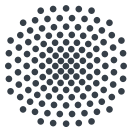

University of Stuttgart
Germany

Detection of Non-recorded Word Senses

January 26, 2024

Jonathan Lautenschlager

Institute for Natural Language Processing, University of Stuttgart

Intro

- Unknown Sense Detection (USD)
- Support dictionary maintenance
- Realistic few-shot scenario
- English and Swedish
- Modern and historical

Outline

Unknown Sense Detection (USD)

USD is the identification of corpus occurrences that are not covered by a given sense inventory. *(Erk (2006))*

- Related to other popular tasks like Word Sense Detection (WSD)
- A specific task of Natural Language Processing (NLP)

- An underlying sense inventory
- Word usages to examine
- A methodology to recognize and differentiate between the meanings of words (in context)

Lexical Resources (WordNet)

- A well-known and well-established large lexical database for the English language.
- It groups words which are synonymous in one of their meanings into so-called *synsets* that describe distinct concepts or senses.

  `<`**`car`**`, auto, automobile, machine, motorcar>`

  `<cable car, `**`car`**`>`
- A total of 117,000 synsets, all containing a gloss.
- Only a portion includes example usages

Lexical Resources (Svensk Ordbok (SO))

- A dictionary of the Swedish language, created and maintained by the University of Gothenburg (Allén, 1981).
- SO is privately managed
- Lists senses under a headword
- A total of 68,000 senses for over 41,500 headwords
- The majority of the senses are described by a gloss
- All sense entries contain example usages

Corpora (Modern)

	English	Swedish
Name	Leipzig_News	Leipzig_News
Year	2020	2022
Source	Goldhahn, Eckart, and Quasthoff (2012)	
Sentences	1 million	1 million

Table 1: Comparison of modern corpora

Corpora (Historical)

	English	**Swedish**
Name	CCOHA	Kubhist2
Year	1810–1860	1790–1830
Source	Alatrash et al. (2020)	Språkbanken (downloaded in 2019)
Sentences	250 thousand	3.3 million

Table 2: Comparison of historical corpora

Contextualized Embeddings (XL-LEXEME)

- SentenceBERT (SBERT): extends BERT, using siamese and triplet networks to produce meaningful sense embeddings for a single sentence (Reimers & Gurevych, 2019)
- XL-LEXEME is a pretrained contextualized embeddings model that is based on SBERT and fine-tuned on human-labeled Word-in-Context (WiC) data (Cassotti et al., 2023)
- Gives prominence to a target word in the given context.

Models (Sense inventory)

- Solely data from the dictionary
- Gloss and example word usages undergoing different replacement strategies (Table 3)
- Naturally limited in their predictions by gaps in the data, i.e., missing gloss or example usages

option	description	example
0	leaves the sequence as is	a poor salary
1	HEADWORD: SEQUENCE	inadequate: a poor salary
2	SEQUENCE (HEADWORD)	a poor salary (inadequate)
3	SEQUENCE, i.e., HEADWORD	a poor salary, i.e., inadequate
4	replace word	a inadequate salary

Table 3: Replacement strategies on an example of *inadequate* taken from WordNet

Models (Target embeddings and comparision)

- Word usages from corpora
- Original sentence
- Target token substituted by its lemma
- Contextualized embedding using XL-LEXEME

- Calculate distance to all eligible sense embeddings
  - Cosine similarity
  - Spearman's rank correlation coefficient
- Assign based on a threshold

Experiments (Outline)

- Human annotation for model tuning and to establish a reference point
- Tuning of hyperparameters, i.e., the similarity threshold
- Predicting with the most promising models
- Determine the models' performance with human annotation

- Outline human annotations:
  - A total of six annotators, three for each language
  - All participants are students and native speakers of the respective language
  - They received a 30-minute briefing
  - They underwent a short test annotation

Experiments (Outline human annotation)

- Carried out using the PhiTag platform
- Inspired by Erk et al. (2013)'s WSsim method
- Individual assessment of all senses for a word usage
- Ask for binary classification
  - *sense gloss fits* (label "1")
  - *sense gloss does not fit* (label "0")
  - *no specification is possible* (label "-")

Experiments (PhiTag Interface)

usage:
Abigail's sister accompanied her to this meeting because
the sister alleged that she had witnessed some of the childhood
sexual abuse from the relative.

gloss:
estimated by comparison; not absolute or complete

all glosses:
estimated by comparison; not absolute or complete

an animal or plant that bears a relationship to another
(as related by common descent or by membership in the same genus)

a person related by blood or marriage

Figure 1: Illustration of the PhiTag user interface

Experiments (Human annotation: Part 1)

- Human annotation is conducted on a random sample from the corpora
- Identically for both languages
- Word usages were retrieved by lemmatizing the sentences and searching for appearances of a randomly selected headwords
- At most five usages for each headword, chosen at random
- Approximately 1,200 annotation instances, distributed over 500–700 usages
- Only primary synsets are considered

Experiments (Human annotation: Part 1)

- Aggregate judgements per annotation instance by majority
- A usage is considered assigned iff at least one instance has the majority label "1"

	English	Swedish
Instances	1165	1202
Usages	474	706
Label distribution (0, 1, -)	(1840, 1651, 4)	(1294, 2104, 208)
Excluded instances	2	87
Remaining usages	473	674
Assigned	428	562
Unassigned	45 (9.5%)	95 (16.6%)

Table 4: Statistics of the first human annotation phase

Experiments (Threshold tuning)

- 10 rounds of 5-fold cross-validation for every model
- Same simulated data for all models
- Try every threshold $\in \{0.0, 0.01, 0.02, \ldots, 1.0\}$
- Choose threshold to maximize F_β-score (including $\beta = 0.3$)

Experiments (Threshold tuning)

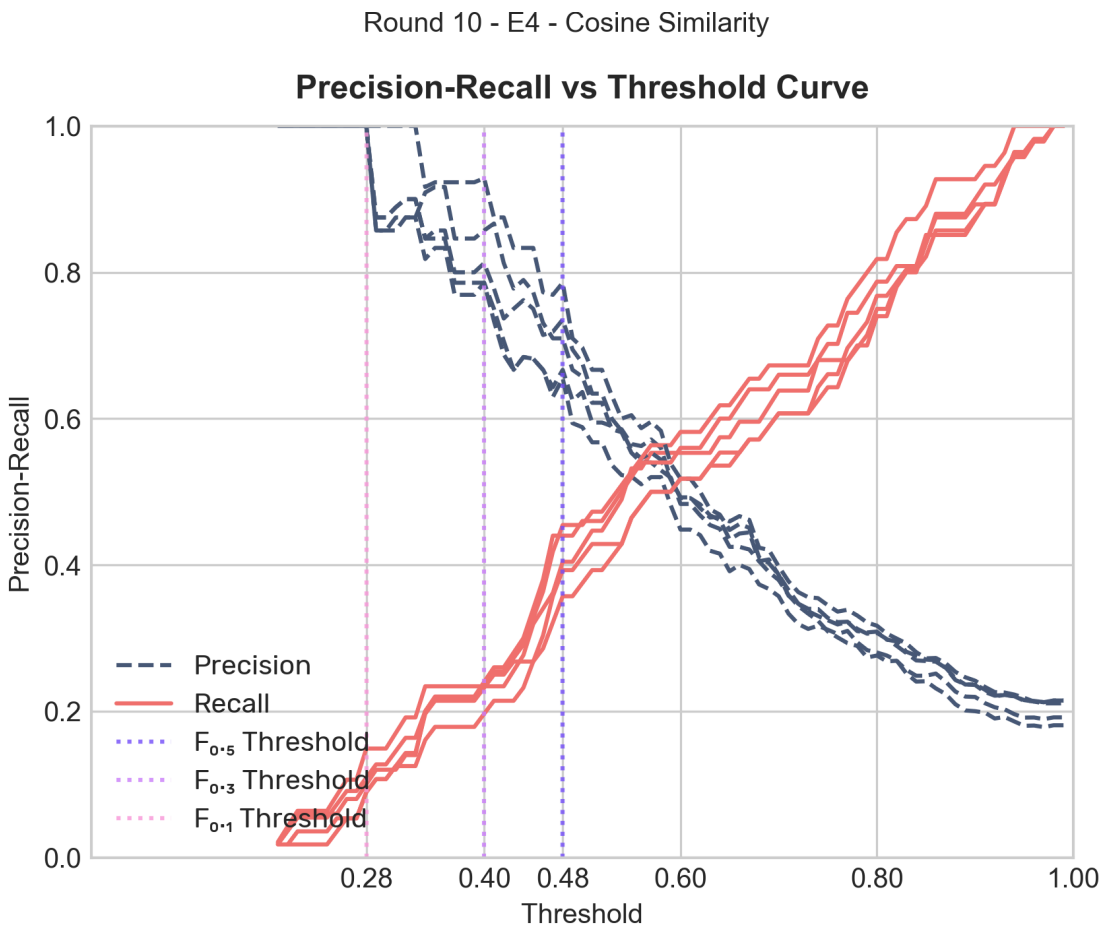

Figure 2: Precisions and recalls of all five folds in the cross-validation round 10 of model `E4_COS` on English data

Experiments (Threshold tuning)

	$F_{0.3}$-score							
	English				**Swedish**			
	DEFAULT		**SUBST.**		**DEFAULT**		**SUBST.**	
	COS	**SPR**	**COS**	**SPR**	**COS**	**SPR**	**COS**	**SPR**
E0	0.466	0.523	0.477	0.452	0.417	0.419	0.413	0.400
E1	0.589	0.590	0.583	0.592	0.451	0.443	0.413	0.399
E2	0.579	0.562	0.566	0.590	0.425	0.431	0.411	0.408
E3	0.494	0.523	0.493	0.489	0.428	0.431	0.397	0.392
E4	**0.613**	0.584	0.593	**0.612**				
G0	0.270	0.280	0.255	0.267	0.349	0.371	0.345	0.340
G1	0.227	0.220	0.226	0.209	0.600	0.606	0.549	0.537
G2	0.260	0.223	0.264	0.245	0.550	0.612	0.564	0.547
G3	0.217	0.259	0.275	0.256	**0.625**	0.617	0.599	0.621

Table 5: Results of cross-validation. Performance is given as average $F_{0.3}$ across rounds and folds

Experiments (Human annotation: Part 2)

- Prediction on 100–150 thousand roughly filtered sentences, from both modern and historical corpora, respectively.
- Every sentence is lemmatized and then searched for headwords represented in the respective model's sense inventory.
- A prediction is made for all usages of the found headwords
- Exclude partially complete headwords (at least one sense incomplete)
- Unassigned usages are sorted by similarity to the nearest sense
- At most eight usages for the same headword are chosen from the top of this list

Experiments (Human annotation: Part 2)

- Aggregate judgements per annotation instance by majority
- A usage is considered assigned iff at least one instance has the majority label "1"

	English	Swedish
Instances	1208	1400
Usages	322	1001
Label distribution (0, 1, -)	(2151, 1462, 11)	(2529, 1218, 456)
Excluded instances	5	109
Remaining usages	322	927
Assigned	277	327
Unassigned	45 (13.98%)	600 (64.725%)

Table 6: Statistics of the second human annotation phase

Evaluation (Comparison)

	Phase 1			Evaluation Phase		
Usages	**All**	**Modern**	**Historical**	**All**	**Modern**	**Historical**
total	474	232	242	322	210	112
excluded	1	1	0	0	0	0
remaining	473	231	242	322	210	112
assigned	428	208	220	277	176	101
unassigned	45 (9.5%)	23 (9.9%)	22 (9.1%)	45 (13.98%)	34 (16.2%)	11 (9.8%)

	Phase 1			Evaluation Phase		
Usages	**All**	**Modern**	**Historical**	**All**	**Modern**	**Historical**
total	706	337	369	1001	478	523
excluded	52	4	28	74	9	65
remaining	674	333	341	927	469	458
assigned	562	293	269	327	224	103
unassigned	112 (16.6%)	40 (12.0%)	72 (21.1%)	600 (64.7%)	245 (52.2%)	355 (77.5%)

Table 7: Comparison of the different corpora of the English and Swedish data

Evaluation (Manual analysis)

- Cases where also a close manual analysis suggests that they are truly non-recorded in our dictionary:

  usage

  No wonder hes up there getting big **baked**.

  senses

  cook and make edible by putting in a hot oven

  prepare with dry heat in an oven

  heat by a natural force

  be very hot, due to hot weather or exposure to the sun

  dried out by heat or excessive exposure to sunlight

  (bread and pastries) cooked by dry heat (as in an oven)

Evaluation (Manual analysis)

- Cases with likely historical meanings that are seemingly non-recorded:

usage

You seem to intend a eulogy , yet leave out whatever was noblest in her , and **blacken** while you mean to praise .

senses

make or become black

burn slightly and superficially so as to affect color

Evaluation (Manual analysis)

- A major problem are multi-word expressions:

  usage

  Im at that age where many of my friends are having children, and a central topic of conversation whenever were together **revolves** around creating the almostscientifically set schedule for their babies.

  senses

  turn on or around an axis or a center

  move in an orbit

  cause to move by turning over or in a circular manner of as if on an axis

Evaluation (Manual analysis)

- Occurrences of ellipsis of a represented multi-word:

usage

Notably, the local Native American tribes were not targeted or **wiped** the new nation embraced them as equals and allowed the tribes a major say in the rule of America.

senses

rub with a circular motion

Evaluation (Manual analysis)

- OCR and spelling errors:

  "[...] mud turtles , and floating timber to say nothing of water snakes , which were far more terrible to me than **shirks** ."

- Likely ambiguous usages:

  "Stolen **bases** are a thing for me, Betts said."

- Incorrect lemmatization:

  virtually → *virtual*

- Ad-hoc innovative word meanings:

  "But sloppy, agendadriven journalism of this sort fans the **flames** of racial division."

Conclusion

- Automatically detect non-recorded word senses in a large corpora, based on a given large, but possibly imperfect sense inventory.
- The method considerably increases the chance to find non-recorded word senses in corpus usages compared to a random baseline
- Predict a large number of unassigned usages that can be used to update WordNet's and SO's sense inventory in the near future
- Manual analysis shows some weaknesses like faulty multi-word detection
- No manual analysis of the Swedish data yet

References I

Alatrash, R., Schlechtweg, D., Kuhn, J., & Schulte im Walde, S. (2020, may). CCOHA: Clean Corpus of Historical American English. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 6958–6966). Marseille, France: European Language Resources Association. Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.859

Allén, S. (1981). The lemma-lexeme model of the swedish lexical data base. In R. Bo (Ed.), *Modersmålet i fäderneslandet. ett urval uppsatser under fyrtio år av sture allén.* (pp. 268–278). Meijerbergs arkiv för svensk ordforskning 25.

Cassotti, P., Siciliani, L., de Gemmis, M., Semeraro, G., & Basile, P. (2023, July). Xl-lexeme: Wic pretrained model for cross-lingual lexical semantic change. In *Proceedings of the 61th annual meeting of the association for computational linguistics.* Online: Association for Computational Linguistics.

Erk, K. (2006, June). Unknown word sense detection as outlier detection. In *Proceedings of the human language technology conference of the NAACL, main conference* (pp. 128–135). New York City, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N06-1017

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

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012, May). Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In N. Calzolari et al. (Eds.), *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)* (pp. 759–765). Istanbul, Turkey: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf

Reimers, N., & Gurevych, I. (2019). *Sentence-bert: Sentence embeddings using siamese bert-networks.*

References II

Språkbanken. (downloaded in 2019). The kubhist corpus, v2 [Computer software manual]. Retrieved from https://spraakbanken.gu.se/korp/?mode=kubhist