Learning Scalar Adjective Intensity from Paraphrases

Anne Cocos[⋆]
acocos@seas.upenn.edu

Skyler Wharton[⋆]
skyler@skylerwharton.com

Ellie Pavlick[†]
epavlick@brown.edu

Marianna Apidianaki[⋆◇]
marapi@seas.upenn.edu

Chris Callison-Burch[⋆]
ccb@upenn.edu

[⋆] Department of Computer and Information Science, University of Pennsylvania
[†] Department of Computer Science, Brown University
[◇] LIMSI, CNRS, Université Paris-Saclay, 91403 Orsay

Abstract

Adjectives like *warm*, *hot*, and *scalding* all describe temperature but differ in intensity. Understanding these differences between adjectives is a necessary part of reasoning about natural language. We propose a new paraphrase-based method to automatically learn the relative intensity relation that holds between a pair of scalar adjectives. Our approach analyzes over 36k adjectival pairs from the Paraphrase Database under the assumption that, for example, paraphrase pair *really hot* ↔ *scalding* suggests that *hot* < *scalding*. We show that combining this paraphrase evidence with existing, complementary pattern- and lexicon-based approaches improves the quality of systems for automatically ordering sets of scalar adjectives and inferring the polarity of indirect answers to *yes/no* questions.

1 Introduction

Semantically similar adjectives are not fully interchangeable in context. Although *hot* and *scalding* are related, the statement *"the coffee was hot"* does not imply the coffee was *scalding*. *Hot* and *scalding* are scalar adjectives that describe *temperature*, but they are not interchangeable because they vary in intensity. A native English speaker knows that their relative intensities are given by the ranking *hot* < *scalding*. Understanding this distinction is important for language understanding tasks such as sentiment analysis (Pang et al., 2008), question answering (de Marneffe et al., 2010), and textual inference (Dagan et al., 2006).

Existing lexical resources such as WordNet (Miller, 1995; Fellbaum, 1998) do not include the relative intensities of adjectives. As a result, there have been efforts to automate the process of learning intensity relations (e.g. Sheinman and Tokunaga (2009), de Melo and Bansal (2013), Wilkinson (2017), etc.). Many existing approaches rely on *pattern-based* or *lexicon-based* methods to predict the intensity ranking of adjectives. Pattern-based approaches search large corpora for lexical patterns that indicate an intensity relationship – for example, *"not just X, but Y"* implies $X < Y$. As with pattern-based approaches for other tasks (such as hypernym discovery (Hearst, 1992)), they are precise but have relatively sparse coverage of comparable adjectives, even when using web-scale corpora (de Melo and Bansal, 2013; Ruppenhofer et al., 2014). Lexicon-based approaches employ resources that map an adjective to a real-valued number that encodes both intensity and polarity (e.g. *good* might map to 1 and *phenomenal* to 5, while *bad* maps to -1 and *awful* to -3). They can also be precise, but may not cover all adjectives of interest.

particularly pleased	↔	*ecstatic*
quite limited	↔	*restricted*
rather odd	↔	*crazy*
so silly	↔	*dumb*
completely mad	↔	*crazy*

Figure 1: Examples of paraphrases from PPDB of the form $RB\ JJ_u \leftrightarrow JJ_v$ which can be used to infer pairwise intensity relationships ($JJ_u < JJ_v$).

We propose paraphrases as a new source of evidence for the relative intensity of scalar adjectives. A paraphrase is a pair of words or phrases with approximately similar meaning, such as *really great* ↔ *phenomenal*. Adjectival paraphrases can be exploited to uncover intensity relationships. A paraphrase pair of the above form, where one phrase is composed of an intensifying adverb and an adjective (*really great*) and the other is a single-word adjective (*phenomenal*), provides evidence that *great* < *phenomenal*. By drawing this evidence from large, automatically-generated

paraphrase resources like the Paraphrase Database (PPDB) [1] (Ganitkevitch et al., 2013; Pavlick et al., 2015), it is possible to obtain high-coverage pairwise adjective intensity predictions at reasonably high accuracy.

We demonstrate the usefulness of paraphrase evidence for inferring relative adjective intensity in two tasks: ordering sets of adjectives along an intensity scale, and inferring the polarity of indirect answers to *yes/no* questions. In both cases, we find that combining the relatively noisy, but high-coverage, paraphrase evidence with more precise but low-coverage pattern- or lexicon-based evidence improves overall quality.

2 Related Work

Noting that adding adjective intensity relations to WordNet (Miller, 1995; Fellbaum, 1998) would be useful, Sheinman et al. (2013) propose a system for automatically extracting sets of same-attribute adjectives from WordNet 'dumbbells' – consisting of two direct antonyms at the poles and satellites of synonymous/related adjectives incident to each antonym (Gross and Miller, 1990) – and ordering them by intensity. The annotations, however, are not in WordNet as of its latest version (3.1).

Work on adjective intensity generally focuses on two related tasks: clustering adjectives based on the attributes they modify, and ranking same-attribute adjectives by intensity. With respect to the former, common approaches involve clustering adjectives by their contexts (Hatzivassiloglou and McKeown, 1993; Shivade et al., 2015). We do not focus on the clustering task in this paper, but concentrate on the ranking task.

Approaches to the task of ranking scalar adjectives by their intensity mostly fall under the paradigms of *pattern-based* or *lexicon-based* approaches. *Pattern-based* approaches work by extracting lexical (Sheinman and Tokunaga, 2009; de Melo and Bansal, 2013; Sheinman et al., 2013) or syntactic (Shivade et al., 2015) patterns indicative of an intensity relationship from large corpora. For example, the patterns "X, but not Y" and "not just X but Y" provide evidence that X is an adjective less intense than Y.

Lexicon-based approaches are derived from the premise that adjectives can provide information about the sentiment of a text (Hatzivassiloglou and McKeown, 1993). These methods draw upon a lexicon that maps adjectives to real-valued scores encoding both sentiment polarity and intensity. The lexicon might be compiled automatically – for example, from analyzing adjectives' appearance in star-valued product or movie reviews (de Marneffe et al., 2010; Rill et al., 2012; Sharma et al., 2015; Ruppenhofer et al., 2014) – or manually. In our experiments we utilize the manually-compiled SO-CAL lexicon (Taboada et al., 2011).

Our paraphrase-based approach to inferring relative adjective intensity is based on paraphrases that combine adjectives with adverbial modifiers. A tangentially related approach is Collex (Ruppenhofer et al., 2014), which is motivated by the intuition that adjectives with extreme intensities are modified by different adverbs from adjectives with more moderate intensities: extreme adverbs like *absolutely* are more likely to modify extreme adjectives like *brilliant* than are moderate adverbs like *very*. Unlike Collex, which requires predetermined sets of 'end-of-scale' and 'normal' adverbial modifiers, our approach learns the identity and relative importance of intensifying adverbs.

Relative intensity is just one of several dimensions of gradable adjective semantics. In addition to intensity scales, a comprehensive model of scalar adjective semantics might also incorporate notions of intensity range (Morzycki, 2015), adjective class (Kamp and Partee, 1995), and scale membership according to meaning (Hatzivassiloglou and McKeown, 1993). In this paper we take the position that relative intensity is worth studying on its own because it is an important component of adjective semantics, usable directly for some NLP tasks such as sentiment analysis (Pang et al., 2008), and as part of a more comprehensive model for other tasks like question answering (de Marneffe et al., 2010).

3 Paraphrase-based Intensity Evidence

Adjectival paraphrases provide evidence about the relative intensity of adjectives. A paraphrase of the form $RB\ JJ_u \leftrightarrow JJ_v$ – where one phrase is comprised of an adjective modified by an intensifying adverb ($RB\ JJ_u$), and the other is a single-word adjective (JJ_v) – is evidence that the first adjective is less intense than the second ($JJ_u < JJ_v$). We propose a new method for encoding this evidence and using it to make pairwise adjective intensity predictions. First, a graph (JJGRAPH) is formed to represent over 36k adjectival paraphrases hav-

[1] www.paraphrase.org

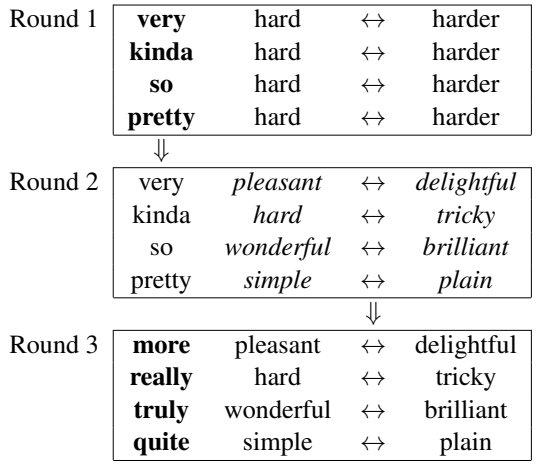

Figure 2: Bootstrapping process for identifying intensifying adverbs. The adverbs found in Rounds 1 and 3 are used to build intensifying edges in JJGRAPH.

ing the specified form. Next, data in the graph are used to make pairwise adjective intensity predictions.

3.1 Identifying Intensifying Adverbs

In JJGRAPH, nodes are adjectives, and each directed edge ($JJ_u \xrightarrow[RB]{} JJ_v$) corresponds to an adjectival paraphrase of the form *RB* $JJ_u \leftrightarrow JJ_v$ – for example, *very tall* ↔ *large* – where one 'phrase' (JJ_v) is an adjective and the other (RB JJ_u) is an adjectival phrase containing an adverb and adjective (see Figure 1 for examples).

Adverbs in PPDB can be intensifying or de-intensifying. An *intensifying* adverb (e.g. *very*, *totally*) strengthens the adjectives it modifies. In contrast, a *de-intensifying* adverb (e.g. *slightly*, *somewhat*) weakens the adjectives it modifies. Since edges in JJGRAPH ideally point in the direction of increasing intensity, the first step in the process of creating JJGRAPH is to identify a set of adverbs that are likely intensifiers to be included as edges.

For this purpose, we generate a set R of likely intensifying adverbs within PPDB using a bootstrapping approach (Figure 2). The process starts with a small seed set of adjective pairs having a known intensity relationship. The seeds are pairs (j_u, j_v) from PPDB-XXL[2] such that j_u is a base-form adjective (e.g. *hard*), and j_v is its **comparative** or **superlative** form (e.g. *harder* or *hardest*). Using the seeds, we identify intensifying adverbs by finding adjectival paraphrases in PPDB of the form ($r_i j_u \leftrightarrow j_v$); because $j_u < j_v$, adverb r_i is inferred to be intensifying (Round 1). All such r_i are added to initial adverb set R_1. The process continues by extracting paraphrases ($r_i j_{u'} \leftrightarrow j_{v'}$) with $r_i \in R_1$, indicating additional adjective pairs ($j_{u'}, j_{v'}$) with intensity direction inferred by r_i (Round 2). Finally, the adjective pairs extracted in this second iteration are used to identify additional intensifying adverbs R_3, which are added to the final set $R = R_1 \cup R_3$ (Round 3).

In all, this process generates a set of 610 adverbs. Examination of the set shows that the process does capture many intensifying adverbs like *very* and *abundantly*, and excludes many de-intensifying adverbs appearing in PPDB like *far less* and *not as*. However, due to the noise inherent in PPDB itself and in the bootstrapping process, there are also a few de-intensifying adverbs included in R (e.g. *hardly*, *kind of*) as well as adverbs that are neither intensifying nor de-intensifying (e.g. *ecologically*). It will be important to take this noise into consideration when using JJGRAPH to make pairwise intensity predictions.

3.2 Building JJGRAPH

JJGRAPH is built by extracting all 36,756 adjectival paraphrases in PPDB of the specified form *RB* $JJ_u \leftrightarrow JJ_v$, where the adverb belongs to R. The resulting graph has 3,704 unique adjective nodes. JJGRAPH is a multigraph, as there are frequently multiple intensifying relationships between pairs of adjectives. For example, the paraphrases *pretty hard* ↔ *tricky* and *really hard* ↔ *tricky* are both present in PPDB. There can also be contradictory or cyclic edges in JJGRAPH, as in the example depicted in the JJGRAPH subgraph in Figure 3, where the adverb *really* connects *tasty* to *lovely* and vice versa. Self-edges are allowed (e.g. *really hard* ↔ *hard*).

3.3 Pairwise Intensity Prediction

Examining the directed adverb edges between two adjectives j_u and j_v in JJGRAPH provides evidence about the relative intensity relationship between them. However, it has just been noted that JJGRAPH is noisy, containing both contradictory/cyclic edges and adverbs that are not uniformly intensifying. Rather than try to eliminate cycles, or manually annotate each adverb with a weight corresponding to its intensity and polarity

[2]PPDB comes in six increasingly large sizes from S to XXXL; larger collections have wider coverage but lower precision. Our work uses XXL.

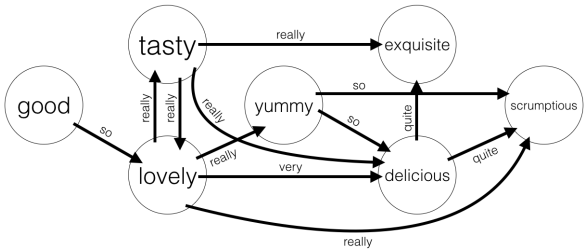

Figure 3: A subgraph of JJGRAPH, depicting its directed graph structure.

(Ruppenhofer et al., 2015; Taboada et al., 2011), we aim to learn these weights automatically in the process of predicting pairwise intensity.

Given adjective pair (j_u, j_v), we build a classifier that outputs a score from 0 to 1 indicating the predicted likelihood that $j_u < j_v$. Its binary features correspond to adverb edges from j_u to j_v and from j_v to j_u in JJGRAPH. The feature space includes only adverbs from R that appear at least 10 times in JJGRAPH, resulting in features for $m = 259$ unique adverbs in each direction (i.e. from j_u to j_v and vice versa) for $2m = 518$ binary features total. Note that while all adverb features correspond to predicted intensifiers from R, there are some features that are actually de-intensifying due to the noise inherent in the bootstrapping process (Section 3.1).

We train the classifier on all 36.7k edges in JJGRAPH, based on a simplifying assumption that all adverbs in R are indeed intensifiers. For each adjective pair (j_u, j_v) with one or more direct edges from j_u to j_v, a positive training instance for pair (j_u, j_v) and a negative training instance for pair (j_v, j_u) are added to the training set. A logistic regression classifier is trained on the data, using elastic net regularization and 10-fold cross validation to tune parameters.

The model parameters output by the training process are in a feature weights vector $w \in \mathbb{R}^{2m}$ (with no bias term) which can be used to generate a paraphrase-based score for each adjective pair:

$$score_{\text{pp}}(j_u, j_v) = \frac{1}{1 + \exp^{-wx_{uv}}} - 0.5 \quad (1)$$

where x_{uv} is the binary feature vector for adjective pair (j_u, j_v). The decision boundary 0.5 is subtracted from the sigmoid activation function so that pairs predicted to have the directed relation $j_u < j_v$ will have a positive score, and those predicted to have the opposite directional relation will have a negative score.

4 Other Intensity Evidence

Our experiments compare the proposed paraphrase approach with existing pattern- and lexicon-based approaches.

4.1 Pattern-based Evidence

We experiment with the pattern-based approach of de Melo and Bansal (2013). Given a pair of adjectives to be ranked by their intensity, de Melo and Bansal (2013) cull intensity patterns from Google *n*-Grams (Brants and Franz, 2009) as evidence of their intensity order. Specifically, they identify 8 types of *weak-strong* patterns (e.g. *"X, but not Y"*) and 7 types of *strong-weak* patterns (e.g. *"not X, but still Y"*) that are used as evidence about the directionality of the intensity relationship between adjectives. Given an adjective pair (j_u, j_v), an overall pattern-based *weak-strong* score is calculated:

$$score_{\text{pat}}(j_u, j_v) = \frac{(W_u - S_u) - (W_v - S_v)}{\text{count}(j_u) \cdot \text{count}(j_v)} \quad (2)$$

where W_u and S_u quantify the pattern evidence for the *weak-strong* and *strong-weak* intensity relations respectively for the pair (j_u, j_v), and W_v and S_v quantify the pattern evidence for the pair (j_v, j_u). W_u and S_u are calculated as:

$$W_u = \frac{1}{P_1} \sum_{p_1 \in P_{ws}} \text{count}(p_1(j_u, j_v))$$
$$S_u = \frac{1}{P_2} \sum_{p_2 \in P_{sw}} \text{count}(p_2(j_u, j_v)) \quad (3)$$

W_v and S_v are calculated similarly by swapping the positions of j_u and j_v. For example, given pair (*good*, *great*), W_u might incorporate evidence from patterns *"good, but not great"* and *"not only good but great"*, while S_v might incorporate evidence from the pattern *"not great, just good"*. P_{ws} denotes the set of *weak-strong* patterns, P_{sw} denotes the set of *strong-weak* patterns, and P_1 and P_2 give the total counts of all occurrences of any pattern in P_{ws} and P_{sw} respectively. The score is normalized by the frequencies of j_u and j_v in order to avoid bias due to high-frequency adjectives. As with the paraphrase-based scoring mechanism (Equation 1), scores output by this method can be positive or negative, with positive scores being indicative of a *weak-strong* relationship from j_u to j_v. Note that $score(j_u, j_v) = -score(j_v, j_u)$.

4.2 Lexicon-based Evidence

We use the manually-compiled SO-CAL[3] lexicon as our third, lexicon-based method for inferring intensity. The SO-CAL lexicon assigns an integer weight in the range $[-5, 5]$ to 2,826 adjectives. The sign of the weight encodes sentiment polarity (positive or negative), and the value encodes intensity (e.g. *atrocious*, with a weight of -5, is more intense than *unlikable*, with a weight of -3). SO-CAL is used to derive a pairwise intensity prediction for adjectives (j_u,j_v) as follows:

$$score_{\text{socal}}(j_u, j_v) = |L(j_v)| - |L(j_u)|, \quad \text{iff } \text{sign}(j_u) = \text{sign}(j_v) \tag{4}$$

where $L(j_v)$ gives the lexicon weight for j_v. Note that $score_{socal}$ is computed only for adjectives having the same polarity direction in the lexicon; otherwise the score is undefined. This is because adjectives belonging to different half scales, such as *freezing* and *steaming*, are frequently incomparable in terms of intensity (de Marneffe et al., 2010).

4.3 Combining Evidence

While the pattern-based and lexicon-based pairwise intensity scores are known to be precise but low-coverage (de Melo and Bansal, 2013; Ruppenhofer et al., 2015), we expect that the paraphrase-based score will produce higher coverage at lower accuracy. Thus we also experiment with scoring methods that combine two or three score types. When combining two metrics **x** and **y** to generate a score for a pair (j_u, j_v), we simply use the first metric **x** if it can be reliably calculated for the pair, and back off to metric **y** otherwise. More formally, the combined score for metrics **x** and **y** is given by:

$$score_{\text{x+y}}(j_u, j_v) = \alpha_{\text{x}} \cdot g_{\text{x}}(score_{\text{x}}(j_u, j_v)) + (1 - \alpha_{\text{x}}) \cdot g_{\text{y}}(score_{\text{y}}(j_u, j_v)) \tag{5}$$

where $\alpha_{\text{x}} \in \{0, 1\}$ is a binary indicator corresponding to the condition that $score_{\text{x}}$ can be reliably calculated for the adjective pair, and $g_{\text{x}}(\cdot)$ is a scaling function (see below). If $\alpha_{\text{x}} = 1$, then $score_{\text{x}}$ is used. Otherwise, if $\alpha_{\text{x}} = 0$, then we default to $score_{\text{y}}$. When combining three metrics x, y, and z, the combined score is given by:

$$score_{\text{x+y+z}}(j_u, j_v) = \alpha_{\text{x}} \cdot g_{\text{x}}(score_{\text{x}}(j_u, j_v)) + (1 - \alpha_{\text{x}}) \cdot score_{\text{y+z}}(j_u, j_v) \tag{6}$$

The criteria for having $\alpha_{\text{x}} = 1$ varies depending on the metric type. For pattern-based evidence (**x**='pat'), $\alpha_{\text{x}} = 1$ when adjectives j_u and j_v appear together in any of the intensity patterns culled from Google *n*-grams (e.g. a pattern like "j_u, but not j_v" exists). For lexicon-based evidence (**x**='socal'), $\alpha_{\text{x}} = 1$ when both j_u and j_v are in the SO-CAL vocabulary, and have the same polarity (i.e. are both positive or both negative). For paraphrase-based evidence (**x**='pp'), $\alpha_{\text{x}} = 1$ when j_u and j_v have one or more edges directly connecting them in JJGRAPH.

Since the metrics to be combined may have different ranges, we use a scaling function $g_{\text{x}}(\cdot)$ to make the scores output by each metric directly comparable:

$$g_{\text{x}}(w) = \text{sign}(w) \cdot \left(\frac{\log(|w|) - \mu_{\text{x}}}{\sigma_{\text{x}}} + \gamma \right) \tag{7}$$

where μ_{x} and σ_{x} are the estimated population mean and standard deviation of $\log(score_{\text{x}})$ (estimated over all adjective pairs in the dataset), and γ is an offset that ensures positive scores remain positive, and negative scores remain negative. In our experiments we set $\gamma = 5$.

5 Ranking Adjective Sets by Intensity

The first experimental application for the different paraphrase evidence is an existing model for predicting a global intensity ordering within a set of adjectives. Global ranking models are useful for inferring intensity comparisons between adjectives for which there is no explicit evidence. For example, in ranking three adjectives like *warm*, *hot*, and *scalding*, there may be direct evidence indicating *warm* < *hot* and *hot* < *scalding*, but no way of directly comparing *warm* to *scalding*. Global ranking models infer that *warm* < *scalding* based on evidence from the other adjective pairs in the scale.

5.1 Global Ranking Model

We adopt the mixed-integer linear programming (MILP) approach of de Melo and Bansal (2013) for generating a global intensity ranking. This model takes a set of adjectives $A = \{a_1, \ldots, a_n\}$

[3] https://github.com/sfu-discourse-lab/SO-CAL

Dataset	# of Scales	Min/Max/Mean Scale Size	# of Unordered (unequal) Pairs	Example Scale
deMelo	87	3 / 8 / 3.90	524 (466)	{clean} < {spotless, immaculate}
Crowd	79	2 / 8 / 3.18	293 (250)	{low} < {limited} < {scarce}
Wilkinson	21	2 / 5 / 2.81	61 (61)	{dry} < {arid} < {parched}

Table 1: Characteristics of the scalar adjective datasets used for evaluation. The deMelo scale example shows an instance of an equally-intense pair (*spotless*, *immaculate*).

and directed, pairwise adjective intensity scores $score(a_i, a_j)$ as input, and assigns each adjective a_i a place along a linear scale $x_i \in [0, 1]$. The adjectives' assigned values define the global ordering. If the predicted weights used as input are inconsistent, containing cycles, the model resolves these by choosing the globally optimal solution.

Recall that all pairwise scoring metrics produce a positive score for adjective pair (j_u, j_v) when it is likely that $j_u < j_v$, and a negative score otherwise. Consequently, the MILP approach should result in $x_u < x_v$ when $score(j_u, j_v)$ is positive, and $x_u > x_v$ otherwise. This goal is achieved by maximizing the objective function:

$$\sum_{u,v} \text{sign}(x_v - x_u) \cdot score(j_u, j_v) \qquad (8)$$

de Melo and Bansal (2013) propose a MILP formulation for maximizing this objective, which we utilize in our experiments. Note that while de Melo and Bansal (2013) incorporate synonymy evidence from WordNet in their ranking method, we do not implement this part of the model.

5.2 Experiments

We experiment with using each of the paraphrase-, pattern-, and lexicon-based pairwise scores as input to the global ranking model in isolation. To examine how the scoring methods perform when used in combination, we also test all possible ordered combinations of 2 and 3 scores.

Experiments are run over three distinct test sets (Table 1). Each dataset contains ordered sets of scalar adjectives belonging to the same *scale*. In general, scalar adjectives describing the same attribute can be ordered along a full scale (e.g. *freezing* to *sweltering*), or a half scale (*warm* to *sweltering*); all three test sets group adjectives into half scales. The three datasets are described here, and their characteristics are given in Table 1.

deMelo (de Melo and Bansal, 2013)[4]. 87 adjective sets are extracted from WordNet 'dumbbell' structures (Gross and Miller, 1990), and partitioned into half-scale sets based on their pattern-based evidence in the Google N-Grams corpus (Brants and Franz, 2009). Sets are manually annotated for intensity relations ($<$, $>$, and $=$).

Wilkinson (Wilkinson and Oates, 2016). Twelve adjective sets are generated by presenting crowd workers with small seed sets (e.g. *huge, small, microscopic*), and eliciting similar adjectives. Sets are automatically cleaned for consistency, and then annotated for intensity by crowd workers. While the original dataset contains full scales, we manually sub-divide these into 21 half-scales for use in this study. Details on the modification from full- to half-scales are in the Supplemental Material.

Crowd. We also crowdsourced a new set of adjective scales with high coverage of the PPDB vocabulary. In a three-step process, we first asked crowd workers whether pairs of adjectives describe the same attribute (e.g. temperature) and therefore should belong along the same scale. Second, sets of same-scale adjectives were refined over multiple rounds. Finally, workers ranked the adjectives in each set by intensity. The final dataset includes 293 adjective pairs along 79 scales.

We measure the agreement between the gold standard ranking of adjectives along each scale and the predicted ranking using three commonly-used metrics:

Pairwise accuracy. For each pair of adjectives along the same scale, we compare the predicted ordering of the pair after global ranking ($<$, $>$, or $=$) to the gold-standard ordering of the pair, and report overall accuracy of the pairwise predictions.

Kendall's tau (τ_b). This metric computes the rank correlation between the predicted ($r_P(J)$) and gold-standard ($r_G(J)$) ranking permutations of each adjective scale J, incorporating a correction for ties. Values for τ_b range from -1 to 1, with extreme values indicating a perfect negative

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

		Score Accuracy (before ranking)		Global Ranking Results			
Test Set	Score Type	Coverage	Pairwise Acc.	Pairwise Acc.	Avg. τ_b	ρ	Example Predicted Scale
deMelo	$score_{pat}$	0.48	**0.844**	0.650	**0.633**	**0.583**	{clean} < {spotless, immaculate}*
	$score_{pp}$	0.33	0.458	0.307	0.071	0.090	{immaculate, clean} < {spotless}
	$score_{socal}$	0.28	0.546	0.246	0.110	0.019	{clean} < {spotless} < {immaculate}
	$score_{pat+socal}$	0.61	0.757	**0.653**	0.609	0.533	{clean} < {spotless} < {immaculate}
	$score_{pat+socal+pp}$	**0.70**	0.722	0.644	0.564	0.482	{clean} < {spotless} < {immaculate}
Crowd	$score_{pat}$	0.11	**0.784**	0.321	0.203	0.221	{limited, low, scarce}
	$score_{pp}$	0.74	0.676	0.597††	0.437†	0.405	{low} < {limited} < {scarce}*
	$score_{socal}$	0.35	0.757	0.421	0.342	0.293	{limited, low, scarce}
	$score_{socal+pp}$	0.81	0.687	0.621††	0.470††	0.465	{low} < {limited} < {scarce}*
	$score_{socal+pat+pp}$	**0.82**	0.694	**0.639**††	**0.495**††	**0.480**	{low} < {limited} < {scarce}*
Wilkinson	$score_{pat}$	0.44	0.852	0.475	0.441	0.435	{quick} < {speedy, fast}
	$score_{pp}$	0.80	0.753	0.639	0.419	0.450	{quick} < {fast} < {speedy}*
	$score_{socal}$	0.31	**0.895**	0.312	0.317	0.422	{fast} < {speedy} < {quick}
	$score_{pat+pp}$	**0.89**	0.833	0.738††	0.605	0.564	{quick} < {fast} < {speedy}*
	$score_{pat+socal+pp}$	**0.89**	0.833	**0.754**††	**0.638**	**0.611**	{quick} < {fast} < {speedy}*

††: $p \leq .01$ †: $p \leq .05$

Table 2: Pairwise relation prediction and global ranking results for each score type in isolation, and for the best-scoring combinations of 2 and 3 score types on each dataset. For the global ranking accuracy and average τ_b results, we denote with the † symbol scores for metrics incorporating paraphrase-based evidence that significantly out-perform both $score_{pat}$ and $score_{socal}$ under the paired Student's t-test, using the Anderson-Darling test to confirm that scores conform to a normal distribution (Fisher, 1935; Anderson and Darling, 1954; Dror et al., 2018). Example output is also given, with correct rankings starred.

or positive correlation, and a value of 0 indicating no correlation between predicted and gold rankings. We report τ_b as a weighted average over scales in each dataset, where weights correspond to the number of adjective pairs in each scale.

Spearman's rho (ρ). We report the Spearman's ρ rank correlation coefficient between predicted ($r_P(J)$) and gold-standard ($r_G(J)$) ranking permutations. For each dataset, we calculate this metric just once by treating each adjective in a particular scale as a single data point, and calculating an overall ρ for all adjectives from all scales.

5.3 Experimental Results

The results of the global ordering experiment, reported in Table 2, are organized as follows: *Score Accuracy* pertains to performance of the scoring methods alone – prior to global ranking – while *Global Ranking Results* pertains to performance of each scoring method as used in the global ranking algorithm. Within *Score Accuracy* there are two metrics. *Coverage* gives the percent of unique same-scale adjective pairs from the test set that can be directly scored using the given method. For $score_{pat}$, covered pairs are all those that appear together in any recognized pattern; for $score_{pp}$, covered pairs are those directly connected in JJGRAPH by one or more direct edges; for $score_{socal}$, covered pairs are all those for which both adjectives are in the SO-CAL lexicon and the metric is defined. *Pairwise Accuracy* gives the accuracy of the scoring method (before global ranking) on *just the covered pairs*, meaning that the subset of pairs scored by each method varies. Within *Global Ranking Results*, we report pairwise accuracy, weighted average τ_b, and ρ calculated over *all pairs* after ranking – including both pairs that are covered by the scoring method, and those whose pairwise intensity relationship has been inferred by the ranking algorithm.

The results indicate that the pairwise score accuracies (before ranking) for $score_{pat}$ and $score_{socal}$ are higher than those of $score_{pp}$ for all datasets, but that their coverage is relatively limited. The one exception is the deMelo dataset, where $score_{pat}$ has high coverage because the dataset was compiled specifically by finding adjective pairs that matched lexical patterns in the corpus. For all datasets, highest coverage is achieved using one of the combined metrics that incorporates paraphrase-based evidence.

The impact of these trends is visible on the

Global Ranking Results. When using pairwise intensity scores to compute the global ranking, higher coverage by a metric drives better results, as long as the metric's accuracy is reasonably high. Thus the paraphrase-based $score_{\text{pp}}$, with its high coverage, gets better global ranking results than the other single-method scores for two of the three datasets. Further, we find that boosting coverage with a combined metric that incorporates paraphrase evidence produces the highest post-ranking pairwise accuracy scores overall for all three datasets, and the highest average τ_b and ρ on the Crowd and Wilkinson datasets. We conclude that incorporating paraphrase evidence can improve the quality of this model for ordering adjectives along a scale because it gives high coverage with reasonably high quality.

The performance trends on the deMelo dataset differ from those on the Crowd and Wilkinson datasets. In particular, $score_{\text{pp}}$ and $score_{\text{socal}}$ have substantially lower pre-ranking pairwise accuracy on the pairs they cover in the deMelo dataset than they do for Crowd and Wilkinson: $score_{\text{pp}}$ has an accuracy of just 0.458 on covered pairs in the deMelo dataset, compared with 0.676 and 0.753 on the Crowd and Wilkinson datasets, and score differences for $score_{\text{socal}}$ are similar. The near-random prediction accuracies of $score_{\text{pp}}$ and $score_{\text{socal}}$ on deMelo before ranking lead to near-zero correlation values on this dataset after global ranking. To explore possible reasons for these results, we assessed the level of human agreement with each dataset in terms of pairwise accuracy. For each test set, we asked five crowd workers to classify the intensity direction for each adjective pair (j_u, j_v) in all scales as less than ($<$), greater than ($>$), or equal ($=$). We found that humans agreed with the 'gold standard' direction 65% of the time on the Bansal dataset, versus 70% of the time on the Crowd and Wilkinson datasets. It is possible that the more difficult nature of the Bansal dataset, coupled with its method of compilation (i.e. favoring adjective pairs that co-occur with pre-defined intensity patterns), lead to the lower coverage and lower accuracy of $score_{\text{pp}}$ and $score_{\text{socal}}$ on this dataset.

6 Indirect Question Answering

The second task that we address is answering indirect *yes* or *no* questions. de Marneffe et al. (2010) observed that answers to such polar questions frequently omit an explicit *yes* or *no* response. In some cases the implied answer depends on the relative intensity of adjective modifiers in the question and answer. For example, in the exchange:

> Q: *Was he a successful ruler?*
> A: *Oh, a tremendous ruler.*

the implied answer is *yes*, which is inferred because *successful* $\leq$ *tremendous* in terms of relative intensity. Conversely, in the exchange:

> Q: *Does it have a large impact?*
> A: *It has a medium-sized impact.*

the implied answer is *no* because *large* $>$ *medium-sized*.

de Marneffe et al. (2010) compiled an evaluation set for this task by extracting 123 examples of such indirect question-answer pairs (IQAP) from dialogue corpora. In each exchange, the implied answer (annotated by crowd workers to be *yes* or *no*[5]) depends on the relative intensity relationship between modifiers in the question and answer texts. In their original paper, the authors utilize an automatically-compiled lexicon to make a polarity prediction for each IQAP.

6.1 Predicting Answer Polarity

Our goal is to see whether paraphrase-based scores are useful for predicting the polarity of answers in the IQAP dataset. As before, we compare the quality of predictions made using the paraphrase-based evidence with predictions made using pattern-based, lexicon-based, and combined scoring metrics.

To use the pairwise scores for inference, we employ a decision procedure nearly identical to that of de Marneffe et al. (2010). If j_q and j_a are scorable (i.e. have a scorable intensity relationship along the same half-scale), then $j_q \leq j_a$ implies the answer is *yes* (first example above), and $j_q > j_a$ implies the answer is *no* (second example). If the pair of adjectives is not scorable, then the predicted answer is *no*, as the pair could be antonyms or completely unrelated. If either j_q or j_a is missing from the scoring vocabulary, the adjectives are impossible to compare and therefore the prediction is *uncertain*. The full decision procedure is given in Figure 4.

[5]The original dataset contains two additional examples where the answer is annotated as *uncertain*, but de Marneffe et al. (2010) exclude them from the results and so do we.

Given: A dialogue exchange consisting of a polar question and answer, where the answer depends on the relative intensities of distinct modifiers j_q and j_a in the question and answer respectively:

1. if j_q or j_a are missing from the score vocabulary, predict "UNCERTAIN"
2. else, if $score(\text{JJ}_q, \text{JJ}_a)$ is undefined, predict "NO"
3. else, if $score(\text{JJ}_q, \text{JJ}_a) \geq 0$, predict "YES"
4. else, predict "NO"
5. If the question or answer contains negation, map a "YES" answer to "NO and a "NO" answer to "YES"

Figure 4: Decision procedure for using pairwise intensity scores for predicting polarity of an IQAP instance, based on de Marneffe et al. (2010).

6.2 Experiments

The decision procedure in Figure 4 is carried out for the 123 IQAP instances in the dataset, varying the score type. We report the accuracy, and macro-averaged precision, recall, and F1-score of the 85 *yes* and 38 *no* instances, in Table 3 alongside the percent of instances with adjectives out of vocabulary. Only the combined scores for the two best-scoring combinations, $score_{\text{socal+pp}}$ and $score_{\text{socal+pat+pp}}$, are reported.

Method	%OOV	Acc.	P	R	F
all-"YES"	.00	**.691**	.346	.500	.409
deMarneffe (2010)	.02	.610	.597	.594	.596
$score_{\text{socal}}$	.33	.504	**.710**	.481	.574
$score_{\text{pp}}$	.09	.496	.568	.533	.550
$score_{\text{pat}}$	.07	.407	.524	.491	.507
$score_{\text{socal+pp}}$	.09	.634	.690	.663	.676
$score_{\text{socal+pat+pp}}$	.06	.642	.684	**.683**	**.684**

Table 3: Accuracy and macro-averaged precision (P), recall (R), and F1-score (F) over *yes* and *no* responses on 123 question-answer pairs. The percent of pairs having one or both adjectives out of the score vocabulary is listed as %OOV.

The simplest baseline of predicting all answers to be "YES" gets highest accuracy in this imbalanced test set, but all score types perform better than the all-"YES" baseline in terms of precision and F1-score. Bouyed by its high precision, the $score_{\text{socal}}$ – which is derived from a manually-compiled lexicon – scored higher than $score_{\text{pp}}$ and $score_{\text{pat}}$. But it mis-predicted 33% of pairs as *uncertain* because of its limited overlap with the IQAP vocabulary. Meanwhile, $score_{pp}$ had relatively high coverage and a mid-level F-score, while $score_{pat}$ scored poorly on this dataset due to its sparsity; while all modifiers in the IQAP dataset are in the Google N-grams vocabulary, most do not have observed patterns and therefore return predictions of "NO" (item 2 in Figure 4). As in the global ranking experiments, the paraphrase-based evidence is complementary to the lexicon-based evidence, and thus the combined $score_{\text{socal+pp}}$ and $score_{\text{socal+pat+pp}}$ produce significantly better accuracy than any score in isolation (McNemar's test, $p < .01$), and also out-perform the original expected ranking method of de Marneffe et al. (2010) (although they do not beat the best-reported score on this dataset, F-score=0.706 (Kim and de Marneffe, 2013)).

7 Conclusion

We have proposed adjectival paraphrases as a new source of evidence for predicting intensity relationships between scalar adjectives. While paraphrase-based intensity evidence produces pairwise predictions that are less precise than those produced by pattern- or lexicon-based evidence, the coverage is substantially higher. Thus paraphrases can be successfully used as a complementary source of information for reasoning about adjective intensity.

Acknowledgments

This material is based in part on research sponsored by the following organizations: the Allen Institute for Artificial Intelligence (AI2) Key Scientific Challenges program, the Google Ph.D. Fellowship program, the French National Research Agency under project ANR-16-CE33-0013, and DARPA under grant numbers FA8750-13-2-0017 (the DEFT program) and HR0011-15-C-0115 (the LORELEI program). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA and the U.S. Government.

We are grateful to our anonymous reviewers for their thoughtful and constructive comments.

References

Theodore W Anderson and Donald A Darling. 1954. A test of goodness of fit. *Journal of the American statistical association*, 49(268):765–769.

Thorsten Brants and Alex Franz. 2009. Web 1T 5-gram, 10 European languages version 1. *Linguistic Data Consortium, Philadelphia*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment: First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, chapter The PASCAL Recognising Textual Entailment Challenge. Springer Berlin Heidelberg, Berlin, Heidelberg.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1383–1392.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Ronald Aylmer Fisher. 1935. The design of experiments.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, pages 758–764, Atlanta, Georgia.

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 the 31st Annual Meeting on Association for Computational Linguistics (ACL)*, pages 172–182, Columbus, Ohio.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2 (COLING)*, pages 539–545, Nantes, France.

Hans Kamp and Barbara Partee. 1995. Prototype theory and compositionality. *Cognition*, 57(2):129–191.

Joo-Kyung Kim and Marie-Catherine de Marneffe. 2013. Deriving adjectival scales from continuous space word representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1625–1630, Seattle, Washington.

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)*, pages 167–176, Uppsala, Sweden.

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.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Marcin Morzycki. 2015. *Modification*. Cambridge University Press.

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

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL) (Volume 2: Short Papers)*, pages 425–430, Beijing, China.

Sven Rill, J. vom Scheidt, Johannes Drescher, Oliver Schütz, Dirk Reinel, and Florian Wogenstein. 2012. A generic approach to generate opinion lists of phrases for opinion mining applications. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)*, Beijing, China.

Josef Ruppenhofer, Jasper Brandes, Petra Steiner, and Michael Wiegand. 2015. Ordering adverbs by their scaling effect on adjective intensity. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 545–554, Hissar, Bulgaria.

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 (EACL)*, Gothenburg, Sweden.

Raksha Sharma, Mohit Gupta, Astha Agarwal, and Pushpak Bhattacharyya. 2015. Adjective intensity and sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods for Natural Language Processing (EMNLP)*, Lisbon, Portugal.

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.

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

Chaitanya P. Shivade, Marie-Catherine de Marneffe, Eric Fosler-Lussier, and Albert M. Lai. 2015. Corpus-based discovery of semantic intensity scales. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies (NAACL HLT)*, pages 483–493, Denver, Colorado.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Bryan Wilkinson. 2017. *Identifying and Ordering Scalar Adjectives Using Lexical Substitution*. Ph.D. thesis, University of Maryland, Baltimore County.

Bryan Wilkinson and Tim Oates. 2016. A gold standard for scalar adjectives. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, Portoro, Slovenia.