E-024

携帯端末用新聞記事とWeb新聞記事による重要個所の分析と自動要約

大森 岳史† 増田　英孝† 中川 裕志‡

東京電機大学工学部† 東京大学情報基盤センター‡

1　はじめに

我々はWeb記事を自動要約し、既存の携帯記事と比較することにより要約の評価を行なった[1]。要約手法は係り受け解析の結果にTF・IDFを用いて文節の重みを算出し、枝の先端の重みが低い文節を削除している。その結果、より細かい重みの設定が必要となった。

本稿では、携帯端末用新聞記事(以下、携帯記事)とインターネット上で配信されているデスクトップPCなど、大画面向の閲覧を対象とした新聞記事(Web新聞記事)を比較して重要個所の分析を行う。

2　使用するデータ

携帯記事とWeb記事の重要個所を分析するために、対応文データ[2]を使用した。携帯記事は1～2文で構成されている。また、Web新聞記事は複数の文から構成されている。対応文とは携帯記事の1文に対応するWeb新聞記事の一文を対としたデータである。携帯記事の対応文を携帯記事文、Web記事文と呼ぶ。

3　重要個所の分析

Web記事文の重要個所の特定のために、次の手法を用いる。

Setp:1 携帯記事文の名詞を形態素解析器により抽出する。

Setp:2 Web記事文から係り受け解析器を用いて係り受けの結果を得る。

Setp:3 携帯記事文の名詞とWeb記事文の名詞を文節毎に比較する。

Setp:4 携帯記事文の名詞が一致したWeb記事文の文節を重要個所とみなして、係り受けの深さや名詞直後の助詞を分析する。

重要個所の特定の手法を次の例文を用いて説明する。

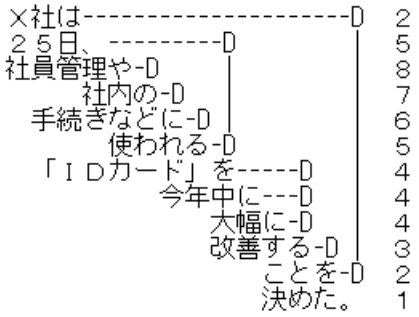

図 1: 係り受けの解析結果

携帯記事文：X社は「IDカード」の改善を決定。

Web記文：X社は25日、社員管理や社内の手続きなどに使われる「IDカード」を今年中に大幅に改善することを決めた。

はじめに、携帯記事文から名詞を取り出すと、「X社」、「IDカード」、「改善」、「決定」となる。次にWeb記事文を係り受け解析器にかけ、図1のような係り受け結果を得る。右側に記されている数字は文節の係り受けの深さを示している。係り受けの深さは文末の文節を1として葉に近付くにしたがい増えていく。係り受け解析の結果から得られた文節毎に携帯記事文から抽出した名詞を比較する。携帯記事文の名詞が使用されているWeb記事の文節は重要個所であると仮定して、この文節の係り受けの深さとこの名詞に付属する助詞を分析する。

4　評価

重要個所の評価としては、470組の対応文データを使用した。図2は文を構成する文節数を調査した結果である。Web記事対応文を係り受け解析したところ、文節数が10～14の間で最も多く、対応文470文中193文がこの区間に位置していた。

4.1　重要個所の名詞と係り受けの深さ

携帯記事文の名詞とWeb記事文の名詞で共通な名詞が使われているWeb記事対応文の文節を重要とし

Alignment between News Articles for PCs and Cell Phones on the Web and Web News Article Summarization

†Takefumi OOMORI, †Hidetaka MASUDA, ‡Hiroshi NAKAGWA

†School of Engineering Tokyo Denki University, ‡Information Technology Center The University of Tokyo

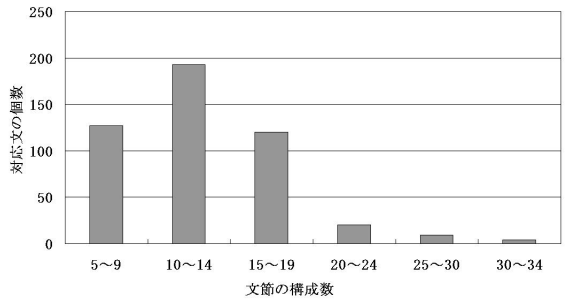

図 2: 対応文の構成数

ている。図 3 は重要個所とした文節の係り受けの深さの分布図である。特徴としては、５〜９文節、１０〜１４文節、１５〜１９文節の全てで係り受けの深さ２の位置で重要個所の名詞が多く出現している。

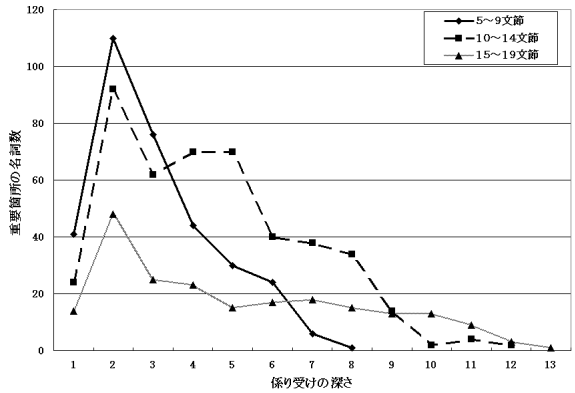

図 3: 重要個所の名詞と係り受けの深さ

4.2 共通な名詞に付属する助詞

Web 記事文の文節で携帯記事文の名詞と共通な名詞に付属する助詞について評価を行った。重要個所に出現した助詞の係り受けの深さと出現頻度の分布を示す。図 4 は重要個所の助詞「は」と係り受けの深さの分布である。助詞「は」は係り受けの深さ２で頻繁に現れた。図 5 は重要個所の助詞「の」と係り受けの深さの分布である。助詞「の」は係り受けの深さ２ではほとんど現れず、深さ３以降で頻繁に分布している。

5 まとめ

本稿では、携帯記事文と Web 記事文の対応文データを使用して重要個所の特定を行った。重要個所の特定は、携帯記事文の名詞と共通な Web 記事文の名詞を抽出して、共通な名詞が含まれる Web 記事の文節

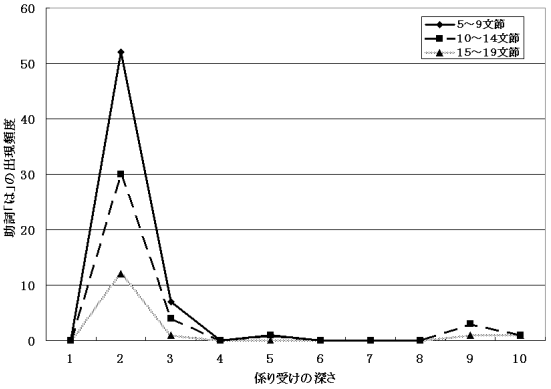

図 4: 重要個所の助詞「は」の分布

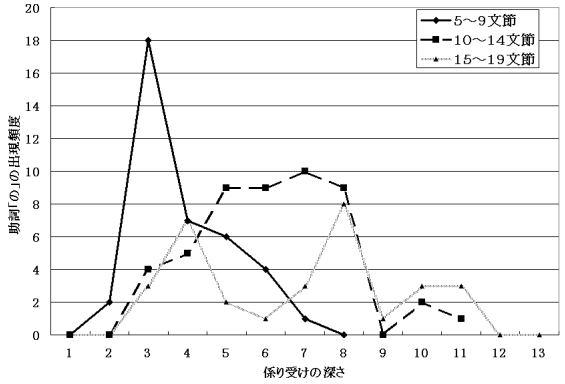

図 5: 重要個所の助詞「の」の分布

とした。重要個所の文節の係り受けの構造での位置を特定した。また、重要個所の助詞を抽出して、助詞と係り受けの関係を示した。今後は本稿の結果を重みの尺度として新たに取り入れ、Web 記事の自動要約を行う予定である。

参考文献

[1] 大森岳史, 増田英孝, 中川裕志：Web 新聞記事の携帯表示のための自動要約, 言語処理学会第９年次大会, pp. 202–206 (2003).

[2] 佐藤大, 増田英孝, 中川裕志：音声出力を目指した聞きやすいニュース読み上げの評価実験, 言語処理学会第９回年次大会, pp. 683–686 (2003).