清華大學 Tsinghua University

Paper Sharing

Zhu Liu

2024.11.21

Two papers

- To Word Senses and Beyond: Inducing Concepts with Contextualized Language Models (EMNLP'24)
- Patterns of Lexical Ambiguity in Contextualised Language Models (EMNLP'21)

Lexical semantics in EMNLP'24

- Using Language Models to **Disambiguate Lexical** Choices in Translation
- FOOL ME IF YOU CAN! An Adversarial Dataset to Investigate the Robustness of LMs in **Word Sense Disambiguation**
- Can Large Language Models Faithfully Express Their Intrinsic **Uncertainty** in **Words**?
- Statistical **Uncertainty** in Word **Embeddings**: GloVe-V
- Automatically Generated Definitions and their utility for Modeling **Word Meaning**
- Encourage or Inhibit **Monosemanticity**? Revisit Monosemanticity from a Feature Decorrelation Perspective

To Word Senses and Beyond: Inducing Concepts with Contextualized Language Models

Bastien Liétard and **Pascal Denis** and **Mikaela Keller**
University of Lille, Inria, CNRS, Centrale Lille,
UMR 9189 - CRIStAL, F-59000 Lille, France
first_name.last_name@inria.fr

EMNLP'24

Background

- Multiplicity of meanings and forms
  - polysemy vs. synonymy
  - word-centric: WSD (disambiguation) & WSI (Word Sense Induction)
  - meaning-centric?
- from WSI to CI (Concept Induction)
  - Concepts: sense/synsets/meaning
  - Induction: in an unsupervised manner
  - Soft clustering: one word can be assigned to different concept labels
  - polysemy: intra-clusters; synonym: inner-clusters

Contributions

- Bi-level clustering
- Evaluation of the clustering
  - Intrinsic: Compared to WordNet synsets
  - Extrinsic: WiC

Related Work

- lexical resources for concepts: WordNet
- Word senses with language models: WSD & WSI
  - embedding + clustering
  - Substitute tokens from LM (Amrami and Goldberg, 2019; Eyal et al., 2022)
- Structures of Meaning in CLM
  - Polysemy vs. homonymy Haber and Poesio (2024)
  - hypernymy vs. hoponyms Hanna and Mareˇcek (2021)
  - Automatic WordNet for Filipino Velasco et al. (2023)
  - Other structure: Layer distributions E-2019/CE-2020
  - Types of polysemy pattern

Notations

1. Sense is word specific: One sense maps only one word
2. An occurrence only maps to one sense/concept; not vice versa
3. Sense is the least nodes for concepts (Only be merged, not splitted)
4. Occurences corresponding to one sense maps only one concept
5. Different senses for one word must be different concepts

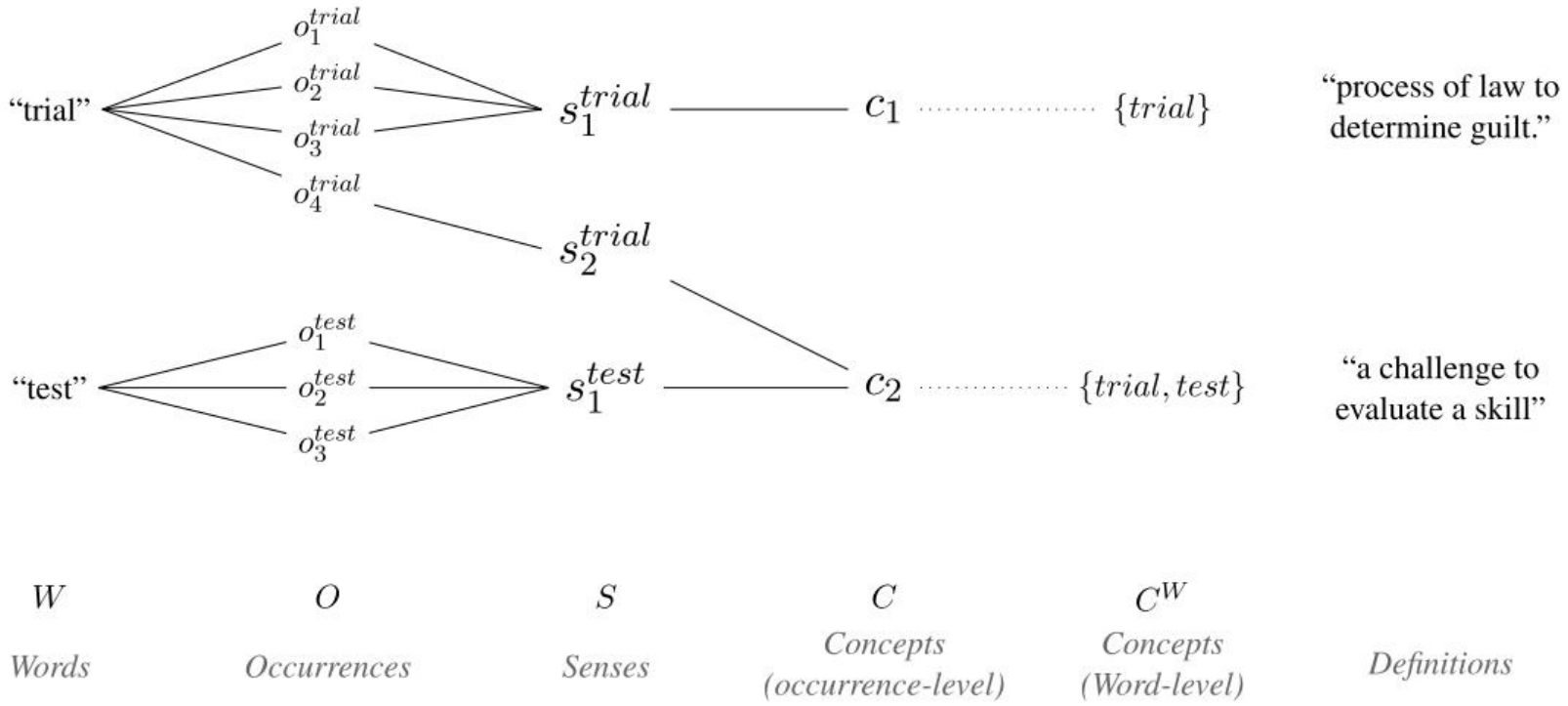

Figure 1: Illustration of our framework. The words "trial" is polysemous and has two senses corresponding to two different concepts, and is synonym with "test" for this second meaning.

Methods

- Bi-level method
  - Local (lemma-centric) clustering: WSI for each word
  - Global (cross-lexicon) clustering: Clustering using the averaged local representations
  - Error by local
- Local only
- Global only

Experiments

- Data
  - SemCor annotated by Wordnet
  - 1560 lemmas; 52997 occurrences; 3855 concepts (synsets)
  - Full data vs. synsets with multiple words
- Model
  - BERT-large + averaged four layers
- Clustering Algorithms
  - Kmeans; Agglomerative clustering
- Evaluation
  - BCubed metrics motivated by WSI: Precision; Recall; F1
- Baselines
  - Lemmas vs. Oracle WSI

Results

- Bi-level > Global
- Agglo > Kmeans
- Global > Local > baseline (In Synon.)

Concept Induction	Full data P	Full data R	Full data F_1	Synon. P	Synon. R	Synon. F_1
Baselines						
Lemmas	1.0	.43	.61	1.0	.61	.50
Oracle WSI	1.0	.75	.86	1.0	.39	.56
Local-only Systems						
Kmeans Local	.73	.70	.71	.67	.38	.49
Agglo Local	.92	.53	.67	.92	.35	.50
Eyal et al. (2022)	.31	.75	.44	.37	.39	.38
CI Systems						
Kmeans Global	.48	.65	.56	.68	.54	.60
Kmeans Bi-level	.70	.59	.64	.82	.47	.59
Agglo Global	.61	.60	.60	.82	.50	.62
Agglo Bi-level	.75	.60	.66	.86	.49	.62

Table 1: Concept Induction BCubed Precision (P), Recall (R) and F_1 on the SemCor data averaged over 5 runs.

Human judgments

	Cluster size		
	2	3	4+
Nb. of annotated clusters	50	50	23
Category (% of annotated clusters)			
Synonyms	42	38	17
Near-synonyms	24	24	35
Related	26	36	48
Invalid	08	02	0

Table 2: Qualitative manual evaluation of obtained word clusters of size ≥ 2.

- Cluster size: number of words in a cluster
- S+N+R: highly correlated to humans

Helpful to WSI

	WSI F_1	ρ
Local-only Systems		
Kmeans Local	.61	NA
Agglo Local	.77	.04
Eyal et al. (2022)	.46	.51
CI Systems		
Kmeans Global	.76	.51
Kmeans Bi-level	.78	.30
Agglo Global	.80	.53
Agglo Bi-level	.80	.46

Table 3: WSI BCubed F_1 and sense number correlation coefficient ρ on SemCor full data. Not computed for Kmeans because the number of cluster is constant.

- Further merged by global clustering
- Correlation with sense number

Helpful to WiC

Model	Acc.
Eyal et al. (2022) (CBOW)	59.3
Eyal et al. (2022) (Skip-Grams)	**61.9**
Ours (Agglo global)	58.8
Ours (Agglo bi-level)	59.7

Table 4: Accuracy scores on the nouns of the WiC test dataset (Pilehvar and Camacho-Collados, 2019).

WiC: Binary classification
Distance from Concept vector

Conclusion

- A local and global complementary view
- Bi-level clustering
- Helpful to tasks of WSI and WiC

Patterns of Lexical Ambiguity in Contextualised Language Models

Janosch Haber and **Massimo Poesio**
Queen Mary University of London and The Alan Turing Institute
{j.haber|m.poesio}@qmul.ac.uk

EMNLP'21

Background

- Degraded word sense similarity
  - Linguistic observation: homonymy and polysemy
  - Correlation between linguistic classification and a similarity score
- Structured polysemic sense
  - Metonymic, metaphoric ...
  - Different patterns within polysemy (e.g., newspaper)
- Uniform treatment of polysemic sense
  - Generative Lexicon

- Similarity and acceptability using co-predication

Contributions

- A extended dataset built on their previous two work
  - More data (target words; contexts)
- Similarity pattern and polysemy types
- Correlation with human annotators
  - Within different categories and similarity patterns
  - Language models
- varying distances between polysemic word sense
- tentative evidence for similarity pattern within the same type

Dataset

- 10 types of logical metonymy, each with multiple words
- Each sense has two contexts

animal/meat: lamb, chicken, pheasant, seagull; **food/event**: lunch, dinner; **container-for-content**: glass, bottle, cup; **content-for-container**: beer, wine, milk, juice; **opening/physical**: window, door; **process/result**: building, construction, settlement; **physical/information**: book, record; **physical/information/organisation**: newspaper, magazine; **physical/information/medium**: CD, DVD; **building/pupils/directorate/institution**: school, university

Dataset

1a The newspaper fired its editor in chief.
1b The newspaper was sued for defamation.
2a The newspaper lies on the kitchen table.
2b The newspaper got wet from the rain.
3a The newspaper wasn't very interesting.
3b The newspaper is rather satirical today.

1ab The newspaper fired its editor in chief and was sued for defamation.

- also include homonymic alternations (e.g., magazine: print medium vs. storage type) but no extra info.
- Human labels pair-wise similarity and co-prediction scores

Results

- Similarity and acceptability score distributions for different types
- For different language models (Word2Vec; Elmo; Bert Base/Large)
- Similarity patterns
- Sense clustering

Human annotations

- 16.5 annotations per item
- IAA: 0.62 Kri
- Similarity
- Acceptability
- High disagreement for Polysemy (Bonferroni correction)

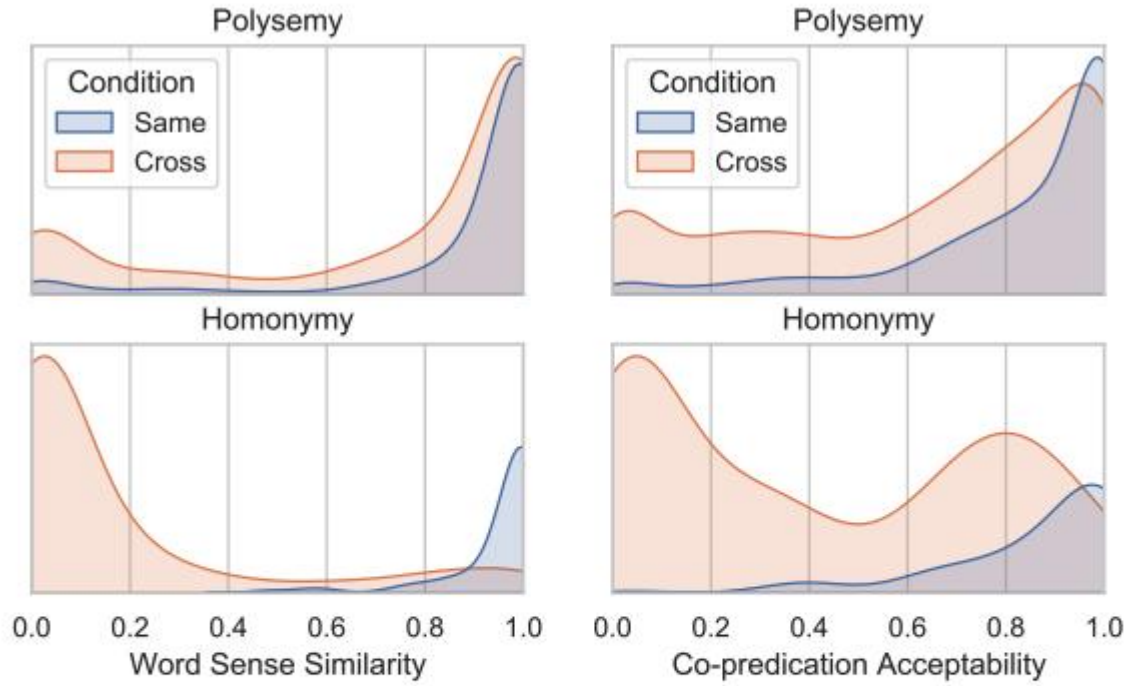

Figure 1: Distributions of explicit word sense similarity ratings and co-predication acceptability ratings given to same-sense (blue) and cross-sense (orange) samples with polysemic and homonymic alternations.

Computational Ratings

Measure	Same-Sense Pol.	Same-Sense Hom.	Same-Sense p	Cross-Sense Pol.	Cross-Sense Hom.	Cross-Sense p
Similarity	0.89	0.96	0.03	0.73	0.17	<0.05
Acceptability	0.83	0.86	0.10	0.64	0.41	<0.05
Word2Vec	0.60	0.65	0.12	0.55	0.58	0.06
ELMo	0.90	0.87	0.14	0.87	0.82	<0.05
BERT Base	0.91	0.93	0.22	0.88	0.78	<0.05
BERT Base (L4)	0.93	0.95	0.27	0.91	0.82	<0.05
BERT Large	0.79	0.85	0.15	0.72	0.44	<0.05
BERT Large (L4)	0.88	0.91	0.18	0.84	0.64	<0.05

Table 1: Word sense similarity distribution means for the different measures investigated in this study. p-values calculated through Mann-Whitney *U*.

- Same-S > Cross-Sense
- Cross-Sense: P > H (except for Word2Vec) human is more obvious
- Same-S: H > P (except for elmo)

Computational Ratings

Combination		Correlation		Ordinary Least Squares (OLS) Regression Analysis					
First Measure	Second Measure	r	p	Coef.	R^2	F-stat.	Prob.	Omnib.	Prob.
Similarity	Acceptability	0.698	1.09E-25	0.484	0.487	156.571	1.09E-25	9.733	0.008
Acceptability	Similarity	0.698	1.09E-25	1.005	0.487	156.571	1.09E-25	0.967	0.617
Word2Vec	Similarity	0.206	0.008	0.675	0.042	7.309	0.008	31.562	0
Word2Vec	Acceptability	0.311	4.39E-05	0.707	0.097	17.625	4.39E-05	9.668	0.008
ELMo	Similarity	0.515	1.11E-12	2.863	0.265	59.475	1.11E-12	10.43	0.005
ELMo	Acceptability	0.523	4.39E-13	2.018	0.273	61.973	4.39E-13	6.552	0.038
BERT Base	Similarity	0.641	1.02E-20	4.070	0.411	115.185	1.02E-20	3.496	0.174
BERT Base	Acceptability	0.560	3.43E-15	2.469	0.314	75.521	3.43E-15	2.07	0.355
BERT Large	Similarity	0.687	1.22E-24	2.181	0.472	147.361	1.22E-24	15.96	0
BERT Large	Acceptability	0.550	1.40E-14	1.212	0.302	71.520	1.40E-14	5.324	0.07

Table 2: Correlations between measures of contextualised word sense similarity. The first set of columns displays pairwise correlation based on Pearson's *r*, the second set shows the key statistics obtained from an OLS regression analysis. BERT results for summing over the last four hidden states.

R^2: Larger, more goodness-of-fit
contextual > statistic

Correlation

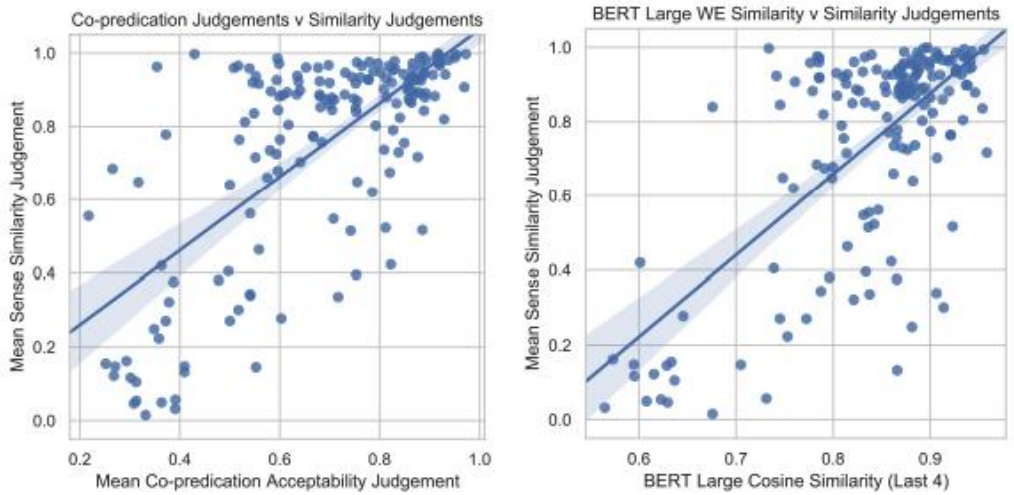

Figure 3: Correlations of co-predication v word sense similarity ratings (left) and BERT Large cosine similarity scores v word sense similarity ratings (right), together with the best linear fit. Scaling of x-axis adjusted for clarity. BERT results for summing over the last four hidden states.

Similarity patterns across words

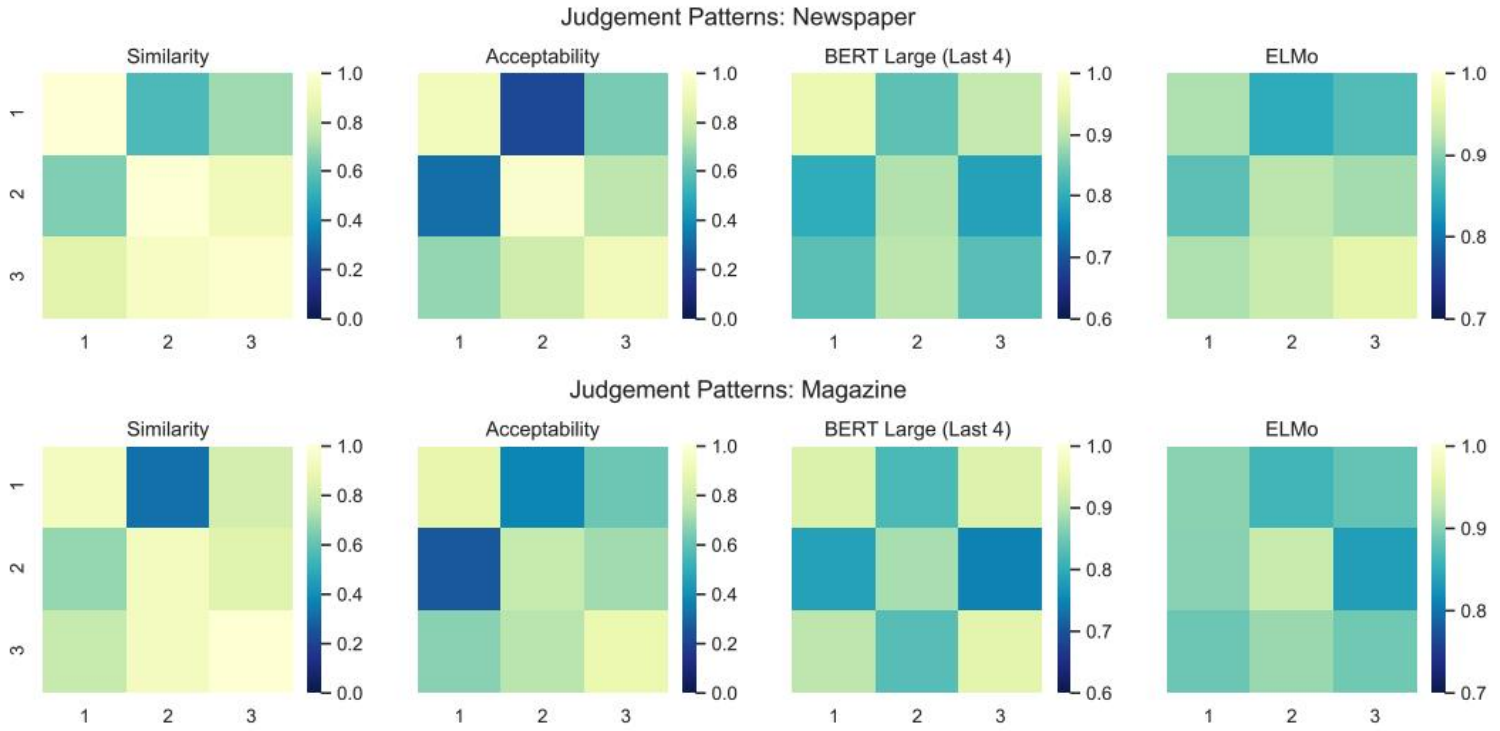

Figure 4: Similarity patterns in the sense similarity ratings for polysemes *newspaper* and *magazine*. Senses: 1-physical, 2-information, 3-organisation. Colour scales adjusted for computational measures.

Correlation between word pairs in the same type

sim: 0.89
co-prediciton: 0.95

Similarity patterns

Measure	Pairwise r	Pairwise p <0.05	Overall r	Overall p
Similarity	0.44	3/24 (12.5%)	0.53	8.260e-10
Acceptability	0.44	4/24 (16.7%)	0.62	5.306e-14
ELMo	0.14	0/24 (0%)	0.21	0.025
BERT Large	0.28	1/24 (4.2%)	0.27	0.003

Table 3: Mean Pearson correlation of polysemic word sense similarity patterns across different target words allowing the same alternation of senses, number of significant comparisons, and overall pattern correlation.

pairwise: for each type (same alternation of senses)
Overall: Concat each pair before the correlation.

Sense clustering

- Whether clustering is similar to the actual patterns?

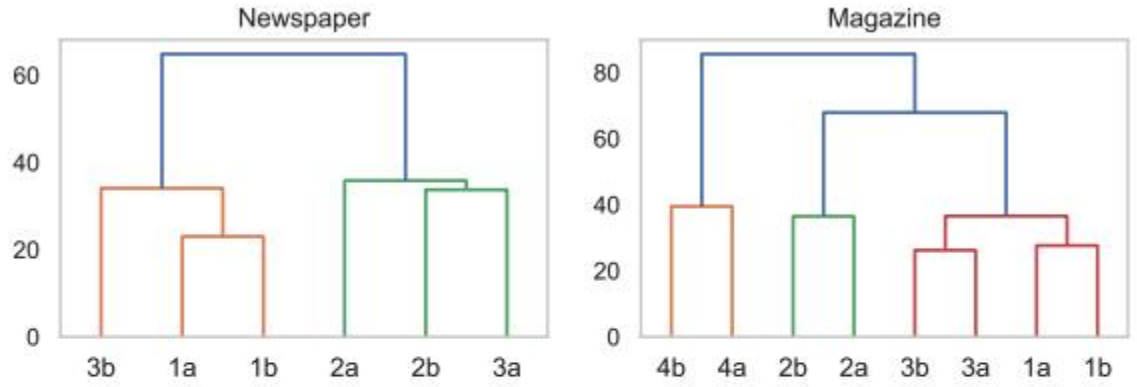

Figure 5: Dendograms of BERT Large contextualised embedding similarity for a selection of target words. Numbers indicate traditional sense distinctions.

Criterion	*t*	#C	NMI	F1	P	R
Inconsistency	<0.7	3.54	0.60	0.77	0.86	0.71
Distance	31	4.21	0.75	0.75	0.90	0.64

Table 4: Best-performing settings for inconsistency and distance-based hierarchical Ward clustering of target word senses. #C is the average number of clusters produced per target.

Conclusion

- A graded word sense similarity for 28 seminal, lexically ambiguous word forms
- Similarity distribution both for human annotators and models
- Analyze the similarity patter for polysemy of different types.

Reference

A local and global view on word senses

- Local: Polysemy
- Gobal: Lexical relations within a lexicon
- Mixed local and gobal: concept space
- How (L)LMs know these knowledge?
  - Know: internal representations / Mechanistic Discovery
  - How: They can well do it! / Intervention (manipulation)

Local View

- Rated similarity between different senses (homo. vs. poly.)
- Workshop in Coling
  - CoMeDi: Context and Meaning—Navigating Disagreements in NLP Annotations https://comedinlp.github.io/
  - Two subtasks: predict majority score and disagreement score for WiC
  - An extension of Uncertainty work in ACL'23 findings
  - Model ensembling to model uncertainty/disagreement
  - Competition + Evaluation + Tech paper writing https://arxiv.org/pdf/2411.12147
- Future Work
  - Other lexical factors to influence disagreements; LLMs

Global View

- Is there any structural features for all the token embeddings in LLMs, if yes, anything to do with scaling law?
- 32000 tokens for LLMs of different scales
- Graph construction
  - tokens as nodes and similarity as edges
  - Strategies to sparify the dense network? (how to keep fair?)
  - Any significant differences of graph statistics for models of different scales? (connectivity; density...)
  - Naunced geometric structures? (word analogy quadrilateral...)
  - Ongoing...

Mixed local and global views

- To construct a (cross-lingual)conceptual space (semantic maps)
- Word-meaning occurrence table
- A graph (concepts as nodes; proximity) meeting:
  - Meanings within a word should be connected (bounded by a region) (local)
  - Multiple related words (global)
- A Top-down Graph-based Tool for Modeling Classical Semantic Maps: A Case Study of Supplementary Adverbs
  - submitted to ARR
  - Future work: more cases to verify the effectiveness.

Q & A

THANK YOU

Note