自動要約に向けた新聞社説の文書構造タグの定式化

橋本喜代太[#*]　　松本裕治[*]
聖和大学人文学部[#]　奈良先端科学技術大学院大学情報科学研究科[*]
{kiyota-h, matsu}@is.aist-nara.ac.jp

1 はじめに

一般に要約と呼ばれるものは元々の文章とは構造的に違った文によって構成される「要約」と、元々の文章で見つかる文章内容上重要と思われる文を抽出して得られる「抄録」とに分かれる。いずれにせよ、これまでさまざまな研究によって次のようなものを手掛かりとすることが模索されてきた。[1]

(1) a. キーワード(タイトルを含む)
b. テキスト中での位置情報
c. テキスト中の文間の類似性
d. テキスト中の手掛かり表現
(接続語、特定の文末表現など)
e. テキスト中の文間の結束関係
f. 段落・パラグラフの構造

要約・抄録はその対象となる文章のジャンルや利用目的によってさまざまな形態を取る。まずその多様性を統一するといった試みは有用ではない。近年、日本語の新聞データが電子的に研究できるようになったが、ここでも通常の報道記事と社説や論説とは大きく異なる。前者がいわば各個の情報を適切に抽出することに焦点が置かれるべきであるのに対して、社説や論説はそこで提示される結論とそれを支持する根拠を適切に抽出し、それを関係づけることに焦点が置かれるべきであろう。この観点から見ると、社説や論説の要約研究は一般に数多い意見提示・論説の文章の典型であると考えられ、その点で重要度が高いと思われるが、これまでいくつかの例外を除き、そう利用されてきていない。

本発表では新聞の社説の自動要約を最終的な目標として、その前段階として特に(1f)の観点から新聞の社説に適した文書構造タグを考察する。その背後にある発想は、このような文書構造タグを(半)自動的に付けられるようにするというものであり、この観点から必要最低限のタグを考察する。

2 なぜ文書タグか

新聞社説は基本的にはその他の新聞記事に比べて具体的な事実の提示よりも主張とその根拠に重心がある。このような文書の場合、その要約・抄録は一意のものでなく、ニーズによって動的にその長さ、深さが変わってくると考えられる。結論だけ欲しい場合もあれば、結論と共にもっとも重要な根拠が欲しい場合もある。時にはもっと詳しく論点を整理したものを必要とする場合もあるだろう。

このような多様なニーズを満たすことを目的の一つとすると、初めから生のテキストを(1)のような観点で分析して一意の要約を得るシステムを考えるよりは、文書の構造などを示すタグを付与したり、必要な統計情報を得たりすることと、実際の要約・抄録の作成とは別の処理としたい。

一方、このような談話・テキストレベルのタグについてはほとんどの場合、人手で付けるのが普通であるが、主観性を排除するのが難しく、限定されたものであれ自動的に付与できるならば好ましい。本発表では汎用のタグを開発することは目的ではない。また、すべての文にタグを付けることも目的としない。要約・抄録は文をいかに削るかということであり、必要な情報の抽出にとって不要である文は積極的に無視していくことで処理の単純化を図りたい。

参考文献

[1] 『言語処理学会第4回年次大会ワークショップ論文集』言語処理学会. 1998年3月
[2] M. A. K. Halliday & R. Hasan. *Cohesion in English*. London: Longman. 1976.
[3] 佐久間まゆみ・杉戸清樹・半澤幹一. 『文章・談話のしくみ』. おうふう. 1997.