Features of Compositionality in English and German Noun-Noun Compounds

Anna Hätty, Sabine Schulte im Walde, Stefan Bott, Nana Khtisavrishvili

Noun-Noun Compounds and Compositionality

1. **Noun-Noun Compounds**: complex words with two simplex nouns as constituents
   - left: modifier \( \Rightarrow \) fish soup
   - right: morphological head \( \Rightarrow \) fish soup

2. **Compositionality**: expresses that the meaning of a compound can be obtained by the meaning of its constituents
   - leather trousers / Lederhose: highly compositional
   - Jailbird / Knastbruder: highly compositional w.r.t. the modifier
   - Sun flower / Sonnenblume: highly compositional w.r.t. the head
   - Scapegoat / Sünderbock: non-compositional

Goal

How do compound features influence the prediction of compositionality with a distributional model?

E.g.: Are compounds with a high-frequent head more easily/difficult to predict than compounds with a low-frequent head?

Features

1. **Corpus frequency**
   - frequencies of compound, modifier and head in the web corpora EN-/DECW14A (Schäler and Bildhauer, 2012) (en-/decow)

2. **Constituent family size**
   - denotes either the number of compound types in en-/decow which have the same modifier or the same head
   - e.g. modifier family size of *game*: game inventor, game console, ...
   - e.g. head family size of *game*: ball game, video game, ...

3. **Ambiguity**
   - number of senses of the modifier and head from WordNet/GermaNet

4. **Semantic relations**
   - define how two nouns link to each other in a compound, e.g. kitchen door \( \Rightarrow \) kitchen have door
   - Relation annotation scheme used: by Ō Séaghdha (2007)

Gold Standards

All compound datasets include compositionality ratings and information about the features.

1. newly created compound sets:
   - **Ghost-NN S** (German): balanced for modifier family size and head ambiguity (180 compounds)
   - **Ghost-NN XL** (German): extended Ghost-NN S, enriched with compounds of the same modifier and head families like in Ghost-NN S (668 compounds)

2. existing datasets enriched with missing features:
   - Schulte im Walde et al. (2013) (German)
   - Reddy et al. (2011) (English)
   - Ō Séaghdha (2007) (English) (part of 396 compounds)

Distributional Model of Compositionality

1. compute vectors for compounds and each of its constituents \( \Rightarrow \) search for context words (nouns) in a window of words around the target (compound and constituents) and count frequencies
2. association measure: local mutual information (LMI) (Evert, 2005)
3. compute cosine similarities: between compound and modifier, and modifier and head vectors
4. compute Spearman’s rank correlation coefficient (Siegel and Castellan, 1988): correlation between manually annotated compositionality scores and those computed by the system

Evaluation of features: extract min/max 60

To distinguish between low and high feature values for evaluation:

1. sort all compounds once for each feature (their corpus frequency, the corpus frequency of their head, the constituent family size of their head ...)
2. compare 60 lowest with 60 highest examples (exception: Reddy et al. (2011): 45 compounds)

Results

Figure 1: Illustration for semantic space model of compositionality

Figure 2: results for a) compound b) modifier and c) head corpus frequency

Figure 3: results for a) modifier and b) head constituent family size

Figure 4: results for a) modifier and b) head ambiguity

Figure 5: results for relations of a) Ghost-NN and b) Ō Séaghdha (2007)

Conclusion

→ Prediction of compositionality with a semantic space model is easier if:
   - compound corpus frequency is high
   - corpus frequency, family size and ambiguity of the head are low
   - corpus frequency, family size and ambiguity of the modifier are irrelevant

References


