テキストからの用語とその定義文の抽出

西野 文人[†]，橋本 三奈子[‡]，落谷 亮[†]
[†]富士通研究所，[‡]富士通
{nisino@flab,hasimoto@aisys.se,ochi@flab}.fujitsu.co.jp

1 はじめに

社会の変化が激しい今日，それにともない新しい用語が使われるようになったり，あるいは既存の語に新しい解釈が与えられたりする．そのようなことから，人が用語を知り理解するための辞書を効率よくタイムリーに作成したいとか，企業内で使われる用語の定義などをコストをかけずに作成したいというような要望があり，テキストから新しい語をその語の意義とともに自動的に抽出することが求められている．このような語とその意義の抽出は，人のための辞書作成というだけでなく，様々な言語処理に利用するという点でも期待されている．我々は表層のパターンマッチング手法に基づいてテキストから用語とその定義文を抽出する実験を行った．本稿では，用語定義文を抽出する手法を述べ，実際の抽出結果を示し，用語定義文抽出における問題点を明らかにする．

2 用語定義文抽出の意義

テキストから言語知識を獲得しようという試みはこれまでにも数多く行われており，機械翻訳やかな漢字変換，情報検索などの応用に向けた知識（表記，品詞，意味分類，対訳語など）を獲得するための手法がいろいろ提案されている．例えば，大量のコーパスに対して，共起データを収集し，相互情報量などの統計量を用いてクラスタリングすることで類似語を集めることなどが行われている．しかし，このような統計処理にはいくつかの問題がある．一つは新語や専門用語のようなものは個々の単語の頻度はそれほど高くはないので，言語知識獲得に充分な統計量が集まらないという問題である．その結果，頻度の低い新語や専門用語に関しては期待したような知識を獲得できないこともある．また，意味分類とその語の用法とは違うという問題もある．たとえば，知識獲得の結果として，「歩」，「香」，「桂」，「銀」などは将棋の駒であるとか，「フライパン」，「包丁」，「皮むき器」などは調理器具であるとかというような知識が獲得できることを期待したいが，同じクラスの中に含まれるものであってもそれぞれ役割が異なるので共起する語も異なっている．例えば，「銀」は「引く」ことができるが「歩」は「引く」ことができないし，「包丁」で「切る」が「フライパン」で「切る」ことはない．その結果，共起情報を利用して統計処理をしても期待するような知識が得られないことも多い．そして，人が問題解決を図ったり知識を蓄えるという用途，あるいは将来の知識ベースを利用した高度な言語処理システム用の辞書作成という場合には，単にある分野における重要語句や専門用語を抽出したり，おおまかなカテゴリを求めるというのではなく，語句の実際の意義・定義を抽出したいという問題である．その場合には統計的処理だけでは不十分であり，言語的な処理がどうしても必要なのである．

我々は新語や専門用語はどこかで必ずその語の意味付けがなされているはずであり，それを利用しない手はないと考え，テキストから用語（被定義語）とその用語の解説（定義部）を抽出することを試みた．なお，抽出結果を言語処理システムが利用するには，定義部に関して何らかの知識記述言語に変換することも必要であろうが，これらの研究そのものは色々研究されているので，とりあえずは定義部は自然言語文そのまま抽出することとした．

3 用語定義文抽出の手法

用語定義文に注目した研究としては [1] があり，辞典の定義文を同義語文，内包的定義文，外延的定義文

に区別し，それぞれをパターン化し，文字列照合によって定義文を取り出している．しかしこれは辞典を対象として各エントリ語の定義文の構造化を試みたものであり，一般文から定義文を抽出しようとしたものではない．一方，本のような大きなサイズの一般テキストに対して索引語の重要説明箇所を求めようとした研究もある [2]。この論文では，重要語の出現密度分布を用いた手法を提案しており，表層パターンを利用する手法はパターンを網羅的に用意することが困難であると指摘している．ここでの指摘のように，確かに一般文書からその文書中の重要語に対して，それを最もよく説明している部分を取り出すのには表層パターンを利用するのは困難かもしれない．しかしこれは語そのものの定義を取り出そうというのではなく，ある語がその文書でどう扱っているかを抽出しようとしているからである．すなわち [2] では，例えば『中国とソ連』という本の中で「貿易」というキーワードに関して説明している部分を求めているのであり，一般に「貿易」という言葉がどういう意味なのかを求めているのではない．特定の文書中の重要語の説明部分というのではなく，専門用語や新語などの定義ということであれば，読者に正確に情報を伝達する必要上，明確な表現がなされているはずである．そうだとすれば，このような定義表現のパターンはある程度限定されていることが予想できる．

我々はこれまで新聞記事を対象として表層パターンに基づいて新製品販売や企業合併のような情報を構造化して抽出したり [3]，人物情報や企業情報を抽出する [4] ことが有効であることを示してきた．これらは読み手に新しい情報をきっちりと伝達するためには，読み手の深い知識を前提とせず，書き手は表層的な工夫によって新しい情報を正確に伝えるはずであるという仮定に基づいている．これらと同様に専門用語や新語の定義文の抽出においても，表層的な手がかりがおおいに利用できるであろうと考え，表層的なパターンマッチングに基づく手法を採用した．

4 定義文と主題文

我々は用語定義文を抽出するに際して，特定の用語に注目するというよりも，用語定義文であると思われる表層パターンに注目して用語定義に関する情報を抽出することにした．これにより，あらかじめ注目していなかった語に対しても用語の定義が取り出せ，新たな発見の支援にも役立つと考えられたからである．

用語定義文としては，「xx とは～である」，「xx というのは～である」というパターンや，「～を xx と呼ぶ」のようないくつかのパターンがあるが，今回は，<被定義語><被定義語提題表現><定義部><定義述部表現> というパターン，特に<被定義語提題表現>として典型的なものである「とは」に的を絞って考察・実験を行った．

「とは」は [5] によれば『「というのは」の意。定義・命題などの主題を示す。』とあるが，実際の文章中には他の「と」の用法と「は」の組み合わせとして定義・命題を示す文以外に様々な主題を示す用法が存在する．表層パターンが同じでも用語定義文ではない文を取り除くことが必要である．前記文献にしたがって「とは」の用法を類型すると次のようになる．なお，以下に示す例および実験結果は毎日新聞 CD-ROM'91～'97 年から抽出したものである．

1. 定義・命題の主題
   例）「ＧＡＯ」とは米国の会計検査院の略称
   　　カルネとは、フランス語で馬肉とかふしだらな女の意味。
2. 引用
   例）「沖縄の痛みを知れ」とは聞こえのよい言葉だ。
   　　「ご反応」とはいかにもこなれない日本語である
3. 特殊な具体的な事物に関する主題の提示
   例）裏書の堀丹後守とは直寄のこと
   　　女王とはあの＜美空ひばり＞さんのこと
4. 動作・作用・状態の内容の主題化
   例）党議決定とはみなせない。
   　　何とか交流を図ろうとはしている。
   　　夫婦とはいいものである。
   　　四番打者とはツライものである。
5. 動作・作用・状態の相手・共同者の主題化
   例）あの人とはもう十数年来のお付き合いだ
6. 比較の基準の主題化
   例）民主主義とは違った性格を持つ
   　　民主主義とは正反対に位置するものである
7. 転化する帰着点の主題化
   例）「優勝祝い」とはならなかった
8. 伝聞の主題化
   例）異常なことが起きたとは聞いていない
9. 副詞の主題化
   例）すんなりとはいかなかった
10. 並列の主題化
   例）規制緩和と既得権益とは対立用語である

5 用語定義文の抽出

単純に文字列の「とは」を抽出し，「ことは」や「もともとは」のような本来の「とは」と関係のないものを取り除いただけでは，非常に多くの非用語定義文が混在してしまう．「とは」の用法としては，定義を示すものと，主題化を示すものとに大別できるが，この主題化を示す「とは」を取り除く必要がある．そこで，これらの用法を観察すると，定義を示す場合には通常被定義語の意味内容が客観的に説明される．これに対して主題化を示す場合には取り上げた事物に対しての観察・思考・意向・決心・命名・言表などの精神作用や，相手・共同者を伴う動作・作用，比較，転化に関する表現が定義部に現れることになる．これらのパターンは前述した分類に沿って細かく分析・収集することも可能であるが，これと同時にある程度機械的に収集することを考えた．例えば，主題化された場合の述語は繰り返し出現すると考えられるので，「とは」ないし「とは、」に後接する高頻度の単語を調べれば，除去パターンとすることができるであろうと考えた．実際に図1にこれらの単語を示す[1]．

7461	いえ（る）	894	「
7294	思（う）	841	限（る）
4301	言（う）	805	い（う）
3927	別	747	裏腹
3018	違（う）	730	。
1589	考え（る）	720	全く
1347	異な（る）	655	」
1333	何か	646	無縁
1320	何	635	無関係
1010	対照的	568	関係な（い）

表 1: 「とは」に後接する頻出単語

この結果を眺めると，“「” を除いて，どれも排除すべきパターンと考えることができる．以下，下位の方まで多くの排除すべき述語を収集することができた．

このような除去パターンによってあきらかな非定義文を取り除いてもまだ多くの非定義文が残る．そこで，実際の定義文をもう少し詳しく分析するために，この中から定義文可能性の高いパターンも収集することにした．具体的には，<定義述部表現> が「である」であるものだけを抽出した．その結果の一部を表 2に示す．

[1]これは毎日新聞 CD-ROM 7 年分から単語としての「とは」を含む約 76,000 件の文からの抽出結果である

ここには本来の辞書の定義・命題だけでなく様々なバリエーションがあることがわかる．長尾は専門用語をどのように定義するかについて整理し，定義文を同義語定義文，内包的定義文，外延的定義文などに類別している [6]．これは辞書定義文を定義の仕方によって類別したものであるが，一般文からの定義文の抽出ではこのような辞書的な定義の仕方とは別に，どんな状況での定義なのかという観点からの類別方法も必要である．そこで，得られた定義文を類別すると以下のようなものがあることがわかる．

事象・概念の説明 まだポピュラーになっていない新しい事象や概念に対して，それらの意味が説明される．一般の辞書記述文である．

例）「インフレ」　通貨の値打ちが下がること
「カオス」　人知を超えた混乱状態のこと

名称の説明 単なる名称そのものの説明．例えば，言語の違いや略称であるために解説が加えられる．「～語で～のこと」，「～の略」などのパターンとして抽出できる．

例）「エイズ」　後天性免疫不全症候群の略
「エクウス」　ギリシャ語で馬のこと

解釈文 特定の人物・組織が，抽象的概念のようなものに対して，従来にない新しい意味や特殊な意味づけをして，自分なりの解釈を述べたものである．解釈文は特定の人物の組織による解釈を示したものであるので，誰の解釈なのかが重要な意味を持つ．逆に言えば，解釈文には通常発言者も同時に示されていると考えられる．

例）「天才」　わずかに我々と一歩を隔てたもののこと（芥川龍之介）

具体的な事物に対する主題化 一般的な定義ではなく，特定の事物の説明である．「～の言う～とは」など被定義語に限定がついていることもあるが，定義と区別のつきにくいものもある．

例）「野蛮人」　セルビア人のこと

引用文 特定の人物・組織の発言に対する解説である．

例）「日本国は女の地獄なり」　福沢諭吉の名言
「調査なくして発言なし」　かの毛沢東の数少ない名言

「まともな仕事」	安定した収入があり、家族を養い、定年近くには家も建つ、そんな職業につくこと
「やつら」	アナタのこと
「ゆとりある生活」	レジャー産業振興の意味
「アカシ」	明石康・旧ユーゴ問題国連事務総長特別代表のこと
「アルマアタ」	カザフ語で「りんごの父」という意味
「アンブレラ」	その字のとおり傘のこと
「インフレ」	通貨の値打ちが下がること
「エイズ」	後天性免疫不全症候群の略
「エクウス」	ギリシャ語で馬のこと
「エチオピア」	「日に焼けた人」という意味
「エッセ（ｅｓｓｅ）」	「生きる」を意味し、「ベネッセ」は「良く生きる」という意味の造語
「カオス」	人知を超えた混乱状態のこと
「カストラート」	その昔、去勢によって大人になってからもボーイソプラノの純な声を保ち続けた歌手のこと
「キムタク」	男性アイドルグループＳＭＡＰの木村拓哉クンのこと
「クーリングオフ」	もともと「冷却期間」を意味する言葉

表 2: 定義文抽出例

上の類別の分析から名称の説明など，定義文の確度の高いパターンとして収集することができるものもありそうである．これらは，<定義述部表現> を「～である」以外にも拡張することで，より多くの定義文を集めることができるであろう．

一方，この中には引用文や具体的な事物に対する主題化など用語の定義ではないものも含まれている．引用文には単独で抽出しても価値のないものも多いが，しかし，名言などは抽出しておいてもいいだろう．具体的な事物に対する主題化は，本来の定義文と境界が微妙であり，これも一部は抽出する価値がありそうである．ただし，単純に用語定義として抽出するのではなく，状況説明文とともに抽出することが必要そうである．

6 今後の課題

今回，用語定義文抽出に関して，新聞記事を対象として簡単な実験を行なった．抽出した結果を見ると，単に用語の定義だけでなくいろいろな種類のものが抽出されている．本来の辞書的な用語の定義だけでなく引用や具体的な事物の主題化などでも抽出しておく価値のあるものは多いようである．ただし，その際，どういうものを抽出する価値があるのかを明らかにしておくことは必要であろう．また，このようなものを抽出する際には，単純に用語とその定義の抽出だけでなく，その状況（発言者や，指している具体物など）も含めて抽出するようにしておくことも必要であろう．

用語とその定義文を抽出する手法としては表層のパターンマッチング手法に基づいたものであり，用語定義文であるかどうかを判定するための根拠は個別に求めてそれをパターンとして個別に利用している．今後としては，被定義語提題表現，被定義語引用記号，定義述部表現，被定義語の種類，定義部の種類に基づいて総合的な評価をするために，その文章が用語定義文の可能性が高いかどうかを示す何らかの指標を導入することを考えるべきであろう．

今回，新聞記事を対象としてパターンを蓄えたが，これらは語を定義しているという意味で，他の情報源（論文，特許，Web 文書など）でも共通に利用できるものであると考えている．今後，社内文書や特定の専門分野での専門用語とその説明文の抽出という応用を考えている．

参考文献

[1] 黒橋禎夫, 長尾眞, 佐藤理史, 村上雅彦：専門用語辞典の自動的ハイパーテキスト化の方法，人工知能学会誌, Vol. 7, No. 2, pp. 336–345 (1992).

[2] 黒橋禎夫, 白木伸征, 長尾眞：出現密度分布を用いた語の重要説明箇所の特定，情報処理学会論文誌, Vol. 38, No. 4, pp. 845–854 (1997).

[3] 西野文人, 落谷亮, 木田敦子, 乾裕子, 桑畑和佳子, 橋本三奈子：トップダウンなパターン解析に基づく情報抽出，情処研報, NL124-13, pp. 95–102 (1998).

[4] 西野文人, 落谷亮：新聞記事からの人物・企業情報の抽出，情処研報, NL127-17, pp. 125–132 (1998).

[5] 国立国語研究所（編）：現代語の助詞・助動詞，秀英出版 (1951).

[6] 長尾眞：辞典形式での専門分野の知識の体系的構成法，人工知能学会誌, Vol. 7, No. 2, pp. 320–328 (1992).