Assigning Grammatical Relations with a Back-off Model

Erika F. de Lima
GMD - German National Research Center for Information Technology
Dolivostrasse 15
64293 Darmstadt, Germany
delima@darmstadt.gmd.de

Abstract

This paper presents a corpus-based method to assign grammatical subject/object relations to ambiguous German constructs. It makes use of an unsupervised learning procedure to collect training and test data, and the back-off model to make assignment decisions.

1 Introduction

Assigning a parse structure to the German sentence (1) involves addressing the fact that it is syntactically ambiguous:

(1) Eine hohe Inflationsrate erwartet die Ökonomin.
a high inflation rate expects the economist
'The economist expects a high inflation rate.'

In this sentence it must be determined which nominal phrase is the subject of the verb. The verb *erwarten* ('to expect') takes, in one reading, a nominative NP as its subject and an accusative NP as its object. The nominal phrases preceding and following the verb in (1) are both ambiguous with respect to case; they may be nominative or accusative. Further, both NPs agree in number with the verb, and since in German any major constituent may be fronted in a verb-second clause, both NPs may be the subject/object of the verb. In this example, morpho-syntactical information is not sufficient to determine that the nominal phrase [$_{NP}$ die Ökonomin] ('the economist') is the subject of the verb, and [$_{NP}$ Eine hohe Inflationsrate] ('a high inflation rate') its object.

Determining the subject/object of an ambiguous construct such as (1) with a knowledge-based approach requires (at least) a lexical representation specifying the classes of entities which may serve as arguments in the relation(s) denoted by each verb in the vocabulary, as well as membership information with respect to these classes for all entities denoted by nouns in the vocabulary. One problem with this approach is that it is usually not available for a broad-coverage system.

This paper proposes an approximation, similar to the empirical approaches to PP attachment decision (Hindle and Rooth, 1993; Ratnaparkhi, Reynar, and Roukos, 1994; Collins and Brooks, 1995). These make use of unambiguous examples provided by a treebank or a learning procedure in order to train a model to decide the attachment of ambiguous constructs. In the current setting, this approach involves learning the classes of nouns occurring unambiguously as subject/object of a verb in sample text, and using the classes thus obtained to disambiguate ambiguous constructs.

Unambiguous examples are provided by sentences in which morpho-syntactical information suffices to determine the subject and object of the verb. For instance in (2), the nominal phrase [$_{NP}$ der Ökonom] with a masculine head noun is unambiguously nominative, identifying it as the subject of the verb. In (3), both NPs are ambiguous with respect to case; however, the nominal phrase [$_{NP}$ Die Ökonomen] with a plural head noun is the only one to agree in number with the verb, identifying it as its subject.

(2) Eine hohe Inflationsrate erwartet der Ökonom.
a high inflation rate expects the economist
'The economist expects a high inflation rate.'

(3) Die Ökonomen erwarten eine hohe Inflationsrate.
the economists expect a high inflation rate
'The economists expect a high inflation rate.'

This paper describes a procedure to determine the subject and object in ambiguous German constructs automatically. It is based on shallow parsing techniques employed to collect training and test data from (un)ambiguous examples in a text corpus,

and the back-off model to determine which NP in a morpho-syntactically ambiguous construct is the subject/object of the verb, based on the evidence provided by the collected training data.

2 Collecting Training and Test Data

Shallow parsing techniques are used to collect training and test data from a text corpus. The corpus is tokenized, morphologically analyzed, lemmatized, and parsed using a standard CFG parser with a hand-written grammar to identify clauses containing a finite verb taking a nominative NP as its subject and an accusative NP as its object.

Constructs covered by the grammar include verb-second and verb-final clauses. Each clause is segmented into phrase-like constituents, including nominative (NC), prepositional (PC), and verbal (VC) constituents. Their definition is non-standard; for instance, all prepositional phrases, whether complement or not, are left unattached. As an example, the shallow parse structure for the sentence in (4) is shown in (4′) below.

(4) Die Gesellschaft erwartet in diesem Jahr
the society expects in this year
in Südostasien einen Umsatz
in southeast Asia a turnover
von 125 Millionen DM.
from 125 million DM
'The society expects this year in southeast Asia a turnover of 125 million DM.'

(4′) [$_S$ [$_{NC_{3,s,\{nom,acc\}}}$ Die Gesellschaft]
[$_{VC_{3,s}}$ erwartet]
[$_{PC}$ in diesem Jahr]
[$_{PC}$ in Südostasien]
[$_{NC_{3,s,acc}}$ einen Umsatz]
[$_{PC}$ von 125 Millionen DM]
]

Nominal and verbal constituents display person and number information; nominal constituents also display case information. For instance in the structure above, *3* denotes third person, *s* denotes singular number, *nom* and *acc* denote nominative and accusative case, respectively. The set $\{nom, acc\}$ indicates that the first nominal constituent in the structure is ambiguous with respect to case; it may be nominative or accusative.

Test and training tuples are obtained from shallow structures containing a verbal constituent and two nominative/accusative nominal constituents. Note that no subcategorization information is used; it suffices for a verb to occur in a clause with two nominative/accusative NCs for it to be considered testing/training data.

Training data consists of tuples (n_1, v, n_2, x), where v is a verb, n_1 and n_2 are nouns, and $x \in \{1,0\}$ indicates whether n_1 is the subject of the verb. Test data consists of ambiguous tuples (n_1, v, n_2) for which it cannot be established which noun is the subject/object of the verb based on morpho-syntactical information alone.

The set of training and test tuples for a given corpus is obtained as follows. For each shallow structure s in the corpus containing one verbal and two nominative/accusative nominal constituents, let n_1, v, n_2 be such that v is the main verb in s, and n_1 and n_2 are the heads of the nominative/accusative NCs in s such that n_1 precedes n_2 in s. In the rules below, $i, j \in \{1,2\}, j \neq i$, and $g(i) = 1$ if $i = 1$, and 0 otherwise. Note that the last element in a training tuple indicates whether the first NC in the structure is the subject of the verb (1 if so, 0 otherwise).

Case Nominative Rule. If n_i is masculine, and the NC headed by n_i is unambiguously nominative[1], then $(n_1, v, n_2, g(i))$ is a training tuple,

Case Accusative Rule. If n_i is masculine, and the NC headed by n_i is unambiguously accusative, then $(n_1, v, n_2, g(j))$ is a training tuple,

Agreement Rule. If n_i but not n_j agrees with v in person and number, then $(n_1, v, n_2, g(i))$ is a training tuple,

Heuristic Rule. If the shallow structure consists of a verb-second clause with an adverbial in the first position, or of a verb-final clause introduced by a conjunction or a complementizer, then $(n_1, v, n_2, 1)$ is a training tuple (see below for examples),

Default Rule. (n_1, v, n_2) is a test triple.

For instance, the training tuple *(Gesellschaft, erwarten, Umsatz, 1)* ('society, expect, turnover') is obtained from the structure (4′) above with the Case Accusative Rule, since the NC headed by the masculine noun *Umsatz* ('turnover') is unambiguously accusative and hence the object of the verb. The training tuple *(Inflationsrate, erwarten, Ökonom, 0)* ('inflation rate, expect, economist') and *(Ökonom, erwarten, Inflationsrate, 1)* ('economist, expect, inflation rate') are obtained from sentences (2) and (3) with the Case Nominative and Agreement Rules, respectively, and the test tuple *(Inflationsrate, erwarten, Ökonomin)* ('inflation rate, expect, economist') from the ambiguous sentence in (1) by the Default Rule.

[1]Only NCs with a masculine head noun may be unambiguous with respect to nominative/accusative case in German.

The Heuristic Rule is based on the observation that in the constructs stipulated by the rule, although the object may potentially precede the subject of the verb, this does not (usually) occur in written text. (5) and (6) are sentences to which this rule applies.

(5) In diesem Jahr erwartet die Ökonomin
in this year expects the economist
eine hohe Inflationsrate.
a high inflation rate
'This year the economist expects a high inflation rate."

(6) Weil die Ökonomin eine hohe Inflationsrate
because the economist a high inflation rate
erwartet, ...
expects
'Because the economist expects a high inflation rate, ...'

Note that the Heuristic Rule does not apply to verb-final clauses introduced by a relative or interrogative item, such as in (7):

(7) Die Rate, die die Ökonomin erwartet, ...
the rate which the economist expects, ...

3 Testing

The testing algorithm makes use of the back-off model (Katz, 1987) in order to determine the subject/object in an ambiguous test tuple. The model, developed within the context of speech recognition, consists of a recursive procedure to estimate n-gram probabilities from sparse data. Its generality makes it applicable to other areas; the method has been used, for instance, to solve prepositional phrase attachment in (Collins and Brooks, 1995).

3.1 Katz's back-off model

Let w_1^n denote the n-gram $w_1, \ldots, w_n$, and $f(w_1^n)$ denote the number of times it occurred in a sample text. The back-off estimate computes the probability of a word given the $n-1$ preceding words. It is defined recursively as follows. (In the formulae below, $\alpha(w_1^{n-1})$ is a normalizing factor and d_r a discount coefficient. See (Katz, 1987) for a detailed account of the model.)

$$P_{bo}(w_n|w_1^{n-1}) = \begin{cases} \tilde{P}(w_n|w_1^{n-1}), & \text{if } \tilde{P}(w_n|w_1^{n-1}) > 0 \\ \alpha(w_1^{n-1}) P_{bo}(w_n|w_2^{n-1}), & \text{otherwise,} \end{cases}$$

where $\tilde{P}(w_n|w_1^{n-1})$ is defined as follows:

$$\tilde{P}(w_n|w_1^{n-1}) = \begin{cases} d_{f(w_1^n)} \frac{f(w_1^n)}{f(w_1^{n-1})}, & \text{if } f(w_1^{n-1}) \neq 0 \\ 0, & \text{otherwise.} \end{cases}$$

3.2 The Revised Model

In the current context, instead of estimating the probability of a word given the $n-1$ preceding words, we estimate the probability that the first noun n_1 in a test triple (n_1, v, n_2) is the subject of the verb v, i.e., $P(S = 1|N_1 = n_1, V = v, N_2 = n_2)$ where S is an indicator random variable ($S = 1$ if the first noun in the triple is the subject of the verb, 0 otherwise).

In the estimate $P_{bo}(w_n|w_1^{n-1})$ only one relation—the precedence relation—is relevant to the problem; in the current setting, one would like to make use of two implicit relations in the training tuple—subject and object—in order to produce an estimate for $P(1|n_1, v, n_2)$. The model below is similar to that in (Collins and Brooks, 1995).

Let $\mathcal{L}$ be the set of lemmata occurring in the training triples obtained from a sample text, and let $c(n_1, v, n_2, x)$ denote the frequency count obtained for the training tuple (n_1, v, n_2, x) $(x \in \{0, 1\})$. We define the count $f_{so}(n_1, v, n_2) = c(n_1, v, n_2, 1) + c(n_2, v, n_1, 0)$ of n_1 as the subject and n_2 as the object of v. Further, we define the count $f_s(n_1, v) = \sum_{n_2 \in \mathcal{L}} f_{so}(n_1, v, n_2)$ of n_1 as the subject of v with any object, and analogously, the count $f_o(n_1, v)$ of n_1 as the object of v with any subject. Further, we define the counts $f_s(v) = \sum_{n_1, n_2 \in \mathcal{L}} c(n_1, v, n_2, 1)$ and $f_o(v) = \sum_{n_1, n_2 \in \mathcal{L}} c(n_1, v, n_2, 0)$. The estimate $P_i(1|n_1, v, n_2)$ $(0 \leq i \leq 3)$ is defined recursively as follows:

$$P_0(1|n_1, v, n_2) = 1.0$$

$$P_i(1|n_1, v, n_2) = \begin{cases} \frac{c_i(n_1, v, n_2)}{t_i(n_1, v, n_2)}, & \text{if } t_i(n_1, v, n_2) > 0 \\ P_{(i-1)}(1|n_1, v, n_2), & \text{otherwise,} \end{cases}$$

where the counts $c_i(n_1, v, n_2)$, and $t_i(n_1, v, n_2)$ are defined as follows:

$$c_i(n_1, v, n_2) = \begin{cases} f_{so}(n_1, v, n_2), & \text{if } i = 3 \\ f_s(n_1, v) + f_o(n_2, v), & \text{if } i = 2 \\ f_s(v), & \text{if } i = 1 \end{cases}$$

$$t_i(n_1, v, n_2) = \begin{cases} f_{so}(n_1, v, n_2) + f_{so}(n_2, v, n_1), & \text{if } i = 3 \\ f_s(n_1, v) + f_o(n_1, v) + f_s(n_2, v) + f_o(n_2, v), & \text{if } i = 2 \\ f_s(v) + f_o(v), & \text{if } i = 1 \end{cases}$$

The definition of $P_3(1|n_1, v, n_2)$ is analogous to that of $P_{bo}(w_n|w_1^{n-1})$. In the case where the counts are

positive, the numerator in the latter is the number of times the word w_n followed the n-gram w_1^{n-1} in training data, and in the former, the number of times n_1 occurred as the subject with n_2 as the object of v. This count is divided, in the latter, by the number of times the n-gram w_1^{n-1} was seen in training data, and in the former, by the number of times n_1 was seen as the subject or object of v with n_2 as its object/subject respectively.

However, the definition of $P_2(1|n_1, v, n_2)$ is somewhat different; it makes use of both the subject and object relations implicit in the tuple. In $P_2(1|n_1, v, n_2)$, one combines the evidence for n_1 as the subject of v (with any object) with that of n_2 as the object of v (with any subject).

At the P_1 level, only the counts obtained for the verb are used in the estimate; although for certain verbs some nouns may have definite preferences for appearing in the subject or object position, this information was deemed on empirical grounds not to be appropriate for all verbs.

When the verb v in a test tuple (n_1, v, n_2) does not occur in any training tuple, the default $P_0(1|n_1, v, n_2) = 1.0$ is used; it reflects the fact that constructs in which the first noun is the subject of the verb are more common.

3.3 Decision Algorithm

The decision algorithm determines for a given test tuple (n_1, v, n_2), which noun is the subject of the verb v. In case one of the nouns in the tuple is a pronoun, it does not make sense to predict that it is subject/object of a verb based on how often it occurred unambiguously as such in a sample text. In this case, only the information provided by training data for the noun in the test tuple is used. Further, in case both heads in a test tuple are pronouns, the tuple is not considered. The algorithm is as follows.
If n_1 and n_2 are both nouns, then n_1 is the subject of v if $P_3(1|n_1, v, n_2) \geq 0.5$, else its object.
In case n_2 (but not n_1) is a pronoun, redefine c_i and t_i as follows:

$$c_i(n_1, v, n_2) = \begin{cases} f_s(n_1, v), & \text{if } i = 2 \\ f_s(v), & \text{if } i = 1 \end{cases}$$

$$t_i(n_1, v, n_2) = \begin{cases} f_s(n_1, v) + f_o(n_1, v), & \text{if } i = 2 \\ f_s(v) + f_o(v), & \text{if } i = 1 \end{cases}$$

and calculate $P_2(1|n_1, v, n_2)$ with these new definitions. If $P_2(1|n_1, v, n_2) \geq 0.5$, then n_1 is the subject of the verb v, else its object. We proceed analogously in case n_1 (but not n_2) is a pronoun.

3.4 Related Work

In (Collins and Brooks, 1995) the back-off model is used to decide PP attachment given a tuple (v, n_1, p, n_2), where v is a verb, n_1 and n_2 are nouns, and p a preposition such that the PP headed by p may be attached either to the verb phrase headed by v or to the NP headed by n_1, and n_2 is the head of the NP governed by p.

The model presented in section 3.2 is similar to that in (Collins and Brooks, 1995), however, unlike (Collins and Brooks, 1995), who use examples from a treebank to train their model, the procedure described in this paper uses training data automatically obtained from sample text. Accordingly, the model must cope with the fact that training data is much more likely to contain errors. The next section evaluates the decision algorithm as well as the training data obtained by the learning procedure.

4 Results

The method described in the previous section was applied to a text corpus consisting of 5 months of the newspaper *Frankfurter Allgemeine Zeitung* with approximately 15 million word-like tokens. The learning procedure produced a total of 24,178 test tuples and 47,547 training triples.

4.1 Learning procedure

In order to evaluate the data used to train the model, 1000 training tuples were examined. Of these tuples, 127 were considered to be (partially) incorrect based on the judgments of a single judge given the original sentence. Errors in training and test data may stem from the morphology component, from the grammar specification, from the heuristic rule, or from actual errors in the text.

4.1.1 Subcategorization Information

The system works without subcategorization information; it suffices for a verb to occur with a possibly nominative and a possibly accusative NC for it to be considered training/test data. Lack of subcategorization leads to errors when verbs occurring with an (ambiguous) dative NC are mistaken for verbs which subcategorize for an accusative nominal phrase. For instance in (7) below, the verb *gehören* ('to belong') takes, in one reading, a dative NP as its object and a nominative NP as its subject. Since the nominal constituent [$_{NC}$ Bill] is ambiguous with respect to case and possibly accusative, the erroneous tuple *(Wagen, gehören, Bill, 1)* ('car, belong, Bill') is produced for this sentence.

(7) Der Wagen gehört Bill.
the car belongs Bill
'The car belongs to Bill.'

Another source of errors is the fact that any accusative NC is considered an object of the verb. For instance in sentence (8), the verb *trainieren* ('to train') occurs with two NCs. Since the NC preceding the verb is unambiguously nominative and the one following the verb possibly accusative, the training tuple *(Tennisspieler, trainieren, Jahr, 1)* ('tennis player, train, year') is produced for this sentence, although the second NC is not an object of the verb.

(8) Der Tennisspieler trainiert das ganze Jahr.
the tennis player trains the whole year

4.1.2 Homographs

In sentence (9) below, the word *morgen* ('tomorrow') is an adverb. However, its capitalized form may also be a noun, leading in this case to the erroneous training tuple *(Morgen, trainieren, Tennisspieler, 0)* (since $[_{NC}$ der Tennisspieler] is unambiguously nominative).

(9) Morgen trainiert der Tennisspieler.
tomorrow trains the tennis player
'The tennis player will train tomorrow.'

4.1.3 Separable Prefixes

In German, verb prefixes can be separated from the verb. When a finite (separable prefix) main verb occupies the second position in the clause, its prefix takes the last position in the clause core. For example in sentence (10) below, the prefix *zurück* of the verb *zurückweisen* ('to reject') follows the object of the verb and a subordinate clause with a subjunctive main verb. This construct is not covered by the current version of the grammar. However, due to the grammar definition, and since *weisen* is also a verb (without a separable prefix) in German, $[_C$ Er weist die Kritik der Prinzessin] is still accepted as a valid clause, leading to the erroneous training tuple $(er, weisen, Kritik, 1)$ ('he, point, criticism'). Such errors may be avoided with further development of the grammar.

(10) Er weist die Kritik der Prinzessin, seine
he rejects the criticism the princess his
Ohren seien zu groß, zurück.
ears are too big PRT
'He rejects the princess' criticism that his ears are too big."

4.1.4 Constituent Heads

The system is not always able to determine constituent heads correctly. For instance in sentence (11), all words in the name *Mexikanische Verband für Menschenrechte* are capitalized. Upon encountering the adjective *Mexikanische*, the system takes it to be a noun (nouns are capitalized in German), followed by the noun *Verband* "in apposition". Sentence (11) is the source of the erroneous training tuple *(Mexikanisch, beschuldigen, Behörde, 1)* ('Mexican, blame, public authorities').

(11) Der Mexikanische Verband für Menschen-
the Mexican Association for Human
rechte beschuldigt die Behörden.
Rights blames the public authorities
'The Mexican Association for Human Rights blames the public authorities.'

4.1.5 Multi-word lexical units

The learning procedure has no access to multi-word lexical units. For instance in sentence (12), the first word in the expression *Hand in Hand* is considered the object of the verb, leading to the training tuple *(Architekten, arbeiten, Hand, 1)* ('architect, work, hand'). Given the information the system has access to, such errors cannot be avoided.

(12) Alle Architekten sollen Hand in Hand arbeiten.
all architects should hand in hand work
'All architects should work hand in hand.'

4.1.6 Source Text

Not only spelling errors in the source text are the source of incorrect tuples. For instance in sentence (13), the verb *suchen* ('to seek') is erroneously in the third person plural. Since *Reihe* ('series') in German is a singular noun, and *Kontakte* ('contacts') plural, the actual object, but not the subject, agrees in number with the verb, so the incorrect tuple *(Reihe, suchen, Kontakt, 0)* ('series, seek, contact') is obtained from this sentence.

(13) *Eine Reihe von Staaten suchen geschäftliche
a series from states seek business
Kontakte zu der Region.
contacts to the region
'*A series of states seek contacts to the region.'

Finally, a large number of errors, specially in test tuples, stems from the fact that soft constraints are used for words unknown to the morphology.

4.2 Decision Algorithm

In order to evaluate the accuracy of the decision algorithm, 1000 triples were selected from the set of test triples. Of these, 285 contained errors, based

P_n	Number	Percent of test tuples	Number correct	Accuracy
P_3	2	0.28	2	100.00
P_2	204	28.53	194	95.10
P_1	486	67.97	431	88.68
P_0	23	3.22	20	86.96
Total	715	100.00	647	90.49

Figure 1: The accuracy of the system at each level

on the judgements of a single judge given the original sentence[2]. The results produced by the system for the remaining 715 tuples were compared to the judgements of a single judge given the original text. The system performed with an overall accuracy of 90.49%.

A lower bound for the accuracy of the decision algorithm can be defined by considering the first noun in every test tuple to be the subject of the verb (by far the most common construct), yielding for these 715 tuples an accuracy of 87.83%.

The above figure shows how many of the 715 evaluated test tuples were assigned subject/object based on the values P_n, and the accuracy of the system at each level.

The accuracy for P_2 and P_3 exceeds 95%. However, their coverage is relatively low (28.81%). Since the procedure used to collect training data runs without supervision, increasing the size of the training set depends only on the availability of sample text and should be further pursued.

One reason for the relatively low coverage is the fact that German compound nouns considerably increase the size of the sample space. For instance, the head of the nominal constituent [$_{NC}$ Der Tennisspieler] ('the tennis player') is considered by the system to be the compound noun *Tennisspieler* ('tennis player'), instead of its head noun *Spieler* ('player'). Consistently considering the head of putative compound nouns to be the head of nominal constituents may in some cases lead to awkward results. However, reducing the size of the sample space by morphological processing of compound nouns should be considered in order to increase coverage.

4.2.1 Examples

Following are examples of test tuples for which a decision was made based on values of P_2. All sentences below stem from the corpus.

Sentence (14) was the source for the test tuple *(Ausstellung, zeigen, Spektrum)* ('exhibition, show, spectrum'). This tuple was correctly disambiguated with $P_2 = 0.87$, with, among others, the training tuples *(Ausstellung, zeigen, Bild, 1)* ('exhibition, show, painting'), *(Ausstellung, zeigen, Beispiel, 1)* ('exhibition, show, example'), and *(Ausstellung, zeigen, Querschnitt, 1)* ('exhibition, show, cross-section') obtained with the Agreement (sentences (15) and (16)) and Case Rules (sentence (17)), respectively.

(14) Die Ausstellung zeigt das Spektrum jüdischer
the exhibition shows the spectrum jewish
Buchkunst von den Anfängen [...]
book art from the beginnings
'The exhibition shows the spectrum of jewish book art from the beginnings [...].'

(15) die letzte Ausstellung vor der Sommerpause
the last exhibition before the summer pause
zeigt Bilder und Zeichnungen von Petra
shows paintings und drawings from Petra
Trenkel zum Thema "Dorf".
Trenkel to the subject village
'The last exhibition before the summer pause shows paintings and drawings by Petra Trenkel on the subject "village".'

(16) Die Ausstellung im Museum für Kunst-
the exhibition in the museum for arts and
handwerk zeigt Beispiele seiner vielfältigen
crafts shows examples his manifold
Objekt-Typen [...]
object types
'The exhibition in the museum for arts and crafts shows examples of his manifold object types [...]'

(17) Eine vom französischen Kulturinstitut
a from the French culture institute
mit Unterstützung des Börsenvereins
with support the Börsenverein
in der Zentralen Kinder- und Jugendbibliothek
in the central children and youth library

[2]The higher error rate for test tuples is due to the soft constraints used for words unknown to the morphology.

im Bürgerhaus Bornheim
in the community center Bornheim

eingerichtete Ausstellung zeigt
organized exhibition shows

einen interessanten Querschnitt.
an interesting cross-section

'A exhibition in the central children's and youth library in the community center Bornheim, organized by the French culture institute with support of the Börsenverein, shows an interesting cross-section.'

Sentence (18) below was the source for the test tuple *(Altersgrenze, nennen, Gesetz)* ('age limit, mention, law'). The system incorrectly considered the noun *Altersgrenze* to be the subject of the verb.

(18) Eine Altersgrenze nennt das Gesetz nicht.
an age limit mentions the law not
'The law does not mention an age limit.'

There were no training tuples in which the compound noun *Altersgrenze* occurred as the subject/object of the verb. However, the noun *Gesetz* occurred more frequently as the object of the verb *nennen* than as its subject, leading to the erroneous decision.

5 Conclusion

This paper describes a procedure to automatically assign grammatical subject/object relations to ambiguous German constructs. It is based on an unsupervised learning procedure to collect test and training data and the back-off model to make assignment decisions. The system was implemented and tested on a 15-million word newspaper corpus.

The overall accuracy of the decision algorithm was almost 3% higher than the baseline of 87.83% established. The accuracy of the procedure for tuples for which a decision was made based on training pairs/triples (P_2 and P_3) exceeded 95%.

In order to increase the coverage for these cases as well as the overall performance of the procedure, the sample space should be reduced by morphologically processing German compound nouns, and the size of the training set should be increased. Further, in the experiment described in this paper, the model was trained with data obtained by an unsupervised procedure which performs with an accuracy of approximately 87% for training data. Further development of the morphology component and grammar definition should lead to improved results.

6 Acknowledgments

I would like to thank Michael Könyves-Tóth, who developed the parser engine used in the experiment described in this paper, for his support. I would also like to thank Martin Böttcher and the anonymous reviewers for many helpful comments on an earlier version of the paper.

References

Collins, Michael and James Brooks. 1995. Prepositional phrase attachment through a backed-off model. In *Proceedings of the Third Workshop on Very Large Corpora.*

Hindle, Donald and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1).

Katz, S. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3).

Ratnaparkhi, A., J. Reynar, and S. Roukos. 1994. A maximum entropy model for prepositional phrase attachment. In *Proceedings of the ARPA Workshop on Human Language Technology.*