OBT+Stat: Evaluation of a combined CG and statistical tagger

Janne Bondi Johannessen, Kristin Hagen and Anders Nøklestad
Text Lab, ILN,
University of Oslo, Norway
{jannebj, kristiha, noklestad} @iln.uio.no

André Lynum
Text Lab, ILN, University of Oslo and IDI, Norwegian University of Science and Technology – NTNU, Norway
andrely@idi.ntnu.no

Abstract

We have created a statistical POS tagger from existing development corpora and use it as a postprocessor to fully disambiguate the detailed morphological and lexical output of a Constraint Grammar tagger. In this article we discuss some of the challenges in unifying these two data-driven and knowledge-based approaches along with the possibilities and challenges that present themselves when using data-driven techniques to disambiguate candidates from a rule-based system. We then present an empirical evaluation that shows how the statistical disambiguation component improves the performance of the rule-based tagger. Our analysis of the results shows the potential for correcting the remaining errors and how the two tagger components interact in the disambiguation task.

1 Introduction

Compared to statistical methods, rule-based systems for natural language processing (NLP) often have a detailed focus on lingustic analysis, and naturally this is reflected in the output from such systems. The constraints and reasoning embedded in the system are often evident in the detail of the system output and ambiguities it leaves unresolved. The Oslo Bergen Tagger (OBT), originally developed at the University of Oslo and University of Bergen[1], is such a system, where linguistic detail has motivated both the particular form of the output and the decision not to fully disambiguate all readings for each token. The precision of the grammar rules and the richness of the underlying lexicon have been both a great resource and a challenge in the later stages of development of the OBT.

1.1 The original OBT output

The OBT is a rule-based tagger based on the Constraint Grammar (CG) formalism (Karlsson et al. 1995), which from the beginning has focused on linguistically precise descriptions. This focus is reflected in the more than 358 morphological tags (with a further 2000 or so for full morphosyntactic analysis) used by the system. Originally the primary users of the tagger were linguists, who are often studying subtle or marginal phenomena, and the recall of the system was considered particularly important; such users want to find all possibly relevant items in a corpus. The CG rules in essence disambiguate between the set of possible readings for a token and have been carefully constructed so that they do not compromise these overall goals of detail and comprehensiveness. As such it was in very few cases seen as acceptable to fuse different tags into portmanteau tags. One illustrative example is the rather large class of nouns that can be treated as either masculine or feminine. We considered it important to be able to identify these as feminine in the few contexts where they are definitely feminine even if that meant the consequence was gender ambiguity in other contexts.

[1] The Oslo-Bergen Tagger was originally developed by the Tagger Project at the University of Oslo in 1996 and written in a rule-based Constraint Grammar (CG1) framework with an interpreter from Lingsoft AB. An electronic lexicon (now available as *Norsk ordbank*) and a preprocessor were also part of the development performed by this project. In 2000 the commercial rule interpreter and the original preprocessor were replaced by components written in Allegro Common Lisp, developed in Bergen at Aksis (now Uni Digital). At this point the tagger was named The Oslo-Bergen Tagger (OBT). In 2008 the OBT was ported to the VISLCG3 framework in order to use the more recent CG3 formalism developed at The University of Southern Denmark in Odense. Uni Digital made a new stand-alone preprocessor based on *Norsk ordbank*, and The Text Laboratory converted the linguistic rules to the CG3 format.

A concrete example that illustrates this and also shows how the OBT works, is the sentence given in (1)[4].

```
   (1) Ei jente drakk. (‘A girl drank.’)
```

This sentence is initially rendered by the OBT as in (2), listing all possible readings by lexical analysis. All readings have been supplied by a preprocessor using a rich lexicon, including those that are ambiguous. Note the two possible genders for *jente* ‘girl’.

```
   (2)
‟<ei/a>″
  ‟ei″    adv
  ‟ei″    pron pers sg hum
  ‟eie″   verb imp
  ‟en″    det quant fem sing <Correct!>
″<jente/girl>″
  ‟jente″ noun fem com sg indef <Correct!>
  ‟jente″ noun masc com sg indef
″<drakk/drank>″
  ‟drikke″ verb past tense <Correct!>
```

The Constraint Grammar module (Hagen et al. 2000) uses the VISL CG3 rule interpreter to remove all ambiguity that is detectable from the grammatical context, in (2) taking advantage of the preceding feminine determiner. In this example, only those readings marked here with <*Correct!*> are left, as shown in (3).

```
   (3)
‟<ei/a>″
  ‟en″    det quant fem sg <Correct!>
″<jente/girl>″
  ‟jente″ noun fem com sg indef <Correct!>
″<drakk/drank>″
  ‟drikke″ verb past tense <Correct!>
```

The gender ambiguity for *jente* appears in the sentence *Det er bra å være jente.* (‘It is good to be (a) girl.’), rendered by the OBT as in (4).

```
   (4)
‟<det/it>″
  ‟det″ pron pers 3 sg neut <Correct!>
″<er/is>″
  ‟være″ verb present <Correct!>
″<bra/good>″
  ‟bra″ adj pos neut indef sg <Correct!>
″<å/to>″
  ‟å″ inf marker <Correct!>
″<være/be>″
  ‟være″ verb inf <Correct!>
″<jente/girl>″
  ‟jente″ noun fem com sg indef  <Correct!>
  ‟jente″ noun masc com sg indef <Correct!>
```

We see that *jente* ‘girl’ became unambiguous in (3), given agreement rules that refer to the feminine determiner, while (4) remains ambiguous since there is nothing in the context to determine the gender of the word. This result was considered acceptable at the time when the OBT was developed, since it would mean that the interested linguist could search a corpus for all instances of feminine nouns, and find ambiguous examples in addition to those where the gender had been fully determined. There are many ambiguities similar to this. Another example is found in the large group of neuter singular indefinite and neuter plural indefinite nouns, which can be disambiguated in some contexts but not all. Also in this case the remaining ambiguity is left in the result on purpose.

OBT only concentrates on grammatical ambiguity and does not look at ambiguity between lemmas. The result is that OBT leaves lemma ambiguities like *fare* ‘danger’ or *far* ‘father’ for the ambiguous word form *faren.*

The ambiguity of the OBT is not only of the type we have just described, it also includes a number of unfortunate ambiguities where the CG rules in the OBT do not have enough coverage. In Section 6, we discuss the different kinds of ambiguities more thoroughly.

1.2 Towards unambiguous output in OBT+Stat

While leaving ambiguity in the output can be reasonable from a linguistic standpoint, using the ambiguous output of the tagger as input for other research or engineering purposes constitutes a problem. These often require the output to contain a single reading for each token. Most notably concerning the OBT, the construction of large-scale Norwegian corpora such as *Norsk aviskorpus*[5] and NoWaC (Guevara 2010) requires a high quality, fully disambiguating tagger. Such requirements motivated us to look into how we could make the OBT suitable for this use.

Since continued rule writing activity gave diminishing returns and resources were limited, the need for a statistical module to complement the OBT increased. This motivated the implementation of OBT+Stat, a statistical module that removes all remaining ambiguity from the OBT output, both the grammatical ambiguities and the lemma ambiguities originally left on purpose – and the unfortunate ambiguities.

In this article, we discuss the changes made to the OBT in order to create an effective system,

[4] Tag abbreviations have been translated from Norwegian to English in all examples.

[5] http://avis.uib.no/

OBT+Stat, for fully disambiguated output that still maintains linguistic detail and comprehensiveness.

2 Related work

The combination of CG rule sets with statistical methods was attempted even during earlier work with CG, notably in the combination of the early Xerox XT statistical tagger with the EngCG rules (Tapanainen et al. 1994). This work was reported as showing some promise but does not appear to have been followed up.

Later work described in Hajič et al. (2001) and Spoustová et al. (2007) combines a CG tagger for Czech with statistical models in a sophisticated manner, including decoding that was constrained by the CG tagger output. Their work goes further than the work presented in this article, but since the Czech CG rule set focuses heavily on full recall at the expense of precision, our work is different in scope. Our results are interesting to other CG based systems that aim for high precision at the cost of some recall.

3 OBT+Stat

Since the OBT reaches a high f-score[6] (97.2), it is desirable to keep all complete disambiguations made by the OBT and only add the statistical disambiguation module as a post-processing step in those cases where the OBT leaves some ambiguity. We chose to run a statistical tagger independently of the OBT in a manner that will be described later in this section, and to combine the results instead of attempting more sophisticated modeling based on selecting candidates directly. This results in a simple pipeline where the statistical model is independent of changes in the CG rule set or the lexical preprocessor.

Since there are only a limited number of annotated corpora available for the development of Norwegian NLP tools, we decided to use the corpora already collected and annotated during the development of the OBT. This consists of 122 523 words in 8178 sentences in a hand-annotated development corpus and a corresponding corpus reserved for evaluation with 32 677 words and 2213 sentences. These corpora contained a number of tokens where several readings were marked as correct. Those parts of the corpora had to be annotated again in a fully disambiguated manner[7] - either by combining several tags into one or by keeping the original set of tags and making specific decisions about which reading is considered the correct one. Since combining tags for all readings would erase linguistic detail, and adding such combined tags for the ambiguous tokens would only add considerable complexity to the already large tag set, we chose instead to establish a set of annotation guidelines for the ambiguous cases.

These guidelines by necessity have to reflect some arbitrary decisions. For example, for words like *jente* 'girl', which we recall from Section 1 can be either masculine or feminine, we always choose the feminine tag in ambiguous contexts. In other cases, like words that are ambiguous between singular and plural, we let the human annotator take into account factors like semantic interpretation, knowledge of mass and count distinctions etc. to decide which reading to choose. We discuss the consequences of these guidelines further in Section 5.

As noted, the statistical disambiguation module works by running a statistical tagger independently of the CG-based disambiguation rules. The two results are then unified when the CG rule set leaves more than one reading for a token. If the reading produced by the statistical tagger agrees with one of the readings left by the CG tagger, that reading is selected. If the two taggers do not agree, we attempt to disambiguate using the lemma if possible (as explained at the end of this section), or, failing that, select an arbitrary reading. As we will see in Section 4, the statistical tagger we use covers nearly 80% of the ambiguous readings, with the lemma disambiguation covering most of the remaining ones, leaving only 0.63% to arbitrary selection.

Hidden Markov Model (HMM) based taggers are well known and widely covered in the literature (see e.g. Brants 2000). The HMM model conditions its sequence labelling decisions on the previous one or two labels and the current token to emit. HMM models also include mature and empirically founded models for unknown words when used with most European languages. In our experience, the available HMM taggers, while currently outperformed by more sophisticated models, still provide robust and competitive results on real world data. We have chosen to use the HMM tagger HunPos (Halácsy 2007), which gave good performance on the disambiguation

[6] We use a standard balanced f-score, which is defined as 2 * precision * recall / (precision + recall).

[7] All these corpora were annotated by Arne Martinus Lindstad.

task in addition to being robust and open software with very fast training and decoding performance.

One may ask if the simplistic manner in which we use the statistical tagger is the right way to model the statistical disambiguation process, but it is our intuition that with the scarce resources available more advanced and specialized models, such as constrained decoding or direct discriminative modeling on the disambiguation task itself, may not necessarily yield consistent improvements. The off-the-shelf HMM models offer robust handling of unknown words and well-understood hidden sequence labeling which we regard as more cost-effective in terms of effort and results than a model specifically designed for the task.

Our lemma disambiguation scheme is also simple. It uses the recently created NoWaC corpus of Norwegian documents published on the internet. Our idea is that most lemmas will appear as words in a large corpus since Norwegian lemmas correspond to uninflected words forms. Motivated by this we use a word frequency list derived from NoWaC and select the lemma with the highest frequency in the corpus. This part of the statistical disambiguator considers the output of the CG tagger rather than being run independendently like the POS tagger. We will discuss the lemma disambiguation further in Section 6.2.

4 Comparison with the OBT

Comparing the earlier published evaluation results for the OBT with the fully disambiguated results presents some difficulties since the tasks are different in some aspects and are usually evaluated differently in the relevant literature. The OBT tagger, like other CG based taggers, has previously focused on removing readings that it knows to be incorrect according to the linguistic knowledge embedded in the rules, while leaving the remaining readings in the result. As we have shown in earlier sections, several correct or incorrect tags may be left after disambiguation, and both the precision and recall are reported in evaluations, combined into a standard balanced f-score. CG taggers often make a trade-off between precision and recall where aggressively eliminating readings will increase precision after leaving fewer incorrect readings while reducing recall by unintentionally removing correct readings. Keeping lost recall to a minimum had a high priority during the development of the OBT since it was deemed important that relevant linguistic examples should not be lost when searching a corpus. The OBT for Norwegian *Bokmål*[8] achieved a standard f-score of 97.2, with a recall of 99.0% and a precision of 95.4%, a highly competitive result, but including some ambiguities as we explained in Section 1.

In contrast, when we perform fully disambiguated tagging, the notion of several correct readings disappears, and precision and recall become identical. Results from this kind of tagging are therefore usually reported as a single token wise accuracy score. This accuracy score is not directly comparable to the older precision score since the evaluation corpus is now fully disambiguated and several readings that were previously considered correct are now considered incorrect. One could view the fully disambiguated evaluation as having 100% recall, but the f-score is not a linear function and we can only speculate on the precision of the OBT tagger if rules were developed to raise the recall to this level.

Still we maintain that the two scores can be compared within reason. We will mainly compare the earlier precision score with the accuracy of the new results, considering a slightly lower accuracy for the fully disambiguated result to be at the same level as the older results, and comparable or higher accuracy as an improved result.

5 Evaluation

The training and test corpora for the OBT are drawn from a variety of text types: newspapers (with headlines, by-lines etc.), journal articles, magazines, government reports, and fiction. Combined with the focus on detailed linguistic representation this makes both the underlying data and the resulting analysis more diverse than what is usually the case (much language technology development has been done on for example the homogenous Wall Street Journal corpus). Having presented this reservation, we will first show some statistics measuring the amount of ambiguity in the corpus before presenting the results proper.

After the CG-driven disambiguation on the fully disambiguated test corpus described in Section 3, the amount of ambiguities is as summarized in Table **1**. The still ambiguous tokens now constitute 8.6% of the test corpus, and as seen from the table and the previous discussion, they

[8] The tagger for the Norwegian language variety *Nynorsk* is not discussed in this paper.

result mainly from an inability to choose between two readings.

	Total readings/ tokens	Ratio
Ambiguity over all words	35639 / 32677	1.09
Ambiguity over ambiguous words only	5778 /2816	2.05

Table 1 Amount of ambiguity left by the OBT in the evaluation corpus.

The overall accuracy of the OBT+Stat tagger is measured at 96.56%. We consider this to be a good result considering the fact that we have removed all remaining ambiguity while at the same time kept a large, detailed tag set and disambiguated lemmas with identical tags.

Breaking down the results further, we see from Table 2 that the POS/morphological disambiguation accuracy is slightly higher than the overall accuracy, which includes disambiguation of identical tags with different lemmas in addition to the pure POS/morphological disambiguation.

	Correct readings / tokens	Ratio
Overall accuracy	31552 / 32677	96.56%
Tagging accuracy	31614 / 32677	96.74%
Lemmatization accuracy	32131 / 32677	98.33%

Table 2 Accuracy scores measuring the performance of the fully disambiguating tagger. Scores are shown for tagging and lemmatization separately and combined.

As shown in Table 3, the overlap between the statistical tagger and the OBT for the tokens left ambiguous by the OBT is quite good: nearly 80% of the cases in question. For these words the accuracy of the statistical tagger is 81.70%. Since these are the most difficult tagging decisions left over by the OBT we find this performance of the statistical tagger to be quite good. Errors include some tokens out of coverage where OBT has mistakenly eliminated the correct reading, making it impossible for the OBT+Stat system to make the right decision. The corresponding statistics for the lemma disambiguation are harder to analyze since this model is only used in very specific circumstances, but the coverage and precision indicate that the lemma model is effective and has some impact on the overall result.

	Ratio
Statistical tagger coverage	79.39% (2273/2863)
Statistical tagger accuracy	81.70% (1857/2273)
Lemma model coverage	54.23% (551/1016)
Lemma model accuracy	86.71%

Table 3 The coverage and accuracy of the statistical disambiguation module on the ambiguous tokens for the statistical tagger and lemma disambiguator respectively.

In addition to evaluating the OBT tagger as it is currently in use we also constructed a CG rule set where we removed rules that had been written in order to remove spurious ambiguity by means of heuristics. The premise is that the ambiguities covered by those rules should now be covered by the statistical module. The results of removing these rules are shown in Table **4**.

	correct readings / tokens	Ratio
Overall accuracy	31332/32677	95.88%
Tagging accuracy	31459/32677	96.27%
Lemmatization accuracy	32187/32677	98.50%

Table 4 Accuracy scores for the fully disambiguating tagger with a modified CG rule set where heuristically disambiguating rules are removed.

The overall accuracy using this rule set is slightly lower, 0.68% in absolute and about 20% in relative terms. The change in tagging accuracy is about the same, while the lemma accuracy is a bit higher. This shows that the statistical model does not necessarily handle some of the ambiguities covered by the heuristic rules as well as the rules themselves do. Including the heuristic rules is still beneficial when using the statistical disambiguator.

We also evaluated the statistical module with a CG rule set that attempts to fix some of the disambiguation decisions that have now been annotated in a consistent manner in the training and development corpora based on the new annotation guidelines. Those rules should hopefully consistently determine some ambiguities which the statistical disambiguation module may not resolve in consistent manner. The rules mostly concern ambiguous masculine/feminine nouns, which are disambiguated as feminine in contexts that lack gender agreement (as illustrated in example (2) in the introduction). The results are shown in Table **5**.

	Correct readings / tokens	Ratio
Overall accuracy	31668/32677	96.52%
Tagging accuracy	31668/32677	96.91%
Lemmatization accuracy	32181/32677	98.48%

Table 5 Accuracy scores for the fully disambiguating tagger with a modified CG rule set that attempts to fix disambiguities now resolved in the annotation guidelines.

The results using these rules are roughly the same as the results for the main CG rule set, indicating that the CG rules and statistical module perform the disambiguation of those cases about equally effectively.

6 Discussion of the results

In this section we will look at some successful and some unsuccessful results, both with respect to grammatical tags and with respect to lemma disambiguation.

6.1 Grammatical disambiguation

We will have a look at the contribution of the statistical module in isolation. We first examine some successful examples.

```
(5)
Offentlige etater har ansvar for…
Public institutions have responsibility for…

Resolved ambiguity:
″<ansvar/responsibility>″
    ‶ansvar″ noun neut com sg indef
            <Correct!> <SELECTED>
    ‶ansvar″ noun neut com pl indef
```

In (5), the OBT had left the word *ansvar* 'responsibility' ambiguous between singular and plural (recall the discussion in Section 1). This particular kind of ambiguity accounts for 615 of the remaining ambiguities before statistical disambiguation, or 21.8%. The statistical module has correctly identified this word as singular, and over all such ambiguities, 418, or 68.0%, are disambiguated correctly by the statistical module.

In example (6) there are actually two ambiguities, both between parts of speech, which have been resolved: one between adverb and preposition reading, and one between verb and noun. The adverb/preposition ambiguity is fairly marginal with only two occurrences in the test corpus, both disambiguated correctly and identical to the one shown in the example. The noun/verb ambiguity is of a more interesting type with 58 (2.1%) occurrences in the corpus. It is often fairly complex with over half of the occurrences having three or more readings to disambiguate and six occurrences having five or more readings. Still, the accuracy of the statistical module for this kind of ambiguity is good, 82.8% (48 correctly resolved occurrences).

```
(6)
... at for mange typer informasjon vil de
elektroniske mediene etter hvert bli enerå-
dende.
... that for many types (of) information the
electronic media will slowly become dominant.

Resolved ambiguity I:
″<for>″
    ‶for″ adv
    ‶for″ prep <Correct!> <SELECTED>

Resolved ambiguity II:
″<typer>″
    ‶type″ noun masc com pl indef <Correct!>
           <SELECTED>
    ‶type″ verb present
```

We now turn to the less successful choices made by the statistical module. In the first quarter of the test corpus, OBT+Stat makes a total of 105 errors. The distribution of those errors with respect to grammatical categories is summarized in **Table 6**.

Singular or plural in neuter nouns (i)	41
Other singular or plural errors (ii)	5
Gender agreement (iii)	15
Nouns ending in -s, genitive or not genitive (iv)	4
Adjective gender (m,f,sg,pl) chosen instead of adj neut/adv (v)	13
Lack of imperative (vi)	3
Other errors (vii)	24

Table 6 Counts of disambiguation errors caused by the statistical module.

Not surprisingly, the largest group of errors by far is due to the difficulty of assigning singular or plural readings to neuter nouns that have the same form in both indefinite singular and plural. As discussed in Section 3, the decisions for the training and test corpus were left to the human annotator, based on linguistic knowledge. The patterns may not in all cases be clear enough for the statistical module to make correct decisions. One may ask whether keeping this distinction as separate tags apart is a good choice, or whether they should have been collapsed.

It turns out that when classifying errors by grammatical type, they usually have very little in common within the classified category. We will now show a few selected examples, beginning with type (i) in (7):

```
(7)
Generelt om informasjon og særtrekk ved of-
fentlig informasjon
Generally about information and characteris-
tics in public information

Wrongly resolved ambiguity:
"<særtrekk/characteristic>"
       "særtrekk" noun neut com sg indef
                  <SELECTED> <ERROR>
       "særtrekk" noun neut com pl indef
<Correct!>
```

In this example we see that the word *særtrekk* 'characteristic' has been tagged as a singular noun, while the test corpus assigns it a plural interpretation. There is very little in the grammatical surroundings that would give a hint as to the correct interpretation. The most likely (but incorrect) clue would have been the fact that this word is a conjunct in a coordination phrase, and that the first conjunct is a singular noun. However, in the present case, the first conjunct is a mass noun, while the second conjunct is a count noun, and it is only when this is taken into account that the right plural tag can be applied. The mass/count distinction is not marked in the tags. A larger training corpus could possibly allow OBT+Stat to disambiguate this case properly.

Our example (8) is of type (ii), but still one that deals with singular and plural.

```
(8)
Selv om utvalget kanskje særlig viser noen av
Hjemme-PCs favoritter
Even if the selection perhaps especially show
some of Hjemme-PC's favourites

Wrongly resolved ambiguity:
"<noen/some>"
       "noen" pron pers 3 sg masc fem
              <SELECTED> <ERROR>
       "noen" pron pers 3 pl <Correct!>
```

In (8), the error is in the assignment for number for the word *noen* 'some/any'. Like with the wrongly annotated word in (7), it is not in an agreement context, which is most likely the reason that the OBT did not disambiguate it. The interpretation of *noen* as singular is actually only possible when it is used as a negative polarity item, which is obviously not the case here. However, since there are a variety of constructions that license negative polarity items in Norwegian (Lindstad 1999), the CG tagger was not able to make the correct choice.

A third type shown in example (9), type (vi), illustrates the problem of having short headlines in the test corpus.

```
(9)
Fly
Fly

Wrongly resolved ambiguity:
"<fly>"
"fly" noun neut com sg indef
      <SELECTED> <ERROR>
"fly" noun neut com pl indef
"fly" verb imp  <Correct!>
```

Even for a human annotator it may be difficult to interpret what the headline is actually meant to be; imperative or noun. Imperatives happen to be few in written texts, and as a result, a statistical module will almost invariably fail in this task.

At the end of this section we would like to point out that while it could have been conceivable that the statistical module made many of its errors due to already faulty output from the OBT (i.e., with the correct tag missing), a quick count shows this not to be the case. Out of our 749 errors made by the statistical module, a modest 74 are due to errors made by the OBT tagger.

6.2 Lemma disambiguation

The OBT was never developed with the intention of doing lemma disambiguation. Often ambiguous lemmas have different grammatical characteristics and they would effectively be disambiguated anyway, but this is not always the case. There were 515 ambiguous lexical lemmas, out of which 395 (76.70%), were correctly resolved. Looking at the 28 errors in the first quarter of the test corpus we will first give some examples of successfully resolved lemmas, and then some less fortunate ones.

```
(10)
I alle deler av den offentlige forvaltningen
In all parts of the public administration

Resolved lemma ambiguity:
"<deler>"
    "del/part" noun masc com pl indef
               <Correct!> <SELECTED>
    "dele/border" noun neut com pl indef
```

In (10), the ambiguity is between the lemmas *del* 'part' and *dele* 'border', and it is only the first one that is correct, correctly disambiguated by OBT+Stat. Given that the correct word has a much higher frequency than the other, we would expect that the tagger chose correctly.

```
(11)
En positiv utvikling i statlig informasjons-
virksomhet de siste årene
A positive development in state information
practice the last years

Resolved lemma ambiguity:
”<årene>”
    “år/paddling ore” noun fem com pl def
    “år/paddling ore” noun masc com pl def
    “år/year” noun neut com pl def
            <Correct!> <SELECTED>
    “åre/paddling ore, vein” noun fem com pl
                             def
    “åre/paddling ore, vein” noun masc com pl
                             def
```

In (11), with all the lemmas to choose between, it is satisfying that OBT+Stat made the right choice. Given that the right word is the most general one, and hence occurs in a wide variety of texts, this is the most frequent word form found in NoWaC and subsequently chosen by the tagger.

However, the texts in the test corpus do not always deal with the most general topics. Hence our first illustration of an error shows the same word form, but now with a different lemma as the correct one. Consider (12).

```
(12)
Fysikeren ville f.eks. studere sammenhenger
mellom blodtrykk, størrelse av årene og blod-
tilførsel.
The physicist wanted e.g. (to) study connec-
tions between blood pressure, size of the
veins and blood supply.

Wrongly resolved lemma:
”<årene>”
”år/paddling ore” noun fem com pl def
”år/paddling ore” noun masc com pl def
”år/year” noun neut com pl def
        <SELECTED> <ERROR>
”åre/paddling ore, vein” noun fem com pl def
                         <Correct!>
”åre/paddling ore, vein” noun masc com pl def
```

The text is about a medical topic, and here the less general lemma meaning 'vein' is the correct one, which the statistical module did not find. We have looked at the test corpus, and the word form *årene* occurs six times, five of them in the 'year' meaning. Sometimes, however, the frequency does not seem to be so equally distributed in NoWaC and the test corpus. Consider (13).

```
(13)
Faren min var mye ute og reiste
My father was much out and travelled
(=travelled a lot)

Wrongly resolved:
”<faren>”
    “far” noun masc com sg def <Correct!>
    “fare” noun masc com sg def
           <SELECTED> <ERROR>
```

The statistical module should disambiguate between the lemma meaning 'father' ("far") and that meaning 'danger' ("fare"). One would guess that in both corpora the word meaning 'danger' would be more frequent, and in fact in (13) the statistical module has picked 'danger' due to an evidently higher frequency of the word form *fare* than of *far* in NoWaC. For this example the NoWaC and OBT corpus disagree in the distribution of the semantics of this word form. In fact, out of four occurrences of the word form *faren* in the test corpus, three had the meaning 'father' and only one the meaning 'danger'.

Our conclusion with respect to the lemma module is that it seems to work quite well, since most lemmas have large differences in frequency as word forms, and the differences seem to correspond fairly well between NoWaC and the test corpus. Furthermore, the hypothesis that uninflected word form frequencies in a large corpus can be used as an indication of lemma frequencies seems to bear out in practice. It is still possible that some rule-based approach would improve our lemma disambiguation. For example, the word form meaning 'father' very often occurs together with a possessive pronoun, and this kind of knowledge could have been put into the system.

7 Conclusion

We have improved an originally rule-based tagger, the OBT, with a statistical HMM tagger. The latter has been applied on the still ambiguous output of the OBT. The resulting OBT+Stat system performs well and has two added advantages compared with the original tagger in that it gives unambiguous output and it performs lemma disambiguation.

References

Alfred. V. Aho and Jeffrey D. Ullman. 1972. The Theory of Parsing, Translation and Compiling, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. Publications Manual. American Psychological Association, Washington, DC.

Association for Computing Machinery. 1983. Computing Reviews, 24(11):503-512.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. Journal of the Association for Computing Machinery, 28(1):114-133.

Brants, T. 2000. TnT: a statistical part-of-speech tagger. In Proceedings of the Sixth Conference on Applied Natural Language Processing. Seattle, Washington.

Emiliano Guevara. 2010. NoWaC: a large web-based corpus for Norwegian. In Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop. Los Angeles.

Giesbrecht, Eugenie and Stefan Evert. 2009. Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. In Proceedings of the Fifth Web as Corpus Workshop (WAC5). San Sebastian, Spain.

Hagen, Kristin, Janne Bondi Johannessen and Anders Nøklestad. 2000. A Constraint-based Tagger for Norwegian. In 17th Scandinavian Conference of Linguistics, [Odense Working Papers in Language and Communication 19], Carl-Erik Lindberg and Steffen Nordahl Lund (eds), pp. 31-48. University of Southern Denmark, Odense.

Jan Hajicˇ , Pavel Krbec, Pavel Kveˇtonˇ , Karel Oliva and Vladim´ır Petkevicˇ . 2001. Serial Combination of Rules and Statistics: A Case Study in Czech Tagging. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. CNRS – Institut de Recherche en Informatique de Toulouse and Universite´ des Sciences Sociales, pp. 260–267. Toulouse.

Péter Halácsy, András Kornai, Csaba Oravecz. 2007. HunPos - an open source trigram tagger. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Companion Volume Proceedings of the Demo and Poster Sessions. Association for Computational Linguistics, Prague, Czech Republic.

Johannessen, Janne Bondi and Kristin Hagen. 2003. Parsing Nordic Languages (PaNoLa) norsk versjon. In Nordisk Sprogteknologi 2002, pp. 89-95. Museum Tusculanums Forlag, University of Oslo.

Karlsson, Fred, Atro Voutilainen, Juho Heikkilä and Arto Anttila. 1995. Constraint Grammar, A language independent system for parsing unrestricted text. Mouton de Gruyter. Berlin; New York.

Lindstad, Arne Martinus. 1999. Issues in the Syntax of Negation and Polarity in Norwegian. A Minimalist Analysis. Cand.philol thesis, University of Oslo.

Spoustová, D., J. Hajič, J. Votrubec, P. Krbec and P. Květoň. 2007. The best of two worlds: cooperation of statistical and rule-based taggers for Czech. In Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies. Prague, Czech Republic.

Tapanainen, P. and A. Voutilainen. 1994. Tagging accurately: don't guess if you know. In Proceedings of the Fourth Conference on Applied Natural Language Processing. Stuttgart, Germany.

VISL CG3: http://beta.visl.sdu.dk/cg3.html