IEICE TRANS. INF. & SYST., VOL.E92–D, NO.8 AUGUST 2009

PAPER

# AdjScales: Visualizing Differences between Adjectives for Language Learners

**Vera SHEINMAN**[†a)], ***Nonmember and*** **Takenobu TOKUNAGA**[†], ***Member***

**SUMMARY** 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 between 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 may be useful for language learners that try to distinguish between similar words. Additionally, learners might gain from a simple visualization of these differences using unidimensional scales. The method is evaluated by comparison with annotation on a subset of adjectives from WordNet by four native English speakers. It is also compared against two non-native speakers of English. The collected annotation is an interesting resource in its own right. This work is a first step toward automatic differentiation of meaning between similar words for language learners. AdjScales can be useful for lexical resource enhancement.

***key words:*** *language education, semantic relations, adjective scales, lexical semantics, natural language processing*

## 1. Introduction

In the process of vocabulary learning, language learners encounter situations where they need to choose an appropriate word to use from a set of near-synonymous words. The subtle differences between the words and the fact that the meaning of near-synonyms between the native language and the target language usually overlap only partially makes 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?

WordNet [1] is a widely-used lexicon that represents concepts by *synsets* of synonymous words and encodes the semantic relations between them. For instance, "small" and "big, large" are linked by the semantic relation of *antonymy* in WordNet.

A Linguistic Scale is a set of words of the same grammatical category, which can be linearly ordered by their semantic strength or degree of informativeness [2]. Not limited to a particular part-of-speech, an example of a linguistic scale for adverbs is ⟨*may, should, must*⟩, an example for adjectives (*adjective scale*) is ⟨*lukewarm, warm, hot*⟩ [3]. Existing linguistic resources and dictionaries rarely contain information on adjectives being part of a scale, or being of a particular strength.

This information may be deduced in some cases from the word definition. For instance, the definition for the word "tiny" in WordNet is "very small", and it may be deduced that "tiny" is *stronger-than* "small" in the sense of *smallness*. However, it is not always so, and lacks the convenience of a single visual scale like *infinitesimal*→*tiny*→*small*→*smallish*. DeCapua [4] recommends that teachers of English use scales to visualize the difference in certainty in English modals, such as *may be*→*might be*→*could be*→*must be*, we follow this recommendation for visualization of difference in strength between adjectives.

Gradation is a related term describing variation of strength between adjectives that describe the same property. Fellbaum [5] describes gradation as a semantic relation organizing lexical memory for adjectives and provides six examples of gradation for six properties. One of the examples is a gradation of adjectives for the property *size*, ⟨*astronomical, huge, large, standard, small, tiny, infinitesimal*⟩.

According to Fellbaum, gradation is rarely lexicalized in English. Adverbial expressions such as "very" and "slightly" or comparative expressions such as "more" and "less" are usually preferred. For this reason it is not encoded in WordNet. While acknowledging this situation, we believe that having a method for grading adjectives that are lexicalized is important and in particular beneficial for learners that struggle with similar adjectives. Moreover, with the Web available as a corpus, this information may be extracted with less effort than before.

Descriptive adjectives have antonyms, describe a certain property and tend to be gradable. WordNet encodes descriptive adjectives in clusters (adjective-sets). Two antonymous representative synsets (head-words) are connected to a noun they describe (attribute). Each one of the *head-word* adjectives is connected to similar adjectives. There is no encoding of the relations between the *similar* adjectives, and there is no encoding of the differences between the *similar* connections. In the example illustrated in Fig. 1, there is a clear difference between the adjective "smallish" that is slightly less small than "small", and the adjective "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

Manuscript received January 19, 2009.
Manuscript revised March 20, 2009.

[†]The authors are with the Graduate School of Information Science and Engineering in Tokyo Institute of Technology, Tokyo, 152–8552 Japan.
a) E-mail: vera46@gmail.com

DOI: 10.1587/transinf.E92.D.1542

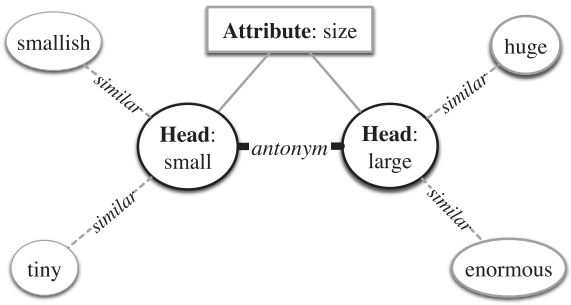


**Fig. 1** Example of descriptive adjectives encoding in WordNet.

of meaning that differentiates it from others. In addition to *strength*, there are also 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 learners trying to search for appropriate words among similar words. 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 in English, AdjScales that incorporates recent advances in Natural Language Processing. In the choice of the computational approach suitable for the task, we aimed for the simple and freely accessible approaches that do not require any specialized corpora, parsing or tagging.

The novelty of AdjScales is in automatic construction of adjective scales from several examples, in the language learner as the target user, and in its evaluation.

AdjScales reduces information load on learners by presenting useful differentiation information using simple unidimensional scales. Possible contributions of this work are in the fields of language learning tools, lexical resource enhancement and textbook authoring. Automatic acquisition of adjective scales is also beneficial in construction of ratings for questionnaires for interface design.

## 2. Proposed Method: AdjScales

### 2.1 Pattern Extraction

Pattern extraction is a preparatory step for AdjScales. At this step we extract patterns from the Web that serve AdjScales for detection of the *stronger-than* relation in the Scaling step described in Sect. 2.2.5.

Similarly to the approach proposed in [6], we use pattern-extraction-queries of the form "a ∗ b" to find patterns where $a$, $b$ are referred to as seed words, and "∗" denotes a wildcard†. We extract binary patterns of the form $p$ = [$\text{prefix}_p$ $x$ $\text{infix}_p$ $y$ $\text{postfix}_p$] from the snippets of the query results using a search engine, where $x$ and $y$ are slots for words or multiword expressions. A pattern $p$ can be instantiated by a pair of words $w_1$, $w_2$ to result in a phrase $p(w_1, w_2)$ = "$\text{prefix}_p$ $w_1$ $\text{infix}_p$ $w_2$ $\text{postfix}_p$" or it can be instantiated by a word $w_1$, and a wildcard to result in a phrase "$\text{prefix}_p$ $w_1$ $\text{infix}_p$ ∗ $\text{postfix}_p$" to search for words cooccurring with the word $w_1$ in a pattern.

Let's consider an example pattern $p_1$ where $\text{prefix}_{p_1} = \phi$, $\text{infix}_{p_1}$ = "if not", and $\text{postfix}_{p_1} = \phi$, if we instantiate it with the pair of words (good, great) we will get a phrase $p_1$(good, great) = "good if not great". Instantiating it with (∗, good) will result in a phrase $p_1$(∗, 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 when querying it with a pattern-extraction-query, we refer to it as $p$ is supported-by $(w_1, w_2)$.

Snippets are a good source for patterns, because they contain the direct context of the query text. For the extraction purposes snippets are split into sentences and are cleaned from all kinds of punctuation.

Davidov and Rappoport [6] introduce a generic approach to relation extraction using the Web. Differently from them we choose the seed word pairs in a supervised manner, so that $seed_2$ is *stronger-than* $seed_1$. For the experimental settings described in this work we used 10 seed word pairs selected from the adjective scale examples provided by [5]. For instance, one of the seed word pairs we have used was (cold, frigid), where "frigid" is *stronger-than* "cold". 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 pair is only on the left side of the pattern (before the infix words) or so that the weaker word is only 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] may be extracted while supported only by seed words describing the *size* property, such as (huge, astronomical), (big, huge), (tiny, infinitesimal).

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

Davidov and Rappoport [6] extract only patterns with function words or frequent words. For instance, the words "is", "but" and "not" in the pattern [is $x$ but not $y$] are function words, while "children" in [$x$ children $y$] is a content word. In our settings, patterns with content words normally

†Denotes 0 to several words that may appear in its place. In reality, search engines, usually use the notation of ∗ for a single-word, and we used several queries: "a b", "a ∗ b", "a ∗ ∗ b" for each pattern-extraction-query.

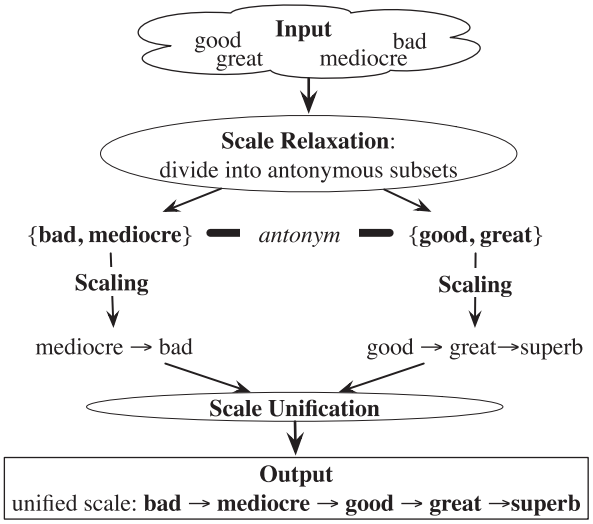


**Fig. 2** General illustration of AdjScales method.

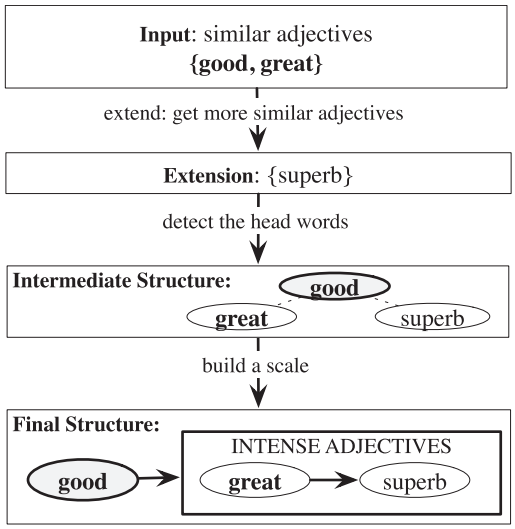


**Fig. 3** AdjScales core.

do not match our requirements for selection, and therefore, content words do not get selected as part of the patterns.

### 2.2 Method Steps

AdjScales method comprises several steps listed below with Scaling being its core. As it is shown in Fig. 2, we divide the input adjectives into two subsets in a Scale Relaxation 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 outputting 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 scaling is interesting for the user, only by the given examples. Similar adjectives for our purposes are adjectives that describe the same property.

In the example illustrated in Fig. 2, the input words are "bad", "good", "mediocre", and "great".

#### 2.2.2 Scale Relaxation

According to [3], in the case of adjective scales, the total scale is commonly relaxed, so that the 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 by using WordNet, since the information about antonymy and similarity is already encoded there. If the input words belong to the same adjective-set structure in WordNet they are divided by their similarity to the head 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, not all the words appear in WordNet, or they are not encoded in the same adjective-set structure. In this work, we assume that in such cases the words belong to a single subscale.

In the running example, the input words are divided into the antonymous subsets {bad, mediocre} and {good, great}. The following steps taken for a single subset of {good, great} are illustrated in Fig. 3.

#### 2.2.3 Extension

At the Extension step, AdjScales attempts to provide the user with other adjectives, similar to the input adjectives. This step is conducted using WordNet. 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, when some of the input adjectives are not in WordNet, or when the input adjectives do not appear as part of the same adjective-set, no extension is currently performed.

For example purposes (in Fig. 3) the adjective "superb" that is one of the words encoded as *similar* to "good" in WordNet is added as an extension to the subset {good, great}.

#### 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 in AdjScales and call it an Intermediate Structure. For cases where the input adjectives do not appear in WordNet, we select the most frequent adjectives as the head-words. Frequency information of that kind can be approximated by search engine page hit counts. 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 useful words in the context of the adjective scale in question.

In the running example, the adjective "good" is encoded as the head-word in WordNet, and the consequent Intermediate Structure includes head-words = {good} and similar-words = {great, superb}.

#### 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 (in Sect. 2.1) as $P$. For each pair (head-word, similar-word) from the Intermediate Structure, we instantiate (as described in Sect. 2.1) each pattern $p$ in $P$ to obtain phrases $s_1 = p$(head-word, similar-word) and $s_2 = p$(similar-word, head-word). We estimate document frequency, $df(s_i)$, by using the page hit counts 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 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 each one of the subsets (*mild* and *intense*), the values are reset, and 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 recorded in a Final Structure that can be visualized as a scale *mildest words* $\rightarrow \cdots \rightarrow$ *least mild words* $\rightarrow$ *head-words* $\rightarrow$ *least intense words* $\rightarrow \cdots \rightarrow$ *most intense words*.

To illustrate the process with the running example, let's assume that $P = \{p_1 = [x \text{ if not } y]\}$. The Intermediate Structure in the running example contains head-words = {good}, and similar-words = {great, superb}. We instantiate $s_1 = p_1$(good, great) = "good if not great", $s_2 = p_1$(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[†]. $p_1$ 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 as a result, both, "great" and "superb" are classified as *intense*. Then, the values of "great" and "superb" are reset to 0, and scaling is performed within the intense subset. $df(p_1(\text{great, superb}) > 3 \times df(p_1(\text{superb, great}))$ reducing a point to "great" and adding a point to "superb". There are no *mild* or *unconfirmed* words in this example, resulting in the final structure:


**Fig. 4** Example of unified scale.

{head-words = {good},
intense words = {great(−1) → superb(1)}},

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

#### 2.2.6 Scales Unification

Subscales may be unified into a single structure. Sometimes 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 scale of adjectives that describe the property *size*, we reverse the direction of links in one of the subscales, resulting in the unified scale *bad* → *mediocre* → *good* → *great* → *superb*. The example unified scale with the actual links between the adjectives appears in Fig. 4 including the links between the adjectives.

Markedness refers to relationships between two complementary or antonymous terms which can be distinguished by the presence or an absence of a property. In particular, markedness is applicable to antonymous gradable adjectives, such as "tall" (unmarked, presence of the property *height*) and "short" (marked, absence of *height*) [7]. Normally, adjective scales are ordered from the *marked* side to the *unmarked*, (marked head-word→unmarked head-word). The authors in [7] suggest that unmarked item is usually more frequent, and that frequency alone can be quite accurate test to make that distinction. Following their conclusion we unify the subscales, so that the subscale with the less frequent head-word (presumably marked) is on the left.

## 3. Evaluation

For evaluation, we preselected 16 patterns (11 *intense* and 5 *mild* patterns) in the manner described in Sect. 2.1[††]. The extracted patterns are listed in Table 1. The conducted experiments evaluating the scaling step are described in the sections below.

### 3.1 WordNet-Based Corpus

We extracted 298 descriptive adjective-sets from WordNet as the input to our system for evaluation of the scaling step. They comprise 757 head-words (645 distinct words) and 6,607 similar-words (5,378 distinct words). Each set was

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

[††]It is important to note that due to the limit search engines' APIs impose on the amount of accessible snippets, some differences in the extracted patterns may occur depending on the query date. We used yahoo search API [8] for all the experiments.

**Table 1** Patterns extracted for the evaluation (*y* is *stronger-than x*).

| Intense Patterns | Mild Patterns |
|---|---|
| *x* even *y* | *y* very *x* |
| *x* if not *y* | not *y* but *x* enough |
| *x* almost *y* | *y* unbelievably *x* |
| *x* no *y* | *y* not even *x* |
| *x* perhaps *y* | *y* but still very *x* |
| extremely *x y* | |
| is *x* but not *y* | |
| are *x* but not *y* | |
| are very *x y* | |
| is very *x y* | |
| *x* sometimes *y* | |

divided into two antonymous subsets. Four native English speakers (2 Americans and 2 British[†]), all male students from engineering departments annotated the input in terms of scaling for comparison.

The subset of an adjective-set represented by the synset {lean, thin} for attribute *body weight* comprises 51 similar words. In cases of such big subsets, words are too difficult to place on a scale for humans. We downloaded snippets for queries of the type $p$(head-word, $*$) and $p(*$, head-word) for each pattern $p$ from the preselected patterns and for all the head-words resulting in 625 MB of data. As the next step, we extracted the list of words that appeared in the extracted phrases in the slot of the wildcard. If a word in an adjective-set was not included in that list it was pruned. The reasoning behind the decision to prune the words this way is as following. In the current stage of our work we experiment with preselected patterns. If a certain word does not appear in any patterns, our method cannot provide a decision on its scaling. So, to test this approach only the words that are potentially applicable may be considered. Also, it is likely that such words are not applicable or are rare in the first place. Finally, all the subsets that comprised head-words only and no similar-words were discarded. The final dataset for evaluation contained 308 subsets with 763 similar-words to be scaled in total.

Each annotator performed the task independently from others. For each subset from the 308 the annotator was presented with the head-words, attribute, the antonymous head-words and a set of similar-words. The head-words were fixed as *neutral* and we asked the annotators to classify each one of the similar-words into one of 5 types (*neutral*, *mild*, *very mild*, *intense*, and *very intense*) while keeping scaling by *strength* in mind. When not sure about a certain word or thinking that it is not applicable for scaling in the given context the annotator was requested to classify it as additional *not sure* or *not applicable* types respectively. When a certain word seemed *stronger* than the head-words it was to be classified as *intense* or *very intense*. When it seemed *weaker* than the head-word, we asked the annotator to classify it as one of the *mild* or *very mild* types. Words of similar intensity to the head-words were to be classified as *neutral*.

We measure the agreement between two annotators, and between AdjScales and an annotator in the following manner. First, general agreement is measured as shown in Table 2. If a word $w$ in subset $s$ is selected as *mild* or as *very mild* by annotator $A$, we will denote it as $w \in gen\text{-}mild_A$, the notation is straightforward rewriting for *intense*. Two annotators $A$ and $B$ agree if

$$w \in gen\text{-}mild_A \wedge w \in gen\text{-}mild_B$$

or if

$$w \in gen\text{-}intense_A \wedge w \in gen\text{-}intense_B.$$

In the case of our task there were many words that ended up undetermined (*not sure*, *not applicable*, or *unconfirmed* for AdjScales), so it was important to also measure the *general disagreement* explicitly. For each two annotators $A$ and $B$ we measure precision of $A$ compared to $B$ defined as

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

Similarly, we define

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

and

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

for general agreement for words selected as *generally mild*. We follow the same notation for *generally intense*.

A simple baseline is to assign the most frequent classification choice to each word. In the WordNet-based corpus, most frequently words were classified by annotators as *intense*, and therefore, our baseline method classifies any given adjective as *intense*.

We averaged the pairwise agreement between the annotators. Similarly, we averaged the pairwise agreement of AdjScales with each one of the annotators, and the pairwise agreement of the baseline with each one of the annotators. From experiments with our training data, we selected the parameters of AdjScales to be 15 for *weight*, and 20 pages for *threshold*. In order to reduce the search engine queries required for computation of each scale, we grouped the queries of patterns into 4 *united queries* unifying each subgroup of $m$ pattern instances by the operator *OR*:

"$p_1(w_1, w_2)$" OR $\cdots$ OR "$p_m(w_1, w_2)$".

The comparison between the annotators, AdjScales, and

**Table 2** General agreement for WordNet adjectives.

| | #words mild | #words intense |
|---|---|---|
| annotator$_1$ | 137 | 358 |
| annotator$_2$ | 99 | 301 |
| annotator$_3$ | 89 | 290 |
| annotator$_4$ | 141 | 313 |
| All annotators | 22 | 163 |

[†]We have observed no particular differences between the British and the Americans in their annotations.

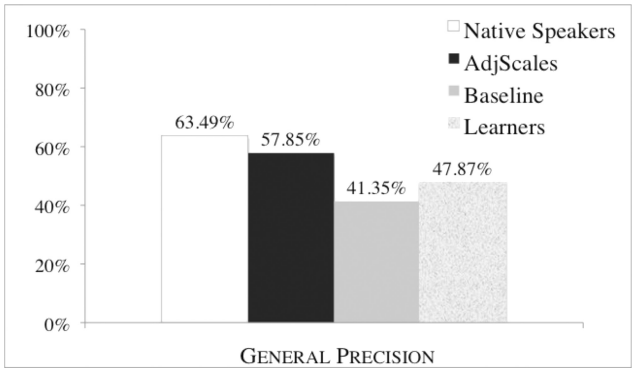


**Fig. 5** Comparison of general precision between native speakers, AdjScales, baseline and non-native speakers (learners) pairwise.

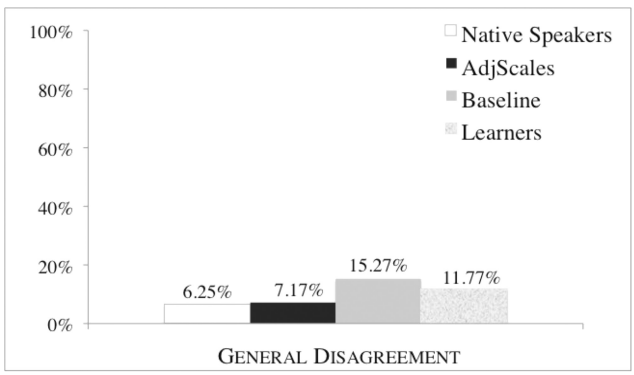


**Fig. 6** Comparison of general disagreement between native speakers, AdjScales, baseline and non-native speakers (learners) pairwise.

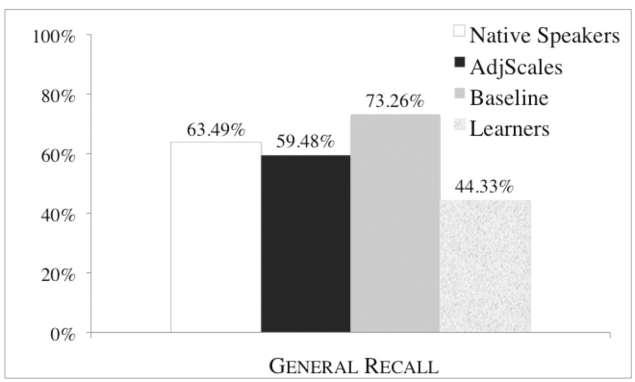


**Fig. 7** Comparison of general recall between native speakers, AdjScales, baseline and non-native speakers (learners) pairwise.

baseline is presented in the charts comparing precision, disagreement and recall in Fig. 5, Fig. 6, and Fig. 7. An additional column represented in the charts labeled "Learners" refers to the performance of non-native English speakers in the experiment as described in Sect. 3.2. It is important to note that while for precision, the higher the values are the better, for disagreement the general objective is to reduce it to 0. Although the baseline is useful for precision and disagreement evaluation, for recall, it is not as informative, as it produces 100% for the intense adjectives by definition, and being the most frequent choice by the annotators, it achieves a very high recall of 73.26% in total. Though, the results presented in the charts represent pairwise averages, when compared against the annotator that agreed the least with other annotators, the precision is very similar, being 57.65% for the annotator and 57.85% for AdjScales.

**Table 3** Order agreement (native-speaker annotators vs. AdjScales vs. non-native speakers pairwise).

| | **native speakers** | **AdjScales** | **non-native speakers** |
|---|---|---|---|
| **mild** | 86.11% | 86.11% | 60.36% |
| **intense** | 88.74% | 70.20% | 90.07% |

We also compared AdjScales to the answers that were generally agreed upon by all 4 annotators. AdjScales disagreed with the four annotators consensus only for one word, "bright" that AdjScales classified as *mild* compared to the head-word "light" and the native annotators classified as *intense*.

Additionally, to understand the finer agreement on ordering adjectives as mild or even milder on a scale, we measure order agreement, similarly to the measurement reported in [9]. Annotators $A$ and $B$ agree on order of a pair $w_1, w_2$ if $A$ and $B$ both classified $w_1$ and $w_2$ as *milder* than the head-word or as *neutral* and if $A$ classified $w_1$ as strictly milder than $w_2$ and so did $B$. 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 respectively. The annotators tend to agree on the order they choose between words within similar intensity as it is shown in Table 3.

### 3.2 Comparison with Non-native English Speakers

In an experiment toward evaluating the usefulness of AdjScales for language learners, we asked 2 Japanese students of advanced[†] level of English from engineering departments to perform the same task that was described in Sect. 3.1. We checked general agreement between each one of the non-native speakers to each one of the native English speaker annotators pairwise, along the same lines of evaluation. The charts in Fig. 5, Fig. 6, and Fig. 7 compare their average performance against the agreement of native English speakers, AdjScales and the baseline.

The non-native speakers selected many more words as *not applicable* or *not sure* (average of 245 words) compared to native speakers (average of 111 words). Compared to the results that were agreed upon among all four native speaker annotators, the non-native annotators misselected 2.5% of adjectives. For instance, the adjective "spotless" that was selected by all the native annotators and by AdjScales as more *intense* than "clean", was selected as *milder* by one of the non-native annotators. In another example, AdjScales and all native annotators agreed on the subscale: *comparable→same*, while a non-native annotator selected "comparable" being more intense than "same".

[†]The grades of the students were within the 2 highest categories on TOEIC test.

**Table 4** Additional adjective scales.

| AdjScales | precision | disagreement | recall |
|---|---|---|---|
| **mild** | 100% | 0% | 26.67% |
| **intense** | 90.48% | 9.52% | 79.17% |
| **total** | 92% | 8% | 58.97% |

The superiority of the results of AdjScales compared to non-native annotators and the baseline, may suggest that it can be useful for learners of English to study about the differences in strength between adjectives.

### 3.3 Additional Adjective Scales

The adjectives in WordNet are not necessarily clustered having a single scale in mind. In order to test AdjScales against confirmed adjective scales, we gathered a special dataset of 22 adjective scales in the following manner.

In independent experimental settings we requested 2 native English speakers and 3 non-native English speakers to produce as many linguistic scales as possible [10]. After the production step the subjects cross-verified the results, and only the scales agreed upon by all of them remained in the dataset. 9 of the scales in the dataset were adjective scales.

We added the adjective scale example from [3], ⟨*lukewarm, warm, hot*⟩, and the example ⟨*good, great, fantastic*⟩ from [11].

Additionally, we extended our dataset with 4 adjective scales from a teaching resource [12] that teaches language learners the shades of meaning. In the exercises in the suggested activity, several verb and adjective scales are provided, where the students are requested to order them by strength.

We relaxed each of the scales in our dataset manually into 2 antonymous subsets, when there were two antonymous components and performed scaling. We compared the scaling results by AdjScales with the same parameters as in Sect. 3.1 to the expected scales as shown in Table 4.

In a surprising application, the authors in [13] report on choosing the adjective ladder ⟨*abysmal, awful, bad, poor, mediocre, fair, good, great, excellent, amazing, phenomenal*⟩ as the most widely agreed upon among native and non-native English speakers in their survey for rating attributes or skills for role-playing games. Providing the items of the scale to AdjScales resulted in the following output, while the words "poor" and "fair" remained unconfirmed. For the positive subscale, *good*→*great*→{*phenomenal, excellent*}→*amazing*, and for the negative subscale *mediocre*→*bad*→{*awful, abysmal*}. AdjScales may be used to evaluate validity of such ladders.

## 4. Related Work

Inkpen and Hirst [14], following the work by [15] introduce a method to acquire information about the differences among sets of near-synonyms, such as "error", "mistake", "slip", and "blunt" using a machine-readable dictionary for near-synonyms. Using automatic methods to differentiate between near-synonyms to enhance existing lexical resources and in the context of language learners is the objective of our research, and in this sense both works are relevant to ours. Adding distinctions on the subtleties to existing language resources is needed. We focus only on the adjective-scaling that may be viewed as differentiation of near-synonymous adjectives by strength, adding further types of differentiation (such as formality level of words) and distinguishing which types of near-synonyms can be scaled is a needed extension of our work in the future. We differ in our corpus being the Web, and in our focus on scaling groups of near-synonyms.

A study much relevant to ours [3] establishes the first step toward automatic identification of adjective scales. It provides an excellent background on adjectives and a general plan to identify adjective scales, though, it concentrates only on clustering of adjectives that describe the same property.

Using patterns extracted from big corpora like the Web in order to learn semantic relations between words is a common approach in computational linguistics pioneered by Hearst [16] and further extended, generalized and improved by others [6], [17]–[19]. AdjScales belongs to this school, as adjective scales comprise a fine-grained asymmetric relation of *stronger-than* between adjectives that describe the same property.

VerbOcean [18], explores fine-grained relations, *stronger-than* being one of them. Their work is very similar to ours in relating associated verbs to one another rather than organizing them in semantic classes and 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, in our evaluation procedure (it is production based, and with a scale in mind), and in our full reliance on free-text for the identification of *stronger-than* relation (they use a smaller part-of-speech tagged corpus). VerbOcean is an important step toward providing differentiation between similar verbs and it should be considered in the context of language learners.

A large body of research [9], [20], [21] has been conducted in the fields of *opinion mining* or *sentiment analysis*. An important distinction for the work in opinion mining is *semantic orientation* of words and utterances. It is essential to determine whether they are *positive* or *negative*. In this work we do not distinguish between the positive or the negative sense of adjectives, but rather make a general distinction of the extent of adjectival descriptive strength. Also, the objective of this work is different. We aim to provide linguistic distinction between similar adjectives for learners, while the research in opinion mining concentrates on strength of subjectivity and sentiment of words, phrases and texts.

Typically, adjectives and relations between them play a central role in understanding opinion from texts. In this

aspect this field is related to ours. According to [20] semantic orientation of a word, in addition to its direction also comprises intensity, *mild* or *strong*. They compute intensity using statistical association with a set of positive and negative paradigm words, but concentrate on detection of semantic orientation of words. Ranking by strength is evaluated only marginally. OPINE [11], [21], a system for product reviews mining related to the work by [20] ranks opinion words by their strength as one of its subtasks. OPINE uses 8 patterns, bootstrapped from the pattern [$x$, $*$ even $y$] in a Web-based manner to rank descriptions of a feature. Opinion phrases with intensifiers, such as "very" or "somewhat" are ranked by the strength of their intensifiers. They report on 73% precision, and the evaluation reported in their work is verification based. Every pair where *strength*(*opinion*′) > *strengh*(*opinion*) is determined by OPINE is verified by a human judge. Differences between lexicalized and non-lexicalized descriptions are not reported. We differ in our focus on lexicalized adjectives. Our evaluation is much more extensive (4 judges, rather than one), and it is production-based (human annotators provide scales and do not verify the results of our system). Verification-based evaluation is more prone for bias.

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

## 5. Conclusion and Future Work

We have presented AdjScales, a method to construct adjective scales from several examples of similar adjectives. Providing the differences between similar adjectives in a form of a compact unidimensional scale reduces the information load on learners. The system is created using a state-of-the-art methodology of extracting relations using patterns from the Web. It is quite 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 humans (−5.5% for precision and +1% for disagreement) and much better than the baseline (+16% for precision and −7% for disagreement). AdjScales performs similarly to the human that performs in the least pairwise agreement with others. AdjScales 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 59% for the additional scales), in particular the items on the milder side of the head-words (recall of only 27% for the additional examples). There is only one disagreement of the system with answers that all native speakers agree upon, suggesting that in cases where scales are clear and thus suitable for learning, AdjScales will be more accurate.

A surprising observation from our experiments is the asymmetry between the adjectives on the mild side and the intense side of the head-words in WordNet. Annotators and AdjScales consistently selected fewer words as *mild*, and also agreed less well within the *mild* selection. They agreed on 163 adjectives being *intense* and only on 22 adjectives being *mild*. There may be several reasons for this asymmetry. 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 (the antonymous words are used for that). We have also observed from analysis of the results by each one of the patterns separately, that some patterns perform better for *mild* words while others do better in identification of *intense* words. This direction will be further investigated in the future.

The comparison with non-native English speakers showed less agreement with the native English speakers than did AdjScales, but they performed better than the baseline. This trend is consistent for precision (−10% from AdjScales, +6% from the baseline), disagreement (+4% from AdjScales, −4% from the baseline), and recall (−15% from AdjScales). The recall levels are substantially lower as almost twice as more words were selected as *not sure* or *not applicable* compared to native speakers. These results may suggest that although AdjScales performs slightly less well than native English speakers, learners can still potentially learn from it, as it performs better than them. AdjScales being a pattern-based system, it may also suggest generally that students may learn from querying search engines for patterns similar to the patterns extracted by AdjScales to enhance their knowledge about the *stronger-than* relation between adjectives.

In some cases similar-words are perceived as equally strong to the head-word, as can be seen from the classification *neutral* by native-speaker annotators. This kind of distinction is not available in AdjScales, which classifies similar-words into *intense*, *mild*, or *unconfirmed*. Incorporation of a measure of *equally strong* into AdjScales and its evaluation against the existing WordNet-based corpus will be a valuable addition to AdjScales 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 annotators 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 cannot necessarily be scaled along a single dimension. We plan to invest further effort to detect adjectives that belong to the same scale as a pre-scaling step.

Granularity of scaling is another issue raised by annotators. 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

[1] G.A. Miller, "Wordnet: A lexical database for English," ACM, vol.38, no.11, pp.39–41, 1995.
[2] S.C. Levinson, "Conversational implicature," Pragmatics, 2000 ed., pp.132–134, Cambridge University Press, 1983.

[3] V. Hatzivassiloglou and K.R. McKeown, "Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning," 31st annual meeting on Association for Computational Linguistics, ACL-93, pp.172–182, Association for Computational Linguistics, 1993.
[4] A. DeCapua, "Grammar for teachers: A guide to American English for native and Non-native speakers," ch. Modal Auxiliary Verbs and Related Structures, p.215, Springer, 2008.
[5] C. Fellbaum, D. Gross, and K. Miller, "Adjectives in wordnet," in Five papers on WordNet, 1993.
[6] D. Davidov and A. Rappoport, "Unsupervised discovery of generic relationships using pattern clusters and its evaluation by automatically generated SAT analogy questions," ACL-08: HLT, Columbus, Ohio, pp.692–700, Association for Computational Linguistics, June 2008.
[7] V. Hatzivassiloglou and K. McKeown, "A quantitative evaluation of linguistic tests for the automatic prediction of semantic markedness," 33rd annual meeting on Association for Computational Linguistics (ACL-95), pp.197–204, Association for Computational Linguistics, 1995.
[8] Y. Inc, "Yahoo," http://www.yahoo.com, 2008.
[9] T. Wilson, J. Wiebe, and R. Hwa, "Just how mad are you? Finding strong and weak opinion clauses," AAAI, pp.761–779, 2004.
[10] V. Sheinman, N. Rubens, and T. Tokunaga, "Commonly perceived order within a category," OntoLex Workshop at 6th International Semantic Web Conference (ISWC 07), Busan, Korea, Nov. 2007.
[11] A.M. Popescu, Information Extraction from Unstructured Web Text, Ph.D. thesis, University of Washington, 2007.
[12] D. Cadman, "Shades of meaning," www.primaryresources.co.uk/english/pdfs/8shades.pdf, 2008.
[13] F. Hicks, L. Valentine, J. Morrow, and I. McDonald, http://www.ianm.eclipse.co.uk/storytelling/lichert_article.htm, 2006.
[14] D. Inkpen and G. Hirst, "Building and using a lexical knowledge base of near-synonym differences," Computational Linguistics, vol.32, no.2, pp.223–262, 2006.
[15] P. Edmonds, Semantic Representation of Near-Synonyms for Automatic Lexical Choice, Ph.D. thesis, University of Toronto, 1999.
[16] M. Hearst, "Automatic acquisition of hyponyms from large text corpora," 14th Conference on Computational Linguistics (COLING-92), pp.539–545, 1992.
[17] E. Riloff and R. Jones, "Learning dictionaries for information extraction by multi-level bootstrapping," 16th National Conference on Artificial Intelligence (AAAI-99), 1999.
[18] T. Chklovski and P. Pantel, "Verbocean: Mining the web for fine-grained semantic verb relations," Conference on Empirical Methods in Natural Language Processing (EMNLP-04), pp.33–40, Barcelona, Spain, 2004.
[19] P.D. Turney, "A uniform approach to analogies, synonyms, antonyms, and associations," 22nd International Conference on Computational Linguistics (COLING-08), Manchester, UK, Aug. 2008.
[20] P.D. Turney and M.L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," ACM Transactions on Information Systems, vol.21, pp.315–346, 2003.
[21] A.M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," The conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT/EMNLP-05, 2005.

**Vera Sheinman** Ph.D. candidate in Tokyo Institute of Technology, Department of Computer Science, Tokyo, 152–8552 Japan.

**Takenobu Tokunaga** Professor of Department of Computer Science. He received the B.S., M.S. and Dr. Eng. degrees from Tokyo Institute of Technology in 1983, 1985 and 1991, respectively. His current interests are natural language processing, in particular, building and managing language resources, application of language technologies to intelligent information access, and dialogue systems. Tokyo Institute of Technology, Department of Computer Science, Tokyo, 152–8552 Japan.