Structural Ambiguity and Lexical Relations

Donald Hindle and Mats Rooth

AT&T Bell Labs
600 Mountain Ave.
Murray Hill, NJ 07974

Introduction

From a certain (admittedly narrow) perspective, one of the annoying features of natural language is the ubiquitous syntactic ambiguity. For a computational model intended to assign syntactic descriptions to natural language text, this seem like a design defect. In general, when context and lexical content are taken into account, such syntactic ambiguity can be resolved: sentences used in context show, for the most part, little ambiguity. But the grammar provides many alternative analyses, and gives little guidance about resolving the ambiguity.

Prepositional phrase attachment is the canonical case of structural ambiguity, as in the time worn example,

(1) I saw the man with the telescope

The problem arises because the grammar provides several sources for prepositional phrases. The prepositional phrase *with the telescope* has two central attachment possibilities (the seeing is by means of a telescope or the man has a telescope), licensed by two different phrase structure rules, namely

$VP \rightarrow NP\ PP$

and

$NP \rightarrow N'\ PP$

(The prepositional phrase might also attach to the subject noun phrase *I*; in this paper we will concentrate on the most important binary choice between attachment to the adjacent Noun Phrase, and attachment to the preceding Verb.)

The existence of such ambiguity raises problems for understanding and for language models. It looks like it might require extremely complex computation to determine what attaches to what. Indeed, one recent proposal suggests that resolving attachment ambiguity requires the construction of a discourse model in which the entities referred to in a text must be reasoned about (Altmann and Steedman 1988).

Of course, if attachment ambiguity demands reference to semantics and discourse models, there is little hope in the near term of building computational models for unrestricted text to resolve the ambiguity.

Structure based ambiguity resolution

There have been several structure-based proposals about ambiguity resolution in the literature; they are particularly attractive because they are simple and don't demand calculations in the semantic or discourse domains. The two main ones are:

- Right Association - a constituent tends to attach to another constituent immediately to its right (Kimball 1973).

- Minimal Attachment - a constituent tends to attach so as to involve the fewest additional syntactic nodes (Frazier 1978).

For the particular case we are concerned with, attachment of a prepositional phrase in a verb + object context as in sentence (1), these two principles – at least in the version of syntax that Frazier assumes – make opposite predictions: Right Association predicts noun attachment, while Minimal Attachment predicts verb attachment.

Unfortunately, these structure-based disambiguation proposals seem not to account for attachment preferences very well. A recent study of attachment of prepositional phrases in a sample of written responses to a "Wizard of Oz" travel information experiment shows that niether Right Association nor Minimal Attachment account for more than 55% of the cases (Whittemore et al. 1990). And experiments by Taraban and McClelland (1988) show that the structural models are not in fact good predictors of people's behavior in resolving ambiguity.

Resolving ambiguity through lexical associations

Whittemore et al. (1990) found lexical preferences to be the key to resolving attachment ambiguity. Similarly, Taraban and McClelland found lexical content was key in explaining people's behavior. Various previous proposals for guiding attachment disambiguation by the lexical content of specific words have appeared (e.g. Ford, Bresnan, and Kaplan 1982; Marcus 1980). Unfortunately, it is not clear where the necessary information about lexical preferences is to be found. In the Whittemore et al.

study, the judgement of attachment preferences had to be made by hand for exactly the cases that their study covered; no precompiled list of lexical preferences was available. Thus, we are posed with the problem: how can we get a good list of lexical preferences.

Our proposal is to use cooccurrence of with prepositions in text as an indicator of lexical preference. Thus, for example, the preposition *to* occurs frequently in the context *send NP* —, i.e., after the object of the verb *send*, and this is evidence of a lexical association of the verb *send* with *to*. Similarly, *from* occurs frequently in the context *withdrawal* —, and this is evidence of a lexical association of the noun *withdrawal* with the preposition *from*. Of course, this kind of association is, unlike lexical preference, a symmetric notion. Cooccurrence provides no indication of whether the verb is selecting the preposition or vice versa. We will treat the association as a property of the pair of words. It is a separate matter, which we unfortunately cannot pursue here, to assign the association to a particular linguistic licensing relation. The suggestion which we want to explore is that the association revealed by textual distribution – whether its source is a complementation relation, a modification relation, or something else – gives us information needed to resolve the prepositional attachment.

Discovering Lexical Association in Text

A 13 million word sample of Associated Press new stories from 1989 were automatically parsed by the Fidditch parser (Hindle 1983), using Church's part of speech analyzer as a preprocessor (Church 1988). From the syntactic analysis provided by the parser for each sentence, we extracted a table containing all the heads of all noun phrases. For each noun phrase head, we recorded the following preposition if any occurred (ignoring whether or not the parser attached the preposition to the noun phrase), and the preceding verb if the noun phrase was the object of that verb. Thus, we generated a table with entries including those shown in Table 1.

VERB	HEAD NOUN	PREP
blame	PASSIVE	for
	money	for
	development	
control	government	
	military	
	accord	
enrage	radical	
	WHPRO	
	it	
grant	concession	to

Table 1: A sample of the Verb-Noun-Preposition table.

In this Table 1, the first line represents a passivized

instance of the verb *blame* followed by the preposition *for*. The second line is an instance of a noun phrase whose head is *money*; this noun phrase is not an object of any verb, but is followed by the preposition *for*. The third line represents an instance of a noun phrase with head noun *development* which neither has a following preposition nor is the object of a verb. The fourth line is an instance of a noun phrase with head *government*, which is the object of the verb *control* but is followed by no preposition. The last line represents an instance of the ambiguity we are concerned with resolving: a noun phrase (head is *concession*), which is the object of a verb (*grant*), followed by a preposition (*to*).

From the 13 million word sample, 2,661,872 noun phrases were identified. Of these, 467,920 were recognized as the object of a verb, and 753,843 were followed by a preposition. Of the noun phrase objects identified, 223,666 were ambiguous verb-noun-preposition triples.

Estimating attachment preferences

Of course, the table of verbs, nouns and prepositions does not directly tell us what the lexical associations are. This is because when a preposition follows a noun phrase, it may or may not be structurally related to that noun phrase (in our terms, it may attach to that noun phrase or it may attach somewhere else). What we want to do is use the verb-noun-preposition table to derive a table of bigrams, where the first term is a noun or verb, and the second term is an associated preposition (or no preposition). To do this we need to try to assign each preposition that occurs either to the noun or to the verb that it occurs with. In some cases it is fairly certain that the preposition attaches to the noun or the verb; in other cases, it is far less certain. Our approach is to assign the clear cases first, then to use these to decide the unclear cases that can be decided, and finally to arbitrarily assign the remaining cases. The procedure for assigning prepositions in our sample to noun or verb is as follows:

1. No Preposition - if there is no preposition, the noun or verb is simply counted with the null preposition.

2. Sure Verb Attach 1 - preposition is attached to the verb if the noun phrase head is a pronoun.

3. Sure Verb Attach 2 - preposition is attached to the verb if the verb is passivized (unless the preposition is *by*. The instances of *by* following a passive verb were left unassigned.)

4. Sure Noun Attach - preposition is attached to the noun, if the noun phrase occurs in a context where no verb could license the prepositional phrase (i.e., the noun phrase is in subject or pre-verbal position.)

5. Ambiguous Attach 1 - Using the table of attachment so far, if a t-score for the ambiguity (see below) is

greater than 2.1 or less than -2.1, then assign the preposition according to the t-score. Iterate through the ambiguous triples until all such attachments are done.

6. Ambiguous Attach 2 - for the remaining ambiguous triples, split the attachment between the noun and the verb, assigning .5 to the noun and .5 to the verb.

7. Unsure Attach - for the remaining pairs (all of which are either attached to the preceding noun or to some unknown element), assign them to the noun.

This procedure gives us a table of bigrams representing our guess about what prepositions associate with what nouns or verbs, made on the basis of the distribution of verbs nouns and prepositions in our corpus.

The procedure for guessing attachment

Given the table of bigrams, derived as described above, we can define a simple procedure for determining the attachment for an instance of verb-noun-preposition ambiguity. Consider the example of sentence (2), where we have to choose the attachment given verb *send*, noun *soldier*, and preposition *into*.

> (2) Moscow sent more than 100,000 soldiers into Afganistan ...

The idea is to contrast the probability with which *into* occurs with the noun *soldier* with the probability with which *into* occurs with the verb *send*. A t-score is an appropriate way to make this contrast (see Church et al. to appear). In general, we want to calculate the contrast between the conditional probability of seeing a particular preposition given a noun with the conditional probability of seeing that preposition given a verb.

$$t \equiv \frac{P(prep \mid noun) - P(prep \mid verb)}{\sqrt{\sigma^2(P(prep \mid noun)) + \sigma^2(P(prep \mid verb))}}$$

We use the "Expected Likelihood Estimate" (Church et al., to appear) to estimate the probabilities, in order to adjust for small frequencies; that is, we simply add 1/2 to all frequency counts (and adjust the denominator appropriately). This method leaves the order of t-scores nearly intact, though their magnitude is inflated by about 30%. To compensate for this, the 1.65 threshold for significance at the 95% level should be adjusted up to about 2.15.

Consider how we determine attachment for sentence (4). We use a t-score derived from the adjusted frequencies in our corpus to decide whether the prepositional phrase *into Afganistan* is attached to the verb (root) *send/V* or to the noun (root) *soldier/N*. In our corpus, *soldier/N* has an adjusted frequency of 1488.5, and *send/V* has an adjusted frequency of 1706.5; *soldier/N* occurred in 32 distinct preposition contexts, and *send/V* in 60 distinct preposition contexts; f(*send/V into*) = 84, f(*soldier/N into*) = 1.5.

From this we calculate the t-score as follows:[1]

$$t \equiv \frac{P(w|soldier/N) - P(w|send/V)}{\sqrt{\sigma^2(P(w|soldier/N)) + \sigma^2(P(w|send/V))}}$$

$$\approx \frac{\frac{f(soldier/N\ into)+1/2}{f(soldier/N)+V/2} - \frac{f(send/V\ into)+1/2}{f(send/V)+V/2}}{\sqrt{\frac{f(soldier/N\ into)+1/2}{(f(soldier/N)+V/2)^2} + \frac{f(send/V\ into)+1/2}{(f(send/V)+V/2)^2}}}$$

$$\approx \frac{\frac{1.5+1/2}{1488.5+70/2} - \frac{84+1/2}{1706.5+70/2}}{\sqrt{\frac{1.5+1/2}{1488.5+70/2)^2} + \frac{84+1/2}{1706.5+V/2)^2}}} \approx -8.81$$

This figure of -8.81 represents a significant association of the preposition *into* with the verb *send*, and on this basis, the procedure would (correctly) decide that *into* should attach to *send* rather than to *soldier*.

Testing Attachment Preference

We have outlined a simple procedure for determining prepositional phrase attachment in a verb-object context. To evaluate the performance of this procedure, we need a graded set of attachment ambiguities. First, the two authors graded a set of verb-noun-preposition triples as follows. From the AP new stories, we randomly selected 1000 test sentences in which the parser identified an ambiguous verb-noun-preposition triple. (These sentences were selected from stories included in the 13 million word sample, but the particular sentences were excluded from the calculation of lexical associations.) For every such triple , each author made a judgement of the correct attachment on the basis of the three words alone (forced choice - preposition attaches to noun or verb). This task is in essence the one that we will give the computer - i.e., to judge the attachment without any more information than the preposition and the head of the two possible attachment sites, the noun and the verb. This gave us two sets of judgements to compare the algorithms performance to.

Judging correct attachment

We also wanted a standard of correctness for these test sentences. To derive this standard, each author independently judged the attachment for the 1000 triples a second time, this time using the full sentence context.

It turned out to be a surprisingly difficult task to assign attachment preferences for the test sample. Of course, many decisions were straightforward, but more than 10% of the sentences seemed problematic to at least one author. There are two main sources of such difficulty. First, it is unclear where the preposition is attached in idiomatic phrases such as :

259

(3) But over time , misery has *given way to* mending .

(4) The meeting will *take place in* Quantico

A second major source of difficulty arose from cases where the attachment either seemed to make no difference semantically or it was impossible to decide which attachment was correct, as

(5) We don't *have* preventive *detention in* the United States.

(6) Inaugural officials reportedly were trying to *arrange* a *reunion for* Bush and his old submarine buddies ...

It seems to us that this difficulty in assigning attachment decisions is an important fact that deserves further exploration. If it is difficult to decide what licenses a prepositional phrase a significant proportion of the time, then we need to develop language models that appropriately capture this vagueness. For our present purpose, we decided to force an attachment choice in all cases, in some cases making this choice arbitrarily.

In addition to the problematic cases, a significant number (111) of the 1000 triples identified automatically as instances of the verb-object-preposition configuration turned out in fact to be other constructions. These misidentifications were mostly due to parsing errors, and in part due to our underspecifying for the parser exactly what configuration to identify. Examples of these misidentifications include: identifying the subject of the complement clause of *say* as its object, as in (7), which was identified as (*say ministers from*); misparsing two constituents as a single object noun phrase, as in (8), which was identified as (*make subject to*); and counting non-object noun phrases as the object as in (9), identified as (*get hell out_of*).

(7) Ortega also *said* deputy foreign *ministers from* the five governments would meet Tuesday in Managua, ...

(8) Congress made a deliberate choice to *make this commission subject to* the open meeting requirements ...

(9) Student Union, *get the hell out of* China!

Of course these errors are folded into the calculation of associations. No doubt our bigram model would be better if we could eliminate these items, but many of them represent parsing errors that obviously cannot be identified by the parser, so we proceed with these errors included in the bigrams.

After agreeing on the "correct" attachment for the sample of 1000 triples, we are left with 889 verb-noun-preposition triples (having discarded the 111 parsing errors). Of these, 568 are noun attachments and 321 verb attachments.

Evaluating performance

First, consider how the simple structural attachment preference schemas do at predicting the outcome in our test set. Right Association, which predicts noun attachment does better, since there are more noun attachments, but it still has an error rate of 36%. Minimal Attachment, interpreted to mean verb attachment has the complementary error rate of 64%. Obviously, neither of these procedures is particularly impressive. For our sample, the simple strategy of attaching a prepositional phrase to the nearest constituent is the more successful strategy.

Now consider the performance of our attachment procedure for the 889 standard test sentences. Table 2 shows the results on the test sentences for the two human judges and for the attachment procedure.

	choice		% correct		
	N	V	N	V	total
Judge 1	583	306	90.0	78.5	86.2
Judge 2	518	371	84.0	87.2	85.2
LA	565	354	82.6	70.0	78.1

Table 2: Performance on the test sentences for 2 human judges and the lexical association procedure (LA).

First, we note that the task of judging attachment on the basis of verb, noun and preposition alone is not easy. Both human judges had overall error rates of nearly 15%. (Of course this is considerably better than always choosing the nearest attachment site.) The lexical association procedure based on t-scores is somewhat worse than the human judges, with an error rate of 22%, again an improvement over simply choosing the nearest attachment site.

260

If we restrict the lexical association procedure to choose attachment only in cases where its confidence is greater than about 95% (i.e., where t is greater than 2.1), we get attachment judgements on 608 of the 889 test sentences, with an overall error rate of 15% (Table 3). On these same sentences, one human judge also showed slight improvement.

	choice		% correct		
	N	V	N	V	total
Judge 1	435	173	91.0	82.7	88.7
Judge 2	396	212	94.4	67.5	85.0
LA	407	201	87.1	80.2	85.0

Table 3: Performance on the test sentences for 2 human judges and the lexical association procedure (LA) for test triples where $t > 2.1$

Comparison with a Dictionary

The idea that lexical preference is a key factor in resolving structural ambiguity leads us naturally to ask whether existing dictionaries can provide useful information for disambiguation. To investigate this question, we turn to the Collins Cobuild English Language Dictionary (Sinclair et al. 1987). This dictionary is appropriate for comparing with the AP sample for several reasons: it was compiled on the basis of a large text corpus, and thus may be less subject to idiosyncrasy than more arbitrarily constructed works; and it provides, in a separate field, a direct indication of prepositions typically associated with many nouns and verbs.

From a machine-readable version of the dictionary, we extracted a list of 1535 nouns associated with a particular preposition, and of 1193 verbs associated with a particular preposition after an object noun phrase. These 2728 associations are many fewer than the number of associations found in the AP sample. (see Table 4.)

Of course, most of the preposition association pairs from the AP sample end up being non-significant; of the 88,860 pairs, fewer than half (40,869) occur with a frequency greater than 1, and only 8337 have a t-score greater than 1.65. So our sample gives about three times as many significant preposition associations as the COBUILD dictionary. Note however, as Table 4 shows, the overlap is remarkably good, considering the large space of possible bigrams. (In our bigram table there are over 20,000 nouns, over 5000 verbs, and over 90 prepositions.) On the other hand, the lack of overlap for so many cases – assuming that the dictionary and the significant bigrams actually record important preposition associations – indicates that 1) our sample is too small, and 2) the dictionary coverage is widely scattered.

First, we note that the dictionary chooses attachments in 182 cases of the 889 test sentences. Seven of these are cases where the dictionary finds an association between the preposition and both the noun and the verb. In these cases, of course, the dictionary provides no information to help in choosing the correct attachment.

Looking at the 175 cases where the dictionary finds one and only one association for the preposition, we can ask how well it does in predicting the correct attachment. Here the results are no better than our human judges or than our bigram procedure. Of the 175 cases, in 25 cases the dictionary finds a verb association when the correct association is with the noun. In 3 cases, the dictionary finds a noun association when the correct association is with the verb. Thus, overall, the dictionary is 86% correct.

It may be unfair to use a dictionary as a source of disambiguation information. There is no reason to expect that the dictionary aims to provide information on all significant associations; it may record only associations that are interesting for some reason (perhaps because they are semantically unpredictable.) But from the standpoint of a language model, the fact that the dictionary provides no help in disambiguation for about 80% of the ambiguous triples considerably diminishes its usefulness.

Conclusion

Our attempt to use lexical associations derived from distribution of lexical items in text shows promising results. Despite the errors in parsing introduced by automatically analyzing text, we are able to extract a good list of associations with preposition, overlapping significantly with an existing dictionary. This information could easily be incorporated into an automatic parser, and additional sorts of lexical associations could similarly be derived from text. The particular approach to deciding attachment by t-score gives results nearly as good as human judges given the same information. Thus, we conclude that it may not be necessary to resort to a complete semantics or to discourse models to resolve many pernicious cases of attachment ambiguity.

It is clear however, that the simple model of attachment preference that we have proposed, based only on the verb, noun and preposition, is too weak to make correct attachments in many cases. We need to explore ways to enter more complex calculations into the procedure. In particular, it will be necessary to include information about the object of the preposition, which will allow us to determine for example whether the preposition *in* is functioning as a temporal or locative modifier in (10). And information about the premodifiers of the object noun phrase will help decide disambiguation in cases like (11), where the *as* phrase depends in the prenominal modifier *such*.

(10) Jefferson Smurfit Inc. of Alton , Ill. , bought the company *in* 1983 ...

(11) The guidelines would affect *such* routine tasks *as* using ladders to enter manholes ...

References

[1] Altmann, Gerry, and Mark Steedman. 1988. Interaction with context during human sentence processing. *Cognition*, 30, 191-238.

[2] Church, Kenneth W. 1988. A stochastic parts program and noun phrase parser for unrestricted text, *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, Texas.

[3] Church, Kenneth W., William A. Gale, Patrick Hanks, and Donald Hindle. (to appear). Using statistics in lexical analysis. in Zernik (ed.) *Lexical acquisition: using on-line resources to build a lexicon.*

[4] Ford, Marilyn, Joan Bresnan and Ronald M. Kaplan. 1982. A competence based theory of syntactic closure, in Bresnan, J. (ed.) *The Mental Representation of Grammatical Relations.* MIT Press.

[5] Frazier, L. 1978. *On comprehending sentences: Syntactic parsing strategies.* PhD. dissertation, University of Connecticut.

[6] Hindle, Donald. 1983. User manual for fidditch, a deterministic parser. Naval Research Laboratory Technical Memorandum 7590-142.

[7] Kimball, J. 1973. Seven principles of surface structure parsing in natural language, *Cognition*, 2, 15-47.

[8] Marcus, Mitchell P. 1980. *A theory of syntactic recognition for natural language.* MIT Press.

[9] Sinclair, J., P. Hanks, G. Fox, R. Moon, P. Stock, et al. 1987. *Collins Cobuild English Language Dictionary.* Collins, London and Glasgow.

[10] Taraban, Roman and James L. McClelland. 1988. Constituent attachment and thematic role assignment in sentence processing: influences of content-based expectations, *Journal of Memory and Language*, 27, 597-632.

[11] Whittemore, Greg, Kathleen Ferrara and Hans Brunner. 1990. Empirical study of predictive powers of simple attachment schemes for post-modifier prepositional phrases. *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, 23-30.

Source	Total	NOUN	VERB
COBUILD	2728	1535	1193
AP sample	88,860	64,629	24,231
AP sample ($f > 1$)	40,869	31,241	9,628
AP sample ($t > 1.65$)	8,337	6,307	2,030
COBUILD ∩ AP	1,931	1,147	784
COBUILD ∩ AP($t > 1.65$)	1,040	656	384

Table 4: Count of noun and verb associations for COBUILD and the AP sample