Features of Compositionality in English and German Noun-Noun Compounds

Institut für Maschinelle Sprachverarbeitung

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

Noun-Noun Compounds and Compositionality

- **Noun-Noun Compounds:** complex words with two simplex nouns as constituents
  → left: **modifier** ⇒ ***fish*** *soup*
  → right: morphological **head** ⇒ *fish* ***soup***
- **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* / *Sündenbock*: 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

- **Corpus frequency**
  frequencies of compound, modifier and head in the web corpora EN-/DECOW14A (Schäfer and Bildhauer, 2012) (*en-/decow*)
- **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*, ...
- **Ambiguity**
  number of senses of the modifier and head from *WordNet/GermaNet*
- **Semantic relations**
  define how two nouns link to each other in a compound, e.g. *kitchen door* → *kitchen* HAVE *door*
  Relation annotation scheme used: by Ó 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 (868 compounds)
2. existing datasets enriched with missing features:
   - Schulte im Walde et al. (2013) *(German)*
   - Reddy et al. (2011) *(English)*
   - Ó Séaghdha (2007) *(English)* (part of 396 compounds)

Distributional Model of Compositionality

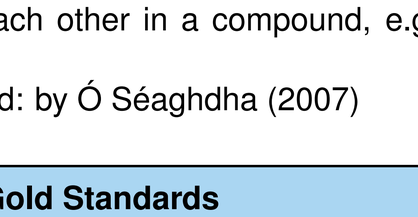

Figure 1: *Illustration for sematic space model of compositionality*

1. compute vectors for compounds and each of its constituents ⇒ 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 compound 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:

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

Results

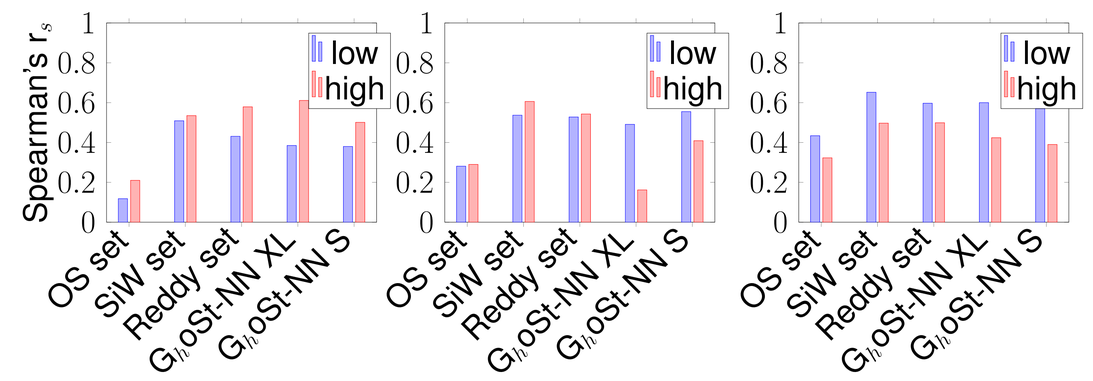

Figure 2: *results for a) compound b) modifier and c) head* ***corpus frequency***

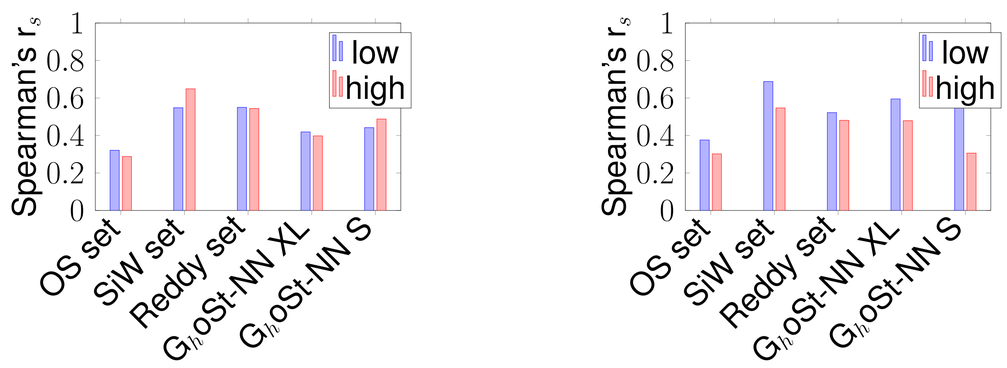

Figure 3: *results for a) modifier and b) head* ***constituent family size***

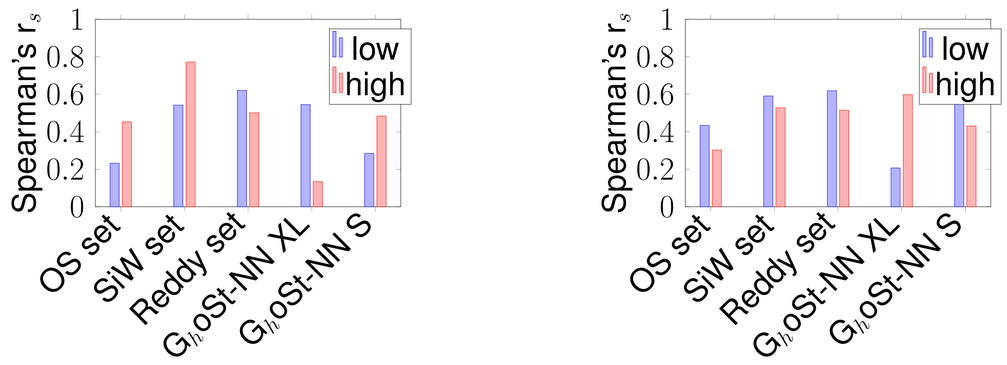

Figure 4: *results for a) modifier and b) head* ***ambiguity***

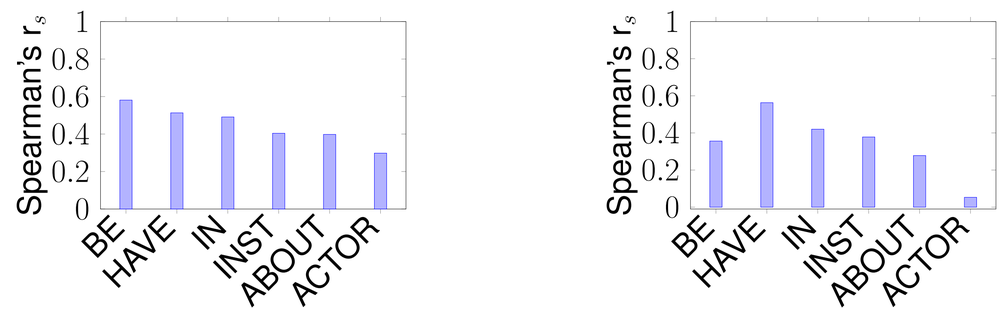

Figure 5: *results for* ***relations*** *of a) Ghost-NN and b) Ó Séaghdha (2007)*

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

Stefan Evert. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, Institute for Natural Language Processing (IMS), University of Stuttgart, 2005.

Diarmuid Ó Séaghdha. Designing and evaluating a semantic annotation scheme for compound nouns. In *Proceedings of the 4th Corpus Linguistics Conference*, Birmingham, UK, 2007.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. An empirical study on compositionality in compound nouns. In *Proceedings of The 5th International Joint Conference on Natural Language Processing*, 2011.

Roland Schäfer and Felix Bildhauer. Building large corpora from the web using a new efficient tool chain. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, 2012.

Sabine Schulte im Walde, Stefan Müller, and Stephen Roller. Exploring Vector Space Models to Predict the Compositionality of German Noun-Noun Compounds. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*, 2013.

Sidney Siegel and N. John (Jr) Castellan. *Nonparametric statistics for the behavioral sciences*. 1988.