動詞辞書の提案とその利用についての一考察

3Ｈ－4

中挟知延子　島田静雄
埼玉大学工学部

１、はじめに

本論文では、我々の作成した電子化日本語動詞辞書の構成内容を述べ、その辞書を利用して行った字面のみによる、日本語文章からの動詞の抽出実験の結果を示す。日本語文章において動詞を代表とする用言は文章を理解する上で重要な役割を担っている。そのため、文章中から動詞を正確に抽出できれば、文章の要点をつかむことができ、機械翻訳をする場合にも役に立つ。

我々のねらいは、日本語文章から動詞を抽出して、そのまま機械翻訳処理にかけるのではなく、むしろ機械翻訳処理が効率良く行われるように、オリジナルの日本語文章を前もって校正しておくことにある。日本語動詞には複合動詞や語尾に「する」の付いたものがあり、これらに対応する訳語は、英語だと１つの単語ではなく２語以上の動詞句の形である場合が多く、対応する訳語として前置詞も登録しなければならない。また、「書く」・「書ける」・「書かせる」のように同じ語幹でも、語尾の活用が違っていると、対応する訳語は異なる。いずれの場合にも翻訳のための辞書はかなり大きくなってしまう。そこで、前もって日本語文章を校正して、同義のものや冗長な言い回しを簡潔な表現に統一しておけば、機械翻訳の際に辞書を参照する回数が減り、処理効率が増すと考えられる。たとえば「書き留める」の英訳は "write down" であるが、「記録する」にも同じ英訳があてはまる。もしも文中にこれら２つの動詞が出てきたら、どちらかの動詞に統一しておけば適切であろう。しかも、そのために必要な動詞の抽出を、形態素解析をせずに字面のみでできれば、抽出のための処理の時間や処理システムの規模も少なくて済み、機械翻訳処理の前処理としてシステム全体に対して占める負荷の割合は大きくならないであろう。

我々は、自家製の動詞辞書を利用するために、「動詞抽出ツール」を作成し、文章中から複合動詞・「する」動詞を含めた動詞の抽出を試みた。抽出は字面のみで行い、辞書を含めたツールの大きさも、フロッピーディスク１枚に収まる程度にしてパソコン上で実現している。

今回述べる動詞辞書は漢字で始まる動詞を中心に作成し、抽出も漢字を用いる動詞にしぼっている。動詞辞書には、ひらがなで始まる動詞も含まれているが、ひらがなのものについては次回の発表で行う。

以下、２章で動詞辞書の構成について述べ、３章で「動詞抽出ツール」について述べたあと、実際にツールを用いて文中から動詞を抽出した結果を示す。４章では、３章の抽出結果を考察し、５章にまとめを述べる。

The Proposal of lexicon of Japanese verb and
a consideration about its use
Chieko Nakabasami,Shizuo Shimada
Saitama University
255 Shimoohkubo,Urawa,Saitama 338,Japan

２、動詞辞書の構成

辞書を以下の項目で構成した。

１）見出し語

動詞に用いる単漢字・ひらがな動詞を登録した。漢字に関しては、ＪＩＳ第１水準１４５０字・第２水準８１字の合計１５３１字を含む。

２）読み

見出し語が漢字の場合、動詞で用いるときにとりうる読みを登録した。「する」動詞になるときの読みも含む。

３）活用

見出し語を動詞で用いるときにとりうる活用の種類を登録した。たとえば、見出し語が「増」の場合には、「増やす」のときには「さ行五段」であり、「増える」のときには「わ行下一段」のように複数ある。また、自動詞・他動詞の区別も登録した。活用の種類は簡略にするため記号化して登録した。

４）送りがな

動詞に漢字を用いるとき、活用変化をしない語幹の部分で、ひらがなをもつものがある。たとえば、「増やす」は、「す」が活用変化をし、「増や」が語幹である。このとき、「や」を送りがなとして登録した。

５）後続可能語句

見出しから作られる動詞が、別の動詞と組み合わさって複合動詞を作るとき、後続可能な漢字を登録した。また、「する」動詞になるとき後続して２字熟語になれる漢字も含む。たとえば、見出し語が「増」のときには、「増え続ける」・「増殖する」などがあるので、「続」・「殖」を後続語句として含む。

次項の（図１）に動詞辞書の実際の画面を示す。

３、動詞抽出ツールと抽出結果

２章で述べた動詞辞書を利用するために作成した「動詞抽出ツール」（以下、ツールとする）の概要を述べる。このツールは、文章中から漢字で始まる動詞を活用の情報も含めて抽出するものであり、抽出実験は参考文献【１】でも報告してあるが、今

回の抽出については、【1】のときに動詞と判断して抽出してしまった、助詞相当句・名詞を、4）・5）により抽出しないようにした。

1）音便形、受身形、使役形の抽出

動詞を活用の種類によって未然形から命令形を抽出するのに加えて、「てフォーム」といわれる、「て」を後ろに伴ったときにとりうる音便形と、受身・使役の助動詞を伴ったときの形も抽出する。

2）可能形の抽出

五段活用の動詞のなかで、活用語尾を下一段にすると、可能の意味を持ち、文章中で多用される形のものがある。たとえば、「書く」の可能形の「書ける」などである。ツールでは、五段活用の動詞に限って可能形も抽出する。

3）常用漢字かどうかのチェック

文章中に現れる漢字は動詞で用いる場合も含めて、常用漢字が望ましいと考え、常用漢字でなければ、警告を出す。

4）助詞相当語句の排除

「に従って」のように、「従う」の活用形が前後に格助詞を伴うと、動詞の形であっても、動詞としてよりも接続詞・副詞の働きをしている語句がある。これらは英語の場合の前置詞句にあたり、動詞・動詞句には対応しないので、「従う」という動詞の連用形の音便変化したものという抽出はしない。

5）同形の名詞の排除

たとえば、「これは一つの試みです」という文で、「試み」は「試みる」という動詞の連用形と同じ形をした名詞であり、動詞ではない。そのため、名詞になる場合に後続する接尾辞・格助詞などの語句の一覧を用意し、もしそれらの語句が後続していれば、動詞として抽出しない。

ツールを用いて行った抽出結果を（表1）に示す。結果として、情報検索でよく用いられる「再現率」と「抽出率」を算出して提示した。なお、実験は2万文字文章で行い、処理時間は原稿用紙1枚あたり約10秒であった。

（表1）漢字で始まる動詞の抽出結果

再現率 （候補中にある抽出すべき対象の数/文章中の抽出すべき対象の数）	99.8%
適合率 （候補中にある抽出すべき対象の数/抽出された候補の数）	99.6%

4、抽出結果の考察

1）再現率が100%にならないのは、動詞辞書に登録していないものが文章中にあったためである。動詞辞書を点検し、内容の充実を図るようにしたい。

2）適合率が100%にならないのは、3章の5）で述べた同形の名詞をすべて排除できなかったことにある。通常の処理で排除できない場合については、例外知識として別に扱うようにしたい。

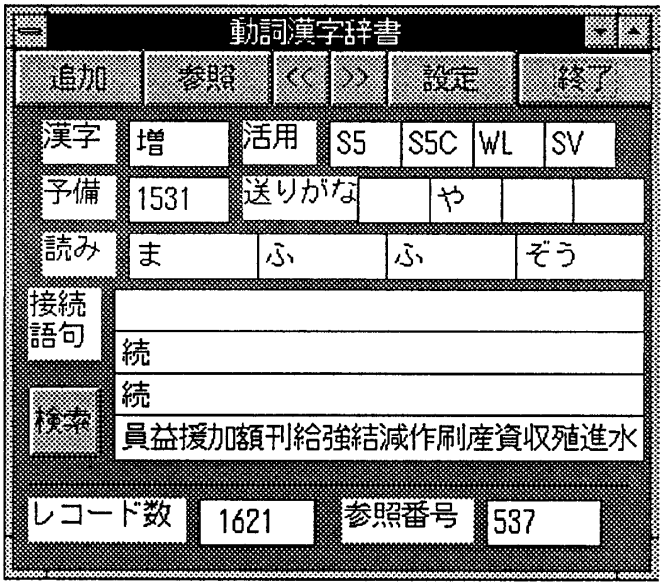
5、まとめ

字面のみの処理で、例外を除けば漢字で書かれる動詞を100%に近い割合で抽出できた。ただもっとサンプル文章を増やして抽出実験を繰り返し、ツールの性能を評価する必要がある。また、動詞かどうかの判断が難しいものについては、抽出をしたうえでユーザと対話をして、決定していきたい。目下、ひらがなで書かれる動詞の抽出についても研究をすすめている。

参考文献

【1】中挟知延子、島田静雄：外国人のための日本語文章校正システム、ＴＣシンポジウム論文集、(1995) 掲載予定

【2】新村　出編：広辞苑、岩波書店（1955）



（図1）　動詞辞書画面