Integrating Symbolic and Statistical Methods for Prepositional Phrase Attachment

Sanda M. Harabagiu and **Marius Paşca**
Department of Computer Science and Engineering
Southern Methodist University
Dallas, TX 75275-0122
{sanda,mars}@seas.smu.edu

Abstract

This paper[1] presents a novel methodology of resolving prepositional phrase attachment ambiguities. The approach consists of three phases. First, we rely on a publicly available database to classify a large corpus of prepositional attachments extracted from the Treebank parses. As a by-product, the arguments of every prepositional relation are semantically disambiguated. In the second phase, the thematic interpretation of the prepositional relations provides additional knowledge. The third phase is concerned with learning attachment decisions from word class knowledge and relation type features. The learning technique builds upon some of the most popular current statistical techniques.

We have tested this methodology on (1) Wall Street Journal articles, (2) textual definitions of concepts from a dictionary and (3) an ad-hoc corpus of Web documents, used for conceptual indexing and information extraction.

Introduction

The problem of prepositional attachments generates one of the major causes of ambiguity in natural language. For example, sentences (S1-4) illustrate the two possibilities of attaching the prepositional phrase from the phrasal context [VP NP for-PP]:

(S1) The executives [joined]$_{VP}$[the president]$_{NP}$ [for the evening]$_{PP}$.
(S2) Last spring Nelson Mandela [was proposed]$_{VP}$ [president]$_{NP}$[for life]$_{PP}$.
(S3) President Bush [has approved]$_{VP}$[duty-free treatment]$_{NP}$[for the Canadian imports]$_{PP}$.
(S4) The chairman [has adjusted]$_{VP}$[all the interests]$_{NP}$[for inflation prevention]$_{PP}$.

In the case of (S1), the prepositional phrase is attached to the verb phrase, as it indicates the duration of joining the president. In sentence (S2), the prepositional phrase, expressing a period of time, is attached to the noun phrase, since it is the position of president that is proposed to be for life. In (S3), the prepositional phrase is an adjunct of the noun phrase [duty-free treatment] because the Canadian imports are the object of the nominalization treatment. However, in (S4), [the interests]$_{NP}$ represents also a nominalization, but the prepositional phrase is attached to the verb phrase, since the adjustment is done with the goal of preventing the inflation. The difference comes from the fact that noun interest has several semantic senses, and the only one that is a nominalization (i.e. sense 5 from WordNet (Fellbaum 1998)) is not the correct sense of interest in the context of (S4).

From these examples we see that the disambiguation of prepositional attachments is based on lexical, thematic and world knowledge. Most of this knowledge is not directly available, therefore we need to rely only on partial knowledge, brought forward by empirical methods operating on large sets of attached phrases.

Recent work in prepositional attachment uses:
(1) statistical approaches to the problem,
(2) knowledge cues derived from lexical databases or
(3) a combination of supervised learning methods and context-based disambiguation algorithms.
Corpus-based statistical prepositional attachment ambiguity resolution was first reported in (Hindle and Rooth 1993). A corpus of 200,000 [VP NP PP] triplets helped devise an unsupervised method of deciding the PP attachment. The decision is based on comparing the co-ocurrance probabilities of a preposition with nouns and verbs from triplets. This method performs at 80% accuracy on a test set of 880 examples.

Another promising approach is the transformation-based rule derivation reported in (Brill and Resnik 1994). Initially, all attachments are assumed adjectival (i.e. the PP is attached to the NP). Rules of transformation from adjectival into adverbial attachments are learned, based on the features of a training corpus. This method achieves an 81.8% success rate on 500 randomly selected sentences.

The current state-of-the art statistical method is the backed-off model proposed in (Collins and Brooks 1995), providing overall an accuracy of 84.5%. Like most of the statistical methods, it suffers from *the sparse data problem.* All these methods are based on matching words from a test set against those from a training set. Unfortunately, many triplets may appear in the test data without ever being encountered in the training data. Brill and Resnik were the first to find a possible solution to this problem, by using word semantic classes instead of words for direct matching. This entails the integration of the word sense disambiguation problem with the PP attachment decision.

A different, knowledge-based method proposed in (Harabagiu 1996), starts by categorizing the arguments of prepositional relations collected from the *Wall Street Journal* corpus from Treebank (Marcus et al.1993) against the semantic classes defined by WordNet (Fellbaum 1998). Then, inferential heuristics establish thematic features of prepositional relations. Unfortunately, the paper does not report on the accuracy of the prepositional attachment, specifying only an overall disambiguation rate of 72.3% that comprises both word sense discrimination and prepositional attachments.

Considering the problem of word sense disambiguation and the prepositional attachment as two interacting processes, Stetina and Nagao report in (Stetina and Nagao 1997) a novel supervised learning method for prepositional attachment. Their method performs word sense disambiguation using a semantic distance between WordNet concepts. WordNet hierarchies are used also to devise the decision trees that provide the PP attachment decisions. Their approach scores the best current performance of prepositional attachment, with an average precision of 88.1%.

In this paper we integrate knowledge resources with statistical techniques to decide upon PP attachments. First of all, we extend the classification algorithm presented in (Harabagiu 1996). We add the Gazetteers files as knowledge resources for proper names, and therefore obtain an overall disambiguation precision of 87.2%. Then we devise a different way of recognizing the thematic features of prepositional relations. The novelty of our method consists in the fact that we incorporate in the learning phase not only word class information, as was done in (Brill and Resnik 1994) and (Stetina and Nagao 1997), but also thematic features of prepositional relations. We evaluate this approach on three different kinds of texts: (1) the *Wall Street Journal* corpus from Treebank, (2) a corpus of conceptual definitions provided by WordNet and (3) an ad-hoc collection of Web documents used for indexing and Information Extraction tasks.

Classes of prepositional relations

The method of classifying prepositions introduced in (Harabagiu 1996) considers only the attachments that obey the principle of locality (cf. (Wertmer 1991)), e.g. the PP is considered to be always attached to the immediately preceding phrase. In (Hindle and Rooth 1993) it was shown that this principle doesn't work well on real world texts, therefore we chose to consider the attachments derived from the parses of a test set of articles extracted from the *Wall Street Journal* corpus. We scanned the PP attachments and filtered the phrase heads to create an ad hoc collection of sequences <*noun prep noun*> and <*verb prep noun*>. We have also filtered out all the prepositional structures that are matched in the Gazetteers files. Such structures represent names of companies or locations. This filter produces better scores of disambiguation than those reported in (Harabagiu 1996). The rest of the collection is divided into classes of prepositional relations, using the following definitions:

Definition 1: Two prepositional attachments $<noun_1$ $prep\ noun_2>$ and $<noun_3\ prep\ noun_4>$ belong to the same class when there are two relations [$noun_1\ r_1$ $noun_3$] and [$noun_2\ r_2\ noun_4$] representing one of the cases listed in Table 1. We assume $word_1$=$noun_1$ and $word_2$=$noun_3$ or $word_1$=$noun_2$ and $word_2$=$noun_4$ respectively.

Case	Relation
(a)	$word_1$ is a *synonym* of $word_2$
(b)	$word_1$ is a *hypernym* of $word_2$
(c)	$word_1$ is a *hyponym* of $word_2$
(d)	$word_1$ and $word_2$ *belong to the same hierarchy*
(e)	$word_1$ is the *genus of the gloss* of $word_2$
(f)	$word_1$ is the *genus of the gloss* of $word_2$
(g)	$word_1$ is the *genus of the gloss* of one of the concepts in the hierarchy of $word_2$
(h)	$word_2$ is the *genus of the gloss* of one of the concepts in the hierarchy of $word_1$

Table 1: WordNet-based relations defining classes of prepositional attachments

Definition 2: Two prepositional attachments $< verb_1$ $prep\ noun_1 >$ and $< verb_2\ prep\ noun_2 >$ belong to the same class when there are two relations [$noun_1$ $r_1\ noun_3$] and [$noun_2\ r_2\ noun_4$] representing one of the cases listed in Table 1. We assume $word_1$=$verb_1$ and $word_2$=$verb_2$ or $word_1$=$noun_1$ and $word_2$=$noun_2$ respectively.

The immediate benefit of grouping prepositional relations into classes is semantic disambiguation of their arguments. The relations from cases (a)-(d) in Table 1 identify the WordNet synonym or the hierarchy containing the arguments of r_1 and r_2, therefore defining the semantic senses of $word_1$ and $word_2$. Cases (e) and (g) identify only the semantic sense of $word_2$. Similarly, cases (f) and (h) identify only the semantic sense of $word_1$. The sense resolution of the other word amounts to disambiguating the genus of the gloss of a known synset.

Prep	Nr. of triplets	Multiple element classes	% Classified relations
of	3,790	125	72.3%
for	803	37	74.3%
from	870	31	76.6%
as	306	14	73%

Table 2: Distribution of prepositional relations classes

Previous work on automatically building hierarchies from dictionary definitions (e.g. (Klavans et al.1990)) indicates that empirical methods can disambiguate successfully the genus of the WordNet glosses. We have developed the following heuristics that disambiguate $word_1$, the genus of the gloss of $word_2$:

Heuristic 1: If there is a sense s of $word_1$ such that it has the genus of its gloss (or the genus of any of its hypernyms) in the same hierarchy as $word_2$, then disambiguate $word_1$ to sense s.

Example: Given the two prepositional structures <*retirement from office*> and <*withdrawal from position*>, we find that the gloss of `retirement`, sense 2, is *(withdrawal from position or occupation)*. Therefore $word_j$=*retirement* with the sense 2, and $word_1$=*withdrawal*. To find the sense of *withdrawal* we notice that the gloss of sense 1 of *withdrawal* is *(retraction from position)*. The genus of this gloss (i.e. *retraction*) is in the same hierarchy with sense 2 of *retirement*. We conclude therefore that *withdrawal* must have sense number 2 from WordNet.

Heuristic 2: If there is a sense s of $word_1$ such that its gloss contains the same prepositional relation and one argument belongs to the same hierarchy as $word_1$, then the semantic sense of $word_1$ is s.

Example: Given two prepositional structures <*explore for knowledge*> and <*examine for sake*>, we see that sense 4 of verb *explore* has in its gloss a prepositional relation <*examine for purpose*>. We also find sense 3 of `examine` to belong to the same hierarchy as `explore`, sense 4, subsumed by the synset {`analyze, analyse, study, examine`}.

Heuristic 3: Let $\{H_i\}$ denote the immediate hypernyms of all senses of $word_1$ and $genus_2$ the genus of the hypernym of $word_2$. To find the sense of $word_1$ apply the recursive procedure $gloss\_search(genus_2, 1, \{H_2\})$. The pseudocode of this procedure is:

```
Procedure gloss_search(genus_j, depth, {H_i})
if (depth == 4) return 0;
if genus_j is found in any H_j (or its geni) return j;
else for every sense s of genus_j
  retrieve new_genus(s) = the gloss genus of genus_j
      having the semantic sense=s
  apply gloss_search(s) (new_genus, depth+1, {H_i})
  if (result is not 0) return result;
```

Example: Given the two prepositional structures <*chairman of company*> and <*leader of computing*>, we retrieve sense 1 of noun *chairman* having the gloss genus the noun *leader*. Noun *leader* has four senses, and hence four hypernyms. The gloss genus of the hypernym of *chairman* is `person`, which is the gloss genus of the hypernym of the first sense of *leader* as well.

Prep	Nr. of case a	Nr. of case b|c|d	Nr. of case e|g	Nr. of case f|h
of	34	283	675	1509
for	16	113	229	445
from	12	129	237	492
as	6	41	86	173

Table 3: Distribution of relations between preposional structures. The cases are those listed in Table 1.

The classification produces two kinds of classes: some containing only one prepositional structure, and others containing multiple, disambiguated structures. Tables 2 and 3 list the classification results. Discarding the one-element classes, as it was chosen in (Harabagiu 1996), increases the chance of the sparse data problem. Consequently, we chose to apply a semantic similarity metric between classes of prepositinal relation and append the classes having the largest semantic similarity. The process is repeated until there are no unique relation classes left. We have employed the semantic density measure defined as:

□ For any one-element class $\mathcal{E}$={<$word_1$ *prep* $word_2$>}, with $word_1$ a noun or a verb and $word_2$ always a noun, we compute the semantic similarity to other classes in the following way:

□ Given a class $\mathcal{C}$ with multiple prepositional structures <$word_i$ *prep* $word_{i+1}$>, the semantic similarity of $\mathcal{E}$ to $\mathcal{C}$ is given by $d = \frac{1}{2}\sqrt{(d_1^2 + d_2^2)}$, where :

- $d_1 = \sum_i$(nr. of common non-stop words in the glosses of: $word_1$, $word_i$ and their hypernyms), and
- $d_2 = \sum_i$(nr. of common non-stop words in the glosses of: $word_2$, $word_{i+1}$ and their hypernyms)

Finding disambiguated classes of prepositional relations allows for the inference of additional features of prepositional attachments.

Thematic features of prepositional attachments

Riloff notes in (Riloff and Schmelzenback 1998) that a thematic relation (e.g. agent, object, instrument) can be lexicalized by a variety of prepositional relations.

This property accounts for one of the main difficulties in acquiring linguistic patterns for the Information Extraction task, therefore selectional constraints have an important role in the disambiguation of prepositional relations. Our approach to deriving the thematic features of prepositional relations is based on:
(1) derivational morphology encoded in WordNet,
(2) phrasal parses of the conceptual glosses,
(3) lists of typical objects of agents, provided in the synset glosses, and
(4) a special treatment of the time, space and quantity expressions, as they have been imposed by the information extraction tasks (.cf (MUC-6)).

WordNet 1.6 encodes a wealth of lexemes obtained by derivational morphology. The largest part of them is contained in the noun semantic class that describes actions. Rules of word formation (cf. (Bauer 1983)) indicate the thematic role of a lexeme with respect to its root. For example, a `successor` is the *agent* that `succeeds` in a position, an `acquisition` represents the action of `acquiring` and a `cutter` is an instrument used for `cutting`. We have implemented these rules and have added morpho-thematic relations between WordNet synsets. In addition we haved parsed the glosses and their examples with a finite-state phrasal parser, detecting *agent* and *object* thematic roles.

Additional thematic roles (e.g. *instrument*, *consequence*) were recognized by testing whether preposional arguments are subsumed by several WordNet synsets (e.g. {*instrumentality, instrumentation*} or {*consequence, effect, outcome, result*}). Thematic features are obtained by the following procedure:

```
for every <word1 prep word2>
if word1 or word2 represent time, location or quantity
          goto ready;
if word1 is an action
   if word2 is agent or object goto ready;
   else for every theme known
      if word2 is theme goto ready;
   create new theme;
else if theme1=theme(word1∈{agent,object,known_theme}
   for every known theme theme2 ≠ theme1
      if theme2==theme(word2) goto ready;
   create new theme;
ready: end;
```

Using this methodology we have obtained the same thematic interpretation of prepositional relations as the one reported in (Harabagiu 1996), and listed in Table 4. For example, the interpretation of the prepositional relations <`acquisition of company`> is recognized as an action exercised upon its object because: 1) the nominalization `acquisition` is morphologically derived from the verb `acquire` and it represents the same action.
2) noun `company` is the *object* of acquire, since verb `acquire` subsumes the synset {`take over, buy out, buy up`}, which lists companies and corporations as possible objects.

The selection of the thematic roles was determined initially by manual inspection of the classes of prepositional relations. As Table 4 shows, the resulting features are either (a) thematic roles or (b) combinations of thematic roles.

Features for $< N1 >$ of $< N2 >$	Example
$N2$=*object of* action($N1$)	acquisition of company
$N2$=*agent of* action($N1$)	approval of authorities
$N1$=*agent of* action with *object*=$N2$	author of paper
$N1$=*agent of* action with *purpose* action($N2$)	activists of support
$N1$=*result of* action whose *agent*=$N2$	record of athlete
$N2$=action with *theme*=$N1$	allegations of fraud
$N1$=*location of* activity($N2$)	place of business

Table 4: Thematic features of prepositional relations

Learning PP attachment decisions

Prepositional attachment decisions over quadruples $[VP\ NP_1\ prep\ NP_2]$ from unseen sentences are based on three kinds of features:
(a) the result of classification tests of $[VP\ prep\ NP_2]$ and $[NP_1\ prep\ NP_2]$ against any of the classes of prepositional relations;
(b) the thematic features resulting for both triplets $[VP\ prep\ NP_2]$ and $[NP_1\ prep\ NP_2]$, and
(c) the semantic similarity to classes of prepositional relations, when the classification tests fail.

We consider three established supervised learning algorithm to obtain the attachment decisions:
C4.5(Quinlan 1992): an algorithm that automatically builds decision trees based on the feature values of positive and negative examples of attachments from the training set. Attachment decisions are made by traversing the decision tree from the root to a leaf that indicates adjectival or adverbial attachment. The traversal is determined by the features resulting from classification, thematic similarity and semantic similarity.
CN2(Clark and Niblett 1989): A rule induction algorithm that selects the attachment rules that cover the largest possible classes of prepositional relations from the training examples, as measured by a Laplace error estimate.
PEBLS(Cost and Salzberg 1993): A k nearest-neighbor algorithm where classification is performed by assigning a test instance to the majority class of the k closes examples (in our case classes of prepositional relations). When using k=1, we obtain a standard nearest-neighbor classifier, which is most appropriate for data where all features are relevant.

We have modified all these algorithms to better fit the characteristics of PP attachment problems. Similarly to the learning phase presented in (Stetina and Nagao 1997), we have modified the C4.5 algorithm by allowing a traversal to a new node in the decision tree only when a special condition is satisfied. Instead of a semantic distance, we have chosen to use the condition that the next node has to maintain the values of the thematic features (i.e. no new thematic roles are learned).

For the CN2 algorithm, we measure the Laplace error estimate only between prepositional attachments that have the same thematic features. Finally, for the PEBLS program, the closest examples were considered those having the largest semantic similarity.

Method	C_1	C_2	C_3
Alaways adjectival	57.2%	63.1%	58.1%
Most likely	70.3%	68.5%	66.8%
C4.5	91.3%	90.5%	90.2%
CN2	90.5%	89.6%	89.9%
PEBLS	88.6%	85.9%	87.3%
Modified decision tree (Stetina and Nagao 1997)	90.8%	90.2%	90.7%
Back-off model	88.1%	74.3%	77.8%
Combining WordNet classes with similarity measures	93.2%	93.3%	94.7%

Table 5: Precision of the PP attachment methods

Table 5 illustrates the results of PP attachment performed on (a) corpus C_1=1000 unseen sentences from the *Wall Street Journal* corpus, (b) corpus C_2=the glosses of 1000 synsets from WordNet and (c) corpus C_3=1000 sentences of Web documents retrieved when querying for <*noun*$_1$ *prep noun*$_2$>, a random element from the largest populated class of prepositional relations.

Discussion and evaluation

The most computationally expensive part of the system is the classification of the training examples of prepositional relations. Every new attachment had to be tested first against each class. When it did not belong to any existing class, it had to be tested against prior attachments that could not be classified. Running these tests on an ALPHA DEC 300 MHz machine took up to an hour. This is however faster than calculating the frequency tables used in (Hindle and Rooth 1993). The above experiments have confirmed the expectations that using thematic features in combination with word class information will improve the precision of the attachments.

Although our method exhibits good accuracy, we feel that there is a lot of work to be done, especially in measuring the interaction between word sense disambiguation, thematic feature discrimination and PP attachment. At the moment we study the effects of different semantic similarity measures on the overall precision. We also contemplate a larger range of learning techniques.

References

L. Bauer. English word-formation. Cambridge University Press, 1983.

E. Brill and P. Resnik. A rule based approach to PP attachment disambiguation. In *Proceedings of 15th International Conference on Computational Linguistics (COLING-94)*, Kyoto, Japan, pages 1198–1204, 1994.

P. Clark and T. Niblett. The CN2 induction algorithm. *Machine Learning*, 3(4):261–283, 1989.

M. Collins and J. Brooks. PP Attachment through a backed-off model. In *Proceedings Third Workshop on Very Large Corpora*, 1995.

S. Cost and S. Salzberg. A weigthed nearest neighbor algorithm with symbolic features. *Machine Learning*, 10(1):57–78, 1993.

C. Fellbaum. WordNet - An electronic lexical database. MIT Press, 1998.

S. Harabagiu. An application of WordNet to Prepositional Attachment. In *Proceedings of 34th Annual Meeting of the Association for Computational Linguistics (ACL-96)*, Santa Cruz, CA, 360–362, 1996.

H. Hindle and M. Rooth. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120, 1993.

J. Klavans, M. Chodorow and N. Wacholder. From dictionary to knowledge base via taxonomy. In *Electronic Text Research*, University of Waterloo, Center for the New OED and Text Research, Waterloo, Canada, pages 110–132, 1990.

M. Marcus, B. Santorini and M.A. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Proceedings of the Sixth Message Understnading Conference (MUC-6). San Francisco. Morgan Kaufmann, 1995.

J. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, San MAteo, CA, 1992.

E. Riloff and M. Schmelzenback. An empirical approach to conceptual case frame acquisition. In *Proceedings of the 6th Workshop on Very Large Corpora*, 1998.

J. Stetina and M. Nagao. Corpus based PP attachment ambiguity resolution with a semantic dictionary. In *Proceedings of the 4th Natural Language Processing Pacific Rim Symposium*, pages 66–80, 1997.

S. Wertmer. Integration of Semantic and Syntactic Constraints for Structural Noun Phrase Disambiguation. In *Proceedings of 12th International Joint Conference on Artificial Intelligence (IJCAI-91)*,pages 1486–1491, 1991.