Compiling a Suitable Level of Sense Granularity in a Lexicon for AI Purposes:

The Open Source COR Lexicon

Bolette S. Pedersen, Nathalie C. H. Sørensen, Sanni Nimb, Ida Flörke, Sussi Olsen, Thomas Troelsgård
Table of Contents

Introducing COR

The semantic component: COR-S

Principles for defining core senses

Automatic sense clustering
Introducing COR

• Companies in Denmark are right now entering the field of language-centered AI – and are therefore working intensively with Danish language data from an NLP perspective

• In this context, there is an increasing request for a standardised machine usable lexicon of Danish with basic morphology and semantics (core senses, sentiment etc.)

• The government has initiated a general effort to support AI in Denmark – COR is part of this initiative funded by the Agency for Digitisation under the Ministry of Finance
Which partners and which background?

- Danish Language Council
- Society for Danish Language and Literature
- Centre for Language Technology (CST) at the University of Copenhagen

COR is based on existing Danish dictionaries

- We take advantage of the very rich and socially contextualised information on word meaning already described in traditional lexica
- In other words on high-quality, locally anchored knowledge about the Danish language and society!
The Semantic Component: COR-S

COR-S
A coarse-grained sense inventory for Danish

The Danish Dictionary (DDO)

The Danish Thesaurus (DT)

The Danish WordNet (DanNet)

FrameNet
The Danish FrameNet Lexicon
How to achieve a suitable level of sense granularity?

Our aim:

• To establish generalized principles of **lexical semantic coreness**
• To reduce the DDO sense inventory accordingly
• Thereby achieve a **core sense inventory** which is potentially relevant for modern texts and **distinguishable on distributional grounds** and thus more suitable for NLP, still, however, capturing the central/relevant senses

Our method:

• To develop a hand-coded and extensive **gold standard** for highly polysemous and more average parts of the vocabulary
• To apply automatic methods for the rest of the vocabulary based on this standard
Related work on sense granularity (for factual references see the paper)

In **lexicography** and **lexical semantics**, the discussion of **sense granularity** has been ongoing for decades.

A typical, slightly simplified, categorisation of lexicographers into being either **lumpers** or **splitters**

Where very rich sense descriptions seem to correspond well to the needs of human users, very subtle sense descriptions tend to cause **notorious problems for NLP and WSD**

In fact, this has been the case to an extent where traditional dictionaries have been **deemed somewhat useless** in relation to NLP.

The **ELEXIS and COR projects** are trying to remedy this problem.
Principles for sense structure in DDO

• A close semantic relationship between a main sense and its sub-senses

• While sub-senses denote either a **broader**, a **narrower** or a **figurative** nuance of its main sense, **main senses are in principle semantically unrelated to each** other although etymologically deriving from the same lemma

• However, in order to avoid deep sense structures in the printed dictionary, senses that in fact could have been classified as sub-senses from the above criteria, **are actually sometimes found to be described as main senses**

• In other words: idiosyncracies have to be taken into account
COR principles of ‘coreness’

**Delete** a DDO main or sub-sense if it:

- is marked as rare, historic, very domain-specific, colloquial, or slang in DDO (and/or has a very low sense weight)

**Merge** a DDO sub-sense with its main sense, unless a sub-sense is:

- Marked with a different ontological type in the wordnet
- Marked as figurative sense in DDO

In some specific cases: **Merge** semantically close main senses
An example: *Hær* (army..)  

**DDO senses:**

**COR senses for hær:**

**Sense 1**: Army/military forces (HUMAN_GROUP)

**Sense 2**: A big quantity of something (ABSTRACT)
The gold standard

The gold standard consists of two parts:

• Part I contains 3,500 highly polysemous lemmas (~15,000 senses in DDO)
• Part II: 2,700 average polysemous lemma

Inter-annotator agreement

• We use Cohen’s $k$
• The average agreement of 0.82
• The principles are actually manageable

43% sense reduction (4.3 senses in DDO to 2.4 senses in COR)
Experiments with automatic sense clustering
The task

Can we replicate the hand annotations with an automatic method?

Use dictionary and wordnet information to partition the set of a lemma’s non-deleted senses into \( k \) clusters.

\( k = \text{number of senses in COR} \)
Challenges

• **Varying** and **unknown** numbers of clusters for different lemmas
  – 2-step method inspired by
    – dataset alignment (McCrae & Buitelaar, 2018)
    – ELEXIS Clusty tool (Martelli et al, 2019)

• How to **model** information from **dictionaries**?
  – **Text-based:** definition & quotes with a word embedding model (**word2vec**, **BERT**)
  – **Rule-based:** Hand-selected features from **DDO** and **DanNet**
The 2 step method

**Step 1:** Calculate pairwise sense proximity score using a model

**Step 2:** Clustering based on the sense proximity score
1. Semantic Textual Similarity (STS) using BERT or word2vec
2. Use the principles (rule-based)
Text-based models

**Word2vec**

- Centroid embedding of the bag-of-words from definitions + quotes (punctuation + stopwords removed)

- Cosine distance as similarity measure
Text-based models

BERT

- Input: two sentences / contexts (quote and/or definition)
- target lemma marked with [TGT] token
- Output: similarity score
- Fine-tune on 80% of the main annotation
Clustering based on similarity scores

1 merged COR sense

2 COR senses (blue, orange)
Results

Primary evaluation sets

Test word classes

<table>
<thead>
<tr>
<th></th>
<th>Main Test</th>
<th>Main Test Reduced</th>
<th>Average Vocab</th>
<th>DT keywords</th>
<th>Test Nouns</th>
<th>Test Verbs</th>
<th>Test Adjectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>rule-based</td>
<td>0.79</td>
<td>0.78</td>
<td>0.76</td>
<td>0.79</td>
<td>0.79</td>
<td>0.77</td>
<td>0.82</td>
</tr>
<tr>
<td>BERT</td>
<td>0.66</td>
<td>0.65</td>
<td>0.68</td>
<td>0.74</td>
<td>0.68</td>
<td>0.62</td>
<td>0.61</td>
</tr>
<tr>
<td>word2vec</td>
<td>0.62</td>
<td>0.60</td>
<td>0.65</td>
<td>0.61</td>
<td>0.62</td>
<td>0.62</td>
<td>0.61</td>
</tr>
</tbody>
</table>
Results

Primary evaluation sets

<table>
<thead>
<tr>
<th>Method</th>
<th>Main Test</th>
<th>Main Test Reduced</th>
<th>Average Vocab</th>
<th>DT keywords</th>
<th>Test Nouns</th>
<th>Test Verbs</th>
<th>Test Adjectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>rule-based</td>
<td>0.79</td>
<td>0.78</td>
<td>0.76</td>
<td>0.79</td>
<td>0.79</td>
<td>0.77</td>
<td>0.82</td>
</tr>
<tr>
<td>BERT</td>
<td>0.66</td>
<td>0.65</td>
<td>0.68</td>
<td>0.74</td>
<td>0.68</td>
<td>0.62</td>
<td>0.61</td>
</tr>
<tr>
<td>word2vec</td>
<td>0.62</td>
<td>0.60</td>
<td>0.65</td>
<td>0.61</td>
<td>0.62</td>
<td>0.62</td>
<td>0.61</td>
</tr>
</tbody>
</table>
Conclusion
Conclusion

• With DDO as starting point we establish a notion of “coreness” and establish principles for merging senses

• Sense reduction of 43% from DDO to COR with an intercoder agreement of 0.82 -> in other words, the principles seem sound and manageable

• Rule-based model shows promise for automatic sense reduction of the remainder of the vocabulary

• Word classes should be treated differently

• Text-based approaches struggle with highly polysemous lemmas – why hand annotation are still necessary
Thank you – and acknowledgements

• The COR development project is funded by the Danish Agency for Digitisation as part of an AI initiative embarked by the Danish Government in 2020

• The research behind the COR project also relies on the European Lexicographic Infrastructure (ELEXIS) project under the European Union's Horizon 2020 research and innovation programme (grant agreement No 731015)