

類似記事集合の主要イベントに注目した 事件の構図を表現するグラフ生成手法

岩橋 篤矢[†] 井上 悦子[‡] 中川 優[‡]

[†] 和歌山大学大学院システム工学研究科 〒640-8510 和歌山県和歌山市栄谷 930

[‡] 和歌山大学システム工学部 〒640-8510 和歌山県和歌山市栄谷 930

E-mail: [†] s111065@sys.wakayama-u.ac.jp, [‡] {etsuko, nakagawa}@sys.wakayama-u.ac.jp

あらまし 本研究では、類似記事集合から主要なイベント群を抽出し、事件の構図を表現するグラフの生成手法を提案する。本手法では、主要なイベントとして、重要な動詞とその主語又は目的語となる名詞を、その動詞に対する係り受け関係と格助詞に基づいて抽出する。そして、主要なイベント群の名詞をノード、動詞をリンクとしたグラフを生成する。提案手法を評価するために、主要イベント群の抽出精度を確認した結果、主語の正解率は91.0%、目的語の正解率は73.5%となった。また、事件の構図を表現したグラフが新聞記事の理解を支援できるかを評価するために、被験者にグラフと記事をセットで見てもらい、アンケートに回答してもらった。その結果、新聞記事をあまり読まない被験者からは、グラフが新聞記事の理解支援に役立つとの意見が得られ、提案手法の有用性が確認できた。

キーワード 情報抽出、係り受け解析、有向グラフ、新聞記事

1. はじめに

近年、活字離れが進み、新聞記事を読まない人が増加している。内閣府の調べによると、12歳から29歳までの男女のうち「新聞記事を全く読まない」と回答した割合が、2001年の調査では3,486人中22.3%であったのに対し、2006年の調査では2,468人中47.7%へと変化しているという調査結果が報告されている[1]。

また、2009年に株式会社アイシェアが実施した新聞記事に対する意識調査では、20代から40代の男女510人中51.6%の人が「新聞記事を理解するのは難しい」と感じているという結果が報告されている[2]。これらの調査結果から、日頃あまり新聞記事を読んでいない人が多く、約半数の人が新聞記事を理解するのが困難だと感じていることがわかる。

この問題に対し、新聞記事を理解するのが難しいと感じる原因として、新聞記事の文章から主体間にどのような関係があるのかといった情報を、事件の構図として頭にうまく思い描くことができないことが理由ではないかと考えた。そうであるならば、事件の構図を頭に思い描くことを支援できれば、新聞記事が理解しやすくなると期待できる。

そこで本研究では、新聞記事の理解支援を目的として、記事本文から重要な動詞とその主語及び目的語で構成された主要イベント群を抽出し、事件の構図を表現するグラフを生成する手法を提案する。なお、本手法では、主要イベント群の動詞に対する主語及び目的語を係り受け関係に基づいて抽出するが、1記事からは十分な情報が抽出できないため、同一事件を報じた

類似記事を入力とする。また、事件の経緯を把握しやすくするために、主要イベントの発生順も表示する。

本稿の構成は以下の通りである。2章では関連研究について紹介する。3章では本研究で生成するグラフ及び各要素の抽出方針を述べ、4章ではグラフの生成手法について説明する。5章で評価実験の結果を示し、6章で結びとする。

2. 関連研究

新聞記事を対象とした情報抽出の研究は盛んに行われている。大竹らは、新聞記事から要約を自動生成することを目的とし、新聞記事内の単語の出現頻度等から各文を重み付けし、重要な文を出力することを行っている[3]。要約された新聞記事は元の記事よりも理解しやすくなると考えられるが、要約文に含まれる主体間の関係から事件の構図を思い浮かべる作業は読み手自身が頭の中で行う必要がある。

金山らは、新聞記事から事件において関連する人間の相互関係として意味構造（容疑者・被疑者・それぞれのプロフィール・犯罪の動機等）を定義し、抽出することを行っている[4]。この研究では、対象とする事件の構造を限定することで、内容に即した構造化を実現している。これに対し、本研究は対象とする事件の構造を限定せず、より多くの事件に適用できる。

また、新聞記事以外を対象とした情報抽出の研究も行われている。数原らは、ブログ記事集合内の文を構文解析し、同じ述語を持つものを重ね合わせる事で、動作関係をグラフ形式で抽出することを行っている

[5]. 動作関係をグラフ形式で抽出するという点では本研究と類似しているが、二者間に複数の動作関係があった場合、その順序関係はわからない。多くの新聞記事では主体間に複数の動作関係が存在するため、本研究ではこれらの動作関係の順序関係も重要な情報であると考え、主要イベントの発生順についても抽出する。

3. 本研究で生成するグラフ

3.1. 事件の構図を表現するグラフ

本節では、本研究で新聞記事の理解支援のために生成する、事件の構図を表現するグラフについて説明する。なお、本研究では社会記事を中心に、幅広いニュースをまとめて「事件」として扱うものとする。

本研究で生成するグラフの例を図 1 に示す。図 1 は、「山田太郎は犬を拾ったので、その犬を飼った。その犬は成長した。」という文に対し、そこに含まれるイベントで形成される構図を、グラフで表現したものである。本研究で生成するグラフでは、主要なイベント群の主語又は目的語をノードとし、主要イベント群の動詞をその動詞の主語から目的語に向かう有向リンクとし、そのリンクラベルとして動詞とその発生順を付加して表現する。ただし、主要イベント群の動詞が自動詞の場合は目的語を持たないため、主語からダミーノードに向かうリンクで表現する。図 1 の「③成長する」は自動詞であるため、主語「犬」からダミーノードへのリンクとなっている。

ここで、事件の構図を表現するグラフを、主要イベント群の動詞及びその主語・目的語で構成する理由を説明する。本研究では、事件を 1 つ以上のイベントの集合で構成される社会的な出来事と考え、事件の構図をそれらの主要なイベントを組み合わせてひとまとまりに捉えたものと考え、グラフを用いて表現することとした。また、グラフを単純にするために、主要イベントの要素を、新聞記事の内容を表す上で重要な動詞とその動詞に対する主語及び目的語のみに限定する。

しかしながら、事件の構図を理解するには、イベントの発生順が重要となるが、事件の構図をグラフで表現すると、記事本文で描写されているイベントの発生順がわからなくなり、グラフから情報を読み取ることが困難になるという問題が考えられる。そこで本研究では、図 2 に示すように、主要イベント群の動詞に対してその発生順を付加しておき、それによってグラフを段階的に表示することで、この問題を回避する。

このような事件の構図を表現するグラフを生成するためには、主要イベント群の動詞とその発生順と、各動詞に対する主語及び目的語を抽出する必要がある。主要イベントの各要素を抽出するための方針について、次節以降で説明する。

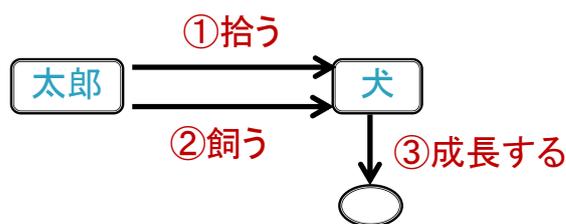


図 1 本研究で生成するグラフ

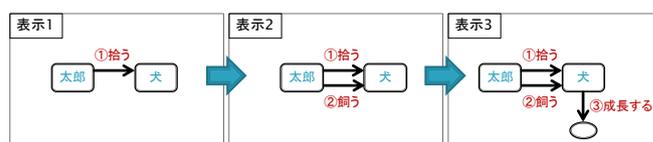


図 2 本研究で生成するグラフの表示方法

3.2. 主要イベント群の動詞の抽出方針

本節では、主要イベント群の動詞とその発生順を抽出するための方針について説明する。主要イベント群とは事件に含まれるイベントの中でも重要なイベントであり、ここで抽出対象となるのはそれらのイベントを表す一要素である動詞である。

多くの新聞記事では、読者に読みやすくするために、1 段落目に事件の概要について説明し、2 段落目以降に事件の経過や詳細について説明するという構成で書かれている。このことから、新聞記事の 1 段落目は、その事件における重要な事柄が順序立てて述べられており、その中に登場する動詞は重要であると考えられる。そこで本研究では、記事の 1 段落目に出現する動詞を主要イベント群の動詞として抽出する。

また、動詞の発生順については、正確に抽出するためには、動詞に対する時間情報に基づいて推定しなければならないが、必ずしも動詞に対する時間情報が明記されていないわけではないため、動詞の発生順を推定するのは困難であると考えられる。そこで本研究では、動詞の出現順を動詞の発生順とする。

3.3. 主要イベント群の主語及び目的語の抽出方針

主要イベント群の各動詞に対する主語及び目的語となる名詞は、同一事件を報じた類似記事集合の本文中から、それぞれの動詞に対する係り受け関係と格助詞に基づいて抽出する。詳細な抽出条件は 4.2 節で述べる。本節では、主要イベント群の各動詞に対する主語及び目的語を、できるだけ漏れなく抽出するための各要素の抽出方針について説明する。

前節で述べた方針で主要イベント群の各動詞を決定しても、それらに対する主語及び目的語を漏れなく抽出するのは容易ではない。なぜならば、日本語では主語や目的語が省略されることは少なくないからである。また、固有名詞等の一部を省略した表記、代名詞、

同義語や類義語など、同一の事柄であっても表記の異なる要素での表現も多用されるが、同一の事柄を異なる要素として抽出してしまうと、図 3 のように同一人物が別々の人物である等の誤った解釈ができてしまうという問題が生じる。

そこで本研究では、同一事件を報じた極めて類似性の高い複数記事から主要イベント群を抽出することでこれらの問題を解決する。類似記事集合の中から 1 記事を基準の記事として 3.2 節に述べた方針で主要イベント群の動詞およびその発生順を抽出した後、その基準の記事を含めた類似記事全体から、主要イベントの各動詞に対する主語及び目的語を抽出する。このように同様の情報を異なる書き手が書いた記事集合を用いることで、文中で主語や目的語が省略されていた場合でも、これらの要素を補うことができると考えた。

一方、同一の事柄が表記の異なる要素で表現される問題は、異なる書き手による記事集合を扱うことでより深刻になる。これに対して我々は、類似記事集合に書かれているのは同一事件に対するほぼ同じ情報であるため、表記が似ていて、かつ同一要素に対して共通の係り受け関係を持つ要素であれば、同一の事柄を言い換えた表現であると考えた。さらに、主要イベントの各動詞に対する主語及び目的語については、係り受け関係と格助詞の条件を満たせば、表記がまったく違っていても、事件に置ける位置づけが等しいことから同一の要素であると考えた。

そこで、主要イベントの動詞に対して係り受けのある名詞に限らず、類似記事集合の本文から名詞及び動詞とこれらの要素間の係り受け関係をすべて抽出し、同一の事柄だと考えられる名詞をあらかじめ統合した後、主要イベントの各動詞に対する主語及び目的語を決定する。なお、このようにすることで、主要イベント群の各動詞に対する主語や目的語である名詞が複数になる場合が生じる。最終的に生成するグラフでは、1 つ以上の要素が共通している名詞群を同一の事柄とみなしてグループ化し、1 つのノードで表現するものとする。

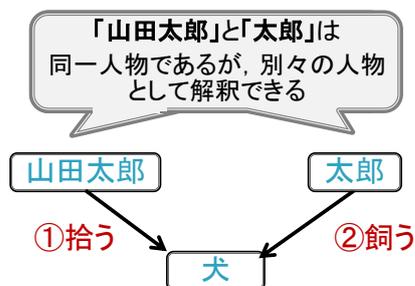


図 3 名詞の表記の違いによる問題

4. 事件の構図を表現するグラフの生成手法

4.1. グラフ生成の流れ

本研究では、類似記事集合から事件の構図を表現するグラフを、3 章で述べた方針に従って生成する手法を提案する。本節では、グラフを生成する流れについて説明する。それぞれの手順についての詳しい説明は次節以降で行う。

図 4 に示すように、提案手法では以下の(1)~(5)の手順でグラフを生成する。

- (1) 名詞・動詞・係り受け関係の抽出
- (2) 係り受け関係の追加
- (3) 名詞の統合
- (4) 主要イベント群の抽出
- (5) 主要イベント群のグラフ化

手順 (1) では、入力として与えられた類似記事集合内の記事の本文から、名詞及び動詞とこれらの要素間の係り受け関係を全て抽出する。動詞を抽出する際、3.2 節で述べたように、類似記事集合内の基準となる記事の 1 段落目に登場する動詞を、主要イベント群の動詞として動詞の発生順と共にこれらの情報を付与しておく。なお、基準となる記事は、類似記事集合の中で 1 段落目に登場する動詞が最も多い記事とする。また、係り元の単語の主語・目的語の判定を、それぞれの動詞に対する係り受け関係と格助詞に基づいて行う。

手順 (2) では、手順 (3) の前処理として、係り受け関係に基づく判定で主語が抽出されなかった動詞に対して、直前の動詞の主語をその動詞の主語として係り受け関係を追加する。

手順 (3) では、同一の事柄が表記の異なる名詞で表現される問題を解決するために、3.3 節で述べた方針に基づいて、表記が似ていて、かつ同一要素に対する共通の係り受け関係を持つ名詞群を統合する。

手順 (4) では、手順 (1) で主要イベントの動詞として情報を付与された動詞に対して、係り受け関係に基づいて主要イベント（動詞・主語・目的語）を選出する。なお、同一の動詞に対する主語や目的語が複数抽出されている場合は、それぞれグループ化してまとめておく。

手順 (5) では、手順 (4) で抽出された主要イベント群を 3.1 節で述べた方針に従って、主語または目的語をノード、動詞を有向リンクで表したグラフを作成する。手順 (4) で主語や目的語をグループ化したものについても 1 つのノードで表すが、グループに含まれる名詞が 1 つでも同一のものが別のノードに含まれている場合は、それらのノードを 1 つに統合する。

以上の手順 (1) から手順 (5) の処理により、事件の構図を表現するグラフが生成できる。

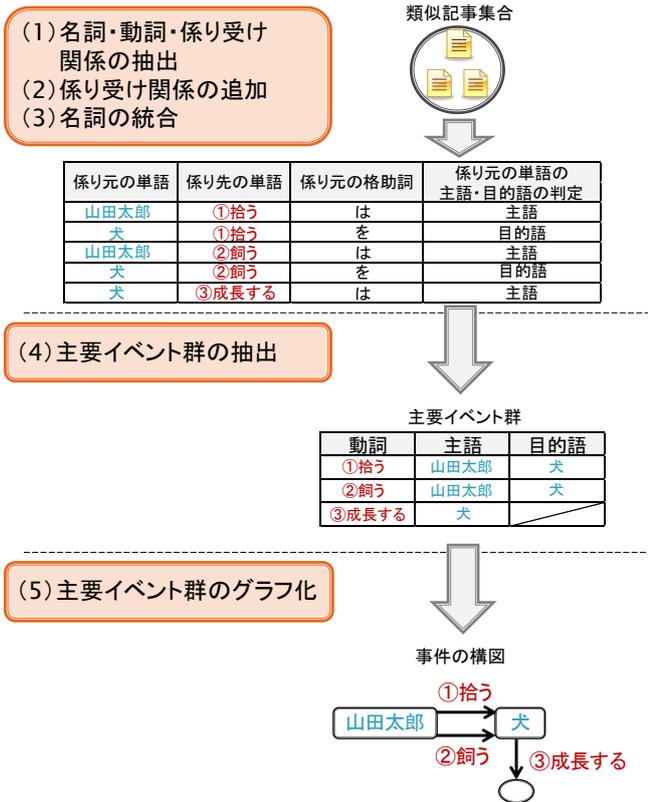


図 4 グラフ生成の流れ

4.2. 名詞・動詞・係り受け関係の抽出

本節では、類似記事集合から名詞及び動詞とこれらの要素間の係り受け関係を全て抽出する流れについて説明する。

名詞及び動詞の抽出に関しては、類似記事集合内の記事の本文に対して、1文ずつ形態素解析器 MeCab[6]を用いて形態素解析を行い、抽出する。この際、3.2節で述べたように、類似記事集合内の基準となる記事の1段落目に登場する動詞は、主要イベント群の動詞として、動詞の発生順と共に抽出する。

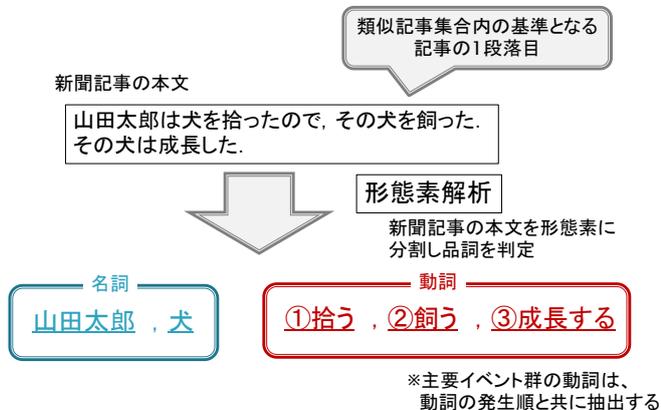


図 5 名詞及び動詞の抽出例

なお、図 5 のように、動詞に関しては全て原形で抽

出し、名詞に関しては複数の単語から意味を成すものは1つの複合名詞として抽出する。さらに、サ変名詞に関しては、サ変名詞の直後に動詞の原形の「する」が続く場合、もしくは、サ変名詞の直後に句点が続く場合、サ変名詞に動詞の原形の「する」を結合し、動詞として抽出する。

係り受け関係の抽出に関しては、類似記事集合内の記事の本文に対して、1文ずつ係り受け解析器 Cabocha[7]を用いて係り受け解析を行い、抽出された単語（名詞もしくは動詞）を含む文節間の係り受け関係をその単語間の係り受け関係として抽出する。

なお、図 6 のように、係り受け関係を抽出する際は、係り元の単語と係り先の単語の他に、係り元の格助詞を抽出しておき、これらの抽出結果を基に係り元の単語の主語・目的語の判定を行う。係り先の単語が動詞であり、かつ係り元の格助詞が「は」もしくは「が」である場合、係り元の単語を主語とし、係り先の単語が動詞であり、係り元の格助詞が「を」である場合、係り元の単語を目的語とする。ただし、係り先の動詞の直後に助動詞の「れる」もしくは「られる」が続く場合、その動詞を受動態と判断し、係り元の格助詞の「に」は「は」に、「は」は「を」に変更し、判定する。

しかしながら、図 6 の係り受け解析の結果を見てみると、動詞の「飼う」は主語となる名詞との係り受け関係を持っていないため、動詞の「飼う」の主語は存在しないという結果となっている。次節ではこの問題を解決するための係り受け関係を追加する処理について説明する。



図 6 係り受け関係の抽出例

4.3. 係り受け関係の追加

本項では、主語を持たない動詞に対しては、直前の動詞の主語との係り受け関係を追加する処理について説明する。また、4.4 節で言及する名詞の統合では、係り受け関係の情報を用いるため、名詞の統合よりも

先に係り受け関係の追加を行う。図 7 に係り受け関係を追加する例を示す。

図 7 は、「山田太郎は犬を拾ったので、その犬を飼った。」という文に対して、係り受け解析を行った結果である。図 7 のように、「飼った。」は主語を持っていないため、「拾ったので、」の主語である「山田太郎は」との係り受け関係を追加する。

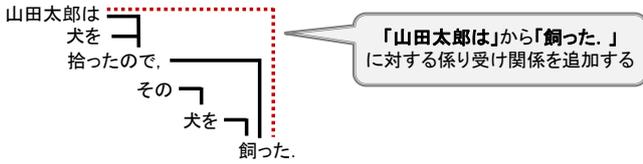


図 7 係り受け関係の追加する例

4.4. 名詞の統合

本項では、3.3 節に述べた同一の事柄が表記の異なる名詞で表現される問題を解決するために、表記が似ていて、かつ同一要素に対して共通の係り受け関係を持つ名詞同士を統合する処理について説明する。

類似記事集合内で、表記が似ていて、かつ同一要素に対して共通の係り受け関係を持つ名詞同士は、同一の事柄を言い換えた表現であると考え、文字数の多い方の名詞に統合し、文字数の少ない方の名詞の係り受け関係の情報を統合後の名詞に追加する。

また、名詞の統合を行う場合、表記の異なる名詞で表現される場合は以下の 3 種類考えられる。

- (i) 単名詞と複合名詞
- (ii) 複合名詞と複合名詞
- (iii) 単名詞と単名詞

この 3 つの内、表記に部分一致が見られるのは図 8 のように (i) 単名詞と複合名詞と (ii) 複合名詞と複合名詞であるため、本手法では、この 2 つの場合のみ、名詞の統合を行う。

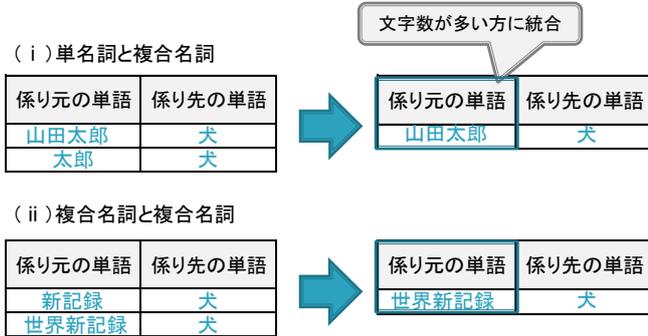


図 8 名詞の統合を行う例

4.5. 主要イベント群の抽出

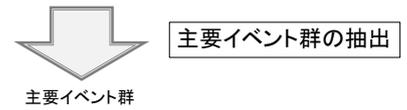
本項では、4.2 節で述べた手順 (1) で抽出された名詞及び動詞の中から、主要イベント群 (動詞・主語・目的語) を抽出する処理について説明する。

主要イベント群の動詞に関しては、4.2 節述べた手順 (1) ですでに抽出しているため、ここでは主要イベント群の主語及び目的語のみ抽出を行う。図 9 に主要イベント群を抽出する流れについて示す。

図 9 のように、主要イベント群の主語及び目的語を抽出する際は、主要イベント群の動詞と係り受け関係がある主語と判定されている名詞を主要イベント群の主語とし、目的語と判定されている名詞を主要イベント群の目的語として抽出する。

また、図 10 のように同じ主要イベント群の動詞に対して、主語や目的語が複数存在する場合、表記がまったく違っていても、事件に置ける位置づけが等しいことから同一の要素であると考え、それぞれグループ化する。

係り元の単語	係り先の単語	係り元の格助詞	係り元の単語の主語・目的語の判定
山田太郎	①拾う	は	主語
犬	①拾う	を	目的語
山田太郎	②飼う	は	主語
犬	②飼う	を	目的語
犬	③成長する	は	主語



動詞	主語	目的語
①拾う	山田太郎	犬
②飼う	山田太郎	犬
③成長する	犬	

図 9 主要イベント群を抽出する流れ

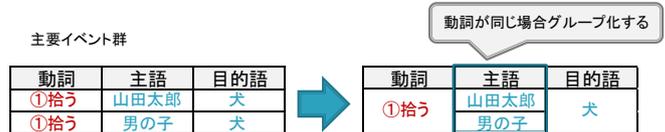


図 10 主語及び目的語をグループ化する例

4.6. 主要イベント群のグラフ化

本項では、4.5 節で述べた手順 (4) で抽出された主要イベント群をグラフ化する処理について説明する。

3.1 節で述べたように、主要イベント群の主語又は目的語をノードとし、主要イベント群の動詞をその動詞の発生順を付加した形で主語から目的語に向かうリンクとした有向グラフで表す。ただし、主要イベント群の動詞が自動詞の場合、目的語を持たないため、主語からダミーノードに向かうリンクで表現する。

また、主要イベント群の主語グループもしくは目的語グループの内、図 11 のように共通する名詞が 1 つでもあればそれらを統合し、主要イベント群の関連付けを行う。統合する際は、グループ内の名詞の数が多いほうに統合し、グループ内の名詞の数が少ないほうの係り受け関係の情報を統合後のグループに追加する。

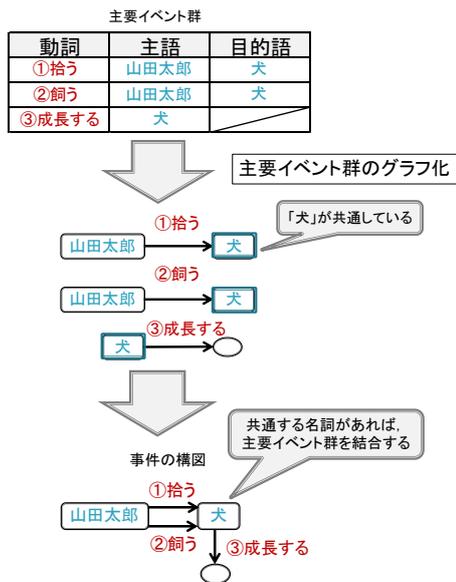


図 11 主要イベントの動詞に基づいて主語・目的語グループを統合する例

5. 評価

5.1. 評価実験の概要

提案手法を評価するために、主要イベント群を正しく抽出できているのかについて、提案手法による事件の構図を表現したグラフが新聞記事の理解を支援できるのかについての2つの評価実験を行う。

実験1では、本手法が主要イベント群を正しく抽出できているのかについて確認するために、主要イベント群の主語及び目的語の抽出精度と、主要イベント群の動詞の抽出精度について確認する。なお、主要イベント群の主語及び目的語の抽出精度については、類似記事集合を用いることと、係り受け関係の追加処理を行うことで、抽出精度にどのような効果があるかについても確認する。

実験2では、事件の構図を表現したグラフが新聞記事の理解を支援できるのかについて確認するために、事件の構図を表現するグラフと新聞記事をセットで閲覧することでどのくらい新聞記事の理解支援に役立つかを被験者による主観評価で行う。なお、提案手法で生成したグラフによる理解支援は、新聞記事をあまり読まない人に対してより効果があると考え、被験者の新聞記事を読む頻度の違いによる結果の差についても確認する。

5.2. 実験1

実験1では、本手法が主要イベント群を正しく抽出できているのかについて確認するために、主要イベント群の主語及び目的語の抽出精度と、主要イベント群の動詞の抽出精度について確認する。

主要イベント群の主語及び目的語の抽出精度については、表1に示す4種類の生成手法で主要イベント群の主語及び目的語を抽出して正解と比較し、手法ごとに主要イベント群の主語及び目的語の抽出精度（正解率）を算出する。

主要イベント群の動詞の抽出精度については、4.2節で述べた手順(1)で抽出した主要イベント群の動詞及び各動詞の発生順を正解と比較し、適合率および再現率を算出する。

なお、今回の実験では、5つの事件に対して行う。各事件の類似記事集合内の記事は5件ずつである。

表1 実験1で用いる生成手法の種類

生成手法	使用するデータ	係り受け関係の追加処理
A	1記事	なし
B	1記事	あり
C	類似記事集合	なし
D	類似記事集合	あり

それぞれの正解データの生成方法について簡単に説明する。筆者が所属する研究室の学生5名に各事件の基準となる1記事を読んでもらい、基準の記事から抽出した主要イベント群の各動詞に対して、動詞の発生順と、各動詞に対する主語及び目的語を記入してもらい、主語及び目的語が存在しないと判断した場合、「なし」と記入することとした。回答を集計する際、同じ事柄を指すと考えられる回答については、被験者に確認した上で同じ事柄として統合する。要素ごとに過半数の被験者が選んだ回答を正解とする。次に、主要イベントとして抽出すべき正解の動詞群を決定するために、被験者自身が記事内で重要と考える動詞を選出してもらい、過半数を超えた回答を正解の動詞群とする。

5.3. 実験2

実験2では、事件の構図を表現したグラフが新聞記事の理解を支援できるのかについて確認するために、被験者にグラフと新聞記事をセットで見てもらい、グラフが理解を支援できるのかについて評価してもらい、被験者は、実験1に参加していない筆者が所属する研究室の学生12名である。実験手順は以下の通りである。

まず、主要イベント群の抽出精度が主観評価に影響しないように、実験1で作成した正解となる主要イベント群を参考に、一部修正したグラフを被験者に提示し、事件の構図を把握してもらうこととした。修正したのは、事件ごとに主語や目的語を3箇所程度である。

次に、図12に示すような、各事件の基準となる1記事と先に提示したグラフをセットで読んでもらい、自分自身が新聞記事を理解するために、グラフがどの

くらい役立ったかを5段階で評価してもらう。

最後に、グラフを使って新聞記事を理解する際の問題点などを自由記述で回答してもらう。

トラに襲われ飼育員死亡＝おりで見合いさせようと・京都市動物園
 7日午前9時50分ごろ、京都市左京区岡崎法勝寺町の京都市動物園のおりの中で、男性飼育員がアムールトラに襲われ血を流して倒れているのを女性客が見つけ、動物園事務所を通じて119番した。飼育員は病院に運ばれたが、右の首やお、後頭部などに引っかき傷があり、間もなく死亡した。
 府警川端署の調べによると、飼育員は伊藤淳さん(40)で、襲ったのは11歳の雄「ビクトル」。同動物園はビクトルと3歳の雌のアムールトラ「アオイ」を飼育。おりは3つあり、伊藤さんはビクトルとアオイの見合いをさせるため、トラのいないおりにアオイをおびき寄せようとしていた。この際、ビクトルのおりの鍵が開いており、出てきたビクトルに襲われたという。同動物園は、事故を受けて営業を中止し臨時休園とした。
 産経新聞 2008/06/07掲載

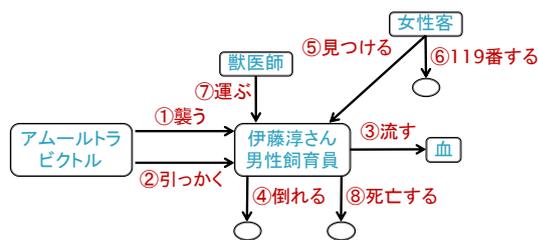


図 12 被験者に提示した記事とグラフの例

5.4. 実験結果

まず、実験1の結果を説明する。表2は各生成手法で抽出した主要イベント群の主語及び目的語の正解率の結果をまとめた表である。表2の括弧内の値は、誤りを含む正解も正解とした場合の正解率を表す。本手法で抽出した主要イベント群の主語及び目的語は、1つの動詞に対して主語及び目的語が複数抽出した場合に統合しているため、抽出結果に正解以外のものを含んでいる場合がある。そのため、正解以外のものを含んでいる場合は、「誤りを含む正解」として正解・不正解と区別して集計している。4種類の生成手法の内、類似記事集合を用いて、さらに係り受け関係の追加処理を行った生成手法Dが主語と目的語共に最も正解率が高く、誤りを含む正解も正解とした場合、主語の正解率が91.0%、目的語の正解率が73.5%となった。図13は、生成手法Dで抽出した主語及び目的語の正解率を、事件ごとに集計したものである。この結果から、事件によって正解率にばらつきがあることがわかる。表3は、実験1の主要イベント群の動詞の抽出精度を事件ごとに集計したものである。適合率の平均が69.5%、再現率の平均が92.7%であった。各動詞の発生順については、5つの事件中3つの事件に対して正解した。不正解の事件の内、1つは、被験者の回答にばらつきがあり、正解定まらなかったため、評価の対象外とし、もう1つは1箇所動詞の順序が入れ替わっていた程度の間違いであった。

次に、実験2の主観評価の結果を表4に示す。表4は、実験2の主観評価に対する結果を類似記事集合ご

とに集計したものである。表4のX、Yは被験者を新聞記事を読む頻度でグループ化したもので、Xは「新聞記事をよく読む人のグループ」、Yは「新聞記事をあまり読まない人のグループ」である。被験者全体の平均評価は3.6であった。一方、新聞記事をよく読む人の事件全体の平均評価は3.0で、新聞記事をあまり読まない人の事件全体の平均評価は4.4であった。また、自由記述での回答については、新聞記事をあまり読まない被験者からは「新聞記事を読む前にグラフによって事件の構図を頭に思い浮かべておくことで、その事件の構図に当てはめながら記事を読み進めていけるため、理解支援に役立った」等の肯定的な意見が見られた。新聞記事をよく読む被験者からは「グラフがなくても記事を読むことで事件の構図を思い浮かべることができるため、グラフは新聞記事を理解支援にはあまり役に立たなかった」等の否定的な意見が見られた。さらに、新聞記事をよく読む被験者からの回答で、「主要イベントだけでは情報が少なく感じる」という意見もあった。

表 2 各生成手法で抽出した主語・目的語の正解率

生成手法	主語の正解率[%]	目的語の正解率[%]
A	34.0 (34.0)	21.0 (21.0)
B	51.5 (61.5)	38.0 (42.0)
C	48.5 (59.5)	43.5 (48.5)
D	61.5 (91.0)	62.0 (73.5)

※ 括弧内の値は、誤りを含む正解も正解とした場合の正解率を表す。

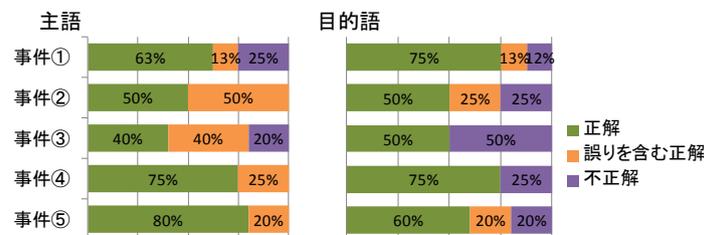


図 13 生成手法Dで抽出した主語・目的語の正解率

表 3 主要イベント群の動詞の抽出精度

	事件					①から⑤の平均
	①	②	③	④	⑤	
適合率 [%]	62.5	75.0	30.0	100.0	80.0	69.5
再現率 [%]	83.3	100.0	100.0	80.0	100.0	92.7

表 4 新聞記事の理解支援に役立つかの主観評価

	事件					①から⑤の平均
	①	②	③	④	⑤	
Xの平均	3.0	2.6	2.7	3.3	3.6	3.0
Yの平均	4.6	4.0	4.0	4.4	4.8	4.4
被験者全員の平均	3.7	3.2	3.3	3.8	4.1	3.6

※ X:新聞記事をよく読む被験者のグループ(7名)

Y:新聞記事をあまり読まない被験者のグループ(5名)

5.5. 考察

本節では、本手法が主要イベント群を正しく抽出できているのかについて、事件の構図を表現したグラフが新聞記事の理解を支援できるのかについて考察する。

まず、実験 1 より、本手法が事件の構図を表現するグラフを正しく生成できているのかについて考察する。

主要イベント群の主語及び目的語の抽出精度は、表 2 の結果から、生成手法 D が主語と目的語共に最も正解率が高かった。これは、類似記事集合を用いたことと、係り受け関係の追加処理を行ったことが、抽出精度の向上に役立ったといえる。

また、誤りを含む正解も正解とした場合の生成手法 D で抽出した主語の正解率は 91.0%と概ね良好であった。しかし、目的語の正解率が 73.5%と高いとは言えない結果となった。

主語の正解率より目的語の正解率が低い点については、目的語に対する係り受け関係の追加処理を行っていないことが原因であると考えられる。よって、目的語に対する係り受け関係の追加するためのルールを検討する必要があると考える。

さらに、誤りを含む正解が多い点については、係り受け関係の追加処理の際の間違った係り受け関係を主語として追加してしまうことと、動詞が受動態である場合に場所を表す名詞が主語になってしまうことが原因として挙げられる。よって、係り受け関係を追加する処理のルールの改善や、動詞が受動態である場合のルールの改善が必要であると考えられる。

主要イベント群の動詞の抽出精度は、表 3 の結果から、再現率の平均は 92.7%と概ね良好であったが、適合率の平均は 69.5%と高いとは言えない結果となった。つまり、類似記事集合内の基準となる記事の 1 段落目に登場する動詞を全て、主要イベント群の動詞として抽出するのではなく、その中から主要イベント群の動詞を選び出さなければならない。

また、動詞の発生順の正解率は 75%と高い結果となった。つまり、類似記事集合内の基準となる記事の 1 段落目に登場する動詞の出現準を動詞の発生順とするという方法は、有効であったといえる。

次に、実験 2 より、事件の構図を表現したグラフが新聞記事の理解を支援できるのかについて考察する。

表 4 の結果から、新聞記事をよく読む人の平均は 3.0 と高いとは言えない結果であったが、新聞記事を読み取らない人の平均は 4.4 と概ね良好であった。この評価の差について t 検定を行ったところ、P 値が 0.6×10^{-3} となり、有意水準 1%で、有意差があることが確認できた。

また、自由記述の回答で、新聞記事をよく読む人の回答の中に、「主要イベントだけでは情報が少なく感じ

る」という意見もがあった。

つまり、事件の構図を表現したグラフは新聞記事を読み取らない人にとっては、新聞記事の理解支援に役立つといえるが、新聞記事をよく読む人にとっては、主要イベントからの情報だけでは新聞記事を理解するためには不十分であることがわかった。

今後は、主要イベント以外の情報を関連付ける方法についての検討や、係り受け関係を追加する処理のルールの改善、受動態への対応、また、主要イベント群の動詞の抽出方法についての検討を行う予定である。

6. おわりに

本研究では、新聞記事の理解支援を目的とし、類似記事集合から主要イベントとして、重要な動詞とその動詞に対する主語及び目的語を抽出し、抽出した主要イベント群を基に事件の構図を表現するグラフ生成手法を提案した。提案手法を評価するために、主要イベント群の抽出精度を確認した結果、主語の正解率は 91.0%、目的語の正解率は 73.5%となった。また、事件の構図を表現したグラフが新聞記事の理解を支援できるかを評価するために、被験者にグラフと記事をセットで見てもらい、アンケートに回答してもらった。その結果、新聞記事を読み取らない被験者からは、グラフが新聞記事の理解支援に役立つとの意見が得られ、提案手法の有用性が確認できた。

今後は、主要イベント群の抽出精度の向上と主要イベント以外の情報を主要イベントに関連付ける方法について検討し、手法の改良を行う予定である。

参考文献

- [1] 内閣府：第 5 回情報化社会と青少年に関する調査の概要、
<http://www8.cao.go.jp/youth/kenkyu/jouhou5/gaiyou.pdf>
- [2] 株式会社アイシェア：漫画新聞に関する意識調査、
<http://release.center.jp/2009/12/0802.html>
- [3] 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．データベース 43(SIG_2(TOD_13))，pp. 37-47, 2002.
- [4] 金山淳一，北条孝，田村直良，“文章の構造解析による新聞記事からの事件情報抽出”，情報処理学会研究報告．自然言語処理研究会報告，pp. 1-6, 2002.
- [5] 数原良彦，戸田浩之，櫻井彰人，“ブログ記事を用いた複数話題語間の動作関係抽出手法”，電子情報通信学会論文誌．D，情報・システム J91-D(3)，pp. 619-627, 2008.
- [6] MeCab/和布蕪:Yet Another Part-of-Speech and Morphological Analyzer,
<http://mecab.sourceforge.net/>
- [7] CaboCha/南瓜:Yet Another Japanese Dependency Structure Analyzer,
<http://chasen.org/~taku/software/cabocho/>