Machine learning approaches to analyzing German synthetic compounds

Carlotta J. Hübener, Institut für Germanistik, Universität Hamburg • carlotta.huebener@uni-hamburg.de

1 Introduction

Synthetic compounds are syntactically shaped words with an **internal argument structure** (e.g., *heart-warming*).

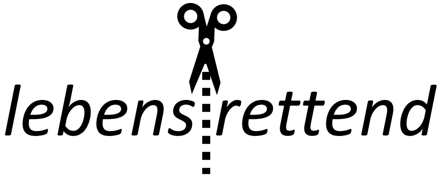
In German, the first constituents correspond to all kinds of objects, for instance:

- accusative: *herzerwärmend*
- dative: *zweckentsprechend*

How can we determine their distribution? I present the integration of a **neural parsing model** in the analysis of synthetic compounds through their base verb valencies.

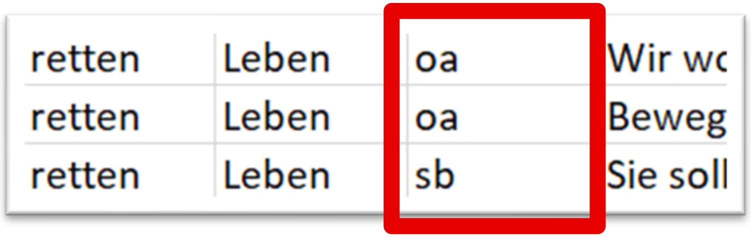
2 Method

1. extract first and second constituent


lebens *rettend*

2. query corpus for noun + verb automatically

3. dependency parsing with de_core_news_sm from spaCy (Honnibal & Montani 2017)



retten	Leben	oa	Wir wc
retten	Leben	oa	Beweg
retten	Leben	sb	Sie soll

3 Results

Sample of 404 noun-participle combinations:

- accuracy 0.94
- precision$_\mu$ 0.99
- recall$_\mu$ 0.89
- F1$_\mu$ 0.94

References

- Digitales Wörterbuch der deutschen Sprache. Berlin-Brandenburgische Akademie der Wissenschaften. DIE ZEIT corpus. https://www.dwds.de/d/korpora/zeit.
- Digitales Wörterbuch der deutschen Sprache. Berlin-Brandenburgische Akademie der Wissenschaften. Core corpus. https://www.dwds.de/d/k-referenz#kern.
- Honnibal, Matthew, and Montani, Ines. 2017. spaCy 2. Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

A special thanks to Melanie Andresen and Janis Pagel