

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

6Y-03

大森岳史,[†] 金田崇宏,[†] 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

本研究ではインターネット新聞記事から携帯端末向け新聞記事に自動要約を行うことを目的としている。人間が手動でインターネット新聞記事と携帯端末向け新聞記事の両方を作成するには時間とコストがかかってしまう。

携帯端末向け新聞記事を作成する際、インターネット新聞記事を元にして自動要約をすることにより記事作成のコストを削減することができる。

携帯端末向け新聞記事を自動生成した場合に、その結果が正解に近いかどうかを判定する為に正解データが必要となる。そこで、既存のインターネット新聞記事と携帯端末向け新聞記事を収集し、自動的にこれらの間の対応付けを行った結果を報告する。

2 評価の基準

2001年4月26日から2001年8月31日にかけて収集した携帯端末向け新聞記事(以下、記事Cとする)とインターネット新聞記事(以下、記事P)を使用する。政治、経済、国際、社会の4つのジャンルからなる記事の総数は記事Cが7,425個、記事Pが16,378個である。

名詞 形態素解析システム「茶筌」[1]を使用して記事の名詞を抽出する。新聞の品詞分布の特徴は名詞比率が非常に高い[2]ことから名詞に注目する。記事Cから名詞を抽出し、これをその日の各記事Pから抽出した名詞と比較する。この名詞が一致した場合、その記事に得点(配点は後に述べる)を付ける。そして、その得点が最も高い記事Pを正解とする。

数字と単位の扱い 数字を扱う際に数字部分のみを比較するのではなく、直後の単位や文字を数字と組にして比較する。例えば、記事Cに「5日」と

いう文字が含まれていたと仮定する。ある記事Pに「5日」という文字が、別の記事Pには「5人」という文字が存在する場合、数字だけで比較すると同じ得点を与えてしまう。この問題を避ける為に、数字の直後の単位や言葉を数字と組にして比較する。

配点の方法 記事Cの名詞が記事Pのどの場所の名詞と一致したかで得点を変化させる。記事Pは紙媒体の新聞記事よりもわずかに長い見出し、その後本文の構成となっている事が多い。見出しの内容は図1のように、最初にそのニュースのキーワード(例では金融)、その後に見出しの文となっている。見出しに使用される名詞は重要度が高いと仮定する。

携帯端末向け新聞記事 C

C_1, C_2, \dots

インターネット新聞記事 P

・見出しキーワード Pt1	$Pt1_1, Pt1_2, \dots$
例) 金融	
・見出し文 Pt2	$Pt2_1, Pt2_2, \dots$
例) 為替(東京) 14日終値 1\$=123円23銭	
・本文 Pb	Pb_1, Pb_2, \dots
例) 14日の終値は・・・(中略)・・・影響している。	

図 1: 各新聞記事の名詞の数

図1に示すように記事Cから抽出した個々の名詞を C_i とする。同様に記事Pの名詞を抽出する。見出し中の最初のキーワードの個々の名詞を $Pt1_j$ 、見出しの文の個々の名詞を $Pt2_k$ 、本文中の個々の名詞を Pb_l とする。以上を用いて次式に従い得点の評価を行った。

得点 $Sim(C, P)$

$$\begin{aligned}
 &= W1 \times (C \text{ と } Pt1 \text{ の名詞が一致した個数}) \\
 &+ W2 \times (C \text{ と } Pt2 \text{ の名詞が一致した個数}) \\
 &+ (C \text{ と } Pb \text{ の名詞が一致した個数}) \quad (1)
 \end{aligned}$$

Correspondence between News Articles for PCs and Cell Phones on the Web

[†]Takefumi OOMORI, [†]Takahiro KANATA, [†]Hidetaka MASUDA, [‡]Hiroshi NAKAGAWA

[†]Department of Electrical Engineering Tokyo Denki University, [‡]Information Technology Center The University of Tokyo

W1, W2 の値を変化させ記事 P の見出し名詞部分の重要度を決定する。

記事長による得点の正規化について 記事 C は 50 文字程度にまとめられている。記事 P は記事の長さにはばらつきはある。しかし、対応付け後の正解記事 P を調査してみると、第 1 段落と第 2 段落に記事 C に使用された名詞が使われており、その合計文字数は 250 文字程度である。第三段落以降には対応する名詞はほとんど出現しない。よって、得点は記事に依存せず、文字数比は、

$$\frac{\text{length}(P)}{\text{length}(C)} \doteq 5 \quad (2)$$

でほぼ一定となるため、点数をそのまま取り扱う。

3 結果の評価

3.1 記事対応付け結果

Sim の値が何点以上で正確な対応付けデータとなっているのかを調査するために、7 日分の対応付けられた新聞記事を人手によって正解と不正解に分類した。記事 C の総数は 605 個であった。

新聞記事の正確さを表す precision を求めるにあたっては下記に従った。

$$\text{precision} = \frac{(\text{抽出した対応付け正解記事数})}{(\text{抽出した全記事数})} \quad (3)$$

図 2 より 3 0 点以上の対応付けされた新聞記事はほぼ正確であるという結果となった。しかし、確実な正解データが必要なので 3 5 点以上で対応付けされた新聞記事のみを正確であると定義した。

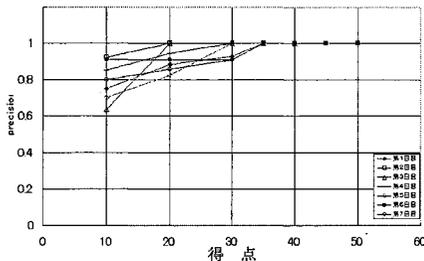


図 2: 対応付けされた新聞記事の precision

3.2 見出しの得点変化の結果

式 (1) の見出し部分の得点 W を記事対応付けでは 3 点と設定した。見出しの点によって対応付けされた新聞記事の正解率の変化を調べる。見出し部分の得点 (W1, W2) を (3, 3)、(6, 3)、(6, 6) と設定した。設定得点 W が 1 点より大きいということは、本文で一致する名詞よりも見出し中の名詞の方が重要な意味を持つことを表す。

図 3 より見出しの得点を (6, 6) にした時は正解率が低下している部分がある。しかし、見出しの得点が (3, 3) や (6, 3) の時は正解率に変化は生じなかった。

原因としては記事 P の見出し部分は似ているものが多いので、W を大きくしすぎると誤った記事を正解としてしまうためである。

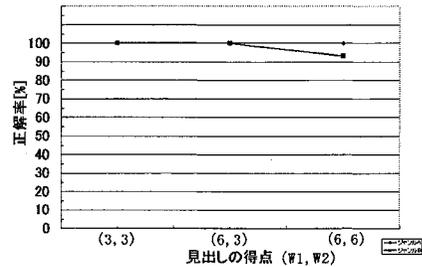


図 3: 配点の変化と正解率

以上により、605 個の記事 Cの中から 481 個の正解データを得ることができ、対応付けの成功率は 76.19 % という結果となった。

4 おわりに

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を行って対応付けされた正解データを得ることができた。この正解データを元にしてインターネット新聞記事の自動要約を行う予定である。

参考文献

- [1] 奈良先端科学技術大学院大学自然言語処理学講座 形態素解析システム「茶筌」, <http://chasen.aist-nara.ac.jp/>
- [2] 金城由美子, 熊野正, 西脇正道, 柏岡秀紀, 田中英輝: “ニュース文のスタイルに関する基礎的調査”, 言語処理学会第 7 回年次大会発表論文集 (2001), pp. 197-200