4P−2

最長一致法と接続表を用いた形態素解析による語彙情報の決定方式

上田一人、瀧口伸雄、小谷善行
東京農工大学工学部電子情報工学科コンピュータサイエンスコース

1．はじめに

本研究は、日本語の文章を形態素解析する事を目的とする。

自然言語処理では、まず最初に文字列を単語に分かち書きしなければならないが、日本語は文章中に単語の切れ目となるはっきりした目印が存在しないため、単語の切り出しのための形態素解析が必要である。

本研究では、右方向最長一致法[2]を用いた形態素解析を行ない、単語の切り出しを試みる。単語辞書のみを用いた形態素解析では、次のような問題が生じる。

(1)文字列を切る位置を誤る。例えば、助詞である文字から始まる単語が辞書に存在すると、助詞は切り出しにくい。

(2)多品詞語の場合は、どの品詞を選択するべきか分からない。

(3)日本語は複合語を作りやすく外来語も多く存在するため、すべての単語を辞書に登録するのは困難である。そのため数多くの未知語が存在する。

これらの問題を解決するため、接続表を使用する。

接続表は文献から得ずに、新聞の文章を正しく形態素解析して作成した。

また、これまでの研究では二つの品詞間の接続表（二連接続表と呼ぶことにする）が使われてきたが、本研究では、三つの品詞間の接続表（三連接続表と呼ぶことにする）を使用し、二連接続表と比較する。

2．文字列長拡張型右方向最長一致法

右方向最長一致法には、文字列を長くしながら辞書引きしていく方法[2]と、文字列を短くしながら辞書引きしていく方法[5]があるが、辞書に存在する単語の長さが短いものが多いときは前者の方が効率がよい。そこで、本研究では前者の方法を用いるが、以下これを「文字列長拡張型右方向最長一致法」と呼ぶ。

この方法は、まず一文字目を辞書で引き、その文字で始まる単語があれば、二文字目までから始まる単語を辞書で引く。これを辞書に文字列とマッチする単語がなくなるまで行う。

次に、文字列長拡張型右方向最長一致法での未知語の切り出し方法について述べる。現在切りだした文字列から始まる単語が辞書に存在していない場合、それまでに確認した単語の最も長いものを出力して、残りの文字列について同様の処理をする。もし、確認した単語がなければ、文字列のはじめの文字を未知語素とする。

例：自然言語理解の研究をする
（自然言語処理，言語は既知語で、その他はすべて未知語とする。）

「自」　　「自」から始まる単語が辞書に登録されているので処理を続ける。
「自然」　　「自然」から始まる単語が辞書に登録されているので処理を続ける。
「自然言語」　「自然言語」から始まる単語が辞書に登録されているので処理を続ける。
「自然言語理」
ここで、「自然言語理」から始まる単語が辞書に存在しないので、今までに確認した一番長い単語を出力する。

しかし、この場合は確認した単語がないので、「自」を未知語素として、残りの「然言語理解の研究をする」から処理を続ける。
「然」
ここで、「然」から始まる単語が辞書に登録されていないので、今までに確認した一番長い単語を出力する。

しかし、この場合は確認した単語がないので、「然」を未知語素として、残りの「言語理解の研究をする」から処理を続ける。
「言」　　「言」から始まる単語が辞書に登録されているので処理を続ける。
「言語」　　「言語」から始まる単語が辞書に登録されているので処理を続ける。この文字列そのものが、単語として登録されていることを確認する。
「言語理」
ここで、「言語理」から始まる単語が辞書に登録されていないので、今までに確認した一番長い単語を出力する。ここでは、「言語」を出力し、残りの「理解の研究をする」から処理を続ける。

・
・
・

3．接続表

接続表とは、それぞれの品詞がお互いに接続可能かどうかを示した表である[2]。次の三つの場合に効果がある。

① 最長一致法による解析で図3.1のような解析結果となった場合に、「バ行五段連体」が「句点」に接続しないという情報がもしもあれば、解析結果の「遊ぶ」の品詞を「バ行五段活用終止」に絞り込める。

Identifying Vocabulary Information by Morphological Analysis on the Basis of Maximum Length Matching and Connection Table.
Kazuhito Ueda, Nobuo Takiguchi, Yoshiyuki Kotani
Tokyo University of Agriculture and Technology

（例文）

私はこれらのゲームで遊ぶ。

（解析結果）

*単語	*品詞
私	名詞
はこれ	名詞
れ	助動詞未然，助動詞連用
ら	接尾語
の	格助詞
ゲーム	未知語
で	格助詞
遊ぶ	バ行五段活用終止，バ行五段活用連体
。	句点

図3.1　接続表の使用例1

二番目に図3.1の解析結果で、「はこれ」の部分を「は」と「これ」に分けたい場合、「名詞」と「助動詞」は接続しないという情報がもしもあれば、切り間違いを直すことができる。

三つ目に未知語の品詞を決定するのに使用する。図3.1の解析結果で、格助詞の後ろに接続できる品詞の集合ををＨ１、格助詞の前に接続できる品詞の集合をＨ２とすると、「格助詞」と「格助詞」にはさまれた未知語「ゲーム」の品詞はＨ１とＨ２の積集合となる。この方法では、異なった文章に出現する同じ未知語を、繰り返し解析することにより未知語の品詞を徐々に絞り込むことができる。

本研究では、接続表を新聞の文章を正しく形態素解析したものから獲得して、それをさらに同じ新聞の他の文章に適用した。これは、文献から品詞の接続関係を調べた場合、少しでも接続するものは「接続可」となってしまい、接続表が冗長になってしまう。すると、品詞の絞り込みが効果的に行えなくなる。そこで、実際の文章から接続表を獲得する。これは、ある種類の文章（新聞の時事面）では、特徴のありそうな文章が多くみられるので、品詞の接続関係にも特徴がみられるのではないかという考えを用いたものである。

接続関係は、２つの単語間の接続（二連接続表）だけでなく３つの単語間の接続関係（三連接続表）についても調べる。

例えば、図3.4のような品詞をもつ三つの単語が並んでいる場合、図3.5の接続許可情報だけでは３つ目の単語の品詞は絞れないが、図3.6の許可情報があれば三つ目の単語の品詞を「形容動詞語幹」に絞れるわけである。

［名詞］［助詞］［形容動詞語幹，名詞］

図3.4　三つの単語の並び
（三つ目の単語は二つの品詞のものがある）

名詞　－　助詞
助詞　－　形容動詞語幹
助詞　－　名詞

図3.5 二連接続の許可情報

名詞　－　助詞　－　形容動詞語幹

図3.6 三連接続の許可情報

三連接続表は二連接続表と同様に、先ほどの新聞の文章を正しく解析したものを用いれば、作成することは容易である。図3.7に三連接続表の構成をヒストグラムに示す。データは、92文3956単語を使用した。横軸は、同じ接続関係が現れた回数、縦軸はその頻度を表している。両軸ともに対数である。

最も多く現れた接続関係は、

格助詞　－　名詞　－　格助詞

で3772個中78回であった。

二連接続表でも同じような傾向のグラフができるが、一回しか現れないものの数が三連のばあい927個なのに対して二連では247個となっている。少数回しか現れないものは他の文章に当てはめにくいので、この点が三連接続表の欠点である。

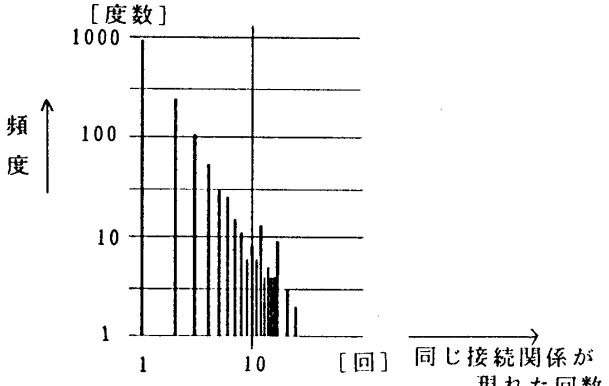

図3.7 三連接続表の構成

4．まとめ

最長一致法と接続表を用いて行う形態素解析において、新聞の文章から獲得した三連接続表を使用する方法を示した。

実験の結果、接続表に関しては、二連接続表を使用した場合よりも、三連接続表を使用した場合の方が、、単語の切り間違いが少なくなることが確認できた。接続表は簡単な構文解析とみなせるが、調べる接続関係の数が多くなればなるほど、構文解析に近づくと考えられ、より精度の高い解析結果が得られる。また、これは未知語の品詞を決定する際にも、非常に有効であることを確かめた。

5．参考文献

[1] 河上芳輝，形態素解析による辞書学習，東京農工大学工学部数理情報工学科卒業論文，1990
[2] 田中穂積，自然言語解析の基礎，産業図書，1989
[3] 芳賀綏，新訂日本文法教室，教育出版，1982
[4] 築島裕・白藤禮幸，新編　国語の文法　改訂新版，明治書院，1982
[5] 西村恕彦，電子技術総合研究所研究報告－機械翻訳プログラムの作成，電子技術総合研究所，pp.68-78，1970