Extracting Ontological Selectional Preferences for Non-Pertainym Adjectives from the Google Corpus

John J. Tanner and **Fernando Gomez**
University of Central Florida
{jjtanner,gomez}@eecs.ucf.edu

Abstract

While there has been much research into using selectional preferences for word sense disambiguation (WSD), much difficulty has been encountered. To facilitate study into this difficulty and aid in WSD in general, a database of the selectional preferences of non-pertainym prenomial adjectives extracted from the Google Web 1T 5-gram Corpus is proposed. A variety of methods for computing the preferences of each adjective over a set of noun categories from WordNet have been evaluated via simulated disambiguation of pseudo-homonyms. The best method of these involves computing for each noun category the ratio of single-word common (i.e. not proper) noun lemma types which can co-occur with a given adjective to the number of single-word common noun lemmata whose estimated frequency is greater than a threshold based on the frequency of the adjective. The database produced by this procedure will be made available to the public.

Introduction

A relatively unexplored avenue for improving noun sense disambiguation is the study of the selectional preferences of the adjectives that modify them: in other words, how strongly the adjective favors modifying the various categories of noun senses found in a general ontology such as WordNet (Fellbaum 1999). In spite of this lack of interest, humans can disambiguate nouns based on the adjective modifying them: consider the phrases "crispy batter," "tasty batter," and "ferocious bat," all of which involve ambiguous nouns, and all of which native English speakers have little trouble understanding. We argue this is due to the selectional preferences of the adjectives: "crispy" and "tasty" have stronger preferences for food than for people, and "ferocious" prefers animals to implements. This article describes the construction of a database of these preferences of the top 3000 non-pertainym adjectives (i.e. adjectives that are defined as "pertaining to X") mentioned in the Google Corpus (Brants and Franz 2006) for various noun categories found in WordNet; pertainyms are excluded as there are pre-existing methods of performing disambiguation with them: to disambiguate "vessel" in "abdominal vessel," a program can chose the sense of "vessel" most closely related to "abdomen," the noun "abdominal" pertains to. Several methods for computing such a database have been evaluated via simulated disambiguation of pseudo-homonyms: the best method of these involves computing for each noun category the ratio of single-word common (i.e. not proper) noun lemma types which can co-occur with a given adjective to the number of single-word common noun lemmata whose estimated frequency is greater than a threshold based on the frequency of the adjective. The database will be made publicly available[1], allowing researchers to easily leverage these preferences in NLP applications and, furthermore, analyze the preferences themselves to discover even better methods of selectional preference extraction: for example, a bootstrapping approach could be devised in which the noun lemmata in the various adjective-noun bigrams analyzed by the system are pre-disambiguated. A simple online tool for disambiguating nouns in adjective-noun bigrams will also be provided to demonstrate the capabilities and potential uses of this database.

The Background section provides the related work. Our approach is explained in the subsection "Scaled Lemmata Fraction" within the Method section. The paper ends with sections on Testing, Results, Discussion and Conclusions.

Background

Much progress has been reported in Web-based/teraword corpora-based "local" context systems, such as that proposed in (Schwartz and Gomez 2008), so why investigate the disambiguation of nouns by modifying adjectives specifically? Such systems have trouble exploiting very local context for disambiguation, that is, context consisting of a direct grammatical relation such as adjective-noun modification. Indeed, (Schwartz and Gomez 2008) uses five or more words of context around the word. As a result, the number of examples for a given context that are long enough to be useful to these systems can be rather small. In the last half decade, disambiguation systems that do handle such very local context tend to focus on verb-direct object pairs (see (Erk 2007) for an example) , leaving adjective-noun bigram techniques in an arrested state.

A good example of research on the selectional preferences of verbs is (Gomez 2004); there, Gomez has successfully

[1] at http://www.cs.ucf.edu/~jjtanner/

used such preferences to disambiguate verbs as well as the nouns that are their arguments.

Many recent works such as (Erk 2007) and (Padó, Padó, and Erk 2007) gather selectional preferences for the roles of various predicates by clustering nouns based on their observed syntactic environments. However, the clusters formed using such techniques tend to be muddled. This confusion can in theory be cleared up via an ontology such as WordNet. Approaches for automatically extracting ontological selectional preferences include (Resnik 1997) and (McCarthy and Carroll 2003). A newer report, (McCarthy, Venkatapathy, and Joshi 2007), suggests a method based on lemmata fractions (as described below) that is even superior to the method of (McCarthy and Carroll 2003); however it only examines verb-direct object preferences, and even at that only indirectly in an experiment concerning compositionality.

Method

We strive to construct a database of selectional preferences for the 3000 adjectives, but first we must select the best method for calculating selectional preferences for the various ontological categories from a collection of candidates described below. Some of the candidates may decide to omit categories. Omitted categories in the hierarchy should be considered to inherit the selection preference of their closest ancestor with a computed score in the ontology. Since WordNet is not quite a tree, care must be taken in finding this "closest ancestor." Here, the ancestors are searched via a breadth-first search, where for each category, the (immediate) normal hypernyms are expanded first in order of appearance in the category's entry in the WordNet 2.1 database, followed by instance hypernyms, the latter to pick up geographical nouns such as "isle" and "tropics" which have no ordinary hypernyms: consult (Fellbaum 1999) for the difference between the two types of hypernym.

Materials

The methods presented in this report use the Google Web 1T 5-gram Corpus (Brants and Franz 2006), which contains 2- through 5-grams occurring at least forty times in Google's index of the Web, pulled from $95 \cdot 10^9$ sentences. It also contains a list of 1-grams occurring at least 200 times. Punctuation is included, as are special markers for the start and end of sentences. An attempt is made to limit the sentences to the English language, with understandably mixed results. This corpus has already been used in several investigations into categorization (Carlson, Mitchell, and Fette 2008) and discovery of metaphors (Krishnakumaran and Zhu 2007), among many others. For a general ontology, WordNet (Fellbaum 1999) is employed. WordNet 3.0, being the latest version when this report was written, was used to select adjectives for testing. WordNet 1.6 was then used in extracting developmental and testing data, for legacy reasons. However, most recent works in WSD that use WordNet use WordNet 2.1, and thus the noun ontology from WordNet 2.1 is used to actually extract preferences. The NLTK toolkit (Bird 2005) was used to access WordNet; it also provided a list of stopwords. To gather testing data, Wikipedia was employed: a data dump from October 26, 2009 (http://download.wikimedia.org/enwiki/20091026/enwiki-20091026-pages-articles.xml.bz2) was downloaded for offline searching.

The top 3000 adjectives according to the Google Corpus unigram frequencies, excluding pertainyms, nouns, forms of verbs according to the WordNet 3.0 Morphy procedure (via NLTK (Bird 2005)) and words in the English stopwords of NLTK, were selected. The adjectives were divided into three bands consisting of the first, second, and third thousand adjectives by frequency. From each band one hundred adjectives were selected uniformly at random, and corresponding adjective-noun bigrams are extracted from Wikipedia to form a developmental data set. A similar process was used to construct a testing set, as detailed in the Testing section.

Conventions

Let $a \cdot l$ be the event of seeing a bigram consisting of lemma a followed by lemma l in a corpus, and $a \cdot \_$ be the event of seeing a bigram beginning with lemma a, and so on. Let $Fr(e)$ be the frequency of seeing event e. Also, let c be the set of senses in an ontological category and c' be the corresponding set of lemmata. The synset that is the parent to a category is considered the *hypernym* of the category. Finally, let $vetted\_lemmata(n)$ be the set of all lemmata in WordNet version n that neither contain any capital letters, spaces, or hyphens (i.e. no proper nouns or multi-word expressions) nor are in the stopwords list of NLTK. Currently lemmata that could be inflected forms of other lemmata are not included in $vetted\_lemmata(n)$. For example, occurrences of the word "bars" are not considered in calculating any of the candidates even though it is a single word lemma in WordNet since "bars" is the plural of "bar."

Candidates

The candidates for calculating selectional preferences are listed below. For each candidate the additive inverse of the natural logarithm of the estimate is used because in general it leads to easier-to-interpret preference scores. Lower scores indicate stronger preferences.

1. First we use $S_{LKM}(a, c) = -\ln F_{LKM}(a, c)$, where F_{LKM} is the following frequency estimate (adapted as mentioned above) from (Lapata, Keller, and McDonald 2001), which is called LKM in this paper,

$$F_{LKM}(a, c) = \frac{\sum_{k|c \subseteq k} Fr'_{LKM}(a, k)}{\mid \{category\ k | c \subseteq k\} \mid} \qquad (1)$$

where

$$Fr'_{LKM}(a, c) = \sum_{l' \in c'} \frac{Fr(a \cdot l')}{\mid \{category\ k | l' \in k'\} \mid} \qquad (2)$$

as a preferences measure. Since Lapata, Keller, and McDonald claim their measure correlates well with human judgements of the plausibility of adjective-noun combinations, it stands to reason that this would measure the plausibility of the combinations of adjectives and noun

senses well. As this correlates to human judges' ratings of adjective-noun plausibility far better than Resnik's method, we use it in place of Resnik's method for comparisons. This method addresses the problem of high-frequency polysemous lemmata by averaging the frequency estimate of a category with the estimates of its super-categories; thus if the system sees "coaxial charger" in the data, it would at first, when calculating Fr'_{LKM}, give improper weight to the category "horse#1", but as there are few other nouns which co-occur with "coaxial" that could refer to animals, the Fr'_{LKM} score for the category "horse#1" would be balanced by rather poor Fr'_{LKM} scores for "placental mammal#1", "vertebrate#1", "animal#1", and so on, leading to a poor score for S_L("coaxial", "horse").

2. However, informal checks of the preferences generated by LKM reveal an inordinate number of categories very high in WordNet with strong selectional preferences. A glance at LKM reveals the problem. Counts accumulate in categories that contain large portions of the hierarchy more than in their subcategories, even if it is only their subcategories for which the adjective has strong preferences. While the averaging method mentioned in the description of LKM attempts to address this for words lower in the hierarchy, a notable number of word senses lie in categories only a handful of levels away from the root and are thus given abnormally strong preferences via LKM. To alleviate this, we devised a normalized version called normalized LKM:

$$F_{nLKM}(a,c) = \frac{F_{LKM}(a,c)}{F_{LKM}(\text{“}the\text{”},c)} \quad (3)$$

where $F_{LKM}(\text{“}the\text{”},c)$ is a measure of the inherent frequency of the nouns in category c. As can be seen in Equations1 and 2, this employs the pattern "the l'" for l' in c', which is used to filter out verbal uses of the lemmata in c'. Due to the normalization, a category will not get a stronger score simply because it is larger or more frequent in general. Rather, it must co-occur *with the adjective* more often to receive a strong score. An analogy can be drawn with pointwise mutual information.
As above, the actual preference score is computed as $S_{nLKM}(a,c) = -\ln F_{nLKM}(a,c)$. Some trials were run on the developmental data set using the scaling procedure mentioned in the Scaled Lemmata Fraction Method subsection below to limit the lemmata analyzed in computing the denominator; however no improvement could be found.

3. The lemmata fraction $S_{LF} = -\ln F_{LF}$ as described below. Since McCarthy, Venkatapathy, and Joshi's work suggests that this method is superior to McCarthy and Carroll's method, we omit McCarthy and Carroll's method in the current study.

4. The scaled lemmata fraction $S_{SLF} = -\ln F_{SLF}$, again as described below. This is the method mentioned in the introduction.

Lemmata Fraction Method

Due to the rapid decay of co-occurrence frequencies for an adjective and lemmata in a given category, selectional preference methods that sum co-occurrence frequencies can be dominated by a small number of frequent lemmata. Since frequent lemmata also tend to be polysemous, this can lead to considerable noise, which can be circumvented by counting lemma types instead of lemma tokens, as in McCarthy, Venkatapathy and Joshi's work: to get a selectional preferences score between an predicate (or here, adjective) and a category, simply count the number of vetted lemmata in the category that co-occur with the predicate and divide by the total number of vetted lemmata in the category: call this score $F_{LF}(a,c)$, where $v\_l = vetted\_lemmata(2.0)$.

$$F_{LF}(a,c) = \frac{\mid \{m \in c' \cap v\_l \mid Fr(a \cdot m) > 0\} \mid}{\mid \{m \in c' \cap v\_l\} \mid}$$

Note that McCarthy, Venkatapathy, and Joshi originally required that at least two lemmata from a category co-occur with the predicate; we do not.

Scaled Lemmata Fraction Method

However, estimation of the lemmata ratio can be difficult as even data from the Google Corpus is limited: the lemmata that can be reasonably combined with the adjective will often occur less than forty times, the cutoff for inclusion in the Google Corpus. Simple investigation reveals that the frequencies of words in various categories decay at notably different rates, which allows the cutoff to distort the ordering of selectional strengths. Consider that the frequencies of lemmata in "sports equipment#1" decay much more rapidly than those for "placental mammal#1"; the slow decay of "placental mammal#1" means that, when analyzing the selectional preferences for "angry," McCarthy, Venkatapathy and Joshi's method must consider many lemmata in the category "placental mammal#1" that due to their inherent low frequency would not have a chance to co-occur with "angry." The category "sports equipment#1" does not have nearly as many such lemmata, due to its quicker decay, leading the system to conclude that "angry" selects for nouns in "sports equipment#1" more strongly than it does for those in "placental mammal#1": an angry bat, according to this system, is more likely a stick of wood than it is a placental mammal, in contrast to native English speakers' expectations. To remedy this, one can scale the denominator of the lemmata fraction to account for the low frequency lemmata; this scaled lemmata fraction then becomes

$$F_{SLF}(a,c) = \frac{\mid m \in c' \cap v\_l \mid Fr(a \cdot m) > 0 \mid}{\mid m \in c' \cap v\_l \mid \sum_{j:j \in J} Fr(j \cdot m) > k(b) \mid}$$

where J is the set of all adjectives in WordNet 3.0. Recall that the adjectives analyzed in this work are divided into three 1000-adjective bands: now $b(a)$ indicates the band a belongs to, and $k(b(a))$ is a threshold determined for that band, computed as $k(b(a)) = q(b(a))\frac{N}{Fr(a \cdot \_)}$, where N is the total number of adjective-noun pair tokens for all adjectives and noun and $q(b)$ is adjusted for adjective frequency:

less frequent adjectives require lower values of q as small denominators lead to unacceptable noise. After some informal experiments run on the developmental data set, we choose $q = 2$ for the first band, $q = 1$ for the second, and $q = 0.5$ for the third.

Only categories with at least five lemmata types from $vetted\_lemmata(2.1)$ with frequencies in the unigrams list greater than the above cutoff were analyzed to help alleviate noise: as mentioned above, fewer lemmata types in the denominator cause the presence or absence of a single erroneous lemma type in combination with the adjective to produce unacceptably large errors in the computed preference score. In addition for each adjective a category must have at least one lemma that co-occurs with it or be omitted.

Testing

As mentioned above, 3000 adjectives were selected and split into three frequency bands of a thousand words each. To form the testing data set, one hundred adjectives were selected uniformly at random from each band in a separate sample from that used to construct the developmental data set. For each adjective, adjective-noun combinations were selected from Wikipedia where the noun was in $vetted\_lemmata(1.6)$ and had at least three letters (to avoid abbreviations), and the combination was not followed by another noun: that is, either the combination was followed by either a non-alphabetical symbol other than a hyphen, comma, or space; or by one space and an alphabetical string that is not a noun in WordNet 1.6. Of these, pairs that occurred only once were discarded.

In an effort to get a accurate and reliable measure of the quality of the selectional preferences, we followed the example of (Rooth et al. 1999) and (Erk 2007) by simulating the disambiguation of *pseudo-homonyms*. While it is theoretically preferable to ask volunteers to disambiguate nouns in adjective-noun pairs culled from real-world sources, thousands of examples would have to be disambiguated by multiple volunteers. Positive results from the disambiguation of pseudo-homonyms would provide a quick and inexpensive method for motivating more intensive testing in the future.

Disambiguation of Pseudo-homonyms

In this procedure, for each attested adjective-noun pair a pseudo-homonym is constructed having two broad "senses" corresponding to the noun from the pair and a confounder that cannot be found in combination with the adjective: in this paper both the Google Corpus bigrams list and Wikipedia were searched to generate the confounder. For each hypothetical combination of an adjective and a pseudo-homonym, the system attempts to determine the correct noun by test score: for each WordNet sense of both actual nouns comprising the pseudo-homonym the system looks for the score of the closest ancestor of the sense in the adjective's selectional preference list, as defined in the beginning of the Method section, to each sense of each actual noun in the pseudo-homonym. If no category can be found, the sense is given a score of infinity; this should only happen rarely. As mentioned above, lower scores are better. Each noun is assigned the score of its best sense, and the actual noun with the best score is declared to be the proper sense of the pseudo-homonym. However, the system only attempts to disambiguate pseudo-homonyms if its constituent nouns have different scores; in other words, it refuses to give an answer for ties.

For example, consider the bigram "crispy batter." To construct a pseudo-homonym for this combination, the system will choose a noun that does not co-occur with "crispy," such as "school." It will then attempt to disambiguate this pseudo-homonym "batter/school" as used in the hypothetical bigram "crispy batter/school" by checking the selectional preferences of "crispy" for the various senses of "batter" and "school." If the system is functioning properly, it will find that the best sense belongs to "batter," and thus select "batter" as the proper sense of "batter/school."

Reviewers have noted that this procedure is not necessarily equivalent to the disambiguation of real homonyms, let alone of polysemous words. However, it is still informative: better results indicate that the preference lists are ordered correctly, and that for each adjective more probable noun categories are given stronger preference scores than less probable categories.

Experiment

For each of the three frequency bands b_i, ten sets of adjective-noun-pair *tokens* corresponding to the adjectives selected from that band for testing each sufficiently large enough to contain $\frac{|lemmatype\ t|\exists j \cdot t\ where\ adjective\ j \in b_i|}{25}$ of the total of all adjective-noun-pair *types* for every adjective in the appropriate band were selected at random with replacement according to the frequency of each pair from the data culled from Wikipedia as described above, where the noun is in $vetted\_lemmata(1.6)$ and is neither an adjective according to WordNet 1.6 nor is in the stopwords list provided by NLTK. For each pair, a pseudo-homonym is constructed as per above, where the confounder, a noun in $vetted\_lemmata(2.1)$ that is neither an adjective according to WordNet 3.0 nor a member of the stopwords list provided by NLTK, is selected at random with replacement according to the frequencies from the bigrams list of the Google Corpus of the bigrams formed by prepending each candidate with the word "the": by the nature of the Google Corpus, this means that such a bigram must exist at least forty times even to be considered.

The adjective-noun pairs were chosen according to frequency in an attempt to reflect the probability of encountering such combinations of concepts. Nonetheless, one must realize that the frequencies of the adjectives in each band decrease rapidly, just like the lemmata in the noun categories, causing the top adjectives to dominate: for each band, roughly the top ten adjectives account for half of the adjective-noun pairs. Even if the adjective-noun pairs were selected uniformly by type, similar results would be seen, since the more frequent adjectives tend to have a greater number of adjective-noun-pair types as well as tokens. With this in mind, we decided to report the results for each band separately, to demonstrate the performance of the system

on less frequent adjectives that would otherwise be virtually completely dominated.

Experiment 1

For each set of data, each algorithm was trained on the Google Corpus with the bigrams selected for testing removed along with their bigram occurrences; however the frequencies for their constituent adjectives and nouns in other contexts were left the same. Preferences were acquired for categories in WordNet 2.1 with at least thirty lemmata with three or more letters (again, to avoid abbreviations) having no capital letters, underscores, hyphens, or spaces. Using these preferences, the system attempted disambiguation of the pseudo-homonyms.

Thirty trials were conducted and the F1 value, the harmonic mean of the precision (number of correct disambiguations / number of attempted disambiguations) and the recall (number of correct disambiguations / total number of pseudo-homonym tokens) of each trial, was recorded. The means and sample standard deviations of these results for each combination of method and frequency band are reported in Table 1.

Experiment 2

We examined the best performing method, the Scaled Lemmata Fraction method, in more detail using the same testing data: for each attempted disambiguation, the difference of the preference scores is computed, and a tally is kept for various score difference bands: (0,0.25), [0.25,0.5), [0.5,0.75), [0.75,1.0), [1.0,1.5), [1.5,2.0), [2.0,2.5), [2.5,3.0), and [3.0,Infinity).

For each frequency band, the precision for each score-difference band is recorded, as is the percentage of the total pseudo-homonyms whose preference score difference lay within that band, and the means of these results for the first frequency band (for space reasons) are reported in Table 2, as are their sample standard deviations.

Results

The summarized results for Experiment 1 are reported in Table 1. In addition, paired *t*-tests conducted on the actual results reveal the differences between Normalized LKM and the Lemmata Fraction methods for the second frequency band to be significant, but with a p-value of only 0.046; all other differences between methods within the same band are significant with p-values less than $1.2 \cdot 10^{-12}$, a much more comfortable margin. From comparing the results for the Lapata, Keller, and McDonald method (LKM) and Normalized LKM it is apparent that the normalization term is very important, as expected. The Scaled Lemmata Fraction method likewise clearly improves on the plain Lemmata Fraction method, as predicted. However, note that the Lemmata Fraction method fared only somewhat better than Normalized LKM; yet even this is illuminating, given that the Lemmata Fraction method is simpler than Normalized LKM. The values for each individual method vary erratically among the various frequency bands, especially when considering the second frequency band. One possibility is that the F1 score varies with the type of adjectives that happened to be selected for testing: adjectives are not equally selective, as can be seen by pondering the nouns that can replace X in "nice X" and "multivariate X." Further investigation is warranted, but nonetheless the differences *within* each frequency band are encouraging. The suggestions in this report do lead to improved results.

The results of Experiment 2 (see Table 2) show that accuracy does indeed improve as the difference in preference score between the correct and incorrect senses of the pseudo-homonym increases; however at the same time the percentage of pseudo-homonyms with such a difference decreases dramatically.

Discussion

We have suggested that the success of methods such as the Scaled Lemmata Fraction method is due to resiliency to noise caused by polysemous lemmata, yet these methods do not overcome this noise completely. Often this noise is due to systemic polysemy: the names of many fruits such as oranges and apples can also refer to the trees they grow on, and many animal body parts can be considered food items, to name just a few examples. Another source of odd preferences is the presence of adjective-noun-noun phrases such as "airtight cookie tin" in the Google Corpus. Other problems include odd abbreviations, foreign phrases and the somewhat "artificial" titles of bands, movies and so on. See (Carlson, Mitchell, and Fette 2008) and (Kilgarriff 2007) for further problems with the Google Corpus. Even WordNet has odd organizational idiosyncrasies: almond is listed under "edible nut#1", yet this category is not listed under any category corresponding with food.

Even with these problems, the results in the paper suggest that the Scaled Lemmata Fraction method is a simple promising method for estimating ontological selectional preferences. Informal inspection of the lists of preferences generated by this method shows that the categories with the best preference scores for an adjective are indeed strongly associated with that adjective, and the categories with the worst scores have very weak or non-existent connections to the adjective. However, preferences in the middle of the lists are less reliable: for the adjective "sly," good categories such as "businessman#1," "manner#2" (personality), and "crime#1" have preference scores close to those of implausible categories such as "convex_shape#1," "animal_product#1," and "weather#1." This is not unexpected, as relatively moderate amounts of co-occurrences of an adjective with a noun category could be the result of stray lemmata, *sub*-categories that happen to be strongly selected by the adjective, or actual moderate preferences.

Conclusion

Humans have a fascinating ability to disambiguate adjective-noun bigrams with little context. A system that can perform such disambiguation well is an advance, and a relatively simple system should prove inspiring. Such a system, selectional preferences as computed by the Scaled Lemmata Fraction method, is presented in this report.

Band	Lapata, Keller, and McDonald(LKM)	Normalized LKM	Lemmata Fraction	Scaled Lemmata Fraction
< 1K	72.1 *0.3*	75.4 *0.6*	79.5 *0.3*	83.6 *0.3*
$1 - 2$K	65.4 *0.8*	75.3 *0.6*	75.7 *0.9*	80.1 *0.9*
$2 - 3$K	62.2 *1.3*	75.7 *1.1*	78.2 *1.1*	83.6 *0.9*

Values are in percentage points. Means of the F1 values for each trial in normal font; *sample standard deviations of the F1 values in italics*. 30 trials conducted.

Table 1: Experiment 1 – F1 average of precision and recall for the various methods over the various frequency bands of adjectives.

Score difference	0-0.25	0.25-0.5	0.5-0.75	0.75-1.0	1.0-1.5	1.5-2.0	2.0-2.5	2.5-3.0	> 3.0
Mean precision	65.5 *0.6*	84.4 *0.5*	92.0 *0.5*	94.4 *0.5*	96.3 *0.3*	98.0 *0.4*	98.8 *0.5*	99.4 *0.7*	99.9 *0.5*
% of total pseudo-homonyms	28.8 *0.4*	23.6 *0.3*	16.5 *0.3*	11.3 *0.2*	11.8 *0.2*	4.7 *0.2*	1.9 *0.1*	0.7 *0.1*	0.2 *

Values in percentage points for the first 1000 adjectives using the Scaled Lemmata Fraction method. Means in normal font; *sample standard deviations in italics*. 30 trials conducted. * means < 0.05.

Table 2: Experiment 2 – Mean precision and percentage of total pseudo-homonyms attempts as a function of selectional preference score difference

This method can also shed light on the difficult task of adjective classification and adjective semantics. In addition, it is hoped that such a method can be extended to other relationships such as verb-direct-object or verb-subject, leading to a high-coverage very-local-context disambiguation procedure. Finally, it should be noted that the improvements in this paper came not from applying a new generic machine learning algorithm such as a new clustering or SVM technique, but out of observations of the frequency distributions of categories of noun lemmata and of combinations of these lemmata with various adjectives. This supports our guiding idea that improvements in NLP can come from investigations of linguistic behavior. Further investigation should produce even better results.

Acknowledgments

This research was supported in part by the NASA Engineering and Safety Center under Grant/Cooperative Agreement NNX08AJ98A.

References

Bird, S. 2005. NLTK-Lite: Efficient scripting for natural language processing. In *Proceedings of the 4th International Conference on Natural Language Processing (ICON)*, 11–18. Kanpur, India. New Delhi: Allied Publishers.

Brants, T., and Franz, A. 2006. Web 1T 5-gram version 1. Linguistic Data Consortium, Philiadelphia.

Carlson, A.; Mitchell, T. M.; and Fette, I. 2008. Data analysis project: Leveraging massive textual corpora using n-gram statistics. Technical report, Carnegie Mellon.

Erk, K. 2007. A simple, similarity-based model for selectional preferences. In *ACL-07*, 216–223.

Fellbaum, C. 1999. *WordNet: An Electionic Lexical Database*. MIT Press.

Gomez, F. 2004. Building verb predicates: A computational view. In *ACL-04*.

Kilgarriff, A. 2007. Googleology is bad science. *Computational Linguistics* 33(1):147–151.

Krishnakumaran, S., and Zhu, X. 2007. Hunting elusive metaphors using lexical resources. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, 13–20. Rochester, New York: Association for Computational Linguistics.

Lapata, M.; Keller, F.; and McDonald, S. 2001. Evaluating smoothing algorithms against plausibility judgements. In *ACL-01*, 354–361.

McCarthy, D., and Carroll, J. 2003. Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Computational Linguistics* 29(4):639–654.

McCarthy, D.; Venkatapathy, S.; and Joshi, A. 2007. Detecting compositionality of verb-object combinations using selectional preferences. In *EMNLP-CoNLL-07*, 369–379.

Padó, S.; Padó, U.; and Erk, K. 2007. Flexible, corpus-based modelling of human plausibility judgements. In *EMNLP-CoNLL-07*, 400–409.

Resnik, P. 1997. Selectional preference and sense disambiguation. In *Proceedings of ACL SIGLEX Workshop on Tagging Text with Lexical Semantics, Why, What and How?*

Rooth, M.; Riezler, S.; Prescher, D.; Carroll, G.; and Beil, F. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *ACL-09*, 104–111.

Schwartz, H. A., and Gomez, F. 2008. Acquiring knowledge from the web to be used as selectors for noun sense disambiguation. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, 105–112. Association for Computational Linguistics.