文章セグメントの単一化による多文書自動要約

浅野 秀胤† 田村 直良††

†横浜国立大学 大学院 工学研究科 電子情報工学専攻
††横浜国立大学 大学院 環境情報研究院
{nocchi,tam}@tamlab.eis.ynu.ac.jp

1 はじめに

本研究では複数の文章を一文章にまとめることを目標とし、文章の事件展開構造化、内容が重複する文の単一化、パス選択による要約文生成からなる要約モデルを提唱し、論じる。

電子化された文書が爆発的に増加している中で、必要な情報を素早く、効率的に入手するために自然言語処理技術に対する要求が高まっている。特に新聞記事やニュース記事のような出来事について書かれた文章に注目してみると、一つの出来事に関する記述が複数の文章にまたがることが多い。また、そのような場合には単に出来事の続きが記述されているだけでなく、視点を変えた文章なども存在する。さらに、複数の文書にわたった新聞記事はまとめて読まれることは考慮されていないので、内容が重複する部分が存在する。この結果、出来事全体の把握が難しくなっている。以上のような特徴を持つ出来事に関する文章に関しては、冗長性の低減を考慮した複数文書要約の実現により、効率的な情報の取得が可能になり、有効であると考えられる。

このような新聞記事について、小倉 [1] は Allen の時空間モデルを用い、文章に含まれる文すべての時間関係を求めることで文章中の文を時系列順に整列する手法を提案し、この手法の有効性を示している。

そこで本研究でも、文章をまとめる観点として時間を用いて複数文書自動要約を実現する手法を提案する。文章中に出現する時間表現を手がかりに、事件展開構造を生成する。ここから、様々な観点でパスを選択する事によって、多様な要約文を生成することができる。

2 文章の事件展開構造化

2.1 事件展開構造の定義

文章中に出現する時間情報に注目してみると、明らかに時間的な前後関係がわかるものと、わからないものがある。前後関係がわからない場合は、文中に出現する時間情報が異なっているが、比較できる部分まで同一である場合である。詳細さが異なる時間情報は、どちらが先に起きた出来事かについては曖昧であるため、単純に比較することは意味がない。また、時間情報と文の記述している出来事について調べてみると、この二つには関連性があることがわかる。一般に、時間情報が詳細なほど文の記述も詳細であり、逆に大まかな時間情報しかない文では出来事を大まかに述べている場合がほとんどである。このことから、時間情報の詳細さが異なる文は並列に扱う事が妥当であると考えられる。

以上より、文章の時間的な関係は図 1の様に、部分的に並列になっていると考えられる。そこで、これを事件展開構造として定義する。事件展開構造とは、時間情報を元に事件記事を構造化したものである。

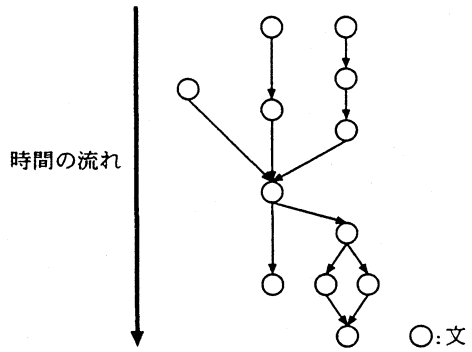

図 1: 文章と時間の関係

定義 1 事件展開構造とは、文をノードとし、文が表す事象のの時間的な前後関係をリンクとして文章を構造化したものである。ノードの時間情報に前後関係があるものは時系列順に整列し、時間情報が同一、もしくは前後関係が不確かなものは並列に並べることにより構造を生成する。

2.2 時間セグメント

時間表現やテンス・アスペクトなどといった時間情報に注目して文章を調べると、文章内では文の時

間情報に関連性がある文は連続して出現していることがわかる。これを時間セグメントと定義する。

定義 2 時間セグメントとは、文章中の時間的な境界から次の時間的な境界の間に出現している文の集合である。そのため、同じセグメントに属する文は何らかの時間的な関連性を持って起きた事象について記述されている。

ここで、時間情報は何らかの時点を明示的に示している。そのため、時間情報がある文の直前で時間の連続性が途切れている。よって、時間セグメントに時間情報は多くとも一つしか存在しない。

時間セグメントの種類は大きく分けて以下の2種類ある。

- 現在のセグメント

  時間情報を持たず、時制によって現在の出来事や状態について書かれた文と判別できる文が集まったセグメントである。このセグメントでは文章が書かれた時点での状況や、行われていることが記述されている。

- 過去のセグメント

  過去の出来事について書かれた文が集まっているセグメントである。このセグメントには過去の出来事や、そのときの状態を表す文が含まれている。そのため、このセグメントには時間情報がある文、もしくはアスペクトによって過去と判別できる文が属している。このため、過去のセグメントには時間情報があるものとないものが存在する。

文章中には時間セグメントの配置順がそこに書かれている時間情報と無関係に出現している。そのため、時間セグメントを時系列順に整列することにより、文の時間順整列が実現できると考えられる。

2.3 時間セグメント生成

ある文の前が時間セグメントの区切りとなるのは以下の場合である。

1. 文中に時間表現が出現した場合
2. 過去の文の次に現在の文が出現した場合
3. 現在の文の次に過去の文が出現した場合
4. 段落が変わった場合

1の場合、時間表現によって文の関連性が途切れるため、新しいセグメントとする。2,3については、過去の出来事についての記述か、文章が書かれた時の状態や出来事であるかが変化している部分である。そのため、これも新しいセグメントとする。4の段落が変わった場合であるが、このような時筆者が意識して文の区切りとしている。そのため、時間的な関連性も途切れている。よって、ここでもセグメントを改める。

また、以下の場合は例外として新たなセグメントを生成しない。

- 文頭に接続詞がある場合
- 段落の最初の文に時間情報がなく、次の文に時間情報がある場合

次に、時間セグメントが持つ時間情報を定める。

1. 記事が書かれた時間

   現在のセグメントには、記事が書かれたときどのような状況であるかが記述されているので、このような場合には記事が書かれた時間を時間セグメントの時間情報とする。

2. 時間表現などを元にした時間情報

   時間情報を持った文が出現する場合、文の時間情報をセグメントの時間情報とする。

3. 知識を元にした時間情報

   個々の事件にもよるが、事件の大まかな流れはあまり変化せず、一定である場合が多い。そのため、一般的な事件の流れと変わらない部分については、時間情報が省略される場合がある。そこで、一般的な出来事の成り立ちを知識として定義する。図2は一般的な事件の流れを表している知識の例である。これらについて表1のような手がかり語をそれぞれ用意し、用いることで時間情報を補完する。

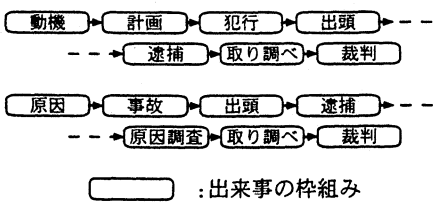

図 2: 一般的な事件の流れの例

このような知識を用いてもまだ時間情報がないセグメントが存在する。このようなセグメントは、図3のような場合が考えられる。

以下、それぞれの場合について述べる。

(1) については、過去のセグメントａとｃの間に挿入的に現在のセグメントが表出している場合である。この場合、セグメントａとｃは本来一つのセ

計画	標的にした
犯行	誘拐した 殺害した
出頭	出頭した 自首した
逮捕	逮捕
取り調べ	押収した

表 1: 出来事と手がかり語の例

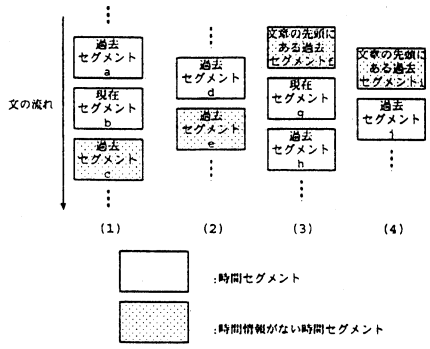

図 3: 時間情報がないセグメントが出現する場合

グメントであったと考えられる。しかし、セグメントaの最後に記述されている事象について、現在の状況などを補足している時などにこのような状態になる。

(2) ついては、段落が切り替わった場合などによりセグメントが分割されたが、本来は同じセグメントであった場合である。この場合、文章構成上新たな段落となっているが、時間的な流れは直前のセグメントを継承している。

(3) と (4) は (1) と (2) がそれぞれ文章の先頭であった場合である。

よって、以上のように時間情報がないセグメントは、本来直前の過去セグメントと同一であったものが、文章の構成上分割されたために発生していると考えられる。よって、このようなセグメントは以下のルールで統合する。

- 文章中の過去のセグメントに時間情報がなかった場合、そのセグメントの前に出現した過去のセグメントに統合する。
- 文章の最初に出現した過去のセグメントに時間情報がなかった場合、その次に出現する過去のセグメントに統合する。

2.4 事件展開構造の生成

事件展開構造は、時間セグメントA、Bを以下のルールで整列することにより生成できる。

- Aが持つ時間情報よりBが持つ時間情報が明らかに後であるならば、A→Bの順にセグメントを整列する。
- AとBの時間情報が全く同じであるか、もしくは細かさが異なる時間情報のうち比較できる部分までは同一であるならば、並列に並べる。
- Aについての前後関係を知識を用いて調べた場合、次に起きる出来事のうちもっとも最初に出現するセグメントBの直前にセグメントをおく。
- Aが持つ時間情報が時点であり、Bが持つ時間情報時区間の時に、両方の時間情報が完全に一致する場合がある。時区間は時間情報の時点から、もしくは時間情報の時点までの出来事である。そこで、時間情報の時間からの場合は時点の時間情報より後の出来事とし、時間情報の時間までの場合は時点の時間情報より前の出来事であるとする。

3 内容が同一な文の単一化

関連文書を一つにまとめた時、内容が重複する文がある。このような文を単一化することにより、冗長性の低減する。文の内容は格フレームを用いる。

格フレームで表現された文について、すべてのスロットが同一か、言い換えられた形であれば単一化していく。

また、事件展開構造における単一化された文の扱いを決定する必要がある。ここで、文が属している時間セグメントが、時間的に関連がある文の集合であることを考えると、文がどちらか片方にのみ出現するようにすると文の流れが途切れてしまう。よって、双方の時間セグメントに文が存在していることが妥当である。そこで、単一化された文にはリンクを生成し、同一の文であることを明示しておく。

4 パス選択による要約文生成

前節までで生成した事件展開構造の例を図4に示す。この構造から、リンクをたどりながらノードを選択していくことで、時間に沿った文章を生成することができる。本節では、このようなパス選択による要約文の生成について述べる。

要約文を生成する時、どのような条件でパスを選ぶかを決定する必要がある。ここで用いる条件により、生成される要約文が変化する。これは、条件の選択と同時に要約文の観点を選択しているためである。例えば、最短パスを選択すると、もっとも要約率が高く、出来事を大まかに述べた要約が生成され

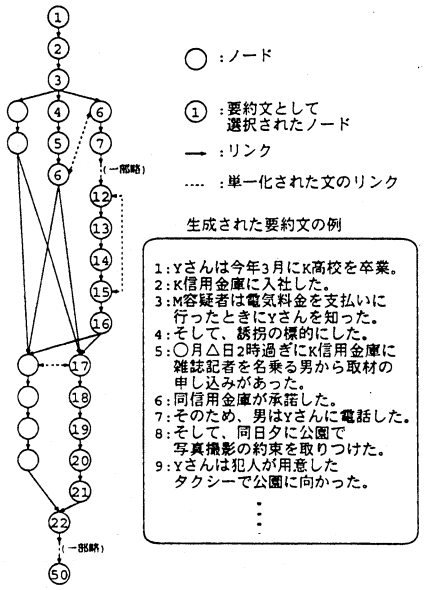

図 4: 事件展開構造の例

る。また、最長パスを選択すると要約率は低いが、出来事を細かく述べた要約を生成することができる。そこで、これらの要約を生成し、評価を行う。

	事件 1	事件 2
全文数	59	24
生成された要約文の数	14	1
最短要約文数 (要約率)	37(62.7%)	24(100%)
最長要約文数 (要約率)	50(84.7%)	24(100%)

表 2: 本手法での要約率

本研究では実験データとして、1993 年毎日新聞のコーパスから 2 文書にわたる 2 つの事件について人手で事件展開構造を生成し、パス選択を行うことによって要約文を生成した。表 2はそれぞれについて最短パスを求めた要約と最長パスを求めた要約の要約率を示したものである。

ここで、事件 2 については時間的に並列な部分がなく、文の単一化も起らなかったためにすべての文が要約として用いられたため、要約率が 100% になっている。

これらの 3 つの要約文それぞれについてについて以下のようなアンケートを行った。この結果を表 3に示す。

1. 文の意味的に文章の繋がりが誤っている部分があるか
2. 事件 1 の最短要約と最長要約のどちらが詳細な要約であると感じたか
3. 出来事全体を時間を追って把握できたと感じたか

		事件 1		事件 2
		最短要約	最長要約	
被験者 1	繋がりの誤り数	0	1	0
	出来事を把握できたか	○	○	○
被験者 2	繋がりの誤り数	0	1	0
	出来事を把握できたか	○	○	○
被験者 3	繋がりの誤り数	0	1	1
	出来事を把握できたか	○	○	○
被験者 4	繋がりの誤り数	0	1	0
	出来事を把握できたか	○	○	○

表 3: 要約文についての調査

それぞれの要約文の平均誤り率は、事件 1 の最短要約で 0%、最長要約で 2.1%、事件 2 の要約で 1.4% であった。被験者が誤りとした文は被験者により異なり、あまり一致が見られなかった。これは、要約文の表層的な整形を行わなかったため、文の意味の取り方が被験者によって異なっているためだと考えられる。

また、すべての被験者が最短要約より最長要約の方が詳細な要約であると回答した。このことから、要約文の観点とパス選択の条件付けに関わりがあると考えられる。

さらに、すべての被験者が出来事を把握できたと感じている。よって、この手法は適切な要約文の生成に有効な手段であると考えられる。

5 まとめと今後の展望

複数文書を時間を観点としてまとめ、事件展開構造の概念を元にパス選択による要約文の生成を行う手法について提案、実装し、評価を行った。

今後の展望としては、要約文生成の際に表層的な整形を行うことで、よりわかりやすい要約を生成することができる。また、本手法で定義した事件展開構造を、時間情報だけでなく主題や文章中の登場人物の視点などで生成することで、事件記事以外の文書についても要約を生成することができると考えられる。

参考文献

[1] 小倉牧人, 田村直良. 文間の時間制約モデルと事象の時系列化への応用に関する研究. 情報処理学会自然言語処理研究会研究報告 NL-140, 2000.