Exploiting the WWW as a corpus to resolve PP attachment ambiguities

Martin Volk
University of Zurich
Department of Computer Science, Computational Linguistics Group
Winterthurerstr. 190, CH-8057 Zurich
volk@ifi.unizh.ch

1. Introduction

Finding the correct attachment site for prepositional phrases (PPs) is one of the hardest problems when parsing natural languages. An English sentence consisting of a subject, a verb, and a nominal object followed by a prepositional phrase is a priori ambiguous. The PP in sentence 1 is a noun attribute and needs to be attached to the noun, but the PP in 2 is an adverbial and thus part of the verb phrase.

(1) Peter reads a book *about computers*.

(2) Peter reads a book *in the subway*.

If the subcategorisation requirements of the verb or the competing noun are known the ambiguity can sometimes be resolved. But many times there are no clear requirements. Therefore, there has been a growing interest in using statistical methods that reflect attachment tendencies.

This new line of research was kicked off by Hindle and Rooth (1993). They tackled the PP-attachment ambiguity problem (for English) by computing *lexical association scores* over a partially parsed corpus. If a sentence contains the sequence V+NP+PP the triple V+N+P is observed with N being the head noun of the NP and P being the head of the PP. The probabilities are estimated from co-occurrence counts of V+N and of N+P. They evaluated their method on manually disambiguated verb-noun-preposition triples. It resulted in 80% correct attachments.

In the meantime the method has been improved and extended. The best reported results are from Stetina and Nagao (1997: up to 88% correct attachment). They use a supervised learning approach (they train the disambiguator over the Penn-Treebank) and a semantic dictionary to cluster the words.

We applied unsupervised statistical methods to German. Since there is no large German treebank available we first worked with a partially parsed corpus. The gathering of co-occurrence data is more complicated for German because of its variable constituent ordering. In Langer et al. (1997) we show that we can achieve around 76% attachment accuracy for the decidable cases. But many cases cannot be decided because of sparse data.

Therefore we have experimented with using the WWW, a corpus that is orders of magnitude larger than our locally accessible corpora. With the help of a WWW search engine we obtain frequency values ("number of pages found"). In querying a search engine we lose some precision compared to corpus analysis. Our hypothesis is that the size of the WWW will compensate our rough queries.

Our method for determining co-occurrence values is based on a simple formula. We use the frequency of a word co-occurring with a given preposition against the overall frequency of this word. For example, if some noun N occurs 100 times in a corpus and this noun co-occurs with the preposition P 60 times then the co-occurrence value of N+P will be 60/100 = 0.6. The general formula is (where X can be either a noun N or a verb V):

$$freq(X,P) / freq(X) = cooc(X,P)$$

In Volk (2000) we have explored this formula in detail. We have shown that the WWW frequencies can be used for the resolution of PP attachment ambiguities if the difference between the competing co-occurrence values is above a certain threshold. In this way the co-occurrence values served to decide 58% of our test cases with an attachment accuracy of 75%.

In the more successful experiments for PP attachment in English (Stetina and Nagao 1997, Collins and Brooks 1995) the co-occurrence statistics included the noun within the PP. The motivation behind this becomes immediately clear if we compare the PPs in the example sentences 3 and 4. Since both PPs start with the same preposition only the noun within the PP helps to find the correct attachment.

(3) Peter saw the thief *with his own eyes*.

(4) Peter saw the thief *with the red coat*.

In a new round of experiments we have included the head noun of the PP into the queries. This means we are now working with the extended formula:

$$freq(X, P, N2) / freq(X) = cooc(X, P, N2)$$

Let us look at an example sentence from our corpus:

(5) Unisource hat die Voraussetzungen *für die Gründung* eines Betriebsrates geschaffen.
Unisource has set up the prerequisites for the foundation of a work council.

	freq(X,P,N2)	freq(X)	cooc(X,P,N2)
X=N1	freq(Voraussetz.,für,Gründung) 274	freq(Voraussetzungen) 255'010	cooc(Voraussetz.,für,Gründung) 0.001074
X=V	freq(geschaffen,für,Gründung) 139	freq(geschaffen) 172'499	cooc(geschaffen,für,Gründung) 0.000805

The co-occurrence value cooc(N1,P,N2) is higher than cooc(V,P,N2), and thus the model correctly predicts noun attachment for the PP.

2. Preparation of the test corpus

We manually compiled a treebank as a test suite for the evaluation of our method. We semi-automatically disambiguated and annotated 3000 sentences. In order to be compatible with the German NEGRA treebank we used the same annotation scheme as Skut et al. (1997).

We selected our evaluation sentences from the 1996 volume of the *ComputerZeitung*, a weekly computer magazine that is available on CD-ROM (Konradin-Verlag 1998). We tagged the text and selected 3000 sentences that contained

1. at least one full verb and
2. at least one sequence of a noun followed by a preposition.

With these conditions we restricted the sentence set to those sentences that contain a prepositional phrase in an ambiguous position.

Manually assigning a complete syntax tree to a sentence is a labour-intensive task. This task can be facilitated if the most obvious phrases are automatically parsed. We used our chunk parser for NPs and PPs to speed up the manual annotation. We also used the NEGRA Annotate-Tool (Brants et al. 1997) to semi-automatically assign syntax trees to all (preparsed) sentences. This tool comes with a built-in parser that can suggest categories over selected nodes. The sentence structures were judged by two linguists to minimize errors. Finally, completeness and consistency checks were applied to ensure that every constituent was included into the sentence structure.

We then used a Prolog program to build the nested structure and to recursively work through the annotations in order to obtain sextuples with the relevant information for the PP classification task:

1. the full verb (a separated verbal prefix is reattached),
2. the real head noun N1 (the noun which the PP is attached to),
3. the possible head noun N1 (the noun that immediately precedes the PP; this noun leads to the attachment ambiguity),
4. the preposition of the PP,
5. the core noun of the PP (called N2), and
6. the attachment decision (as given by the human annotators).

Let us illustrate this with some example sentences.

(6) Das Dorfmuseum gewährt nicht nur einen Einblick *in den häuslichen Alltag vom Herd bis zum gemachten Bett*.
The village museum allows not only insights into the everyday life from the oven to the bed.

(7) ... nachdem dieses wichtige Feld *seit 1985* brachlag.
... since this important field lay idle since 1985.

(8) Das trifft auf alle Waren *mit dem berüchtigten "Grünen Punkt"* zu.
This holds true for all goods with the ill-famed "Green Dot".

These corpus sentences will lead to the following sextuples:

verb	real N1	possible N1	prep.	N2 (in PP)	function of the PP
gewährt	*Einblick*	*Einblick*	*in*	*Alltag*	postnominal modifier
gewährt	*Alltag*	*Alltag*	*vom*	*Herd*	postnominal modifier
gewährt	*Alltag*	*Herd*	*bis*	*Bett*	postnominal modifier
brachlag	/	*Feld*	*seit*	*1985*	verb modifier
zutrifft	*Waren*	*Waren*	*mit*	*Punkt*	postnominal modifier

Each sextuple represents a PP with the preposition occurring in a position where it can be attached either to the noun or to the verb. Note that the PP *auf alle Waren* in 8 is not in such an ambiguous position and thus does not appear in the test cases.

In sentence 6 we observe the difference between the real head noun and the possible head noun. The PP *bis zum gemachten Bett* is not attached to the possible head noun *Herd* but to the preceding noun *Alltag*. Obviously, there is no real head noun if the PP attaches to the verb (as in 7). In the following tests we use the real reference noun N1 if it is present else the possible reference noun N1.

Our test corpus consists of 4383 test cases, out of which 63% are noun attachments and 37% verb attachments.

3. Disambiguating with WWW frequencies

We queried AltaVista in order to obtain the frequency data for our co-occurrence values. For all queries we use AltaVista advanced search restricted to German documents. For co-occurrence frequencies we use the NEAR operator.

- For nouns and verbs we query for the word form by itself.
- For co-occurrence frequencies we query for `Verb NEAR preposition NEAR N2` and `N1 NEAR preposition NEAR N2` again using the verb forms and noun forms as they appear in the corpus. The NEAR operator in AltaVista restricts the search to documents in which its argument words co-occur within 10 words.

We then compute the co-occurrence values for all cases in which both the word form frequency and the co-occurrence frequency are above zero. We evaluate these co-occurrence values against our test corpus using the following disambiguation algorithm.

```
if (cooc(N1,P,N2) && cooc(V,P,N2)) then
  if (cooc(N1,P,N2) > cooc(V,P,N2)) then
    noun attachment
  else
    verb attachment
else
  noun attachment
```

If both co-occurrence values exist, the attachment decision is based on the higher value. If one or both co-occurrence values are missing we decide in favour of noun attachment since 63% of our test cases are noun attachment cases. The disambiguation result is summarized in table 1.

	correct	incorrect	accuracy
noun attachment	2553	1129	69.34%
verb attachment	495	206	70.61%
total	1800	1335	69.54%

Table 1: Attachment accuracy for the complete test corpus

The attachment accuracy is improved by 6.5% compared to pure guessing. But it is way below the accuracy that we computed for the decidable cases in earlier experiments. Even in the WWW many of our test triples do not occur. Only 2422 (55%) of the 4383 test cases can be decided by using both co-occurrence values. The attachment accuracy for these test cases is 74.32% and thus about 5% higher than when forcing a decision on all cases (cf. table 2)

	correct	incorrect	accuracy
noun attachment	1305	416	75.83%
verb attachment	495	206	70.61%
total	1800	622	74.32%

Table 2: Attachment accuracy when requiring both cooc(N1,P,N2) and cooc(V,P,N2)

3.1. Using the co-occurrence values against a threshold

A way of tackling the sparse data problem lies in using partial information. Instead of insisting on both *cooc(N1,P,N2)* and *cooc(V,P,N2)* values, we can back off to either value for those cases with only one value available. Comparing this value against a given threshold we decide on the attachment. If, for instance, *cooc(N1,P,N2)* is available (but no *cooc(V,P,N2)* value), and if this value is above the threshold then we decide on noun attachment. If *cooc(N1,P,N2)* is below the threshold we take no decision. Thus we extend the disambiguation algorithm as follows:

```
if (cooc(N1,P,N2) && cooc(V,P,N2)) then
  if (cooc(N1,P,N2) > cooc(V,P,N2)) then
    noun attachment
  else
    verb attachment
elseif (cooc(N1,P,N2) > threshold) then
  noun attachment
elseif (cooc(V,P,N2) > threshold) then
  verb attachment
```

Now the problem arises on how to set the thresholds. It is obvious that the attachment decision gets more reliable the higher we set the thresholds. At the same time the number of cases that are decidable decreases. We suggest to set the threshold in such a way that using this partial information is not worse than using both the *cooc(N1,P,N2)* and *cooc(V,P,N2)* values. That means that we set the threshold so that we keep the overall attachment accuracy at around 75%.

	correct	incorrect	accuracy
noun attachment	1448	446	76.45%
verb attachment	629	245	71.97%
total	2077	691	75.04%

Table 3: Attachment accuracy when requiring either cooc(N1,P,N2) or cooc(V,P,N2)

We thus set the threshold to 0.001 and obtain the result in table 3. The attachment rate (the number of decidable cases) has risen from 55% to 63%; 2768 out of 4383 cases can be decided based on either both co-occurrence values or on the comparison of one co-occurrence value against the threshold. Noun attachment is still better than verb attachment.

3.2. Using the co-occurrence values of word forms and base forms

The above frequencies were based on word form counts. But German is a highly inflecting language for verbs, nouns and adjectives. If a rare verb form (e.g. a conjunctive verb form) or a rare noun form

(e.g. a new compound form) appears in the test corpus it often results in a zero frequency for the triple. We may safely assume that the co-occurrence tendency is constant over the different verb forms. We may therefore substitute the rare verb form with a more frequent form of this verb. We decided to query with the given verb form and with the corresponding verb lemma (the infinitive form).

For nouns we also query for the lemma. As a special case we reduce compound nouns to the last compound element and we compute the lemma for the last element (e.g. *Informationssystemen* → *System*). We do the same for hyphenated compounds (e.g. *GI-Kongresses* → *Kongress*). We also reduce company names ending in *GmbH* or *Systemhaus* to these keywords and use them in place for the lemma (e.g. *CSD Software GmbH* → *GmbH*).

The co-occurrence value is thus computed as (X is the verb V or the reference noun N1):

$$\frac{freq(X_{form},P,N2) + freq(X_{lemma},P,N2)}{freq(X_{form}) + freq(X_{lemma})} = cooc(X,P,N2)$$

The disambiguation algorithm is the same as above and we use the same threshold of 0.001. As table 4 shows, the attachment accuracy stays at around 75% but the attachment rate increases from 63% to 71% (3109 out of 4379 test cases can be decided).

	correct	incorrect	accuracy
noun attachment	1615	459	77.87%
verb attachment	735	300	71.01%
total	2350	759	75.59%

Table 4: Attachment accuracy including threshold and lemmas

In order to complete the picture we evaluate without using the threshold. We get an attachment accuracy of 74.72% at an attachment rate of 65%. This a 10% increase to the result we computed for word forms (cf. table 2). If, in addition, we use any single co-occurrence value (i.e. we set the threshold to 0), the attachment accuracy slightly decreases to 74.23% at an attachment rate of 85%. This means that for 85% of our test cases we have at least one co-occurrence value from the WWW frequencies. If we default the remaining cases to noun attachment we end up with an accuracy of 73.08% which is significantly higher than our initial result of 69.54% (reported in table 1).

3.3. Conclusion

The most important lesson from these experiments is that triples (X,P,N2) are much more reliable than tuples (X,P) for deciding the PP attachment site. Using a large corpus such as the WWW helps to obtain frequency values for many triples and thus provides co-occurrence values for most cases.

Furthermore, we have shown that querying for word form and lemma substantially increases the set of decidable cases and thus the attachment rate without any loss in the attachment accuracy. The accuracy is 74% for all decidable test cases and 73% for all test cases. We can further enhance the co-occurrence frequencies by querying for all word forms, as long as the WWW search engines index every word form separately.

If we are interested only in highly reliable disambiguation cases (80% accuracy) we may lower the number of decidable cases by increasing the threshold (or by requiring a minimal distance between *cooc(V,P,N2)* and *cooc(N1,P,N2)* as we have shown for tuples in Volk, 2000).

When using frequencies from the WWW the number of decidable cases should be higher for English since the number of English documents in the WWW by far exceeds the number of German documents. Still the problem remains that querying for co-occurrence frequencies with WWW search engines using the NEAR operator allows only for very rough queries. For instance, the query `P NEAR N2` does not guarantee that the preposition and the noun co-occur within the same PP. It matches even if the noun precedes the preposition. There are various possibilities for improved queries.

1. `X NEAR "P DET N2"` with an appropriate determiner DET will query for the sequence "P DET N2" and thus for P and N2 co-occurring in a standard PP.

2. `X NEAR (P NEXT 3 N2)` will query for N2 as one of the three tokens following P. The NEXT operator is often available in information retrieval systems but not in the WWW search engines that we are aware of. This query is more flexible than querying for a standard PP.
3. `"N1 P" NEXT 3 N2` will query for noun N1 and preposition P immediately following each other as is most often the case if the PP is attached to N1.
4. `V SAME_SENTENCE (P NEXT 3 N2)` will query for the verb V co-occurring within the same sentence as the PP. From a linguistic point of view this is the minimum requirement for the PP being attached to the verb. In fact, to be linguistically precise we must require the verb to co-occur within the same clause as the PP. But none of these operators is available in current search engines.

Obviously, any of these constraints will reduce the frequency counts and may thus lead to sparse data. We will therefore have to counterbalance this with querying for words that behave similarly with respect to PP attachment, for instance, words from the same semantic class.

Acknowledgement

We thank Charlotte Merz for comments and corrections on earlier versions of this paper.

References

Brants T, Skut W, Krenn B 1997 Tagging grammatical functions. In *Proc. of EMNLP-2*, Providence, RI.

Collins M, Brooks J 1995 Prepositional phrase attachment through a backed-off model. In *Proc. of the Third Workshop on very large corpora*.

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

Konradin-Verlag 1998 Computer Zeitung auf CD-Rom. Volltextrecherche aller Artikel der Jahrgänge 1993 bis 1998. Leinfelden-Echterdingen, Konradin-Verlag.

Langer H, Mehl S, Volk M 1997 Hybride NLP-Systeme und das Problem der PP-Anbindung. In Wermter S, Busemann S, Harbusch K (eds), *Berichtsband des Workshops "Hybride konnektionistische, statistische und symbolische Ansätze zur Verarbeitung natürlicher Sprache" auf der 21. Deutschen Jahrestagung für Künstliche Intelligenz, KI-97 (auch erschienen als DFKI-Document D-98-03)*, Freiburg.

Skut W, Krenn B, Brants T, Uszkoreit, H 1997 An annotation scheme for free word order languages. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, DC, pp 88-95.

Stetina J, Nagao M 1997 Corpus based pp attachment ambiguity resolution with a semantic dictionary. In Zhou J, Church K (eds), Proc. of the 5th Workshop on very large corpora, Beijing and Hongkong, pp 66-80.

Volk M 2000 Scaling up. Using the WWW to resolve PP attachment ambiguities. In *Proc. of Konvens-2000. Sprachkommunikation*, Ilmenau, VDE Verlag, pp 151-156.