Disambiguation of Prepositional Phrases in Automatically Labelled Technical Text

Lois Boggess Rajeev Agarwal Ron Davis

Department of Computer Science
Mississippi State University
Mississippi State, MS 39762
lboggess@cs.msstate.edu
rajeev@cs.msstate.edu
rdavis@cs.msstate.edu

Abstract

A system is described for semi-automatically tagging a large body of technical English with domain-specific syntactic/semantic labels. These labels have been used to disambiguate prepositional phrase attachments for a 10,000-word body of text containing more than 1,000 prepositions, and to provide case role information for about half of the phrases.

Introduction

It is our contention that, given a coherent body of text that is large enough, the text itself should be used to interpret the text. That is, the source contains considerable useful information that should be taken advantage of. This paper describes two portions of a natural language processing system embedded in a larger research effort that extracts information from a 700,000 word technical manual (the *Merck Veterinary Manual*).

The larger system of which this research is a part was initially concerned with the augmentation of an existing knowledge base in a particular domain with information taken from technical text in that domain. It has become obvious, however, that the techniques that allow a system to be augmented also lend themselves to the bootstrapping of a system from a relatively small initial state.

This paper reports our experiences in semi-automatically generating syntactic and semantic labels for a chapter of the manual that is slightly more than 10,000 words long, and then using the probabilities derived from that labelled text to label other text segments taken from elsewhere in the manual. We also describe a rule-based system that uses the syntactic and semantic labels of the 10,000 word corpus to attach prepositional phrases in that text to the appropriate sentential component.

* This work was supported in part by the National Science Foundation under grant number IRI-9002135.

Automatically generating syntactic/semantic labels

In our previous work with single source (that is, single author) texts of moderate size (20,000 to 300,000 words of English), we have noted that the body of work of a single writer differs significantly both from other writers and from norms for English text derived from large, multiple-source corpora [English and Boggess, 1986; Boggess, 1988].

Although our current text deals with a single, fairly constrained topic (veterinary medicine), it clearly was written by multiple authors; there are changes in style from chapter to chapter and even within a single chapter. Nevertheless, a single, coherent large text is likely to differ in significant ways from the body of English as a whole. Not surprisingly, it uses a specialized vocabulary, where the specialization extends not only to vocabulary which is not found in general English, but also to the fact that some general English words are used only in restricted contexts and with restricted senses. (See [Sager, 1987], for a general discussion of specialized language constructs in restricted domains.) We quickly discovered for ourselves the inadvisability of using our standard dictionary when a parse failed because the word *like* was taken as a noun (e.g., "we shall never see his like again"). Rather than hand build our own dictionary as Sager did, we chose a probabilistic method to, in effect, associate with each word that actually occurs in the text only those parts of speech with which the word is actually used in the text.

The method which we used is derived from that discussed in [Derouault and Merialdo, 1986]. They describe a technique for labelling (tagging) each word in text with its part of speech. We chose to use their technique to assign semantic information as well. Hence, in addition to having labels such as aux, conj\subord, prep, noun, and so on, we also have such labels as noun\body_fluid, noun\measure, noun\symptom, adj\quantity, adj\body_part, adj\time. Some of the labels are domain-specific, while others are general. Currently we use 79 labels, though that number may grow slightly; almost a third of the labels are

singular/plural variants of noun labels; virtually all of the semantic tags are associated with adjectives and nouns.

The process is described in [Davis, 1990]. We hand labelled an initial text of 2,000 words, and from this we built Markov bigram and trigram source models of the label sequences. (A Markov bigram model gives the probability of a two-label sequence, given the presence of the first label; a Markov trigram model gives the probability of a three-label sequence, given the presence of the first two labels.) The next 6,000 words of text were labelled automatically, using the probabilities of the Markov models to calculate which sequence of labels was most probable for a given sentence. The 6,000 newly labelled words of text were hand-corrected to give an 8,000 word body of text, which was then used to generate new bigram and trigram Markov models. Next the remainder of the chapter was labelled automatically and hand-corrected.

In tagging new text, the probability that a sentence will be labelled by a given sequence of tags is taken to be the product of the probabilities of the individual words' being given their respective tags. If we were using only a trigram model to compute the latter, they would be calculated as follows: the probability that word w_i would be labelled with tag t_i given that the preceding tags were t_{i-2} and t_{i-1} is computed as

$$p(w_i|\, t_{i-2},t_{i-1}) = p(t_i|\, t_{i-2},t_{i-1}) \times p(t_i|\, w_i).$$

This expression differs in the final term from the more traditional $p(w_i|\, t_{i-2},t_{i-1}) = p(t_i|\, t_{i-2},t_{i-1}) \times p(w_i|\, t_i)$ used by Derouault and Merialdo and others, and was suggested by Church [Church, 1988]. Over a sample of five test sets, the Church method consistently gave us lower error rates - usually an improvement of 3% (e.g. 16% instead of 19%).

Following the lead of Derouault and Merialdo, we weighted the predictions of the bigram and trigram models, with rather more emphasis given the bigram model. This is necessary since a trigram model of n labels could conceivably have $n^2 + n^1 + n^0$ states and $O(n^3)$ connections. For n = 79, the 6321 states could hardly have all been visited by 2,000 words of input.

The revised formula becomes

$$p(w_i|t_{i-2},t_{i-1}) = [\lambda_1 p(t_i|t_{i-2},t_{i-1})+\lambda_2 p(t_i|t_{i-2},t_{i-1})]\times p(t_i|w_i),$$

where λ_1 and λ_1 are calculated by an iterative process identical to that of Derouault and Merialdo.

Estimating $p(t_i|\, w_i)$

When a "known" word is encountered in the new text, we estimate $p(t_i|\, w_i)$ using the known distribution of tags for that word. That is, if word w_i has occurred with label t_i x times and word w_i has occurred y times, then $p(t_i|\, w_i)$ is approximately x/y. (We say "approximately" because we include a small probability that w_i may occur with a label that we have not yet seen.)

For unknown words we use two heuristics: the simplest is a "last resort" measure, used when there are no other cues

$$p(t_i|\, w_i) = \frac{\text{(number of occurrences of } t_i)}{\text{(occurrences of words outside the closed sets)}}.$$

More often, however, in dealing with an unknown word we use suffix cues. A list of all tags encountered in words ending in a two-letter pair, for all two-letter combinations encountered (not just traditional suffixes), is used to estimate the probability for an unknown word ending in that suffix. For example, if x occurrences of words ending in *-ie* were labelled noun\plural\disease_agent and if y occurrences of words ending in *-ie* have been encountered, then $p(t_i|\, w_i)$ for words ending in *-ie* is estimated to be approximately x/y, where again allowances are made for the fact that the suffix may occur with a label that we have not yet seen for it.

It should be mentioned that in dealing with new text, the system is prohibited from hypothesizing that a new, unknown word belongs to one of the closed classes. Hence, adding a word to the closed classes requires human intervention.

pronouns
relative pronouns
possessive pronouns
auxiliaries
determiners
coordinate conjunctions
subordinate conjunctions
correlative conjunctions
prepositions

Figure 1: Closed Classes

We have had to add to the "closed" classes rather more often than we expected. For example, a number of apparent compound prepositions were discovered in our text, some of which may be prepositions only in our opinion, but which clearly fit the use patterns of prepositions in our source. Moreover, we have added the word "following" to the set of prepositions, after our strong suspicions were confirmed by the supplement to the Oxford English Dictionary, which cites such usages beginning in the late 1940's.

as soon as
as to
as well as
because of
due to
in addition to
prior to
rather than
regardless of
such as

Figure 2: "compound" prepositions

Success rates and sources of error

Several 200-word excerpts of automatically tagged text were examined. The results are as shown below. The first sample was produced on the basis of Markov models of the initial hand-tagged 2,000 word text; the second sample was produced on the basis of Markov models of 8,000 words of hand-corrected text. Six additional samples were produced on the basis of Markov models of the full 10,000-word chapter of hand-corrected text. These latter samples came from text tens of thousands of words apart and far from the initial chapter, such that the subject matter was radically different and the authors were almost certainly different as well.

	Part of Speech	Combined Part of Speech/ Semantic Label
Sample 1	6%	14%
Sample 2	6%	9%
Sample 3	4%	8%
Sample 4	10%	17%
Sample 5	5%	13%
Sample 6	5%	16%
Sample 7	7%	16%
Sample 8	2%	7%

Figure 3: Error rates for labelled text

The error rate reported in [Derouault and Merialdo, 1986], based on probabilities built from at least 47,000 words of text that had been hand-corrected, was "less than 5 per cent". The labels applied appear to have been entirely syntactic in nature. Our own labels have a part-of-speech component, and if that portion is the only consideration, then our present error rate is not much higher than theirs - an average of 5.6% over the eight samples, based on a body of only 10,000 words of text. Church [1988] reports an error rate of 1% to 5%, depending on what one considers an error, for a program performing syntactic labelling. It seems likely that, with models based on more than 10,000 labelled words, our error rate for the syntactic labels will fall within the latter range.

However, given the potential benefits of labels that are both syntactic and semantic in nature, we are particularly interested in the errors that occur in the semantic labelling. It should be mentioned that determining what should be called an "error" is not a straightforward process. In general, we counted an error whenever a label was applied by the system that differed from the label applied by a human. There were two specific exceptions to this rule. One had to do with the two semantic classifications "disorder" and "symptom"; most terms which can be classified as "disorder" in some contexts can be classified as "symptom" in other contexts, and in a great many contexts humans end up essentially flipping a coin in attaching one or the other of the two labels. (A veterinarian and an M.D. have confirmed that the difficulty in deciding between the two semantic categories is not due to lack of expertise on the part of the labellers.) The two terms were treated as synonymous in calculating error rates. The other "break" that we extended to the system was that on those relatively few occasions when the system labelled a term with a semantically more general label than the human, we treated the label as correct. The reverse, however, was not true. In the context of the larger system within which this research was taking place, giving a word a correct but overly general label (<noun> rather than <noun/treatment>, for example), leaves the system with rather more work to do than the more specific label would, but seldom would it cause the system to make a mistake. On the other hand, a too-specific label (<noun/diagnostic-aid> when a human would have labelled the word <noun>) might well lead the overall system astray.

As it happens, the current version of the labelling system is far more likely to commit the latter error. For the samples reported above, more than one-third of all errors reported above (and hence two-thirds of the semantic labelling errors) have been due to over-specification of semantic type. Since the samples based on the full 10,000-word probabilistic models were all taken from a different chapter than the basis of the models, a fairly large proportion of these errors were due to the fact that the correct, specific semantic label had not been created. For example, the label <noun/bodyfunction> had not yet been created for the system, because it had not been observed as a useful label in the first chapter examined. Yet numerous words that should have received such a label occurred in the samples taken from outside that chapter. The human who was labelling those words at the time simply chose to label them <noun>. But the labelling system in almost every case gave them more specific, hence erroneous, semantic labels.

We expect to improve the error rates, then, by the following means: Syntactic errors should decrease as a result of larger bases for the Markov models. We are experimenting with minor modifications of our heuristics for estimating probabilities for unknown words. Adding semantic categories that are clearly missing will lower the semantic labelling errors to a significant degree, and we also expect to address directly the question of how to determine that a general label is preferable to a specific one. We also are in the process of examining the degree of improvement given by hand-labelling a small excerpt from a new chapter, to be added to the larger Markov models, prior to automatically labelling the rest of the new chapter. All in all, we anticipate improving the error rates substantially.

Using the labels to disambiguate prepositional phrases

As described in [Agarwal,1990], the first task assigned to the syntactic/ semantic labels was that of disambiguating prepositional phrases by attaching them to the appropriate sentential components and assigning case roles to the

```
sentence (word (if, conj\subord),
        noun_phrase (       the,det
                            signs,noun\plural)
        verb_phrase (       correlate, verb)
        prep_phrase (       with,prep
                            noun_phrase (       the,det
                                                extra,adj\quantity
                                                cilia, noun\plural\body_part))
        word (comma,punc)
        noun_phrase (       excision\noun\treatment)
        prep_phrase (       of,prep
                            noun_phrase (       cilia,noun\plural\body_part))
        verb_phrase (       is,aux
                            indicated,verb\past_p))
```

Figure 4: Sample output from semi-parser.

resulting structures. The "cases" used in our system are an extension of the more standard cases of case grammar, since in our source the standard case roles account for very few of the roles taken by prepositional phrases. Hence, we added to such traditional roles as *location*, *time*, and *instrument* such domain-specific roles as *population*, *disorder*, *treatment*, and the somewhat more general *part_whole* case.

For example, the occurrence of <verb> *in* <noun\body_part> receives a role designation of location, while <verb> *in* <noun\patient> is designated population, <noun\treatment> *of* <noun\patient> is considered treatment, and <adj\body_part>,<noun> *of* <noun\body_part> is designated part_whole. (The last pattern illustrates the requirement that a body-part adjective precede a general noun prior to the preposition *of*.

Prior to disambiguation of the prepositional phrases, the labelled text is passed to what we call a semi-parser. This simple parser has as its only task the identification of the most fundamental phrases - noun phrases, prepositional phrases, gerund phrases and compound verbs with adjacent adverbs. These simple phrases are not embedded within each other, excepting that a prepositional phrase is defined as a preposition followed by a single other kind of phrase, and a gerund phrase may include a noun phrase object. The output of the semi-parser is a very flat-looking sentential structure. (See Figure 4, above.)

The most common format for determining the proper attachment for a prepositional phrase is as follows: The system looks at the content words preceding the preposition (the pre-frame) and the phrase that is the object of the preposition (the post-frame). In examining the pre-frame, the attachment program looks first at the content word nearest the preposition, and if necessary works its way farther from the preposition, in the direction of the beginning of the sentence. In examining the post-frame, it begins with the headword of the phrase, often the farthest word from the preposition. If the labels for these content words match a rule for the preposition, such as *<noun\patient> with <noun/disorder>*, or *<noun/patient> with <noun/medication>*, then an attachment is made and, when possible, a case is assigned. If the labels for the content words do not match a rule for the preposition, then the text preceding the preposition is scanned further backwards to find the next content word and another match is attempted, and so on.

The foregoing is the normal procedure, but a number of the prepositions have special rules specific to the word that immediately precedes the preposition. For example, in our text about 21% of the occurrences of the preposition *of* are immediately preceded by words ending in *-tion* and *-sion*, where the usages of these words are verbal in nature. A special rule for *of* specifies that if the immediately preceding word is a-*tion* or *-sion* word that does not belong to a small set of common *-tion* words such as "junction," then the preposition is to be removed and the object of the preposition is to be marked as the object of the verbal form that preceded *of*. Most of the other prepositions also have from one to three rules specific only to the immediately preceding word or label. There are also provisions for the cases where the prepositional phrase precedes the sentential component to which it should be attached, as, for instance, when the prepositional phrase occurs in sentence-initial position.

A Prolog program using a surprisingly small set of rules (an average of 15 per preposition, for the nine prepositions that occur more than 10 times in the 10,000 word chapter of labelled text) has enabled the correct placement of 944 of the 1029 phrases headed by those nine prepositions in the chapter. The same rules assigned appropriate case roles to 46% of the prepositional phrases; the case roles of virtually all of the remaining 54% were designated "unknown" by the system.

The performance of the prepositional phrase attacher is summarized in the table that follows.

preposition	number of rules	attachments		case assignments	
of	24	335/341	98.24%	155/341	45.45%
in	24	269/291	92.44%	159/291	54.64%
with	16	80/84	95.24%	29/84	34.52%
by	13	64/71	90.14%	37/71	52.11%
for	7	55/68	80.88%	23/68	33.82%
to	11	55/68	80.88%	13/68	19.11%
from	15	45/54	83.33%	37/54	68.52%
on	12	21/29	72.41%	12/29	41.38%
at	14	20/23	86.96%	9/23	39.13%
Totals	136	944/1029	91.74%	474/1029	46.06%

Figure 5: Success rates in attaching prepositional phrases

Most of our failures to attach a prepositional phrase to the correct component of the sentence are associated with a mishandled conjunction ("and" is the second most frequent word in the chapter analyzed). If the sentences currently causing prepositional attachment errors are any indication, there is cause to believe that the labels will be extremely helpful in correctly handling even complicated conjoined constructs. In almost every case, the error involves an "and" followed by a prepositional phrase that should be conjoined with an earlier but distant prepositional phrase, with multiple intervening noun phrases and even intervening prepositional phrases. Frequently, under those circumstances, the two phrases that ought to be conjoined have noun objects belonging to the identical semantic category, and that category is generally different from the semantic categories of the intervening noun phrases. Many of these distant but coupled prepositional phrases repeat the same preposition as well.

One of our next projects will be to investigate how much the labels can accomplish for us in the complex task of disambiguating conjoined phrases and clauses. We expect this coordination specialist to be independent from the preposition handler. As a matter of fact, one of the reasons we favor the flat nature of our semi-parser is that it leaves all the elements of the sentence relatively accessible to any of the specialists that we design. The specialists themselves do not restructure the sentence so much as leave notes on where phrases should be attached in the final analysis, and a mopping-up segment of the natural language processor actually produces the final structure that is passed to the knowledge analyzer of the larger system.

References

Agarwal, R., 1990. Disambiguation of Prepositional Phrase Attachments in English Sentences using Case Grammar Analysis. MS thesis, Dept. of Computer Science, Mississippi State University.

Boggess, L., 1988. Two Simple Prediction Algorithms to Facilitate Text Production. In Proceedings of the Second Conference on Applied Natural Language Processing: 33-40. Austin, Texas: Association for Computational Linguistics.

Church, K., 1988. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In Proceedings of the Second Conference on Applied Natural Language Processing: 136-143. Austin, Texas: Association for Colmputational Linguistics.

Davis, R., 1990. Automatic Text Labelling System. MCS project report, Dept. of Computer Science, Mississippi State University.

Derouault, A. and Merialdo, B. 1986. Natural Language Modeling for Phoneme-to-text Transcription. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-8(6) 742-749.

English, T. and Boggess, L. 1986. A Grammatical Approach to Reducing the Statistical Sparsity of Language Models in Natural Domains. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing: 1141-1144. Tokyo, Japan:

Sager, N., Friedman, C. and Lyman, M., 1987. *Medical Language Processing*. Addison-Wesley.