines.

3.1 Results

The results of the runs submitted by each team in the STS and RTE tasks are shown in Table 2, together with three baselines.

Considering the Pearson correlation (ρ), the best result in STS (0.826) was achieved by the first run submitted by the IPR team, although the best MSE

[11] http://comissoes.sbc.org.br/ce-pln/stil2019/

[12] https://sites.google.com/view/pop2-propor2020

was by the second run of Stilingue (0.39). We highlight that these were the only teams with ρ higher than 0.8. Although Pearson ρ was used as the main metric, this metric and the MSE are two different ways of analysing the results. A high ρ means that the ranking of most similar pairs is closer to the one in the gold standard, while a low MSE means that the similarity scores are closer to the gold ones. Both the best MSE and the best values of ρ are significantly better than the best results achieved in ASSIN 1, both in the official evaluation (ρ=0.73 [9]) and in post-evaluation experiments (ρ=0.75 [2]).

On RTE, Deep Learning Brasil had the best run (second run), considering both F1 and Accuracy, though not very far from IPR, Stilingue and NILC. Again, the values achieved are higher than the best official results in ASSIN 1.

The globally higher performances suggest that, when compared to ASSIN 1, ASSIN 2 was an easier task. This might, indeed, be true, especially considering that, for RTE, the ASSIN 2 collection only used two labels, due to *Paraphrases* being labelled as *Entailment* and thus not "competing". ASSIN 2 data was also aimed to be easier and not having complex linguistic phenomena. Another point to keep in mind when comparing ASSIN 1 and ASSIN 2 is that in this edition, competitors had access to a balanced corpus. This might have also contributed to the better performance of systems in ASSIN 2 data. Still, we should also consider that, in the last two years, NLP had substantial advances when it comes to the representation of sentences and their meaning, which lead to significant improvements in many tasks.

3.2 Approaches

Approaches followed by the participants show that the Portuguese NLP community is quickly adopting the most recent trends, with several teams (IPR, Deep Learning Brasil, L2F/INESC, Stilingue and NILC), including those with the best results, somehow exploring BERT [8] contextual embeddings, some of which (IPR, NILC) fine-tuned for ASSIN 2. Some teams combined the previous with other features commonly used in STS / RTE, including string similarity measures (e.g., Jaccard for tokens, token n-grams and character n-grams), agreement in negation and sentiment, lexical-semantic relations (synonyms and hyponymy), as well as pre-trained classic word embeddings (e.g., word2vec, GloVe, fastText, all available for Portuguese as part of the NILC embeddings [15]). Besides BERT, non-pretrained neural models, namely LSTM Siamese Networks (PUCPR) and Transformers (Stilingue), were also used, while a few teams (ASAPPpy, ASAPPj) followed a more classic machine learning approach, and learned a regressor from some of the previous features. Models were trained not only in the ASSIN 2 train collection, but also in data from ASSIN 1.

Towards the best Pearson ρ, the IPR team relied on a pre-trained multilingual BERT model, freely available by the developers of BERT, which they fine-tuned with large Portuguese corpora. A neural network was built by adding one layer to the resulting BERT model and trained with ASSIN 1 (train and test) and ASSIN 2 (train) data.

Stilingue relied on the exploration of Transformers, trained with BERT [8] features, plus a set of 18 additional features covering sentiment and negation agreement, synonyms and hyponyms according to Onto.PT [13] and VerbNet [26], similarity, gender and number agreement, Jaccard similarity of shared tokens, verb tense, presence of the conjunction 'e' (and), similar and different tokens, sentence subject, and cosine of sentence embeddings computed with FastText [15].

For RTE, the best run, by Deep Learning Brasil, was based on an ensemble of multilingual BERT and RoBERTa [17], which improves on the results of BERT, both fine-tuned for the ASSIN 2 data. However, for RoBERTa, this data was previously translated to English, with Google Translate. The IPR team also relied on BERT and used ASSIN 1 data to fine-tune the model.

Our first baseline was the word overlap, which had very competitive results in ASSIN 1. It counts the ratio of overlapping tokens in both the first and second sentence, and trains a logistic/linear regressor (for RTE/STS) with these two features. A second baseline is inspired by Gururangan et al. [14] and trains the same algorithms on bag-of-words features extracted only from the second sentence of each pair. It aims to detect biases in the construction of the dataset. For RTE, a third baseline was considered, namely, Infernal [10], a system based on hand designed features, which has state-of-the-art results on ASSIN 1.

3.3 Results in Harder Pairs

Similar to Gururangan et al. [14], we took all the pairs misclassified by our second baseline and called them a *hard* subset of the data. In other words, these pairs were not correctly classified by only looking at the hypothesis, the second sentence of the pair. In order to provide an alternative view on the results, we analysed the participants' results on this subset.

Results are shown in table 3. Though worse than the performance in the full collection, in table 2, the differences are not as large as those reported by Gururangan et al. [14]. This is not surprising, as the second baseline had an F1 score only marginally above chance level[13], indicating that the dataset does not suffer from annotation artifacts as seriously as SNLI.

A particular outlier is the second run of IPR in STS. But the highly differing value is due to their already very low Pearson ρ in the original data. Still, eight runs had a decrease of more than 15% in RTE, suggesting they might have been exploiting some bias in the collection.

4 Conclusions

ASSIN 2 was the second edition of ASSIN, a shared task targeting Recognizing Textual Entailment / Natural Language Inference and Semantic Textual Similarity in Portuguese. It had nine participating teams, from Portugal and Brazil,

[13] For comparison, Gururangan et al. [14] had 67% accuracy in a dataset with three classes.

Team	Run	STS			RTE		
		ρ^*	MSE	ρ diff	F1	Acc	F1 diff
ASAPPj	1	0.567	0.75	-11.68%	0.526	64.02	-6.07%
	2	0.578	0.70	-11.35%	0.558	64.59	-7.92%
	3	0.586	0.94	-4.87%	0.551	64.40	-4.34%
ASAPPpy	1	0.630	0.75	-13.22%	0.302	35.79	-24.69%
	2	0.635	0.75	-13.01%	0.591	59.49	-9.91%
	3	0.655	0.77	-11.49%	0.587	59.49	-9.55%
IPR	1	0.764	0.70	-7.51%	0.798	81.02	-8.90%
	2	0.018	13.43	-148.65%	0.788	79.89	-9.74%
	3	0.734	0.86	-9.27%	0.781	79.32	-10.23%
LIACC	1	0.368	1.20	-25.35%	0.630	63.36	-18.18%
	2	0.383	1.09	-16.56%	0.574	57.41	-20.28%
	3	0.348	1.15	-24.02%	0.581	58.07	-20.74%
NILC	1	0.632	0.88	-13.31%	0.777	78.94	-10.79%
	2	0.632	0.88	-13.31%	0.774	78.56	-10.83%
	3	0.632	0.88	-13.31%	0.760	77.05	-12.14%
PUCPR	1	0.528	1.23	-17.88%	—	—	—
	2	0.562	1.17	-17.11%	—	—	—
	3	0.528	1.25	-18.27%	—	—	—
L2F/INESC	1	0.593	0.77	-15.04%	0.639	66.76	-11.98%
	2	0.677	0.73	-12.19%	0.644	65.34	-16.90%
	3	0.684	0.71	-12.08%	0.658	66.95	-16.07%
Deep Learning Brasil	1	0.659	1.54	-12.25%	0.681	68.56	-16.54%
	2	0.718	0.75	-8.54%	0.804	81.40	-8.95%
	3	0.579	0.78	-11.87%	0.244	32.20	-26.73%
Stilingue	1	0.666	0.68	-10.96%	0.747	77.24	-5.20%
	2	0.718	0.52	-10.25%	0.777	79.13	-10.28%
	3	0.744	0.66	-8.94%	0.777	79.13	-10.28%

Table 3. Results of each submitted run in the harder pairs and difference towards the full results.

and, differently from the previous ASSIN edition [11], most of the systems outperformed the proposed baselines. We believe that the effort of having a simpler task in ASSIN 2 was beneficial, not only because systems could do better in this edition, but also because the ASSIN 2 corpus has a sound annotation strategy, comparable with previous shared tasks for English. Looking at the participation, it seems that the Portuguese processing community is now more interested in the proposed tasks.

On the results achieved, it is notable that systems based on transfer learning had better results in the competition for both tasks. A note should be added on the Deep Learning Brasil team, which achieved the best scores for RTE with a

strategy based on translating the data to English, to make possible the use of more powerful models. However, it is possible that the nature of the data, which is a translated and adapted version of SICK, makes this strategy more sound than it would be in real-world scenarios. After all, ASSIN 2 results may indicate how the pre-trained language models used, namely BERT and RoBERTa, rapidly improved the state-of-the-art of a given task. For the future, we would like to discuss new ways of evaluating the generalization power of the proposed systems, since intrinsic metrics, considering only a subset of the data that follows exactly the same format of the training data, seems nowadays not to be enough to effectively evaluate the systems' performance.

References

1. Agirre, E., Diab, M., Cer, D., Gonzalez-Agirre, A.: SemEval-2012 task 6: A pilot on semantic textual similarity. In: Proc. 1st Joint Conf. on Lexical and Computational Semantics-Vol. 1: Proc. of main conference and shared task, and Vol. 2: Proc. of Sixth Intl. Workshop on Semantic Evaluation. pp. 385–393. Association for Computational Linguistics (2012)
2. Alves, A., Gonçalo Oliveira, H., Rodrigues, R., Encarnação, R.: ASAPP 2.0: Advancing the state-of-the-art of semantic textual similarity for Portuguese. In: Proceedings of 7th Symposium on Languages, Applications and Technologies (SLATE 2018). OASIcs, vol. 62, pp. 12:1–12:17. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany (June 2018)
3. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 632–642. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015)
4. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics (2015)
5. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.: Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In: Proceedings of 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 1–14. Association for Computational Linguistics (2017)
6. Crouch, R., Karttunen, L., Zaenen, A.: Circumscribing is not excluding: A response to manning. http://web.stanford.edu/ laurik/publications/reply-to-manning.pdf
7. Dagan, I., Glickman, O., Magnini, B.: The pascal recognising textual entailment challenge. Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognizing Textual Entailment. (2006)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc 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)
9. Fialho, P., Marques, R., Martins, B., Coheur, L., Quaresma, P.: INESC-ID@ASSIN: Medição de similaridade semântica e reconhecimento de inferência textual. Linguamática **8**(2), 33–42 (2016)

10. Fonseca, E., Aluísio, S.M.: Syntactic knowledge for natural language inference in Portuguese. In: Villavicencio, A., Moreira, V., Abad, A., Caseli, H., Gamallo, P., Ramisch, C., Gonçalo Oliveira, H., Paetzold, G.H. (eds.) Computational Processing of the Portuguese Language. pp. 242–252. Springer, Cham (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. Giampiccolo, D., Magnini, B., Dagan, I., Dolan, B.: The third PASCAL recognizing textual entailment challenge. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. pp. 1–9. Association for Computational Linguistics, Prague (Jun 2007)
13. Gonçalo Oliveira, H., Gomes, P.: ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically. Language Resources and Evaluation **48**(2), 373–393 (2014)
14. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., Smith, N.A.: Annotation artifacts in natural language inference data. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 107–112. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018)
15. Hartmann, N.S., Fonseca, E.R., Shulby, C.D., Treviso, M.V., Rodrigues, J.S., Aluísio, S.M.: Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In: Proceedings the 11th Brazilian Symposium in Information and Human Language Technology. STIL 2017 (2017)
16. Kalouli, A.L., Real, L., de Paiva, V.: Correcting contradictions. In: Proceedings of Computing Natural Language Inference (CONLI) Workshop, 19 September 2017 (2017)
17. 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. arXiv preprint arXiv:1907.11692 (2019)
18. Manning, C.: Local textual inference: It's hard to circumscribe, but you know it when you see it – and nlp needs it. https://nlp.stanford.edu/ manning/papers/LocalTextualInference.pdf (2006)
19. 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: Proc. of 8th Intl. Workshop on Semantic Evaluation (SemEval 2014). pp. 1–8. Association for Computational Linguistics, Dublin, Ireland (2014)
20. de Marneffe, M.C., Rafferty, A.N., Manning, C.D.: Finding contradictions in text. In: Proceedings of ACL-08: HLT. pp. 1039–1047. Association for Computational Linguistics, Columbus, Ohio (Jun 2008)
21. Nangia, N., Williams, A., Lazaridou, A., Bowman, S.: The RepEval 2017 shared task: Multi-genre Natural Language Inference with sentence representations. In: Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP. pp. 1–10. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017)
22. Negri, M., Marchetti, A., Mehdad, Y., Bentivogli, L., Giampiccolo, D.: Semeval-2012 task 8: Cross-lingual textual entailment for content synchronization. In: Proceedings of *SEM (2012)
23. Negri, M., Marchetti, A., Mehdad, Y., Bentivogli, L., Giampiccolo, D.: Semeval-2013 task 8: Cross-lingual textual entailment for content synchronization. In: Proceedings of *SEM (2013)

24. Real, L., Fonseca, E., Oliveira, H.G.: The assin 2 shared task: a quick overview. In: Computational Processing of the Portuguese Language - 13th International Conference, PROPOR 2020, Évora, Portugal, March 2-4, 2020, Proceedings. p. in press. LNCS, Springer (2020)
25. Real, L., Rodrigues, A., Vieira, A., Albiero, B., Thalenberg, B., Guide, B., Silva, C., de Oliveira Lima, G., C. S. Câmara, I., Stanojević, M., Souza, R., De Paiva, V.: SICK-BR: A Portuguese corpus for inference. In: Proceedings of 13th PROPOR (2018)
26. Scarton, C., Aluısio, S.: Towards a cross-linguistic verbnet-style lexicon for brazilian portuguese. In: Proceedings of LREC 2012 Workshop on Creating Cross-language Resources for Disconnected Languages and Styles (2012)
27. Voorhees, E.M.: Contradictions and justifications: Extensions to the textual entailment task. In: Proceedings of ACL-08: HLT. pp. 63–71. Association for Computational Linguistics, Columbus, Ohio (Jun 2008)
28. Zaenen, A., Karttunen, L., Crouch, R.: Local textual inference: Can it be defined or circumscribed? In: Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment. pp. 31–36. Association for Computational Linguistics, Ann Arbor, Michigan (Jun 2005)