Building Corpora in Portuguese

Livy Real

Grupo de Linguística Computacional (GLiC) - University of São Paulo (USP)
Outline

1. Summary of Previous Works
2. Bosque-UD
3. SICK-BR
4. Challenges
Livy and Corpora

- IBM Research
- Lionbridge/Appen
- GLiC/ USP
- Grammatical Framework
- Abstract Meaning Representation
- Corpus design for Natural Language Inference
- Open Corpora
Bosque-UD

- 2016
- OpenWordNet-PT team* (Valeria de Paiva (Nuance Communications), Alexandre Rademaker (IBM Research / FGV), Fabricio Chalub (IBM Research), Claudia Freitas (PUC/RJ))
- Universal Dependencies Project (http://universaldependencies.org/)

Universal Dependencies

- Universal Dependencies: the promise of greater parallelism between languages
- Universal dependencies not too far from semantic dependencies
- Dependencies are useful in many applications, e.g. IE, IR, etc.
- Open corpora for more than 60 languages now
- Some open tools
Motivation

- Improve the quality of open corpora
- We wanted to have an open, golden standard UD-corpus for Portuguese
- We wanted to contribute to the UD guidelines for Portuguese
Bosque-UD

- Goal: high quality and open corpus
- Restriction: time and staff
- Difficulty: convince employers that we want an open resource
- Solution: work on a conversion of an open golden corpus for Portuguese: Bosque 8.0 (Linguateca Team - https://www.linguateca.pt/Floresta/corpus.html)
Bosque-UD

- ‘Bosque’ means ‘woods’ in Portuguese
- Golden corpus of morpho-syntactic analysis for both European and Brazilian Portuguese
- Annotated with PALAVRAS parsing and revised by linguists
- Largely used by Portuguese and Brazilian communities
- Bosque-UD has 9,368 sentences, from 1,000 newspapers extracts, and 227,653 tokens, with 18,140 unique lemmas.
- Available at: https://github.com/UniversalDependencies/UD_Portuguese-Bosque
- License: CC BY-SA 4.0
Bootstrapping the Bosque-UD creation

- The conversion grammar ultimately used for the first conversion of Bosque to UD contained some 530 rules.
- Manual review motivated by differences between PALAVRAS and UD guidelines.
- Appositions, clitics, MWEs, participles, particle ‘se’, negation, ellipsis, gender annotation.
- Since PALAVRAS was created for Portuguese and UD are language independent, many PALAVRAS annotations didn’t have a place in Bosque-UD; we kept them in MISC field.
Summary of Previous Works

Example: MWEs handling - Bosque 8.0

Livy Real

Grupo de Linguística Computacional (GLiC) - University of São Paulo (USP)

Building Corpora in Portuguese
Example: MWEs handling - Bosque-UD

# text = Produtores da Paraíba, por exemplo, venderam abacaxi a um grupo de empresários espanhóis, no valor de US$ 323 mil.
# source = CETENFolha n=675 cad=Agrofolha sec=agr sem=94a
# sent_id = GF675-2
# id = 2835

1 Produtores produtor NOUN _ Gender=Masc|Number=Plur 9 nsubj _ _
2 de de ADP _ _ 4 case _ _
3 a o DET _ Definite-Def|Gender=Fem|Number=Sing|PronType=Art 4 det _ _
4 Paraíba Paraíba PROPN Gender=Fem|Number=Sing 1 nmod _ SpaceAfter=No
5 , , PUNCT 7 punct _ _
6 por por ADP _ _ 9 advmod _ MWE=por exemplo|MWEPOS=ADV
7 exemplo exemplo NOUN _ Gender=Masc|Number=Sing 6 fixed _ SpaceAfter=No
8 , , PUNCT 7 punct _ _
9 venderam vender VERB _ Mood=Ind|Number=Plur|Person=3|VerbForm=Fin 0 root _ _
10 abacaxi abacaxi NOUN _ Gender=Masc|Number=Sing 9 obj _ _
11 a a ADP _ _ 13 case _ _
12 um um DET _ Definite=Ind|Gender=Masc|Number=Sing|PronType=Art 13 det _ _
13 grupo grupo NOUN _ Gender=Masc|Number=Sing 9 obl _ _
14 de de ADP _ _ 15 case _ _
15 empresários empresário NOUN _ _ Gender=Masc|Number=Plur 13 nmod _ _
16 espanhóis espanhol ADJ _ Gender=Masc|Number=Plur 15 amod _ SpaceAfter=No
17 , , PUNCT _ _ 9 punct _ _
18-19 no _ _ _ _ _ _ _ _
19 em em ADP _ _ 20 case _ _
20 valor valor NOUN _ Gender=Masc|Number=Sing 9 obl _ _
21 de de ADP _ _ 22 case _ _
22 US$ US$ SYM _ Gender=Masc|Number=Plur 20 nmod _ _
23 323 323 NUM _ NumType=Card 24 nummod _ _
24 mil mil NUM _ NumType=Card 22 nummod _ SpaceAfter=No
25 . . PUNCT _ _ 9 punct _ _
Assessment

- CL-conllu library and an online CoNLL-U validation service
- Syntactic validation
Uses of Bosque-UD

- Part of the data used in two tasks: CoNLL 2017 and CoNLL 2018 shared tasks
- Training Freeling’s dependency parser for Portuguese
- Cross-validation of temporal annotation using UD syntactic dependency labels + HeidelTime
  (https://github.com/own-pt/portuguese-time/

Ongoing work, started 2018


External collaborators: Valeria de Paiva (Nuance) and Milos Stanojevic (University of Edinburgh)

Natural Language Inference (NLI) for Portuguese
Previous work

- ASSIN (Fonseca et al., 2016): only one PT corpus annotated for inference (and similarity)
- Some issues: overlapping labels, no contradictions, more suitable for ML approaches
- In ASSIN shared task, no one could do better than the baseline, suggesting the need for a simpler corpus
Why SICK?

- Sentences Involving Compositional Knowledge
- English benchmark for Compositional Distributional Semantic Models
- 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 (477 unique verb lemmas, 290 unique adjectives, 142 unique adverbs and 1099 unique nouns)
- corpus used at the SemEval 2014
Examples

- AcBBcA; 3.8
  A = Two children are lying in the snow and are making snow angels.
  B = There is no child lying in the snow and making snow angels.

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

1. Keep the inference labels of SICK
2. Keep the relatedness labels
3. Have a naturally sounding corpus in Portuguese
SICK-BR steps

- Pre-processing and Machine Translation
- Guidelines and Annotators training
- Post-processing and Reconstruction
- Checking labels
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 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 (all linguists) reviewed 600 sentences and difficult cases were checked by an experienced annotator

- Glossary
- Everyone sees everyone’s work
- ”I don’t know” is a possible answer
- Ask for double checking
- Online forum (more than 2k messages!)
Checking labels

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

Two main conclusions:
(i) relatedness labels are very subjective
(ii) some SICK inference labels are wrong
### Relatedness labels

- **4305** A woman is not riding a horse/A woman is riding a horse **CONTRACTION 4.5**
- **4587** A woman is riding a horse/A woman is not riding a horse **CONTRACTION 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 touch the labels for now.
SICK-BR results

- Our hypothesis that it is possible to re-use the semantic annotation (insisting on linguistic strategies for translation and adaptation) has been confirmed.
- We have an open Portuguese NLI corpus.
- Aligned to English SICK.
- We could correct ungrammatical and non-sensical sentences, typos and managements mistakes, therefore SICK-BR seems to have a better quality.
- However, we still have labels we don’t agree with.
Challenges

- Funding for open projects is hard
- Integration of Computer Scientists and Linguists can be complicated
- Training (very few Linguistics under-graduate courses and almost no NLP courses) is scarce
- Open tools for annotation are required
Thanks!

livyreal@gmail.com
References


(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), 22 September 2017. Held in Montpellier, France


(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)

Previous open UD corpora for Portuguese

- UD_Portuguese 1.2: subset of Bosque, automatically converted to CoNLL (by HamleDT project, 2011). Converted again to UD in 2015.
- UD 1.3 one additional corpus, Portuguese-BR (from Google’s treebanks), a conversion of the original work of (McDonald et al., 2013)
- Both have many mistakes and some loss of information due to the conversions
- None of them was revised
Bootstrapping the Bosque-UD creation: conversion

- The conversion grammar ultimately used for the first conversion of Bosque to UD contained some 530 rules;
- 70 were simple feature mapping rules, 130 were local MWE splitting rules, the remaining rules handled UD-specific dependency and function label changes in a context-dependent fashion;
- Manually reviewed by a team
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 Knowledge; Marelli et al. (2014)) to use it as a baseline
- We = Livy Real, Valeria de Paiva (Nuance), Katerina Kalouli (Univ. Konstanz)
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.
SICK: Pre-processing and Machine Translation

- 10k sentence pairs, 6k unique sentences
- State-of-the-art machine translation system
Guidelines and Annotators training

- 10 annotators with linguistic training and Brazilian Portuguese native speakers
- 55 example sentences annotated individually
- Discussion
- Guidelines
- Glossary
Post-processing and Reconstruction

- Use of Glossary to make sure lexical choices are uniform
- Grammar and speller checkers
- Corpus reconstruction (pairing sentences as in the original corpus)
- Rechecking same-sentences pair (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**