Determining the Syntactic Structure of Medical Terms in Clinical Notes

Bridget T. McInnes
Dept. of Computer Science
and Engineering
University of Minnesota
Minneapolis, MN, 55455
bthomson@cs.umn.edu

Ted Pedersen
Dept. of Computer Science
University of Minnesota Duluth
Duluth, MN, 55812
tpederse@d.umn.edu

Serguei V. Pakhomov
Dept. of Pharmaceutical Care
and Health Systems Center
for Health Informatics
University of Minnesota
Minneapolis, MN, 55455
pakh0002@umn.edu

Abstract

This paper demonstrates a method for determining the syntactic structure of medical terms. We use a model-fitting method based on the Log Likelihood Ratio to classify three-word medical terms as right or left-branching. We validate this method by computing the agreement between the classification produced by the method and manually annotated classifications. The results show an agreement of 75% - 83%. This method may be used effectively to enable a wide range of applications that depend on the semantic interpretation of medical terms including automatic mapping of terms to standardized vocabularies and induction of terminologies from unstructured medical text.

1 Introduction

Most medical concepts are expressed via a domain specific terminology that can either be explicitly agreed upon or extracted empirically from domain specific text. Regardless of how it is constructed, a terminology serves as a foundation for information encoding, processing and exchange in a specialized sub-language such as medicine. Concepts in the medical domain are encoded through a variety of linguistic forms, the most typical and widely accepted is the noun phrase (NP). In some even further specialized subdomains within medicine, such as nursing and surgery, an argument can be made that some concepts are represented by an entire predication rather than encapsulated within a single nominalized expression. For example, in order to describe someone's ability to lift objects 5 pounds or heavier above their head, it may be necessary to use a term consisting of a predicate such as [LIFT] and a set of arguments corresponding to various thematic roles such as <PATIENT> and <PATH> (Ruggieri et al., 2004). In this paper, we address typical medical terms encoded as noun phrases (NPs) that are often structurally ambiguous, as in Example 1, and discuss a case for extending the proposed method to non-nominalized terms as well.

$$\text{small}_1 \text{ bowel}_2 \text{ obstruction}_3 \tag{1}$$

The NP in Example 1 can have at least two interpretations depending on the syntactic analysis:

$$[[\text{small}_1 \text{ bowel}_2] \text{ obstruction}_3] \tag{2}$$

$$[\text{small}_1 [\text{bowel}_2 \text{ obstruction}_3]] \tag{3}$$

The term in Example 2 denotes an obstruction in the small bowel, which is a diagnosable disorder; whereas, the term in Example 3 refers to a small unspecified obstruction in the bowel.

Unlike the truly ambiguous general English cases such as the classical "American History Professor" where the appropriate interpretation depends on the context, medical terms, such as in Example 1, tend to have only one appropriate interpretation. The context, in this case, is the discourse domain of medicine. From the standpoint of the English language, the interpretation that follows from Example 3 is certainly plausible, but unlikely in the context of a medical term. The syntax of a term only shows

BioNLP 2007: Biological, translational, and clinical language processing, pages 9–16,
Prague, June 2007.

what interpretations are possible without restricting them to any particular one. From the syntactic analysis, we know that the term in Example 1 has the potential for being ambiguous; however, we also know that it does have an intended interpretation by virtue of being an entry term in a standardized terminology with a unique identifier anchoring its meaning. What we do not know is which syntactic structure generated that interpretation. Being able to determine the structure consistent with the intended interpretation of a clinical term can improve the analysis of unrestricted medical text and subsequently improve the accuracy of Natural Language Processing (NLP) tasks that depend on semantic interpretation.

To address this problem, we propose to use a model-fitting method which utilizes an existing statistical measure, the Log Likelihood Ratio. We validate the application of this method on a corpus of manually annotated noun-phrase-based medical terms. First, we present previous work on structural ambiguity resolution. Second, we describe the Log Likelihood Ratio and then its application to determining the structure of medical terms. Third, we describe the training corpus and discuss the compilation of a test set of medical terms and human expert annotation of those terms. Last, we present the results of a preliminary validation of the method and discuss several possible future directions.

2 Previous Work

The problem of resolving structural ambiguity has been previously addressed in the computational linguistics literature. There are multiple approaches ranging from purely statistical (Ratnaparkhi, 1998), to hybrid approaches that take into account the lexical semantics of the verb (Hindle and Rooth, 1993), to corpus-based, which is the approach discussed in this paper. (Marcus, 1980) presents an early example of a corpus-based approach to syntactic ambiguity resolution. One type of structural ambiguity that has received much attention has to do with nominal compounds as seen in the work of (Resnik, 1993), (Resnik and Hearst, 1993), (Pustejovsky et al., 1993), and (Lauer, 1995).

(Lauer, 1995) points out that the existing approaches to resolving the ambiguity of noun phrases fall roughly into two camps: adjacency and dependency. The proponents of the adjacency model ((Liberman and Sproat, 1992), (Resnik, 1993) and (Pustejovsky et al., 1993)) argue that, given a three word noun phrase XYZ, there are two possible analyzes [[XY]Z] and [X[YZ]]. The correct analysis is chosen based on the "acceptability" of the adjacent bigrams A[XY] and A[YZ]. If A[XY] is more acceptable than A[YZ], then the left-branching analysis [[XY]Z] is preferred.

(Lauer and Dras, 1994) and (Lauer, 1995) address the issue of structural ambiguity by developing a dependency model where instead of computing the acceptability of A[YZ] one would compute the acceptability of A[XZ]. (Lauer, 1995) argues that the dependency model is not only more intuitive than the adjacency model, but also yields better results. (Lapata and Keller, 2004) results also support this assertion.

The difference between the approaches within the two models is the computation of acceptability. Proposals for computing acceptability (or preference) include raw frequency counts ((Evans and Zhai, 1996) and (Lapata and Keller, 2004)), Latent Semantic Indexing ((Buckeridge and Sutcliffe, 2002)) and statistical measures of association ((Lapata et al., 1999) and (Nakov and Hearst, 2005)).

One of the main problems with using frequency counts or statistical methods for structural ambiguity resolution is the sparseness of data; however, (Resnik and Hearst, 1993) used conceptual associations (associations between groups of terms deemed to form conceptual units) in order to alleviate this problem. (Lapata and Keller, 2004) use the document counts returned by WWW search engines. (Nakov and Hearst, 2005) use the χ^2 measure based on statistics obtained from WWW search engines to compute values to determine acceptability of a syntactic analysis for nominal compounds. This method is tested using a set of general English nominal compounds developed by (Lauer, 1995) as well as a set of nominal compounds extracted from MEDLINE abstracts.

The novel contribution of our study is in demonstrating and validating a corpus-based method for determining the syntactic structure of medical terms that relies on using the statistical measure of association, the Log Likelihood Ratio, described in the following section.

3 Log Likelihood Ratio

The Log Likelihood Ratio (G^2) is a "goodness of fit" statistic first proposed by (Wilks, 1938) to test if a given piece of data is a sample from a set of data with a specific distribution described by a hypothesized model. It was later applied by (Dunning, 1993) as a way to determine if a sequence of N words (N-gram) came from an independently distributed sample.

(Pedersen et al., 1996) pointed out that there exists theoretical assumptions underlying the G^2 measure that were being violated therefore making them unreliable for significance testing. (Moore, 2004) provided additional evidence that although G^2 may not be useful for determining the significance of an event, its near equivalence to mutual information makes it an appropriate measure of word association. (McInnes, 2004) applied G^2 to the task of extracting three and four word collocations from raw text.

G^2, formally defined for trigrams in Equation 4, compares the observed frequency counts with the counts that would be expected if the words in the trigram (3-gram; a sequence of three words) corresponded to the hypothesized model.

$$G^2 = 2 * \sum_{x,y,z} n_{xyz} * log(\frac{n_{xyz}}{m_{xyz}}) \qquad (4)$$

The parameter n_{xyz} is the observed frequency of the trigram where x, y, and z respectively represent the occurrence of the first, second and third words in the trigram. The variable m_{xyz} is the expected frequency of the trigram which is calculated based on the hypothesized model. This calculation varies depending on the model used. Often the hypothesized model used is the independence model which assumes that the words in the trigram occur together by chance. The calculation of the expected values based on this model is as follows:

$$m_{xyz} = n_{x++} * n_{+y+} * n_{++z} / n_{+++} \qquad (5)$$

The parameter, n_{+++}, is the total number of trigrams that exist in the training data, and n_{x++}, n_{+y+}, and n_{++z} are the individual marginal counts of seeing words x, y, and z in their respective positions in a trigram. A G^2 score reflects the degree to which the observed and expected values diverge. A G^2 score of zero implies that the observed values are equal to the expected and the trigram is represented perfectly by the hypothesized model. Hence, we would say that the data 'fits' the model. Therefore, the higher the G^2 score, the less likely the words in the trigram are represented by the hypothesized model.

4 Methods

4.1 Applying Log Likelihood to Structural Disambiguation

The independence model is the only hypothesized model used for bigrams (2-gram; a sequence of two words). As the number of words in an N-gram grows, the number of hypothesized models also grows. The expected values for a trigram can be based on four models. The first model is the independence model discussed above. The second is the model based on the probability that the first word and the second word in the trigram are dependent and independent of the third word. The third model is based on the probability that the second and third words are dependent and independent of the first word. The last model is based on the probability that the first and third words are dependent and independent of the second word. Table 1 shows the different models for the trigram XYZ.

Table 1: Models for the trigram XYZ

Model 1	P(XYZ) / P(X) P(Y) P(Z)
Model 2	P(XYZ) / P(XY) P(Z)
Model 3	P(XYZ) / P(X) / P(YZ)
Model 4	P(XYZ) / P(XZ) P(Y)

Slightly different formulas are used to calculate the expected values for the different hypothesized models. The expected values for Model 1 (the independence model) are given above in Equation 5. The calculation of expected values for Model 2, 3, 4 are seen in Equations 6, 7, 8 respectively.

$$m_{xyz} = n_{xy+} * n_{++z} / n_{+++} \qquad (6)$$

$$m_{xyz} = n_{x++} * n_{+yz} / n_{+++} \qquad (7)$$

$$m_{xyz} = n_{x+z} * n_{+y+} / n_{+++} \qquad (8)$$

The parameter n_{xy+} is the number of times words x and y occur in their respective positions, n_{+yz} is the number of times words y and z occur in their respective positions and n_{x+z} is the number of times that words x and z occur in their respective positions in the trigram.

The hypothesized models result in different expected values which results in a different G^2 score. A G^2 score of zero implies that the data are perfectly represented by the hypothesized model and the observed values are equal to the expected. Therefore, the model that returns the lowest score for a given trigram is the model that best represents the structure of that trigram, and hence, best 'fits' the trigram. For example, Table 2 shows the scores returned for each of the four hypothesized models for the trigram "small bowel obstruction".

Table 2: Example for the term "small bowel obstruction"

Model	G^2 score	Model	G^2 score
Model 1	11,635.45	Model 2	**5,169.81**
Model 3	8,532.90	Model 4	7,249.90

The smallest G^2 score is returned by Model 2 which is based on the first and second words being dependent and independent of the third. Based on the data, Model 2 best represents or 'fits' the trigram, "small bowel obstruction". In this particular case that happens to be the correct analysis.

The frequency counts and G^2 scores for each model were obtained using the N-gram Statistics Package [1] (Banerjee and Pedersen, 2003).

4.2 Data

The data for this study was collected from two sources: the Mayo Clinic clinical notes and SNOMED-CT terminology (Stearns et al., 2001).

4.2.1 Clinical Notes

The corpus used in this study consists of over 100,000 clinical notes covering a variety of major medical specialties at the Mayo Clinic. These notes document each patient-physician contact and are typically dictated over the telephone. They range in length from a few lines to several pages of text and represent a quasi-spontaneous discourse where the dictations are made partly from notes and partly from memory. At the Mayo Clinic, the dictations are transcribed by trained personnel and are stored in the patient's chart electronically.

4.2.2 SNOMED-CT

SNOMED-CT (Systematized Nomenclature of Medicine, Clinical Terminology) is an ontological resource produced by the College of American Pathologists and distributed as part of the Unified Medical Language System[2] (UMLS) Metathesaurus maintained by the National Library of Medicine. SNOMED-CT is the single largest source of clinical terms in the UMLS and as such lends itself well to the analysis of terms found in clinical reports.

SNOMED-CT is used for many applications including indexing electronic medical records, ICU monitoring, clinical decision support, clinical trials, computerized physician order entry, disease surveillance, image indexing and consumer health information services. The version of SNOMED-CT used in this study consists of more than 361,800 unique concepts with over 975,000 descriptions (entry terms) (SNOMED-CT Fact Sheet, 2004).

4.3 Testset of Three Word Terms

We used SNOMED-CT to compile a list of terms in order to develop a test set to validate the G^2 method. The test set was created by extracting all trigrams from the corpus of clinical notes and all three word terms found in SNOMED-CT. The intersection of the SNOMED-CT terms and the trigrams found in the clinical notes was further restricted to include only simple noun phrases that consist of a head noun modified with a set of other nominal or adjectival elements including adjectives and present and past participles. Adverbial modification of adjectives was also permitted (e.g. "partially edentulous maxilla"). Noun phrases with nested prepositional phrases such as "fear of flying" as well as three word terms that are not noun phrases such as "does not eat" or "unable to walk" were excluded from the test set. The resulting test set contains 710 items.

The intended interpretation of each three word term (trigram) was determined by arriving at a

[1]http://www.d.umn.edu/ tpederse/nsp.html

[2]Unified Medical Language System is a compendium of over 130 controlled medical vocabularies encompassing over one million concepts.

consensus between two medical index experts (kappa=0.704). These experts have over ten years of experience with classifying medical diagnoses and are highly qualified to carry out the task of determining the intended syntactic structure of a clinical term.

Table 3: Four Types of Syntactic Structures of Trigram Terms

left-branching ((XY)Z): [[urinary tract] infection] [[right sided] weakness]
right-branching (X(YZ)): [chronic [back pain]] [low [blood pressure]]
non-branching ((X)(Y)(Z)): [[follicular][thyroid][carcinoma]] [[serum][dioxin][level]]
monolithic (XYZ): [difficulty finding words] [serous otitis media]

In the process of annotating the test set of trigrams, four types of terms emerged (Table 3). The first two types are left and right-branching where the left-branching phrases contain a left-adjoining group that modifies the head of the noun phrase. The right-branching phrases contain a right-adjoining group that forms the kernel or the head of the noun phrase and is modified by the remaining word on the left. The non-branching type is where the phrase contains a head noun that is independently modified by the other two words. For example, in “follicular thyroid carcinoma”, the experts felt that “carcinoma” was modified by both “follicular” and “thyroid” independently, where the former denotes the type of cancer and the latter denotes its location. This intuition is reflected in some formal medical classification systems such as the Hospital International Classification of Disease Adaptation (HICDA) where cancers are typically classified with at least two categories - one for location and one for the type of malignancy. This type of pattern is rare. We were able to identify only six examples out of the 710 terms. The monolithic type captures the intuition that the terms function as a collocation and are not decomposable into subunits. For example, “leg length discrepancy” denotes a specific disorder where one leg is of a different length from the other. Various combinations of subunits within this term result in nonsensical expressions.

Table 4: Distribution of term types in the test set

Type	Count	%total
Left-branching	251	35.5
Right-branching	378	53.4
Non-branching	6	0.8
Monolithic	73	10.3
Total	708	100

Finally, there were two terms for which no consensus could be reached: “heart irregularly irregular” and “subacute combined degeneration”. These cases were excluded from the final set. Table 4 shows the distribution of the four types of terms in the test set.

5 Evaluation

We hypothesize that general English typically has a specific syntactic structure in the medical domain, which provides a single semantic interpretation. The patterns observed in the set of 710 medical terms described in the previous section suggest that the G^2 method offers an intuitive way to determine the structure of a term that underlies its syntactic structure.

Table 5: G^2 Model Descriptions

left-branching	Model 2	[[XY] Z]
right-branching	Model 3	[X [YZ]]

The left and right-branching patterns roughly correspond to Models 2 and 3 in Table 5. Models 1 and 4 do not really correspond to any of the patterns we were able to identify in the set of terms. Model 1 would represent a term where words are completely independent of each other, which is an unlikely scenario given that we are working with terms whose composition is dependent by definition. This is not to say that in other applications (e.g., syntactic parsing) this model would not be relevant. Model 4 suggests dependence between the outer edges of a term and their independence from the

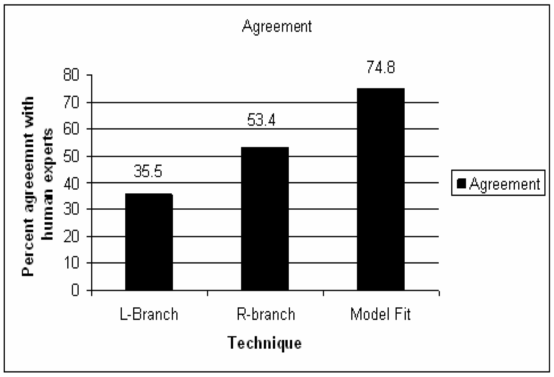

Figure 1: Comparison of the results with two baselines: L-branching and R-branching assumptions

middle word, which is not motivated from the standpoint of a traditional context free grammar which prohibits branch crossing. However, this model may be welcome in a dependency grammar paradigm.

One of the goals of this study is to test an application of the G^2 method trained on a corpus of medical data to distinguish between left and right-branching patterns. The method ought to suggest the most likely analysis for an NP-based medical term based on the empirical distribution of the term and its components. As part of the evaluation, we compute the G^2 scores for each of the terms in the test set, and picked the model with the lowest score to represent the structural pattern of the term. We compared these results with manually identified patterns. At this preliminary stage, we cast the problem of identifying the structure of a three word medical term as a binary classification task where a term is considered to be either left or right-branching, effectively forcing all terms to either be represented by either Model 2 or Model 3.

6 Results and Discussion

In order to validate the G^2 method for determining the structure of medical terms, we calculated the agreement between human experts' interpretation of the syntactic structure of the terms and the interpretation suggested by the G^2 method. The agreement was computed as the ratio of matching interpretations to the total number of terms being interpreted. We used two baselines, one established by assuming that each term is left-branching and the other by assuming that each term is right-branching. As is clear from Table 4, the left-branching baseline is 35.5% and the right-branching baseline is 53.4% meaning that if we simply assign left-branching pattern to each three word term, we would agree with human experts 35.5% of the time. The G^2 method correctly identifies 185 trigrams as being left-branching (Model 2) and 345 trigrams as being right-branching (Model 3). There are 116 right-branching trigrams incorrectly identified as left-branching, and 62 left-branching trigrams incorrectly identified as right- branching. Thus the method and the human experts agreed on 530 (75%) terms out of 708 (kappa=0.473), which is better than both baselines (Figure 1). We did not find any overlap between the terms that human experts annotated as non-branching and the terms whose corpus distribution can be represented by Model 4 ([[XZ]Y]). This is not surprising as this pattern is very rare. Most of the terms are represented by either Model 2 (left-branching) or Model 3 (right-branching). The monolithic terms that the human experts felt were not decomposable constitute 10% of all terms and may be handled through some other mechanism such as collocation extraction or dictionary lookup. Excluding monolithic terms from testing results in 83.5% overall agreement (kappa=0.664).

We observed that 53% of the terms in our test set are right-branching while only 35% are left-branching. (Resnik, 1993) found between 64% and 67% of nominal compounds to be left-branching and used that finding to establish a baseline for his experiments with structural ambiguity resolution. (Nakov and Hearst, 2005) also report a similar percentage (66.8%) of left-branching noun compounds. Our test set is not limited to nominal compounds, which may account for the fact that a slight majority of the terms are found to be right-branching as adjectival modification in English is typically located to the left of the head noun. This may also help explain the fact that the method tends to have higher agreement within the set of right-branching terms (85%) vs. left-branching (62%).

We also observed that many of the terms marked as monolithic by the experts are of Latin origin such as the term in Example 9 or describe the functional status of a patient such as the term in Example 10.

$$\text{erythema}_1\ \text{ab}_2\ \text{igne}_3 \tag{9}$$

$$\text{difficulty}_1\ \text{swallowing}_2\ \text{solids}_3 \tag{10}$$

Example 10 merits further discussion as it illustrates another potential application of the method in the domain of functional status terminology. As was mentioned in the introduction, functional status terms may be be represented as a predication with a set of arguments. Such view of functional status terminology lends itself well to a frame-based representation of functional status terms in the context of a database such as FrameNet [3] or PropBank[4]. One of the challenging issues in representing functional status terminology in terms of frames is the distinction between the core predicate and the frame elements (Ruggieri et al., 2004). It is not always clear what lexical material should be part of the core predicate and what lexical material should be part of one or more arguments. Consider the term in Example 10 which represents a nominalized form of a predication. Conceivably, we could analyze this term as a frame shown in Example 11 where the predication consists of a predicate [DIFFICULTY] and two arguments. Alternatively, Example 12 presents a different analysis where the predicate is a specific kind of difficulty with a single argument.

$$\begin{array}{l}\text{[P:DIFFICULTY]}\\ \quad\text{[ARG1:SWALLOWING}_{<\text{ACTIVITY}>}\text{]}\\ \quad\text{[ARG2:SOLIDS}_{<\text{PATIENT}>}\text{]}\end{array} \tag{11}$$

$$\begin{array}{l}\text{[P:SWALLOWING DIFFICULTY]}\\ \quad\text{[ARG1: SOLIDS}_{<\text{PATIENT}>}\text{]}\end{array} \tag{12}$$

The analysis dictates the shape of the frames and how the frames would fit into a network of frames. The G^2 method identifies Example 10 as left-branching (Model 2), which suggests that it would be possible to have a parent DIFFICULTY frame and a child CLIMBING DIFFICULTY that would inherit form its parent. An example where this is not possible is the term “difficulty staying asleep” where it would probably be nonsensical or at least impractical to have a predicate such as [STAYING DIFFICULTY]. It would be more intuitive to assign this term to the DIFFICULTY frame with a frame element whose lexical content is “staying asleep”. The method appropriately identifies the term “difficulty staying asleep” as right-branching (Model 3) where the words “staying asleep” are grouped together. This is an example based on informal observations; however, it does suggest a utility in constructing frame-based representation of at least some clinical terms.

7 Limitations

The main limitation of the G^2 method is the exponential growth in the number of models to be evaluated with the growth in the length of the term. This limitation can be partly alleviated by either only considering adjacent models and limiting the length to 5-6 words, or using a forward or backward sequential search proposed by (Pedersen et al., 1997) for the problem of selecting models for the Word Sense Disambiguation task.

8 Conclusions and Future Work

This paper presented a simple but effective method based on G^2 to determine the internal structure of three-word noun phrase medical terms. The ability to determine the syntactic structure that gives rise to a particular semantic interpretation of a medical term may enable accurate mapping of unstructured medical text to standardized terminologies and nomenclatures. Future directions to improve the accuracy of our method include determining how other measures of association, such as dice coefficient and χ^2, perform on this task. We feel that there is a possibility that no single measure performs best over all types of terms. In that case, we plan to investigate incorporating the different measures into an ensemble-based algorithm.

We believe the model-fitting method is not limited to structural ambiguity resolution. This method could be applied to automatic term extraction and automatic text indexing of terms from a standardized vocabulary. More broadly, the principles of using distributional characteristics of word sequences derived from large corpora may be applied to unsupervised syntactic parsing.

[3] http://www.icsi.berkeley.edu/framenet/

[4] http://www.cis.upenn.edu/ ace/

Acknowledgments

We thank Barbara Abbott, Debra Albrecht and Pauline Funk for their contribution to annotating the test set and discussing aspects of medical terms.

This research was supported in part by the NLM Training Grant in Medical Informatics (T15 LM07041-19). Ted Pedersen's participation in this project was supported by the NSF Faculty Early Career Development Award (#0092784).

References

S. Banerjee and T. Pedersen. 2003. The design, implementation, and use of the Ngram Statistic Package. In *Proc. of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, February.

A.M. Buckeridge and R.F.E. Sutcliffe. 2002. Disambiguating noun compounds with latent semantic indexing. *International Conference On Computational Linguistics*, pages 1–7.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

D.A. Evans and C. Zhai. 1996. Noun-phrase analysis in unrestricted text for information retrieval. *Proc. of the 34th conference of ACL*, pages 17–24.

D. Hindle and M. Rooth. 1993. Structural Ambiguity and Lexical Relations. *Computational Linguistics*, 19(1):103–120.

M. Lapata and F. Keller. 2004. The Web as a Baseline: Evaluaing the Performance of Unsupervised Web-based Models for a Range of NLP Tasks. *Proc. of HLT-NAACL*, pages 121–128.

M. Lapata, S. McDonald, and F. Keller. 1999. Determinants of Adjective-Noun Plausibility. *Proc. of the 9th Conference of the European Chapter of ACL*, 30:36.

M. Lauer and M. Dras. 1994. A Probabilistic Model of Compound Nouns. *Proc. of the 7th Australian Joint Conference on AI.*

M. Lauer. 1995. Corpus Statistics Meet the Noun Compound: Some Empirical Results. *Proc. of the 33rd Annual Meeting of ACL*, pages 47–55.

M. Liberman and R. Sproat. 1992. The stress and structure of modified noun phrases in English. *Lexical Matters, CSLI Lecture Notes*, 24:131–181.

M.P. Marcus. 1980. *Theory of Syntactic Recognition for Natural Languages*. MIT Press Cambridge, MA, USA.

B.T. McInnes. 2004. Extending the log-likelihood ratio to improve collocation identification. Master's thesis, University of Minnesota.

R. Moore. 2004. On log-likelihood-ratios and the significance of rare events. In Dekang Lin and Dekai Wu, editors, *Proc. of EMNLP 2004*, pages 333–340, Barcelona, Spain, July. Association for Computational Linguistics.

P. Nakov and M. 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 (CoNLL-2005)*, pages 17–24, Ann Arbor, Michigan, June. Association for Computational Linguistics.

T. Pedersen, M. Kayaalp, and R. Bruce. 1996. Significant lexical relationships. In Howard Shrobe and Ted Senator, editors, *Proc. of the Thirteenth National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference, Vol. 2*, pages 455–460, Menlo Park, California. AAAI Press.

T. Pedersen, R. Bruce, and J. Wiebe. 1997. Sequential model selection for word sense disambiguation. In *Proc. of the Fifth Conference on Applied Natural Language Processing*, pages 388–395, Washington, DC, April.

J. Pustejovsky, P. Anick, and S. Bergler. 1993. Lexical semantic techniques for corpus analysis. *Computational Linguistics*, 19(2):331–358.

A. Ratnaparkhi. 1998. *Maximum Entropy Models for Natural Lnaguage Ambiguity Resolution*. Ph.D. thesis, University of Pennsylvania.

P. Resnik and M. Hearst. 1993. Structural Ambiguity and Conceptual Relations. *Proc. of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, June*, 22(1993):58–64.

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

A.P. Ruggieri, S. Pakhomov, and C.G. Chute. 2004. A Corpus Driven Approach Applying the "Frame Semantic" Method for Modeling Functional Status Terminology. *Proc. of MedInfo*, 11(Pt 1):434–438.

M.Q. Stearns, C. Price, KA Spackman, and AY Wang. 2001. SNOMED clinical terms: overview of the development process and project status. *Proc AMIA Symp*, pages 662–6.

S. S. Wilks. 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, March.