Combining distributional semantics and structured data to study lexical change

Astrid van Aggelen, Laura Hollink, and Jacco van Ossenbruggen

Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

Abstract. Statistical Natural Language Processing (NLP) techniques allow to quantify lexical semantic change using large text corpora. Word-level results of these methods can be hard to analyse in the context of sets of semantically or linguistically related words. On the other hand, structured knowledge sources represent such relationships explicitly, but ignore the problem of semantic change. We aim to address these limitations by combining the statistical and symbolic approach: we enrich WordNet, a structured lexical database, with quantitative lexical change scores provided by HistWords, a dataset produced by distributional NLP methods. We publish the result as Linked Open Data and demonstrate how queries on the combined dataset can provide new insights.

Keywords: lexical semantics, NLP, Knowledge bases, Linked Open Data

1 Introduction

How words have been used in discourse over time, have adopted new senses or changed their meaning is studied in the humanities and social sciences (e.g., [1–3]) and information sciences (e.g., [4, 6]). We make a case for interlinking structured knowledge bases with the outcomes of Natural Language Processing (NLP) methods for the purpose of studying language change over time.

Semantic change in words is increasingly modelled using distributional NLP methods (word embeddings). These techniques represent the meaning of a word in terms of its tendency to co-occur with other words in the lexicon, as observed in large text corpora. Since this results in vectors, cosine distances can be used to quantify the correspondence between two such representations. When vectors are assembled for the lexicon in separate time spans, the notion of distance can be applied to find a word's nearest neighbours within a time frame, or to calculate the degree of change a word underwent from one time interval to the next.

However, word embeddings alone are not sufficient to gain insight into the dynamics of the lexicon. They operate on the level of individual terms, often without metadata, making it hard to see patterns and connections. It is thinkable, though, that language change affects not just individual terms but also clusters of (related) terms, that show interaction in their motions of semantic drift. Also, some types of words might change more than others. Structured knowledge sources can help derive such insights. For instance, lexical resources allow to group and connect findings for individual terms by their relation.

At the same time, structured knowledge bases (KBs) benefit from enrichment with semantic change information derived from word embeddings. The largest diverse collection of open knowledge, the Linked Open Data cloud, contains billions of facts about entities and their relations. However, shifts in meaning of these entities are not explicitly encoded, and thus unavailable to these applications. For instance in digital library applications, this could cause problems when mapping contemporary user queries to the metadata vocabulary of archived documents. Khan et al.[7] have introduced a vocabulary, LemonDIA, to describe lexical semantic shifts in KBs from a linguistic perspective. This vocabulary is compatible with, and the expressed knowledge is complementary to, the data curated here.

This paper is a step towards the goal of a structured, interconnected knowledge source of diachronic lexical semantics. It presents an interlinking effort between HistWords, a unique corpus of (open) lexical change data, and WordNet, a lexical database that is part of the Linked Open Data cloud. This combination results in a knowledge graph were concepts, linguistic data elements such as lexemes, and semantic change scores can be queried together. By publishing the data in RDF, we aim to contribute to the (re-)usability of these open corpora.

In the remainder of this paper, we discuss how we linked the HistWords data to lexical entries in WordNet and how the result was represented in an RDF data model. With example queries on this aggregated dataset we demonstrate the use as well as the limitations of the approach.

2 Source data

HistWords. HistWords is a research project of 'Word embeddings for Historical Text' at Stanford University that has produced sets of word embeddings and cross-decade lexical change scores. We used all ready-made lexical change scores for English[1], i.e., for the 10.000 most frequent, non-proper-noun words from the English Google N-Grams dataset[2]. The entries in HistWords are not lemmatised, disambiguated or part-of-speech tagged, hence each similarity score reflects all senses and grammatical functions in which the word can occur. The linking effort to WordNet, which does distinguish between different parts of speech, does not solve this issue, but does make it more explicit. The similarity scores are given between discreet decades. They were calculated as the cosine similarity between the vector for a term derived from corpus material in one decade, and the vector for the same term derived from materials from the other decade. Training was done by the ("word2vec") skip-gram method with negative sampling [9].

Figures are available for every two consecutive decades between 1810 and 1990; i.e., the degree of semantic stability of a lexical term from the 1810s to the 1820s, the 1820s to 1830s, and so on, up to 1980s-1990s. As an example, the word *gay* seems to have underwent semantic change between the 1980s and 1990s, where the cosine similarity between the two term representations fell to 0.91 (from 0.96 for the 1970s-1980s). In addition, there are figures for every

[1] `http://snap.stanford.edu/historical_embeddings/eng-all_sgns.zip`, fullstats
[2] `http://storage.googleapis.com/books/ngrams/books/datasetsv2.html`

decade vs. the 1990s, i.e., for 1810s vs.1990s until 1980s vs. 1990s. These can be used to express the overall change of a lexeme in, for instance, the 20th century (1900s-1990s), or over the entire dataset (1810s-1990s). Due to corpus characteristics, some entries have (some) missing values, which were left out.

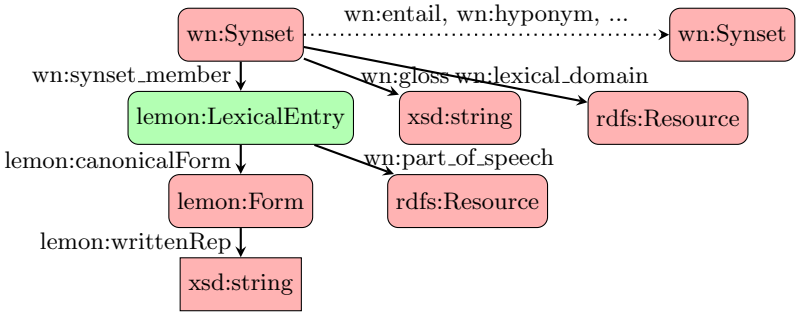

Fig. 1. The basic types of the WordNet RDF model.

WordNet. WordNet [10] is a lexical database of English. It is based on the idea of synsets, synonymous terms of a given grammatical category that express the same concept. One term hence can appear in multiple synsets; e.g., *gay*(adj.) is part of a synset of adjectives to denote "homosexual or arousing homosexual desires" (alongside *homophile* and *queer*) and a synset of adjectives for "bright and pleasant; promoting a feeling of cheer" (alongside *cheery* and *sunny*).

The RDF conversion of WordNet [8, 11] (henceforth *RDF-WordNet*) used in this project is based on the Lemon vocabulary. The basic resource types in RDF-WordNet are shown in Figure 1. A `lemon:lexicalEntry` represents a single lemma of some grammatical type, of which RDF-WordNet counts 158K. The unique base form of each lemma (of type `Lemon:Form`) is pointed to by `lemon:canonicalForm`. The grammatical type is indicated through property `wn:part_of_speech`. A `lemon:LexicalEntry` instance connects to one or more senses (`wn:Synset`) through `wn:synset_member`. Property `wn:gloss` relates a `wn:Synset` instance to its definition. When applicable, synsets are interrelated through semantic relations such as hyponymy, entailment, and meronymy. Additionally, each synset is categorised (using `wn:lexical_domain`) into one of 46 semantic-grammatical types such as `noun.artifact` and `verb.emotion`.

3 Approach

The sourced similarity scores were transformed into change data and connected to WordNet through (stemming and) string matching. The result was represented in RDF and OWL and made available as a Turtle download[3].

[3] `www.github.com/aan680/SemanticChange`

Deriving semantic change scores. The scores were converted to distance measures as we care about the degree of change rather than the degree of stability of the words' meaning. This was done with an arc-cosine transformation rather than by the formula $1 - cosine\_similarity$ to stretch the scale of the change interval and trace more fine-grained differences. The semantic change rate thus lies between 0 and $\pi/2$ (in our dataset, between 0.09 and 1.48). For instance, between the 1980s and 1990s the change values ranged from 0.11 (*pepper*) to 1.12 (*web*). The rates for a larger period are generally higher than those for consecutive decades, e.g. 0.97 for *gang* between the 1810s and 1990s. The change scores have no clear absolute meaning but can be used contrastively between terms or time frames.

Linking HistWords to WordNet. The words in HistWords were mapped onto `lemon:LexicalEntry` instances in RDF-WordNet. First, we merged on an exact match between a word in HistWords and the value of the `lemon:writtenRep` property of the `lemon:Form` corresponding to the `lemon:LexicalEntry` instance. Since the HistWord words are not part-of-speech specific, they were mapped onto all lexical matches in WordNet, irrespective of grammatical type. This step resulted in 7.365 matches for the 10.000 source words.

Second, unmapped HistWords entries were Porter stemmed and re-matched based on an exact match of the stem and a WordNet entry. We included the matches as new `lemon:lexicalEntry` instances with their unstemmed form as the canonical form, and connected them to their WordNet `lemon:lexicalEntry` counterparts through the `lemon:lexicalVariant` property. This brought the total number of mappings to 8.878 out of 10.000 source entries, connected to 12.469 `lemon:LexicalEntry` instances. In future work, it is likely that more words can be matched by refining our stem-and-match technique.

Data model. The resulting data, i.e., the tuples {lexical entry, decade1, decade2, change value}, were represented in RDF. Existing vocabularies were used where possible; newly introduced classes and properties are recognisable by the `cwi-sc` prefix. Figure 2 illustrates how a `lemon:LexicalEntry` is connected to a (blank) node of type `cwi-sc:SemanticChange` for each data tuple with a value and an onset and offset decade. The latter two were modelled, in accordance with OWL-Time[4], as intervals with a start and an end date.

Following OWL-Time ensures interoperability and supports temporal reasoning, but complicates queries for the semantic change of a word between two specified decades. For this reason we introduced a shortcut property for each set of decades, which directly connects a `lemon:LexicalEntry` instance to the semantic change value. The property URI encodes the decades it contrasts, e.g., `cwi-sc:semantic_change_1910s-1920s` leads to the change score between the 1910s and the 1920s.

Note that instead of at the `lemon:LexicalEntry` level, we could have linked the HistWords entries to the `lemon:Form` level, representing the lexeme. We

[4] `http://www.w3.org/TR/owl-time/`

decided against this since it would greatly complicate the queries that we anticipate at the `LexicalEntry` or `Synset` level, while yielding only 334 mappings to inflectional variants, part of which were now matched in the second mapping step.

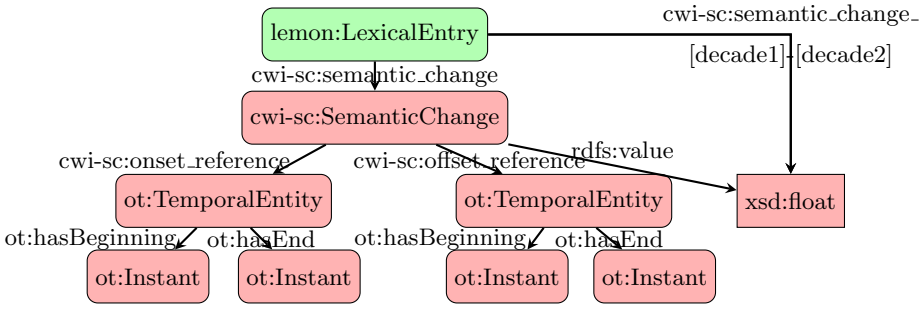

Fig. 2. A model for connecting WordNet entries to cross-decade scores of lexical change. Prefix `ot` stands for OWL-Time and `cwi-sc` for the purpose-built vocabulary.

4 Usage examples

We used the semantic web server ClioPatria [12] to query the RDF dataset of semantic change scores in combination with RDF-WordNet. Below we show example queries that exploit the connection to WordNet as a background source.

Example 1: average change per semantic/linguistic category. We collected the change rate between the decades 1810s and 1990s for all lexical entries as a proxy for their overall change score (alternatively, we could have averaged over all subsequent-decade scores), and related these scores to, first, their part of speech property, and second, the WordNet domain they belong to. Recall that the HistWord index consists of raw word forms; thanks to WordNet, we can annotate these with grammatical and semantic information.

Figure 3 summarises the results and shows the spread of the change scores grouped by the parts of speech distinguished in WordNet. It shows that the change rates are evenly distributed over the grammatical categories. Looking at the distribution over parts of speech of the word entries themselves (Table 1), though, we see that our dataset contains relatively many verbs and adjectives and few nouns as compared to WordNet.

Table 2 shows examples of semantic domains and the mean change score of their lexical entries. Words referring to processes, phenomena and events have seen a higher degree of change than words for food, feelings, or the weather.

Example 2: the relationship between polysemy and semantic change. The synset structure of WordNet provides a simple way to quantify the degree of polysemy of a word. Hamilton et al.[5] find a positive correlation between the

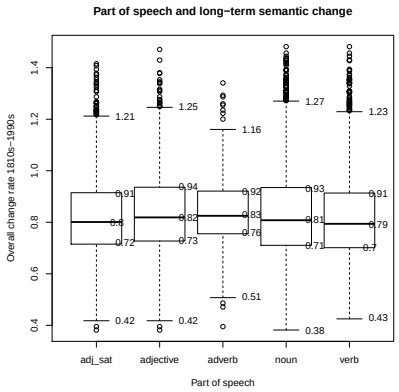

Fig. 3. The spread of the change score of lexical entries between the 1810s and 1990s by part of speech.

POS	dataset	%	RDF-WN	%
noun	4410	44	118303	75
verb	2021	20	11540	7
adjective	1111	11	8358	5
adjective satellite	1941	19	15068	10
adverb	504	5	4475	3
TOTAL	9987	100	157744	100

Table 1. The distribution over parts of speech of entries with a change score between 1810s and 1990s in our dataset and in RDF-WN.

Mean change score	**Domain**	**Mean change score**	**Domain**
0.909	noun.process	0.814	verb.body
0.872	noun.phenomenon	0.791	noun.animal
0.869	noun.event	0.784	noun.food
0.867	noun.act	0.778	noun.feeling
0.86	noun.possession	0.737	verb.weather

Table 2. Examples of WordNet domains and the average change between the 1810s and 1990s (Avg) of their lexical entries.

degree of change of words and their polysemy. They quantify polysemy using a co-occurrence network derived from a large text corpus, under the assumption that polysemous words tend to co-occur with words that do not tend to mutually co-occur. We were curious if we found the same effect when quantifying polysemy directly based on WordNet, as the number of senses (synsets) related to a word.

We plot the change score for 1810s-1990s of each word form (again, as a proxy for the overall change, as do [5]) against the number of synsets related to that word form (Figure 4). One complicating factor is that a word form can be related to several lexical entries, for several parts of speech. Therefore, we also plot the change rate of lexical entries (rather than word forms) against their corresponding number of synsets. With neither of these tests, however, were we able to replicate the results of [5]: on our data we found just a very weak positive correlation (Kendall = 0.06 and 0.05 for words and lexical entries, respectively).

Example 3: exploring senses responsible for semantic drift. Upon browsing the dataset, we came across the word *yellow*. While this term did not display a great degree of change for most decades, we noticed a local peak in change for time period 1910s-1920s, where the score went from 0.25 (for 1900s-1910s) to 0.28

to then fall back to 0.23 (1920s-1930s) and climb up again to 0.25 (1930s-1940s). Clicking through to the senses of the word *yellow*, as RDF-WordNet allows one to do, we found a sense unknown to us. In addition to the colour, *yellow* is an adjective meaning *easily frightened*, with synonyms such as *chickenhearted*. Maybe the word was used in the two World Wars to refer to not-so-brave soldiers? This would explain the observed peaks. Since the change scores are not part-of-speech-, let alone sense-disambiguated, the answer is not in our dataset. For conclusions we would need to go back to the underlying (open source) text corpus, Google N-grams, and have a close look at the term's occurrences. This example illustrates that our dataset is an addition to close reading methods.

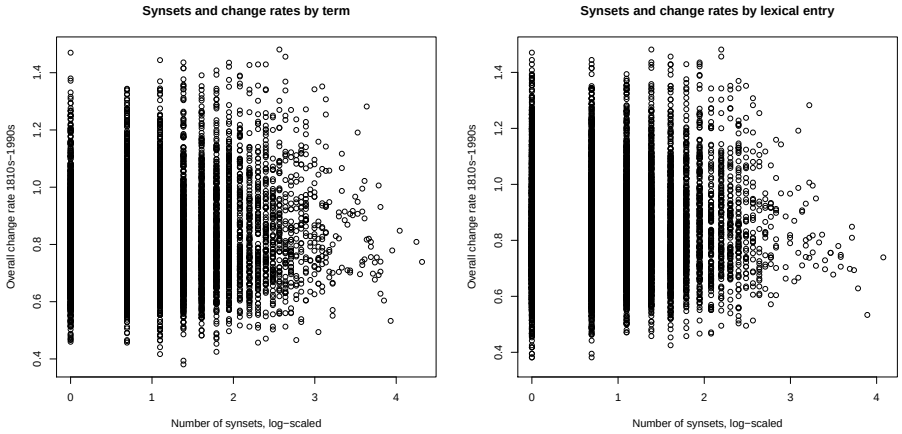

Fig. 4. Number of synsets and overall change rate by term (left) and by lexical entry (right).

5 Discussion and future work

Taking HistWords and WordNet as data sources, this paper demonstrated why and how statistical findings of lexical semantic change can benefit from a connection to a structured knowledge base. Support for the opposite claim, that knowledge bases benefit from lexical change information, is left for future work.

We have shown how this project enables us to aggregate semantic change scores over semantic and linguistic categories. The queries also highlighted some limitations. Lexical change scores, although useful, are heavily refined. Derived from word vectors, distances no longer tell us anything about the nature of the contrasts between two words. The word vectors in turn are derived from mentions in a text corpus that are not included in the dataset itself. An open question remains what sort of data researchers in the field would like to see curated and integrated to benefit from a single source.

In future work, we intend to enrich the dataset in several directions. Envisaged are a cross-lingual dictionary such as Babelnet, to see if other languages display parallels in their lexical patterns of change, and a frame-semantic source like Framenet as an alternative ground for grouping term-level findings. Another addition we aspire is a score set based on part-of-speech-tagged word vectors.

6 Acknowledgments

This work was partially supported by H2020 project VRE4EIC under grant agreement No 676247.

References

1. Andreas Blank. Words and concepts in time: towards diachronic cognitive onomasiology. 2003.
2. Peter De Bolla. *The architecture of concepts: the historical formation of human rights*. Oxford University Press, 2013.
3. Costas Gabrielatos and Paul Baker. Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996-2005. *Journal of English linguistics*, 36(1):5–38, 2008.
4. Kristina Gulordava and Marco Baroni. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, pages 67–71. Association for Computational Linguistics, 2011.
5. William L Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*, 2016.
6. Tom Kenter, Melvin Wevers, Pim Huijnen, and Maarten de Rijke. Ad hoc monitoring of vocabulary shifts over time. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pages 1191–1200. ACM, 2015.
7. Fahad Khan, Javier E Díaz-Vera, and Monica Monachini. Representing polysemy and diachronic lexico-semantic data on the Semantic Web. In *Proceedings of the Second International Workshop on Semantic Web for Scientific Heritage co-located with 13th Extended Semantic Web Conference (ESWC 2016)*, 2016.
8. John P McCrae, Christiane Fellbaum, and Philipp Cimiano. Publishing and linking WordNet using lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*, 2014.
9. T Mikolov and J Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 2013.
10. George A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.
11. Mark Van Assem, Aldo Gangemi, and Guus Schreiber. Conversion of WordNet to a standard RDF/OWL representation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC06), Genoa, Italy*, pages 237–242, 2006.
12. Jan Wielemaker, Wouter Beek, Michiel Hildebrand, and Jacco van Ossenbruggen. ClioPatria: A logical programming infrastructure for the Semantic Web. *Semantic Web Journal*, 2015.