Scaling up: using the WWW to resolve PP attachment ambiguities

Volk, Martin

Scaling up. Using the WWW 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

Abstract

We have developed a method to resolve ambiguities in prepositional phrase (PP) attachment in German. We measure on the one hand the cooccurrence strength between nouns (N) and prepositions (P) and on the other hand between verbs (V) and prepositions. The competing values of N+P versus V+P are used to decide whether to attach a prepositional phrase to the noun or to the verb. We calculate the cooccurrence strengths from frequencies given by a WWW search engine. We show that our method is easy to use and assigns a confidence level to the attachments depending on the distance between the N+P and the V+P cooccurrence value.

1 Introduction

In a first approach these cooccurrence values were derived from text corpora [Mehl et al. 98]. There, we achieved a rate of around 75% correct attachments for a domain specific Computer Magazine corpus (2.7 million tokens). But many of our cooccurrence values were based on rather low frequencies. We saw that attachment precision increases if these low frequencies are excluded from the tests. But then the number of attachment ambiguities that can be resolved decreases drastically.

In this paper we investigate a corpus that is many orders of magnitude larger than our Computer Magazine corpus, we compute the cooccurrence values from frequencies in the world wide web (WWW). Some search engines such as AltaVista (www.altavista.com) provide a frequency ('number of pages found') for every query. We owe this idea to [Grefenstette 99] who has shown that WWW frequencies can be used to find the correct translation of German compounds if the possible translations of their parts are known.

We show that cooccurrence values based on such WWW frequencies can easily be obtained and that they provide reliable attachment heuristics for a subclass of attachment ambiguities. We evaluated the attachment decisions against the NEGRA treebank [Skut et al. 98].

2 Computing the cooccurrence values

Our method for determining cooccurrence values is based on using the overall frequency of a word against the frequency of that word cooccurring with a given preposition. For example, if some noun N occurs 100 times in a corpus and this noun cooccurs with the preposition P 60 times then the cooccurrence value of N+P will be 60/100 = 0.6. The general formula:

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

where X can be either a noun N or a verb V.

We interpret the notion of cooccurrence from a linguistic point of view. A PP can only attach to a noun if it follows that noun. In most cases the PP will immediately follow the noun. Therefore, in the past, we used the frequency of N immediately followed by P to compute the cooccurrence value of N+P. For verbs the issue is more complicated. A PP attached to a (full) verb can occur anywhere in the same clause as the verb. We therefore used a clause boundary detector to segment sentences into one-verb clauses. The frequency of a verb V cooccurring with a preposition P within the same clause was used to compute the cooccurrence value of V+P. This approach had the disadvantage that it contains a bias towards verb attachment. For a verb all possible prepositions are taken into account whereas for a noun only the most likely preposition is considered. We needed a noun factor to work against this bias. With this kind of corpus analysis one can achieve around 75% correct attachments.

When using the AltaVista frequencies from the WWW we cannot restrict cooccurrence of N+P and V+P as precisely as when using a locally accessible corpus. Our hypothesis is that the size of the WWW will compensate our rough queries.

For all queries we used AltaVista's Advanced Search facility restricted to German documents. In a first set of experiments we assumed that all forms of a noun

(or of a verb) behave in the same way towards prepositions and we therefore query the web only for the base forms.

- For nouns we used the nominative singular form in the queries.
- For verbs we used the infinitive form in the queries.
- For cooccurrence frequencies we queried for 'noun NEAR preposition' and 'verb NEAR preposition'. The NEAR operator in AltaVista restricts the search to cases where the two words cooccur within 10 words.

For example, Altavista provided the following frequencies which led to the cooccurrence values in column 5.

noun	prep.	freq (N+P)	freq(N)	cooc (N+P)
Umgang	mit	218'025	218'948	0.9957
Einblick	in	83'554	89'294	0.9357
…				
Feld	seit	1076	179'407	0.0060
Sekunde	hinter	214	52'675	0.0045

In this way we computed frequencies for all N+P pairs and V+P pairs occuring in our test corpus.

3 Evaluating against the NEGRA treebank

In 1999 the NEGRA treebank [Skut 98] was made available. It contains 10'000 manually annotated sentences for German. In this treebank every PP is annotated with one of the following functions:

- 'postnominal modifier' or 'pseudo-genitive'. We count these as noun attachments.
- 'modifier' (of a verb) or 'passivised subject'. We count these as verb attachments.
- seldom: some other function such as 'comparative complement' or 'measure argument of adjective'. We disregard these functions.

We converted the sentences from NEGRA's export format into a Prolog format. We then used a Prolog program to work through the nested annotations in order to obtain quadruples with the relevant word forms. The two corpus sentences

> *Das Dorfmuseum gewährt nicht nur einen Einblick in den häuslichen Alltag …*
>
> *… nachdem dieses wichtige Feld seit 1985 brachlag.*

will lead to the following quadruples:

verb form	noun form	prep.	function of PP
gewährt	*Einblick*	*in*	postnominal modifier
brachlag	*Feld*	*seit*	verb modifier

Every quadruple represents a PP starting with the preposition P occuring in a position where it can either be attached to the noun or to the verb. The function field contains the attachment decision given by the NEGRA annotators. From the complete NEGRA corpus we obtain 5266 such quadruples (2237 with a verb attachment and 3029 with a noun attachment). The two above examples are correctly disambiguated on the basis of our cooccurrence values by following the higher value.

verb form	noun form	prep.	cooc(X+P)
	Einblick	*in*	**0.93**
gewährt		*in*	0.35
	Feld	*seit*	0.006
brachlag		*seit*	**0.11**

3.1 Evaluation results for lemmas

As in our corpus studies we used the Gertwol system to lemmatize the verb forms and the noun forms. The lemma of a compound noun is the lemma of its last constituent (*Sprengstoffanschläge* → *Anschlag*). Contracted prepositions are reduced to their base forms (*im* → *in, zur* → *zu*). Pronominal adverbs are reduced to their preposition stem (*darunter* → *unter, dazu* → *zu*). If Gertwol provides more than one lemma for a given wordform we use our filter to determine the correct lemma [Volk 99].

If a lemma cannot be determined (e.g. if a word form is unknown to Gertwol, as is often the case for proper names) the word form itself is assumed to be the lemma. The 5266 lemmatized quadruples constitute the basis for our first evaluation.

The cooccurrence values will be applied in two steps. First, if both cooc(N+P) and cooc(V+P) are available, the higher value decides the attachment. Second, for the few cases (around 3%) where one of the values is not available (e.g. the word was not found by AltaVista), the other value is compared to a threshold. If cooc(X+P) is above that threshold, the attachment is decided in favor of X. If the value is below, then no

attachment is made. Obviously, if both values are unknown the attachment cannot be decided.

If we select 0.1 as the minimum cooccurrence threshold we reach a decision for 99% of the attachment ambiguities from the NEGRA treebank. Out of the 5266 cases only 48 cannot be decided. That is, by using the WWW the attachment rate is almost complete. Unfortunately, these settings result in only 68% correct attachments. Although this attachment quality is much better than pure guessing (50% chance) it is way below the values we reached with domain specific corpus processing (around 75% correct attachments).

3.1.1 Cooccurrence value above threshold

Therefore we need to find an appropriate subset of the attachment cases where the attachment quality is at least equal to that of our corpus studies. We observe that high cooccurrence values are strong indications of a specific attachment. If, for instance, we require either cooc(N+P) or cooc(V+P) to be above a cooccurrence threshold of 0.5 we reach 82% correct attachments. But then the attachment rate drops down to 7.7%. If we reduce the threshold to 0.3 we reach 75% correct attachments with an attachment rate of 36.7%. This means that we can solve a little over a third of all attachments with about the same attachment quality as when doing corpus analysis.

3.1.2 Minimal distance between cooccurrence values

As an alternative to a minimal cooccurrence threshold we introduced a minimal distance between cooc(N+P) and cooc(V+P). It is obvious that an attachment decision is better founded the larger this distance. With a distance value of 0.07 we again reached 75% correct attachments. But now the attachment rate is at 50%. So the minimal distance is superior to the minimal cooccurrence threshold in that it allows to resolve half of the ambiguous cases with an attachment rate comparable to detailed corpus analysis.

3.2 Evaluation results for full forms

In the first evaluation we had lemmatized all noun and verb forms as we usually do when we compute cooccurrence values over our local corpora. The intention is to reduce the number of values to be computed by folding every word form to its lemma. We also reduce the sparse data problem in this way.

Obviously the lemmatization introduces a number of potential errors. First, some word forms are ambiguous towards their lemma (e.g. *rasten* can be a form of either *rasen* - to race - or *rasten* - to rest). When filtering for the correct lemma we may pick the wrong one.[1]

Second, different word forms of a lemma may behave differently with respect to a given preposition. For instance, the noun *Zusammenarbeit* has a high rate of cooccurrence with the preposition *mit* since they often cooccur in the idiomatic phrase *in Zusammenarbeit mit* (English: in collaboration with). But the plural form *Zusammenarbeiten* cannot be used in the idiomatic phrase and therefore shows much lower values.

noun	prep	freq (N+P)	freq(N)	cooc (N+P)
Zusammenarbeit	*mit*	286’441	390’199	0.73
Zusammenarbeiten	*mit*	799	2’398	0.33

In addition, the goal of reducing the sparse data problem by using lemmas rather than word forms cannot be achieved with AltaVista searches since AltaVista does not use a lemmatized index but rather full forms. And it is by no means clear that the lemma is the most frequently used form. The following table shows the AltaVista frequencies for the most important forms of the verbs *denken* (to think) and *zeigen* (to show).

	denken	*zeigen*
1st sg. present / imperative sg.	denke 107’348	zeige 42’224
2nd sg. present	denkst 17’496	zeigst 2’315
3rd sg. present / 2nd pl. present / imperative pl.	denkt 101’486	zeigt **446’642**
1st and 3rd pl. present / infinitive	denken **228’928**	zeigen 366’287
past participle	gedacht 150’153	gezeigt 192’543

For *denken* the frequency is highest for the infinitive form but for *zeigen* the frequency of the 3rd singular form (which also functions as 2nd plural and imperative plural form) is higher than for the infinitive.

[1] Note, however, that some word forms might have homonyns that spoil the frequency values whereas their lemma is unambiguous. As an example think of the English verb form *saw* with its noun homonym whereas searching the lemma *see* does not suffer from such interference.

Therefore we ran a second evaluation querying AltaVista with the full forms as they appear in the NEGRA corpus. Only two small modifications were taken over from our first set of experiments. In the case of hyphenated compounds we use only the last component (*Berlin-Umzug* → *Umzug*). And a separated prefix of a separable prefix verb is attached (*deutete … an* → *andeutete*) since the prefixed verb often behaves differently from its non-prefixed mother.

Now we are querying AltaVista with 3840 noun forms and 2335 verb forms (previously: 2482 noun lemmas and 1491 verb lemmas). We also query with 5209 N+P pairs and 4967 V+P pairs (previously: 4374 noun lemma + preposition and 3773 verb lemma + preposition pairs). This means we save the lemmatization step but we need 30% more queries to the WWW search engine.

3.2.1 Cooccurrence value above threshold

If we again require either cooccurrence value to be above a certain threshold the picture is better than for the lemmatized experiment with a high cooccurrence threshold. With a threshold of 0.5 we reach 86% correct attachments (formerly 82%) and an attachment rate of 8.2% (formerly 7.7%). But if again we choose to compare at a rate of 75% correct attachments we have to lower the threshold to 0.3 and end up with an attachment rate of only 32.5% (formerly 36.7%). So, there is no real gain in using full forms with a threshold. Again, the result is more positive for the minimal distance method.

3.2.2 Minimal distance between cooccurrence values

With the same settings as above (i.e. a minimal cooccurrence distance of 0.7) we achieve 77% correct attachments at an attachment rate of 48.6%. If we adjust our evaluation at 75% correct attachments we increase the attachment rate to 58% (at a minimal cooccurrence distance of 0.5). This means that querying the WWW with full forms (i.e. without lemmatization) leads to better results than querying with lemmas.

This means also that we can assign a confidence level to an attachment based on the distance between cooc(N+P) and cooc(V+P). For some relevant values this is:

cooc distance	confidence
0.04	74%
0.06	76%
0.08	78%
0.10	79%
0.12	80%
0.14	81%

This means, for instance, that if we rely on N+P and V+P cooccurrence values from the WWW that are at least 0.04 apart from each other, our decision will be correct in 74% of the cases. If the cooccurrence distance increases, so will the confidence in the decision.

4 Conclusions

We have shown that frequency values easily obtainable from WWW search engines can be used to resolve PP attachment ambiguities. We use the frequencies to compute cooccurrence values and then we apply the competing cooccurrence values of N+P and V+P to decide the attachments. We have evaluated the method with and without lemmatization against 5266 ambiguous PP cases from the NEGRA treebank. It turned out that using full forms leads to better results (75% correct attachments for a subset covering 58% of the test cases).

The sparse data problem almost disappears when using the WWW as a linguistic resource for this type of attachment problem. From our corpus only 3% of all noun form tokens and less than 1% of all verb form tokens get a frequency of less than 10 from AltaVista (1.5% of nouns and 0.3% of verbs are unknown).

Some recent studies report better attachment quality, but compared to our wide coverage (no limitation on verbs, nouns and prepositions) those results were computed under 'clinical' conditions. [Hartrumpf 99] reports on 81.7% to 91.7% correct attachments by combining statistical information and interpretation rules. But these rules work only for 6 prepositions. [Stetina and Nagao 97] achieve up to 88% correct attachment for English. But they are training over the Penn-Treebank and they are using a semantic dictionary to cluster the words. Our method is easier to use and widely applicable.

Our method might be extended to include the head noun within the PP. Including the head noun has been shown to have positive effects on the attachment quality [Collins and Brooks 95]. This entails moving from a quadruple to a quintuple (V, N, P, head-N, PP-function) which often posed sparse data problems. With the help of the WWW we hope to overcome these problems.

In the future we will also look into combining detailed corpus analysis and WWW frequencies to get the precision of the former and the wide coverage of the latter.

Acknowledgement
We wish to thank Simon Clematide for helpful comments on earlier versions of this paper.

5 References

[Collins and Brooks 95] Michael Collins and James Brooks. 1995. *Prepositional Phrase Attachment through a backed-off model*. In: Proc. of the Third Workshop on Very Large Corpora.

[Grefenstette 99] Gregory Grefenstette. 1999. *The world wide web as a resource for example-based machine translation tasks*. In Proc. of Aslib Conference on Translating and the Computer 21, London, November.

[Hartrumpf 99] Sven Hartrumpf. 1999. *Hybrid disambiguation of prepositional phrase attachment and interpretation*. In Proc. of Seventh Workshop on Very Large Corpora.

[Mehl et al. 98] S. Mehl, H. Langer, and M. Volk. 1998. *Statistische Verfahren zur Zuordnung von Präpositionalphrasen*. In: B. Schröder, W. Lenders, W. Hess, and T. Portele, editors, *Computers, Linguistics, and Phonetics between Language and Speech.* Proc. of the 4th Conference on Natural Language Processing. KONVENS-98, pages 97-110, Bonn. Peter Lang. Europäischer Verlag der Wissenschaften.

[Skut et al. 98] Wojciech Skut, Thorsten Brants, Brigitte Krenn, and Hans Uszkoreit. 1998. *A linguistically interpreted corpus of German newspaper text*. In Proc. of ESSLLI-98 Workshop on Recent Advances in Corpus Annotation, Saarbrücken.

[Stetina and Nagao 97] 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, pages 66-80, Beijing and Hongkong.

[Volk 99] Martin Volk. 1999. *Choosing the right lemma when analysing German nouns*. In: Multilinguale Corpora: Codierung, Strukturierung, Analyse. 11. Jahrestagung der GLDV. Frankfurt. 1999. 304-310.