Determining PP Attachment through Semantic Associations and Preferences

Michael Niemann
Department of Linguistics & Applied Linguistics
The University of Melbourne

niemann@cs.mu.oz.au

Abstract[1]

The problem of determining the correct attachment sites for PPs when parsing natural language is investigated. Semantic information is introduced into the parsing process by grouping lexical items according to the semantic associations given in WordNet V5.1. Data extracted from a partially parsed corpus is tagged with the semantic groupings. The resultant attachment preference patterns displayed in the data are used to assist the parsing.

1. Introduction

A common problem in computationally parsing natural language has been deciding the most suitable syntactic structure for each sentence. The grammar used by a parser may create more than one possible structure for a sentence. For instance, the phrase structure rules in (1) will parse (2a,b) correctly. However, they can also give incorrect structures (3a,b).

(1) a. S → NP VP
b. NP → NP PP
c. VP → VP NP
d. VP → VP NP PP

(2) a. I [saw [$_{NP}$ [$_{NP}$ the girl] [$_{PP}$ with a basketball]]].
b. I [bought [$_{NP}$ a book] [$_{PP}$ on Sunday]].

(3) a. I [saw [$_{NP}$ the girl] [$_{PP}$ with a basketball]].
b. I [bought [$_{NP}$ [$_{NP}$ a book] [$_{PP}$ on Sunday]]].

This problem of determining the correct attachment of PPs is yet to be completely resolved. When a natural speaker determines the structure for the sentence in (2a), the structure in (3a) is dismissed as the speaker knows it is impossible to see *with a basketball*. Such common-sense is not available to any computer program. However, a parser will need a similar sort of guidance in its decision making if it is to parse a single sentence structure.

2. Approaches to the Problem

One preposition or PP can have very many different meanings. For instance, the preposition *with* can indicate that its object is a companion (4a), adjacent to (4b), in control (4c), in support (4d) or a part (4e) (Smith 1991:330).

(4) a. Joe [with his friend]
b. flowers [with ferns]
c. She left the letter [with me].
d. Are you [with me] ?
e. the hand [with the broken finger]

The sense of a preposition could be said to be the type of relationship it has with its object.

Such relationships relate to the semantic role that the PP plays in the sentence. As indicated above, the structure in (3a) is implausible because *with a basketball* cannot take an instrument role for the verb *saw*. Yet it can take the role of a possession PP for *the girl*. This is the type of semantic role that is expected by a natural speaker for such a PP. Taraban & McClelland (1988) found that natural speakers regularly anticipate such roles.

Therefore, a parser should try and determine which PP attachment is most plausible and will have the expected semantic role. The semantic role will determine the appropriate sense of the preposition.

Hindle & Rooth (1993) use a system of lexical preferences to gauge the plausibility of the attachments. For any given preposition, and possible attachment to a verb or a noun (as for *saw the girl with a basketball*), a lexical association score is determined, using the structures in a hand-tagged corpus as reference. If the preposition (e.g. *with*) is more often attached in the corpus to the noun (*girl*) than the verb (*saw*), then verb attachment is more likely.

Collins & Brooks (1995) take a similar approach, but they also include the object of the PP in their scoring.

[1] This paper is a summary of the research undertaken for my BA Honours thesis (Niemann 1997) in Linguistics.

Michael Niemann (1998) **Determining PP Attachment through Semantic Associations and Preferences.** In D. Estival (ed.) *Abstracts for the ANLP Post Graduate Workshop,* pp 25-32.

This is a vital component of any plausibility evaluation. As shown by examples in (5a,b), a different object can easily affect the sense of the preposition and the PP's semantic role. Collins & Brooks also use a different formula to score the possible attachments.

(5) a. 1 [bought [a book] [on Sunday]].
b. 1 [bought [a book [on linguistics]]].

While lexical preferences certainly are one way to try and determine which attachments are most plausible, they can be computationally expensive. Both Hindle & Rooth and Collins & Brooks are restricted by relying on relationships between lexical items. The data extracted from the training corpora is finite but it cannot easily handle relationships not present in the corpus. A parser needs to be able handle any possible attachment ambiguity. Merely expanding the corpus to include more words in the training data is not an efficient solution.

Table 1: WordNet groups for nouns

action	animal	artifact
attribute	body	cognition
communication	event	feeling
food	group	location
motive	object	person
phenomenon	plant	possession
process	quantity	relation
shape	state	substance
time		

Table 2: WordNet groups for verbs

body	change	cognition
communication	competition	consumption
contact	creation	emotion
motion	perception	possession
social	stative	weather

3. Semantic Association and Preference

If all of the lexical items in the data can be categorized into a finite number of groups then this problem can be eliminated. The WordNet lexical database (Miller, Beckwith, Fellbaum, Gross & Miller, 1993; Miller & Fellbaum, 1991) contains data associating words with their antonyms, synonyms and hypernyms. The nouns and verbs are grouped within a finite number of sets according to semantic concepts (see Tables 1 and 2). The nouns are grouped in 25 hypernym trees, such that for each member of a group, there is another member that is its hypernym. At the top of the trees are "unique beginners" that are hypernyms of all other members of the tree. For instance, *domestic_dog* is the hypernym of *poodle* because a poodle is a domestic dog. They are both members of the *animal* group because the *animal* hypernym tree contains a branch that goes *poodle* → *domestic_dog* → *dog* → ... → *animal*. Verbs are grouped according to 15 "troponym" trees (Fellbaum 1993).[2] For example, *limp* is a troponym of *walk* as *to limp* is *to walk in a certain manner* such that *limping* entails *walking*.

Therefore, as WordNet V5.1 has over 132,000 senses for nouns and verbs, most objects and attachment sites can be categorized into a particular WordNet group. This allows a finite set of attachment preferences corpus that doesn't rely on lexical items to be calculated from a training.

For this research, sentences from the *DSO Corpus of Sense-tagged English Nouns and Verbs* were tagged by Brill's part-of-speech tagger (Brill, 1994) then partially parsed by Abney's CASS partial parser (Abney, 1991;1997). For most prepositions, CASS uses a very general rule which is sometimes incorrect. If a PP is within a VP according to the grammar's rules, the PP is attached to the VP. No distinction is made between adjuncts and complements. For some sentences, CASS can not provide a complete structure as it does not have rules that place every phrase in the sentence into a single structure. For this reason, some phrases are left unattached. This commonly occurs with PPs.

The aim of this research is to demonstrate a way in which a parser may be assisted in resolving PP attachment ambiguities. Therefore, 1097 of the PPs left unattached by CASS were extracted from the corpus. For example (6), if CASS left the PP *in the future* unattached, then the PP could be extracted and the noun *establishment* would be hand-tagged as the correct attachment site. No effort was made to correct the inaccuracies in CASS' grammar.

(6) Without dissent, senators passed a bill authorizing establishment in the future of a school.

Each noun, verb, adjective or adverb in the extracted data was hand-tagged, when possible, with a number corresponding to the correct sense of the word in WordNet. For instance, the *establishment* in (6) is sense 1 of the *establishment* noun in WordNet which is glossed as "the act of forming something". Once all the data is so tagged, the nouns and verbs can simply be

[2] For this research, both the Prolog and the UNIX versions of WordNet were used. The Prolog version does not include any information on which group the verbs belong to.

retagged with an identification number that corresponds to the relevant WordNet group. *establishment* would be tagged with the number 100016649 which indicates that it is a member of the *action* group of nouns.

If the particular sense could not be determined then the first sense for the word encountered in WordNet was used. While this was the case for about 10% of the words, many words had only one sense, or had multiple senses as members of the same WordNet group..

Resnik (1993) also investigated the use of WordNet groups for resolving PP attachment ambiguity. However, he only grouped nouns and did not removed sense ambiguity from any of his data. Therefore, he did not establish whether the WordNet groups actually assist handling PP attachment when there is no sense ambiguity at all. This is precisely what this research intended to do.

All lexical items were also hand-tagged with a SYNCODE to indicate their syntactic role in the structure (see Table 3). For example, both *establishment* and *future* would be tagged with *n* for common noun.

Table 3: The syntactic tags (SYNCODEs)

n	common noun
ni	proper noun
nn	number
pr	pronoun
X*p*/xxx	SYNCODE X, complement of preposition xxx
v	verb
vc	verb taking the PP as a complement
va	verb taking the PP as an adjunct
a	adjective
r	adverb
s	sentence or S'
p	preposition
0	no object for this preposition

For some lexical items, like proper nouns, the sense number that is tagged is independent of WordNet. All non-date proper nouns were given the SYNCODE *ni* and the sense 0. Pronouns (SYNCODE *pr*) were given sense 1 if they referred to humans, 2 for animals, 3 for other tangible items and 4 for all other concepts. Numbers (SYNCODE *nn*) were tagged according to their length. For instance, the number 123 would be given the 'sense' number 3. Numbers with length 4 (i.e. years) were tagged as belonging to the *time* noun group and changed to SYNCODE *n*.

The resultant data can be investigated in various ways. Table 5 shows that different object groups prefer different attachment sites. When the contents of these tables are compared to the number of occurrences of each WordNet group in the data (Table 4), the preferences are more easily seen. For instance, the *change* verb group is the second most frequent group for attachment sites. However, PP with *artifact* or *person* nouns as objects are rarely attached to this group. Likewise, PPs with *action* object nouns are not frequently attached to *motion* nouns. The fact that *time* PPs most commonly attach to *action* nouns goes against the commonly used heuristic rule that *time* PPs attach to verbs.

Table 4: Ten most frequent groups for objects and attachment sites

OBJECTS		**CORRECT ATTACHMENT SITES**		
NOUN GROUPS	**NO.**	**GROUP**	**POS**	**NO.**
proper noun	133	action	N	103
artifact	113	change	V	77
action	109	motion	V	69
time	84	stative	V	60
cognition	75	communication	V	51
person	61	prepositions	P	46
group	57	communication	N	44
communication	54	person	N	43
attribute	52	creation	V	40
location	44	social	V	39
TOTAL	1097	TOTAL		1097

There is also a high preference for PPs to be attached to the same type of noun group as their object. For example, *artifact* PPs are frequently attached to *artifact* nouns. Thus a PP with a NP object is often semantically similar to the attachment noun. This explains why proper noun PPs are frequently attached to *person* and *group* nouns and supports the use of WordNet groups for semantic associations.

The five most common prepositions in the data (*in*, *to*, *for*, *with*, *on*) have preferences as to which WordNet groups are their objects (Table 6) or their attachment sites (Table 7). The preferences seem to follow the ranking in Table 4 but there are some notable exceptions. For example, few *time* nouns are objects in *to* PPs; there is a low frequency of verbs of *change* as *for* PP attachment sites and a low frequency of proper nouns as objects for *for* and *with* PPs. These may be due to the semantic roles given to the PPs.

While it is sometimes hard to define semantic roles, having looked at the data, it appears that *to* PPs have a preference for a 'locative' type of role. For this reason, the more common attachment site/object pairs for *to* PPs have an object that refers to a location, grouped as a proper noun, *artifact* or *location* (see (7)). There is

Table 5: The five most frequent attachment sites for the primary object WordNet groups
(N= noun group, V= verb group)

OBJECT		1		2		3		4		5	
	Total	**No.**	**Group**	**No.**	**Group**	**No.**	**Group**	**No.**	**Group**	**No.**	**Group**
PROPER NOUN	133	24	action N	11	person N	9	group N	8	creation V	5	motion V stative V communication V communication N
ARTIFACT	113	14	motion V	11	artifact N	10	action N	9	contact V	6	communication V
ACTION	109	12	action N	9	change V	8	stative V	7	communication V	5	social V possession V communication N
TIME	84	9	action N	8	stative V	6	time N	5	motion V change V possession V social V	4	body V communication V
COGNITION	75	8	change V	6	stative V action N	5	cognition N		*same no. for more than four groups*		
PERSON	61	6	stative V action N	5	communication V communication N	4	motion V person N	3	cognition N		

Table 6: Top five attachment sites for a selection of prepositions
(N= noun group, V= verb group)

PREP		1		2		3		4		5	
	Total	**No.**	**Group**	**No.**	**Group**	**No.**	**Group**	**No.**	**Group**	**No.**	**Group**
in	250	29	action N	21	change V	18	stative V	15	person N	9	event N group N
to	167	23	action N	18	motion V	10	change V stative V	9	communication N	7	communication V
for	100	9	cognition N	8	communication N	7	possession V action N	6	stative V	4	social V
with	97	10	change V	9	communication V	8	action N	7	stative V	5	creation V artifact N
on	86	7	change V	6	creation V motion V action N	5	communication V artifact N phenomenon N	4	cognition V cognition N communication N	3	possession N attribute N

Table 7: Top five object nouns for a selection of prepositions

PREP		1		2		3		4		5	
	Total	**No.**	**Group**	**No.**	**Group**	**No.**	**Group**	**No.**	**Group**	**No.**	**Group**
in	250	41	proper noun	27	artifact action time	16	cognition	14	location state	13	attribute
to	167	25	proper noun	16	artifact	15	cognition	13	location	12	action
for	100	18	action	11	artifact	10	person time	7	cognition possession	6	group
with	97	15	person	13	cognition	9	artifact	8	attribute	7	action communication
on	86	10	cognition action	9	artifact	8	communication	7	time	6	proper noun

also a preference for attachment to prepositions, *action* nouns and *motion* verbs for this semantic role. Therefore, the attachment preferences may certainly be due to a preference for a particular role.

(7) a. over to Phil (preposition/proper noun)
b. go to crossroads (*motion* verb/*artifact* noun)
c. circle to lawn (*motion* verb/ *location* noun)

However, a semantic role cannot be as easily defined for *with* PPs as these PPs have no clear role preferences. Except for the *artifact*/*artifact* pair (see (8)), it is hard to associate particular attachment-site/object pairs with specific roles. However, such a relationship may occur if there were more data to inspect.

(8) a. building with dome
b. room with dais

The preferences of the prepositions and the different relationships between the various object and attachment site groups, should be taken into consideration by parser. Therefore, it is expected that the most accurate parser will consider all the semantic and syntactic information about the PPs and their possible attachment sites.

4. Experiments with Parsing

The PP parser Lexass was developed to show that the accuracy of parsing can be improved through the use of the WordNet semantic associations and the data extracted from the corpus. For 372 of the 1097 PPs in the data, a complete list of possible attachment sites was added to the data by hand. Moving left through the sentences, all nouns were included as possible sites until a verb or the start of the sentence was encountered. For instance, while *establishment* is the correct attachment site for (6), *authorize*, a *communication* verb, is a possible attachment site. The job of Lexass is to use the 1097 example PP attachments from the corpus[3] and the WordNet groupings to determine which of these sites *in the future* should be attached to.

Table 8: A Selection of the Levels used by Lexass

LEVEL 33	all data
LEVEL 38	ignore attachment site WordNet group
LEVEL 40	ignore object WordNet group
LEVEL 16	ignore preposition
default	select left-most attachment site

[3] Of course, not including the PP being attached.

For each PP, Lexass moves through a sequence of scoring levels, trying to select the most likely attachment site. Each level accesses a file that contains the OCCURRENCENUMs for various patterns (see Table 8). An OCCURRENCENUM corresponds to the number of occurrences in the data of certain attachment patterns.

For instance, LEVEL 33 examines the number of times in the data that a particular preposition with a particular object is attached to each possible attachment site. The site with the highest OCCURRENCENUM for this pattern is the site to which the PP is attached more frequently in the data than any other possible sites.

Other levels generalize the counting of the OCCURRENCENUMs. LEVEL 40 ignores the WordNet group for the preposition's object, only including its SYNCODE. Therefore, for LEVEL 40, Lexass would use the OCCURRENCENUMs for an *in* PP with any noun object attached to a *communication* verb or an *action* noun.

To handle small frequencies and similar OCCURRENCENUMs, a cut-off score was used. If the top OCCURRENCENUM for a level was not more than the cut-off, or the difference between the two highest OCCURRENCENUMs was not greater than the cut-off, then the next level in the sequence was attempted (see Table 9).

Table 9: The algorithm for Lexass

- Read the level from the sequence parameter.
  - Score each attachment site according to the level.
  - If there is a valid top score ,
    select that attachment score.
  - Else, go to next level of scoring.
- If no scoring levels give valid top scores,
  use the default method of selection.

The final level in every sequence was the default level of selection. The accuracy of the default level was used as a benchmark to compare all other selection methods against. By default, PPs were attached to left-most site in the sentence. As this is normally the highest site in the sentence structure, the default method resembles the method of Minimal Attachment, except that it does not actually check for minimality.

This scoring method eliminates another of the problems with Hindle & Rooth and Collins & Brooks. Their methodologies are restricted to deciding between PP attachment to a verb or attachment a noun and their scoring formulae refer to all possible attachment sites in the one formula. They do not try and handle phrases

Table 10: Correct % at individual levels

	All PPs		DOUBLES		5PREPS		5DOUBLES	
Level Includes	**No. Correct**	**% Correct**	**No. Correct**	**% Correct**	**No. Correct**	**% Correct**	**No. Correct**	**% Correct**
Default (1)	*206*	*55 %*	*151*	*63 %*	*138*	*56 %*	*101*	*62 %*
All data (33)	16	76 %	8	80 %	13	72 %	7	78 %
No Object (40)	41	64 %	28	72 %	31	57 %	22	66 %
No Site (38)	49	64 %	33	72 %	35	57 %	25	66 %
No Prep (16)	131	48 %	89	52 %	80	43 %	53	47 %

Table 11: Correct % of entire sequences
(Sequence = Level/s, Default)

	All PPs		DOUBLES		5PREPS		5DOUBLES	
Sequence Includes	**No. Correct**	**% of 372**	**No. Correct**	**% of 239**	**No. Correct**	**% of 246**	**No. Correct**	**% of 162**
1	*206*	*55 %*	*151*	*63 %*	*138*	*56 %*	*101*	*62 %*
33, 1	214	58 %	156	65 %	143	58 %	105	65 %
40, 1	213	57 %	156	65 %	140	57 %	104	64 %
38, 1	216	58 %	156	65 %	142	58 %	103	64 %
16, 1	184	49 %	130	54 %	114	46 %	82	51 %
33,38,40,1	218	59 %	157	66 %	142	58 %	103	64 %

like (9). These PPs can be handled by Lexass as, by using OCCURRENCENUMS, it scores each possible site individually. Therefore, it is not restricted to any particular number of possible attachments.[4]

(9) The man in the store with a white hat on his head

The experiments tried a range of sequences. The accuracy of chosen attachments varied depending on the levels in the sequences. As well as parsing all 372 of the PPs, various subgroups of the PPs were also parsed separately. DOUBLES is all the PPs with only two possible attachment sites. This allows some sort of comparison to be made with Hindle & Rooth and Collins & Brooks. 5PREPS is the selection of PPs which contain the five most common prepositions. 5DOUBLES is the selection of PPs from 5PREPS that have two possible attachment sites.

While the best sequence [33,38,40,1] selects the correct attachment only 59% of the time, this is clearly better than the 55% accuracy of the default method (see Table 11). If the accuracy of the selections made at individual levels is investigated (see Table 10), a number of further points can be made. Not surprisingly, the most accurate level is LEVEL 33 (76%) as it includes all of the semantic data about the PP and its possible attachment sites. However, only 21 of the 372 PPs were attached by LEVEL 33.[5] This is due to small size of the training data. Any further studies should include a larger set of data. The small amount of data also resulted in the best cut-off score being 0. Therefore, at LEVEL 33, any attachment site/preposition/object pattern in the training data provides a valid OCCURRENCENUM.

In contrast, LEVEL 40 and LEVEL 38 are able to attach more of the PPs due to the generalization of the patterns they were counting in the data. They are not as successful as LEVEL 33 (only 64% accurate) but are a clear improvement on the default level (55%).

The failure of LEVEL 16 indicates the importance of the preposition in the parsing process. The preferences of the preposition must be considered when parsing. The attachment should be guided by the choice of preposition involved as it is the sense of the preposition that will establish the semantic role of the PP. While this is still partially dependent on other components of the sentence, it is a vital factor that should not be ignored.

If the results for DOUBLES, 5PREPS and 5DOUBLES are investigated, the same sort of accuracies occur. Furthermore, the 80% accuracy of LEVEL 33 for DOUBLES is comfortably similar to the accuracy reached by Hindle & Rooth and Collins & Brooks.

To test the efficiency of this parsing methodology with non-sense tagged texts, the 372 PPs were reprocessed in such a way that all the WordNet senses were presumed to be ambiguous. Hence, the lexical

[4] The 372 PPs had on average 2.5 possible attachment sites. Therefore, guesswork would choose the correct attachment 40% of the time.

[5] I.e. the top OCCURRENCENUM for these PPs was a valid score.

items were placed in the WordNet group of the first sense encountered in WordNet. When parsed, the sequence [33,1] was only 57% accurate, with LEVEL 33 reducing from 76% accuracy to 72%. While these results are greater than the default level's benchmark of 55% accuracy, there is a clear reduction in accuracy when there is word sense ambiguity. However, as indicated in Section 3, this ambiguity does not always affect which WordNet groups some words are placed in. Hence the methodology recommended by this paper may be suitable even when the WordNet sense of words is unknown.

5. Conclusion

The results from this research demonstrate that the semantic associations given as hypernym and troponym trees in WordNet V5.1 can be used to categorize lexical items when parsing PP attachment. Clear preference patterns have been shown to exist, demonstrating relationships between prepositions, their objects and their attachment sites. The use of these preference patterns in the parsing was reasonably successful, provided the preposition was always included in the pattern.

The results also support the use of partial parsers. Partial parsing can be used to provide the basic phrasal structures in a sentence then semantic preferences, like those described in this paper, can be used to determine the most suitable phrasal attachments, given their context. Due to the varying levels of preferences, a system of weights, like those used for neural networks, may be required in order to balance the preferences in some way.

6. Acknowledgments

Thanks to Dominique Estival for her guidance. Thanks to the Department of Linguistics & Applied Linguistics, University of Melbourne, for purchasing and allowing me to use the DSO corpus and the Department of Computer Science for the computer facilities used for this research.

7. References

Abney, Steven. 1997. *The SCOL Manual: Version 0.1b*. http://www.sfs.nphil.uni-tuebingen.de/~abney/.

Abney, Steven. 1991. Parsing by Chunks. In Robert Berwick, Steven Abney & Carol Tenny (Eds.), *Principle-Based Parsing*. Dordrecht: Kluwer Academic Publishers. Also http://www.sfs.nphil.uni-tuebingen.de/~abney/.

Collins, Michael & Brooks, James. 1995. Prepositional Phrase Attachment through a Backed-Off Model. *Proceedings of the Association for Computational Linguistics Third Workshop on Very Large Corpora*. Also http://www.lanl.gov/archive /cmp-lg/9506021 .

Fellbaum, Christine. 1993. English Verbs as a Semantic Net. In *Five Papers on WordNet* (CSL Report 43, Revised). Also http:// www.cogsci.princeton.edu/~wn/ .

Hindle, Donald & Rooth, Mats. 1993. Structural Ambiguity and Lexical Relations. *Computational Linguistics*, 19(1), 103-120. Also http:// www2.ims.uni-stuttgart.de/~mats/ .

Miller, George A; Beckwith, Richard; Fellbaum, Christian; Gross, Derek & Miller, Katherine. 1993. Introduction to WordNet: An On-Line Lexical Database. In *Five Papers on WordNet* (CSL Report 43, Revised). Also http://www.cogsci.princeton.edu /~wn/ .

Miller, George A. & Fellbaum, Christian. 1991. Semantic networks of English. *Cognition*: *International Journal of Cognitive Science*, 41(1-3). Reprinted Beth Levin & Steven Pinker (Eds.), *Lexical & Conceptual Semantics*. Cambridge: Blackwell Publishers. 197-229.

Niemann, Michael. 1997. *Determining Prepositional Phrase Attachment by Semantic Association and Preferences*. Honours thesis. Department of Linguistics & Applied Linguistics, the University of Melbourne, Melbourne.

Smith, George W. 1991. *Computers and Human Language*. New York: Oxford University Press.

Taraban, Roman & McClelland, James L. 1988. Constituent Attachment and Thematic Role Assignment in Sentence Processing: Influences of Content-Based Expectations. *Journal of Memory and Language*. 27, 597-632.