Detecting Novel Compounds: The Role of Distributional Evidence

Mirella Lapata
Department of Computer Science
University of Sheffield
Regent Court, 211 Portobello Street
Sheffield S1 4DP, UK
mlap@dcs.shef.ac.uk

Alex Lascarides
School of Informatics
The University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
alex@inf.ed.ac.uk

Abstract

Research on the discovery of terms from corpora has focused on word sequences whose recurrent occurrence in a corpus is indicative of their terminological status, and has not addressed the issue of discovering terms when data is sparse. This becomes apparent in the case of noun compounding, which is extremely productive: more than half of the candidate compounds extracted from a corpus are attested only once. We show how evidence about established (i.e., frequent) compounds can be used to estimate features that can discriminate rare valid compounds from rare nonce terms in addition to a variety of linguistic features than can be easily gleaned from corpora without relying on parsed text.

1 Introduction

The nature and properties of compounds have been studied at length in the theoretical linguistics literature. It is a well-known fact that compound noun formation in English is relatively productive (see (1)). Although compounds are typically binary (see (1a,b)), they can be also longer than two words (see (1e)). Compounds are commonly written as a concatenation of words (see (1a,b)), or as single words (see (1c)), sometimes a hyphen is also used (see (1e)).

(1) a. income tax
b. AT & T headquarters
c. bathroom
d. public-relations
e. income-tax relief

The use of noun compounds is frequent not only in technical writing and newswire text (McDonald, 1982) but also in fictional prose (Leonard, 1984), and spoken language (Liberman and Sproat, 1992). Novel compounds are used as a text compression device (Marsh, 1984), i.e., to pack meaning into a minimal amount of linguistic structure, as a deictic device, or as a means to classify an entity which has no specific name (Downing, 1977).

Computational investigations of compound nouns have concentrated on their automatic acquisition from corpora, syntactic disambiguation (i.e., determine the structure of compounds like *income tax relief*), and semantic interpretation (i.e., determine the semantic relation between *income* and *tax* in *income tax*). The acquisition of compound nouns is usually subsumed under the general discovery of terms from corpora. Terms are typically acquired by either symbolic or statistical means. Under a symbolic approach, candidate terms are extracted from the corpus using surface syntactic analysis (Lauer, 1995; Justeson and Katz, 1995; Bourigault and Jacquemin, 1999) and sometimes are further submitted to experts for manual inspection. The approach typically assumes no prior terminological knowledge, although Jacquemin (1996) proposed the detection of terminological variants in a corpus by making use of lists of existing terms.

The main assumption underlying the statistical approach to term acquisition is that lexically associated words tend to appear together more often than expected on the basis of their individual occurrence frequencies. Once candidate terms are detected in the corpus, statistical tests (e.g., mutual information, the log-likelihood ratio) are used to determine which co-occurrences are valid terms (see Daille, 1996 and Manning and Schütze, 1999 for overviews).

Most of the statistical tests proposed in the literature rely on the fact that candidate terms will occur frequently in the corpus (Justeson and Katz, 1995) or, when hypothesis testing is applied, on the assumption that two words form a term when they co-occur more often than chance (Church and Hanks, 1990). This means that statistical tests cannot be applied reliably for candidate compounds

CoocF	BNC	Sample	Acc
> 4	52,832	800	93.5
> 1	160,214	800	82.0
≥ 1	510,673	800	71.0
$= 1$	350,459	800	57.7

Table 1: Relation of noun co-occurrence frequency with accuracy

with co-occurrence frequency of one and cannot be used to distinguish rare but valid noun compounds from rare but nonce noun sequences (compare (2b) and (2a) which are extracted from the British National Corpus; both bracketed terms were found in the corpus once.).

(2) a. Although no one will doubt their possibilities for elegance and robustness, sitting on a solid [wood$_N$ seat$_N$] can test the limits of comfort after quite a short time and woven seats are little better.

b. The use of the [term$_N$ shilling$_N$] derives from a 19th century system of invoicing beer according to its gravity.

In this paper we present a method that attempts to distinguish compounds from non-compounds in cases where very little direct evidence is found in the corpus and therefore the assumptions underlying lexical association scores do not hold. We restrict our attention to compounds formed by a concatenation of two nouns (see (1a)) and investigate how surface syntactic and semantic cues can be used to discriminate valid compounds from rare nonce terms.

2 Compound Noun Extraction

The extraction of two word compounds (as opposed to terms) from a corpus has been previously addressed by Lauer (1995) who proposed a heuristic which simply looks for consecutive pairs of nouns which are neither preceded nor succeeded by a noun (see (3)).

(3) $C = \{(w_2, w_3) \mid w_1\ w_2\ w_3\ w_4; w_1, w_4 \notin N; w_2, w_3 \in N\}$

Here, $w_1\ w_2\ w_3\ w_4$ denotes the occurrence of a sequence of four words in the corpus and N is a predefined set of unambiguous nouns. Lauer (1995) used a corpus derived from the Grolier Multimedia Encyclopedia (8M words) for his study and a predefined list of 90,000 nouns which had no part-of-speech ambiguity. He reports an accuracy of 97.9% on a sample of 1,068 noun-noun sequences. Note that the above heuristic incorrectly classifies (2b) as a valid compound.

We replicated Lauer's (1995) study 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 (Burnard, 1995). An important difference, however, between our study and Lauer's is that we used a POS-tagged version of the BNC. Noun sequences were identified 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. Gsearch was run on a lemmatised version of the BNC in order to compile a comprehensive count of all nouns occurring in a head-modifier relationship. Tokens containing noun sequences of length two were classified as candidate compounds unless: (a) the two consecutive nouns were preceded or succeeded by a noun (e.g., *light bulb phobia*, see (3)) and (b) either noun was a number (e.g., *flour 100g*). This procedure resulted in a total of 1,624,915 tokens consisting of 510,673 distinct types of candidate compounds.

We evaluated Lauer's (1995) heuristic as follows: 800 tokens were randomly selected from the noun-noun sequences that were classified as compounds; accordingly, a random sample of 800 tokens was selected from the sequences that were discarded as non-compounds (in order to examine whether valid compounds are missed). The noun sequences contained in the samples were manually inspected within context using the corpus concordance tool Xkwic (Christ, 1995) and classified as to whether they formed a valid compound or not. Lauer's heuristic expectedly achieved a lower accuracy on the POS-tagged corpus. This was 71% using CLAWS4 (Leech et al., 1994), a probabilistic part-of-speech tagger, with error rate ranging from 3% to 4% and 70.3% using Elworthy's (1994) HMM part-of-speech tagger, with an error rate of approximately 4%. The heuristic reached an accuracy of 98.8% in rejecting noun sequences as non-compounds.

We further examined how the accuracy of the heuristic varies when different thresholds are imposed on the frequency of the candidate compounds (see Table 1). For example, when we consider noun-noun sequences that appear in the BNC more than once (CoocF $>$ 1) the heuristic's accuracy is increased by 11.0%. However, the number of potential compounds is reduced by a factor of three. The majority of the candidate compounds extracted from the corpus are hapaxes (i.e., words that occurred only once). These represent 68.6% of the noun-noun sequences retrieved from the BNC; 57.7% of the hapaxes are valid compounds. Analysis of the misclassifications in the case of hapaxes revealed that 61.9% are tagging errors

	$f(n_1)$	$f(n_2)$	$P(\mathrm{H}\|n_1)$	$P(\mathrm{M}\|n_2)$	$f(c_1,c_2)$
cocaine customer	71	159	1	.18	285.85
baby calf	740	22	.91	.15	35.13
people excitement	1,823	9	.45	1	4.98
may push	0	35	0	.43	76.93

Table 2: Feature values for noun-noun sequences (with CoocF = 1)

(i.e., if tagging was perfect these sequences would have been excluded), 30.6% are due to the absence of structural information (i.e., they would have been ruled out if accurate parsing information was available), 5.30% are acronyms, and 2.20% are foreign terms or typographical mistakes.

In the next sections we turn to hapaxes and propose a method that distinguishes valid compounds from nonce noun sequences by modeling the distributional tendencies observed in lexicalized (i.e., frequent) compounds. In Section 3 we present and motivate these features. Section 4 details our machine learning experiments and Section 5 discusses our results.

3 Features for Discovering Compounds

In this section we introduce the features used in the machine learning experiments described in Section 4 and the motivation behind their selection. In our experiments we make use of numeric features (i.e., frequency, probability) as well as categorical features (i.e., the context surrounding candidate noun-noun sequence). All the numeric features detailed below were estimated from a corpus consisting of noun-noun sequences extracted from the POS-tagged BNC (via Lauer's 1995 heuristic) with CoocF greater than four (52,832 in total, see Table 1). 93.5% of these sequences are valid compounds and can therefore provide reliable information about the likelihood of a given noun as a compound head or modifier.

Noun frequency. Given a noun-noun sequence $n_1 n_2$ we look at whether the frequency of the head n_2, $f(n_2)$, or the frequency of the modifier n_1, $f(n_1)$, are reliable indicators for distinguishing compounds from non-compounds. Consider for example the compound *cocaine customer* which is attested in the BNC only once. The word *cocaine* is attested as a modifier 71 times and the word *customer* is attested as a head 159 (see Table 2). Compare now *cocaine customer* to *people excitement* which is not a valid compound and is also found in the BNC once (the sequence is attested in the sentence *For some people excitement is only possible outside marriage.*). The modifier frequency $f(\mathit{people})$ is 1,823 whereas the head frequency $f(\mathit{excitement})$ is nine. Clearly, *excitement* is less likely to be a compound head when compared to *customer* (see Table 2).

Probability. Given a noun-noun sequence $n_1 n_2$ we investigate whether it is likely for n_2 to be a head and for n_1 to be a modifier. We express these quantities as follows:

$$P(\mathrm{M}|n_2) = \frac{f(\mathrm{M},n_2)}{f(n_2)} \tag{4}$$

$$P(\mathrm{H}|n_1) = \frac{f(n_1,\mathrm{H})}{f(n_1)} \tag{5}$$

Here, $f(\mathrm{M},n_2) = \sum_{n_1} f(n_1,n_2)$ and $f(n_1,\mathrm{H}) = \sum_{n_2} f(n_1,n_2)$. Equation (4) expresses the likelihood of n_2 as a head (preceded by any noun modifier) and equation (5) expresses the likelihood of n_1 as a modifier (followed by any noun head). We estimate $f(\mathrm{M},n_2)$ and $f(n_1,\mathrm{H})$ from the reliable noun-noun sequences attested previously in the corpus (CoocF > 4). The frequencies $f(n_1)$ and $f(n_2)$ are the number of times we see n_1 and n_2 in our estimation corpus independently of their position (i.e., independently of whether they are heads or modifiers).

Consider the compounds *cocaine customer* and *baby calf* in Table 2. The likelihood of the words *cocaine* and *baby* to be found in a modifier position is very high (1 and .91, respectively). Contrast this with the sequence *may push* which is the result of a tagging mistake (i.e., both *may* and *push* are annotated as nouns in the sentence *Their different responsibilities in relation to the public may push them in opposite directions*): the likelihood of the word *may* to be found in a modifier position is zero. Note further that *push* can be a noun (denoting the act of pushing) and therefore it is not entirely unlikely to be found in a head position (see Table 2). Note also that the fact that *may push* is classified as a potential compound indicates that the preceding word *public* was mistagged as well.

Concept frequency. Linguistic models of compound noun formation typically involve a hierarchical structure of lexical rules, which capture the regularities of compound noun formation while

$\langle c_1, c_2 \rangle$	$f(c_1, c_2)$	Examples
$\langle$`substance, object`$\rangle$	604.7	iron table
$\langle$`act, social group`$\rangle$	403.0	mining family
$\langle$`entity, location`$\rangle$	382.4	girls school
$\langle$`group, relation`$\rangle$	267.6	world language
$\langle$`communication, act`$\rangle$	231.1	speech treatment
$\langle$`person, artefact`$\rangle$	162.1	developer's kit
$\langle$`institution, person`$\rangle$	38.7	bank spokesman

Table 3: Estimated concept pair frequencies

also ruling out certain compounds as candidates (Pustejovsky, 1995; Copestake and Lascarides, 1997). Each lexical rule takes a pair of nouns of certain semantic type as input, and the output of the rule is a compound noun whose semantic representation stipulates the relation between a modifier and its head. For example, the compounds *metal tube*, *leather belt* and *tin cup* are the result of a lexical rule that combines a noun denoting a substance and a noun denoting an artefact to yield a compound denoting the artefact *made of* the substance.

The noun frequency and probability features do not capture meaning regularities concerning the compounding process. For example, we would expect the combination of the concepts representing *cocaine* and *customer* to be more likely than the combination of the concepts representing *people* and *excitement*. A way to obtain such likelihoods is by substituting the head and modifier by the concepts with which they are represented in a taxonomy. The frequency of the concept pair $f(c_1, c_2)$ could then be estimated by counting the number of times c_1 corresponding to n_1 was observed as the modifier of c_2 corresponding to the head n_2. Concept combination frequencies can be thought of as potential lexical rules which capture regularities and constraints on noun compound formation.

Counting concept frequencies 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 compounds 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 class. Nouns in WordNet (Miller et al., 1990) correspond to an average of 11.5 concepts (the word *return* belongs to 104 distinct conceptual classes), whereas nouns in Roget's thesaurus correspond to an average of 1.7 concepts (the word *point* has 18 distinct concepts). Because a head or a modifier can generally be the realization of one of several conceptual classes, counts of modifier-head configurations must be constructed for all potential concept combinations.

To give a concrete example consider again the compound *cocaine customer*. The word *cocaine* has one sense in WordNet and belongs to six conceptual classes ($\langle$`hard drug`$\rangle$, $\langle$`narcotic`$\rangle$, $\langle$`drug`$\rangle$, $\langle$`artefact`$\rangle$, $\langle$`object`$\rangle$, $\langle$`entity`$\rangle$). The word *customer* has also one sense in WordNet and belongs to five conceptual classes ($\langle$`consumer`$\rangle$, $\langle$`person`$\rangle$, $\langle$`life form`$\rangle$, $\langle$`causal agent`$\rangle$, $\langle$`entity`$\rangle$). Since we do not know which particular instantiation of these conceptual classes *cocaine* and *customer* are, we will distribute the attested frequency of *cocaine customer* over all pairwise concept combinations. We formally define the set of concept combinations as follows:

$$c(n_1, n_2) = \{\langle c_i, c_j \rangle \mid c_i \in classes(n_1), \\ c_j \in classes(n_2), c_i \neq c_j\} \quad (6)$$

Here, $c(n_1, n_2)$ is the set of distinct concept pairs a given noun-noun sequence is an instantiation of. Note that we impose a restriction on the type of concept pairs we generate, namely we disallow pairs with identical concepts (see (6)). The motivation for this restriction is twofold: first, we want to avoid overly general concept pairs that could potentially represent any noun-noun combination (e.g., $\langle$`entity, entity`$\rangle$, $\langle$`artefact, artefact`$\rangle$); second, it is implicitly assumed in the theoretical linguistics literature (Levi, 1978) that compounds are derived through combinations of distinct concepts[1].

For each compound in our corpus we generate the set of concept pairs it is potentially an instantiation of. The compound *cocaine customer* generates 29 concept pairs (e.g., $\langle$`artefact, consumer`$\rangle$, $\langle$`artefact, person`$\rangle$). We estimate the frequency of a concept pair $f(c_1, c_2)$ by summing over all noun-noun sequences $n_1 n_2$ that are representative of the concept combination $\langle c_1, c_2 \rangle$. We divide the contribution of each compound $n_1 n_2$ by the number of concept combinations it represents (Resnik, 1993; Lauer, 1995):

$$f(c_1, c_2) \approx \sum_{\langle n_1, n_2 \rangle \in \langle c_1, c_2 \rangle} \frac{f(n_1, n_2)}{|c(n_1, n_2)|} \quad (7)$$

Here, $f(n_1, n_2)$ is the number of times a given noun-noun sequence was observed in the estimation corpus and $|c(n_1, n_2)|$ is the number of conceptual pairs $n_1 n_2$ has. Assuming that we want to take the compound *cocaine customer* into account for estimating the frequency of the

[1] Dvanda or appositional compounds (e.g., *mother child*, *player coach*) are a notable exception.

concept pair $\langle\texttt{artefact,person}\rangle$, we will increment the observed co-occurrence count of $\langle\texttt{artefact,person}\rangle$ by $\frac{1}{29}$, since *cocaine customer* is represented by 29 distinct concept pairs. Table 3 shows a random sample of the derived concept pairs and their estimated frequencies.

Assume now that we want to decide whether the sequence *people excitement* is a valid compound or not. We generate all pairs of conceptual classes represented by *people excitement* (see (6)). The word *people* has four senses and belongs to 6 conceptual classes; *excitement* has also four senses and belongs to 15 classes. This means that *people excitement* is potentially represented by 90 concept pairs (*people* and *excitement* have no concepts in common), the frequency of which can be estimated from our corpus of valid compounds using (7). Since we do not know the actual classes for the nouns *people* and *excitement* in the corpus, we weight the contribution of each class pair by taking the average of the estimated frequencies for all 90 class pairs:

$$f(n'_1, n'_2) = \frac{\sum\limits_{\langle c_1, c_2\rangle \in \, c(n'_1, n'_2)} f(c_1, c_2)}{|c(n'_1, n'_2)|} \qquad (8)$$

As shown in Table 2 *people excitement* is much less likely than *cocaine customer*. Also note that *may push* is considered fairly likely (in fact more likely than *baby calf* which is a valid compound) since both *May* and *push* can be nouns and are listed as such in the WordNet taxonomy. The estimation of the concept frequencies in (7) relies on the simplifying assumption that a given noun is equally likely to be represented by any of its conceptual classes. As a result, the occurrence frequency of a compound is evenly distributed across all possible concept combinations representing the nouns forming the compound, since we cannot assess (without access to a corpus annotated with class information) which concept combinations are likely and which are not.

Context. Although the numerical features described above encode important information with respect to modifier-head relations and their properties, they are blind to contextual information that could potentially make up for tagging errors or the lack of structural information. Consider again the noun-noun sequence *may push* from Table 2, which is attested in sentence (9a). In this case, the context strongly indicates that *may push* is not a compound given that *push* is followed by a personal pronoun (personal pronouns typically precede compound nouns but never follow them).

We encode contextual information as the words preceding and succeeding the noun-noun sequence in question. In order to capture grammatical and syntactic dependencies we reduce words to their parts of speech and encode their positions to the left or right of the candidate compound. An example of this type of feature-encoding is given in (9b) which represents the context surrounding *may push* in sentence (9a). The feature-vector in (9b) consists of the candidate compound *may push*, represented by its parts of speech (NN1 and NN1, respectively) and a context of four words to its right and four words to its left, also reduced to their parts of speech.[2]

(9) a. Their different responsibilities in relation to the public may push them in opposite directions.
b. [NN2, PRP, AT0, AJ0, NN1, NN1, PNP, PRP, AJ0, NN2]

In the following we explore how the two types of features (i.e., numerical and categorical) perform independently as well as in combination.

4 Experiments

4.1 Machine Learning

The different features were combined using the C4.5 decision tree learner (Quinlan, 1993). Decision trees are among the most widely used machine learning algorithms. They perform a general to specific search of a feature space, adding the most informative features to a tree structure as the search proceeds. The objective is to select a minimal set of features that efficiently partitions the feature space into classes of observations and assemble them into a tree. For our experiments, the classification is binary, a noun-noun sequence is a compound or not. For comparison we also report the performance of the *Naive Bayes classifier* (Duda and Hart, 1973). The latter classifier does not perform a search through the feature space in order to build a model for classifying future examples. Instead all features are included in the classification. The learner is based on the simplifying assumption that each feature is conditionally independent of all other features, given the class of a given noun-noun sequence. We use the Weka (Witten and Frank, 2000) implementations of the C4.5 decision tree and Naive Bayes learner.

The classifiers were trained and tested using 10-fold cross-validation on 1,000 noun-sequences which were attested in the BNC only once. The

[2] The part-of-speech NN1 stands for singular common nouns, NN2 stands for plural common nouns, AT0 stands for determiners, PRP for prepositions, PNP for pronouns, and AJ0 for adjectives.

data was annotated by two judges. They were instructed to decide whether a noun-noun sequence is a compound or not and given a page of guidelines but had no prior training. The candidate compounds were classified in context: the judges were given the corpus sentence in which the noun-noun sequence occurred together with the previous and following sentence. Using the Kappa coefficient (Cohen, 1960) the judges' agreement[3] on the classification task was $K = .80$ ($N = 1000$, $k = 2$). This translates into a percentage agreement of 89%.

4.2 Experimental Results

Table 4 shows how accuracy varies when the learners (decision tree (DT) and Naive Bayes (NB)) use individual numeric features. For the concept frequency feature we experimented with two hierarchies, Roget's thesaurus and WordNet. As can be seen in Table 4 the best feature is concept frequency using WordNet ($f_{wn}(n_1,n_2)$), with an accuracy of 66.7% (for DT), a significant improvement over the baseline ($p < .05$) which was measured as the most frequent class (i.e., compound) in our data set (56.3%). Note that WordNet outperforms Roget's thesaurus even though both dictionaries contain taxonomic information. This fact may be due to the size of the taxonomies. WordNet contains twice as many noun entries as Roget (47,302 versus 20,448). Another 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 further examined the accuracy on the classification task when solely contextual features are used. We evaluated the influence of context by varying both the position and the size of the window of words (i.e., parts of speech) surrounding the candidate compound. The window size parameter was varied between one and four words before and after the candidate compounds. We use symbols l and r for left and right context, respectively and number to denote the window size. For example, $l = 2$, $r = 4$ represents a window of two words to the left and four words to the right of the candidate noun-noun sequence. Table 5 shows the performance of the two classifiers for some of the contextual feature sets we examined.

Good performances are attained by both learners. For DT, the best accuracy (69.1%) is obtained with windows of three or four words to the left of the candidate noun-noun sequence (see $l = 4$ and $l = 3$ in Table 5). NB performs best (70.8% and 69.8%) with small window sizes (see $l = 1$, and $l = 1$, $r = 1$ in Table 5). All three performances are a significant improvement over the baseline ($p < .05$). In general, better performance is achieved when one type of context is used (either left or right) instead of their combination (with the exception of $l = 1$, $r = 1$ and $l = 2$, $r = 1$ for NB). Our results suggest that even though context is encoded naively as parts of speech without preserving any structural or semantic knowledge, it retains enough information to distinguish compounds from non-compounds. This is an important result given that the best numerical predictor (i.e., $f_{wn}(n_1,n_2)$) relies heavily on taxonomic information. The contextual features are straightforward to obtain—all we need is a concordance of the candidate compound annotated with parts of speech.

Table 6 shows various combinations of numeric features, but also the interaction between numeric and contextual features. Again, we report some (i.e., the most informative) of the feature sets we examined. When only numeric features are used, the best accuracy for DT is attained with the combination of $f_{wn}(n_1,n_2)$ with $P(\mathrm{H}|n_1)$ (67.3%) or with $f_{ro}(n_1,n_2)$ (67.4%). Similar accuracies are obtained when $f_{wn}(n_1,n_2)$ is combined with two or three features (see Table 6). For the NB classifier, the best overall accuracy (72.3%) is attained for the feature set $\{f_{wn}(n_1,n_2),\ P(\mathrm{H}|n_1),\ l = 1\}$. This set of features yields signifiant improvement over the baseline ($p < .05$) and outperforms any other feature combinations including any other pairings with contextual information.

The DT learner's performance is consistently better when numeric features are combined with contextual ones. For all feature combinations shown in Table 6 the inclusion of context yields better results and accuracies around 70%. Generally, a small context (e.g., $l = 1$ or $r = 1$) yields better results (over a larger context) when combined with numeric features. A smaller context captures local syntactic dependencies such as the fact that compound nouns are typically preceded by determiners, verbs, or adjectives and succeded by verbs, prepositions or function words (e.g., *and*, *or*). On the other hand, widening the context tends to proliferate global syntactic ambiguity making local syntactic dependencies harder to learn. The DT learner achieves its best performance (72.0%) for the feature sets $\{f(n_1), f(n_2), P(\mathrm{H}|n_1), f_{wn}(n_1,n_2), f_{ro}(n_1,n_2), l = 2\}$ and $\{P(\mathrm{M}|n_2), f_{wn}(n_1,n_2), f_{ro}(n_1,n_2), f(n_1), l = 1\}$. It is worth noting that the second best performance (71.7%) is attained by the feature set $\{P(\mathrm{H}|n_1), P(\mathrm{M}|n_2), l = 1\}$. This is an important result given

[3] Cases of disagreement were excluded from the data on which the classifiers were trained and tested.

Features	DT	NB
Baseline	56.3	56.3
$f(n_1)$	60.7	48.9
$f(n_2)$	57.2	55.3
$P(\mathrm{H} \mid n_1)$	59.7	59.9
$P(\mathrm{M} \mid n_2)$	61.6	60.0
$f_{wn}(n_1,n_2)$	**66.7**	**62.3**
$f_{ro}(n_1,n_2)$	58.9	50.2

Table 4: Numeric Features

Features	DT	NB
Baseline	56.3	56.3
$l = 4$	**69.1**	63.9
$l = 3$	**69.1**	66.2
$l = 2$	68.5	67.9
$l = 1$	66.7	**70.8**
$r = 4$	64.7	65.0
$r = 3$	63.3	65.7
$r = 2$	64.3	66.6
$r = 1$	66.5	69.3
$l = 1, r = 1$	63.4	**69.8**
$l = 2, r = 1$	63.5	68.1
$l = 3, r = 1$	65.1	66.2
$l = 2, r = 3$	63.5	65.9
$l = 3, r = 4$	63.5	63.3
$l = 2, r = 3$	64.3	66.5
$l = 4, r = 4$	65.3	62.8

Table 5: Categorical Features

Features	DT	NB
Baseline	56.3	56.3
$f(n_1),P(\mathrm{M} \mid n_2)$	62.5	60.4
$f(n_1),P(\mathrm{M} \mid n_2),\ l = 1$	71.1	71.1
$f(n_1),P(\mathrm{M} \mid n_2),\ r = 1$	71.0	68.3
$P(\mathrm{H} \mid n_1),P(\mathrm{M} \mid n_2)$	62.1	63.2
$P(\mathrm{H} \mid n_1),P(\mathrm{M} \mid n_2),\ l = 1$	**71.7**	69.9
$P(\mathrm{H} \mid n_1),P(\mathrm{M} \mid n_2),\ r = 1$	69.7	70.5
$P(\mathrm{H} \mid n_1),f_{wn}(n_1,n_2)$	67.3	63.9
$P(\mathrm{H} \mid n_1),f_{wn}(n_1,n_2),\ l = 1$	70.8	**72.3**
$P(\mathrm{H} \mid n_1),f_{wn}(n_1,n_2),\ r = 1$	70.4	**70.8**
$f_{wn}(n_1,n_2),f_{ro}(n_1,n_2)$	67.4	55.0
$f_{wn}(n_1,n_2),f_{ro}(n_1,n_2),\ l = 1$	71.5	65.6
$f_{wn}(n_1,n_2),f_{ro}(n_1,n_2),\ r = 1$	71.4	66.5
$f_{wn}(n_1,n_2),f_{ro}(n_1,n_2),\ f(n_1)$	67.0	53.7
$f_{wn}(n_1,n_2),f_{ro}(n_1,n_2),\ f(n_1),l = 1$	70.4	65.0
$f_{wn}(n_1,n_2),f_{ro}(n_1,n_2),\ f(n_1),r = 1$	70.3	65.5
$P(\mathrm{M} \mid n_2),f_{wn}(n_1,n_2),f_{ro}(n_1,n_2)$	67.3	55.2
$P(\mathrm{M} \mid n_2),f_{wn}(n_1,n_2),f_{ro}(n_1,n_2),l = 1$	71.4	63.1
$P(\mathrm{M} \mid n_2),f_{wn}(n_1,n_2),f_{ro}(n_1,n_2),r = 1$	71.4	67.0
$P(\mathrm{M} \mid n_2),f_{wn}(n_1,n_2),f_{ro}(n_1,n_2),f(n_1)$	67.1	55.2
$P(\mathrm{M} \mid n_2),f_{wn}(n_1,n_2),f_{ro}(n_1,n_2),f(n_1),\ l = 1$	**72.0**	60.1
$P(\mathrm{M} \mid n_2),f_{wn}(n_1,n_2),f_{ro}(n_1,n_2),f(n_2),\ r = 2$	70.6	65.6
$P(\mathrm{H} \mid n_1),P(\mathrm{M} \mid n_2),f_{wn}(n_1,n_2),f_{ro}(n_1,n_2)$	66.9	56.0
$P(\mathrm{H} \mid n_1),P(\mathrm{M} \mid n_2),f_{wn}(n_1,n_2),f_{ro}(n_1,n_2),\ l = 1$	68.6	68.8
$P(\mathrm{H} \mid n_1),P(\mathrm{M} \mid n_2),f_{wn}(n_1,n_2),f_{ro}(n_1,n_2),\ r = 2$	69.8	67.1
$f(n_1),f(n_2),P(\mathrm{H} \mid n_1),f_{wn}(n_1,n_2),f_{ro}(n_1,n_2)$	66.9	55.3
$f(n_1),f(n_2),P(\mathrm{H} \mid n_1),f_{wn}(n_1,n_2),f_{ro}(n_1,n_2),\ l = 2$	**72.0**	61.4
$f(n_1),f(n_2),P(\mathrm{H} \mid n_1),f_{wn}(n_1,n_2),f_{ro}(n_1,n_2),\ r = 2$	71.2	62.0
$f(n_1),f(n_2),P(\mathrm{H} \mid n_1),P(\mathrm{M} \mid n_2),f_{wn}(n_1,n_2),f_{ro}(n_1,n_2)$	66.7	54.9
$f(n_1),f(n_2),P(\mathrm{H} \mid n_1),P(\mathrm{M} \mid n_2),f_{wn}(n_1,n_2),f_{ro}(n_1,n_2),\ l = 1$	70.5	64.3
$f(n_1),f(n_2),P(\mathrm{H} \mid n_1),P(\mathrm{M} \mid n_2),f_{wn}(n_1,n_2),f_{ro}(n_1,n_2),\ r = 1$	71.5	64.6

Table 6: Combination of numeric and categorical features

that these three features can be simply estimated from the corpus without recourse to taxonomic information.

When compared, the two learners yield similar performances. The NB classifier yields better results with smaller numbers of features, whereas the DT's performance remains steadily good, presumably because the most informative features are selected during the learning process.

5 Discussion

In this paper we focused on noun-noun sequences for which little evidence is found in the corpus and attempted to distinguish those which are valid compounds from nonce terms. The automatic acquisition of compound nouns (as opposed to terms) from unrestricted wide-coverage text has not received much attention in the literature. Lauer's (1995) study was conducted on a corpus exhibiting a uniform register and was furthermore biased in favor of syntactically unambiguous nouns. It cannot therefore be considered representative of part-of-speech tagged domain independent text.

Our results are encouraging considering the simplicity of the features we took into account and the fact that no structural information was used. Our experiments revealed that surface features such as the frequency of the compound head/modifier, the likelihood of a word as a head/modifier, or the context surrounding a candidate compound perform almost as well as features that are estimated on the basis of existing taxonomies such as WordNet. Our approach achieved an accuracy of 72% on the compound detection task. Although this performance is a significant improvement over the baseline (56.3%), it is 16.7% lower than the upper bound of 89% established in our agreement study (see Section 4.1). The task of deciding whether two nouns form a compound or not crucially depends on a variety of factors such as world-knowledge, the situation at hand, and the speaker's and hearer's communicative goals, none of which are directly represented by our features. We demonstrated that a machine learning approach can overcome the problem of sparse data which is closely related to the productivity of compounding. In particular, by exploiting information about frequent compounds or frequent contexts (which can be easily retrieved from the corpus) we can *indirectly* recreate evidence about the likelihood of two nouns to form a valid compound without necessarily relying on parsed text.

Our approach is conceptually close to Jacquemin (1996): in both cases a list of terms is used for the acquisition task. The crucial difference is that our approach does not presuppose the availability of a list of established terms external to the corpus for the acquisition to take place. We rely solely on the corpus for the discovery of reliable compounds (i.e., noun-noun sequences with CoocF>4) from which our numerical features are estimated. Another difference is that we discover novel compounds, whereas Jacquemin's (1996) method can only discover variants of already existing terms.

In the future we plan to experiment with better estimation schemes for the concept frequency feature that are appropriate for finding the the right level of generalisation in a concept hierarchy (Clark and Weir, 2002) and with smoothing techniques that *directly* recreate the frequencies of word combinations. We will also investigate in more depth the effect of context (represented as word-forms and word-lemmas) by taking into account bigger windows and use learners that are particularly suited for handling large numbers of features (e.g., Support Vector Machines, AdaBoost).

References

Didier Bourigault and Christian Jacquemin. 1999. Term extraction and term clustering: An integrated platform for computer aided terminology. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pages 15–21, Bergen, Norway.

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

Oliver Christ, 1995. *The XKWIC User Manual*. Institute for Computational Linguistics, University of Stuttgart.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Stephen Clark and David Weir. 2002. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2):187–206.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Ann Copestake and Alex Lascarides. 1997. Integrating symbolic and statistical representations: The lexicon pragmatics interface. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 136–243, Madrid, Spain.

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.

Béatrice Daille. 1996. Study and implementation of combined techniques for automatic extraction of terminology. In Judith Klavans and Philip Resnik, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pages 49–66. The MIT Press, Cambridge, MA.

Pamela Downing. 1977. On the creation and use of English compound nouns. *Language*, 53(4):810–842.

Richard O. Duda and Peter E. Hart. 1973. *Pattern Classification and Scene Analysis*. Wiley, NY.

David Elworthy. 1994. Does Baum-Welch re-estimation help taggers? In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pages 53–58, Stuttgart, Germany.

Christian Jacquemin. 1996. A symbolic and surgical acquisition of terms through variation. In Stefan Wermter, Ellen Riloff, and Gabriele Scheler, editors, *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language*, Lecture Notes in Artificial Intelligence, pages 425–438. Springer, Berlin.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27.

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

Geoffrey Leech, Roger Garside, and Michael Bryant. 1994. The tagging of the British national corpus. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 622–628, Kyoto, Japan.

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

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

Mark Liberman and Richard Sproat. 1992. The stress and structure of modified noun phrases in english. In Ivan Sag and Ann Szabolcsi, editors, *Lexical Matters*, pages 131–281. CSLI Publications, Stanford, CA.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA.

Elaine Marsh. 1984. A computational analysis of complex noun phrases in navy messages. In *Proceedings of the 10th International Conference on Computational Linguistics*, pages 505–508, Stanford, CA.

David McDonald. 1982. *Understanding Noun Compounds*. Ph.D. thesis, Carnegie Mellon University.

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.

James Pustejovsky. 1995. *The Generative Lexicon*. The MIT Press, Cambridge, MA.

Ross J. Quinlan. 1993. *C4.5: Programs for Machine Learning*. Series in Machine Learning. Morgan Kaufman, San Mateo, CA.

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

Ian H. Witten and Eibe Frank. 2000. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufman, San Francisco, CA.