Simulating Lexical Semantic Change from Sense-Annotated Data

May 25, 2019

Dominik Schlechtweg, Sabine Schulte im Walde
Institute for Natural Language Processing, University of Stuttgart, Germany

Motivation

- obtain **evaluation data** for evaluation of Lexical Semantic Change (LSC) detection models

Existing Work

- small test sets **annotated** by humans (e.g. Schlechtweg et al., 2018)
- **synthetic** data sets (pseudo-change) (e.g. Rosenfeld & Erk, 2018)

Lexical Semantic Change

- new and old senses are semantically related (Blank, 1997)
- polysemy is the synchronic result of LSC (Blank, 1997; Bybee, 2015)

→ **synchronic** word senses are a good basis to simulate **diachronic** LSC

Corpus

- **SemCor** is a sense-tagged corpus of English
- consists of a subset of the Brown Corpus
- 700,000 words, with more than 200,000 sense-annotated
- is lemmatized and POS-tagged
- there exist similar corpora in different languages

Sense Frequency Distributions

- a Sense Frequency Distribution (SFD) encodes how often a word *w* occurs in each of its senses
  - **sense 1**: plant, works, industrial plant (buildings for carrying on industrial labor); "they built a large plant to manufacture automobiles"[1]
  - **sense 2**: plant, flora, plant life (botany: a living organism lacking the power of locomotion)

[1] `https://wordnet.princeton.edu/`

Sense-Annotated Corpus

This reduces the number of expensive **plant** shutdowns and startups.	**(s1)**
The pilot **plant** was equipped with a 3-hp. turbine aerator (Figure 2).	**(s1)**
Remove about half the branches from each **plant**, leaving only the strongest with the largest buds.	**(s2)**
"On the side toward the horizon – the southern hemisphere – it is spring; **plants** are being taught to grow".	**(s2)**
Can you share medical facilities and staff with neighboring **plants**??	**(s1)**

Table 1: Corpus sample for the noun *plant*. SFD: $T = (3, 2)$

Split Corpus

t_1	t_2
0000 remove about half the branch from each **plant** leave only the strong with the largest bud **(s2)**	1111 the pilot **plant** was equip with a 3 hp turbine aerator figure 2 **(s1)**
0000 on the side toward the horizon the southern_hemisphere it be spring **plant** are being teach to grow **(s2)**	1111 this reduce the number of expensive **plant** shutdown and startup **(s1)**
0000 can you share medical facility and staff with neighboring **plant** **(s1)**	

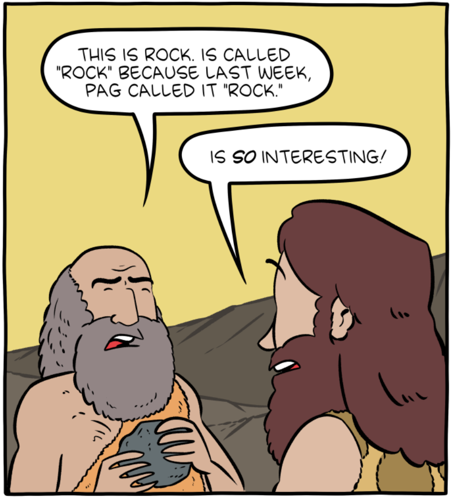
Table 2: $T_1 = (1, 2)$, $T_2 = (2, 0)$. Condition 3.

Experiments

Dataset	Representation	best
SemCor	SGNS	**0.444**
	CNT	0.385
	SVD	0.367
	RI	0.277
	PPMI	0.268

Table 3: Best ρ score across parameter settings for cosine distance in different semantic vector space types. See Schlechtweg et al. (2019) for more information. SemCor was split weakly (condition 3), we tested only on verbs above frequency of 100.

Last Slide


THIS IS ROCK. IS CALLED "ROCK" BECAUSE LAST WEEK, PAG CALLED IT "ROCK."
IS *SO* INTERESTING!

In the early days, etymology was much easier.

Figure 1: Found on `http://languagelog.ldc.upenn.edu`.

Bibliography

Bentivogli, L., & Pianta, E. (2005). Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. *Natural Language Engineering*, *11*(3), 247–261.

Blank, A. (1997). *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*. Tübingen: Niemeyer.

Bybee, J. L. (2015). *Language change*. Cambridge, United Kingdom: Cambridge University Press.

Francis, W. N., & Kucera, H. (1979). *Brown corpus manual* (Tech. Rep.). Department of Linguistics, Brown University, Providence, Rhode Island, US. Retrieved from `http://icame.uib.no/brown/bcm.html`

Henrich, V., & Hinrichs, E. (2013). Extending the tüba-d/z treebank with germanet sense annotation. In I. Gurevych, C. Biemann, & T. Zesch (Eds.), *Language processing and knowledge in the web* (pp. 89–96). Berlin, Heidelberg: Springer Berlin Heidelberg.

Langone, H., Haskell, B. R., & Miller, G. A. (2004). Annotating wordnet. In *Proceedings of the workshop frontiers in corpus annotation@hlt-naacl 2004, boston, ma, usa, may 6, 2004.*

Rosenfeld, A., & Erk, K. (2018). Deep neural models of semantic shift. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, NAACL-HLT 2018, new orleans, louisiana, usa, june 1-6, 2018, volume 1 (long papers)* (pp. 474–484).

Schlechtweg, D., Hätty, A., del Tredici, M., & Schulte im Walde, S. (2019). A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In *Proceedings of the 57th annual meeting of the association for computational linguistics (volume 1: Long papers).* Florence, Italy: Association for Computational Linguistics.

Schlechtweg, D., Schulte im Walde, S., & Eckmann, S. (2018). Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (p. 169-174). New Orleans, Louisiana.