University of Stuttgart
Germany

Human and Computational Measurement of Lexical Semantic Change

March 24, 2022

Dominik Schlechtweg
Supervisor: apl. Prof. Dr. Sabine Schulte im Walde
Institute for Natural Language Processing, University of Stuttgart, Germany

Introduction

- human language changes over time
- semasiological vs. onomasiological

(1) *Der zweyte Theil vom Bauernrechte ist schon lange aus der Presse;*
'The second part of Farmers' Rights already left the press;'

(2) *Alle Freiheiten suspendirt! die persönliche Freiheit wie die der Presse!* 'All freedoms suspended! the personal freedom as well as the one of the press!'

- Lexical Semantic Change Detection (LSCD)
  1. the "digital turn" in the humanities
  2. new computational models of word meaning (word embeddings)
- motivation: support historical semanticists to find semantic changes (more and faster)
- problem: **correctness**

Previous Research

When it comes to evaluating methods and systems, there is a general lack of standardized evaluation practices. Different papers use different datasets and testset words, making it difficult or impossible to compare the proposed solutions. Proper evaluation metrics for semantic change detection and temporal analog detection have not been yet established.

(Tahmasebi, Borin, & Jatowt, 2018)

Aim & Contribution

- provide solid evaluation for LSCD, including
  - definition of basic concepts and tasks
    - (Schlechtweg et al., 2020; Schlechtweg & Schulte im Walde, 2020)
  - annotation schemes
    - (Schlechtweg et al., 2018; Schlechtweg, Tahmasebi, et al., 2021)
  - multilingual benchmark test set
    - (Hätty et al., 2019)
  - model evaluation
    - first systematic evaluation of type-based models (Schlechtweg, Hätty, et al., 2019)
    - first SemEval shared task (Schlechtweg et al., 2020)
    - analysis of BERT cluster biases (Laicher et al., 2021)
  - application
    - discovery of new changes for historical semantics/lexicography (Kurtyigit et al., 2021)

Lexical Semantic Change

- in historical semantics word meaning change is generally defined as changes in word senses
- a common definition is Blank (1997)'s distinguishing two main types:
  - **innovative meaning change**: emergence of a full-fledged additional sense of a word
  - **reductive meaning change**: loss of a full-fledged sense of a word

Human Measurement of Lexical Semantic Change

A	1824	and taking a knife from her pocket, she opened a vein in her little **arm**,	☺
B	1842	And those who remained at home had been heavily taxed to pay for the **arms**, ammunition;	⚔
C	1860	and though he saw her within reach of his **arm**, yet the light of her eyes seemed as far off	☺
		...	
D	1953	overlooking an **arm** of the sea which, at low tide, was a black and stinking mud-flat	
E	1975	twelve miles of coastline lies in the southwest on the Gulf of Aqaba, an **arm** of the Red Sea.	
F	1985	when the disembodied **arm** of the Statue of Liberty jets spectacularly out of the	☺

Table 1: Sample of diachronic corpus.

Word Use Pairs

(A) [. . .] and taking a knife from her pocket, she opened a vein in her little **arm**, and dipping a feather in the blood, wrote something on a piece of white cloth, which was spread before her.

(D) It stood behind a high brick wall, its back windows overlooking an **arm** of the sea which, at low tide, was a black and stinking mud-flat [. . .]

Semantic Proximity Scale

↑	4: Identical
	3: Closely Related
	2: Distantly Related
	1: Unrelated

Table 2: DURel relatedness scale.

Graph representation

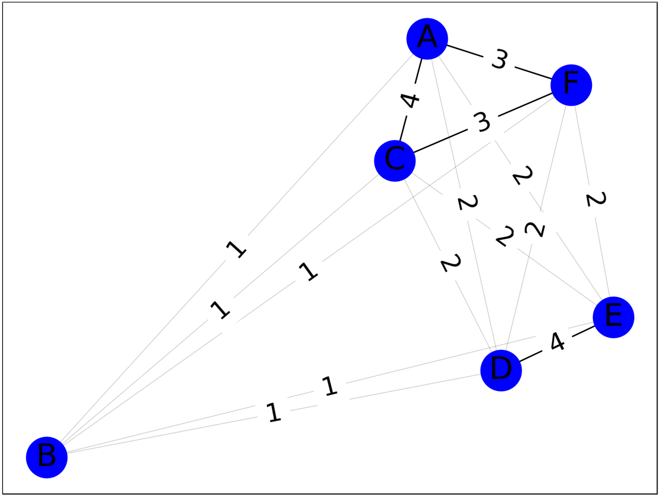


Figure 1: Word Usage Graph of English *arm*.

Clustering

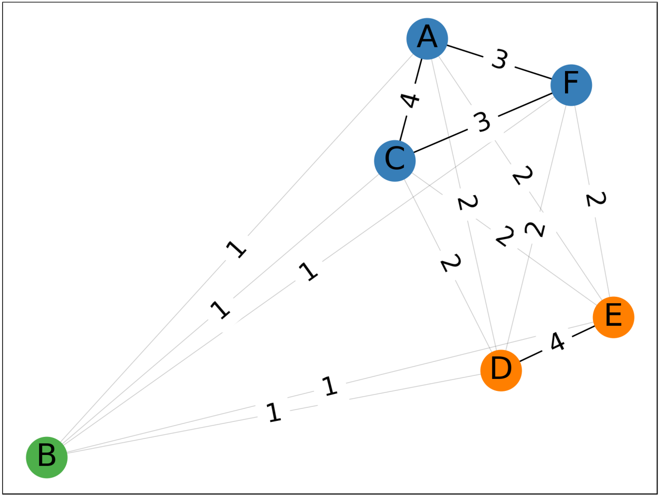
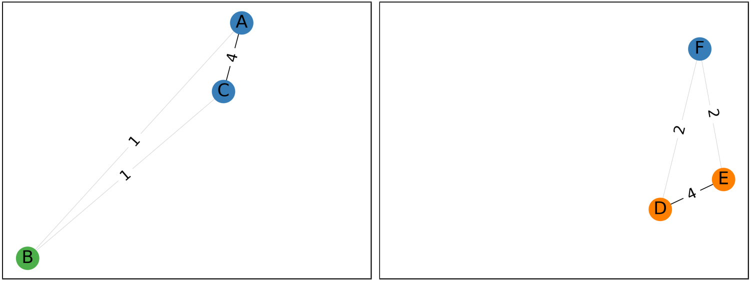


Figure 2: Word Usage Graph of English *arm*. $D = (3, 2, 1)$.

Lexical Semantic Change



$t_1, \ D_1 = (2, 0, 1)$ $t_2, \ D_2 = (1, 2, 0)$

Change Scores

- **binary change** (loss and gain of senses)
- **graded change** (changes in sense probabilities)

Validation

- **agreement** is reasonably high (.51 – .83)
- correspondence to traditional **sense assignments** moderately high (.65)
- problem: sparsity
  - adding more annotations improves correspondence and causes variations in change scores

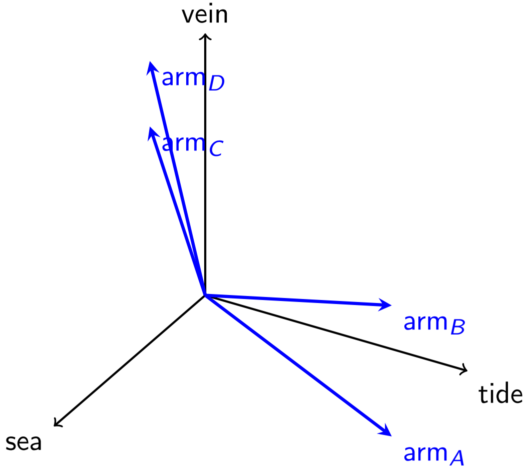
Computational Measurement of Lexical Semantic Change

- unsupervised
- distributional
- vector space models

Token-based VSMs

- model the human measurement process
- contextualized embeddings (BERT, ELMo)
- one vector per use
- composed by
  1. semantic representation per word use (token)
  2. clustering method (optional)
  3. change measure

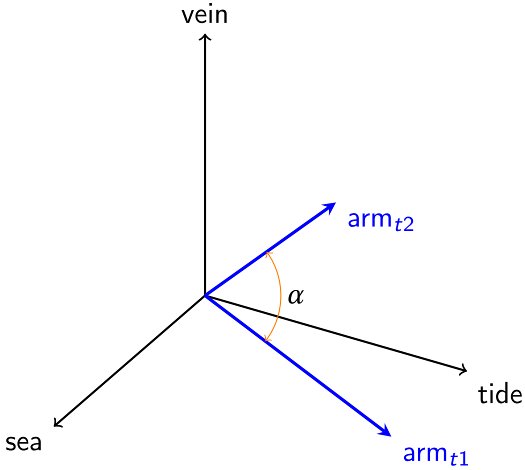
Simple Model



Type-based VSMs

- do not model the human measurement process
- one average vector per word (Word2Vec, GloVe)
- composed by
  1. semantic representation per word (type vector)
  2. alignment
  3. measure

Simple Model



Evaluation

Subtask 1 Binary classification: for a set of target words, predict the binary change score

Subtask 2 Ranking: rank a set of target words according to their graded change score

Results

System	Binary	Graded
type	**.63**	**.33**
token	.60	.26

Table 3: Average performance of best submissions per subtask for different system types.

Analysis

Measure	Raw	Preprocessed
form bias	**.44**	.15
performance	.12	**.62**

Table 4: Cluster bias and performance on graded change on DWUG DE.

Application

Discovery Task Classification: Decide for a **large set of unseen** words which ones lost or gained senses

Results

System	Performance
type	**.71**
token	.62
random	.35

Table 5: Performance type- and token-based compared to random baseline.

Discovered Change

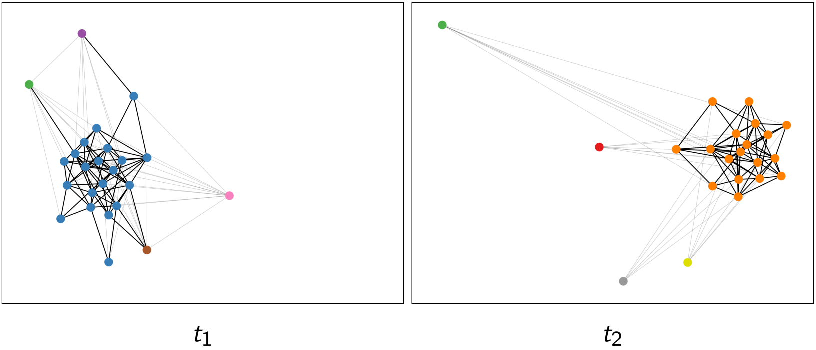


Figure 4: Word Usage Graph of German *Zehner*.

Conclusion

- complete evaluation framework for LSCD
- **humans**:
  - simple annotation strategy
  - humans show reasonable agreement
  - clusterings reflect traditional sense distinctions
  - → LSC can be measured **inter-subjectively** with humans
- **computers**:
  - models show medium to high performance
  - type embeddings dominate token embeddings
  - preprocessing has a major influence token embeddings
  - both model types discover new semantic changes with above-random probability
  - → LSCD is a **valid and meaningful NLP task** which can be solved reasonably well with computers

Discussion

- range of **follow-up** studies:
  - shared tasks for Italian, Russian and Spanish
  - human annotation framework
  - annotation modeling and optimization
- annotation style has inspired transfer of **WiC models** to LSCD
  - quantum leap in performance
  - recent dominance of token embeddings

Future Research

- improve data quality:
  1. add more annotations
  2. clean existing data sets
  3. use alternative annotation strategies
- multiple time points
- fine-tune token embeddings on semantic proximity judgments

Bibliography I

Alatrash, R., Schlechtweg, D., Kuhn, J., & Schulte im Walde, S. (2020, may). CCOHA: Clean Corpus of Historical American English. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 6958–6966). Marseille, France: European Language Resources Association. Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.859

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

Dubossarsky, H., Hengchen, S., Tahmasebi, N., & Schlechtweg, D. (2019). Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 457–470). Florence, Italy: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P19-1044/

Hätty, A., Schlechtweg, D., Dorna, M., & Schulte im Walde, S. (2020). Predicting Degrees of Technicality in Automatic Terminology Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Seattle, Washington: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.258/

Hätty, A., Schlechtweg, D., & Schulte im Walde, S. (2019). SURel: A gold standard for incorporating meaning shifts into term extraction. In *Proceedings of the 8th Joint Conference on Lexical and Computational Semantics* (pp. 1–8). Minneapolis, MN, USA. Retrieved from https://www.aclweb.org/anthology/S19-1001/

Hengchen, S., Tahmasebi, N., Schlechtweg, D., & Dubossarsky, H. (2021). Challenges for Computational Lexical Semantic Change. In N. Tahmasebi, L. Borin, A. Jatowt, Y. Xu, & S. Hengchen (Eds.), *Computational approaches to semantic change* (Vol. Language Variation, chap. 11). Berlin: Language Science Press. Retrieved from https://arxiv.org/abs/2101.07668v1

Kaiser, J., Kurtyigit, S., Kotchourko, S., & Schlechtweg, D. (2021, apr). Effects of pre- and post-processing on type-based embeddings in lexical semantic change detection. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume* (pp. 125–137). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.eacl-main.10 doi: 10.18653/v1/2021.eacl-main.10

Bibliography II

Kaiser, J., Schlechtweg, D., Papay, S., & Schulte im Walde, S. (2020). IMS at SemEval-2020 Task 1: How low can you go? Dimensionality in Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation.* Barcelona, Spain: Association for Computational Linguistics. Retrieved from https://arxiv.org/abs/2008.03164

Kaiser, J., Schlechtweg, D., & Schulte im Walde, S. (2020). OP-IMS @ DIACR-Ita: Back to the Roots: SGNS+OP+CD still rocks Semantic Change Detection. In V. Basile, D. Croce, M. Di Maro, & L. C. Passaro (Eds.), *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020).* Online: CEUR.org. Retrieved from https://arxiv.org/abs/2011.03258 (Winning Submission!)

Kurtyigit, S., Park, M., Schlechtweg, D., Kuhn, J., & Schulte im Walde, S. (2021, aug). Lexical semantic change discovery. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers).* Online: Association for Computational Linguistics.

Laicher, S., Baldissin, G., Castaneda, E., Schlechtweg, D., & Schulte im Walde, S. (2020). CL-IMS @ DIACR-Ita: Volente o Nolente: BERT does not outperform SGNS on Semantic Change Detection. In V. Basile, D. Croce, M. Di Maro, & L. C. Passaro (Eds.), *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020).* Online: CEUR.org. Retrieved from https://arxiv.org/abs/2011.07247

Laicher, S., Kurtyigit, S., Schlechtweg, D., Kuhn, J., & Schulte im Walde, S. (2021, apr). Explaining and improving BERT performance on lexical semantic change detection. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Student research workshop* (pp. 192–202). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.eacl-srw.25 doi: 10.18653/v1/2021.eacl-srw.25

Schlechtweg, D., Castaneda, E., Kuhn, J., & Schulte im Walde, S. (2021, aug). Modeling sense structure in word usage graphs with the weighted stochastic block model. In *Proceedings of *sem 2021: The tenth joint conference on lexical and computational semantics* (pp. 241–251). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.starsem-1.23 doi: 10.18653/v1/2021.starsem-1.23

Bibliography III

Schlechtweg, D., Eckmann, S., Santus, E., Schulte im Walde, S., & Hole, D. (2017). German in flux: Detecting metaphoric change via word entropy. In *Proceedings of the 21st Conference on Computational Natural Language Learning* (pp. 354–367). Vancouver, Canada. Retrieved from https://www.aclweb.org/anthology/K17-1036/

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* (pp. 732–746). Florence, Italy: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P19-1072/

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020). SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation.* Barcelona, Spain: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.semeval-1.1/

Schlechtweg, D., Oguz, C., & Schulte im Walde, S. (2019). Second-order co-occurrence sensitivity of skip-gram with negative sampling. In *Proceedings of the 2019 ACL workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP* (pp. 24–30). Florence, Italy: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W19-4803/

Schlechtweg, D., & Schulte im Walde, S. (2018). Distribution-based prediction of the degree of grammaticalization for German prepositions. In C. Cuskley, M. Flaherty, H. Little, L. McCrohon, A. Ravignani, & T. Verhoef (Eds.), *The Evolution of Language: Proceedings of the 12th International Conference (EVOLANGXII).* Online at http://evolang.org/torun/proceedings/papertemplate.html?p=169.

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* (pp. 169–174). New Orleans, Louisiana. Retrieved from https://www.aclweb.org/anthology/N18-2027/

Schlechtweg, D., & Schulte im Walde, S. (2020). Simulating Lexical Semantic Change from Sense-Annotated Data. In A. Ravignani et al. (Eds.), *The Evolution of Language: Proceedings of the 13th International Conference (EvoLang13).* Retrieved from http://brussels.evolang.org/proceedings/paper.html?nr=9 doi: 10.17617/2.3190925

Bibliography IV

Schlechtweg, D., Tahmasebi, N., Hengchen, S., Dubossarsky, H., & McGillivray, B. (2021, nov). DWUG: A large resource of diachronic word usage graphs in four languages. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 7079–7091). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.567

Schütze, H. (1998, March). Automatic word sense discrimination. *Computational Linguistics*, *24*(1), 97–123.

Shwartz, V., Santus, E., & Schlechtweg, D. (2017). Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain* (pp. 65–75). Retrieved from https://www.aclweb.org/anthology/E17-1007/

Tahmasebi, N., Borin, L., & Jatowt, A. (2018). Survey of Computational Approaches to Diachronic Conceptual Change. *arXiv e-prints*.