格構造解析を用いた新語抽出支援システムに関する研究

1N－5

阿部さつき[†1]
NTTアドバンステクノロジ株式会社[†1]

津田和彦[‡2]
筑波大学大学院 経営システム科学専攻[‡2]

1. はじめに

自然言語処理を用いたシステム、例えば、文献検索や機械翻訳システムなどでは、一般的に形態素解析を利用する.

形態素解析の解析精度はその辞書の品質に大きく左右される. 大量情報時代である昨今はこれまでになく次々に多くの新語が生まれてくる.

この新語をすばやく的確に辞書に登録することは重大な課題となっている.

本稿では、格構造解析を用いて日本語の形態素解析用の辞書に新語を効率よく収録するための手法を提案する.

2. 従来の手法

従来の代表的な新語の抽出方法には、以下の３つの方法がある.

(a) 形態素解析の結果、解析エラー（未知語）となった箇所を新語の候補とする.

(b) 既存の新語辞典（書籍, 電子ブック）等を元にする.

(c) 字面の統計情報（n-gram 等）を用いる.

形態素解析は、字種の区切れと隣り合った単語の連接関係を用いているものが多いため、偶然連接する単語が並んでいる場合、それに引きずられて解析エラーを起こす場合がある. (a)は、このようなノイズ（解析エラー）が多く、形態素解析で未知語にならない新語は見落とされる恐れがある. (b)は書籍に出版されるまでタイムラグがあり、即時性に欠ける. (c)は大量の文書がないと適切な単語の抽出ができない. 以上のようにいずれの方法にも問題がある.

A Study of a News Words Extracting Support System using Care Structure Analysis
Satsuki Abe
NTT Advanced Technology Corp.
90-6 Kawakami-cho, Totsuka-ku, Yokohama, JAPAN

3. 格構造解析を用いた新語候補の抽出

我々は、日本語の文章の中に多少見知らぬ単語（新語）があっても、おおよその文意をとることができる. これは、人間は文章を解読する際に無意識に動詞を中心とした単語の格に着目して、文の骨格（フレーム）を把握し、骨格内の要素に未知語が存在しても、不確定要素を予測できるからである.

本稿では、この人間が新語を認識する仕組みに着目し、文の格構造から主語や目的語となる単語を新語の候補とする方法を提案する.

3.1 前処理（単文への分割）

まず、格構造ルールのマッチングを効果的に行うために、全ての文章を単文に分割する. 尚、単文に分割し、用言をマーキング必要があるため、格構造ルールを使用する際に前処理に形態素解析（今回はALTJAWS を使用）を行う。

以下の手順で文章を単文に分割する.

① 日本文に対して、形態素解析を行う.

② 助詞相当語を意味的に等価の助詞などに置き換える. [例]「からは」→「は」

③ 不要語（副詞、連体詞、名詞に修飾する形容詞・形容動詞等）の削除を行う.

④ 用言にマーキングを行う.

⑤ マーキングした用言をキーにして、文を単文に分割する.

尚、埋め込み文はそのまま単文に分割した. ただし、埋め込み文が形式名詞にかかる場合は、形式名詞を含む文節は削除する.

3.2 新語候補の抽出

以下の手順で新語候補を抽出する.

① 格構造ルールのマッチング

3.1 で生成した単文に存在する各々の用言の格構造ルールをマッチングする[1].

② 新語候補のピックアップ

マッチングした結果、主語、目的語等にあたる文字列をピックアップする.

③ 既登録語の排除

ピックアップした単語のうち既存の辞書にあ

る単語をリストから削除する.

尚、名詞句，だ文は対象外とした.

3.3 実験の結果

日経産業新聞の新製品情報の 108 文をサンプルデータとし、4 の手法の検証を行った.

《分析対象》日経産業新聞（1999.2.10～2.14）

新製品 NEWS＝10 記事（108 文）

(1) 対象文の単文分割

① 対象文章の単文化 108 文→233 文

② 対象文の絞込み* 233 文→178 文

*①から名詞句＋だ文(55 文)を引いた文数

(2) 新語の抽出

① 新語候補のピックアップ…338 件

② 既登録語の排除

新語候補＝198 件（述べ）

-181 件（異なり）

表 抽出された新語の例

データ	用言	格1	格2
松下電器産業	発売する	が	
同社	発売する	が	
記憶容量	発売する	が	
コンパックコンピュータ	発売する	が	
日本アイ・ビー・エム	発売する	が	
スマートメディア	発売する	で	
3月上旬	発売する	に	
モニター一体型DVD(デジタル・ビデオディスク)プレーヤー「DVD—L50」	発売する	を	
「M—32P」	発売する	を	
「MAFP—2」(希望小売価格1万2000円)	発売する	を	
新型「アルマダ1500c Basicアドバンテージモデル」4機種	発売する	を	
新機種	発売する	を	
上位機種「プロライアント5500」「同6000」「同6500」「同7000」	発売する	を	
2月下旬	発売する		、
9日	発売する		、
32メガバイト	発売する		の
ノート型パソコン	発売する		の
基幹業務システム向けパソコンサーバー	発売する		の
中型サーバー	発売する		の
「AS／400eサーバー」	発売する		の

4. 提案手法の検証

4.1 効果が見られた点

・格構造を用いることにより、形態素解析の結果に左右されにくい.

［例］スマート／メディア

・複数種類の字種が使用されている新語が収集可能である.

［例］DSTN（デュアル超ねじれネマティック）

・複合単語の抽出、外来語、複合語、略語等が抽出可能である.

［例］レセプタクル，省スペース，ＵＳＶ規格，ＳＭＰ（対称型マルチプロセッシング）

検証の結果、いくつかの改善すべき点はあるが、新語の抽出に一定の効果が見られた.

しかし、以下の問題も見られた.

(a) 対象文に格要素と同じ助詞が含まれる と不適切な箇所で切ってしまう.（5 件）

［例］5 分の／1

(b) 格構造ルールになかった格（で，から等の格=27 件）や用言直前の副詞的用法の名詞等(8 件)が抽出できなかった.

4.2 効果

3.3(1)②で縛り込まれた対象文から、抽出すべき単語[3.3(2)①＋4.1(a)(b)＝378 件]のうち、抽出の成功率は 89.4%であるが、4.1(b)の一部(27 件)はデフォルトの格要素を追加すると抽出される見込みが高い. これを含めて考えると成功率は 96.5%に上昇すると考えられる.

5. おわりに

今回行った実験では、サンプルとした分野が偏っている可能性がある. 特に、埋め込み文や複文に関しては、他の用言と同様に一律に分割し、特に大きな問題は発生しなかった. 今後、他の分野での検証が必要である.

また、新語の検出率を上げるためには、単文に用言が出現しない場合のルールを追加し、対象となる文の範囲を広げる必要がある.

また、新語抽出の結果を同一の用言から抽出された新語と格の組み合わせに着目して、ソーティングすると、意味的なまとまりが得やすいという利点も見られた.

《参考文献》

[1] 池原他. “日本語語彙体系 5：構文体系”. 岩波書店, 1997.

[2] 松本他. “岩波講座言語の科学 3：単語と辞書”, 岩波書店, 1998.

[3] 影山太郎. “動詞意味論——言語と認知の接点”, くろしお出版, 1996.

[4] 宮島達夫. “語彙論研究”, むぎ書房, 1994.

[5] 影山太郎, 由本洋子. “語形成と概念構造”, 研究者出版, 1997.

[6] 影山太郎. “文法と語形成”, ひつじ書房, 1993.