Classifying the Semantic Relations in Noun Compounds via a Domain-Specific Lexical Hierarchy

Barbara Rosario and Marti Hearst
School of Information Management & Systems
University of California, Berkeley
Berkeley, CA 94720-4600
{*rosario,hearst*}*@sims.berkeley.edu*

Abstract

We are developing corpus-based techniques for identifying semantic relations at an intermediate level of description (more specific than those used in case frames, but more general than those used in traditional knowledge representation systems). In this paper we describe a classification algorithm for identifying relationships between two-word noun compounds. We find that a very simple approach using machine learning algorithms and a domain-specific lexical hierarchy successfully generalizes from training instances, performing better than a baseline consisting of training on the words themselves.

1 Introduction

We are exploring empirical methods of determining semantic relationships between constituents in natural language. Our current project focuses on biomedical text, both because it poses interesting challenges, and because it should be possible to make inferences about propositions that hold between scientific concepts within biomedical texts (Swanson and Smalheiser, 1994).

One of the important challenges of biomedical text, along with most other technical text, is the proliferation of noun compounds. A typical article title is shown below; it consists a cascade of four noun phrases linked by prepositions:

> *Open-labeled long-term study of the efficacy, safety, and tolerability of subcutaneous sumatriptan in acute migraine treatment.*

Clearly the real concern in analyzing such a title is in determining the relationships that hold between different concepts, rather than on finding the appropriate attachments (which is especially difficult given the lack of a verb). And clearly before we tackle the prepositional phrase attachment problem, we must find a way to analyze the meanings of the noun compounds.

Our goal is to extract propositional information from text, and as a step towards this goal, we classify constituents according to which semantic relationships hold between them For example, in the medical title example above, we are interested in converting *acute migraine treatment* into propositions such as:

- Therapy(migraine,treatment)
- Degree(acute,migraine), and
- Therapy(degree(acute,migraine),treatment)

By contrast, the term *intranasal migraine treatment* would have the relationships

- Therapy(migraine,treatment),
- AdministrationMethod(intranasal,treatment), and
- AdministrationMethod(intranasal, Therapy(migraine,treatment)).

(Some propositions will have arity greater than two.) A representation like this allows for control over the detail of the representation, since some systems may care how the sumatriptin is administered, while others may not, and some may care about treatments for any kind of migraine, while others will want to distinguish between acute and other forms of migraine.

These relations are intended to be combined to produce larger propositions that can then be used in a variety of interpretation paradigms, such as abductive reasoning (Hobbs et al., 1993) or inductive logic programming (Ng and Zelle, 1997).

Note that because we are concerned with the semantic relations that hold between the concepts, as opposed to the more standard, syntax-driven computational goal of determining left versus right association (or noun versus verb attachment, in the case of prepositional phrase disambiguation), this has the fortuitous effect of changing the problem into one of classification, amenable to standard machine learning classification techniques.

We have found that we can use such algorithms to classify relationships between two-word noun compounds with a surprising degree of accuracy, if we take advantage of lexical ontologies to generalize

over the words in the training sets. Yes-no (binary) classification using decision trees yielded 15% over the baseline, and a one-out-of-thirteen classification using a neural net achieves accuracies as high as 70%. Thus, we think this is a promising approach for a variety of semantic labeling tasks.

The reminder of this paper is organized as follows: Section 2 describes related work, Section 3 describes the semantic relations and how they were chosen, and Section 4 describes the data collection and ontologies. In Section 5 we describe the algorithms for assigning semantic relations to noun compounds, and report the results of experiments using this algorithm. Section 6 concludes the paper and discusses future work.

2 Related Work

Several approaches have been proposed for empirical noun compound interpretation. (Lauer and Dras, 1994) points out that there are three components to the problem: identification of the compound from the rest of 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 Sproat, 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 constructed 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.

In the related subarea of information extraction (Cardie, 1997; Riloff, 1996), the main goal is to find every instance of particular entities or events of interest. These systems use empirical techniques for learning which terms signal entities of interest, to fill in a pre-defined frame. Our goals are more general than those of information extraction, and so should be helpful for that task. However, our approach requires that terms in question appear in a lexical hierarchy, while often the terms that are most important for information extraction are proper nouns or other terms that are not classified in advance.

There have been several efforts to incorporate lexical hierarchies into statistical processing, primarily for the problem of prepositional phrase (PP) attachment. The current standard formulation is: given a verb followed by a noun and a prepositional phrase, represented by the tuple $v, n1, p, n2$, determine which of v or $n1$ the PP consisting of p and $n2$ attaches to, or is most closely associated with. Because the data is sparse, empirical methods that train on word occurrences alone (Hindle and Rooth, 1993) have been supplanted by algorithms generalize one or both of the nouns according to class-membership measures (Resnik, 1993; Resnik and Hearst, 1993; Brill and Resnik, 1994; Li and Abe, 1998), but the statistics are computed for the particular preposition and verb.

It is not clear how to use the results of such analysis after they are found; the semantics of the relationship between the terms must still be determined. In our framework we would cast this problem as finding the relationship $R(p, n2)$ that best characterizes the preposition and the NP that follows it, and then seeing if the categorization algorithm determines their exists any relationship $R'(n1, R(p, n2))$ or $R'(v, R(p, n2))$.

The algorithms used by this related work reflect the fact that they condition probabilities on a particular verb and noun. (Resnik, 1993; Resnik, 1995) use classes in Wordnet (Fellbaum, 1998) and a measure of conceptual association to generalize over the nouns. (Brill and Resnik, 1994) use Brill's transformation-based algorithm along with simple counts within a lexical hierarchy in order to generalize over individual words. (Li and Abe, 1998) use a minimum description length-based algorithm to find an optimal tree cut over WordNet for each classification problem, finding improvements over both lexical association (Hindle and Rooth, 1993) and conceptual association, and equaling the transformation-based performance. Our approach differs from these in that we are using machine learning techniques to determine which level of the lexical hierarchy is appropriate for the nouns that participate in each relation type.

3 Noun Compound Relations

The problem remains of determining what the appropriate kinds of relations are. In theoretical linguistics, there are contradictory views regarding the semantic properties of noun compounds. Levi (Levi, 1978) argues that there exists a small set of semantic relationships that NCs may imply. Downing (Downing, 1977) argues that the semantics of NC cannot be exhausted by any finite listing of relationships.

In this work we aim for a representation that is intermediate in generality between standard case roles (such as Agent, Patient, Topic, Instrument), and the specificity required for information extraction. We have created a set of relations that are sufficiently

general to cover a significant number of noun compounds, but that can be domain specific enough to be useful in analysis. We want to support relationships between entities that are shown to be important in cognitive linguistics, in particular we intend to support the kinds of inferences that arise from Talmy's force dynamics (Talmy, 1985). It has been shown that relations of this kind can be combined in order to determing the "directionality" of a sentence (e.g,., whether or not a politician is in favor of, or against, a proposal) (Hearst, 1990).

(Vanderwende, 1995) summarizes a set of relations that have been studied in theoretical linguistics (Downing, 1977; Levi, 1978), formulating the relations in terms of what kind of *wh*-question they answer. The relations we have produced so far parallel most of these, although some have domain-specific refinements, and some are outright additions. Table 1) shows the set for which an adequate number of examples (at least 50) were found in the current collection, although we have identified about 40 relations. For example, the notion of damage-to or obstruction-of are important relationship in the medical domain.

Two categories are especially problematic. Some compounds are noncompositional or lexicalized, such as *vitamin k* and *guinea pig*, and these are placed in a catch-all category. A second such category consisted of those for which a well-defined relationship could not be determined, they are simply subtypes of one another. This group includes *migraine headache*, *amnesia attack*, and *ulcer disease*. Not surprisingly, these two categories were the worst performers for the algorithm.

We also had to contend also with terms that are not strictly medical, such as *family life*, *ice water*, *cat ownership*, but we only retained those compounds for which both words could be found in the medical ontology.

The relations were found by iterative refinement based on looking at 1858 extracted compounds and finding commonalities among them. We expect to continue development and refinement of these relationship types, based on what ends up clearly being useful "downstream" in the analysis.

4 Collection and Lexical Resources

To create a colletion of noun compounds, we performed searches from Medline, which contains references and abstracts from 4300 biomedical journals. We used several query terms and we considered only the titles and the abstracts of the retrieved documents.

There are rich (English) lexical resources available for biomedical subject domain, in particular the MeSH subject headings (which are manually assigned to Medline journal articles), and the National Library of Medicine's UMLS metathesaurus (Humphreys et al., 1998). The UMLS is comprised of three resources: the *Metathesaurus*, the *Semantic Network* and the *SPECIALIST lexicon*. For this work we considered only the Metathesaurus, which contains semantic information about biomedical concepts, their various names, and the relationships among them. It is organized by concept, and links alternative names and views of the same concept together. The 2000 edition of the Metathesaurus includes about 730,000 concepts and 1.5 million concept names.

MeSH (Lowe and Barnett, 1994) is one of the source vocabularies of UMLS and its concepts are identified by unique concept identifiers that have a hierarchical structure. There are 15 main trees in MeSH, each corresponding to a major branch of medical ontology. For example, Tree A corresponds to Anatomy, tree B to Organisms, and so on. There are about 19,000 unique main terms in MeSH, and additional modifiers. The longer the tree position, 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, a C (a disease), C10 (Nervous System Diseases), C10.228 (Central Nervous System Diseases) and so on.

We mapped the noun compounds into MeSH using synonym information drawn from the UMLS. The mapping of the noun compounds to Metathesaurus concepts is a difficult problem in itself. For this study, we implemented a very simple mapping: each word was mapped to a UMLS concept (called a CUI), and whenever multiple concepts were found, only the first one was considered. We considered only those noun compounds for which both nouns can be mapped into MeSH terms.

5 Algorithm and Results

Because we have defined the problem as a classification problem, we can make use of standard classification algorithms. In particular, we used a neural network to classify across all relations simultaneously, and decision trees to make binary (yes-no) classifications on a per-relation basis.

We ran the experiments creating models that used different levels of MeSH. For example, Model 2 meant that both words in the noun compound were represented by a descriptor that showed the top level MeSH category as well as the label for the subcategory just below it. To illustrate how we created the feature vectors for each level of description, consider the following example:

- NC : flu vaccination (relation class 5)
  - CUIs: C0016366|C0042196
  - MeSH: D4.808.54.79.429.154.349|G3.770.670.310.890

	N	Name	Examples
1	170	Lexicalized	cd8 mab, hpv antigen, vitamin b
2	124	Subtype of	headaches migraines, family doctors,weight peptides
3	39	Activity	gallbladder absorption, bile secretion, virus transmission
4	41	Production (genetic)	cmv mrna, haptoglobin gene
5	53	Purpose	asthma therapy, headache medication, tumor treatment
6	47	Person afflicted	aids patient, hiv persons, headache sufferer
7	64	Study Instrument	blood analyses, headache history, chest radiograph
8	82	Cause	food allergy, asthma hospitalizations, varicella pneumonia
9	97	Location	lung mucosa, brain tissue, hospital beds
10	45	Measure	attack duration, headache incidence, hospitalisation rate
11	48	Damage	growth abnormalities, artery aneurysm, enzyme deficiencies
12	101	Instrument	light endoscopy, cyclosporin treatment
13	59	Procedure	brain biopsy, cell culture, asthma diagnosis

Table 1: The lexical relations, showing the number of labeled examples for each, and some example compounds.

1. Level 2: main tree and first level down (flu vaccination → D 4 G 3)
2. Level 3: main tree and two levels down (flu vaccination → D 4 808 G 3 770)
3. Level 4: main tree and three levels down (flu vaccination → D 4 808 54 G 3 770)
4. Level 5: main tree and four levels down (flu vaccination → D 4 808 54 79 G 3 770 670)
5. Level 6: main tree and five levels down (flu vaccination → D 4 808 54 79 429 G 3 770 670 310)

Whenever a word was mapped to a general MeSH term (like treatment, Y11) we added zeros for the missing values (so that treatment in level 3 is Y 11 0, in level 4, Y 11 0 0 and so on). The NCs were used as the input for both neural networks and decision trees. The numbers in the MeSH descriptors are categorical values; we represented them with indicator variables. That is, for each variable we calculated the number of possible categories c and then represented an observation of the variable as a sequence of c binary variables in which one binary variable was one and the remaining $c-1$ binary variables were zero.

We also considered a representation in which the words themselves were used as categorical input variables (we call this representation "lexical"). For this collection of NCs there were 1336 unique nouns and therefore the feature vector for each noun had 1336 components.

In Table 2 we report the length of the feature vectors for one noun for each model. The whole NC was then described by concatenating the feature vectors for the two nouns in sequence.

5.1 Neural Networks and Multi-way Classification

The NC represented in this fashion were the input to a neural network. We used a feed-forward network trained with conjugate gradient. The network had one hidden layer, in which a hyperbolic tangent function was used, and an output layer representing the 13 classes. A logistic sigmoid function was used in the output layer to map the outputs into the interval $(0, 1)$.

Model	Feature Vectors
2	83
3	463
4	938
5	1270
6	1443
Lexical	1336

Table 2: Length of the feature vectors for different numbers of MeSH descriptors.

The number of units of the output layer was the number of classes (13) and therefore fixed; as for the hidden layer we trained the network for several choices of numbers of hidden units and chose the best-performing networks based on training set error for each of our models. We subsequently tested these networks on held-out testing data (the data was split into 80% training and 20% testing for each relation).

We compared our results with a baseline in which logistic regression was used on the lexical features. Given our indicator variable representation of these features, this logistic regression essentially forms a table of log-odds for each lexical item. We also compared to a method in which the lexical indicator variables were used as input to a neural network. This approach provides the neural network with the maximum amount of information, and it is of interest to see to what extent, if any, our MeSH-based features lead to a reduction of performance. Note also that this lexical-neural-network approach is feasible in our setting given that the number of unique words

is limited (1336)—such an approach would not scale to larger problems.

In Table 3 and in Figure 1 we report the results from these experiments. The results indicate that the best results are obtained for level 6 of description using the MeSH terms to represent the nouns, but that reasonable performance is also obtained for only two MeSH terms.

Multi-class classification is a difficult problem (Vapnik, 1998). In particular, the guessing baseline yields less than 10 percent accuracy in this problem. We see that our method is a significant improvement over the tabular logistic-regression-based approach, which yields an accuracy of only 43 percent. Moreover, our results also show that despite the significant reduction in raw information content from the lexical representation to the MeSH representation, the MeSH-based neural network performs as well as the lexical-based neural network. (And we again stress that the lexical-based neural network is not a viable option for larger domains).

We also tested the robustness of our MeSH-based model to the case of unseen words. To do this, we partitioned the testing set into 3 subsets:

- Case 1: NCs for which one noun (or, rather, one MeSH description of the noun) was not present in the training set
- Case 2: NCs for which both nouns were present in the training set
- Case 3: NCs for which an identical representation was in the training set.[1]

In Table 4 and in Figure 2 we present the accuracy of these test set partition for 3 levels (2, 3 and 6). As expected, the accuracy of the data for which only one noun was seen is lower and we get the maximum accuracy for the identical testing points. However, at the best level of description, level 6, the accuracies are very similar. This seems to indicate that our method is quite robust to the case of unseen examples.[2]

[1]This does not occur for the lexical case, because our collection has only unique NCs. However, the mapping through the metathesaurus allows some NCs get mapped to the same MeSH descriptors. Also, while we had NCs in this testing set for which *neither* word was present in the training set, the mapping again allows these words to be represented with a feature vector present in the training set. This is another important advantage of our method. The mapping is the first step of generalization, and if the mapping is correct, the machine learning techniques will take advantage of this.

[2]Note that for unseen words, the baseline lexical-based logistic regression approach, which essentially builds a tabular representation of the log-odds for each class, reduces to random guessing.

Model	Whole test set	Case 1	Case 2	Case 3
2	0.5670	0.2000	0.4157	0.6600
3	0.5979	0.5536	0.6364	0.7321
6	0.7010	0.7284	0.6790	0.8333

Table 4: Test accuracy in the case of unseen examples

5.2 Decision Trees and Binary Classifications

We also ran experiments using a decision tree (CART (Breiman et al., 1984)). The are advantages and disadvantages of using a decision tree versus a neural network. A decision tree allows to investigate which features are the most important and we wanted to explore this. Decision trees are very natural for binary classification and this allowed us to see which classes were harder/easier to classify.

Unfortunately, CART and other off-the-shelf decision tree software does not allow for very long categorical feature vectors of the kind described in the previous session. For this reason, we used numerical values (except for the two main tree)[3]. This had however the advantage of being able to look up easily in the CART trees the most important variables. A systematic analysis of this issue remains to be done.

Our input to CART was then (referring to the example "flu vaccination" of Section 5 of the kind *D 4 808 54 79 429 G 3 770 670 310* for level 6 or *D 4 G 3* for level 2. The variables that indicate the main trees (*D* and *G*) were categorical, and the other numerical. For each class we trained and tested that class against all the others.

In Table 5 we report the accuracies for each class using the MeSH descriptors, the levels of the MeSH hierarchy for which we get the highest accuracy, and the accuracies for the lexical case. (In future work the best level will be determined via cross-validation.) In Figure 3 appear the corresponding plots. We can see that the poorest performances are for very general classes, class 1 and 2 (see definitions in Table 1) and for class 5. This NC classified under this class seem to be quite varied.

6 Conclusions and Future Work

We have presented a simple approach to corpus-based assignment of semantic relations for noun compounds. The main idea is to define a set of relations that can hold between the terms and use standard machine learning techniques and lexical hierarchy to generalize from training instances to new examples. The initial results are quite promising.

[3]We did run CART with categorical features for models 2 and 3 and the results were not significally different

Model	Accuracy1	Accuracy2	Accuracy3	Hidden Units
Logistic Regression on Lexical	0.4213	0.4719	0.4944	–
Neural Network on Lexical	0.6573	0.7472	0.7528	55
2	0.5670	0.6495	0.7165	20
3	0.5979	0.7165	0.7887	30
4	0.6546	0.7474	0.7990	45
5	0.6237	0.7577	0.7938	50
6	0.7010	0.7732	0.8196	10

Table 3: Test accuracy for each model, where model number corresponds to the level of the MeSH hierarchy used for classification. Accuracy1 refers to how often the correct relation is the top-scoring relation, Accuracy2 refers to how often the correct relation is one of the top two according to the neural net, and so on. The average accuracy for guessing is 0.077.

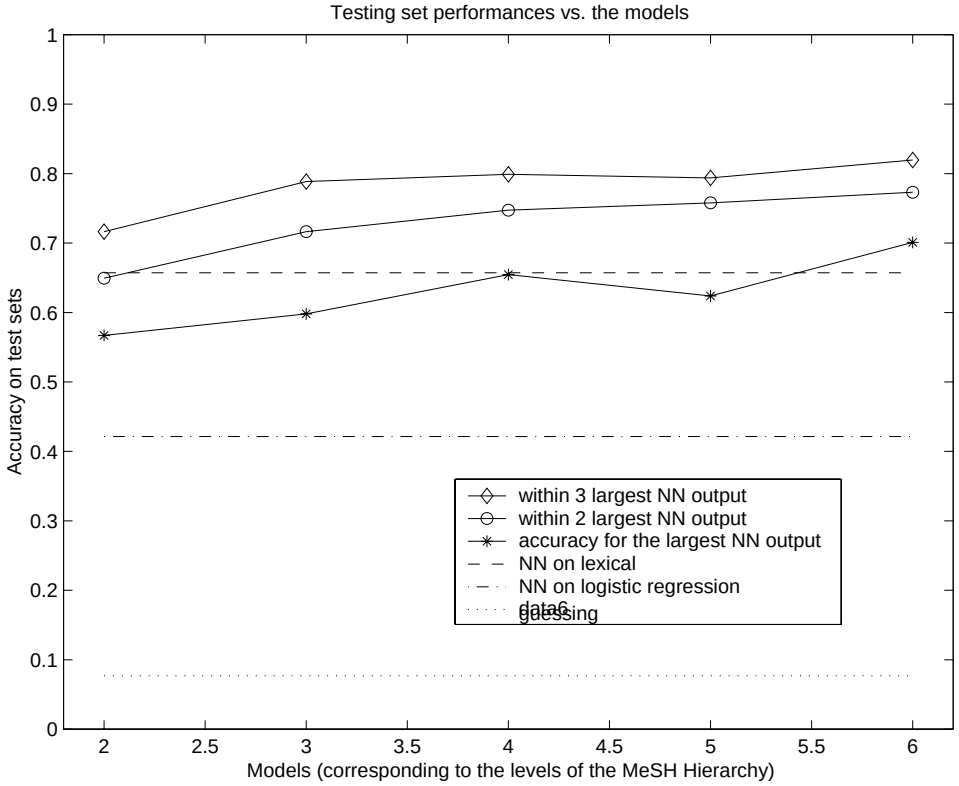

Figure 1: Accuracies on the test sets for all the models. The dotted line at the bottom is the accuracy of guessing (the inverse of the number of classes). The dash-dot line above this is the accuracy of logistic regression on the lexical data. The last flat line (the dashed one) is the performance of the neural network on lexical inputs. The solid line with asterisks is the accuracy of our representation, when only the maximum output value from the network is considered. The solid line with circles if the accuracy of getting the right answer within the two largest output values from the neural network and the last solid line with diamonds is the accuracy of getting the right answer within the first three outputs from the network.

In future we plan to train the algorithm allowing different levels for each noun in the compound, and compare the results to the tree cut algorithm reported in (Li and Abe, 1998), which allows different levels to be identified for different subtrees. We also plan to tackle the proble of noun compounds containing more than two terms.

Acknowledgments

We would like to thank Nu Lai for the help of with the classification of the noun compounds.

References

L. Breiman, J. Friedman, R. Olsen, and C. Stone. 1984. *Classification and Regression Trees.*

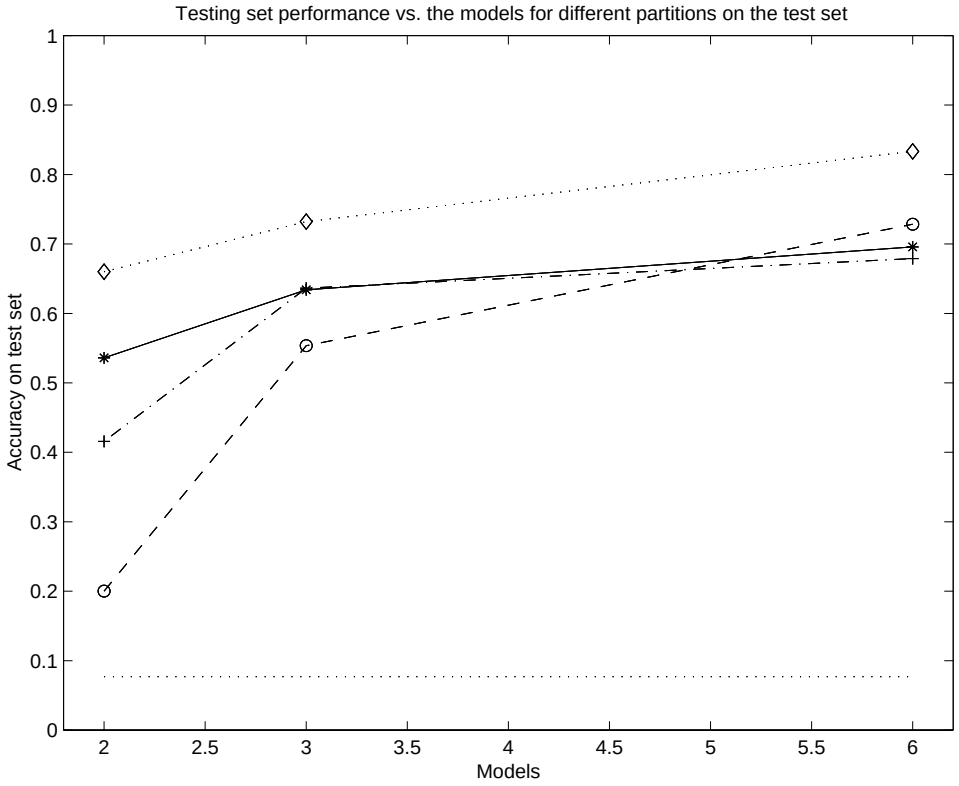

Figure 2: The solid line with asterisks is the accuracy of the entire test set. The dashed line with circles is for the case of one unseen NC, dash-dot line with plus is for the case of both NCs present in the training set and finally, the dotted line at the top is for the NCs for which identical descriptions were found in the training set. All these curves refer to the case of getting exactly the right answer.

Class	Lexical	Class-based	MeSH Level
3	0.71	0.78	4
4	0.55	0.67	6
5	0.94	1.00	3
7	0.88	0.93	4
14	0.86	0.85	2
15	0.90	1.00	2
18	0.65	0.88	2
20	0.56	0.88	4
21	0.62	0.76	2
23	0.82	0.89	4
27	0.56	0.89	2
33	0.72	0.90	5
36	0.82	0.95	3
avg	0.74	0.87	–

Table 5: Decision tree accuracy on binary classification judgements for each class. Lexical is the average score for the decision tree using words alone, Best Class is the average score for the decision tree on the level at which it does best for that class, and Level shows the best level for that class. The optimal level will be determined in future via cross-validation.

Wadsworth.

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

Claire Cardie. 1997. Empirical methods in information extraction. *AI Magazine*, 18(4).

P. Downing. 1977. On the creation and use of english compound nouns. *Language*, (53):810–842.

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.

Marti A. Hearst. 1990. A hybrid approach to restricted text interpretation. In Paul S. Jacobs, editor, *Text-Based Intelligent Systems: Current Research in Text Analysis, Information Extraction, and Retrieval*, pages 38–43. GE Research & Development Center, TR 90CRD198.

Donald Hindle and Mats Rooth. 1993. Structual ambiguity and lexical relations. *Computational Linguistics*, 19(1).

Jerry R. Hobbs, Mark Stickel, Douglas Appelt, and Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63(1-2).

L. Humphreys, D.A.B. Lindberg, H.M. Schoolman, and G. O. Barnett. 1998. The unified medical language system: An informatics research collaboration. *Journal of the American Medical Informatics Assocation*, 5(1):1–13.

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.

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

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

Mark Liberman and Richard Sproat. 1992. The stress and structure of modified noun phrases in english. In I.l Sag and A. Szabolsci, editors, *Lexical Matters*. CSLI Lecture Notes No. 24, University of Chicago Press.

Henry J. Lowe and G. Octo Barnett. 1994. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *Journal of the American Medical Association (JAMA)*, 271(4):1103–1108.

Hwee Tou Ng and John Zelle. 1997. Corpus-based approaches to semantic interpretation in natural language processing. *AI Magazine*, 18(4).

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. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania, December. (Institute for Research in Cognitive Science report IRCS-93-42).

Philip Resnik. 1995. Disambiguating noun groupings with respect to WordNet senses. In *Third Workshop on Very Large Corpora*. Association for Computational Linguistics.

Ellen Riloff. 1996. Automatically generating extraction patterns from untagged text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference*, Menlo Park. AAAI Press / MIT Press.

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).

Don R. Swanson and N. R. Smalheiser. 1994. Assessing a gap in the biomedical literature: Magnesium deficiency and neurologic disease. *Neuroscience Research Communications*, 15:1–9.

Len Talmy. 1985. Force dynamics in language and thought. In *Parasession on Causatives and Agentivity*, University of Chicago. Chicago Linguistic Society (21st Regional Meeting).

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

Lucy Vanderwende. 1995. Ambiguity in the acquisition of lexical information. In *Working Notes of the 1995 AAAI Spring Symposium on Representation and Acquisition of Lexical Knowledge*, pages 174–179.

V. Vapnik. 1998. *Statistical Learning Theory*. Oxford University Press.

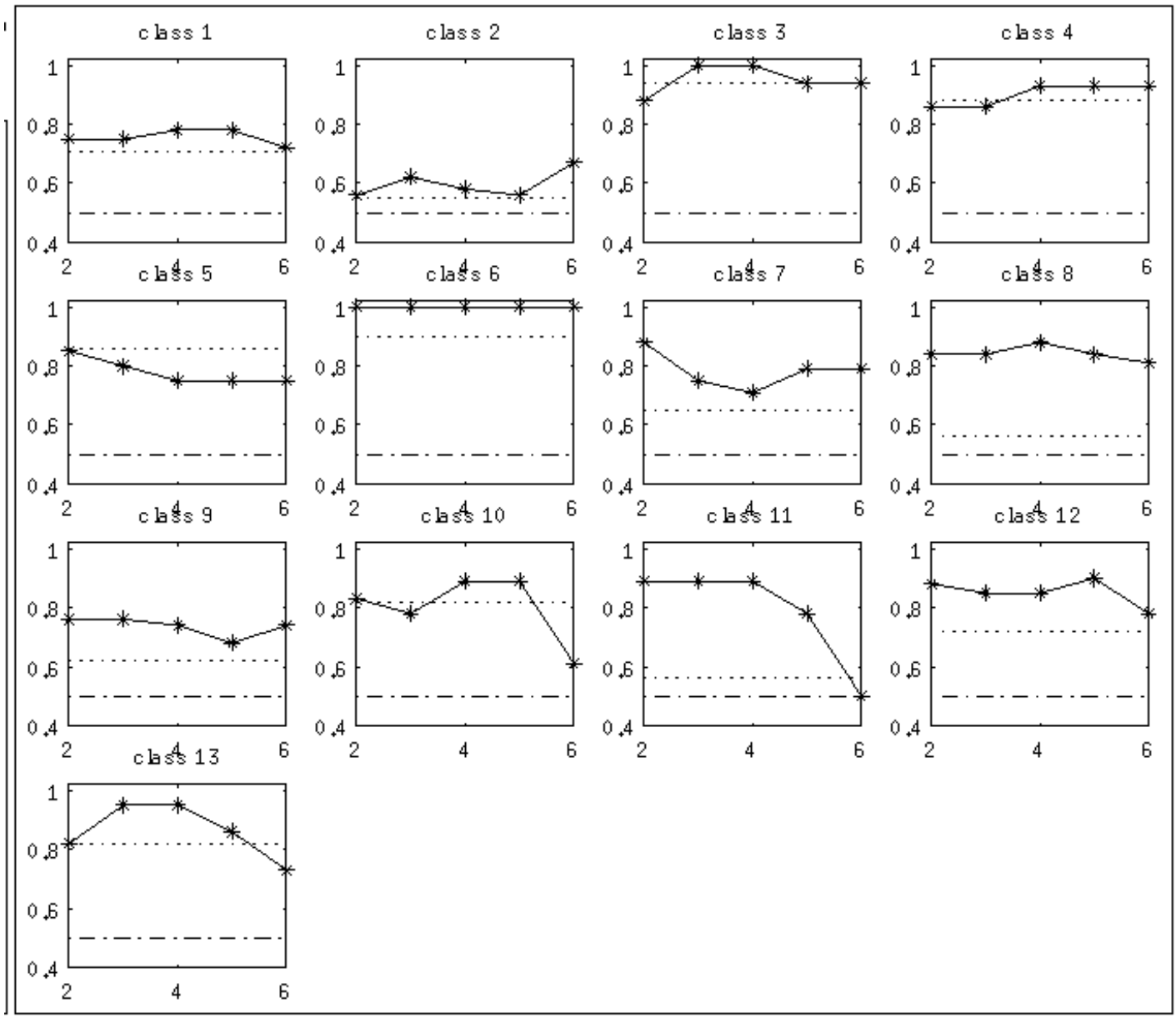

Figure 3: Accuracies for each class with CART.