Darja Fišer

Faculty of Arts, University of Ljubljana

Aškerčeva 2, SI-1000 Ljubljana

Jožef Stefan Institute

Jamova cesta 39, SI 1000 Ljubljana

Subject:	Scientific Report on Short Term Scientific Mission
Reference:	COST Action IS1305
Host institution:	Natural Language Processing Centre at the Faculty of Informatics, Masaryk University, Brno
Period:	10/10/2016 to 10/12/2016
Reference code:	COST-STSM-ECOST-STSM-IS1305-101016-080738
Amount up to:	EUR 700

Semantic shift detection in Slovene netspeak

a. Background and motivation

Meanings of words are not fixed but undergo changes, either due to the advent of new word senses or due to established word senses taking new shades of meaning or becoming obsolete (Mitra et al. 2015). These semantic shifts typically occur systematically (Campbell 2004), resulting in a meaning of a word to either expand/become more generalized, narrow down to include fewer referents or shift/transfer to include a new set of referents (Sagi et al. 2009). There are also many cases in which words acquire new positive or negative connotations, processes that lexical semanticists call amelioration and pejoration (Cook and Stevenson 2009).

While automatic discovery of word senses has been studied extensively (Spark-Jones 1986; Ide and Veronis 1998; Schütze 1998; Navigli 2009), changes in the range of meanings expressed by a word have received much less attention, despite the fact that it is a very important challenge in lexicography where it is needed to keep the description of dictionary entries up-to-date. Apart from lexicography, up-to-date semantic inventories are also required for a wide range of human-language technologies, such as question-answering and machine translation. As more and more diachronic, genre- and domain-specific corpora are becoming available, it is becoming an increasingly attainable goal.

b. Goals of the proposed project

On the STSM we investigated semantic shift detection in Slovene Twitter corpora with respect to the reference corpus. We believe that user-generated content is an ideal resource to detect semantic shifts due to its increasing popularity and heterogeneous use(r)s, the language of which is all the more valuable because it is not covered by any of the existing traditional authoritative lexical and language resources.

c. Methodology and corpora

We used the Janes corpus of tweets that was developed within the basic national research project on building resources, methods and tools for the processing of non-standard Slovene (http://nl.ijs.si/janes/english/) containing 100 million tokens. As the reference corpus, we used the 1-billion token Gigafida (Logar et al. 2012).

We tested the suitability of using word embeddings to identify semantic shifts in user-generated content. This is a simple approach that relies on the basic principles of distributional semantics suggesting that one can model the meaning of a word by observing the contexts in which it appears (Firth 1957). Vector models position words in a semantic space given the contexts in which the words appear, making it possible to measure the semantic similarity of words as the distance between the positions in the semantic space, with CBOW and skip-gram (Mikolov et al., 2013) being nowadays the most widely used models.

In addition, we developed a preliminary supervised method for our problem by running a series of supervised learning experiments, trying to discriminate between (1) noise (preprocessing errors) and (2) signal (lexemes of interest), as well as (1) lexemes experiencing semantic shift and (2) lexemes without semantic shift among the manually annotated candidates annotated.

d. Results and contribution of the project

We performed linguistic analysis on the top-ranking 200 lemmas from the reference and the Twitter corpus which display the most differences in their contexts.

A detailed comparative analysis was performed by comparing Word Sketches of the same lemma in both corpora in the Sketch Engine concordancer (Kilgarriff et. al. 2014). The analysis of semantic shifts was performed in three steps. First, we tried to determine whether any semantic shift can be detected. If yes, we further tried to determine whether the shift is minor or major. Finally, they were then classified into three subcategories each. The results are summarized in the table below.

Table 1: Types of semantic shifts in Slovene tweets

	No.	%
No shift	28	25%
Minor shift	21	19%
Semantic narrowing	3	3%
Usage pattern	6	5%
Redistribution of senses	12	11%
Major shift	61	56%
CMC-specific	6	5%
Colloquial	23	21%
Events	32	29%

The analysis shows that apart from the noise due to preprocessing errors (45%) that are easy to spot, the approach yields a lot of highly valuable semantic shift candidates, especially the novel senses occurring due to daily events and the ones produced in informal communication settings. The results of this experiment will be used in the development of the dictionary of Slovene Twitterese (Gantar et al. 2016).

The contribution of the project is two-fold. First, it resulted in a valuable source of lexicographic information of contemporary Slovene which is lacking a comprehensive, updated and corpus-based description of the lexical inventory. Second, a semantic shift detection algorithm was developed, enabling other researchers to perform first insights into the evolution of the languages on the web. The approach is data-driven and language-independent, so that it can be applied to other languages and different genres and registers as well.

The results of the preliminary supervised approach were just slightly better than the most frequent class baseline, which made us drop the supervised approach, at least for now. Annotating more data in the future could be a way forward.

e. Publication

Based on the results of the work performed on the STSM, we published a paper that was presented at the 10th Workshop on Recent Advances in Slavonic Natural Language Processing:

- Fišer, D., Ljubešić, N. 2016. Detecting Semantic Shifts in Slovene Twitterese. Proceedings of the Tenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2016. Brno, Czech Republic. https://nlp.fi.muni.cz/raslan/2016/paper10-Fiser_Ljubesic.pdf

Bibliography

- Blei, DM.M., A.Y. Ng, and M.I. Jordan. 2003. “Latent Dirichlet Allocation.” The Journal of Machine Learning Research 3: 993 - 1022.
- CAMPBELL, L. (2004) Historical linguistics: An introduction. Cambridge, MA: The MIT Press.
- COOK, Paul, STEVENSON, Suzanne (2009) CALC '09 Proceedings of the Workshop on Computational Approaches to Linguistic Creativity, pp. 71-78.
- Firth, JR. 1957. “A Synopsis of Linguistic Theory.”Studies in Linguistic Analysis
- Ide, Nancy, and Jean Véronis. "Introduction to the special issue on word sense disambiguation: the state of the art." Computational linguistics 24.1 (1998): 2-40.
- Mikolov, T.; Yih, W.; Zweig, G. (2013). "Linguistic Regularities in Continuous Space Word Representations.". HLT-NAACL: pp. 746–751.

- Mitchell T.M., 1997, Machine Learning, McGraw-Hill, Inc. New York, NY, USA.
- Mitra, Sunny, et al. "An automatic approach to identify word sense changes in text media across timescales." Natural Language Engineering 21.05 (2015): 773-798.
- Navigli, Roberto. "Word sense disambiguation: A survey." ACM Computing Surveys (CSUR) 41.2 (2009): 10.
- SAGI, E., Kaufmann, S., and Clark, B. (2009) Semantic Density Analysis: Comparing Word Meaning across Time and Phonetic Space. In Proceedings of the EACL 2009 Workshop on GEMS: GEometical Models of Natural Language Semantics, pages 104-111. Athens, Greece.
- Schütze, Hinrich. "Automatic word sense discrimination." Computational linguistics 24.1 (1998): 97-123.
- Spark-Jones, K. (1986). Synonym and Semantic Classification. Edinburgh Information Technology Series. Edinburgh University Press, Edinburgh.