ADJSCALES: DIFFERENTIATING BETWEEN SIMILAR ADJECTIVES FOR LANGUAGE LEARNERS

Vera Sheinman and Takenobu Tokunaga
Department of Computer Science, Tokyo Institute of Technology, Ookayama 2-12-1, Meguro-ku Tokyo 152-8552, Japan

Keywords: Adjective-scales, Patterns, Web-based, Second language learning, Gradation, Computational linguistics.

Abstract: In this study we introduce AdjScales, a method for scaling similar adjectives by their strength. It combines existing Web-based computational linguistic techniques in order to automatically differentiate similar adjectives that describe the same property by strength. Though this kind of information is rarely present in most of the lexical resources and dictionaries, it might be useful for language learners that try to distinguish between similar words and that want to capture the differences from a single structure. Additionally, AdjScales might be used by constructors of lexical resources in order to enrich them. The method is evaluated by comparison with annotation on a subset of adjectives from WordNet by four native English speakers. The collected annotation is an interesting resource by its own right. This work is a first step towards automatic differentiation of meaning between similar words for learners.

1 INTRODUCTION

In the process of building their vocabulary, language learners sometimes need to choose an appropriate word to use from a set of near-synonymous words. The subtle differences between words and the fact that the semantics of near-synonyms between the native language and the second language usually overlap only partially make it all more difficult. Consider for example the sentences, “This film is **good**”, “This film is **great**”, “This film is **superb**”. All of these give a positive evaluation of a film, but in which one and under what circumstances will the film be perceived by a native speaker of English as the best? How is the learner to know?

A **Linguistic Scale** is a set of words of the same grammatical category, which can be ordered by their semantic strength or degree of informativeness (Levinson, 1983). Linguistic scales are lexicalized for various parts of speech. For instance, ⟨*surprise, startle, shock*⟩ is a verbal scale(Chklovski and Pantel, 2004).

Existing linguistic resources and dictionaries rarely contain information on adjectives being part of a scale, or being of a particular strength. Though, this information may be deduced in some cases from the word definition, such as “very small” for “tiny” in WordNet (Miller, 1995), it is not always so, and lacks the convenience of a single visual scale like *infinitesimal→ tiny→small→smallish*.

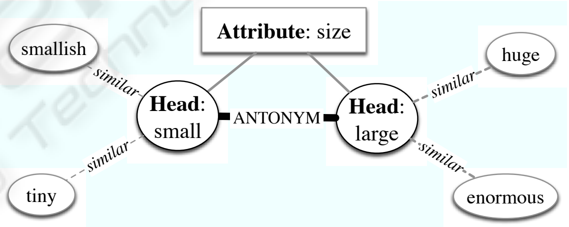

Figure 1: Descriptive adjectives encoding in WordNet.

Gradation is a related term describing variation of strength between adjectives. (Fellbaum et al., 1993) describes gradation as a semantic relation organizing lexical memory for adjectives and provides six examples of gradation for attributes SIZE, WHITENESS, AGE, VIRTUE, VALUE, and WARMTH. For instance, the example gradation for SIZE is ⟨*astronomical, huge,* ***large,*** *standard,* ***small,*** *tiny, infinitesimal*⟩. According to Fellbaum, gradation is rarely lexicalized in English, and thus it is not encoded in WordNet. Adverbial expressions like “slightly” or comparative expressions like “more” are usually preferred. While agreeing with this claim, we believe that having a method for grading adjectives that are lexicalized is important and beneficial for learners that struggle with similar adjectives. More-

Sheinman V. and Tokunaga T. (2009).
ADJSCALES: DIFFERENTIATING BETWEEN SIMILAR ADJECTIVES FOR LANGUAGE LEARNERS.
In *Proceedings of the First International Conference on Computer Supported Education*, pages 228-234
DOI: 10.5220/0001972602280234
Copyright © SciTePress

over, using the Web as a corpus, this information may be extracted with less effort than before.

Descriptive adjectives describe a property and tend to be scalar. WordNet encodes them in clusters (**adjective-sets**). Two antonymous representative synsets (**head-words**) are linked to a noun they describe (**attribute**). Each *head* adjective is linked to **similar** adjectives. Relations between the *similar* adjectives and differences between the SIMILAR connections are not encoded. In the example encoding in Figure 1, there is a clear difference between "smallish" that is slightly less small than "small", and "tiny" that is normally perceived to be *smaller* than "small". In this work, our objective is to identify such cases and to provide this kind of distinction.

The similar adjectives in each adjective-set in WordNet are not identical, and usually each synset provides a nuance of meaning that differentiates it from others. In addition to STRENGTH, there are others, such as INFORMAL-LANGUAGE-OF relation that holds between "teeny-weeny" and "small". Detecting these kinds of relations is also important in the context of lexical choice by learners. Gradation being very central in adjectives, other possible relations are left out of the scope of this work.

We introduce an automatic Web-based approach to extract strength information for adjectives, AdjScales, that incorporates recent advances in Natural Language Processing. In choosing the suitable methods for this task, our goal was simple and freely accessible methods that do not require any special corpora, parsing or tagging. The novelty of AdjScales is in its automatic construction of adjective-scales from several examples, in the language learner as the target user, and in its evaluation. This work can contribute to improving existing language resources, textbook authoring, and tools for learners.

2 PROPOSED METHOD: ADJSCALES

2.1 Pattern Extraction

Pattern extraction is a preparatory step for AdjScales. Similarly to (Davidov and Rappoport, 2008), we use **pattern-extraction-queries** of the form "a $*$ b" to find patterns where *a*, *b* are **seed words**, and "$*$" denotes a wildcard[1]. We extract binary patterns of the form

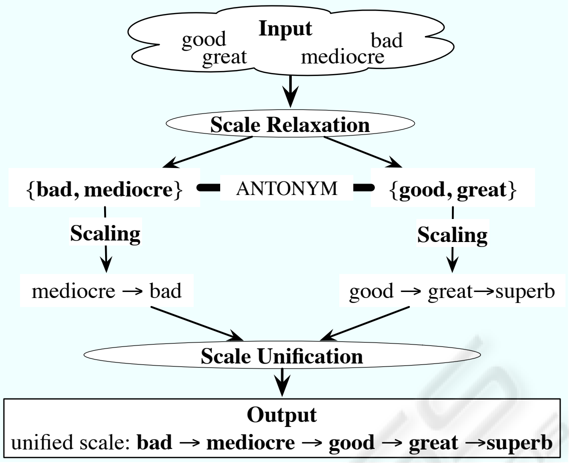

Figure 2: General Illustration of the Proposed Method.

$$p = [\text{prefix}_\text{p} \quad x \quad \text{infix}_\text{p} \quad y \quad \text{postfix}_\text{p}]$$

from the snippets of the query results returned by a search engine[2]. Snippets are a good source for patterns, because they contain the direct context of the query text[3]. A pattern p can be instantiated by a pair of words w_1, w_2 to result in a phrase

$$p(w_1,w_2) = \text{“prefix}_\text{p} \quad w_1 \quad \text{infix}_\text{p} \quad w_2 \quad \text{postfix}_\text{p}\text{”},$$

or similarly it can be instantiated by a word w_1, and a wildcard to result in a phrase "prefix$_\text{p}$ w_1 infix$_\text{p}$ $*$ postfix$_\text{p}$" to search for words cooccurring with the word w_1 in a pattern.

Let's consider an example pattern p_1 where prefix$_{\text{p}_1}$ $= \phi$, infix$_{\text{p}_1}$ = "if not", and postfix$_{\text{p}_1}$ $= \phi$, if we instantiate it with the pair of words (good, great) we will get a phrase $p_1(\text{good, great}) =$ "good if not great". Instantiating it with ($*$, good) will result in a phrase $p_1(*, \text{good}) =$ "$*$ if not good" that can be used to search for items appearing on the left side of the pattern p_1 with the word "good".

If $p(w_1, w_2)$ appears in snippets that are returned by a search engine for a pattern-extraction-query, we refer to it as p is **supported-by** (w_1, w_2).

Differently from (Davidov and Rappoport, 2008), we choose the seed word pairs in a supervised manner, so that *seed*$_2$ STRONGER-THAN *seed*$_1$. For the experimental settings described in this work we used 10 seed word pairs selected from the adjective scale examples in (Fellbaum et al., 1993). The relation STRONGER-THAN is asymmetric, therefore, we select only the **asymmetric patterns** that are extracted consistently so that the weaker word in each supporting

[1] $*$ denotes 0 or more terms that may appear in its place. In practice, search engines, usually use the notation of $*$ for a single-word, and we used the queries "a b", "a $*$ b", "a $*$ $*$ b" for each pattern-extraction-query.

[2] We use Yahoo search API (Yahoo Inc., 2008) throughout the experiments described in this paper.

[3] For the extraction purposes snippets are split into sentences and are cleaned from all kinds of punctuation.

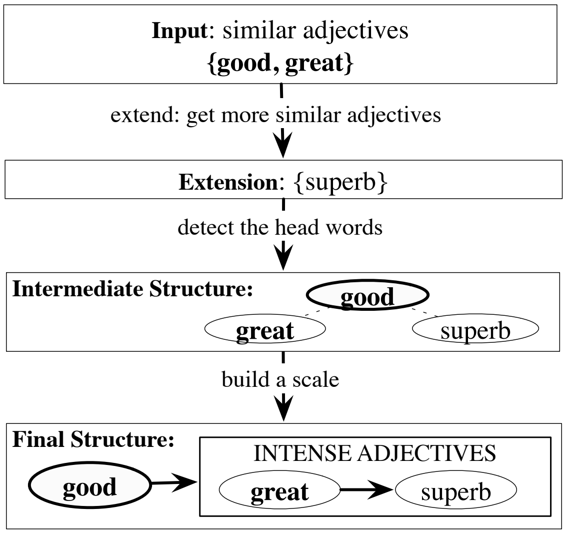

Figure 3: AdjScales Core.

pair is on the left side of the pattern (before the infix words) or so that the weaker word is on the right side of the pattern (after the infix words). If not all the supporting pairs of words share the same direction, the pattern is discarded. We define the former selected patterns as **intense**, and the latter as **mild**.

We select only the patterns supported by at least 3 seed pairs and we require a pattern instance by each supporting pair to repeat at least twice in the sentences extracted from the snippets to increase reliability. We also require the patterns to be supported by adjectives describing different properties. This constraint is important, because patterns that are supported by seeds that describe the same property tend to appear in very specific contexts and are not useful for other properties. For instance, [*x* even *y* amount] might be extracted while supported only by seeds describing the SIZE property, such as (huge, astronomical), (big, huge), (tiny, infinitesimal).

To exclude patterns that are too short and tend to be too generic, if pattern *p* is included in pattern *q*, and both of them match the other requirements, we select only the longer pattern, *q*.

2.2 Method Steps

AdjScales method, outlined in Figure 2, comprises several steps listed below with **Scaling** (Section 2.2.5) being its core. We divide the input adjectives into two subsets in a **Scale Relaxation** (Section 2.2.2) step. Then, the rest of the method is performed on each of the subsets separately until the results are unified in the final step of **Scale Unification** (Section 2.2.6) outputing an adjective scale.

2.2.1 Input

AdjScales expects at least 2 similar adjectives as the input. One adjective leaves the task of scaling open for too many interpretations, while two adjectives give a good clue on what scale is interesting for the user. Similar adjectives for our purposes are adjectives that describe the same property.

2.2.2 Scale Relaxation

According to (Hatzivassiloglou and McKeown, 1993), for adjectives, the total scale is commonly relaxed, so that elements of the scale can be partitioned into several subscales. Consider the adjective scale ⟨*cold, lukewarm, warm, hot*⟩. It is not clear what is the scale relationship between antonyms, such as “cold” and “hot”. A total order by the relation of strength within the subscale ⟨*lukewarm, warm, hot*⟩ is, however, evident.

In the Scale Relaxation step, AdjScales divides the input into two antonymous subsets using antonymy and similarity information from WordNet. If the input words belong to the same adjective-set structure in WordNet they are divided by their similarity to the representative antonyms in the set. If the input words all belong to the same subset they will remain in the same set for the next steps. In other cases, if not all the words appear in WordNet, or they are not encoded in the same adjective-set structure, we currently assume that the input words belong to a single subscale.

2.2.3 Extension

AdjScales attempts to provide the user with further similar adjectives that do not appear in the input. Adjectives that are encoded as SIMILAR to the input adjectives in WordNet are added to the subset as an extension. For cases where WordNet is not applicable no extension is currently performed.

2.2.4 Intermediate Structure

WordNet encodes adjectives by selecting the head adjectives in each adjective-set and connecting the other adjectives to them with similarity links. The relation between the head adjectives is antonymous. We keep this type of encoding and call it an **Intermediate Structure**. For cases where the input adjectives do not appear in WordNet, we select the most frequent words sharing the same context as others as the head words. The Intermediate Structure allows us to reduce the pairwise computations in the Scaling step. It also allows the learner using the system to recognize the most basic words in the scale.

2.2.5 Scaling

The Scaling step depends only on availability of a search engine that estimates page counts. For this step, we refer to the set of patterns preselected by Pattern Extraction (Section 2.1) as P. For each pair (head-word, similar-word) from the Intermediate Structure, we instantiate each pattern p in P to obtain phrases $s_1 = p(\text{head-word, similar-word})$ and $s_2 = p(\text{similar-word, head-word})$. We estimate document frequency, $df(s_i)$, by using the estimated page count hits returned by the search engine. We run the resulting 2 phrases as 2 separate queries and check whether $df(s_1) > weight \times df(s_2)$ and whether $df(s_1) > threshold$. The higher the values are for the $threshold$ and $weight$ parameters, the more reliable are the results, and the fewer there are. If p is of the type *intense*, then a positive value is added to the similar-word, otherwise if p is of the type *mild* a negative value is added. When all the patterns are tested, similar-words with positive values are classified as *intense*, while the similar-words with negative values are classified as *mild*. Words that do not receive any points are classified as *unconfirmed*. For each pair of words in the each one of the subsets (*mild* and *intense*), the same procedure is repeated, creating further subsets of *mildest* words that have the most negative values within the *mild* subset, and *most intense* words for the words with the highest positive values within the *intense* subset. The information is recored in a **Final Structure** that can be visualized as a scale *mildest words* → ⋯ → *least mild words* → *head-words* → *least intense words* → ⋯ → *most intense words*.

To illustrate this process, consider the example shown in Figure 3. Assume that $P = \{p_1 = [x \text{ if not } y]\}$. The Intermediate Structure in the example contains head-words={good}, and similar-words={great, superb}. We instantiate $s_1 = p_1(\text{good, great}) = $ "good if not great", $s_2 = p_1(\text{great, good})$. Choosing $weight = 3$ and $threshold = 100$ pages, we run the queries s_1, s_2. Google estimates $df(s_1)$ as $353,000$ and $df(s_2)$ as 108[4]. p_1 is a pattern of type *intense*, therefore a point will be added to the word "great". Similarly, $df(p_1(\text{good, superb}) > 3 \times df(p_1(\text{superb, good}))$, and $df(p_1(\text{great, superb}) > 3 \times df(p_1(\text{superb, great}))$. There are no *mild* or *unconfirmed* words in this example, resulting in the final structure:
{head-words={good},
intense words={great (-1) →superb (1)}}

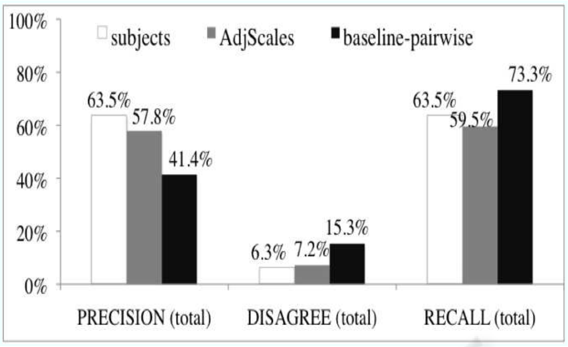

Figure 4: General pairwise agreement between AdjScales compared to human subjects.

or simply *good* → *great* → *superb*.

2.2.6 Scales Unification

Subscales may be unified into a single adjective scale simply by adding a link between the mildest words on both sides. The internal STRONGER-THAN links point from the mildest words towards the extremities. In fact, different properties are measured for each subscale. For instance, the words "good", "great", and "superb" in our running example measure GOODNESS, while their opponents "bad" and "mediocre" measure BADNESS. To present a unified scale of adjectives that describe the property VALUE we reverse the direction of links in one of the scales.

3 EVALUATION

For evaluation, we preselected 16 patterns (11 intense and 5 mild) in the manner described in Section 2.1.

3.1 WordNet-based Corpus

We extracted descriptive adjective-sets[5] from WordNet 3.0 as the input to our system for evaluation of the scaling step, and divided them into antonymous subsets. Four native English speakers (2 Americans and 2 British[6]), all male students from engineering departments scaled the adjective subsets.

Some subsets in WordNet are too big to be scaled by human subjects[7], and they need pruning. We downloaded snippets for queries of the type

[4] These figures are true for a search performed on 6th of December, 2008 and might change slightly depending on the date and the location of the search.

[5] The extracted adjective-sets comprise 757 head words (645 distinct words) and 6,607 similar words (5,378 distinct words).

[6] We have observed no particular differences between British and American subjects in the evaluation

[7] The subset of "lean, thin" for attribute BODY WEIGHT comprises 51 similar words, for instance.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Table 1: Subjects selections for WordNet adjectives.

	#words mild	#words intense
subject$_1$	137	358
subject$_2$	99	301
subject$_3$	89	290
subject$_4$	141	313
All subjects	22	163

$p(\text{head-word}, *)$ and $p(*, \text{head-word})$ for each pattern p from the preselected patterns and for all the head-words resulting in 625MB of data. If a word in an adjective-set was not among the words appearing in the wildcard slots in the extracted snippets, it was pruned. The reasoning behind pruning is that currently our method cannot provide scaling decisions for words that do not appear in any patterns, so, to test this approach only the words that are potentially applicable are considered. The final dataset for evaluation contained 308 subsets with 763 similar words to be scaled.

Each subject scaled adjectives independently. For each subset in the dataset, the subject was presented with the head-words, attribute, and a set of similar words. The head-words were fixed as **neutral** and we asked the subjects to classify each similar word as one of 5 types, compared to the head-words and to the other words. When a word seemed stronger than the head-words it was to be classified as **intense** or **very intense**. When it seemed weaker than the head, it was classified as **mild** or **very mild**. Words of similar intensity to the head words were to be classified as *neutral*. When not sure about a certain word or thinking that it is not applicable for scaling, the subject classified it as **not sure** or **not applicable**, respectively.

We measure agreement between two subjects or between AdjScales and a subject as follows. First, **general agreement** is measured. If a word w in subset s is selected as *mild* or as *very mild* by subject A (or selected as *mild* by AdjScales), we will denote it as $w \in gen - mild_A$ (likewise for *intense*). Two subjects A and B agree if $w \in gen - mild_A \wedge w \in gen - mild_B$ or if $w \in gen - intense_A \wedge w \in gen - intense_B$. There are many words that were undetermined (*not sure, not applicable,* or *unconfirmed* for AdjScales), so it was important to also measure the **general disagreement** explicitly. For each two subjects A and B we measure

$$precision = \frac{|gen - mild_A \cap gen - mild_B|}{gen - mild_A},$$

$$disagreement = \frac{|gen - mild_A \cap gen - intense_B|}{gen - mild_A},$$

$$recall = \frac{|gen - mild_A \cap gen - mild_B|}{gen - mild_B}$$

Table 2: Additional adjective scales.

	precision	disagreement	recall
AdjScales	91.30%	8.70%	56.76%

for general agreement for words selected as *generally mild*. Same notation holds for *generally intense.* Table 1 shows the *generally mild* and the *generally intense* selections made by the subjects, and the number of selections all four of them agreed upon.

We averaged the pairwise agreement between the subjects, and we averaged pairwise agreement of AdjScales with each one of the subjects. Additionally, we ran a baseline method that selected the most frequently chosen classification, *intense* for all words, and compared it in a similar manner. The comparison between the subjects, AdjScales[8], and baseline is shown in the chart in Figure 4.

We also compared AdjScales to the answers that were generally agreed upon by all 4 subjects. AdjScales disagreed with the four subjects consensus for only one word. Additionally, to understand the finer agreement on ordering adjectives on a scale, we measure **order agreement**. Subjects A and B agree on order of a pair $w_1,\ w_2$ if A and B both classified w_1 and w_2 as generally mild or as neutral and if both A and B classified w_1 as milder than w_2. They disagree if A classified w_1 as milder than w_2 while B put them in the inverse order. The same is true for the intense side likewise. The subjects tend to agree on the order between words (for 86.11% of word pairs within the mild side, and for 88.74% of word pairs within the intense side). AdjScales scores 86.11% and 70.20% for order agreement for the mild and the intense sides respectively.

3.2 Additional Adjective Scales

In independent experimental settings we requested 2 native English speakers and 3 non-native English speakers to produce as many linguistic scales as possible. After the production step the subjects cross-verified the results, and only the scales agreed upon by all of them remained. We selected 9 of the scales that were adjective scales for our dataset. We ex-

[8] We used AdjScales parameters $weight = 15,\ threshold = 20$ set empirically. In order to reduce the search engine queries required for computation of each scale, we grouped the queries of patterns into 4 subgroups unifying m patterns instances in each subgroup by the operator *OR*:

$$\text{“}p_1(w_1, w_2)\text{”}\ OR\ \text{“}p_2(w_1, w_2)\text{”}\ OR\ \ldots\ OR\ \text{“}p_m(w_1, w_2)\text{”}$$

tended this dataset by adding several adjective scales from examples in the literature, such as 4 adjective scales from a teaching resource (Cadman, 2008). In the exercises in the suggested resource, several verb and adjective scales are provided, where the students are requested to order them by strength. The dataset of this kind is less confusing than adjective-sets in WordNet that are not organized as scalable-sets to begin with.

We relaxed each of the scales in our dataset manually into 2 antonymous subsets, where there were two antonymous components and performed scaling, resulting in 21 subsets. We compared the scaling results by AdjScales with the same parameters as in section 3.1 to the expected scales as shown in Table 2.

4 RELATED WORK

A major work in differentiation between near-synonyms in computational linguistics by (Inkpen and Hirst, 2006) provides a list of types of nuances of meaning that need to be differentiated, such as stylistic, attitudinal etc. Using automatic methods to differentiate between near-synonyms is the objective of our research, and in this sense this work is relevant to ours. Language learners have difficulty in perceiving the differences among near-synonyms and adding distinctions on these subtleties to existing language resources is needed. Currently, we focus only on differentiation of similar adjectives by strength. Adding further types of differentiation is a much needed extension of our work in the future. Differently from Inkpen and Hirst that use a collection of machine readable dictionaries, we use the Web as a corpus.

(Hatzivassiloglou and McKeown, 1993) established the first step towards automatic identification of adjective scales. They provide an excellent background on adjectives and a general plan to identify adjective scales, though, they focus only on clustering of similar adjectives.

Using patterns extracted from large corpora in order to learn semantic relations between words is a common approach in computational linguistics. The pioneering work (Hearst, 1992) extracted hyponym (IS-A*)* and meronym (PART-OF) relations. Further studies (Chklovski and Pantel, 2004; Davidov and Rappoport, 2008; Turney, 2008) intensively extend this methodology, further relations are explored, supervised and unsupervised methods are introduced. Our work belongs to this school of relation extraction.

VerbOcean (Chklovski and Pantel, 2004), explores fine-grained relations between verbs, STRONGER-THAN being one of them. Their work is very similar to ours in using lexico-syntactic patterns extracted from the Web. Their selection of patterns is manual, and it is based on training on 50 verb pairs, with a total of 8 patterns selected for the STRONGER-THAN relation. We utilize the asymmetry of the STRONGER-THAN relation in a similar manner to VerbOcean. We differ in our focus on adjectives and in our evaluation procedure. VerbOcean, providing differentiation between similar verbs, should be considered in the context of language learners.

A large body of research (Turney and Littman, 2003; Popescu and Etzioni, 2005) has been conducted in the field of **opinion mining**. An important distinction for opinion mining is **semantic orientation** (**positive, negative**, or **neutral**) of words and utterances. In this work we do not distinguish between the positive or the negative senses of adjectives, but rather make a more general distinction of the extent of adjectival descriptive strength. We also have a different objective to provide linguistic distinction between synonymous adjectives for learners, while the research in opinion mining concentrates on strength of subjectivity and sentiment of words, and texts.

One of the main approaches in opinion mining is extraction of semantic information from the Web, and typically adjectives play a central role in understanding opinion from texts. In these aspects, this field is related to our work. According to (Turney and Littman, 2003) semantic orientation of a word, in addition to its direction also comprises intensity, **mild** or **strong**. They compute intensity in a combined computation of the direction, using statistical association with a set of positive and negative paradigm words. OPINE (Popescu and Etzioni, 2005), a system for product reviews mining ranks opinion words by their strength as one of its subtasks. Both of these works focus on detection of semantic orientation of words and report on a very limited evaluation of ranking by strength.

No previous work that we are aware of proposes an automatic method to identify adjective-scales for language learners.

5 DISCUSSION

We have presented AdjScales, a method to build adjective scales from several examples of similar adjectives using a state-of-the-art methodology of extracting relations using patterns over the Web. It is simple, and the only required resource (for Scaling step, which is the main focus of this work) is access to a search engine. Overall, as can be seen from the evaluation, AdjScales scales similar adjectives only slightly less well than human subjects and much better than the baseline[9]. It also performs quite well on examples that seem more relevant in the context of a language learner, although quite a few words still remain unconfirmed by the system (recall of only 56% for the additional examples). There is only one disagreement of the system with answers that all human subjects agree upon, suggesting that in cases where a scale is clear and thus suitable for learning, AdjScales will be more accurate. The usefulness of the system for learners is in the area of differentiation between similar words using a simple structure (scale) to visualize it.

A surprising observation from our experiments is an unexpected asymmetry between the adjectives on the mild side and the intense side of the head words in WordNet. Subjects and AdjScales consistently selected less words as mild, and also performed less well for their mild selection. It may suggest that WordNet structure or even language structure itself, is such that there are many more words to intensify the common head-words rather than weaken them. We have also observed from analysis of the results that some patterns perform better for mild words while others do better in identification of intense words. This direction will be further explored in the future.

Similar adjectives in general and adjectives in the same adjective-set in WordNet differ in more than one way. In many cases the subjects faced a difficulty in scaling similar words that were presented to them, because they were different in several aspects. This suggests that the similar adjectives in adjective-sets in WordNet are not necessarily in the same **scalable-adjectives** set. We plan to study how to detect adjectives that are on the same scale.

Some adjectives are much more intense than others, while others are only slightly so. Estimating the distances between the links on a scale seems to be an interesting task that may be a useful visualization for learners.

REFERENCES

Cadman, D. (2008). Shades of meaning. www.primaryresources.co.uk/ english/ pdfs/ 8shades.pdf.

Chklovski, T. and Pantel, P. (2004). VerbOcean: Mining the web for fine-grained semantic verb relations. In *EMNLP-04*, pages 33–40, Barcelona, Spain.

Davidov, D. and Rappoport, A. (2008). Unsupervised discovery of generic relationships using pattern clusters and its evaluation by automatically generated SAT analogy questions. In *ACL-08: HLT*, pages 692–700, Columbus, Ohio.

Fellbaum, C., Gross, D., and Miller., K. (1993). Adjectives in wordnet. In *Five papers on WordNet*. Princeton, USA.

Hatzivassiloglou, V. and McKeown, K. R. (1993). Towards the automatic identification of adjectival scales: clustering adjectives according to meaning. In *ACL-93*, pages 172–182.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *COLING-92*, pages 539–545.

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

Levinson, S. C. (1983). *Pragmatics*, chapter Conversational Implicature, pages 132–134. Cambridge University Press, 2000 edition.

Miller, G. A. (1995). Wordnet: a lexical database for English. *ACM*, 38(11):39–41.

Popescu, A.-M. and Etzioni, O. (2005). Extracting product features and opinions from reviews. In *HLT/EMNLP-05*.

Turney, P. D. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. In *COLING-08*, Manchester, UK.

Turney, P. D. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21:315–346.

Yahoo Inc. (2008). http://www.yahoo.com.

[9] Recall is substantially higher for the baseline due to its nature of classifying all the words as intense.