Word Epoch Disambiguation:
Finding How Words Change Over Time

Rada Mihalcea
Computer Science and Engineering
University of North Texas
rada@cs.unt.edu

Vivi Nastase
Institute for Computational Linguistics
University of Heidelberg
nastase@cl.uni-heidelberg.de

Abstract

In this paper we introduce the novel task of "word epoch disambiguation," defined as the problem of identifying changes in word usage over time. Through experiments run using word usage examples collected from three major periods of time (1800, 1900, 2000), we show that the task is feasible, and significant differences can be observed between occurrences of words in different periods of time.

1 Introduction

Most current natural language processing works with language as if it were a constant. This however, is not the case. Language is continually changing: we discard or coin new senses for old words; metaphoric and metonymic usages become so engrained that at some point they are considered literal; and we constantly add new words to our vocabulary. The purpose of the current work is to look at language as an evolutionary phenomenon, which we can investigate and analyze and use when working with text collections that span a wide time frame.

Until recently, such task would not have been possible because of the lack of large amounts of non-contemporary data.[1] This has changed thanks to the Google books and Google Ngrams historical projects. They make available in electronic format a large amount of textual data starting from the 17th century, as well as statistics on word usage. We will exploit this data to find differences in word usage across wide periods of time.

The phenomena involved in language change are numerous, and for now we focus on word usage in different time epochs. As an example, the word *gay*, currently most frequently used to refer to a sexual orientation, was in the previous century used to express an emotion. The word *run*, in the past used intransitively, has acquired a transitive sense, common in computational circles where we run processes, programs and such.

The purpose of the current research is to quantify changes in word usage, which can be the effect of various factors: changes in meaning (addition/removal of senses), changes in distribution, change in topics that co-occur more frequently with a given word, changes in word spelling, etc. For now we test whether we can identify the epoch to which a word occurrence belongs. We use two sets of words – one with monosemous words, the other with polysemous ones – to try and separate the effect of topic change over time from the effect of sense change.

We use examples from Google books, split into three epochs: 1800+/-25 years, 1900+/-25, 2000+/-25. We select open-class words that occur frequently in all these epochs, and words that occur frequently only in one of them. We then treat each epoch as a "class," and verify whether we can correctly predict this class for test instances from each epoch for the words in our lists. To test whether word usage frequency or sense variation have an impact on this disambiguation task, we use lists of words that have different frequencies in different epochs as well as different polysemies. As mentioned before, we also compare the performance of monosemous – and thus (sensewise) unchanged through time – and polysemous words, to verify whether we can in fact predict sense change as opposed to contextual variation.

[1] While the Brown corpus does include documents from different years, it is far from the scale and time range of Google books.

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 259–263,
Jeju, Republic of Korea, 8-14 July 2012. ©2012 Association for Computational Linguistics

2 Related Work

The purpose of this paper is to look at words and how they change in time. Previous work that looks at diachronic language change works at a higher language level, and is not specifically concerned with how words themselves change.

The historical data provided by Google has quickly attracted researchers in various fields, and started the new field of *culturomics* (Michel et al., 2011). The purpose of such research is to analyse changes in human culture, as evidenced by the rise and fall in usage of various terms.

Reali and Griffiths (2010) analyse the similarities between language and genetic evolution, with the transmission of frequency distributions over linguistic forms functioning as the mechanism behind the phenomenon of language change.

Blei and Lafferty (2006) and Blei and Lafferty (2007) track changes in scientific topics through a discrete dynamic topic model (dDTM) – both as types of scientific topics at different time points, and as changing word probability distributions within these topics. The "Photography" topic for example has changed dramatically since the beginning of the 20th century, with words related to digital photography appearing recently, and dominating the most current version of the topic.

Wang and McCallum (2006), Wang et al. (2008) develop time-specific topic models, where topics, as patterns of word use, are tracked across a time changing text collection, and address the task of (fine-grained) time stamp prediction.

Wijaya and Yeniterzi (2011) investigate through topic models the change in context of a specific entity over time, based on the Google Ngram corpus. They determine that changes in this context reflect events occurring in the same period of time.

3 Word Epoch Disambiguation

We formulate the task as a disambiguation problem, where we automatically classify the period of time when a word was used, based on its surrounding context. We use a data-driven formulation, and draw examples from word occurrences over three different epochs. For the purpose of this work, we consider an epoch to be a period of 50 years surrounding the beginning of a new century (1800+/-25 years, 1900+/-25, 2000+/-25). The word usage examples are gathered from books, where the publication year of a book is judged to be representative for the time when that word was used. We select words with different characteristics to allow us to investigate whether there is an effect caused by sense change, or the disambiguation performance comes from the change of topics and vocabulary over time.

4 Experimental Setting

Target Words. The choice of target words for our experiments is driven by the phenomena we aim to analyze. Because we want to investigate the behavior of words in different epochs, and verify whether the difference in word behavior comes from changes in sense or changes in wording in the context, we choose a mixture of polysemous words and monosemous words (according to WordNet and manually checked against Webster's dictionary editions from 1828, 1913 and the current Merriam-Webster edition), and also words that are frequent in all epochs, as well as words that are frequent in only one epoch.

According to these criteria, for each open class (nouns, verbs, adjectives, adverbs) we select 50 words, 25 of which have multiple senses, 25 with one sense only. Each of these two sets has a 10-5-5-5 distribution: 10 words that are frequent in all three epochs, and 5 per each epoch such that these words are only frequent in one epoch. To avoid part-of-speech ambiguity we also choose words that are unambiguous from this point of view. This selection process was done based on Google 1gram historical data, used for computing the probability distribution of open-class words for each epoch. [2]

The set of target words consists thus of 200 open class words, uniformly distributed over the 4 parts of speech, uniformly distributed over multiple-sense/unique sense words, and with the frequency based sample as described above. From this initial set of words, we could not identify enough examples in the three epochs considered for 35,[3] which left us with a final set of 165 words.

Data. For each target word in our dataset, we collect the top 100 snippets returned by a search on Google Books for each of the three epochs we consider.

[2] For each open class word we create ranked lists of words, where the ranking score is an adjusted *tfidf* score – the epochs correspond to documents. To choose words frequent only in one epoch, we choose the top words in the list, for words frequent in all epochs we choose the bottom words in this list.

[3] A minimum of 30 total examples was required for a word to be considered in the dataset.

All the extracted snippets are then processed: the text is tokenized and part-of-speech tagged using the Stanford tagger (Toutanova et al., 2003), and contexts that do not include the target word with the specified part-of-speech are removed. The position of the target word is also identified and recorded as an offset along with the example.

For illustration, we show below an example drawn from each epoch for two different words,
dinner:

> 1800: On reaching Mr. Crane's house, dinner was set before us ; but as is usual here in many places on the Sabbath, it was both **dinner** and tea combined into a single meal.
> 1900: The average **dinner** of today consists of relishes; of soup, either a consomme (clear soup) or a thick soup.
> 2000: Preparing **dinner** in a slow cooker is easy and convenient because the meal you're making requires little to no attention while it cooks.

and *surgeon*:

> 1800: The apothecaries must instantly dispense what medicines the **surgeons** require for the use of the regiments.
> 1900: The **surgeon** operates, collects a fee, and sends to the physician one-third or one-half of the fee, this last transaction being unknown to the patient.
> 2000: From a New York plastic surgeon comes all anyone ever wanted to know–and never imagined–about what goes on behind the scenes at the office of one of the world's most prestigious plastic **surgeons**.

Disambiguation Algorithm. The classification algorithm we use is inspired by previous work on data-driven word sense disambiguation. Specifically, we use a system that integrates both local and topical features. The *local features* include: the current word and its part-of-speech; a local context of three words to the left and right of the ambiguous word; the parts-of-speech of the surrounding words; the first noun before and after the target word; the first verb before and after the target word. The *topical features* are determined from the global context and are implemented through class-specific keywords, which are determined as a list of at most five words occurring at least three times in the contexts defining a certain word class (or epoch). This feature set is similar to the one used by (Ng and Lee, 1996). The features are then integrated in a Naive Bayes classifier (Lee and Ng, 2002).

POS	No. words	Avg. no. examples	Baseline	WED
Noun	46	190	42.54%	66.17%
Verb	49	198	42.25%	59.71%
Adjective	26	136	48.60%	60.13%
Adverb	44	213	40.86%	59.61%
AVERAGE	165	190	42.96%	61.55%

Table 1: Overall results for different parts-of-speech.

Evaluation. To evaluate word epoch disambiguation, we calculate the average accuracy obtained through ten-fold cross-validations applied on the data collected for each word. To place results in perspective, we also calculate a simple baseline, which assigns the most frequent class by default.

5 Results and Discussion

Table 1 summarizes the results obtained for the 165 words. Overall, the task appears to be feasible, as absolute improvements of 18.5% are observed. While improvements are obtained for all parts-of-speech, the nouns lead to the highest disambiguation results, with the largest improvement over the baseline, which interestingly aligns with previous observations from work on word sense disambiguation (Mihalcea and Edmonds, 2004; Agirre et al., 2007).

Among the words considered, there are words that experience very large improvements over the baseline, such as "computer" (with an absolute increase over the baseline of 42%) or "install" (41%), which are words that are predominantly used in one of the epochs considered (2000), and are also known to have changed meaning over time. There are also words that experience very small improvements, such as "again" (3%) or "captivate" (7%), which are words that are frequently used in all three epochs. There are even a few words (seven) for which the disambiguation accuracy is below the baseline, such as "oblige" (-1%) or "cruel" (-15%).

To understand to what extent the change in frequency over time has an impact on word epoch disambiguation, in Table 2 we report results for words that have high frequency in all three epochs considered, or in only one epoch at a time. As expected, the words that are used more often in an epoch are also easier to disambiguate.[4] For instance, the

[4] The difference in results does not come from difference in

verb “reassert” has higher frequency in 2000, and it has a disambiguation accuracy of 67.25% compared to a baseline of 34.15%. Instead, the verb “conceal,” which appears with high frequency in all three epochs, has a disambiguation accuracy of 44.70%, which is a relatively small improvement over the baseline of 38.04%.

POS	No. words	Avg. no. examples	Baseline	WED
High frequency in all epochs				
Noun	18	180	42.31%	65.77%
Verb	19	203	43.45%	56.43%
Adjective	7	108	46.27%	57.75%
Adverb	17	214	40.32%	56.41%
AVERAGE	61	188	42.56%	59.33%
High frequency in one epoch				
Noun	28	196	42.68%	66.42%
Verb	30	194	41.50%	61.80%
Adjective	19	146	49.47%	61.02%
Adverb	27	213	41.20%	61.63%
AVERAGE	104	191	43.20%	62.86%

Table 2: Results for words that have high frequency in all epochs, or in one epoch at a time

The second analysis that we perform is concerned with the accuracy observed for polysemous words as compared to monosemous words. Comparative results are reported in Table 3. Monosemous words do not have sense changes over time, so being able to classify them in different epochs relies exclusively on variations in their context over time. Polysemous words’s context change because of both changes in topics/vocabulary over time, and changes in word senses. The fact that we see a difference in accuracy between disambiguation results for monosemous and polysemous words is an indication that word sense change is reflected and can be captured in the context.

To better visualize the improvements obtained with word epoch disambiguation with respect to the baseline, Figure 1 plots the results.

6 Conclusions

In this paper, we introduced the novel task of word epoch disambiguation, which aims to quantify the changes in word usage over time. Using examples collected from three major periods of time, for 165 words, we showed that the word epoch disambiguation algorithm can lead to an overall absolute improvement of 18.5%, as compared to a baseline that picks the most frequent class by default. These results indicate that there are significant differences between occurrences of words in different periods of time. Moreover, additional analyses suggest that changes in usage frequency and word senses contribute to these differences. In future work, we plan to do an in-depth analysis of the features that best characterize the changes in word usage over time, and develop representations that allow us to track sense changes.

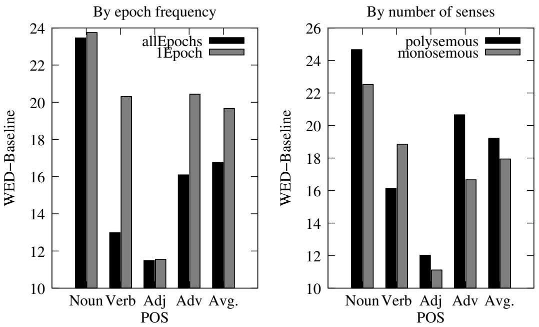

Figure 1: Word epoch disambiguation compared to the baseline, for words that are frequent/not frequent (in a given epoch), and monosemous/polysemous.

POS	No. words	Avg. no. examples	Baseline	WED
Polysemous words				
Noun	24	191	41.89%	66.55%
Verb	25	214	42.71%	58.84%
Adjective	12	136	45.40%	57.42%
Adverb	23	214	39.38%	60.03%
AVERAGE	84	196	41.94%	61.16%
Monosemous words				
Noun	22	188	43.25%	65.77%
Verb	24	181	41.78%	60.63%
Adjective	14	136	51.36%	62.47%
Adverb	21	213	42.49%	59.15%
AVERAGE	81	183	44.02%	61.96%

Table 3: Results for words that are polysemous or monosemous.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation CAREER award #0747340. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

size in the data, as the number of examples extracted for words of high or low frequency is approximately the same.

References

E. Agirre, L. Marquez, and R. Wicentowski, editors. 2007. *Proceedings of the 4th International Workshop on Semantic Evaluations*, Prague, Czech Republic.

D. Blei and J. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*.

D. Blei and J. Lafferty. 2007. A correlated topic model of Science. *The Annals of Applied Science*, 1(1):17–35.

Y.K. Lee and H.T. Ng. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Philadelphia, June.

J.-B. Michel, Y.K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, January.

R. Mihalcea and P. Edmonds, editors. 2004. *Proceedings of SENSEVAL-3, Association for Computational Linguistics Workshop*, Barcelona, Spain.

H.T. Ng and H.B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An examplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL 1996)*, Santa Cruz.

F. Reali and T. Griffiths. 2010. Words as alleles: connecting language evolution with bayesian learners to models of genetic drift. *Proceedings of the Royal Society*, 277(1680):429–436.

K. Toutanova, D. Klein, C. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of Human Language Technology Conference (HLT-NAACL 2003)*, Edmonton, Canada, May.

X. Wang and A. McCallum. 2006. Topics over time: A non-Markov continuous-time model of topical trends. In *Conference on Knowledge Discovery and Data Mining (KDD)*.

C. Wang, D. Blei, and D. Heckerman. 2008. Continuous time dynamic topic models. In *International Conference on Machine Learning (ICML)*.

D. Wijaya and R. Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proc. of the Workshop on Detecting and Exploiting Cultural Diversity on the Social Web (DETECT) 2011*.