# Combining unsupervised and supervised methods for PP attachment disambiguation

**Martin Volk**
University of Zurich
Schönberggasse 9
CH-8001 Zurich
vlk@zhwin.ch

## Abstract

Statistical methods for PP attachment fall into two classes according to the training material used: first, unsupervised methods trained on raw text corpora and second, supervised methods trained on manually disambiguated examples. Usually supervised methods win over unsupervised methods with regard to attachment accuracy. But what if only small sets of manually disambiguated material are available? We show that in this case it is advantageous to intertwine unsupervised and supervised methods into one disambiguation algorithm that outperforms both methods used alone.[1]

## 1 Introduction

Recently, numerous statistical methods for prepositional phrase (PP) attachment disambiguation have been proposed. They can broadly be divided into unsupervised and supervised methods. In the unsupervised methods the attachment decision is based on information derived from large corpora of raw text. The text may be automatically processed (e.g. by shallow parsing) but not manually disambiguated. The most prominent unsupervised methods are the Lexical Association score by Hindle and Rooth (1993) and the cooccurrence values by Ratnaparkhi (1998). They resulted in up to 82% correct attachments for a set of around 3000 test cases from the Penn treebank. Pantel and Lin (2000) increased the training corpus, added a collocation database and a thesaurus which improved the accuracy to 84%.

In contrast, the supervised methods are based on information that the program learns from manually disambiguated cases. These cases are usually extracted from a treebank. Supervised methods are as varied as the Back-off approach by Collins and Brooks (1995) and the Transformation-based approach by Brill and Resnik (1994). Back-off scored 84% correct attachments and outperformed the Transformation-based approach (80%). Even better results were reported by Stetina and Nagao (1997) who used the WordNet thesaurus with a supervised learner and achieved 88% accuracy.

All these accuracy figures were reported for English. We have evaluated both unsupervised and supervised methods for PP attachment disambiguation in German. This work was constrained by the availability of only a small German treebank (10,000 sentences). Under this constraint we found that an intertwined combination of using information from unsupervised and supervised learning leads to the best results. We believe that our results are relevant to many languages for which only small treebanks are available.

## 2 Our training resources

We used the NEGRA treebank (Skut et al., 1998) with 10,000 sentences from German newspapers and extracted 4-tuples $(V, N_1, P, N_2)$ whenever a PP with the preposition $P$ and the core noun $N_2$ immediately followed a noun $N_1$ in a clause headed by the verb $V$. For example, the sentence

*In Deutschland ist das Gerät über die Bad Homburger Ergos zu beziehen.*
*[In Germany the appliance may be ordered from Ergos based in Bad Homburg.]*

leads to the 4-tuple (*beziehen, Gerät, über, Ergos*). In this way we obtained 5803 4-tuples with the human judgements about the attachment of the PP (42% verb attachments and 58%

[1]This research was supported by the Swiss National Science Foundation under grant 12-54106.98.

noun attachments). We call this the NEGRA test set.

As raw corpus for unsupervised training we used four annual volumes (around 5.5 million words) of the "Computer-Zeitung" (CZ), a weekly computer science magazine. This corpus was subjected to a number of processing steps: sentence recognition, proper name recognition for persons, companies and geographical locations (cities and countries), part-of-speech tagging, lemmatization, NP/PP chunking, recognition of local and temporal PPs, and finally clause boundary recognition.

3000 sentences of the CZ corpus each containing at least one PP in an ambiguous position were set aside for manual disambiguation. Annotation was done according to the same guidelines as for the NEGRA treebank. From these manually annotated sentences we obtained a second test set (which we call the CZ test set) of 4469 4-tuples from the same domain as our raw training corpus.

## 3 Results for the unsupervised methods

We explored various possibilities to extract PP disambiguation information from the automatically annotated CZ corpus. We first used it to gather frequency data on the cooccurrence of pairs: nouns + prepositions and verbs + prepositions.

The cooccurrence value is the ratio of the bigram frequency count $freq(word, preposition)$ divided by the unigram frequency $freq(word)$. For our purposes $word$ can be the verb $V$ or the reference noun $N_1$. The ratio describes the percentage of the cooccurrence of $word + preposition$ against all occurrences of $word$. It is thus a straightforward association measure for a word pair. The cooccurrence value can be seen as the attachment probability of the preposition based on maximum likelihood estimates. We write:

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

with $W \in \{V, N_1\}$. The cooccurrence values for verb $V$ and noun $N_1$ correspond to the probability estimates in (Ratnaparkhi, 1998) except that Ratnaparkhi includes a back-off to the uniform distribution for the zero denominator case. We will add special precautions for this case in our disambiguation algorithm. The cooccurrence values are also very similar to the probability estimates in (Hindle and Rooth, 1993).

We started by computing the cooccurrence values over word forms for nouns, prepositions, and verbs based on their part-of-speech tags. In order to compute the pair frequencies $freq(N_1, P)$, we search the training corpus for all token pairs in which a noun is immediately followed by a preposition. The treatment of verb + preposition cooccurrences is different from the treatment of N+P pairs since verb and preposition are seldom adjacent to each other in a German sentence. On the contrary, they can be far apart from each other, the only restriction being that they cooccur within the same clause. We use the clause boundary information in our training corpus to enforce this restriction. For computing the cooccurrence values we accept only verbs and nouns with a occurrence frequency of more than 10.

With the N+P and V+P cooccurrence values for word forms we did a first evaluation over the CZ test set with the following simple disambiguation algorithm.

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

We found that we can only decide 57% of the test cases with an accuracy of 71.4% (93.9% correct noun attachments and 55.0% correct verb attachments). This shows a striking imbalance between the noun attachment accuracy and the verb attachment accuracy. Obviously, the cooccurrence values favor verb attachment. The comparison of the verb cooccurrence value and the noun cooccurrence value too often leads to verb attachment, and only the clear cases of noun attachment remain. This points to an inherent imbalance between the cooccurrence values for verbs and nouns. We will flatten out this imbalance with a noun factor.

The noun factor is supposed to strengthen the N+P cooccurrence values and thus to attract more noun attachment decisions. What is the rationale behind the imbalance between noun cooccurrence value and verb cooccurrence value? One influence is certainly the well-known

fact that verbs bind their complements stronger than nouns.

The imbalance between noun cooccurrence values and verb cooccurrence values can be quantified by comparing the overall tendency of nouns to cooccur with a preposition to the overall tendency of verbs to cooccur with a preposition. We compute the overall tendency as the cooccurrence value of all nouns with all prepositions.

$$cooc(all\_N, all\_P) = \frac{\sum_{(N_1,P)} freq(N_1, P)}{\sum_{N_1} freq(N_1)}$$

The computation for the overall verb cooccurrence tendency is analogous. For example, in our training corpus we have found 314,028 N+P pairs (tokens) and 1.72 million noun tokens. This leads to an overall noun cooccurrence value of 0.182. The noun factor ($nf$) is then the ratio of the overall verb cooccurrence tendency divided by the overall noun cooccurrence tendency:

$$nf = \frac{cooc(all\_V, all\_P)}{cooc(all\_N, all\_P)}$$

In our training corpus this leads to a noun factor of $0.774/0.182 = 4.25$. In the disambiguation algorithm we multiply the noun cooccurrence value with this noun factor before comparing the product to the verb cooccurrence value. This move leads to an improvement of the overall attachment accuracy to 81.3% (83.1% correct noun attachments and 76.9% correct verb attachments).

We then went on to increase the attachment coverage, the number of decidable cases, by using lemmas, decompounding (i.e. using only the last component of a noun compound), and proper name classes. These measures increased the coverage from 57% to 86% of the test cases. For the remaining test cases we used a threshold comparison if either of the needed cooccurrence values ($cooc(N_1, P)$ or $cooc(V, P)$) has been computed from our training corpus. This raises the coverage to 90%. While coverage increased, accuracy suffered slightly and at this stage was at 78.3%.

This is a surprising result given the fact that we counted all PPs during the training phases. No disambiguation was attempted so far, we counted ambiguous and non-ambiguous PPs in the same manner. We then added this distinction in the training, counting one point for a PP in a non-ambiguous position and only half a point for an ambiguous PP, in this way splitting the PP's contribution to verb and noun attachment. This move increased the accuracy rate by 2% (to 80.5%).

So far we have used bigram frequencies over word pairs, $(V, P)$ and $(N_1, P)$, to compute the cooccurrence values. Some of the previous research (e.g. (Collins and Brooks, 1995) and (Pantel and Lin, 2000)) has shown that it is advantageous to include the noun from within the PP (called $N_2$) in the calculation. But moving from pair frequencies to triple frequencies will increase the sparse data problem. Therefore we computed the pair frequencies and triple frequencies in parallel and used a cascaded disambiguation algorithm to exploit the triple cooccurrence values and the pair cooccurrence values in sequence.

In analogy to the pair cooccurrence value, the triple cooccurrence value is computed as:

$$cooc(W, P, N_2) = freq(W)/freq(W, P, N_2)$$

with $W \in \{V, N_1\}$. With the triple information $(V, P, N_2)$ we were able to identify support verb units (such as *in Angriff nehmen, unter Beweis stellen*) which are clear cases of verb attachment. We integrated this and the triple cooccurrence values into the disambiguation algorithm in the following manner.

```
if ( support_verb_unit(V,P,N2) )
  then verb attachment
elsif (cooc(N1,P,N2) && cooc(V,P,N2))
  then if ((cooc(N1,P,N2) * nf)
            >= cooc(V,P,N2))
    then noun attachment
    else verb attachment
elsif (cooc(N1,P) && cooc(V,P)) then
  if ((cooc(N1,P) * nf) >= cooc(V,P))
    then noun attachment
    else verb attachment
elsif (cooc(N1,P) > threshold(N))
  then noun attachment
elsif (cooc(V,P) > threshold(V))
  then verb attachment
```

The noun factors for triple comparison and

| | factor | correct | incorrect | accuracy | threshold |
|---|---|---|---|---|---|
| noun attachment | 5.47; 5.97 | 2213 | 424 | 83.92% | 0.020 |
| verb attachment | | 1077 | 314 | 77.43% | 0.109 |
| total | | 3290 | 738 | 81.67% | |
| decidable test cases | | 4028 (of 4469) coverage: 90.13% | | | |

Table 1: Attachment accuracy for the CZ test set using cooccurrence values from unsupervised learning.

| decision level | number | coverage | accuracy |
|---|---|---|---|
| support verb units | 97 | 2.2% | 100.00% |
| triple comparison | 953 | 21.3% | 84.36% |
| pair comparison | 2813 | 62.9% | 79.95% |
| $cooc(N_1, P) >$ threshold | 74 | 1.7% | 85.13% |
| $cooc(V, P) >$ threshold | 91 | 2.0% | 84.61% |
| total | 4028 | 90.1% | 81.67% |

Table 2: Attachment accuracy for the cooc. method split on decision levels.

pair comparison are computed separately. The noun factor for pairs is 5.47 and for triples 5.97.

The attachment accuracy is improved to 81.67% by the integration of the triple cooccurrence values (see table 1). A split on the decision levels reveals that triple comparison is 4.41% better than pair comparison (see table 2).

The 84.36% for triple comparison demonstrates what we can expect if we enlarge our corpus and consequently increase the percentage of test cases that can be disambiguated based on triple cooccurrence values.

The accuracy of 81.67% reported in table 1 is computed over the decidable cases. If we force a default decision (noun attachment) on the remaining cases, the overall accuracy is at 79.14%.

## 4 Results for the supervised methods

One of the most successful supervised methods is the Back-off model as introduced by Collins and Brooks (1995). This model is based on the idea of using the best information available and backing off to the next best level whenever an information level is missing. For the PP attachment task this means using the attachment tendency for the complete quadruple $(V, N_1, P, N_2)$ if the quadruple has been seen in the training data. If not, the algorithm backs off to the attachment tendency of triples. All triples that contain the preposition are considered: $(V, N_1, P)$; $(V, P, N_2)$; $(N_1, P, N_2)$. The triple information is used if any of the triples has been seen in the training data. Else, the algorithm backs off to pairs, then to the preposition alone, and finally to default attachment.

The attachment tendency on each level is computed as the ratio of the relative frequency to the absolute frequency. Lacking a large treebank we had to use our test sets in turn as training data for the supervised learning. In a first experiment we used the NEGRA test set as training material and evaluated against the CZ test set. Both test sets were subjected to the following restrictions to reduce the sparse data problem.

1. Verbs, nouns and contracted prepositions were substituted by their base forms. Compound nouns were substituted by the base form of their last component.
2. Proper names were substituted by their name class tag (PERSON, LOCATION, COMPANY).
3. Pronouns and numbers (in PP complement position) were substituted by a pronoun tag or number tag respectively.

This means we used 5803 NEGRA quadruples with their given attachment decisions as training material for the Back-off model. We then

| | correct | incorrect | accuracy |
|---|---|---|---|
| noun attachment | 2291 | 677 | 77.19% |
| verb attachment | 1015 | 486 | 67.62% |
| total | 3306 | 1163 | 73.98% |
| decidable test cases | 4469 (of 4469) coverage: 100% | | |

Table 3: Attachment accuracy for the CZ test set using supervised learning over the NEGRA test set based on the Back-off method.

| decision level | number | coverage | accuracy |
|---|---|---|---|
| quadruples | 8 | 0.2% | 100.00% |
| triples | 329 | 7.3% | 88.75% |
| pairs | 3040 | 68.0% | 75.66% |
| preposition | 1078 | 24.1% | 64.66% |
| default | 14 | 0.3% | 64.29% |
| total | 4469 | 100.0% | 73.98% |

Table 4: Attachment accuracy for the Back-off method split on decision levels.

applied the Back-off decision algorithm to determine the attachments for the 4469 test cases in the CZ test set. Table 3 shows the results. Due to the default attachment step in the algorithm, the coverage is 100%. The accuracy is close to 74%, with noun attachment accuracy being 10% better than verb attachment.

A closer look reveals that the attachment accuracy for quadruples (100%) and triples (88.7%) is highly reliable (cf. table 4) but only 7.5% of the test cases can be resolved in this way. The overall accuracy is most influenced by the accuracy of the pairs (that account for 68% of all attachments with an accuracy of 75.66%) and by the attachment tendency of the preposition alone which resolves 24.1% of the test cases but results in a low accuracy of 64.66%.

We suspected that the size of the training corpus has a strong impact on the disambiguation quality. Since we did not have access to any larger treebank for German, we used cross validation on the CZ test set in a third experiment. We evenly divided this test corpus in 5 parts of 894 test cases each. We added 4 of these parts to the NEGRA test set as training material. The training material thus consists of 5803 quadruples from the NEGRA test set plus 3576 quadruples from the CZ test set. We then evaluated against the remaining part of 894 test cases. We repeated this 5 times with the different parts of the CZ test set and summed up the correct and incorrect attachment decisions.

The result from cross validation is 5% better than using the NEGRA corpus alone as training material. This could be due to the enlarged training set or to the domain overlap of the test set with part of the training set. We therefore did another cross validation experiment taking only the 4 parts of the CZ test set as training material. If the improved accuracy were a result of the increased corpus size, we would expect a worse accuracy for this small training set. But in fact, training with this small set resulted in around 77% attachment accuracy. This is better than training on the NEGRA test set alone. This indicates that the domain overlap is the most influential factor.

## 5 Intertwining unsupervised and supervised methods

Now, that we have seen the advantages of the supervised approaches, but lack a sufficiently large treebank for training, we suggest combining the unsupervised and supervised information. With the experiments on cooccurrence values and the Back-off method we have worked out the quality of the various decision levels within these approaches, and we will now order the decision levels according to the reliability of the information sources.

We reuse the triple and pair cooccurrence values that we have computed for the experiments

with our unsupervised method. That means that we will also reuse the respective noun factors and thresholds. In addition, we use the NEGRA test set as supervised training corpus for the Back-off method.

The disambiguation algorithm will now work in the following manner. It starts off with the support verb units as level 1, since they are known to be very reliable. As long as no attachment decision is taken, the algorithm proceeds to the next level. Next is the application of supervised quadruples (level 2), followed by supervised triples (level 3). In section 4 we had seen that there is a wide gap between the accuracy of supervised triples and pairs. We fill this gap by accessing unsupervised information, i.e. triple cooccurrence values followed by pair cooccurrence values (level 4 and 5). Even threshold comparisons based on one cooccurrence value are usually more reliable than supervised pairs and therefore constitute levels 6 and 7. If still no decision has been reached, the algorithm continues with supervised pair probabilities followed by pure preposition probabilities. The left-over cases are handled by default attachment. Below is the complete disambiguation algorithm in pseudo-code:

```
if ( support_verb_unit(V,P,N2) )
  then  verb attachment
### level 2 ###
elsif ( supervised(V,N1,P,N2) ) then
  if ( prob(noun_att | V,N1,P,N2) >= 0.5)
    then  noun attachment
  else  verb attachment
### level 3 ###
elsif ( supervised(triple) ) then
  if ( prob(noun_att | triple) >= 0.5 )
    then  noun attachment
  else  verb attachment
### level 4 ###
elsif ( cooc(N1,P,N2) && cooc(V,P,N2) )
  then
  if ((cooc(N1,P,N2)*nf) >= cooc(V,P,N2))
    then  noun attachment
  else  verb attachment
### level 5 ###
elsif ( cooc(N1,P) && cooc(V,P) ) then
  if ((cooc(N1,P) * nf) >= cooc(V,P))
    then  noun attachment
  else  verb attachment
### levels 6 / 7 ###
elsif ( cooc(N1,P) > threshold(N) )
  then  noun attachment
elsif ( cooc(V,P) > threshold(V) )
  then  verb attachment
### level 8 ###
elsif ( supervised(pair) ) then
  if ( prob(noun_attach | pair) >= 0.5)
    then  noun attachment
  else  verb attachment
### level 9 ###
elsif ( supervised(P) ) then
  if ( prob(noun_attach | P) >= 0.5 )
    then  noun attachment
  else  verb attachment
### level 10 ###
else default  verb attachment
```

And indeed, this combination of unsupervised and supervised information leads to an improved attachment accuracy. For complete coverage we get an accuracy of 80.98% (cf. table 5). This compares favorably to the accuracy of the cooccurrence experiments plus default attachment (79.14%) reported in section 3 and to the Back-off results (73.98%) reported in table 3. We obviously succeeded in combining the best of both worlds into an improved behavior of the disambiguation algorithm.

The decision levels in table 6 reveal that the bulk of the attachment decisions still rests with the cooccurrence values, mostly pair value comparisons (59.9%) and triple value comparisons (18.9%). But the high accuracy of the supervised triples and, equally important, the graceful degradation in stepping from threshold comparison to supervised pairs (resolving 202 test cases with 75.74% accuracy) help to improve the overall attachment accuracy.

We also checked whether the combination of unsupervised and supervised approaches leads to an improvement for the NEGRA test set. We exchanged the corpus for the supervised training (now the CZ test set) and evaluated over the NEGRA test set. This results in an accuracy of 71.95% compared to 68.29% for pure application of the supervised Back-off method. That means, the combination leads to an improvement of 3.66% in accuracy.

## 6 Conclusions

We have shown that unsupervised approaches to PP attachment disambiguation are about as

| | factor | correct | incorrect | accuracy | threshold |
|---|---|---|---|---|---|
| noun attachment | 5.47; 5.97 | 2400 | 469 | 83.65% | 0.020 |
| verb attachment | | 1219 | 381 | 76.19% | 0.109 |
| total | | 3619 | 850 | 80.98% | |
| decidable test cases | | 4469 (of 4469) coverage: 100% | | | |

Table 5: Attachment accuracy for the combination of Back-off and cooccurrence values for the CZ test set (based on training over the NEGRA test set).

| decision level | | number | coverage | accuracy |
|---|---|---|---|---|
| 1 | support verb units | 97 | 2.2% | 100.00% |
| 2 | supervised quadruples | 6 | 0.1% | 100.00% |
| 3 | supervised triples | 269 | 6.0% | 86.62% |
| 4 | cooccurrence triples | 845 | 18.9% | 84.97% |
| 5 | cooccurrence pairs | 2677 | 59.9% | 80.39% |
| 6 | $cooc(N_1, P) >$ threshold | 71 | 1.6% | 85.51% |
| 7 | $cooc(V, P) >$ threshold | 81 | 1.8% | 82.72% |
| 8 | supervised pairs | 202 | 4.5% | 75.74% |
| 9 | supervised prepositions | 210 | 4.7% | 60.48% |
| 10 | default | 11 | 0.3% | 54.55% |
| total | | 4469 | 100.0% | 80.98% |

Table 6: Attachment accuracy split on decision levels for the combination of Back-off and cooccurrence values.

good as supervised approaches over small manually disambiguated training sets. If only small manually disambiguated training sets are available, the intertwined combination of unsupervised and supervised information sources leads to the best results.

In another vein of this research we have demonstrated that cooccurrence frequencies obtained through WWW search engines are useful for PP attachment disambiguation (Volk, 2001). In the future we want to determine at which decision level such frequencies could be integrated.

## References


E. Brill and P. Resnik. 1994. A rule-based approach to prepositional phrase attachment disambiguation. In *Proceedings of COLING*, pages 1198–1204, Kyoto. ACL.

M. Collins and J. Brooks. 1995. Prepositional phrase attachment through a backed-off model. In *Proc. of the Third Workshop on Very Large Corpora*.

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

P. Pantel and D. Lin. 2000. An unsupervised approach to prepositional phrase attachment using contextually similar words. In *Proc. of ACL-2000*, Hongkong.

Adwait Ratnaparkhi. 1998. Statistical models for unsupervised prepositional phrase attachment. In *Proceedings of COLING-ACL-98*, Montreal.

W. Skut, T. Brants, B. Krenn, and H. Uszkoreit. 1998. A linguistically interpreted corpus of German newspaper text. In *Proc. of ESSLLI-98 Workshop on Recent Advances in Corpus Annotation*, Saarbrücken.

J. Stetina and M. Nagao. 1997. Corpus based PP attachment ambiguity resolution with a semantic dictionary. In J. Zhou and K. Church, editors, *Proc. of the 5th Workshop on Very Large Corpora*, Beijing and Hongkong.

Martin Volk. 2001. Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In *Proc. of Corpus Linguistics 2001*, Lancaster, March.