Distributed Distributional Similarities of Google Books over the Centuries

Martin Riedl, Richard Steuer and Chris Biemann

FG Language Technology
Computer Science Department
Technische Universität Darmstadt
Hochschulstrasse 10
D-64289 Darmstadt, Germany
{riedl,biem}@cs.tu-darmstadt.de, ri.st@freenet.de

Abstract

This paper introduces a distributional thesaurus and sense clusters computed on the complete Google Syntactic N-grams, which is extracted from Google Books, a very large corpus of digitized books published between 1520 and 2008. We show that a thesaurus computed on such a large text basis leads to much better results than using smaller corpora like Wikipedia. We also provide distributional thesauri for equal-sized time slices of the corpus. While distributional thesauri can be used as lexical resources in NLP tasks, comparing word similarities over time can unveil sense change of terms across different decades or centuries, and can serve as a resource for diachronic lexicography. Thesauri and clusters are available for download.

Keywords: Distributional Thesaurus, Semantics, Large-Scale Distributional Methods, Word Similarity, Lexical Resources

1. Motivation

With the availability of large text data from the web or from content providers, and the surge in parallel computation, processing huge amounts of data has more and more become feasible. More affordable storage as well as the introduction of paradigms like MapReduce (Dean and Ghemawat, 2004) allows us to apply big-data techniques on natural language data.

Here, we introduce a distributional thesaurus (DT) and word sense clusters computed on the Google Syntactic N-grams (Goldberg and Orwant, 2013). Distributional similarities have been demonstrated to increase in quality for increased corpus size (Riedl and Biemann, 2013), which in turn leads to improvements in NLP applications (Miller et al., 2012; Szarvas et al., 2013). The resource described here, which we distribute[1] freely under a permissive license, is accompanied by an API and a demonstrator, and contains distributional similarities for a vocabulary of millions of words. Apart from a DT computed on the entirety of Google Books Syntactic N-grams, we also provide DTs restricted to certain time spans, which give rise to diachronic studies on sense change (Mitra et al., 2014).

2. Methodology

In Biemann and Riedl (2013) we have introduced the JoBimText[2] framework to compute distributional similarities between terms using an efficient and effective approach. Our method, implemented with Apache Hadoop[3] and Apache Pig[4], does not only scale to very large amounts of data but also outperforms standard similarity measures (Lin, 1997; Curran, 2004) when using large data (Riedl and Biemann, 2013).

In our approach we first extract terms and their context features, which could be e.g. the neighboring words or dependency parses. We then calculate the frequencies of the terms, the context and the terms with their context. After this step, we remove all context features that occur with more than w words, as these context features are too general and do not contribute to word similarity. Then we compute the Lexicographer's mutual information (LMI, (Evert, 2005)): $LMI(term, feature) = p(term, feature)\log_2(\frac{p(term,feature)}{p(term)p(feature)})$. After that step we only keep the top-ranked p features of each term and count the number of context features two terms share, without considering word-context counts or significance scores. This results in a distributional thesaurus (DT), where all sufficiently frequent words in the vocabulary have an entry that consists of a ranked list of similar words..

We also provide sense clusters on DT entries using the Chinese Whispers graph clustering algorithm (Biemann, 2010). This clustering algorithm has the advantage that the number of clusters is detected automatically – thus it is not forced to yield several senses for terms that have only one meaning. As previously noted in Biemann (2010), sense clusters rather correspond to different usages of words than different referents in the real world – e.g. the body part sense of *hip* is frequently split into a usage related to clothing (dressing the hip) and a usage related to surgery (hip replacement). Figure 1 illustrates the sense clustering on *bar* as a noun (tag:NN) for the very large DT as described below.

3. Google Books DT

We processed dependency parses extracted from Google Books (Goldberg and Orwant, 2013). This dependency

[1] sf.net/p/jobimtext/wiki/LREC2014_Google_DT/

[2] ASL 2.0, sf.net/projects/jobimtext/

[3] http://hadoop.apache.org/

[4] https://pig.apache.org/

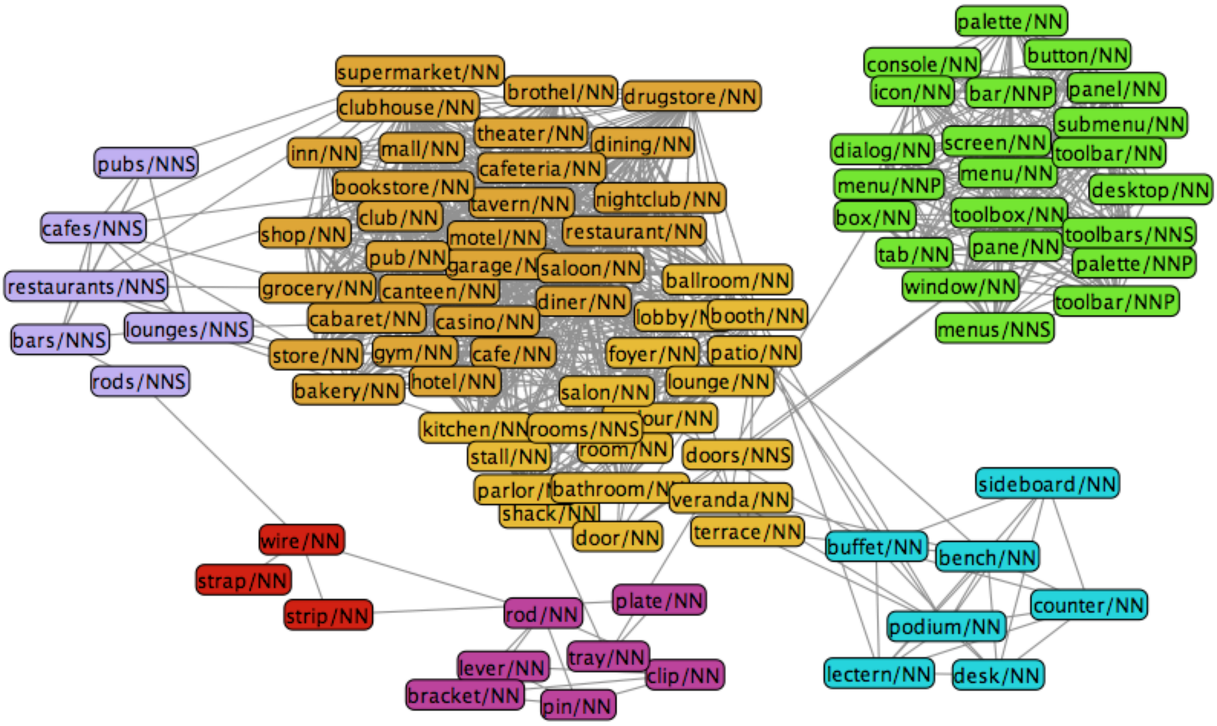
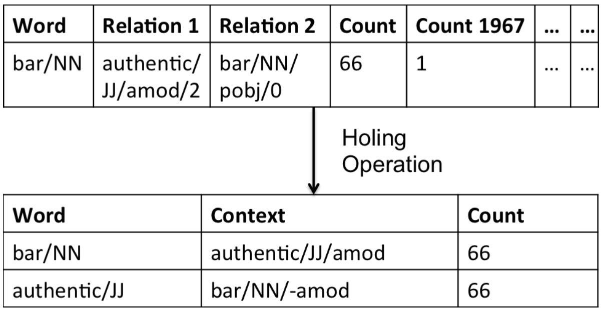


Figure 1: Sense clustering for ”bar/NN” in Google Books DT. Different senses and usages become apparent: the location sense in its usages pub/restaurant, salesplace and room type, the desk/board sense of bar, the building material sense and the GUI toolbar sense.

parse fragment corpus was aggregated over 17.6 billion sentences, collected from books in the time period of 1520 to 2008. For the generation of our DT, we use the top 1000 ranked features per term ($p = 1000$), cf. Section 2. From the format as provided in this corpus, it is straightforward to produce pairs of terms and context features (cf. Sect. 2.): For each dependency parse tree fragment, we perform a holing operation (Biemann and Riedl, 2013), which yields, for each term, a pair of term and the remainder of the dependency tree fragment. Figure 2 shows an example for a syntactic bigram[5].



Word	Relation 1	Relation 2	Count	Count 1967	...	...
bar/NN	authentic/ JJ/amod/2	bar/NN/ pobj/0	66	1	...	...

Word	Context	Count
bar/NN	authentic/JJ/amod	66
authentic/JJ	bar/NN/-amod	66

Figure 2: Holing operation on a syntactic bigram from Goldberg and Orwant (2013) to extract context features that characterize terms

3.1. Very Large DT

With our scalable method, we are able to process the entire data and compute a distributional thesaurus in about one day on a medium-sized Hadoop cluster. The resulting DT is of high quality and showed much better results in comparison to two other thesauri computed on 120 million sentences of news data and a recent dump of Wikipedia of 35 million sentences (see Table 1).

	Corpus	P@1	Path@5	Path@10
frequent nouns	Newspaper	0.709	0.3277	0.2906
	Wikipedia	0.703	0.3365	0.2968
	Google Books	0.764	0.3712	0.3217
infrequent nouns	Newspaper	0.516	0.2577	0.2269
	Wikipedia	0.514	0.2565	0.2265
	Google Books	0.641	0.2989	0.2565

Table 1: Comparing DT quality for different corpora. Cf. (Riedl and Biemann, 2013)

The evaluation is performed for the same 1000 frequent and 1000 infrequent nouns as used by Weeds et al. (2004). We evaluate the thesaurus against a combination of manually created thesauri following Curran (2004) and Riedl and Biemann (2013). The P@1 (precision at 1) measure checks whether the most similar term of the DT entry for a target term is contained in the gold standard thesaurus. Additionally, we evaluated the thesauri against WordNet using the WordNet::path measure (Pedersen et al., 2004), which ist the inverse of the shortest path between WordNet (Miller, 1995) synsets containing the two terms. While the path measures are well suited for relative comparison and are easy to implement due to the availability of WordNet data and APIs, the absolute scores are somewhat hard to interpret. Nevertheless, we can observe a large improvement

[5] These bigrams are called *arcs* in the Google Books corpus.

across all measures when comparing the Google Books DT with the two others. Notably when looking at P@1, in more than 3/4 of cases for frequent nouns and in almost 2/3 cases for infrequent nouns, the most similar term per target is found in a manually compiled thesaurus. Regarding Path@10, we see that the 10 most similar terms are on average 2 (frequent) resp. 3 hops (infrequent) away in WordNet. Since lexical resources are always incomplete, these estimations should rather be understood as lower bounds on DT quality.

3.2. Time slices

Additionally, the Google Books Syntactic N-grams also contain counts of dependency tree fragments for each year in the period of 1520 to 2008, which we can utilize for defining DTs for specific time slices. For this paper, we chose the time slices to contain about equal volume[6], as shown in Table 2. Since an increasing amount of books has been published over time, the length of the time span decreases, as we get closer to the present day. While we chose this setup for the purpose of showing some diachronic analyses of sense change, it would be straightforward to compute DTs on different time slices, especially for recent ones where plenty of data is available.

From	To	Token-dependency relation sum	Percentage
1520	1908	22,524,932,140	(13.17%)
1909	1953	22,161,642,430	(12.95%)
1954	1972	21,684,032,743	(12.68%)
1973	1986	22,548,838,767	(13.18%)
1987	1995	20,840,577,921	(12.18%)
1996	2001	20,929,306,474	(12.23%)
2002	2005	21,657,680,778	(12.66%)
2006	2008	18,725,389,920	(10.95%)

Table 2: "Token" (we sum the counts of terms with all available dependency relations) with different dependency subtrees for a single token for the different time slices used in this work.

Furthermore, we can observe a change of the vocabulary over the time. Table 3 shows that for most cases, the number of terms grows larger in each century, as most of the terms are also used in subsequent years. Losses in vocabulary are partially caused by transcription errors from the optical character recognition (OCR) process. E.g. the long "s", used until the mid of the 19th century is often recognized as "f", which changes "absolute" to "abfolute".

The different time slices can be used to analyze the change of the vocabulary, as well as to analyze the change of the meaning of terms across different centuries. During the 15th to 19th century, the term *bar* (shown in Table 4) occurs mostly with the meaning of *lattice bar* and is therefore similar to the terms like *rod* or *wire*. Additionally, we see that the term *bar* has also the meaning of *tribunal* and *court*, as still present in today's "to pass the *bar*" when passing the lawyer's exam. These meanings vanish in the DTs starting from 1954, at least in the most similar terms as shown here. Whereas between 1954 and 1986 we still observe the term *rod* within the top similar words, the *toolbar* sense, as used in computer GUIs becomes increasingly popular. Furthermore, the similar terms to *bar* are dominated by the sense of *bar* as in *pub, restaurant* starting from 1954. Regarding the ranking of the terms we can also observe a trend: starting from the DT covering the years 1954 to 1972, the term *tavern* falls out of use, whereas the term *pub* receives more popularity in the recent past.

Figure 1 illustrates the use of sense clusters for the term *bar* for the complete corpus. Here we directly observe that the *menu* is not related to *pub* but to *toolbar* and thus relates to elements used in software interfaces: while *menus* are found in bars and pubs and hence co-occur with these words, these menus are not similar (a.k.a. second order co-occurrence) to bars and pubs. Comparing the clusters for the time span of 1520 to 1908 (see left side of Figure 3), we observe several senses that are not detected in the time span between 1996 to 2001 (see right side of Figure 3): In this time span, the term *bar* mostly appears in the location sense and in the GUI toolbar sense.

This exemplifies that sense induction for time spans can unveil changes in the sense distribution, which is covered in more depth in (Mitra et al., 2014).

4. Conclusion

We have described automatically computed distributional thesauri (DTs) that were computed on the very large Google Books Syntactic N-gram corpus. Scaling DT computation to corpora with hundreds of billions of words does not only lead to a broad vocabulary coverage, but also to a lexical resource of very high quality. To our knowledge, this constitutes the largest freely available distributional thesaurus available today. The large DT and the DTs for time slices as laid out above have been made available for download[7], along with the pipeline to produce them, under a permissive license. While the large DT is primarily targeted for the use in NLP applications, such as word sense disambiguation, information retrieval or summarization, the time-sliced DTs along with their sense clusters serve as a firm basis for conducting studies on diachronic sense change in more linguistically-oriented projects.

5. Acknowledgments

This work has been supported by the Hessian research excellence program "Landes-Offensive zur Entwicklung Wissenschaftlich-Ökonomischer Exzellenz" (LOEWE) as part of the research center "Digital Humanities" and by IBM under a Shared University Research Grant.

6. References

Chris Biemann and Martin Riedl. 2013. Text: Now in 2D! a framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, 1(1):55–95.

[6] The corpus does not contain counts of single words . We use the so-called *nodes* files that contain information about how often a word occurs with different dependency relations. To compute the "count" of a term, we sum up all counts for all its dependency relations per time span.

[7] sf.net/p/jobimtext/wiki/LREC2014_Google_DT/, http://www.lt.informatik.tu-darmstadt.de/de/data/distributional-thesauri/

	1520 - 1908	1909 - 1953	1954 - 1972	1973 - 1986	1987 - 1995	1996 - 2001	2002 - 2005	2006 - 2008
1520 - 1908	8.872.808	7.390.052	7.259.803	7.014.053	6.598.621	6.562.215	6.678.304	6.911.093
1909 - 1953		10.471.576	8.970.991	8.687.845	8.175.730	8.027.127	8.030.927	8.000.229
1954 - 1972			11.955.022	10.275.843	9.691.779	9.424.066	9.265.873	8.940.478
1973 - 1986				12.944.323	10.905.427	10.543.896	10.260.647	9.693.174
1987 - 1995					12.883.775	10.923.698	10.569.848	9.861.095
1996 - 2001						12.981.690	10.922.994	10.135.419
2002 - 2005							12.983.086	10.354.845
2006 - 2008								12.330.560

Table 3: Vocabulary overlap (types) for the different time slices.

1520-1908	1909 - 1953	1954 - 1972	1973 - 1986	1987 - 1995	1996 - 2001	2002 - 2006	2006- 2008
bar/NN	bar/NN	bar/NN	bar/NN	bar/NN	bar/NN	bar/NN	bar/NN
bar/NNP	bars/NNS	bars/NNS	bars/NNS	restaurant/NN	bars/NNS	bars/NNS	bars/NNS
bars/NNS	bar/NNP	bar/NNP	restaurant/NN	bars/NNS	restaurant/NN	restaurant/NN	restaurant/NN
rod/NN	rod/NN	rod/NN	lounge/NN	cafe/NN	cafe/NN	cafe/NN	**pub/NN**
wire/NN	wire/NN	restaurant/NN	cafe/NN	lounge/NN	lounge/NN	**pub/NN**	cafe/NN
beam/NN	beam/NN	cafe/NN	bar/NNP	bar/NNP	bar/NNP	bar/NNP	bar/NNP
plate/NN	plate/NN	lounge/NN	**tavern/NN**	**pub/NN**	**pub/NN**	toolbar/NN	counter/NN
lever/NN	bracket/NN	**tavern/NN**	**pub/NN**	**tavern/NN**	**tavern/NN**	lounge/NN	lounge/NN
obstacle/NN	counter/NN	counter/NN	saloon/NN	shop/NN	toolbar/NN	**tavern/NN**	cafeteria/NN
tribunal/NN	lever/NN	saloon/NN	cafeteria/NN	cafeteria/NN	saloon/NN	club/NN	toolbar/NN
court/NN	girder/NN	**pub/NN**	shop/NN	saloon/NN	shop/NN	shop/NN	shop/NN
bedside/NN	rods/NNS	desk/NN	counter/NN	nightclub/NN	nightclub/NN	saloon/NN	club/NN
table/NN	rail/NN	shop/NN	desk/NN	counter/NN	cafeteria/NN	counter/NN	menu/NN
magnet/NN	bolt/NN	wire/NN	nightclub/NN	club/NN	club/NN	menu/NN	**tavern/NN**
polls/NNS	pin/NN	plate/NN	rod/NN	restaurants/NNS	counter/NN	cafeteria/NN	desk/NN
impediment/NN	frame/NN	cafeteria/NN	booth/NN	pool/NN	menu/NN	nightclub/NN	nightclub/NN
desk/NN	desk/NN	hotel/NN	parlor/NN	salon/NN	palette/NN	palette/NN	hotel/NN
bolt/NN	shaft/NN	buffet/NN	hotel/NN	desk/NN	hotel/NN	pane/NN	gym/NN
needle/NN	strip/NN	rods/NNS	club/NN	buffet/NN	gym/NN	desk/NN	diner/NN
strip/NN	strut/NN	lever/NN	salon/NN	toolbar/NN	pool/NN	hotel/NN	palette/NN

Table 4: Similar terms for the word ”bar” in different time slices

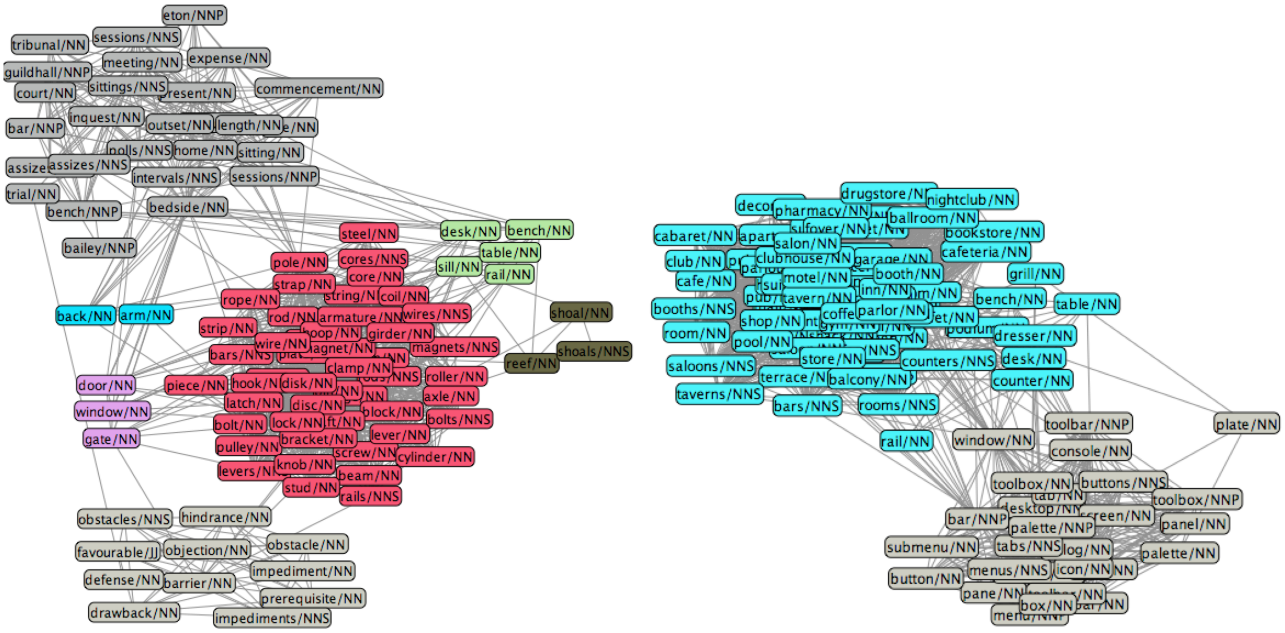


Figure 3: Sense clustering for ”bar/NN” in the Google Books DT from 1520 - 1908 (left) and 1996 - 2002 (right). While some senses fall out of use, the GUI toolbar sense is gaining popularity.

Chris Biemann. 2010. Co-occurrence cluster features for lexical substitutions in context. In *Proceedings of TextGraphs-5*, pages 55–59, Uppsala, Sweden.

James R. Curran. 2004. *From Distributional to Semantic Similarity*. Ph.D. thesis, University of Edinburgh.

Jeffrey Dean and Sanjay Ghemawat. 2004. MapReduce: Simplified Data Processing on Large Clusters. In *Proceedings of Operating Systems, Desing & Implementation (OSDI) '04*, pages 137–150, San Francisco, CA, USA.

Stefan Evert. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations.* Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic N-grams over time from a very large corpus of English books. In *Second Joint Conference on Lexical and*

*Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 241–247, Atlanta, Georgia, USA.

Dekang Lin. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, pages 64–71, Madrid, Spain.

Tristan Miller, Chris Biemann, Torsten Zesch, and Iryna Gurevych. 2012. Using distributional similarity for lexical expansion in knowledge-based word sense disambiguation. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING '12)*, pages 1781–1796, Mumbai, India.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38:39–41.

Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. That?s sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL-2014)*, Baltimore, MD, USA.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet::similarity - measuring the relatedness of concepts. In *Proceedings of the 2004 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies HLT-NAACL 2004: Demonstration Papers*, pages 38–41, Boston, Massachusetts, USA.

Martin Riedl and Chris Biemann. 2013. Scaling to large^3 data: An efficient and effective method to compute distributional thesauri. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*, Seattle, WA, USA.

György Szarvas, Chris Biemann, and Iryna Gurevych. 2013. Supervised all-words lexical substitution using delexicalized features. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 1131–1141, Atlanta, GA, USA.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, pages 1015–1021, Geneva, Switzerland.