Automatic learning of discourse relations in Swedish

Stefan Karlsson
Department of Computer Science
Lund University
f98sk@efd.lth.se

Abstract

This report describes some experiments with a statistical technique that extracts discourse relations in Swedish. Three types of relations are used: Cause-Explanation-Evidence (CEV), Contrast and Elaboration. The method is evaluated by building two-way classifiers, with the results: Contrast vs. CEV 64.5%, Contrast vs. Elaboration 56.6% and CEV vs. Elaboration 54.9%. The conclusion is that the technique, with improvements or modifications, seems to be usable to extract discourse relation in Swedish, but further investigations are necessary.

1 Credits

Most of the corpora used in this project have been provided by Lars Aronsson from the Runeberg project. As a tool to identify nouns and verbs the Granska grammatical tool was used, which was provided by Jonas Sjöbergh at KTH. Finally the Stockholm-Umeå corpus was needed to build the Granska tool, and was provided by Sofia Gustavfson-Capkova also at KTH.

2 Introduction

There are relations in discourse, for example “I did put my coat on this morning, because it was cold”. The second clause of the sentence is describing a cause of the fact stated in the first clause. Another example is the contrast: “The car is green on one side, but red on the other”. No good techniques have yet been developed to automatically identify different types of discourse relations. A useful application would perhaps be to extract all causes from discourses and put them into a knowledge base. For example you could ask: “What causes crops to grow?”, and a knowledge base agent would answer “the sun”, “the earth” and “rain”.

In many cases there are markers that indicate a particular discourse relation, as in the examples above “because” and “but”. However you could also say for example “The car is green on one side, and red on the other”. A human might conclude that the car being red on one side is in contrast to the car being green on the other side, but since it’s not explicitly marked, some more elaborate reasoning strategy is needed.

This project work was an attempt to implement a algorithm that makes a choice whether two word spans in Swedish can be classified as together constituting a particular discourse relation. The algorithm has been developed by Marku and Echihabi (2002), and is here implemented for and tested on Swedish discourse.

3 A statistical model

The approach taken by Marcu and Echihabi (2002) was to build a simplistic statistical model. Basically there might be some word pairs that are frequent in contrasts, for example “green” and “red” as in the example above, and other pair in explanations, i.e. “cold” and “coat”. To capture these patterns a table can be made, where the word pairs formed from each different combination of words from the first and second part are counted. This is the cartesian product of the text spans W_1 and W_2, defined as $(w_i, w_j) \in W_1 \times W_2$. The table becomes non-commutative since the first text span, so to speak, ends up in the columns of the table and the second text span ends up in the rows.

The probability that two text spans forms a particular relation can be calculated as follows:

$$P(r_k|W_1, W_2) = \frac{P(W_1, W_1|r_k)P(r_k)}{P(W_1, W_1)}. \quad (1)$$

The factor $P(W_1, W_1|r_k)$ can be estimated with $\prod P((w_i, w_j)|r_k)$, were w_i and w_j symbolizes the words in each span. $P((w_i, w_j)|r_k)$ is directly calculated from the table.

Of course not all possible word pairs, that might be encountered in for example contrasts, will be counted in the table, which makes it necessary to shift some probability mass to these previously unseen pairs. Marcu and Echihabi used the Laplace method.

4 Extraction of sentences

First of all three discourse relations were used in the experiment. These are Cause-Explanation-Evidence (CEV), Contrast and Elaboration. [1]

To be able to experiment with the technique, Swedish corpora were attained from the Runeberg project (45 million words) and the European Parliament (16 million words). The first difficulty was to find good Swedish markers for extraction of training examples. By inspection of sentences the extraction patterns in table 1 were judged to be good enough. Since a great portion of the corpora was from Nordisk Familjebok from the end of the 19th century and the beginning of the 20th century, some old markers were used (“ty” and “ehuru”), together with some more modern ones (“eftersom” and “trots att”).

The corpus was divided into a training set (99.5%) and a test set (0.5%). Results from Marcu and Echihai (2002) indicate that using only nouns and verbs makes a steeper training curve, and since quite a small corpus was used this approach was taken. For this purpose the Granska grammatical tool [2] was used. All non-nouns and non-verbs were identified and marked, and later discarded in all experiments. The Granska tool also helped to identify sentences in the corpora.

Contrast
[BOS ...][, men ... EOS]
[BOS ...][, ehuru ... EOS]
[BOS ...][, fastän ... EOS]
[BOS ...][, trots att ... EOS]
Cause-Explanation-Evidence
[BOS ...][, därför att ... EOS]
[BOS ...][, eftersom ... EOS]
[BOS ... EOS][BOS Alltså ... EOS]
[BOS ...][, alltså ... EOS]
[BOS ... EOS][BOS Således ... EOS]
[BOS ...][, således ... EOS]
[BOS ... EOS][BOS Sålunda ... EOS]
[BOS ...][, sålunda ... EOS]
[BOS ...][, ty ... EOS]
[BOS ... EOS][BOS Ty ... EOS]
[BOS ... EOS][BOS Därför ... EOS]
Elaboration
[BOS ...][vilket ... EOS]

Table 1: Swedish extraction patterns used in the experiments.

Finally the training examples were extracted with 156762 Contrasts, 43159 CEV and 21072 Elaborations, and the testing set with 771 Contrasts, 176 CEV and 79 Elaborations.

5 Experiments

5.1 Methods of evaluation

To evaluate the technique, two-way classifiers were built to distinguish Contrast vs. CEV, Contrast vs. Elaboration and CEV vs. Elaboration. A decision is made by taking the maximum of $P(r_k|W_1, W_2)$ for each relation, where $P(r_k|W_1, W_2)$ is calculated from the table. In

[1] For a description of these, see Marku and Echihabi (2002) and literature referenced to from there.

[2] see www.nada.kth.se/theory/projects/granska/

equation 1, $P(W_1, W2)$ can be discarded, since it's the same for all relations.

An approach that would eliminate the factor $P(r_k)$, would be to put the same amount of sentences from each type of relation in the test set. However, since there were as few as 71 Elaborations in the test set, and as many as 771 Contrasts, this was instead emulated by taking the mean of the amount of correctly classified sentences from each relation. In this way the result is more statistically validated. So, for example in the case of Contrast vs. CEV, all contrast sentences from the test set were classified and the percentage of correctly classified sentences was calculated. The same thing was done for CEV sentences and finally the mean of these percentages taken as the result. This is similar to having the same proportions of contrasts and CEVs, meaning that $P(r_{contrast}) = P(r_{CEV})$ and this factor can be discarded.

During early experiments it was discovered that the Laplace method seemed to shift too much mass of probability to unseen word pairs. In one table there were 900 times as many zeros as actual counts in the table. Therefore Lidstone's rule was used instead, which amounts to setting:

$$P((w_1, w_2)|r_k) = \frac{(count + \lambda)}{(total + \lambda \cdot cardinal)}, \quad (2)$$

where cardinal is the number of entries in the table. It was found that a lambda of 0.05 seemed to maximize the accuracy of the classifiers; a value that was kept during all subsequent experiments.

5.2 Results and an improvement

The accuracy of the classifiers are presented in table 2. A maximum result of 64.5% in the Contrast vs. CEV condition is in the same realm as the results from Marcu and Echihai (2002), who had between 60% and 70% in most conditions.The results for Elaboration were worse, with 57% in the Contrast vs. Elaboration case.

During these experiments one thing wasn't considered however. The method so far has been consisting of using a left and a right part in the training examples. A problem with this can be illustrated with the two sentences "I put my coat on, because it is cold" and "It is cold, thus I put my coat on". Both are causal and basically state the same thing, but the order of the clauses are reversed. In the first case the training procedure would take "I put my raincoat on" as the first clause of the sentence and "it is cold" as the second, but in the opposite order in the second case. Perhaps it would be better to identify the cause and the effect in each sentence and train the table with these instead of using the left and right part of the sentences. To test this hypothesis the training with CEV relations was done again, but with the word spans put in logical order, that is causes to the left and effects to the right. The result was that 65.1% of the sentences were correctly classified in the contrast vs. CEV case.

	Contrast	Elaboration
CEV	64.5 %	54.9%
Contrast		56.6%

Table 2: The result of each evaluation.

5.3 A source of bias

Since texts from different time-spans were used, i.e. Nordisk Familjebok and Svenska Familj-journalen from the end of the 19th and beginning of 20th century, and modern discourse from the European Parliament, the results above might have been biased. At worse the classifiers only differentiates between old and new sentences, and not at all between different kinds of discourse relations. The problem is clear since in older Swedish, as in Nordisk Familjebok, "v" is often spelled "f" or "fv". For example "av" becomes "af" and "silver" becomes "silfver". Other differences might also influence the classifiers. It was actually found that using the contrast vs. CEV classifier on old vs. new sentences gave a result of 77%.

In an attempt to get around this problem, the tests sets were further divided. Sentences from Nordisk Familjebok and Svenska Familj-journalen were separated from sentences from the European Parliament, and old and new sets were thus formed. The CEV vs. contrast classifier was evaluated for each case. The results were 60% correctly classified sentences from the old set and 58% from the new.

6 Conclusions

Results around 60% clearly indicated that the classifier is better than random assignment of text spans to each class. Of course the accuracy is not high, but since marker words themselves are not used to train the classifiers, the accuracy can probably be increased tremendously. The corpora used were also quite small, and more training examples are needed to increase accuracy.

The results with elaboration are not very good, which first of all can be accounted to the fact that only 21072 training examples were used. Also, as Marcu and Echihabi (2002) states, there is some dispute regarding the existence of a well formed category of elaboration as a discourse relation. Perhaps this can be used as a way to experimentally find evidence for well formed categories.

The fact that two quite different types of discourse was used, that is old and new, had a great impact. Since the Contrast vs. CEV classifier was better at discriminating between old and new sentences than Contrasts and CEVs, the bias introduced was quite large. A conclusion might be that the method works better on restricted discourses. For example a classifier trained on technical reports would be very good at identifying discourse relations in other technical reports.

There are some simple improvements that could be made. For example, since there seems to be no intrinsic order in contrast relations the table should be made commutative, that is training both “forward” and “backward”. This was however not judged as a critical point, since there were more than 150000 training examples of contrasts. Another thing is to find the value of lambda in Lidstones rule, that maximized the accuracy. Using the value of 0.05, instead of 1 as in the Laplace method, greatly improved the classifiers, and more improvement can probably be made.

To sum up, some evidence is presented that this constitutes a feasible technique for automatic extraction of discourse relations in Swedish, at least with some improvements, but further investigations would be necessary to accurately evaluate this contention.

References

Daniel Marcu and Abdessamad Echihabi. 2002. An Unsupervised Approach to Recognizing Discourse Relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics ACL-2002* (available at www.isi.edu/m̃arcu/papers/relations-acl2002.pdf), Philadelphia, PA, July 7-12