



University of Stuttgart
Germany



UNIVERSITY OF
CAMBRIDGE

SemEval 2020 Task 1: Unsupervised Lexical Semantic Change Detection

November 2, 2020

Dominik Schlechtweg,[♣] Barbara McGillivray,^{◇,♡} Simon
Hengchen,[♠] Haim Dubossarsky,[♡] Nina Tahmasebi[♠]
semeval2020lexicalsemanticchange@turing.ac.uk

♣University of Stuttgart, ◇The Alan Turing Institute, ♡University of Cambridge
♠University of Gothenburg

The
Alan Turing
Institute

SPRÅKBANKENTEXT



Introduction

- ▶ evaluation is currently our most pressing problem
- ▶ SemEval is competition-style semantic evaluation series¹
- ▶ SemEval 2020 Task 1 on Unsupervised Lexical Semantic Change Detection (Schlechtweg, McGillivray, Hengchen, Dubossarsky, & Tahmasebi, 2020)²
- ▶ datasets for 4 languages with 100,000 human judgments
- ▶ 2 subtasks
- ▶ 33 teams submitted 186 systems

¹<https://semeval.github.io/>

²<https://languagechange.org/semeval/>

Tasks

- ▶ comparison of two time periods t_1 and t_2
 - (i) reduces the number of time periods for which data has to be annotated
 - (ii) reduces the task complexity
- ▶ two tasks:
 - ▶ **Subtask 1** – Binary classification: for a set of target words, decide which words lost or gained senses between t_1 and t_2 , and which ones did not.
 - ▶ **Subtask 2** – Ranking: rank a set of target words according to their degree of LSC between t_1 and t_2 .
- ▶ defined on **word sense frequency distributions**

Sense Frequency Distributions (SFDs)

	t₁			t₂		
Senses	Chamber	Biology	Phone	Chamber	Biology	Phone
# uses	12	18	0	1	11	18

Figure 1: An example of a sense frequency distribution for the word *cell* in two time periods.

	t_1	t_2
English	CCOHA 1810-1860	CCOHA 1960-2010
German	DTA 1800-1899	BZ+ND 1946-1990
Latin	LatinISE -200-0	LatinISE 0-2000
Swedish	Kubhist 1790-1830	Kubhist 1895-1903

Table 1: Time-defined subcorpora for each language.

Annotation

- ▶ 100–200 changing words selected from etymological dictionaries (OED, 2009; Paul, 2002; Svenska Akademien, 2009)
- ▶ pre-annotation (rough filtering by one annotator)
- ▶ adding of control words with similar frequency properties
- ▶ sample 100 uses (30 for Latin) of each word per time period
- obtain SFDs all samples by annotation
- ▶ graded word sense annotation (Erk, McCarthy, & Gaylord, 2013)
- ▶ mostly based on **DURel** (Schlechtweg, Schulte im Walde, & Eckmann, 2018)

Scale


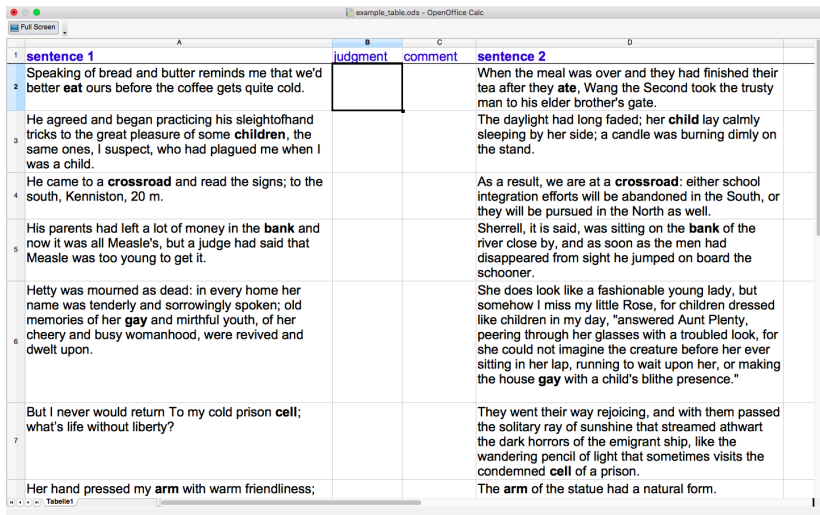
- 
- 4: Identical
 - 3: Closely Related
 - 2: Distantly Related
 - 1: Unrelated
- 0: Cannot decide

Table 2: Four-point scale of relatedness (Schlechtweg et al., 2018).

Data



The image shows a screenshot of an OpenOffice Calc spreadsheet titled "example_table.ods". The spreadsheet is in "Full Screen" mode and displays a table with 7 rows and 4 columns. The columns are labeled "A", "B", "C", and "D". The rows are numbered 1 through 7. The table content is as follows:

	A	B	C	D
1	sentence 1	judgment	comment	sentence 2
2	Speaking of bread and butter reminds me that we'd better eat ours before the coffee gets quite cold.			When the meal was over and they had finished their tea after they ate , Wang the Second took the trusty man to his elder brother's gate.
3	He agreed and began practicing his sleightofhand tricks to the great pleasure of some children , the same ones, I suspect, who had plagued me when I was a child.			The daylight had long faded; her child lay calmly sleeping by her side; a candle was burning dimly on the stand.
4	He came to a crossroad and read the signs; to the south, Kenniston, 20 m.			As a result, we are at a crossroad : either school integration efforts will be abandoned in the South, or they will be pursued in the North as well.
5	His parents had left a lot of money in the bank and now it was all Measle's, but a judge had said that Measle was too young to get it.			Sherrell, it is said, was sitting on the bank of the river close by, and as soon as the men had disappeared from sight he jumped on board the schooner.
6	Hetty was mourned as dead: in every home her name was tenderly and sorrowingly spoken; old memories of her gay and mirthful youth, of her cheery and busy womanhood, were revived and dwelt upon.			She does look like a fashionable young lady, but somehow I miss my little Rose, for children dressed like children in my day, "answered Aunt Plenty, peering through her glasses with a troubled look, for she could not imagine the creature before her ever sitting in her lap, running to wait upon her, or making the house gay with a child's blithe presence."
7	But I never would return To my cold prison cell ; what's life without liberty?			They went their way rejoicing, and with them passed the solitary ray of sunshine that streamed athwart the dark horrors of the emigrant ship, like the wandering pencil of light that sometimes visits the condemned cell of a prison.
	Her hand pressed my arm with warm friendliness;			The arm of the statue had a natural form.

Table 3: Annotation Table.

Diachronic Data

- (1) 1830 but I am bound and thrown into a dark **cell**.
- (2) 1851 ...be fit to burn in a jail; no, not in a condemned **cell**.
- ...
- (3) 1990 But I never would return To my cold prison **cell**.
What's life without liberty?
- (4) 2006 She searched the bag for her **cell** as we headed toward the door.

Diachronic Data

	A	B	C	D
1	sentence 1	judgment	comment	sentence 2
2	but I am bound and thrown into a dark cell .			be fit to burn in a jail; no, not in a condemned cell .
3	but I am bound and thrown into a dark cell .			But I never would return To my cold prison cell ; what's life without liberty?
4	but I am bound and thrown into a dark cell .			She searched the bag for her cell as we headed toward the door.
5	be fit to burn in a jail; no, not in a condemned cell .			But I never would return To my cold prison cell ; what's life without liberty?
6	be fit to burn in a jail; no, not in a condemned cell .			She searched the bag for her cell as we headed toward the door.
7	She searched the bag for her cell as we headed toward the door.			But I never would return To my cold prison cell ; what's life without liberty?

Table 4: Annotation Table.

Diachronic Data

	A	B	C	D
1	sentence 1	judgment	comment	sentence 2
2	but I am bound and thrown into a dark cell .	4		be fit to burn in a jail; no, not in a condemned cell .
3	but I am bound and thrown into a dark cell .	4		But I never would return To my cold prison cell ; what's life without liberty?
4	but I am bound and thrown into a dark cell .	2		She searched the bag for her cell as we headed toward the door.
5	be fit to burn in a jail; no, not in a condemned cell .	4		But I never would return To my cold prison cell ; what's life without liberty?
6	be fit to burn in a jail; no, not in a condemned cell .	2		She searched the bag for her cell as we headed toward the door.
7	She searched the bag for her cell as we headed toward the door.	2		But I never would return To my cold prison cell ; what's life without liberty?

Table 5: Annotation Table.

From DUREl to SFDs

	A	B	C	D
1	sentence 1	judgment	comment	sentence 2
2	but I am bound and thrown into a dark cell .	4		be fit to burn in a jail; no, not in a condemned cell .
3	but I am bound and thrown into a dark cell .	4		But I never would return To my cold prison cell ; what's life without liberty?
4	but I am bound and thrown into a dark cell .	2		She searched the bag for her cell as we headed toward the door.
5	be fit to burn in a jail; no, not in a condemned cell .	4		But I never would return To my cold prison cell ; what's life without liberty?
6	be fit to burn in a jail; no, not in a condemned cell .	2		She searched the bag for her cell as we headed toward the door.
7	She searched the bag for her cell as we headed toward the door.	2		But I never would return To my cold prison cell ; what's life without liberty?

Table 6: Annotation Table.

Word Usage Graphs (WUGs)

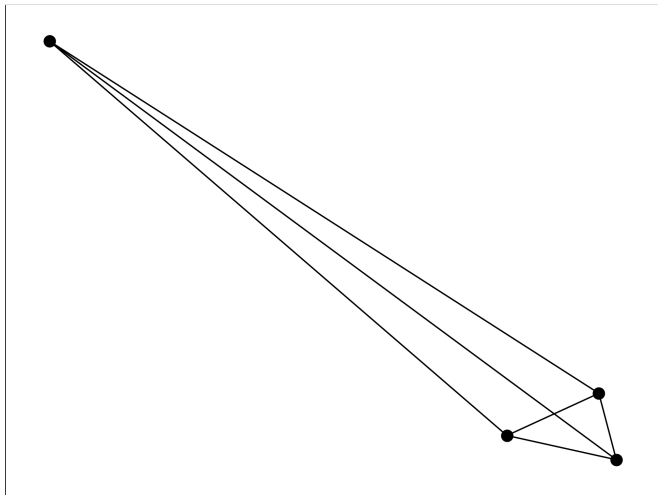


Figure 2: Graph visualization four uses of *cell*.

Word Usage Graphs (WUGs)

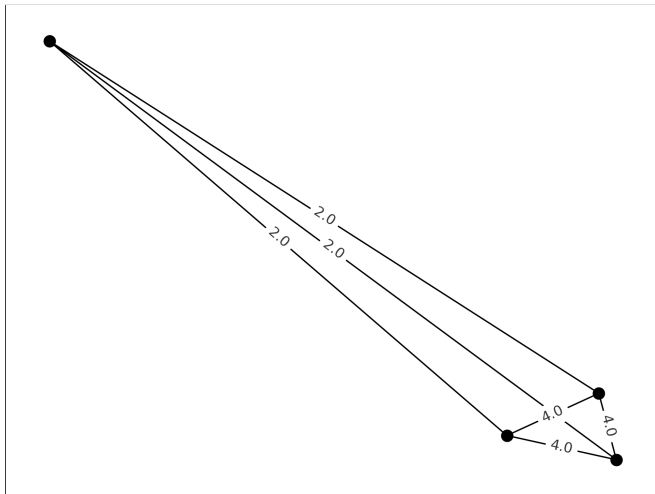


Figure 3: Graph visualization four uses of *cell*.

Clustering

- ▶ correlation clustering (Bansal, Blum, & Chawla, 2004)
- ▶ optimization criterion: **reduce (weighted) number of cluster-edge conflicts**
 - (i) finds the optimal number of clusters on its own
 - (ii) handles missing information (non-observed edges)
 - (iii) robust to errors by using the global information
 - (iv) respects the gradedness of word meaning
 - (v) dominated in simulation study

Clustering

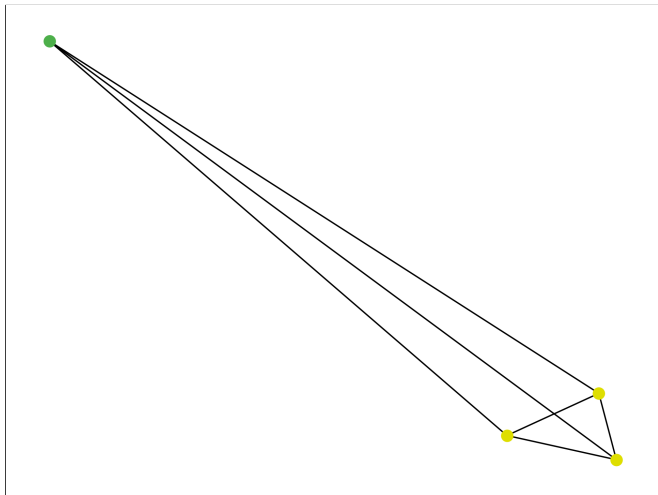


Figure 4: Graph visualization for uses of cell $D = (3, 1)$.

Clustering

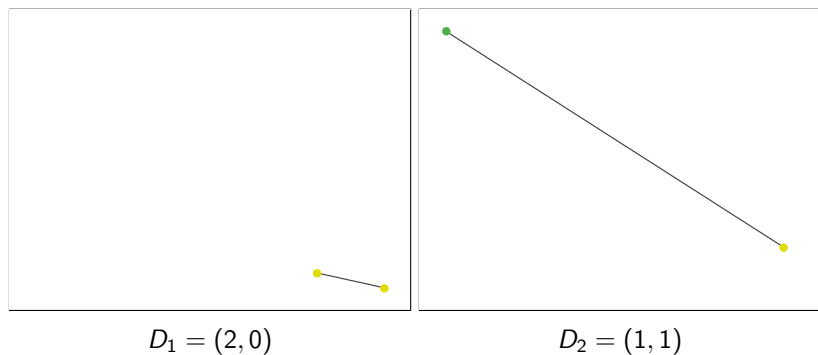
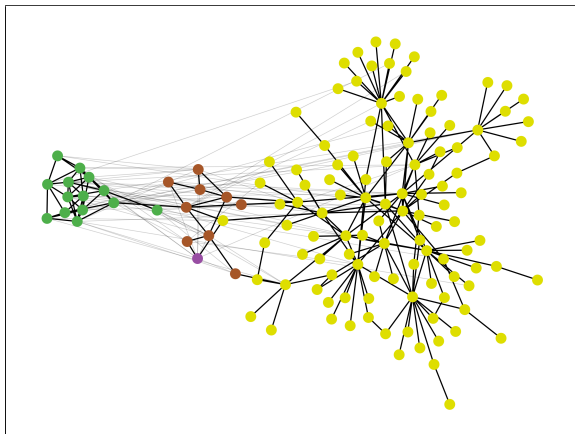
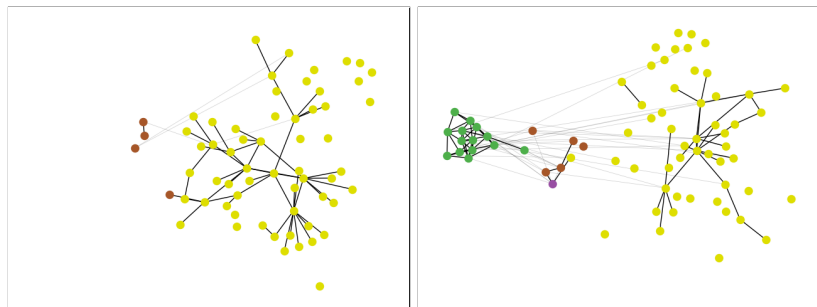


Figure 5: Graph visualization for uses of *cell*. $B(w) = 1$ and $G(w) = 0.5$



$$D = (110, 14, 9, 1)$$

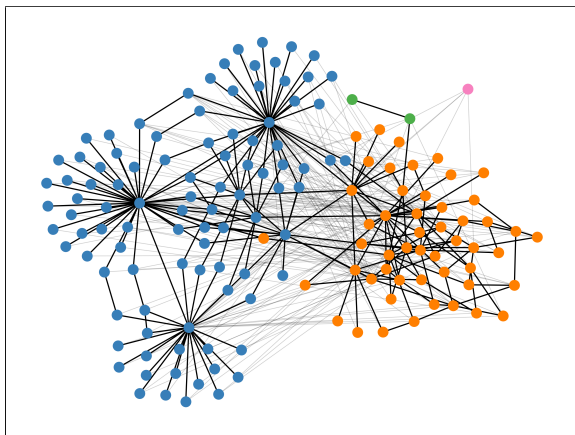
Figure 6: Usage graph of Swedish *ledning*.



$$D_1 = (58, 0, 4, 0)$$

$$D_2 = (52, 14, 5, 1)$$

Figure 7: Usage graph of Swedish *ledning*. $B(w) = 1$ and $G(w) = 0.34$.



$$D = (97, 51, 1, 2)$$

Figure 8: Usage graph of German *Eintagsfliege*.

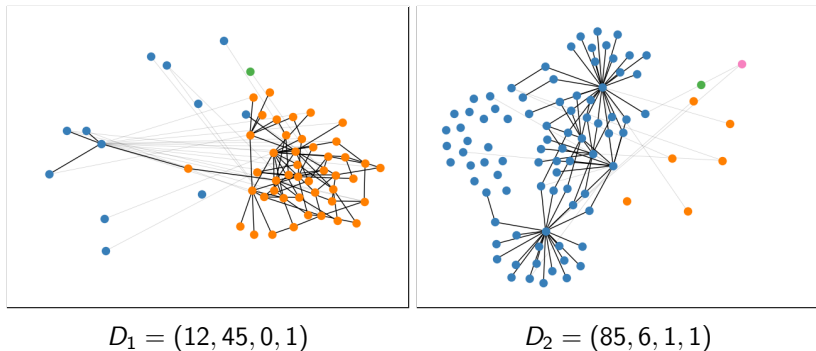


Figure 9: Usage graph of German *Eintagsfliege*. $B(w) = 0$ and $G(w) = 0.66$.

Advantages

- ▶ guarantee that changes are reflected in data
 - ▶ yields high inter-annotator agreement of non-experts
 - ▶ relies on intuitive linguistic concept of **semantic relatedness**
 - ▶ it is well-grounded in cognitive semantic theory
 - ▶ avoids assignment of particular sense to a word use
- requires only minimal preparation efforts
- ▶ annotation interface³
 - ▶ small variation of procedure for Latin

³<https://www.ims.uni-stuttgart.de/en/research/resources/tools/durel-annotations-tool/>

Annotated Data

	n	N/V/A	AGR	JUD
English	37	33/4/0	.69	30k
German	48	32/14/2	.59	38k
Latin	40	27/5/8	-	9k
Swedish	31	23/5/3	.58	20k

Evaluation and Results

- ▶ Subtask 1: target words are classified into two hidden/true classes for binary change
- ▶ Subtask 2: target words are ranked yielding a hidden/true ranking for greater change
- ▶ participants had to predict the true classification and the true ranking in the evaluation phase
- ▶ they were allowed to submit up to 10 submissions from which we selected the best for the final ranking
- ▶ performance was measured with Accuracy and Spearman

Subtask 1 (Binary change)

Team	Subtask 1					System	Model	Threshold
	Avg.	EN	DE	LA	SV			
UWB	.687	.622	.750	.700	.677	type	SGNS+CCA+CD	✓
Life-Language	.686	.703	.750	.550	.742	type	SGNS	✓
Jiaxin & Jinan	.665	.649	.729	.700	.581	type	SGNS+TR+CD	✓
RPI-Trust	.660	.649	.750	.500	.742	type		
UG_Student_Intern	.639	.568	.729	.550	.710	type		
DCC	.637	.649	.667	.525	.710	type		
NLP@IDSIA	.637	.622	.625	.625	.677	token		
JCT	.636	.649	.688	.500	.710	type		
Skurt	.629	.568	.562	.675	.710	token		
Discovery_Team	.621	.568	.688	.550	.677	ens.		
Count Bas.	.613	.595	.688	.525	.645	-		
TUE	.612	.568	.583	.650	.645	token		
Entity	.599	.676	.667	.475	.581	type		
IMS	.598	.541	.688	.550	.613	type		
cs2020	.587	.595	.500	.575	.677	token		
UiO-UvA	.587	.541	.646	.450	.710	token		
NLPCR	.584	.730	.542	.450	.613	token		
Maj. Bas.	.576	.568	.646	.350	.742	-		
cbk	.554	.568	.625	.475	.548	token		
Random	.554	.486	.479	.475	.774	type		
UoB	.526	.568	.479	.575	.484	topic		
UCD	.521	.622	.500	.350	.613	graph		
RIJP	.511	.541	.500	.550	.452	type		
Freq. Bas.	.439	.432	.417	.650	.258	-		

Subtask 2 (Graded change)

Team	Subtask 2					System	Model
	Avg.	EN	DE	LA	SV		
UG_Student_Intern	.527	.422	.725	.412	.547	type	SGNS+OP+ED
Jiaxin & Jinan	.518	.325	.717	.440	.588	type	SGNS+TR+CD
cs2020	.503	.375	.702	.399	.536	type	SGNS+OP+CD
UWB	.481	.367	.697	.254	.604	type	
Discovery_Team	.442	.361	.603	.460	.343	ens.	
RPI-Trust	.427	.228	.520	.462	.498	type	
Skurt	.374	.209	.656	.399	.234	token	
IMS	.372	.301	.659	.098	.432	type	
UiO-UvA	.370	.136	.695	.370	.278	token	
Entity	.352	.250	.499	.303	.357	type	
Random	.296	.211	.337	.253	.385	type	
NLPCR	.287	.436	.446	.151	.114	token	
JCT	.254	.014	.506	.419	.078	type	
cbk	.234	.059	.400	.341	.136	token	
UCD	.234	.307	.216	.069	.344	graph	
Life-Language	.218	.299	.208	-.024	.391	type	
NLP@IDSIA	.194	.028	.176	.253	.321	token	
Count Bas.	.144	.022	.216	.359	-.022	-	
UoB	.100	.105	.220	-.024	.102	topic	
RIJP	.087	.157	.099	.065	.028	type	
TUE	.087	-.155	.388	.177	-.062	token	
DCC	-.083	-.217	.014	.020	-.150	type	
Freq. Bas.	-.083	-.217	.014	.020	-.150	-	
Maj. Bas.	-	-	-	-	-	-	

Type versus token embeddings

System	Subtask 1		Subtask 2	
	Avg.	Max.	Avg.	Max.
type embeddings	0.625	0.687	0.329	0.527
ensemble	0.621	0.621	0.442	0.442
token embeddings	0.598	0.637	0.258	0.374
topic model	0.526	0.526	0.100	0.100
graph	0.521	0.521	0.234	0.234

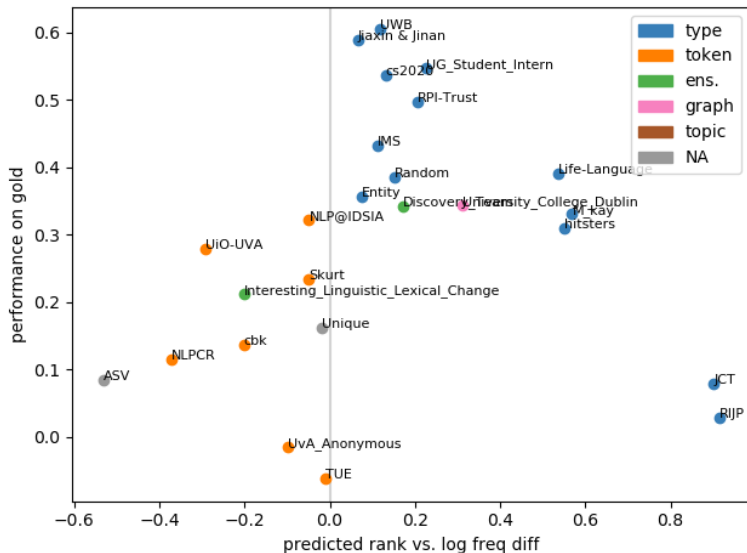
Table 7: Average and maximum performance of best submissions per subtask for different system types. Submissions that corresponded exactly to the baselines or the sample submission were removed.

Type versus token embeddings

- ▶ we suggest these reasons for low performance of (contextualized) token embeddings
 - (i) they are new and lack proper usage conventions
 - (ii) they carry additional, and possibly irrelevant, information that may mask true diachronic changes
 - (iii) restricted context in task corpora
 - (iv) lemmatization in task corpora
- ▶ In order to make the input more suitable for token-based models, we also provide the raw corpora after the evaluation phase and will publish the annotated uses of the target words with additional context⁴

⁴<https://www.ims.uni-stuttgart.de/data/sem-eval-ulscd>

The influence of frequency



Conclusion

- ▶ type embeddings dominate token embeddings
- ▶ type embeddings are strongly influenced by frequency
- ▶ SGNS is dominant type-based embedding architecture
- ▶ OP, TR and CCA are dominant type-based alignment strategies
- ▶ CD is dominant measure for semantic change
- ▶ thresholding instead of clustering works well for Subtask 1 (binary change)
- ▶ results summarized in Schlechtweg et al. (2020)

How solid are these results?: DIACR-Ita shared task

- ▶ Italian data for Subtask 1 (Binary change) (Basile, Caputo, Caselli, Cassotti, & Varvara, 2020)
- ▶ access to full corpus in linear order
- ▶ Subtask 1 should in theory favor sense-differentiating systems (as e.g. token embeddings)
- ▶ but: results reproduce SemEval results

DIACR-Ita results

Rank	Team	Accuracy	Model	Threshold	Type
1.	OP-IMS	0.944	SGNS+OP+CD	✓	type
...					
3.	VI-IMS	0.778	SGNS+VI+CD	✓	type
4.	CL-IMS	0.722	BERT+APD	✓	token
...					
6.	SBM-IMS	0.611	BERT+WSBM	✗	token

Why OP?

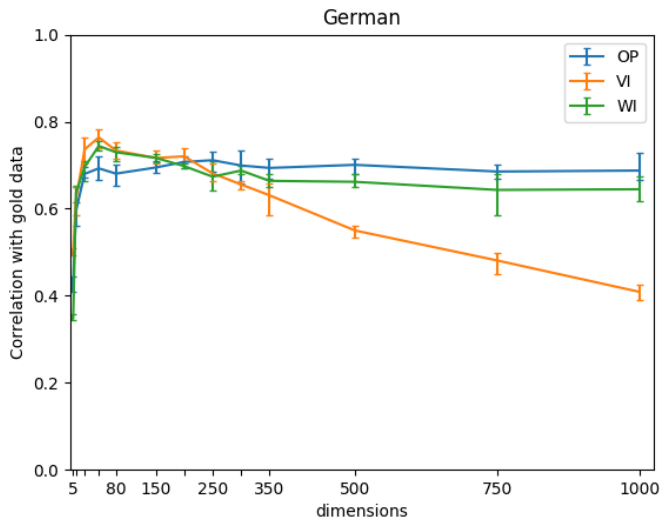


Figure 11: Experiments from Kaiser et al. (2020).

Why OP?

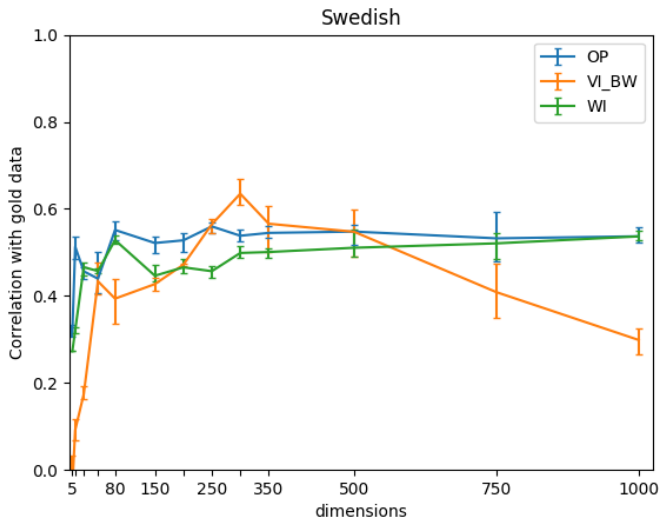


Figure 12: Experiments from Kaiser et al. (2020).

Bibliography

- Bansal, N., Blum, A., & Chawla, S. (2004). Correlation clustering. *Machine Learning*, 56(1-3), 89–113. doi: 10.1023/B:MACH.0000033116.57574.95
- Basile, P., Caputo, A., Caselli, T., Cassotti, P., & Varvara, R. (2020). Overview of the EVALITA 2020 Diachronic Lexical Semantics (DIACR-Ita) Task. 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.
- Erk, K., McCarthy, D., & Gaylord, N. (2013). Measuring word meaning in context. *Computational Linguistics*, 39(3), 511–554.
- 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>
- OED. (2009). *Oxford english dictionary*. Oxford University Press.
- Paul, H. (2002). *Deutsches Wörterbuch: Bedeutungsgeschichte und Aufbau unseres Wortschatzes* (10. ed.). Tübingen: Niemeyer.
- 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.
- 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.
- Svenska Akademien. (2009). *Contemporary dictionary of the Swedish Academy*. The changed words are extracted from a database managed by the research group that develops the Contemporary dictionary.