ASAPP: Automatic Semantic Alignment for Phrases applied to Portuguese

Ana Alves[1,2],Hugo Oliveira[1],Ricardo Rodrigues[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

COGNITIVE & MEDIA SYSTEMS

CISUC

U C

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

- Machine Learning and Natural Language Processing
- Past participation in SemEval14 [Alves et al.2014] and SemEval15 [Alves et al.2015] on Semantic Textual Similarity tasks: ASAP (Automatic Semantic Alignment for Phrases)
- 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 redundacy 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... (I)

- Lexical Features
  - number of common lemmas
  - number of negative words and expressions in each sentence (Cn_t and Cn_h)
  - and their absolute difference ($|Cn_t - Cn_h|$)
- Morphosyntactic Features
  - number of Noun, Verb and Prepositional Phrases in each sentence (Cnp_t, Cvp_t, Cpp_t, Cnp_h, Cvp_h and Cpp_h)
  - and their absolute difference ($|Cnp_t - Cnp_t|$, $|Cvp_t - Cvp_h|$, $|Cpp_t - Cpp_h|$)

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 ="Além de Ishan, a polícia pediu ordens de detenção de outras 11 pessoas, a maioria deles estrangeiros.",
      h ="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

Run	Entailment	Similarity
1	Majority voting from 3 classifiers [Kittler et al.1998, Kuncheva2004]	Aditive Regression by *Boosting* [Friedman1999]
2	Majority voting from 5 classifiers [Kittler et al.1998, Kuncheva2004]	Multischeme selection [Hall et al.2009]
3	Automatic Feature Selection [Hall et al.2009]	Simple Gaussian Process [Mackay1998]

10-fold cross validation over built models

Run	Entailment Accuracy	F1	Similarity Pearson	MSE
1 - PTBR	79.87%	**0.767**	0.620	0.677
1 - PTPT	78.27%	**0.766**	0.715	0.613
2 - PTBR	**80.77%**	0.765	0.622	0.677
2 - PTPT	**78.73%**	0.765	0.716	0.612
3 - PTBR	76.50%	0.759	**0.635**	**0.668**
3 - PTPT	77.77%	0.775	**0.723**	**0.606**

Test: final results of ASAPP' runs

Run	Entailment Accuracy	F1	Similarity Pearson	MSE
1 - PTBR	81.20%	0.5	**0.65**	**0.44**
1 - PTPT	77.75%	0.57	**0.68**	**0.70**
2 - PTBR	**81.56%**	0.47	0.65	0.44
2 - PTPT	**78.90%**	0.58	0.67	0.71
3 - PTBR	77.10%	**0.5**	0.65	0.44
3 - PTPT	74.35%	**0.59**	0.68	0.73

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 I

Alves, Ana, David Simões, Hugo Gonçalo Oliveira, e Adriana Ferrugento.
2015.
Asap-ii: From the alignment of phrases to textual similarity.
Em *Proceedings of 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 184–189, Denver, Colorado, June, 2015. ACL Press.

Alves, Ana O., Adriana Ferrugento, Mariana Lourenço, e Filipe Rodrigues.
2014.
Asap: Automatic semantic alignment for phrases.
Em *SemEval Workshop, COLING 2014, Ireland*.

de Paiva, Valeria, Alexandre Rademaker, e Gerard de Melo.
2012.
OpenWordNet-PT: An open Brazilian wordnet for reasoning.
Em *Proceedings of 24th International Conference on Computational Linguistics*, COLING (Demo Paper).

Friedman, J.H.
1999.
Stochastic gradient boosting.
RelatÃșrio tÃľcnico, Stanford University.

Gonçalo Oliveira, Hugo.
2016.
CONTO.PT: Groundwork for the Automatic Creation of a Fuzzy Portuguese Wordnet.
Em *Proceedings of 12th International Conference on Computational Processing of the Portuguese Language (PROPOR 2016)*, pp. in press, Tomar, Portugal, July, 2016. Springer.

References II


Gonçalo Oliveira, Hugo, Diana Santos, Paulo Gomes, e Nuno Seco.
2008.
PAPEL: A dictionary-based lexical ontology for Portuguese.
Em *Proceedings of Computational Processing of the Portuguese Language – 8th International Conference (PROPOR 2008)*, volume 5190 of *LNCS/LNAI*, pp. 31–40, Aveiro, Portugal, September, 2008. Springer.


Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, e Ian H. Witten.
2009.
The weka data mining software: An update.
SIGKDD Explor. Newsl., 11(1):10–18.


Kittler, J., M. Hatef, Robert P.W. Duin, e J. Matas.
1998.
On combining classifiers.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3):226–239.


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


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

References III

Maziero, Erick G., Thiago A. S. Pardo, Ariani Di Felippo, e Bento C. Dias-da-Silva.
2008.
A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil.
Em *VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pp. 390–392.

Rodrigues, Ricardo, Hugo Gonçalo-Oliveira, e Paulo Gomes.
2014.
LemPORT: a High-Accuracy Cross-Platform Lemmatizer for Portuguese.
Em Maria João Varanda Pereira, José Paulo Leal, e Alberto Simões, editores, *Proceedings of the 3rd Symposium on Languages, Applications and Technologies (SLATE '14)*, OpenAccess Series in Informatics, pp. 267–274, Germany, June, 2014. Schloss Dagstuhl — Leibniz-Zentrum für Informatik, Dagstuhl Publishing.

Simões, Alberto, Álvaro Iriarte Sanromán, e José Jo ao Almeida.
2012.
Dicionário-Aberto: A source of resources for the Portuguese language processing.
Em *Proceedings of 10th International Conference on the Computational Processing of the Portuguese Language (PROPOR 2012)*, volume 7243 of *LNCS*, pp. 121–127, Coimbra Portugal, April, 2012. Springer.

Simoes, Alberto e Xavier Gómez Guinovart.
2014.
Bootstrapping a Portuguese wordnet from Galician, Spanish and English wordnets.
Em *Advances in Speech and Language Technologies for Iberian Languages*, volume 8854 of *LNCS*, pp. 239–248.

ASAPP: Automatic Semantic Alignment for Phrases applied to Portuguese

Ana Alves[1,2],Hugo Oliveira[1],Ricardo Rodrigues[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

COGNITIVE & MEDIA SYSTEMS

CISUC

U C