ASAPP: Automatic Semantic Alignment for Phrases applied to Portuguese

Ana Alves¹,², Hugo Oliveira¹, Ricardo Rodrigues¹,²

¹ - CISUC, Dep. of Informatics Engineering, University of Coimbra, Portugal
² - IPC, Polytechnic Institute of Coimbra, Portugal
This work was supported by the REMINDS project - FCT-UTAP-ICDT/EEI-CTP/0022/2014

July 2016
1 Introduction

2 NLP Resources and Tools

3 Feature Extraction

4 ASAPP Pipeline

5 Results

6 Conclusions
Semantic similarity between sentences in Portuguese (both Brazilian and European)
Recognition of textual entailment

Background:
- Machine Learning and Natural Language Processing
- Semantic Similarity as a function of lexical, syntactic, semantic and distributional features (explained soon)
Approach:

- Adapting ASAP to Portuguese resources and tools
- Application of supervised machine learning
  - Preprocessing: extract the same kind of features (as in ASAP) for both Portuguese variations
  - Regression Analysis: discover a semantic similarity function
  - New challenge: build a classifier for textual entailment given train dataset
Resources I:

- **PAPEL** [Gonçalo Oliveira et al.2008], relations extracted from Porto Editora’s Dicionário da Língua Portuguesa, using grammars based on regularities in the definitions;
- **Dicionário Aberto** [Simões, Sanromán e ao Almeida2012], relations extracted using the grammars of PAPEL;
- **Wikcionário.PT**, relations extracted using the grammars of PAPEL;
- **TeP** [Maziero et al.2008], thesaurus that groups words with their synonyms + antonymy relations;
Resources II:

- **OpenThesaurus.PT**, similar to the previous, but smaller and without antonymy;
- **OpenWordNet-PT** [de Paiva, Rademaker e de Melo2012], open Portuguese wordnet;
- **PULO** [Simoes e Guinovart2014], another Portuguese wordnet, smaller than the previous;
- **CONTO.PT** [Gonçalo Oliveira2016], a fuzzy wordnet based on the redundancy of previous resources.
Tools

- Apache OpenNLP\(^1\) with already trained maximum entropy models
  - Tokenization
  - POS-tagging

- LemPORT [Rodrigues, Gonçalo-Oliveira e Gomes2014], rule-based lemmatizer that uses a lexicon of lemmas and derivative words.
  - Lematization

- NE recognizer: an Apache OpenNLP model trained by our team over Amazónia\(^2\)
  - Named Entity Recognition

- Chunker: same as previous, an Apache OpenNLP model trained by our team, now over Bosque 8.0.

\(^1\)http://opennlp.apache.org/
\(^2\)http://www.linguateca.pt/floresta/corpus.html
Given two sentences $t$ and $h$... (1)

- **Lexical Features**
  - number of common lemmas
  - number of negative words and expressions in each sentence ($C_{nt}$ and $C_{nh}$)
  - and their absolute difference ($|C_{nt} - C_{nh}|$)

- **Morphosyntactic Features**
  - number of Noun, Verb and Prepositional Phrases in each sentence ($C_{npt}$, $C_{vpt}$, $C_{ppt}$, $C_{nph}$, $C_{vph}$ and $C_{pph}$)
  - and their absolute difference ($|C_{npt} - C_{npt}|$, $|C_{vpt} - C_{vph}|$, $|C_{ppt} - C_{pph}|$)
Given two sentences $t$ and $h$... (II)

- **Semantic Features:**
  - all semantic similarity metrics computed by the RECICLAGEM system, are considered as individual features
  - additionally, simple accounting of four semantic relations present in PAPEL, such as: synonyms, hypernyms/hyponyms, antonyms and others.

- **Example:**
  
  \[t = \text{"Além de Ishan, a polícia pediu ordens de detenção de outras 11 pessoas, a maioria deles estrangeiros."},
  
  \[h = \text{"Além, de Ishan, a polícia deu ordem de prisão para outras 11 pessoas, a maioria estrangeiros."} \]

- **Synonyms** = 3 – \{(polícia, ordem), (ordem, polícia), (detenção, prisão)\}
- **Hyponyms** = 1 – \{(estrangeiro, pessoa)\}
- **Antonyms** = 0
- **Others** = 2 – \{(polícia SERVE_PARA ordem), (ordem FAZ_SE_COM polícia)\}
Given train dataset ...

1. **Preprocessing**: Tokenization, PoS tagging, Lemmatization, Chunking and NER applied to each sentence

2. **Feature Extraction**: Negative expressions, NPs, VPs, PPs, NEs, RECICLAGEM’ semantic metrics applied to each pair of sentences

3. **Train**: Using Weka\(^3\) to perform regression analysis over similarity metric, and classification over textual entailment

4. **Evaluation**: Using 10-fold cross validation and ensemble learning approach ...
   - To select the top-3 best regression ensemble algorithms to build two models which computes similarity for both PT-PT and PT-BR
   - To choose the top-3 best classification ensemble algorithms to build two classifiers which predict entailment for both PT-PT and PT-BR

\(^3\)http://www.cs.waikato.ac.nz/ml/weka/
Given test dataset ...

6. Test
   - To compute *Pearson* correlation and MSE over semantic similarity results for each of the 3x2 regression models previously selected.
   - To compute accuracy and *F1* over textual entailment predicted for each of the 3x2 classifiers previously selected.
Selected ensemble algorithms during train phase

<table>
<thead>
<tr>
<th>Run</th>
<th>Entailment</th>
<th>Similarity</th>
</tr>
</thead>
</table>
10-fold cross validation over built models

<table>
<thead>
<tr>
<th>Run</th>
<th>Entailment Accuracy</th>
<th>F1</th>
<th>Similarity Pearson</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - PTBR</td>
<td>79.87%</td>
<td>0.767</td>
<td>0.620</td>
<td>0.677</td>
</tr>
<tr>
<td>1 - PTPT</td>
<td>78.27%</td>
<td>0.766</td>
<td>0.715</td>
<td>0.613</td>
</tr>
<tr>
<td>2 - PTBR</td>
<td>80.77%</td>
<td>0.765</td>
<td>0.622</td>
<td>0.677</td>
</tr>
<tr>
<td>2 - PTPT</td>
<td>78.73%</td>
<td>0.765</td>
<td>0.716</td>
<td>0.612</td>
</tr>
<tr>
<td>3 - PTBR</td>
<td>76.50%</td>
<td>0.759</td>
<td>0.635</td>
<td>0.668</td>
</tr>
<tr>
<td>3 - PTPT</td>
<td>77.77%</td>
<td>0.775</td>
<td>0.723</td>
<td>0.606</td>
</tr>
</tbody>
</table>
## Test: final results of ASAPP’ runs

<table>
<thead>
<tr>
<th>Run</th>
<th>Entailment Accuracy</th>
<th>F1</th>
<th>Similarity Pearson</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - PTBR</td>
<td>81.20%</td>
<td>0.5</td>
<td>0.65</td>
<td>0.44</td>
</tr>
<tr>
<td>1 - PTPT</td>
<td>77.75%</td>
<td>0.57</td>
<td>0.68</td>
<td>0.70</td>
</tr>
<tr>
<td>2 - PTBR</td>
<td>81.56%</td>
<td>0.47</td>
<td>0.65</td>
<td>0.44</td>
</tr>
<tr>
<td>2 - PTPT</td>
<td>78.90%</td>
<td>0.58</td>
<td>0.67</td>
<td>0.71</td>
</tr>
<tr>
<td>3 - PTBR</td>
<td>77.10%</td>
<td>0.5</td>
<td>0.65</td>
<td>0.44</td>
</tr>
<tr>
<td>3 - PTPT</td>
<td>74.35%</td>
<td>0.59</td>
<td>0.68</td>
<td>0.73</td>
</tr>
</tbody>
</table>
In short

- Our team participated with two systems RECICLAGEM and ASAPP
- ASAPP is a supervised learning which considers NLP features from 3 dimensions: Lexical, Morphosyntactic, Semantic
- Combining classifiers and regression analyzers of known existing algorithms: ensemble learning
- Each run was composed of results from single classification method and a regression analysis for both Portuguese variations
- ASAPP showed the best accuracy for Brazilian Portuguese
Future work

- To use more tools and resources for Portuguese not considered on this first implementation of ASAPP.
- To consider a 4th NLP dimension on feature extraction: distributional features [Alves et al. 2014].
- To apply feature selection over all extracted features in order to select the most relevant.
References


Gonçalo Oliveira, Hugo, Diana Santos, Paulo Gomes, e Nuno Seco.
2008.
PAPEL: A dictionary-based lexical ontology for Portuguese.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, e Ian H. Witten.
2009.
The weka data mining software: An update.

1998.
On combining classifiers.

Kuncheva, Ludmila I.
2004.
Combining Pattern Classifiers: Methods and Algorithms.
Wiley-Interscience.

Mackay, David J.C.
1998.
Introduction to gaussian processes.


ASAPP: Automatic Semantic Alignment for Phrases applied to Portuguese

Ana Alves\textsuperscript{1,2}, Hugo Oliveira\textsuperscript{1}, Ricardo Rodrigues\textsuperscript{1,2}

1 - CISUC, Dep. of Informatics Engineering, University of Coimbra, Portugal
2 - IPC, Polytechnic Institute of Coimbra, Portugal
This work was supported by the REMINDS project - FCT-UTAP-ICDT/EEI-CTP/0022/2014

July 2016