A Connectionist Approach to Prepositional Phrase Attachment for Real World Texts

Josep M. Sopena and **Agusti LLoberas** and **Joan L. Moliner**
Laboratory of Neurocomputing
University of Barcelona
Pg. Vall d'Hebron, 171
08035 Barcelona (Spain)
e-mail: {pep,agusti,joan}@axon.psi.ub.es

Abstract

In this paper we describe a neural network-based approach to prepositional phrase attachment disambiguation for real world texts. Although the use of semantic classes in this task seems intuitively to be adequate, methods employed to date have not used them very effectively. Causes of their poor results are discussed. Our model, which uses only classes, scores appreciably better than the other class-based methods which have been tested on the Wall Street Journal corpus. To date, the best result obtained using only classes was a score of 79.1%; we obtained an accuracy score of 86.8%. This score is among the best reported in the literature using this corpus.

1 Introduction

Structural ambiguity is one of the most serious problems faced by Natural Language Processing (NLP) systems. It occurs when the syntactic information does not suffice to make an assignment decision. Prepositional phrase (PP) attachment is, perhaps, the canonical case of structural ambiguity. What kind of information should we use in order to solve this ambiguity? In most cases, the information needed comes from a local context, and the attachment decision is based essentially on the relationships existing between predicates and arguments, what Katz y Fodor (1963) called selectional restrictions. For example, in the expression: *(V accommodate) (NP Johnson's election) (PP as a director)*, the PP is attached to the NP. However, in the expression: *(V taking) (NP that news) (PP as a sign to be cautious)*, the PP is attached to the verb. In both expressions, the attachment site is decided on the basis of verb and noun selectional restrictions. In other cases, the information determining the PP attachment comes from a global context. In this paper we will focus on the disambiguation mechanism based on selectional restrictions.

Previous work has shown that it is extremely difficult to build hand-made rule-based systems able to deal with this kind of problem. Since such handmade systems proved unsuccessful, in recent years two main methods have appeared capable of automatic learning from tagged corpora: automatic rule based methods and statistical methods. In this paper we will show that, providing that the problem is correctly approached, an NN can obtain better results than any of the methods used to date for PP attachment disambiguation.

Statistical methods consider how a local context can disambiguate PP attachment estimating the probability from a corpus:

$p(verb\ attach|v\ NP1\ prep\ NP2)$

Since an NP can be arbitrarily complex, the problem can be simplified by considering that only the heads of the respective phrases are relevant when deciding PP attachment. Therefore, ambiguity is resolved by means of a model that takes into account only phrasal heads: $p(verb\ attach|verb\ n1\ prep\ n2)$. There are two distinct methods for establishing the relationships between the verb and its arguments: methods using words (lexical preferences) and methods using semantic classes (selectional restrictions).

2 Using Words

The attachment probability

$$p(verb\ attach|verb\ n1\ prep\ n2)$$

should be computed. Due to the use of word cooccurrence, this approach comes up against the serious problem of data sparseness: the same 4-tuple *(v n1 prep n2)* is hardly ever repeated across the corpus even when the corpus is very large. Collins and Brooks (1995) showed how serious this problem can be: almost 95% of the 3097 4-tuples of their test set do not appear in their 20801 training set 4-tuples. In order to reduce data sparseness, Hindle and Rooth (1993) simplified the context, by considering only verb-preposition ($p(prep|verb)$), and n1-preposition ($p(prep|n1)$) co- occurrences. $n2$ was ignored in spite of the fact that it may play an important role. In the test, attachment to verb was decided if $p(prep|verb) > p(prep|noun)$; otherwise attachment to $n1$ is decided. Despite these limitations, 80% of PP were correctly assigned.

Another method for reducing data sparseness has been introduced recently by Collins and Brooks

(1995). These authors showed that the problem of PP attachment ambiguity is analogous to n-gram language models used in speech recognition, and that one of the most common methods for language modelling, the backed-off estimate, is also applicable here. Using this method they obtained 84.5% accuracy on WSJ data.

3 Using Classes

Working with words implies generating huge parameter spaces for which a vast amount of memory space is required. NNs (probably like people) cannot deal with such spaces. NNs are able to approximate very complex functions, but they cannot memorize huge probability look-up tables. The use of semantic classes has been suggested as an alternative to word co-occurrence. If we accept the idea that all the words included in a given class must have similar (attachment) behaviour, and that there are fewer semantic classes than there are words, the problem of data sparseness and memory space can be considerably reduced.

Some of the class-based methods have used WordNet (Miller et al., 1993) to extract word classes. WordNet is a semantic net in which each node stands for a set of synonyms (*synset*), and domination stands for set inclusion (IS-A links). Each *synset* represents an underlying concept. Table 1 shows three of the senses for the noun *bank*. Table 2 shows the accuracy of the results reported in previous work. The worst results were obtained when only classes were used. It is reasonable to assume a major source of knowledge humans use to make attachment decisions is the semantic class for the words involved and consequently there must be a class-based method that provides better results. One possible reason for low performance using classes is that WordNet is not an adequate hierarchy since it is hand-crafted. Ratnaparkhi et al. (1994), instead of using hand-crafted semantic classes, uses word classes obtained via Mutual Information Clustering (MIC) in a training corpus. Table 2 shows that, again, worse results are obtained with classes. A complementary explanation for the poor results using classes would be that current methods **do not use class information very effectively** for several reasons: 1.-In WordNet, a particular sense belongs to several classes (a word belongs to a class if it falls within the IS-A tree below that class), and so determining an adequate level of abstraction is difficult. 2.- Most words have more than one sense. As a result, before deciding attachment, it is first necessary to determine the correct sense for each word. 3.- None of the preceding methods used classes for verbs. 4.- For reasons of complexity, the complete 4-tuple has not been considered simultaneously except in Ratnaparkhi et al.(1994). 5.- Classes of a given sense and classes of different senses of different words can have complex interactions and the preceding methods cannot take such interactions into account.

4 Encoding and Network Architecture.

Semantic classes were extracted from Wordnet 1.5. In order to encode each word we did not use WordNet directly, but constructed a new hierarchy (a subset of WordNet) including only the classes that corresponded to the words that belonged to the training and test sets. We counted the number of times the different semantic classes appear in the training and test sets. The hierarchy was pruned taking these statistics into account. Given a threshold h, classes which appear less than $h\%$ were not included. In this way we avoided having an excessive number of classes in the definition of each word which may have been insufficiently trained due to a lack of examples in the training set. We call the new hierarchy obtained after the cut *WordNet'*. Due to the large number of verb hierarchies, we made each verb lexicographical file into a tree by adding a root node corresponding to the file name. According to Miller et al. (1993), verb *synsets* are divided into 15 lexicographical files on the basis of semantic criteria. Each root node of a verb hierarchy belongs to only one lexicographical file. We made each old root node hang from a new root node, the label of which was the name of its lexicographical file. In addition, we codified the name of the lexicographical file of the verb itself.

There are essentially two alternative procedures for using class information. The first one consists of the simultaneous presentation of all the classes of all the senses of all the words in the 4-tuple. The input was divided into four slots representing the verb, n1, prep, and n2 respectively. In slots n1 and n2, each sense of the corresponding noun was encoded using all the classes within the IS-A branch of the *WordNet'* hierarchy, from the corresponding hierarchy root node to its bottom-most node. In the verb slot, the verb was encoded using the IS_A_WAY_OF branches. There was a unit in the input for each node of the WordNet subset. This unit was **on** if it represented a semantic class to which one of the senses of the word to be encoded belonged. As for the **output**, there were only two units representing whether the PP attached to the verb or not.

The second procedure consists of presenting all the classes of each sense of each word serially. However, the parallel procedure have the advantage that the network can detect which classes are related with which ones in the same slot and between slots. We observed this advantage in preliminary studies.

Feedforward networks with one hidden layer and

Table 1: WordNet information for the noun 'bank'.

Sense 1	*group* → *people* → *organization* → *institution* → *financial_institut.*
Sense 2	*entity* → *object* → *artifact* → *facility* → *depository*
Sense 3	*entity* → *object* → *natural_object* → *geological_formation* → *slope*

Table 2: Test size and accuracy results reported in previous works. 'W' denotes words only, 'C' class only and 'W+C' words+classes.

Author	W	C	W+C	Classes	Test size
Hindle and Rooth (93)	80	-	-	-	880
Resnik and Hearst(93)	81.6	79.3	83.9	WordNet	172
Resnik and Hearst (93)	-	-	75[a]	WordNet	500
Ratnaparkhi et al. (94)	81.2	79.1	81.6	MIC	3097
Brill and Resnik (94)	80.8	-	81.8	WordNet	500
Collins and Brooks (95)	84.5	-	-	-	3097
Li and Abe (95)	-	85.8[b]	84.9	WordNet	172

[a]Accuracy obtained by Brill and Resnik (94) using Resnik's method on a larger test.
[b]This accuracy is based on 66% coverage.

a full interconnectivity between layers were used in all the experiments. The networks were trained with backpropagation learning algorithm. The activation function was the logistic function. The number of hidden units ranged from 70 to 150. This network was used for solving our classification problem: attached to noun or attached to verb. The output activation of this network represented the bayesian posterior probability that the PP of the encoded sentence attaches to the verb or not (Richard and Lippmann (1991)).

5 Training and Experimental Results.

21418 examples of structures of the kind 'VB N1 PREP N2' were extracted from the Penn-TreeBank Wall Street Journal (Marcus et al. 1993). WordNet did not cover 100% of this material. Proper names of people were substituted by the WordNet class *someone*, company names by the class *business_organization*, and prefixed nouns for their stem (co-chairman → chairman). 788 4-tuples were discarded because of some of their words were not in WordNet and could not be substituted. 20630 codified patterns were finally obtained: 12016 (58.25%) with the PP attached to N1, and 8614 (41.75%) to VB.

We used the cross-validation method as a measure of a correct generalization. After encoding, the 20630 patterns were divided into three subsets: training set (18630 patterns), set A (1000 patterns), and set B (1000 patterns). This method evaluated performance (the number of attachment errors) on a pattern set (validation set) after each complete pass through the training data (epoch). Series of three runs were performed that systematically varied the random starting weights. In each run the networks were trained for 40 epochs. In each run the weights of the epoch having the smallest error with respect to the validation set were stored. The weights corresponding to the best result obtained on the validation test in the three runs were selected and used to evaluate the performance in the test set. First, we used set A as validation set and set B as test, and afterwards we used set B as validation and set A as test. This experiment was replicated with two new partitions of the pattern set: two new training sets (18630 patterns) and 4 new validation/test sets of 1000 patterns each.

Results showed in table 3 are the average accuracy over the six test sets (1000 patterns each) used. We performed three series of runs that varied the input encoding. In all these encodings, three tree cut thresholds were used: *10%, 6% and 2%*. The number of semantic classes in the input encoding ranged from 139 (10% cut) to 475 (2%) In the first encoding, the 4-tuple without extra information was used. The results for this case are shown in the 4-tuple column entry of table 3. In the second encoding, we added the prepositions the verbs select for their internal arguments, since English verbs with semantic similarity could select different prepositions (for example, *accuse* and *blame*). Verbs can be classified on the basis of the kind of prepositions they select. Adding this classification to the *WordNet'* classes in the input encoding improved the results

(4-tuple$^+$ column entry of table 3).

The 2% cut results were significantly better ($p < 0.02$) than those of the 6% cut for 4-tuple and 4-tuple$^+$ encodings. Also, the results for the 4-tuple$^+$ condition were significanly better ($p < 0.01$).

For all simulations the momentum was 0.8, initial weight range 0.1. No exhaustive parameter exploration was carried out, so the results can still be improved.

Some of the errors committed by the network can be attributed to an inadequate class assignment by WordNet. For instance, names of countries have only one sense, that of *location*. This sense is not appropriate in sentences like: *Italy increased its sales to Spain*; *locations* do not sell or buy anything, and the correct sense is *social_group*. Other mistakes come from what are known as reporting and aspectual verbs. For example in expressions like *reported injuries to employees* or *initiated talks with the Soviets* the n1 has an argumental structure, and it is the element that imposes selectional restrictions on the PP. There is no good classification for these kinds of verbs in WordNet. Finally, collocations or idioms, which are very frequent, (e.g. *take a look, pay attention*), are not considered lexical units in the WSJ corpus. Their idiosyncratic behaviour introduces noise in the selectional restrictions acquisition process. Word-based models offer a clear advantage over class-based methods in these cases.

6 Discussion

When sentences with PP attachment ambiguities were presented to two human expert judges the mean accuracy obtained was 93.2% using the whole sentence and 88.2% using only the 4-tuple (Ratnaparkhi et al., 1994). Our best result is 86.8%. This accuracy is close to human performance using the 4-tuple alone. Collins and Brooks (1995) reported an accuracy of 84.5% using words alone, a better score than those obtained with other methods tested on the WSJ corpus. We used the same corpus as Collins and Brooks (WSJ) and a similar sized training set. They used a test set size of 3097 patterns, whereas we used 6000. Due to this size, the differences between both results (84.5% and 86.81%) were probably significant. Note that our results were obtained using only class information. Ratnaparkhi et al. (1994)'s results are the best reported so far using only classes (for 100% coverage): 79.1%. From these results we can conclude that improvements in the syntactic disambiguation problem will come not only from the availability of better hierarchies of classes but also from methods that use them better. NNs seem especially well designed to use them effectively.

How do we account for the improved results? First, we used verb class information. Given the set of words in the 4-tuple and a way to represent senses and semantic class information, a syntactic disambiguation system (SDS) must find some regularities between the co-occurrence of classes and the attachment point. Presenting all of the classes of all the senses of the complete 4-tuple simultaneously, assuming that the training set is adequate, the network can detect which classes (and consequently which senses) are related with which others. As we have said, due to its complexity, current methods do not consider the complete 4-tuple simultaneously. For example, Li and Abe (1995) use $p(verb\ attach|v\ prep\ n2)$ or $p(verb\ attach|v\ n1\ prep))$. The task of selecting which of the senses contributes to making the correct attachment could be difficult if the whole 4-tuple is not simultaneously present. A verb has many senses, and each one could have a different argumental structure. In the selection of the correct sense of the verb, the role of the object ($n1$) is very important. Deciding the attachment site by computing $p(verb\ attach|v\ prep\ n2)$ would be inadequate. It is also inadequate to omit $n2$. Rule based approaches also come up against this problem. In Brill and Resnik (1994), for instance, for reasons of run-time efficiency and complexity, rules regarding the classes of both $n1$ and $n2$ were not permitted. Using a parallel presentation it is also possible to detect complex interactions between the classes of a particular sense (for example, exceptions) or the classes of different senses that cannot be detected in the case of current statistical methods. We have detected these interactions in studies on word sense disambiguation we are currently carrying out. For example, the behavior of verbs which have the senses of *process* and *state* differs from that of verbs which have the sense of *process* but not of *state*, and viceversa.

A parallel presentation (of classes as well of senses) gives rise to a highly complex input. A very important characteristic of neural networks is their capability of dealing with multidimensional inputs (Barron, 1993). They can compute very complex statistical functions and they are model free. Compared to the current methods used by the statistical or rule-based approaches to natural language processing, NNs offer the possibility of dealing with a much more complex approach (non-linear and high dimensional).

References.

Barron, A. (1993). Universal Approximation Bounds for Superposition of a Sigmoidal Function. *IEEE Transactions on Information Theory*, 39:930–945.

Brill, E. & Resnik, P. (1994). A Rule-Based Approach to Prepositional Phrase Attachment Disambiguation. In *Proceedings of the Fifteenth International Conferences on Computational Linguistics (COLING-94)*.

Collins, M. & Brooks, J. (1995). Prepositional Phrase

Table 3: Accuracy results for different input encoding and tree cuts.

Cut	4-tuple	4-tuple$^+$
10%	83.17 ±0.9	85.15 ±0.8
6%	84.07 ±0.7	85.32 ±0.9
2%	85.12 ±1.0	86.81 ±0.9

attachment. In *Proceedings of the 3rd Workshop on Very Large Corpora.*

Hindle, D. & Rooth, M. (1993). Structural Ambiguity and Lexical Relations. *Computational Linguistics*, 19:103–120.

Katz, J. & Fodor, J. (1963). The Structure of Semantic Theory. *Language, 39*: 170-210.

Li, H. & Abe, N. (1995). Generalizing Case Frames using a Thesaurus and the MDL Principle. In *Proceedings of the International Workshop on Parsing Technology.*

Marcus, M., Santorini, B. & Marcinkiewicz, M. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330.

Miller, G., Beckwith, R., Felbaum, C., Gross, D. & Miller, K. (1993). Introduction to WordNet: An Online Lexical Database. Anonymous FTP, internet: clarity.princeton.edu.

Ratnaparkhi, A., Reynar, J. & Roukos, S. (1994). A Maximum Entropy Model for Prepositional Phrase Attachment. In *Proceedings of the ARPA Workshop on Human Language Technology.*

Resnik, P. & Hearst, M. (1993). Syntactic Ambiguity and Conceptual Relations. In *Proceedings of the ACL Workshop on Very Large Corpora.*