Good, Great, Excellent: Global Inference of Semantic Intensities

Gerard de Melo
ICSI, Berkeley
demelo@icsi.berkeley.edu

Mohit Bansal
CS Division, UC Berkeley
mbansal@cs.berkeley.edu

Abstract

Adjectives like *good*, *great*, and *excellent* are similar in meaning, but differ in intensity. Intensity order information is very useful for language learners as well as in several NLP tasks, but is missing in most lexical resources (dictionaries, WordNet, and thesauri). In this paper, we present a primarily unsupervised approach that uses semantics from Web-scale data (e.g., phrases like *good but not excellent*) to rank words by assigning them positions on a continuous scale. We rely on Mixed Integer Linear Programming to jointly determine the ranks, such that individual decisions benefit from global information. When ranking English adjectives, our global algorithm achieves substantial improvements over previous work on both pairwise and rank correlation metrics (specifically, 70% pairwise accuracy as compared to only 56% by previous work). Moreover, our approach can incorporate external synonymy information (increasing its pairwise accuracy to 78%) and extends easily to new languages. We also make our code and data freely available.[1]

1 Introduction

Current lexical resources such as dictionaries and thesauri do not provide information about the intensity order of words. For example, both WordNet (Miller, 1995) and Roget's 21st Century Thesaurus (*thesaurus.com*) present *acceptable*, *great*, and *superb* as synonyms of the adjective *good*. However, a native speaker knows that these words represent varying *intensity* and can in fact generally be ranked by intensity as *acceptable* $<$ *good* $<$ *great* $<$ *superb*. Similarly, *warm* $<$ *hot* $<$ *scorching* are identified as synonyms in these resources. Ranking information, however, is crucial because it allows us to differentiate e.g. between various intensities of an emotion, and is hence very useful for humans when learning a language or judging product reviews, as well as for automatic text understanding and generation tasks such as sentiment and subjectivity analysis, recognizing textual entailment, question answering, summarization, and coreference and discourse analysis.

In this work, we attempt to automatically rank sets of related words by intensity, focusing in particular on adjectives. This is made possible by the vast amounts of world knowledge that are now available. We use lexico-semantic information extracted from a Web-scale corpus in conjunction with an algorithm based on a Mixed Integer Linear Program (MILP). Linguistic analyses have identified phrases such as *good but not great* or *hot and almost scorching* in a text corpus as sources of evidence about the relative intensities of words. However, pure information extraction approaches often fail to provide enough coverage for real-world downstream applications (Tandon and de Melo, 2010), unless some form of advanced inference is used (Snow et al., 2006; Suchanek et al., 2009).

In our work, we address this sparsity problem by relying on Web-scale data and using an MILP model that extends the pairwise scores to a more complete joint ranking of words on a continuous scale, while maintaining global constraints such as transitivity and giving more weight to the order of word pairs with higher corpus evidence scores. Instead of considering intensity ranking as a pairwise decision process, we thus exploit the fact that *individual decisions may benefit from global information*, e.g. about how two words relate to some third word.

Previous work (Sheinman and Tokunaga, 2009; Schulam and Fellbaum, 2010; Sheinman et al., 2012) has also used lexico-semantic patterns to or-

[1] http://demelo.org/gdm/intensity/

Transactions of the Association for Computational Linguistics, 1 (2013) 279–290. Action Editor: Lillian Lee.
Submitted 11/2012; Revised 1/2013; Published 7/2013.

der adjectives. They mainly evaluate their algorithm on a set of pairwise decisions, but also present a partitioning approach that attempts to form scales by placing each adjective to the left or right of pivot words. Unfortunately, this approach often fails because many pairs lack order-based evidence even on the Web, as explained in more detail in Section 3.

In contrast, our MILP jointly uses information from all relevant word pairs and captures complex interactions and inferences to produce intensity scales. We can thus obtain an order between two adjectives even when there is no explicit evidence in the corpus (using evidence for related pairs and transitive inference). Our global MILP is flexible and can also incorporate additional synonymy information if available (which helps the MILP find an even better ranking solution). Our approach also extends easily to new languages. We describe two approaches for this multilingual extension: pattern projection and cross-lingual MILPs.

We evaluate our predicted intensity rankings using both pairwise classification accuracy and ranking correlation coefficients, achieving strong results, significantly better than the previous approach by Sheinman & Tokunaga (32% relative error reduction) and quite close to human-level performance.

2 Method

In this section, we describe each step of our approach to ordering adjectives on a single, relative scale. Our method can also be applied to other word classes and to languages other than English.

2.1 Web-based Scoring Model

2.1.1 Intensity Scales

Near-synonyms may differ in intensity, e.g. *joy* vs. *euphoria*, or *drizzle* vs. *rain*. This is particularly true of adjectives, which can represent different degrees of a given quality or attribute such as size or age. Many adjectives are gradable and thus allow for grading adverbial modifiers to express such intensity degrees, e.g., a house can be *very big* or *extremely big*. Often, however, completely different adjectives refer to varying degrees on the same scale, e.g., *huge*, *gigantic*, *gargantuan*. Even adjectives like *enormous* (or *superb*, *impossible*) that are considered non-gradable from a syntactic perspective can be placed on a such a scale.

Weak-Strong Patterns	Strong-Weak Patterns
⋆ (,) but not ⋆	not ⋆ (,) just ⋆
⋆ (,) if not ⋆	not ⋆ (,) but just ⋆
⋆ (,) although not ⋆	not ⋆ (,) still ⋆
⋆ (,) though not ⋆	not ⋆ (,) but still ⋆
⋆ (,) (and/or) even ⋆	not ⋆ (,) although still ⋆
⋆ (,) (and/or) almost ⋆	not ⋆ (,) though still ⋆
not only ⋆ but ⋆	⋆ (,) or very ⋆
not just ⋆ but ⋆	

Table 1: Ranking patterns used in this work. Among the patterns represented by the regular expressions above, we use only those that capture less than or equal to five words (to fit in the Google n-grams, see Section 2.1.2). Articles (*a, an, the*) are allowed to appear before the wildcards wherever possible.

2.1.2 Intensity Patterns

Linguistic studies have found lexical patterns like '⋆ but not ⋆' (e.g. *good but not great*) to reveal order information between a pair of adjectives (Sheinman and Tokunaga, 2009). We assume that we have two sets of lexical patterns that allow us to infer the most likely ordering between two words when encountered in a corpus. A first pattern set, P_{ws}, contains patterns that reflect a weak-strong order between a pair of word (the first word is weaker than the second), and a second pattern set, P_{sw}, captures the strong-weak order. See Table 1 for the adjective patterns that we used in this work (and see Section 4.1 for implementation details regarding our pattern collection). Many of these patterns also apply to other parts of speech (e.g. 'drizzle but not rain', 'running or even sprinting'), with significant discrimination on the Web in the right direction.

2.1.3 Pairwise Scores

Given an input set of words to be placed on a scale, we first collect evidence of their intensity order by using the above-mentioned intensity patterns and a large, Web-scale text corpus.

Previous work on information extraction from limited-sized raw text corpora revealed that coverage is often limited (Hearst, 1992; Hatzivassiloglou and McKeown, 1993). Some studies (Chklovski and Pantel, 2004; Sheinman and Tokunaga, 2009) used hit counts from an online search engine, but this is unstable and irreproducible (Kilgarriff, 2007). To avoid these issues, we use the largest available

(good, great)	(great, good)	(small, minute)
good , but not great → 24492.0	not great , just good → 248.0	small , almost minute → 97.0
good , if not great → 1912.0	great or very good → 89.0	small , even minute → 41.0
good , though not great → 504.0	not great but still good → 47.0	
good , or even great → 338.0		
not just good but great → 181.0		
good , almost great → 156.0		

Table 2: Some examples from the Web-scale corpus of useful intensity-based phrases on adjective pairs.

static corpus of counts, the Google n-grams corpus (Brants and Franz, 2006), which contains English n-grams ($n = 1$ to 5) and their observed frequency counts, generated from nearly 1 trillion word tokens and 95 billion sentences.

We consider each pair of words (a_1, a_2) in the input set in turn. For each pattern p in the two pattern sets (weak-strong P_{ws} and strong-weak P_{sw}), we insert the word pair into the pattern as $p(a_1, a_2)$ to get a phrasal *query* like "*big* but not *huge*". This is done by replacing the two wildcards in the pattern by the two words in order. Finally, we scan the Web n-grams corpus in a batch approach similar to Bansal and Klein (2011) and collect frequencies of all our phrase queries. Table 2 depicts some examples of useful intensity-based phrase queries and their frequencies in the Web-scale corpus. We also collect frequencies for the input word unigrams and the patterns for normalization purposes. Given a word pair (a_1, a_2) and a corpus count function cnt, we define

$$W_1 = \frac{1}{P_1} \sum_{p_1 \in P_{ws}} cnt(p_1(a_1, a_2))$$
$$S_1 = \frac{1}{P_2} \sum_{p_2 \in P_{sw}} cnt(p_2(a_1, a_2))$$
$$W_2 = \frac{1}{P_1} \sum_{p_1 \in P_{ws}} cnt(p_1(a_2, a_1))$$
$$S_2 = \frac{1}{P_2} \sum_{p_2 \in P_{sw}} cnt(p_2(a_2, a_1)) \quad (1)$$

with

$$P_1 = \sum_{p_1 \in P_{ws}} cnt(p_1)$$
$$P_2 = \sum_{p_2 \in P_{sw}} cnt(p_2), \quad (2)$$

such that the final overall *weak-strong score* is

$$score(a_1, a_2) = \frac{(W_1 - S_1) - (W_2 - S_2)}{cnt(a_1) \cdot cnt(a_2)}. \quad (3)$$

Here W_1 and S_1 represent Web evidence of a_1 and a_2 being in the weak-strong and strong-weak relation, respectively. W_2 and S_2 fit the reverse pair (a_2, a_1) in the patterns and hence represent the strong-weak and weak-strong relations, respectively, in the opposite direction. Hence, overall, $(W_1 - S_1) - (W_2 - S_2)$ represents the total weak-strong score of the pair (a_1, a_2), i.e. the score of a_1 being on the left of a_2 on a relative intensity scale, such that $score(a_1, a_2) = -score(a_2, a_1)$. The raw frequencies in the score are divided by counts of the patterns and by individual word unigram counts to obtain a pointwise mutual information (PMI) style normalization and hence avoid any bias in the score due to high-frequency patterns or word unigrams.[2]

2.2 Global Ordering with an MILP

2.2.1 Objective and Constraints

Given pairwise scores, we now aim at producing a global ranking of the input words that is much more informative than the original pairwise scores. Joint inference from multiple word pairs allows us to benefit from global information: Due to the sparsity of the pattern evidence, determining how two adjectives relate to each other can sometimes e.g. only be inferred by observing how each of them relate to some third adjective.

We assume that we are given N input words $A = a_1, \ldots, a_N$ that we wish to place on a linear scale, say $[0, 1]$. Thus each word a_i is to be assigned a position $x_i \in [0, 1]$ based on the pairwise weak-strong weights $score(a_i, a_j)$. A positive value for

[2] In preliminary experiments on a development set, we also evaluated other intuitive forms of normalization.

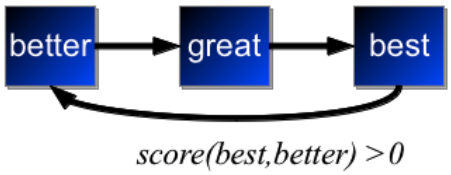

Figure 1: The input weak-strong data may contain one or more cycles, e.g. due to noisy patterns, so the final ranking will have to choose which input scores to honor and which to remove.

$score(a_i, a_j)$ means that a_i is supposedly weaker than a_j and hence we would like to obtain $x_i < x_j$. A negative value for $score(a_i, a_j)$ means that a_i is assumed to be stronger than a_j, so we would want to obtain $x_i > x_j$. Therefore, intuitively, our goal corresponds to maximizing the objective

$$\sum_{i,j} \text{sgn}(x_j - x_i) \cdot score(a_i, a_j) \qquad (4)$$

Note that it is important to use the signum function sgn() here, because we only care about the relative order of x_i and x_j. Maximizing $\sum_{ij}(x_j - x_i) \cdot score(a_i, a_j)$ would lead to all words being placed at the edges of the scale, because the highest scores would dominate over all other ones. We do include the score magnitudes in the objective, because they help resolve contradictions in the pairwise scores (e.g., see Figure 1). This is discussed in more detail in Section 2.2.2.

In order to maximize this non-differentiable objective, we use Mixed Integer Linear Programming (MILP), a variant of linear programming in which some but not all of the variables are constrained to be integers. Using an MILP formalization, we can find a globally optimal solution in the joint decision space, and unlike previous work, we jointly exploit global information rather than just individual local (pairwise) scores. To encode the objective in a MILP, we need to introduce additional variables d_{ij}, w_{ij}, s_{ij} to capture the effect of the signum function, as explained below.

We additionally also enable our MILP to make use of any external equivalence (synonymy) information $E \subseteq \{1, \ldots, N\} \times \{1, \ldots, N\}$ that may be available. In this context, two words are considered synonymous if they are close enough in meaning to be placed on (almost) the same position in the intensity scale. If $(i, j) \in E$, we can safely assume that a_i, a_j have near-equivalent intensity, so we should encourage x_i, x_j to remain close to each other. The MILP is defined as follows:

maximize

$$\sum_{(i,j) \notin E} (w_{ij} - s_{ij}) \cdot score(a_i, a_j) - \sum_{(i,j) \in E} (w_{ij} + s_{ij})\, C$$

subject to

$$\begin{array}{ll}
d_{ij} = x_j - x_i & \forall i, j \in \{1, \ldots, N\} \\
d_{ij} - w_{ij} C \leq 0 & \forall i, j \in \{1, \ldots, N\} \\
d_{ij} + (1 - w_{ij}) C > 0 & \forall i, j \in \{1, \ldots, N\} \\
d_{ij} + s_{ij} C \geq 0 & \forall i, j \in \{1, \ldots, N\} \\
d_{ij} - (1 - s_{ij}) C < 0 & \forall i, j \in \{1, \ldots, N\} \\
x_i \in [0, 1] & \forall i \in \{1, \ldots, N\} \\
w_{ij} \in \{0, 1\} & \forall i, j \in \{1, \ldots, N\} \\
s_{ij} \in \{0, 1\} & \forall i, j \in \{1, \ldots, N\}
\end{array}$$

The difference variables d_{ij} simply capture differences between x_i, x_j. C is any very large constant greater than $\sum_{i,j} |score(a_i, a_j)|$; the exact value is irrelevant. The indicator variables w_{ij} and s_{ij} are jointly used to determine the value of the signum function $\text{sgn}(d_{ij}) = \text{sgn}(x_j - x_i)$. Variables w_{ij} become 1 if and only if $d_{ij} > 0$ and hence serve as indicator variables for weak-strong relationships in the output. Variables s_{ij} become 1 if and only if $d_{ij} < 0$ and hence serve as indicator variables for a strong-weak relationship in the output. The objective encourages $w_{ij} = 1$ for $score(a_i, a_j) > 0$ and $s_{ij} = 1$ for $score(a_i, a_j) < 0$.[3] When equivalence (synonymy) information is available, then for $(i, j) \in E$ both $s_{ij} = 0$ and $w_{ij} = 0$ are encouraged.

2.2.2 Discussion

Our MILP uses intensity evidence of all input pairs together and assimilates all the scores via global transitivity constraints to determine the positions of the input words on a continuous real-valued scale. Hence, our approach addresses drawbacks

[3] In order to avoid numeric instability issues due to very small $score(a_i, a_j)$ values after frequency normalization, in practice we have found it necessary to rescale them by a factor of 1 over the smallest $|score(a_i, a_j)| > 0$.

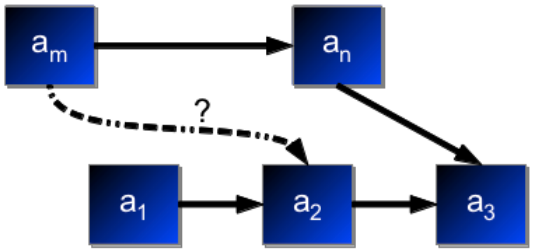

Figure 2: Equivalence Information: Knowing that a_m, a_2 are synonyms gives the MILP an indication of where to place a_n on the scale with respect to a_1, a_2, a_3

of local or divide-and-conquer approaches, where adjectives are scored with respect to selected pivot words, and hence many adjectives that lack pairwise evidence with the pivots are not properly classified, although they may have order evidence with some third adjective that could help establish the ranking. Optional synonymy information can further help, as shown in Figure 2.

Moreover, our MILP also gives higher weight to pairs with higher scores, which is useful when breaking global constraint cycles as in the simple example in Figure 1. If we need to break a constraint violating triangle or cycle, we would have to make arbitrary choices if we were ranking based on $\text{sgn}(score(a, b))$ alone. Instead, we can choose a better ranking based on the magnitude of the pairwise scores. A stronger score between an adjective pair doesn't necessarily mean that they should be further apart in the ranking. It means that these two words are attested together on the Web with respect to the intensity patterns more than with other candidate words. Therefore, we try to respect the order of such word pairs more in the final ranking when we are breaking constraint-violating cycles.

3 Related Work

Hatzivassiloglou and McKeown (1993) presented the first step towards automatic identification of adjective scales, thoroughly discussing the background of adjective semantics and a means of discovering clusters of adjectives that belong on the same scale, thus providing one way of creating the input for our ranking algorithm.

Inkpen and Hirst (2006) study near-synonyms and nuances of meaning differentiation (such as stylistic, attitudinal, etc.). They attempt to automatically acquire a knowledge base of near-synonym differences via an unsupervised decision-list algorithm. However, their method depends on a special dictionary of synonym differences to learn the extraction patterns, while we use only a raw Web-scale corpus.

Mohammad et al. (2013) proposed a method of identifying whether two adjectives are antonymous. This problem is related but distinct, because the degree of antonymy does not necessarily determine their position on an intensity scale. Antonyms (e.g., *little*, *big*) are not necessarily on the extreme ends of scales.

Sheinman and Tokunaga (2009) and Sheinman et al. (2012) present the most closely related previous work on adjective intensities. They collect lexico-semantic patterns via bootstrapping from seed adjective pairs to obtain pairwise intensities, albeit using search engine 'hits', which are unstable and problematic (Kilgarriff, 2007). While their approach is primarily evaluated in terms of a local pairwise classification task, they also suggest the possibility of ordering adjectives on a scale using a pivot-based partitioning approach. Although intuitive in theory, the extracted pairwise scores are frequently too sparse for this to work. Thus, many adjectives have no score with a particular headword. In our experiments, we reimplemented this approach and show that our MILP method improves over it by allowing individual pairwise decisions to benefit more from global information. Schulam and Fellbaum (2010) apply the approach of Sheinman and Tokunaga (2009) to German adjectives. Our method extends easily to various foreign languages as described in Section 5.

Another related task is the extraction of lexico-syntactic and lexico-semantic intensity-order patterns from large text corpora (Hearst, 1992; Chklovski and Pantel, 2004; Tandon and de Melo, 2010). Sheinman and Tokunaga (2009) follows Davidov and Rappoport (2008) to automatically bootstrap adjective scaling patterns using seed adjectives and Web hits. These methods thus can be used to provide the input patterns for our algorithm.

VerbOcean by Chklovski and Pantel (2004) extracts various fine-grained semantic relations (including the stronger-than relation) between pairs of verbs, using lexico-syntactic patterns over the Web.

Our approach of jointly ranking a set of words using pairwise evidence is also applicable to the VerbOcean pairs, and should help address similar sparsity issues of local pairwise decisions. Such scales will again be quite useful for language learners and language understanding tools.

de Marneffe et al. (2010) infer yes-or-no answers to questions with responses involving scalar adjectives in a dialogue corpus. They correlate adjectives with ratings in a movie review corpus to find that *good* appears in lower-rated reviews than *excellent*.

Finally, there has been a lot of work on measuring the general sentiment polarity of words (Hatzivassiloglou and McKeown, 1997; Hatzivassiloglou and Wiebe, 2000; Turney and Littman, 2003; Liu and Seneff, 2009; Taboada et al., 2011; Yessenalina and Cardie, 2011; Pang and Lee, 2008). Our work instead aims at producing a large, unrestricted number of individual intensity scales for different qualities and hence can help in fine-grained sentiment analysis with respect to very particular content aspects.

4 Experiments

4.1 Data

Input Clusters In order to obtain input clusters for evaluation, we started out with the satellite cluster or 'dumbbell' structure of adjectives in WordNet 3.0, which consists of two direct antonyms as the poles and a number of other satellite adjectives that are semantically similar to each of the poles (Gross and Miller, 1990). For each antonymy pair, we determined an extended dumbbell set by looking up synonyms and words in related (satellite adjective and 'see-also') synonym sets. We cut such an extended dumbbell into two antonymous halves and treated each of these halves as a potential input adjective cluster.

Most of these WordNet clusters are noisy for the purpose of our task, i.e. they contain adjectives that appear unrelatable on a single scale due to polysemy and semantic drift, e.g. *violent* with respect to *supernatural* and *affected*. Motivated by Sheinman and Tokunaga (2009), we split such hard-to-relate adjectives into smaller scale-specific subgroups using the corpus evidence[4]. For this, we consider an undirected edge between each pair of adjectives that has a non-zero intensity score (based on the Web-scale scoring procedure described in Section 2.1.3). The resulting graph is then partitioned into connected components such that any adjectives in a subgraph are at least indirectly connected via some path and thus much more likely to belong to the same intensity scale. While this does break up partitions whenever there is no corpus evidence connecting them, ordering the adjectives within each such partition remains a challenging task. This is because the Web evidence will still not necessarily directly relate all adjectives (in a partition) to each other. Additionally, the Web evidence may still indicate the wrong direction. Figure 3 shows the size distribution of the resulting partitions.

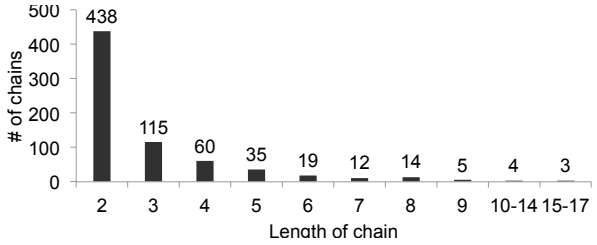

Figure 3: The histogram of cluster sizes after partitioning.

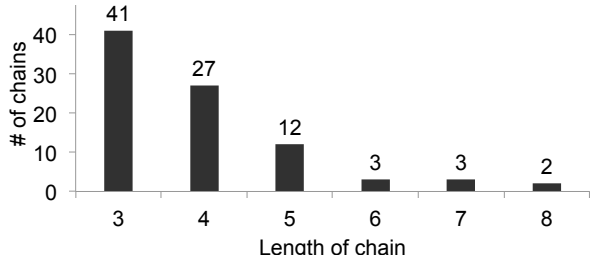

Figure 4: The histogram of cluster sizes in the test set.

Patterns To construct our intensity pattern set, we started with a couple of common rankable adjective seed pairs such as (*good, great*) and (*hot, boiling*) and used the Web-scale n-grams corpus (Brants and Franz, 2006) to collect the few most frequent patterns between and around these seed-pairs (in both directions). Among these, we manually chose a

[4]Note that we do not use the WordNet dataset of Sheinman and Tokunaga (2009) for evaluation, as it does not provide full scales. Instead, their annotators only made pairwise comparisons with select words, using a 5-way classification scheme (*neutral*, *mild*, *very mild*, *intense*, *very intense*).

small set of intuitive patterns that are linguistically useful for ordering adjectives, several of which had not been discovered in previous work. These are shown in Table 1. Note that we only collected patterns that were not ambiguous in the two orders, for example the pattern '$\star$ *, not* $\star$' is ambiguous because it can be used as both '*good, not great*' and '*great, not good*'. Alternatively, one can easily also use fully-automatic bootstrapping techniques based on seed word pairs (Hearst, 1992; Chklovski and Pantel, 2004; Yang and Su, 2007; Turney, 2008; Davidov and Rappoport, 2008). However, our semi-automatic approach is a simple and fast process that extracts a small set of high-quality and very general adjective-scaling patterns. This process can quickly be repeated from scratch in any other language. Moreover, as described in Section 5.1, the English patterns can also be projected automatically to patterns in other languages.

Development and Test Sets Section 2.1 describes the method for collecting the intensity scores for adjective pairs, using Web-scale n-grams (Brants and Franz, 2006). We relied on a small development set to test the MILP structure and the pairwise score setup. For this, we manually chose 5 representative adjective clusters from the full set of clusters.

The final test set, distinct from this development set, consists of 569 word pairs in 88 clusters, each annotated by two native speakers of English. Both the gold test data (and our code) are freely available.[5] To arrive at this data, we randomly drew 30 clusters each for cluster sizes 3, 4, and 5+ from the histogram of partitioned adjective clusters in Figure 3. While labeling a cluster, annotators could exclude words that they deemed unsuitable to fit on a single shared intensity scale with the rest of the cluster. Fortunately, the partitioning described earlier had already separated most such cases into distinct clusters. The annotators ordered the remaining words on a scale. Words that seemed indistinguishable in strength could share positions in their annotation.

As our goal is to compare scale formation algorithms, we did not include trivial clusters of size 2. On such trivial clusters, the Web evidence alone determines the output and hence all algorithms, including the baseline, obtain the same pairwise accuracy (defined below) of 93.3% on a separate set of 30 random clusters of size 2.

Figure 4 shows the distribution of cluster sizes in our main gold set. The inter-annotator agreement in terms of Cohen's κ (Cohen, 1960) on the pairwise classification task with 3 labels (weaker, stronger, or equal/unknown) was 0.64. In terms of pairwise accuracy, the agreement was 78.0%.

4.2 Metrics

In order to thoroughly evaluate the performance of our adjective ordering procedure, we rely on both pairwise and ranking-correlation evaluation metrics. Consider a set of input words $A = \{a_1, a_2, \ldots, a_n\}$ and two rankings for this set – a gold-standard ranking $r_G(A)$ and a predicted ranking $r_P(A)$.

4.2.1 Pairwise Accuracy

For a pair of words a_i, a_j, we may consider the classification task of choosing one of three labels ($<$, $>$, $=?$) for the case of a_i being weaker, stronger, and equal (or unknown) in intensity, respectively, compared to a_2:

$$L(a_1, a_2) = \begin{cases} < & \text{if } \; r(a_i) < r(a_j) \\ > & \text{if } \; r(a_i) > r(a_j) \\ =? & \text{if } \; r(a_i) = r(a_j) \end{cases}$$

For each pair (a_1, a_2), we compute gold-standard labels $L_G(a_1, a_2)$ and predicted labels $L_P(a_1, a_2)$ as above, and then the pairwise accuracy $PW(A)$ for a particular ordering on A is simply the fraction of pairs that are correctly classified, i.e. for which the predicted label is same as the gold-standard label:

$$PW(A) = \frac{\sum\limits_{i<j} \mathbf{1}\{L_G(a_i, a_j) = L_P(a_i, a_j)\}}{\sum\limits_{i<j} 1}$$

4.2.2 Ranking Correlation Coefficients

Our second type of evaluation assesses the rank correlation between two ranking permutations (gold-standard and predicted). Many studies use *Kendall's tau* (Kendall, 1938), which measures the total number of pairwise inversions, while others prefer *Spearman's rho* (Spearman, 1904), which measures the L1 distance between ranks.

[5] http://demelo.org/gdm/intensity/

Kendall's tau correlation coefficient We use the τ_b version of Kendall's correlation metric, as it incorporates a correction for ties (Kruskal, 1958; Dou et al., 2008):

$$\tau_b = \frac{P - Q}{\sqrt{(P + Q + X_0) \cdot (P + Q + Y_0)}}$$

where P is the number of concordant pairs, Q is the number of discordant pairs, X_0 is the number of pairs tied in the first ranking, Y_0 is the number of pairs tied in the second ranking. Given the two rankings of an adjective set A, the gold-standard ranking $r_G(A)$ and the predicted ranking $r_P(A)$, two words a_i, a_j are:

- *concordant* iff both rankings have the same strict order of the two elements, i.e., $r_G(a_i) > r_G(a_j)$ and $r_P(a_i) > r_P(a_j)$, or $r_G(a_i) < r_G(a_j)$ and $r_P(a_i) < r_P(a_j)$.
- *discordant* iff the two rankings have an inverted strict order of the two elements, i.e., $r_G(a_i) > r_G(a_j)$ and $r_P(a_i) < r_P(a_j)$, or $r_G(a_i) < r_G(a_j)$ and $r_P(a_i) > r_P(a_j)$.
- *tied* iff $r_G(a_i) = r_G(a_j)$ or $r_P(a_i) = r_P(a_j)$.

Spearman's rho correlation coefficient For two n-sized ranked lists $\{x_i\}$ and $\{y_i\}$, the Spearman correlation coefficient is defined as the Pearson correlation coefficient between the ranks of variables:

$$\rho = \frac{\sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \cdot \sum_i (y_i - \bar{y})^2}}$$

Here, $\bar{x}$ and $\bar{y}$ denote the means of the values in the respective lists. We use the standard procedure for handling ties correctly. Tied values are assigned the average of all ranks of items sharing the same value in the ranked list sorted in ascending order of the values.

Handling Inversions While annotating, we sometimes observed that the ordering itself was very clear but the annotators disagreed about which end of a particular scale was to count as the strong one, e.g. when transitioning from *soft* to *hard* or from *alpha* to *beta*. We thus also report average *absolute* values of both correlation coefficients, as these properly account for anticorrelations. Our test set only contains clusters of size 3 or larger, so there is no need to account for inversions in clusters of size 2.

4.3 Results

In Table 3, we use the evaluation metrics mentioned above to compare several different approaches.

Web Baseline The first baseline simply reflects the original pairwise Web-based intensity scores. We classify (with one of 3 labels) a given pair of adjectives using the Web-based intensity scores (as described in Section 2.1.3) as follows:

$$L_{baseline}(a_1, a_2) = \begin{cases} < & \text{if } score(a_i, a_j) > 0 \\ > & \text{if } score(a_i, a_j) < 0 \\ =? & \text{if } score(a_i, a_j) = 0 \end{cases}$$

Since $score(a_i, a_j)$ represents the weak-strong score of the two adjectives, a more positive value means a higher likelihood of a_i being weaker ($<$, on the left) in intensity than a_j.

In Table 3, we observe that the (micro-averaged) pairwise accuracy, as defined earlier, for the original Web baseline is 48.2%, while the ranking measures are undefined because the individual pairs do not lead to a coherent scale.

Divide-and-Conquer The divide-and-conquer baseline recursively splits a set of words into three subgroups, placed to the left (weaker), on the same position (no evidence), or to the right (stronger) of a given randomly chosen pivot word.

While this approach shows only a minor improvement in terms of the pairwise accuracy (50.6%), its main benefit is that one obtains well-defined intensity scales rather than just a collection of pairwise scores.

Sheinman and Tokunaga The approach by Sheinman and Tokunaga (2009) involves a similar divide-and-conquer based partitioning in the first phase, except that their method makes use of synonymy information from WordNet and uses all synonyms in WordNet's synset for the headword as neutral pivot elements (if the headword is not in WordNet, then the word with the maximal unigram frequency is chosen). In the second phase, their method performs pairwise comparisons within the more intense and less intense subgroups. We reimplement their approach here, using the Google N-Grams dataset instead of online Web search engine hits. We observe a small improvement over the Web baseline in terms of pairwise accuracy. Note that the

Method	Pairwise Accuracy	Avg. τ	Avg. $\|\tau\|$	Avg. ρ	Avg. $\|\rho\|$
Web Baseline	48.2%	N/A	N/A	N/A	N/A
Divide-and-Conquer	50.6%	0.45	0.53	0.52	0.62
Sheinman and Tokunaga (2009)	55.5%	N/A	N/A	N/A	N/A
MILP	69.6%	0.57	0.65	0.64	0.73
MILP with synonymy	78.2%	0.57	0.66	0.67	0.80
Inter-Annotator Agreement	78.0%	0.67	0.76	0.75	0.86

Table 3: Main test results

		Predicted Class		
		Weaker	Tie	Stronger
True Class	Weaker	117	127	15
	Tie	5	42	15
	Stronger	11	122	115

Table 4: Confusion matrix (Web baseline)

		Predicted Class		
		Weaker	Tie	Stronger
True Class	Weaker	177	29	53
	Tie	9	24	29
	Stronger	15	38	195

Table 5: Confusion matrix (MILP)

rank correlation measure scores are undefined for their approach. This is because in some cases their method placed all words on the same position in the scale, which these measures cannot handle even in their tie-corrected versions. Overall, the Sheinman and Tokunaga approach does not aggregate information sufficiently well at the global level and often fails to make use of transitive inference.

MILP Our MILP exploits the same pairwise scores to induce significantly more accurate pairwise labels with 69.6% accuracy, a 41% relative error reduction over the Web baseline, 38% over Divide-and-Conquer, and 32% over Sheinman and Tokunaga (2009). We further see that our MILP method is able to exploit external synonymy (equivalence) information (using synonyms marked by the annotators). The accuracy of the pairwise scores as well as the quality of the overall ranking increase even further to 78.2%, approaching the human inter-annotator agreement. In terms of average correlation coefficients, we observe similar improvement trends from the MILP, but of different magnitudes, because these averages give small clusters the same weight as larger ones.

4.4 Analysis

Confusion Matrices For a given approach, we can study the confusion matrix obtained by cross-tabulating the gold classification with the predicted classification of every unique pair of adjectives in the ground truth data. Table 4 shows the confusion matrix for the Web baseline. We observe that due to the sparsity of pairwise intensity order evidence, the baseline method predicts too many ties.

Table 5 provides the confusion matrix for the MILP (without external equivalence information) for comparison. Although the middle column still shows that the MILP predicts more ties than humans annotators, we find that a clear majority of all unique pairs are now correctly placed along the diagonal. This confirms that our MILP successfully infers new ordering decisions, although it uses the same input (corpus evidence) as the baseline. The remaining ties are mostly just the result of pairs for which there simply is no evidence at all in the input Web counts. Note that this problem could for instance be circumvented by relying on a crowdsourcing approach: A few dispersed tie-breakers are enough to allow our MILP to correct many other predictions.

Predicted Examples Finally, in Table 6, we provide a selection of real results obtained by our algorithm. For instance, it correctly inferred that *terrifying* is more intense than *creepy* or *scary*, although the Web pattern counts did not provide any explicit information about these words pairs. In some cases, however, the Web evidence did not suffice to draw the right conclusions, or it was misleading due to issues like polysemy (as for the word *funny*).

Accuracy	Prediction	Gold Standard
Good	hard < painful < hopeless	hard < painful < hopeless
	full < stuffed < (overflowing, overloaded)	full < stuffed < overflowing < overloaded
	unusual < uncommon < rare < exceptional < extraordinary	uncommon < unusual < rare < extraordinary < exceptional
Average	creepy < scary < sinister < frightening < terrifying	creepy < (scary, frightening) < terrifying < sinister
Bad	(awake, conscious) < alive < aware	alive < awake < (aware, conscious)
	strange < (unusual, weird) < (funny, eerie)	(strange, funny) < unusual < weird < eerie

Table 6: Some examples (of bad, average and good accuracy) of our MILP predictions (without synonymy information) and the corresponding gold-standard annotation.

While we show results on gold-standard chains here for evaluation purposes, in practice one can also recombine two $[0, 1]$ chains for a pair of antonymic clusters to form a single scale from $[-1, 1]$ that visualizes the full spectrum of available adjectives along a dimension, from *adjacent* all the way to *removed*, or from *black* to *glaring*.

5 Extension to Multilingual Ordering

Our method for globally ordering words on a scale can easily be applied to languages other than English. The entire process is language-independent as long as the required resources are available and a small number of patterns are chosen. For morphologically rich languages, the information extraction step of course may require additional morphological analysis tools for stemming and aggregating frequencies across different forms.

Alternatively, a cross-lingual projection approach is possible at multiple levels, utilizing information from the English data and ranking. As the first step, the set of words in the target language that we wish to rank can be projected from the English word set if necessary – e.g., as shown in de Melo and Weikum (2009). Next, we outline two projection methods for the ordering step. The first method is based on projection of the English intensity-ordering patterns to the new language, and then using the same MILP as described in Section 2.2. In the second method, we also change the MILP and add cross-lingual constraints to better inform the target language's adjective ranking. A detailed empirical evaluation of these approaches remains future work.

5.1 Cross-Lingual Pattern Projection

Instead of creating new patterns, in many cases we obtain quite adequate intensity patterns by using cross-lingual projection. We simply take several adjective pairs, instantiate the English patterns with them, and obtain new patterns using a machine translation system. Filling the wildcards in a pattern, say '$\star$ but not $\star$', with *good*/*excellent* results in 'good but not excellent'. This phrase is then translated into the target language using the translation system, say into German 'gut aber nicht ausgezeichnet'. Finally, put back the wildcards in the place of the translations of the adjective words, here *gut* and *ausgezeichnet*, to get the corresponding German pattern '$\star$ aber nicht $\star$'. Table 7 shows various German intensity patterns that we obtain by projecting from the English patterns as described. The process is repeated with multiple adjective pairs in case different variants are returned, e.g. due to morphology. Most of these translations deliver useful results.

Now that we have the target language adjectives and the ranking patterns, we can compute the pairwise intensity scores using large-scale data in that language. We can use the Google n-grams corpora for 10 European languages (Brants and Franz, 2009), and also for Chinese (LDC2010T02) and Japanese (LDC2009T08). For other languages, one can use available large raw-text corpora or Web crawling tools.

5.2 Crosslingual MILP

To improve the rankings for lesser-resourced languages, we can further use a joint MILP approach for the new language we want to transfer this process to. Additional constraints between the English

Weak-Strong Patterns		Strong-Weak Patterns	
English	German	English	German
⋆ but not ⋆	⋆ aber nicht ⋆	not ⋆ just ⋆	nicht ⋆ gerade ⋆
⋆ if not ⋆	⋆ wenn nicht ⋆	not ⋆ but just ⋆	nicht ⋆ aber nur ⋆
⋆ and almost ⋆	⋆ und fast ⋆	not ⋆ though still ⋆	nicht ⋆ aber immer noch ⋆
not just ⋆ but ⋆	nicht nur ⋆ sondern ⋆	⋆ or very ⋆	⋆ oder sehr ⋆

Table 7: Examples of German intensity patterns projected (translated) directly from the English patterns.

words and their corresponding target language translations, in combination with the English ranking information, allow the algorithm to obtain better rankings for the target words whenever the non-English target language corpus does not provide sufficient intensity order evidence.

In this case, the input set A contains words in multiple languages. The Web intensity scores $score(a_i, a_j)$ should be set to zero when comparing words across languages. We instead link them using a translation table $T \subseteq \{1, \ldots, N\} \times \{1, \ldots, N\}$ from a translation dictionary or phrase table. Here, $(i, j) \in T$ signifies that a_i is a translation of a_j. We do not require a bijective relationship between them (i.e., translations needn't be unique). The objective function is augmented by adding the new term

$$\sum_{(i,j)\in T} (w'_{ij} + s'_{ij}) C_T \tag{5}$$

for a constant $C_T > 0$ that determines how much weight we assign to translations as opposed to the corpus count scores. The MILP is extended by adding the following extra constraints.

$$\begin{array}{ll}
d_{ij} - w'_{ij} C_T < -d_{\max} & \forall i, j \in \{1, \ldots, N\} \\
d_{ij} + (1 - w'_{ij}) C_T \geq -d_{\max} & \forall i, j \in \{1, \ldots, N\} \\
d_{ij} + s'_{ij} C_T > d_{\max} & \forall i, j \in \{1, \ldots, N\} \\
d_{ij} - (1 - s'_{ij}) C_T \leq d_{\max} & \forall i, j \in \{1, \ldots, N\} \\
w'_{ij} \in \{0, 1\} & \forall i, j \in T \\
s'_{ij} \in \{0, 1\} & \forall i, j \in T
\end{array}$$

The variables $d_{i,j}$, as before, encode distances between positions of words on the scale, but now also include cross-lingual pairs of words in different languages. The new constraints encourage translational equivalents to remain close to each other, preferably within a desired (but not strictly enforced) maximum distance $d_{\max}$. The new variables w'_{ij}, s'_{ij} are similar to w_{ij}, s_{ij} in the standard MILP. However, the w'_{ij} become 1 if and only if $d_{ij} \geq -d_{\max}$ and the s'_{ij} become 1 if and only if $d_{ij} \leq d_{\max}$. If both w'_{ij} and s'_{ij} are 1, then the two words have a small distance $-d_{\max} \leq d_{ij} \leq d_{\max}$. The augmented objective function explicitly encourages this for translational equivalents. Overall, this approach thus allows evidence from a language with more Web evidence to improve the process of adjective ordering in lesser-resourced languages.

6 Conclusion

In this work, we have presented an approach to the challenging and little-studied task of ranking words in terms of their intensity on a continuous scale. We address the issue of sparsity of the intensity order evidence in two ways. First, pairwise intensity scores are computed using linguistically intuitive patterns in a very large, Web-scale corpus. Next, a Mixed Integer Linear Program (MILP) expands on this further by inferring new relative relationships. Instead of making ordering decisions about word pairs independently, our MILP considers the joint decision space and factors in e.g. how two adjectives relate to some third adjective, thus enforcing global constraints such as transitivity.

Our approach is general enough to allow additional evidence such as synonymy in the MILP, and can straightforwardly be applied to other word classes (such as verbs), and to other languages (monolingually as well as cross-lingually). The overall results across multiple metrics are substantially better than previous approaches, and fairly close to human agreement on this challenging task.

Acknowledgments

We would like to thank the editor and the anonymous reviewers for their helpful feedback.

References

Mohit Bansal and Dan Klein. 2011. Web-scale features for full-scale parsing. In *Proceedings of ACL 2011*.

Thorsten Brants and Alex Franz. 2006. The Google Web 1T 5-gram corpus version 1.1. *LDC2006T13*.

Thorsten Brants and Alex Franz. 2009. Web 1T 5-gram, 10 European languages, version 1. *LDC2009T25*.

Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the web for fine-grained semantic verb relations. In *Proceedings of EMNLP 2004*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Dmitry Davidov and Ari Rappoport. 2008. Unsupervised discovery of generic relationships using pattern clusters and its evaluation by automatically generated sat analogy questions. In *Proceedings of ACL 2008*.

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 ACL 2010*.

Gerard de Melo and Gerhard Weikum. 2009. Towards a universal wordnet by learning from combined evidence. In *Proceedings of CIKM 2009*.

Zhicheng Dou, Ruihua Song, Xiaojie Yuan, and Ji-Rong Wen. 2008. Are click-through data adequate for learning web search rankings? In *Proc. of CIKM 2008*.

Derek Gross and Katherine J. Miller. 1990. Adjectives in WordNet. *International Journal of Lexicography*, 3(4):265–277.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1993. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of ACL 1993*.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of ACL 1997*.

Vasileios Hatzivassiloglou and Janyce M. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of COLING 2000*.

Marti Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING 1992*.

Diana Inkpen and Graeme Hirst. 2006. Building and using a lexical knowledge base of near-synonym differences. *Computational Linguistics*, 32(2):223–262.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Adam Kilgarriff. 2007. Googleology is bad science. *Computational Linguistics*, 33(1).

William H. Kruskal. 1958. Ordinal measures of association. *Journal of the American Statistical Association*, 53(284):814–861.

Jingjing Liu and Stephanie Seneff. 2009. Review sentiment scoring via a parse-and-paraphrase paradigm. In *Proceedings of EMNLP 2009*.

George A. Miller. 1995. WordNet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Said M. Mohammad, Bonnie J. Dorr, Graeme Hirst, and Peter D. Turney. 2013. Computing lexical contrast. *Computational Linguistics*.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, January.

Peter F. Schulam and Christiane Fellbaum. 2010. Automatically determining the semantic gradation of german adjectives. In *Proceedings of KONVENS 2010*.

Vera Sheinman and Takenobu Tokunaga. 2009. AdjScales: Visualizing differences between adjectives for language learners. *IEICE Transactions on Information and Systems*, 92(8):1542–1550.

Vera Sheinman, Takenobu Tokunaga, I. Julien, P. Schulam, and C. Fellbaum. 2012. Refining WordNet adjective dumbbells using intensity relations. In *Proceedings of Global WordNet Conference 2012*.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of COLING/ACL 2006*.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101.

Fabian M. Suchanek, Mauro Sozio, and Gerhard Weikum. 2009. SOFIE: a self-organizing framework for information extraction. In *Proceedings of WWW 2009*.

Maite Taboada, Julian Brooke, Milan Tofiloskiy, and Kimberly Vollz. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*.

Niket Tandon and Gerard de Melo. 2010. Information extraction from web-scale n-gram data. In *Proceedings of the SIGIR 2010 Web N-gram Workshop*.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, 21(4):315–346, October.

Peter D. Turney. 2008. A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of COLING 2008*.

Xiaofeng Yang and Jian Su. 2007. Coreference resolution using semantic relatedness information from automatically discovered patterns. In *Proceedings of ACL 2007*.

Ainur Yessenalina and Claire Cardie. 2011. Compositional matrix-space models for sentiment analysis. In *Proceedings of EMNLP 2011*.