属性を用いた文節重要度に基づくニュース文要約

和田裕二*、 奥村明俊**、 浦谷則好***、 白井克彦****
*通信・放送機構（ＴＡＯ）渋谷上原リサーチセンター wada@shibuya.tao.go.jp
**ＴＡＯ／日本電気 ***ＴＡＯ／ＮＨＫ ****ＴＡＯ／早稲田大学

１．はじめに

現在、聴覚障害者がテレビ放送を楽しめるように字幕付きテレビ放送（字幕放送）が行われている。「情報バリアフリー」環境の整備を促進している総務省は平成 10 年度版通信白書において「2007 年（平成 19 年）までに、午前 7 時～午後 12 時までの字幕付与可能な全ての番組に字幕を付与する」という、字幕放送普及のための指針を掲げた。

しかし字幕放送の普及率は平成 12 年度でＮＨＫが 67.6%、民放関東キー5 局合計で 8.6%と、米国 4 大ネットワークの 70%以上に比べて非常に低い（平成 13 年 6 月 28 日の総務省報道資料）。こうした状況を打開するために、通信・放送機構は平成 8 年 4 月より聴覚障害者向け字幕付きテレビ放送番組の制作を支援するための研究を行っており、その要素研究としてニュース番組を対象にした自動要約の研究を行ってきた。[1]

ニュース番組を対象にした要約手法は過去に重要文抽出[2]や言い換えによる要約[3]といった研究がなされている。しかし重要文抽出法は採用されなかった文に含まれる情報が欠落する可能性が高く、また言い換えによる要約は高い要約率が得られないという欠点がある。[3]本稿ではより細やかな要約を行うために、文節や形態素の持つ属性より文節重要度を求め、その値を元に文節単位の要約を行う手法を開発したので、それについて報告する。

２．属性の抽出

提案する手法はニュース番組を対象にしていることから、ニュース文独特の性質を利用して属性の抽出を行った。ここでは次の 4 つを文節の重要度を決定する属性と定義した。

a.第１文
b.固有名詞等の特定単語
c.具体的な数値・日付
d.繰り返し出現する単語

ａは「ニュースは最初にそのおおまかな内容を述べる」という特徴に起因している。ｂ、ｃはニュース中のキーワードとなりやすい語である。ｄは通常の重要文抽出法においても重要とされる単語である。若尾ら[4]により重要文抽出の精度は TF-IDF 法より頻度の高い名詞性語句を用いた方が良いという結果が得られているので、ここでも同様に単純に頻度 2 以上の名詞性語句を重要とした。

次に文章としての意味が損なわれないようにするために、以下の属性を含む箇所も重要とした。

e.格助詞「が」
f. 格助詞「を」
g.係助詞「は」

以上を重要属性として抽出することとした。また石埼ら[5]を参考に、

h.重複表現

となっている箇所は逆に「重要でないもの」として抽出することとした。

3．属性コストと実験

形態素解析、構文解析済みのニュース記事に対し、以下の手順で処理を行うシステムを作成した。

(1)全ての文節重要度の初期値を0とする。

(2)2 章で述べた属性を抽出し、抽出された属性を含む文節の文節重要度に、各属性に設定されているコストを加算する。

(3)係り受け関係を元に、文節重要度の修正を行う。

手順(2)で用いられる、2 章で述べた属性に対するコストを表 1 に示す。コストの設定値はいくつかのニュース記事に対し実験を繰り返した結果、設定した値である。例えば属性ｂのコストは、人手による要約実験から属性ｂを含む文節が「削除されない」割合と、属性ｂを含まない文節のそれを計測した結果を用いて決定した。要約実験は記者経験のある 2 名を含めた 4 名の専門家により、4 年分のテレビニュース番組合計 800 記事を対象に行った。結果は表 2 に示すように、属性ｂを含む文節の方がそうでない文節より「削除されない」割合が高い。つまり、属性ｂを含む文節の方が含まない文節よりも重要度が高いと考えられるので、属性ｂを含む文節のコストを他のコストとのバランスから 120 と決めた。また属性 d のコストも同様に同じ要約データにおける、頻度2以上の名詞性語句が「削除されない」割合を計測した結果（図 1 参照）を用いて決めた。「削除されない」割合は総出現回数にはあまり依存しないのに、登場順序とは強い関連が認められたのでコスト値はそれを反映するように決めた。なおニュースは文数、単語数がそれほど多くないこと[2]や、形態素解析間違いによるカウントの増加もあることから、図1の対象単語は有効総登場回数を3～6とした。

手順(3)は、文節Ａが文節Ｂに係る場合に、Ｂの文節重要度がＡの文節重要度より低いとＢだけ削除され文構造が壊れてしまうという事態が起こり得るため、それを回避するための処理である。具体的にはＡの文節重要度をａ、Ｂの文節重要度をｂとした場合、ａ＞ｂならばｂ＝ａ＋１とし、係り先の重要度の方がわずかに大きくなるようにしている。

毎日新聞社の新聞記事に対し処理を行った要約結果を図 2 に、その元となった文節重要度の計算結果を図 3 に示す。なお係り受け解析の間違いにより文節重要度の修正がうまく行われていない箇所があるが、係り受け解析の問題は本稿で論じる内容ではないので、ここでは割愛する。

属性	コスト
a	文中の文節全てに150
b	120
c	200
d	50×√N／√n
e、f	50
g	150
h	−50

属性ｂのコストは、1 記事中の総登場回数をN、登場順序をｎとする。

表1　属性に対するコスト

属性ｂを含む文節

	91	93	94	95	平均
○	75.38	67.27	80.26	82.87	76.26
×	24.62	32.73	19.74	17.13	23.74

属性ｂを含まない文節

	91	93	94	95	平均
○	66.82	57.04	73.16	75.52	67.84
×	37.18	42.96	24.84	24.48	32.16

○は「削除されない」（要約者 2 名とも残した）、×はそれ以外（2 名とも削除した、または 1 名が削除した）を表す。

表2　属性ｂにおける文節削除の割合

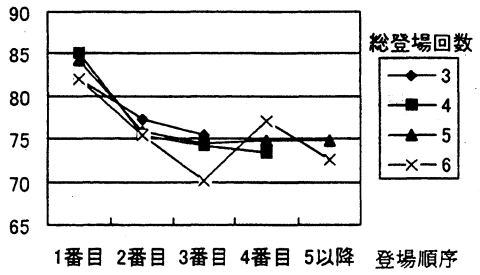

図1 出現順序と「削除されない」割合

牛肉とオレンジの輸入が自由化されますが、牛肉の関税が引き上げられることなどから当面、輸入価格の値下がりは多くは期待できないようです。輸入自由化は昭和６３年の日米合意などに基づいて実施されるもので、牛肉については輸入数量制限が撤廃されるとともに、保護する目的で輸入価格に上乗せされてきた調整金が廃止されます。従来２５%だった関税が、７０%に引き上げられることになっています。関税は６０%平成５年度は５０%と順次引き下げられることになっていますが、いずれにしても関税がかかるため、値段は、下がらないのではみられています。 高級品については、調整金の割合が低かったことから７０%もの関税がかかる今年度は心配もあるということです。過剰気味になっていることから自由化はされるものの、減るという見方が広がっています。オレンジについては関税は据え置かれるものの主力のカリフォルニア産の減産でひっ迫していることから、関係者の間では輸入オレンジの小売価格は高くなるとの見方が一般的です。

図2 原文の80%要約結果

4. おわりに

ニュース番組におけるアナウンサーの発話速度は 400 文字／分程度であるが、人間が字幕として読み取れる最高速度は 300 文字／分程度と言われていることから、7 割程度の要約を行う必要がある。しかし本文で述べた手法のみで 7 割に要約を行うと記事の意味が通らなくなってくることから、本手法による要約は 8 割程度にとどめ、言い換えによる要約を併用することが好ましいと考える。

今後の課題としては抽出する属性の追加や詳細化と、それに伴う属性コストの設定・修正が挙げられる。そのためにはより多くのニュース記事に対しての実験と、その結果に対する客観評価が必要であり、今後行っていきたい。またニュース以外の番組への適用についても検討したい。

参考文献

[1]江原、沢村、若尾、阿部、白井「聴覚障害者のための字幕つきテレビ放送制作への自然言語処理の応用」言語処理学会第 3 回年次大会 pp.489-492

[2]加藤、浦谷「放送ニュースを対象にした重要文抽出」言語処理学会第 6 回年次大会 pp.237-240

[3]山崎、三上、増山、中川「聴覚障害者用字幕生成のための言い替えによるニュース文要約」言語処理学会第 4 回年次大会 pp.646-649

[4]若尾、江原、白井「テレビニュースのための自動要約」言語処理学会第 4 回年次大会ワークショップ pp.7-13

[5]石垣、片岡、増山、山本、中川「係り受け関係を用いた重複表現削除」自然言語処理 Vol.7 No.4 pp.119-142

150 きょうから
261 牛肉と
262 オレンジの
442 輸入が
443 自由化されますが、
229 牛肉の
442 関税が
444 引き上げられることなどから
445 当面、
562 輸入価格の
563 値下がりは
300 多くは
564 期待できないようです。

64 牛肉と
70 オレンジの
410 輸入自由化は
320 昭和６３年の
321 日米合意などに
322 基づいて
411 実施されるもので、
205 牛肉については
351 輸入数量制限が
412 撤廃されるとともに、
0 これまで
120 国内の
170 畜産農家を
413 保護する
414 目的で
355 輸入価格に
415 上乗せされてきた
416 調整金が
417 廃止されます。

340 関税は
150 来年度は
200 ６０％
590 平成５年度は
200 ５０％と
201 順次
591 引き下げられることに
592 なっていますが、
593 いずれに
593 しても
120 比較的高水準の
231 関税が
594 かかるため、
50 牛肉の
270 値段は、
0 あまり
595 下がらないのでは
0 ないかと
596 みられています。

0 特に
0 サーロインなどの
390 高級品については、
150 これまでは
290 調整金の
291 割合が
391 低かったことから
200 ７０％もの
224 関税が
392 かかる
393 今年度は
120 逆に、
0 これまでよりも
0 高くなる
120 心配も
394 あると
395 いうことです。

図３ 実験結果（抜粋）