自然言語処理 119－6
（1997．5．27）

テレビニュース番組電子化原稿を題材とした自動要約手法の大規模評価

若尾 孝博　江原 暉将　村木一至　白井克彦
通信･放送機構 (TAO)　NHK／TAO　NEC／TAO　早稲田大学／TAO
渋谷・上原リサーチセンター
wakao@shibuya.tao.or.jp

概要

電子化された情報が増大し、テキストの要約や重要文の抽出に関する研究が、最近注目されるようになって来ている。本研究では、テレビニュース番組の電子化原稿を題材として、テキストから重要文を選び出す基本的な手法の評価を行った。基本的手法としては、重要文抽出の伝統的手法である重要語密度法、及び情報検索分野で知られているＴＦ・ＩＤＦ法に基づいた重要文抽出法を用いた。テストデータは１万件のテレビニュース番組電子化原稿で、第１文が最も重要であるというニュース原稿の特徴を利用して自動評価を行った。評価の結果は、全体的に重要度密度法がＴＦ・ＩＤＦ法に基づいた手法よりも良い結果を示すことが判明した。

Evaluation of Methods of Sentence Extraction on TV News Texts

Takahiro Wakao　Terumasa Ehara　Kazunori Muraki　Katsuhiko Shirai
TAO of Japan　NHK / TAO　NEC / TAO　Waseda University / TAO
wakao@shibuya.tao.or.jp

Abstract

We describe and evaluate methods for automatically selecting important sentences in TV news texts in Japanese. High-frequency word ratio method and a method based on TF-IDF are used to rank the sentences in a text. One of the features of TV news texts is that the first sentence is the most important. We take advantage of the feature and evaluate the methods automatically by using 10000 texts. On the whole, the high-frequency word ratio is better than the TF-IDF based method.

1. はじめに

近年電子化されたテキストの量が増大し、テキストの要約や、そこからの重要文の抽出をする研究が盛んになってきている。
本研究では、聴覚障害者のための字幕付きテレビ放送制作への自然言語処理の応用の一環として（[1]）、テレビニュース番組の電子化原稿を対象にした重要文の抽出のための基本的手法を自動評価をした。用いた手法は、重要文抽出の伝統的手法である重要語密度法と、情報検索分野で知られているＴＦ・ＩＤＦを用いた重みづけを利用した重要文抽出法の２つの手法である。
評価で用いたテキストは、ＴＶニュース番組の電子化原稿１万件である。同時にこの１万件のテキストから「政治」「経済」「国際」「スポーツ」の４分野の原稿を選び出し、分野ごとでの評価も行った。

2. 重要文抽出手法

今回用いた重要文の抽出のための手法は、重要語密度法とＴＦ・ＩＤＦに基づいた抽出法であるが、重要文を抽出にあったての共通の手順は以下のとおりである。

1 形態素解析
　インプットされたテキスト（１記事）を形態素解析する。
2 形態素の重みの計算
　自立語（名詞、固有名詞、形容詞、動詞、未知語）を選び出し、重みを計算する。
3 各文の重要度の計算
　各文の重要度は、その文中にある自立語の重みの総和に基づいて計算する。
4 文の重要度による順位付け
　１テキスト中の文を算出された重要度の高いものの順にソートする。

次に、本研究で用いられた手法を詳しく説明する。

2.1 重要語密度法

この手法は、Luhn（[2]）が提唱し、古くから重要文の抽出に使われて来た。計算の方法は、文中の重要形態素を数え、その数を文中の形態素数で割ったものをその文の重要度とするものである。本研究では、自立語で、テキスト中に２回以上現れるものを重要形態素とした。つまり、重みで言うと、頻度が２以上の自立語は 1.0 であり、残りの形態素については０とした。

2.2 ＴＦ・ＩＤＦに基づいた抽出法

第２の重要文抽出の手法は、ＴＦ・ＩＤＦを用いて重要形態素の重みを算出して、重要文を決定するものである。ＴＦはその形態素のテキスト中の頻度であり、ＩＤＦは、そのテキストが属するコレクションで、コレクション全体のテキスト数を該当する形態素が現れたテキスト数で割ったものの対数をとったものである。各形態素の重みを式で示すと、次のようになる。

$$weight = tf \times \log (N/n)$$

tf：形態素のテキスト中の頻度
Ｎ：コレクション全体のテキスト数
ｎ：形態素が現れたテキスト数
となる。

各文の重要度は、重要語密度法と同じで、各形態素の重みを加算し、それを文中の形態素の数で割ったものである（[3]）。

3. ＴＶニュース電子化原稿

次に、テキストを紹介する。使用したテキストは、ＮＨＫ放送データベースから選択さた１万２千件、１９９２年から１９９５年まで各年３０００件のテレビニュース番組の電子化原稿である。このうち、２千件は、システムの訓練用に用い、残りの１万件（各年２５００件）を評価用に用いた。同時に、この１万件のテキストの中から、「政治」「経済」「国際」「スポーツ」の４分野の原稿を選び出し、分野ごとでの評価も行った。各分野のテキスト数を表１に示す。

分野	テキスト数
政治	１６４９
経済	１１６３
国際	１４３９
スポーツ	７８０

表１　４分野のテキスト数

対象となるテレビニュース原稿を分析した結果、以下のような特徴を備えていることが分かった。

1. １つのテキスト中の文数は、平均で5.65 であり、１万件中に１文または２文だけのテキストが５１６件あった。比較のために、新聞記事１０００件を調べてみると、１記事当たり平均の文数は、7.07 であった。つまり、テレビニュース原稿のほうが新聞記事よりも１記事当たりの文数が少ない。
2. １テキスト当たりのバイト数は、新聞記事、テレビニュースでほとんど差はなく、テレビニュース原稿での各文が新聞記事の文よりも長いことが判明した。
3. テレビニュース番組テキストでは、第１文がほとんどの場合、最も重要である。第２文以降は、第１文で述べられた事柄についての詳細を説明する形となっている。
4. また、ニュース番組テキストでは、段落が１つしかなく、ディスコースレベルでの構造は単純である。

4. 評価と結果

4.1 評価の方法

評価は完全に自動で行われた。ニュース番組テキストでは、テキスト中の第１文が最も重要であるという特徴があり、この特徴を生かして第１文が各手法を用いて算出された重要度において第１位となれば正解とし精度を計算した。同時に、第１位または第２位になる場合も正解として、正解率（精度）を算出した。

4.2 重要語密度法の結果

重要語密度法を用いた場合の精度（パーセント表示）以下の通りである。

年	テキスト	1位正解	1 or 2位
1992	2500件	66.68	88.36
1993	2500件	68.76	89.32
1994	2500件	71.40	89.72
1995	2500件	68.60	88.40
合計	10000件	68.86	88.95

表2　重要語密度法の結果（1）

4分野ごとの精度は表3のとおりである。

分野	テキスト	1位正解	1 or 2位
政治	1649件	68.36	90.03
経済	1163件	70.33	90.34
国際	1439件	74.10	93.45
スポーツ	780件	41.41	69.87

表3 重要語密度法の結果（2）

4.3 ＴＦ・ＩＤＦに基づいた手法の結果

ＴＦ・ＩＤＦ法に基づく重要文抽出手法では、２種類のＩＤＦを用い、その結果を評価した。２種類のＩＤＦは、各年度毎のコレクション（２５００件）から算出したものと、１万件全部を使って算出したものである。

まず、各年のＩＤＦを用いた場合の精度は表4と表5のとおりである。

年	テキスト	1位正解	1 or 2位
1992	2500件	53.40	79.28
1993	2500件	54.84	81.32
1994	2500件	56.80	83.08
1995	2500件	52.08	78.00
合計	10000件	54.28	80.42

表4　TF-IDFに基づく手法の結果（1）

分野	テキスト	1位正解	1 or 2位
政治	1649件	59.58	87.19
経済	1163件	55.18	81.51
国際	1439件	58.82	85.93
スポーツ	780件	42.43	71.41

表5　TF-IDFに基づく手法の結果（2）

次に、１万件の記事から求めたＩＤＦを用いた場合の精度は表6と表7のとおりである。

年	テキスト	1位正解	1 or 2位
1992	2500件	52.76	78.84
1993	2500件	54.40	81.88
1994	2500件	56.88	83.24
1995	2500件	52.04	78.72
合計	10000件	54.02	80.67

表6　TF-IDFに基づく手法の結果（3）

分野	テキスト	1位正解	1 or 2位
政治	1649件	61.13	87.70
経済	1163件	57.89	82.42
国際	1439件	62.82	88.90
スポーツ	780件	39.10	68.85

表7　TF-IDFに基づく手法の結果（4）

5. 考察

5.1 1万件の評価のまとめ

1万件のテキストを使っての評価の結果をまとめると次のとおりである。

1. 全体的に見て、重要語密度法のほうがＴＦ・ＩＤＦに基づく手法よりも良い結果を示した。
2. ＴＦ・ＩＤＦに基づく手法では、２種類のＩＤＦを用いたが、差がほとんど見られなかった。
3. 分野別での精度にばらつきが見られた。特に、スポーツのテキストに対する精度が両方の手法で、他の分野のテキストと比べて低かった。
4. 「政治」「経済」「国際」の分野のテキストでは、精度が各年で分けてみた場合よりも良い傾向を示した。特に、国際ニューステキストに対しては、良い精度が得られた。

5.2 誤りの分析と今後の課題

Paice ([4]) がまとめているように、重要文抽出の従来の研究では、見出し、段落など、テキスト中での「位置」を手がかりにして重要語を探すものが幾つかある。しかし、テレビニュース原稿は、テキスト中に段落が１つしかなく、その上、本研究では見出しの情報も利用しなかったため、これらのヒューリスティックに基づいた従来の手法は用いることが出来なかった。

重要語密度法で不正解となったテキストを分析してみると、失敗の原因は以下のようであることが分かった。

1. 幾つかのテキストの第１文は短く、しかも話題を提供するだけのものであった。例としては、「今日の大相撲の結果です。」などのような第１文である。この種の第１文では、重要語が少なく、基本的に語の頻度に基づいた今回の両手法では、重要度が上がらず重要な文とならない。
2. スポーツ番組のテキストでは「マッチプレー」、「バーディー」などのスポーツ分野での用語が多く現れ、これらは大抵の場合未知語となった。未知語は、自立語の扱いとなり、頻度の高いものは重要語となる。第１文が前述のような「話題」提供の文である場合、専門語が出てくるスポーツテキストにおいては、第２文以降に、その専門語また固有名詞などを含んだ実質的な情報が含まれており、それらの文のほうが重要語を含むことになり、重要度が第１文よりも高くなることになる。

これらの分析から分かるように、今後は、話題提供の短い第１文への対処、および、未知語のより適切な取り扱いなどが課題である。また、本研究はニュース番組原稿を短縮化して字幕として見せることに役立てるための基礎研究であり、今後は、実際に字幕にすること考慮して、ニュース電子化原稿をいかに短縮していくかを研究して行く予定である。

参考文献

[1] 江原 暉将、沢村 英治、若尾 孝博、阿部 芳春、 白井 克彦 1997 「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回年次大会発表論文集 pp 489-492.

[2] H.P. Luhn 1958 The automatic creation of literature abstracts. In *IBM Journal for Research and Development* 2(2), pp 59-165.

[3] Gerald Salton, Christopher Buckley 1988 Term-Weighting Approaches in Automatic Text Retrieval, In *Information Processing & Management*, 24(5), pp 513-523.

[4] Chris D. Paice 1990 Constructing literature abstracts by computer: techniques and prospects. In *Information Processing & Management*, 26(1), pp 171-186.