1C−6

形態素解析による辞書学習

河上芳輝　小谷善行　西村恕彦
東京農工大学　工学部　電子情報工学科

1．はじめに

日本語は膠着（こうちゃく）語の文である。膠着語の文は文字列が羅列してあるだけで、文の中の単語がはっきりしていない。そのため、文字列の中から単語を切り出すとき、文中に未知語が出くることがあり、未知語に対して品詞の情報を得るための処理を行なわなければならない。その処理を、未知語が出てきたときにその都度処理をしていたのでは面倒である。例えば、新聞の記事に頻繁に出てくる固有名詞など未知語になることが多い。その固有名詞に対し折角未知語処理を行っても、それを記憶せずに捨ててしまうのはかなり無駄なことである。そこで、品詞獲得の処理で得られた情報を、補助の辞書に保持しておき、獲得される品詞情報を数度の解析によって同じ情報を得られたときに、システムの辞書に登録するという方法考察する。

2．形態素解析の手法

形態素解析の手法として最長一致法を使用する。
最長一致法によって切られた実行結果を図１に示す。

（ａ）例文
　スタルヒンが登板する
（ｂ）解析結果

スタルヒン	未知語
が	ガ行５段未然１
が	格助詞が
が	接続助詞が
登	ラ行５段
登	固有名詞
板	名詞
する	サ変動詞（する型）

図１　最長一致法による解析結果

ここで、未知語となっているのは、”スタルヒン”であり、辞書に”登板”という単語が登録されていないので、切り出し方として理想でないのは、”登板”の間で切れ目が入っていることである。そこで、まず、理想の切り出しでなく、切りすぎた単語をどのようにして理想の単語の切り出しにまとめるかを考察する。そして、未知語となるものの品詞の付け方について考察する。

3．理想の切り出しでないときの処理

まず、理想でない単語の切り出しを例１、例２に示す。例１は、”国連安全保証理事会”という単語は辞書になく、それぞれ部分的に”国連”、”安全”、”保証”、”理事”、”会”という単語が辞書にあるので５つの単語に分けられてしまう。例２の”フライドチキン”も辞書になく、”フライ”、”チキン”は辞書にあるのだが”ド”という単語は辞書になく未知語となり３つの単語に分けられてしまう。

例１　”国連安全保証理事会”
　国連　名詞
　安全　名詞（形容動詞語幹）
　保証　サ変動詞（名詞形）
　理事　名詞
　会　　名詞（接尾語）

例２　”フライドチキン”
　フライ　名詞
　ド　　未知語
　チキン　名詞

これを次のような複合名詞の規則にまとめることにする。

複合名詞の規則
　名詞句：：−名詞｜固有名詞｜
　　名詞句　名詞句｜
　　形容動詞語幹　名詞句｜
　　名詞句　形容動詞語幹｜
　　名詞句　サ変動詞（名詞形）｜
　　サ変動詞（名詞形）　名詞句｜

この規則に従って複合名詞の処理を行い、最終的に名詞句となった単語には品詞として名詞をつけてから、補助の辞書に登録をする。

4．未知語になっている文字の処理

未知語となるもの処理の規則を以下に示す。

learning of lexicon by morpheme analyze
Yoshiteru Kawakami, Yoshiyuki kotani,
hirohiko nishimura
Tokyou University of Agriculture and Tecnology

未知語が英文字であるときその英文字に続く単語が、英文字列であったなら、まとめて一つの英文字列（名詞）とする。そして英文字列がこれ以上続かなくなったときに品詞を名詞とし、補助辞書に登録する。規則としては、

名詞句：：－英文字*

となる。

ここで”Ｊａｐａｎｅｓｅ”という英文字列を使って説明する。この単語が辞書に登録されていないが、”Ｊａｐａｎ”は、名詞で登録されているとする。最長一致法によって解析すると、”ｅｓｅ”は未知語で”Ｊａｐａｎ”は名詞となる。未知語の”ｅｓｅ”をここで英文字列であるので名詞としては辞書に登録しては得たい情報ではなく、意味がない。英文字列の一部に品詞がついていてもその前後に英文字列があれば単語として得たいのは、それらを含んだ英文字列である。だから、この英文字の未知語の場合、重要なのは、一部に品詞がついていてもそれは意味がないことであり、続いている英文字列が一つの単語であるということである。

例３

Ｊａｐａｎ	名詞			
ｅｓｅ	未知語	→	名詞	無意味
Ｊａｐａｎｅｓｅ	名詞	得たい情報		

カタカナ文字列の場合も同様である。例２を見ると、”ド”が未知語となっている。しかし、前後に、カタカナ文字列があるので、まとめて”フライドチキン”とし、名詞という品詞をつける。規則としては、

名詞句：：－カタカナ文字*

である。

５．サ変動詞（名詞形）の品詞の付け方

図１を見ると”登板する”という文字列の場合は”登／板／する”というように切り出される。”板”が名詞であるので、”登”はラ行５段でないことがわかり、名詞（固有名詞）である。ここで、上で述べた規則を用いると”登板”は名詞ということで新しく補助辞書の方へ登録されるが、その後ろにサ変動詞（する形）がきているので品詞はサ変動詞（名詞形）となる。そこで、補助の辞書の方へ品詞としてサ変動詞（名詞形）を持つ単語も登録する。

以上４，５，６の処理から図１で最長一致法によって切り出された単語は次のように再解析され、新しく辞書に登録される。

再解析の結果

スタルヒン	名詞
が	ガ行５段未然１
が	格助詞が
が	接続助詞が
登板	サ変動詞（名詞形）
する	サ変動詞（する型）

補助辞書に登録された単語

スタルヒン	名詞
登板	名詞
登板	サ変動詞（名詞形）

６．補助辞書への単語登録

本方法によって新しく付けられた品詞は補助辞書に登録する。この理由は、正しい品詞が付いているの判断をするためである。正しいと判断できないものをいきなりシステムの辞書に登録することは危険である。そこで、補助の辞書にその品詞情報を保存しておく。その場合、その単語に数字で印を付けておく。そして、その数字は他の文章で解析されたときにまたその単語がその品詞として切り出されたときにその数字を上げていく。そこで、何度かその単語で正しく切り出されたとき、システム側の辞書に登録するようにした。この方式で、できるだけ品詞として正しいものだけをシステムの辞書に登録するようにした。

７．課題

本研究は形態素解析を元に単語の切り出しや未知語の処理を行ってきた。そのため、文字種などで未知語の処理をしたが、名詞に関しての処理はかなりできたのだが、その他の品詞に対しての処理はサ変動詞くらいしかできなかった。やはりその他の未知語に関しての新しい処理を考察しなければならないだろう。

参考文献

１．丸山、梅田、他；"A Japanese sentence analyzer"、IBM Journal of Reseach and Development

２．田中穂積；”自然言語解析の基礎”