Towards Tracking Semantic Change by Visual Analytics

Christian Rohrdantz[1] **Annette Hautli**[2] **Thomas Mayer**[2]
Miriam Butt[2] **Daniel A. Keim**[1] **Frans Plank**[2]
Department of Computer Science[1] Department of Linguistics[2]
University of Konstanz

Abstract

This paper presents a new approach to detecting and tracking changes in word meaning by visually modeling and representing diachronic development in word contexts. Previous studies have shown that computational models are capable of clustering and disambiguating senses, a more recent trend investigates whether changes in word meaning can be tracked by automatic methods. The aim of our study is to offer a new instrument for investigating the diachronic development of word senses in a way that allows for a better understanding of the nature of semantic change in general. For this purpose we combine techniques from the field of Visual Analytics with unsupervised methods from Natural Language Processing, allowing for an interactive visual exploration of semantic change.

1 Introduction

The problem of determining and inferring the sense of a word on the basis of its context has been the subject of quite a bit of research. Earlier investigations have mainly focused on the disambiguation of word senses from information contained in the context, e.g. Schütze (1998) or on the induction of word senses (Yarowsky, 1995). Only recently, the field has added a diachronic dimension to its investigations and has moved towards the computational detection of sense development over time (Sagi et al., 2009; Cook and Stevenson, 2010), thereby complementing theoretical investigations in historical linguistics with information gained from large corpora. These approaches have concentrated on measuring general changes in the meaning of a word (e.g., narrowing or pejoration), whereas in this paper we deal with cases where words acquire a new sense by extending their contexts to other domains.

For the scope of this investigation we restrict ourselves to cases of semantic change in English even though the methodology is generally language independent. Our choice is on the one hand motivated by the extensive knowledge available on semantic change in English. On the other hand, our choice was driven by the availability of large corpora for English. In particular, we used the New York Times Annotated Corpus.[1] Given the variety and the amount of text available, we are able to track changes from 1987 until 2007 in 1.8 million newspaper articles.

In order to be able to explore our approach in a fruitful manner, we decided to concentrate on words which have acquired a new dimension of use due to the introduction of computing and the internet, e.g., *to browse, to surf, bookmark*. In particular, the Netscape Navigator was introduced in 1994 and our data show that this does indeed correlate with a change in use of these words.

Our approach combines methods from the fields of Information Visualization and Visual Analytics (Thomas and Cook, 2005; Keim et al., 2010) with unsupervised techniques from Natural Language Processing (NLP). This combination provides a novel instrument which allows for tracking the diachronic development of word meaning by visualizing the contexts in which the words occur. Our overall aim is not to replace linguistic analysis in

[1] `http://http://www.ldc.upenn.edu/`

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 305–310,
Portland, Oregon, June 19-24, 2011.

this field with an automatic method, but to guide research by generating new hypotheses about the development of semantic change.

2 Related work

The computational modeling of word senses is based on the assumption that the meaning of a word can be inferred from the words in its immediate context ("context words"). Research in this area mainly focuses on two related tasks: Word Sense Disambiguation (WSD) and Word Sense Induction (WSI). The goal of WSD is to classify occurrences of polysemous words according to manually predefined senses. One popular method for performing such a classification is Latent Semantic Analysis (LSA) (Deerwester et al., 1990), with other methods also suitable for the task (see Navigli (2009) for an extensive survey).

The aim of WSI is to learn word senses from text corpora without having a predefined number of senses. This goal is more difficult to achieve, as it is not clear beforehand how many senses should be extracted and how a sense could be described in an abstract way. Recently, however, Brody and Lapata (2009) have shown that Latent Dirichlet Allocation (LDA) (Blei et al., 2003) can be successfully applied to perform word sense induction from small word contexts.

The original idea of LSA and LDA is to learn "topics" from documents, whereas in our scenario word contexts rather than documents are used, i.e., a small number of words before and after the word under investigation (bag of words). Sagi et al. (2009) have demonstrated that broadening and narrowing of word senses can be tracked over time by applying LSA to small word contexts in diachronic corpora. In addition, we will use LDA, which has proven even more reliable in the course of our investigations.

In general, the aim of our paper is to go beyond the approach of Sagi et al. (2009) and analyze semantic change in more detail. Ideally, a starting point of change is found and the development over time can be tracked, paired with a quantitative comparison of prevailing senses. We therefore suggest to visualize word contexts in order to gain a better understanding of diachronic developments and also generate hypotheses for further investigations.

3 An interactive visualization approach to semantic change

In order to test our approach, we opted for a large corpus with a high temporal resolution. The New York Times Annotated Corpus with 1.8 million newspaper articles from 1987 to 2007 has a rather small time depth of 20 years but provides a time stamp for the exact publication date. Therefore, changes can be tracked on a daily basis.

The data processing involved context extraction, vector space creation, and sense modeling. As Schütze (1998) showed, looking at a context window of 25 words before and after a key word provides enough information in order to disambiguate word senses. Each extracted context is complemented with the time stamp from the corpus. To reduce the dimensionality, all context words were lemmatized and stop words were filtered out.

For the set of all contexts of a key word, a global LDA model was trained using the MALLET toolkit[2] (McCallum, 2002). Each context is assigned to its most probable topic/sense, complemented by a specific point on the time scale according to its time stamp from the corpus. Contexts for which the highest probability was less than 40% were omitted because they could not be assigned to a certain sense unambiguously. The distribution of senses over time was then visualized.

3.1 Visualization

Different visualizations provide multidimensional views on the data and yield a better understanding of the developments. While plotting every word occurrence individually offers the opportunity to detect and inspect outliers, aggregated views on the data are able to provide insights on overall developments.

Figure 1 provides a view where the percentages of word contexts belonging to different senses are plotted over time. For the verbs *to browse* and *to surf* seven senses are learned with LDA. Each sense corresponds to one row and is described by the top five terms identified by LDA. The higher the gray area at a certain x-axis point, the more of the contexts of the corresponding year belong to the specific sense. Each shade of gray represents 10% of the overall data, i.e., three shades of gray mean that between

[2] http://mallet.cs.umass.edu/

Figure 1: Temporal development of different senses concerning the verbs *to browse* (left) and *to surf* (right)

20% and 30% of the contexts can be attributed to that sense. For each year one value has been generated and values between two years are linearly interpolated.

Figure 2 shows the development of contexts over time, with each context plotted individually. The more recent the context, the darker the color.[3] Each axis represents one sense of *to browse*, in each subfigure different combinations of senses are plotted. A random jitter has been introduced to avoid overlaps. Contexts in the middle (not the lower left corner, but the middle of the graph, e.g., see *e* vs. *f*) belong to both senses with at least 40% probability. Senses that share many ambiguous contexts are usually similar. By mousing over a colored dot, its context is shown, allowing for an in depth analysis.

3.2 Case studies

In order to be able to judge the effectiveness of our new approach, we chose key words that are likely candidates for a change in use in the time from 1987 to 2007. That is, we concentrated on terms relating to the relatively recent introduction of the internet. The advantage of these terms is that the cause of change can be located precisely in time.

Figure 1 shows the temporal sense development of the verbs *to browse* and *to surf*, together with the descriptive terms for each sense. Sense *e* for *to browse* and sense *k* for *to surf* pattern quite similarly. Inspecting their contexts reveals that both senses appear with the invention of web browsers, peaking shortly after the introduction of Netscape Navigator (1994). For *to browse*, another broader sense (sense *f*) concerning browsing in both the internet and digital media collections shows a continuous increase over time, dominating in 2007.

The first occurrences assigned to sense *f* in 1987 are "browse data bases", "word-by-word browsing" in databases and "browsing files in the center's library", referring to physical files, namely photographs. We speculate that the sense of browsing physical media might haven given rise to the sense which refers to browsing electronic media, which in turn becomes the dominating sense with the advent of the web.

Figure 2 shows pairwise comparisons of word senses with respect to the contexts they share, i.e., contexts that cannot unambiguously be assigned to one or the other. Each context is represented by one dot colored according to its time stamp. It can be seen that senses *d* (animals that browse) and *e* (browsing the web) share no contexts at all. Senses *d* (animals that browse) and *f* (browsing files) share only few contexts. In turn, senses *e* and *f* share a fair number of contexts, which is to be expected, as they are closely related. Single contexts, each represented by a colored dot, can be inspected via a

[3]The pdf version of this paper contains a bipolar color map.

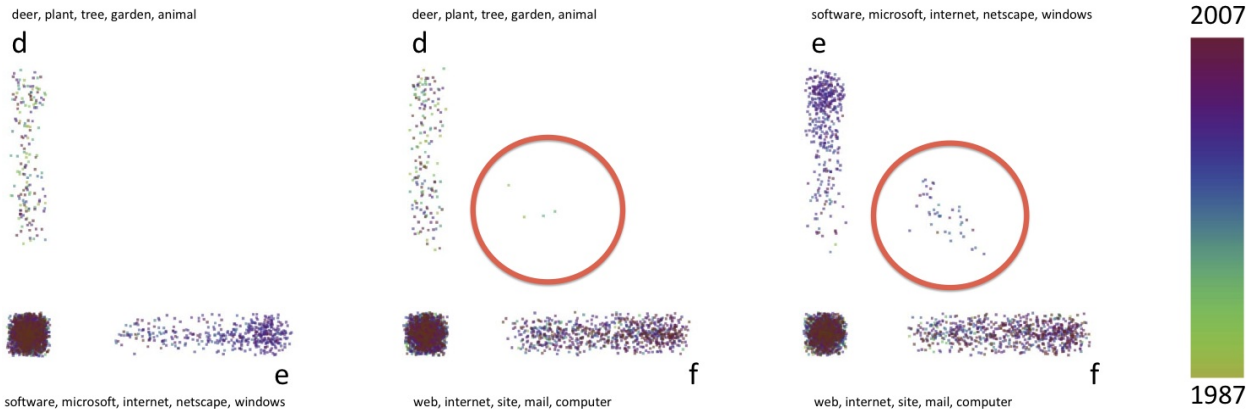

Figure 2: Pairwise comparisons of different senses for the verb "to browse". In each subfigure different combinations of LDA dimensions are mapped on the axes.

	LSA dimensions
1	web 0.40, internet 0.38, software 0.36, microsoft 0.28, windows 0.18
2	microsoft 0.24, software 0.23, windows 0.13, internet 0.13, netscape 0.12
3	microsoft 0.27, store 0.22, shop 0.20, windows 0.19, software 0.16
4	shop 0.32, netscape 0.23, web 0.23, store 0.19, software 0.19
5	book 0.48, netscape 0.26, software 0.17, world 0.13, communication 0.12
6	internet 0.58, shop 0.25, service 0.16, computer 0.13, people 0.11
7	make 0.39, shop 0.34, site 0.16, windows 0.13, art 0.08
...	...
15	find 0.30, people 0.22, year 0.19, **deer** 0.16, day 0.15

Table 1: Descriptive terms for the top LSA dimensions for the contexts of *to browse*. For each dimension the top 5 positively associated terms were extracted, together with their value in the corresponding dimension.

mouse roll over. This allows for an in-depth look at specific data points and a better understanding how the data points relate to a sense.

3.3 LSA vs. LDA

In comparison, Table 1 shows the LSA dimensions learned from the contexts of the verb *to browse*. The top five associated terms for each dimension have been extracted as descriptor. The dimensions are heavily dominated by senses strongly represented in the corpus (e.g., browsing the web). Infrequent senses (e.g., animals that browse) only occur in very low-ranked dimensions and are mixed with other senses (see the bold term *deer* in dimension 15).

4 Evaluation

We compared the findings provided by our visualization with word sense information coming from various resources, namely the 2007 Collins dictionary (COLL), the English WordNet[4] (WN) (Fellbaum, 1998) and the Longman Dictionary (LONG) from 1987. Senses that evolved later than 1987 should not appear in LONG, but should appear in later dictionaries.

However, we are well aware that dictionaries are by no means good gold standards as lexicographers themselves vary greatly when assigning word senses. Nevertheless, this comparison can provide a first indication as to whether the results of our tool is in line with other methods of identifying senses.

In the case of *to browse*, COLL and WordNet suggest the senses "shopping around; not necessarily buying", "feed as in a meadow or pasture" and "browse a computer directory, surf the internet or the world wide web." These senses are also identified in our visualizations, which even additionally differentiate between the senses of "browsing the web" and "browsing a computer directory." A WordNet sense that cannot be detected in the data is the meaning "to eat lightly and try different dishes."

Table 2 shows the results of comparing dictionary word senses (DIC) with the results from our visualization (VIS). What can be seen is that our method is able to track semantic change diachronically and

[4] http://wordnetweb.princeton.edu

	to browse		to surf		messenger		bug		bookmark	
	# of word senses		*# of word senses*		*# of word senses*		*# of word senses*		*# of word senses*	
	DIC	VIS	DIC	VIS	DIC	VIS	DIC	VIS	DIC	VIS
1987 (LONG)	2	3	1	1	1	2	6	3	1	1
1998 (WN)	5	4	3	3	1	3	5	3	1	2
2007 (COLL)	3	4	3	2	1	3	5	3	2	2

Table 2: A comparison of different word senses as given in dictionaries with the visualization results across time

in the majority of cases, the number of our senses correspond to the information coming from the dictionaries. In some cases we are even more accurate in discriminating them. In the case of "messenger", the visualizations suggest another sense related to "instant messaging" that arises with the advent of the AOL instant messenger in 1997. This leads us to the conclusion that our method is appropriate from a historical linguistic point of view.

5 Discussion and conclusions

When dealing with a complex phenomenon such as semantic change, one has to be aware of the limitations of an automatic approach in order to be able to draw the right conclusions from its results. The first results of the case studies presented in this paper show that LDA is useful for distinguishing different word senses on the basis of word contexts and performs better than LSA for this task. Further, it has been demonstrated by exemplary cases that the emergence of a new word sense can be detected by our new methodology

One of the main reasons for an interactive visualization approach is the possibility of being able to detect conspicuous patterns at-a-glance, yet at the same time being able to delve into the details of the data by zooming in on the occurrences of particular words in their contexts. This makes it possible to compensate for one of the major disadvantages of generative and vector space models, namely their functioning as "black boxes" whose results cannot be tracked easily.

The biggest problem in dealing with a corpus-based method of detecting meaning change is the availability of suitable corpora. First, computing semantic information on the basis of contexts requires a large amount of data in order to be able to infer reliable results. Second, the words in the context from which the meanings will be distinguished should be both semantically and orthographically stable over time so that comparisons between different stages in the development of the language can be made. Unfortunately, both requirements are not always met. On the one hand words do change their meaning, after all this is what the present study is all about. However, we assume that the meanings in a certain context window are stable enough to infer reliable results provided it is possible that the forms of the same words in different periods can be linked. This of course limits the applicability of the approach to smaller time ranges due to changes in the phonetic form of words. Moreover, in particular for older periods of the language, different variants for the same word, either due to sound changes or different (or rather no) spelling conventions, abound. For now, we circumvent this problem by testing our tool on corpora where the drawbacks of historical texts are less severe but at the same time interesting developments can be detected to prove our approach correct.

For future research, we want to test our methodology on a broader range of terms, texts and languages and develop novel interactive visualizations to aid investigations in two ways. As a first aim, the user should be allowed to check the validity and quality of the visualizations by experimenting with parameter settings and inspecting their outcome. Second, the user is supposed to gain a better understanding of semantic change by interactively exploring a corpus.

Acknowledgments

This work has partly been funded by the Research Initiative "Computational Analysis of Linguistic Development" at the University of Konstanz and by the German Research Society (DFG) under the grant GK-1042, Explorative Analysis and Visualization of Large Information Spaces, Konstanz. The authors would like to thank Zdravko Monov for his programming support.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Samuel Brody and Mirella Lapata. 2009. Bayesian word sense induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 103–111, Stroudsburg, PA, USA. Association for Computational Linguistics.

Paul Cook and Suzanne Stevenson. 2010. Automatically Identifying Changes in the Semantic Orientation of Words. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 28–34, Valletta, Malta.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Daniel A. Keim, Joern Kohlhammer, Geoffrey Ellis, and Florian Mansmann, editors. 2010. *Mastering The Information Age - Solving Problems with Visual Analytics*. Goslar: Eurographics.

Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):1–69.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 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.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

James J. Thomas and Kristin A. Cook. 2005. *Illuminating the Path The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics (ACL '95)*, pages 189–196, Cambridge, Massachusetts.