

ShadowSense: a Multi-annotated Dataset for Evaluating Word Sense Induction

Ondřej Herman, Miloš Jakubíček
Lexical Computing, Czechia

ShadowSense: Dataset for Evaluating WSI

- **Word senses of the noun **band****
 - Music band
 - Wedding band
 - Frequency band
 - Hair band
 - Rubber band
 - ...
- **Word Sense Induction**
 - Identify the different senses a word can take on from raw corpus text
 - Fundamental within NLP
 - Open problem, thought to be AI-complete

- What granularity is the “right” one?

Existing Evaluation Strategies

- Limited, overly opinionated or imprecise.
- SemEval shared tasks
 - SemEval 2007 Task 2
 - SemEval 2010 Task 14
 - SemEval 2013 Task 13
- Dictionary Data
- Synthetic Approaches
 - Word replacement

Annotation Methodology

- We selected 25 known-polysemous words for Czech and English
- We then extracted 150 most salient collocations using Sketch Engine for each
 - Word Sketch grammar based on morphosyntactically defined features
- The annotators were presented with:
 - The collocate
 - The longest commonest match – a sequence of words, representative of the collocation
 - A link to the concordance, should they need to see more context
- Each annotator grouped the collocations by the word sense
 - Different senses for every annotator
 - No assigned sense is a valid annotation

Annotation Results

- 5 Annotators for the Czech dataset.
 - 6.5 M total concordance lines
 - 3.6 senses per word on average
- 10 annotators for the English dataset.
 - 6 native speakers
 - 4 English as a second language speakers
 - 1.5 M total concordance lines
 - 8.11 senses per word on average

Structure of the Dataset

```
head  sense1 sense2 sense3 text
band-n a1.s1 a2.s1 a3.s1 The Beatles are arguably the most famous <band> in rock and roll history.
band-n a1.s1 a2.s1 a3.s1 This <band's> music was pop influenced.
band-n a1.s2 a2.s2 a3.s2 Three <bands> at 5 GHz have been allocated for WiFi and similar services.
band-n a1.s7 a2.s6 a3.s9 Put two large rubber <bands> around the base of the cup.
band-n a1.s3 a2.s6 a3.s10 If he put the gold <band> on your finger, he likes you pretty well.
band-n a1.s1 a2.sx a3.sx Hang out the bunting, strike up the <band>, Tom and Ben are home.
band-n a1.s6 a2.sx a3.sx This <band> is an abandoned band.
```

- UTF-8, TAB separated columnar format
- Available at <https://github.com/lexicalcomputing/shadowsense/>

Inter-Annotator Agreement

We evaluated the IAA using the Adjusted Rand Index.

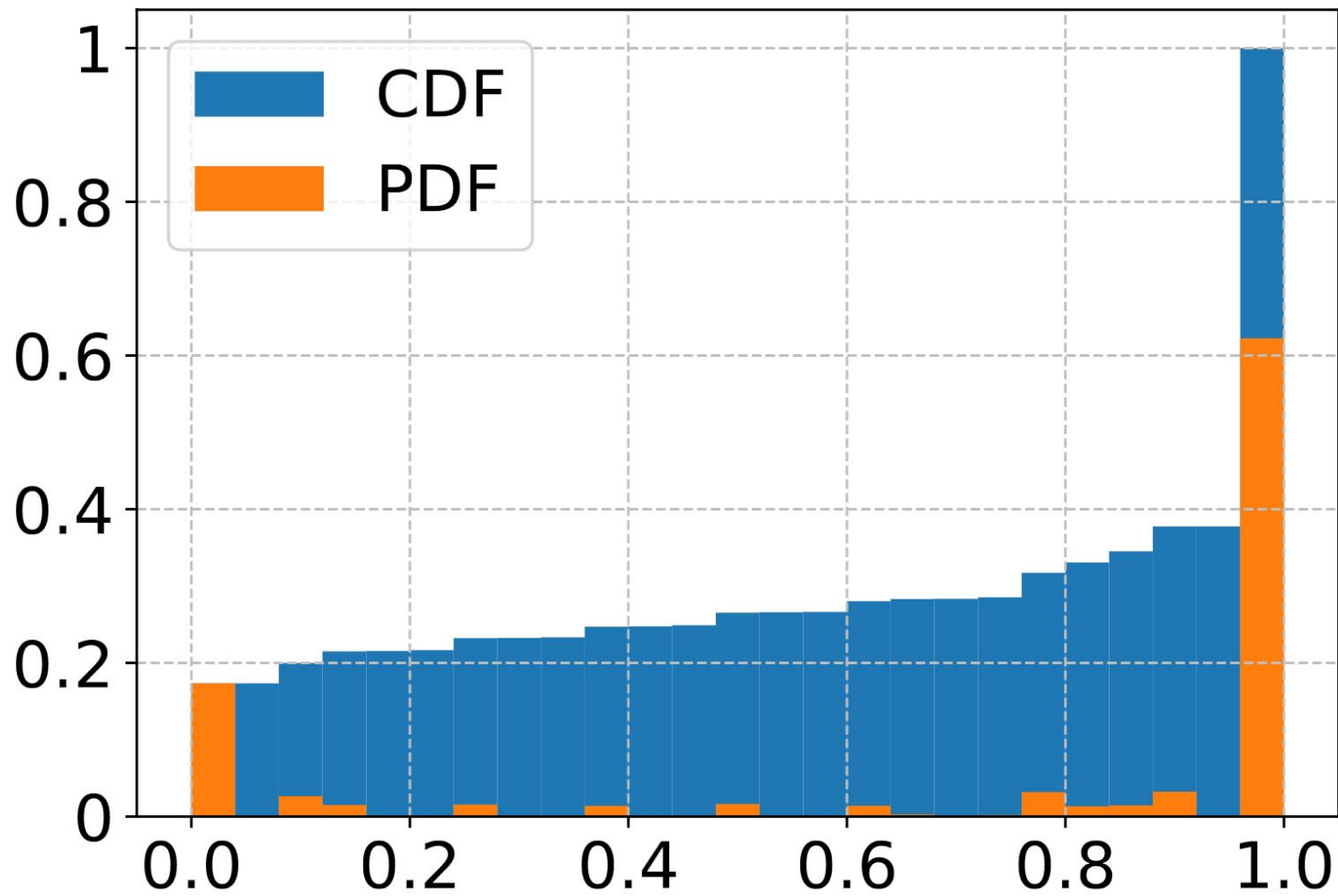
| | 2 | 3 | 4 | 5 |
|----------|----------|----------|----------|----------|
| 1 | .88 | .94 | .87 | .92 |
| 2 | | .92 | .90 | .91 |
| 3 | | | .92 | .94 |
| 4 | | | | .91 |

Czech (0.914)

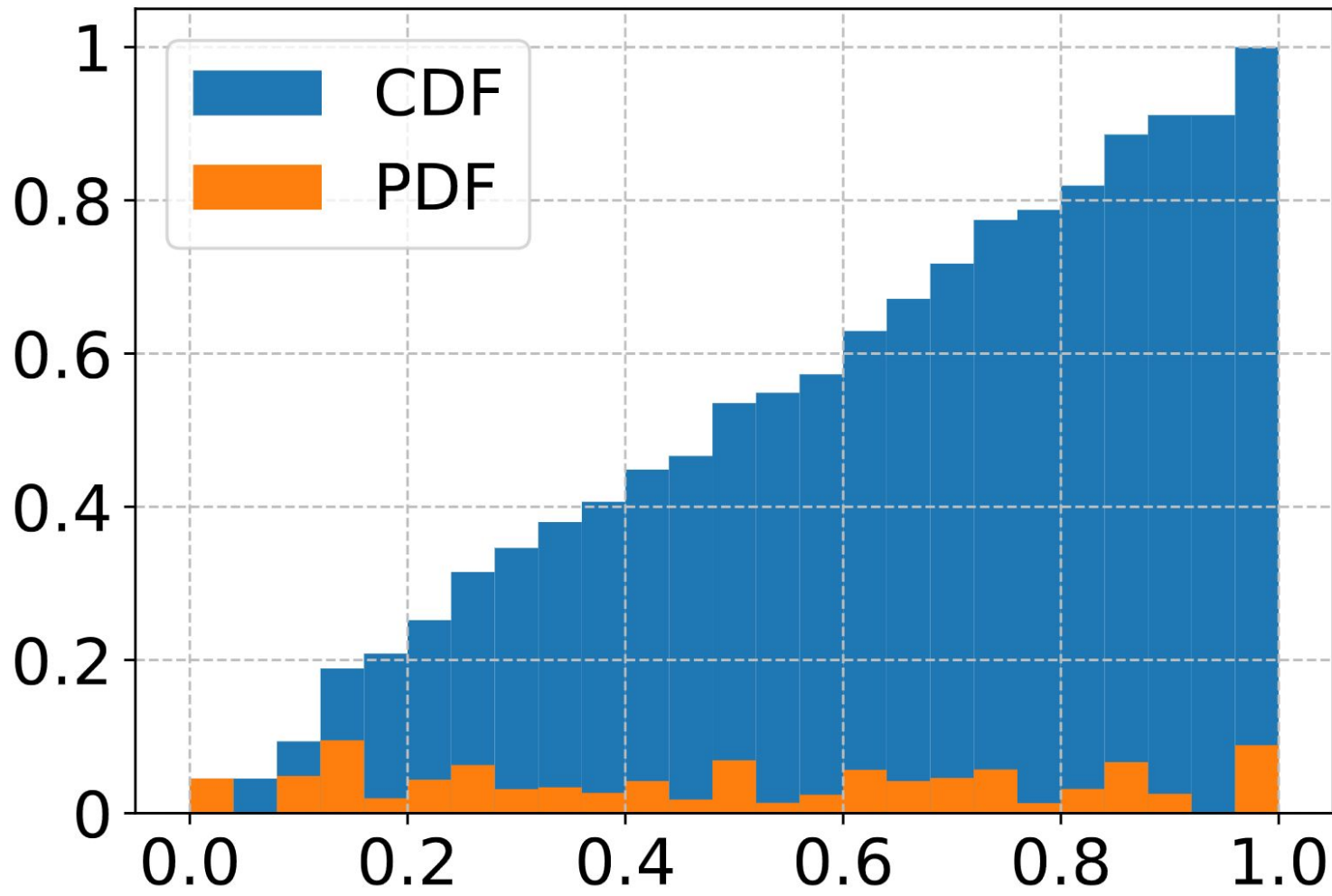
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----------|----------|----------|----------|----------|----------|----------|----------|----------|-----------|
| 1 | .79 | .84 | .82 | .84 | .86 | .84 | .82 | .84 | .85 |
| 2 | | .79 | .81 | .81 | .80 | .81 | .81 | .78 | .81 |
| 3 | | | .81 | .81 | .82 | .82 | .81 | .81 | .83 |
| 4 | | | | .82 | .81 | .82 | .82 | .83 | .83 |
| 5 | | | | | .83 | .81 | .80 | .82 | .82 |
| 6 | | | | | | .83 | .86 | .80 | .84 |
| 7 | | | | | | | .84 | .83 | .86 |
| 8 | | | | | | | | .82 | .83 |
| 9 | | | | | | | | | .84 |

English (0.817)

bark-n



club-n



Evaluation Methodology: Motivation

1. Some pairs of occurrences carry the same sense.
2. Some pairs of occurrences do not carry the same sense.
3. Some pairs of occurrences are inconclusive.

We want to evaluate only on the pairs from points 1 and 2, while ignoring the pairs from the point 3, only using the data on which most annotators can agree.

Evaluation Methodology: Metrics

- Based on Rand Index (~ what proportion of pairs of observations are clustered in the same way as the gold data is clustered)
- Adjusted Rand Index
 - Adjusted for chance – better range and expressive power
- Shadow Rand Index (sRI)
 - Only take the clear pairs into account (> 75 % agreement)
 - n annotators, m occurrences of a word
 - c_x is the x -th word gold label
 - s_{xy} is the x -th word annotation of the y -th annotator

$$r_{ij} = \frac{\sum_{k=1}^n [s_{ik} = s_{jk} \wedge s_{ik} \neq \perp]}{\sum_{k=1}^n [s_{ik} \neq \perp \wedge s_{jk} \neq \perp]}$$

$$\text{sRI} = \frac{2(tp \cdot tn - fp \cdot fn)}{(tn + fn)(tp + fp) + (tn + fp)(tp + fn)}$$

$$tp = \sum_{i=1}^m \sum_{j=1}^{i-1} [c_i = c_j \wedge r_{ij} \geq 0.75]$$

$$tn = \sum_{i=1}^m \sum_{j=1}^{i-1} [c_i \neq c_j \wedge r_{ij} < 0.25]$$

$$fp = \sum_{i=1}^m \sum_{j=1}^{i-1} [c_i = c_j \wedge r_{ij} < 0.25]$$

$$fn = \sum_{i=1}^m \sum_{j=1}^{i-1} [c_i \neq c_j \wedge r_{ij} \geq 0.75]$$

Evaluation of Sample WSI Systems

- We evaluated the English dataset:
 - SymPatternWSI, BertWSI and Adaptive Skip-gram.
 - Using SemEval 2013 data and metrics
 - Using ShadowSense data and metrics

| | SemEval2013 | | ShadowSense | |
|---|-------------|------|-------------|-------|
| | fNMI | fBC | sRI | wsRI |
| SymPatternWSI | .115 | .572 | -.004 | -.004 |
| BertWSI | .209 | .641 | .757 | .761 |
| AdaGram ($\alpha = 0.1, d = 256, w = 4$) | .065 | .455 | .285 | .282 |

Future Work

- Extension of the dataset for more languages
- Introduce different parts of speech

Conclusions

- We created a novel multi-annotated dataset for evaluating WSI systems
 - For Czech and English, 25 nouns for each language
- We devised two metrics
 - Which take the not well-defined nature of word-senses into account
- We evaluated the dataset using our metrics
- The dataset and an efficient scorer are available at <https://github.com/lexicalcomputing/shadowsense>