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

[2] University of Campinas

[3] University of Edinburgh

[4] Nuance Communications

September 25, 2018

- Natural Language Inference (NLI)/ Recognizing Textual Entailment (RTE)
- A man is dancing on a roof $\rightarrow$ Somebody is moving
- Core of semantic reasoning (Cooper et al., 1996)
- Applied tasks: QA, summarization, knowledge extraction, plagiarism detection
- Symbolic approaches (Kalouli et al 2018c; Abzianidze, 2017)
- Portuguese

Previous work: ASSIN

- Open annotated corpus + Shared task (PROPOR 2016)
- ASSIN (Fonseca et al. 2016): the only PT corpus annotated for inference (and similarity) so far
- Some issues: overlapping labels, no contradictions at all
- Sentences from Google news, suitable for ML approaches
- In the ASSIN shared task, no one could do better than the baseline for NLI, suggesting the need for a simpler corpus

SICK

- Sentences Involving Compositional Knowledge (Marelli et al, 2014)
- English testsuite for Compositional Distributional Semantics
- Suitable for symbolic approaches
- Created from captions of pictures, contains literal, non-abstract, common-sense concepts
- No NEs, MWEs, temporal expressions, reported speech, complex verbs, etc (in principle...)
- 9840 English sentence pairs, 6076 sentences, but only 1886 unique lemmas
- corpus used at the SemEval 2014

Example

- AeBBnA; 4.5
  A = A man is singing and playing a guitar.
  B = A guitar is being played by a man.

SICK-BR - Strategic goals

- Portuguese corpus aligned to SICK
- Our hypothesis: logical phenomena in both languages should be similar and entailment and contradiction relations between sentences should work the 'same way'
- Reuse of SICK annotation
    1. Keep the inference labels of SICK
    2. Keep the relatedness labels
    3. 6k sentences vs 30k labels

SICK-BR steps

- Pre-processing and Machine Translation
- Manual translation checking: guidelines, annotators training, golden standard set
- Post-processing and Reconstruction
- Checking labels

Pre-processing and Machine Translation

- 10k sentence pairs, 6k unique sentences
- State-of-the-art machine translation system
- A man is mindlessly slicing the carrot with a machine
  Um homem está **sem** cortar a cenoura com uma máquina
- A man is parking a car in a garage
  Um homem **estao** estacionando um carro em uma garagem
- A group of children is **playing** maracas and tambourines
  Um grupo de crianças está jogando maracas e pandeiros

Manual Translation Checking

- 10 annotators: Portuguese native speakers, with linguistic training
- 55 example sentences annotated individually
- Discussion
- Guidelines
- Glossary

Guidelines

The guidelines are to be followed in this order.

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

Annotation strategies

Each annotator reviewed 600 sentences and difficult cases were checked by an experienced annotator

- Glossary
- Everyone sees everyone else work
- "I don't know" is a possible answer
- Ask for double checking
- Online forum (more than 2k messages!)

Post-processing and Reconstruction

- Use of Glossary to make sure lexical choices are uniform
- Grammar and spelling checkers
- Corpus reconstruction: pairing sentences as in the original corpus

Checking labels

- Checked 400 relatedness labels
- Checked 800 inference labels
- Pairs chosen randomly but equally distributed between the different label types

Two issues:
(i) relatedness labels are very subjective
(ii) some SICK inference labels are wrong (3/800)

However, all labels checked are consistent!

Relatedness labels

- **4305** A woman is not riding a horse/A woman is riding a horse CONTRADICTION **4.5**
- **4587** A woman is riding a horse/A woman is not riding a horse CONTRADICTION **3.8**
- SICK-BR: Uma mulher não está andando a cavalo / Uma mulher está andando a cavalo

Inference labels

- A menina loira está dançando atrás do equipamento de som / A menina loira está dançando em frente ao equipamento de som NEUTRAL 3.9 A_contradicts_B B_neutral_A
- The blond girl is dancing behind the sound equipment / The blond girl is dancing in front of the sound equipment NEUTRAL 3.9 A_contradicts_B B_neutral_A

We would annotate it differently, but we don't change the labels (for now. at least)

SICK-BR results

- The thesis that one can re-use the semantic annotations (insisting on linguistic strategies for translation and adaptation) has been confirmed
- we have an open Portuguese NLI corpus
- this corpus is aligned to English SICK
- We corrected ungrammatical and non-sensical sentences, typos and managements mistakes, therefore SICK-BR seems slightly better quality
- However we still have labels we don't agree with

SICK-BR example

151 | Um cervo selvagem está pulando uma cerca | Um cervo está pulando uma cerca | ENTAILMENT | 4.8 | A_entails_B | B_neutral_A | A wild deer is jumping a fence | A deer is jumping a fence | FLICKR | FLICKR | TEST

id | Sent_A | Sent_B | NLI | REL | A_to_B | B_to_A | **Eng_SICK_A** | **Eng_SICK_B** | Original_DataSet | Original_DataSet | SemEval2014_set

Further work: tasks

- Investigate different approaches to RTE using SICK-BR
  - Guide (2018, to appear) obtained 90% accuracy on NLI task using BOW + Similarity labels
  - apply the approaches of Fonseca (2018)
  - Kalouli et al. (2018b) almost 100% accuracy on SICK 'one-word difference pairs using Princeton Wordnet — Check results on SICK-BR and OpenWordNet-PT (de Paiva, 2012); Thalenberg (2018, to appear)
  - Investigate how Graphical Knowledge Representation (GKR) models for SICK can be used for SICK-BR (Kalouli et al., 2018c)
  - Natural logic proof system (Abzianidze, 2017)

Further work: resources

- Obtain a subset of SICK-BR with labels agreed by Portuguese speakers (ongoing work)
- Construct a new corpus for NLI, with sentences originally produced in Portuguese
- Investigate bias of 'artifacts' in SICK-BR (Gururangan et al., 2018)
- New corpus design and guidelines for NLI/RTE task

Thanks!

livyreal@gmail.com

References

M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi and R. Zamparelli (2014). A SICK cure for the evaluation of compositional distributional semantic models. Proceedings of LREC 2014.
Cooper, R. et al. (1996). Using the framework, The FraCaS consortium. Fonseca, E. et al. ASSIN, PROPOR 2016.
Fonseca, E. Reconhecimento de implicação textual em português, 2018
Kalouli et al. (2017a) Aikaterini-Lida Kalouli, Livy Real, Valeria de Paiva. Textual Inference: getting logic from humans. 12th International Conference on Computational Semantics (IWCS)
Kalouli et al. (2017b): Correcting Contradictions. Computing Natural Language Inference (CONLI) Workshop.
Kalouli et al. (2018a) . Annotating Logic Inference Pitfalls. Workshop on Data Provenance and Annotation in Computational Linguistics, co-located with the 16th Treebanks and Linguistic Theory conference (TLT16)
Kalouli et al. (2018b). WordNet for "Easy" Textual Inferences. GLOBALEX, co-located with LREC 2018.
Kalouli et al. (2018c). Graphical Knowledge Representations for SICK, NLCS.
Abzianidze, L. (2017): LangPro: Natural Language Theorem Prover. EMNLP.
Thalenberg et al. (2018). Miçangas e lambretas. (to appear)
de Paiva, V. at al. OpenWordNet-PT: An Open Brazilian Wordnet for Reasoning. COLING 2012
Guide, B. (2018). Testando classificadores Naive Bayes no corpus SICK-br. (to appear)
Gururangan et al. (2018) Annotation Artifacts in Natural Language Inference Data. CoRR.

SICK: Previous Project Motivation

- Logic based Natural Language Inference
- Aim: a controlled system that can split different linguistic phenomena and deal with them using different linguistic approaches
- We need a baseline
- Revisions to SICK (Sentences Involving Compositional Knowldedge; Marelli et al. (2014)) to use it as a baseline
- We = Livy Real, Valeria de Paiva (Nuance), Katerina Kalouli (Univ. Konstanz)

SICK Construction

Idea was to simplify the linguistic structure, and to create comparisons of different linguistic phenomena (synonymy, active/passive, negation, agentives, relative clauses, etc)

- Sentences describing the same pictures were normalised
- Applied a 3-step generation on 500 normalised sentences (negations/modifiers/etc)
- A native English speaker reviewed all the sentences
- Pairs were annotated by Amazon Turkers
- Instructions described the task only through examples of relatedness and entailment

‘Bad’ SICK Examples

AcBBnA A = A black and white dog is carrying a small stick on the green grass.
B = A black and white dog is carrying a huge stick on the green grass.

AcBBnA A= A man is parking a car in a garage.
B = A man is getting into a car.

Alignment SICK - SICK-BR

- Sentence level: each sentence in SICK would be a sentence in SICK-BR
- A= Kids in red shirts are playing in the leaves.
  B= Children in red shirts are playing in the leaves.
- Crianças de camisas vermelhas estão brincando nas folhas.
- kid/child/infant - no ontological gender
- boy/girl/toddler/baby/teen/teenager/young

Alignment SICK - SICK-BR

- Pair level: each pair of sentences in SICK has a corresponding pair in SICK-BR
- A= Kids in red shirts are playing in the leaves.
  B= Children in red shirts are playing in the leaves.
- A= Meninos de camisas vermelhas estão brincando nas folhas.
  B= Garotos de camisas vermelhas estão brincando nas folhas.

Checking same-sentences pairs

- Corpus reconstruction: pairing sentences as in the original corpus
- Rechecking same-sentences pairs (a/one = um)
- Example: **One** man is leading the race; **A** man is leading the race ENTAILMENT 5
- SICK-BR: **Um** homem está liderando a corrida; **O** homem está liderando a corrida ENTAILMENT 5