Interpretation of Compound Nominalisations using Corpus and Web Statistics

Jeremy Nicholson and Timothy Baldwin
Department of Computer Science and Software Engineering
University of Melbourne, VIC 3010, Australia

and

NICTA Victoria Research Laboratories
University of Melbourne, VIC 3010, Australia
`{jeremymn,tim}@csse.unimelb.edu.au`

Abstract

We present two novel paraphrase tests for automatically predicting the inherent semantic relation of a given compound nominalisation as one of subject, direct object, or prepositional object. We compare these to the usual verb–argument paraphrase test using corpus statistics, and frequencies obtained by scraping the Google search engine interface. We also implemented a more robust statistical measure than maximum likelihood estimation — the confidence interval. A significant reduction in data sparseness was achieved, but this alone is insufficient to provide a substantial performance improvement.

1 Introduction

Compound nouns are a class of multiword expression (MWE) that have been of interest in recent computational linguistic work, as any task with a lexical semantic dimension (like machine translation or information extraction) must take into account their semantic markedness. A compound noun is a sequence of two or more nouns comprising an $\bar{N}$, for example, *polystyrene garden-gnome*. The productivity of compound nouns makes their treatment equally desirable and difficult. They appear frequently: more than 1% of the words in the British National Corpus (BNC: Burnard (2000)) participate in noun compounds (Tanaka and Baldwin, 2003). However, unestablished compounds are common: almost 70% of compounds identified in the BNC co-occur with a frequency of only one (Lapata and Lascarides, 2003).

Analysis of the entire space of compound nouns has been hampered to some degree as the space defies some regular set of predicates to define the implicit semantics between a modifier and its head. This semantic underspecification led early analysis to be primarily of a semantic nature, but more recent work has advanced into using syntax to predict the semantics, in the spirit of the study by Levin (1993) on diathesis alternations.

In this work, we examine compound nominalisations, a subset of compound nouns where the head has a morphologically–related verb. For example, *product replacement* has an underlying verbal head *replace*, whereas *garden-gnome* has no such form. While compound nouns in general have a set of semantic relationships between the head and modifier that is potentially non-finite, compound nominalisations are better defined, in that the modifier fills a syntactic argument relation with respect to the head. For example, *product* might fill the direct object slot of the verb *to replace* for the compound above. Compound nominalisations comprise a substantial minority of compound nouns, with figures of about 35% being observed (Grover et al., 2005; Nicholson, 2005).

We propose two novel paraphrases for a corpus statistical approach to predicting the relationship for a set of compound nominalisations, and investigate how using the World Wide Web as a corpus alleviates the common phenomenon of data sparseness, and how the volume of data impacts on the classification results. We also examine a more robust statistical approach to interpretation of the statistics than maximum likelihood estimates, called the confidence interval.

The rest of the paper is structured as follows: in Section 2, we present a brief background for our work, with a listing of our resources in Section 3. We detail our proposed method in Section 4, the corresponding results in Section 5, with a discus-

sion in Section 6 and a brief conclusion in Section 7.

2 Background

2.1 Compound Noun Interpretation

Compound nouns were seminally and thoroughly analysed by Levi (1978), who hand–constructs a nine–way set of semantic relations that she identifies as broadly defining the observed relationships between the compound head and modifier. Warren (1978) also inspects the syntax of compound nouns, to create a somewhat different set of twelve conceptual categories.

Early attempts to automatically classify compound nouns have taken a semantic approach: Finin (1980) and Isabelle (1984) use "role nominals" derived from the head of the compound to fill a slot with the modifier. Vanderwende (1994) uses a rule–based technique that scores a compound on possible semantic interpretations, while Jones (1995) implements a graph–based unification procedure over semantic feature structures for the head. Finally, Rosario and Hearst (2001) make use of a domain–specific lexical resource to classify according to neural networks and decision trees.

Syntactic classification, using paraphrasing, was first used by Leonard (1984), who uses a prioritised rule–based approach across a number of possible readings. Lauer (1995) employs a corpus statistical model over a similar paraphrase set based on prepositions. Lapata (2002) and Grover et al. (2005) again use a corpus statistical paraphrase–based approach, but with verb–argument relations for compound nominalisations — attempting to define the relation as one of subject, direct object, or a number of prepositional objects in the latter.

2.2 Web–as–Corpus Approaches

Using the World Wide Web for corpus statistics is a relatively recent phenomenon; we present a few notable examples. Grefenstette (1998) analyses the plausibility of candidate translations in a machine translation task through Web statistics, and avoids some data sparseness within that context. Zhu and Rosenfeld (2001) train a language model from a large corpus, and use the Web to estimate low–density trigram frequencies. Keller and Lapata (2003) show that Web counts can obviate data sparseness for syntactic predicate–

argument bigrams. They also observe that the noisiness of the Web, while unexplored in detail, does not greatly reduce the reliability of their results. Nakov and Hearst (2005) demonstrate that Web counts can aid in identifying the bracketing in higher–arity noun compounds. Finally, Lapata and Keller (2005) evaluate the performance of Web counts on a wide range of natural language processing tasks, including compound noun bracketing and compound noun interpretation.

2.3 Confidence Intervals

Maximum likelihood statistics are not robust when many sparse vectors are under consideration, i.e. naively "choosing the largest number" may not be accurate in contexts when the relative value across samplings may be relevant, for example, in machine learning. As such, we apply a statistical test with confidence intervals (Kenney and Keeping, 1962), where we compare sample z-scores in a pairwise manner, instead of frequencies globally.

The confidence interval P, for z-score n, is:

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

t is chosen to normalise the curve, and P is strictly increasing on n, so we are only required to find the largest z-score.

Calculating the z-score exactly can be quite costly, so we instead use the binomial approximation to the normal distribution with equal prior probabilities and find that a given z-score Z is:

$$Z = \frac{f - \mu}{\sigma} \qquad (2)$$

where f is the frequency count, μ is the mean in a pairwise test, and σ is the standard deviation of the test. A more complete derivation appears in Nicholson (2005).

3 Resources

We make use of a number of lexical resources in our implementation and evaluation. For corpus statistics, we use the written component of the BNC, a balanced 90M token corpus. To find verb–argument frequencies, we parse this using RASP (Briscoe and Carroll, 2002), a tag sequence grammar–based statistical parser. We contrast the corpus statistics with ones collected from the

Web, using an implementation of a freely available Google "scraper" from CPAN.[1]

For a given compound nominalisation, we wish to determine all possible verbal forms of the head. We do so using the combination of the morphological component of CELEX (Burnage, 1990), a lexical database, NOMLEX (Macleod et al., 1998), a nominalisation database, and CATVAR (Habash and Dorr, 2003), an automatically–constructed database of clusters of inflected words based on the Porter stemmer (Porter, 1997).

Once the verbal forms have been identified, we construct canonical forms of the present participle (+*ing*) and the past participle (+*ed*), using the morph lemmatiser (Minnen et al., 2001). We construct canonical forms of the plural head and plural modifier (+*s*) in the same manner.

For evaluation, we have the two–way classified data set used by Lapata (2002), and a three–way classified data set constructed from open text. Lapata automatically extracts candidates from the British National Corpus, and hand–curates a set of 796 compound nominalisations which were interpreted as either a subjective relation SUBJ (e.g. *wood appearance* "wood appears"), or a (direct) objective relation OBJ (e.g. *stress avoidance* "[SO] avoids stress". We automatically validated this data set for consistency, removing:

1. items that did not occur in the same chunk, according to a chunker based on fnTBL 1.0 (Ngai and Florian, 2001),

2. items whose head did not have a verbal form according to our lexical resources, and

3. items which consisted in part of proper nouns,

to end up with 695 consistent compounds. We used the method of Nicholson and Baldwin (2005) to derive a small data set of 129 compound nominalisations, also from the BNC, which we instructed three unskilled annotators to identify each as one of subjective (SUB), direct object (DOB), or prepositional object (POB, e.g. *side show* "[SO] show [ST] on the side"). The annotators identified nine prepositional relations: {about,against,for,from,in,into,on,to,with}.

4 Proposed Method

4.1 Paraphrase Tests

To derive preferences for the SUB, DOB, and various POB interpretations for a given compound nominalisation, the most obvious approach is to examine a parsed corpus for instances of the verbal form of the head and the modifier occurring in the corresponding verb–argument relation. There are other constructions that can be informative, however.

We examine two novel paraphrase tests: one prepositional and one participial. The prepositional test is based in part on the work by Leonard (1984) and Lauer (1995): for a given compound, we search for instances of the head and modifier nouns separated by a preposition. For example, for the compound nominalisation *leg operation*, we might search for *operation on the leg*, corresponding to the POB relation *on*. Special cases are *by*, corresponding to a subjective reading akin to a passive construction (e.g. *investor hesitancy*, *hesitancy by the investor* ≡ "the investor hesitates"), and *of*, corresponding to a direct object reading (e.g. *language speaker*, *speaker of the language* ≡ "[SO] speaks the language").

The participial test is based on the paraphrasing equivalence of using the present participle of the verbal head as an adjective before the modifier, for the SUB relation (e.g. *the hesitating investor* ≡ "the investor hesitates"), compared to the past participle for the DOB relation (*the spoken language* ≡ "[SO] speaks the language"). The corresponding prepositional object construction is unusual in English, but still possible: compare *?the operated-on leg* and *the lived-in village*.

4.2 The Algorithm

Given a compound nominalisation, we perform a number of steps to arrive at an interpretation. First, we derive a set of verbal forms for the head from the combination of CELEX, NOMLEX, and CATVAR. We find the participial forms of each of the verbal heads, and plurals for the nominal head and modifier, using the morph lemmatiser.

Next, we examine the BNC for instances of the modifier and one of the verbal head forms occurring in a verb–argument relation, with the aid of the RASP parse. Using these frequencies, we calculate the pairwise z-scores between SUB and DOB, and between SUB and POB: the score given to the SUB interpretation is the greater of the two.

We further examine the RASP parsed data for instances of the prepositional and participial tests for the compound, and calculate the z-scores for these as well.

We then collect our Google counts. Because the Web data is unparsed, we cannot look for syntactic structures explictly. Instead, we query a number of collocations which we expect to be representative of the desired structure.

For the prepositional test, the head can be singular or plural, the modifier can be singular or plural, and there may or may not be an article between the preposition and the modifier. For example, for the compound nominalisation *product replacement* and preposition *of* we search for all of the following: (and similarly for the other prepositions)

> *replacement of product*
> *replacement of the product*
> *replacement of products*
> *replacement of the products*
> *replacements of product*
> *replacements of the product*
> *replacements of products*
> *replacements of the products*

For the participial test, the modifier can be singular or plural, and if we are examining a prepositional relation, the head can be either a present or past participle. For *product replacement*, we search for, as well as other prepositions:

> *the replacing product*
> *the replacing products*
> *the replaced product*
> *the replaced products*
> *the replacing–about product*
> *the replacing–about products*
> *the replaced–about product*
> *the replaced–about products*

We comment briefly on these tests in Section 6.

We choose to use *the* as our canonical article because it is a reliable marker of the left boundary of an NP and number-neutral; using *a/an* represents a needless complication.

We then calculate the z-scores using the method described in Section 2, where the individual frequency counts are the maximum of the results obtained across the query set.

Once the z-scores have been obtained, we choose a classification based on the greatest-valued observed test. We contrast the confidence interval–based approach with the maximum likelihood method of choosing the largest of the raw frequencies. We also experiment with a machine learning package, to examine the mutual predictiveness of the separate tests.

5 Observed Results

First, we found majority-class baselines for each of the data sets. The two–way data set had 258 SUBJ–classified items, and 437 OBJ–classified items, so choosing OBJ each time gives a baseline of 62.9%. The three–way set had 22 SUB items, 63 of DOB, and 44 of POB, giving a baseline of 48.8%.

Contrasting this with human performance on the data set, Lapata recorded a raw inter-annotator agreement of 89.7% on her test set, which corresponds to a Kappa value $\kappa = 0.78$. On the three–way data set, three annotators had a agreement of 98.4% for identification and classification of observed compound nominalisations in open text, and $\kappa = 0.83$. For the three-way data set, the annotators were asked to both identify and classify compound nominalisations in free text, and agreement is thus calculated over all words in the test. The high agreement figure is due to the fact that most words could be trivially disregarded (e.g. were not nouns). Kappa corrects this for chance agreement, so we conclude that this task was still better-defined than the one posed by Lapata. One possible reason for this was the number of poorly–behaved compounds that we removed due to chunk inconsistencies, lack of a verbal form, or proper nouns: it would be difficult for the annotators to agree over compounds where an obvious well–defined interpretation was not available.

5.1 Comparison Classification

Results for classification over the Lapata two–way data set are given in Table 1, and results over the open data three–way set are given in Table 2. For these, we selected the greatest raw frequency count for a given test as the intended relation (Raw), or the greatest confidence interval according to the z-score (Z-Score). If a relation could not be selected due to ties (e.g., the scores were all 0), we selected the majority baseline. To deal with the nature of the two–way data set with respect to our three–way selection, we mapped compounds that we would prefer to be POB to OBJ, as there are

Paraphrase	Default	Corpus Counts		Web Counts	
		Raw	Z-Score	Raw	Z-Score
Verb–Argument	62.9	67.9	68.3	–	–
Prepositional	62.9	62.1	62.4	62.6	63.0
Participial	62.9	63.0	63.2	61.4	58.8

Table 1: Classification Results over the two–way data set, in %. Comparison of raw frequency counts vs. confidence–based z-scores, for BNC data and Google scrapings shown.

Paraphrase	Default	Corpus Counts		Web Counts	
		Raw	Z-Score	Raw	Z-Score
Verb–Argument	48.8	54.3	55.0	–	–
Prepositional	48.8	48.4	50.0	59.7	58.9
Participial	48.8	43.2	45.4	43.4	38.0

Table 2: Classification results over the three-way data set, in %. Comparison of raw frequency counts vs. confidence-based z-scores, for BNC data and Google scrapings shown.

compounds in the set (e.g. *adult provision*) that have a prepositional object reading ("provide for adults") but have been classified as a direct object OBJ.

The verb–argument counts obtained from the parsed BNC are significantly better than the baseline for the Lapata data set ($\chi^2 = 4.12, p \leq 0.05$), but not significantly better for the open data set ($\chi^2 = 0.99, p \leq 1$). Similar results were reported by Lapata (2002) over her data set using backed–off smoothing, the most closely related method.

Neither the prepositional nor participial paraphrases were significantly better than the baseline for either the two–way ($\chi^2 = 0.00, p \leq 1$), or the three–way data set ($\chi^2 = 3.52, p \leq 0.10$), although the prepositional test did slightly improve on the verb–argument results.

5.2 Machine Learning Classification

Although the results were not impressive, we still believed that there was complementing information within the data, which could be extracted with the aid of a machine learner. For this, we made use of TiMBL (Daelemans et al., 2003), a nearest-neighbour classifier which stores the entire training set and extrapolates further samples, as a principled method for combination of the data. We use TiMBL's in-built cross-validation method: 90% of the data set is used as training data to test the other 10%, for each stratified tenth of the set. The results it achieves are assumed to be able to generalise to new samples if they are compared to the current training data set.

The results observed using TiMBL are shown

	Corpus Counts	Web Counts
Two–way Set	72.4	74.2
Three–way Set	51.1	50.4

Table 3: TiMBL results for the combination of paraphrase tests over the two–way and three–way data sets for corpus and Web frequencies

in Table 3. This was from the combination of all of the available paraphrase tests: verb–argument, prepositional, and participial for the corpus counts, and just prepositional and participial for the Web counts. The results for the two–way data set derived from Lapata's data set were a good improvement over the simple classification results, significantly so for the Web frequencies ($\chi^2 = 20.3, p \leq 0.01$). However, we also notice a corresponding decrease in the results for the three–way open data set, which make these improvements immaterial.

Examining the other possible combinations for the tests did indeed lead to varying results, but not in a consistent manner. For example, the best combination for the open data set was using the participial raw scores and z-scores (58.1%), which performed particularly badly in simple comparisons, and comparatively poorly (70.2%) for the two–way set.

6 Discussion

Although the observed results failed to match, or even approach, various benchmarks set by Lapata (2002) (87.3% accuracy) and Grover et al. (2005) (77%) for the subject–object and subject–

direct object–prepositional objects classification tasks respectively, the presented approach is not without merit. Indeed, these results relied on machine learning approaches incorporating many features independent of corpus counts: namely, context, suffix information, and semantic similarity resources. Our results were an examination of the possible contribution of lexical information available from high–volume unparsed text.

One important concept used in the above benchmarks was that of statistical smoothing, both class–based and distance–based. The reason for this is the inherent data sparseness within the corpus statistics for these paraphrase tests. Lapata (2002) observes that almost half (47%) of the verb–noun pairs constructed are not attested within the BNC. Grover et al. (2005) also note the sparseness of observed relations. Using the immense data source of the Web allows one to circumvent this problem: only one compound (*anarchist prohibition*) has no instances of the paraphrases from the scraping,[2] from more than 900 compounds between the two data sets. This extra information, we surmise, would be beneficial for the smoothing procedures, as the comparative accuracy between the two methods is similar.

On the other hand, we also observe that simply alleviating the data sparseness is insufficient to provide a reliable interpretation. These results reinforce the contribution made by the statistical and semantic resources used in arriving at these benchmarks.

The approach suggested by Keller and Lapata (2003) for obtaining bigram information from the Web could provide an approach for estimating the syntactic verb–argument counts for a given compound (dashes in Tables 1 and 2). In spite of the inherent unreliability of approximating long–range dependencies with n-gram information, results look promising. An examination of the effectiveness of this approach is left as further research. Similarly, various methods of combining corpus counts with the Web counts, including smoothing, backing–off, and machine learning, could also lead to interesting performance impacts.

Another item of interest is the comparative difficulty of the task presented by the three–way data set extracted from open data, and the two–way data set hand–curated by Lapata. The baseline

of this set is much lower, even compared that of the similar task (albeit domain–specific) from Grover et al. (2005) of 58.6%. We posit that the hand–filtering of the data set in these works contributes to a biased sample. For example, removing prepositional objects for a two–way classification, which make up about a third of the open data set, renders the task somewhat artificial.

Comparison of the results between the maximum likelihood estimates used in earlier work, and the more statistically robust confidence intervals were inconclusive as to performance improvement, and were most effective as a feature expansion algorithm. The only obvious result is an aesthetic one, in using "robust statistics".

Finally, the paraphrase tests which we propose are not without drawbacks. In the prepositional test, a paraphrase with *of* does not strictly contribute to a direct object reading: consider *school aim* "school aims", for which instances of *aim by the school* are overwhelmed by *aim of the school*. We experimented with permutations of the available queries (e.g. requiring the head and modifier to be of different number, to reflect the pluralisability of the head in such compounds, e.g. *aims of the school*), without observing substantially different results.

Another observation is the inherent bias of the prepositional test to the prepositional object relation. Apparent prepositional relations can occur in spite of the available verb frames: consider *cash limitation*, where the most populous instance is *limitation on cash*, despite the impossibility of **to limit on cash* (for *to place a limit on cash*). Another example, is *bank agreement*: finding instances of *agreement with bank* does not lead to the pragmatically absurd [SO] *agrees with the bank*.

Correspondingly, the participial relation has the opposite bias: constructions of the form *the lived-in flat* "[SO] lived in the flat" are usually lexicalised in English. As such, only 17% of compounds in the two–way data set and 34% of the three-way data set display non-zero values in the prepositional object relation for the participial test. We hoped that the inherent biases of the two tests might balance each other, but there is little evidence of that from the results.

7 Conclusion

We presented two novel paraphrase tests for automatically predicting the inherent semantic relation of a given compound nominalisation as one of subject, direct object, or prepositional object. We compared these to the usual verb–argument paraphrase test, using corpus statistics, and frequencies obtained by scraping the Google search engine. We also implemented a more robust statistical measure than the insipid maximum likelihood estimates — the confidence interval. A significant reduction in data sparseness was achieved, but this alone is insufficient to provide a substantial performance improvement.

Acknowledgements

References

Ted Briscoe and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings 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.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2003. *TiMBL: Tilburg Memory Based Learner, version 5.0, Reference Guide*. ILK Technical Report 03-10.

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

Gregory Grefenstette. 1998. The World Wide Web as a resource for example-based machine translation tasks. In *Proceedings of the ASLIB Conference on Translating and the Computer*, London, UK.

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 *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the ACL*, pages 17–23, Edmonton, Canada.

Pierre Isabelle. 1984. Another look at nominal compounds. In *Proceeedings of the 10th International Conference on Computational Linguistics and 22nd Annual Meeting of the ACL*, pages 509–516, Stanford, USA.

Bernard Jones. 1995. Predicating nominal compounds. In *Proceedings of the 17th International Conference of the Cognitive Science Society*, pages 130–5, Pittsburgh, USA.

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

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

Mirella Lapata and Frank Keller. 2005. Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing*, 2(1).

Mirella Lapata and Alex Lascarides. 2003. Detecting novel compounds: The role of distributional evidence. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computional Linguistics*, pages 235–242, Budapest, Hungary.

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, the Netherlands.

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

Beth Levin. 1993. *English Verb Classes and Alternations*. The University of Chicago Press, Chicago, USA.

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

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–23.

Preslov Nakov and Marti Hearst. 2005. Search engine statistics beyond the n-gram: Application to noun compound bracketing. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 17–24, Ann Arbor, USA.

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

Jeremy Nicholson and Timothy Baldwin. 2005. Statistical interpretation of compound nominalisations. In *Proceeding of the Australasian Langugae Technology Workshop 2005*, Sydney, Australia.

Jeremy Nicholson. 2005. Statistical interpretation of compound nouns. Honours Thesis, University of Melbourne, Melbourne, Australia.

Martin Porter. 1997. An algorithm for suffix stripping. In Karen Sparck Jones and Peter Willett, editors, *Readings in information retrieval*. Morgan Kaufmann, San Francisco, USA.

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

Takaaki Tanaka and Timothy Baldwin. 2003. Noun-noun compound machine translation: A feasibility study on shallow processing. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 17–24, Sapporo, Japan.

Lucy Vanderwende. 1994. Algorithm for automatic interpretation of noun sequences. In *Proceedings 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.

Xiaojin Zhu and Ronald Rosenfeld. 2001. Improving trigram language modeling with the World Wide Web. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 533–6, Salt Lake City, USA.