SICK-BR: a Portuguese corpus for inference

Livy Real[1], Ana Rodrigues[1], Andressa Vieira e Silva[1], Beatriz Albiero[1], Bruna Thalenberg[1], Bruno Guide[1], Cindy Silva[1], Guilherme de Oliveira Lima[1], Igor C. S. Câmara[2], Miloš Stanojević[3], Rodrigo Souza[1], Valeria de Paiva[4]

[1] University of São Paulo
livyreal@gmail.com; bruna@ime.usp.br
{ana2.rodrigues,andressa.vieira.silva,beatriz.albiero,bruno.guide,
cindy.silva,guilherme.oliveira.lima,rodrigo.aparecido.souza}@usp.br
[2] University of Campinas
igor0csc@gmail.com
[3] University of Edinburgh
m.stanojevic@ed.ac.uk
[4] Nuance Communications
valeria.depaiva@gmail.com

Abstract. We describe SICK-BR, a Brazilian Portuguese corpus annotated with inference relations and semantic relatedness between pairs of sentences. SICK-BR is a translation and adaptation of the original SICK, a corpus of English sentences used in several semantic evaluations. SICK-BR consists of around 10k sentence pairs annotated for neutral/contradiction/entailment relations and for semantic relatedness, using a 5 point scale. Here we describe the strategies used for the adaptation of SICK, which preserve its original inference and relatedness relation labels in the SICK-BR Portuguese version. We also discuss some issues with the original corpus and how we might deal with them.

Keywords: Portuguese · Open Corpus · Textual Inference · NLI · Semantic Relatedness

1 Introduction

Determining semantic relationships between sentences is essential for machines that understand and reason with natural language. While the task of detecting Natural Language Inference (NLI) may be considered implicit in all the work done in Natural Language Semantics, one could argue that the strategies currently used for NLI are superficial and unable to deal with the real problem: reasoning with semantic content. We are working towards reasoning with Portuguese texts and have been working for a while on different strategies to obtain open systems that can help with processing Portuguese [13–15]. Inference can be seen as one of the most basic tasks for semantic reasoning. Our long term research goal is the development of symbolic and statistics based approaches that can deal with inference in Portuguese.

Much work has been done for inference in English, symbolic or otherwise, and not so much for Portuguese. The initial Recognizing Textual Entailment (RTE)[5] challenges (from 2005 to 2013) have drawn attention to the problems of detecting inference. Nowadays, with the success of deep learning techniques, the inference tasks are again in the spotlight for Natural Language Processing, now under the label of Natural Language Inference (NLI). Large datasets have been constructed to serve as supervised data for systems that want to learn to perform inference e.g. amongst others SICK [9], SNLI [1], MultiNLI [16]. However, it is not clear how trustworthy these datasets are, how much they codify the flexible human intuitions of inference. Also, it is still debatable how these datasets should be constructed, since it has been shown [5] that neural systems can be largely influenced by the way the training corpus is assembled. This makes it possible, for example, that a system assigns an inference relation between two sentences only by looking at the first sentence, thereby missing completely the point of the inference. For this and other reasons, it is particularly important to have different corpora available to train and test systems that detect inference in different languages.

1.1 Related Work

There is already one Portuguese corpus for textual inference publicly available, the ASSIN corpus [3]. This was released for a shared task[6], associated with the conference PROPOR2016. The tasks in the competition were both Semantic Similarity and Textual Entailment. Competitors could chose to participate in only one of the tasks. Only 4 of the 6 teams participating in the task decided to work on textual entailment, while all the 6 teams worked on semantic similarity.

ASSIN was the first attempt to gather the Portuguese NLP community to discuss textual entailment and semantic similarity. However, since the ASSIN corpus was built over international news, there is a large amount of temporal expressions, named entities and other complex phenomena that make the reasoning over the sentences very difficult. No system could do better than the offered baselines. This led the ASSIN creators to suggest that maybe a simpler corpus for NLI, organized in the style of SICK[9], with less subjectivity was needed [3, 4]. Thus, in some ways, this work can be seen as a continuation of the ASSIN work.

However, we can also see problems with the ASSIN corpus. The ASSIN corpus is annotated for entailment, paraphrase and neutral relations. We agree with [2] and [10] that *contradictions* are an essential component of human reasoning, so we want to have a corpus which annotates contradictions too. The ASSIN corpus has also the label 'paraphrase' and paraphrases are also entailments, so one could say that these two ASSIN categories overlap with each other. This is something that we would like to avoid. A corpus with a restricted scope of the semantic and syntactic phenomena it intends to tackle seems a sensible idea and SICK is exactly that.

[5] `https://aclweb.org/aclwiki/Recognizing_Textual_Entailment`
[6] `http://propor2016.di.fc.ul.pt/?page_id=381`

1.2 Our goal

Given our long term goal of doing semantic reasoning in Portuguese and the sensible suggestion of the need for a simplified corpus, we decided to take on the task of building a NLI corpus in Portuguese.

The corpus SICK concentrates on compositional phenomena and it is annotated with the kind of inference we want to be able to distinguish. The SICK construction process, coming up from picture captions, restricts its scope to concrete actions and scenes. Moreover some care was taken to restrict the amount of world knowledge required to perform the intended inferences. Finally, because the inferences aimed at were fairly basic, they held the promise of common sense reasoning, which we believe is somewhat universal.

Finally there were practical considerations. Building a resource like SICK is very time and money consuming. Bootstrapping the creation of a Portuguese corpus via automatic translating and adapting SICK for Portuguese, giving rise to SICK-BR, seemed the easier route. This approach has also the added value of producing a parallel resource, since the pairs of SICK and SICK-BR are aligned and offer the same labels, both for relatedness and inference relations.

2 SICK

The corpus SICK[7] was conceived to provide a benchmark for compositional distributional semantic models [9]. The corpus SICK consists of English sentence pairs annotated to account for inference relations (entailment, contradiction and neutral) and relatedness (on a 5-point rating scale). The corpus is simplified in aspects of language processing not fundamentally related to composionality: there are no named entities, the tenses have been simplified to the progressive, there are few modifiers, few compounds, few pronouns, etc. The data set consists of 9840 English sentence pairs (composed from some 6k unique sentences), generated from existing sets of captions of pictures.

The authors of SICK randomly selected a subset from the caption sources and applied a 3-step generation process to obtain the pairs. After a normalization phase, when undesirable phenomena were excluded or re-written, desirable phenomena were added, as negation and active/passive alternation. After these generation steps, the sentences in pairs were sent to Amazon 'mechanical turkers' who annotated them for inference and relatedness.

Inference annotation led to 5595 neutral pairs, 1424 contradiction pairs, and 2821 entailment pairs, hence 4245 informative pairs in total. SICK was the resource used in the SemEval 2014 task 1.

3 SICK-BR

We bootstrapped the creation of a simplified corpus for NLI in Portuguese by making use of the human annotations in the original SICK. We start from a

[7] `http://clic.cimec.unitn.it/composes/sick.html`

basic machine translation of the corpus. We want to be sure that our translated pairs get exactly the same truth-conditional semantics as the original ones. We also want to have, as much as possible, the same kind of linguistic phenomena that SICK discusses. Another parallel goal is to keep the relatedness between the paired sentences, which imposes challenges on lexical choices. Here we explain some of our strategies to keep the translations of SICK-BR as close to the original SICK as possible and we describe the phases of the construction of SICK-BR. The process of building SICK-BR had the following phases: 1. pre-processing and machine translation; 2. guidelines creation, training and translation checking; 3. post-processing and reconstruction; 4. label checking. We discuss each of these steps in this section.

3.1 Pre-processing and Machine Translation

Firstly, we got all the (6076) unique sentences that are part of the 9480 SICK pairs and translated them to Portuguese using a state-of-the-art online tool. As expected the output of the automatic translation is full of mistakes. For example, in SICK, the most used verb, apart from the verb *to be*, is the verb *to play*. This needs to be translated by different verbs in Portuguese *tocar/play an instrument, brincar/play with other kids* and *jogar/play sports*, etc. So we expected this to be difficult for machine translation, and it was. We also found many spelling mistakes in the translation, such as *estao* or *estáo* for *estão*, which is easy to correct, but that can cause trouble when processing the corpus, since most of the systems would just not recognize these misspelled forms.

3.2 Guidelines, training and checking

As discussed in [8, 6], many mistakes in SICK are due to the lack of clear guidelines for annotators and to the fact that they did not have linguistic training. To avoid introducing mistakes in the corpus and to try to ensure the quality of SICK-BR, concerted effort was put in this phase.

Since we worked on a translation to produce a new corpus, SICK-BR, which should keep the same labels (for relatedness and inference) from the previous one, SICK, we call here 'annotators' the linguists that worked in the translation, rather than the people who actually annotated the original labels of the pairs. Also, we call 'guidelines' the instructions to be used for translation checking, rather than instructions to actually label the relations within a given pair. In this translation phase, ten annotators took part in the work, all of them native Brazilian Portuguese speakers, proficient in English and all have linguistic training.

Once we had an automated machine translation of the corpus, two of us selected 55 sentences that showcased the intended linguistic phenomena in SICK and also other phenomena that may be difficult to translate from English. These sentences were given to the 10 annotators without the machine translation and, after a detailed discussion, an agreement was reached on how to translate these sentences. We then compared these 'golden' translations to the ones produced

automatically and got some insights on where the machine translation system systematically goes wrong.

Considering our main goals — (i.) keep the inference labels of SICK, (ii.) the relatedness labels and (iii.) having a naturally sounding corpus in Portuguese — and the results of this initial discussion, we reached our main guidelines:

- 1. Translated sentences should keep the same truth values as the original sentences;
- 2. We try to maintain, over the Portuguese corpus, the same lexical choices for the same English expressions within reason;
- 3. We keep, as much as possible, the same phenomena that we believe the original sentence was showcasing;
- 4. We keep naturally sounding Portuguese sentences, as much as possible;
- 5. We keep word alignment, whenever possible.

The guidelines are to be followed in this order, which tells us that keeping the same labels as the sentences have in SICK is more important than to keep the naturalness of the Portuguese sentences, for example. Although we tried to not have sentences that sound odd our main goal was to keep the labels aligned.

We also produced and updated during the project a glossary[8] for the most used terms, such as the many multiword expressions (MWEs) we found in SICK, despite the original SICK creators efforts to not have any MWEs. This might be useful to scholars interested in MWE and named entities recognition in Portuguese, since these choices are informative.

The 6076 unique sentences of the corpus were equally distributed among the 10 annotators. We used an online platform for the checking. This made it possible for annotators to look at each others' work when translating their sentences. We also kept an online forum for discussing issues, where more than 2k messages were exchanged during this work. The glossary was always updated when a solution was reached. Each annotator could also mark out complex sentences that they thought needed further review. Differently from other corpora creation processes we know about, our annotators could always say they were not able to annotate something, an easy strategy that helps to ensure the quality of the work. Finally, an experienced annotator double checked all sentences considered complex and proposed a final translation for these sentences.

Rethinking our steps During this phase, we realized that one of our previous goals was not reachable. We would like to have SICK and SICK-BR aligned as parallel corpora at a **sentence level**. This would mean that each sentence of SICK would be translated in SICK-BR by one sentence. However, some translation issues showed us that it was an impossible goal.

One third of the pairs in SICK differ by only one or two words, for example the pair *A= Kids in red shirts are playing in the leaves. B= Children in red shirts are playing in the leaves.* Since we could not 'perfectly' translate this pair

[8] Available at `https://github.com/livyreal/SICK-BR/tree/master/Glossary`.

of sentences into Portuguese (Portuguese has no two words for 'child' that have no (ontological) gender attached) keeping exactly the same referent, we would not keep the original NLI labels. We looked at these pairs[9] , called one-word-apart pairs, and, for most of them, we could find words in Portuguese that kept the same truth value for the sentences.

However, the scenario changes when we consider pairs as *kid* and *child*. We have many words for child in Portuguese, but most of them have (ontological) gender attached to them, only *criança* can be used for boys and girls. Considering the many pairs in SICK based on *child/kid* difference, we could not translate both of them to *criança*, without ending up with sentences that were literally the same. Also in SICK, there are many pairs based on the difference between *kid/child* and gender specific words such as *boy/girl*, therefore translating *kid/child* for pairs as *garoto/menino* would not solve the problem, but would rather create a new one. Because of that, we decided to have a new step in our corpus building. We translated both *kid* and *child* by *criança* and had a new step to make sure there is no sentence pair in the corpus with exactly the same sentence repeated. In the sentence translation phase, so, both *A= Kids in red shirts are playing in the leaves. B= Children in red shirts are playing in the leaves.* were translated by *Crianças de camisas vermelhas estão brincando nas folhas.* After in the corpus construction, these exactly-the-same pairs were reanalyzed and re-translated by: *A= Meninos de camisas vermelhas estão brincando nas folhas. B= Garotos de camisas vermelhas estão brincando nas folhas.* With this solution, we still keep the inference label for the pair (A_entails_B, B_entails_A).

These choices had two main consequences: we needed a new phase of corpus construction; and we did not have a corpus aligned at a sentence level. However, we still have a corpus aligned to the original corpus SICK at the **level of paired sentences**. The new corpus keeps exactly the same labels for the intended tasks and these are possible to trace since the `id` pair in SICK-BR is the same as the one in SICK. The pairs in SICK and SICK-BR are aligned and have the same labels, but one sentence in SICK may be translated by more than one sentence in SICK-BR. Therefore, SICK-BR has the same amount of pairs as SICK, but SICK-BR has a slightly bigger number of sentences than SICK.

3.3 Post-processing and Reconstruction

This phase was concerned with making sure we were not introducing new mistakes to SICK-BR. We ran a state-of-the-art speller and grammar checkers on all the 6k unique sentences. We also made sure that we had no extra spaces and final periods in the sentences. Although SICK pairs in general do not have any punctuation, a few sentences still have it and for people interested in syntactic parsing having punctuation or not in a sentence can change the parsing. We then used the glossary we prepared for the annotators, checking to make sure no one missed an agreed lexical choice during the translation phase.

[9] We thank Katerina Kalouli for the processing of original SICK, made public available in `https://github.com/kkalouli/SICK-processing`.

Finally, the corpus was reconstructed: the sentences were paired as the original ones and the original labels were assigned to the Portuguese pairs. We then reviewed all the 'same sentences' pairs.

Bellow, one example of an entry in SICK-BR:

```
580 | Um grupo de meninos está brincando com uma bola em frente
a uma porta grande feita de madeira | Um monte de meninos está
brincando com uma bola em frente a uma porta grande feita de
madeira | ENTAILMENT | 4.9 | A_entails_B | B_entails_A | A group
of boys are playing with a ball in front of a large door made of
wood | A bunch of boys are playing with a ball in front of a large
door made of wood | FLICKR | FLICKR | TEST
```

The first field (1) is the ID pair. The next two fields (2, 3) are the human proposed version of the original sentences. The four following fields (4, 5, 6, 7) are the original SICK labels we reused in our corpus. The next two fields (8, 9) are the original SICK pair. The following two fields (10, 11) indicate the dataset where the original sentences in SICK came from. Finally, the last field (12) indicates the set of SEMEVAL 2014 dataset split this pair was part of.

3.4 Checking labels

We then verified how well the original SICK labels fitted our translated pairs. We checked 400 labels for relatedness and 800 labels for inference relations, chosen randomly but equally distributed between the different label types. This step showed that we do not always agree with the original SICK labels. The annotation for 'semantic relatedness' is especially problematic. This is a subtle classification, that was presented by the original SICK annotators only through examples, therefore the labels are not always consistent. Since our goal was not to re-annotate SICK, but rather to think of strategies that would keep the original human annotation, the 'mistakes' we found in SICK labels are also present in SICK-BR.

The lack of clear guidelines on what would be considered a related pair made impossible for us checking the relatedness scores without considering the original English pair. We compared the relatedness score of the Portuguese pairs checking the pairs in English in parallel. We found that when a certain score was given to an English pair, this score still holds in Portuguese. Since relatedness is a so subtle phenomenon, very difficult to annotate, only huge discrepancies would be considered mistakes and, over 400 checked labels, we didn't find any. Although SICK-BR is a translation of SICK, that could make the relatedness between the sentences different, the fact that SICK is a simplified corpus and that we kept as much as possible the same lexical choices over the whole corpus, made feasible the reuse of semantic relatedness scores from SICK to SICK-BR. Despite of the fact that 100% of the checked relatedness scores were reasonable when applied to the Portuguese pairs, we recall that this annotation is not (in both languages) as reliable as we would like it to be.

Checking the consistency of these labels made us realize some new issues with the original corpus. For example, the sentences *A woman is not riding a*

horse./A woman is riding a horse are part of two pairs with different `ids`. So in SICK, we have both the pair *id=4305 A= A woman is not riding a horse. B= A woman is riding a horse* and the pair *id=4587 A= A woman is riding a horse. B= A woman is not riding a horse.* Since all the pairs were annotated for inference in both directions (whether A entails B and also B entails A), it does not make sense to have repeated pairs. The situation gets worse when we consider that these two pairs have different labels for relatedness in SICK: while the first pair has a 4.5 relatedness score, the second one is scored as only 3.8. This clearly shows how the relatedness score is subjective and debatable.

We also found some inconsistency on inference labels as [8, 7] have already shown. From the 800 pairs checked for inference, 20 do not hold for Portuguese. All these 20 pairs are labeled as ENTAILMENT. From our analysis (double checked by two native English speakers), 14 of those 20 inference labels were already wrong in SICK. As pointed in [8] some sentences in SICK are non-sensical or ungrammatical. For example *A motorcycle is riding standing up on the seat of the vehicle* or *The players is maneuvering for the ball*[10]. The other six pairs are debatable labels even in English. It happens that one of the sentences is ambiguous (as *The kid is still in the snow.*) or that the entailment among the pairs is not obvious but possible (is a shore always by the beach? Is a lady a girl?).

In SICK-BR, we corrected all the ungrammatical sentences (18 sentences), but we do not correct (35) non-sensical sentences since correcting them would mean radically changing their interpretation. We listed 35 sentences as non-sensical, such as *A woman is bowling two eggs to a break dancer* and *A man is pouring a pot of cheese sauce into a shredded plate.* It seems that these sentences are the result of the expansion phase of the SICK creation process. They were created by scrambling the original words. Although this way of generating new sentences might have seemed a good idea, it created a lot of noise in SICK. Almost always these non-sensical and ungrammatical sentences are part of pairs that were labeled as neutral for inference, suggesting that when the annotators could not judge the sentences, they just annotated them as neutral.

4 Results

The Portuguese pairs of SICK-BR can be downloaded in `https://github.com/livyreal/SICK-BR`. SICK-BR has the sentences in Portuguese and keeps the same identifiers `id` and labels for inference and relatedness as the original corpus.

Our hypotheses that the logical phenomena in both languages would be similar and that entailment and contradiction relations between sentences would work the same way both in English and in Portuguese have been mostly confirmed. From 800 inference labels checked, we disagree on only 20 in SICK-BR: 14 of them were already wrong in SICK and the other 6 are somehow debatable in the original resources as well. Considering 400 relatedness score, we confirmed

[10] These analyses are available in `https://github.com/livyreal/SICK-BR`.

that, for all the checked pairs, if the English label was reasonable, ours was also reasonable. This makes SICK-BR as reliable as SICK for the relatedness task.

Of course, this translation, preserving relations, was possible because of the simplification of data that SICK aimed for.

Some of the issues found in SICK are also present in our corpus. We still have, for example, sentences that are not common sense such as *Um hamster está cantando* (translated from *A hamster is singing*). However, given the need to manually verify all the translations, we have managed to correct non-grammatical sentences, sentences with smaller typos and such like. We have decided to keep the sentences lacking commonsense, to keep the parallelism between the corpora. Many of the goals stated by the SICK creators were not really fully realized. For instance, not all sentences are in the progressive. We found around 90 sentences that were not in progressive, such as *A topless boy has a clean face.* To preserve as much as possible the original label assignments, we also kept some of the ambiguity and bias from the original corpus.

SICK-BR is more uniform than SICK as far as punctuation goes. We also corrected some spelling and processing mistakes. SICK has sentences, such as *A black dog is jumping from* ***n*** *hay ball to another hay ball* (should *n* be *an*?) It also has sentences such as *The man is not adding seasoning to* ***the/some*** *water in a bowl* and *A piece of bread, which is big, is having butter spread upon it by a man* ***OR A piece of bread, which is big, is being spread with butter by a man***. In these cases, it seems that some steps of the construction phase were messy and left behind the choice markers used by the SICK creators. For all these cases, we have a single and grammatical sentence in SICK-BR.

5 Conclusions and Future Work

We described the construction of a Natural Language Inference (NLI) corpus for Portuguese, SICK-BR, which is based on and aligned to the English corpus SICK. We focused on linguistic strategies to guarantee (i) the reuse of the original NLI and relatedness labels of SICK into SICK-BR; (ii) a natural register of Portuguese and (iii) the existence and discussion of the same linguistic phenomena found in SICK. The issues found with the labels in SICK-BR were almost all already found in the original SICK. Due to some specificities of the languages involved, it was impossible to keep SICK-BR aligned to SICK at the sentence level, instead we have SICK and SICK-BR aligned at the pair level. We leave to future work the investigation of different approaches to automatically detecting inference relations in SICK-BR. Concentrating on SICK-BR, we would like to make sure that existing lexical resources for Portuguese, such as OpenWordNet-PT [13], are capable of dealing with the information in SICK-BR. Finally, we would like to investigate the phenomena of implicatives and factives in Portuguese, following up on the work of Kartunnen and others [11], [12].

References

1. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326 (2015)
2. Condoravdi, C., Crouch, D., De Paiva, V., Stolle, R., Bobrow, D.: Entailment, intensionality and text understanding. In: HLT-NAACL 2003 workshop on Text meaning (2003)
3. Fonseca, E., Borges dos Santos, L., Criscuolo, M., Aluisio, S.: Visao geral da avaliacao de similaridade semantica e inferencia textual. Linguamatica **8**(2) (2016)
4. Fonseca, E.R.: Reconhecimento de implicação textual em português. Ph.D. thesis, ICMC-USP (2018)
5. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S.R., Smith, N.A.: Annotation artifacts in natural language inference data. CoRR **abs/1803.02324** (2018), `http://arxiv.org/abs/1803.02324`
6. Kalouli, A.L., Real, L., De Paiva, V.: Annotating logic inference pitfalls. Workshop on Data Provenance and Annotation in Computational Linguistics (2018)
7. Kalouli, A.L., Real, L., de Paiva, V.: Correcting contradictions. In: Proceedings of Computing Natural Language Inference (CONLI) Workshop (2017)
8. Kalouli, A.L., Real, L., de Paiva, V.: Textual inference: getting logic from humans. In: Proceedings of the 12th International Conference on Computational Semantics (IWCS) (2017)
9. Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R.: A SICK cure for the evaluation of compositional distributional semantic models. In: Proceedings of LREC 2014 (2014)
10. de Marneffe, M.C., Rafferty, A.N., Manning, C.D.: Finding contradictions in text. In: Proceedings of ACL-08 (2008)
11. de Melo, G., de Paiva, V.: Sense-specific implicative commitments. In: 15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2014) (2014)
12. Nairn, R., Condoravdi, C., Karttunen, L.: Computing relative polarity for textual inference. Inference in Computational Semantics (ICoS-5) pp. 20–21 (2006)
13. de Paiva, V., Rademaker, A., de Melo, G.: Openwordnet-pt: An open brazilian wordnet for reasoning. In: COLING 2012: Demonstration Papers (2012)
14. de Paiva, V., Real, L., Rademaker, A., de Melo, G.: Nomlex-pt: A lexicon of portuguese nominalizations. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014). Reykjavik, Iceland (May 2014)
15. Real, L., Rademaker, A., Chalub, F., V de Paiva, V.: Towards temporal reasoning in portuguese. In: LREC2018 Workshop Linked Data in Linguistics (2018)
16. Williams, A., Nangia, N., Bowman, S.R.: A broad-coverage challenge corpus for sentence understanding through inference. arXiv (2017), `http://arxiv.org/abs/1704.05426`