タグ無しコーパスからの複合語データの自動抽出

5B－4

細田 春美　相川 勇之　鈴木 克志
三菱電機（株）情報技術総合研究所

１．はじめに

日本語処理における複合語の解析は、困難な問題であることが従来より指摘されている。これは、複合語は単語を組合せることで無限に生成でき、すべてを辞書に登録することは不可能であるため、十分な解析精度を得ることができないからである。

複合語の数の多さを解決する方法として、複合語を構成する単語を概念に置換し、その概念間の共起情報（複合語概念対）を用いて複合語を解析する方法がいくつか提案されている[1][2][3][4]。

複合語解析において、木構造を成す概念情報を利用する場合、下位の階層を用いた方が細かな意味の違いを解析に反映できる。しかし実際には、大量の複合語概念対の獲得は困難なため、解析時には上位の階層の概念情報しか利用できない[1]。高精度の複合語解析の実現には、下位の階層の概念までを含む大量の複合語概念対を獲得する必要がある。

本稿では、タグ付きコーパスから抽出した複合語概念対を利用して、タグ無しコーパスから複合語概念対を自動抽出する方法を提案する。

２．複合語概念対抽出の問題点

近年、電子化文書の入手は容易になってきた。しかし、ほとんどは形態素区切りや品詞などの情報のないタグ無しコーパスである。タグ付きコーパスは、生成が困難なため、電子化文書のうちのごくわずかである。そこで、大量に存在するタグ無しコーパスを形態素解析して、複合語概念対の抽出対象とする。

ところが、タグ無しコーパスからの複合語概念対の抽出には、形態素区切りの曖昧性と概念の曖昧性という２つの曖昧性の問題がある。

Automatic Compoud Noun Extraction
from Untagged Corpus
Harumi HOSODA, Takeyuki AIKAWA, Katsushi SUZUKI
Mitsubishi Electric Corporation.
5-1-1 Ofuna, Kamakura, Kanagawa 247, JAPAN

形態素区切りの曖昧性とは「高/価格」「高価/格」のように、形態素区切りの候補が複数あるものである。[1][4]は、四文字漢字語の約７割が２つの二文字漢字語から構成されることを利用して、四文字漢字語から複合語の共起情報を抽出し、形態素区切りの曖昧性に対処している。しかし、四文字漢字語の約３割は二文字漢字語から構成されるものではないため、高精度の共起情報を抽出できない。

また、概念の曖昧性とは、複合語を構成する単語が複数の概念を持つ場合に生じる問題である。例えば、「相手」という単語は、<敵>、<相棒>の２種類の異なる概念を持つ。「対戦相手」の「相手」は、<敵>という概念であり、「対談相手」の「相手」は、<相棒>という概念である。ところが「相手国」の場合は、<敵>、<相棒>のいずれの概念にもなりえるため、曖昧性が生じる。[1]では、概念が一意である複合語概念対のみを抽出対象とすることによって、この問題を回避している。

３．問題点への対処と複合語概念対の抽出

２つの曖昧性の問題に対処し、タグ無しコーパスから複合語概念対を抽出する。ただし、抽出対象は２語の名詞類（名詞、サ変名詞）からなる名詞複合語に限定する。３語以上からなる複合語は、更に構造の曖昧性の問題が生じるので扱わない。

３．１．形態素区切りの曖昧性への対処

タグ無しコーパスを形態素解析した結果から複合語を抽出する際に、コーパス中に単独で出現する複合語のみを抽出する。つまり、複合語の前後が名詞、サ変名詞以外の語であるような複合語のみを抽出する。

次に、形態素区切りや品詞に曖昧性のあるものと、複合語を構成する単語の文字列長がいずれも１であるものを排除する。このように処理対象を限定することにより、形態素区切りの曖昧性を回避する。

3．2．概念の曖昧性への対処

単語の概念として EDR 電子化辞書[5]の概念識別子（概念 ID）を使用する。EDR 電子化辞書の概念 ID は木構造を成し、多重継承を許している。

3．2．1．タグ付きコーパスからの概念対抽出

概念の曖昧性に対処するための知識源として、形態素解析済みのタグ付きコーパスから抽出した複合語概念対（第１次複合語概念対）を用いる。

タグ付きコーパスとして、21 万文の EDR コーパスを使用し、次の手順で第１次複合語概念対を求めた。

(1)コーパス中に単独で出現する複合語を抽出
(2)複合語を構成する２単語を各の概念 ID に置換
(3)複合語概念対の頻度を算出

EDR コーパスから得た第１次概複合語概念対は、概念組合せの異なり個数で約 54000 種類であった。

3．2．2．タグ無しコーパスからの概念対抽出

複合語を構成する２単語の概念 ID が、いずれも一意である場合は、曖昧性の問題は生じない。したがって、これは[1]と同様に、新たに得られた複合語概念対（第２次複合語概念対）とする。

さらに、概念 ID が複数ある場合に対しては、概念 ID の組合せ候補のうち、第１次複合語概念対に存在する候補のみを第２次複合語概念対とし、頻度を更新する。このとき頻度は、候補数で分配した値を加算する。第１次複合語概念対に存在しない候補は、曖昧性が高いので排除する（図１）。

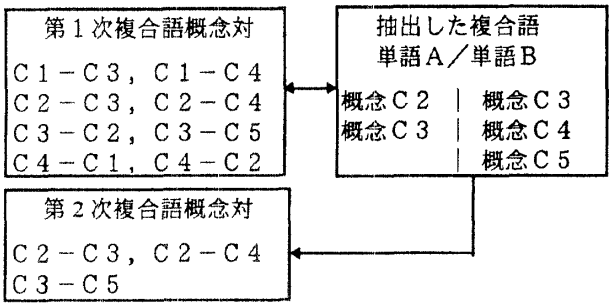

図１ 概念 ID が複数ある場合

このように、概念の曖昧性のない第１次概複合語概念対を参照して概念 ID の組合せを決定することにより、概念の曖昧性の問題に対処する。

4．EDR コーパスと新聞記事による実験結果

タグ無しコーパスとして「朝日新聞」記事１年分（朝日新聞社提供、約 78MB）のデータを使用し、複合語概念対の抽出実験を行なった。形態素解析は、JUMAN2.0[6]の形態素解析エンジンに約 11 万語の辞書を結合したシステムを用いた。

その結果、新聞１年分のコーパスから約 26000 種類の複合語概念対を新たに獲得することができた。また、第１次複合語概念対を用いることにより、概念の曖昧性に対処し、頻度の更新をした複合語概念対は、約 10000 種類であった（図２）。

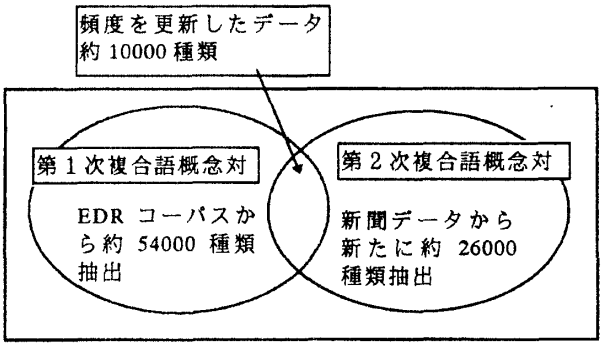

図２ 抽出した複合語概念対

5．まとめと今後の課題

タグ付きコーパスから抽出した複合語概念対を知識源として用いて、タグ無しコーパスから複合語概念対を自動抽出する方法を提案した。

EDR コーパスから抽出した約 54000 種類の複合語概念対を知識源とし、新聞１年分のコーパスから複合語概念対を抽出した結果、新たに約 26000 種類を獲得し、約 10000 種類の頻度情報の強化をすることができた。今後は更に大量のデータを獲得すると共に、概念階層を考慮した複合語解析を検討する。

参考文献

[1] 小林，山本，徳永，田中：語の共起を用いた複合名詞の解析，情報処理学会研究報告 自然言語処理 Vol.94，No.47，pp.1-8(1994)．
[2] 太田，宮崎：複合語用例データベースを用いた複合名詞の構造的曖昧さの絞り込み，情報処理学会 第53回全国大会 No.2-9(1996)．
[3] 宮崎，池原，横尾：複合語の構造化に基づく対訳辞書の単語結合型辞書引，情報処理学会論文誌 Vol.34，No.4，pp.743-754(1993)．
[4] 筏井，横山，佐久間：シソーラスを用いた複合名詞の生成・解析，情報処理学会 第52回全国大会，No.3-7(1996)．
[5] 日本電子化辞書研究所：EDR 電子化辞書仕様説明書，(1993)．
[6] 松本，黒橋，宇津木，妙木，長尾：日本語形態素解析システム JUMAN 仕様説明書 Version2.0，(1994)．