

テキストの結論重視型要約の生成

Conclusion-Oriented Text summarization

谷川信弘 砂山渡
 Nobuhiro Tanikawa Wataru Sunayama

広島市立大学院 情報科学研究科
 Graduate School of Information Sciences, Hiroshima City University

Conclusion is the most important in a text, which is described in order as backgrounds, expansion and conclusion especially in Japanese texts. In the case of information retrievals, texts are often evaluated by their conclusions. In this paper, sentences strongly related to the conclusions are extracted as a summary according to keywords in a title of a text, words frequency, and their co-occurrence information.

1. 序論

近年、Webの発展により、膨大な量のテキストの情報がWeb上から得ることが可能となった。テキストには様々な種類や長さのものがあり、そこから必要な情報を得る為に、重要な文を抽出し自動要約をする研究が盛んに行われている。一般的にテキストの最も重要な部分は結論部分であることが多く、結論を把握するには、テキストの背景や内容を理解してから結論にたどり着くことが多い。しかしその為には多くの時間を費やしてしまう。また、テキストは無数に存在する為、文数が比較的短いコラム記事などでも、複数のテキストの内容を理解するには多くの時間がかかる。そこで、テキストの題名に対する結論のみを抽出することができれば、テキストの結論が短時間で把握できるだけでなく、テキストの内容や重要部分の理解の支援もできると考えられる。

本研究では、テキストの題名に対する結論となる文を抽出することで、結論重視型の要約を生成するシステムを提案する。

2. 関連研究

本研究では、テキストの結論文を抽出し、結論重視型の要約を生成するシステムを提案する。まず、テキストの結論抽出に関連する研究として、話題の継続に着目した国会会議録要約[1]がある。この研究では、国会会議録を要約するにあたって、話題に対する導入段落と結論段落を要約として出力することを目的としている。結論部分を抽出するという点は本研究と同じだが、本研究では結論を文単位で評価し出力する。また対象とするテキストの種類も異なる。

次に重要文抽出する手法を用いて新聞記事を要約する研究[2]がある。この研究では、重要度を決定する際に複数の記事の語彙特性を用いている点で本研究と異なっている。また文中の主題や焦点を用いて重要文を抽出し要約を生成する研究[3]がある。この研究は重要文を抽出する際に、文中の単語の出現情報を用いる点は共通している。しかしこれら従来の研究では重要文を抽出することが目的だが、本研究では結論文を抽出することが目的である。重要文が必ずしも結論ではない為、従来の研究とは異なると言える。

3. 結論重視型要約システム

図1に提案システムの構成図を示す。本システムでは、文章に一貫性があり、話の広がりもあるテキスト(コラム記事など)と、テキストの題名中の名詞「観点語」を入力とする。文中の名詞と観点語との共起頻度を元に、名詞へ話の展開を表す「展開レベル」を与える。名詞の展開レベルを用いて文に評価値を与え、評価値の高い文を「結論文」として抽出する。

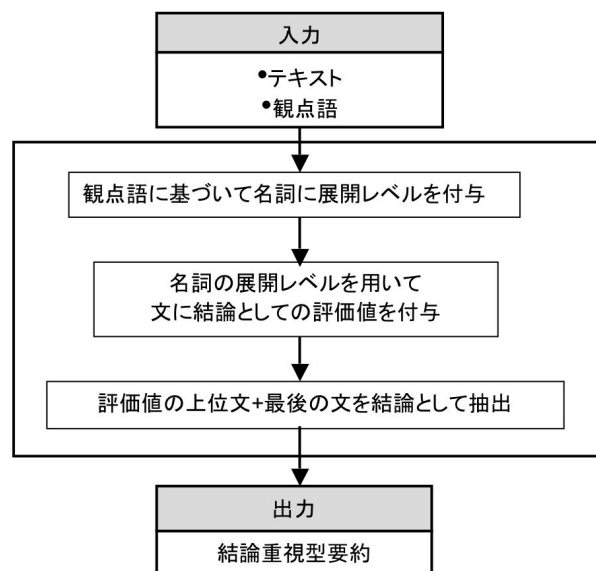


図1: 提案システムの構成

3.1 入力：テキスト・観点語

本システムへの入力、文章に一貫性があり、話の広がりがあるテキストとその観点語とする。「一貫性」とは、テキストの内容が観点語に沿っていること、「広がり」とは、観点語を含まない具体例や関係のない文があることを表す。

3.2 名詞への話の展開レベルの付与

「茶釜」[5]を用いて形態素解析を行い、テキストの題名と本文から名詞を抽出する。文中の名詞と観点語の共起関係により、表1のように名詞に展開レベルを与え、分類する。展開レ

連絡先: 谷川信弘, 広島市立大学情報科学研究科,
 tani@sys.info.hiroshima-cu.ac.jp

ベルとは、名詞が示す「話の展開具合」の基準である。各展開レベルの持つ意味を表2に示す。

表 1: 名詞の展開レベルによる分類

展開レベル 0	テキスト中で初めて出現した名詞
展開レベル 1	観点語と同文に初めて出現した名詞
展開レベル 2	観点語と同文に再度出現した名詞
展開レベル 3	観点語の無い文に出現した、レベル 2 の名詞
展開レベル 4	観点語と同文に出現したレベル 3 の名詞

表 2: 各展開レベル名詞の意味

展開レベル 0	話題の始まりや、新しい意見を表す名詞 観点とは関係のない名詞
展開レベル 1	観点語に関する話題や新しい意見を表す名詞
展開レベル 2	観点語に関する話題の中心となっている名詞
展開レベル 3	話が広げられた名詞
展開レベル 4	観点語に関係し、話の結論が得られた名詞

3.3 文の評価値の算出

観点語と名詞の展開レベルを用いて、1文ごとに「題名に対する結論」としての重要度を示す評価値を算出する。「題名に関する結論」とは、題名に関連し、かつ題名に関連した意見が述べられている文であると考えた。

3.3.1 評価式

式(1)を文*i*の最終的な評価値 $final_score(i)$ として計算する。

$$final_score(i) = (score(i) - N) \times W \quad (1)$$

式(1)により、結論らしさの強い文ほど高い評価値となるような値を与えることができる。 $Score(i)$ を式(2)、 N を式(3)、 W を式(4)により計算する。

$$Score(i) = 100 \times (P_n + \sum_{j=0}^4 P_j) + x_1 + x_2 + x_4 - (x_n + x_3) \quad (2)$$

$$\text{ただし} \begin{cases} 0 & P_j & 1 \\ \sum_{j=0}^4 P_j & = & 1 \end{cases}$$

$$\begin{cases} P_j = K_j & (x_i \leq \bar{x}_i) \\ P_j = 0 & (\text{上記以外}) \end{cases}$$

評価式に用いる各パラメータについて説明する。まず、使用するテキストの観点語 n の一文平均個数 \bar{x}_n と、レベル j の名詞の一文平均個数 \bar{x}_j を計算する。次に、先頭から i 番目の文の観点語数と各レベルの名詞数をカウントし、その個数により P_n と P_j に値を与える。予備実験を行い、式(2)の P_n, P_j の値を $K_n = 0.5, K_0 = 0.3, K_1 = K_4 = 0.1, K_2 = K_3 = 0$ のように決定した。 K_n の値は観点語がテキストの一文平均個数より多い文に与えられる。すなわち、他の文よりも題名に対

する話が多い文を評価できるパラメータである。 K_j についても同様で、各展開レベルの名詞がテキストの一文平均個数より多い文には高い評価値を与える。展開レベル j の名詞が多い場合に評価するためのパラメータである。

式(2)を用いて、観点語や各展開レベルの名詞の個数が、テキストの一文平均個数より文より多い文を評価する。 P_n と $\sum_{j=0}^4 P_j$ の和を 100 倍としたことで、観点語や各展開レベルの名詞がテキストの平均より多い文では、その個数にかかわらず高い評価値を与えることができる。このままではテキストの平均以上に名詞が出現する文では評価値の差がでないので、展開レベル 1, 2, 4 の名詞の出現個数を加えた。これらの名詞は題名に関連しているため、結論文には多く含まれると考え、加点した。ここで観点語の出現個数を引いているのは、結論文が題名についての話(観点語)ばかりではないと考えたためである。反対に話が広がり、題名の内容から逸れている文も結論文ではないと考え、 x_3 も値から引いた。

3.3.2 補正值

式(2)のみでは、観点語、レベル 1, 2, 4 の名詞が多くなる長文ほど評価値が高くなってしまふ。そのため、 i 番目の文がレベル 4 の名詞を含まず、かつ名詞数 num が 14 個以上の文を対象に、以下の式(3)による補正值 N を与える。これは長文であってもレベル 4 の名詞を含むことにより、結論に関する話がされていると考えたためである。また、ここで基準とする名詞数は予備実験の結果を基に決定した。

$$N = \begin{cases} (num - 14) \times 2 \\ 0 \end{cases} \quad (3)$$

3.3.3 文の出現位置による重み

結論とは、テキストの最後に存在する可能性が高いと考え、テキスト中の後半の文ほど評価値が高くなるような重みをつける。文章によっては冒頭部分に結論が存在することもあるが、その場合でも最後にまとめとして結論が存在することが多い。テキストの全文数 all の中で i 番目の文の位置による重み W を以下の式(4)によりつける。

$$W = \frac{i}{all} + 1 \quad (4)$$

3.4 出力：要約

式(1)により得られた評価値の高い上位 9 文を抽出する。更に、テキストの最後の文を評価値に関わらず抽出する。これらを含めた計 10 文を結論文として出力する。ただし文章中の結論文自体が少ないテキストも存在することを考慮し、評価値が 85 点以下の文は出力しない。この評価値の閾値は予備実験を基に算出した。

出力は html 形式で、ページ前半に要約文を表示し、ページ後半にはテキストの原文を表示する。また、要約文として出力された文は原文中で赤く表示し、結論文の位置が把握しやすいようにした。出力例として、本論文に提案システムを適用して得られた結論文を表 3 に示す。

4. 評価実験

本章では、提案システムによる結論文抽出の精度を評価する為に行った実験について説明する。

表 3: 提案システムの出力例

まず、使用するテキストの観点語 n の一文平均個数 x_n と、レベル j の名詞の一文平均個数 x_j を計算する。
5. 提案システムの適用範囲本節では提案システムがコラム記事以外のテキストにも適用可能であるかを検証した。ニュース記事については精度の良い結果が得られなかったが、1 テキストの平均文数が 17。
このことから本システムは掲示板のような形態のテキストでの結論抽出には適さないと考えた。
表 7 より、条件を満たすテキストにおいては結論文抽出の再現率が 0。本システムは、題名に使用されている名詞に注目し、この名詞とテキスト中に出現するその他の名詞との共起頻度を用いて、各名詞を評価する。これらの名詞を用いてテキストを文単位で評価し、評価の高い文を結論として抽出する。
評価実験を行い、結論文が抽出できることを確認した。
また今後の課題として、適用条件を満たさないテキストに対しての抽出精度が挙げられる。
解決策として、観点語をシステムにより推薦、追加することで一貫性レートをシステムの適用範囲内にすることが考えられる。

4.1 実験概要

本実験は提案システムによる結論文抽出が可能であることの確認を目的とする。実験では 20 個のコラム記事を用意し、大学生、大学院生の計 20 人の被験者にシステムによる出力が結論であるか判断してもらった。用いたコラム記事は、観点語がある程度含まれているが、観点語の出現頻度が高すぎず、先述の入力条件を満たすテキストである。比較システムとして展望台システム [4] を用い、提案システムと比較した。展望台システムは話の観点に基づいて重要文を抽出するシステムである。提案、比較システム共に観点に基づいた文抽出を行うシステムの為、展望台システムは比較システムとして妥当であると考えた。

1 つのコラム記事について、各システムを 5 人に利用してもらい、過半数が結論とした文を正解とした。各システムにより出力された結論文と正解の文を比較した。実験の手順を以下に示す。

1. システムが出力した結論文を読み、その後原文を最後まで読んでもらう。
2. 結論文の中から結論として不要な文を選んでもらう。
3. 原文の中から、結論文以外に、結論に加えるべき文を選んでもらう。

手順 2 において過半数の人が除外しなかった文と、手順 3 において過半数が追加した文を正解文とした。

4.2 実験結果

被験者による正解文とシステムが出力した結論文との適合率、再現率のテキスト平均を表 4 に示す。また提案システムにおいてテキスト最後の文を出力しない場合の適合率、再現率の平均も表 4 に示す。結論文の中で、被験者が結論ではないと判断した文の 1 人当たりの平均個数を図 2 に示す。被験者が結論として追加した文の 1 人当たりの平均個数を図 3 に示す。

表 4: 各システムが出力した結論文と正解文の適合率、再現率の平均

	適合率	再現率
提案システム	0.775	0.914
提案 (最後の文出力除く)	0.697	0.768
比較システム	0.600	0.773

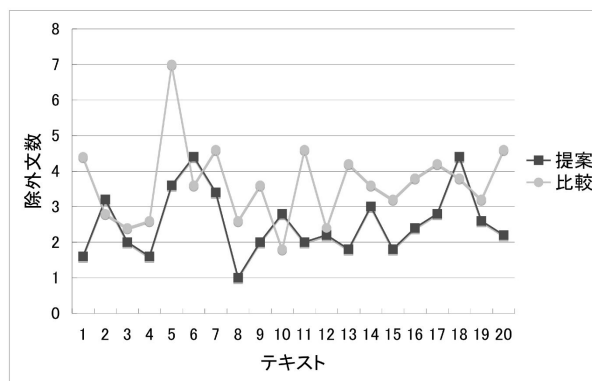


図 2: 被験者により除外された文数

4.3 考察

表 4 より、提案システムの方が再現率が高くなった。提示した文に観点語が存在することで、結論であると感じた被験者が多かった。評価値を付けるにあたって、観点語のパラメータ P_n に他の名詞より高い値を与え、「観点語を含む文」に高い評価値を与えて多く出力したことが、精度が良くなった要因と考えられる。それに加え、重みを用いて「テキスト後半の観点語を含む文」を出力したことで、結論文が抽出できたと考えられる。適合率が低いことについては、結論に説明や具体例などが抽出されたことが原因だと考えられる。詳しい原因については以下で述べる。

図 2 では、ほとんどのテキストにおいて、提案システムの方が除外された文が少ないことが分かる。しかしテキスト 2, 6, 10, 18 の 4 つにおいては提案の方が除外された文が多く、結論文以外の文を多く出力してしまった。除外された文の特徴として次の 2 点が挙げられる。

1. 題名に対する具体例が述べられた文。
2. 結論の説明となる文。

まず 1. について、具体例が述べられた文には様々な名詞が出現する為、評価値が高くなり抽出されたことが考えられる。しかし題名によっては具体例が結論として必要になることもあるので、不必要な文とは一概に言えないと考えられる。具体例が述べられた文の出力を減らすには、長文に対する補正值 N を大きくし、具体例の文の評価値を低くすればよいと考えられる。次に 2. についても名詞数が多いことで出力される原因となった。この文は結論の理解に役立つ文だが最終的な結論ではない為、被験者によっては不必要と判断する人がいたことが挙げられる。

図 3 では、ほとんどのテキストにおいて提案システムの方が追加された文が少ないことが分かる。しかしテキスト 7, 9, 14, 18 の 4 つにおいては提案システムの方が追加された文が多く、結論文を出力できていなかった。追加された文の特徴として次の 3 点が挙げられる。

1. 結論の説明となる文。
2. 指示語が用いられ、名詞数が少ない文。
3. 結論が複数の文に渡って箇条書きされている文。

まず 1. については、先述の「結論から除外された理由」と同じく被験者によって判断が分かれたことが挙げられる。次に 2. において、特に観点語が指示語に置き換えられている場合は、評価値がかなり低くなり抽出されない。また 3. においても、箇条書きされることにより観点語や名詞数が少なくなり評価値が低くなっていた。

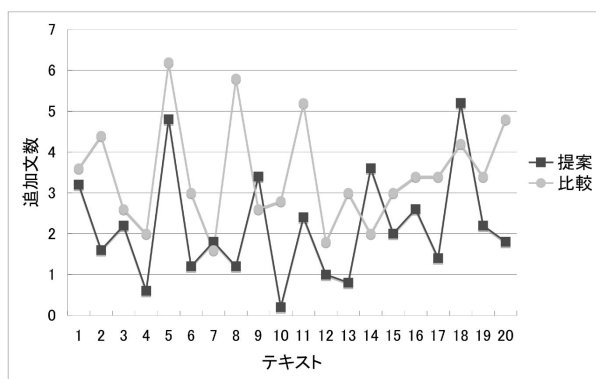


図 3: 被験者により追加された文数

5. 提案システムの適用範囲

本章では提案システムがコラム記事以外のテキストにも適用可能であるかを検証した。用いたテキストは表 5 のようにジャンル分けし、各 20 テキスト、2 ちゃんねる 10 テキスト用いた。本システムでは、「文章に一貫性があり、話の広がりもあるテキスト」という特徴を持つテキストを用いた。このような特徴を一貫性レート R として以下の式 (5) で算出した。 V はテキスト中で観点語を 1 つ以上含む文数、 all はテキスト中の全文数を表す。

$$R = \frac{V}{all} \quad (5)$$

本システムを適用したときの結論文抽出の再現率が 0.66 以上、すなわちテキストの結論の 3 分の 2 以上を出力できることを目標に設定した。目標が達成できたテキストの割合を表 6 に示す。ニュース記事については精度の良い結果が得られなかったが、1 テキストの平均文数が 17.7 文と短いことから要約を行う必要性は低いと考えられる。また 2 ちゃんねるについては、他のテキストに比べ、文数が極端に多く一貫性が低い。更に複数の人が意見を述べるだけの内容である場合が多い。このことから本システムは掲示板のような形態のテキストの結論抽出には適さないと考えた。このことからシステムの適用条件を「テキストの文数が 30 文以上 200 文以下」とした。更に「一貫性と広がり」のテキスト条件として「 $0.4 \leq R \leq 0.7$ 」を条件に加える。これらの条件を満たすテキストにおいて、目標が達成できたテキストの割合を表 7 に示す。

表 5: 一貫性レートを確認したテキスト

使用テキスト	一貫性レート平均	テキスト平均文数
ニュース記事	0.574	17.7
論文	0.541	90.3
コラム記事	0.403	50.5
ブログ記事	0.307	29.0
2 ちゃんねる	0.255	656.4

表 7 より、条件を満たすテキストにおいては結論文抽出の再現率が 0.66 以上の目的を達成したテキストの割合が高いことが分かる。よって提案システムの適用範囲の条件の妥当性が確認できた。

表 6: 結論文と正解文の再現率が 0.66 以上のテキストの割合

テキストジャンル	目標達成割合	達成テキスト数
ニュース記事	0.15	(1/20)
論文	0.60	(12/20)
コラム記事	0.85	(17/20)
ブログ記事	0.40	(8/20)
2 ちゃんねる	0.00	(0/20)

表 7: 条件を満たすテキストでの結論文と正解文の再現率が 0.66 以上のテキストの割合

テキストジャンル	目標達成割合	達成テキスト数
ニュース記事	1.00	(1/1)
論文	0.72	(8/11)
コラム記事	0.94	(15/16)
ブログ記事	0.50	(2/4)
2 ちゃんねる		(0/0)

6. 結論

本研究では、テキストの題名に対する結論となる文を抽出することで、結論重視型の要約を生成するシステムを提案した。本システムは、題名に使用されている名詞に注目し、この名詞とテキスト中に出現するその他の名詞との共起頻度を用いて、各名詞を評価する。これらの名詞を用いてテキストを文単位で評価し、評価の高い文を結論として抽出する。評価実験を行い、結論文が抽出できることを確認した。また今後の課題として、適用条件を満たさないテキストに対しての抽出精度が上げられる。解決策として、観点語をシステムにより推薦、追加することで一貫性レートをシステムの適用範囲内にすることが考えられる。

参考文献

- [1] 川端正法, 山本和英: 話題の継続に着目した国会会議録要約, 言語処理学会第 13 回年次大会, pp.696-699, (2007)
- [2] 大竹清敬, 岡本大吾, 児玉充, 増山繁: 重要文抽出, 自由作成要約に対応した新聞記事要約システム YELLOW, 情報処理学会論文誌, Vol.43, No.SIG_2, pp.37-47, (2002)
- [3] 横山晶一, 菅野崇, 西原典孝: 主題・焦点リンクを用いた重要文抽出システム, 情報処理学会研究報告, Vol.2003, No.76(20030725) pp. 1-6, (2003)
- [4] 砂山渡, 谷内田正彦: 観点に基づいて重要文を抽出する展望台システムとそのサーチエンジンへの実装, 人工知能学会論文誌, Vol.17, No.12, pp.14-22, (2002)
- [5] 松本裕治, 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』, Version2.3.3, 使用説明書, (2002)