ニュース原稿を利用した用語集作成の検討

1L－1

山田一郎　柴田正啓
NHK放送技術研究所

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」、木田らは「αはβ」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た。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、ニュース原稿から新たな用語を説明する定型的な用語説明構造を解析し、そこで最も多く見られる連体修飾により用語を定義する節の抽出実験を行った。以下にその内容を報告する。

2 ニュースにおける用語定義の種類

ニュース原稿中で用語を定義する場合、その表現は以下の3通りに分類できる。

A) 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。

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) 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、・・・

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	41（7.7%）
B	94（17.7%）
C	397（74.6%）

表1．用語を定義するパターン

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Cが多いことがわかる。パターンAは、文献[1][2]により考察が行われ、また、パターンBは「という」「と呼ばれる」といった特定の表現が用いられるため、その定義部を容易に抽出できる。そこで、以下ではパターンCの場合についての考察を行う。

3 用語定義節抽出実験

パターンCの用語を連体修飾するすべて節が、その定義をしているとは限らない。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統語構造が多く使われるため[3]、用語定義節も、用語補足節とは異なる定型表現が用いられると考えられる。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、共通項は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

*A study on generating a term dictionary using news articles.
Ichiro Yamada, Masahiro Shibata
NHK Science & Technical Research Laboratories

「逆ざや」	長引く低金利によって資産の運用利回りが契約者に約束した「予定利率」を下回る
「京都議定書」	地球温暖化を食い止めるために先進国に二酸化炭素の排出削減を義務付けた
「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
「産業廃棄物税」	企業などが出す産業廃棄物に県が独自に課税する
「司法制度改革推進法」	司法制度改革を推進するための体制などを定める
「自動車窒素酸化物削減法」	大都市の深刻な大気汚染を改善するためディーゼル車の排気ガスに含まれる有害な粒子状物質を新たに規制することを柱とした
「御祭文」	天皇陛下のお使いの侍従が陵所の完成を告げる
「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
「プルサーマル」	取り出したプルトニウムを一般の原発で燃やす
「マネーロンダリング」	麻薬取引などで得た不正な資金の出どころを判らなくする
「レッドデータブック」	絶滅の恐れのある野鳥を記録した

表２．　用語定義節抽出結果（一部）

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界があるため、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を利用した。

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 $weight(v,p)$ は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。この値を利用して、用語定義節を抽出した。用語は、強調を意図する鍵括弧で囲まれた名詞句[5]を処理対象とした。2001 年 6 月のニュース原稿から抽出した結果の一部を表２に示す。この処理では、学習データとして 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判断した。

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)であり、１ヶ月間のニュース原稿から 195 種類の用語に対する用語定義節が自動抽出できた。この用語定義節は用言の連体形で終わっているため、そのままの形では定義文としては不自然である。辞書に登録する際には、節末に「こと」「もの」などの名詞を自動選択して追加する処理が必要となる。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。

4 まとめ

本論文では、ニュース原稿の定型性を利用して、連体修飾により用語を定義する節を抽出した。その結果、連体修飾節に含まれる動詞と助詞の２項組の定型性が指標の一つとなることがわかった。

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回年次大会論文集, 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」情報処理学会研究報告, 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回年次大会論文集,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回年次大会論文集, 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)