Comparing methods for deriving intensity scores for adjectives

Josef Ruppenhofer*, **Michael Wiegand**†, **Jasper Brandes***
*Hildesheim University
Hildesheim, Germany
{ruppenho|brandesj}@uni-hildesheim.de
†Saarland University
Saarbrücken, Germany
michael.wiegand@lsv.uni-saarland.de

Abstract

We compare several different corpus-based and lexicon-based methods for the scalar ordering of adjectives. Among them, we examine for the first time a low-resource approach based on distinctive-collexeme analysis that just requires a small predefined set of adverbial modifiers. While previous work on adjective intensity mostly assumes one single scale for all adjectives, we group adjectives into different scales which is more faithful to human perception. We also apply the methods to both polar and non-polar adjectives, showing that not all methods are equally suitable for both types of adjectives.

1 Introduction

Ordering adjectives by strength (e.g. *good* < *great* < *excellent*) is a task that has recently received much attention due to the central role of intensity classification in sentiment analysis. However, the need to assess the relative strength of adjectives also applies to non-polar adjectives. We are thus interested in establishing prior or lexical intensity scores and rankings for arbitrary sets of adjectives that evoke the same scale.[1] We do not address contextualized intensity, i.e. the fact that e.g. negation and adverbs such as *very* or *slightly* impact the perceived intensity of adjectives.

We work with four scales of adjectives (cf. Table 1). Our polar adjectives include 29 adjectives referring to quality and 18 adjectives relating to intelligence. Our non-polar adjectives include 8 dimensional adjectives denoting size and 22 denoting duration. The adjectives are taken, in part, from FrameNet's (Baker et al., 1998) frames for DESIRABILITY, MENTAL PROPERTY, SIZE and DURATION DESCRIPTION. These scales are used because they are prototypical and have multiple members on the positive and negative half-scales.

We evaluate several corpus- and resource-based methods that have been used to assign intensity scores to adjectives. We compare them to a new corpus-based method that is robust and of low complexity, and which directly uses information related to degree modification of the adjectives to be orderered. It rests on the observation that adjectives with different types of intensities co-occur with different types of adverbial modifiers.[2]

POLAR ADJECTIVES	
Intelligence Adjs.	**Intensity Level**
brilliant	very high positive
ingenious	high positive
brainy, intelligent	medium positive
smart	low positive
bright	very low positive
daft	very low negative
foolish	low negative
inane	lower medium negative
dim	upper medium negative
dim-witted, dumb, mindless	high negative
brainless, idiotic, imbecillic, moronic, stupid	very high negative
Quality Adjs.	**Intensity Level**
excellent, extraordinary, first-rate, great, outstanding, super, superb, superlative, tip-top, top-notch	very high positive
good	high positive
decent	upper medium positive
fine, fair	lower medium positive
okay, average	low positive
so-so	very low positive
mediocre	very low negative
second-rate, substandard	low negative
inferior	lower medium negative
bad, crappy, lousy, poor, third-rate	medium negative
rotten	upper medium negative
awful	high negative
shitty	very high negative
DIMENSIONAL ADJECTIVES	
Size Adjs.	**Intensity Level**
colossal, enormous, gargantuan, giant, gigantic, gi-normous, humongous	high positive
big, huge, immense, large, oversize, oversized, vast	medium positive
outsize, outsized	low positive
diminutive, little, puny, small	low negative
tiny	medium negative
microscopic	high negative
Duration Adjs.	**Intensity Level**
long	high positive
lengthy	medium positive
extended	low positive
momentaneous	low negative
brief, fleeting, momentary	medium negative
short	high negative

Table 1: Adjectives used grouped by human gold standard intensity classes

[1] As there has been previous work on how to group adjectives into scales (Hatzivassiloglou and McKeown, 1993), we consider this grouping as given.

[2] The ratings we collected and our scripts are available at www.uni-hildesheim.de/ruppenhofer/data/DISA_data.zip.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 117–122, Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

2 Data and resources

Table 2 gives an overview of the different corpora and resources that we use to produce the different scores and rankings that we want to compare. The corpora and ratings will be discussed alongside the associated experimental methods in §4.1 and §4.2.

Corpora	Tokens	Reference
BNC	∼112 M	(Burnard, 2007)
LIU reviews	∼1.06 B	(Jindal and Liu, 2008)
ukWaC	∼2.25 B	(Baroni et al., 2009)
Resources	**Entries**	**Reference**
Affective norms	∼14 K	(Warriner et al., 2013)
SoCAL	∼ 6.5 K	(Taboada et al., 2011)
SentiStrength	∼ 2.5 K	(Thelwall et al., 2010)

Table 2: Corpora and resources used

3 Gold standard

We collected human ratings for our four sets of adjectives. All items were rated individually, in randomized order, under conditions that minimized bias. Participants were asked to use a horizontal slider, dragging it in the desired direction, representing polarity, and releasing the mouse at the desired intensity, ranging from -100 to $+100$.

Through Amazon Mechanical Turk (AMT), we recruited subjects with the following qualifications: US residency, a HIT-approval rate of at least 96% (following Akkaya et al. (2010)), and 500 prior completed HITs. We collected 20 ratings for each item but had to exclude some participants' answers as unusable, which reduced our sample to 17 subjects for some items. In the raw data, all adjectives had different mean ratings and their standard deviations overlapped. We therefore transformed the data into sets of equally strong adjectives as follows. For a given pair of adjectives of identical polarity, we counted how many participants rated adjective A more intense than adjective B; B more intense than A; or A as intense as B. Whenever a simple majority existed for one of the two unequal relations, we adopted that as our relative ranking for the two adjectives.[3] The resulting rankings (intensity levels) are shown in Table 1.

4 Methods

Our methods to determine the intensity of adjectives are either corpus- or lexicon-based.

[3]In our data, there was no need to break circular rankings, so we do not consider this issue here.

4.1 Corpus-based methods

Our first method, **distinctive-collexeme analysis (Collex)** (Gries and Stefanowitsch, 2004) assumes that adjectives with different types of intensities co-occur with different types of adverbial modifiers (Table 3). End-of-scale modifiers such as *extremely* or *absolutely* target adjectives with a partially or fully closed scale, such as *brilliant* or *outstanding*, which occupy extreme positions on the intensity scale. "Normal" degree modifiers such as *very* or *rather* target adjectives with an open scale structure (in the sense of Kennedy and McNally (2005)), such as *good* or *decent*, which occupy non-extreme positions.

To determine an adjective's preference for one of the two constructions, the Fisher exact test (Pedersen, 1996) is used. It makes no distributional assumptions and does not require a minimum sample size. The direction in which observed values differ from expected ones indicates a preference for one construction over the other and the p-values are taken as a measure of the preference strength. Our hypothesis is that e.g. an adjective A with greater preference for the end-of-scale construction than adjective B has a greater inherent intensity than B. We ran distinctive-collexeme analysis on both the ukWaC and the BNC. We refer to the output as **Collex**$_{ukWaC}$ and **Collex**$_{BNC}$. Note that this kind of method has ***not*** yet been examined for automatic intensity classification.

end-of-scale	"normal"
100%, fully, **totally**, absolutely, **completely**, perfectly, entirely, utterly, **almost**, partially, half, mostly	all, as, awfully, enough, extremely, fairly, highly, how, least, less, much, pretty, quite, rather, **so**, somewhat, sort of, terribly, **too**, **very**, well

Table 3: Domain independent degree modifiers (3 most freq. terms in the **BNC**; 3 most freq. terms in the ukWaC)

Another corpus-based method we consider employs **Mean star ratings (MeanStar)** from product reviews as described by Rill et al. (2012). Unlike Collex, this method uses no linguistic properties of the adjectives themselves. Instead, it derives intensity from the star rating scores that reviewers (manually) assign to reviews. We count how many instances of each adjective i (of the set of adjectives to classify) occur in review titles with a given star rating (score) S_j within a review corpus. The intensity score is defined as the weighted mean of the star ratings $SR_i = \frac{\sum_{j=1}^{n} S_j^i}{n}$.

Horn (1976) proposes **pattern-based diagnos-**

Pattern	Any	Int.	Qual.	Size	Dur.
X or even Y	4118	1	34	9	3
X if not Y	3115	1	0	29	0
be X but not Y	2815	0	74	3	1
not only X but Y	1114	0	3	0	0
X and in fact Y	45	0	0	0	0
not X, let alone Y	4	0	0	0	0
not Y, not even X	4	0	1	0	0

Table 4: Phrasal patterns in the ukWaC

tics for acquiring information about the scalar structure of adjectives. This was validated on actual data by Sheinman and Tokunaga (2009). A pattern such as *not just/only X but Y* implies that [Y] must always be stronger than [X] (as in *It's not just good but great.*).

The pattern-based approach has a *severe* coverage problem. Table 4 shows the results for 7 common phrasal patterns in the larger of our two corpora, the ukWaC. The slots in the patterns are typically *not* filled by adjectives from the same scale. For example, the most frequent pattern *X or even Y* has 4118 instances in the ukWaC. Only 34 of these have quality adjectives in both slots. Though de Melo and Bansal (2013) have shown that the coverage problems can be overcome and state-of-the-art results obtained using web scale data in the form of Google n-grams, we still set aside this method here because of its great resource need.

4.2 Manually compiled lexical resources

In addition to the corpus methods, we also consider some manually compiled resources. We want to know if the polarity and intensity information in them can be used for ordering polar adjectives.

One resource we consider are the affective ratings (elicited with AMT) for almost 14,000 English words collected by Warriner et al. (2013). They include scores of valence (*unhappy* to *happy*), arousal (*calm* to *aroused*) and dominance (*in control* to *controlled*) for each word in the list. This scoring system follows the dimensional theory of emotion by Osgood et al. (1957). We will interpret each of these dimensions as a separate intensity score, i.e. **War**$_{Val}$, **War**$_{Aro}$ and **War**$_{Dom}$.

Beyond Warriner's ratings, we consider the two polarity lexicons **SentiStrength** (Thelwall et al., 2010) and **SoCAL** (Taboada et al., 2011) which also assign intensity scores to polar expressions.

5 Experiments

For our evaluation, we compute the similarity between the gold standard and every other ranking we are interested in in terms of Spearman's rank correlation coefficient (Spearman's ρ).

	Polar		Dimensional	
Data set	**Intelligence**	**Quality**	**Duration**	**Size**
MeanStar	0.886	0.935	0.148	-0.058
SoCAL	0.848	0.953	NA	0.776
SentiStrength	0.874	0.880	NA	NA
Collex$_{ukWaC}$	0.837	0.806	0.732	0.808
Collex$_{ukWaC*}$	0.845	0.753	0.732	0.940
Collex$_{BNC}$	0.834	0.790	0.732	0.733
Collex$_{BNC*}$	0.705	0.643	0.834	0.700
War$_{Val}$	0.779	0.916	-0.632	-0.031
War$_{Aro}$	0.504	-0.452	0.316	0.717
War$_{Dom}$	0.790	0.891	0.632	0.285

Table 5: Spearman rank correlations with the human gold standard (*: only the 3 most frequent modifiers are used (see Table 3))

5.1 Data transformation

For the word lists with numeric scores (MeanStar (§4.1); SentiStrength, SoCAL, War$_{Val}$, War$_{Aro}$ and War$_{Dom}$ (§4.2)) we did as follows: Adjectives not covered by the word lists were ignored. Adjectives with equal scores were given tied ranks.

For the experiments involving distinctive collexeme analysis in our two corpora (§4.1) we proceeded as follows: The adjectives classified as distinctive for the end-of-scale modification constructions were put at the top and bottom of the ranking according to polarity; the greater the collostructional strength for the adjective as denoted by the p-value, the nearer it is placed to the top or bottom of the ranking. The adjectives that are distinctive for the normal degree modification construction are placed between those adjectives distinctive for the end-of-scale modification construction, again taking polarity and collostructional strength into account. This time, the least distinctive lemmas for the normal modification construction come to directly join up with the least distinctive lemmas for the end-of-scale construction. In between the normal modifiers, we place adjectives that have no preference for one or the other construction, which may result from non-occurrence in small data sets (see §5.2).

5.2 Results

The results of the pairwise correlations between the human-elicited gold standard and the rankings derived from various methods and resources are shown in Table 5. For polar adjectives, most rankings correlate fairly well with human judgments. Warriner's arousal list, however, performs poorly on quality adjectives, whereas MeanStar and Warriner's dominance and valence lists perform better on quality than on intelligence adjectives. For MeanStar, this does not come as a surprise as quality adjectives are much more frequent in prod-

uct reviews than intelligence adjectives. Overall, it seems that MeanStar most closely matches the human judgments that we elicited for the intelligence adjectives. SentiStrength also produces high scores. However, we do not have full confidence in that result since SentiStrength lacks many of our adjectives, thus leading to a possibly higher correlation than would have been achieved if ranks (scores) had been available for all adjectives.

The picture is very different for the dimensional (non-polar) adjectives. While Collex still gives very good results, especially on the ukWaC, the MeanStar method and most Warriner lists produce very low positive or even negative correlations. This shows that estimating the intensity of non-polar adjectives from metadata or ratings elicited in terms of affect is not useful. It is much better to consider their actual linguistic behavior in degree constructions, which Collex does. SentiStrength has no coverage for size or duration adjectives. SoCAL covers 14 of the 22 size adjectives.

Although it never gives the best result, Collex produces stable results across both corpora and the four scales. It also requires the least human effort by far. While all other rankings are produced with the help of heavy human annotation (even MeanStar is completely dependent on manually assigned review scores), one has only to specify some *domain-independent* degree and end-of-scale modifiers. Table 5 also shows that normally a larger set of modifiers is necessary: only considering the 3 most frequent terms (Table 3) results in a notably reduced correlation. As there is no consistent significant difference between Collex_{BNC} and Collex_{ukWaC} even though the ukWaC is 20 times larger than the BNC (Table 2), we may conclude that the smaller size of the BNC is already sufficient. This, however, raises the question whether even smaller amounts of data than the full BNC could already produce a reasonable intensity ranking. Figure 1 plots the Spearman correlation for our adjectives using various sizes of the BNC corpus.[4] It shows that further reducing the size of the corpus causes some deterioration, most significantly on the intelligence adjectives. The counter-intuitive curve for duration adjectives is explained as follows. Collex produces ties in the middle of the scale when data is lacking (see §5.1). Because the smallest corpus slices contain no or very few instances and because the gold standard does include several ties, the results for duration adjectives are inflated initially, when data is lacking.

[4]For each size, we average across 10 samples.

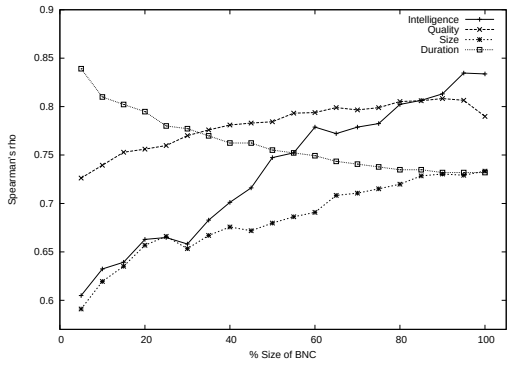

Figure 1: Reducing the size of the BNC

6 Related work

Sentiment analysis on adjectives has been extensively explored in previous work, however, most work focussed on the extraction of subjective adjectives (Wiebe, 2000; Vegnaduzzo, 2004; Wiegand et al., 2013) or on the detection of polar orientation (Hatzivassiloglou and McKeown, 1997; Kamps et al., 2004; Fahrni and Klenner, 2008).

Intensity can be considered in two ways, as a contextual strength analysis (Wilson et al., 2004) or as an out-of-context analysis, as in this paper.

Our main contribution is that we compare several classification methods that include a new effective method based on distinctive-collexeme analysis requiring hardly any human guidance and which moreover can solve the problem of intensity assignment for all, not only polar adjectives.

7 Conclusion

We compared diverse corpus-based and lexicon-based methods for the intensity classification of adjectives. Among them, we examined for the first time an approach based on distinctive-collexeme analysis. It requires only a small predefined set of adverbial modifiers and relies only on information about individual adjectives rather than co-occurrences of adjectives within patterns. As a result, it can be used with far less data than e.g. the Google n-grams provide. Unlike the mean star approach, it needs no extrinsic meta-data and it can handle both polar and non-polar adjectives. Accordingly, it appears to be very promising for cases where only few resources are available and as a source of evidence to be used in hybrid methods.

Acknowledgments

Michael Wiegand was funded by the German Federal Ministry of Education and Research (BMBF) under grant no. 01IC12SO1X. The authors would like to thank Maite Taboada for providing her sentiment lexicon (SoCAL) to be used for the experiments presented in this paper.

References

Cem Akkaya, Alexander Conrad, Janyce Wiebe, and Rada Mihalcea. 2010. Amazon Mechanical Turk for Subjectivity Word Sense Disambiguation. In *NAACL-HLT 2010 Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk*, pages 195–203, Los Angeles, CA, USA.

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.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetti. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Lou Burnard, 2007. *Reference Guide for the British National Corpus*. Research Technologies Service at Oxford University Computing Services, Oxford, UK.

Gerard de Melo and Mohit Bansal. 2013. Good, Great, Excellent: Global Inference of Semantic Intensities. *Transactions of the Association for Computational Linguistics*, 1:279–290.

Angela Fahrni and Manfred Klenner. 2008. Old Wine or Warm Beer: Target Specific Sentiment Analysis of Adjectives. In *Proceedings of the Symposium on Affective Language in Human and Machine*, pages 60–63, Aberdeen, Scotland, UK.

Stefan Th. Gries and Anatol Stefanowitsch. 2004. Extending collostructional analysis: a corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics*, 9(1):97–129.

Vasileios Hatzivassiloglou and Kathleen McKeown. 1993. Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 172–182, Columbus, OH, USA.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the Semantic Orientation of Adjectives. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 174–181, Madrid, Spain.

Laurence Robert Horn. 1976. *On the Semantic Properties of Logical Operators in English*. Indiana University Linguistics Club.

Nitin Jindal and Bing Liu. 2008. Opinion Spam and Analysis. In *Proceedings of the international conference on Web search and web data mining (WSDM)*, pages 219–230, Palo Alto, USA.

Jaap Kamps, M.J. Marx, Robert J. Mokken, and Maarten De Rijke. 2004. Using Wordnet to Measure Semantic Orientations of Adjectives. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, pages 1115–1118, Lisbon, Portugal.

Christopher Kennedy and Louise McNally. 2005. Scale Structure, Degree Modification, and the Semantics of Gradable Predicates. *Language*, 81(2):345–338.

Charles E. Osgood, George Suci, and Percy Tannenbaum. 1957. *The Measurement of Meaning*. University of Illinois Press.

Ted Pedersen. 1996. Fishing for exactness. In *Proceedings of the South-Central SAS Users Group Conference*, Austin, TX, USA.

Sven Rill, Johannes Drescher, Dirk Reinel, Joerg Scheidt, Oliver Schuetz, 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.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, and Di Cai. 2010. Sentiment Strength Detection in Short Informal Text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558.

Stefano Vegnaduzzo. 2004. Acquisition of Subjective Adjectives with Limited Resources. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, Stanford, CA, USA.

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.

Janyce M. Wiebe. 2000. Learning Subjective Adjectives from Corpora. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 735–740, Austin, TX, USA.

Michael Wiegand, Josef Ruppenhofer, and Dietrich Klakow. 2013. Predicative Adjectives: An Unsupervised Criterion to Extract Subjective Adjectives. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 534–539, Atlanta, GA, USA.

Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. 2004. Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 761–767, San Jose, CA, USA.