Ordering adverbs by their scaling effect on adjective intensity

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

Abstract

In recent years, theoretical and computational linguistics has paid much attention to linguistic items that form scales. In NLP, much research has focused on ordering adjectives by intensity (*tiny* < *small*). Here, we address the task of automatically ordering English adverbs by their intensifying or diminishing effect on adjectives (e.g. `extremely` *small* < `very` *small*).

We experiment with 4 different methods: 1) using the association strength between adverbs and adjectives; 2) exploiting scalar patterns (such as *not only X but Y*); 3) using the metadata of product reviews; 4) clustering. The method that performs best is based on the use of metadata and ranks adverbs by their scaling factor relative to unmodified adjectives.

1 Introduction

Being able to recognize the intensity associated with scalar expressions is a basic capability needed for tackling any NLP task that can be reduced to textual entailment. For instance, as illustrated by de Marneffe et al. (2010), when interpreting dialogue (A: *Was it good?* B: *It was ok / great / excellent.*), a yes/no question involving a gradable predicate may require understanding the entailment relations between that predicate and another contained in the answer. Another application is within sentiment analysis, where assessing the strength of subjective expressions (e.g. *good* < *great* < *excellent*) is one of the central tasks besides subjectivity detection and polarity classification (Rill et al., 2012b; Sheinman et al., 2013; de Melo and Bansal, 2013; Ruppenhofer et al., 2014, *inter alia*). It is also well known that subjective adjectives are frequently modified by adverbs that increase (`very` *expensive*) or decrease (`fairly` *expensive*) their intensity. As Benamara et al. (2007) have shown, it is useful to take such adverbial intensification into account when predicting document-level sentiment scores. However, Benamara et al. (2007) used human-assigned scores to model adverbs' effect on adjectives.

As far as we know, there is no well-established automatic method that can determine for degree adverbs what their effect will be on the intensity of various adjectives. In this paper, we explore several methods on English data that might be used towards that purpose, evaluating them against a new gold standard data set that we collected. All new resources that were created in the context of our investigation will be made publicly available.

The remainder of this paper is structured as follows. We present our data in §2. We describe the construction of our gold standard in §3 and the methods we use in §4. This is followed by the presentation of our experiments and results in §5. We discuss related work in §6 and conclude in §7.

2 Data

For our experiments we use two large corpora (Table 1). The first is a large set of Amazon reviews, which consist of numerical star ratings and textual assessments. Since both express the writers' evaluation, they are strongly correlated. Accordingly, we project the numerical star ratings onto the adjectives and adverbs in the texts as intensity scores (cf. §4.2). Second, we also use the ukWaC webcorpus, which is even larger than the review corpus, as general language data on which we compute association measures (cf. §4.1) and which we mine for linguistic patterns (cf. §4.3, §4.4).

Corpora	Tokens	Reference
Amazon reviews	∼1.06 B	Jindal and Liu (2008)
ukWaC	∼2.25 B	Baroni et al. (2009)

Table 1: Corpora used

Proceedings of Recent Advances in Natural Language Processing, pages 545–554,
Hissar, Bulgaria, Sep 7–9 2015.

3 Construction of human gold standard

To be able to assess adverb rankings produced by automatic methods, we collected human ratings for adverb and adjective combinations through an online survey. All combinations were rated individually, in randomized order, under conditions intended to minimize the effects of bias, habituation, fatigue etc. on the results. 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 $+100$. To indicate the intended word sense of each item, the scale was labeled accordingly. For instance, we specified that *cool* should be interpreted in terms of Temperature (*cool day*) rather than Desirability (*cool app*).

Through Amazon Mechanical Turk (AMT), we recruited subjects with the following qualifications: US residency, a HIT-approval rate of at least 97%, and 500 prior completed HITs. We collected 20 ratings per item but had to exclude some participants' answers as unusable, which reduced our sample for some items.

3.1 Adjectives

The adjectives we used – shown in Table 2 – cover four semantic areas, two of them (more or less) objective, namely Duration and Temperature, and two of them subjective, namely Quality and Intelligence. They are a subset of those used by Ruppenhofer et al. (2014) for ordering adjectives by intensity (cf. §4.1). Following Paradis (1997; 2001), we classify adjectives into three types. **Scalar adjectives** are ones that combine with scalar degree adverbs (`fairly` *long*, `very` *good*, `terribly` *nasty*). The mode of oppositeness that characterizes scalar adjectives is antonymy (e.g. *good* - *bad*). **Extreme adjectives** combine with reinforcing totality adverbs (`absolutely` *terrible*, `totally` *brilliant*, `utterly` *disastrous*). Like scalar adjectives, these adjectives are also antonymic (*hot* - *cold*) and they are conceptualized according to a scale. However, extreme adjectives do not represent a range on a scale but an (end-)point on the scale. The third type, **limit adjectives**, also combines with totality adverbs (`completely` *dead*, `absolutely` *true*, `almost` *identical*). This type differs from the others in that it is not associated with a scale but conceptualized in terms of either-or. It is not represented in our data elicitation but it is used by one of the automatic ranking methods (cf. §4.1, §5.1.)

Adjective	Scale	Polarity	Type
dumb	Intelligence	neg	scalar
smart	Intelligence	pos	scalar
brainless	Intelligence	neg	extreme
brainy	Intelligence	pos	extreme
bad	Quality	neg	scalar
good	Quality	pos	scalar
mediocre	Quality	neg	scalar
super	Quality	pos	extreme
cool	Temperature	neg	scalar
warm	Temperature	pos	scalar
frigid	Temperature	neg	extreme
hot	Temperature	pos	extreme
short	Duration	neg	scalar
long	Duration	pos	scalar
brief	Duration	neg	scalar
lengthy	Duration	pos	scalar

Table 2: Adjectives used and their classification

Maximizer	Booster
absolutely	awfully
completely	extremely
perfectly	very
quite	highly
Moderator	**Diminisher**
quite	slightly
fairly	a little
pretty	somewhat
Approximator	**Control**
almost	*none*

Table 3: Adverbs used and their classification

3.2 Adverbs

The adverbs in our surveys as well as their classification are inspired by Paradis (1997). The adverbs belong to five types plus a control condition as shown in Table 3. As Table 4 shows, maximizers and approximators are totality adverbs, they target adjectives that belong to the limit or extreme class. The other adverb classes are scalar adverbs that target scalar adjectives. In the control condition (*none*), subjects rate the unmodified adjective.

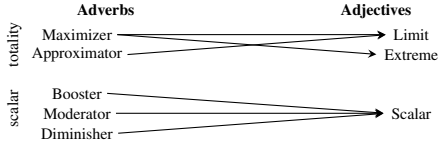

Table 4: Prototypical associations between adverb and adjective types according to Paradis (1997)

3.3 Design

We designed four parallel surveys, each eliciting data for degree modification of four adjectives, to be completed by non-overlapping sets of participants (enforced via AMT Worker IDs). In each

survey, participants first were asked for metadata such as age, residency, native language etc. Subsequently, pairs of main and distractor block followed until at the end feedback on difficult survey items was solicited. Each main block used one adjective, which participants first had to rate unmodified before giving ratings for seven combinations of the adjective with half the available adverbs.

Each main block was followed by a distractor block in which participants had to match verbs to related adjectives. As the combinations of an adjective with all adverbs were spread out over two main blocks, each survey had a total of 8 main blocks. The adverbs used with the first main block for an adjective were sampled randomly from our list, the remaining adverbs were put into the second main block featuring the adjective.

Note that we elicited data for all possible combinations of adjective and degree adverb. As shown by Desagulier (2014) and Erman (2014) for moderators and maximizers, respectively, some adverb-adjective combinations are highly entrenched, while others are likely to be rare or unfamiliar and thus possibly more difficult to rate.

3.4 Final ranking

Table 5 shows the ranking of adverb-adjective combinations, generalized over all 16 adjectives. The score per combination is the sum of all absolute scores for the adverb with any adjective across all participants, renormalized into the range [0,100]. Note that rank 8 is occupied by the cases where the relevant adjectives are not modified by any adverb. The results closely match expectations based on linguistic theory. We have booster and maximizer adverbs occupying ranks higher than the unmodified adjective, while we find moderators and diminishers occupying lower ranks. The ranking for the ambiguous *quite* seems to reflect its moderator use more than its maximizer use. The ordering among the moderators (*quite* > *pretty* > *fairly*) matches that reported as expert linguistic analysis by Paradis (1997, 148-155).

We next apply the method for building a gold standard described above to the combinations of all adverbs with each single adjective. The correlations between the 14 different resulting adverb rankings are high throughout with Spearman values >0.900. This argues that the ranking that we get when summing over all adjectives (cf. Table 5) also applies to the adjectives individually.

Finally, we constructed a relative ranking based on the number of raters for whom the combination of adverb A with a given adjective had a higher score than the combination involving adverb B. That method produces essentially the same result: the Spearman rank correlation with the absolute ranking in Table 5 is ρ=0.993. Due to space limitations, we only report results relative to the absolute gold standard in the remainder of the paper.

#	score	adverb	#	score	adverb
1	91.1	extremely †	9	59.9	quite †,○
2	89.2	absolutely ⋆	10	52.5	pretty ○
3	84.2	completely ⋆	11	42.1	fairly ○
4	79.3	highly †	12	35.9	somewhat ▷
5	78.6	very †	13	30.5	slightly ▷
6	75.2	awfully †	14	27.4	almost ♣
7	74.8	perfectly ⋆	15	26.7	a little ▷
8	62.7	*none*			

Table 5: Gold standard ranking of adverb-adjective intensity, based on absolute scores (†=maximizer, ⋆=booster, ○=moderator, ▷=diminisher, ♣=approximator)

In order to be able to experiment with more than the 14 prototypical and frequent adverbs that we could collect ratings for, we make use of the intensity ratings for 93 adverbs provided by Taboada et al.'s (2011) SoCaL resource. While various lexical resources provide polarity scores for nouns, verbs, and adjectives (Wilson et al., 2005; Thelwall et al., 2010; Taboada et al., 2011, *inter alia*), few resources cover and assign scores to degree adverbs. The adverb ranking obtained from the SoCaL resource for our 14 adverbs correlates strongly with our two gold standards, with coefficients of 0.969 against the absolute gold standard and 0.976 against the relative one. This gives us confidence that we can use the SoCaL ratings as an extended gold standard. Note that the set of 93 adverbs from SoCaL contains many adverbs that are less frequent and less grammaticized than the 14 adverbs from the smaller set.

4 Methods

Our methods to determine the intensifying effect of adverbs on adjectives are all corpus-based.

4.1 Collostructional analysis (Collex)

Our first method, **distinctive-collexeme analysis** (**Collex**) (Gries and Stefanowitsch, 2004) has previously been successfully applied to the intensity ordering of both subjective and objective adjectives (Ruppenhofer et al., 2014), with stable correlation results as evaluated against a human gold standard (Spearman's ρ of 0.732-0.837).

For the task of ordering adverbs according to their intensifying effect on an adjective, we assume that adverbs with different intensifying effects co-occur with different types of adjectives, as shown by Table 4 in §3.2. We identify two different constructions an adverb can occur in: modification of scalar adjectives such as *dumb* or modification of limit and extreme adjectives such as *brainless*. Booster, moderator, and diminisher adverbs co-occur with scalar adjectives (e.g. `very/rather` *dumb*), while limit and extreme adjectives are modified by maximizer and approximator adverbs (e.g. `absolutely/almost` *brainless*). Our hypothesis is that if adverb A has a higher preference for the limit and extreme adjective construction than adverb B, then A has a greater scaling effect than B. An adjective's preference for occurring in either construction is used to derive an ordering of the given adverbs by their effect on the intensity of adjectives. This preference is determined using the Fisher exact test (Fisher, 1922; Pedersen, 1996). It makes no distributional assumptions and does not require a minimum sample size. The direction in which observed frequencies differ from expected ones is taken to indicate the preference for one of the two constructions and is measured by the p-value.

We ran a distinctive-collexeme analysis for both the smaller and the larger set of degree adverbs on ukWaC with two different settings. First, we used the 16 adjectives from the survey differentiated into the two types *scalar* and *extreme* as presented in Table 2. We refer to the output as **Collex**$_{surveyAdj}$. Second, we used a larger set of 188 adjectives culled from the literature (Paradis, 1997; Erman, 2014; Desagulier, 2014). The adjectives are distributed across the three classes as follows: 26 extreme (**xtrm**), 123 limit (**lim**) and 39 **scalar**. We refer to the output as **Collex**$_{moreAdj}$.

4.2 Mean star ratings (MeanStar)

Another method we evaluate employs **Mean star ratings** (**MeanStar**) from product reviews as described by Rill et al. (2012b). Unlike Collex, this method uses no linguistic properties of words or phrases. Instead, it derives intensity values for words or phrases in review texts from the numeric star ratings that reviewers (manually) assign to products. The star ratings encode a polar score on the document level. Since the ratings are not binary but on a five-point scale, they can also be used as source for deriving intensity information. The basic idea is to count how many instances of a word or phrase occur in reviews with a given star rating (score) within a review corpus.

Following Rill et al. (2012b)'s model for simple adjectives, we generically define the intensity score for a word or phrase as the mean of the star ratings $SR_i = \frac{\sum_{j=1}^{n} S_j^i}{n}$, where i designates a distinct word or phrase, j is the j-th occurrence of the word or phrase, S_j^i is the star rating associated with i in j, and n is the number of observed instances of i. We experiment with three methods that are based on MeanStar. They differ a) in how the item *i* that is to be scored is defined (as a word or phrase) and b) in whether the resulting scores are used directly to generate a ranking or only after further processing.

Adverbs only In the simplest application of MeanStar, we calculate for each adverb the average star level of the reviews it occurs in, and then rank the adverbs by these scores.

Adjective-specific In a different mode of using the star-based scores, we do not build a general ordering of adverbs. Instead, we only order combinations of adverbs with specific adjectives. Accordingly, we perform a rank correlation of adverb-adjective combinations against the gold standard *per adjective* and report the average of the absolute Spearman rank correlation results.

Scaling factor The third method, **Scaling**, builds a global ranking of adverbs by comparing the MeanStar scores of adverb-adjective combinations to those of unmodified adjectives. The benefit of this is that we can make use of each adverb-adjective combination independently of any other and do not need to rely only on adjectives that are attested with all or many of the adverbs that we need to rank, which is rarely the case. The algorithm works as presented in Algorithm 1.

An important facet of the algorithm is the filtering in step 4. In order to get clearly polar cases, we retain only combinations with a score >=3.75 ('positive') or with a score <=2.5 ('negative'). It is known that the average review tends to have a slightly higher score than three. For that reason, the threshold for positive reviews is slightly more extreme than that for negative reviews. We discard combinations: 1) that are observed only once; 2) where the adjective contains characters other than letters or a hyphen; or 3) where the adjective never occurs unmodified in the corpus.

Algorithm 1 Rank by scaling factor (sf)

```
1: take a stratified random sample of n items from the set of adverbs
2: for each adverb adv in the sample do
3:     retrieve all combinations of adv with any adjective
4:     filter combinations
5:     sort combinations
6:     for combination in top k combinations do
7:         calculate scaling factor relative to unmodified adjective
8:         classify as intensifying or diminishing
9:     end for
10:    if length(intensifying_uses) > length(diminishing_uses) then
11:        if length(pos_intensifying_uses) / length(neg_intensifying_uses)
       > Threshold then
12:            average_sf=mean(pos_intensifying_uses)
13:        else:
14:            average_sf=mean(pos_intensifying_uses+neg_intensifying_uses)
15:        end if
16:    else if length(diminishing_uses) > length(intensifying_uses) then
17:        if length(neg_diminishing_uses) / length(pos_diminishing_uses)
       > Threshold then
18:            average_sf=mean(neg_diminishing_uses)
19:        else
20:            average_sf=mean(pos_diminishing_uses+neg_diminishing_uses)
21:        end if
22:    end if
23: end for
24: rank adverbs by their average scaling factor (average_sf)
```

In steps 7 and 8, we look at the *k* most frequent combinations per adverb. For each combination, we calculate a scaling factor in the interval [-1,+1] relative to the unmodified adjective. For intensifying adverbs we measure what fraction of the distance from the simple adjective to the highest score (5 for positive adjectives) or lowest score (1 for negative adjectives) the adjective has been 'pushed' by the adverb. For diminishing adverbs, we measure what fraction of the unmodified adjective's distance to the neutral score (3) the adjective has been 'pushed'. For each adverb, we keep track of the scaling factors for all *k* combinations. The classification as intensifying or diminishing is corpus-driven: an adverb in combination with a specific adjective is intensifying/diminishing, if the combination's value is more/less extreme than that of the unmodified adjective.

In lines 10-22, we perform two levels of checks before deciding how to assign the final scaling factor to the adverb. On the first level, we discard whichever type of uses is in the minority, intensifying or diminishing uses. On the second level, we identify whether the uses retained in the previous step have mostly been observed with positive adjectives or with negative ones. If the quotient exceeds a certain threshold, we again choose to ignore the evidence from the minority class. With both checks, the idea is to obtain a clearer signal of what the adverb's effect is.

Finally, we rank all adverbs by their aggregate scaling factor and perform a rank correlation test against a gold standard.

	Adjectives in X and Y	
Pattern	**Any**	**Identical**
X(,) and in fact Y	0	0
X(,) or even Y	15	3
X(,) if not Y	64	1
be X(,) but not Y	60	5
not only X(,) but Y	7	0
not X, let alone Y	0	0
not Y, not even X	0	0
$\sum$	146	9

Table 6: Phrasal patterns in the ukWaC

4.3 Horn patterns

Horn (1976) put forth a set of **pattern-based diagnostics** for acquiring information about the relative intensity of linguistic items that express different degrees of some shared property. The complete set is shown in the first column of Table 6. For all patterns, the item in the Y slot needs to be stronger than that in the X slot. The two slots can be filled by different types of expressions such as nouns, verbs, and adjectives. We are interested in the case, shown in sentences 1 and 2, where adverb-adjective combinations occupy both slots.

(1) This is [`very` *good*], if not [`extremely` *good*].
(2) It's not just [`mildly` *entertaining*] but [`very` *entertaining*].

As shown above, we can apply Horn patterns to our task by requiring X and Y to be adverb-adjective combinations where the adjective is identical and the adverbs are two distinct items from the 93 adverbs from SoCaL. Based on the frequencies with which different adverbs occur in the X and Y slots, we can induce a ranking of the adverbs. Table 6 shows the number of matches one gets when querying the ukWaC for instances of the 7 patterns with the above constraints. We get only 146 unique hits overall. Moreover, we get only 9 where the adjective in slot X is identical to the one in slot Y. The coverage problem we observe is familiar from earlier work on ordering adjectives, where it could be overcome through the use of web-scale n-grams and a sophisticated interpolation technique by de Melo and Bansal (2013). However, in the case of adverbs the problem is more severe. Furthermore, looking for the patterns in web-scale n-grams is not possible since the instances of these diagnostic patterns all exceed 5 tokens when X and Y are complex adjective phrases: at this time, no web-scale n-gram collection for n $>$ 5 is available.

4.4 Cluster analysis

Cluster analysis aims to group data objects into different groups based on object-specific features.

Gold	Configuration	Corr.
Ours	$\text{Collex}_{surveyAdj}$	0.055
	$\text{Collex}_{moreAdj-xtrm+scalar}$	-0.099
	$\text{Collex}_{moreAdj-lim+scalar}$	0.165
	$\text{Collex}_{moreAdj-xtrm+lim+scalar}$	0.191
SoCaL	$\text{Collex}_{surveyAdj}$	0.003
	$\text{Collex}_{moreAdj-xtrm+scalar}$	0.152
	$\text{Collex}_{moreAdj-lim+scalar}$	-0.188
	$\text{Collex}_{moreAdj-xtrm+lim+scalar}$	-0.154

Table 7: Spearman rank correlations for Collex

While it does not produce a ranking of adverbs according to their intensifying/diminishing effect, we can consider it a fallback method in case no robust ranking method can be found. The aim would be to obtain groups of adverbs that have a similar intensifying/diminishing effect on a modified adjective. Potentially, the clusters could subsequently be converted into a ranking (with tied ranks) by another method.

The features we use to cluster the adverbs are the co-occurrence frequencies with the top 35 adjectival collocates of each adverb, following Desagulier (2014). The adjectival collocates of each adverb are determined via Collexeme analysis (cf. Gries and Stefanowitsch, 2004). Furthermore, we use the *Canberra* distance measure (Lance and Williams, 1966) and *Ward.D* clustering algorithm (Ward, 1963), as this setting has produced clusters that are coherent with Paradis' (1997) classification of degree adverbs (Desagulier, 2014).

We performed hierarchical cluster analysis on both the 14 adverbs from our gold standard as well as on 93 single-term degree adverbs that are included in Taboada et al.'s (2011) SoCaL resource. We refer to the output as $\textbf{Cluster}_{surveyAdj}$ and $\textbf{Cluster}_{SoCaLAdj}$, respectively.

5 Experiments

For our evaluation, we compute the similarity between a gold standard ranking – either that based on our data elicitation (cf. Table 5), or that based on the degree adverbs in SoCaL (cf. §3.4) – and any other ranking that we are interested in, as Spearman's rank correlation coefficient (Spearman's ρ).

5.1 Collex

For the output of Collex, we constructed a ranking of the adverbs as follows: The adverb with the highest preference for extreme adjectives was placed at the top of the ranking. The remaining adverbs with preference for extreme adjectives were placed below that, ordered by descending p-values. Then, we continued with the adverb that had the lowest preference for scalar adjectives and added the remaining adverbs, placing the adverb with the highest preference for scalar adjectives at the bottom of the ranking. This approach of building a ranking has produced good results for the intensity ordering of adjectives (Ruppenhofer et al., 2014) and we adopt it with the idea of now exploiting the connection between adjectives and adverbs in the reverse direction.

The results of the pairwise Spearman rank correlations between the gold standard of either of the two adverb sets and the rankings derived from Collex are shown in Table 7. $\textbf{Collex}_{surveyAdj}$, the adverb ranking obtained from a distinctive-collexeme analysis performed on the 16 adjectives from our survey, produces no correlation with either gold standard. $\textbf{Collex}_{moreAdj}$, the adverb ranking derived from a distinctive-collexeme analysis ran on a larger set of adjectives, yields minimal positive and negative correlations against both gold standards. One way to interpret this result has to do with the associations between adjectives and adverbs as shown in Table (4). In the earlier work of Ruppenhofer et al. (2014) on ordering adjectives, maximizers and approximators were grouped as one pole of attraction for adjectives, and boosters, moderators, and diminishers as another. The gold ranking to be matched for adjectives has a relatively simple structure since extreme adjectives (e.g. *brilliant*) are simply more intensive than scalar adjectives (e.g. *smart*). When we go in the opposite direction, such a clear delimitation is not the case: as Table 5 shows, some boosters actually have a higher scaling effect than maximizers. Similarly, we have a problem in that approximators push intensity towards neutrality whereas maximizers push towards the extreme: Collex treats them as if they are pushing in the same direction. The structural properties of the adjective-adverb interaction may thus make Collex only suitable in one direction.

5.2 MeanStar

Table 8 shows the results for the three variants of MeanStar. Note that an asterisk in the last column marks experiments where results are averaged over 10 runs and each run is based on a stratified random sample of adverbs from SoCaL. The first four rows for **Adv** in Table 8 show the results for the adverb-only approach: while the cor-

Method	Configuration	Gold	Corr.	Adverbs
Adv	MeanStar$_{global-any}$	Ours	0.283	14
	MeanStar$_{global-title}$	Ours	0.446	14
	MeanStar$_{global-any}$	SoCaL	0.311	*30
	MeanStar$_{global-title}$	SoCaL	0.531	*30
Spec	MeanStar$_{specific-any}$	Ours	-0.091	14
	MeanStar$_{specific-title}$	Ours	0.203	14
Scaling	MeanStar$_{global-any}$	Ours	0.382	14
	MeanStar$_{global-title}$	Ours	0.787	14
	MeanStar$_{global-any}$	SoCaL	0.780	*30
	MeanStar$_{global-title}$	SoCaL	0.930	*30

Table 8: Spearman rank correlations for MeanStar
(* experiments involve adverbs randomly selected from SoCaL)

relation results are not very high, performance is better when using data from review titles alone (0.446 against our gold standard; 0.531 against SoCaL). This was to be expected since titles reflect the tenor of the star rating more directly than sentences in the body of a review.

The results for the adjective-specific variant of MeanStar are shown in the two rows marked **Spec**. We cannot evaluate against the larger set of adverbs in SoCaL because SoCaL contains no information on specific adverb-adjective combinations. For the results shown, we use only adverb-adjective combinations that occur at least twice. Regardless of whether we use only titles or full reviews, we face data sparsity problems as we do not see instances of all combinations between our adjectives and the adverbs. Coverage is better, if we use the reviews as a whole (11.5 vs. 4.4). By contrast, the correlation results, though low overall, are better if we use titles only (0.203 vs. -0.091). If we used the absolute values of the correlations, then the average correlation would be higher for full reviews (0.644 vs. 0.612).

As we can see, the Scaling method performs very well, even without having been optimized. For instance, the 2:1 margin for the second-level check is not based on any work with a development set but simply a rough guess. Omitting the second-level checks on steps 11 and 17 of the algorithm drops the score for **MeanStar**$_{global-title}$ with 30 random adverbs from 0.930 to 0.880 and for **MeanStar**$_{global-any}$ from 0.780 to 0.720, which are still good levels of performance.

5.3 Cluster analysis

To assess the quality of the clustering, we report on an external cluster validation performed against an expert classification of the adverbs. For the 14 gold standard adverbs we use the classification by Paradis (1997), while for the 93 adverbs from SoCaL (Taboada et al., 2011), we use a grouping of these adverbs into Paradis' (1997) five adverb classes that two of the authors worked out collaboratively. Results are shown in Table 9.

Degree adverbs	N Adverbs	N Clusters	ARI	Purity
Cluster$_{surveyAdj}$	14	5	0.572	0.857
Cluster$_{SoCaLAdj}$	93	5	-0.066	0.623

Table 9: External cluster evaluation for a cluster analysis based on Canberra distance measure and Ward.D clustering algorithm

The quality of the clustering results is measured by the *adjusted Rand index* (ARI) and *Cluster purity* (Purity). ARI measures the accuracy of the clustering, that is the percentage of correctly clustered objects based on the given classes and corrects the basic Rand Index (RI) for chance (Hubert and Arabie, 1985). For Purity, in turn, we assign each cluster to the adverb class that is most frequent in the cluster. Then, the accuracy of this assignment, i.e. the percentage of the correctly assigned adverbs is measured (Manning et al., 2008, 356-360). Purity can take values between 0 and 1, where 0 represents a "bad clustering" and a value of 1 indicates a perfect fit with a given (manual) classification. For ARI, the interpretation of the [0,1] range is the same. However, ARI can sometimes produce negative values when the original RI is smaller than the expected index. These negative values also represent bad clusterings. It is easy for Purity to achieve a value of 1 - as is the case when each object has its own cluster (Manning et al., 2008, 357). We therefore report results for both evaluation metrics.

By using the top adjectival collocates of each adverb as clustering features, we get a good clustering for the 14 degree adverbs for which we elicited human ratings as compared to the classification of Paradis (1997). For the larger set of 93 adverbs from SoCaL, we obtain very poor results. Figure 1 illustrates the clustering result for the smaller set of adverbs, **Cluster**$_{surveyAdj}$.

5.4 Summary

We found the MeanStar method that computes a scaling factor to perform best. Unlike the adverb-only variant of MeanStar, it makes use of the fact that the score of an adverb-adjective combination also depends on the adjective. And unlike the adjective-specific version of MeanStar, it builds a global ranking and is able to combine evidence from adverb-adjective combinations independently of which other combinations have been

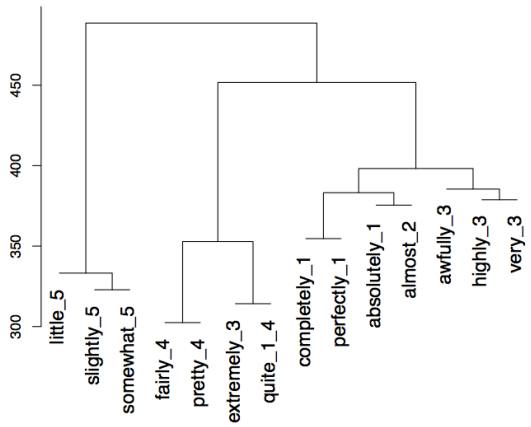

Figure 1: Dendrogram for the 14 adverbs from the survey. Indices show Paradis' (1997) classes.

observed. Somewhat surprisingly, the methods that are more directly grounded in linguistic theory performed worse (collostructional analysis) or proved unusable (Horn patterns). One possible reasons for the inferior Collex and clustering results may be that the relation between adverbs and adjectives is asymmetric to begin with, and easier to exploit in one direction than the other. Another is that the 5-way classification of adverbs and the assumptions about their common interaction with three types of adjectives cannot readily be extended beyond the set of well-known and highly grammaticized degree adverbs such as *very, quite, absolutely* to the much larger set of less grammaticized cases such as *mind-bogglingly* or *blisteringly*. The metadata approach, notably makes no assumptions about adverb or adjective classes.

6 Related work

Benamara et al. (2007) show the usefulness of taking adverbial intensification of adjectives into account when predicting document-level sentiment scores for news articles and blog posts. They divide adverbs into 5 classes based on the work of Quirk et al. (1985) and Bolinger (1972). The various scoring functions they explore for the adverb-adjective combinations are sensitive both to an adverb's class and to its score. The score of an adverb could lie between 0 and 1, with 0 meaning that the adverb has no impact on an adjective and 1 signifying that the adverb pushes the score of the combination to the minimum or maximum of the [-1,+1] scale. However, Benamara et al. (2007) lack an automatic way of scoring adverbs and rely on scores gathered from annotators.

Rill et al. (2012a) present a method for gathering opinion-bearing words and phrases, including adjective-phrases, from Amazon review data and assigning polarity scores on a continuous range between -1 and +1 to the entries based on the star ratings associated with the reviews. In subsequent work, Rill et al. (2012b) mention ways to infer the scores of unobserved adverb-adjective combinations based on observed combinations involving other, similar adjectives. However, the authors do not implement and evaluate these ideas.

Finally, a great deal of research on intensity has focused on acquiring prior polarity scores for individual words, and specifically adjectives. Various methods have been explored, including phrasal patterns (Sheinman et al., 2013; de Melo and Bansal, 2013); the use of star ratings (Rill et al., 2012b); extracting knowledge from lexical resources Gatti and Guerini (2012); and collostructional analysis (Ruppenhofer et al., 2014).

7 Conclusion

We examined various methods for ranking degree adverbs by their effect on the intensity of adjectives. We evaluated the methods against a new carefully-built gold standard that we collected experimentally as well as against a larger expert-constructed gold standard that we found to correlate well with ours for the overlapping members. While we found one method, Horn surface patterns, to currently not be workable at all due to the lack of suitable n-gram resources, we developed a MeanStar-based method that produces very good results using ratings metadata from product reviews to compute a scaling factor for adverb-adjective combinations relative to unmodified adjectives. Conspicuously, this scaling method makes no assumptions about any inherent properties of adverbs or adjectives, unlike the Collex and clustering approaches. In future work, we plan on looking more closely into the low results for the collostructional analysis approach, which had produced good results on the adjective ordering task, to ascertain if the asymmetries in the adverb-adjective associations (cf. §5.1) really are what prevents better results. Similarly, we plan on revisiting the typologies of adverbs and adjectives that we adopted from linguistic theory in order to see if they could be extended or revised in a way to give better clustering results.

Acknowledgments

The authors were partially supported by the German Research Foundation (DFG) under grants RU 1873/2-1 and WI 4204/2-1.

References

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.

Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato, and VS Subrahmanian. 2007. Sentiment Analysis: Adjectives and Adverbs are Better than Adjectives Alone. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

Dwight Bolinger. 1972. *Degree words*. Mouton, the Hague.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2010. "Was It Good? It Was Provocative." Learning the Meaning of Scalar Adjectives. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 167–176, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Guillaume Desagulier, 2014. *Corpus Methods for Semantics*, chapter Visualizing distances in a set of near-synonyms, pages 145–178. John Benjamins Publishing Company, Amsterdam, Philadelphia.

Britt Erman. 2014. There is no such thing as a free combination: a usage-based study of specific construals in adverb-adjective combinations. *English Language and Linguistics*, 18:109–132.

R. A. Fisher. 1922. On the Interpretation of 2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, 85(1):87–94, January.

Lorenzo Gatti and Marco Guerini. 2012. Assessing Sentiment Strength in Words Prior Polarities. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 361–370, Mumbai, India.

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.

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

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2(1):193–218.

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.

G. N. Lance and W. T. Williams. 1966. Computer programs for hierarchical polythetic classification ("similarity analyses"). *The Computer Journal*, 9(1):60–64.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*, volume 1. Cambridge University Press, Cambridge,UK.

Carita Paradis. 1997. *Degree modifiers of adjectives in spoken British English*, volume 92. Lund University Press.

Carita Paradis. 2001. Adjectives and boundedness. *Cognitive Linguistics*, (12):47–65.

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

Randolph Quirk, Sydney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman.

Sven Rill, Sven Adolph, Johannes Drescher, Dirk Reinel, Jörg Scheidt, Oliver Schütz, Florian Wogenstein, Roberto V. Zicari, and Nikolaos Korfiatis. 2012a. A phrase-based opinion list for the German language. In Jeremy Jancsary, editor, *Proceedings of KONVENS 2012*, pages 305–313. ÖGAI.

Sven Rill, Johannes Drescher, Dirk Reinel, Jörg Scheidt, Oliver Schütz, Florian Wogenstein, and Daniel Simon. 2012b. 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.

Josef Ruppenhofer, Michael Wiegand, and Jasper Brandes. 2014. Comparing methods for deriving intensity scores for adjectives. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 117–122, Gothenburg, Sweden, April. Association for Computational Linguistics.

Vera Sheinman, Christiane Fellbaum, Isaac Julien, Peter Schulam, and Takenobu Tokunaga. 2013. Large, huge or gigantic? Identifying and encoding intensity relations among adjectives in WordNet. *Language Resources and Evaluation*, 47(3):797–816.

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.

Joe H Ward. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 347–354, Vancouver, BC, Canada.