HiER 2013

Griesbaum, Heuwing, Ruppenhofer, Werner (Hrsg.)

HiER 2013

Proceedings des 8. Hildesheimer Evaluierungs- und Retrievalworkshop

Hildesheim, 25.–26. April 2013

J. Griesbaum, B. Heuwing, J. Ruppenhofer, K. Werner (Hrsg.):
HIER 2013 - Proceedings des 8. Hildesheimer Evaluierungs-
und Retrievalworkshop, Hildesheim 2013

Defining a gold standard for polar intensity ordering

Jasper Brandes, Josef Ruppenhofer

Universität Hildesheim
Institut für Informationswissenschaft und Sprachtechnologie
Marienburger Platz 22, 31141 Hildesheim
{jasper.brandes,josef.ruppenhofer}@uni-hildesheim.de

Abstract

In this paper, we report on an effort to develop a gold standard for the intensity ordering of subjective adjectives. Rather than pursue a complete order as produced by paying attention to the mean scores of human ratings only, we take into account to what extent assessors consistently rate pairs of adjectives relative to each other. We show that different available automatic methods for producing polar intensity scores produce results that correlate well with our gold standard, and discuss some conceptual questions surrounding the notion of polar intensity.

1 Introduction

As illustrated prominently by Hatzivassiloglou & Wiebe (2000), the correlation between gradability and evaluativeness can be used to harvest evaluative adjectives based on their occurrence in patterns that relate to gradability (e.g.*more/less entertaining*). The task that concerns us specifically is how to order gradable evaluative predicates that refer to the same scale relative to each other: how can one find out whether e.g. moronic is more negative than *stupid*?

In this paper, we address two aspects of the problem of intensity ordering. The first one is the fundamental question what a gold standard for this task should look like. Various researchers have proposed ways to automatically

assign valence scores to evaluative predicates. The methods used tend to produce distinct scores, and thereby absolute orderings, for all predicates. However, it is far from clear that humans would agree on complete orders between predicates such as *dumb*, *smart*, *stupid*. In fact, Kennedy & McNally (2005) suggest that by one set of linguistic criteria one can only distinguish between two major types of scale structures for adjectives: open and closed scales, depending on whether adjectives can be modified by intensifiers such *very* and *quite* (open scales) or by intensifiers such *utterly* and *absolutely* (closed scales).

We develop a gold standard that forms a sort of compromise between the complete orders that natural language processing techniques tend to produce and the bipartition that theoretical linguistics suggests. The second interest of our work is to compare how closely the results of different automatic methods of assigning intensity scores correlate with the human gold standard. Both our proposal for producing a gold standard and the evaluation of available methods are preliminary in that we have worked so far only with one scalar semantic domain, namely the degrees of intelligence associated with 18 English adjectives (see section 3 for a list).

2 Related Work

Acquiring knowledge about the polar strength of subjective expressions (*how positive/negative?*) has been pursued in three ways.

One obvious way is corpus-based, using the distributional properties of the scalar items to be scored or ranked in order to acquire intensity ratings. This approach has two important subtypes. In one, represented e.g. by Rill et al. (2012), scores for subjective expressions are derived from an extrinsic source, the star ratings metadata associated with reviews. In the second subtype, represented by Sheinman and Tokunaga (2009), language-intrinsic properties are used to assign intensity scores.

A second way of inducing scores or rankings consists in exploiting the knowledge inherent in lexical resources. SentiWordNet (Bacchianella et al. 2010), based on WordNet (Miller et al. 1990), is a prominent example of this approach. The main limitation of this approach is that it has to rely on the correctness and coverage of the existing resource.

A third way to acquire intensity ratings consists in re-purposing so-called affective norm data elicited from human subjects by psychologists. The elicitation usually involves the rating of individual words presented out of context. When done at large scale by way of crowd-sourcing to non-experts, as practiced by e.g. Warriner et al. (2013) using Amazon Mechanical Turk, affective norms promise good coverage. A practical limitation of this approach is, however, that no crowd may be available for the language of interest.

3 Constructing a Human Gold Standard

16 native speakers and 2 near-natives[1] were asked to assess the intensity of each of 18 intelligence-related adjectives in a fixed context. We used this setup so as to derive a ranking that would contain much less noise than automatic approaches, where factors like ambiguity and (lack of) coverage can play quite a substantial role. The adjectives we worked with are mainly culled from FrameNet's (Baker et al. 1998) MENTAL_PROPERTIES frame:

brainless, brainy, bright, brilliant, daft, dim, dimwitted, dumb, foolish, idiotic, imbecilic, inane, ingenious, intelligent, mindless, moronic, smart, stupid.

The survey was conducted online using a local installation of the Limesurvey[2] tool and consisted of three groups of questions. The first collected demographic information; the second elicited the ratings for the survey items (the adjectives); the third gave participants the chance to give feedback on the task. Ratings were elicited for each adjective individually, in randomized order, and under conditions intended to minimize bias. Specifically, to assess

1 The majority of the participants (n=9) claimed to speak American English (AE); 6 identified with the Australian variety (AuE) and 1 with New Zealand English (NZE).

2 www.limesurvey.org

the intensity of an adjective, the participant was asked to use a slider which could be moved from left to right by clicking the slider, dragging it in the desired direction and releasing the mouse click at the desired intensity. The slider scale ranged from -100 on the left to +100 on the right. The scores' meaning was described as follows: "Complete lack of intellectual characteristics equals -100, whereas extremely high intellectual characteristics equal +100." Participants were also given the following instruction on how to contextualize the meaning of the adjective: "Please rate the following adjective describing the intellectual characteristics of utterances and persons according to the intensity conveyed by it."

The results of our data elicitation from 18 subjects are shown in Figure 1.[3] Two key aspects of the data are that a) each adjective has a different mean rating and b) the overlapping standard deviations suggest that not all adjectives can really be distinguished from each other with respect to intensity.

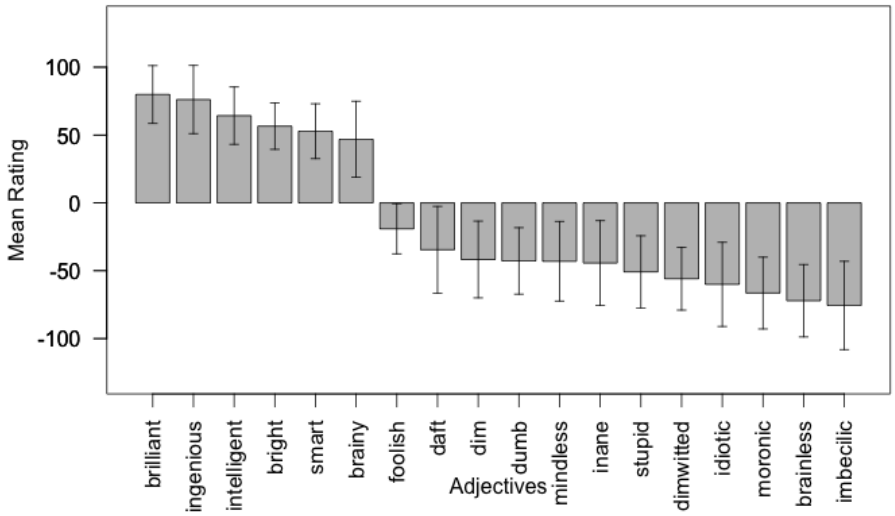

Figure 1 Human Evaluation: Means and Standard Deviations

3 24 subjects participated in the online survey. Three participants terminated the survey early, resulting in 21 full responses. Out of these 21 full responses, three were not included in the subsequent analysis. One response was a test run. For the two others it was evident from the ratings and the feedback on the survey that the participants had not entirely understood what they were expected to do.

In order to obtain some distinctions between, and grouping of, our adjectives by strength despite the overlapping standard deviations of the mean values, we decided to look at how consistently pairs of adjectives were rated across all participants.[4] Thus, for a given pair of adjectives we counted how many participants rated adjective A more intense than adjective B; adjective B more intense than adjective A; or adjective A equally intense as adjective B. With this data in hand, we defined a threshold of 12 out of the 18 raters who had to agree on an unequal rating so that we would adopt that as our relative ranking for the two adjectives.[5] Based on the predominant ordering relations for the pairwise comparison of the adjectives derived from the ratings for the individual adjectives, the six positive adjectives can be divided into three subgroups and the negative adjectives into 4, as shown in Figure 2.[6]

Intensity	**Subgroup of adjectives**
high positive	brilliant & ingenious
medium positive	intelligent
low positive	brainy, bright & smart
⇕	
low negative	daft & foolish
medium low negative	dim, dumb, inane, mindless & stupid
medium high negative	dim-witted & idiotic
high negative	brainless, imbecilic & moronic

Figure 2 Adjective Intensity Subgroups

4 Note that the comparison between two adjectives was not part of the survey but was constructed after the survey ended using simple combinatorics; only combinations of two adjectives with identical polarity needed to be investigated.

5 The number of participants needed for a qualified majority of raters is a tuneable parameter. In this first exploration, we decided to err on the side of being conservative and set it higher than the minimum number of 10.

6 In our data-set, we had no need to deal with any circular strength relations. But we feel that for wider application, we will need to be able to handle such cases.

4 A Comparison of Methods for Deriving Intensity scores

As discussed earlier, multiple methods for producing intensity scores have been developed. Here, we want to compare two of them relative to our gold standard for intelligence adjectives, namely one that relies on intensity scores derived extrinsically, from review metadata, and one that re-purposes ratings elicited by psychologists as affective norms.

4.1 Mean star ratings

The basic intuition of this method is that words in a review and the score of the review (or, its star rating) correlate: positive words are more likely to occur in positive reviews, and negative words more likely to occur in negative reviews. Thus, along the lines of Rill et al. (2012), we counted how many instances *n* of each adjective *i* of the set of intelligence adjectives occur in reviews with a given star rating *j* (score) in a corpus of Amazon reviews (Prettenhofer & Stein 2010).

$$SR_i = \frac{\sum_{j=1}^{n} S_j^i}{n}$$

Note that we slightly modify the approach of Rill et al. (2012): while they only considered the language in the review titles, we used only the body of the reviews. The reason for this is that we hope to benefit from a) greater coverage and b) more data points per adjective by using the body of the reviews rather than the titles.

4.1 Arousal scores

The affective ratings for almost 14,000 English words collected by Warriner et al. (2013) include scores of valence (from unhappy to happy), arousal (from calm to aroused) and dominance (from in control to controlled) for each word in the list. Scales from 1 to 9 were used for each of the ranges. This three-variable scoring system follows the dimensional theory of emotion by Osgood et al. (1957).

The scores for the words were elicited on the crowd-sourcing site Amazon Mechanical Turk from participants who needed to be US-residents. In each of the assignments (a set of about 350 words to be rated), the participants rated the words only along one of the three distinct dimensions (valence, arousal, dominance). Note that lemmas were assessed and not word senses. For our purposes, we will interpret the valence score as a relevant clue to the polar intensity score. The valence score was elicited as follows:

> You are invited to take part in the study that is investigating emotion, and concerns how people respond to different types of words. You will use a scale to rate how you felt while reading each word. There will be approximately 350 words. The scale ranges from 1 (happy) to 9 (unhappy).

As the focus on how *happy* the subjects *felt* indicates, the valence construct in theory is not defined identically to the notion of intensity we used since ours targets the degree of presence/lack of the underlying valued characteristic (intelligence).

4.3 Results

Table 1 shows the results of a Spearman rank correlation analysis between the gold standard (using the version in which some adjectives are grouped together into ordered ranks based on rater consistency), the mean star ratings and the affective norm valence scores, the latter two suitably transformed into ranks.[7]

Table 1: Correlations between intensities derived from different data

Data Sets	Correlation
Amazon Mean Star Ratings – Our data	0.837
Affective Norms Valence – Our data	0.673
Amazon Mean Star Rating – Affective Norms Valence	0.749

7 The correlation coefficient *rho* ranges between -1 and +1 for perfect negative and perfect positive correlation, respectively.

The results of the analysis show that both ways of deriving intensity scores produce good correlations with our human ratings. Interestingly, the reviews yield a better correlation than the affective norms. This is somewhat surprising in two respects. In the affective norms elicitation the relevance of the rating to specific lexical items was part of the design whereas it was indirect in the review data. Second, our use of occurrences in the text body involves an even more indirect relation between the star rating and the word occurrences than that used by Rill et al. (2012). Those authors focused specifically on occurrences of words in review titles rather than in the review text body based on the assumption that the words in the review titles would be more likely to have a clear connection to the star rating. Nonetheless, our overall results by using text bodies seem to be very good. In future work, we will seek to compare the usefulness of the body text of reviews relative to that of review titles by running parallel scoring experiments for multiple sets of adjectives from different semantic domains.

Finally, we also show the correlation between the valence scores from the affective norms dataset and the amazon star ratings. Considering this correlation is of interest for the following reason. Recall that the affective norms data elicitation used a different setup and asked specifically about how participants felt upon reading the word, whereas our instructions asked subjects to "[p]lease rate the following adjective describing the intellectual characteristics of utterances and persons according to the intensity conveyed by it". Our elicitation method may thus have focused more on the "objective" quality of intelligence rather than its evaluation. If that were so, then maybe Warriner et al.'s valence score should be considered the gold standard and our elicitation method one way of approximating it. On that understanding, our approach still fares well but worse than the approach that derives polar intensity scores from Amazon reviews (0.673 vs. 0.749).

Conclusion

In this paper, we presented an approach to deriving a human gold standard for polar intensity scores, using adjectives related to intelligence as our test case. Rather than producing a total order by only using the mean scores of our elicited responses, and rather than not making any distinctions due to the

relatively large standard deviations for all adjectives, we induce some subgroups by paying attention to the cases where a qualified majority of raters observe an asymmetric relation between the intensities associated with different adjectives.

We then used our gold standard to compare two ways of harvesting polar intensity scores. One uses review metadata and one interprets valence scores collected in a large affective norming experiment as polar intensity scores. Both methods and data sources produce good correlations with our gold standard, with the corpus-based approach being even better. What is attractive about the corpus-based approach is that it can run automatically and does not require the time and effort of elicitation. Nevertheless, since the review-based approach relies on a large review collection, such reviews need to be available in the first place. Further, and more significantly, there may be problems of coverage within collections of review data when one is interested in other domains or even general language vocabulary.

While the two approaches investigated here were shown to work for intelligence adjectives, we have not demonstrated their usefulness for other semantic domains. Further investigation of these and other methods for deriving polar intensity scores is needed.

Finally, we threw up a methodological and conceptual problem that is usually not directly addressed: when collecting human ratings for use as a gold standard, how should the task be stated to the participants? In most cases, having more of a desirable property correlates directly with more positive affect about the degree to which the property is exhibited. However, that need not be so: somebody described as a *know-it-all* in English or as *neunmalklug* in German may exhibit a very high degree of informedness but might still not be appreciated for it. For that reason, we will compare in future work for our intelligence adjectives and further sets of adjectives how well different ways of eliciting intensity information correlate.

References

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, May. European Language Resources Association (ELRA).

Baker, C.F., Fillmore, C.J. and Lowe, J.B. (1998) The Berkeley Framenet Project. In: Proceedings of the 17th international conference on Computational linguistics, Vol 1.86--90.

Hatzivassiloglou, V. and Wiebe, J. (2000). Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In:Proceedings of the International Conference on Computational Linguistics (COLING-2000).

Kennedy, C. and McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. Language, pages 345–381.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. International journal of lexicography, 3(4), 235-244.

Osgood, C.E., Suci, G.J. and Tannenbaum, P.H. (1957). The measurement of meaning. Vol. 47. Urbana: University of Illinois Press, 1957.

Prettenhofer, P. and Stein, B. (2010). Cross-Language Text Classification using Structural Correspondence Learning. In: 48th Annual Meeting of the Association of Computational Linguistics (ACL 10), 1118-1127, July 2010.

Rill, S., Scheidt, J., Drescher, J., Schütz, O., Reinel, D., and Wogenstein, F. (2012). A generic approach to generate opinion lists of phrases for opinion mining applications. In: Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, WISDOM '12, pages 7:1–7:8. ACM.

Sheinman, V. and Tokunaga, T. (2009). AdjScales: Differentiating between Similar Adjectives for Language Learners. CSEDU (1): 229-235.

Warriner, A., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. http://crr.ugent.be/papers/Warriner_et_al_affective_ratings.pdf (accessed March 25, 2013).