文書解析と要約のための支援環境

松本 裕治, 宮田 高志 (奈良先端大), 野本 忠司 (国文学研究資料館),
徳永 健伸 (東京工業大学), 大林 正晴 (管理工学研究所)
{matsu,takashi}@is.aist-nara.ac.jp, nomoto@nijl.ac.jp,
take@is.titech.ac.jp, obayashi@kthree.co.jp

概要

文書構造を考慮した解析および要約技術の研究を支援するためのシステムを紹介する。本システムは形態素解析、係り受け解析、文書構造解析の三つのモジュールとそれらの結果を修正するためのGUIシステムおよび各解析結果を総合して文書を要約するモジュールからなる。モジュール間のデータの受け渡しはXMLタグ付けされたテキストで行なうので、モジュールごとに置換えが可能である。また、現在構築中の各モジュールは言語によらないので、多言語に適用することが可能である。デモでは主に修正のためのGUIシステムを通じて全体の構成を説明する。

Document Analysis and Summarization Support Environment

Yuji Matsumoto, Takashi Miyata (NAIST)
Tadashi Nomoto (National Institute of Japanese Literature)
Takenobu Tokunaga (Tokyo Institute of Technology)
Masaharu Obayashi (Kanrikogaku Kenkyusho, Ltd)

Abstract

We introduce our environment which is currently being constructed for supporting the study of analysis and summarization techniques considering the structure of target documents. Our system consists of the five modules; morphological, dependency, and document structure analyzers, editing system which corrects analysis errors through a graphical user interface, and summarization system which integrates the results of other modules. Since each module exchanges their data in XML text format, they can be easily replaced with another one. The fact that the modules currently under construction work independently of language allows the user to apply the system to multi-lingual domains. We will explain the overview of our system mainly through the editing GUI system.

近年、電子化文書の急速な増大によりそれらの中から必要な情報をできるだけ手間をかけずに早く抽出したいという要求が高まっている。また、自然言語処理に関する研究において統計に基づく解析手法がいくつか提案され、実世界の文書に対して実用的なコストと精度で解析を行なうことが可能になりつつある。
統計的手法を用いた解析や要約においては、(1) どのような**素性**に対して、(2) どのような**確率モデル**を仮定するかが重要である。とくに素性の選択に関しては、研究者の直観だけで有効そうなものを選ぶような方法ではすでに性能の向上は望めない状況になってきており、系統的・網羅的に素性を取捨選択することを支援するようなシステムが必要である。
このような背景をふまえ、我々は文書に対する各種解析および文書要約の研究を行なうための支援環境を構築している。システムは形態素解析、係り受け解析、文書構造解析の三つのモジュールとそれらの結果を修正するためのGUIシステムおよび各解析結果を総合して文書を要約するモジュールからなる。モジュール間のデータの受け渡しはXMLタグ付けされたテキストで行なうので、モジュールごとに置換えることができる。また、現在構築中の各モジュールは言語によらないので、多言語に適用することが可能である。入力された文書は各種解析の後、句や文ごとに何種類かの優先度が付与される。要約部ではパラメータとして設定された閾値や条件によって句や文を選択する。また、要約する前の文書がどのように解析されたかもGUIを通じて参照可能なので、辞書やコーパスを修正・構築し、各種解析が要約結果に与える影響を分析することができる。