Summary of Thesis Contents

Thesis Title:
A Study on Semi-Supervised Approaches for Discourse Analysis
（談話解析への半教師付き学習アプローチ）

エルノー　ユーゴ

The understanding of discourse structure has several important applications in Natural Language Processing. In particular, it facilitates the automatic creation of text summaries, enables dialogue generation, and supports question-answering. However, the development of automatic discourse analysis systems is facing several challenges. First, natural language texts must be accurately segmented into elementary text units. Second, a data structure should be built, representing the way all the units of a text are organized. Finally, and most important, discourse relations indicating in what manner pairs of units of the text are related, have to be determined. This thesis addresses the problems of (1) segmenting a text efficiently into units, (2) developing a full-fledged discourse analyzer able to produce tree discourse structures, and (3) creating semi-supervised discourse relation classifiers, which is an important perspective for creating discourse analyzers working on domains with a lack of annotated training data.

In the first section of the thesis, we introduce a sequential discourse segmentation method based on Conditional Random Fields. Segmenting a text into elementary discourse units is the first step of all discourse analyzers. Because improper segmentation jeopardizes the rest of the discourse analysis process, this task is of paramount importance. We employ Conditional Random Fields to train a discourse segmenter on the RST Discourse Treebank, one of the major annotated discourse corpora, using a set of lexical and syntactic features. The proposed method is compared to other statistical and rule-based segmenters, including one based on Support Vector Machines. Experimental results indicate that the proposed sequential model outperforms current state-of-the-art discourse segmenters, with an F-score of 0.94. This performance level is close to the human agreement F-score of 0.98.

In the second section of the thesis, we present an implemented discourse analyzer based on Support Vector Machine classification. Previous supervised approaches were aimed at producing sentence level analysis or at describing partially-implemented systems. By contrast, our system targets discourse

structure at text level. Specifically, we created a fully-implemented, extensively-evaluated system.

In the next sections of the thesis, we focus more particularly on semi-supervised methods for performing discourse relation classification, which is the core task of a discourse analyzer. For this task, most researchers have employed fully-supervised machine learning methods. In these approaches, a large annotated discourse corpus is employed, and used as a basis to train a discourse relation classifier able to recognize which discourse relation holds between two units of text given as input. However, only three annotated discourse corpora are available for training discourse relation classifiers. Since different applications and domains might require employing a custom set of discourse relations, it becomes necessary to create each time a new training corpus, which is costly and time-consuming. To tackle this issue, we propose to employ semi-supervised machine learning methods, whereby a small amount of labeled training data is combined with a large amount of freely-available, low-cost unlabeled training data, in order to train a classifier with increased performance.

In particular, in the third section of the thesis, we introduce a semi-supervised discourse relation classification method based on the analysis of co-occurring features in unlabeled data. This information is then taken into account for extending the feature vectors given to a classifier. Our experimental results on the RST Discourse Treebank corpus and Penn Discourse Treebank indicate that the proposed method brings a significant improvement in classification accuracy and macro-average F-score when small training datasets are used. For instance, with training sets of ca. 1000 labeled instances, the proposed method brings improvements in accuracy and macro-average F-score up to 50% compared to a baseline classifier. We believe that the proposed method is a first step towards improving classification performance for small datasets, with potentially infrequent discourse relations, which is useful for domains with a lack of annotated data.

In the fourth section of the thesis, we present a different semi-supervised approach to discourse relation classification, based on learning from multiple auxiliary problems (Structural Learning). First, we solve a set of auxiliary classification problems using unlabeled data. Second, the learned classifiers are used to extend feature vectors and train a discourse relation classifier. We show that, when using the same feature set and unlabeled data set as in the co-occurrence-based method introduced in the third section, Structural Learning-based discourse relation classification reaches similar performance levels. Compared to the co-occurrence-based discourse relation classification method, the Structural Learning-based method has the advantage of requiring few additional features, typically ca. 50, while the co-occurrence-based method requires extensive dimension increase of ca. 15000. This is another positive prospect for training discourse relation classifiers on domains where little labeled training data is available, thus potentially enabling novel applications.

Finally, in the last section of the thesis, we conclude our study, summarize its main points, and discuss future work and potential research directions.