Statistical Interpretation of Compound Nominalisations

Jeremy Nicholson[†] and **Timothy Baldwin**[‡]

[†‡] Dept. of Computer Science and Software Engineering
University of Melbourne
Victoria 3010 Australia

[‡] NICTA Victoria Lab
University of Melbourne
Victoria 3010 Australia

{jeremymn,tim}@cs.mu.oz.au

Abstract

This paper presents a method for detecting compound nominalisations from open data, and providing a semantic intepretation. It uses a statistical model based on confidence intervals over frequencies extracted from a large, balanced corpus. Using three paraphrases of the given compound nominalisation, and interpretation preferences of its components, the algorithm achieves about 70% accuracy in classifying the semantic relationship as one of SUBJECT, and OBJECT, and 57% between SUBJECT, DIRECT OBJECT, and PREPOSITIONAL OBJECT.

1 Introduction

Compound nouns have been a thorn in the side of computational linguistics since its inception, as it is almost impossible to avoid the issue of compound noun interpretation in any language task with a semantic or lexical semantic dimension. For example, an information extraction task may need to predict the semantic divergences between the compound nouns *news print* ("cheap paper on which newspapers are printed"), *thumb print* ("impression of the pattern on a thumb"), and *colour print* ("printed matter in colour").

Interpreting these divergences has become yet another instance of disaccord between the empiricists and rationalists of theoretical and computational linguistics. While the rationalists contend that compound nouns can be semantically described by some small, hand-crafted set of relations, the empiricists point to discordant examples which defy such natural sets, and rely on data sets for the necessary description. The rationalists then call this approach biased and brittle.

Needless to say, the argument is not one which will be resolved here. What we do hope to shed some light on is the applicability of an empiricist approach to the interpretation of an important subclass of compound nouns: **compound nominalisations**, or those compound nouns whose head derives from a verb. One example is *pattern generation*, where *generation* is derived from *generate*. These compounds tend to have a clearer semantic definition and are well-suited to techniques based on corpus statistics.

We propose a method for taking English[1] raw text input, detecting the compound nouns within, and applying a semantic interpretation for the compound nominalisations.

Section 2 details some of the previous work performed in the task. Section 3 outlines the resources used in our study. Section 4 describes the method we applied, with an analysis of the results in Section 5, and a discussion in Section 6.

2 Background

The first notable work on interpreting compound nouns focussed on the development of discrete semantic classes with which to classify all compound nouns: Levi (1978) proposed a nine-way classification, Warren (1978) identified a basic set of twelve paraphrases, Leonard (1984) developed an eight-way typology, and Finin (1980) put forth a much larger number of possible roles. This early research established a basic dichotomy in classification approaches: semantic class-based approaches (e.g. *linguistics textbook* $\equiv$ TOPIC(*textbook*) = *linguistics*) and syntactic paraphrase-based approaches (e.g. *linguistics textbook* = *textbook* on *linguistics*). A basic assumption underlying both approaches is that all compound nouns can be classified according to a finite set of relations, although researchers rarely agree on the number and ele-

[1] Although this paper focusses on English compounds, the phenomenon occurs readily in other languages, such as German, Modern Greek, Japanese, and Welsh. With some caveats for morphology and syntax, our concepts still apply.

Proceedings of the Australasian Language Technology Workshop 2005, pages 152–159,
Sydney, Australia, December 2005.

ments. Many identify nominalisations as a subclass, or include subjective and objective relations, which imply deverbal forms.

Automatically interpreting compound nouns has usually taken one of two approaches: one conceptual, in that the modifier fills a slot according to the concept of the head (Finin, 1980; McDonald, 1982); the other rule-based, in that the relation is selected by the first applicable rule taken from a series (Leonard, 1984; Vanderwende, 1994). Few, however, include evaluation of their systems — notable exceptions are the rule-based systems of Leonard (1984), who achieves 76% accuracy over a training set, Vanderwende (1994), who reports 52% over a test set for thirteen relations, and the concept-based work of Rosario and Hearst (2001), scoring 70% in a specific domain.

The first notable work on statistically interpreting compound nouns was that of Lauer (1995), who used prepositional paraphrases as an interpretation model, similar to those of Leonard (1984). Another useful element of this work was that of automatic compound noun bracketing, which has allowed work since to dismiss ternary and higher-arity compounds as a solved problem, and reduce the task to considering interesting binary compounds only. Lauer's interpretations get 47% accuracy over a set of twelve paraphrases, explicitly excluding subjective and objective relations. Lapata (2002), performs much better using a combination of a probabilistic method with a decision tree learner, to achieve 86% accuracy, albeit on a far better-defined and far simpler two-way classification task. Moldovan et al. (2004) get an F-score of 43% using statistical techniques across a wide semantic space. Finally, Grover et al. (2005) use a technique similar to that of Lapata to achieve 77% accuracy in a domain-specific setting over a broader thirteen item set.

Lapata (2002) and Grover et al. (2005) provide the usual statistical interpretation of a given compound nominalisation: corpus frequencies are derived for the verb-argument pair of a given deverbal head and modifer across a list of semantic relations. The selected relation is that which has the most attested instances in the corpus. To counter the problem of data sparseness, they examine the influence of backing-off, class-based, and distance-weighted smoothing, which are not found to perform significantly differently.

Automatic detection of compound nouns and compound nominalisations has had much less analysis than their interpretation, partly because detecting simple compounds is usually considered trivial, and reliably detecting nominalisations requires a semantic interpretation. Consider *corner piece* (of a puzzle): a naive system can certainly identify this contiguous noun sequence in an NP as a compound noun, but correctly dismissing a nominalisation interpretation requires semantic analysis of the absurdity of "*corner pieces [ST]", "*[ST] pieces the corner", "??[SO] pieces together the corner", and so on. Nonetheless, examination performed by Leonard (1984) notes the increasing frequency of compound noun usage over the past 250 years, and Grover et al. (2005) note that 72% of a small sample of sentences contained one or more compound nouns in a domain in which they are prevalent. 35% of these are compound nominalisations.

3 Resources

3.1 Tools

Similarly to both Lapata (2002) and Grover et al. (2005), we use the British National Corpus (BNC: Burnard (2000)), but only the 90M-token written component. We parsed the corpus using RASP (Briscoe and Carroll, 2002), a tag sequence grammar-based stochastic parser, in order to extract the corpus frequencies for use in disambiguation. We use tagging and chunking tools built with fnTBL 1.0 (Ngai and Florian, 2001) over the BNC independently, and for use in the detection of compound nominalisations.

3.2 2-Way Classification

We attempt to replicate the experiment performed by Lapata (2002), who manually extracted and annotated a sample of 796 binary compound nominalisations out of about 170,000 candidates automatically identified in the BNC.

In the original Lapata data, the underlying verb form of the head noun was uniquely identified using a combination of CELEX (Burnage, 1990) and NOMLEX (Macleod et al., 1998) data, sometimes resulting in sub-optimal results (e.g. the base verb of *question* is given as *quest*). 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 (Habash and Dorr, 2003) 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. *divination* is listed as all of *divine*, *divinise* and *divinize*).

To validate the data for consistency, we removed those nominalisations which were not associated with a context sentence in the data set, those which did not occur in the same chunk in that sentence, according to the chunker above, those for which we did not find a verbal form (e.g. *decision-maker*), and those consisting of one or more proper nouns. We were left with 695 items which were classified as one of SUBJ or OBJ interpretations.

3.3 3-Way Classification

We also attempt to replicate the experiment performed by Grover et al. (2005), but in a domain-inspecific environment. To do so, we extract 1000 sentences randomly from the BNC which are then examined for compound nouns. About 32% of these contained at least one compound, much lower than the number in the biomedical domain of Grover et al. (2005).

We overtly exclude compounds which were of higher arity than two, (e.g. *silk jersey halterneck evening dress*) and those consisting at least in part of proper nouns, similarly to the two-way task. They represented about 7.5% and 25% of the total compounds in the sample space respectively. The rest we classified according to the relations in Table 1: that of subject, direct object, prepositional object, not verbal (where the head does not have a verbal form), and not applicable (where the modifier is not the argument of the verbal head in an acceptable paraphrase).

We thereby collated a small data set, that of 129 items which occurred in a nominalisation relationship. The kappa coefficient, where the raw agreement is corrected for chance agreement (Carletta, 1996), between three annotators was $\kappa = 0.83$ ($N = 129, k = 3$) in detecting noun compounds.[2] This corresponds to a unigram agreement between the judges of 98.4%. Compared to the gold standard, the annotators had a mean precision of 92.5% and recall of 84.8% in detection of compound nouns, and 78.8% mean accuracy in semantic classiification of the compound nominalisations within.

4 Proposed Method

We propose an algorithm to detect compound nominalisations based on the output of a chunker, and then interpret each detected compound nominalisation by way of corpus evidence.

[2] $\kappa \geq 0.8$ indicates good agreement.

4.1 Detection of Compound Nominalisations in Open Data

To detect compound nominalisations in open data, we examine sequences of nouns that occur with the same chunk. Hence, we chunk parse a given sentence, and check for noun chunks with common noun modifiers immediately preceding the chunk head.

Next, we perform a table lookup over the head of each compound noun to check if it is contained in the combined set of deverbal nouns mined from NOMLEX, CELEX and CATVAR. If the head noun is not found to be deverbal, we conclude that the compound noun is not a compound nominalisation.

While our combined set of deverbal nouns provides excellent coverage, it suffers from low precision, largely because of CATVAR lacking explicit word-to-word derivational information. That is, we are able to access clusters of words which share the same stem, but have no way of checking for direct derivational correspondence between a given noun and verb. As a result, the output of the filter tends to have excellent recall, but diminished precision. We combat this effect by additionally checking for the plausibility of a subject or object interpretation against corpus data and thresholding over the probability for each interpretation type.

We evaluated the detection method by attempting to re-extract the gold standard two-way classification data from the BNC, and over our annotated data set for the three-way classification.

4.2 Paraphrase Tests

Lapata (2002) and Grover et al. (2005) provide the usual statistical interpretation of a compound nominalisation: that of the most attested relation in corpus frequencies for the verb-argument pair.

Other paraphrases are also used, however: as Lauer (1995) notes, the system by Leonard (1984) has paraphrasing as a goal, whereby *mountain vista* is interpreted via paraphrasing as *vista of a mountain or mountains*. Lauer himself also attempts to paraphrase compounds based on corpus statistics.

We notice that the other direction is also productive: instances of *vista of mountains* occur in the corpus, and they supply evidence for the reading "view mountains". We can thus form paraphrase tests to influence our interpretations.

Table 1: Classes of Compounds in the Sample Data

Class	Example	Frequency	
SUBJ	*eyewitness report*	22	(6.4%)
DOBJ	*eye irritation*	63	(18.2%)
POBJ	*side show*	44	(12.8%)
NV	*scout hut*	58	(16.8%)
NA	*memory size*	158	(45.8%)

We can search plain text for instances where the head noun is separated from the modifier by the preposition, and count corpus frequencies from these. For the preposition *by*, we assume a subject interpretation, as samples like *passage by animals* for *animal passage* imply that it is the animals that are passing. For the preposition *of*, we assume an direct object interpretation, as samples like *speaker of language* for *language speaker* imply that there is someone or something that speaks the language. Other prepositions separating the head and modifier contribute to a prepositional object interpretation, as samples like *operation on leg* imply that someone or something operates on a leg.

A second, related paraphrase test is for adjectival participles of the verbal form of the head connected to the modifier noun. In this case, present participles like *[the] passing animals* contribute to the subject interpretation, and past participles like *[a] spoken language* contribute to the direct object interpretation. There are no sensible cases in this test for prepositional objects, as *?operating on leg* would almost certainly be termed an indirect object relation by RASP, and included in our standard frequency counts.

A possible drawback to the prepositional test is losing phrasal verbs which legitimately take *by* or *of*. As well, paraphrases in this form 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 *child's behaviour*. Despite examples such as this, this test is indicative of most paraphrases in the language.

4.3 Statistical interpretation

We interpret compound nominalisations by considering pairwise SUBJ vs. DOBJ and SUBJ vs. POBJ. First, we make the null hypothesis that the probabilities of all relations are equal, i.e. $P(rel_A \mid (rel_A \cup rel_B)) = 0.5$. We then consider each occurrence of a verb–noun pair to be a normally-distributed binomial trial for the two relations under consideration.

We derive our selection preferences based on the largest confidence interval between that of the SUBJ-DOBJ comparison (as Lapata (2002)), and that of the SUBJ-POBJ comparison.

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

$$P = \frac{2}{\sqrt{\pi}} \int_0^{n/\sqrt{2}} e^{-t^2} dt \qquad (1)$$

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.

For a large set, calculating the z-score exactly is very costly. Instead, we estimate the sample z-scores for our observed trial by way of the binomial approximation to the normal distribution. Considering two relations at a time, having equal probability from the null hypothesis, our sample mean is the arithmetic mean of the frequencies, and our sample standard deviation is half of the square root of twice this number. The z-scores are then: $Z = \frac{f-\mu}{\sigma}$.

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 2.

From this, the highest z-score is Z_{PS}, for the prepositional object interpretation, which coincides with the correct reading "provide for adults", and the gold-standard tag OBJECT. The correct reading here would not have been

Verb-noun	SUBJ	DOBJ	POBJ	Z_{SD}	Z_{DS}	Z_{SP}	Z_{PS}
adult provision	7	5	18	0.58	-0.58	-2.20	2.20

Table 2: Z-scores for sample verb–noun pairs extracted from the BNC

captured by a simple SUBJ-DOBJ comparison, as Lapata would perform.

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 construct a list of prepositional verbs derived from WordNet, COMLEX, the ERG, and the Longman Phrasal Verb Dictionary, and we can choose to apply the SUBJ-POBJ z-scores if the verb in question coincides with one of these.

4.4 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, we automatically assume that the opposite interpretation is correct. For example, in *license holder*, the head absorbs the SUBJ relation, so we are left with OBJ. In *business employee*, *employee* absorbs the DOBJ relation, so we consider only SUBJ or POBJ.

This occurs for 8.9% of our compounds in the binary set, with all but one of them accurate (*woman referee*, who does not referee women, but is a woman who referees). In the ternary set, 6.2% of the compounds have such an interpretation, again with all but one accurate (*immigrant worker*, who is an immigrant who works).

This is similar to the suffix indications used by Lapata (2002), and the affixes used by Grover et al. (2005). Lapata identifies 12.9% of her set as having one of *-er, -or, -ant* suffixes, leading to an OBJECT interpretation, or an *-ee* suffix, leading to a SUBJ interpretation. Grover et al. identify that *-er* affixes receive an DOBJ relation, and *-or, -our* a SUBJ. These are also features to a decision tree learner. NOMLEX only captures a portion of these, but a head can have one of the endings without demanding such an interpretation. For example, *transfer* ends with *-er*, but does not take a DOBJ relation in *bank transfer*.

Next, we normalise the lemmas and attempt an interpretation. We acquire subject, direct object, and prepositional object counts for the modifier and verbal head pair, for each individual verbal form, as well as counts for the prepositional and participial paraphrase tests. We then calculate each of the four z-scores $Z_{SD}, Z_{DS}, Z_{SP}, Z_{PS}$ for the three tests, and select the interpretation having the highest z-score from the set.

If the best z-score 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:

$$P(rel \mid n_1, n_2) = \alpha \frac{f(rel, n_1)}{f(n_1)} \qquad (2)$$

The reason for this being superior to backing-off to the verb counts is not immediately clear, so we compare backing-off to those counts as well. We also examine another form of "backing-off" — that of the deverbal head counts, which cannot be directly examined from the corpus. Instead, we mine the BNC for sentences which contain the head in an instance which we can interpret using corpus frequencies, and count frequencies based on the number interpreted as SUBJ, DOBJ or POBJ.

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 to in the corpus, as well as 36% of the open data set.

We implement backing-off by examining the interpretation preferences, 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.

5 Experimental Results

5.1 Detection

We evaluated the detection method first by running it over the contextual sentences and seeing how many of the compound nominalisations in the normalised Lapata data set were detected by our method. On this data set, we were able to detect 88.8% of the instances.

On the open text data set, our algorithm detected 69.8% of the SUBJ, DOBJ, and POBJ compounds. The more general compound nouns were detected with a precision of 86.6% and a recall of 77.0%, comparable to the human annotators.

The primary causes of data instances being missed by our method were that 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 chunker. Many of the latter errors were caused by poorly punctuated sentences in the corpus (e.g. *citizen charter* in *Ministers ' views were set out in the citizens charter*), with some mistakes made by the POS tagger (e.g. calling *covers* a verb in *leopardskin seat covers*).

As for the various relations in our set above, the detection algorithm discovers NV relations with a precision of 58.4% and a recall of 77.6%, NA relations with a precision of 65.1% and a recall of 53.2%, and SUBJ/ DOBJ/ POBJ with a precision of 57.1% and a recall of 49.6%.

Recalling that CATVAR lacks derivational information, and therefore tends to broaden coverage at the cost of precision, we examine the detection procedure without CATVAR in the deverbal filter. This algorithm discovers NV with a precision of 22.2% and a recall of 82.8%, NA with a precision of 26.5% and a recall of 5.7%, and nominalised relations with a precision of 72.3% and a recall of 26.4%. Indeed, the precision in detecting nominalisations increases, but at a substantial cost of recall.

Errors cascade in this definition, so that a noun incorrectly given a verbal form causes a false negative in NV and a false positive in NA, and so on. This explains the loss of precision in the latter classifier, when a compound is classified as NV from not recognising that the head is verbal and is an NA relation.

These figures occur for a baseline classifier, where NV implies that the noun was not in our deverbal list, NA implies that the verb-noun pair was not attested in the corpus, and one of SUBJ, DOBJ, or POBJ otherwise.

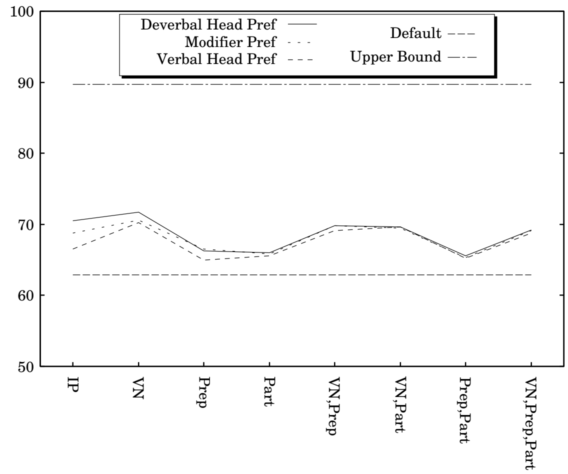

Figure 1: Disambiguation Accuracy for the 2-Way Classification Task

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. So, the baseline of choosing the OBJECT relation each time has a performance of 62.9%.

Figure 1 shows the performance for the three paraphrase tests: using verb–noun pair counts (*VN*), using participial paraphrases (*Part*), and using prepositional paraphrases (*Prep*). We can back-off to the verbal head, modifier noun, or deverbal head in each case. We also contrast these with the performance of the baseline (*Default*) and the interpretation preferences when used without the paraphrase tests (*IP*).

The prepositional and 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.

The verb–noun counts 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 examination of the corpus. Also, the best

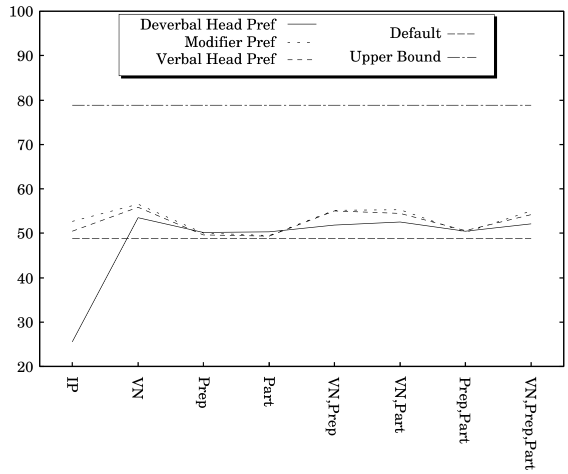

Figure 2: Disambiguation Accuracy for the 3-Way Classification Task

performance occurs for the verb–noun pairs using backing-off to the deverbal head, but including the paraphrases does not give results that are significantly different to these.

5.3 3-Way Classification

Our collated data set had a baseline of 48.8%, that of selecting DOBJ each time. Figure 2 shows the results of our experiments, similarly to the two-way classification.

Again, the later paraphrases do not perform significantly better than the baseline ($\chi^2 = 0.39, p \leq 1$), while the verb–counts do perform better ($\chi^2 = 4.01, p \leq 0.05$)), and slightly improve on 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.

6 Discussion

We presented a method for detecting compound nominalisations and deriving an interpretation in open data for a two-way classification task between an underlying subject or object semantic relation between a head noun and its modifier, and for a three-way task between subject, direct object, and prepositional object relations. This 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 the relations. We investigated the performance of three paraphrase tests across the BNC: the frequency of the modifier noun and verbal head pair, the frequency of prepositions separating instances of the head and modifier nouns, and the frequency of the verbal head occuring as a participial adjective connected to the modifier noun. We also examined the interpretation preferences of the modifier noun independent of the verbal head and the head as both verbal and deverbal, and used these for backing-off the paraphrase counts. Interestingly, the performance of the different tests was not altogether dissimilar: the best-performing set for the two-way classification task was the deverbal interpretation preferences for the verb–noun counts, and verb–noun counts backed-off to the modifier in the three-way task.

This method was useful in that it can act over open data to detect and interpret compound nominalisations. It performed comparably to the human annotators in detecting compound nominalisations, and the generosity of CatVar in classifying a deverbal noun was avoided to some extent by thresholding over the probabilities of interpretation in the corpus frequencies.

Our method also extended the scope of the interpretation of nominalisations away from the need for pre-filtered data, such as was necessary for the two statistical works of interpretation using corpus frequencies, that of Lapata (2002) and Grover et al. (2005). Our method also does not presuppose hand-crafted parsed data, which is necessary for both of these investigations. It also 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.

The utility of the two proposed paraphrases tests, using prepositions and participles, is that they do not require parsed data to acquire corpus frequencies. This allows us to take counts for these tests from the Web, which we believe will alleviate the data sparseness problem for these tests to some extent.

The fact that the performance of the algorithm (70% and 57% for the two tasks) does not match the state-of-the-art performance by these works (86% and 77% respectively) does not worry us too much, as they match the simple performance of the works, and these better figures included a variety of class-based smoothing tasks, contextual features, and machine learning tools.

Acknowledgements

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

References

Ted Briscoe and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proc. of the 3rd International Conference on Language Resources and Evaluation*, pages 1499–1504, Las Palmas, Canary Islands.

Gavin Burnage. 1990. CELEX: A guide for users. Technical report, University of Nijmegen.

Lou Burnard. 2000. *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Tim Finin. 1980. The semantic interpretation of nominal compounds. In *Proc. of the First National Conference on Artificial Intelligence*, pages 310–315, Stanford, California. AAAI Press.

Claire Grover, Mirella Lapata, and Alex Lascarides. 2005. A comparison of parsing technologies for the biomedical domain. *Journal of Natural Language Engineering*, 11(01):27–65.

Nizar Habash and Bonnie Dorr. 2003. A categorial variation database for English. In *Proc. of the 2003 Human Language Technology Conference of the North American Chapter of the ACL*, pages 17–23, Edmonton, Alberta.

John F. Kenney and E. S. Keeping, 1962. *Mathematics of Statistics, Pt. 1*, chapter 11.4, pages 167–9. Van Nostrand, Princeton, New Jersey, 3rd edition.

Maria Lapata. 2002. The disambiguation of nominalizations. *Computational Linguistics*, 28(3):357–388.

Mark Lauer. 1995. *Designing Statistical Language Learners: Experiments on Noun Compounds*. Ph.D. thesis, Macquarie University, Sydney, Australia.

Rosemary Leonard. 1984. *The Interpretation of English Noun Sequences on the Computer*. Elsevier Science, Amsterdam.

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

Catherine Macleod, Ralph Grishman, Adam Meyers, Leslie Barrett, and Ruth Reeves. 1998. NOMLEX: A lexicon of nominalizations. In *Proc. of the 8th International Congress of the European Association for Lexicography*, pages 187–193, Liege, Belgium.

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

Dan Moldovan, Adriana Badulescu, Marta Tatu, Daniel Antohe, and Roxana Girju. 2004. Models for the semantic classification of noun phrases. In *Proc. of HLT-NAACL 2004 Workshop on Computational Lexical Semantics*, pages 60–7, Boston, USA.

Grace Ngai and Radu Florian. 2001. Transformation-based learning in the fast lane. In *Proc. of the 2nd Annual Meeting of the North American Chapter of the ACL*, pages 40–7, Pittsburgh, USA.

Barbara Rosario and Marti Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proc. of the 6th Conference on Empirical Methods in Natural Language Processing*, Pittsburgh, USA.

Lucy Vanderwende. 1994. Algorithm for automatic interpretation of noun sequences. In *Proc. of the 15th International Conference on Computational Linguistics*, pages 782–788, Kyoto, Japan.

Beatrice Warren. 1978. *Semantic Patterns of Noun-Noun Compounds*. Acta Universitatis Gothoburgensis, Göteborg, Sweden.