An Assessment of Experimental Protocols for Tracing Changes in Word Semantics Relative to Accuracy and Reliability

Johannes Hellrich[1] **& Udo Hahn**[2]

1 Research Training Group "The Romantic Model. Variation - Scope - Relevance"

2 Jena University Language & Information Engineering (JULIE) Lab

Friedrich-Schiller Universität Jena, Jena, Germany

Tracing Word Semantics with WORD2VEC

- Skip-gram models perform very well for word similarity/relatedness tasks
- Can be used with historical corpora (such as Google Books Ngram) to track diachronic semantic changes
- Competing training protocols (sampling, continuous vs. independent, training algorithm)
- Reliability problem: slightly different results for repeated training on the same corpus

→ Problem for qualitative interpretation, is *romantic* about *melancholies, fanciful* or *lazzaroni* ?

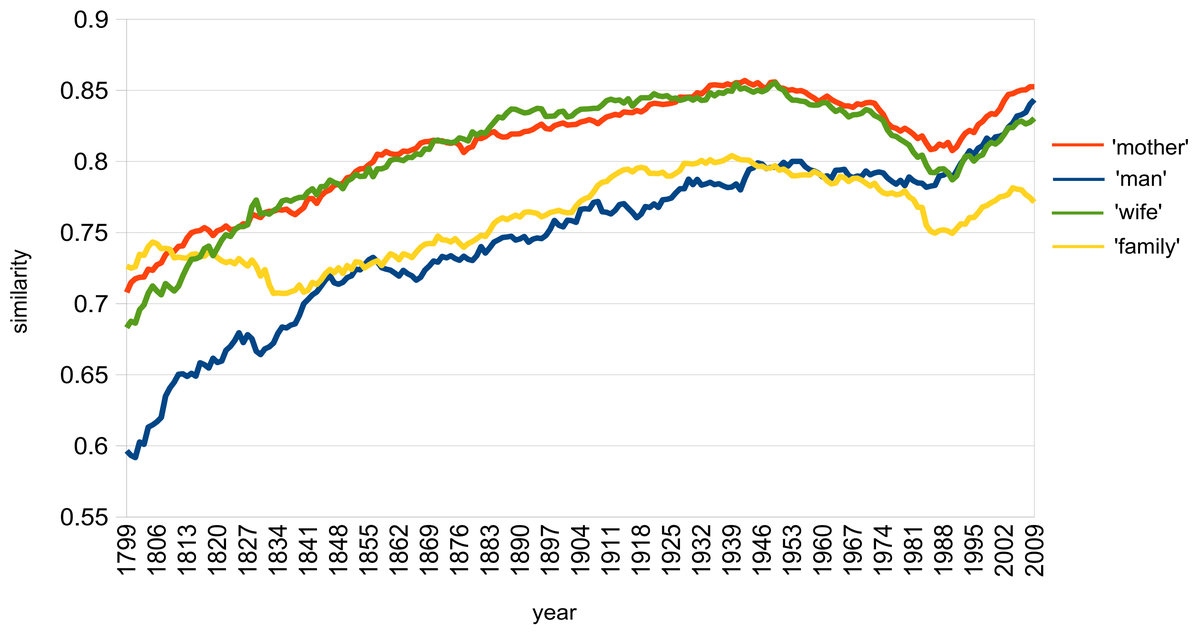

Figure 1: German *Frau 'woman'* during the last 200 years as an example for tracing word semantics. See Hellrich & Hahn, DH2016 for details.

Quantifying Reliability by Comparing Models

- Trained three models for each training protocol on English Fiction Google Books 5-grams for the beginning of the 20th century
- Reliability r := lexical overlap between word neighborhoods for each word
- See Figures 2–4 for more details on the best protocol from Table 1, i.e., all texts from 1900–1904

$$r@n := \frac{1}{t * n} \sum_{j=1}^{t} || \bigcap_{k=1}^{3} \{W_{1 \leq i \leq n, j, k}\} ||$$

Table 1: Top-n reliability and accuracy for different training protocols after 10 epochs.

Description of training protocol			top-n Reliability 1	2	3	4	5	Accuracy
independent	negative	in all texts	0.40	0.41	0.41	0.40	0.40	0.38
		in 10M sample	0.45	0.48	0.50	0.51	0.52	0.25
		between 10M samples	0.09	0.10	0.10	0.10	0.10	0.26
	hierarchical	in all texts	0.33	0.34	0.34	0.34	0.34	0.28
		in 10M sample	0.38	0.40	0.42	0.42	0.43	0.22
		between 10M samples	0.09	0.09	0.10	0.10	0.10	0.22 ± 0.01
continuous	negative	in 10M sample	0.54	0.55	0.56	0.56	0.57	0.25
		between 10M samples	0.21	0.21	0.22	0.22	0.22	0.25
	hierarchical	in 10M sample	0.31	0.32	0.32	0.32	0.33	0.22
		between 10M samples	0.12	0.13	0.13	0.13	0.13	0.23

Recommendations

- Use negative sampling instead of hierarchical softmax
- Avoid samples, use complete corpora (training time favors independent training)
- 4 up to 6 epochs provide quasi optimal reliability and accuracy
- Be skeptical when confronted with word neighborhoods

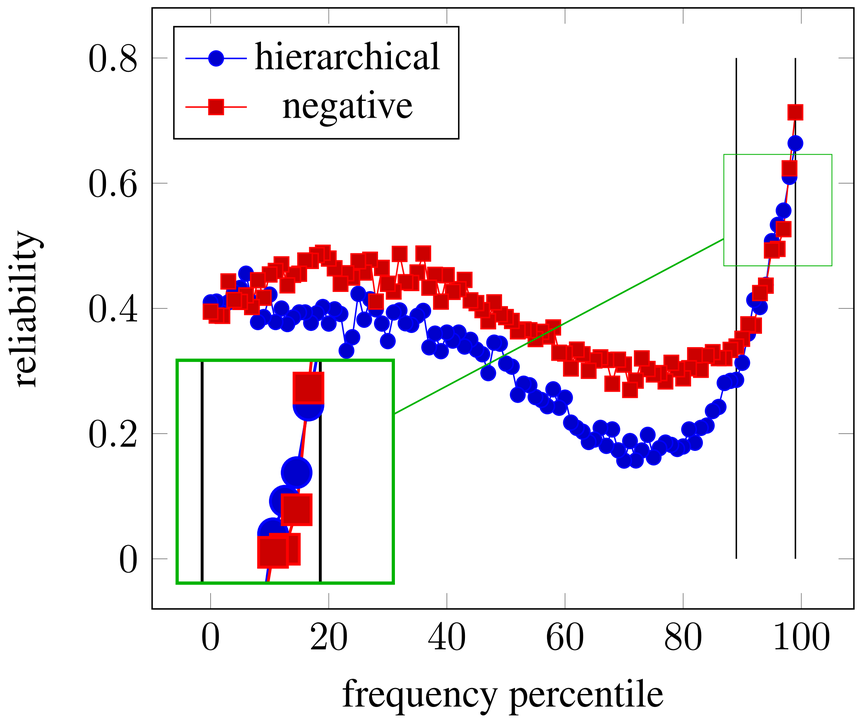

Figure 2: Influence of frequency on reliability. Vertical lines mark frequency of words known to have changed during the 20th century.

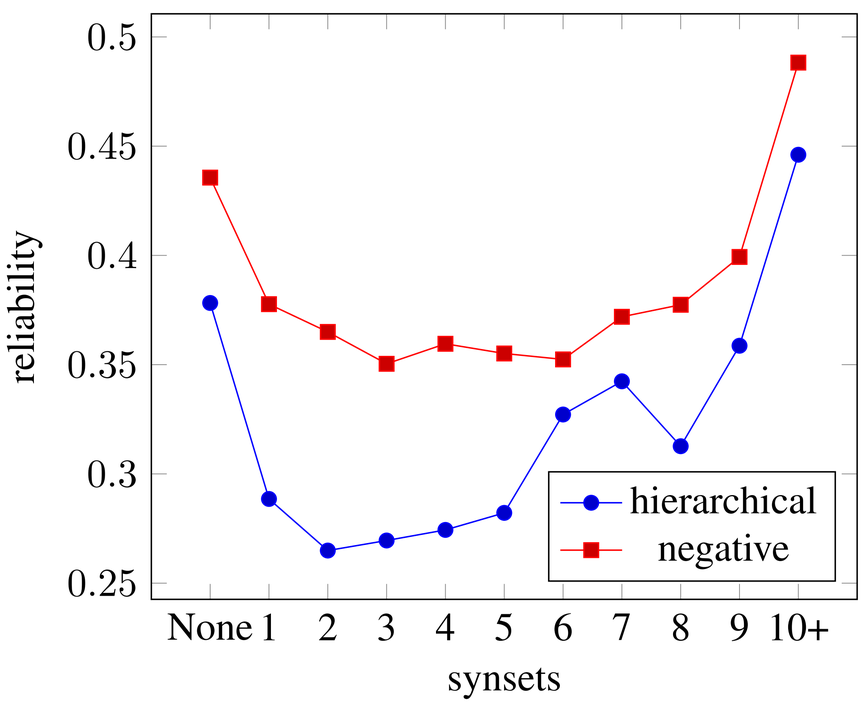

Figure 3: Influence of ambiguity (WORDNET synsets) on reliability.

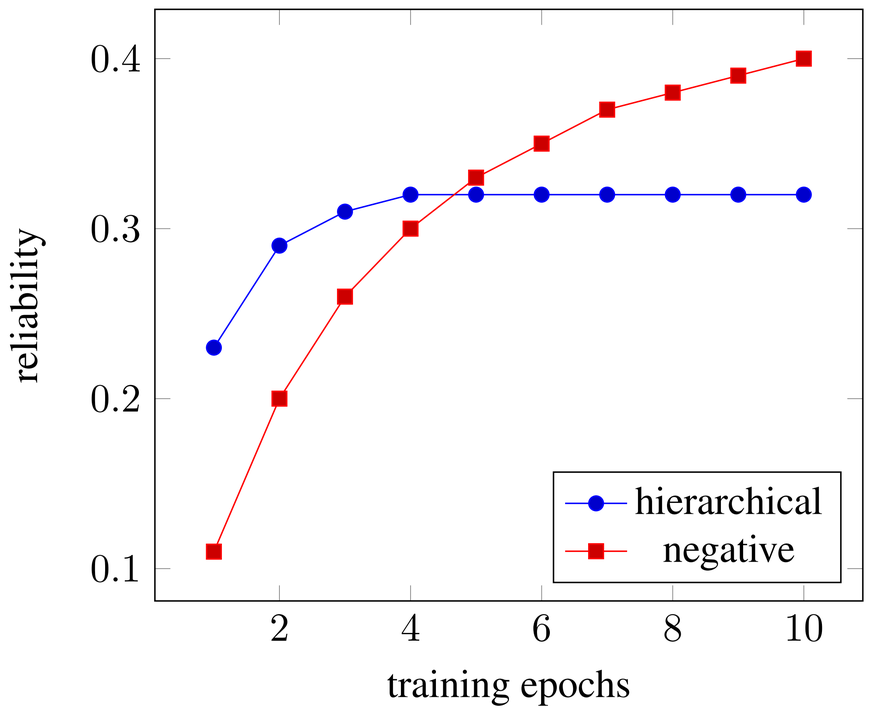

Figure 4: Influence of the number of training epochs on reliability.

Acknowledgments

This research was conducted within the Research Training Group „The Romantic Model. Variation – Scope – Relevance" supported by grant GRK 2041/1 from the Deutsche Forschungsgemeinschaft (DFG).

www.modellromantik.uni-jena.de