Statistical Interpretation of Compound Nouns

A thesis presented
by
Jeremy Nicholson
to
The Department of Computer Science and Software Engineering
in partial fulfillment of the requirements
for the degree of
Bachelor of Computer Science (Honours)

University of Melbourne
Melbourne, Australia
October 2005
Statistical Interpretation of Compound Nouns

Abstract

We present a method for detecting compound nominalisations in open data, and deriving an interpretation for them. Discovering the semantic relationship between the modifier and head noun in a compound nominalisation is first construed as a two-way disambiguation task between an underlying subject or object semantic relation between a head noun and its modifier, and second as a three-way task between subject, direct object, and prepositional object relations.

The detection method achieves about 89% recall on a data set annotated by way of Celex and Nomlex, and about 70% recall on a randomly-sampled data set based on the British National Corpus, with 77% recall on detecting a more general set of compound nouns from this data.

The interpretation method achieved about 72% accuracy in the two-way task, and 57% in the three-way task, using a statistical measure based on z-scores — the confidence interval — in selecting one of the relations. Our proposed method has the advantage over previous research in that it can act over open data to detect and interpret compound nominalisations, as opposed to only operating in a limited domain or requiring hand-selection or hand-tuning.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title Page</td>
<td>i</td>
</tr>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>iv</td>
</tr>
<tr>
<td>Citations to Previously Published Work</td>
<td>vi</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>vii</td>
</tr>
<tr>
<td>Dedication</td>
<td>viii</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Compound Nouns</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Compound Nominalisations</td>
<td>3</td>
</tr>
<tr>
<td>1.3 Overview</td>
<td>4</td>
</tr>
<tr>
<td>2 Background</td>
<td>5</td>
</tr>
<tr>
<td>3 Resources</td>
<td>12</td>
</tr>
<tr>
<td>3.1 Tools</td>
<td>12</td>
</tr>
<tr>
<td>3.1.1 The BNC</td>
<td>12</td>
</tr>
<tr>
<td>3.1.2 Pre-processing Tools</td>
<td>12</td>
</tr>
<tr>
<td>3.1.3 Lexical Resources</td>
<td>13</td>
</tr>
<tr>
<td>3.2 Data Sets</td>
<td>15</td>
</tr>
<tr>
<td>3.2.1 2-Way Classification</td>
<td>15</td>
</tr>
<tr>
<td>3.2.2 3-Way Classification</td>
<td>16</td>
</tr>
<tr>
<td>4 Proposed Method</td>
<td>18</td>
</tr>
<tr>
<td>4.1 Detection of Compound Nominalisations</td>
<td>18</td>
</tr>
<tr>
<td>4.2 Interpretation of Compound Nominalisations</td>
<td>18</td>
</tr>
<tr>
<td>4.2.1 Paraphrase Tests</td>
<td>18</td>
</tr>
<tr>
<td>4.2.2 Robust Statistical Interpretation</td>
<td>20</td>
</tr>
<tr>
<td>4.2.3 The Algorithm</td>
<td>22</td>
</tr>
<tr>
<td>5 Experimental Results</td>
<td>25</td>
</tr>
<tr>
<td>5.1 Evaluation of the Detection Method</td>
<td>25</td>
</tr>
<tr>
<td>5.2 2-Way Classification</td>
<td>27</td>
</tr>
</tbody>
</table>

iv
5.3 3-Way Classification ........................................... 29

6 Discussion .................................................................. 32
  6.1 Summary ......................................................... 32
  6.2 Future Work ....................................................... 33
  6.3 Two Final Words .................................................. 34

A Samples of Lexical Resources ................................. 39
Citations to Previously Published Work

Portions of Chapter 4 and Chapter 5 have appeared in the following paper:

Acknowledgments

We would thank Mirella Lapata for generously contributing her data set to the study.

Also, many thanks to Amy, Andy, and Fi for donating their time in a hectic period to annotate the open data set.

And Tim, for his constant patience.
Pour B,

Sans laquelle rien de bon ne passe jamais.
Chapter 1

Introduction

People are brought up to believe that to be successful in science you have first to have your right cerebral hemisphere obliterated; and the people (like a lot of teachers) who perpetrate this nonsense should be fried slowly in rancid yak fat.

In this sentence taken from the British National Corpus, what does the compound noun yak fat mean? Does it mean “yak living in fat”, like bird cage? Or “yak shaped like fat”, like pork medallion? Or perhaps “fat made at a yak”, like creamery butter?

Almost any English speaker will instantly reject these paraphrases in favour of something along the lines of “fat taken from a yak”. They have a broad knowledge base that makes the alternative proposals seem absurd: fat, unlike a cage, is not a container inside which something can live; yaks cannot take the form of fat (c.f. ?yak medallion\(^1\)); and yak, unlike a creamery, is not a location at which one can make fat.

Consider attempting to automatically propose such paraphrases and make judgements on them, perhaps for capturing meaning in an information extraction system. A wide-coverage knowledge base is clearly infeasible, as the pragmatic relationships between real-world entities are innumerable, and therefore intractable. And even if such a wide-coverage knowledge base were available outside a narrow domain, each instance would introduce errors in interpretations that would have been correct in a more limited domain (see Chapter 2).

This automatic interpretation of these compounds is usually performed as a disambiguation task, along the lines of the procedure above: a set of classes is proposed, and the most preferable is selected. These classes usually take one of two forms: a semantic conceptual description or a syntactic paraphrase. For example, polystyrene garden-gnome might have a conceptual description “polystyrene(x) ∧ garden-gnome(y) ∧ material(y, x)” or a syntactic paraphrase “garden-gnome of polystyrene”. We take

\(^1\)We adopt, as is standard practice, the informal Chomskian gradation of ungrammaticality. An asterisk (*) indicates an ungrammatical example, while a question mark (?) indicates a marginally grammatical, but “marked” usage.
the stance that these are, in fact, equivalent, subject to the appropriate selection of paraphrases. This is in line with works such as Levin (1993), who examines the interaction and associated predictability of syntax and semantics in diathesis alternations, where the syntax can be predicted from the semantics, and vice versa.

To simplify this interpretation task of compound nouns, we adopt the common ploy of restricting the set of compounds under consideration. First, however, let us consider what a compound noun is, and how it can be interpreted.

1.1 Compound Nouns

A **compound noun** is a sequence of two or more nouns comprising an \( N \) (i.e. a noun phrase without a determiner). All of the examples from above (yak fat, pork medallion, bird cage, and creamery butter) are noun compounds, as is noun phrase from the previous sentence, as indeed is compound noun. In open language, and especially technical language, compound nouns are productive, in that novel instances can be readily formed and understood in context, as attested to by Lapata and Lascarides (2003). Of 510,673 noun sequences found in the British National Corpus (Burnard 2000), almost 70% co-occur with a frequency of only one, implying a novel compound. The commonality of compound nouns is also evidence for their productivity: Tanaka and Baldwin (2003) note that more than 1% of the words in this corpus participate in such compounds.

Compound nouns occur in a number of different languages across language families. We find instances in German, Japanese, modern Greek, and Welsh, among other languages, with morphological and syntactic variations for each. For example, in German *der Vogelkäfig* “the bird cage” is derived from *der Vogel* “the bird” and *der Käfig* “the cage”.\(^2\) We will only consider English compounds in this work, but the techniques we propose apply equally in the consideration of other languages, with caveats for these variations.

Implicit in these compounds is a (possibly hierarchical) semantic relationship between the **modifier** and the **head** — the rightmost noun. (For example, in yak fat, we have the head fat and modifier yak: it is fat modified in some way by yak.) The comparisons for yak fat that we have listed above could be classified as for, of, from, and from, corresponding to the preposition in a minimal paraphrase of the compounds: “cage for bird”, “medallion of pork”, “butter from creamery”, and “fat from yak”. This classification set collapses the semantic distinction between the last two compounds: getting butter from a creamery and getting fat from a yak are

\(^2\)These unsegmented compounds are standard in German. English does display a similar phenomenon, when the compound is lexicalised. bird cage has a preference for being segmented, while nightgown has entered the language as a distinct entity, and has a preference for being fused. wicket-keeper prefers an intermediate hyphenated form. For a more complete study, see Bauer (1983). We explicitly exclude fused and hyphenated noun compounds in our study.
Chapter 1: Introduction

perhaps semantically distinct, although syntactically equivalent in this context.

Deriving a total set of classes to describe the full semantic extent of productive
noun compounds has been a stumbling block in efforts aimed at their automatic in-
terpretation. Mirroring tensions in linguistics, psychology, and artificial intelligence,
two approaches have been taken to create such a set: a rationalist approach and an
empirical approach. The rationalists contend that humans, as oracles, can intuitively
construct a natural list that achieves this. The empiricists contend that any such
proposed list is inherently biased and can therefore only be constructed in the con-
text of enumerating the discovered relations in some data set. Both sets suffer from
productivity, the former in the existence of unusual samples: Downing (1977) gives
the example apple-juice seat, “a seat at which a glass of apple juice is placed”; this
and other outliers do not satisfactorially fit in most proposed class sets. The empirical
approach, on the other hand, suffers from brittleness: the enumerated list only covers
the relations peculiar to the given data. Any such list is inherently biased and is
particular to a given corpus or domain; to have complete coverage, one would need
to consider every instance of language ever used, and even this would not account for
novel usages.

1.2 Compound Nominalisations

An important subclass of compound nouns is compound nominalisations: those compounds whose head is morphologically derived from a verb. An exam-
ple of this is human consumption, which can be loosely paraphrased as “humans consume”. In English, verbalisation of nouns is productive, so many compounds may
be viewed as nominalisations, although the semantic connection can be somewhat
tenuous. Consider creamery butter from above: the verbalisation to butter, which
came into use much later than the noun butter,\(^3\) leads to the spurious interpretations
“*creamery buttered [st]”,\(^4\) “*[so] buttered the creamery” (compare ?bread butter
“butter used to butter bread”).

Nevertheless, a significant proportion of nouns do have a legitimate verbal form
that can be used to form paraphrases. The advantage of this is that the semantic
set of relationships, difficult to establish across all compound nouns, is reasonably
well-defined. A modifier may be the subject of the deverbal head, or its object.
Furthermore, the object paraphrase may be further broken down into direct objects
and prepositional objects. The compound nominalisation example above human con-
sumption “humans consume” is an example of the subject relationship, film criticism
“*[so] criticises (a) film” is an example of a direct object relationship, and welfare report
“*[so] reports on welfare” is an example of a prepositional object relation, for

\(^3\)The Oxford English Dictionary (OED) gives nominal examples of butter from "1000, and verbal
usage since 1496.

\(^4\)In this work, st and so refer to “something” and “someone” respectively.
the preposition on.

1.3 Overview

The classification of semantic relationships of compound nouns has been performed variously, with varying degrees of success (Leonard 1984; Vanderwende 1994; Lauer 1995; Rosario and Hearst 2001; Lapata 2002; Moldovan et al. 2004; Kim and Baldwin 2005; Grover et al. 2005). The motivation for this is that almost any natural language processing (NLP) task with a semantic or lexical semantic leaning requires such interpretation, such as machine translation (Johnston and Busa 1996), or question answering (Leonard 1984; Lauer 1995). Our aim in this work is to examine the novel task of detecting compound nominalisations in open text, and provide a robust selection technique across a set of subject, direct object, and prepositional object relations.

In Chapter 2, we review the relevant literature on interpreting both compound nouns and compound nominalisations. In Chapter 3, we give a short description of the resources we employed. In Chapter 4, we describe our approaches toward identification and interpretation of compound nominalisations, with results of experimentation in Chapter 5, and a brief discussion and summary in Chapter 6.
Chapter 2

Background

There are two works that form the foundation of this research: Lapata (2002) and Grover et al. (2005). Both of these perform, as we do, statistical interpretation of compound nominalisations. However, let us first examine the earlier proposals of methodologies for interpreting compound nouns, and systems which applied those methodologies.

The almost legendary examination of the grammaticality — syntax and semantics — of nominalisations is that of Chomsky (1970). A detailed description of the heavy linguistic theory does not belong here; however, the readings that he proposes for nominalisations appear in various forms throughout the more recent work on them. One example of this is John’s refusal of the offer (for offer refusal). Another, which will interest us later, is the equality of several of John’s proofs of the theorem and John proved the theorem (for theorem proof). Chomsky emphasises the genitive (indicating possession, e.g. ‘s in English) in his work, which we employ in Section 4.2.1.

The seminal work on the description and interpretation of compound nouns is that of Levi (1978), who takes a rational approach. She examines a plethora of identified and hand-curated instances of compound nouns and their implied meanings. She distinguishes between compound nouns and “complex nominals”, which also include predicating adjectives — these have not usually been considered since. There are nine “Recoverably Deletable Predicates (RDPs)” by which she delimits the general space of compound nouns: \{cause, have, make, use, be, in, for, from, about\}. Their readings correspond to “[HEAD] causes/has/makes/uses/is [MOD]”, and “[HEAD] in/for/from/about [MOD]”. Even though these categories seem broad, we already lack a wholly satisfactory classification for the noun compound pork medallion from above: the closest we find is “medallion is pork”, a somewhat questionable paraphrase.

Levi also explicitly notes compound nominalisations as a well-defined subclass of compound nouns, and identifies four types: Act, Product, Agent, and Patient nominalisations, which can be classified as one of Subjective and Objective classes. For example, dream analysis is Act and Objective, for “the act of analysing dreams”.

From the same era, and also applying a rational approach, Warren (1978) exam-
Chapter 2: Background

ines the syntactic structure of compound nouns and proposes her own twelve conceptual semantic classes for categorisation: \{Source-Result, Copula, Resemblance, Whole-Part, Part-Whole, Goal-OBJ, Place-OBJ, Time-OBJ, Origin-OBJ, Purpose, Activity-Actor\}, without explicitly considering nominalisations. There is some overlap between these classes and those of Levi, with Copula corresponding to BE, and Purpose to FOR, as well as some generalisation and specificity. However, this set still fails to define pork medallion more satisfactorially than the copular paraphrase “medallion is pork”.

Taking a different, and arguably even more rational approach to the problem, Finin (1980) defines interpretation in terms of role nominals: all nouns have a verbal form, and these need not be morphologically related. Selection of a paraphrase relies on a concept-based approach across a hand-crafted set of slot templates. An example of the “role” in vegan chocolate would be that chocolate has an underlying verb to eat, and the compound’s interpretation becomes “vegan eats”. The choice of a given verb is not always clear however: consider milk chocolate and glove box chocolate, where readings of “milk eats” or “glove box eats” are clearly absurd. Another unfortunate property of role nominals is that their number is essentially bounded only by the power set of verbs, which is itself an open class — and is therefore potentially infinite (Lapata 2002). A similar work is that of Isabelle (1984), except that he works specifically with nominalisations, as we do. For him, the slots to be filled derive from the verb frame of the head of a given compound.

Copestake and Lascarides (1997) assert that any approach based wholly on lexical grammar or wholly on pragmatics will be inadequate. They examine the lexical preference for a given semantic sense of a word, and attempt to integrate that with a set of knowledge resources and hand-crafted rules, finding a balance between rational and empirical approaches. Similar to this, with slot-filling according to prior probabilities, is work performed by McDonald (1982) and Lehner (1988), the latter of which analyses the performance of a knowledge-based system by Lebowitz (1983) at classification according to Levi’s RDP’s. McDonald’s work also notes a high relative frequency of noun compounds in his data set of newspaper articles — one novel instance every second sentence — but this has been surmised to be a peculiarity of that data set (Leonard 1984). No formal evaluation for any of these systems is given.

The first evaluated work on automatically interpreting compound nouns is the rational work of Leonard (1984), who develops a typology for classification — \{Sentence, Locative Sentence, Locative, Annex, Equative, Material, Additive, Reduplicative\} — and notes that it loosely agrees with earlier sets, especially that of Warren (1978). Interpretation is performed using a rule-based approach: a paraphrase is selected by applying a prioritised series of rules, and halting when one is satisfied. These paraphrases are in the same vein as those identified by Chomsky: e.g. mountain vista is “view of a mountain or mountains”, which is equivalent to a Sentence interpretation, for “[sO] views a mountain”. Her features for rule-matching make use of substantial linguistic pre-processing of the data set. She concludes that 76% of her paraphrases
were generously possible in English — although this was analysis on the training set.

In addition, Leonard also examines the upsurge in compound noun usage over the past 250 years as motivation for her analysis. In 20,000 words from each of five prose fiction works between 1759 and 1962, she observes a monotonic increase from eight types and ten tokens to 198 types and 229 tokens. Interestingly, the ratio of types to tokens stays more or less constant, implying that the proportion of novel instances remains about the same. This gives further evidence toward the productivity of compounds, as this ratio is quite high.

Vanderwende (1994) also uses and evaluates a rule-based technique for the interpretation of so-called “noun sequences”, using a weighted scoring procedure to avoid dependency on the order of rule application. If a compound encounters a rule that it satisfies before considering a later, preferable rule, an incorrect classification can be given — a feature of Leonard’s system that Vanderwende circumvents by always applying every rule. Her thirteen rationally derived categories, \{Subject, Object, Locative, Time, Possessive, Whole-Part, Part-Whole, Equative, Instrument, Purpose, Material, Causes, Caused-By\}, are a combination of those proposed by Warren and Levi. She also explicitly examines compound nominalisations with the Levi classes of Subject and Object. Deriving optimal weights from a training set leads to the correct interpretation of 52% of the compounds in a small held-out test set.

A different rational concept-based approach from those above is used by Jones (1995): after creating a feature structure for the head concept, a graph-based unification procedure is performed over the given semantic relations to achieve an interpretation. The method was not explicitly evaluated.

Rosario and Hearst (2001) explore compound noun interpretation in an empirical way, using concepts in a specific biomedical domain. They make use of a lexical resource in the domain (the UMLS metathesaurus) for deriving the concepts of the given words, and use neural networks and decision trees to achieve 70% classification accuracy over a moderate-sized test set. They have thirteen semantic relations, \{Lexicalized, Subtype of, Activity, Production (genetic), Purpose, Person afflicted, Study Instrument, Cause, Location, Measure, Damage, Instrument, Procedure\}, with some overlap with other sets, but several relations particular to their domain.

The work of Lauer (1995) is the first notable attempt to use a statistical model to interpret the semantics of noun compounds empirically. He attempts to paraphrase a compound noun in the same manner as Leonard above, by proposing a minimal preposition-based reading for the compound, based on the set \{OF, FOR, IN, ABOUT, WITH, FROM, ON, AT\}, explicitly excluding nominalisations. He achieves accuracy of 47% over a small test set, but in a domain-independent setting.

Lauer also points out an information retrieval study by Gay and Croft (1990) that examines the effect of adding a slot-filling compound noun interpretation program on the performance of retrieval in a limited domain. The conclusion is that there is an accuracy reduction associated with broadening the knowledge base for a conceptual system, and therefore building such a base has an unjustifiable cost.
This result was reinforced by the work of Fan et al. (2003), who conclude that detailed knowledge of the noun constituting the compound is not required to form a correct interpretation.

An interesting subtask in this work was that of **noun compound bracketing**, which had been attempted with only limited success previously Lauer (1995). Lauer, noting that such compounds are right-headed, correctly identifies that the required bracketing depends on the preference of the modifier for collocating with the head or head modifier. For example, the ternary noun compound *horseback riding school* should be bracketed as [[[horseback riding] school]] (and not [horseback [riding school]]) as *horseback riding* is preferable to *horseback school* (compare this with *velvet evening* and *velvet dress* from [[velvet [evening dress]])]. As a consequence of this, most work since has dismissed ternary and higher-arity compounds as a solved problem, and reduced the task to considering interesting binary compounds only. We take the same stance, but ensure ourselves that they occur with low frequency.

More recent statistical work has been performed by Moldovan et al. (2004) and Kim and Baldwin (2005) on 35 and 20 semantic relations and achieving an F-score of 43% and an accuracy of 53% respectively. Moldovan et al. use a semantic learning approach, while Kim and Baldwin make use of semantic similarities in the WordNet lexical database (Fellbaum 1998).

One of the few works specifically devoted to compound noun detection is Lapata and Lascarides (2003). They mine the British National Corpus for high-frequency noun sequences, which they note as often being compound nouns (93.5%), and use these to improve discrimination between low-frequency compound nouns and low-frequency noun sequences that are not compounds (e.g. *term cancer* in usage of the term *cancer*). They use a statistical approach along with a machine learner to raise accuracy from a baseline of 56% to 72%.

Spärck Jones’s (1985) take on the issue is a pessimistic one: she contends that detailed world knowledge is required to interpret noun compounds, and that the structural and semantic problems are too difficult for automatic systems to overcome. In the face of this, Copestake (2003) reanalyses many of the issues in compound noun interpretation, and tentatively concludes that a weak conceptual approach combined with a statistical approach based on corpus frequencies has the potential to satisfactorially make interpretations on noun compounds, even in a domain-independent environment.

So far in our examination of the literature, we have examined many works that have considered the interpretation of compound nouns, with several that have identified compound nominalisations, like Levi (1978), Isabelle (1984), and Vanderwende (1994). Works like Lauer (1995) have taken a statistical approach. Let us now consider the works specifically on the statistical interpretation of compound nominalisations: Lapata (2002) with a limited semantic set, and Grover et al. (2005) in a specific domain.
The greatest raw accuracy for the compound noun interpretation task, 86%, has been achieved by Lapata (2002), albeit on a far better-defined and far easier two-way classification task. Lapata only considers binary compound nominalisations in her data set, hand-curating a sample of 796 out of about 170,000 candidates automatically extracted from the British National Corpus. Another simplification of the task is that she only has the interpretation set of \{SUBJ, OBJ\}: any candidate whose relation of the modifier to the head noun was something other than subject or object (specifically, the Levi predicates) was discarded. She interprets an objective relation as a direct object only, also discarding compounds whose head bears a prepositional relation to the modifier. The Levi predicate and prepositional object compounds respectively account for 28.0% and 9.2% of a sample of the candidate set.

Lapata uses a statistical interpretation of a given compound nominalisation: corpus frequencies are acquired for the verb-argument pair of each relation for a given deverbial head and modifier. Let us work through the method based on the compound market acceptance. First, counts of “market accept \[st\]” and “\[st\] accept market” are acquired. Interpretation amounts to selecting the most likely of these two relations. The probability of a relation given the noun compound \(n_1n_2\) is:

\[
P(\text{rel} \mid n_1, n_2) = \frac{f(n_2, \text{rel}, n_1)}{f(n_1, n_2)} \quad (2.1)
\]

\[
\approx \frac{f(v_{n_2}, \text{rel}, n_1)}{\sum_i f(v_{n_2}, \text{rel}_i, n_1)} \quad (2.2)
\]

From equation (2.1), we choose the relation with the greatest probability. However, the counts \(f(n_2, \text{rel}, n_1)\) cannot be read directly from the corpus (c.f. "acceptance SUBJ market", as a noun cannot have a syntactic subject relationship to another noun). Instead, we make use of the verb-argument frequencies \(f(v_{n_2}, \text{rel}, n_1)\) and find equation (2.2).

\(f(n_1, n_2)\) (or, equivalently, \(\sum_i f(v_{n_2}, \text{rel}_i, n_1)\)) is the same for each of the two relations, so choosing the greatest probability is only a matter of choosing which of SUBJ or OBJ has more attested instances in the corpus for a given compound. In our case, \(f(\text{accept,SUBJ,market}) = 6\) and \(f(\text{accept,OBJ,market}) = 8\), so an OBJ interpretation is chosen.

As the probability is not meaningful when \(f(v_{n_2}, \text{rel}, n_1) = 0\), and not defined when \(\sum_i f(v_{n_2}, \text{rel}_i, n_1) = 0\) (i.e. the verb-noun pair does not occur in the corpus), Lapata makes use of different kinds of smoothing: approximating the count of unseen instances by holding out some probability mass and reassigning it to unseen events. The simplest of these, backing-off techniques, examines shorter “contexts” in the absence of occurrences of the entire instance. Lapata assumes that the probability can be approximated by backing off to the counts of the modifier noun, where
equation (2.1) and equation (2.2) become:

\[ P(\text{rel} \mid n_1, n_2) \approx \frac{\alpha f(\text{rel}, n_1)}{f(n_1)} \]

\[ \approx \frac{\alpha f(\text{rel}, n_1)}{\sum_i f(\text{rel}_i, n_1)} \]  

where \( \alpha \) is a normalisation constant used to ensure that the probabilities sum to one — a well-formed probability distribution.

Two other kinds of smoothing examined are **class-based smoothing**, where frequencies are recreated across similar words using taxonomies such as WordNet (Fellbaum 1998) or a publicly available version of Roget’s thesaurus, and **distance-based smoothing**, where the similar words are found from co-occurrence in the corpus. An in-depth analysis is unnecessary here, as Lapata does not find a significant performance difference between the three kinds of smoothing, although backing-off generally performs slightly worse than the others.

Lapata, noting the pragmatic dependency of compound nouns on context, considers sentential context in the form of part-of-speech tags (see Chapter 3) and lemmas (normalised word form, usually by stripping morphological affixes) of the previous and following five words. She integrates this into the interpretation algorithm using Ripper (Cohen 1996), a decision-tree learner, along with the judgements given by the different smoothing algorithms. All in all, she notes that optimal performance (87.3%) occurs when WordNet class-based smoothing is combined with the two following lemmas, and performs significantly better (\( \chi^2 = 30.64, p < .05 \)) than the baseline.

Grover et al. (2005) perform a similar disambiguation task to Lapata, but in a biomedical domain similar to Rosario and Hearst (2001). Instead of Lapata’s binary classification task, they expand the set of compounds under consideration to include heads which have a prepositional relation to the modifier noun. Their semantic set has thirteen items: \{SUBJ, OBJ, WITH, TO, ON, IN, FROM, ABOUT, AGAINST, BY, INTO, OF\}.

Selecting a relation is performed in the same way as Lapata: i.e. choosing the one which has the greatest probability according to equation (2.2). Again, this corresponds to the relation with the greatest number of attested instances from the corpus. The only difference is that thirteen relations are under consideration, instead of two.

Class-based smoothing is again employed for data sparseness, using both the UMLS metathesaurus, as in Rosario and Hearst (2001), and WordNet, although the comparative performance of these two resources is not clear from the given results. They also examine the impact of affixes on the derivation of the deverbal form of the head (e.g. -er for walker), which Lapata dismisses as ineffective.

The final performance figure of 77% accuracy is achieved again by way of a decision-tree learner (this time C4.5 (Quinlan 1993)), with features being the corpus
frequencies smoothed using both WordNet and UMLS, as well as affixes. Previous
and following context in the form of POS tags or lemmas was not observed to improve
performance.

Another feature of this biomedical work is an examination of the relative frequency
of compound nouns and nominalisations in this domain, as well as the accuracy of
automatic collection of these. From a small set of sample sentences, the authors
note that 72% of them contained one or more compound nouns, 35% of which were
nominalisations. This frequency is much greater than observed previously, which is
surmised to be an artefact of this technical domain. Finally, they perform a brief
analysis of the automatic collection of compounds via lookup on the deverbal head
and observe an accuracy of 95.6%.

Lapata (2002) has some undesirable features: the data set was filtered in such
a way as to make the task artificial, as the likelihood of a test set consisting only
of subject and direct object relations is low. Also, little consideration was given to
the derivation of this data set from text. Grover et al. (2005) circumvents these
shortcomings by examining a larger set of interpretations, and acquiring the data
set automatically, at a cost of accuracy and domain-specificity. This work, however,
performs substantial linguistic pre-processing of the data, in hand-tuning the parser
for marking-up compounds so as to simplify identification and interpretation, for
maximum numerical performance.

A criticism, then, of both of these works is that neither approach can perform
on open text. Lapata’s requires a manually-selected data set of limited scope, while
the work of Grover et al. requires substantial pre-processing of the data. In an
environment where this is impossible, such as the World Wide Web, both methods
will be ineffective.
Chapter 3
Resources

3.1 Tools

Many lexical resources have been constructed for natural language processing. We make use of several in our research for data sets, pre-processing, and during analysis of a compound noun.

3.1.1 The BNC


The BNC is a 100M token balanced corpus that is generally considered to be representative of the English language, as it contains both spoken and written text samples from a broad variety of sources. Unlike the authors above, we restrict ourselves to the 90M token written component of the corpus.

3.1.2 Pre-processing Tools

RASP (Briscoe and Carroll 2002) is a tag sequence grammar-based statistical parser. It attempts to find a full syntactic parse for a given sentence based on the sequence of part-of-speech (POS) tags on the words in the sentence (i.e. singular common noun NN1, past tense of regular verb VVD, etc.). It uses statistical analysis of the constituents in the sentence fragments to find partial parses, and attempts to construct a spanning parse.

The parser discovers its own POS tags and chunks (i.e. noun phrases, verb phrases, etc.). We also wished to find these independently across the BNC, which we did by way of tools built with fnTBL 1.0 (Ngai and Florian 2001), a fast transformation-based learning system. The reason for this is that RASP is effective at finding a syntactic
parse for a sentence, but slow and not attuned for POS tagging and chunking. fnTBL is so attuned, and much faster.

We POS tag, chunk, and parse the corpus using these tools, so as to be able to obtain the corpus frequencies for interpretation, and to be able to identify compound nouns within it. Given the output of a chunk parser, a naive approach to the discovery of compound nouns would be to search for noun sequences heading a noun chunk (e.g. car wash in “[the car wash]NP [on the corner]PP”). To get corpus frequency counts, we make use of the fact that the parse returned by RASP for each sentence lists the syntactic relations contained within the most probable parse tree, which we use to find verb-argument pairs or noun collocations. In the sentence The Jabberwock, with eyes of flame, came whiffling through the tulgey wood, and burbled as it came!, we find relations such as:

ncsubj(burble+ed,Jabberwock,_)  
ncmod(of,eye+s,flame)  
ncmod(_,wood,tulgey)

corresponding to the fact that Jabberwock is the subject of burble (in the past tense +ed), eye of flame is a noun sequence separated by of (for plural eye +s), and tulgey wood is a noun sequence (perhaps erroneously so, as this text is subject to interpretation).

### 3.1.3 Lexical Resources

We use the combination of CELEX, NOMLEX and CATVAR to create a list of deverbal nouns, and then use this list to see if the head of a given compound is deverbal. The entry for the word acceptance in each of these is given for illustration in Appendix A.

CELEX (Burnage 1990) is a database of English, German, and Dutch containing phonological, morphological, syntactic, and frequency data on each lexical entry, with English lemmas derived by annotators for the most part from the 1974 Advanced Oxford’s Learner Dictionary and the 1978 Longman Dictionary of Contemporary English. We make use of the morphological information to search for nouns with verbal forms (or, equivalently, verbs with nominal forms), but choose not to use the given frequencies.

NOMLEX (Macleod et al. 1998) is a database of 1025 nominalisations in English containing information about their complementation, or by which prepositions they can be followed in a prepositional phrase. These were derived mostly from frequent occurrences in the Brown and Wall Street Journal corpora. We extract the verbal form for each noun, and observe whether some relation has been absorbed in the compound itself.\footnote{In holder, the subject relation has been absorbed, so that licence holder cannot mean that “the}
Chapter 3: Resources

CatVar (Habash and Dorr 2003) is a database of clusters of uninflected words and their variants, without specific morphological relatedness; words were stemmed using the Porter Stemmer (Porter 1997), and then clustered with other words having the same stem. Words were extracted from the LCS Verb and Preposition Databases, an English morphological lexicon (ENGLEX), WordNet 1.6, and other sources.

CatVar’s list is comprised of 13,730 deverbal nouns, compared to 1,012 for NomLEX and 2,130 for CELEX (2,570 when these two are combined). This imbalance occurs because the words in CatVar were automatically extracted, while those in CELEX and NomLEX were hand-annotated.

The total number of nouns in the deverbal list is about 14,000, and is extensive. In general, coverage over a data set is very good, but with many false positives, due to nouns that are not deverbal being included in the combined set. These are often due to the inclusion of CatVar: a verbal form extracted from CELEX or NomLEX is certainly morphologically related due to the hand-annotation, but this is not necessarily the case with CatVar.

For example, the noun first is not deverbal, as correctly attested in CELEX. CatVar, however, clusters it with the verb fire because they happen to share the same stem fire-. This unusual stem is given to first as the stemmer removes the apparent superlative affix -st and replaces it with -e.

This method creates many erroneous deverbal noun–verb pairs, which we wish to avoid. In principle, we can apply interpretation techniques to a compound identified as deverbal so as to discard some of these compounds incorrectly identified as being so. We compare these interpretation techniques with simply removing CatVar from the deverbal list for avoiding false detection of compounds in Section 5.1.

We used COMLEX, WordNet, the English Resource Grammar (ERG), and the Longman Phrasal Verb Dictionary to create a list of phrasal verbs to decide when to apply the analysis of a prepositional paraphrase for the verbal head of a given compound.

COMLEX (Grishman et al. 1994) is a dictionary of English designed specifically for NLP, hand-constructed from an number of sources including the Brown and Wall Street Journal corpora. WordNet (Fellbaum 1998) is a lexical reference system where items are combined based on synonymous concepts. The ERG (Copestake and Flickinger 2000) is a broad-coverage, linguistically precise grammar of English. The Longman Phrasal Verb Dictionary (Dignen et al. 2000) is a corpus-based listing of phrasal verbs. From these, we extracted a list of 618 distinct phrasal verbs.
3.2 Data Sets

3.2.1 2-Way Classification

Our primary data set for the 2-way classification task is the 796-item data set used by Lapata (2002). For each item, we have a gold-standard tag, the lemma of the modifier, the lemma of the head, the base verbal form of the head, and the sentence in which the item occurred. Lapata divides the data set into training and test sets, to optimise parameters for her distance-weighted smoothing. This is unnecessary for us, as we do not use a parametrised method.

In the original Lapata data, the underlying verb form of the head noun was identified using a combination of CELEX and NOMLEX data. As a unique form was chosen, sub-optimal results occurred (e.g. the base verb of position is given as posit). In order to ameliorate such quirks in morphological analysis and expand the coverage of our method, we mined CELEX and NOMLEX, and also the word clusters in CatVAR for morphologically-related noun–verb pairs. This culminated in a total of about 14,000 deverbal nouns, many of which are listed with multiple base verb forms (e.g. position is listed as all of pose, posit and position).

To validate the Lapata data for consistency, we removed those nominalisations for which we were missing one of: the lemma of the head, the base verbal form of the head, or (especially) the context sentence. We also removed nominalisations that did not occur in the same chunk when we parsed the sentence using the tools described above, as a number of the data items were not compound nouns (e.g. girl cry in the sentence ‘Sometimes even big girls cry,’ he mused..., as cry is a verb). Items whose deverbal form did not appear in our list of deverbal nouns were also removed (e.g. transport decision-maker, as the supposed verb ??decision-make does not appear in the language resources). As well, nominalisations with a proper noun as the head or modifier were removed. The relative frequency of each of these is shown in Table 3.1. We were left with 695 consistent items classified as one of SUBJ or OBJ.

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Item</td>
<td>avalanche recording</td>
<td>16 (2.0%)</td>
</tr>
<tr>
<td>Chunk Boundary</td>
<td>record conversation</td>
<td>22 (2.8%)</td>
</tr>
<tr>
<td>Not Deverbal</td>
<td>transport decision-maker</td>
<td>4 (0.5%)</td>
</tr>
<tr>
<td>Proper Noun</td>
<td>china relation</td>
<td>59 (7.4%)</td>
</tr>
<tr>
<td>Consistent</td>
<td>plant failure</td>
<td>695 (87.3%)</td>
</tr>
</tbody>
</table>
Chapter 3: Resources

3.2.2 3-Way Classification

We also wish to duplicate the experiment of Grover et al., but in a domain-independent environment. The Lapata data set is not domain-specific, but had been constructed specifically to exclude prepositions and is thereby not useable for this task. Also, we wish to have text that is not strongly dependent on the selection procedure. As such, we construct a data set from open text to be annotated with prepositional object relations.

Again turning to the BNC as a lexical resource, we extract 1000 sentences at random which are then examined manually by annotators for compound nouns. About 32% of these contained at least one compound noun, much lower than the number in the biomedical domain of Grover et al. (2005). The total number of compound nouns in the data set was 464, but 119 of them (25.6%) consisted of one or more proper nouns, and were excluded so as to be consistent with the automatic system. 129 of the remaining 345 (37.4%) had one of SUBJ, DOBJ, or POBJ relations.

Another interesting point is the relative frequency of ternary and higher-arity compounds (i.e. consisting of three or more nouns), as a large number would demand some procedure for dealing with them. However, only about 7.5% of the 464 compounds were high-arity (e.g. silk jersey halter-neck evening dress), allowing us to exclude them from our research.

To examine the performance of humans on the detection task, we analysed the precision and recall of the human annotators compared to the gold-standard tags. Briefly, a standard evaluation metric for NLP as well as information retrieval and other areas is a balance between recall and precision.

\[
\text{recall} = \frac{\text{# of correct answers given by system}}{\text{total # of possible correct answers}} \quad (3.1)
\]

\[
\text{precision} = \frac{\text{# of correct answers given by system}}{\text{total # of answers given by system}} \quad (3.2)
\]

Informally, recall measures coverage and the number of false negatives. Precision measures accuracy and the number of false positives.

The three annotators had a mean precision of 92.5% and a recall of 84.8% on detecting the 345 compounds in the 1000 randomly-generated sentences. Most incorrectly tagged compounds contained adjectives (e.g. green beret or phonemic association). Their raw inter-annotator agreement over unigrams was 98.4%. The kappa coefficient, \(\kappa\) (Carletta 1996), is a statistical measure of agreement, where the raw agreement is corrected for the probability of agreeing by chance. In our case, \(\kappa\) is 83.0%. The agreement of the annotators was good (\(\kappa > 0.8\)) but not perfect, indicating the difficulty of the task.

These 345 compounds were classified according to the relations in Table 3.2: that of subject (SUBJ), direct object (DOBJ), prepositional object (POBJ), not verbal (NV: where the head does not have a verbal form), and not applicable (NA: where the
modifier is not the argument of the verbal head in an acceptable paraphrase).

We thereby collated a small data set of 129 compound nouns that occurred in a nominalisation relationship and could be analysed using a statistical interpretation method.

Table 3.2: Classes of Compounds in the Sample Data

<table>
<thead>
<tr>
<th>Class</th>
<th>Example</th>
<th>Frequency</th>
<th>Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUBJ</td>
<td>eyewitness report</td>
<td>22</td>
<td>(6.4%)</td>
</tr>
<tr>
<td>DOBJ</td>
<td>eye irritation</td>
<td>63</td>
<td>(18.2%)</td>
</tr>
<tr>
<td>POBJ</td>
<td>side show</td>
<td>44</td>
<td>(12.8%)</td>
</tr>
<tr>
<td>NV</td>
<td>scout hut</td>
<td>58</td>
<td>(16.8%)</td>
</tr>
<tr>
<td>NA</td>
<td>memory size</td>
<td>158</td>
<td>(45.8%)</td>
</tr>
</tbody>
</table>
Chapter 4

Proposed Method

We first propose an algorithm to detect compound nominalisations based on an automatically acquired broad list of deverbal nouns and the output of a chunker. We then present a system for interpreting each detected compound nominalisation by way of corpus evidence.

4.1 Detection of Compound Nominalisations

To detect compound nominalisations in open data, we examine sequences of nouns that occur within the same chunk. To do so, we use the tools mentioned in Section 3.1.2 to POS tag and then chunk parse a given sentence, and check for noun chunks with common noun modifiers immediately preceding the chunk head.

For example, the sentence *He drank most of the extensive wine list* would be tagged and chunked as *[he (PRP)]\_NP [drink (VBD)]\_VP [most (RBS)]\_AdvP [of (IN)]\_PP [the (DT) extensive (JJ) wine (NN) list (NN)]\_NP*. The noun sequence *wine list* heads a chunk, and is therefore a candidate for being a compound noun.

Next, we perform a table lookup over the head of each compound noun to see whether it is contained in the combined set of deverbal nouns mined from NOMLEX, CELEX and CATVAR, constructed in Section 3.1.3. If the head noun is found not to be deverbal, we conclude that the compound noun is not a compound nominalisation, and of the class \textit{nv} above.

For a detailed analysis of this identification method, see Section 5.1.

4.2 Interpretation of Compound Nominalisations

4.2.1 Paraphrase Tests

We analyse three paraphrase tests: Verb–Argument, Prepositional, and Participial. The Verb–Argument paraphrase test is the one described in the consideration
Chapter 4: Proposed Method

Table 4.1: Paraphrase tests for the compound nominalisation market acceptance

<table>
<thead>
<tr>
<th>Relation</th>
<th>Verb–Argument Pair</th>
<th>Prepositional</th>
<th>Participial</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUBJ</td>
<td>market accepts [ST]</td>
<td>acceptance by market</td>
<td>accepting market</td>
</tr>
<tr>
<td>DOBJ</td>
<td>[ST] accepts market</td>
<td>acceptance of market</td>
<td>accepted market</td>
</tr>
<tr>
<td>POBJ</td>
<td>[ST] accepts for market</td>
<td>acceptance for market</td>
<td>-</td>
</tr>
</tbody>
</table>

of Lapata (2002) in Chapter 2. The corpus is mined for occurrences of the modifier and verbal head in one of the subject or direct object relations. Grover et al. (2005) also search for occurrences of prepositional object relations. Examples for market acceptance are given in Table 4.1.

We propose two other paraphrase tests: that of a preposition separating the head and modifier, and that of the deverbal head occurring in a participial form before the modifier. For the Prepositional test, we note that collocations like passage by animals imply “animals pass”, speaker of language implies “[SO] speaks a language”, and operation on leg implies “[SO] operates on a leg”. For the Participle test, we note that the present participle indicates a subject usage, like passing animals for “animals pass”, and a past participle indicates an object usage, like spoken language for “[SO] speaks a language”. Examples of market acceptance for these tests are also given in Table 4.1.

The advantage of these other paraphrase tests is that discovering instances of the prepositional and participial paraphrases does not require parsed text. Determining whether the modifier and verbal head occur in a verb–argument relation requires syntactic parsing like that performed by RASP, as these relationships may be long-distance (e.g. the market would accept changes, changes were accepted by the market, or the market in the city of London, England decided that it was going to accept changes for the SUBJ relation of market acceptance). Prepositional and participial forms generally can be read directly from plain text. This avoids potential errors in the parsing, and is useful in instances where parsing cannot be performed, like Web data. We do not explicitly examine Web data in this research, but these paraphrases allow such frequencies to be derived in future work.

We are not unique in exploring the prepositional paraphrases, as Leonard (1984) uses them for semantic classification, and Lauer (1995) collects frequencies of the prepositional paraphrases, and uses them to predict semantic behaviour. The participial paraphrases have not yet been used in this manner.

There are some possible drawbacks to this: we admittedly lose phrasal verbs which legitimately take by or of, like remote control1 “[SO] controls [ST] by remote”. Searching for [a] control by [a] remote leads to counts for SUBJ.2 It is also true that,

1Remote control is a legitimate compound nominalisation, as although remote was an adjective at the time of coining, the noun remote has come into use in phrases like “Pass me the remote”.
2Other paraphrases for this compound are also possible, which do not display this phenomenon,
in some instances, these paraphrases blur somewhat in “current” English. Consider child behaviour, where a child behaves. Instances of behaviour by child are greatly overwhelmed by occurrences of behaviour of child and the genitive child’s behaviour.

In fact, genitive examples such as child’s behaviour are common in the corpus data. But, as Chomsky (1970) observes, the behaviour of the genitive is quite complex. As such, we do not include these counts in our study.

One other drawback is that counts for the prepositional object relation in the Participial test are not well-defined, as instances like operating on leg would almost certainly be termed leg as an indirect object of operate by RASP, and included in our standard frequency counts. We can therefore only use this paraphrase test on the 2-way classification test.

4.2.2 Robust Statistical Interpretation

The simple maximum likelihood estimate used by both Lapata (2002) and Grover et al. (2005) is not statistically robust when sparse vectors of many relations are under consideration. They perform smoothing to avoid this problem, but smooth only over the relations which are unattested — possibly biasing the counts.

We interpret compound nominalisations in a more statistically robust manner by using a statistical concept, confidence intervals, in a pairwise analysis over the relations. We first make the null hypothesis that the probabilities of all relations are equal, that is:

\[ P(\text{rel}_A \mid (\text{rel}_A \cup \text{rel}_B)) = 0.5 \] (4.1)

We then consider each occurrence of a verb–noun pair to be a normally-distributed binomial trial for the pair of relations under consideration.

We derive our selection preferences based on the largest confidence interval between that of the SUBJ-DOBJ interpretation (c.f. Lapata (2002)) and that of the SUBJ-POBJ comparison. This is a statistical technique which allows us to select the relation over which we are most confident in the statistical significance, as opposed to naively selecting the relation with the greatest frequency.

A Confidence Interval \( P \) is the region under a normal curve with mean \( \mu \) and standard deviation \( \sigma \) between \([\mu - n\sigma, \mu + n\sigma]\), where \( n \) is the z-score of a trial. Kenney and Keeping (1962) show that:

\[ P = \frac{2}{\sqrt{\pi}} \int_0^{n/\sqrt{\pi}} e^{-t^2} dt \] (4.2)

like “[so] controls [st] with [a] remote”. Also, nominal verbs that take these phrasings often do not form compounds: consider /st/ reeks of dishonesty “dishonesty reek”, or have active readings: consider rail transport, where “[so] transports [st] by rail” ≡ “rail transports [st]”. This tendency gives the test more logistic viability.
where \( t \equiv (x - \mu) / \sqrt{2\sigma} \), so as to normalise the curve. We observe that \( P \) is strictly increasing on \( n \), so choosing the largest confidence interval from a set is simply a matter of choosing the largest \( z \)-score.

Choosing the largest \( z \)-score is quite intuitive, as a high \( z \)-score is indicative of a low-probability event in a distribution: an event which is unlikely to occur randomly. We can see this in an informal setting: say we assume that the set of all pumpkin weights is normally distributed with mean \( \mu \) kg and standard deviation \( \sigma \) kg. We would like to choose one of a number of patches, where we expect that the selected patch has especially heavy pumpkins, based on a single trial of selecting one pumpkin from the patch. If we take a pumpkin from one patch, and that pumpkin has a weight of \( \mu + 3\sigma \) kg, this patch is less likely to be normally distributed than another patch from which we select a pumpkin with weight \( \mu + 0.5\sigma \). The confidence interval is the total area under the bell curve between \( \mu + 3\sigma \) and \( \mu - 3\sigma = 0.997 \) for the first patch and 0.383 for the second. We are confident with probability 99.7% that the first patch is not normally distributed (and contains heavy pumpkins), and only confident with probability 38.3% that the second patch is not normally distributed.\(^3\)

For us, the interpretation with the highest \( z \)-score is the one that is least likely to be binomially distributed, and holds least to the assumption that both interpretations are equally valid for that event.

For a large set, calculating the \( z \)-score exactly is very costly. Instead, we estimate the sample \( z \)-scores for our observed trial between relation A with count \( |A| = f(v_{n_2}, rel_A, n_1) \), and relation B with count \( |B| = f(v_{n_2}, rel_B, n_1) \). Recalling the null hypothesis from equation (4.1), we use the binomial approximation to the normal distribution to find that our sample mean \( \mu \) and our sample standard deviation \( \sigma \) are:

\[
\mu = \frac{1}{2} \cdot (|A| + |B|), \quad \sigma = \frac{1}{2} \cdot \sqrt{(|A| + |B|)}
\]  

(4.3)

Maximising \( P \) from equation (4.2) is a matter of integrating over \( t \), where \( t = (x - \mu) / \sqrt{2\sigma} \). Using these parameters to find the largest \( z \)-score, we again use the binomial approximation to the normal distribution to find:

\[
Z_A = \frac{|A| - \mu}{\sigma}, \quad Z_B = \frac{|B| - \mu}{\sigma}
\]  

(4.4)

For example, consider the compound nominalisation from the Lapata data set adult provision found in the BNC in the following context: ...protecting someone’s rights in the justice system (for example, appropriate adult provision). We attempt to interpret the compound nominalisation, relative to the verbal forms provide and provision. provision adult is not productive, while provide adult gives the counts seen in Table 4.2.\(^3\)

\(^3\)Heavy pumpkins are desirable, because they make the best pumpkin pie.
Based on these counts, we calculate $Z_{SD}$ (the z-score of the SUBJ-DOBJ test), e.g., as follows:

$$
Z_{SD} = \frac{|\text{SUBJ}| - \mu}{\sigma} = \frac{7 - \frac{1}{2} \cdot (7 + 5)}{\frac{1}{2} \cdot \sqrt{7 + 5}} \approx \frac{1}{1.73} \approx 0.58
$$

The highest z-score in Table 4.2 is $Z_{PS}$, corresponding to the prepositional object interpretation (i.e. the correct reading “provide for adults”).

We apply a 3-way classification via the pairwise SUBJ-DOBJ and SUBJ-POBJ classifications. It is, however, not the case that we wish to examine prepositional object interpretations in every instance. If a verb does not take any prepositional objects at all, they will not occur in the data, and calculating the SUBJ-POBJ comparison will not be meaningful, and may introduce incorrect interpretations if it has a higher z-score than the SUBJ-DOBJ interpretation. As such, we use our list of phrasal verbs constructed in Section 3.1.3 and choose to apply the SUBJ-POBJ z-scores if the verb in question coincides with one of these.

Furthermore, we require an approach for aligning our 3-way classification to the 2-way classified Lapata data set. This set was only annotated with SUBJ and OBJ tags: for instances where the data erroneously contains an item with a POBJ interpretation (e.g. adult provision), we map the class POBJ to OBJ.

### 4.2.3 The Algorithm

For a given detected compound nominalisation, we perform a number of steps to attempt to arrive at an interpretation.

First, we derive a set of verbal forms for the head using the table lookup from NOMLEX, CELEX, and CATVAR, as mentioned above, and note whether any of the forms occur in our set of prepositional verbs. If NOMLEX indicates that the head absorbs one of the possible interpretations (see Section 3.1.3), we automatically reject that interpretation in our consideration. For example, in license holder, the head absorbs the SUBJ relation, so we are left with DOBJ or POBJ (which both correspond to OBJ in the tagged data set).

NOMLEX indicates absorption for 8.9% of the head nouns in the binary set, and all but one of them lead to a correct interpretation (the one error is for woman referee,
who does not referee women, but is a woman who referees, even though SUBJ is absorbed). In the ternary set, 6.2% of the compounds have such an absorption, and again, all but one of them correctly exclude a relation (the error is for immigrant worker who is an immigrant who works, even though SUBJ is absorbed).

Lapata (2002) and Grover et al. (2005) identify a heuristic approach that is similar to this: Lapata notes that items in the set having a suffix of -er, -or, or -ant have an OBJ interpretation, while those with the suffix -ee have a SUBJ interpretation; Grover et al. note that -er compounds have a DOBJ interpretation, while -or and -our compounds have a SUBJ interpretation. Although our approach with NOMLEX does not identify all of these compounds (Lapata classifies 12.9% of her set with this heuristic), it is more principled: the absorbed relations have been identified by human annotators, while a compound may have one of these endings without demanding such an interpretation (c.f. bank transfer “bank transfers [ST]”, which is not DOBJ).

Next, we attempt an interpretation. We acquire subject, direct object, and prepositional object counts for the modifier and verbal head pair, for each individual verbal form. The z-scores \[ \{Z_{SD}, Z_{DS}, Z_{SP}, Z_{PS}\} \] are chosen as the greatest z-score for that relation across all of the verbal forms. We also calculate z-scores for the prepositional and participial paraphrase tests. We then select the interpretation having the highest z-score from across the set.

If the best z-scores for two differing interpretations are equal, we employ the simplest smoothing method from Lapata (2002): backing-off. Lapata assumes that the ratio of the counts can be approximated by backing-off to the counts of the modifier noun:

\[
P(\text{rel} \mid n_1, n_2) = \alpha \frac{f(\text{rel}, n_1)}{f(n_1)}
\]  

(4.5)

The reason for this being superior to backing-off to the verb counts is not immediately clear, however. Another valid back-off estimate would be:

\[
P(\text{rel} \mid n_1, n_2) = \alpha \frac{f(n_2, \text{rel})}{f(v_{n_2})}
\]  

(4.6)

We also examine another form of “backing-off” — that of the deverbal head counts. This time, equation (4.5) becomes:

\[
P(\text{rel} \mid n_1, n_2) = \alpha \frac{f(n_2, \text{rel})}{f(n_2)}
\]  

(4.7)

Of course, \(f(n_2, \text{rel})\) cannot be directly calculated from the corpus. Instead, we mine the BNC for sentences which contain the head \(n_2\) in a compound nominalisation, and attempt to identify these using corpus frequencies. \(f(n_2, \text{rel})\) corresponds to the count of the compounds classified with a given relation, and \(f(n_2)\) corresponds to all of the classified compounds with that head.
Regardless of the chosen method, the need for backing-off occurs quite often in practice, as some 16% of the Lapata data set has no instances of the verb–noun pair attested in the corpus, as well as 36% of the open data set.

When implementing backing-off, we disregard the earlier frequencies and instead examine the interpretation preferences for the backed-off counts from equation 4.5, 4.6, or 4.7, again using confidence intervals. The preference for the modifier noun or verbal head is the greatest z-score from the counts of all instances of that noun or verb occurring as or with a subject, direct object, or prepositional object. The preference for the deverbal head is the greatest z-score from the counts of all instances of that head occurring with a modifier for which we can provide an interpretation of subject, direct object, or prepositional object using corpus frequencies.
Chapter 5

Experimental Results

5.1 Evaluation of the Detection Method

We evaluated the performance of the detection method over the version of the Lapata data set that we had normalised for consistency (see Section 3.2.1). On this set, we were able to detect 88.8% of the data set.

Similarly, we examined how many of the annotated subj, dobj, and pobj samples from the open text data set were detected using our method. This was less comprehensive, with only 69.8% of compounds correctly detected. However, it was quite accurate in detecting all of the 464 annotator-selected compounds (including NA and NV) in the open data set, with a precision of 86.6% and a recall of 77.0%. This performance was comparable to that of the human annotators.

There were two primary causes of data instances being missed by our method. Items from either set would not have been found if the head noun was not contained as a nominalisation in our combined lexicon (e.g. decision-maker), or the input had been misanalysed by the tagger (where a noun was tagged as something other than a noun) or the chunker (where a compound noun that headed a chunk was not found to do so).

To analyse the sources of error, we examine two approaches for augmenting our collection of compounds: removing the filter of deverbal nouns so as to ascertain whether the error was in the deverbal noun listing, and examining the entire BNC to determine whether a given compound could be discovered in any sentence within the corpus, so as to indicate that the error was in processing the sentences. We also wished to examine the coverage of the filter using all of Celex, Nomlex, and CatVAR as compared to using only one of these, or excluding CatVAR from the filter. The results are summarised in Table 5.1.

Simply removing the filter for either data set is not effective: there is no performance improvement over the filtered approach for the Lapata set, and about 3% over the open data set, at the cost of many false positives. Exhaustively mining the BNC allows us to recover a small number of compounds that were incorrectly analysed by
Table 5.1: Detection accuracy over the 695-item Lapata data set and the 129-item open data set.

<table>
<thead>
<tr>
<th>Filter</th>
<th>Lapata</th>
<th>Open</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unfiltered</td>
<td>617 (88.8%)</td>
<td>94 (72.9%)</td>
</tr>
<tr>
<td>BNC Unfiltered</td>
<td>627 (90.2%)</td>
<td>113 (87.6%)</td>
</tr>
<tr>
<td>BNC Filtered</td>
<td>627 (90.2%)</td>
<td>108 (83.7%)</td>
</tr>
<tr>
<td>Filtered</td>
<td>617 (88.8%)</td>
<td>90 (69.8%)</td>
</tr>
<tr>
<td>CatVAR</td>
<td>614 (88.3%)</td>
<td>90 (69.8%)</td>
</tr>
<tr>
<td>CELEX</td>
<td>486 (69.9%)</td>
<td>31 (24.0%)</td>
</tr>
<tr>
<td>NOMLEX</td>
<td>501 (72.1%)</td>
<td>43 (33.3%)</td>
</tr>
<tr>
<td>CELEX + NOMLEX</td>
<td>594 (85.5%)</td>
<td>53 (41.1%)</td>
</tr>
</tbody>
</table>

the tagger or chunker in the given sentence; for example, *group communication*, which had not been found to be the head of a noun chunk due to coordination ambiguity in the phrase *group communication and community activism*.

The remaining 10% of the Lapata data set and 12% of the open data set is composed of singleton compounds that were misparsed, or more productive compounds where the lexical tools make common mistakes (e.g. *subject* in compounds such as *subject guidance*, which is generally tagged as an adjective in a pre-nominal position).

As for the relative performances of the lexical resources, CatVAR performs well even on its own, while NOMLEX and CELEX perform poorly by themselves, but fairly well in combination for the Lapata data set, although still more than 25% worse than CatVAR on the open data set. The reason for this is clear: in the collection of the Lapata data set, candidates were proposed using CELEX and NOMLEX, and annotators selected correct examples. For the open data set, sentences were taken entirely at random, and annotators were asked to detect the compounds themselves.

Finally, we develop a simple metric with which to evaluate the detection procedure over the open data set, which was annotated with NA and NV compounds. The relation NV was selected when a candidate noun sequence was discovered, but where the head was not in the list of deverbal nouns, or was in the list, but was not attested as a verb in the corpus (thereby making all of the smoothed probabilities equal 0). The relation NA was selected when a candidate noun sequence had a head that occurred in the deverbal list, but where no instances of the verb–noun pair were attested in the corpus. Otherwise, the compound is identified as a compound nominalisation, and takes one of SUBJ, DOBJ, or POBJ relations.

The majority of errors in assigning an interpretation in detection were mistakes made by the POS tagger (e.g. calling *covers* a verb in *leopardskin seat covers*) and mistakes made by the chunker (e.g. *year contract* in *$5 million a year contract*), which were often caused by poorly punctuated sentences in the corpus *citizens charter* instead of *citizen’s charter* in *Ministers’ views were set out in the citizens charter*).

As for the various relations in our set above, we can also compare the relative
performance of the lexical resources in the domain. We are interested in whether removing CATVAR gives a precision gain, at the cost of recall. Errors cascade in precision and recall, so that a noun incorrectly given a verbal form causes a false negative in NV and a false positive in NA, and so on.

Table 5.2: Precision and recall of compound nominalisation detection in the open data set for the deverbal list filter and the filter without CATVAR, with annotator performance for comparison.

<table>
<thead>
<tr>
<th></th>
<th>Filtered</th>
<th>CELEX + NOMLEX</th>
<th>Annotators</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>NV</td>
<td>58.4%</td>
<td>77.6%</td>
<td>22.2%</td>
</tr>
<tr>
<td>NA</td>
<td>65.1%</td>
<td>53.2%</td>
<td>26.5%</td>
</tr>
<tr>
<td>Nominalisation</td>
<td>57.1%</td>
<td>49.6%</td>
<td>72.3%</td>
</tr>
</tbody>
</table>

Removing CATVAR from the deverbal head filter causes a precision increase in the identification of compound nominalisations, agreeing with our expectations according to nature of CATVAR. Otherwise, performance greatly decreases (the small recall gain for NV occurs because the full filter causes some NV compounds to be classified as NA). Human performance is unsurprisingly better, except in NV, where the deverbal list performs more reliably.

5.2 2-Way Classification

The data set from Lapata had 695 compound nominalisations: of these, 258 had a SUBJ interpretation and 437 had a OBJECT interpretation. Thus, the baseline of choosing the OBJECT relation each time has a performance of 62.9%.

Figure 5.1 shows the performance for the three paraphrase tests in the 2-way classification task, when applying the deverbal filter to decide for which compounds we consider the POBJ relation. We use Verb–Noun pair counts (VN), Participial paraphrases (Part), and Prepositional paraphrases (Prep). For each of these, we can back-off to the preference according to the verbal head, modifier noun, or deverbal head. We contrast the performance of these tests with the performance of the baseline (Default), the upper bound (Upper Bound) of inter-annotator agreement that was calculated by Lapata to be 89.7%, and the interpretation preferences when used without the paraphrase tests (IP).

The prepositional or participial paraphrases when used on their own do not perform significantly better than the baseline ($\chi^2 = 1.97, p \leq 0.2$). This is not overly surprising, as coverage over the data set is quite poor: only 40% could be given an interpretation using one test, and 58% for both tests — far lower than the 84% for the verb–noun pairs.
Chapter 5: Experimental Results

Figure 5.1: Disambiguation Accuracy for the 2-Way Classification Task using the Phrasal Verb Filter

Figure 5.2: Disambiguation Accuracy for the 2-Way Classification Task considering no Verbs as Phrasal
The verb–noun counts (70.6% accuracy) are significantly better than the baseline \( \chi^2 = 9.45, p \leq 0.01 \), and also slightly improve upon the figures recorded by Lapata for backing-off — namely, 69.6% over the test set and 68.0% over the entire data set.

Interestingly, backing-off to the deverbal head is consistently slightly better than backing-off to the modifier noun or verbal head, at the cost of extra examinations of the corpus. Adding the prepositional and participial paraphrases do not greatly reduce performance.

Also of interest is the fact that considering no verbs as phrasal (Figure 5.2) is uniformly slightly better than considering verbs as phrasal according to the filter (Figure 5.1), which is itself slightly better than considering all verbs as phrasal (Figure 5.3). This occurs because when we identify a verb as phrasal, we examine the \text{SUBJ-POBJ} z-scores — as compounds having prepositional relationships between the head and modifier nouns were specifically excluded from this data set, there is an artificially low number of them. As such, applying the z-score test may give an erroneously high weighting to the \text{SUBJ} relation, as we would compare its relatively high frequencies with comparatively low \text{POBJ} frequencies. Our best figure of 72.3% accuracy was obtained by regarding no verbs as phrasal.

### 5.3 3-Way Classification

Our collated data set had a baseline of 48.8%, that of selecting DOBJ each time. Figure 5.4 shows the results of our experiments using the phrasal verb filter, similarly to the two-way classification. We also consider the results of applying the \text{POBJ} test regardless of whether we have identified the verb as phrasal in Figure 5.6, and never applying the \text{POBJ} test in Figure 5.5.
Chapter 5: Experimental Results

Figure 5.4: Disambiguation Accuracy for the 3-Way Classification Task using the Phrasal Verb Filter

Figure 5.5: Disambiguation Accuracy for the 3-Way Classification Task considering no Verbs as Phrasal
Chapter 5: Experimental Results

Figure 5.6: Disambiguation Accuracy for the 3-Way Classification Task considering all Verbs as Phrasal

Again, the prepositional and participial paraphrases do not perform significantly better than the baseline ($\chi^2 = 0.39, p \leq 1$), while the verb–argument counts do perform better ($\chi^2 = 4.01, p \leq 0.05$), and slightly improve upon the figures reported by Grover et al. (2005) using frequency counts and affixes.

In this case, backing-off to the modifier noun proves better than to either the verbal or deverbal head, and the further paraphrases do not improve the performance of the frequency counts. Also of note is the fact that the deverbal head preferences on their own perform quite poorly here, in stark contrast to their performance on the binary task.

Contrastingly, using the phrasal verb filter in the 3-way classification task (Figure 5.4) performs better than both of considering all verbs as phrasal and no verbs as phrasal. This is intuitive, as we only wish to apply the SUBJ-POBJ $z$-score comparison for those verbs which can legitimately take a prepositional object. Since this data set is seeded with those compounds, identifying them improves performance. Applying the phrasal verb filter using verb–noun frequency counts and backing-off to the modifier noun gives us our best performance of 57.4% in this task.

The fact that the performance of the algorithm (72% and 57% for the two tasks) does not match the state-of-the-art performance by Lapata (2002) and Grover et al. (2005) (86% and 77% respectively) does not worry us too much, as they match the simple performance of the works, and these works included a variety of class-based smoothing tasks, contextual features, and machine learning tools. We attest that we can similarly implement the other smoothing techniques and analysis of the context, and machine learning to achieve similar results. An discussion is future work is given in Section 6.2.
Chapter 6

Discussion

6.1 Summary

We presented a method for detecting compound nominalisations in open data, and deriving an interpretation for them. Discovering the semantic relation between the modifier and head noun in a compound nominalisation was first construed as a two-way classification task between an underlying subject or object semantic relation between a head noun and its modifier, and second as a three-way task between subject, direct object, and prepositional object relations.

The detection method achieved about 89% recall on a data set annotated by way of CELEX and NOMLEX, and about 70% recall on a data set extracted at random from the BNC, with 77% recall on detecting a more general set of compound nouns from this data. We also examined the effect of CATVAR, a broader, automatically constructed resource: it decreased precision, as we expected, but the increase in recall justified this.

The interpretation method achieved about 70% accuracy in the two-way task, and 57% in the three-way task, using a statistical measure based on z-scores — the confidence interval — in selecting one of the relations.

We investigated the performance of three paraphrase tests across the BNC: corpus frequencies of the modifier noun and verbal head pair, corpus frequencies of prepositions separating instances of the head and modifier nouns, and corpus frequencies of the verbal head occurring as a participial adjective connected to the modifier noun.

Interpretation preferences were examined: of the modifier noun independent of the verbal head and of the head as both verbal and deverbal, and these were used for backing-off the paraphrase counts. As well, a phrasal verb filter was applied to decide for which compounds to consider the prepositional object relation.

The best-performing interpretation method for the two-way classification task was using deverbal interpretation preferences for the verb–noun counts without considering the prepositional object relation. Contrastingly, the best-performing method for the three-way classification task was using verb–noun counts backed-off to the
modifier, while applying our phrasal verb filter. Applying the phrasal verb filter unsurprisingly helped in the three-way task, as the data set did contain prepositional objects, which were excluded from the data set in the two-way interpretation task.

The proposed method had the advantage over previous research in that it can act over open data to detect and interpret compound nominalisations. It was able to automatically capture the majority of types in the data sets, with some limitations across the resources, due mostly to the generosity of CatVAR in classifying a deverbal noun. These often do not lead to productive compounds in our corpus, however, and can be extracted as such.

Our method also extended the scope of the interpretation of nominalisations away from the need for pre-filtered or hand-tuned data, such as was necessary for the two statistical works of interpretation using corpus frequencies, that of Lapata (2002) and Grover et al. (2005). This method does not presuppose a hand-tuned parser, and can operate more or less independently of the domain in which it is used, as we demonstrated in sampling random sentences over a balanced corpus.

### 6.2 Future Work

To improve upon the most obvious limitation of the work — the fact that the results fail to improve on the state-of-the-art — the natural continuation would be to follow the works of Lapata (2002) and Grover et al. (2005), and apply further techniques to a machine learner so as to improve accuracy of interpretation. One method for this is to incorporate class-based smoothing, as Lapata noted that it performs better than the backed-off smoothing which we examined. Another method would be to integrate contextual features from the sentence in which a compound occurs, as the pragmatic dependency of a compound on context has been well-documented in works such as Copestake and Lascarides (1997).

A criticism of the work would be that we have contracted the 13-way classification performed by Grover et al. to a seemingly simpler 3-way classification. This disparity could be balanced using an approach to select a preposition for each compound labelled as a prepositional object relation. This extension is quite natural: the phrasal verb filter we apply contains expected prepositions by which a verb could be followed, and we could make use of the corpus frequencies for the prepositional object relation of the Verb–Argument and Prepositional paraphrase tests that we performed.

Finally, one more natural extension of this work would be to incorporate data from the World Wide Web, as we would expect that this would improve the accuracy of the Prepositional and Participial paraphrase tests, in that a greater volume of data would be considered — and we would expect this to be congruent with English usage (although there are certainly caveats to using Web data for robust linguistic consideration). The fact that these paraphrase tests can generalise to open data, while Verb–Argument paraphrases cannot, makes an analysis of this procedure even
more desirable.

6.3 Two Final Words

To conclude, let us recall *yak fat*, and follow our proposed method in deriving an interpretation. There are two deverbal forms for the head: *fat* (c.f. *to fat a calf*) and *fatten*. The verb *fat* is not attested in the corpus data. The verb *fatten* is attested, but with no instances of “yak fattens [ST]”, “[SO] fattens yak”, and so on. Our algorithm therefore concludes that *yak fat* is NA — agreeing with our intuition on the matter (regardless of whether or not the yak has been fattened earlier).

This, we contend, is our novel contribution: the obvious deficiency of many compound noun and compound nominalisation interpretation systems in handling *yak fat* may be done away with, so that one might truly give due consideration to *yak fat*. 
Bibliography


Fan, James, Ken Barker, and Bruce Porter. 2003. The knowledge required to interpret noun compounds. Technical Report UT-AI-TR-03-301, University of Texas at Austin, Department of Computer Sciences.


Appendix A

Samples of Lexical Resources

We give the entry of the noun *acceptance* from each of Celex, Nomlex, and CatVar to illustrate the morphological information of which we make use in determining the verbal form of a given deverbal noun.

In Celex, we have an index number, word form, frequency count, and morphological information:

```
213\acceptance\411\C\1\N\N\N\N\Y\accept+ance\3x
\SA\N\N\N\#\N\N\SA\((accept\)\[V\],(ance)\[N\|V\.]\)\[N]\N\N\N
```

In Nomlex, we have a word form, verbal form, nominal type (which describes whether some relation is absorbed), and a detailed list of the complements (post-nominal modifiers) taken by the word:

```
(NOM :ORTH "acceptance"
 :PLURAL "acceptances"
 :PLURAL-FREQ "not rare"
 :VERB "accept"
 :NOM-TYPE ((VERB-NOM))
 :VERB-SUBJ ((N-N-MOD)
  (DET-POSS)
  (PP :PVAL ("by")))
 :SUBJ-ATTRIBUTE ((COMMUNICATOR))
 :VERB-SUBC ((NOM-NP :SUBJECT ((N-N-MOD)
  (DET-POSS)
  (PP :PVAL ("by")))
 :OBJECT ((DET-POSS)
  (N-N-MOD)
  (PP :PVAL ("of")))
 :REQUIRED ((OBJECT)))
 (NOM-NP-PP :SUBJECT ((N-N-MOD)
```

39
In CatVar, we simply have a list of words (and their part-of-speech types) that have the same stem according to the Porter Stemmer, and have thus been clustered together:

accept_V#acceptor_N#accepted_AJ#acceptance_N#acceptant_AJ#
accepting_AJ#acceptive_AJ#acceptable_AJ#acceptably_AV#
acceptation_N#acceptability_N#acceptableleness_N