The Descent of Hierarchy, and Selection in Relational Semantics *

Barbara Rosario
SIMS
UC Berkeley
Berkeley, CA 94720-4600

Marti Hearst
SIMS
UC Berkeley
Berkeley, CA 94720-4600

Charles Fillmore
ICSI
UC Berkeley
Berkeley, CA 94720

ABSTRACT

In many types of technical texts, meaning is embedded in noun compounds. A language understanding program needs to be able to interpret these in order to ascertain sentence meaning. We explore the possibility of using the top levels of an existing lexical hierarchy for the purpose of placing words from a noun compound into categories, and then using this category membership to determine the relation that holds between the words. Since lexical hierarchies are not necessarily ideally suited for this task, we pose the question: how far down the hierarchy must the algorithm descend before all the terms within the subhierarchy behave uniformly with respect to the semantic relation in question? We present the results of an analysis on two-word noun compounds from the biomedical domain which suggests that we can obtain classification accuracy of approximately 90% while generalizing well over the data.

Introduction

A major difficulty for the interpretation of sentences from technical texts is the complex structure of noun phrases and noun compounds. Consider, for example, these titles taken from biomedical journal articles:

> *Congenital anomalies of tracheobronchial branching patterns.*
>
> *Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) is a genetically linked neurologic disease characterized by recurrent strokes and progressive or stepwise dementia, with or without migraine-like headaches, seizures, and pseudobulbar palsy.*

An important step towards being able to interpret such technical sentences is to analyze the meaning of noun compounds, and noun phrases more generally.

Interpretation of noun compounds (NCs) is highly dependent on lexical information. Thus we explore the use of a large corpus and a large lexical hierarchy to determine the relations that hold between the words in noun compounds. We use counts from a corpus to characterize the problem space, and thus determine which parts of the space are densely occupied.

We make a surprising finding with respect to the use of the lexical ontology. We find that we can simply use the juxtaposition of category membership within the hierarchy to determine the relation that holds between pairs of nouns. For example, for the NCs *leg paresis, skin numbness,* and *hip pain*, the first word of the NC falls into the A01 (Body Regions) category, and the second word falls into the C10 (Nervous System Diseases) category. From these we can declare that the relation that holds between the words is "location-of". Similarly, for *adenoma risk* and *cancer statistics*, the first word falls under C04 (Neoplasms) and the second is found in H01.548 (Mathematics), yielding the "measurement-of" relation. We find that we obtain 90% accuracy overall.

In some sense, this is a very old idea, dating back to the early days of semantic nets and semantic grammars. The critical difference now is that large lexical resources and corpora have become available, thus allowing some of those old techniques to become feasible in terms of coverage. However, the success of such an approach depends on the structure and coverage of the underlying lexical ontology.

In the remainder of the paper we discuss the linguistic motivations behind our approach, characteristics of the lexical ontology MeSH, the method of using the corpus to examine the problem space, the method of determining the relations and evaluating the results, related work, and conclusions.

Linguistic Motivation

One way to understand the relations between the words in a two-word noun compound is to cast the words into a head-modifier relationship, and assume that the head

*With apologies to Charles Darwin.

noun has an argument structure, much the way verbs do. Then the meaning of the head noun determines what kinds of things can be done to it, what it is made of, what it is a part of, and so on.

For example, consider the noun *knife*. Knives are created for particular activities or settings, can be made of various materials, and can be used for cutting or manipulating various kinds of things. A set of relations for knives, and example NCs exhibiting these results is shown below:

(Used-in): kitchen knife, hunting knife
(Material-of): steel knife, plastic knife
(Instrument-for): carving knife
(Used-on): meat knife, putty knife
(Used-by): chef's knife, butcher's knife

Note that some relationships apply to only certain classes of nouns, and the semantic structure of the head noun determines the range of possibilities. Thus if we can capture regularities about the behaviors of the constituent nouns, we should also be able to predict which relations will hold between them.

We propose using the categorization provided by a lexical hierarchy for this purpose. Furthermore, we avoid the need to enumerate in advance all of the relations that may hold. Rather, the corpus tells us which combinations actually occur.

The Lexical Hierarchy: MeSH

MeSH (Medical Subject Headings) is the National Library of Medicine's controlled vocabulary thesaurus; it consists of set of terms arranged in a hierarchical structure. There are 15 main sub-hierarchies (trees) in MeSH, each corresponding to a major branch of medical terminology. For example, tree A corresponds to Anatomy, tree B to Organisms, tree C to Diseases and so on. Every branch has several sub-branches; Anatomy, for example, consists of Body Regions (A01), Musculoskeletal System (A02), Digestive System (A03) etc., which we refer to as "level 0". We also refer to these as the main MeSH categories.

These nodes have children, for example, Abdomen (A01.047), Back (A01.176) are level 1 children of Body Regions. The longer the ID of the MeSH term, the longer the path from the root and the more precise the description. For example migraine is C10.228.140.546.800.525, that is, C (a disease), C10 (Nervous System Diseases), C10.228 (Central Nervous System Diseases) and so on. There are over unique 35,000 IDs in MeSH 2001, although many words are assigned more than one MeSH ID and thus occur in more than one location within the hierarchy; thus under some interpretations MeSH is actually a network.

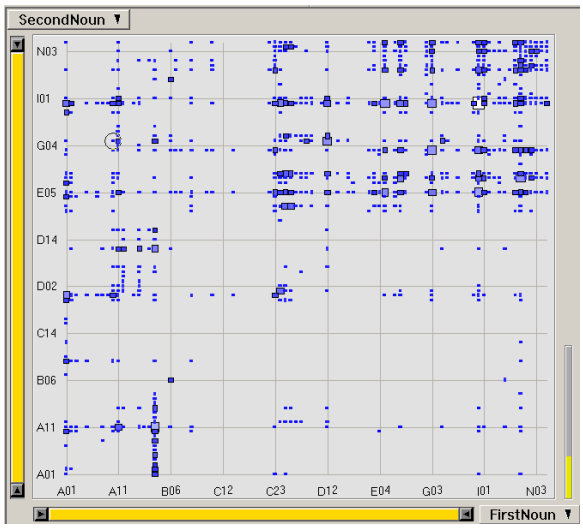

Figure 1: Distribution of Level 0 Category Pairs. Mark size indicates the number of unique NCs that fall under the CP. Only those for which > 50 NCs occur are shown.

Some of the categories are more heterogeneous than others. The tree A (Anatomy) for example, seems to be quite homogeneous; at level 1, the nodes are all *part of* (meronymic to) Anatomy; the Digestive (A03), Respiratory (A04) and the Urogenital (A05) Systems are all part of anatomy; the Biliary Tract (A03.159) and the Esophagus (A03.365) are part of the Digestive System and so on. Thus we assume that every node is a (body) part of the parent node (and all the nodes above it).

Tree C for Diseases is also homogeneous; the children nodes are a *kind of* (hyponym of) the disease at the parent node: Neoplasms (C04) is a *kind of* Disease C and Hamartoma (C04.445) is a *kind of* Neoplasms.

Other trees are more heterogeneous, in the sense that the relationships among the nodes are more diverse. Information Science (L01), for example, contains, among others, Communications Media (L01.178), Computer Security (L01.209) and Pattern Recognition (L01.725). Another heterogeneous sub-hierarchy is Natural Science H01. Among the children of H01 we find Chemistry (parent of Biochemistry), Electronics (parent of Amplifiers and Robotics), Mathematics (Fractals, Game Theory and Fourier Analysis). In other words, we find a wide range of concepts that are not related by a simple relationship.

Our hypothesis is that once we have descended to a homogeneous level within a subhierarchy, words falling into that subhierarchy behave similarly with respect to relation assignment.

Counting Noun Compounds

In this and the next section, we describe how we investigated the hypothesis:

> For all two-word noun compounds (NCs) that can be characterized by a category pair (CP), a particular semantic relationship (R) holds between the nouns comprising those NCs.

The kinds of relations we find are like those described in Section . Note that, in this analysis we focused on determining which sets of NCs fall into the same relation, without explicitly assigning names to the relations themselves. Furthermore, the same relation may be described a many different category pairs.

First, we extracted two-word noun compounds from approximately 1M titles and abstracts from the Medline collection of biomedical journal articles, resulting in about 1M NCs. Of these, 79677 are unique. The NCs were extracted by finding adjacent word pairs in which both words appear in the MeSH hierarchy, and neither word before or after the pair appears in MeSH. (Clearly this simple approach results in some erroneous extractions, but we identified nonsensical pairs during our analysis.)

Next we used MeSH to characterize the NCs according to the semantic category(ies) each noun is assigned to. For example, the NC *fibroblast growth* would be categorized into A11.329.228 (Fibroblasts) and G07.553.481 (Growth). Note that the same words can be represented at different levels of description; thus *fibroblast growth* can also be described by A11 (Cell) G07 (Physiological Processes) or A11.329 (Connective Tissue Cells) G07.553 (Growth and Embryonic Development). If a noun falls under more than one MeSH ID, we made multiple versions of this categorization. We will refer to this renaming as a category pair (CP).

We placed these CPs into a two-dimensional table, with the MeSH category for the first noun on the X axis, and the MeSH category for the second noun on the Y axis. Each intersection indicates the number of NCs that are classified under the corresponding two MeSH categories.

A visualization tool (Ahlberg and Shneiderman, 1994) allowed us to explore the dataset to see which areas of the category space are most heavily populated, and to get a feeling for if the distribution is uniform or not (see Figure 1). If our hypothesis holds (that NCs that fall within the same category pairs are assigned the same relation), then if most of the NCs fall within only a few category pairs then we only need to determine which relations hold between a subset of the possible pairs. Thus, the more clumped the distribution, the potentially more easy our task is.

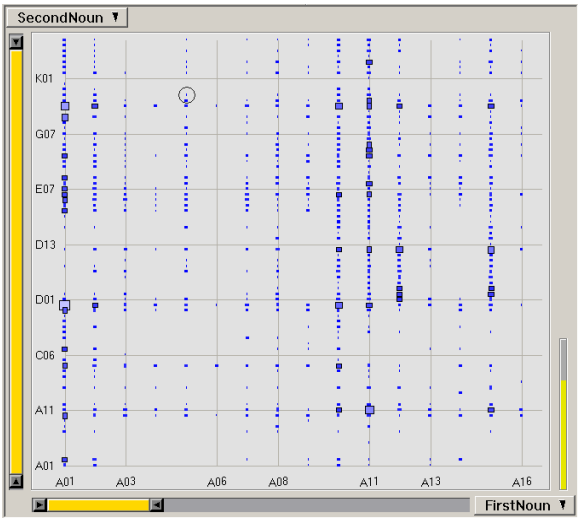

Figure 2: Distribution of Level 0 Category Pairs in which the first noun is from the A (Anatomy) category. Mark size indicates the number of unique NCs that fall under the CP.

Figure 1 shows that some areas in the CP space with a higher concentration of unique NCs (the Anatomy, and the E through N sub-hierarchies, for example), especially if we focus on those for which at least 50 unique NCs are found. Figure 2 focuses on the distribution of NCs for which the first noun can be classified under the Anatomy category. Note that many of the possible second noun categories are sparsely populated, again potentially reducing the space of the problem.

Labeling NC Relations

Given the promising nature of the NC distributions, the question remains as to whether or not the hypothesis holds. To answer this, we examined a subset of the CPs to see if we could find positions within the subhierarchies for which the relation assignments for the member NCs are always the same, and then assess the accuracy of this evaluation.

Method

We first selected a subset of the CPs to examine in detail. For each of these we examined, by hand, 20% of the NCs they cover, paraphrasing the relation between the nouns, and seeing if that paraphrase is the same for all the NCs in the group. If it is the same, then the current levels of the CP were the correct levels of description. If, on the other hand, several different paraphrases were found, then the analysis must descend one level of the hierarchy to see if the resulting partition of the NCs results in uniform relation assignments.

For example, all the following NCs were mapped to the same CP, A01 (Body Regions) and A07 (Cardiovascular System): *scalp arteries, heel capillary, shoulder artery, ankle artery, leg veins, limb vein, forearm arteries, fin-*

A01 H01 (Natural Sciences):
A01 H01 *abdomen x-ray, ankle motion*
A01 H01.770 (Science): *skin observation*
A01 H01.548 (Mathematics): *breast risk*
A01 H01.939 (Weights and Measures): *head calibration*
A01 H01.181 (Chemistry): *skin iontophoresis*
A01 H01.671 (Physics)
A01 H01.671.538 (Motion): *shoulder rotations*
A01 H01.671.100 (Biophysics): *shoulder biomechanics*
A01 H01.671.691 (Pressure): *eye pressures*
A01 H01.671.868 (Temp.): *forehead temperature*
A01 H01.671.768 (Radiation): *thorax x-ray,*
A01 H01.671.252 (Electricity): *chest electrode*
A01 H01.671.606 (Optics): *skin color*

Figure 3: Levels of descent needed for NCs classified under A01 H01.

ger capillary, eyelid capillary, forearm microcirculation, hand vein, forearm veins, limb arteries, thigh vein, foot vein.

All these NCs are "similar" in the sense that the relationships between the two words are the same. We did not therefore need to descend either hierarchy, and in future when we see an NC mapped to this CP, we will assign to it this semantic relationship. On the other hand, the following NCs, having the CP A01 (Body Regions) and M01 (Persons), did not have the same relationships between the component words: *abdomen patients, arm amputees, chest physicians, eye patients, skin donor.* The relationships are different depending on whether the person is a patient, a physician or a donor. We therefore descend the M01 sub-hierarchy, obtaining the following clusters of NCs:
A01 M01.643 (Patients): *abdomen patients, ankle inpatient, eye outpatient*
A01 M01.526 (Occupational Groups): *chest physician, eye nurse, eye physician*
A01, M01.898 (Donors): *eye donor, skin donor*
A01, M01.150 (Disabled Persons): *arm amputees, knee amputees.*
In other words, for correctly assigning a relationship to these NCs, we needed to descend one level for the second word. Figure 3 shows one CP for which we needed to descend 3 levels.

In our collection we have a total of 2627 of CPs at level 1 with at least 10 unique NCs, 798 (30%) of which have an A (Anatomy) as either the first or the second noun. We analyzed 250 of such CPs (randomly selected). The other word ranged across all MeSH categories.

We also analyzed 21 of the 90 CPs for which the second noun was H01 (Natural Sciences); we decided to analyze this portion of the MeSH hierarchy because the NCs with H01 as second noun are very frequent in our collection, and because we wanted to test the hypothesis that we do indeed need to descend farther for heterogeneous parts of MeSH.

We started with the CPs at level 0 for both words, descending when the corresponding clusters of NCs were not homogeneous and stopping when they were. We did this for 20% of the NCs in each CP. The results were as follows.

For 187 of 250 (74%) CPs with a noun in the Anatomy category, the classification remained at level 0 for both words (for example, A01 A07). For 55 (22%) of the CPs we had to descend 1 level (e.g., A01 M01: A01 M01.898, A01 M01.643) and for 7 CPs (2%) we descended 2 levels. We descended one level most of the time for the sub-hierarchies E (Analytical, Diagnostic and Therapeutic Techniques), G (Biological Sciences) and N (Health Care) (around 50% of the time for these categories combined). We never descended for B (Organisms) and did so only for A13 (Animal Structures) in A. This was to be able to distinguish a few non-homogeneous subcategories (e.g., milk appearing among body parts, thus forcing a distinction between *buffalo milk* and *cat forelimb*).[1]

For CPs with H01 as the second noun, of the 21 CPs analyzed, we observed the following (level number, count) pairs: (0, 1) (1, 8) (12, 2).

In all but three cases, the descending was done for the second noun only. This may be because the second noun usually plays the role of the head noun in two-word noun compounds in English, thus requiring more specificity. Alternatively, it may reflect the fact that for the examples we have examined so far, the more heterogeneous terms dominate the second noun. Further examination is needed to answer this decisively.

Evaluation

We tested the resulting classifications by developing a randomly chosen test set (20% of the NCs for each CP), entirely distinct from the labeled set, and used the classifications found above to automatically predict which relations should be assigned to the member NCs. We then checked these predictions by hand (this was done by an independent evaluator with biomedical training who is not part of the research team), and found impressive accuracies: For the CPs which contained a noun in the Anatomy domain, the assignments of new NCs were 94.2% accurate computed via intra-category averaging, and 91.3% accurate with extra-category averaging. For the CPs in the Natural Sciences we found 81.6% accuracy via intra-category averaging, and 78.6% accuracy with extra-category averaging.

[1] Although we began with 250 CPs in the A category, when a descend operation is performed, the CP is split into two or more CPs at the level below. Thus the total number of CPs after all assignments are made was 416.

We also tested 3 CP pairs in category C (Diseases); the most frequent CP with in terms of the total number of NCs (with repetitions) is C04 (Neoplasms) A11 (Cells), with 30606 NCs; the second CP was A10 C04 (27520 total NCs) and the fifth most frequent, A01 C04, with 20617 total NCs. Again, the testing was done with respect to unique NCs. On this dataset we obtained 100% accuracy.

The total accuracy across the portions of the A, H01 and C04 hierarchies that we analyzed were 89.6% via intra-category averaging, and 90.8% via extra-category averaging.

The lower accuracy for the Natural Sciences category indicates how our results depend on the properties of the lexical hierarchy. We can generalize well if the subhierarchies are in a well-defined semantic relation with their ancestors. If they are a list of "unrelated" topics, we cannot use the generalization of the higher levels; most of the mistakes for the Natural Sciences CPs occurred in fact when we failed to descend for broad terms such as Physics. Performing this evaluation allowed us to find such problems and update the rules;o the resulting categorization should now be more accurate.

Generalization

An important issue is whether this is an economic way of classifying the NCs. The advantage of the high level description is, of course, that we need assign by hand many fewer NCs than if we used all CPs at their most specific levels. Our approach provides generalization over the "training" examples in two ways. First, we find that we can use the juxtaposition of categories in a lexical hierarchy to identify semantic relationships. Second, we find we can use the higher levels of these categories for the assignments of these relationships.

To assess the degree of this generalization we calculated how many CPs are accounted for by the classification rules created above for the Anatomy categories. In other words, if we know that A01 A07 unequivocally determines a relationship, how many possible (i.e., present in our collection) CPs are there that are "covered by" A01 A07 and that we do not need to consider explicitly? It turns out that our 415 classification rules cover 46001 possible CP pairs. This, and the fact that we achieve high accuracies with these classification rules, show that we successfully use MeSH to generalize over unique NCs.

Related Work

Noun Compound Relation Assignment

Several approaches have been proposed for empirical noun compound interpretation. Lauer and Dras (1994) point out that there are three components to the problem: identification of the compound from within the text, syntactic analysis of the compound (left versus right association), and the interpretation of the underlying semantics. Several researchers have tackled the syntactic analysis (Lauer, 1995; Pustejovsky et al., 1993; Liberman and Church, 1992), usually using a variation of the idea of finding the subconstituents elsewhere in the corpus and using those to predict how the larger compounds are structured.

We are interested in the third task, interpretation of the underlying semantics. Most related work relies on hand-written rules of one kind or another. Finin (1980) examines the problem of noun compound interpretation in detail, and constructs a complex set of rules. Vanderwende (1994) uses a sophisticated system to extract semantic information automatically from an on-line dictionary, and then manipulates a set of hand-written rules with hand-assigned weights to create an interpretation. Rindflesch et al. (2000) use hand-coded rule-based systems to extract the factual assertions from biomedical text. Lapata (2000) classifies nominalizations according to whether the modifier is the subject or the object of the underlying verb expressed by the head noun.

Barker and Szpakowicz (1998) describe noun compounds as triplets of information: the first constituent, the second constituent, and a marker that can indicate a number of syntactic clues. Relations are initially assigned by hand, and then new ones are classified based on their similarity to previously classified NCs. However, similarity at the lexical level means only that the same word occurs; no generalization over lexical items is made. The algorithm is assessed in terms of how much it speeds up the hand-labeling of relations.

Rosario and Hearst (2001) demonstrate the utility of using a lexical hierarchy for assigning relations to two-word noun compounds. They use machine learning algorithms and MeSH to successfully generalize from training instances, achieving about 60% accuracy on an 18-way classification problem using a very small training set. Their approach is bottom up and requires good coverage in the training set; the approach described in this paper is top-down, characterizing the lexical hierarchies explicitly rather than implicitly through machine learning algorithms.

Using Lexical Hierarchies

Many approaches attempt to automatically assign semantic roles (such as case roles) by computing semantic similarity measures across a large lexical hierarchy; primarily using WordNet (Fellbaum, 1998). Resnik's well-known conceptual association algorithm uses WordNet for this purpose (Resnik, 1996).

However, it is uncommon to simply use the hierarchy directly for generalization purposes. Many researchers have noted that WordNet's words are classified into senses

that are too fine-grained for standard NLP tasks. For example, Buitelaar (1997) notes that the noun *book* is assigned to seven different senses, including *fact* and *section, subdivision*. Thus most users of WordNet must contend with the sense disambiguation issue in order to use the lexicon.

Another reason may be that the problems to which lexical hierarchies have been applied have not been well-suited for classification via membership in the hyponymy relation. For example, Resnik has applied his conceptual association algorithm to the prepositional phrase attachment problem (Resnik and Hearst, 1993; Brill and Resnik, 1994). As set up by Hindle and Rooth (1991), the problem is to determine, for a sequence V N P N, whether the second noun attaches to the verb or the first noun. Consider the following minimal pair:

(i) *eat spaghetti with a fork*
(ii) *eat spaghetti with sauce*

In the PP attachment problem, one has to determine which is a more likely association: fork and eat, or fork and spaghetti. If a hierarchical lexicon is used to compute similarity between fork/spaghetti and fork/eat, because these are not linked *hierarchically*, the generalization provided by the hierarchy is not sufficient, and more general kinds of associations must be computed.

One difficulty with the standard PP attachment problem formulation is that fork/spaghetti and sauce/eat are both related to each other, but they are related to each other in two different ways. We instead are asking: what is the relationship between fork and spaghetti? Between sauce and spaghetti? We argue that obtaining the answer to these types of questions is more useful than obtaining an answer to the attachment question.

The most closely related use of a lexical hierarchy that we know if is that of Li and Abe (1998), which uses an information-theoretic measure to make a cut through the top levels of the noun portion of WordNet. This is then used to determine acceptable classes for verb argument structure, and for the PP attachment problem described above, and is found to perform as well as or better than existing algorithms.

Additionally, Boggess et al. (1991) "tag" veterinary text using a small set of semantic labels, assigned in much the same way a parser works, and describe this in the context of PP attachment.

Conclusions and Future Work

We have provided evidence that the upper levels of a lexical hierarchy can be used to accurately classify the relations that hold between two-word technical noun compounds. It is surprising that this technique works as well as it does. In this paper we focus on biomedical terms using the biomedical lexical ontology MeSH. It may be that such technical, domain-specific terminology is better behaved than NCs drawn from a more general domain; we will have to assess the technique in other domains to fully assess its applicability.

Several issues need to be explored further. First, we need to ensure that this technique works across the full spectrum of the lexical hierarchy. We have demonstrated the likely usefulness of such an exercise, but all of our analysis was done by hand. It may be useful enough to simply complete the job manually; however, it would also be nice to automate some or all of the analysis. There are several ways to go about this. One approach would be to use existing statistical similarity measures (e.g., (Lin, 1998), (Resnik, 1996)) to attempt to identify which subhierarchies are homogeneous. Another approach would be to see if, after analyzing more CPs, that those categories found to be heterogeneous should be assumed to be heterogeneous across classifications, and similarly for those that seem to be homogeneous.

The second major issue to address is how to extend the technique to multi-word noun compounds. We will need to distinguish between NCs such as *acute migraine treatment* and *oral migraine treatment*, and handle the case when the relation must first be found between the left-most words. Thus additional steps will be needed; one approach is to compute statistics to indicate likelihood of the various CPs. The third pressing issue is to determine if this technique works for non-technical NCs.

Finding noun compound relations is part of our larger effort to investigate what we call statistical semantic parsing (as in (Burton and Brown, 1979); see Grishman (1986) for a nice overview). For example, we would like to be able to interpret titles terms of semantic relations, for example, transforming *Congenital anomalies of tracheobronchial branching patterns* into a form that allows questions to be answered such as "What kinds of irregularities can occur in lung structure?" We hope that by compositional application of relations to entities, such inferences will be possible.

Acknowledgements We thank Kaichi Sung for her work on the relation labeling. This research was supported by a grant from ARDA.

References

Christopher Ahlberg and Ben Shneiderman. 1994. Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In *Proceedings of ACM CHI'94*, pages 313–317.

Ken Barker and Stan Szpakowicz. 1998. Semi-

automatic recognition of noun modifier relationships. In *Proceedings of COLING-ACL '98*, Montreal, Canada.

Lois Boggess, Rajeev Agarwal, and Ron Davis. 1991. Disambiguation of prepositional phrases in automatically labelled technical text. In *AAAI 91*, pages 155–159.

Eric Brill and Philip Resnik. 1994. A rule-based approach to prepositional phrase attachment disamibuation. In *Proceedings of COLING-94*.

P. Buitelaar. 1997. A lexicon for underspecified semantic tagging. In *Proceedings of ANLP 97, SIGLEX Workshop*, Washington DC.

R. R. Burton and J. S. Brown. 1979. Toward a natural-language capability for computer-assisted instruction. In H. O'Neil, editor, *Procedures for Instructional Systems Development*, pages 273–313. Academic Press, New York.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Timothy W. Finin. 1980. *The Semantic Interpretation of Compound Nominals*. Ph.d. dissertation, University of Illinois, Urbana, Illinois.

Ralph Grishman. 1986. *Computational Linguistics*. Cambridge University Press, Cambridge.

Donald Hindle and Mats Rooth. 1991. Structural ambiguity and lexical relations. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*.

Maria Lapata. 2000. The automatic interpretation of nominalizations. In *Proceedings of AAAI*.

Mark Lauer and Mark Dras. 1994. A probabilistic model of compound nouns. In *Proceedings of the 7th Australian Joint Conference on AI*.

Mark Lauer. 1995. Corpus statistics meet the compound noun. In *Proceedings of the 33rd Meeting of the Association for Computational Linguistics*, June.

Hang Li and Naoki Abe. 1998. Generalizing case frames using a thesaurus and the MDI principle. *Computational Linguistics*, 24(2):217–244.

Mark Y. Liberman and Kenneth W. Church. 1992. Text analysis and word pronunciation in text-to-speech synthesis. In Sadaoki Furui and Man Mohan Sondhi, editors, *Advances in Speech Signal Processing*, pages 791–831. Marcel Dekker, Inc.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL98*.

James Pustejovsky, Sabine Bergler, and Peter Anick. 1993. Lexical semantic techniques for corpus analysis. *Computational Linguistics*, 19(2).

Philip Resnik and Marti A. Hearst. 1993. Structural ambiguity and conceptual relations. In *Proceedings of the ACL Workshop on Very Large Corpora*, Columbus, OH.

Philip Resnik. 1996. Selectional constraints: an information theoretical model and its computational realization. *Cognition*, 61:127–159.

Thomas Rindflesch, Lorraine Tanabe, John N. Weinstein, and Lawrence Hunter. 2000. Extraction of drugs, genes and relations from the biomedical literature. *Pacific Symposium on Biocomputing*, 5(5).

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

Lucy Vanderwende. 1994. Algorithm for automatic interpretation of noun sequences. In *Proceedings of COLING-94*, pages 782–788.