Statistical noun chunking for Swedish

Tobias Håkansson
Department of Computer Science
Faculty of Science, Lund University
tobias@rebias.se

January 15, 2007

Abstract

The purpose of this project is to give some limited insight as to how a statistical model can be created for the purpose of analyzing texts and generate noun chunks of these texts. This article describe a program which from a given text parses that text and shows the noun chunks from the text in a simple GUI with some simple information about the text. The chunking annotation used is IOB2. We argue from the results that a statistical noun chunker will work relatively well on Swedish, this based on the results we got from some simple tests. From these results we then motivate actions to be taken to get better results and then argue for the accuracy of the results.

1. Introduction

A chunk is a 'non overlapping phrase which splits a text in parts (chunks) and where the words in the chunk relate to each other'. Chunking is mainly used to give a fast approach to analyze a text, it is not the most accurate method but it is simple and easy to implement. The goal of the project was to implement a program which parsed a corpus and annotated the noun chunks and after that to create a statistical model from the noun chunk annotated text. To have a easy way to overview the chunking we created a simple GUI in which the chunk annotated text was displayed, all noun chunks displayed in red and all others in plain black. For portability the programs were written in Java with the exception of the programs to test and create the statistical model. The program Weka was used to create the statistical model. The goal was to reach a percentage of approximately 65-70 percent, which with the relatively simple approach we use would be good. The chunking annotation used is IOB2, for the line 'Individuell beskattning av arbetsinkomster' (Individual taxation of work income) it would look like this:

Individuell AJ B-NP
beskattning VN I-NP
av PR O
arbetsinkomster NN__SS B-NP

The program which converts the corpus is written in Java and the program which extracts data from the converted corpus is written in Prolog. We got results of approximately 69 percent which we were satisfied with due to the relatively simple method to extract the statistical data. With this result it would not be too difficult to get some more percent with an easy modification so that the model takes more data.

2. The Project

2.1 Converter and GUI

The first thing to do was to choose which nouns we wanted to include in the chunks, this because the corpus made difference between "all" of them. The agreement here was to choose the five most common noun types to get the project going. After further theoretical discussions we agreed that approximately ten to fifteen different noun types should be used in the final testing stage but at this point we settled for five. The data we used was from 'Talbanken' which is a Swedish corpus which dates back to 1970, but has been updated through out the years. 'Talbanken' is split into four types of texts and we chose to use a more rigid text about 'Individual taxation of work income'.

The format of 'Talbanken' is quite extensive so a scale down was a logical thing to do since the statistical model should not be too complex (The format can be viewed in appendix 1). We chose the IOB2 annotation in the new format which means that a noun chunk starts with B, if there are more words in the chunk they get an I and if the word doesn't belong to a chunk it gets an O. The converter program converts the 'Talbanken' format into the new one on the form:

Individuell AJ B-NP
beskattning VN I-NP
av PR O
arbetsinkomster NN__SS B-NP

As can be seen here 'Individuell beskattning' is annotated as a noun chunk beginning with 'Individuell' and ends with 'beskattning' and also that 'arbetsinkomster' is a noun chunk. The word 'av' does not belong to a chunk so it gets an O.

Once the format was correct and worked we added more noun types to ensure a better result. At this time we realized that it was time to implement the simple GUI to get an overview of the converted text. We wrote the GUI in Java and it is very simple. The GUI takes a text file with the new format and reads it one sentence at a time, and then it writes out the sentence in the window marking up all chunks with red and the rest of the text with black.

The two programs, the converter and the GUI, are later merged into a single program. This program takes a text file as input (on the correct 'Talbanken' annotation) and converts it and shows it in the GUI.

2.2 Creating the statistical model

The next step now that the data for the statistical model was created was to actually create the model. For this purpose we used a program called Weka. To create the data for Weka we used a couple of Prolog programs which extracted the Part Of Speech tag (POS-tag) and the chunk from the converters data. The data was then

split into two parts, a training set of approximately 80% of the data and a test set of the remaining 20%. We then used classifier J48 with a cross-validation of 10 to create the model in Weka. Now when the model was created it was time to test it, for this we used first a Prolog program to create data from our statistical model and then we used an evaluation program to measure the results from our new data.

2.3 Results

The results we got from the evaluation program were what we hoped for, with an accuracy of 69.34% it was in the in the range of our hypothesis.

2.4 Improvements

As our accuracy reached approximately 69 percent there are surely room for improvement. One thought is that different types of text can give some differences but we assume that this variation can be neglected. When we create the statistical model now we only use the current words POS-tag as to determine the correct chunk, one way to surely increase the percentage is to look at more words. If one were to look at one word before and one word after we can assume that it would increase the accuracy. Other ways to improve the model might be to train it with a different classifier then J48 which we used for our model.

2.5 Conclusion

The conclusions we can draw from this project is that statistical noun chunking for Swedish can work pretty well and it is easy to create a good model. Even though the model used in this project was very simple it gave relatively high scores so it looks promising further studying.

Acknowledgement:

I would like to thank both Richard Johansson and Pierre Nugues for the help I received during this project.

Reference list:

http://w3.msi.vxu.se/~nivre/research/Talbanken05.html

Appendix

1. The original format from ‘Talbanken’

```
- <sentence id="2" user="" date="">
  <word id="1" form="Genom" postag="PR" head="3" deprel="AA" />
  <word id="2" form="skattereformen" postag="NNDDSS" head="1" deprel="PA" />
  <word id="3" form="införs" postag="VVPSSMPA" head="0" deprel="ROOT" />
  <word id="4" form="individuell" postag="AJ" head="5" deprel="AT" />
  <word id="5" form="beskattning" postag="VN" head="3" deprel="SS" />
  <word id="6" form="(" postag="IR" head="5" deprel="IR" />
  <word id="7" form="särbeskattning" postag="VN  SS" head="5" deprel="AN" />
  <word id="8" form=")" postag="IR" head="5" deprel="JR" />
  <word id="9" form="av" postag="PR" head="5" deprel="ET" />
  <word id="10" form="arbetsinkomster" postag="NN  SS" head="9" deprel="PA" />
  <word id="11" form="." postag="IP" head="3" deprel="IP" />
  </sentence>
```

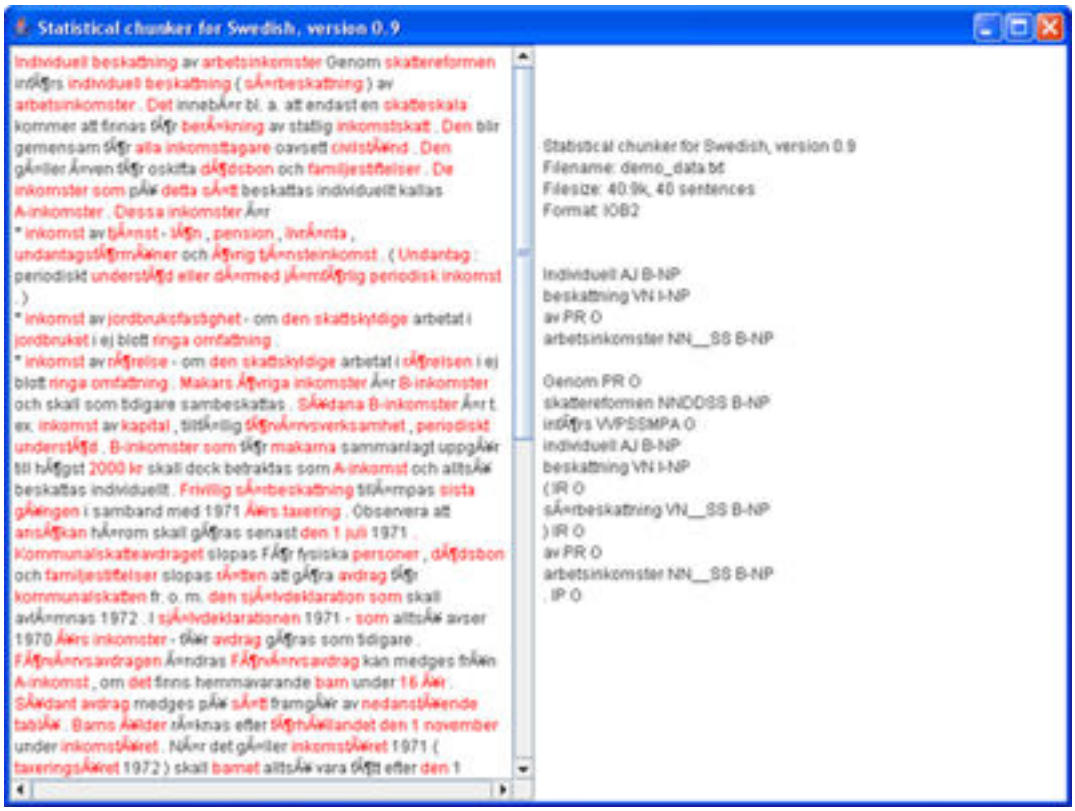
2. The new format

```
Genom PR O
skattereformen NNDDSS B-NP
införs VVPSSMPA O
individuell AJ B-NP
beskattning VN I-NP
( IR O
särbeskattning VN__SS B-NP
) IR O
av PR O
arbetsinkomster NN__SS B-NP
. IP O
```

3. The accuracy results from the statistical model

```
accuracy:  88.33%; precision:  64.20%; recall:  75.38%; FB1:  69.34
                  NP: precision:  64.20%; recall:  75.38%; FB1:  69.34  4182
```

4. The GUI


Statistical chunker for Swedish, version 0.9
Individuell beskattning av arbetsinkomster Genom skattereformen införs individuell beskattning (särbeskattning) av arbetsinkomster . Det innebär bl. a. att endast en skatteskala kommer att finnas för beräkning av statlig inkomstskatt . Den blir gemensam för alla inkomsttagare oavsett civilstånd . Den gäller även för oskifta dödsbon och familjeställelser . De inkomster som på detta sätt beskattas individuellt kallas A-inkomster . Dessa inkomster är
* inkomst av tjänst - lön , pension , livränta , undantagsförmåner och övrig tjänsteinkomst . (Undantag : periodiskt understödd eller därmed jämförlig periodisk inkomst .)
* inkomst av jordbruksfastighet - om den skattskyldige arbetat i jordbruket i ej blott ringa omfattning .
* inkomst av rörelse - om den skattskyldige arbetat i rörelsen i ej blott ringa omfattning . Makars övriga inkomster är B-inkomster och skall som tidigare sambeskattas . Sådana B-inkomster är t. ex. inkomst av kapital , tillfällig förvärvsverksamhet , periodiskt understödd . B-inkomster som för makarna sammanlagt uppgår till högst 2000 kr skall dock betraktas som A-inkomst och alltså beskattas individuellt . Frivillig särbeskattning tillämpas sista gången i samband med 1971 års taxering . Observera att ansökan härom skall göras senast den 1 juli 1971 . Kommunalskatteavdraget slopas FÖR fysiska personer , dödsbon och familjeställelser slopas rätten att göra avdrag för kommunalskatten fr. o. m. den självdeklaration som skall avlämnas 1972 . I självdeklarationen 1971 - som alltså avser 1970 års inkomster - får avdrag göras som tidigare . FÖRvärvsavdragen ändras FÖRvärvsavdrag kan medges från A-inkomst , om det finns hemmavarande barn under 16 år . Sådant avdrag medges på sätt framgår av nedanstående tabell . Barns ålder räknas efter förhållandet den 1 november under inkomståret . När det gäller inkomståret 1971 (taxeringsåret 1972) skall barnet alltså vara född efter den 1
Statistical chunker for Swedish, version 0.9
Filename: demo_data.txt
Filesize: 40.9k, 40 sentences
Format: IOB2
Individuell AJ B-NP
beskattning VN I-NP
av PR O
arbetsinkomster NN__SS B-NP
Genom PR O
skattereformen NNDDSS B-NP
införs VVPSSMPA O
individuell AJ B-NP
beskattning VN I-NP
(IR O
särbeskattning VN__SS B-NP
) IR O
av PR O
arbetsinkomster NN__SS B-NP
. IP O