Identification of rhetorical words in swedish

Petter Bergman, c01pb@efd.lth.se

January 17, 2005

1 Introduction

For some words in Swedish, two distinct usage patterns can be seen. Like, for instance, in the following sentences, the words we shall discuss are written in italics.

1. taket är mörkbrunt *liksom* dom enkla borden
2. jo det skulle *liksom* vara jag det
3. dom är av samma *typ*
4. vi skulle *typ* gå dit och hämta honom

Sentences 1 and 3 examplifies what we would call "normal" use of the words, this use is commonly seen as more "correct" Swedish. Sentences 2 and 4 contain a different use, the word in question is used more as a pause. Note that in sentences 1 and 3, removing the words distort the sentences. In sentences 2 and 4, the removal of the words will leave the meaning of the sentence intact.

From here, the use in sentences 1 and 3 is called "grammatical", and the one in 2 and 4 "rhetorical". Can we classify the words automatically from examining its context in a sentence?

1.1 Part-of-Speech

In sentence 1, "liksom" is a conjunction, in sentence 3 it is an adverb: In sentence 2, "typ" is a noun, in sentence 4 it is an adverb[2].

Using this information we could classify the words, but then we would have to know it's Part-of-Speech.

The task of classifying the words can be seen as a subtask of disambiguating in a Part-of-Speech tagger.

1.2 Spoken vs Written language

Rhetorical words occur almost exclusively in spoken language, and at least grammatical "liksom":s occur almost exclusively in written language. So are we just disambiguating between written and spoken language (or rather, spoken language translated into written language, which has it's own problems)?

2 The Method

To find the use of a word in a text is to find the u which maximizes:

$$P(u|W_1, W_2)$$

$$u \in \{Grammatical, Rhetorical\}$$

because:

$$P(A|B)P(B) = P(B|A)P(A)$$

we can instead maximize:

$$P(u)P(W_1, W_2|u)$$

we pretend that $P(Grammatical) = P(Rhetorical)$ and that words occur independently:

$$\text{argmax} \prod_{(w_1,w_2)\in W_1,W_2} P(w_1, w_2|u)$$

We estimate $P(w_1, w_2|r)$ by counting words from a manually tagged corpus. We would get a lot of zero counts so we use Laplace estimates to cope with the sparse data.

3 Implementation

The implementation consists of three tools written in O'Caml:

collect counts occurences of a word in it's grammatical/rhetorical use from a hand-annotated file.

tag tags occurences of a word using collected data

eval evaluates the output by comparing it to a file hand-annotated with the correct use.

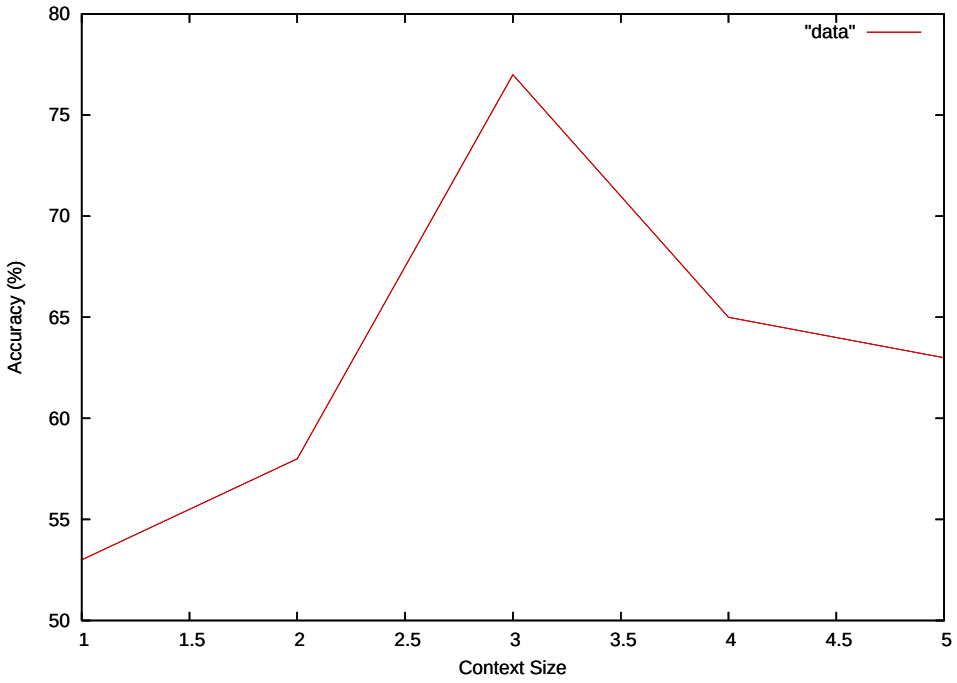

Figure 1: Results of the test-run

4 Results

For evaluation, a training set was compiled from [4] and [3] in such a way that it contained 50 occurences of liksom, hand annotated. Data was then collected from this text, using different context sizes.

In a similiar way, the (not annotated)test set was compiled. The results of running the tagger on the test set for different context sizes is shown in figure 1.

The behaviour is as expected, although slightly better than expected for such sparse data.

5 Final Comments

In spite of the result from the test run there are some concerns.

We make the asumption that "liksom" occurs an equal number of times in the rhetorical use as in the grammatical use, we then proceed to collect a test and training set making this true. What is a balanced corpus in this case? it should be a sample of the body of all swedish, written or spoken. Spoken swedish is is hard to sample, do we mean all utterances ever made in swedish? or just contemporary swedish?

We could look at the POS of the context instead of the words themselves

and reduce the sparse-data problem. But if we know the POS of every word in the text, we also know if our examined word is rhetorical or not.

References

[1] Daniel Marcu and Abdessamad Echihabi, "An Unsupervised Approach to Recognizing Discourse Relations" Information Sciences Institute and Department of Computer Science, University of Southern California

[2] "Nordstedts Svenska Ordbok" Nordstedts Förlag 1990

[3] Allwood J., Björnberg M., Grönqvist L., Ahlsén E., Ottesjö C. "The Spoken Language Corpus at the Linguistics Department" Göteborg University in Forum Qualitative Social Research, vol 1, no 3 - Dec 2000

[4] Philipp Koehn "Europarl: A Multilingual Corpus for Evaluation of Machine Translation" Draft, Unpublished