ニュース原稿を利用した用語集作成の検討

山田一郎　　柴田正啓　　金淵培
ＮＨＫ放送技術研究所
E-mail: { ichiro, shibata, kimyb }@strl.nhk.or.jp

1 はじめに

放送用ニュース原稿は、最新の社会情勢や一般常識といった有益な情報を含み、映像ともリンクするため、教育用コンテンツ素材として有用である。例えば、生徒の質問に答えるQ&Aシステムへの利用が考えられる。この時、ニュースで使われる用語知識を蓄積した用語辞書が必要となる。しかし、毎年多くの新語がニュース原稿中に出現するため、用語辞書の更新には大変な労力を必要とする。そこで、用語の定義を自動獲得する技術が求められる。

新たな用語がニュースで扱われる場合は、視聴者が容易に理解できるよう、用語の説明を伴うことが多い。我々はこの説明を利用して、ニュース原稿中で使われる用語集を自動構築するための研究を進めている。

従来、テキストから用語の定義を抽出する研究として、表層パターンマッチングに基づいた手法が提案されている[1][2]。西野らは「αとはβである」という表現に的を絞り、新聞記事から用語とその定義文を抽出している。また木田らは「αはβ」という表現を分析し、αとβの関係を、その表層表現により整理している。しかし後述するとおり、ニュース原稿では上記の表現を使うことが少なく、用語辞書構築のためには十分でない。

本稿では、ニュース原稿を解析し、用語を説明する定型的な構造をもとに用語の定義を抽出し、用語集を自動作成する手法を提案する。

2 ニュースにおける用語定義の種類

本稿では、NHKの放送用読み原稿として利用されるニュース記事を処理対象とする。このニュース記事は、１日当たり約200記事が作成されている。我々が所有するデータベースには、10年分のニュース約33万記事（200万文）の大量テキストデータが蓄積されている。

ニュース原稿中で用語を説明する場合、その表現は以下の３通りに分類できる。

A) 連体修飾節の係り元に定義部、係り先に用語
例: 寝入りばなに怖い夢を見る「入眠時幻覚」は、・・・

B) 連体修飾節の係り元に用語、係り先に定義部
例：「情報家電」と呼ばれる次世代の高速インターネットに対応した家電製品を・・・

C) 文の主部に用語、述部に定義部
例：「クローン規制法」は、クローン技術を使って同じ遺伝子を持つ人間を人工的に作り出すことを禁止する法律です。

例では鍵括弧内が用語、下線部が定義部に対応している。2001年6月のニュース原稿を対象として、用語を定義する文を手作業により抽出し、どのパターンに属するか調査した。結果を表１に示す。

パターン	出現数
A	397(74.6%)
B	94(17.7%)
C	41(7.7%)

表１．用語を定義するパターンの出現数

この結果から、ニュース原稿中で用語を定義する場合、圧倒的にパターンAが多いことがわかる。放送のニュースでは、多くの情報をできるだけ短い時間で伝えることが重要である[7]。そのため、表１に示すように、１文で用語の説明をするパターンCは避けられ、用語の説明とニュースの主題を同時に記述できるパターンAが多く出現していると考えられる。実際に同じ期間の毎日新聞の記事の第一文とＮＨＫニュース原稿の第一文の文字数を調べたところ、新聞が平均11.3文節（122.4文字）であったのに対し、ニュース原稿は、18.2文節（178.8文字）と新聞より約1.6倍も長い文であり、従属節が多用されていた。そのため、ニュース原稿から用語の定義を抽出する処理では、パターンAの解析が重要となる。

また、パターンBでは、その定義部を構成する文節の数が少ないという特徴がある。実際に上記データを調査したところ、平均2.0個の文節しか存在しなかった（パターンAは平均7.9個、パターンCは平均9.5個）。このパターンは、用語を容易な単語に言い換える目的で使われることが多く、用語を十分に表現する定義として相応しくない。そこで本手法では、この言い換え表現を用語の上位概念として抽出し、定義文生成時に補完する処理において利用する。

パターンCについては、文献[1][2]により考察が行われているので深くは言及しない。本手法では、「αはβです」のパターンのみから定義部を抽出する。

3 用語定義抽出手法

本手法では、図１に示す手順により用語集を作成する。まず、ニュース原稿集合から、定義抽出対象とする用語を抽出する。ここでは、強調を意図する鍵括弧で囲まれた名詞句[5]に限定して抽出した。鍵括弧で囲まれていても一般的な語には定義を抽出する必要が無いため、ここでは辞書に登録されている語は処理対象から除いている。次に、用語定義のパターンを判別し、前章で説明したパターンＡとパターンＣを解析して、用語の定義部分を抽出する。パターンＡから抽出された連体修飾節には、用語の上位概念を抽出して、定義文とする。このとき、パターンＢの構造を利用する。最後に一つの用語に対して複数の定義文が抽出された場合に、最適な定義を選択し、用語集データベースに登録する。以下に各処理について説明する。

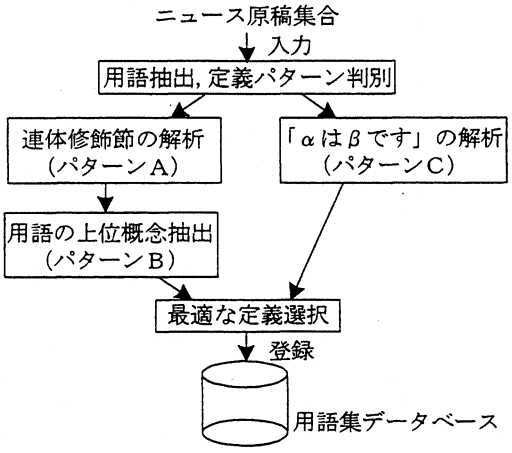


図１．用語集作成手順

3．1 連体修飾節の解析（パターンＡ）

ニュース原稿から用語の定義を抽出するために、パターンＡの解析が重要であることは２章で述べた。しかし、この用語を連体修飾するすべての節が、その定義であるとは限らない。例えば、以下では、前者は「電子政府」を定義した連体修飾節であるが、後者は用語の定義ではない。

- 住民票の取得や、企業が行う許認可などの手続きを役所に出向かなくてもインターネットでできるようにする「電子政府」の実現などの・・・
- 政府が推進する「電子政府」の構想では、・・・

ここでは、前者の用語を定義する節を「用語定義節」、後者の用語の補足的な説明をする節を「用語補足節」を呼ぶ。用語補足節も付加的な情報として有用であるがその役割が用語定義節とは異なるため、ここでは用語定義節と用語補足節の識別を行い、用語定義節のみを抽出対象とした。

ニュース原稿では定型的な表現が多く使われるが[3]、用語定義節にも、用語補足節とは異なる定型表現が用いられる。そこで、手作業により用語定義節を抽出し、用語定義節と、用語語補足節の表現の違いを調査した。各節中の用語に係る動詞は、その24.2%が両者で共通であったのに対し、用語に係る動詞とその直前の助詞の２項組では、共通なものは8.6%であった。そこで、用語定義節ではこの２項組に定型表現が用いられると仮定して、連体修飾節が用語定義節であるか否かの判断を行った。

また、学習データに出現する動詞の種類には限りがあるため、動詞と助詞の２項組データではスパースネスの問題が生じる。そこで、分類語彙表[4]を利用して、類似する動詞の学習データも利用した。

用語に係る連体修飾節中の動詞を v、動詞 v と同じグループに属する動詞集合を $vg1$、動詞 v の親ノードに属する動詞集合を $vg2$、動詞 v の直前の助詞を p、動詞の類似度に対する重み付け係数を w_a、w_b、動詞集合 vg と助詞 p が学習データ中に出現した回数を $n(vg,p)$、その期待値を $e(vg,p)$ としたとき、連体修飾節が用語定義節であるかを判断するための指標を表す $weight(v,p)$ を以下のように定義した。

$$weight(v,p) = w_a \times \frac{(n(vg1,p) - e(vg1,p))^2}{e(vg1,p)} + w_b \times \frac{(n(vg2,p) - e(vg2,p))^2}{e(vg2,p)}$$

ここで、$n(vg1,p)<e(vg1,p)$ の時は上式の第一項を０、$n(vg2,p)<e(vg2,p)$ の時は第二項を０とした。この値がしきい値より大きい場合、用語定義節と判定した。本手法では、学習データとして手作業により抽出した15295 個の用語定義節を与え、実験的に w_a=0.67、w_b=0.33 とし、$weight(v,p)$>1.0 である連体修飾節を用語定義節とした。「とする」「を通過する」の２組について $weight$ を計算したところ、以下に示す値となった。

$weight$（”する”,”と”）= 92.5
$weight$（”通過する”,”を”）= 0

前者が用語定義節の特徴的な表現と判断できる。

3．2 用語に係る節の抽出問題

連体修飾節を利用する場合、用語の前に出現する節の、どこまでが用語の定義となるかを判定しなければならない。例えば下記のニュース原稿では、下線部は用語の定義ではない。

- 群馬県内では、当面、畜産農家が牛を食肉として出荷する際に付ける「生産履歴書」という文書に、・・・

構文解析の結果から判定できそうであるが、複数の構文解析システムによる処理結果では、その全てが、下線部の「群馬県内では、当面、」は「付ける」に係ると解析されてしまった。現状の技術では、この判定は大変難しい。これは、今後の課題となる。本手法では、構文解析結果から、「係助は」「代名詞」「時詞」などを含む節のみを除外する処理のみに留めている。

3.3 用語の上位概念抽出処理

3.1節で抽出した連体修飾節は、動詞の連体形で終わっているため、このままでは、文として不完全である。そこで、用語の上位概念を抽出して、この連体修飾節と統合することにより、定義文を生成する。

本手法では、ニュース原稿の表層的特徴を基に、用語の上位概念を抽出する。用語の最終形態素が上位概念となるもの（「航空安全法案」）、用語の直前に上位概念があるもの（通貨「ユーロ」の～）、用語の直後に「という」「と呼ばれる」といった定型的な言い換えを意図する句があり、その後に上位概念がくるもの（「マルス」と呼ばれるコンピュータシステムは、～）の場合に分けて処理を行った。

まず、抽出された用語を形態素解析し、用語の最終形態素が辞書に含まれている場合は、その形態素を上位概念とした。この処理により、抽出された用語の44.1%に上位概念が与えられた。また、用語が含まれるニュース原稿を構文解析し、用語と並列関係にある直前の名詞を上位概念とした。この処理により、抽出された用語の12.7%に上位概念が与えられた。

用語の直後に言い換えを意図する句がある場合は、2章で説明したパターンBに該当する。上記の例では、「マルス」の上位概念は「コンピュータシステム」と特定できるが、下記の例では、その特定が困難となる。

- ～に溶けていた「ホスゲン」という人体に有毒なガスが、～

この例では、用語「ホスゲン」の上位概念が「人体」であるか「ガス」であるかを、表層的には判断できない。そこで、用語の直前にある動詞との整合性を調べる。この例では、「人体」＋助詞＋「溶ける」の組み合わせと「ガス」＋助詞＋「溶ける」の組み合わせの過去のニュース原稿における出現頻度を調べ、その相互情報量の大きさを基準として、どちらが上位概念になりやすいかを判定した。この処理により、抽出された用語の7.4%に上位概念が与えられた。

これらの3つの処理により抽出された上位概念を検証した結果、適合率は95.7%と良好な結果が得られた。上位概念が得られなかった用語は、「もの（こと）」という一般的な単語をその上位概念とした。

3.4 「αはβです」の解析（パターンC）

文献[1]で考察された「αとはβである」というパターンは、国語学では「真の題目」と呼ばれ、定義抽出では有効な表現となる[2]。しかし、2001年6月のニュース原稿で、「αとはβ」のパターンは出現しなかった。そこで、本手法では「αとはβである」の関係を包含する表現「αはβである」に着目する。ニュース原稿では、語尾に丁寧語が用いられるため、実際には、「αはβです」の表現を対象とした。この表現では、全てのβがαの定義となるとは限らない。下記の例では、βの部分（下線部）は、企業の活動に関する情報となっており、用語の定義ではない。

- 「S社」は、再来年に千葉県袖ヶ浦市に出力五万キロワットの発電所を建設し、他の企業に電力を販売する計画です。

そこで、βの最終形態素にあたる名詞が、「計画」「方針」「意向」といった事態の確実性を表す名詞[2]の場合は、βは主体の行為を表していると判断し、定義として抽出は行わない。

3.5 最適定義文選択

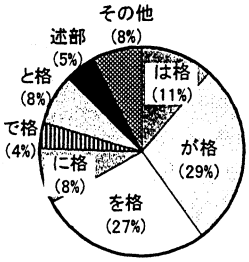
ニュース原稿中に同じ用語が繰り返し出現する場合、用語の定義文も複数抽出される可能性がある。抽出された定義文に優先順位をつける必要が生じる。そこで、同じ用語に対して複数の定義文が生成された場合には、定義文の動詞に係る文節数が多く、かつ新しいニュース原稿から生成されたものほど優先順位を高くして順位付けを行った。最後に、優先順位の高い定義を用語集データベースに登録することにより、用語集システムを構築した。

4 実験

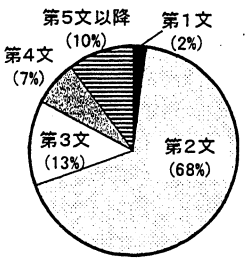
前章までに説明した提案手法を検証するために2001年6月のニュース原稿を対象として用語定義文を抽出する実験を行った。結果の一部を表2に示す。

この結果を検証したところ、適合率が79.2%(304/384)、再現率76.6%(304/397)であった。この正解データに対して、パターンAの構造によって定義される用語が出現した格（格助詞、係助詞を対象）と、パターンCの構造が何文目に出現したかを調査した結果を図2に示す。

この結果では、連体修飾節による用語定義は「が格」と「を格」に多くみられた。この構造がニュース原稿



連体修飾節により
定義された用語の格



「αはβです」に
よる定義の位置

図２．用語の定義が出現する位置の調査結果

の何文目に出現したかを調査したところ、第１文が36%、第２文18%、第３文15%、第４文15%、第５文以降15%と、ニュース原稿の最初のほうに出現しやすいこともわかった。また、パターンＣによる定義は圧倒的に第２文に多いことがわかる。今後、これらの特徴を利用することにより、用語定義抽出の精度向上が期待できる。

また、ここでは鍵括弧で囲まれた用語を重要と判断してその定義を抽出する実験を行ったが、TFIDF 等を利用して用語の重要性を評価することにより、鍵括弧で囲まれていない用語へも応用可能と考えられる。

5 まとめ

本稿では、ニュース原稿を利用した用語定義文生成の手法の検討を行った。ニュース原稿中での用語定義のパターンを整理し、その各パターンを利用した用語の定義文抽出処理を行い、実験で良好な結果を得た。また、調査の結果、ニュース原稿中で用語を定義する位置の特徴を明らかにした。この手法をもとに構築したシステムは、毎日生成されるニュース原稿を処理対象としているので、毎日新しい用語集に自動更新できる。さらに、ニュース原稿に対応した動画も存在するため、これからのマルチメディアコンテンツとして有益と考えている。

今後、本結果を利用したマルチメディア教育支援システム[6]へと進めていく予定である。

【参考文献】

[1] 西野ほか「テキストからの用語とその定義文の抽出」言語処理学会第５回年次大会, pp124-127 (1999)

[2] 木田ほか「新聞記事からの用語集作成のためのテキスト分析」情処学会研究報告, NL134-12, pp85-92 (1999)

[3] 山田ほか「ニュース記事からの話題構成要素抽出の検討～国会審議に関する話題を対象として～」言語処理学会第７回年次大会, pp297-300 (2001)

[4] 中野「分類語彙表形式による語彙分類表(増補版)」国立国語研究所(1996)

[5] 後藤ほか「かぎ括弧で囲まれた表現の種類の自動判別」言語処理学会第６回年次大会, pp35-38 (2000)

[6] 住吉ほか「エージェントを利用した映像検索のためのユーザーインターフェイス」信学技報, OFS2000-24, pp9-14 (2000)

[7] 奥秋義信「ニュース原稿の書き方～その理論と実際」岩崎放送出版社(1970)

用語	生成された定義文
クローン規制法	クローン技術を使って同じ遺伝子を持つ人間を人工的に作り出すことを人間の尊厳や社会の秩序に重大な影響を与えるとして、禁止する法律
顔文字	パソコンや携帯電話でやり取りする電子メールについて記号などを使って感情を表す表現
石油備蓄法	原油価格の高騰といった緊急時に対応するため石油の輸入業者を登録制にして備蓄体制を強化する法律
炭素税	温暖化対策の一つとして、石油や石炭に課税して、燃やすと二酸化炭素を出すエネルギーの消費を抑えようというもの
エコ・ツーリズム	観光客の人数を制限して、専門のガイドの案内で自然を観察してもらおうというもの
タウンミーティング	小泉内閣の閣僚が、国民と直接、政策課題について意見を交わすもの(こと)
世界水フォーラム	干ばつや洪水など地球規模で深刻化する水資源の問題の解決策を話し合おうと国際機関などが参加して三年ごとに開かれているもの
官製談合	国や地方自治体の職員が、業者に入札の最低価格を教えたりして関与する談合
金沢百万石まつり	前田利家と城下町金沢の歴史を偲ぶまつり
確定給付企業年金法	サラリーマンが退職した後に確実に企業年金を受け取れるように、新たに二種類の企業年金制度を設けることなどを柱とした法律

表２．用語の定義文抽出例