University of Stuttgart
Germany

Lexical Semantic Change Discovery

July 7, 2021

Sinan Kurtyigit[♠] Maike Park[♡] Dominik Schlechtweg[♠]
Jonas Kuhn[♠] Sabine Schulte im Walde[♠]
[♠]Institute for Natural Language Processing, University of Stuttgart, Germany
[♡]Leibniz Institute for the German Language, University of Mannheim, Germany

IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Introduction

- Most work in Lexical Semantic Change Detection (LSCD) focuses on developing and analysing models.
- Limited focus on discovering novel instances of semantic change.

We propose a **shift of focus to change discovery.**

Introduction

In this work we

- .. use high quality models to predict novel semantic changes.
- .. validate the model predictions through human annotation.
- .. discover novel instances of semantic changes.
- .. evaluate the usability of the approach from a lexicographers viewpoint.
- .. provide a highly automated framework.[1]

[1] The code is available at `https://github.com/seinan9/LSCDiscovery`

Lexical Semantic Change Discovery

Given a diachronic corpus pair (C_1, C_2), decide for the intersection of their vocabularies which words lost or gained sense(s) between C_1 and C_2.

Discovery Process

Given two corpora C_1 and C_2 from two time periods:

1. Generate word embeddings for words in vocabulary intersection.
2. Measure differences between word embeddings from C_1 and C_2.
3. Calculate a threshold. Mark words with a value greater than or equal to this threshold as changing.
4. Filter out undesirable words.

Approaches

Two approaches to generate graded values:

1. Type-based: **SGNS+OP+CD**
2. Token-based: **BERT+APD/COS**

Population

Generating word embeddings is expensive for token-based approach.

- Only consider a sample for the discovery.[2]
- Here a population of 500 words is used for both approaches.
- Population can be much larger in practice.

[2] This limitation is only necessary so we can experiment with different parameters.

Thresholding

According to the graded values a threshold is calculated:

$$TH = \mu + t \cdot \sigma,$$

where μ is the mean and σ standard deviation.
Words whose graded values are greater than or equal to this threshold, are labeled as changing.

Filtering

Two filters are provided to remove undesirable words:

1. A lemma-level filter.
2. A usage-level filter.

Annotation

The model predictions are validated by human annotation:

1. Usages are uploaded to the DURel interface for annotation and visualization.[3]
2. Annotators judge the semantic relatedness of pairs of word usages.[4]

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

Table 1: DURel relatedness scale.

[3] https://www.ims.uni-stuttgart.de/data/durel-tool

[4] https://www.ims.uni-stuttgart.de/data/wugs

Word Usage Graphs (WUGs)

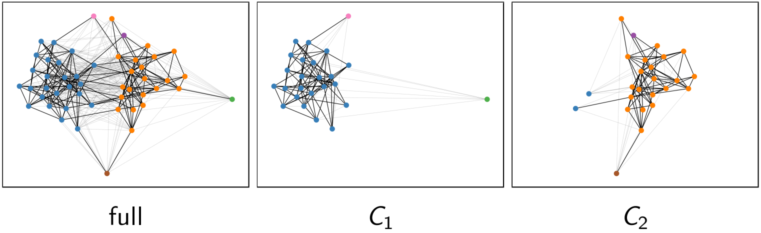

Figure 1: Word Usage Graph of German *Aufkommen* (left), subgraphs for first time period C_1 (middle) and for second time period C_2 (right). **black**/gray lines indicate **high**/low edge weights.

Data

German data set provided by SemEval-2020 shared task:

- Two time-specific Corpora C_1 (DTA, 1800–1899) und C_2 (BZ+ND 1946–1990).
- 48 target words.
- Binary und graded gold data for evaluation and tuning.

Tuning

Solve the SemEval-2020 subtasks to find good parameters:

1. Subtask 2 is solved to optimize the graded value predictions.
2. Afterwards, Subtask 1 is solved to find the best-performing threshold
3. The best parameter configuration for both models are then used to discover changing words.

Predictions

Three sets of predictions:

1. Discovered with type-based approach.
2. Discovered with token-based approach.
3. Randomly sampled from population.

All three sets are annotated and evaluated separately.

Results

Approach	$\sum$	$+$	-	$F_{0.5}$
type-based	27	18 / 67%	9 / 33%	.714
token-based	30	17 / 57%	13 / 43%	.620
random	30	10 / 34%	20 / 66%	.349

Table 2: Number of total/correct/false predictions and $F_{0.5}$-performance for type-based approach, token-based approach and random baseline.

Error Sources

1. **Context Change**: Words where the context in the usages shifts between C_1 and C_2, e.g., *Angriffswaffe* ('offensive weapon'), *aussterben* ('to die out') and *Königreich* ('kingdom').
2. **Context Variety**: Word that can be used in a large variety of contexts, e.g., *neunjährig* ('9-year-old'), *vorjährig* ('of the previous year') and *Bemerken* ('notice').

WUG - Angriffswaffe

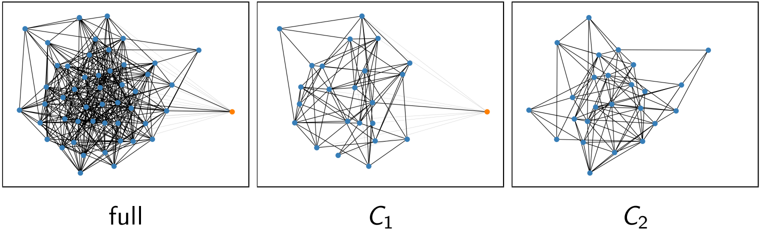

Figure 2: Word Usage Graph of German *Anriffswaffe* (left), subgraphs for first time period C_1 (middle) and for second time period C_2 (right).

Lexicographical Evaluation

- Annotation process can ensure more objective analysis of corpus data.
- Visualization is helpful for analysing purposes.
- Model predictions are promising candidates.

Records of Novel Senses

Comparing 21 correct predictions to existing dictionary contents:

- In most cases, all senses identified by the system are included in a dictionary.
- In 4 cases, at least one novel sense is not included.

A Novel Sense

1. Man sieht also, daß die Striche nach den Tausenden, nach den Hunderten und nach den **Zehnern** gesetzt werden.
   '*So you can see that the strokes are placed after the thousands, after the hundreds, and after the* ***tens****.*'
2. Fußball-Toto : Kein Elfer ; 6 **Zehner** mit je 3778 Mark ; 152 Neuner mit je 298 Mark.
   '*Soccer lottery : No eleven ; 6* ***tens*** *with 3778 marks each ; 152 nines with 298 marks each.*'

WUG - Zehner

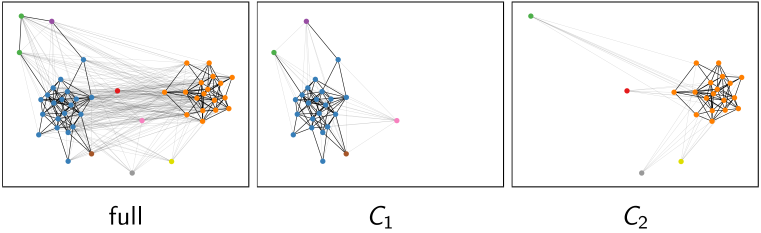

Figure 3: Word Usage Graph of German *Zehner* (left), subgraphs for first time period C_1 (middle) and for second time period C_2 (right).

Conclusion

- We used two LSCD approaches to discover semantic changes in a German corpus pair.
- Both approaches were able to discover semantic changes.
- Validated results through human annotation.
- Provided convenient visualization through Word Usage Graphs.
- Further validated the usefulness from a lexicographers viewpoint.

End

Thank you for your attention.