Towards Temporal Reasoning in Portuguese

Livy Real[4] **Alexandre Rademaker**[1,2]
Fabricio Chalub[1] Valeria de Paiva[3]

[1]IBM Research, Brazil

[2]Nuance Communications, USA

[3]FGV/EMAp, Brazil

[4]PUC-Rio, Brazil

LDL Workshop 2018

Basic Idea

- To reason with temporal information, need first to mark temporal expressions;

Basic Idea

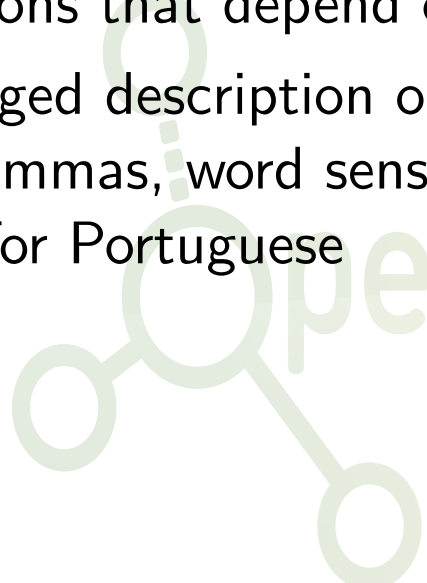
- To reason with temporal information, need first to mark temporal expressions;
- There are several systems for that, but HeidelTime won a competition and has a Portuguese version, so trying it;

Basic Idea

- To reason with temporal information, need first to mark temporal expressions;
- There are several systems for that, but HeidelTime won a competition and has a Portuguese version, so trying it;
- We create a baseline to compare future work to, it serves to start investigating applications that depend on this data;

Basic Idea

- To reason with temporal information, need first to mark temporal expressions;
- There are several systems for that, but HeidelTime won a competition and has a Portuguese version, so trying it;
- We create a baseline to compare future work to, it serves to start investigating applications that depend on this data;
- We aim at a fully fledged description of a temporal logic system, but we need the basics (lemmas, word senses, relationships for temporal expressions) in place for Portuguese

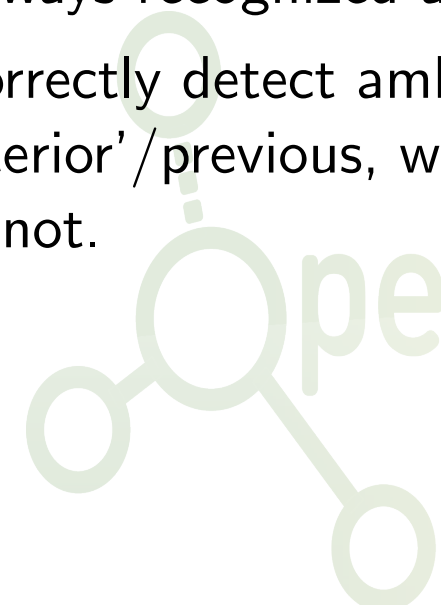


The Experiment I

1. We start by checking how well HeidelTime works for Portuguese and how much of the needed temporal information is in OpenWordNet-PT (OWN-PT);
2. Connecting our lexical resources, we use open linked resources (LLOD); In particular OWN-PT is linked to OMW, which links several other WordNet projects, including TempoWordNet (TempoWN).
3. Contributions:
   - 3.1 Bosque-T, a Portuguese corpus tagged by HeidelTime and a manual assessment of the data produced;
   - 3.2 The improvement of OpenWordNet-PT's synsets related to temporal information;
   - 3.3 An assessment of the quality found in TempoWord-Net and of the usefulness of using its linked knowledge for Portuguese processing.

The Experiment II

4. two-way road: 1) improve the coverage of the lexical resource considering the output of the temporal system; 2) improve the temporal tags, if we have more lexical knowledge.
5. We need to recognize adverbial expressions – such as yesterday, today, tomorrow, respectively 'ontem', 'hoje', 'amanhã' – and these temporal expressions are not always recognized as such;
6. More difficult is to correctly detect ambiguous words, such as 'último'/last and 'anterior'/previous, whether they are used in temporal contexts or not.



OpenWordnet-PT I

http://openwordnet-pt.org

1. Not a simple translation of PWN. Based on PWN architecture, a true thesaurus and dictionary for the Portuguese language.
2. Three language strategies in its lexical enrichment process: (i) translation; (ii) corpus extraction; (iii) dictionaries.
3. Freely available since Dec 2011. Download as RDF files, query via SPARQL or browse via web interface (above).
4. Used by Google Translate, FreeLing, OMW, BabelNet, Onto.PT, etc.
5. Around half the size of PWN, more than twice the size as old Portuguese non-open wordnets
6. The ability to connect the different wordnets helps to complete each one individually.

OpenWordnet-PT II

http://openwordnet-pt.org

7. Due to the construction process, all the original English synsets are present in OWN-PT, but not all of them have Portuguese words and many glosses and examples are still missing.
8. Automatic translations of glosses are available, and they are being manually checked, but the process is ongoing.
9. We are engaged in completing the translation of the empty OWN-PT synsets, long term work, we focus on subsets of synsets related to specific tasks.
10. PWN classifies as temporal nouns in 1028 synsets, the `noun.time` lexicographer file. Of these, around 350 synsets still have no Portuguese translations.

TempoWordNet

1. lexical KB for temporal analysis where each synset of PWN is assigned an intrinsic temporal value.

TempoWordNet

1. lexical KB for temporal analysis where each synset of PWN is assigned an intrinsic temporal value.
2. TempoWN is already linked to OMW, so using its data for improving OWN-PT is easily achieved.

TempoWordNet

1. lexical KB for temporal analysis where each synset of PWN is assigned an intrinsic temporal value.
2. TempoWN is already linked to OMW, so using its data for improving OWN-PT is easily achieved.
3. Each synset of TempoWN is semi-automatically time-tagged with four labels: `atemporal`, `past`, `present` and `future` and a confidence level.

TempoWordNet

1. lexical KB for temporal analysis where each synset of PWN is assigned an intrinsic temporal value.
2. TempoWN is already linked to OMW, so using its data for improving OWN-PT is easily achieved.
3. Each synset of TempoWN is semi-automatically time-tagged with four labels: `atemporal`, `past`, `present` and `future` and a confidence level.
4. In PWN, nouns are easly recognized as temporal, but not other PoS.

TempoWordNet

1. lexical KB for temporal analysis where each synset of PWN is assigned an intrinsic temporal value.
2. TempoWN is already linked to OMW, so using its data for improving OWN-PT is easily achieved.
3. Each synset of TempoWN is semi-automatically time-tagged with four labels: `atemporal`, `past`, `present` and `future` and a confidence level.
4. In PWN, nouns are easly recognized as temporal, but not other PoS.
5. We use TempoWN to check how many temporal adjectives, adverbs and verbs should be in OWN-PT. We aim to detect, amongst the many adjectives, verbs and adverbs that exist in English and that are empty in Portuguese, the ones that are temporally cogent.

HeidelTime

1. multilingual, cross-domain temporal tagger that extracts temporal expressions from documents and normalizes them according to the TIMEX3 annotation standard.

HeidelTime

1. multilingual, cross-domain temporal tagger that extracts temporal expressions from documents and normalizes them according to the TIMEX3 annotation standard.
2. It uses different normalization strategies depending on the domain of the documents that are to be processed, be them news, narratives, colloquial, or scientific.

HeidelTime

1. multilingual, cross-domain temporal tagger that extracts temporal expressions from documents and normalizes them according to the TIMEX3 annotation standard.
2. It uses different normalization strategies depending on the domain of the documents that are to be processed, be them news, narratives, colloquial, or scientific.
3. The tool is a rule-based system and its source code and the resources (patterns, normalization information, and rules) are strictly separated.

UD Portuguese Bosque

1. The Bosque corpus has 9,368 sentences, corresponding to 1,962 different extracts from newspaper text.

UD Portuguese Bosque

1. The Bosque corpus has 9,368 sentences, corresponding to 1,962 different extracts from newspaper text.
2. Since the corpus was extracted from newswire, there are many headlines that are simply noun phrases like ‘PT no governo’ (The Workers Party (PT) in Power).

UD Portuguese Bosque

1. The Bosque corpus has 9,368 sentences, corresponding to 1,962 different extracts from newspaper text.
2. Since the corpus was extracted from newswire, there are many headlines that are simply noun phrases like ‘PT no governo’ (The Workers Party (PT) in Power).
3. There are also dialogues, recognizable through the use of the names of the interlocutors, and answers to questions, which tend not to be full grammatical sentences.

UD Portuguese Bosque

1. The Bosque corpus has 9,368 sentences, corresponding to 1,962 different extracts from newspaper text.
2. Since the corpus was extracted from newswire, there are many headlines that are simply noun phrases like 'PT no governo' (The Workers Party (PT) in Power).
3. There are also dialogues, recognizable through the use of the names of the interlocutors, and answers to questions, which tend not to be full grammatical sentences.
4. Still, Bosque is probably the most used corpus in the Lusophone community, both Brazilian and European Portuguese variants, annotated using several different linguistic theories.

UD Portuguese Bosque

1. The Bosque corpus has 9,368 sentences, corresponding to 1,962 different extracts from newspaper text.
2. Since the corpus was extracted from newswire, there are many headlines that are simply noun phrases like 'PT no governo' (The Workers Party (PT) in Power).
3. There are also dialogues, recognizable through the use of the names of the interlocutors, and answers to questions, which tend not to be full grammatical sentences.
4. Still, Bosque is probably the most used corpus in the Lusophone community, both Brazilian and European Portuguese variants, annotated using several different linguistic theories.
5. Most recently it has been converted to Universal Dependencies version 2.0 (Rademaker et al., 2017). The statistics derived from the UD annotation of the corpus are useful for the work of temporal extraction.

UD Portuguese Bosque

1. The Bosque corpus has 9,368 sentences, corresponding to 1,962 different extracts from newspaper text.
2. Since the corpus was extracted from newswire, there are many headlines that are simply noun phrases like 'PT no governo' (The Workers Party (PT) in Power).
3. There are also dialogues, recognizable through the use of the names of the interlocutors, and answers to questions, which tend not to be full grammatical sentences.
4. Still, Bosque is probably the most used corpus in the Lusophone community, both Brazilian and European Portuguese variants, annotated using several different linguistic theories.
5. Most recently it has been converted to Universal Dependencies version 2.0 (Rademaker et al., 2017). The statistics derived from the UD annotation of the corpus are useful for the work of temporal extraction.

Bosque-T I

1. This is similar to the work on TimeBank-PT but using open and state-of-the-art tools. TimeBank-PT is a translation from EN.
2. Out of the 1962 extracts, HeidelTime says 741 have no time annotations at all.
   'same month last year' and 'daily average': "Em relação ao **mesmo mês do ano passado**, quando os negócios atingiram 139,8 toneladas de ouro, a redução é de 61,37%. A **média diária naquele mês** foi de 6,6 toneladas, segundo dados da Bolsa de Mercadorias e Futuros."
3. Given that HeidelTime is rule-based, we expected that it would be able to detect all expressions composed by digits or expressions that tend to be always related to time, as the names of the months.
   "A cotação para **maio** ficou em 20.000 pontos" and "Empresa funciona das 9h a's 19h, **diariamente**."

Bosque-T II

4. HeidelTime identified 2464 tags, 644 unique ones, of different types. Most of the ones identified were dates. Almost 300 timex occurrences were the word ontem (yesterday). Several temporal expressions were correctly marked, from full dates such as dia 23 de maio de 1972 (day 23 of May of 1972) to some complex phrases such as 'há cerca de 20 anos' (around 20 years ago).
5. Mistakes? "Manifestações espontâneas em protesto contra o facto de Daniel Cohn-Bendit, líder do **Maio de 68**, ter sido proibido de residir em França." The French political movement 'May of 68' is named entity or date?
6. we choose random 20 extracts from Bosque-T to verify. Many temporal expressions are missed or half-marked.

Bosque-T III

7. In “A mudança do local de jogo que deve acontecer também na partida contra o Corinthians, no <TIMEX3>próximo</TIMEX3> dia 17 foi determinada pela CBF, que não viu garantias de segurança no estádio santista.” it missed ‘dia 17’.
8. traditional way of referring to the past in Portuguese is missing altogether from the terms produced.
   In “Monique, 37, disse que descobriu a marquinha, que não é pedra no rim quando se separou do marido, em **junho passado**.” no ‘passado‘ in the annotations.
9. more subtle “Eles se dizem oposição, mas **ainda** não informaram o que vão combater.” (event not happened)
10. “A seca que atingiu as áreas produtoras de grãos não deve causar grandes estragos na safra <TIMEX3>1994</TIMEX3>/95.” missed 1995 year.

Bosque-T IV

11. “Pizzaria oferece cardápio especial para Páscoa.” missed ‘Easter’ holiday that we are adding in OWN-PT.
12. We are now in the process of checking the markings we have and verifying their accuracy. We plan ‘triangulate’ information provided by OWN-PT with the HeidelTime tags in the near future.

Linked Open Data for Temporal Tagging I

1. From TempoWN scores, we considered only the synsets whose probability of being PAST or FUTURE according to TempoWordNet is above 90 percent.
2. This includes more than 3K synsets. TempoWN is not manually curated, we started to manually check the quality of it - we found many labels that we do not agree with and that do not seem very useful for the present task. Too noisy.
3. `PAST:0.998: 00012689-a: ideal | constituting or existing only in the form of an idea or mental image or conception.`
4. Checking most frequent timex expressions in Bosque-T in TempoWN and OWN-PT, we could complete some missing synsets in Portuguese, but we should not use the extra time score offered by TempoWN.

Linked Open Data for Temporal Tagging II

5. The markings of adjectives and adverbs should be useful for reasoning with texts in Portuguese, if the probability assignments are reasonable. Many of them seem good, but how to improve TempoWN scores is future work.
6. Many of the TE found in Bosque-T were missing in OWN-PT `00065748-r | last | most recently`. While in English, this is clearly an adverb, in Portuguese, we need an adverbial phrase *por último* ("by last").
7. For this preliminary work more than 300 temporal synsets were completed in OWN-PT. Many language or culture specific ones are still missing.

Linked Open Data for Temporal Tagging III

8. Typical holidays in the United States, such as the synset `15189982-n` for *Father's Day*. There is a holiday called Father's Day (*Dia dos Pais*) in Portuguese. But it happens at different times and this synset holds a relationship with June, which only makes sense for the English wordnet.
9. smaller differences between the languages. We do not use a prefix like "mid" in the synset `15211711-n` for *mid-May*; we say instead **meados de maio**, a multi-word expression (?), is compositional in Portuguese and therefore it may not necessarily be included in a Portuguese lexical base if multilingual alignment was not a previous goal.

Conclusions I

1. started investigating temporal expressions in PT
2. need to mark temporal ones, used HeidelTime, need to make sure they are in OWN-PT, UDs might help find them - if we can connect the processing of the two, working on that (perhaps) later on.
3. issues at the intersection of multilingual and multicultural aspects of lexical and world knowledge.
4. We are interested in temporal reasoning, not only in temporal IR. As a long term goal, we aim to merge temporal information with other linguistic levels.
5. We plan to use the data in the Portuguese DBPedia to help with some of the culturally specific problems - named holidays.
6. We've got Bosque-T (with some temporal annotations) to play with and improve. Both HeidelTime and Bosque are opensource, whoever wants to improve it, can do it. (even undergrad Linguistics students!)