Evaluation of Computational Models of Semantic Change: An Annotation of Semantic Change based on Usage Similarity

1 Introduction

We see an increasing interest in the automatic detection of semantic change in computational linguistics (Hamilton et al., 2016a; Frermann and Lapata, 2016; Schlechtweg et al., 2017, i.a.). The roots of this interest range from expected performance improvements of practical natural language processing applications to mere theoretical interest in language or cultural change. However, a major obstacle in computational modeling of semantic change is evaluation. This is acknowledged by various authors such as Lau et al. (cf. 2012, p. 597), Cook et al. (cf. 2014, p. 1625) or Frermann and Lapata (cf. 2016, p. 33). The reasons are multiple: For instance, there is no thematically consistent benchmark corpus representing a variety of time intervals and genres (cf. Frermann and Lapata, 2016, p. 33). But, most importantly, there is no standard test set of semantic change for any language. Hence, computational models of semantic change are only superficially evaluated. That this can be harmful is indicated by recent results in the field: Despite their superficial evaluation models of semantic change are applied to large amounts of data. From the models' predictions scholars then draw conclusions and propose statistical laws of semantic change (Hamilton et al., 2016b; Eger and Mehler, 2016) or test independently derived laws (Xu and Kemp, 2015). However, as Dubossarsky et al. (2017) suggest, the proposed models have biases that may explain the putative laws of semantic change.

Drawing conclusions from large amounts of data is appealing, since this is the main advantage of a computational approach to language change: once a sufficiently performing model is established, it may support the historical linguist's job of analyzing language data and finding semantic change, while being much more efficient and able to generalize to give insights into the nature of semantic change. It is thus understandable that scholars apply their models to data as soon as possible. However, as we saw, this can lead to the overlooking of model insufficiencies and hence to false conclusions. This is why we need a more thorough evaluation of models of semantic change before we start to draw empirical conclusions from them. We argue that such an evaluation should rely on a human annotation process rather than artificial simulation of assumed effects or other elegant, but more indirect ways of evaluation. We take further steps into this direction by proposing a structured annotation process of semantic change and some of its subtypes based on usage similarity. We also apply the proposed process on data from the German DTA corpus obtaining the first standard test set of semantic change. The process combines ideas from synchronic research in word sense disambiguation and recent research in metaphoric change. It is language-independent and can easily be transferred, e.g., to English data.

2 Related Work

Previous evaluation procedures of computational models of semantic change include case studies of individual words (Sagi et al., 2009; Jatowt and Duh, 2014; Hamilton et al., 2016a), stand-alone comparison of a few hand-selected words (Wijaya and Yeniterzi, 2011; Hamilton et al., 2016b), comparison of few hand-selected words with semantically stable words (Lau et al., 2012; Cook et al., 2014) or post-hoc evaluation of the predictions of the presented models (Cook and Stevenson, 2010; Kulkarni et al., 2014; Eger and Mehler, 2016). Amongst these, Lau et al. (2012) and Cook et al. (2014) aim at verifying the semantic developments of their targets by a quasi-annotation procedure without reporting inter-annotator agreement or other reliability measures. Gulordava and Baroni (2011), as an exception, conduct an annotation study. However, it is not clear from the description what the annotators judged. It seems that annotators were presented with words and then asked for their intuition without relating this to any data: "Human raters were asked to rank the resulting list according to their intuitions about change in last 40 years" (p. 69).

Many of the described studies focus on contemporary language where rich corpora are available, simply avoiding the huge obstacles presented by the modeling of earlier processes (e.g. Kulkarni et al., 2014). Also, some studies try to evade evaluation problems by simulating the distributional effects they assume to be present with the phenomena they examine, as e.g., Cook and Stevenson (2010) for pejorization and meliorization and Kulkarni et al. (2014) for linguistic shift. Often evaluation is performed on the same canonical examples from standard literature, e.g., the use of *gay* (amongst other words) in Cook and Stevenson (2010), Wijaya and Yeniterzi (2011), Kim et al. (2014), Kulkarni et al. (2014), Jatowt and Duh (2014), Hamilton et al. (2016a) and Hamilton et al. (2016b).

Diverse evaluation methods can be found in the field of word sense induction. Frermann and Lapata (2016), for instance, evaluate their model in a number of ways: (i) they measure how well the word representations of their model for different time slices predict from which time period they are; (ii) they measure how well the new senses their model infers are reflected in Word-

Net as proposed by Mitra et al. (2015); (iii) they evaluate the sense novelty scores of their model against the ranked data from Gulordava and Baroni (2011); and (iv) they apply their model to the SemEval-2015 diachronic text evaluation subtasks (Popescu and Strapparava, 2013). The latter is a set of tasks where the aim is to identify when a piece of text was written, introduced in (Mihalcea and Nastase, 2012). This way of evaluation is also applied in diachronic topic modeling (topics over time) (Wang and Mccallum, 2006; Wijaya and Yeniterzi, 2011). Moreover, Bamman and Crane (2011) exploit aligned translated texts as source of word senses and conduct a very limited annotation study on Latin texts from different time periods.

The only evaluation on data obtained in a structured annotation process is done by Schlechtweg et al. (2017). They create a small test set for metaphoric change, yet achieving statistical significance of certain effects on the test set. Contrary to Gulordava and Baroni (2011) their annotation process requires annotators to judge language data. However, metaphoric change is only a subtype of semantic change which is why their annotation procedure is only partly useful for a general annotation of semantic change. As we will see, we adopt their ideas where they are useful for us.

3 Annotation

We want to build on the annotation procedure adopted in Schlechtweg et al. (2017) who compare pairs of a word's usage contexts for a metaphoric relation. In their study three annotators were asked to compare 20 context pairs of a target word for a metaphoric relation. The context pairs are combined in such a way that they stem from two different time periods so that a high number of annotated metaphoric relations can be interpreted as metaphoric change. For instance, annotators should judge whether *umwälzen* in (1) from 1824 was metaphorically related to *umwälzen* in (2) from 1616 and inversely.

(1) *Kinadon wollte den Staat umwälzen...*
'Kinadon wanted to revolutionize the state...'

(2) *...muß ich mich vmbweltzen / vnd kan keinen schlaff in meine augen bringen*
'...I have to turn around and cannot bring sleep into my eyes.'

Adopting their approach will have the advantage that no knowledge of a theory-laden notion of semantic change has to be presupposed of the annotators. Also, from the annotation of numerous context pairs a fine-grained degree of change can be inferred.

A similar procedure is used in Erk et al. (2009, 2013) for annotation of (synchronic) usage similarity (Usim). Instead of making a binary decision annotators were asked to rate pairs of a word's usage contexts for their degree of semantic similarity on a five-point scale. For instance, annotators should judge how similar the meanings of *run* in (3) and (4) are.

(3) She knows how to run a successful company.

(4) I run a mile every day.

We combine the two approaches by merging Schlechtweg et al.'s idea to combine contexts from different time periods and Erk et al.'s idea to annotate graded word similarity of word uses into an annotation of Diachronic Usage Similarity (DUSim), measuring how strongly uses of a word differ between different time periods. For this, we modify Erk et al.'s guidelines according to our purposes. Basically, people will annotate the USim of pairs of a word's contexts from different time periods, as in (5), (6) and (7) for German *toll*, 'mad' > 'mad, nice'.

(5) *Danckestu also dem Herren deinem Gott / du toll vnd töricht Volck?*
'Hence, you thank the Lord, your God, you mad and foolish folk?'

(6) *Wo bei Tagesanbruch ein so tolles Leben gepulst und gebraust, lagerte jetzt das große Schweigen.*
'Where had been pulsating and racing such a nice life at dawn, was now big silence.'

(7) *Man könnte geradezu toll werden, wenn man dich ansieht.*
'One could become mad looking at you.'

One would expect a combination of contexts (5) from 1603 and (6) from 1924 to receive low usage similarity from annotators with each expressing a different meaning, while a combination of contexts (5) from 1603 and (7) from 1924 should receive high usage similarity with both expressing the same meaning. Low (high) average usage similarity over all context pairs for a word is then interpreted as strong (little) semantic change. We feel justified to make this inference, since this is exactly the way in which historical linguists work when identifying instances of semantic change: by comparing uses of expressions in different time periods and judging whether these uses are different from each other. In our view, hence, semantic change is equal to a change in usage similarity between the old and the new uses.

Word uses are also compared within one time period. Decrease (increase) in average usage similarity is interpreted as meaning innovation (reduction).The items for annotation are taken from Paul (2002)'s diachronic semantic dictionary of German.

References

D. Bamman and G. Crane. 2011. Measuring Historical Word Sense Variation. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*. ACM, New York, NY, USA, JCDL '11, pages 1 – 10.

P. Cook, J. H. Lau, D. McCarthy, and T. Baldwin. 2014. Novel Word-sense Identification. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*. pages 1624 – 1635.

P. Cook and S. Stevenson. 2010. Automatically Identifying Changes in the Semantic Orientation of Words. In N. C. C. Chair, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 1147–1156.

Steffen Eger and Alexander Mehler. 2016. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on word senses and word usages. In *In Proceedings of ACL-09*.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics* 39(3):511–554.

Lea Frermann and Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *TACL* 4:31–45.

K. Gulordava and M. Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of GEMS*.

W. L. Hamilton, J. Leskovec, and D. Jurafsky. 2016a. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In *Emnlp*.

W. L. Hamilton, J. Leskovec, and D. Jurafsky. 2016b. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *CoRR* abs - 1605 - 09096.

Adam Jatowt and Kevin Duh. 2014. A framework for analyzing semantic change of words across time. In *JCDL*.

Y. Kim, Y.-I. Chiu, K. Hanaki, D. Hegde, and S. Petrov. 2014. Temporal Analysis of Language through Neural Language Models. *ArXiv e-prints* .

Peter Koch. 2016. Meaning change and semantic shifts. In Maria Koptjevskaja-Tamm Pänivi Juvonen, editor, *The Lexical Typology of Semantic Shifts*, De Gruyter Mouton.

V. Kulkarni, R. A.-. Rfou, B. Perozzi, and S. Skiena. 2014. Statistically Significant Detection of Linguistic Change. *CoRR* abs - 1411 - 3315.

J. H. Lau, P. Cook, D. McCarthy, D. Newman, and T. Baldwin. 2012. Word Sense Induction for Novel Sense Detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, EACL '12, pages 591 – 601.

R. Mihalcea and V. Nastase. 2012. Word Epoch Disambiguation: Finding How Words Change Over Time. In *Proceedings of the 50th Annual Meeting of ACL*.

S. Mitra, R. Mitra, S. K. Maity, M. Riedl, C. Biemann, P. Goyal, and A. Mukherjee. 2015. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering* 21(5):773 – 798.

H. Paul. 2002. *Deutsches Wörterbuch: Bedeutungsgeschichte und Aufbau unseres Wortschatzes*. Niemeyer, Tübingen, 10 edition.

O. Popescu and C. Strapparava. 2013. Behind the Times: Detecting Epoch Changes using Large Corpora. In *Ijcnlp*.

E. Sagi, S. Kaufmann, and B. Clark. 2009. Semantic Density Analysis: Comparing Word Meaning Across Time and Phonetic Space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*. Association for Computational Linguistics, Stroudsburg, PA, USA, GEMS '09, pages 104 – 111.

Dominik Schlechtweg, Stefanie Eckmann, Enrico Santus, Sabine Schulte im Walde, and Daniel Hole. 2017. German in flux: Detecting metaphoric change via word entropy. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 354–367.

X. Wang and A. Mccallum. 2006. Topics over time: A non-Markov continuous-time model of topical trends. In *In SIGKDD*.

D. T. Wijaya and R. Yeniterzi. 2011. Understanding Semantic Change of Words over Centuries. In *Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web*. ACM, New York, NY, USA, DETECT '11, pages 35–40.

Y. Xu and C. Kemp. 2015. A Computational Evaluation of Two Laws of Semantic Change. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society, CogSci 2015, Pasadena, California, USA, July 22-25, 2015*.