Fine-grained Termhood Prediction for German Compound Terms Using Neural Networks

Anna Hätty
Robert Bosch GmbH

Sabine Schulte im Walde
University of Stuttgart

Motivation

- domain-specific terms = linguistic expressions which characterize a domain
- automatic term extraction and term understandability
  \[ \rightarrow \text{separately researched} \]
- classes of termhood, which naturally include understandability

Background

- Tiers of terminology for different degrees of association to the domain
- Appearance in only this domain (= very specific) \rightarrow not likely to be known outside of domain (= difficult to understand)

Termhood Classes

| Class  | Description                                                                 | Example                                                                 |
|--------|-----------------------------------------------------------------------------|-------------------------------------------------------------------------|
| NONTERM | Not a domain term                                                            | Deutsch „Germany“                                                        |
| SemTerm | Semantically related to the domain                                           | Vitaminebedarf „requirement of vitamins“                                |
| TERM   | Prototypical and understandable term of the domain                         | Schweinebraten „roast pork“                                             |
| SPECTERM | Prototypical and non-understandable term of the domain                       | Blasensalat (blue boiling) specialty kind of boiling fish               |

Compound Examples

- **Perfect matches:**
  - Tomato (Tomato) + Puree (Tomato) \rightarrow Tomatenpüree (Tomato)
  - tomato + puree \rightarrow tomato puree

- **Same component classes, but different compound classes:**
  - Mittel (Mittel) + Alter (Mittel) \rightarrow Mittelalter (Mittel)
  - mit = age \rightarrow Middle Ages
  - Bei (Bei) + Fuß (bei Fuß) \rightarrow Beifuß (bei Fuß)
  - with + foot \rightarrow mugwort

- **Different component classes, but same compound class:**
  - Paprika (Paprika) + Salat (Salat) \rightarrow Paprikasalat (Paprika)
  - sweet pepper + salad \rightarrow sweet pepper salad
  - Paprika (Paprika) + Hälften (Paprika) \rightarrow Paprikahälften (Paprika)
  - sweet pepper + halves \rightarrow sweet pepper halves

Data Extraction & Annotation

- 400 cooking recipes (kochwiki.de, wikibooks cookbook, wikihow)
- 5 native speaker annotators
- 396 compounds

Model Pipeline

- Input: Compound
  - Compound Splitting
  - Feature Extraction (Lex, Word Embeddings)
  - Neural Network Classifier
  - Output: Predicted Class

Models & Features

- Baseline model:
  - Word embeddings: pre-trained on Wikipedia, adapted on cooking recipes
  - Features for components in cooking-domain:
    - Frequency: How frequently does a constituent appear in other expressions?
    - Productivity: How many expressions the constituent is part of?
  - Optimization on heuristically estimated component classes (ConstOpt)

Results

- Better results for models using both compound and component information
- Optimization on heuristically estimated component class improves the results

Conclusion

- New model of fine-grained classes of termhood, representing both the different degree of association to the domain and a domain term’s understandability
- Including and optimizing information about components leads to 0.8 F-score for best model

References

- Faller, W. 2014. Morphological Processing of Compounds for Statistical Machine Translation. Dissertation, Institute for Natural Language Processing (IWR), University of Stuttgart.
- Thormann, Bernhard, and Stephan Wuhrer. Grundlagen der Terminologie- und Lexikographielehre. Mannheim: Bibliographisches Institut. 2013.
- Faller, W., & Thormann, B. 2016. Morphological processing of compounds for automatic terminology extraction. Proceedings of the 11th Workshop on Morphological Processing, MWE@EACL, 2017, pages 151–154, Valencia, Spain.