University of Stuttgart
Germany

# State-of-the-art models in lexical semantic change detection

January 19, 2021

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

# Introduction

- **topic**: Lexical Semantic Change Detection (LSCD)

→ detect sense-divergences for a word over time in textual data

- why interesting?
  - main use: support historical semanticists to find semantic changes (more and faster)
- why text?
  - in many cases only historical language data available
  - relatively cheap resource
  - shown to encode parts of word meaning (Turney & Pantel, 2010)
- what is the current state-of-the-art?

# Tasks

- SemEval 2020 Task 1 on Unsupervised Lexical Semantic Change Detection (Schlechtweg, McGillivray, Hengchen, Dubossarsky, & Tahmasebi, 2020)[1]
- comparison of two time periods $t_1$ and $t_2$
- two tasks:
    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.
    2. **Ranking**: rank a set of target words according to their sense-frequency divergence between $t_1$ and $t_2$.
- defined on **word sense frequency distributions**

[1] https://languagechange.org/semeval/

# Sense Frequency Distributions

| | $t_1$ | | | $t_2$ | | |
|---|---|---|---|---|---|---|
| **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.

# Data

- for each language
  - 2 corpora (one for each time period)
  - set of target words
  - binary and graded labels for target words
    - $\rightarrow$ derived from sense-frequency distributions
    - $\rightarrow$ derived from graded use pair judgments of human annotators

# Corpora

| | $t_1$ | $t_2$ |
|---|---|---|
| **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.

# Target words

- 100–200 changing words selected from etymological dictionaries (OED, 2009; Paul, 2002; Svenska Akademien, 2009)
- adding of control words with similar frequency properties
- sample 100 uses (30 for Latin) of each word per time period

# Labels

- obtain SFDs of corpus samples by annotation
- graded word sense annotation (Erk, McCarthy, & Gaylord, 2013)

# Diachronic Data

| | | |
|---|---|---|
| (1) | 1830 | but I am bound and thrown into a dark prison **cell** in Newgate jail. |
| (2) | 1851 | I had to destroy all the letters in my **cell** when I left the prison. |
| | ... | |
| (3) | 1990 | I call it my prison **cell**, this dark chamber. |
| (4) | 2006 | She grabbed her **cell** and started a call as we headed toward the door. |

# Use Pair Combinations

| Use 1 | Use 2 | relatedness judgment |
|---|---|---|
| (1) | (2) | ? |
| (1) | (3) | ? |
| (1) | (4) | ? |
| (2) | (3) | ? |
| | ... | |

Table 2: Use Pair Combinations.

# Use Pair Judgments

| Use 1 | Use 2 | relatedness judgment |
|---|---|---|
| (1) | (2) | 4 |
| (1) | (3) | 4 |
| (1) | (4) | 1 |
| (2) | (3) | 4 |
| | ... | |

Table 3: Use Pair Combinations.

# Scale

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

0: Cannot decide

# Word Usage Graphs (WUGs)

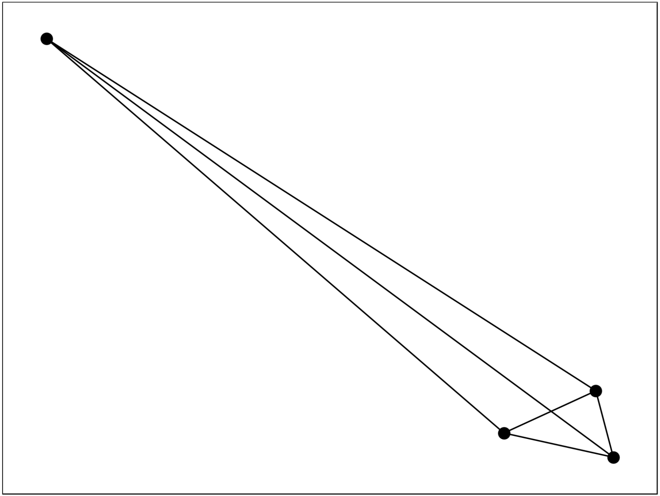


Figure 2: Graph visualization four uses of *cell*.

# Word Usage Graphs (WUGs)

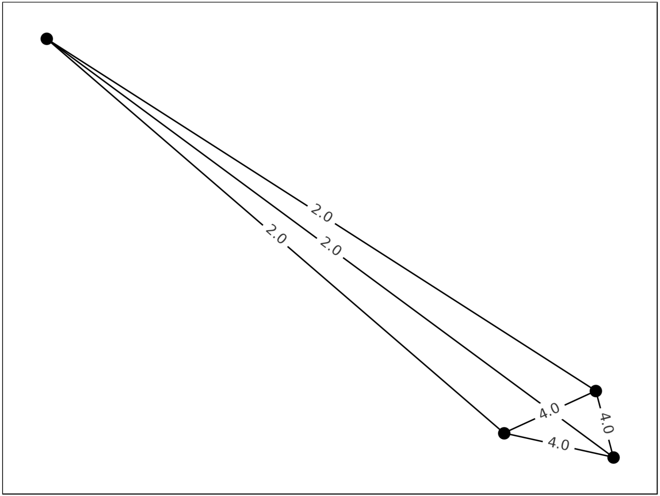



Figure 3: Graph visualization four uses of *cell*.

# Clustering

- correlation clustering (Bansal, Blum, & Chawla, 2004)
- optimization criterion: **reduce (weighted) number of cluster-edge conflicts**

$$L(C) = \sum_{e \in \phi_{E,C}} W(e) + \sum_{e \in \psi_{E,C}} |W(e)| \qquad (1)$$

# Clustering

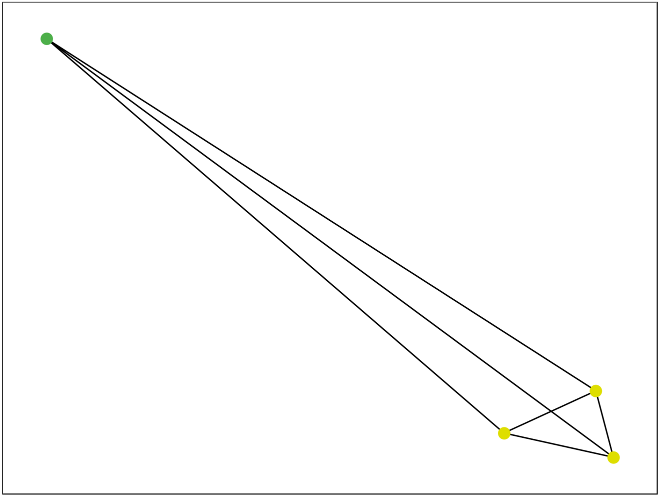
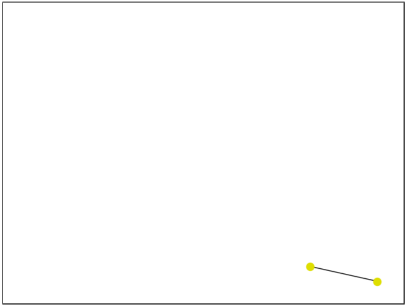
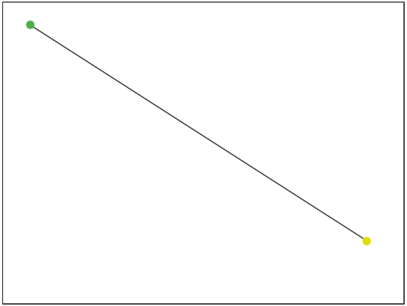


Figure 4: Graph visualization for uses of *cell* $D = (3, 1)$.

# Clustering



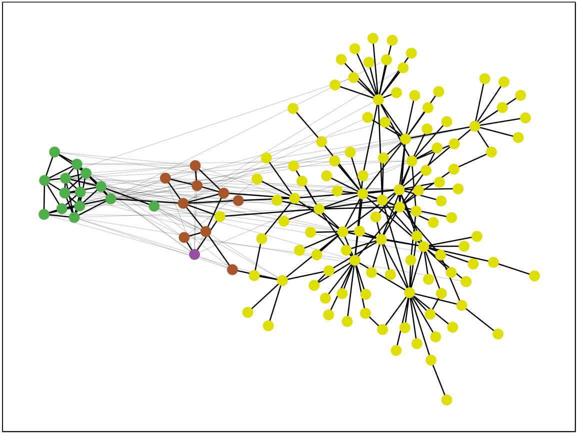
$D_1 = (2, 0)$



$D_2 = (1, 1)$

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

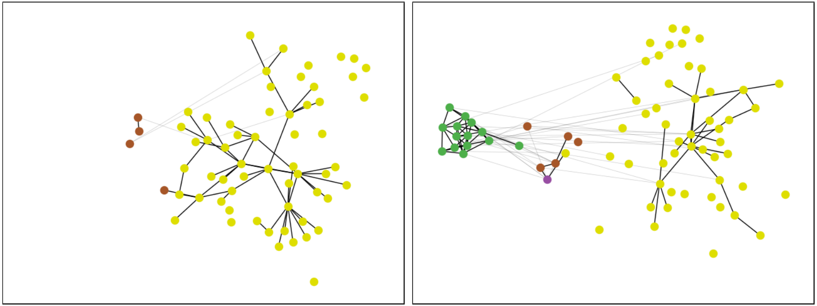
# SemEval WUG



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

Figure 6: Usage graph of Swedish *ledning*.

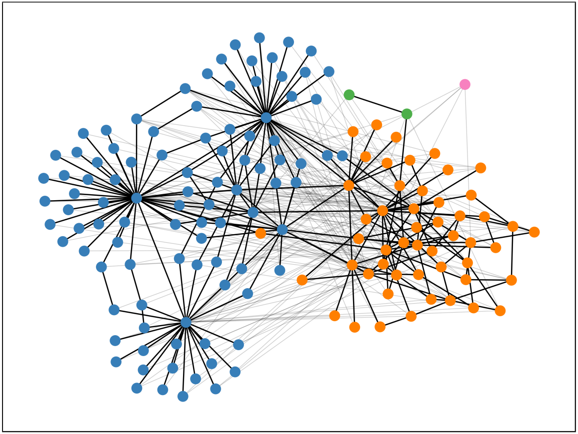
# SemEval WUG



$D_1 = (58, 0, 4, 0)$ $D_2 = (52, 14, 5, 1)$

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

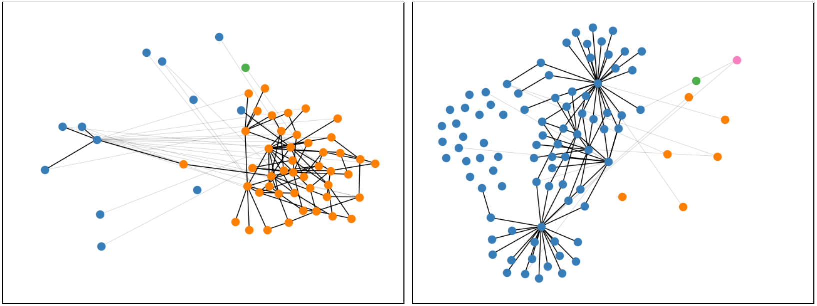
# SemEval WUG



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

Figure 8: Usage graph of German *Eintagsfliege*.

# SemEval WUG



$D_1 = (12, 45, 0, 1)$ $D_2 = (85, 6, 1, 1)$

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

# Models

- unsupervised (no labeled training data)
- distributional (Harris, 1954)
- vector space models
- mostly bag-of-words-based
- most successful ones are neural language models

# Type-based VSMs

- do not model senses
- one average vector per word
- composed by
    1. semantic representation per word (type vector)
    2. alignment
    3. measure

# Simple Model

- co-occurrence count model

# Corpus

| | | |
|---|---|---|
| (1) | 1830 | but I am bound and thrown into a dark prison **cell** in Newgate jail. |
| (2) | 1851 | I had to destroy all the letters in my **cell** when I left the prison. |
| | ... | |
| (3) | 1990 | I call it my prison **cell**, this dark chamber. |
| (4) | 2006 | She received a call on her **cell** as we headed toward the door. |

# Preprocess

| | | |
|---|---|---|
| (1) | 1830 | bound thrown dark prison **cell** jail |
| (2) | 1851 | destroy letters **cell** left prison |
| | ... | |
| (3) | 1990 | call prison **cell** dark chamber |
| (4) | 2006 | received call **cell** headed door |

# Finding Context (Bags of Words)

(1) 1830 bound thrown **dark prison cell jail**

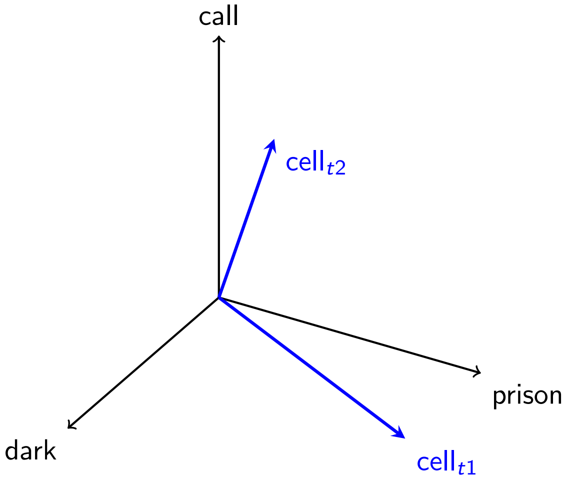
(2) 1851 **destroy letters cell left prison**

...

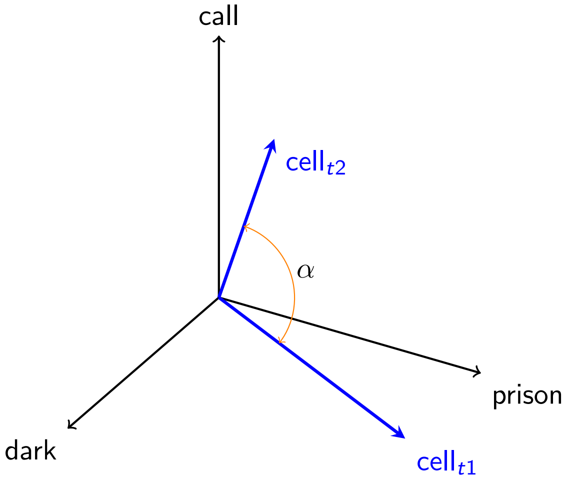
(3) 1990 **call prison cell dark chamber**

(4) 2006 **received call cell headed door**

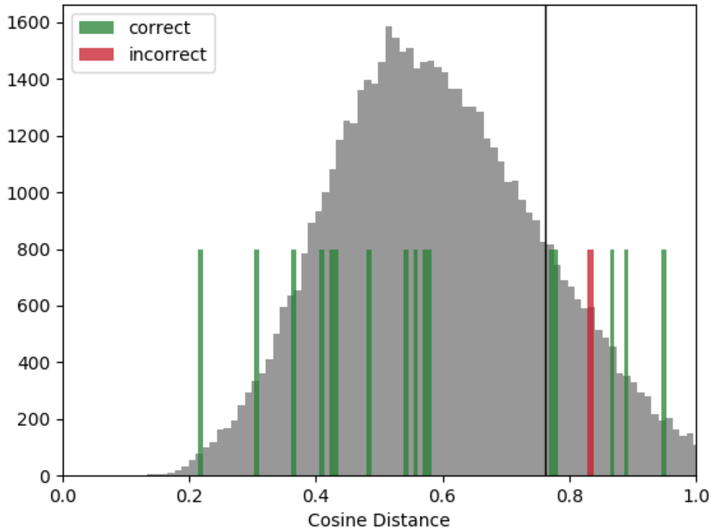
# Vector Space Representation



# Cosine Distance



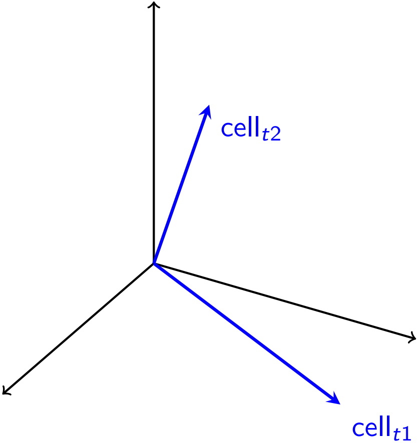
# Thresholding



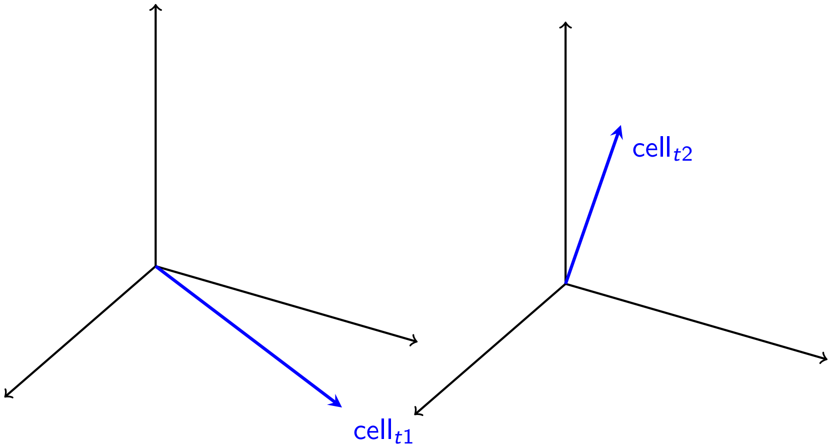
# Best Models

- neural-network-based language models
- trained on context word prediction
- compresses contextual information into low-dimensional vectors
- SGNS+OP+CD (Hamilton, Leskovec, & Jurafsky, 2016)
  1. **Semantic Representation**: Skip-gram with Negative Sampling (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013)
  2. **Alignment**: Orthogonal Procrustes (Schönemann, 1966)
  3. **Change Measure**: Cosine Distance (Salton & McGill, 1983)

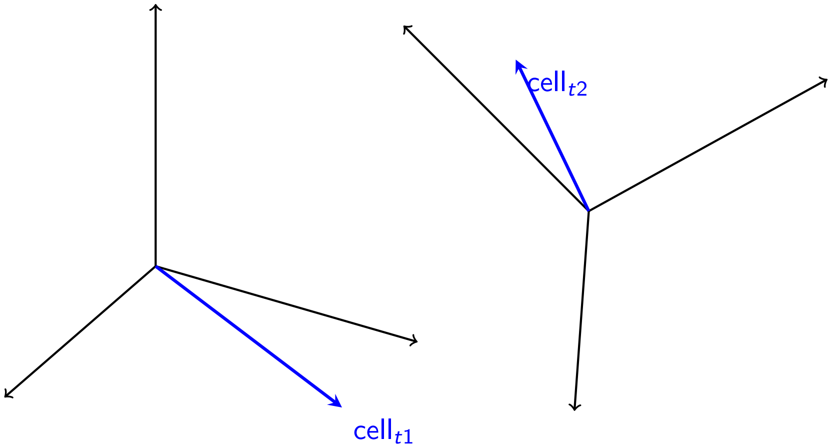
# Non-interpretable dimensions



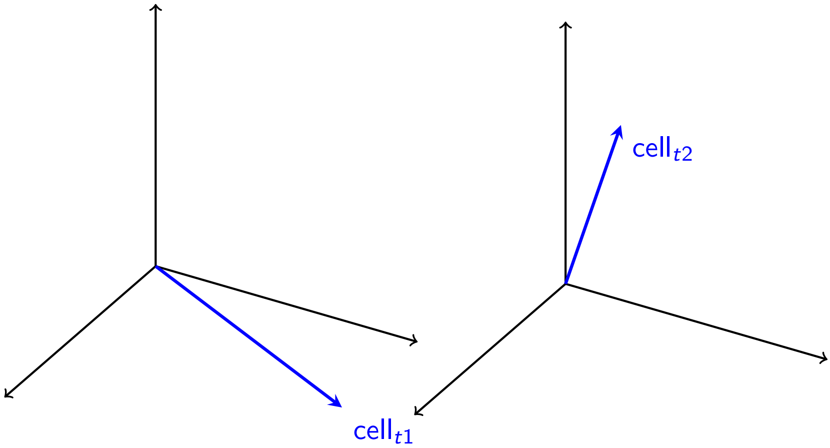
# Alignment



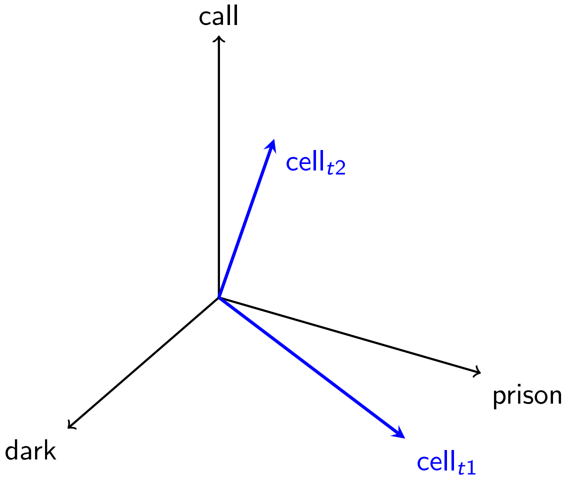
# Alignment



# Alignment



# Common Space



# Token-based VSMs

- word sense discrimination (Schütze, 1998)
- model the human measurement process (one meaning per use, semantic proximity, clustering)
- one vector per use
- composed by
  1. semantic representation per word use (token vector)
  2. (clustering)
  3. change measure

# Simple Model

- co-occurrence count model

# Finding Context (Bags of Words)

(1) 1830 bound thrown **dark prison cell jail**

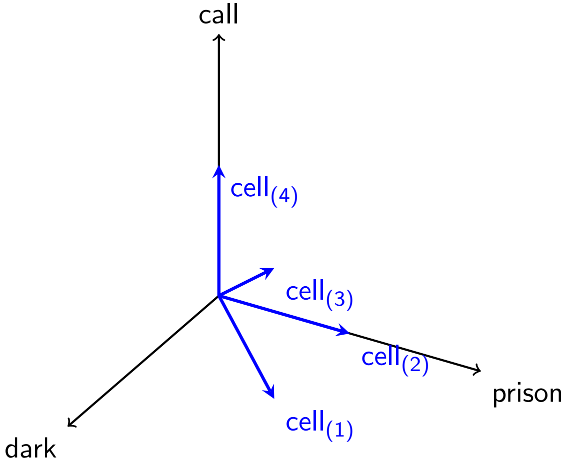
(2) 1851 **destroy letters cell left prison**

...

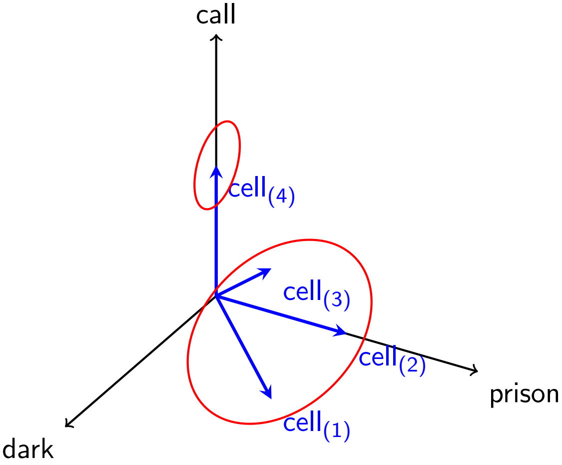
(3) 1990 **call prison cell dark chamber**

(4) 2006 **received call cell headed door**

# Vector Space Representation



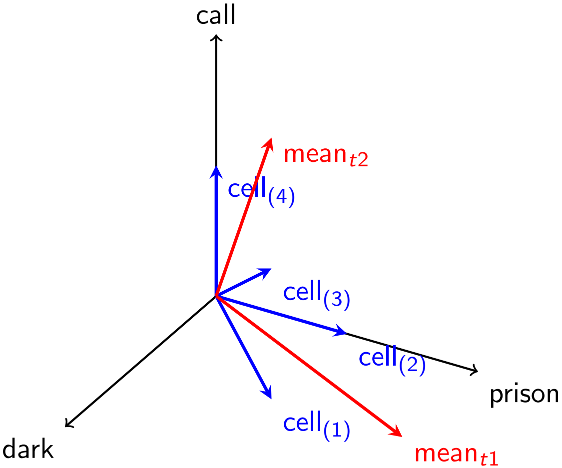
# Clustering



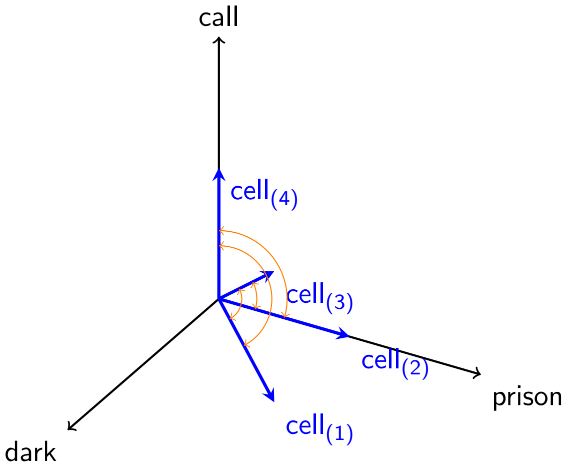
# Sense Frequency Distribution

| | $t_1$ | | $t_2$ | |
|---|---|---|---|---|
| **Senses** | **Chamber** | **Phone** | **Chamber** | **Phone** |
| # uses | 2 | 0 | 1 | 1 |

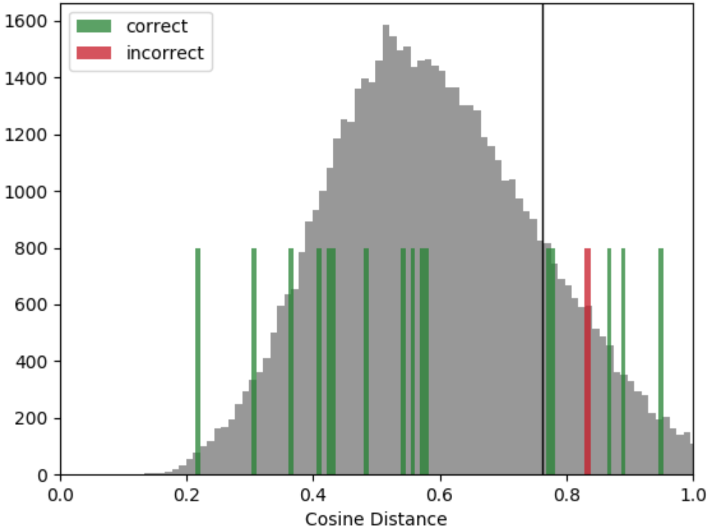
# cos



# APD



# Thresholding



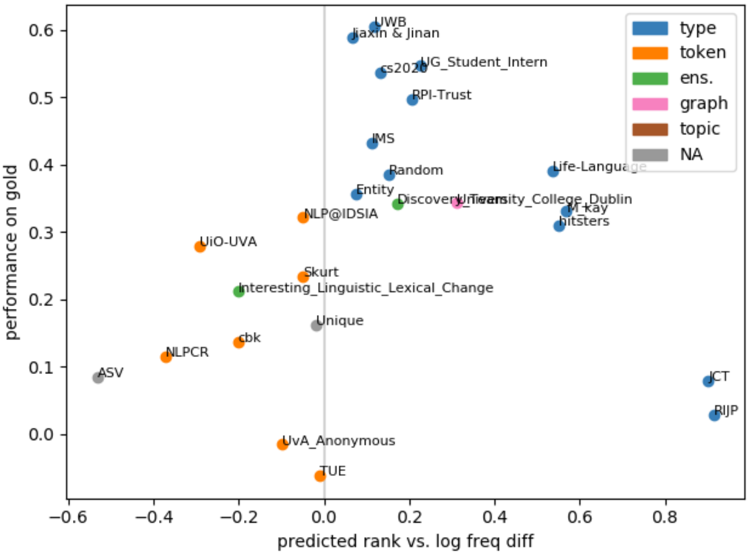
# Best Models

- neural-network-based language models
- contextualized embeddings (Devlin, Chang, Lee, & Toutanova, 2019; Peters et al., 2018)
- trained on context-sensitive context word prediction
- token representations in multiple layers
- pre-trained on modern data
- or perform type embeddings on various modern downstream tasks
- we use
  1. **Semantic Representation**: BERT (Peters et al., 2018)
  2. **Change Measure**: COS/APD/$APD_{norm}$ (Beck, 2020; Kutuzov & Giulianelli, 2020)

# Evaluation

- we now compare the human and computational measurements of change on our data sets
- two tasks:
  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.
  2. **Ranking**: rank a set of target words according to their sense-frequency divergence between $t_1$ and $t_2$.

# SemEval Results (Ranking)



# Summary

- token embeddings are dominated by type embeddings
- SGNS+OP+CD is the overall dominant model
- averaging token embeddings works much better than clustering
- models show medium to high performance (depending on the task and tuning data)

$\rightarrow$ currently it is better not to model the human annotation process

# BERT performance on German

| | DE | prep | Reference |
|---|---|---|---|
| **BERT+APD**$_{norm}$ | .41 | lemma | (Beck, 2020) |
| **BERT+CL+JSD** | .53 | lemma | (Martinc, Montariol, Zosa, & Pivovarova, 2020) |
| **BERT+COS** | .58 | lemma | (Kutuzov & Giulianelli, 2020) |

Table 4: Token embedding performance on ranking task.

# What blocks BERT's performance?

- what do clusters reflect?
  1. sentence position
  2. number of proper names (Martinc et al., 2020)
  3. corpus
  4. word form
- joint work with Severin Laicher and Sinan Kurtyigit

# Cluster bias

| | 1 | 12 | 1+12 | 1+2+3+4 | 9+10+11+12 |
|---|---|---|---|---|---|
| **Position Influence** | .631 | .497 | .629 | .633 | .512 |
| **Position Random** | .384 | .383 | .383 | .382 | .383 |
| **Position Baseline** | .712 | .712 | .712 | .712 | .712 |
| **Name Influence** | .535 | .476 | .538 | .537 | .485 |
| **Name Random** | .378 | .378 | .375 | .381 | .379 |
| **Name Baseline** | .602 | .602 | .602 | .602 | .602 |
| **Corpora Influence** | **.538** | **.566** | **.550** | **.547** | **.564** |
| **Corpora Random** | .531 | .527 | .526 | .531 | .528 |
| **Corpora Baseline** | .522 | .522 | .522 | .522 | .522 |
| **Form Influence** | **.945** | **.667** | **.917** | **.922** | **.670** |
| **Form Random** | .478 | .481 | .481 | .483 | .477 |
| **Form Baseline** | .611 | .611 | .611 | .611 | .611 |

Table 5: ACC scores for influencing factors: English BERT-cased.

# Cluster bias

| | 1 | 12 | 1+12 | 1+2+3+4 | 9+10+11+12 |
|---|---|---|---|---|---|
| **Position Influence** | .549 | .586 | .582 | .571 | .590 |
| **Position Random** | .397 | .403 | .401 | .402 | .396 |
| **Position Baseline** | .670 | .670 | .670 | .670 | .670 |
| **Corpora Influence** | **.613** | **.669** | **.645** | **.633** | **.665** |
| **Corpora Random** | .529 | .528 | .525 | .525 | .530 |
| **Corpora Baseline** | .560 | .560 | .560 | .560 | .560 |
| **Form Influence** | **.775** | **.705** | **.774** | **.770** | **.722** |
| **Form Random** | .278 | .278 | .276 | .282 | .285 |
| **Form Baseline** | .490 | .490 | .490 | .490 | .490 |

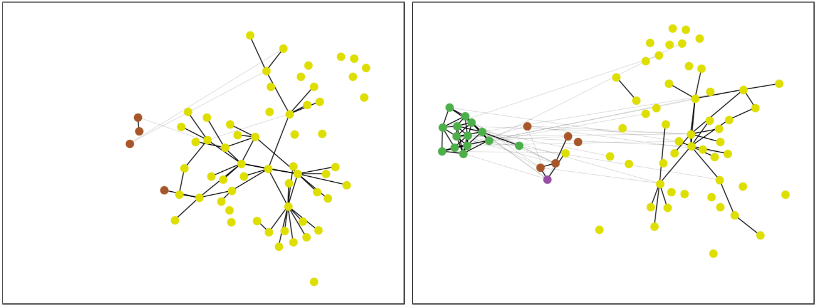
Table 6: ACC scores for influencing factors: German BERT-cased.

# Tuning

| | DE | prep | Reference |
|---|---|---|---|
| **BERT+APD**$_{norm}$ | .41 | lemma | (Beck, 2020) |
| **BERT+CL+JSD** | .53 | lemma | (Martinc et al., 2020) |
| **BERT+COS** (12) | .58 | lemma | (Kutuzov & Giulianelli, 2020) |
| **BERT+COS** (9-12) | .47 | token | |
| **BERT+COS** (9-12) | .69 | lemma | |
| **BERT+COS** (9-12) | .72 | toklem | |
| **ELMo+COS** (12) | .74 | lemma | (Kutuzov & Giulianelli, 2020) |
| **BERT+APD**$_{norm}$ (1+12) | **.83** | toklem | (uses only) |

Table 7: Token embedding tuning on ranking task.

# Polysemy



$D_1 = (58, 0, 4, 0)$ $D_2 = (52, 14, 5, 1)$

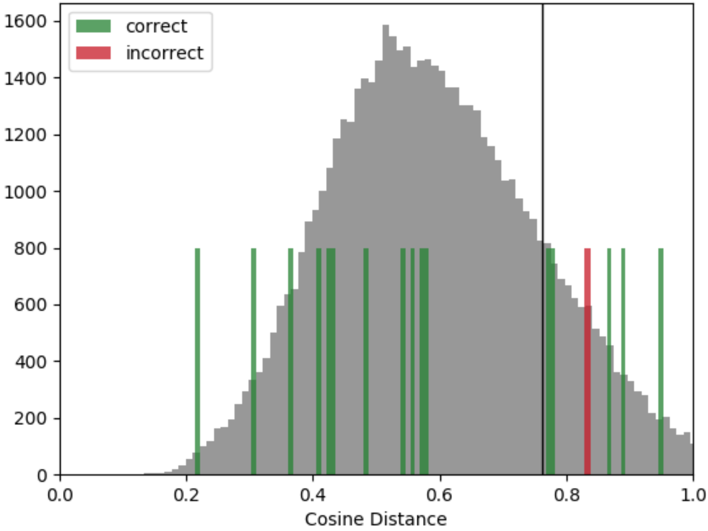
Figure 14: Usage graph of Swedish *ledning*.

# Polysemy

| | EN | DE | SV |
|---|---|---|---|
| **BERT+APD** | **.55**/.45 | .69/.72 | **.65**/.60 |
| **BERT+APD**$_{norm}$ | **.49**/.41 | .74/.83 | **.50**/.48 |
| **BERT+COS** | .09/.19 | .60/.72 | **.12**/.08 |

Table 8: Correlation of true polysemy in $t_1$ vs. true change with predicted change scores by different models (toklem, 1+12).

# Predict



correct
incorrect
1600
1400
1200
1000
800
600
400
200
0
0.0
0.2
0.4
0.6
0.8
1.0
Cosine Distance

# Predict

| | DE | Predict |
|---|---|---|
| **BERT+COS** | .74 | .62 |
| **SGNS+OP+CD** | .74 | **.75** |

Table 9: Results of prediction on German SemEval data for best type and token model (toklem, 1+12).

# Conclusion

- token and type embeddings perform similarly when tuned on the test data
- lemmatizing only the target word for BERT strongly improves performance
- BERT is strongly influenced by surface form of target word (especially in lower layers)
- BERT is moderately influenced by corpus bias
- polysemy often explains BERT performance better than change
- polysemy-controlled change measures still suffer from this
- COS is least influenced by polysemy
- type embeddings outperform embeddings clearly on a prediction task

# Future Research

- clustering
- supervised LSCD
  - learn binary classifier on (e.g. concatenation) of vectors
  - follow hypernymy detection (Shwartz, Santus, & Schlechtweg, 2017)
  - problem: training data

# Bibliography I


Bansal, N., Blum, A., & Chawla, S. (2004). Correlation clustering. *Machine Learning*, *56*(1-3), 89–113. doi: 10.1023/B:MACH.0000033116.57574.95

Beck, C. (2020). DiaSense at SemEval-2020 Task 1: Modeling sense change via pre-trained BERT embeddings. In *Proceedings of the 14th international workshop on semantic evaluation.* Barcelona, Spain: Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. doi: 10.18653/v1/N19-1423

Erk, K., McCarthy, D., & Gaylord, N. (2013). Measuring word meaning in context. *Computational Linguistics*, *39*(3), 511–554.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1489–1501). Berlin, Germany.

Harris, Z. S. (1954, 08). Distributional structure. *Word*, *10*, 146-162.

Kutuzov, A., & Giulianelli, M. (2020). UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection. In *Proceedings of the 14th international workshop on semantic evaluation.* Barcelona, Spain: Association for Computational Linguistics.

Martinc, M., Montariol, S., Zosa, E., & Pivovarova, L. (2020). Discovery Team at SemEval-2020 Task 1: Context-sensitive Embeddings not Always Better Than Static for Semantic Change Detection. In *Proceedings of the 14th international workshop on semantic evaluation.* Barcelona, Spain: Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In Y. Bengio & Y. LeCun (Eds.), *1st international conference on learning representations, ICLR 2013, scottsdale, arizona, usa, may 2-4, 2013, workshop track proceedings.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS.*

OED. (2009). *Oxford english dictionary*. Oxford University Press.

Paul, H. (2002). *Deutsches Wörterbuch: Bedeutungsgeschichte und Aufbau unseres Wortschatzes* (10. ed.). Tübingen: Niemeyer.

# Bibliography II


Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 2227–2237). New Orleans, LA, USA.

Salton, G., & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw - Hill Book Company.

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.

Schönemann, P. H. (1966, Mar 01). A generalized solution of the orthogonal procrustes problem. *Psychometrika*, *31*(1), 1–10.

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).

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.

Turney, P. D., & Pantel, P. (2010, January). From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, *37*(1), 141–188.