Detecting Semantic Shifts in Slovene Twitterese

Darja Fišer[1,2] and Nikola Ljubešić[2,3]

[1] Dept. of Translation, Faculty of Arts, University of Ljubljana,
Aškerčeva cesta 2, SI-1000 Ljubljana, Slovenia
[2] Department of Knowledge Technologies, Jožef Stefan Institute,
Jamova cesta 3, SI-1000 Ljubljana, Slovenia
[3] Department of Information and Communication Sciences, University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb, Croatia
`darja.fiser@ff.uni-lj.si,nikola.ljubesic@ijs.si`

Abstract. This paper presents first results of automatic semantic shift detection in Slovene tweets. We use word embeddings to compare the semantic behaviour of common words frequently occurring in a reference corpus of Slovene with their behaviour on Twitter. Words with the highest model distance between the corpora are considered as semantic shift candidates. They are manually analysed and classified in order to evaluate the proposed approach as well as to gain a better qualitative understanding of the nature of the problem. Apart from the noise due to preprocessing errors (45%), the approach yields a lot of valuable candidates, especially the novel senses occurring due to daily events and the ones produced in informal communication settings.

Key words: semantic shift detection, distributional semantics, word embeddings, user-generated content, tweets

1 Introduction

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). A classic example of expansion is the noun `miška/mouse` which used to refer to the small rodent but is now also used for describing the computer pointing device. The reverse process occurred with the noun `faks/faxs` that used to mean both the machine for telephonic transmission of printed documents and higher education institution, only the latter of which continues to be of use in contemporary colloquial Slovene.

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). Amelioration, which is especially frequent in

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2016, pp. 43–50, 2016. © Tribun EU 2016

slang, can be observed in the use of the adverb `hudo/terrific` which has a strong negative connotation in standard Slovene but has acquired a distinctly positive one in colloquial Slovene. Pejoration, the opposite effect of semantic shifts, can be observed in the use of the noun for `blondinka/blond woman`, which is neutral in standard Slovene but used distinctly pejoratively in informal settings.

2 Related work

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.

Most work in semantic shift detection focuses on diachronic changes in word usage and meaning by utilizing large historical corpora spanning several decades or even centuries (Mitra et al. 2015, Tahmasebi, Risse and Dietze 2011). Since we wish to look at the differences between standard and non-standard Slovene, our work is closer to the approaches conducted over two time points or corpora. Cook et al. (2013) induce word senses and identify novel senses by comparing the new 'focus corpus' with the 'reference corpus' using topic modelling for word sense induction. We instead chose to follow a simpler and potentially more robust approach which does not require us to discriminate specific senses, but which simply relies on measuring contextual difference of a lexeme in two corpora. In this respect, our work is similar to Gulordava and Baroni (2011) who detect semantic change based on distributional similarity between word vectors.

3 Data

In this paper we investigate semantic shifts in the 100-million token corpus of Slovene tweets (Fišer et al. 2016) with respect to the 1-billion token reference corpus Gigafida (Logar et al. 2012). 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.

We define headwords as lowercased lemmata expanded with the first two characters of the morphosyntactic description. The list of headwords of interest is produced by identifying lemmata with over 500 occurrences in our non-standard dataset that are also covered in the Sloleks lexicon[4] and are either common nouns (`Nc`), general adjectives (`Ag`), adverbs (`Rg`) or main verbs (`Vm`). Thereby we produced a list of 5425 lemmas.

4 Method

In this paper we test 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.

We want to build two 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[5]. 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

[4] `https://www.clarin.si/repository/xmlui/handle/11356/1039`

[5] `https://bitbucket.org/yoavgo/word2vecf`

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 $ss(w)$ as

$$ss(w) = 1 - cossim(\boldsymbol{w_s}, \boldsymbol{w_n})$$

where the *cossim* function calculates the cosine similarity of two vectors, $\boldsymbol{w_s}$ is the 200-dimensional representation of the word calculated on the standard corpus data, and $\boldsymbol{w_n}$ the same representation on the non-standard corpus data.

5 Linguistic analysis of the results

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. 90 (45%) of these were preprocessing errors in either corpus due to language identification, tokenisation, normalisation, tagging or lemmatisation errors (e.g. `talka/female hostage` which was a wrongly assigned lemma to the English word `talk`) and were therefore excluded from further analysis. This level of noise is not surprising as we are dealing with highly non-standard data that is difficult to process with high accuracy. At the same time, our analysis shows that this type of noise is highest at the top of the list and steadily decreases.

A detailed comparative analysis of the remaining 110 lemmas 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 indeed 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 that are described in detail in the following section.

5.1 Minor semantic shifts

As the first type of minor shifts we considered those cases in which we identified the same senses in both corpora but with a different frequency distribution (e.g. `odklop`, which predominantly refers to the disconnecting of electricity, the internet etc. in the reference corpus but is most often used metaphorically in the Twitter corpus in the sense of taking a break, going on holiday or off-line to relax from work and every-day routine or `sesalec`, the predominant sense of which in Gigafida is mammal but vacuum cleaner in the Janes corpus).

Second, we also counted the cases in which distinct discrepancies were detected in the patterns in which a word is regularly used, influencing the sense of the target word (e.g. `kvadrat/square`, which is almost exclusively used in the pattern `na kvadrat/squared` on Twitter, or `eter/ether`, which on Twitter is almost exclusively used in the pattern `v eter/on air`).

The third type of minor shifts we detected is the narrowing of a word's semantic repository that is most likely not a sign of a word sense dying out but rather due to a limited set of topics present in Twitter discussions with respect to the set of topics in the reference corpus (e.g. `posodobiti`, which is only used in the IT sense on Twitter, never as `modernise` in general as is frequent in Gigafida, or `podnapis`, with which Twitter users only referes to subtitles, never to captions below images etc. as is frequent in Gigafida).

5.2 Major semantic shifts

As the first type of major shifts we considered novel usage of words that is a direct consequence of daily events, political situations, natural disasters or social circumstances (e.g. `vztrajnik` which traditionally meant flywheel but started being used to refer to the persistent protesters in the period of political and social unrest in 2012-2013 or `pirat` who used to be confined to the sea but can now be found on the internet as well and even in politics as members of the new party, only the latter in distinctly positive contexts). It would be interesting to track whether such semantic shifts are short-lived or which of them become a permanent part of our lexico-semantic repository.

Second, many new senses in the Twitter corpus can be detected because a lot of informal communication is performed via Twitter and colloquial language is frequent (e.g. `optika` which is used to refer to the lense mechanism in different devices, a store that sells glasses or a viewpoint in standard Slovene but is often used to refer to broadband internet in informal settings, or `carski` which is an adjective to refer to emperor but is used as a synonym of great, wonderful in non-standard language).

Finally, some new communication conventions have emerged on social media which resulted in some novel word senses as well (e.g. `sledilec`, a person who follows you on Twitter or other social media, or `opomnik`, a reminder message on the computer or telephone).

5.3 Results and discussion

As can be seen in Table 1, some type of semantic shift was detected in 75% of the cases in the sample that was analysed, suggesting the proposed approach to be quite accurate, given the complexity of the task. A large majority of all the semantic shifts detected were major shifts (74% of all the shifts detected). Unsurprisingly, most semantic shifts can be attributed to discussing daily events and to using informal, colloquial language on Twitter. At the same time, these are also the most interesting cases from the research perspective because they are still missing in all the available lexico-semantic resources of Slovene, which proves the suitability of the proposed approach for the task. In addition, some highly creative attribution of new meaning to common words has also been detected (e.g. `kahla/potty` which refers to a politician Karel Erjavec who cannot pronounce letter r, or `pingvin/penguin`, a derogatory nickname of the leader of a political party), showing that Twitter users play with language skilfully and are quick to adopt new coinages. The results of the performed linguistic analysis thus show that the approach presented in this paper could significantly contribute to regular semi-automatic updates of corpus-based general as well as specialized dictionaries.

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 detected minor shifts systematically show the differences in the focus and range of topics between the two corpora. The fact that many more novel usages than narrowings were detected suggests that the reference corpus could be further enhanced with texts from social media and other less formal and standard communication practices as they contain rich and valuable linguistic material that is now absent in the reference corpus.

6 Conclusion

In this paper we presented the first results of automatic semantic shift detection for the Slovene used in social media. We measured the semantic shift of a word as the distance between the word embedding representation learned from a reference corpus of Slovene and the word embedding learned on a Twitter corpus of Slovene. We performed a manual analysis of 200 words with the highest measurements. 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).

Our future work will focus on (1) extending the manual analysis to lower-ranked candidates, (2) extending the approach to lower-frequency candidates, (3) comparing our method with alternative methods such as representing words as word sketches / syntactic patterns and (4) using supervised learning for detecting semantic shifts, discriminating between specific types of semantic shifts or filtering preprocessing errors.

Acknowledgments. The work described in this paper was funded by the Slovenian Research Agency within the national basic research project "Resources, Tools and Methods for the Research of Nonstandard Internet Slovene" (J6-6842, 2014-2017) and the COST Action "European Network of e-Lexicography" (IS1305, 2013-2017).

References

1. Blei, DM.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. The Journal of Machine Learning Research 3: 993 – 1022 (2003).
2. Campbell, L. Historical linguistics: An introduction. Cambridge, MA: The MIT Press. (2004)
3. Cook, P., Stevenson, S. CALC '09 Proceedings of the Workshop on Computational Approaches to Linguistic Creativity, pp. 71–78 (2009).
4. Cook, P., Lau, J. H., Rundell, M., McCarthy, D., Baldwin, T.: A lexicographic appraisal of an automatic approach for detecting new word senses. In Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex conference, Tallinn, Estonia, 49–65 (2013).
5. Firth, J.R.: A Synopsis of Linguistic Theory. Studies in Linguistic Analysis. (1957).
6. Fišer, D., Erjavec, T., Ljubešić, N.: JANES v0.4: Korpus slovenskih spletnih uporabniških vsebin Slovenščina 2.0 4/2, 67–99 (2016).
7. Gantar P., Škrjanec I., Fišer D., Erjavec T.: Slovar tviterščine. Proceedings of the Conference on Language Technologies and Digital Humanities, Ljubljana, Slovenia, 71–76. (2016)
8. Gulordava, K., Baroni, M.: A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics, 67–71. (2011)
9. Ide, N., Véronis, J.: Introduction to the special issue on word sense disambiguation: the state of the art. Computational linguistics 24/1, 2-40. (1998)
10. Kilgarriff, A., et al.: The Sketch Engine: ten years on. In Lexicography 1—30, (2014).
11. Logar Berginc, N., Grčar, M., Erjavec, T., Arhar Holdt, Š., Krek, S.: Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba. Trojina, zavod za uporabno slovenistiko: Fakulteta za družbene vede. Zbirka Sporazumevanje, Ljubljana, Slovenia. (2012)
12. Mikolov, T., Yih, W., Zweig, G.: Linguistic Regularities in Continuous Space Word Representations. HLT-NAACL, 746—751. (2013)
13. Mitchell, T.M.: Machine Learning, McGraw-Hill, Inc. New York, NY, USA. (1997)
14. Mitra, S., et al.: An automatic approach to identify word sense changes in text media across timescales. Natural Language Engineering 21/05, 773–798. (2015)
15. Navigli, R.: Word sense disambiguation: A survey. ACM Computing Surveys (CSUR) 41/2. (2009)

16. Sagi, E., Kaufmann, S., Clark, B.: 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. (2009)
17. Schütze, H.: Automatic word sense discrimination. Computational linguistics 24/1, 97–123. (1998)
18. Spark-Jones, K.: Synonym and Semantic Classification. Edinburgh Information Technology Series. Edinburgh University Press, Edinburgh. (1986)
19. Tahmasebi, N., Risse, T., Dietze, S.: Towards automatic language evolution tracking, a study on word sense tracking. In Joint Workshop on Knowledge Evolution and Ontology Dynamics. (2011)