Nikola Ljubešić

Faculty of Humanities and Social Sciences, University of Zagreb

Ivana Lučića 3, HR-10000 Zagreb

Subject:	Scientific Report on Short Term Scientific Mission
Reference:	COST Action IS1305
Host institution:	Lexical Computing CZ s.r.o., Brno (CZ)
Period:	10/10/2016 to 10/12/2016
Reference code:	COST-STSM-IS1305-34619
Amount up to:	EUR 1000

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.

d. Results and contribution of the project

Unsupervised method

Our unsupervised method is based on calculating distributional models of lexemes occurring in two (sub)corpora and ranking them via a distance of the two models.

We developed a method for building distributional models for each headword, one representing the headword in the standard language (from the Gigafida reference corpus), the other in non-standard language (from the Janes Twitter corpus).

Learning sparse representations of same words from different corpora is a straightforward task as these representations require context features to be counted and potentially processed with a statistic of choice. On the other hand, dense representations are based on representing each word in a way that maximises the predictability of a word given its context or vice versa. Given that the representation depends on the data available in each of the corpora, the representation learning for both corpora has to be performed in a single process. To do that, a trick has to be applied: encoding whether an occurrence of a headword came from the standard or non-standard dataset in form of a prefix to the headword itself (like s_miška#Nc for the occurrence in standard data and n_miška#Nc for the occurrence in non-standard data). Therefore the representation cannot be learned from running text as headwords need to have corpus information encoded while their contexts have to be free of that information so that they are shared between the two corpora.The only tool that we know to accept already prepared pairs of headwords and context features is word2vecf (https://bitbucket.org/yoavgo/word2vecf). Other tools accept running text only, limiting thereby the headwords and context features to the same phenomena like surface forms or lemmata.As context features we use surface forms, avoiding thereby the significant noise introduced while tagging and lemmatising non-standard texts. The features are taken from a punctuation-free window of two words to each side of the headword. The relative position of each feature to the headword is not encoded. By following the described method, we produced dense vector representations of 200 dimensions for each of the 5425 lemmas for each of the two corpora.We calculate the semantic shift simply as a cosine similarity, transformed to a distance measure, between the dense representation of a word built from standard and from non-standard data. More formally, for each $w \in V$ where w is a word and V is our vocabulary, we calculate the semantic shift of a word ss(w) $asss(w) = 1 - cossim(ws, wn)$where the cossim function calculates the cosine similarity of two vectors, ws is the 200-dimensional representation of the word calculated on the standard corpus data, and wn the same representation on the non-standard corpus data.

The output of the developed method was linguistically analysed by Darja Fišer inside her STSM, which is reported on in her scientific report.

Supervised method

In addition to the unsupervised method, we ran a series of experiments on identifying semantic shift in a supervised manner. Given the significant amount of preprocessing noise that was caught as semantic shift by our unsupervised method, we investigated two tasks:

1. discrimination between lexemes and noise, and

2. discrimination between lexemes that experience a semantic shift and those that do not experience it

The goal of the first tasks would be to remove noise from the list of candidates, while the second task should identify true semantic shifts among all candidates.

The supervised learning experiments were run in scikit-learn, using SVM classifiers with RBF kernels, grid searching for optimal gamma and C hyperparameters, scaling data to zero mean and unit variance, and performing ten-fold cross-validation for evaluation via weighted F1. As features we used the 200 dense dimensions per lexeme, obtained by subtracting the non-standard dense representation of each lexeme from the standard representation of the lexeme.

The results of both experiments have proven for both tasks to be very hard. In both cases we experienced results just slightly better than the most frequent class baseline.

We plan to continue this line of work, first by annotating more candidates, and second by investigating additional features for representing the problem at hand.

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.