GDAタグを用いたテキスト自動要約

野村雄司† 伊藤一成‡ 斎藤博昭†
†慶應義塾大学 理工学部 ‡慶應義塾大学 大学院理工学研究科
{yuji,k_ito,hxs}@nak.ics.keio.ac.jp

1 はじめに

電子化された膨大な情報が溢れている現在、必要な部分だけを入手するために、情報を制御し短い時間で的確に内容を把握するといった自動要約の必要性が高まっている。これまで様々な要約研究が行われてきたが[1][2]、表層的な情報を用いたものが主流である。これらの方法では要約の品質の向上はすぐに限界にきてしまい、今後は内容に基づく処理が必要だと考えられる。近年深層的な情報を用いた研究もなされているが、現在の技術ではまだ十分でない。そこで、GDAタグを用いた首尾一貫性の高い自動要約システムを提案する。また、GDAタグのどのような情報が要約において有用であるかの検証をする。

2 Global Document Annotation

GDA (Global Document Annotation) は多言語間に共通の統語・意味等に関するXMLタグの標準を作って普及させようというプロジェクトである[3]。

GDAの目的は主に以下のようなことである。

1. タグを用いた機械翻訳、情報検索、要約、質問応答、知識発見などを実用化する
2. それによってタギングのメリットを生じさせ、多くのユーザが自分のファイルにタギングするように仕向け、タグを普及させる
3. タグによって構造化されたデータを自然言語処理、人工知能、言語学などの研究に利用する

GDAでは、文法機能 (主語、目的語、間接目的語)、主題役割 (動作主、非動作主、受益者など)、修辞関係 (理由、結果など) や照応関係を表わすことができる。これらの情報を用いることで、要約においてより自然な文を生成することが可能になると考えられる。

3 従来手法の問題点

抜粋、つまり重要文抽出による多くの要約手法では、文 (または形式段落) を1つの単位として、それらに何らかの情報を基に重要度を付与し、その重要度で順序付け、重要な文を選択し、それらを寄せ集めることで作成する。重要度の評価には、単語の出現頻度やテキスト中での位置情報やタイトル情報や文間のつながり情報など様々な情報を用いた研究がされている。

重要文抽出における全体的な問題点としては、抽出した文中に照応詞が含まれている場合、その先行詞が要約文中に存在する保証がないことやテキスト中の色々な箇所から抽出したものを単に集めているため抽出した文間のつながりが悪いことが挙げられる。

照応詞の問題の対策として、照応詞を含む文の前の数文を追加することで対処する方法がとられているが、これは冗長な情報を多く含んでしまい、完全な解決策とはなっていない。より自然で冗長の少ない要約を実現するためには、照応詞の指している語に置き換えるなどの処理が必要だと考えられる。文間のつながりの問題に対しては、接続詞を削除することで部分的な対処がされている。

4 本研究の手法

GDAタグを利用した要約手法として、Nagaoらの活性拡散を用いた手法[4]があるが、本研究では、これまでにある程度成果のでている既存の手法にGDAタグから得られる情報を組合わせることで、より精度が良く、首尾一貫性の高い要約を試みる。

本手法の全体的な流れとしては、重要文抽出、各文に対しての重要箇所抽出、照応詞処理の3段階の処理を行なう。以下にそれぞれの方法について述べる。

4.1 重要文抽出

重要文抽出は、吉見らの提案した手法 [5] に基づいて重要度を決定する。この手法は、タイトルはテキスト中で最も重要な文であり、重要な文へのつながりが強い文ほど重要な文であるという考えに基づき、語のつながりから文間の関連度を求めることで文の重要度を決定する。文の関連度を重要度評価に利用しているため、比較的文間のつながりの良い抽出が可能だと考えられる。吉見らは 1 文前の人称代名詞と先行（代）名詞の照応の検出と 5 文前までの語彙的なつながりの検出によってのみで、文間の関連度を求めていたが、本手法では処理対象文を限定せずに、照応と語彙的なつながりの検出に加え、GDA タグの関係属性の情報から代用、省略の検出も行なう。以下に関係属性による照応、代用、省略の検出について述べる。

- 照応

  <persname id="M">松井</persname>のヤンキース入りが決まった。3 年契約とのことだ。<np eq="M">彼</np>はメジャーでも活躍してくれるだろう。

  この例では、先行詞と照応詞が 1 文以上離れているが、eq 属性と id 属性を参照することによって正確に照応関係を検出することができる。

- 代用

  <persname id="M">松井</persname>のヤンキース入りが決まった。<np eq="M">ゴジラ</np>はメジャーでも活躍してくれるだろう。

  この例では、eq 属性と id 属性を参照することによって、「松井」と「ゴジラ」が同一の語であることがわかる。

- 省略

  <persname id="M">松井</persname>のヤンキース入りが決まった。メジャーでも<vp agt="M">活躍し</vp>てくれるだろう。

  この例では、agt 属性と id 属性を参照することによって、「活躍し」の行為者が「松井」であることがわかる。

4.2 文内の重要箇所抽出

より冗長の少ない要約を行なうために、各文に対して重要箇所だけを抽出する。基本的には、GDA タグの文法機能（主語、目的語、間接目的語）、主題役割（動作主、非動作主、受益者など）、修辞関係（理由、結果など）の情報を利用して得られる文のテキスト構造から各単語に非重要度のスコアを付け、スコアの小さい語のみを抽出することで行なう。まず、文の必須語（主辞、主語、目的語）を抽出する。次に、図 1 のように、各節に対して GDA タグの文法の種類と係り受けに応じて重要度を付与し、ある閾値以上の語を抽出する。今回、閾値は要約率と各文の非重要度の最大値から決定した。

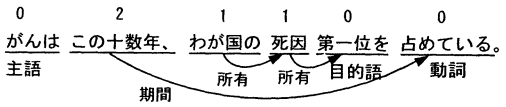

図 1: 重要箇所抽出の例

4.3 照応詞処理

先に述べた照応詞の問題を考慮して、抽出された文において照応詞が存在するがその先行詞が抽出されていない場合、照応詞に対して文間のつながりを自然にするための処理をする。処理としては、GDA タグの関係属性から得られる先行詞の種類に応じて、照応詞を先行詞で置換えるか、先行詞を含む文を追加するか、または照応詞を削除する。これにより、余計な文を追加することなく、より自然で要約率の良い要約が可能だと考えられる。

まず、関係属性から照応詞の先行詞を特定し、先行詞が名詞であったときに限り、照応詞の種類に応じた照応詞の置換えを行なう。以下に、照応詞の置換えを行なう例を示す。

> <np id="C">車</np>が止まった。<adp eq="C">その</adp>ドアが開いた。
>
> → 車のドアが開いた。

この例において、2 文目だけが抽出された場合、関係属性から照応詞「その」の先行詞が「車」であることがわかる。そして、先行詞が名詞であり照応詞が「の」で終わるため、照応詞を「車」＋「の」で置換える。

先行詞が名詞以外であった場合は、無理に置換えるよりも先行詞を含む 1 文を追加する方が自然であると考え、文の追加を行なう。例外的に、文を追加すると重要な文が除かれてしまう場合は、文を追加せずに照応詞を削除する。

また、照応詞処理だけでなく、より文間のつながりをよくするために、文の先頭に位置する接続詞において対象文が存在しない場合、接続詞を削除する。

4.4 ユーザ適応

タイトルによって重み付けし、語のつながりを文の重要度の評価に用いて重要文を抽出しているため、ユーザによって入力されたキーワードにタイトル同様に重み付けすることで、情報検索においてユーザの要求に即した要約を出力として利用することができる。また、要約の長さを調節できるようにすることで動的な要約を可能とした。

5 実験

GDA タグの付けられた 95 年の毎日新聞の 50 記事を要約の対象とした。重要文抽出、重要箇所抽出、照応詞処理の 3 つについて評価実験をした。それぞれの実験について以下に述べる。

5.1 重要文抽出

要約率 20%と 40%に設定し、3 人の人間が行なった抽出結果を正解として正解率を求めることで評価をした。文の先頭から順に抽出する lead 法と Microsoft Word によって重要文抽出したものと比較した結果を表 1 に示す。正解率の参考として、それぞれの要約者に対しても残りの要約者 2 人を正解として正解率を求めた。要約者の平均正解率は、要約率 20%と 40%ともに 72.3%であった。

表 1: 正解率の比較

	要約率 20%			要約率 40%		
	本手法	lead 法	Word	本手法	lead 法	Word
要約者 A	69.7%	57.3%	31.5%	72.2%	54.0%	44.3%
要約者 B	61.8%	59.6%	30.3%	67.0%	57.4%	44.9%
要約者 C	67.4%	60.7%	30.3%	68.8%	56.9%	43.2%
平均	66.3%	59.2%	30.7%	69.3%	56.1%	44.1%

また、GDA の関係属性の情報を用いたときと用いないときの比較した結果を表 2 に示す。

表 2: 関係属性の利用有無による正解率の比較

	関係属性あり	関係属性なし
要約率 20%	66.3%	65.2%
要約率 40%	69.3%	66.1%

5.2 重要箇所抽出

要約率を 40%に設定して、重要箇所抽出を行なった。重要箇所抽出の実験では、各文に対してどのくらい必要な情報を欠落することなく、文を短くすることができるかを評価する必要がある。そのため、この実験ではどの文が重要文として選択されたかは評価せずに、各文で削除された部分に対して、その削除が適切かどうかを判定することで評価をした。要約文として必要な情報が失われていないか、構文的に不自然な文が生成されていないかを不適切と判定する基準とした。削除に対する適合率を表 3 に示す。

表 3: 削除箇所の適合率

	括弧部分の削除	その他の削除
削除数（削除率）	57（5.0%）	182（11.0%）
適切な数	57	139
適合率	100%	76.4%

5.3 照応詞処理

重要箇所抽出した結果に対して、照応詞の処理を行なった。処理した箇所に対して、要約文としての情報量と文のつながりの自然さの 2 点を考慮し、処理前に比べて改善されたかを判定して評価をした。各処理の結果を表 4 に示す。

表 4: 照応詞の処理結果

	文の追加	照応詞置換え	照応詞削除
処理総数	2	2	2
適切な数	2	2	1

6 考察

重要文抽出においては、lead 法、Microsoft Word を大きく上回る結果が得られている。また、文のつながりの評価に関係属性を用いることで、要約率 20%、40%でそれぞれ 1.1、3.2 ポイントの改善が見られた。これは GDA の関係属性から得られる照応、代用、省略の情報が重要度の評価に有効に働いたと言える。また、吉見らは英文テキストを対象として実験をしてい

たが、この手法が日本語テキストに対しても有効であることが確認できた。

重要箇所抽出においては、文字数を16.0%削減することができたが、精度は76.4%であった。この結果は首尾一貫性の高い要約を目指す上で、課題の残る結果となった。不適切に削除された原因について以下に述べる。まず、構文上不自然になってしまった例を以下に示す。

処理前

消防本部によると、この副士長は昨年九月、大阪市東淀川区内のマンション一室を妻（２７）名義で借り、ダイヤルＱ２を二十四回線開設。

処理後

よると、この副士長は大阪市東淀川区内のマンション一室を妻名義で借り、ダイヤルＱ２を二十四回線開設。

本研究では係り受けに応じて非重要度を決定しているため、この例のように関連が強い語の間で削除されると不自然になってしまうことがある。これは、係り受けの情報だけでなく、語義などを用いて単語と単語の意味的な関係を考慮する必要があると考えられる。

また、構文上は正しくても、情報量の面で必要な情報を削除してしまったケースがあったが、この問題を対処するためには、文書全体における各文の役割を考慮する必要があり、これには文と文の関係を示す修辞構造の情報が有用だと考えられる。

照応詞の処理として関係属性を用いて照応詞の置換えをすることで、より自然な冗長の少ない要約結果が得られた。照応詞の置換えを行なった要約結果例を以下に示す。

処理前

それを踏まえた食管法廃止と新食糧法の制定方針だったはずだ。

処理後

今春の「平成コメ騒動」を踏まえた食管法廃止と新食糧法の制定方針だったはずだ。

今回の実験では、関係属性が付与されていなかったために照応詞を削除した結果、処理前に比べて不自然になってしまったケースがあったが、これは関係属性を正確にタグ付けることによって改善できると考えられる。

7 まとめ

GDAタグを利用した重要文抽出、重要箇所抽出、照応詞処理を行なった。重要文抽出においては、要約率20%、40%でそれぞれ正解率66.3%、69.3%と良好な結果が得られた。文のつながり情報を利用した重要文抽出法におけるGDAタグの関係属性の有効性が確認できた。また、文を自然にするための照応詞の処理においても関係属性が有用であることが確認できた。

しかし、重要箇所抽出においては係り受けなどの文章構造情報のみでは良好な結果は得られなかった。今後、文間の修辞構造の情報や語義情報がGDAタグ情報として明確に付与されるか、外部情報として利用することで、改善が可能だと考えられる。

今回は、人手でGDAタグ付けられたテキスト文書のみを対象としたが、機械的に付与されたテキストや表や図の入った文書集合や動画に対しても検討していきたい。

謝辞

本研究では、実験にGDAタグの付いたデータを利用させて頂きました。データを提供して頂いた電子技術総合研究所の方々には大変感謝しております。

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] 奥村学, 難波英嗣：テキスト自動要約に関する最近の話題, 自然言語処理, Vol.9, No.4, pp.97-116, 2002

[3] 大域文書修飾 Global Document Annotation (GDA), http://www.i-content.org/gda/

[4] Katashi Nagao and Koiti Hashida : Automatic text summarization based on the Global Document Annotation, In *Proceedings of COLING-ACL'98*, 1998

[5] 吉見毅彦, 奥西稔幸, 山路孝浩, 福持陽士：表題へのつながりに基づく文の重要度評価, 自然言語処理, Vol.6, No.1, pp.43-57, 1999