E-027

新聞記事コーパスの共通文を利用した重要文抽出方式

Sentence extraction using the common sentence of newspaper articles

松村　繁男† Shigeo Matsumura　**山田　剛一**† Koichi Yamada　**絹川　博之**† Hiroshi Kinukawa　**中川　裕志**‡ Hiroshi Nakagawa

1．はじめに

近年のコンピュータの発展・普及とともに、電子化された文書数は急激な増加の一途を辿っている。膨大なテキストデータから必要な文書を見つけ出す際、多くの文書に目を通さなければならなくなっており、性能の良い検索システム・自動要約システムへのニーズが高まっている。これに伴い、膨大な情報の概要や内容を短時間で把握することが必要となっており、これを補助する技術として、テキスト自動要約技術が注目されている[1]。

従来の研究では、要約手法として、統計的な単語重み付け法と、文書中の表層的な情報を基にした重み補正を組み合わせ、単語重みから重要と判断される文を抽出する重要文抽出型の手法が多く提案されている。

本研究では、新聞記事において、見出しとの共通度の高い文を重要文として抽出する方式を提案する。提案方式は、新聞記事の見出しは、記事の内容を表す究極の要約であり、これに類似している文は重要である可能性が高いということに着目した方式である。

本稿では，提案方式における重要文を含む段落抽出アルゴリズムおよび精度評価結果について述べる。

2．重要文を含む段落抽出方式

重要文を含む段落は、まず新聞記事から、見出しを抽出し、見出しに含まれる単語を用いて、記事本文の各段落との共通度を算出することによって決定される。以下に、「社説記事」の記事構成や見出しにおける特徴と具体的な抽出手法について述べていく。

2.1.「社説記事」の記事構成・見出しの特徴

「社説記事」の構成は、まず、はじめに記事の主題や問題提起が述べられており、次にその主題や問題の詳細な情報や解決方法について論じられ、その次に詳細な情報な情報を受けての問題点や問題解決方法の代替案について論じ、最後にそれらすべてを受けての結論が述べられる、という「起承転結」の形になっていることが多い。したがって、このような形式の記事では、記事の主題を示す「起」の部分と、著者の考えをまとめた「結」の部分は非常に重要であるといえるだろう。

また、「社説記事」における見出しは、見出しの前半部が記事全体の主題にあたる内容を示し、後半部は記事についての著者がもっともいいたいことを集約した部分となっていることが多い。特に、毎日新聞社の社説記事では、主題が分離されており、例えば「［社説］心の教育　言いっ放しにならないか」という見出しがあった場合、この記事は「心の教育」について書かれている記事で、その中で著者がもっともいいたいことは、「言いっ放しにならないか」ということである。このように、見出しには「起承転結」の「起」にあたる部分と「結」にあたる部分に相当する内容の要約が書かれている。

「起承転結」で構成されている文書では、「起」にあたる内容は一番初めの段落に、「結」にあたる内容は最終段落に書かれていることが多い。しかし、必ず一番初めの段落と最終段落に書かれているわけではない。そこで、見出しとの共通度を算出することで、「起承転結」で構成されている文書で重要な「起」と「結」にあたる部分を抽出することができるのではないかと考えた。以下に、見出しと段落の間での共通度の算出方式について述べていく。

2.2.　各段落との共通度算出

見出しと各段落との共通度は、各段落の文に対して、見出しがどの程度の割合で一致しているかで決定される。以下に具体的な手順を述べる。

(1) 見出しを「起」にあたる部分と「結」にあたる部分に分割し、分割した見出しと記事本文を段落ごとに形態素解析を行い、単語に分割する。このとき、共通度算出に用いる単語の品詞は、名詞・形容詞・動詞・未知語とする[2]。

(2) 分割した見出しと記事本文の段落の間で、単語同士のマッチングを行い、共通度を算出する。スコア算出式は (式 1) にて定義する。

$$Score(Title, A) = \frac{match(Title, A)}{ele(Title)} \quad \text{(式 1)}$$

これは、段落 A と見出しの共通要素(match(Title, A))が見出し全体の要素(ele(Title))に対して、どの程度含んでいるかを算出するというものである。

(3) 共通度算出の結果、最も共通度の高い段落を、「起(結)」の段落であるとする。

なお、共通度が等しい段落が出現した場合、「起」の重要段落抽出のときは、より前の段落を、「結」の重要段落抽出のときは、より後ろの段落を取得するようにしている。例えば、「起」の共通度を計算していて、第 1 段落と第 3 段落の共通度が等しくなってしまった場合は、第 1 段落のほうを「起」の要素を含む段落として取得するということである。これは、「起」にあたる内容は初めに、「結」にあたる内容は後ろに書かれていることが経験的に多いとされているからである。

なお、社説記事の中には、見出しの前後が明確に区分されていない場合がある。しかし、見出しの前半部は「起」に、後半部は「結」を表現していることが多いので、見出しの前半部分との共通度の高い段落を「起」の段落、後半部分との共通度が高い段落を「結」の段落とし、段落の抽出を行う。

†東京電機大学大学院　工学研究科　情報通信工学専攻

‡東京大学　情報基盤センター

3． 精度評価実験

本方式の性能を確認するため、NTCIR Workshop2 TSC1 の重要文抽出型要約タスク (Task A1) で用いられたテストコレクションを用いた評価を行った。

本評価では，「起(結)」にあたる部分として抽出した段落に、テストコレクション中で指定された重要文が含まれているかで評価している。したがって精度評価の定義式は (式 2) のようになる。

$$\text{精度} = \frac{\text{重要文を含む段落を抽出した記事の総数}}{\text{評価対象となる記事の総数}} \quad \text{(式 2)}$$

今回評価対象とした記事は、Dry-run に含まれている社説記事 13 件である。評価は、要約率ごとに指定された重要文が含まれた段落を抽出したかで行っている。

3.1. 評価結果

上記に述べたデータを用いて、精度評価を行い、baseline と比較した結果を表 1、2 に示す。baseline には、共通度を算出せずに、「起」として第一段落を取得し、「結」として最終段落を取得した場合を用いている。

表 1 「起」における抽出精度評価結果

	10%	30%	50%	Average
baseline	0.769	0.923	0.923	0.872
提案方式	**0.923**	**1.000**	**1.000**	**0.974**

表 2 「結」における抽出精度評価結果

	10%	30%	50%	Average
baseline	0.385	0.692	0.769	0.615
提案方式	**0.538**	**0.769**	**0.846**	**0.718**

3.2. 考察

表 1、2 の結果より提案手法の方が優れているということがいえる。また、全体では、baseline より約 10%精度が向上していた。このことより、本手法の有効性を示すことができた。しかし、重要文を含む段落を完全に抽出することはできていない。特に「結」においては、まだまだ精度が低いといえるだろう。以下に精度低下の要因について述べていく。また、今回は Dry-run による評価なので、データ量が少ない。今後は、Formal-run や他の多くの社説による実験をする必要がある。

3.2.1. 略語に対する脆弱性

見出しというのは、記事全体の要約である。したがって、記事全体の主題となる単語も省略した形式で書かれていることが少なくない。例えば、「［社説］住銀・大和提携４大証券時代が終焉した」という見出しの記事があった場合、「住銀・大和提携」という部分が記事の主題を表していることになるのだが、記事本文では「住銀」が「住友銀行」などと書かれているために単語として一致させることができないことがある。このため、共通要素が減ることになり、本来抽出したい段落を抜き出すことができないというケースがあった。

コーパスから名詞と略語を自動獲得する研究というのが行われているので、そういったものを利用していけばこの問題は解決できるのではないかと考えている[3]。

3.2.2. 結論の重要性

記事全体の結論を述べている段落は、著者がもっともいいたいことが書かれている段落である。このため、全体から見れば重要だといえるのだが、その段落を見ただけでは、どのような経緯を経て、そのような結論に至ったのか、ということがわからなくなってしまう。

NTCIR の正解データでは、結論を述べている段落の文が必ずしも重要文として抽出されているわけではなかった。特に、要約率 10%の場合では、ほとんど抽出されていなかった。

3.2.3. 見出しの内容のゆらぎ

はじめにも述べた通り、見出しは記事の内容を表す極端な要約である。そして、「社説記事」における見出しは主題についての部分と結論についての部分に分かれていることが多かった。しかし、中には、結論を述べていない見出しも存在していた。このような見出しでは、記事の主題となる部分とそれに対する問題提起の中心となる部分、「起承転結」でいうと、「承」や「転」の冒頭部にあたる部分である。NTCIR の正解データでは、重要文として抽出されていることはほとんどなかった。

4．おわりに

今回、新聞記事コーパスの見出し情報との共通度を用いた重要文を含む段落抽出方式を提案し、その有効性を示すことができた。

今後は、3.2. 節で述べた問題の改善および抽出精度の向上、そして段落単位より細かい文単位での抽出精度の評価を行っていく予定である。また、要約率 10%においては「起」の段落と「結」の内容を含む段落に抽出すべき重要文が含まれていることが多いが、要約率 30%、50%では、そこからだけでは足らず、「承」と「転」の内容を含む段落からも重要文を抽出する必要がある。よって、「承」と「転」の切れ目を判別し、今回提案したシステムと組み合わせ、「起」「承」「転」「結」に対応する部分から重要文を抽出できるようにすることも今後の課題である。

謝辞

毎日新聞記事データの使用許諾をしてくださった毎日新聞社、形態素解析器「茶筌」の各開発者の方々に感謝いたします。

参考文献

[1] 奥村学，難波英嗣:“テキスト自動要約に関する研究動向(巻頭言に代えて),”自然言語処理，Vol.6, No.6, pp.1-26, 1999.

[2] 松本裕治，北内啓，山下達雄，平野善隆，松田寛， 高岡一馬，浅原正幸:“形態素解析システム茶筌．”http://chasen.aist-nara.ac.jp/，2000.

[3] 酒井浩之，増山繁:“コーパスからの名詞と略語の対応関係の自動獲得”，言語処理学会第 9 回年次大会発表論文集，pp.226-229．2003.