DGfS-CL

Jasper Brandes[1], Josef Ruppenhofer[1] and Michael Wiegand[2]
[1]Universität Hildesheim, [2]Universität des Saarlandes

Comparing methods for deriving polar intensity scores for adjectives

Within sentiment analysis, assessing the strength of subjective expressions is one of the central tasks besides subjectivity detection and polarity classification. Our particular interest is in the problem of acquiring prior or lexical polar intensity scores and rankings for sets of predicates that refer to the same scale. We do not address contextualized intensity. We focus on predicates evoking the same scale for the reason that only for these predicates context-independent relations hold: hearers can validly derive that a characterization of something as *great* entails that it is assessed as more positive by the speaker than had it been characterized as *good*. Such inferences are not supported across different scales.

Various approaches have been proposed in the past for automatically acquiring polar intensity scores of scalar predicates. While these methods tend to produce distinct scores, it is far from clear that humans would agree on complete orders for adjectives such as e.g. *bad* and *crappy*.

In this poster, we compare various approaches for deriving polar intensity scores for adjectives. We evaluate them against a human gold standard built from intensity ratings for two sets of adjectives elicited from native speakers recruited through Amazon Mechanical Turk.

In terms of methods, we apply both corpus-based and lexicon-based approaches. While the latter use given intensity scores for adjectives (Taboada et al., 2011; Warriner et al., 2013), the former exploit extrinsic metadata such as the star ratings that come with product reviews (Rill et al., 2012) or use language intrinsic phrase patterns indicating relative intensity relations between adjectives (Sheinman and Tokunaga, 2009). The two sets of adjectives on which we focus are 29 adjectives referring to quality taken from FrameNet's (Baker et al., 1998) DESIRABILITY frame and 18 relating to intelligence taken in part from the MENTAL_PROPERTY frame. Instead of merely relying on mean scores, our gold standard groups predicates of similar intensity.

Finally, we present an overview of the results from the evaluation, as well as discussing potential advantages of some methods over others.

References:

- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, pages 86–90, Montréal, Quebec, Canada.
- Sven Rill, Johannes Drescher, Dirk Reinel, Jörg Scheidt, Oliver Schütz, Florian Wogenstein, and Daniel Simon. 2012. A Generic Approach to Generate Opinion Lists of Phrases for Opinion Mining Applications. In *Proceedings of the KDD-Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)*, Beijing, China.
- Vera Sheinman and Takenobu Tokunaga. 2009. AdjScales: Differentiating between Similar Adjectives for Language Learners. *CSEDU*, 1:229–235.
- Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267 – 307.
- Amy Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. Behavior Research Methods, Online First:1–17.