Web新聞記事の自動要約とiモード記事による評価

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

近年、Webブラウズ機能付き携帯電話やPDAなどの普及に伴い、携帯端末でWebコンテンツを利用する機会が増えている。また、Webコンテンツの中には新聞社がデスクトップPCなどの大画面での利用を前提としてインターネットに配信している新聞記事 (Web記事) も多く存在する。このWeb記事を携帯端末で読もうとした場合、次のような問題がある。Web記事はデスクトップPCなどでは画面が大きいので一つの新聞記事の内容を比較的容易に把握できる。しかし、携帯端末は小画面なために一度に表示可能な文字数に限界がある。

一方、iモードなどの携帯端末向けの新聞記事 (携帯記事) はひとつのニュースを携帯端末の一画面に収まるような文字数で配信している。現在のようにWeb記事と携帯記事を人手によって作成した場合、時間とコストがかかってしまう。そこで、我々はWeb記事の自動要約を目的とした。本研究では携帯端末向け新聞記事を自動生成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。

2 対象とする新聞記事データ

自動要約した結果の正確さを判定するために正解データが必要となる。そこで、毎日新聞社 [1] からインターネットに配信されているWeb記事と携帯記事を用いた。そして、Web記事と携帯記事で同じ内容の記事の対を作成した [2]。

2.1 Web記事の特徴

Web記事の構成を図1に示す。Web記事は新聞記事のキーワード、20文字程度のタイトルの後に本文が続く。本文は数百字でまとめられている。本文は記事によって文字数にばらつきがあり、文字数が多い場合には段落でまとめられている。

2.2 携帯記事の特徴

携帯記事1記事の文字数はおよそ50文字である。どのような形式で携帯記事が書かれているかを知るために、以下に例文を示す。

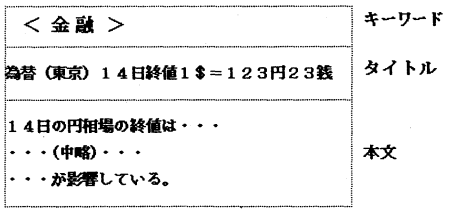

図 1: Web新聞記事の構成

"昨年の国内自動車市場は、普通乗用車の売上が3割増加。ハイブリッドカーの売上は前年比10%で過去最高。"

このように、携帯記事は末尾は体言止めが使用されていることが多い。

3 Web記事の自動要約

3.1 要約の方針

Web記事はジャンルと日付によってまとまりを持っている。したがって、ある日のあるジャンルの記事を文書集合とみなすことができる。すると、この文書集合に対するTF・IDFという尺度を用いれば不要個所の特定ができる。そこで、本研究ではTF·IDF (Term Frequency・Inverse Document Frequency) に基づく要約手法を提案する。各単語のTF·IDFを利用して削除部分を決める方法は文献 [3] にも見られ、文内要約の基本的手法のひとつとなっている。ここで図2に示すように自動要約の対象のWeb新聞記事を記事Aとし、Aの第1段落の文をA1,A2,…,Amとする。文Amを構成する文節をa(m,1),a(m,2),…,a(m,n)とする。また、要約の目標の長さをLとする。Lは50文字程度と100文字程度の2種類を設定した。要約手順としては以下の通りである。

1. 形態素解を行い、名詞を抽出する (未知語も含む)

2. TF・IDF 値を算出する

3. 構文解析を行う

4. 3. の結果の解析木から同定される葉の部分のうち、重要度の低い文節の削除

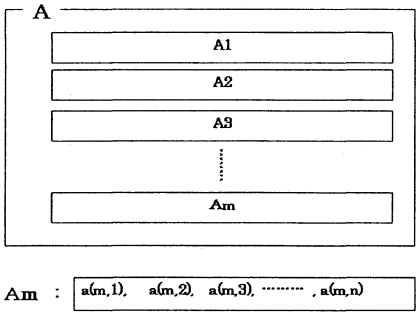

図 2: 要約対象記事の構成

3.2 要約アルゴリズムの詳細

形態素解析器「茶筌」[4] を用いて記事 A を形態素に分割する。この中から名詞と未知語を抽出する。抽出した名詞は次のステップで使用する。

TF・IDF の算出

単語に重みを持たせるために名詞の $TF \cdot IDF$ を算出する。記事 A を A、$TF \cdot IDF$ 値を求める単語を Wi とする。以下の式により記事 A に出現する単語 Wi の $TF \cdot IDF$ 値を求める。

$$TF \cdot IDF(A, Wi) = TF(A, Wi) \cdot IDF(Wi) \quad (1)$$

$TF(A, Wi)$ は記事 A における、単語 Wi の生起頻度である。$IDF(Wi)$ は当日に収集された文書数 N と、N の中で Wi が一回以上生起する文書数 $DF(Wi)$ に関係し、次のように定義する。

$$IDF(Wi) = \log(\frac{N}{DF(Wi)} + 1) \quad (2)$$

ただし、助詞「は」の文節の名詞は記事のトピックなど重要な情報を表すことが多いので、重みを 10 倍にする。

構文解析

記事 A の中から文 A1,A2,A3 を係り受け解析器「南瓜」[5] にかけ、係り受けの情報を得る。A1 が例文"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。"とした場合の係り受け解析結果を図 3 に示す。各文節に出現する名詞に算出した TF・IDF の値を加算する。例文 A1 の TF・IDF を表 1 に示す。

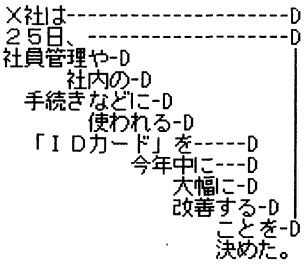

図 3: 係り受け解析の実行結果

表 1: 例文の TF・IDF

文節番号	文節	TF・IDF
a(1,1)	X社は	60.5
a(1,2)	25日、	5.7
a(1,3)	社員管理や	8.3
a(1,4)	社内の	23.2
a(1,5)	手続きなどに	15.3
a(1,6)	使われる	0
a(1,7)	「IDカード」を	21.5
a(1,8)	今年中に	13.3
a(1,9)	大幅に	11.2
a(1,10)	改善する	11.2
a(1,11)	ことを	14.8
a(1,12)	決めた。	0

重要度の低い文節の削除

TF・IDF の小さい枝を刈る要約アルゴリズムを以下 Step:0～Step:4 に示す。文字長 L が 50 文字程度の時は L = 50、100 文字程度の時は L=100 とする。

Step:0 k=0 に初期化（k はくり返しの回数）

Step:1 係り受け解析の結果から、文 A1,A2,A3 のそれぞれの先端の葉に相当する文節を抽出する。例文 A1 の枝の先端を図 4 に示す。抽出した文節のうち重みが最低な名詞、ないしは未知語を含む文節を除去する。

例文 A1,A2,A3 のうち、最低点の文節が A1 の文節番号 a(1,2)"２５日、"であった場合、図 5 のように削除する。

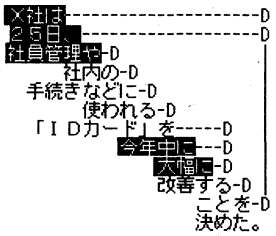

図 4: 例文の枝の先端（反転表示）

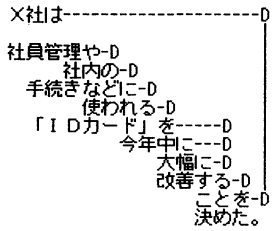

図 5: 枝の先端の削除（”２５日、”）

Step:2 A1,A2,A3 のいずれかで名詞 1 個と用言だけになったときは、その文 Ai を A1,A2,A3 から除去する。

Step:3 Step:1、Step:2 の結果を S(k) とする。

Step:4 A1,A2,A3 が文字長 L よりも長ければ k を 1 増やし、Step:1 へ戻る。短ければ S(k-1) を要約結果として終了。

助詞「は」のつく文節は「大統領は」、「台風は」など文の重要な主語などであることが多く、言語的に重要な要素とみなせるので重みを 10 倍にした。また、実験の結果、名詞 1 個と用言だけの文は「大統領は説明した」のように単独では意味が分からないものが多いので除去する。

4 要約結果の評価

4.1 結果の例

要約の結果についての評価を行なう。原文 A1,A2,A3 と 100 文字程度の要約、50 文字程度の要約の結果の例を示す。原文は次のように構成されている。

”社員登録にコンピューター端末による登録を導入する電子登録記録管理システム特別案が３０日、役員本会議で全会一致で可決、成立した。社内の社員登録事務を大幅に迅速化、省力化するとともに、コストを削減する効果がある。X 社の××社長は同日の会議で「社外ネットワークの問題などを克服できれば、グループ会社にも拡大していく」と述べた。(160 文字)”

次の文は 100 文字程度に要約された文の例である。

”社員登録にコンピューター端末による登録を導入する電子登録記録管理システム特別案が役員本会議で全会一致で可決、成立した。社内の社員登録事務を迅速化、省力化するとともに、コストを削減する効果がある。(97 文字)”

となっている。また、50 文字程度の要約文を以下に示す。

”コンピューター端末による電子登録記録管理システム特別案が成立した。社員登録事務を省力化するとともに、ある。(53 文字)”

元の記事は 3 文であったが、第 3 文が 3.2 節の Step:2 によって削除されている。

4.2 問題点

以下に要約結果に現れた問題点を示す。

重要度が高い文節の削除 次の例文のように重要度の高い文節が削除される場合がある。

”一連の不審火が相次いで発生したことで、民間調査会社も調査に乗り出した。”

この文で、読点直後の文節”民間調査会社も”、”調査に”の重みが低いために要約結果は

”一連の不審火が相次いで発生したことで、乗り出した。”

となった。

置き換え Web 記事と携帯記事で同じ事柄を説明しているが、異なる表現を使用している場合がある。Web 記事の例文では

”A 社の経営が軌道に乗りはじめ、業界トップの B 社を追い上げている。”

となっているのに対して、携帯記事の例文は

”A 社の売上が軌道に乗る。今後、B 社との競争が予想。”

という記事がある。これはWeb記事では"追い上げている"という表現が携帯記事では"競争が予想"という表現に置き換えられている。ただし、このような置き換えでも本質的な意味内容は変化しないことが多い。

代名詞の示す文節の削除 文A1に固有名詞があり、文A2にA1の固有名詞を指し示す"その"や"あの"などの代名詞が出現することがある。A1の固有名詞を含む文節が削除された場合、A2の"その"を示す文節が意味の通らない文になってしまう。

4.3 名詞の一致率

評価は対応付け記事[2]を使用して行った。これはWeb記事と携帯記事で同じ事柄を伝える記事の対である。Web記事の要約結果とそれに対応する携帯記事の名詞一致率を算出した。調査記事数は、政治が241記事、経済は452記事、国際は443記事、社会は362の合計1,498記事である。携帯記事と要約結果に関して精度と再現率を以下の式に従い算出した。

$$\text{精度} = \frac{(\text{両方の記事に共通する名詞数})}{(\text{要約文に含まれる名詞数})} \quad (3)$$

$$\text{再現率} = \frac{(\text{両方の記事に共通する名詞数})}{(\text{携帯記事の名詞数})} \quad (4)$$

表2は要約結果の名詞と携帯記事の名詞との精度を示している。50文字程度の要約の場合の精度は40%台という結果になった。また、表3に要約結果の名詞

表2: 要約結果の精度

	Precision			
	50文字	100文字	Web記事	記事数
政治	45.8%	40.1%	38.3%	241
経済	46.2%	40.3%	38.7%	452
国際	49.6%	42.2%	40.5%	443
社会	41.1%	35.1%	33.1%	362
全体	45.7%	39.4%	37.7%	1,498

と携帯記事の名詞との再現率を示した。これらの評価結果を他の研究と直接比較することは、コーパスの差異もあって困難である。しかし、類似研究との比較として以下のことは言える。すなわち、Berger[6]らのOCELOTでは確率モデルによる要約を行なっており、人手で作った要約との単語のオーバーラップ率を示している。オーバーラップ率は我々の精度にほぼ一致する。彼らの結果では、最大でも40%である。我々の携帯記事との比較結果は50文字、100文字とも40%を超えており、高い性能を持つといえる。

表3: 要約結果の再現率

	Recall			
	50文字	100文字	Web記事	記事数
政治	50.2%	72.3%	85.9%	241
経済	48.5%	69.4%	85.7%	452
国際	52.5%	74.5%	83.7%	443
社会	46.1%	66.7%	79.6%	362
全体	49.3%	70.7%	83.7%	1,498

5 まとめ

本稿では携帯端末向けにWeb新聞記事の要約を行なう手法を提案した。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するものである。今後は言い換えの処理や、重要度の高い文節を残して要約をするための方法を検討する予定である。また、文内要約した結果が、十分読み易いものになるような編集システムも大規模データによって検討し評価する予定である。

参考文献

[1] 毎日新聞社, http://www.mainichi.co.jp/.

[2] 大森ほか：携帯端末向け記事とインターネット新聞記事の対応付け, 情報処理学会第64回全国大会, Vol. 3, pp. 147–148 (2002).

[3] 上田ほか：句表現要約手法に基づく要約システムの開発と評価, 自然言語処理, Vol. 9, No. 4, pp. 75–96 (2002).

[4] 奈良先端科学技術大学院大学自然言語処理学講座：日本語形態素解析システム「茶筌」, http://chasen.aist-nsaara.ac.jp/.

[5] 奈良先端科学技術大学院大学自然言語処理学講座：日本語係り受け解析器「南瓜」, http://cactus.aist-nara.ac.jp/~taku-ku/software/cabocha/.

[6] A.L.Berger, and V.O.Mittal, : OCELOT：A System for Summarizing Web Pages, *23rd ACM SIGIR*, pp. 144–151 (2000).