MiniExperts: An SVM Approach for Measuring Semantic Textual Similarity

Hernani Costa[a*], Hanna Béchara[b*], Shiva Taslimipoor[b], Rohit Gupta[b], Constantin Orăsan[b], Gloria Corpas Pastor[a] and Ruslan Mitkov[b]
{hercos,hanna.bechara,shiva.taslimi,r.gupta, c.orasan,gcorpas,r.mitkov}@{[a]wlv.ac.uk,[b]uma.es}
*These two authors contributed equally to this work.

[a]LEXYTRAD, University of Malaga, Spain
[b]RIILP, University of Wolverhampton, UK

Introduction

- Semantic similarity measures play an important role in a wide variety of Natural Language Processing (NLP) applications
- SemEval2015's Task 2 involves computing how similar two sentences are in English (Subtask 2a) and Spanish (Subtask 2b)
- In order to solve this challenge, the University of Wolverhampton and the University of Malaga submitted an improved and revised version of the system presented last year [1][a]
  - the current version employs a Machine Learning method which exploits available NLP technology, features inspired by deep semantics (e.g. parsing and paraphrasing) with distributional similarity measures, conceptual similarity measures, semantic similarity measures and corpus pattern analysis

[a] https://github.com/rohitguptacs/wlvsimilarity

Data Preprocessing

- POS-Tagger, Lemmatiser, Stemmer
  - Stanford CoreNLP[a], TT4J[b] and Snowball[c]
- Named Entity Recogniser
  - Apache OpenNLP[d]
- Translation Model
  - PB-SMT system Moses [2]
- Resources
  - two lists: Stopwords[e]; and Multiword Expressions (MWEs) with their likelihood scores extracted from the Europarl corpus

[a] http://nlp.stanford.edu/software/corenlp.shtml
[b] https://code.google.com/p/tt4j
[c] http://snowball.tartarus.org
[d] http://opennlp.apache.org
[e] https://github.com/hpcosta/stopwords

Extracted Features

a) Baseline Features

- previously developed for SemEval2014, which consists of 13 features explained in detail in [1]

b) Multiword Expressions

- whenever a `verb+noun` and `verb+particle` combination occurs in a sentence pair, we search for them in the prepared MWEs list
- then, the degree of association of these combinations served as a feature

c) Conceptual Similarity Measures

- we created a conceptual sentence for all input pair of sentences (English and Spanish) using BabelNet[a] [3]
  - for every pair of sentences a two conceptual term lists were built by extracting all the occurrences of the terms in the BabelNet conceptual network
  - then, the terms in sentence 1 were intersected with all the conceptual term lists in sentence 2
  - finally, Jaccard' [4], Lin' [5] and PMI' [6] scores were calculated

[a] http://babelnet.org

d) Semantic Similarity Measures

- Align, Disambiguate and Walk (ADW)[a] library [7]
  - with and without disambiguation: WeightedOverlap, Cosine, Jaccard, KLDivergence and JensenShannon divergence

e) Distributional Similarity Measures

- Spearman's Rank Correlation Coefficient and the Chi-Square [8]
  - both language-independent and independent of text size

[a] http://lcl.uniroma1.it/adw

Results

Subtask 2a – Pearson Correlation for English

	Run-1	Run-2	Run-3
answers-forums	0.6781	0.6454	0.6179
answers-students	0.7304	0.7093	0.6977
belief	0.6294	0.5165	0.3236
headlines	0.6912	0.6084	0.5775
images	0.8109	0.7999	0.7954
mean	***0.7216***	***0.6746***	***0.6353***
rank (out of 74)	***33***	***45***	***55***

Subtask 2b – Pearson Correlation for Spanish

	Run-1	Run-2	Run-3
wikipedia	0.5239	0.4671	0.4402
newswire	0.5076	0.5437	0.5524
mean	***0.5158***	***0.5054***	***0.4963***
rank (out of 17)	***9***	***10***	***11***

Conclusions and Future Work

- This work presents an efficient approach to calculate semantic relatedness for both English and Spanish sentence pairs
- Our system performed:
  - satisfactorily for English ($\rho = 0.7216$, ranked 33 out of 74)
  - less adequately for Spanish ($\rho = 0.5158$, ranked 9 out of 17)
- In the future we plan to extract the conceptual description provided by the BabelNet network in order to match it with the conceptual terms

Acknowledgements

Hernani Costa, Hanna Béchara and Rohit Gupta are supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement nº 317471. Also, the research reported in this work has been partially carried out in the framework of the Educational Innovation Project TRADICOR (PIE 13-054, 2014-2015); the R&D project INTELITERM (ref. nº FFI2012-38881, 2012-2015); and the R&D Project for Excellence TERMITUR (ref. nº HUM2754, 2014-2017).

References

[1] R. Gupta, H. Bechara, I. El Maarouf, and C. Orasan, "UoW: NLP techniques developed at the University of Wolverhampton for Semantic Similarity and Textual Entailment," in *8th Int. Workshop on Semantic Evaluation (SemEval'14)*, (Dublin, Ireland), pp. 785–789, ACL and Dublin City University, 2014.

[2] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," in *45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pp. 177–180, 2007.

[3] R. Navigli and S. Paolo Ponzetto, "BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network," *Artificial Intelligence*, vol. 193, pp. 217–250, 2012.

[4] P. Jaccard, "Étude comparative de la distribution florale dans une portion des alpes et des jura," *Bulletin del la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 547–579, 1901.

[5] D. Lin, "An Information-Theoretic Definition of Similarity," in *15th Int. Conf. on Machine Learning*, ICML'98, (San Francisco, CA, USA), pp. 296–304, Morgan Kaufmann, 1998.

[6] P. D. Turney, "Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL," in *12th European Conf. on Machine Learning*, EMCL'01, (London, UK), pp. 491–502, Springer, 2001.

[7] M. T. Pilehvar, D. Jurgens, and R. Navigli, "Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity," in *51st Annual Meeting of the ACL - Volume 1*, (Sofia, Bulgaria), pp. 1341–1351, ACL, 2013.

[8] A. Kilgarriff, "Comparing Corpora," *Int. Journal of Corpus Linguistics*, vol. 6, no. 1, pp. 97–133, 2001.