Reducing False Positives in the Construction of Adjective Scales

Alice Zhang
Princeton University
Princeton, New Jersey, USA.
alicez@princeton.edu

Abstract

Many adjectives that appear to be synonyms of one another differ in their intensity. Distinguishing the nuances between adjective synonyms is vital to linguistic understanding of a language, but WordNet currently does not encode the relative intensities of adjective synonyms that lie on the scale. Sheinman & Tokunaga (2009) proposed a solution of constructing Adjective Scales by data mining a web corpus. However, this process suffers from some limitations, most notably that of False Positives, which inaccurately suggest that adjective **X** is more or less intense than **Y**.

This paper classifies the types of false positives that Sheinman's method generates, then proposes a method to diminish the quantity of these false positives using linguistic searches in WordNet.

1 Introduction

Adjectives are currently represented in WordNet in a dumbbell structure, such that antonymous adjective pairs like "wet-dry" and "early-late" are connected with a single antonym link. Each word of the antonym pair is represented as one of two centroids on the dumbbells, and each of their synonyms are spread out radially around the centroid. This representation is problematic because 1) it suggests that all adjectives within a synset are equally similar to the centroid and 2) because many similar adjectives are misclassified as members of the same clusters, indicating that they describe the same types of objects, when in reality they are very different.

In their paper *Large, huge or gigantic?: Identifying and encoding intensity relations in WordNet*, Sheinman et al. (2013) proposed a method to uncover the differing intensity relationships amongst a set of adjective synonyms by mining a web corpus. In particular, Sheinman noticed particular patterns that occurred naturally in English speech that already codified the intensity relationships between the adjectives that were used within the pattern. For example, one of these semantic patterns is "X **but not** Y," where **Y** is implied to be more intense than **X**, e.g. "good but not great" implies that "great" is more intense than its synonym "good" based merely on the pattern "**X** but not **Y**" In fact, these patterns occur in both directions, such that while some patterns imply that **X** is more intense than **Y**, while others imply that **X** is less intense than **Y**. By discovering pairs of adjectives that occurred in the natural patterns, Sheinman was able to construct scales of adjective synonyms, where each adjective was listed according to its relative intensity.

While Sheinman's method seems, in large part, successful in constructing adjective scales, it also suffers from limitations of false positives, which appear when certain adjective pairs show up in the linguistic patterns, but do not actually indicate that adjective **Y** is more or less intense than **X**. For instance, the natural phrase "good but not good enough" would seem to suggest that "good" is more intense than "good" based merely on the battern "**X** but not **Y**," even though this is not true. These false positives are significant: a simple Google News search of the phrase "hot but not" will return false positives for over half of the results. This paper classifies the different types of false positives that can be generated and proposes an algorithm that utilizes WordNet to be able to detect these false positives.

2 Type A False Positives

2.1 Classification

Type A false positives are phrases where adjective **Y** is classified as being more intense than adjective

Intense Patterns
(is / are) X but not Y
(is / are) very X Y
extremely X Y
not X (hardy / barely / let alone) Y
X (but / yet / though) never Y
X (but / yet / though) hardly Y
X (even / perhaps) Y
X (perhaps / and) even Y
X (almost / no / if not / sometimes) Y

Mild Patterns
if not X at least Y
but Y but X enough
not Y (just / merely / only) X
not Y not even X
not Y but still very X
though not Y (at least X)
Y (very / unbelievably) X

Table 1 Examples of the linguistic patterns that Sheinman et al. noticed in natural language. X and Y represent adjectives such that X is more intense than Y.

X, even though both **X** and **Y** fall on the same adjective scale, but **Y** is not more intense than **X**. In particular, Type A False Positives can be further classified into three particular types: repetitions, antonyms, and reversals.

Repetitions occur when both adjectives **X** and **Y** are the same word. For example, one naturally occurring English phrase that follows the "X **but not** Y" pattern that Sheinman noted is the phrase "It was good, but not good enough." Another example would be the phrase "It was good, but not as good as it could have been." In both instances, the two adjectives that are being compared cannot have one be more intense than the other because they are the same.

Antonyms occur when **X** and **Y** are direct antonyms of one another - both **X** and **Y** fall on the same scale, but they cannot be synonyms of one another because they lie on opposite ends of the same scale. For example, consider the following sentence: "He is not tall, but not short either." Sheinman's method would falsely classify "short" as a more intense synonym to the word "tall," which is a misclassification.

Finally, **reversals** occur where **X** and **Y** are real adjective synonyms of one another, but **X** is a more intense adjective than **Y**. For example, the sentences "This artifact is ancient, perhaps even old enough to have existed before dinosaurs," "The water was scorching, but not hot enough to kill the bacteria," and "President Taft was extremely obese, fat to the point of getting stuck in his own bathtub" are all instances that would seem, based on Sheinman's method, to suggest that **Y** is more intense than **X**, when in reality, **X** is more intense than **Y**.

2.2 Correcting Type A False Positives

To identify Type A false positives, one only needs an algorithm that can detect instances of repetitions, antonyms, and reversals.

Checking for repetitions is a trivial task: one simply needs to determine if **X** and **Y** are the same word.

To detect antonyms, one can take advantage of the pointers that are built into WordNet to check if any of the direct or indirect antonyms of **X** is equal to **Y**, or alternatively, that any of the direct or indirect antonyms of **Y** is equal to **X**.

Finally, we can fix reversals by taking advantage of a web database. Let the phrase p_1 be the original phrase, and let p_2 be the original phrase with X and Y swapped. After conducting queries on a search engine for both p_1 and p_2, we can determine that the query for which more results appear is the correct intensity ordering of the two adjectives.

3 Type B False Positives

3.1 Classification

The second type of false positives is **Type B** false positives, which are phrases wherein **Y** is inaccurately classified as being a more intense synonym of **X** because **X** and **Y** are adjectives that do not fall on the same scale.

For example, consider the sentence "Stevie Wonder is very good, but not lyrical." Using Sheinman's method of pattern extraction, one would falsely infer that "lyrical" is a more intense synonym to "good," which cannot be true, as "lyrical" is not even a synonym for "good", much less a more intense form of it.

Furthermore, Type B false positives occur frequently in human speech, as it is very common to switch scales when using a particular pattern.

3.2 Correction with Level 1 Searches

The most straightforward way of fixing Type B false positives is to performa simple search, testing to see if **X** falls under the synset - a word's set of synonyms - of **Y**, or if **Y** falls under the synset of **x**. This term can be classified as a Level 1 search.

Level 1 searches are searches conducted in WordNet, wherein only the two synsets of words **X** and **Y** will be explored. They differ from Level 2 searches, which increase the depth of the search. In general, a Level N search searches through a set of words *w*, then a Level N+1 search will search through all synsets for each word contained in *w*. Thus a Level 2 search will search through all the synsets of words contained in synsets of words **X** and **Y**.

The Level 1 searches are successfully able to eliminate a large number of Type B false positives. For instance, TYPEB-LEVEL1 can correctly identify "good but not lyrical" and "tasty but not expensive" as false positives.

These types of false positives are interesting because they reveal innate patterns of cultural thinking. People sometimes associate a given quality or attribute with another, such as price and quality. A phrase such as "wealthy but not arrogant" might seem to suggest that human thinking associates the wealthy as having arrogant qualities, or a phrase such as "fat but not jolly" might seem to suggest that a culture views associates fat people with being jolly. Future work might be to further investigate Type B false positives to extract cultural associations from the linguistic patterns.

The problem is that Level 1 searches overgenerate the number of false positives. The following table lists a collection of instances where the Level 1 searches classify the phrases as a false positive, even though intuition as an English speaker tells us otherwise.

X	Y	LEVEL1(X, Y)
good	wonderful	true
good	awesome	true
good	amazing	true
wonderful	awesome	true
elephantine	monstrous	true
gnomish	pocket-size	true

Table 2 Misclassified examples from a Level 1 search.

As evidenced, these examples suggest that Level-1 searches overgenerate the actual number of false positives. Further investigation allows us to see why: if we take all of the words included in the synset of $good$ and all the words included in the synset of $wonderful$, we can observe that neither word appears in the other's synset.

However, we can observe that triangulation appears in the synsets: $great$ appears as one of the words contained in the synset of $good$, and the words $great$ and $wonderful$ both have the word $extraordinary$ contained in both their synsets. Something that is $good$ must also be $great$, which is also $extraordinary$. Since something $wonderful$ is also $extraordinary$, it follows that $good$ and $wonderful$ are, indeed, true synonyms of one another.

3.3 Correction with Level 2 Searches

We have observed that two synonymous words that differ in intensity may not be included in each other's synsets, but may nonetheless share a common word between the two synsets. The word $wonderful$ does not appear in the synset of $good$ and $good$ does not appear in the synset of $wonderful$, but both $good$ and $wonderful$ share $great$ in their synsets. This leads us to believe that many of the falsely identified false positives could be eliminated by performing a Level 2 search instead of a Level 1 search.

A Level 2 search performs its searches one level deeper. A Level 2 search chooses one of the pair (X, Y) as its base, and then calculates the synset of the other word. For each word in the synset, the Level 2 search performs a Level 1 search against the base word that it chose earlier. Then, it switches the base word and performs the same set of Level 1 searches on the opposite synset. For each Level 1 search, if the the algorithm has found a word common to both X and Y's synsets, the checker identifies the pair as a false positive. Otherwise, if every synset pair has been searched and no word has been found common to both synsets, the algorithm identifies the pair as a false positive. The pseudocode for a Level 2 search is given in Algorithm 1.

3.4 Results

Performing Level-2 searches on Type B false positives eliminates overgeneration of false positives, but also yields the problem of undergeneration because of word sense disambiguation. Each of the

Algorithm 1 This function returns $true$ if phrase X and Y are identified as being a Type B false positive, and returns $false$ otherwise.

procedure TYPEB-LEVEL2(X, Y)
  $synset_x \leftarrow$ GETADJECTIVESYNSET(X)
  for all i in $synset_x$ **do** ▷ Search for Y in the synsets of X
    if TYPEA(i, Y) is $false$ and TYPEB-LEVEL1(i, Y) is $false$ **then**
      return false
  $synset_y \leftarrow$ GETADJECTIVESYNSET(Y)
  for all i in $synset_y$ **do** ▷ Search for X in the synsets of Y
    if TYPEA(i, X) is $false$ and TYPEB-LEVEL1(i, X) is $false$ **then**
      return false
  return true

synsets contain so many different senses that a Level-2 search could easily identify two words as synonyms based off of a faulty "common word." Sample adjective queries are shown in the table below, along with the adjective pair that was found to be a successful Level-1 pair and the word that the two adjectives held in common.

(X, Y)	Adj. Pair	Common Adj.
tall, thin	tall, thin	gangling
fat, smart	fat, intense	thick
short, rich	rich, dumpy	fat
happy, tasty	tasty, prosperous	rich
fat, red	red, rich	colorful
tall, awful	tall, tremendous	large
up, wide	up, broad	high
big, pretty	big, pretty	bad
strong, fat	strong, fat	fertile
good, big	good, large	ample
fat, atomic	fat, little	dumpy
sad, fat	sad, heavy	distressing

Table 3 Misclassified examples from a Level 2 search.

Our goal now is to reconcile the undergeneration of Level 1 searches with the overgeneration of the Level 2 searches. We do not consider searches deeper than a Level 2 search, because a Level 2 search already overgenerates.

4 Attributes

WordNet pointers contain information about a word's attribute, which stores the word's category, e.g. "size" for the adjectives "big" and "small." Adding checks that discard words of different attributes successfuly eliminates all the searches stored in Table 3.

The pseudocode for the altered algorithm, which includes attribute checks, is included as Algorithm 3. Running this altered algorithm corrects all of the results found in Table 3.

Algorithm 2 Returns $true$ if X and Y are Type B false positives, and $false$ otherwise.

procedure TYPEB-LEVEL2-ATTR(X, Y)
  $A_x \leftarrow$ GETATTRIBUTE(X)
  $A_y \leftarrow$ GETATTRIBUTE(Y)
  if A_x is not $null$ and A_y is not $null$ and A_x is not equal to A_y **then**
    return true
  $synset_x \leftarrow$ GETADJECTIVESYNSET(X)
  for all i in $synset_x$ **do**
    $A_i \leftarrow$ GETATTRIBUTE(i)
    if A_y is not $null$ and A_i is not $null$ and A_y is not A_i **then**
      continue
    if TYPEA(i, Y) is $false$ and TYPEB-LEVEL1(i, Y) is $false$ **then**
      return false
  $synset_y \leftarrow$ GETADJECTIVESYNSET(Y)
  for all i in $synset_y$ **do**
    $A_i \leftarrow$ GETATTRIBUTE(i)
    if A_x is not $null$ and A_i is not $null$ and A_x is not A_i **then**
      continue
    if TYPEA(i, X) is $false$ and TYPEB-LEVEL1(i, X) is $false$ **then**
      return false
  return true

4.1 Limitations

The most notable limitation is that the set of adjectives that have attributes is extremely small, and are thus susceptible to all of the pitfalls of the Level-2 searches described in Algorithm 3. In fact, most of the adjectives contained in WordNet do not have attribute pointers. Our algorithm could be substantially improved by encoding the attribute pointer more consistently in WordNet.

There are also a few exceptional cases, where

two adjectives are actually synonyms, but WordNet gives the two words different pointers. For example, the word "good" has an attribute of "quality" whereas the word "extraordinary" has an attribute of "ordinariness." Speakers of the English language can recognize that "good" and "extraordinary" are synonyms, but the algorithm would immediately reject them because they have different attributes.

4.2 Results

To test the algorithm, we ran four adjective pairs on it, selecting the phrases by the following criteria: 1) Returning a high enough number of hits on Google News so that the results can be considered significant, 2) Returning a low enough number of hits on Google News so that it is not overstrenuous to hand-classify each of the results, and 3) Adjectives that could be represented on a scale.

The searches were run by typing the pattern into a Google News search query in quotes (e.g. "hot but not"). Then, each search was classified by running it into the False Positive Checker described in Algorithm 1, and the accuracy of the classifications were checked by hand.

Overall, the False Positive Checker returns robust results for most adjectives. The vast majority of the errors occured because their attribute pointers returned $null$, leaving them susceptible to the Type 2 errors.

Altogether, for the two example searches listed above, the algorithm had 18 misclassifications out of 823 search results, for a total accuracy of 97.81%. All 823 instances described in Table 4 are instances of positives generated by Sheinman's method, but classifying these as true positives or false positives is left up to our algorithm. The high degree of accuracy from from the searches suggests that this algorithm is successfully able to classify Sheinman's phrases a true positives or false positives. If one could encode adjective attributes more consistently in WordNet, most of these errors would be able to be eliminated.

4.3 Limitations of WordNet

All of our searches rely on the ability of WordNet to classify adjectives correctly. However, many of our searches using the False Positive Checker indicate that there are gaps in WordNet's structure. More specifically, limitations on attribute pointers make it difficult to completely eliminate the appearance of false positives in Sheinman's method.

phrase	misclassified/total	percentage
hot but not	9/148	93.92%
big but not	5/423	98.82%
old but not	2/136	98.53%
happy but not	2/116	98.28%
Total	18/823	97.81%

Table 4 Accuracy of phrases searched on Google News.

Furthermore, synset membership is not always consistent with human intuition. For instance, both the words "subatomic" and "gnomish" might be included in the synset for "small," but "subatomic" is used to describe particles, whereas "gnomish" is used to describe people. These flaws suggest that WordNet needs to be more consistent in its attribute pointers for adjectives, as well as in how it links its adjectives together in synsets. In order to consistently be able to detect the false positive errors using Sheinman's method, it is vital for WordNet to be improve the quality of synsets, as well as to vastly expand the coverage of its attribute pointers.

Finally, the dumbbell structure of WordNet as it is renders it difficult to encode adjective scales within each synset. For future use, it would be important to rework the organization of WordNet such that adjective scales could be extracted more easily.

5 Conclusion

Type A false positives suggest that adjectives **X** and **Y** are on the same scale, but that **Y** is not more intense than **X**. There are three types of Type A false positives: repetitions, antonyms, and reversals, and all of these can be corrected relatively easily. **Type B** false positives occur when **X** and **Y** are not synonyms of one another and also do not fall on th same scale. Performing a Level 1 search on WordNet undergenerates false positives, but a Level 2 search overgenerates them. To solve this issue, we must use the attribute pointers, which can accurately classify the category of many of the adjectives contained in WordNet.

After conducting tests, the False Positive Checker accurately classified 97.81% of all phrases conducted through a test. These results could be further improved by improving the structure of WordNet by improving both the precision

and the coverage of its attribute pointer.

All in all, the ability to distinguish the differing intensities of adjective synonyms is vital to being able to master the nuances of the English langauge. By improving the accuracy of Sheinman's method, we can continue to improve our ability to encode these unstated nuances into a lexical tool.

6 Acknowledgments

I would like to thank my advisor, Professor Christiane Fellbaum, for her support and guidance in helping me with this paper. She has provided valuable assistance to the undertaking of the work summarized here. I am also grateful to Princeton University's Student Activities Funding Engine (SAFE) for their generous sponsorship of my travel to the Global WordNet Conference.

References

Fellbaum, Christiane. 1998. *WordNet: an electronic lexical database.* MIT Press WordNet is available from http://www.cogsci.princeton.edu/wn. 2010.

Vera Sheinman, Christiane Fellbaum, Isaac Julien, Peter Schulam, Takenobu Tokunaga. 2013. Large, huge or gigantic? Identifying and encoding intensity relations among adjectives in WordNet. *Language Resources and Evaluation*, 47:1-20.

Vera Sheinman, Takenobu Tokunaga. 2009. Adjscales: Differentiating between similar adjectives for language learners. *Proceedings of the International conference on computer supported education (CSEDU-09), 229-235.*

Mark Finlayson. 2013. Java Wordnet Interface (JWI) 2.2.4. MIT. http://projects.csail.mit.edu/jwi/.