講演文を対象にした重要文抽出

伊藤 山彦† 松本 賢司† 谷田 泰郎‡ 柏岡 秀紀† 田中 英輝†
†(株)エイ・ティ・アール音声言語通信研究所 ‡(株)国際電気通信基礎技術研究所

1 はじめに

情報通信技術の発達に伴い、様々な状況において、短時間での大量の文書のサーベイや要点の把握が要求される場面が増えている。近年では、映像や音声を利用した形態での情報提供も増えており、音声情報に関しても、効果的に情報を整理する技術が必要とされている。

我々は音声で提供される情報として講演に着目し、講演文を対象とした自動要約の研究を進めている。従来、自動要約の研究は、論説文や新聞記事等、主として書き言葉を対象に行われていた。我々は話し言葉としての講演文の特徴を利用することにより、的確な要約の生成が可能な手法を確立することを目標としている。

本研究では、講演文の例として、NHK 番組「あすを読む」の書き起こし原稿を対象に重要文抽出の実験を行った。正解データは複数の人で手作業により作成した。実験は、単語の頻度を利用した手法と、表層の手掛かりを利用した手法の2つを行った。本稿では、実験結果を基に両者の手法を比較評価すると共に、講演文の特徴の分析や、人間同士における重要文の判定の揺れについても述べる。

2 関連研究

重要文抽出において、文書の重要度を判定するための手法として、(1)単語の出現頻度を利用する手法[1][2]や、(2)文の出現位置や手掛かり表現等の表層的な情報を利用する手法[3][4][5]が研究されている。

上記のうち、(1)の手法は適用範囲が広く、講演文にもそのまま適用できる。しかし、単語の頻度値に反映されない講演文の特徴を利用できないという問題点がある。

また、(2)で挙げた文献[3][4][5]は、いずれも論説文や新聞記事など書き言葉を対象としている。段落構造という書き言葉に特有の性質を利用しているため、そのまま講演文に適用することは難しい。

本研究では、(2)の手法を講演文に適用するため、講演文を分析し、講演文に特有の特徴を抽出した。抽出した特徴を属性値として用い、決定木学習による重要文抽出実験を試みた。また(1)の手法も試み、(2)の手法との比較評価を行った。

3 人による重要文抽出

本研究で実験の対象とした「あすを読む」は、主として時事問題をテーマとした 10 分間の番組であり、書き起こすと約3000字程度のテキストとなる。実験の対象とした原稿の数は50件、1文書中含まれる文の数は平均して60.2文である。

人による重要文抽出作業として、3 人の被験者に対し、短い要約として5±2 文、長い要約として 20 ±3 文の抽出を依頼した。重要であることの基準は指示せず、被験者の判断に委ねた。また、長い要約(以下、20 文抽出と呼ぶ)で抽出した文の中に、必ずしも短い要約(以下、5 文抽出と呼ぶ)で抽出した文が含まれる必要はないとした。

結果を表1に示す。値はいずれも平均値である。

表1 人による重要文抽出結果

	5文	20文
1文書中の文の数	60.2文	
1者の平均選択数	5.2文	20.5文
3者の全員が選択	1.5文	9.0文
2者以上が選択	4.2文	20.1文
3者のいずれかが選択	9.9文	32.6文
3者の平均正解率[1]	46.7%	61.9%

[1]被験者A、B、Cの何れかの抽出結果を正解としたときに、他の被験者の抽出結果の再現率および適合率を、全ての2者間の組み合わせに対して求め、平均した値。全ての組み合わせを平均すると、再現率と適合率は同じ値になるため、「正解率」という言葉を用いた。

実験の結果、人間同士の判定においても、正解率は5文抽出で46.7%、20文抽出で61.9%に止まるとの結果を得た。この値は、自動抽出の実験を評価する際の上限値になる。

4 単語の頻度を利用した重要文抽出

単語の頻度を利用した重要文抽出の代表的な手法であるキーワード密度法[1]と tf*idf 法[2]の実験を試みた。これらは文書の種類に依存しない汎用的な手法であり、講演文の要約にそのまま適用することができる。

4.1 実験方法

以下の手順に従い、実験を行った。

(1)文書の形態素解析結果から、助詞、助動詞、代名詞、補助動詞、感動詞、接頭辞、連体詞、接尾辞、接続詞、及び副詞を取り除き、内容語を取り出す。

(2-a)キーワード密度法の処理
抽出された内容語のうち、各文書において頻度値が上位10位以内の語を重要語とし、以下の式に従って文の重要度を計算する。

$$\text{文の重要度} = \frac{(\text{文中に現れる重要語の数})^2}{\text{文中に現れる内容語の数}}$$

(2-b)tf*idf 法の処理
以下の式に従い内容語の tf*idf 値を求め、その合計値を文の重要度とする。

$$tf*idf(w_i) = f(w_i)*\log\frac{Nd}{d(w_i)}$$

ただし $f(w_i)$: 内容語 w_i が現文書に現れる数
Nd : 実験対象の全文書数
$d(w_i)$: w_i が出現する文書の数

(3)文の重要度のスコアの高い上位5文または20文を抽出する。

4.2 実験結果

表2に結果を示す。表の値は、それぞれの手法で抽出した結果に対し、3人の被験者それぞれの抽出結果を正解としたときの再現率、及び適合率の平均値である。

表2 単語の頻度に基づく重要文抽出実験結果

	5文抽出		20文抽出	
	再現率	適合率	再現率	適合率
(a)KW密度法	14.7	15.1	42.1	43.1
(b)tf*idf法	17.4	18.1	45.0	45.9

5文抽出、20文抽出とも、表1に示した人の間の正解率に比べるとかなり低い。特に、5文抽出では人の正解率 46.7%に対して、本手法の再現率／適合率が10%台と、大幅な低下を示している。

5 表層の特徴を利用した重要文抽出

第4節で述べた手法が、文書の種類に依存しない手法であるのに対し、本節では、講演文の特徴を利用した重要文抽出について記す。講演文の特徴を分析し、抽出した特徴を属性として確率決定木[2]による学習を用いた実験を行った。

5.1 学習データの属性

講演文を分析し、文の重要度に関係があると考えられる以下の特徴を抽出した。

(1)接続詞の出現
実験対象の文書中に出現する全ての接続詞を抽出し、各文に対して、抽出した各接続詞の有無を属性値として付与した。

(2)副詞の出現
実験対象の文書中に出現する全ての副詞を抽出し、各文に対して、抽出した各副詞の有無を属性値として付与した。

(3)文書中における文の位置
文書の位置情報を学習に反映させるため、文書の先頭10文に対し、最初の文から順に10から1までの数値を属性値として付与した。また、文書の末尾10文に対し、最後の文から順に10から1までの数値を属性値として付与した。

[2] C4.5の出力に確率値をつけたもの。

(4)文のタイプ

講演文に現れる表現パターンから機械的に分類した文のタイプを属性値とした。以下に、文のタイプ、各文のタイプが文書中に出現した数と割合、及び表現パターンの例を記す。なお、対象の文の総数は3010文である。

(a)総括的説明表現(136文、4.52%)
　～わけです。
(b)理由説明表現(23文、0.76%)
　～からです。
(c)強調的説明表現(14文、0.47%)
　～ということなのであります。
(d)導出的表現(85文、2.82%)
　～ということになります。
(e)問題提示表現(64文、2.13%)
　なぜ～でしょうか。
(f)疑問提示表現(17文、0.56%)
　～で良いのでしょうか。
(g)直接的主張表現(179文、5.95%)
　～と思います。
(h)問い掛け的主張表現(25文、0.83%)
　～ではないでしょうか。
(i)一般見解表現(11文、0.37%)
　～と考えられます。
(j)推量表現(29文、0.96%)
　～でしょう。
(k)主題導入表現(34文、1.13%)
　～についてお伝えします。
(l)宣言的表現(74文、2.46%)
　～てみましょう。
(m)挨拶表現(71文、2.36%)
　こんばんは。
(n)指示表現(7文、0.23%)
　ご覧下さい。
(o)その他(2241文、74.45%)
　上記の分類に当てはまらない文。大部分は事実を述べた文である。

(5)時制

文の末尾が「た」で終わるものを「過去」、その他を「現在」とし、時制の異なりを属性値とした。

(6)列挙表現

「第一に」「第二に」のような列挙表現の有無を属性値とした。

比較のため、上記(1)～(6)の表層の手掛かり表現に加え、tf*idf法による文の重要度を属性値に加えた学習データについても実験も行った。

5.2 実験方法

3人の被験者が50文書に対して作成した正解データ150文書に対して、10分割の交差検定を行った。1つの文書に対して、3つの正解が存在することとなるため、同じ文書に対するデータは1組とし、同じ文書が学習用データと評価用データの両方に現れることがないようにした。

出力は、決定木が判定した重要文である／ないのクラスではなく、重要文である確率値の高い順に上位5文、及び20文とした。これは、第4.2節の結果と比較するためである。

5.3 実験結果

表3に実験結果を示す。4.2節の実験結果と同様、表の値は、本手法の抽出結果に対し、3人の被験者それぞれの抽出結果を正解としたときの再現率と適合率の平均値である。表3の「表層」の欄の値は5.1節に述べた6種類の属性を用いた結果であり、「表層+tf*idf」の欄の値は、学習の属性にtf*idf法による文の重要度の値を追加したものである。

表3 表層の特徴を利用した重要文抽出結果

	5文抽出		20文抽出	
	再現率	適合率	再現率	適合率
表層	29.6	30.1	48.4	49.3
表層+tf*idf	30.8	30.8	47.4	48.3

結果は表3に示す通り、単語の頻度のみを利用した手法(表2)に比べ、抽出精度の向上が見られた。特に5文抽出では大幅な向上が見られた。しかし、学習の属性にtf*idf法による文の重要度の値を加えたことによる精度の向上は見られなかった。

単独の属性のみで決定木学習を行った実験結果から、各属性を重要文の判定への寄与の高い順に並べると、以下のようになる。

5文抽出

文の位置>列挙表現>文のタイプ>時制>接続詞>副詞

20文抽出
文の位置>文のタイプ>副詞>列挙表現>接続詞>時制

6 考察

第4節と第5節の結果を比較すると、単語の頻度情報のみから重要文を抽出するよりも、表層の手掛かりから重要文を抽出する方が良いという結果が得られた。特に5文抽出において顕著であった。

このことは、内容的な情報を考慮しなくても、表層の手掛かりだけでかなりの部分、重要な個所が判定できるということを意味する。特に講演では、紙に書かれた文章と異なり、読み返すことができないため、重要な個所では、話者は明示的に強調的な表現を用いるものと考えられる。

次に、個々の被験者と機械的手法の判定の傾向の違いを調査した。表4は、被験者A、B、Cそれぞれの抽出結果を正解としたときの他の被験者の再現率と適合率であり、表5は、各被験者に対する各手法の再現率と適合率である。なお、表の「表層」の欄の値は、tf*idf法による文の重要度を決定木学習の属性に入れない実験結果の値である。

表4 人-人の再現率・適合率(20文抽出)

	A		B		C	
	再現	適合	再現	適合	再現	適合
A			65.3	74.3	59.3	61.8
B	74.3	65.3			57.9	52.6
C	61.8	59.3	52.6	57.9		

表5 各手法-人の再現率・適合率(20文抽出)

	KW密度		tf*idf		表層	
	再現	適合	再現	適合	再現	適合
A	41.6	45.0	43.8	47.1	48.1	52.0
B	42.2	39.7	45.4	42.9	52.8	49.7
C	42.7	44.5	45.7	47.7	49.7	46.1

表4と表5を比較すると、人同士の間の判定のばらつきに比べて、各手法と人との間の判定のばらつきが小さいことが分かる。同様の傾向は5文抽出でも見られた。本結果を見る限り、機械的な基準で抽出できる割合は、人によらずほぼ一定であるという現象が見られた。

最後に、決定木学習の属性にtf*idfの値を追加しても、効果が得られなかった原因について考察する。本実験では、決定木作成の終了条件を、エントロピーの利得率が0のときと設定した。そのため、必要以上に学習データに適合した細かい枝の木が作成されるという過学習が起き、特に数値データにおいてその影響が大きかったと考えられる。決定木作成の終了条件を緩めることにより、改善が見込める可能性がある。

7 まとめ

本稿では、講演文を対象とした重要文抽出実験について述べた。汎用的で講演文に直接利用可能である、単語の頻度を利用した手法より、講演文の表層の手掛かりを利用して決定木学習を行った手法の方が有効であり、特に5文抽出において大幅な精度の向上が見られた。

本実験では、文の重要度を文ごとに計算し、重要度が高いと判定された上位の文を抽出した結果を評価した。抽出した文全体が要約としてのまとまりを持つか否かについては考慮していない。今後の課題としては、抽出された文同士の関係や講演全体の構成を反映し、全体としてまとまりを持つ要約文の生成について検討することが挙げられる。

参考文献

[1]H.P.Luhn: "The Automatic Creation of Literature Abstracts", IBM Journal of Research and Development, Vol.2, No.2, pp.159-165(1958).

[2]Zechner,K: "Fast Generation of Abstracts from General Domain Text Corpora Extracting Relevant Sentences", Proc. of the 16th International Conference of Computational Linguistics, pp.986-989(1996).

[3]山本和英,増山繁,内藤昭三: "文章内構造を複合的に利用した論説文要約システム GREEN",自然言語処理, Vol.2, No.1, pp.39-55(1995).

[4]Hideo Watanabe: "A Method for Abstracting Newspaper Articles by Using Surface Clues", Proc. of the 16th International Conference on Computational Linguistics,pp.974-979(1996).

[5]野本忠司,松本裕治: "人間の重要文判定に基づいた自動要約の試み", 情報処理学会自然言語処理研究会報告,120-11,pp.71-76(1997).