# 日本文における共起情報を用いた未知語検索

2B−6

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

## 1．はじめに

本研究は、形態素解析において未知語を正しく学習することを目的とするものである。

日本語を形態素解析する場合には単語辞書が必要であるが、全ての単語を辞書に登録するのは非現実的であるので、未知語というものが存在する。前回の発表[6]では、形態素解析において未知語の解析を行なったが、これは単語に含まれる文字が辞書に存在しない場合の処理であった。本稿では、単語に含まれる文字が辞書に存在しないために出現する未知語だけでなく、単語に含まれる文字が辞書にある場合に出現する未知語の検索について、共起情報を用いて解析する。

## 2．従来の未知語検索方法

未知語の定義は、「既知語でない文字の集まり」としている。[1]

つまり、文字列Sが

S＝A　b　c　d　E

（A，Eは既知の単語，b，c，dは文字）

となっているとき，b，c，dの組み合わせの中に既知語がなければb，c，dを一つの単語として、未知語と判断する。例えば、

S＝「長官ワインバーガー氏が来日する。」

のとき、「長官」「氏」が既知語であり、「ワインバーガー」の中に既知語が含まれなければ「ワインバーガー」は未知語とした。

## 3．従来の未知語検索方法の問題点

2．で示した文字列Sの例では、未知語は「b，c，d」であり、実際にそうだとする。しかし、「b，c，d」の並びの中に既知語が含まれていた場合は、「bc、d」のように未知語が分割されてしまう場合がある。例えば、2．のSで、「ワイン」が既知語である場合は、未知語である「ワインバーガー」は「ワイン」と「バーガー」に分割されてしまう。このように間違った未知語が学習されてしまう可能性は、形態素解析においては、多々あることである。

S＝「長官　ワイン　バーガー　氏　が　来日　する。」

名詞　未知語

## 4．問題点の回避

### 4．1　分割された片仮名未知語の結合

前回の発表では、この問題の内、片仮名文字列が分割される場合については、次のような規則を作成することにより対応した。

規則Ⅰ：未知語である片仮名の単語のとなりに片仮名の単語が存在した場合は、二つの単語を一つの未知語とする。

実際に前回の発表では、時事新聞でこの規則を用いた解析を行ったところ、すべての片仮名未知語がうまく学習された。これは、片仮名の単語は、名詞や形容動詞語幹、サ変動詞であることが多く、長い単語でも一つの単語とみなして差し支えないものが多いと思われるからである。

また、ここでは片仮名の場合について述べたが、数詞や記号などの場合についても、同じ方法で良い成果が挙げられる。

### 4．2　共起情報を用いた未知語の検索

4．1では、「片仮名の文字列がとなりあって出現する」という共起情報を用いた。本研究では、これを他の文字種（漢字、平仮名）にも拡張する。この時使用する情報は、次の2つである。

①文字の種類（漢字，平仮名，片仮名）

②文字の長さ

つまり、どのような文字種の文字で、長さがいくつのものが、前後の文字列と共起するかを調べることによって、どの部分が正しい未知語なのかを調べるのである。

### 4．3　文字種別の対策

共起情報を使用する場合に、文字種ごとにある程度の知識が存在する。

### 4．3．1　片仮名

規則Ⅰを用いることにより、片仮名に関しては良い結果が得られる。この規則では文字列が未知語にならなければならないが、未知語が「既知語」「既知語」の組み合わせで分割されることもある。例えば次のような場合。

「ソルトレイク」⟶「ソルト」「レイク」

既知語　既知語

「ソルト」「レイク」それぞれに意味があるがこの場合は「ソルトレイク」でないと意味がない。しかし、

---

Searching Unknown Word by
Information of Concurrency.

Kazuhito Ueda, Nobuo Takiguchi, Yoshiyuki Kotani
Dept. of Computer Science, Tokyo University of Agriculture and Technology

「ハード」　＋　「コピー」
「パターン」　＋　「マッチング」
「トイレット」　＋　「ペーパー」

のように、切り離して考えてもある程度の意味を持つ単語が存在する。このように、ある程度意味が存在する場合には一つ一つの単語の長さが長いものが多いようである。そこで、長さをコストとしたコスト最小法による解析が可能であると考えられる。

### 4．3．2　漢字

日本語では漢字何文字かで構成される漢字熟語が多く存在する。それらは、膨大な数なので未知語であるものも多い。

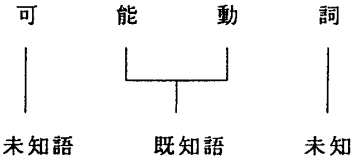
未知語はその中に既知語を含まないのだが、漢字何文字かで構成される未知語の場合、辞書に一文字の漢字から構成される単語が多く含まれていると、バラバラに分割されてしまう。このようになった場合も片仮名を接続する場合のように単純に接続するだけでもうまく行きそうな例が見受けられるが、うまく行かないものも相当数現れる。

そこで、複合漢字の結合パターン[7]を解析に利用する。複合漢字の結合パターンとは、漢字熟語のどの部分で単語が分割できるかを表したものである。例えば、３文字の漢字熟語の結合パターンには、次のようなものがある。

①□□＋□型　　文化－人　機関－車　動物－園
②□＋□□型　　核－実験　党－本部　女－生徒
③□・□・□型　松－竹－梅　市－町－村

①②③に存在する熟語にはそれぞれ割合がある。　①型と②型では、約５倍程度①型に属するものが多い。③型に属するものは、きわめて少ない。これらの割合を、コストとした、コスト最小法により誤って未知語として学習されるものを防ぐことができる。

例えば、「可能動詞」が次のように解析されたとする。




このとき、４文字熟語のコストが、「２文字＋２文字」の方が「１文字＋２文字＋１文字」よりもかなり小さくなるようならば、二つの二字漢字熟語から四文字の熟語が作られているのではないかと推測することが可能である。

### 4．3．3　平仮名

日本語は、漢字かな混じり文で書かれている。平仮名が使われるのは、

①動詞，形容詞の送りがな
②助詞，助動詞，接続詞，感動詞，接辞
③名詞

が良く使われる。このうち、②は辞書に多くが登録されていて、しかも単語の長さが短いため、①③が未知語になる場合に②の単語によって未知語が分割されてしまう場合がある。例えば次の場合。

S＝　政治家　の　う　わ　さ
　　　　　｜　｜　｜
　　　　助動詞　未知語　接辞

助詞、助動詞、接続詞などが並んで平仮名が連続する場合にはあるパターンが見受けられる。そこで、③の名詞がすべて平仮名で形成される場合の判断に品詞の共起パターンを利用する。

## 5．まとめ

以上の方法で、正しい未知語の学習を試みる。まず、

・片仮名文字列の長さ
・複合漢字の結合パターンと結合のコスト
・平仮名文字列の結合のコスト

について正確に調査する必要がある。

未知語を検索する場合、係り受けや、品詞などの知識を利用することは難しいと思われるので、知識を必要としない文字種を使って未知語を解析することは妥当であると思われる。

片仮名の文字列については、これまでもうまく学習されてはいたが、単語が長すぎる傾向があった。長い単語一つを学習することも意味があることだが、辞書が大きくなりすぎる可能性があり、また、汎用性のある辞書を作成するのに妨げとなるだろう。この研究により、実用的で汎用的な辞書を作成したい。

## 参考文献

[1] 河上芳輝，形態素解析による辞書学習，情報処理学会（平成３年前期）全国大会講演論文集，1991
[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
[6] 上田一人，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，情報処理学会（平成４年前期）全国大会講演論文集，pp.3-171,172，1992
[7]水谷静夫　他，「文字・表記と語構成」，朝倉書店，1987