Web新聞記事の携帯端末表示のための自動要約

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

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、携帯端末で新聞記事などの Web コンテンツを利用する機会が増えている。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 記事の自動要約を行い、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]。

3　Web記事の自動要約

Web 記事はジャンルと日付によってまとまりを持つものである。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、この文書集合に対して 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。ここで図 1 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. **名詞を抽出する（未知語も含む）**
   形態素解析器「茶筅」[3] 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。

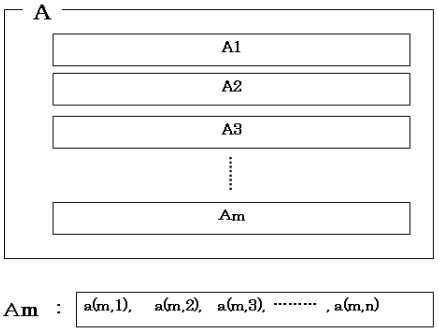

図 1: 要約対象記事の構成

2. **TF・IDF値を算出する**
   単語に重みを持たせるために名詞の TF・IDF を算出する。記事 A を A、TF・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・IDF 値を求める。

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

   $TF(A,Wi)$ 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」の文節の名詞は重みを 10 倍にする。

3. **構文解析を行う**
   記事 A の中から重要文として 3 文 A1,A2,A3 を取り出し、係り受け解析器「南瓜」[4] にかけて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 2 に示す。各文節に出現する名詞に手順 2 で算出した TF・IDF の値を加算する。

4. **重要度の低い文節の削除**

Web News Article Summarization for Moble Terminals
†Takefumi OOMORI, †Hidetaka MASUDA, ‡Hiroshi NAKAGWA
†School of Engineering Tokyo Denki University, ‡Information Technology Center The University of Tokyo

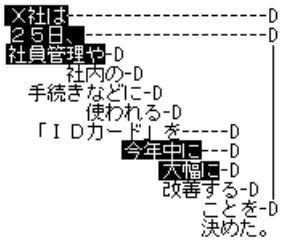

図 2: 係り受け解析の実行結果

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。

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文節を抽出する。図 2 の反転表示の文節が A1 の枝の先端文節である。抽出した文節のうち手順 3 でつけた重みが最低な名詞、未知語を含む文節を除去する。

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 k=k+1;
Step:2 の結果を S（k）とする。

Step:4 A1,A2,A3 の文字長が L より長ければ Step1 へ戻る。短ければ、S（k）を要約結果として終了。

3.1 名詞の一致率

要約結果と携帯記事との名詞の一致率を算出した。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 1 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 2 に要約結果の名詞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。Berger[5] 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 1: 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

表 2: 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

4 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スムージングも大規模データによって検討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] 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「茶筅」, http://chasen.aist-nara.ac.jp/.

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

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