University of Stuttgart
Germany

Explaining and Improving BERT Performance on Lexical Semantic Change Detection

Severin Laicher Sinan Kurtyigit Dominik Schlechtweg Jonas Kuhn Sabine Schulte im Walde

Lexical Semantic Change Detection (LSCD)

- LSCD is the automatic detection of words whose meaning has changed over time.
- Token-based approaches using BERT performed very poorly in the shared tasks SemEval-2020 and DIACR-Ita.

[Schlechtweg, McGillivray, Hengchen, Dubossarsky, and Tahmasebi 2020]

[Basile, Caputo, Caselli, Cassotti, and Varvara 2020]

Research Questions

- **Why do BERT vector clusterings show poor performance?**

→ Due to a strong influence of orthographic information.

- **Can we improve it?**

→ Yes – By removing orthographic differences **only on the target words**.

Exp. 1: Word Sense Clustering Biases

- Clustering of BERT token vectors
- Based on the clustering results we measure LSC (ρ).
- We measure clustering performance and the following biases (ARI):

1. Word Form
2. Target Word Position
3. Corpora

Exp. 2: BERT Token Performance on LSCD

- We compare different text preprocessings and BERT layers on LSCD
- We measure LSC using average measures: APD and COS
- We observe a strong bias of the target word form.
- To reduce the target word form bias we use token sentences and replace the target word by its lemma.
- We considerably improve our results.

Exp. 1: Results

	Layer	Token	Lemma	TokLem
ρ	1	-.265	-.062	-.170
	12	.123	.427	**.624**
	9-12	.122	.420	.533
ARI	1	.033	.002	.003
	12	.119	.159	**.161**
	9-12	.155	.142	.154
Form	1	**.706**	.024	.004
	12	**.439**	**.056**	**.150**
	9-12	**.420**	**.047**	**.094**
Position	1	.005	**.023**	**.027**
	12	-.002	.005	-.002
	9-12	**.009**	**.018**	**.012**
Corpora	1	.074	.003	.005
	12	**.110**	**.095**	**.096**
	9-12	**.107**	.068	**.089**

Table: **Exp. 1:** German clustering scores. Bold font indicates best scores for ρ and ARI (top) or scores above all corresponding baselines for influence variables (bottom).

Exp. 2: Results

		Layer	Token	Lemma	TokLem
GER	APD	12	.359	.303	**.456**
		1+12	.316	.643	**.731**
		9-12	.407	.305	**.516**
	COS	12	.472	.693	**.755**
		1+12	.373	.698	**.729**
		9-12	.446	.689	**.726**

Table: **Exp. 2:** German LSCD scores for different layers and preprocessings for average measures.

Data

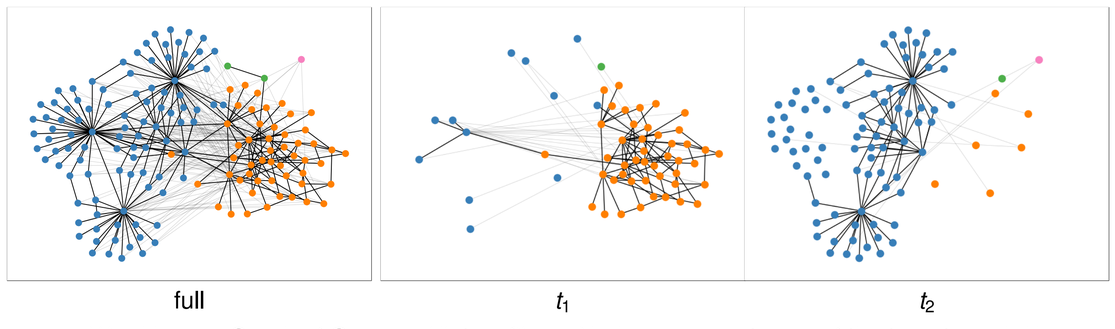

Figure: Word Usage Graph of German *Eintagsfliege*. Nodes represent uses of the target word. Edge weights represent the median of relatedness judgments between uses (**black**/gray lines for **high**/low edge weights). Colors indicate clusters (senses) inferred from the full graph.

Acknowledgments

Dominik Schlechtweg was supported by the Konrad Adenauer Foundation and the CRETA center funded by the German Ministry for Education and Research (BMBF) during the conduct of this study.

References

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. Diacr-ita@ evalita2020: Overview of the evalita2020 diachronic lexical semantics (diacr-ita) task. *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), Online. CEUR. org*, 2020.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. SemEval-2020 task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain, 2020. Association for Computational Linguistics.