Document Analysis and Summarization Workbench

Yuji Matsumoto, Takashi Miyata
Nara Institute of Science and Technology
Email: {matsu,takashi}@is.aist-nara.ac.jp

Tadashi Nomoto
National Institute of Japanese Literature
Email: nomoto@nijl.ac.jp

Takenobu Tokunaga
Tokyo Institute of Technology
Email: take@cl.cs.titech.ac.jp

Makoto Takeda, Masaharu Obayashi
Kanrikogaku Kenkyusho, Ltd.
Email: {ma-take,obayashi}@kthree.co.jp

Keywords: Document Structure Analysis, Text Summarization, XML, Graphical User Interface, Annotation

1 Introduction

Although most of text summarization tasks are carried out on the basis of sentence extraction that relies on the importance measure of lexical items appearing in the sentences, further linguistic analyses are inevitably necessary for better and more comprehensible summarization. In such a more sophisticated system, we have to take into account fine-grained linguistic information such as syntactic dependency between words or phrases, rhetorical structure of documents, coreference relations among entity descriptions. While most of the existing natural language analysis systems are not accurate enough to achieve satisfactory linguistic analysis of real world documents, it is important to know how far such linguistic information are useful and effective in text summarization.

We are currently developing a system for extending the study of language analysis and text summarization techniques considering the structure and fine-grained linguistics analyses of target documents. Recent researches on statistical natural language processing have shown that statistical learning approach is quite useful and can be applied to various applications, such as syntactic dependency analysis (Fujio 1998), coreference resolution (Soon 1999), and rhetorical structure analysis (Marcu 1999)(Nomoto 1999). The crucial point of the approach is the availability of large scale annotated corpora. Besides POS-tagged corpora, few annotated corpora are available which are large enough to extract statistically significant information. The aims of our system is two-fold. First is to provide an environment for consistent construction of the corpora annotated by the following linguistic information:

- Chunking or base phrase analysis
- Syntactic dependency among chunks
- Coreference among entity descriptions
- Rhetorical structure of documents

Once annotation is done automatically or manually to the documents, our second aim is to investigate the use of linguistic clues for text summarization and to provide an integrated environment to support such investigation. While the user has to devise the way to use linguistic information in defining the importance of various entities in a document, the system graphically show the results of summarization in the course of application of user-provided techniques.

While the main target language is Japanese, all the systems are designed as language independent and the system is easily applicable to any other languages.

2 System Overview

The system consists of five modules for text analysis and summarization; a statistical part-of-speech tagger, a statistical dependency analyzer, a document structure analyzer, an editing system which corrects analysis errors through a graphical user interface, and a summarization system which integrates the results of other modules. Those modules are used to give linguistic annotation to the input documents and to perform experiments on text summarization based on linguistic information. Besides those modules, we developed an interface for database access for retrieval of annotated documents.

Since each module exchanges their data in an XML format, they can be easily transformed into another format. The fact that the modules currently under construction work independently of language allows the user to apply the system to multi-lingual domains. Figure 1 illustrates the organization of the system.

3 Module Description

3.1 Parsers

Tokenization and POS tagging is done by Japanese morphological analyzer ChaSen (Matsumoto et al 1999). For syntactic analysis we use statistical dependency parser ChaGake (Fujio 1998). Both systems learn the statistical parameters from an

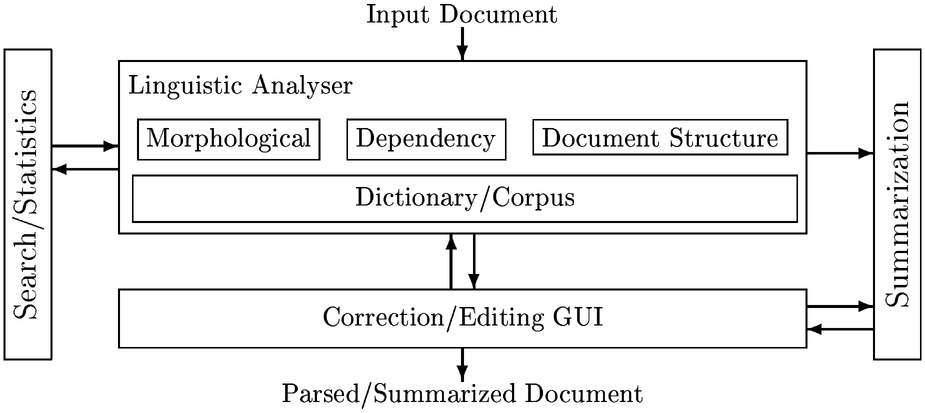

Figure 1: System Overview

annotated corpus. Although automatic rhetorical structure analysis of documents is difficult, we developed a decision tree learning model for local inter-sentence analysis (Nomoto 1999). The output of the systems, tagged in an XML format, is displayed and can be modified through a GUI with simple mouse operations (cf section 3.3).

3.2 Correction/Editing

For each type of annotation (POS tags, dependency, rhetorical and coreference relations), a GUI displays the annotated results produced by the analyzers. The user can browse the results and correct annotation errors by simple mouse operations. The annotated corpus becomes the training data for further improvement of the analyzers as well as the experimental data for the summarization task.

3.3 Summarization

In the current system, text summarization is approximated in two steps; sentence extraction and sentence reduction, both of which are controlled by the importance values allocated to words, chunks and sentences. The definition and the calculation scheme of the importance values are to be given by the user. The system, obtaining the importance value for each constituent of the input text, shows the important parts by eliminating the parts that have lower importance value than a user-defined threshold. The threshold can be altered by a very simple operation of sliding gauges. The facility of the system is quite simple, but it is surely useful for the purpose of experimenting and developing text summarization techniques.

3.4 Search/Statistics

While this module does not directly related with the summarization task, searching and browsing facilities for annotated corpus are useful especially when the corpus is annotated with various kinds of tags into which the original text is buried. We provide a retrieval system for annotated corpora, in which user can describe a query as a linguistic pattern using POS or dependency relations. The XML annotated data is stored into a relational database and a query is compiled into an SQL formula to obtain the matched data. Statistical information such as the count of patterns can be displayed as well.

Acknowledgement

This project is supported by the Next Generation Application Project under the auspices of Information-technology Promotion Agency, Japan (IPA).

References

Fujio, M. and Matsumoto, Y., 1998. Japanese Dependency Structure Analysis based on Lexicalized Statistics. *Proc. 3rd EMNLP*, 88–96.

Marcu, D., “A Decision-based Approach to Rhetorical Parsing,” 37th Annual Meeting of the ACL, pp.365-372, 1999.

Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H. and Asahara, M., 1999. *Japanese Morphological Analysis System ChaSen 2.0 Users Manual, 2nd edition.* Technical Report NAIST-IS-TR99012, Nara Institute of Science and Technology.

Nomoto, T. and Matsumoto, Y., “Learning Discourse Relations with Active Data Selection,” 1999 Joint SIGDAT Conference on EMNLP and VLC, pp.158-167, 1999.

Soon, W.M., Ng, H.T. and Lim, C.Y., “Corpus-based Learning for Noun Phrase Coreference Resolution,” 1999 Joint SIGDAT Conference on EMNLP and VLC, pp.285-291, 1999.