The Automatic Interpretation of Nominalizations

The Automatic Interpretation of Nominalizations

Maria Lapata
Division of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, UK
mlap@cogsci.ed.ac.uk

Abstract

This paper discusses the interpretation of nominalizations in domain independent wide-coverage text. We present a statistical model which interprets nominalizations based on the cooccurrence of verb-argument tuples in a large balanced corpus. We propose an algorithm which treats the interpretation task as a disambiguation problem and achieves a performance of approximately 80% by combining partial parsing, smoothing techniques and domain independent taxonomic information (e.g., WordNet).

Introduction

The automatic interpretation of compound nouns has been a long-standing unsolved problem within Natural Language Processing (NLP). Compound nouns in English have three basic properties which pose difficulties for their interpretation: (a) the compounding process is extremely productive, (b) the semantic relationship between the compound head and its modifier is implicit (this means that it cannot be easily recovered from syntactic or morphological analysis), and (c) the interpretation can be influenced by a variety of contextual and pragmatic factors.

To arrive at an interpretation of the compound *onion tears* (e.g., onions CAUSE tears) it is necessary to identify that *tears* is a noun (and not the third person singular of the verb *tear*) and to use semantic information about *onions* and *tears* (for example the fact that onions cannot be tears or that tears are not made of onions). Even in the case of a compound like *government promotion* where the head noun is derived from the verb *promote* and the modifier *government* is its argument, it is necessary to determine whether *government* is the subject or the object. One might argue that the preferred analysis for *government promotion* is "government that is promoted by someone". However, this interpretation can be easily overridden in context as shown in example (1) taken from the British National Corpus: here it is the government that is doing the promotion.

(1) By the end of the 1920s, *government promotion* of agricultural development in Niger was limited, consisting mainly of crop trials and model sheep and ostrich farm.

The interpretation of compound nouns is important for several NLP tasks, notably machine translation. Consider the compound *satellite observation* which may mean *observation by satellite* or *observation of satellites*. In order to translate *satellite observation* into Spanish, we have to work out whether *satellite* is the subject or object of the verb *observe*. In the first case *satellite observation* translates as *observación por satelite* (observation by satellite), whereas in the latter it translates as *observación de satelites* (observation of satellites).

A considerable amount of work within NLP focused on the interpretation of two word compounds whose nouns are related via a basic set of semantic relations (e.g., CAUSE relates *onion tears*, FOR relates *pet spray*). With the exceptions of Wu (1993) and Lauer (1995) who have proposed probabilistic models for the interpretation of compounds, the majority of proposals are symbolic: most algorithms rely on hand-crafted knowledge bases or dictionaries containing detailed semantic information for each noun; a sequence of rules exploit this information in order to choose the correct interpretation for a given compound (Leonard 1984; Vanderwende 1994). Most of the proposals contain no qualitative evaluation. The exceptions are Leonard (1984) who reports an accuracy of 76% (although on the training set), Vanderwende (1994) whose algorithm attains an accuracy of 52%, and Lauer (1995) who reports an accuracy of 47%. The low accuracy is indicative of the difficulty of the task given the variety of contextual and pragmatic factors which can influence the interpretation of a compound.

In this paper, we focus solely on the interpretation of nominalizations, i.e., compounds whose head noun is a nominalized verb and whose prenominal modifier is derived from either the underlying subject or direct object of the verb (Levi 1978) (see the examples in (2)–(3)).

(2) a. SUBJ child behaviour ⇒ *child behaves*
b. OBJ car lover ⇒ *love cars*
c. OBJ soccer competition ⇒ *compete in soccer*

(3) a. SUBJ|OBJ government promotion
b. SUBJ|OBJ satellite observation

The nominalized verb can either take a subject (cf. (2a)), a direct object (cf. (2b)) or a prepositional object (cf. (2c)). In some cases, the relation of the modifier and the nominalized verb (SUBJ or OBJ) can be predicted either from the subcategorization properties of the verb (cf. (2a) where *child* can only be the subject of the intransitive verb *behave*) or from the semantics of the of the nominalization suffix of the head

noun (cf. (2b) where the agentive suffix *-er* of the noun *lover* indicates that the modifier *car* is the object of *love*). In other cases, the relation of the modifier and the head noun is genuinely ambiguous (see (3)).

The interpretation of nominalizations poses a challenge for empirical approaches since the argument relations between a head and its modifier are not readily available in the corpus. We present a probabilistic algorithm which treats the interpretation task as a disambiguation problem, and show how the severe sparse data problem in this task can be overcome by combining partial parsing, smoothing techniques, and domain independent taxonomic information (e.g., WordNet). We report on the results of five experiments which achieve a combined precision of approximately 80% on the British National Corpus (BNC), a 100 million word collection of samples of written and spoken language from a wide range of sources designed to represent a wide cross-section of current British English, both spoken and written (Burnard 1995).

The model

Given a nominalization, our goal is to develop a procedure to infer whether the modifier stands in a subject or object relation to the head noun. In other words, we need to assign probabilities to the two different relations (SUBJ, OBJ). For each relation *rel* we calculate the simple expression $P(rel|n_1,n_2)$ given in (4) below.

$$P(rel|n_1,n_2) = \frac{f(n_1,rel,n_2)}{f(n_1,n_2)} \tag{4}$$

Since we have a choice between two outcomes we will use a likelihood ratio to compare the two relation probabilities (Mosteller & Wallace 1964). In particular we will compute the log of the ratio of the probability $P(\text{OBJ}|n_1,n_2)$ to the probability $P(\text{SUBJ}|n_1,n_2)$. We will call this log-likelihood ratio the argument relation (*RA*) score.

$$RA(rel,n_1,n_2) = \log_2 \frac{P(\text{OBJ}|n_1,n_2)}{P(\text{SUBJ}|n_1,n_2)} \tag{5}$$

Notice, however, that we cannot read off $f(n_1,rel,n_2)$ directly from the corpus. What we can obtain from a corpus (through parsing) is the number of times a noun is the object or the subject of a given verb. By making the simplifying assumption that the relation between the nominalized head and its modifier noun is the same as the relation between the latter and the verb from which the head is derived, (4) can be rewritten as follows:

$$P(rel|n_1,n_2) \approx \frac{f(v_{n_2},rel,n_1)}{\sum_i f(v_{n_2},rel_i,n_1)} \tag{6}$$

where $f(v_{n_2},rel,n_1)$ is the frequency with which the modifier noun n_1 is found in the corpus as the subject or object of v_{n_2}, the verb from which the head noun is derived. The sum $\sum_i f(v_{n_2},rel_i,n_1)$ is a normalization factor.

Parameter estimation

Verb-argument tuples

A part-of-speech tagged and lemmatized version of the BNC (100 million words) was automatically parsed by Cass (Abney 1996). Cass is a robust chunk parser designed for the shallow analysis of noisy text. We used the parser's built-in function to extract verb-subject and verb-object tuples. The tuples obtained from the parser's output are an imperfect source of information about argument relations. For example, the tuples extractor mistakes adjectives for verbs (cf. (7a)) and nouns for verbs (cf. (7b)).

(7) a. SUBJ isolated people
b. SUBJ behalf whose

In order to compile a comprehensive count of verb-argument relations we discarded tuples containing verbs or nouns with a BNC frequency of one. This resulted in 588,333 distinct types of verb-subject pairs and 615,328 distinct types of verb-object pairs.

The data

It is beyond the scope of the present study to develop an algorithm which automatically detects nominalizations in a corpus. In the experiments described in the subsequent sections compounds with deverbal heads were obtained as follows:

1. Two word compound nouns were extracted from the BNC by using a heuristic which looks for consecutive pairs of nouns which are neither preceded nor succeeded by a noun (Lauer 1995).
2. A dictionary of deverbal nouns was created using: (a) Nomlex (Macleod *et al.* 1998), a dictionary of nominalizations containing 827 lexical entries and (b) Celex (Burnage 1990), a general morphological dictionary, which contains 5,111 nominalizations;
3. Candidate nominalizations were obtained from the compounds acquired from the BNC by selecting noun-noun sequences whose head (i.e., rightmost noun) was one of the deverbal nouns contained either in Celex or Nomlex. The procedure resulted in 172,797 potential nominalizations.

From these candidate nominalizations a random sample of 2,000 tokens was selected. The sample was manually inspected and compounds with modifiers whose relation to the head noun was other than subject or object were discarded. Nominalizations whose heads were derived from verbs taking prepositional objects (cf. (2c)) were also discarded. After manual inspection the sample contained 796 nominalizations. From these, 596 tokens were used as training data for the experiments described in the following sections. The remaining 200 nominalizations were used as test data and also to evaluate whether humans can reliably disambiguate the argument relation between the nominalized head and its modifier.

Agreement

Two judges decided whether the modifier is the subject or object of a nominalized head. The nominalizations were disambiguated in context: the judges were given the corpus sentence in which the nominalization occurred together with the preceding and following sentence. The judges were given a page of guidelines but no prior training We measured agreement using the Kappa coefficient (Siegel & Castellan 1988), which is the ratio of the proportion of times, $P(A)$, that k raters agree to the proportion of times, $P(E)$, that we would

expect the raters to agree by chance (cf. (8)). If there is a complete agreement among the raters, then $K = 1$, whereas if there is no agreement among the raters (other than the agreement which would be expected to occur by chance), then $K = 0$.

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{8}$$

The judges' agreement on the disambiguation task was $K = .78$ ($N = 200$, $k = 2$). The agreement was good given that the judges were given minimal instructions and no prior training. However, note that despite the fact that context was provided to aid the disambiguation task, the annotators were not in complete agreement. This points to the intrinsic difficulty of the task. Argument relations and consequently selectional restrictions are influenced by several pragmatic factors which may not be readily inferred from the immediate context. In the following we propose a method which faces a greater challenge: the interpretation of nominalizations without taking context into account.

Mapping

In order to estimate the frequency $f(v_{n_2}, rel, n_1)$ and consequently the probability $P(rel|n_1, n_2)$, the nominalized heads were mapped to their corresponding verbs. Inspection of the frequencies of the verb-argument tuples contained in the sample (596 tokens) revealed that 372 verb-noun pairs had a verb-object frequency of zero in the corpus. Similarly, 378 verb-noun pairs had a verb-subject frequency of zero. Furthermore, a total of 287 tuples were not attested at all in the BNC either in a verb-object or verb-subject relation. This finding is perhaps not surprising given the productivity of compounds. Considering the ease with which novel compounds are created it is to be expected that some verb-argument configurations will not occur in the training corpus.

We estimated the frequencies of unseen verb-argument pairs by experimenting with three types of smoothing techniques proposed in the literature: back-off smoothing (Katz 1987), class-based smoothing (Resnik 1993) and similarity-based smoothing (Dagan, Lee, & Pereira 1999).

Smoothing

Back-off smoothing

Back-off n-gram models were initially proposed by Katz (1987) for speech recognition but have been also successfully used to disambiguate the attachment site of structurally ambiguous PPs (Collins & Brooks 1995). The main idea behind back-off smoothing is to adjust maximum likelihood estimates like (6) so that the total probability of observed word co-occurrences is less than one, leaving some probability mass to be redistributed among unseen co-occurrences. In general the frequency of observed word sequences is discounted using Good Turing's estimate and the probability of unseen sequences is estimated by using lower level conditional distributions. Assuming that the denominator in (6) $f(v_{n_2}, rel, n_1)$ is zero we can approximate $P(rel|n_1, n_2)$ by backing-off to $P(rel|n_1)$:

$$P(rel|n_1, n_2) = \alpha \frac{f(rel, n_1)}{f(n_1)} \tag{9}$$

where α is a normalization constant which ensures that the probabilities sum to one. If the frequency $f(rel, n_1)$ is also zero backing-off continues by making use of $P(rel)$.

Class-based smoothing

Class-based smoothing recreates co-occurrence frequencies based on information provided by taxonomies such as WordNet or Roget's thesaurus. Taxonomic information can be used to estimate the frequencies $f(v_{n_2}, rel, n_1)$ by substituting the word n_1 occurring in an argument position by the concept with which it is represented in the taxonomy (Resnik 1993). Hence, $f(v_{n_2}, rel, n_1)$ can be estimated by counting the number of times the concept corresponding to n_1 was observed as the argument of the verb v_{n_2} in the corpus.

This would be a straightforward task if each word was always represented in the taxonomy by a single concept or if we had a corpus of verb-argument tuples labeled explicitly with taxonomic information. Lacking such a corpus we need to take into consideration the fact that words in a taxonomy may belong to more than one conceptual classes: counts of verb-argument configurations are reconstructed for each conceptual class by dividing the contribution from the argument by the number of classes it belongs to (Resnik 1993; Lauer 1995):

$$f(v_{n_2}, rel, c) \approx \sum_{n_1' \in c} \frac{count(v_{n_2}, rel, n_1')}{|classes(n_1')|} \tag{10}$$

where $count(v_{n_2}, rel, n_1')$ is the number of times the verb v_{n_2} was observed with noun $n_1' \in c$ bearing the argument relation *rel* (i.e., subject or object) and $|classes(n_1')|$ is the number of conceptual classes n_1' belongs to. The frequency $f(v_{n_2}, rel, c)$ is reconstructed for all classes c with which the argument n_1 is represented in the taxonomy. Since we do not know which is the actual class of the noun n_1 in the corpus we weigh the contribution of each class by taking the average of the reconstructed frequencies for all classes c:

$$f(v_{n_2}, rel, n_1) = \frac{\sum_{c \in classes(n_1)} \sum_{n_1' \in c} \frac{count(v_{n_2}, rel, n_1')}{|classes(n_1')|}}{|classes(n_1)|} \tag{11}$$

Similarity-based smoothing

Similarity-based smoothing is based on the assumption that if a word w_1' is "similar" to word w_1, then w_1' can provide information about the frequency of unseen word pairs involving w_1 (Dagan, Lee, & Pereira 1999). There are several measures of word similarity which can be derived from lexical co-occurrences, providing an alternative to taxonomies such as WordNet (see Dagan, Lee, & Pereira (1999) for an overview).

We have experimented with two measures of distributional similarity derived from co-occurrence frequencies: the Jensen-Shannon divergence and the confusion probability. The choice of these two measures was motivated by work described in Dagan, Lee, & Pereira (1999) where the

Jensen-Shannon divergence outperforms related similarity measures on a word sense disambiguation task which uses verb-object pairs. The confusion probability has been used by several authors to smooth word co-occurrence probabilities (e.g., Grishman & Sterling 1994). In the following we describe these two similarity measures and show how they can be used to recreate the frequencies for unseen verb-argument tuples (for a more detailed description see Dagan, Lee, & Pereira 1999).

Confusion Probability The confusion probability P_C 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 contexts.

$$P_C(w_1|w_1') = \sum_s P(w_1|s)P(s|w_1') \tag{12}$$

where $P_C(w_1'|w_1)$ is the probability that word w_1' occurs in the same contexts s as word w_1, averaged over these contexts. Given a tuple of the form w_1, rel, w_2 we chose to treat rel, w_2 as context and smooth over the verb w_1. By taking verb-argument tuples into consideration (12) is rewritten as follows:

$$\begin{aligned} P_C(w_1|w_1') &= \sum_{rel,w_2} P(w_1|rel,w_2)P(rel,w_2|w_1') \\ &= \sum_{rel,w_2} \frac{f(w_1,rel,w_2)}{f(rel,w_2)} \frac{f(w_1',rel,w_2)}{f(w_1')} \end{aligned} \tag{13}$$

The confusion probability can be computed efficiently as it involves summation only over the common contexts rel, w_2.

Jensen-Shannon divergence The Jensen-Shannon divergence J is a measure of the "distance" between distributions:

$$J(w_1,w_1') = \frac{1}{2}\left[D\left(w_1 \middle\| \frac{w_1+w_1'}{2}\right) + D\left(w_1' \middle\| \frac{w_1+w_1'}{2}\right)\right] \tag{14}$$

$$D(w_1\|w_1') = \sum_{rel,w_2} P(rel,w_2|w_1)\log\frac{P(rel,w_2|w_1)}{P(rel,w_2|w_1')} \tag{15}$$

where D in (14) is the Kullback-Leibler divergence, a measure of the dissimilarity between two probability distributions (cf. equation (15)) and $(w_1+w_1')/2$ is a shorthand for the average distribution:

$$\frac{1}{2}(P(rel,w_2|w_1)+P(rel,w_2|w_1')) \tag{16}$$

Similarly to the confusion probability, the computation of J depends only on the common contexts rel, w_2. Dagan, Lee, & Pereira (1999) provide for the J divergence a weight function $W_J(w,w_1')$:

$$W_J(w_1,w_1') = 10^{-\beta J(w_1,w_1')} \tag{17}$$

The parameter β controls the relative influence of the neighbors (i.e., distributionally similar words) closest to w_1: if β is high, only words extremely close to w_1 contribute to the estimate, whereas if β is low distant words also contribute to the estimate.

We estimate the frequency of an unseen verb-argument tuple by taking into account the similar w_1s and the contexts in which they occur (Grishman & Sterling 1994):

$$f_s(w_1,rel,w_2) = \sum_{w_1'} \mathrm{sim}(w_1,w_1')f(w_1',rel,w_2) \tag{18}$$

Given a nominalization n_1 n_2:

1. map the head noun n_2 to the verb v_{n_2} from which it is derived;
2. retrieve $f(verb_{n_2}, \mathrm{OBJ}, n_1)$ and $f(verb_{n_2}, \mathrm{SUBJ}, n_1)$ from the corpus;
3. **if** $f(verb_{n_2}, \mathrm{OBJ}, n_1) < k$ **then** recreate $f_s(verb_{n_2}, \mathrm{OBJ}, n_1)$;
4. **if** $f(verb_{n_2}, \mathrm{SUBJ}, n_1) < k$ **then** recreate $f_s(verb_{n_2}, \mathrm{SUBJ}, n_1)$;
5. calculate probabilities $P(\mathrm{OBJ}|n_1,n_2)$ and $P(\mathrm{SUBJ}|n_1,n_2)$;
6. compute $RA(rel,n_1,n_2)$;
7. **if** $RA \geq j$ **then** n_1 is the subject of n_2;
8. **else** n_1 is the object of n_2;

Figure 1: Disambiguation algorithm for nominalizations

where $\mathrm{sim}(w_1,w_1')$ is a function of the similarity between w_1 and w_1'. In our experiments $\mathrm{sim}(w_1,w_1')$ was substituted by the confusion probability $P_C(w_1|w_1')$ and the Jensen-Shannon divergence $W_J(w_1,w_1{}')$.

The algorithm

The disambiguation algorithm for nominalizations is summarized in Figure 1. The algorithm uses verb-argument tuples in order to infer the relation holding between the modifier and its nominalized head. When the co-occurrence frequency for the verb-argument relation is zero, verb-argument tuples are smoothed. The sign of the *RA* score (cf. equation (5) and steps 6–8) indicates the relation between the head n_1 and its modifier n_2: a positive *RA* score indicates an object relation, whereas a negative score indicates a subject relation. Depending on the task and the data at hand we can require that an object or subject analysis is preferred only if *RA* exceeds a certain threshold j (see steps 7 and 8 in Figure 1). We can also impose a threshold k on the type of verb-argument tuples we smooth. If for instance we know that the parser's output is noisy, then we might choose to smooth not only unseen verb-argument pairs but also pairs with attested frequencies in the corpus (e.g., $f(verb_{n_2},rel,n_1) \geq 1$, see steps 3 and 4 in Figure 1).

Experiments

The task

The algorithm was trained on 596 nominalizations and tested on 200. The 596 nominalizations were also used as training data for finding the optimal parameters for the two parameterized similarity-based smoothing approaches. In particular we examined whether the size of the vocabulary (e.g., number of verbs used to find the nearest neighbors) has an impact on disambiguation performance and what the best value for the parameter β is. As far as class-based smoothing is concerned we experimented with two concept hierarchies, Roget's thesaurus and WordNet. Although the class-based and back-off methods are not parameterized, we report their performance both on training and test set for completeness.

The algorithm's output was compared to the manual classification and precision was computed accordingly. For 59%

Method	$\text{Accuracy}_{\text{train}}$	$\text{Accuracy}_{\text{test}}$
Default	59.0%	61.8%
Back-off	63.0%	67.0%
Confusion	68.3%	73.7%
Jensen	67.9%	67.0%
WordNet	67.9%	70.6%
Roget	64.6%	66.5%
Ripper	79.7%	78.3%

Table 1: Disambiguation performance

of the nominalizations contained in the train data the modifier was the object of the deverbal head, whereas in the remaining 41% the modifier was the subject. This means that a simple heuristic which defaults to an object relation yields a precision of approximately 59%. Our decision procedure defaults to an object relation when there is no evidence to support either analysis (e.g., when $f(v_{n_2}, \text{OBJ}, n_1) = f(v_{n_2}, \text{SUBJ}, n_1)$).

Results

Before reporting the results of the disambiguation task, we describe our experiments on finding the optimal parameter settings for the two similarity-based smoothing methods.

Figure 2a shows how performance on the disambiguation task varies with respect to the number and frequency of verbs over which the similarity function is calculated. The y-axis in Figure 2a shows how performance on the training set varies (for both P_C and J) when verb-argument pairs are selected from the 1,000 most frequent verbs in the corpus, the 2,000 most frequent verbs in the corpus, etc. (x-axis). The best performance for both similarity functions is achieved using the 2,000 most frequent verbs. Furthermore, performance between J and P_C is comparable (67.9% and 68.3%, respectively). Another important observation is that performance deteriorates less severely for P_C than for J as the number of verbs increases: when all verbs for which verb-argument tuples are extracted from the BNC are used precision for P_C is 66.94%, whereas precision for J is 62.75%. These results are perhaps unsurprising: verb-argument pairs with low-frequency verbs introduce noise due to the errors inherent in the partial parser.

Finally, we analyzed the role of the parameter β. Recall that β appears in the weight function for J and controls the influence of the most similar words. Figure 2b shows how the value of β affects performance on the disambiguation task when the similarity function is computed for the 1,000 and 2,000 most frequent verbs in the corpus. It is clear that performance is low with very high or very low β values (e.g., $\beta \in \{2,9\}$). We chose to set the parameter β to 5 and the results shown in Figure 2a have been produced for this value for all verb frequency classes.

Table 1 shows how the three types of smoothing, back-off, class-based, and similarity-based, influence performance in predicting the relation between a modifier and its nominalized head. For the similarity-based methods we report the results obtained with the optimal parameter settings (β = 5; 2,000 most frequent verbs). Let us concentrate on the

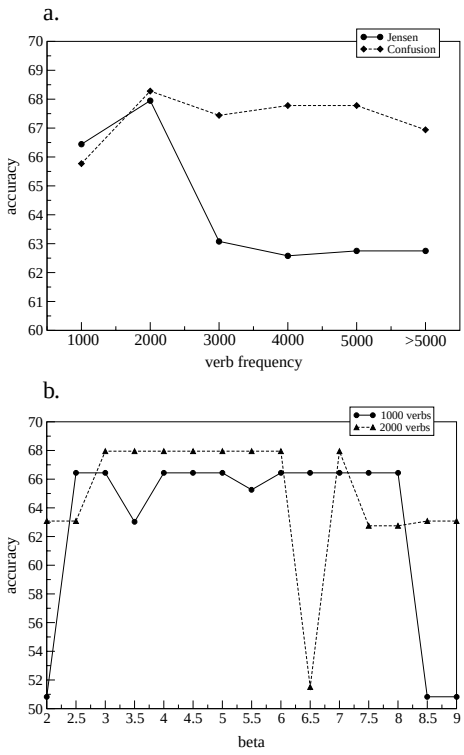

Figure 2: Parameter settings for P_C and J

	Back-off	Jensen	Confusion	WordNet
Jensen	.31			
Confusion	.26	.53		
WordNet	.01	.37	.75	
Roget	.25	.26	.49	0.46

Table 2: Agreement between smoothing methods

training set first. The back-off method is outperformed by all other methods, although its performance is comparable to class-based smoothing using Roget's thesaurus (63% and 64.6%, respectively). Similarity-based methods outperform concept-based methods, although not considerably (accuracy on the training set was 68.3% for P_C and 67.9% for class-based smoothing using WordNet). Furthermore, the particular concept hierarchy used for class-based smoothing seems to have an effect on disambiguation performance: an increase of approximately 2% is obtained by using WordNet instead of Roget's thesaurus. One explanation might be that Roget's thesaurus is too coarse-grained a taxonomy for the task at hand (Roget's taxonomy contains 1,043 concepts, whereas WordNet contains 4,795). We used a χ^2 test to examine whether the the observed performance is better than the simple strategy of always choosing an object relation which yields an accuracy of 59%. The proportion of nominalizations classified correctly was significantly greater than 59% ($p < 0.01$) for all methods but back-off and Roget.

Similar results were observed on the test set. Again P_C outperforms all other methods achieving a precision of 73.7% (see Table 1). The portion of nominalization classified correctly by P_C was significantly greater than 61.8%

Method	SUBJ	OBJ
Back-off	41.6%	78.0%
Jensen	34.7%	91.2%
Confusion	47.3%	82.9%
WordNet	47.8%	80.3%
Roget	50.6%	74.4%

Table 3: Performance on predicting argument relations

($p < 0.05$) which was the percentage of object relations in the test set. The second best method is class-based smoothing using WordNet (see Table 1). The back-off method performs as well as J, reaching an accuracy of 67%.

An interesting question is the extent to which any of the different methods agree in their assignments of subject and object relations. We investigated this by calculating the methods' agreement on the training set using the Kappa coefficient. We calculated the Kappa coefficient for all six pairwise combinations of the five smoothing variants. The results are reported in Table 2. The highest agreement is observed for P_C and the class-based smoothing using the WordNet taxonomy ($K = .75$). This finding suggests that methods inducing similarity relationships from corpus co-occurrence statistics are not necessarily incompatible with methods which quantify similarity using manually crafted taxonomies. Agreement between J and P_C as well as agreement between WordNet and Roget's thesaurus was rather low ($K = .53$ and $K = .46$, respectively). This suggests that different similarity functions or taxonomies may be appropriate for different tasks.

Table 3 shows how the different methods compare for the task of predicting the individual relations for the training set. A general observation is that all methods are fairly good at predicting object relations. Predicting subject relations is considerably harder: no method exceeds an accuracy of approximately 50%. One explanation for this is that selectional constraints imposed on subjects can be more easily overridden by pragmatic and contextual factors than those imposed on objects. J is particularly good at predicting object relations, whereas P_C and class-based smoothing using WordNet seem to yield comparable performances when it comes to predicting subject relations (see Table 3).

An obvious question is whether the precision is increased when combining the five smoothing variants given that they seem to provide complementary information for predicting argument relations. For example, Roget's thesaurus is best for the prediction of subject relations, whereas J is best for the prediction of object relations. We combined the five information sources using a decision tree classifier (Ripper, Cohen 1996). The decision tree was trained on the 596 nominalizations on which the smoothing methods were compared and tested on the 200 unseen nominalizations for which the inter-judge agreement was previously calculated. The average error rate of the decision tree learner was 20.30% $\pm$ 1.65% on the training set and 21.65% $\pm$ 2.96% on the test set. The latter result translates into a precision of 78.3% (cf. Table 1) which is significantly better ($p < 0.01$) than 61.8%, the percentage of object relations in the test set.

Conclusions

The work reported here is an attempt to provide a statistical model of nominalizations occurring in wide coverage text. We showed that a simple algorithm which combines information about the distributional properties of words and domain independent symbolic knowledge (i.e., WordNet) achieves high performance on unseen data. This is an important result considering the simplifications in the system and the sparse data problems encountered in estimating the probability $P(rel|n_1, n_2)$. Finally, we explored the merits and limitations of various smoothing methods and systematically showed how recreated frequencies can be used in a task other than language modeling to produce interesting results.

References

Abney, S. 1996. Partial parsing via finite-state cascades. In Carroll, J., ed., *Workshop on Robust Parsing*, 8–15. Prague: ESSLLI.

Burnage, G. 1990. Celex – a guide for users. Technical report, Centre for Lexical Information, University of Nijmegen.

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

Cohen, W. W. 1996. Learning trees and rules with set-valued features. In *Proceedings of 13th National Conference on Artificial Intelligence*, 709–716. Portland, Oregon: AAAI Press.

1994. *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto: COLING.

Collins, M., and Brooks, J. 1995. Prepositional phrase attachment through a backed-off model. In *Proceedings of the 3rd Workshop on Very Large Corpora*, 27–38. Cambridge, MA: ACL.

Dagan, I.; Lee, L.; and Pereira, F. C. N. 1999. Similarity-based models of word cooccurrence probabilities. *Machine Learning* 1–3(34):43–69.

Grishman, R., and Sterling, J. 1994. Generalizing automatically generated selectional patterns. In COLING (1994), 742–747.

Katz, S. M. 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.

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

Leonard, R. 1984. *The Interpretation of English Noun Sequences on the Computer*. Amsterdam: North-Holland.

Levi, J. N. 1978. *The Syntax and Semantics of Complex Nominals*. New York: Academic Press.

Macleod, C.; Grishman, R.; Meyers, A.; Barrett, L.; and Reeves, R. 1998. Nomlex: A lexicon of nominalizations. In *Proceedings of the 8th International Congress of the European Association for Lexicography*, 187–193. Liège, Belgium: EURALEX.

Mosteller, F., and Wallace, D. L. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley.

Resnik, P. S. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. Dissertation, University of Pennsylvania, Philadelphia, Philadelphia.

Siegel, S., and Castellan, N. J. 1988. *Non Parametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

Vanderwende, L. 1994. Algorithm for automatic interpretation of noun sequences. In COLING (1994), 782–788.

Wu, D. 1993. Approximating maximum-entropy ratings for evidential parsing and semantic interpretation. In *Proceedings of 13th International Joint Conference on Artificial Intelligence*, 1290–1296. Chamberry, France: Morgan Kaufman.