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

Anna Hätty
Robert Bosch GmbH

Sabine Schulte im Walde
University of Stuttgart

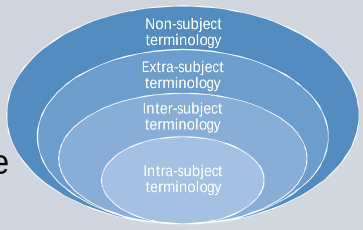
Motivation

- *domain-specific terms* = linguistic expressions which characterize a domain
- automatic term extraction and term understandability → 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) → not likely to be known outside of domain (= difficult to understand)


Non-subject terminology
Extra-subject terminology
Inter-subject terminology
Intra-subject terminology

Roecke (1999)

Termhood Classes

Class	Description	Example
NonTerm	Not a domain term	Deutschland „Germany“
SimTerm	Semantically related to the domain	Vitaminbedarf „requirement of vitamins“
Term	Prototypical and understandable term of the domain	Schweinebraten „roast pork“
SpecTerm	Prototypical and non-understandable term of the domain	Blausud [blue boiling] *special kind of boiling fish*

Compound Examples

- **Perfect** matches:

Tomate (Term) + Püree (Term) → Tomatenpüree (Term)
tomato + puree → tomato puree

- **Same** component classes, but different compound classes:

Mittel (NonTerm) + Alter (NonTerm) → Mittelalter (NonTerm)
mean + age → Middle Ages

Bei (NonTerm) + Fuß (NonTerm) → Beifuß (SpecTerm)
with + foot → mugwort

- **Different** component classes, but same compound class:

Paprika (Term) + Salat (Term) → Paprikasalat (Term)
sweet pepper + salad → sweet pepper salad

Paprika (NonTerm) + Häften (NonTerm) → Paprikahälften (Term)
Sweet pepper + halves → sweet pepper halves

Data Extraction & Annotation

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

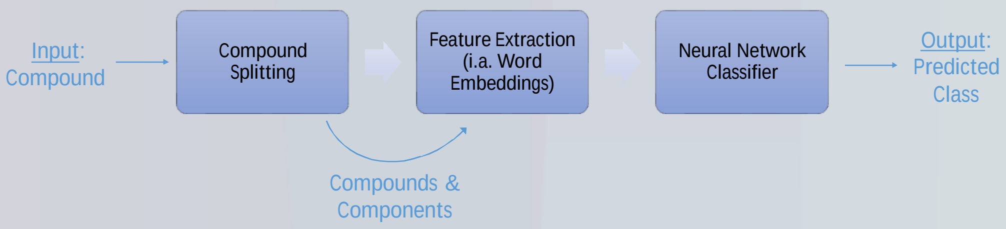
5:0	4:1	3:2
185	116	83

Agreement.

NonTerm	SimTerm	Term	SpecTerm
44	43	250	59

Number of annotated terms per class.

Model Pipeline



Compound Splitting

Combine three splitters:

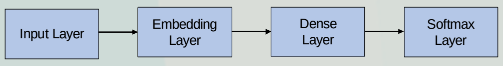
- **CharSplit** (Tuggener, 2016), an n-gram-based splitter
- **CompoST** (Cap, 2014), SMOR-based splitter (high precision)
- **Simple Compound Splitter (SCS)** (Weller-Di Marco, 2017) – handcrafted rules and frequency information (high recall)

Splitters	Wrong Splits	Wrong Side	% correct splits
CharSplit	25	9	91.4%
CompoST + CharSplit	14	9	94.2%
Compost + SCS + CharSplit	9	8	**95.7%**

Splitting performances of the three compound splitters.

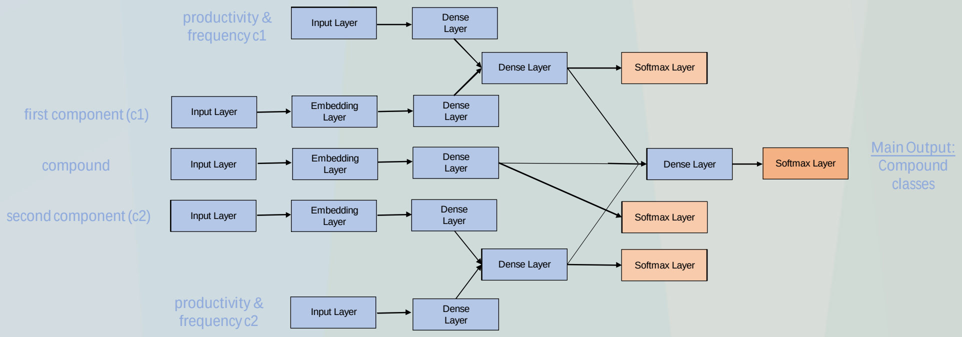
Models & Features

- Baseline model:



Basic architecture.

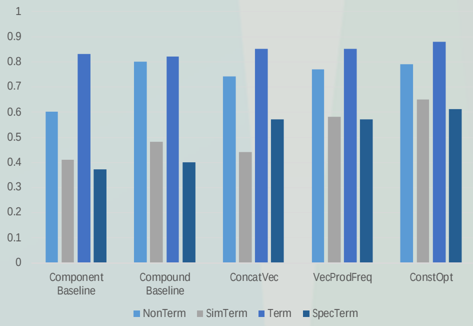
- Word embeddings: pre-trained on Wikipedia, adapted on cooking recipes
- Features for components in cooking-domain:
  - **Frequency**: How frequently does a constituent appear in other expressions?
  - **Productivity**: Of how many expressions the constituent is part of?
- Optimization on heuristically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



ConstOpt Model.

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 on information about components leads to 0.8 F-score for best model

References:

- Fabienne Cap. 2014. Morphological Processing of Compounds for Statistical Machine Translation. Dissertation, Institute for Natural Language Processing (IMS), University of Stuttgart.
- Thorsten Roelcke. 1999. Fachsprachen. Grundlagen der Germanistik. Erich Schmidt Verlag.
- Don Tuggener. 2016. Incremental Coreference Resolution for German. Dissertation, Faculty of Arts, University of Zurich.
- Marion Weller-Di Marco. 2017. Simple compound splitting for German. In Proceedings of the 13th Workshop on Multiword Expressions, MWE@EACL 2017, pages 161–166, Valencia, Spain.

Corporate Research | CR/AEU2 | 2018-08-25 | COLING LAW-MWE 2018

Universität Stuttgart

Institut für Maschinelle Sprachverarbeitung

BOSCH