Evaluating Smoothing Algorithms against Plausibility Judgements

Maria Lapata and **Frank Keller**
Department of Computational Linguistics
Saarland University
PO Box 15 11 50
66041 Saarbrücken, Germany
{mlap, keller}@coli.uni-sb.de

Scott McDonald
Language Technology Group
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
scottm@cogsci.ed.ac.uk

Abstract

Previous research has shown that the plausibility of an adjective-noun combination is correlated with its corpus co-occurrence frequency. In this paper, we estimate the co-occurrence frequencies of adjective-noun pairs that fail to occur in a 100 million word corpus using smoothing techniques and compare them to human plausibility ratings. Both class-based smoothing and distance-weighted averaging yield frequency estimates that are significant predictors of rated plausibility, which provides independent evidence for the validity of these smoothing techniques.

1 Introduction

Certain combinations of adjectives and nouns are perceived as more plausible than others. A classical example is *strong tea*, which is highly plausible, as opposed to *powerful tea*, which is not. On the other hand, *powerful car* is highly plausible, whereas *strong car* is less plausible. It has been argued in the theoretical literature that the plausibility of an adjective-noun pair is largely a collocational (i.e., idiosyncratic) property, in contrast to verb-object or noun-noun plausibility, which is more predictable (Cruse, 1986; Smadja, 1991).

The collocational hypothesis has recently been investigated in a corpus study by Lapata et al. (1999). This study investigated potential statistical predictors of adjective-noun plausibility by using correlation analysis to compare judgements elicited from human subjects with five corpus-derived measures: co-occurrence frequency of the adjective-noun pair, noun frequency, conditional probability of the noun given the adjective, the log-likelihood ratio, and Resnik's (1993) selectional association measure. All predictors but one were positively correlated with plausibility; the highest correlation was obtained with co-occurrence frequency. Resnik's selectional association measure surprisingly yielded a significant negative correlation with judged plausibility. These results suggest that the best predictor of whether an adjective-noun combination is plausible or not is simply how often the adjective and the noun collocate in a record of language experience.

As a predictor of plausibility, co-occurrence frequency has the obvious limitation that it cannot be applied to adjective-noun pairs that never occur in the corpus. A zero co-occurrence count might be due to insufficient evidence or might reflect the fact that the adjective-noun pair is inherently implausible. In the present paper, we address this problem by using smoothing techniques (distance-weighted averaging and class-based smoothing) to recreate missing co-occurrence counts, which we then compare to plausibility judgements elicited from human subjects. By demonstrating a correlation between recreated frequencies and plausibility judgements, we show that these smoothing methods produce realistic frequency estimates for missing co-occurrence data. This approach allows us to establish the validity of smoothing methods independent from a specific natural language processing task.

2 Smoothing Methods

Smoothing techniques have been used in a variety of statistical natural language processing applications as a means to address data sparseness, an inherent problem for statistical methods which rely on the relative frequencies of word combinations. The problem arises when the probability of word combinations that do not occur in the training data needs to be estimated. The smoothing methods proposed in the literature (overviews are provided by Dagan et al. (1999) and Lee (1999)) can be generally divided into three types: *discounting* (Katz, 1987), *class-based smoothing* (Resnik, 1993; Brown et al., 1992; Pereira et al., 1993), and *distance-weighted averaging* (Grishman and Sterling, 1994; Dagan et al., 1999).

Discounting methods decrease the probability of previously seen events so that the total probability of observed word co-occurrences is less

Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter of the Association for Computational Linguistics, 346–353. Toulouse, 2001.

than one, leaving some probability mass to be redistributed among unseen co-occurrences.

Class-based smoothing and distance-weighted averaging both rely on an intuitively simple idea: inter-word dependencies are modelled by relying on the corpus evidence available for words that are similar to the words of interest. The two approaches differ in the way they measure word similarity. Distance-weighted averaging estimates word similarity from lexical co-occurrence information, viz., it finds similar words by taking into account the linguistic contexts in which they occur: two words are similar if they occur in similar contexts. In class-based smoothing, classes are used as the basis according to which the co-occurrence probability of unseen word combinations is estimated. Classes can be induced directly from the corpus (Pereira et al., 1993; Brown et al., 1992) or taken from a manually crafted taxonomy (Resnik, 1993). In the latter case the taxonomy is used to provide a mapping from words to conceptual classes.

In language modelling, smoothing techniques are typically evaluated by showing that a language model which uses smoothed estimates incurs a reduction in perplexity on test data over a model that does not employ smoothed estimates (Katz, 1987). Dagan et al. (1999) use perplexity to compare back-off smoothing against distance-weighted averaging methods and show that the latter outperform the former. They also compare different distance-weighted averaging methods on a pseudo-word disambiguation task where the language model decides which of two verbs v_1 and v_2 is more likely to take a noun n as its object. The method being tested must reconstruct which of the unseen (v_1, n) and (v_2, n) is a valid verb-object combination.

In our experiments we recreated co-occurrence frequencies for unseen adjective-noun pairs using two different approaches: taxonomic class-based smoothing and distance-weighted averaging.[1] We evaluated the recreated frequencies by comparing them with plausibility judgements elicited from human subjects. In contrast to previous work, this type of evaluation does not presuppose that the recreated frequencies are needed for a specific natural language processing task. Rather, our aim is to establish an independent criterion for the validity of smoothing techniques by comparing them to plausibility judgements, which are known to correlate with co-occurrence frequency (Lapata et al., 1999).

In the remainder of this paper we present class-based smoothing and distance-weighted averaging as applied to unseen adjective-noun combinations (see Sections 2.1 and 2.2). Section 3 details our judgement elicitation experiment and reports our results.

2.1 Class-based Smoothing

We recreated co-occurrence frequencies for unseen adjective-noun pairs using a simplified version of Resnik's (1993) selectional association measure. Selectional association is defined as the amount of information a given predicate carries about its argument, where the argument is represented by its corresponding classes in a taxonomy such as WordNet (Miller et al., 1990). This means that predicates which impose few restrictions on their arguments have low selectional association values, whereas predicates selecting for a restricted number of arguments have high selectional association values. Consider the verbs *see* and *polymerise*: intuitively there is a great variety of things which can be seen, whereas there is a very specific set of things which can be polymerised (e.g., ethylene). Resnik demonstrated that his measure of selectional association successfully captures this intuition: selectional association values are correlated with verb-argument plausibility as judged by native speakers.

However, Lapata et al. (1999) found that the success of selectional association as a predictor of plausibility does not seem to carry over to adjective-noun plausibility. There are two potential reasons for this: (1) the semantic restrictions that adjectives impose on the nouns with which they combine appear to be less strict than the ones imposed by verbs (consider the adjective *superb* which can combine with nearly any noun); and (2) given their lexicalist nature, adjective-noun combinations may defy selectional restrictions yet be intuitively plausible (consider the pair *sad day*, where sadness is not an attribute of *day*).

To address these problems, we replaced Resnik's information-theoretic measure with a simpler measure which makes no assumptions with respect to the contribution of a semantic class to the total quantity of information provided by the predicate about the semantic classes of its argument. We simply substitute the noun occurring in the adjective-noun combination with the concept by which it is represented in the taxonomy and estimate the adjective-noun co-occurrence frequency by counting the number of times the concept corresponding to the noun is observed to co-occur with the adjective in the corpus. Because a given word is not always represented by a single class in the taxonomy (i.e., the

[1] Discounting methods were not included as Dagan et al. (1999) demonstrated that distance-weighted averaging achieves better language modelling performance than back-off.

Adjective	Class	$f(a,n)$
proud	⟨`entity`⟩	13.70
proud	⟨`life from`⟩	9.80
proud	⟨`causal agent`⟩	9.50
proud	⟨`person`⟩	9.00
proud	⟨`leader`⟩	.75
proud	⟨`superior`⟩	.08
proud	⟨`supervisor`⟩	.00

Table 1: Frequency estimation for *proud chief* using WordNet

noun co-occurring with an adjective can generally be the realisation of one of several conceptual classes), we constructed the frequency counts for an adjective-noun pair for each conceptual class by dividing the contribution from the adjective by the number of classes to which it belongs (Lauer, 1995; Resnik, 1993):

$$(1) \quad f(a,c) \approx \sum_{n' \in c} \frac{f(a,n')}{|classes(n')|}$$

where $f(a,n')$ is the number of times the adjective a was observed in the corpus with concept $c \in classes(n')$ and $|classes(n')|$ is the number of conceptual classes noun n' belongs to. Note that the estimation of the frequency $f(a,c)$ relies on the simplifying assumption that the noun co-occurring with the adjective is distributed evenly across its conceptual classes. This simplification is necessary unless we have a corpus of adjective-noun pairs labelled explicitly with taxonomic information.[2]

Consider the pair *proud chief* which is not attested in the British National Corpus (BNC) (Burnard, 1995). The word *chief* has two senses in WordNet and belongs to seven conceptual classes (⟨`causal agent`⟩, ⟨`entity`⟩, ⟨`leader`⟩, ⟨`life form`⟩, ⟨`person`⟩, ⟨`superior`⟩, and ⟨`supervisor`⟩) This means that the co-occurrence frequency of the adjective-noun pair will be constructed for each of the seven classes, as shown in Table 1. Suppose for example that we see the pair *proud leader* in the corpus. The word *leader* has two senses in WordNet and belongs to eight conceptual classes (⟨`person`⟩, ⟨`life from`⟩, ⟨`entity`⟩, ⟨`causal agent`⟩, ⟨`feature`⟩, ⟨`merchandise`⟩, ⟨`commodity`⟩, and ⟨`object`⟩). The words *chief* and *leader* have four conceptual classes in common, i.e., ⟨`person`⟩ and ⟨`life form`⟩, ⟨`entity`⟩, and ⟨`causal agent`⟩. This means that we will increment the observed co-occurrence count of *proud* and ⟨`person`⟩, *proud* and ⟨`life form`⟩, *proud* and ⟨`entity`⟩, and *proud* and ⟨`causal agent`⟩ by $\frac{1}{8}$. Since we do not know the actual class of the noun *chief* in the corpus, we weight the contribution of each class by taking the average of the constructed frequencies for all seven classes:

$$(2) \quad f(a,n) = \frac{\sum\limits_{c \in classes(n)} \sum\limits_{n' \in c} \frac{f(a,n')}{|classes(n')|}}{|classes(n)|}$$

Based on (2) the recreated frequency for the pair *proud chief* in the BNC is 6.12 (see Table 1).

2.2 Distance-Weighted Averaging

Distance-weighted averaging induces classes of similar words from word co-occurrences without making reference to a taxonomy. A key feature of this type of smoothing is the function which measures distributional similarity from co-occurrence frequencies. Several measures of distributional similarity have been proposed in the literature (Dagan et al., 1999; Lee, 1999). We used two measures, the Jensen-Shannon divergence and the confusion probability. Those two measures have been previously shown to give promising performance for the task of estimating the frequencies of unseen verb-argument pairs (Dagan et al., 1999; Grishman and Sterling, 1994; Lapata, 2000; Lee, 1999). In the following we describe these two similarity measures and show how they can be used to recreate the frequencies for unseen adjective-noun pairs.

Jensen-Shannon Divergence. The Jensen-Shannon divergence is an information-theoretic measure that recasts the concept of distributional similarity into a measure of the "distance" (i.e., dissimilarity) between two probability distributions.

Let w_1 and w'_1 be an unseen sequence of two words whose distributional similarity is to be determined. Let $P(w_2|w_1)$ denote the conditional probability of word w_2 given word w_1 and $P(w_2|w'_1)$ denote the conditional probability of w_2 given w'_1. For notational simplicity we write $p(w_2)$ for $P(w_2|w_1)$ and $q(w_2)$ for $P(w_2|w'_1)$. The Jensen-Shannon divergence is defined as the average Kullback-Leibler divergence of each of two distributions to their average distribution:

$$(3) \quad J(p,q) = \frac{1}{2}\left[D\left(p \middle\| \frac{p+q}{2}\right) + D\left(q \middle\| \frac{p+q}{2}\right)\right]$$

where $(p+q)/2$ denotes the average distribution:

$$(4) \quad \frac{1}{2}\left(P(w_2|w_1) + P(w_2|w'_1)\right)$$

The Kullback-Leibler divergence is an information-theoretic measure of the dissimilarity of two probability distributions p and q, defined as follows:

$$(5) \quad D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$$

[2] There are several ways of addressing this problem, e.g., by discounting the contribution of very general classes by finding a suitable class to represent a given concept (Clark and Weir, 2001).

In our case the distributions p and q are the conditional probability distributions $P(w_2|w_1)$ and $P(w_2|w_1')$, respectively. Computation of the Jensen-Shannon divergence depends only on the linguistic contexts w_2 which the two words w_1 and w_1' have in common. The Jensen-Shannon divergence, a dissimilarity measure, is transformed to a similarity measure as follows:

$$(6) \quad W_J(p,q) = 10^{-\beta J(p,q)}$$

The parameter β controls the relative influence of the words most similar to w_1: if β is high, only words extremely similar to w_1 contribute to the estimate, whereas if β is low, less similar words also contribute to the estimate.

Confusion Probability. The confusion probability is an estimate of the probability that word w_1' can be substituted by word w_1, in the sense of being found in the same linguistic contexts.

$$(7) \quad P_c(w_1|w_1') = \sum_{w_2} P(w_1|w_2)P(w_2|w_1')$$

where $P_c(w_1'|w_1)$ is the probability that word w_1' occurs in the same contexts w_2 as word w_1, averaged over these contexts.

Let w_2w_1 be two unseen co-occurring words. We can estimate the conditional probability $P(w_2|w_1)$ of the unseen word pair w_2w_1 by combining estimates for co-occurrences involving similar words:

$$(8) \quad P_{\mathrm{SIM}}(w_2|w_1) = \sum_{w_1' \in S(w_1)} \frac{W(w_1,w_1')}{N(w_1)} P(w_2|w_1')$$

where $S(w_1)$ is the set of words most similar to w_1, $W(w_1,w_1')$ is the similarity function between w_1 and w_1', and $N(w_1)$ is a normalising factor $N(w_1) = \sum_{w_1'} W(w_1,w_1')$. The conditional probability $P_{\mathrm{SIM}}(w_2|w_1)$ can be trivially converted to co-occurrence frequency as follows:

$$(9) \quad f(w_1,w_2) = P_{\mathrm{SIM}}(w_2|w_1) f(w_1)$$

Parameter Settings. We experimented with two approaches to computing $P(w_2|w_1')$: (1) using the probability distribution $P(n|a)$, which discovers similar adjectives and treats the noun as the context; and (2) using $P(a|n)$, which discovers similar nouns and treats the adjective as the context. These conditional probabilities can be easily estimated from their relative frequency in the corpus as follows:

$$(10) \quad P(n|a) = \frac{f(a,n)}{f(a)} \qquad P(a|n) = \frac{f(a,n)}{f(n)}$$

The performance of distance-weighted averaging depends on two parameters: (1) the number of items over which the similarity function is computed (i.e., the size of the set $S(w_1)$ denoting the set of words most similar to w_1), and (2) the value of the parameter β (which is only relevant for the Jensen-Shannon divergence). In this study we recreated adjective-noun frequencies using the 1,000 and 2,000 most frequent items (nouns and adjectives), for both the confusion probability and the Jensen-Shannon divergence.[3] Furthermore, we set β to .5, which experiments showed to be the best value for this parameter.

Jensen-Shannon		Confusion Probability	
proud	chief	proud	chief
young	chairman	lone	venture
old	venture	adverse	chairman
dying	government	grateful	importance
wealthy	leader	sole	force
lone	official	wealthy	representative
dead	scientist	elderly	president
rich	manager	registered	official
poor	initiative	dear	manager
elderly	president	deliberate	director

Table 2: The ten most similar adjectives to *proud* and the ten most similar nouns to *chief*

Once we know which words are most similar to the either the adjective or the noun (irrespective of the function used to measure similarity) we can exploit this information in order to recreate the co-occurrence frequency for unseen adjective-noun pairs. We use the weighted average of the evidence provided by the similar words, where the weight given to a word w_1' depends on its similarity to w_1 (see (8) and (9)). Table 2 shows the ten most similar adjectives to the word *proud* and then the ten most similar nouns to the word *chief* using the Jensen-Shannon divergence and the confusion probability. Here the similarity function was calculated over the 1,000 most frequent adjectives in the BNC.

3 Collecting Plausibility Ratings

In order to evaluate the smoothing methods introduced above, we first needed to establish an independent measure of plausibility. The standard approach used in experimental psycholinguistics is to elicit judgements from human subjects; in this section we describe our method for assembling the set of experimental materials and collecting plausibility ratings for these stimuli.

3.1 Method

Materials. We used a part-of-speech annotated, lemmatised version of the BNC. The BNC is a large, balanced corpus of British English, consisting of 90 million words of text and 10 million words of speech. Frequency information obtained

[3] These were shown to be the best parameter settings by Lapata (2000). Note that considerable latitude is available when setting these parameters; there are 151,478 distinct adjective types and 367,891 noun types in the BNC.

Adjective	Nouns		
hungry	tradition	innovation	prey
guilty	system	wisdom	wartime
temporary	conception	surgery	statue
naughty	regime	rival	protocol

Table 3: Example stimuli for the plausibility judgement experiment

from the BNC can be expected to be a reasonable approximation of the language experience of a British English speaker.

The experiment used the same set of 30 adjectives discussed in Lapata et al. (1999). These adjectives were chosen to be minimally ambiguous: each adjective had exactly two senses according to WordNet and was unambiguously tagged as 'adjective' 98.6% of the time, measured as the number of different part-of-speech tags assigned to the word in the BNC. For each adjective we obtained all the nouns (excluding proper nouns) with which it failed to co-occur in the BNC.

We identified adjective-noun pairs by using Gsearch (Corley et al., 2001), a chart parser which detects syntactic patterns in a tagged corpus by exploiting a user-specified context free grammar and a syntactic query. From the syntactic analysis provided by the parser we extracted a table containing the adjective and the head of the noun phrase following it. In the case of compound nouns, we only included sequences of two nouns, and considered the rightmost occurring noun as the head. From the adjective-noun pairs obtained this way, we removed all pairs where the noun had a BNC frequency of less than 10 per million, in order to reduce the risk of plausibility ratings being influenced by the presence of a noun unfamiliar to the subjects. Each adjective was then paired with three randomly-chosen nouns from its list of non-co-occurring nouns. Example stimuli are shown in Table 3.

Procedure. The experimental paradigm was magnitude estimation (ME), a technique standardly used in psychophysics to measure judgements of sensory stimuli (Stevens, 1975), which Bard et al. (1996) and Cowart (1997) have applied to the elicitation of linguistic judgements. The ME procedure requires subjects to estimate the magnitude of physical stimuli by assigning numerical values proportional to the stimulus magnitude they perceive. In contrast to the 5- or 7-point scale conventionally used to measure human intuitions, ME employs an interval scale, and therefore produces data for which parametric inferential statistics are valid.

ME requires subjects to assign numbers to a series of linguistic stimuli in a proportional fashion. Subjects are first exposed to a modulus item, which they assign an arbitrary number. All other stimuli are rated proportional to the modulus. In this way, each subject can establish their own rating scale, thus yielding maximally finegraded data and avoiding the known problems with the conventional ordinal scales for linguistic data (Bard et al., 1996; Cowart, 1997; Schütze, 1996).

	Plaus	Jen$_a$	Conf$_a$	Jen$_n$	Conf$_n$
Jen$_a$	.058				
Conf$_a$	.214*	.941**			
Jen$_n$	.124	.781**	.808**		
Conf$_n$	.232*	.782**	.864**	.956**	
WN	.356**	.222*	.348**	.451**	.444**

*$p < .05$ (2-tailed) **$p < .01$ (2-tailed)

Table 4: Correlation matrix for plausibility and the five smoothed frequency estimates

In the present experiment, subjects were presented with adjective-noun pairs and were asked to rate the degree of adjective-noun fit proportional to a modulus item. The experiment was carried out using WebExp, a set of Java-Classes for administering psycholinguistic studies over the World-Wide Web (Keller et al., 1998). Subjects first saw a set of instructions that explained the ME technique and included some examples, and had to fill in a short questionnaire including basic demographic information. Each subject saw the entire set of 90 experimental items.

Subjects. Forty-one native speakers of English volunteered to participate. Subjects were recruited over the Internet by postings to relevant newsgroups and mailing lists.

3.2 Results

Correlation analysis was used to assess the degree of linear relationship between plausibility ratings (Plaus) and the three smoothed co-occurrence frequency estimates: distance-weighted averaging using Jensen-Shannon divergence (Jen), distance-weighted averaging using confusion probability (Conf), and class-based smoothing using WordNet (WN). For the two similarity-based measures, we smoothed either over the similarity of the adjective (subscript a) or over the similarity of the noun (subscript n). All frequency estimates were natural log-transformed.

Table 4 displays the results of the correlation analysis. Mean plausibility ratings were significantly correlated with co-occurrence frequency recreated using our class-based smoothing method based on WordNet ($r = .356$, $p < .01$).

As detailed in Section 2.2, the Jensen-Shannon divergence and the confusion probability are parameterised measures. There are two ways to smooth the frequency of an adjective-noun combination: over the distribution of adjectives or over the distribution of nouns. We tried both approaches and found a moderate correlation between plausibility and both the frequency recreated using distance-weighted averaging and confusion probability. The correlation was significant both for frequencies recreated by smoothing over adjectives ($r = .214$, $p < .05$) and over nouns ($r = .232$, $p < .05$). However, co-occurrence frequency recreated using the Jensen-Shannon divergence was not reliably correlated with plausibility. Furthermore, there was a reliable correlation between the two Jensen-Shannon measures Jen$_a$ and Jen$_n$ ($r = .781$, $p < .01$), and similarly between the two confusion measures Conf$_a$ and Conf$_n$ ($r = .864$, $p < .01$). We also found a high correlation between Jen$_a$ and Conf$_a$ ($r = .941$, $p < .01$) and Jen$_n$ and Conf$_n$ ($r = .956$, $p < .01$). This indicates that the two similarity measures yield comparable results for the given task.

We also examined the effect of varying one further parameter (see Section 2.2). The recreated frequencies were initially estimated using the $n = 1,000$ most similar items. We examined the effects of applying the two smoothing methods using a set of similar items of twice the size ($n = 2,000$). No improvement in terms of the correlations with rated plausibility was found when using this larger set, whether smoothing over the adjective or the noun: a moderate correlation with plausibility was found for Conf$_a$ ($r = .239$, $p < .05$) and Conf$_n$ ($r = .239$, $p < .05$), while the correlation with Jen$_a$ and Jen$_n$ was not significant.

An important question is how well people agree in their plausibility judgements. Inter-subject agreement gives an upper bound for the task and allows us to interpret how well the smoothing techniques are doing in relation to the human judges. We computed the inter-subject correlation on the elicited judgements using leave-one-out resampling (Weiss and Kulikowski, 1991). Average inter-subject agreement was .55 (Min = .01, Max = .76, SD = .16). This means that our approach performs satisfactorily given that there is a fair amount of variability in human judgements of adjective-noun plausibility.

	WN	Jen$_a$	Conf$_a$	Jen$_n$	Conf$_n$
Actual freq.	.218*	.324**	.646**	.308**	.728**
Plausibility	.349**	.268*	.395**	.247*	.416**

*$p < .05$ (2-tailed) **$p < .01$ (2-tailed)

Table 5: Correlation of recreated frequencies with actual frequencies and plausibility (using Lapata et al.'s (1999) stimuli)

One remaining issue concerns the validity of our smoothing procedures. We have shown that co-occurrence frequencies recreated using smoothing techniques are significantly correlated with rated plausibility. But this finding constitutes only indirect evidence for the ability of this method to recreate corpus evidence; it depends on the assumption that plausibility and frequency are adequate indicators of each other's values. Does smoothing accurately recreate the co-occurrence frequency of combinations that actually do occur in the corpus? To address this question, we applied the class-based smoothing procedure to a set of adjective-noun pairs that occur in the corpus with varying frequencies, using the materials from Lapata et al. (1999).

First, we removed all relevant adjective-noun combinations from the corpus. Effectively we assumed a linguistic environment with no evidence for the occurrence of the pair, and thus no evidence for any linguistic relationship between the adjective and the noun. Then we recreated the co-occurrence frequencies using class-based smoothing and distance-weighted averaging, and log-transformed the resulting frequencies. Both methods yielded reliable correlation between recreated frequency and actual BNC frequency (see Table 5 for details). This result provides additional evidence for the claim that these smoothing techniques produce reliable frequency estimates for unseen adjective-noun pairs. Note that the best correlations were achieved for Conf$_a$ and Conf$_n$ ($r = .646$, $p < .01$ and $r = .728$, $p < .01$, respectively).

Finally, we carried out a further test of the quality of the recreated frequencies by correlating them with the plausibility judgements reported by Lapata et al. (1999). Again, a significant correlation was found for all methods (see Table 5). However, all correlations were lower than the correlation of the actual frequencies with plausibility ($r = .570$, $p < .01$) reported by Lapata et al. (1999). Note also that the confusion probability outperformed Jensen-Shannon divergence, in line with our results on unfamiliar adjective-noun pairs.

3.3 Discussion

Lapata et al. (1999) demonstrated that the co-occurrence frequency of an adjective-noun combination is the best predictor of its rated plausibility. The present experiment extended this result to adjective-noun pairs that do not co-occur in the corpus.

We applied two smoothing techniques in order to recreate co-occurrence frequency and found that the class-based smoothing method was the best predictor of plausibility. This result is inter-

guilty	*dangerous*	*stop*	*giant*
guilty	dangerous	stop	giant
interested	certain	moon	company
innocent	different	employment	manufacturer
injured	particular	length	artist
labour	difficult	detail	industry
socialist	other	page	firm
strange	strange	time	star
democratic	similar	potential	master
ruling	various	list	army
honest	bad	turn	rival

Table 6: The ten most similar words to the adjectives *guilty* and *dangerous* and the nouns *stop* and *giant* discovered by the Jensen-Shannon measure

esting because the class-based method does not use detailed knowledge about word-to-word relationships in real language; instead, it relies on the notion of equivalence classes derived from WordNet, a semantic taxonomy. It appears that making predictions about plausibility is most effectively done by collapsing together the speaker's experience with other words in the semantic class occupied by the target word.

The distance-weighted averaging smoothing methods yielded a lower correlation with plausibility (in the case of the confusion probability), or no correlation at all (in the case of the Jensen-Shannon divergence). The worse performance of distance-weighted averaging is probably due to the fact that this method conflates two kinds of distributional similarity: on the one hand, it generates words that are semantically similar to the target word. On the other hand, it also generates words whose syntactic behaviour is similar to that of the target word. Rated plausibility, however, seems to be more sensitive to semantic than to syntactic similarity.

As an example refer to Table 6, which displays the ten most distributionally similar words to the adjectives *guilty* and *dangerous* and to the nouns *stop* and *giant* discovered by the Jensen-Shannon measure. The set of similar words is far from semantically coherent. As far as the adjective *guilty* is concerned the measure discovered antonyms such as *innocent* and *honest*. Semantically unrelated adjectives such as *injured*, *democratic*, or *interested* are included; it seems that their syntactic behaviour is similar to that of *guilty*, e.g., they all co-occur with *party*. The same pattern can be observed for the adjective *dangerous*, to which none of the discovered adjectives are intuitively semantically related, perhaps with the exception of *bad*. The set of words most similar to the noun *stop* also does not appear to be semantically coherent.

This problem with distance-weighted averaging is aggravated by the fact that the adjective or noun that we smooth over can be polysemous. Take the set of similar words for *giant*, for instance. The words *company*, *manufacturer*, *industry* and *firm* are similar to the 'enterprise' sense of *giant*, whereas *artist*, *star*, *master* are similar to the 'important/influential person' sense of *giant*. However, no similar word was found for either the 'beast' or 'heavyweight person' sense of *giant*. This illustrates that the distance-weighted averaging approach fails to take proper account of the polysemy of a word. The class-based approach, on the other hand, relies on WordNet, a lexical taxonomy that can be expected to cover most senses of a given lexical item.

Recall that distance-weighted averaging discovers distributionally similar words by looking at simple lexical co-occurrence information. In the case of adjective-noun pairs we concentrated on combinations found in the corpus in a head-modifier relationship. This limited form of surface-syntactic information does not seem to be sufficient to reproduce the detailed knowledge that people have about the semantic relationships between words. Our class-based smoothing method, on the other hand, relies on the semantic taxonomy of WordNet, where fine-grained conceptual knowledge about words and their relations is encoded. This knowledge can be used to create semantically coherent equivalence classes. Such classes will not contain antonyms or items whose behaviour is syntactically related, but not semantically similar, to the words of interest.

To summarise, it appears that distance-weighted averaging smoothing is only partially successful in reproducing the linguistic dependencies that characterise and constrain the formation of adjective-noun combinations. The class-based smoothing method, however, relies on a pre-defined taxonomy that allows these dependencies to be inferred, and thus reliably estimates the plausibility of adjective-noun combinations that fail to co-occur in the corpus.

4 Conclusions

This paper investigated the validity of smoothing techniques by using them to recreate the frequencies of adjective-noun pairs that fail to occur in a 100 million word corpus. We showed that the recreated frequencies are significantly correlated with plausibility judgements. These results were then extended by applying the same smoothing techniques to adjective-noun pairs that occur in the corpus. These recreated frequencies were significantly correlated with the actual frequencies, as well as with plausibility judgements.

Our results provide independent evidence for the validity of the smoothing techniques we employed. In contrast to previous work, our evaluation does not presuppose that the recreated frequencies are used in a specific natural language processing task. Rather, we established an independent criterion for the validity of smoothing techniques by comparing them to plausibility judgements, which are known to correlate with co-occurrence frequency. We also carried out a comparison of different smoothing methods, and found that class-based smoothing outperforms distance-weighted averaging.[4]

From a practical point of view, our findings provide a very simple account of adjective-noun plausibility. Extending the results of Lapata et al. (1999), we confirmed that co-occurrence frequency can be used to estimate the plausibility of an adjective-noun pair. If no co-occurrence counts are available from the corpus, then counts can be recreated using the corpus and a structured source of taxonomic knowledge (for the class-based approach). Distance-weighted averaging can be seen as a 'cheap' way to obtain this sort of taxonomic knowledge. However, this method does not draw upon semantic information only, but is also sensitive to the syntactic distribution of the target word. This explains the fact that distance-weighted averaging yielded a lower correlation with perceived plausibility than class-based smoothing. A taxonomy like WordNet provides a cleaner source of conceptual information, which captures essential aspects of the type of knowledge needed for assessing the plausibility of an adjective-noun combination.

References

Ellen Gurman Bard, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language*, 72(1):32–68.

Peter F. Brown, Vincent J. Della Pietra, Peter V. de Souza, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Lou Burnard, 1995. *Users Guide for the British National Corpus*. British National Corpus Consortium, Oxford University Computing Service.

Stephen Clark and David Weir. 2001. Class-based probability estimation using a semantic hierarchy. In *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, PA.

Steffan Corley, Martin Corley, Frank Keller, Matthew W. Crocker, and Shari Trewin. 2001. Finding syntactic structure in unparsed corpora: The Gsearch corpus query system. *Computers and the Humanities*, 35(2):81–94.

Wayne Cowart. 1997. *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Sage Publications, Thousand Oaks, CA.

D. A. Cruse. 1986. *Lexical Semantics*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

Ido Dagan, Lillian Lee, and Fernando Pereira. 1999. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1):43–69.

Ralph Grishman and John Sterling. 1994. Generalizing automatically generated selectional patterns. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 742–747, Kyoto.

Slava M. Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics Speech and Signal Processing*, 33(3):400–401.

Frank Keller, Martin Corley, Steffan Corley, Lars Konieczny, and Amalia Todirascu. 1998. WebExp: A Java toolbox for web-based psychological experiments. Technical Report HCRC/TR-99, Human Communication Research Centre, University of Edinburgh.

Maria Lapata, Scott McDonald, and Frank Keller. 1999. Determinants of adjective-noun plausibility. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pages 30–36, Bergen.

Maria Lapata. 2000. *The Acquisition and Modeling of Lexical Knowledge: A Corpus-based Investigation of Systematic Polysemy*. Ph.D. thesis, University of Edinburgh.

Mark Lauer. 1995. *Designing Statistical Language Learners: Experiments on Compound Nouns*. Ph.D. thesis, Macquarie University, Sydney.

Lilian Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32, University of Maryland, College Park.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190, Columbus, OH.

Philip Stuart Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Carson T. Schütze. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press, Chicago.

Frank Smadja. 1991. Macrocoding the lexicon with co-occurrence knowledge. In Uri Zernik, editor, *Lexical Acquisition: Using Online Resources to Build a Lexicon*, pages 165–189. Lawrence Erlbaum Associates, Hillsdale, NJ.

S. S. Stevens. 1975. *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. John Wiley, New York.

Sholom M. Weiss and Casimir A. Kulikowski. 1991. *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann, San Mateo, CA.

[4] Two anonymous reviewers point out that this conclusion only holds for an approach that computes similarity based on adjective-noun co-occurrences. Such co-occurrences might not reflect semantic relatedness very well, due to the idiosyncratic nature of adjective-noun combinations. It is possible that distance-weighted averaging would yield better results if applied to other co-occurrence data (e.g., subject-verb, verb-object), which could be expected to produce more reliable information about semantic similarity.