Automatic Learning of Discourse Relations in Swedish Using Cue Phrases

Stefan Karlsson and Pierre Nugues

Lund University
Lund Institute of Technology
Department of Computer Science
Box 118
S-221 00 Lund, Sweden
stefan.karlsson.342@student.lth.se, Pierre.Nugues@cs.lth.se

Abstract. This paper describes experiments to extract discourse relations holding between two text spans in Swedish. We considered three relation types: cause-explanation-evidence (CEV), contrast, and elaboration and we extracted word pairs eliciting these relations. We determined a list of Swedish cue phrases marking explicitly the relations and we learned the word pairs automatically from a corpus of 60 million words. We evaluated the method by building two-way classifiers and we obtained the results: Contrast vs. Other 67.9%, CEV vs. Other 57.7%, and Elaboration vs. Other 52.2%.
The conclusion is that this technique, possibly with improvements or modifications, seems usable to capture discourse relations in Swedish.

Key words: rhetorical relations, discourse relations, cue phrases, naïve Bayes classification.

1 Introduction

Rhetorical relations and the *Rhetorical structure theory* [1] form a framework to describe and interpret the organization of a text. In this theory, relations consist of annotated links tying two text spans as, for example, the clauses in the sentence:

> Malaria förekommer framför allt i sumpiga trakter, därför att mygglarverna utvecklas väsentligen i stillastående vattensamlingar.
> “Malaria exists primarily in wetlands, because the mosquito larvae develops in still waters.”
> [2, Uggleupplagan, vol. 8:1490].

The next sentence gives another example of a rhetorical relation between two clauses:

> Till en början utgafs tidningen en gång i veckan, men i dec. 1850 förvandlades den till daglig

> "Initially the newspaper was published once a week, but in Dec. 1850 it was transformed into a daily"
> [2, Uggleupplagan, vol. 3:1157].

Rhetorical relations can be associated with certain cue words or phrases, such as *därför* 'because' with explanations in the first example and *men* 'but' with contrasts in the second one. Nonetheless, cue phrases are often ambiguous. If you change *but* to *and* in the second example, a reader would probably conclude that the relation tying the two spans remains the same. However, since *and* is not a discourse marker as explicit as *but*, it cannot be used in a one-to-one association to identify a relation. A more elaborate strategy is then necessary to extract and label rhetorical relations.

First techniques to automatically identify different types of discourse relations used discourse markers and were based on manually-written rules as in [3] and [4]. Most algorithms described in the literature have only been applied to English or Japanese.

This paper describes a system that decides whether two text spans in Swedish can be classified as being tied by a particular discourse relation. In this system, we implemented and adapted Marcu and Echihabi's algorithm [5], which automatically learns relations from minimally annotated texts. A useful application of the analysis of rhetorical relations would be to extract all causes of a fact and put them into a knowledge base.

2 A Statistical Model

Some word pairs are frequent in contrasts, hypothetically for example, *week* and *daily*, as in the example above, and other pairs in explanations, i.e. *exists* and *develops*. Instead of extracting relations with manual rules, we can try to derive automatically sets of words involved in specific relations from corpora.

Marcu and Echihabi proposed an unsupervised method [5] to train naïve Bayesian classifiers based on this idea. The first step extracts contiguous text spans using a set of predefined markers and forms the Cartesian product of the words in them. Let W_1 and W_2 be two contiguous text spans. The second step counts all the word pairs $(w_i, w_j) \in W_1 \times W_2$ of the contiguous text spans extracted from the corpus.

The probability that two text spans are tied by a particular relation is calculated as follows:

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

Using the naïve Bayes strategy, we estimate $P(W_1, W_2|r_k)$ as $\prod P((w_i, w_j)|r_k)$, where w_i and w_j stand for the words in each span.

3 Experimental Setup

3.1 Extraction of Text Spans

We considered three discourse relations: *cause-explanation-evidence* (CEV), *contrast*, and *elaboration*. We compiled a Swedish corpus using texts from the Runeberg project (45 million words) and the European Parliament proceedings [6] (16 million words), a multilingual corpus, where we used the Swedish source parts. We then inspected the corpus manually and incrementally built the extraction patterns shown in Table 1.

Table 1. Swedish extraction patterns used in the experiments. BOS indicates the beginning of the sentence and EOS, the end of the sentence.

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]
[BOS ...][hvilket ... EOS]

The *Nordisk Familjebok* encyclopedia from the end of the 19th century and the beginning of the 20th century represents a large part of the corpus. This explains why we had to use words like *ty* 'because' and *ehuru* 'in spite of' as markers that do not belong to present day Swedish.

The corpus was randomly divided into a training set (90%) and a test set (10%). To improve training, we used only verbs and nouns [5]. We tagged the corpus words with their part of speech using the Granska tagger [7]. We kept the nouns and the verbs and discarded the rest of the words, including the markers from the patterns.

Finally, we compiled the training examples: 130,796 contrasts, 37,319 CEV, and 43,387 elaborations, and a test set of 14,643 contrasts, 4,107 CEV, and 4,976 elaborations, all extracted using the patterns in Table 1.

3.2 Evaluation Methods

For the evaluation, we built binary classifiers to distinguish:

- Contrast vs. Other,
- CEV vs. Other, and
- Elaboration vs. Other,

where Other stands for an equal amount of relations of the other two types as CEV+Elaboration in the first case. A decision is made by taking the maximum of $P(r_k|W_1, W_2)$ for each relation. In Equation 1, $P(W_1, W2)$ can be discarded, since it is the same in all the relations.

In the evaluation, we build sets of equal proportions to eliminate the factor $P(r_k)$. In the Contrast vs. Other case, we extracted 8,000 contrasts, 4,000 CEVs, and 4,000 elaborations from the test set. In the CEV vs. Other case, we used 4,000 CEVs, 2,000 contrasts, and 2,000 elaborations. Finally in the Elaboration vs. Other case, we used 4,000 elaborations, 2,000 contrasts, and 2,000 CEVs.

We found that the Laplace method shifted too much mass of probability to unseen word pairs. Therefore, we used Lidstone's rule instead, which amounts to setting [8]:

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

where *cardinal* is the number of entries in the table. We found that a lambda of 0.05 seemed to maximize the accuracy of the classifiers. In a similar experiment, [9] used the value of 0.25.

3.3 Results

Table 2 shows the accuracy of the classifiers. A result of 67.9% in the Contrast vs. Other condition is in the same range as the results obtained for English [5], which reported between 60% and 70% for most relations. The results for Elaboration vs. Other that reached 52.2% were significantly lower, however.

4 Conclusions

Results around 60% clearly indicates that the classifier is better than a random assignment of text spans to each class. The result with elaboration is not completely satisfying though, which first of all can be accounted to the fact that we only used 43,387 training examples.

Table 2. The accuracy of each classifier. In each case, the baseline is 50%.

Relation	Accuracy
Contrast vs. Other	67.9%
CEV vs. Other	57.7%
Elaboration vs. Other	52.2%

As perspectives, some simple improvements could be made. Since there is no intrinsic order in contrast relations, the table could be made commutative. However, we did not consider it a critical point, since there were more than 130,000 training examples of contrasts. The most critical point though is to find the best set of cues phrases for each discourse relation. The corpora used in this experiment was quite small for the task and we had to use many cue phrases at one time. With a larger training set, we could determine which phrases contribute most to the model without introducing noise; for example by comparing results obtained by including or excluding training examples from a particular extraction pattern.

Not only the size of the corpora limits the performance of this technique. The example words that indicate a contrast, i.e. *week* and *daily* in the example in Sect. 1 can possibly stand in other types of discourse relations. Such overlapping word pairs will dim the statistical accuracy of the model no matter the size of the corpora. This is a major limitation of the general approach taken and can only be dealt with by introducing other types of classification information to distinguish between the rhetorical relations. In English, possibly WordNet [10, 11] or FrameNet [12] could be used to figure out which word pairs indicate a particular relation.

To sum up, we presented evidence of a feasible technique for the automatic extraction of discourse relations in Swedish. Marcu and Echihabi showed [5] that using this technique as a complement to extracting cue-phrase marked sentences, can increase the number of correctly classified contrasts from 26% to 77%. Further investigations are however necessary to evaluate more accurately the applicability of this algorithm in Swedish.

Acknowledgments. Lars Aronsson provided most of the corpora used from the Runeberg project. Jonas Sjöbergh at Kungliga Tekniska Högskolan provided the Granska grammatical tool that we used to identify nouns and verbs. Finally, the Granska tool was trained on the Stockholm-Umeå corpus [13].

References

1. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: A theory of text organization. Technical Report RS-87-190, Information Sciences Institute (1987)
2. Meijer, B., ed.: Nordisk familjebok. Uggleupplagan edn. Nordisk familjeboks förlags aktiebolag, Stockholm (1904–1926)

3. Kurohashi, S., Nagao, M.: Automatic detection of discourse structure by checking surface information in sentences. In: Proceedings of the 15th International Conference on Computational Linguistics, COLING-94. Volume 2., Kyoto (1994) 1123–1127
4. Corston-Oliver, S.: Computing Representations of the Structure of Written Discourse. PhD thesis, University of California, Santa Barbara (1998)
5. Marcu, D., Echihabi, A.: An unsupervised approach to recognizing discourse relations. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics ACL-2002, Philadelphia (2002) 368–375
6. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proceedings of The Tenth Machine Translation Summit, Phuket, Thailand (2005)
7. Carlberger, J., Kann, V.: Implementing an efficient part-of-speech tagger. Software Practice and Experience **29** (1999) 815–832
8. Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA (1999)
9. Blair-Goldensohn, S., McKeown, K.R., Rambow, O.C.: Building and refining rhetorical-semantic relation models. In: Proceedings of NAACL HLT 2007, Rochester, NY (2007) 428–435
10. Miller, G.A.: WordNet: A lexical database for English. Communications of the ACM **38** (1995) 39–41
11. Fellbaum, C.: WordNet: A Lexical Database for English. MIT Press, Cambridge, MA (1998)
12. Ruppenhofer, J., Baker, C.F., Fillmore, C.J.: The FrameNet database and software tools. In Braasch, A., Povlsen, C., eds.: Proceedings of the Tenth Euralex International Congress. Volume 1., Copenhagen, Denmark (2002) 371–375
13. Ejerhed, E., Källgren, G., Wennstedt, O., Åström, M.: The linguistic annotation system of the Stockholm-Umeå project. Technical report, University of Umeå, Department of General Linguistics (1992)