

E-3

要約文生成のための集合名詞の要素の抽出

Extraction of Elements in Collective Nouns for Summary Sentence Generation

深井 康治[†]
Koji Fukai

関口 芳廣[†]
Yoshihiro Sekiguchi

鈴木 良弥[†]
Yoshimi Suzuki

1. はじめに

筆者らは音声認識の研究をしているが、講演や物語のような長時間の発話を理解するためには文章の要約が欠かせない。特に人物や物事に焦点をあてた要約文を生成できれば、発話内容の適切な理解に役立つと考えられる。

適切な要約文を生成するためには様々な問題を解決しなければならないが、集合名詞に含まれる要素を正しく抽出することも重要な課題である。この問題に対しては、すでに様々な試みがあるが、ここでは、工学的な立場から、できるだけ簡便な手法で、集合名詞に含まれる要素を抽出する手法を検討しているので、その手法を、実例を含めて報告する。

2. 集合名詞とその分類

ここでは、「仲間」のように、複数の人間や動物等をまとめて表現している名詞を集合名詞と呼び、その集合名詞が含んでいる個々の人間や動物を集合名詞に含まれる要素と呼ぶことにする。例えば、

「ロバとイヌとニワトリは仲間になりました。」

という文では、集合名詞「仲間」の要素は、「ロバ、イヌ、ニワトリ」である。

集合名詞の要素を抽出するために、集合名詞を以下の3種類に分類している。

- I. 文中で要素が明示されている集合名詞
例：サルとイヌは「家来」になりました。
- II. 文中で要素が明示されていない集合名詞
例：それから、「みんな」は仲良く暮らしました。
- III. 単数形の名詞に接尾辞がついた集合名詞
例：「子ガニたち」はサルをやっつけてしまいました。

以下に、集合名詞の種類毎に、その要素を抽出する方法を説明する。

3. 文章中の単語の同定処理

集合名詞の要素抽出処理をする前に、文章中出现した「同じ表現の名詞」を同一内容であるか否か判定する。同一か否かの判定は、それらの単語に係っている単語の比較や、主題・焦点の得点 [1] 等を利用して行なう。これを前処理と呼ぶ。

上記の処理を行なって、集合名詞以外の名詞に対して同定の得点が正ならば、比較した単語同士は同一な概念と判定し、集合名詞に対しては、得点づけのみを行なう。

4. 集合名詞の要素の候補の推定方法

文章中に集合名詞の要素が明示されている場合は、それらを要素として確定する。その他の場合は、要素の推定の規則を作り、これらの規則を用いて集合名詞の要素の推定を行なう。また、前処理で行なった単語の同定の結果、同一とみなされた単語は、同一の概念として扱う。

なお、抽出された集合名詞の要素を利用して、集合名詞同士の同定も行なう。

4.1 要素が明示されている場合

(1) 変化を表す動詞の利用

変化を表す動詞が使われた場合、変化のゴールにあたる名詞が集合名詞なら、変化の主体を要素とする。

例：サルは仲間になりました。(サル⇒仲間の要素)

(2) 断定の助動詞「だ」の利用

一般には、助詞「の」を伴って、「A は、C の B (集合名詞) だ」「A は、C の B (集合名詞) です」など使用される場合が多い。この場合は主体 A を集合名詞 B の要素とする。

例：サルは桃太郎の仲間です。

この例のように「C の B (集合名詞)」となる場合、集合名詞の要素を代表するものとして、C を B の「主」と呼ぶことにする。上記の例では、「仲間」の主は桃太郎である。また、(1) の例の様に誰の「仲間」になったのわからない場合は、対象としている集合名詞と同一の文内にある単語、集合名詞の前文の主題または焦点の中で人間、動物の属性を持つ単語から、「主」を見つける。

(3) 前処理の同定の結果の利用

対象としている集合名詞の要素の候補と同定の得点が正であった集合名詞の要素の候補を比較して、両集合名詞間の同定と要素の受渡しを行なう。

4.2 要素が明示されていない場合

集合名詞の要素が明示されていない場合は、下記の方法で、集合名詞に含まれる要素の推定を行なう。なお、ここで出てくる、「探索範囲」とは、対象としている集合名詞と同定の得点が正であった集合名詞（対象としている集合名詞から最も近い集合名詞）を含む文の次の文から対象の集合名詞を含む文の前文までの範囲である。

(1) 対象となる「集合名詞」に合致した要素を探索範囲の中から複数個探す。

ここでは、対象となる集合名詞の前文の主題、または焦点等の情報を用いて候補が複数個になる文まで文単位で探索を行なう。

例：おじいさんとおばあさんは、帰ってきた桃太郎を見て大喜びしました。それから、みんなは仲良く暮らしました。

の場合、集合名詞「みんな」の要素として、人間、動物の属性を持つ単語を探索範囲から探して、「桃太郎、おじいさん、おばあさん」を要素の候補とする。

(2) 対象となる「集合名詞」と同一の文内にある単語を候補から外す。

例：みんなは、力を合わせて、鬼を退治しました。

の場合、集合名詞「みんな」の要素の候補「鬼」の得点を下げる。

(3) 対象としている集合名詞の候補と同定の得点が正

[†]山梨大学工学部、甲府

であった集合名詞が「要素が明示されている集合名詞」の場合

対象としている集合名詞の要素は、同定の得点が正であった集合名詞の要素のみである可能性が高い。しかし、話の流れから要素が変化している可能性も考慮して、この場合は、ここまで与えた要素の候補の得点を下げる。

(4) 前処理の同定の結果の利用

- i. 対象としている集合名詞の要素の候補との同定の得点が正であった集合名詞の要素の候補を比較して、両集合名詞間の同定と要素の候補の得点づけを行なう。
- ii. 対象としている集合名詞との同定の得点が正であった集合名詞を含む文内に存在する人間、動物の属性を持つ単語は、要素の候補としての得点を下げる。

4.3 接尾辞がついた集合名詞の場合

要素の候補の探索範囲は、「要素が明示されていない場合」と同様である。

(1) 対象となる集合名詞と同一の単数形の単語の利用

動詞の格になっており、単数形の単語が集合名詞「Aたち」の「A」である場合、係っている単語等を比較して、「A」を探索範囲から見つけ、その単語を利用して要素の候補の推定を行なう。また、対応する単語が見つからない集合名詞は、4.3(2) 以下は、行なわない(但し、4.2.(4)のiと同様の処理は行なう)。なお、ここで抽出した「A」にあたる単語をこの集合名詞の「主要素」と呼ぶ。

(2) 助詞、格の種類の利用

対象となっている集合名詞の助詞、格の種類を利用して要素の候補の推定を行なう。

(3) 前処理で比較していない集合名詞の利用

対象としている集合名詞と比較しておらず、「主」または「主要素」が一致しない集合名詞の要素である候補は、得点を下げる。

(4) 「主要素」の利用

抽出した「主要素」が別の集合名詞の「主」であった場合、対象としている集合名詞の要素の候補の内、別の集合名詞の要素の候補になっているものに得点を与える。

以下、4.1の(2)、4.2の(2),(3),(4)のiと同様の処理を行なう。

5. 集合名詞同士の同定

ここまでの処理で求めた、前処理の同定の得点、集合名詞同士の要素の候補の比較、また、「主」が存在する場合には、「主」の比較を行ない集合名詞間の同定を行なう。

もし、「主」、「主要素」が存在する集合名詞と同一と判定された単語で、「主」、「主要素」が選ばれていない場合、同一と判定された単語の「主」、「主要素」を継承する。また、「要素が明示されていない」集合名詞が、「要素が明示されている」集合名詞と同一の概念を表すと判定された場合、4.2(1),(2)の処理はリセットされ、これ以降、この集合名詞を「要素が明示されている」集合名詞として扱うようにする。

6. 集合名詞の要素の抽出

集合名詞同士の同定で比較した集合名詞が別の概念を持つ集合名詞であった場合は、要素の候補の得点はもとのままとする。また、同一の概念と判定された場合は、比

較対象にした集合名詞の要素の候補の得点を、対象としている集合名詞の要素の候補の得点に加える。

その結果から、要素の候補の得点が閾値より高いものをその集合名詞の要素として抽出する。ただし、「接尾辞がついた集合名詞」の場合、「主」または「主要素」がなければ要素の抽出は行なわない。また、「要素が明示されている集合名詞」の場合、要素が多少異なっている集合名詞同士の同定の結果、同一の概念を表していると判定された場合、話の流れに対処するために、要素に「第何文から第何文まで」に出現する集合名詞の要素かという情報を付け加えておく。

7. 集合名詞の要素抽出例

参考文献[2]の一部を変更した下記の文を例として実験を行なった。

< 使用した文の例 >

むかしむかし、あるところにおじいさんとおばあさんがいました。おばあさんが川で洗濯をしていると川上から大きな桃が流れてきました。おばあさんはその桃を家に持って帰りました。おじいさんとおばあさんが、桃を切ってみると、桃の中から元気のいい男の子が出てきました。おじいさんとおばあさんは、その男の子に桃太郎と名前をつけて、大事に育てました。桃太郎は何でも食べて、とても大きくなりました。悪いことが嫌いな桃太郎は村を襲う鬼が住む鬼ヶ島へ鬼退治に行くことにしました。桃太郎は、おじいさんに「日本一」ののぼりを、おばあさんにきびだんごを買い、鬼ヶ島に出发しました。しばらく行くと、桃太郎は、猿、犬、雉に出会いました。猿と犬と雉は、きびだんごを買い仲間になりました。そして、桃太郎と仲間たちは船に乗り、鬼ヶ島を目指しました。ほどなく、鬼が住む鬼ヶ島に着きました。鬼の屋敷の前に大きな門がありました。桃太郎は、雉に中の様子を見に行かせました。すると、鬼は酒盛りをしていました。桃太郎は猿に門の鍵を開けさせました。桃太郎は、鬼を投げ飛ばしました。犬は鬼に噛み付き、雉は鬼の目を突き、猿は鬼の顔を引っかきました。桃太郎たちは力を合わせ、鬼を退治しました。鬼は謝罪し、桃太郎たちに宝を差し出しました。桃太郎は、宝を船に積み込み、おじいさんとおばあさんの元へ帰りました。おじいさんとおばあさんは、帰ってきた桃太郎を見て大喜びしました。それから、みんなは仲良く暮らしました。

< 集合名詞とその要素抽出結果 >

(仲間)(主:桃太郎)
 < 要素: 猿, 犬, 雉 >
 (仲間たち)(主:桃太郎)
 < 要素: 猿, 犬, 雉 >
 (桃太郎たち)(主:なし)(主要素:桃太郎)
 < 要素: 桃太郎, 猿, 犬, 雉 >
 (桃太郎たち)(主:なし)(主要素:桃太郎)
 < 要素: 桃太郎, 猿, 犬, 雉 >
 (みんな)(主:なし)
 < 要素: おじいさん, おばあさん, 桃太郎 >

この例では、ほぼ正しく集合名詞の要素を抽出できている。また、この他の物語文[2][3]14話に対しても適切な結果が得られた。

8. おわりに

簡単な方法ではあるが、本手法を用いることによって、集合名詞の要素の抽出と集合名詞同士の同定がある程度できるようになった。今後は、高度な意味情報、指示詞の照応手法等を加えさらに精度を向上させ、要約文生成に役立てたい。

参考文献

- [1] 村田真樹, 長尾真: “用例や表層表現を用いた日本語文章中の指示詞・代名詞・ゼロ代名詞の指示対象の推定”, 自然言語処理(言語処理学会誌), 1997年1月, 4巻, 1号, pp87-pp109.
- [2] イー・こども・ドットコム: “<http://www.e-codomo.com>”.
- [3] 仲田安津子: “3さいの本・おはなし”, 講談社, 1990.