イベントの参照関係に注目した新聞記事の複数文書要約

北海道大学大学院 工学研究科
○吉岡 真治　原口 誠
e-mail:{yoshioka,makoto}@db-ei.eng.hokudai.ac.jp

1　緒言

一連の事象に関して報道された複数の新聞記事から全体の流れが分かるように要約を作成することは、文書の閲覧性の向上に役立つと考えられる。しかし、個々の記事から個別に要約を作成する方法では、別々の記事から同じ情報を持つ要約が作成される場合などがあり、全体としての情報量が下がるという問題がある。よって、記事間の関係を理解し、同じような記述は削除するという複数文書要約の技術が必要となる。本研究では、新聞記事が、主に、ある特定の日時に起きた特定の事象 (以下ではイベントと呼ぶ) に関する報道を行っていることに注目し、イベントの参照関係に基づいて、文書間の重要度を決定する方法を提案する。また、要約文書から共通のイベントに関する記述を削除することにより、冗長な表現を削除した要約の作成を行う。また、本手法を用いて NTCIR の TSC に参加した結果の評価についても述べる。

2　複数新聞記事からのイベントの参照関係の抽出

本研究では、ある事件が起きたという第一報から、その事件に関する影響や、事件の原因が特定されたといった続報などを組み合わせた複数の新聞記事から、全体の流れが理解できる要約の作成を目指している。

このような複数記事の特徴としては、第一報で報じられた事象に関する記述や新たに分かった事象が参照され、あらたな記述が追加されていくという関係があるという点にある。よって、各々の記事で共通に参照されている事象を同定し、その関係を整理することは、複数記事の関係を反映した要約作成に役立つと考えられる。

単一の記事などから、このような話の展開を追っていき、要約に応用する方法としては、語彙的結束性などに基づく研究などが行われているが、複数の記事間の対応関係を取る場合には、一文書内での文間の距離といった情報が利用できないため、よりグローバルな参照関係を取り扱う必要がある。一方で、単純な語彙的重なりにより、参照関係とする場合には、定期的に報告されるような新聞発表のように、ほぼ同じ語彙で表現されているが、違う事象を表しているような事象を同一視してしまうという問題がある。

よって、本研究では、このような参照を行う事象の単位として、ある特定の時期に起きた特定の事象である**イベント**というものを考え、この参照関係に注目する事により、記事間の参照関係の表現に利用する方法を考える。

2.1　新聞記事からのイベント抽出

本研究で考える**イベント**においては、事象に関する記述だけでなく、その事象に関連する重要な日付がある場合には、その日付情報を記録する。

Root イベントを表す中心となる語 (動作を表す動詞や、主体や対象を表す名詞)

修飾語 イベントの内容を修飾する語 (動詞の場合には、主語や目的語など。名詞の場合は、形容詞や連体節など)

Negative Root の動詞に否定を表す助動詞「ない」などがついていた場合に、否定の Flag をつける。

Depth 構文解析木の Root の要素へかかるまでに、どれだけの要素を経由するかを記載する。

Date イベントを特徴づける日付が文から獲得できる場合にはその情報を記載

ArticleDate 記事が発表された日時

Chunks イベントに対応する語の文中の位置 (文の先頭から 0,1,2 という番号をつける) のリスト

本手法では、以下の手順に基づいて、各記事に含まれる全ての文から対応するイベントに関する情報を自動抽出する。

1. Cabocha[1] により各文の構文解析を行い、係り受けの関係を抽出する。
2. 係り受けの関係がある動詞と名詞を、イベントの Root になる候補として抽出する。
3. イベントの Root にかかる語について、語の品詞やの後接している助詞の情報に注目して、修飾語をタイプ毎に分類する。また、修飾語には、直接かかっている語だけではなく、そのかかりに直接関係する全ての語を含める。

4. 文章中に日付の情報が入っている場合には、イベントを特徴づける日付に設定する。

図 1 に、元になった文章と、イベントの例を提示する。

元の文章：
市は「通常は５日前までに通告がある」と話し、県と基地周辺７市で１０日に抗議する。(毎日新聞 1998 年 1 月 10 日の記事より)
抽出したイベント：

Root ある
　　Depth 2
　　Subject 通常, 通告
　　Date
　　ArticleDate 980110
　　Chunks 4,3,2,1
　　助詞-に ５, 日, 前
Root 話す
　　Depth 1
　　Subject
　　Date
　　ArticleDate 980110
　　Chunks 5,4,2,1
　　助詞-と 通常, ５, 日, 前, ある
Root 抗議
　　Depth 0
　　Subject 市
　　Date １０日
　　ArticleDate 980110
　　Chunks 9,8,7,6,10,0
　　助詞-で 県, 基地, 周辺, ７, 市
　　助詞-に １０, 日

図 1: 文章からのイベント抽出の例

2.2 イベント間の参照関係の取り扱い

我々は、既に、異なる文中に存在する共通の語を文間の関係を表すリンクとして捉えて、PageRank [2] のアルゴリズムを適用することにより、重要文抽出を行うアルゴリズムを提案している [3]。

PageRank のアルゴリズムは、Web ページの有効性をリンク構造に基づいて求めるための手法である。具体的には、ユーザが文書中のリンクをランダムにクリックした (ランダムウォークをした) と仮定し、収束状態での各々のページへの滞留確率により、その評価値を求める手法である。

PageRank に基づく重要文抽出のアルゴリズムでは、要約対象となる文書中の各々の文を Web ページに対応付け、異なる文の間に共通の語が含まれる場合に、文間にリンクが存在すると考え、PageRank のアルゴリズムを適用した。ただし、Web ページのリンクと異なり、異なる文間の関係は、共通する語数の大小や、その語の文中での役割などで、重要度が大きく異なると考えられる。よって、このアルゴリズムでは、共通する語数の大小や、その語の文中での役割などに応じて、次の文への伝搬確率を調整した上で、PageRank のアルゴリズムを適用している。

本研究では、このアルゴリズムの拡張により、イベント間の参照関係を取り扱う。既報のアルゴリズムと異なり、本研究では、複数文書を取り扱うことになっている。この複数文書に拡張する方法としては、次の 2 通りが考えられる。

- 各文書内での文の重要度の計算と文書の重要度の計算を階層的に行う
- 複数の文書をひとまとめとした文書を仮定し、重要度の計算を行う。

ここで、新聞記事の特性に基づき、両者の方法について考える。例えば、事件の続報などの場合に、以前に起きた重要なイベント (事件の発生) について、各文書内で 1 度参照されるが、複数回は述べられないようなイベントが各々の文書内に存在することが考えられる。このような文は、以前のイベントを受けるという意味で重要な文と考えられる。しかし、前者の方法では、個別の記事の中では 1 回しか参照されないため、このイベントの存在が記事内での重要度の向上に役立たない。これに対し、後者の方法では、このイベントの存在がリンクの増加につながり、文の重要度の向上に役立つ。

この考察を踏まえ、本研究では、後者の方法により、文書の重要度の計算を行うこととした。

次に、リンクについてであるが、既報のアルゴリズムでは、共通する語の存在をリンクの存在に対応づけた。これに対し、本研究では、イベントを単位としているので、対応するイベントの存在をリンクとして扱うこととした。ただし、イベントに関しては、同じイベントに対して様々な表現が可能であるだけでなく、部分的な省略などが行われ、語のように完全に一致するものの間だけにリンクを仮定すると、ほとんどのイベント間の関係を正確に把握することができない。

よって、次の基準によりイベントが共通のものかどうかを判断する。

1. イベントを構成する語の類似度
   イベントを構成する語について、Root、修飾語の各々について、共通の語が存在する割合を、各語について IDF(Inverted Document Frequency) に基づく重み付けを行った上で計算する。さらに、Root、修飾語の類似度については、Root に重点をおき、重み付けをまとめ、その値が閾値を越える場合は類似イベントの候補とする。
2. 日付によるイベントの共通性の判定
   イベントに日付が設定されている場合に

は、日付の整合性を検証する。この時、年月の情報が不明の場合は、記事の発表年月日の情報に基づき補完を行う。日付の情報がない場合には、新聞記事が掲載された日付に基づき整合性を検証する。これらの結果、不整合が見つかった場合には、類似イベントの候補からイベントを削除する。

この共通なイベントがある時にリンクを生成する。このリンクに基づく伝搬確率を計算する際には、従来の要約システムにおけるリンクと同様に、イベントの重要性を考慮する。ただし、イベントが参照される回数に基づくイベントの重要性はリンクの数という形で表現されるため、文内におけるイベントの重要性を考えることにする。構文解析木は修飾・被修飾などの依存関係を階層的に表現している。ここでは、文の、主なトピックが後側 (例えば、修飾と被修飾の場合の被修飾) に記述されていると考え、構文解析の木が Root から調べて浅いところにあるもののほうが重要なイベントと考え、リンクの重みを重くする。

次に、リンクの方向性の問題については、双方向に同じ重みのリンクがあることとしてモデル化した。このように作成した重みつきリンクの情報に基づき、PageRank のアルゴリズムを適用する。

2.3 イベントの同一性に基づく抽出文書からの文書圧縮

上記のアルゴリズムで重要文の抽出を行った場合、ほぼ同じ内容を記述している文が異なる記事の中に存在した場合に、ほぼ同じ重要度が計算され、一定数の文を抽出する場合に、冗長な文を選択してしまう問題がある。よって、複数文書からの要約作成の場合には、このような冗長な文を検出し、冗長な文の削除、あるいは、冗長な記述の削除を行うことが求められる [4]。

このような、冗長な文を抽出するための方法としては、文書中の語に含まれる共通性を利用する方法などがあるが、本手法では、イベントの同一性に関する情報を利用し、冗長な文や記述を抽出する方法をとる。

文抽出の場合には、文単位で情報を操作することが求められている。ただ、単純に文毎の一対比較で類似文を探すのではなく、複数の文に記載されている情報が一つの文にまとめられている場合も考慮して、以下の手順により、冗長な文の排除を行う。

1. イベント集合の作成
   選択された文中に含まれているイベント集合を要約文中のイベント集合とする。
2. 文の追加
   文を追加する際には、文毎の重要度に基づいて文の追加を行う。この時、追加を検討する文に含まれるイベントと、先の手続きの要約文中のイベントの包含関係を判定する。一定の新しいイベントが含まれていない文は冗長な文と判断し、追加を行わない。文を追加する場合は、手順 1 に戻り、イベント集合を更新し、所定の文数 (ないしは、文字数) になるまで、文の追加を行う。

また、次の基準により、文を並べ替える。

- 同じ記事内の記事の前後関係は保持する。
- 記事の掲載日の前後関係を保持する。
- 同じ掲載日の異なる記事の文については、その文より以前の文に類似していて追加しようとしている文が存在する場合には、類似文を先に要約に含める。
- 上記の関係がたすき掛けになったり、順序を判断する基準が存在しない場合は、文の登録順序を保持して要約文に含める。

これに対し、自由要約文生成の場合には、冗長なイベントのみを削除し、新たなイベントに関する記述を要約文に含めるという操作が可能になる。冗長な情報は、その情報がはじめて出てくるところではなく、2 回目以降で出てくるところで削除したいので、以下の手順で行う。

1. 文の並べ替え
   追加したい文があった場合には、先の基準により文の並べ替えを行い、要約の最初の文から順番に文を追加する。
2. イベント集合の作成
   選択された文中に含まれているイベント集合を要約文中のイベント集合とする。
3. 追加する文の作成
   文を追加する際には、文毎の重要度に基づいて文の追加を行う。この時、追加を検討する文に含まれるイベントの内、先の手続きの要約文中のイベントに含まれているイベントに対応する文中の要素を、次の基準に基づき削除する。
   (a) 残すことが決まっているイベントにかかっている文中の要素については、少なくとも、一つ以上の内容語 (名詞など) を残す。
   (b) 削除した要素に依存関係を持つ語の要素を削除する。
4. 文の追加
   要素を削除した結果、文を構成する要素が残る場合は、文として追加する。

この結果として、残った要素により、文章を構成することにより、冗長な表現が削除され、要約文全体のサイズを圧縮する。この手続きを、規定の文字数に達するまで繰り返し、要約文を作成する。

2.4 文の位置情報と語の情報の利用

一般に、新聞記事では、記事の先頭に重要な事が記載されることが多い。よって、本研究では、文の位置情報に基づく文の重要性を文の初期重みとして与える方法を組み合わせることとした。PageRank において、各 Page に対する初期重みを与える方法として Topic-Sensitive PageRank[5] が提案されている。この手法では、ページ間の滞留確率を表すベクトル $\vec{r}$ の計算をページ間の遷移確率行列 M、初期重みを表すベクトル $\vec{v}$(具体例については、後の要約実験の中で述べる) とパラメータ α によって、次のように表す。

$$\overrightarrow{r_{i+1}} = (1-\alpha) * M\vec{r_i} + \alpha * \vec{v}$$

先にのべた方法では、イベントの参照関係にのみ注目して、リンクを作成したが、実際には異表記、構文解析による誤り、同一ではないが類似するイベントの間の関係などを考慮すると、イベントの参照関係だけではなく、従来の語によるリンク付けと併用するほうが、より適切な場合も考えられる。よって、リンクの重みの計算においては、語のみを利用する場合、主にイベントのみを利用して語の情報を少し利用する場合、イベントのみを利用する場合の 3 種類について実験を行った。

3 要約実験と考察

本研究で作成した要約システムを現在開催中の NTCIR-4 の一環として行われている TSC-3 の要約課題に適用した。TSC-3 では、一連の事象に関するタイトル、想定される質問、毎日新聞と読売新聞の複数記事から適切な要約を作ることが課題として設定されている。

今回のシステムでは、タイトルの情報については、タイトルのみによって構成される 1 記事を追加し、その文は要約には利用しないという形で反映を試みた。また、想定される質問に関する情報は利用しないこととした。

現在、TSC-3 では、自由要約文生成と文抽出の 2 つの課題があるが、現在のところ、文抽出の課題についてのみ評価データが公開されている。この評価データに基づく評価結果は以下の通りである。複数文書からの文抽出では、冗長な文が存在するために、従来の recall に変わって、coverage という評価尺度が利用されている。この coverage とは、冗長な文を排除して、どれだけ、網羅的に必要な文を集めたかを表す指標であり、precision は、選ばれた文が冗長な文も含めた正解の中に度のくらい含まれるかを表した指標である。表 1 に $\alpha = 0.1$ (各文の初期重みは、$1/log(n+1)$ ただし、n は記事中での文番号) として、各々の設定での全体での coverage と precision(30 課題の long,short の平均) を示す。

表 1: 要約実験結果

	イベント	語とイベント	語
coverage	0.309	0.325	0.328
precision	0.523	0.570	0.557

この結果、イベント単独に比べ、語の情報を用いることが coverage の高い文選択に寄与していることが確認できたが、イベントの情報をリンクに使う有効性を積極的に示すデータではない。この原因としては、現在のイベントの同一性の判定の正確さの問題などが考えられ、より一層の手法の洗練化が望まれる。また、本システムの成績は、参加チームの成績の中で中の上くらいであり、リンクの方向性、様々なパラメータチューニングなどについて、更なる検討が必要である。

4 結言

本研究では、新聞記事におけるイベントの参照関係に注目して、重要文抽出、要約の作成を行うシステムを作成した。要約実験の結果、語の情報だけではなく、イベントの情報を用いることが有効である事は確認できたが、さらなる手法の洗練化が求められると考えている。

謝辞

NTCIR の TSC の評価結果を利用させて頂きました。TSC の評価情報などを作成して頂いている TSC の実行委員の方に謝意を表します。また、毎日新聞・読売新聞 1997 年版、1998 年版データを使用させて頂きました。

参考文献

[1] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. Vol. 43, No. 6, pp. 1834–1842, 2002.

[2] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, Vol. 30, No. 1–7, pp. 107–117, 1998.

[3] 四ッ谷雅輝, 溝江彰人, 吉岡真治, 原口誠. 共起語を介した文間の相互依存関係に基づく重要文の多段階抽出法. 言語処理学会第 9 回年次大会発表論文集, pp. 145–148, 2003.

[4] Inderjeet Mani. *Automatic Summarization*. John Benjamin Publishing Company, 2001. (自動要約 (奥村学・難波英嗣・植田禎子) 共立出版 2003).

[5] T. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the Eleventh International World Wide Web Conference*, 2002.