A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus

Kristina Gulordava
DISI, University of Trento
Trento, Italy
kgulordava@gmail.com

Marco Baroni
CIMeC, University of Trento
Trento, Italy
marco.baroni@unitn.it

Abstract

This paper presents a novel approach for automatic detection of semantic change of words based on distributional similarity models. We show that the method obtains good results with respect to a reference ranking produced by human raters. The evaluation also analyzes the performance of frequency-based methods, comparing them to the similarity method proposed.

1 Introduction

Recently a large corpus of digitized books was made publicly available by Google (Mitchel et al., 2010). It contains more than 5 millions of books published between the sixteenth century and today. Computational analysis of such representative diachronic data made it possible to trace different cultural trends in the last centuries. Mitchel et al. (2010) exploit the change in word frequency as the main measure for the quantitative investigation of cultural and linguistic phenomena; in this paper, we extend this approach by measuring the semantic similarity of the word occurrences in two different time points using *distributional semantics model* (Turney and Pantel, 2010).

Semantic change, defined as a change of one or more meanings of the word in time (Lehmann, 1992), is of interest to historical linguistics and is related to the natural language processing task of unknown word sense detection (Erk, 2006). Developing automatic methods for identifying changes in word meaning can therefore be useful for both theoretical linguistics and a variety of NLP applications which depend on lexical information.

Some first automatic approaches to the semantic change detection task were recently proposed by Sagi et al. (2009) and Cook and Stevenson (2010). These works focus on specific types of semantic change, i.e., Sagi et al. (2009) aim to identify widening and narrowing of meaning, while Cook and Stevenson (2010) concentrate on amelioration and pejoration cases. Their evaluation of the proposed methods is rather qualitative, concerning just a few examples.

In present work we address the task of automatic detection of the semantic change of words in quantitative way, comparing our novel distributional similarity approach to a relative-frequency-based method. For the evaluation, we used the Google Books Ngram data from the 1960s and 1990s, taking as a reference standard a ranking produced by human raters. We present the results of the method proposed, which highly correlate with the human judgements on a test set, and show the underlying relations with relative frequency.

2 Google Books Ngram corpus

The overall data published online by Google represent a collection of digitized books with over 500 billion words in 7 different languages distributed in n-gram format due to copyright limitations (Mitchel et al., 2010). An n-gram is a sequence of n words divided by space character; for each n-gram it is specified in which year it occurred and how many times.

For our diachronic investigation we used the American English 2-grams corpus (with over 150 millions 2-grams) and extracted two time slices from the 1960s and 1990s time periods. More precisely, we automatically selected 2-grams with year of occurrence between 1960 and 1964 for the 1960s slice,

and between 1995 and 1999 for the 1990s slice, and summed up the number of occurrences of each 2-gram for both corpora. After preprocessing, we obtained well-balanced 60s and 90s corpora containing around 25 and 28 millions of 2-grams, respectively.

We consider the 60s and 90s to be interesting time frames for the evaluation, having in mind that a lot of words underwent semantic change between these decades due to many significant technological and social movements. At the same time, the 60s are close enough so that non-experts should have good intuitions about semantic change between then and now, which, in turn, makes it possible to collect reference judgments from human raters.

3 Measuring semantic change

3.1 Relative frequency

Many previous diachronic studies in corpus linguistics focused on changes of relative frequency of the words to detect different kinds of phenomena (Hilpert and Gries, 2009; Mitchel et al., 2010). Intuitively, such approach can also be applied to detect semantic change, as one would expect that many words that are more popular nowadays with respect to the past (in our case: the 60s) have changed their meaning or gained an alternative one. Semantic change could explain a significant growth of the relative frequency of the word.

Therefore we decided to take as a competing measure for evaluation the logarithmic ratio between frequency of word occurrence in the 60s and frequency of word occurrence in the 90s[1].

3.2 Distributional similarity

In the distributional semantics approach (see for example Turney and Pantel, 2010), the similarity between words can be quantified by how frequently they appear within the same context in large corpora. These distributional properties of the words are described by a vector space model where each word is associated with its context vector. The way a context is defined can vary in different applications. The one we use here is the most common approach which considers contexts of a word as a set of all other words with which it co-occurs. In our case we decided to use 2-grams, that is, only words that occur right next to the given word are considered as part of its context. The window of length 2 was chosen for practical reasons given the huge size or the Google Ngram corpus, but it has been shown to produce good results in previous studies (e.g. Bullinaria and Levy, 2007). The words and their context vectors create a so called co-occurrence matrix, where row elements are target words and column elements are context terms.

The scores of the constructed co-occurrence matrix are given by local mutual information (LMI) scores (Evert, 2008) computed on the frequency counts of corresponding 2-grams[2]. If words w_1 and w_2 occurred $C(w_1, w_2)$ times together and $C(w_1)$ and $C(w_2)$ times overall in corpus then local mutual information score is defined as follows:

$$LMI = C(w_1, w_2) \cdot \log_2 \frac{C(w_1, w_2)N}{C(w_1)C(w_2)},$$

where N is the overall number of 2-gram in the corpus.

Given the words w_1, w_2 their distributional similarity is then measured as the cosine product of their context vectors $\mathbf{v}_1, \mathbf{v}_2$: $sim(w_1, w_2) = \cos(\mathbf{v}_1, \mathbf{v}_2)$.

We apply this model to measure similarity of a word occurrences in two corpora of different time periods in the following way. The set of context elements is fixed and remains the same for both corpora; for each corpus, a context vector for a word is extracted independently, using counts in this corpus as discussed above. In this way, each word will have a 60s vector and a 90s vector, with the same dimensions (context elements), but different co-occurrence counts. The vectors can be compared by computing the cosine of their angle. Since the context vectors are computed in the same vector space, the procedure is completely equivalent to calculating similarity between two different words in the same corpora; the context vectors can be considered as belonging to one co-occurrence matrix and corresponding to two different row elements $word\_60s$ and $word\_90s$.

[1] The logarithmic ratio helps intuition (terms more popular in the 60s get negative scores, terms more popular in the 90s have similarly scaled positive scores), but omitting the logarithmic transform produced similar results in evaluation.

[2] LMI proved to be a good measure for different semantic tasks, see for example the work of Baroni and Lenci, 2010.

group	examples	sim	freq
more frequent in 90s	users	0.29	-0.94
	sleep	0.23	-0.32
	disease	0.87	-0.3
	card	0.17	-0.1
more frequent in 60s	dealers	0.16	0.04
	coach	0.25	0.12
	energy	0.79	0.14
	cent	0.99	1.13

Table 1: Examples illustrating word selection with similarity (sim) and log-frequency (freq) metric values.

We use the described procedure to measure semantic change of a word in two corpora of interest, and hence between two time periods. High similarity value (close to 1) would suggest that a word has not undergone semantic change, while obtaining low similarity (close to 0) should indicate a noticeable change in the meaning and the use of the word.

4 Experiments

4.1 Distributional space construction

To be able to compute distributional similarity for the words in the 60s and 90s corpora, we randomly chose 250,000 mid-frequency words as the context elements of the vector space. We calculated 60s-to-90s similarity values for a list of 10,000 randomly picked mid-frequency words. Among these words, 48.4% had very high similarity values (> 0.8), 50% average similarity (from 0.2 to 0.8) and only 1.6% had very low similarity (< 0.2). According to our prediction, this last group of words would be the ones that underwent semantic change.

To test such hypothesis in a quantitative way some reference standard must be available. Since for our task there was no appropriate database containing words classified for semantic change, we decided to create a reference categorization using human judgements.

4.2 Human evaluation

From the list of 10,000 words we chose 100 as a representative random subset containing words with different similarities from the whole scale from 0 to 1 and taken from different frequency range, i.e., words that became more frequent in 90s (60%) and words that became less frequent (40%) (see Table 1 for examples). Human raters were asked to rank the resulting list according to their intuitions about change in last 40 years on a 4-point scale (0: no change; 1: almost no change; 2: somewhat change; 3: changed significantly). We took the average of judgments as the reference value with which distributional similarity scores were compared. For the 5 participants, the inter-rater agreement, computed as an average of pair-wise Pearson correlations, was 0.51 ($p < 0.01$). It shows that the collected judgements were highly correlated and the average judgement can be considered an enough reliable reference for semantic change measurements evaluation.

	sim-HR	freq-HR	sim-freq
all words	**0.386****	0.301**	0.380**
frequent in 90s	0.445**	0.184	0.278*
frequent in 60s	0.163	0.310	0.406*

Table 2: Correlation between similarity (sim), frequency (freq) and human ranking (HR) values for all words, words more frequent in 60s and more frequent in 90s. Values statistically significant for $p = 0.01(0.05)$ in one-sample t-test are marked with $^{**}(^{*})$.

5 Results and discussion

To assess the performance of our similarity-based measure, we computed the correlations between the values it produced for our list of words and the average human judgements (Table 2). The Pearson correlation value obtained was equal to 0.38, which is reasonably high given 0.51 inter-rater agreement. The frequency measure had a lower correlation (0.3), though close to the similarity measure performance. Yet, the correlation of 0.38 between the two measures in question suggests that, even if they perform similarly, their predictions could be quite different.

In fact, if we consider separately two groups of words: the ones whose frequency increased in the 90s (*log-freq* < 0), that is, the ones that are more popular nowadays, and those whose frequency instead decreased in the 90s (*log-freq* > 0), that is, the ones that were more popular in the 60s, we can make some interesting observations (see Table 2). Remarkably, similarity performs better for the words that are popular nowadays while the frequency-based measure performs better for the words that

were popular in the 60s.

We can see the origin of this peculiar asymmetry in behavior of similarity and frequency measures in the following phenomenon. As we already mentioned, if a word became popular, the reason can be a new sense it acquired (a lot of technological terms are of this kind: *'disk'*, *'address'*, etc). The change in such words, that are characterized by a significant growth in frequency (*log-freq* $\ll 0$), is detected by the human judges, as well as by the similarity measure. However, other cases such as *'spine'*, *'smoking'* are also characterized by a significant growth in frequency, but no semantic change was reported by raters (nor by the similarity measure). If word frequency instead decreases, intuitively, a change in word meaning is less probable. These intuitions together can explain the behavior of the frequency measure: for the test set as a whole its performance is quite high, as it captures this asymmetrical distribution of words that change meanings, despite its failure to reliably indicate semantic change for independent words. A strong evidence for this interpretation is also that, if the frequency measure is made symmetric, that is, equal for the words that decreased and the ones that increased in frequency, it dramatically drops in performance, showing a correlation of just 0.04 with human ranking.

Some interesting observation regarding the performance of the similarity measure can be made after accurate investigation of 'false-positive' examples — the ones that have low similarity but were ranked as 'not changed' by raters — like *'sleep'* and *'parent'*. It is enough to have a look at their highest weighted co-occurrences to admit that the context of their usage has indeed changed (Table 3). These examples show the difference between the phenomenon of semantic change in linguistics and the case of context change. It is well known that the different contexts that distributional semantics catches do not always directly refer to what linguists would consider distinct senses (Reisinger and Mooney, 2010). Most people would agree that the word *'parent'* has the same meaning now as it had 40 years before, still the social context in which it is used has evidently changed, reflected by the more frequent *'single parent family(ies)'* collocate found in the 90s. The same is true for *'sleep'*, whose usage context did not change radically, but might have a more prominent negative orientation.

	'parent'	'sleep'
60s	p. company 2643	deep s. 3803
	p. education 1905	s. well 1403
	p. corporation 1617	cannot s. 1124
	p. material 1337	long s. 1102
	p. body 1082	sound s. 1101
	p. compound 818	dreamless s. 844
	common p. 816	much s. 770
90s	p. families 17710	REM s. 20150
	single p. 10724	s. apnea 14768
	p. company 8367	deep s. 8482
	p. education 5884	s. disorders 8427
	p. training 5847	s. deprivation 6108
	p. involvement 5591	s. disturbances 5973
	p. family 5042	s. disturbance 5251

Table 3: Examples of the top weighted 2-grams containing 'sleep' and 'parent'.

The distributional similarity measure captures therefore two kinds of phenomena: the semantic change in its linguistic definition, that is, change of meaning or acquiring a new sense (e.g., *'virus', 'virtual'*), but also the change in the main context in which the word is used. The latter, in turn, can be an important preliminary evidence of the onset of meaning change in its traditional sense, according to recent studies on language change (Traugott and Dasher, 2002). Moreover, context changes have cultural and social origins, and therefore the similarity measure can also be used for collecting evidence of interest to the humanities and social sciences.

6 Conclusions

In this paper we introduced and evaluated a novel automatic approach for measuring semantic change with a distributional similarity model. The similarity-based measure produces good results, obtaining high correlation with human judgements on test data. The study also suggests that the method can be suitable to detect both "proper" semantic change of words, and cases of major diachronic context change. Therefore, it can be useful for historical linguistic studies as well as for NLP tasks such as novel sense detection. Some interesting phenomena related to changes in relative frequency were also discovered, and will be the object of further investigations.

References

Marco Baroni, Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721. MIT Press, Cambridge, MA, USA.

John A. Bullinaria, Joseph P. Levy. 2007. Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study. *Behavior Research Methods*, 39: 510-526.

Paul Cook, Suzanne Stevenson. 2010. Automatically identifying changes in the semantic orientation of words. *Proceedings of the 7th International Conference on Language Resources and Evaluation.* Valletta, Malta: 28–34.

Katrin Erk. 2006. Unknown word sense detection as outlier detection. *Proceedings of the Human Language Technology of the North American Chapter of the ACL.* New York, USA: 128–135.

Stefan Evert. 2008. Corpora and collocations. In A. Ldeling and M. Kyt (eds.), *Corpus Linguistics. An International Handbook*, article 58. Mouton de Gruyter, Berlin.

Hilpert Martin, Stefan Th. Gries. 2009. Assessing frequency changes in multi-stage diachronic corpora: applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 34(4): 385-40.

Winfred P. Lehmann. 1992. *Historical linguistics: an introduction.* (3. ed.) Routledge & Kegan Paul, London.

Jean-Baptiste Michel*, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden*. 2010. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* (Published online ahead of print: 12/16/2010).

Joseph Reisinger, Raymond Mooney. 2010. A Mixture Model with Sharing for Lexical Semantics. *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* MIT, Massachusetts, USA: 1173–1182.

Eyal Sagi, Stefan Kaufmann, Brady Clark. 2009. Semantic Density Analysis: Comparing Word Meaning across Time and Phonetic Space. *Proceedings of the EACL 2009 Workshop on GEMS: Geometrical Models of Natural Language Semantics.* Athens, Greece: 104–111.

Elizabeth C. Traugott, Richard B. Dasher. 2002. *Regularity in Semantic Change.* Cambridge University Press.

Peter Turney, Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research (JAIR)*, 37(1):141-188. AI Access Foundation.