見出しからの制約による新聞記事からの重要文抽出

1E－9

瀬戸 喜巳† 井手 一郎†† 坂井 修一†† 田中 英彦††

{yoshimi,ide,sakai,tanaka}@mtl.t.u-tokyo.ac.jp

†東京大学工学部 ††東京大学大学院工学系研究科

1. はじめに

近年、電子化された文書量の増加に伴い、有用な情報を人手により選択することが困難になりつつあり、計算機を用いて効率的に取捨選択する必要が高まっている。

一般に文章には、内容の要約を示す見出しが附随している場合が多いが、見出しのみでは文書の内容を十分に表現しているとは言い難い。そこで、本研究では、見出しと本文全体の中間的な量の情報を得て、より効率よく正確に文書の内容を把握するために、見出しに関連の深い文を重要文として抽出することを目的とする。これにより、見出しに出現しない重要語の獲得や、ユーザーの必要に応じた長さの要録作成が可能となる。

要約に関する研究は多くなされているが[1,2]、文章に附随している見出しを利用した重要文抽出の研究は[3]に見られる程度で、比較的少ない。しかし、[3]の手法では見出し中の名詞の出現を重要度に利用するため、同様の内容を記すのに見出しと本文中の文で異なる語が用いられている場合に問題が生じる。

本稿ではこのような問題を避けるための手法として、文同士の意味的な近さをより正確に判定するために、見出し中の語と本文中の文の語との概念間距離を用いた重要文抽出手法を提案し、実験結果を示す。

2. 見出しの制約による重要文抽出

本手法の概要を図1に示す。黒い楕円部分が本研究で提案する手法に基づく処理である。

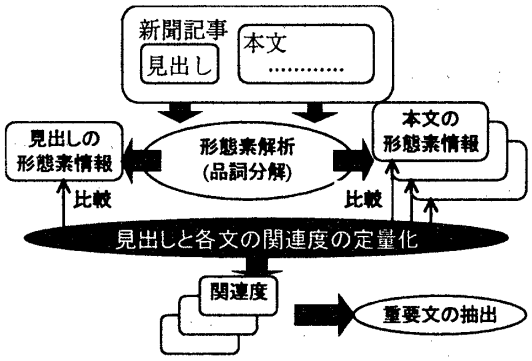

図1: 重要文抽出機構の全体像

2.1. 見出しと各文の関連度の定量化

新聞記事では、同じ内容を表すのに見出しと本文中の文で異なる語が用いられたり、略語が用いられることが多い。そこで、EDR 概念辞

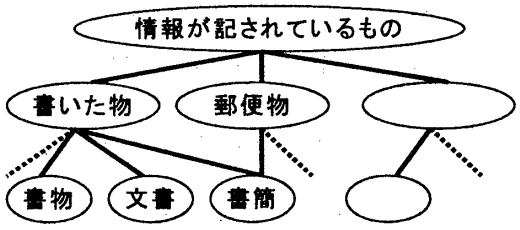

図2: EDR 概念体系の例

書[4]の概念体系 (図 2) を利用し、単語間の意味的な距離の指標として、単語が属する概念のノード間距離 (複数の概念に属する単語はそのうち最短のもの) を用いる。

" Extraction of Important Sentences from Newspaper Articles by Restriction from Headline Sentence "
Yoshimi Seto, Ichiro Ide, Shuichi Sakai, Hidehiko Tanaka
† Faculty of Engineering, ††Graduate School of Engineering, The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

本文中に含まれる単語のうち、JUMAN[5]による形態素解析の結果が名詞 (人名、地名、組織名、アルファベット、普通名詞、サ変名詞、時相名詞) であるものについて、表 1 のように得点基準を定め、文ごとに合計した値をその文の重要度とし、重要度の順位付けを行う。

表 1: 名詞の種別による得点基準

JUMAN の出力	得点基準
人名	・見出し文での同一単語の出現
地名、組織名、アルファベット	・見出し文での同一単語の出現 ・見出し文での同一概念の出現
普通名詞、サ変名詞、時相名詞	・見出し文でのノード間距離が 2 以下の概念の出現

3. 実験と評価

以上の手法を計算機上に実装し、1994 年の毎日新聞の記事 (13 記事、1 記事平均 17 文) に対して実験を行った。そのうち 1 記事の結果を表 2 示す。

表 2: 重要文抽出実験結果の例

見出し	ＩＡＥＡ、査察の機能停止にいら立ち北朝鮮との交渉、焦点――米朝交渉基本合意
順位	文
1 位	【ウィーン３１日高畑昭男】朝鮮民主主義人民共和国（北朝鮮）の核疑惑問題は、米朝が基本合意に達したことで、今後北朝鮮と国際原子力機関（ＩＡＥＡ、本部・ウィーン）の話し合いが焦点になる。
2 位	北朝鮮問題は、平壌郊外の寧辺にある二施設（非申告施設）に対してＩＡＥＡが特別査察を要請したのが発端だが、端緒は同国に対する特定・通常査察を進める中で廃棄物標本から「高濃度のプルトニウムが未申告のまま抽出された疑いが濃厚」という状況証拠が検出された事実にある。
	米朝交渉で、米国は途中から二施設に対する特別査察受け入れを、米朝高官第三ラウンド会談の条件にしなくなった。
4 位	特別査察は、北朝鮮とＩＡＥＡが話し合う問題だとゲタをあずけてしまっており、核疑惑自体は何も解決していない。

一方、6 人の被験者に対して 13 記事の本文から重要な文を上位 5 文挙げるアンケートを行い、得票数上位 5 文と、1 位に挙げられた文のそれぞれに対して適合率と再現率を求めた結果を表 3 に示す。

表 3：重要文抽出結果のアンケートによる評価

アンケート結果	適合率	再現率
1 位の文で評価	43%	71%
上位 5 文で評価	59%	60%

被験者の回答と一致しなかった文は、単語数が少なく、その文以降の記事の内容を述べる文や、今後の見通しなどを述べる文であったため、見出し文との関連性が低かったものであった。

4. おわりに

本稿では、見出しと本文の意味的な関連度を定量化することにより、重要文を抽出する手法を提案し、実験によってその有効性を確認した。今後は、得点を与える際に重み付けを行うなどのチューニングを施すことにより、さらに精度を上げることを目指す。

謝辞

EDR 電子化辞書は (株) 日本電子化辞書研究所の、CD-毎日新聞 94 版は (株) 毎日新聞社の利用許諾のもとに使用した。

参考文献

[1] 任 福継、定永 靖史; 「統計情報と文章構造に基づく重要文の自動抽出」; 情報処理学会技術研究報告 NL125, Vol.98, No.48, pp.71-78, May 1998.

[2] 原 正巳、中島 浩之、木谷 強; 「単語共起と語の部分一致を利用したキーワード抽出法の検討」; 情報処理学会技術研究報告 NL106, Vol.95, No.27, pp.1-6, Mar. 1995.

[3] 仲尾 由雄; 「見出しを利用した新聞・レポートからのダイジェスト情報の抽出」; 情報処理学会技術研究報告 NL117, Vol.97, No.4, pp.121-128, Jan. 1997.

[4] (株)日本電子化辞書研究所; 「EDR 電子化辞書 1.5 版」.

[5] 黒橋 禎夫、長尾 眞; 「日本語形態素解析システム JUMANversion3.5」; Mar. 1998.