4B−3

日本語入力による英文作成支援システム
— 辞書学習 —

安藤真一　赤峯享　村木一至
(NEC C&C情報研究所)

1 はじめに

近年、機械翻訳のより現実的な利用形態として、対話的に翻訳処理を進めるユーザインターフェイスが検討されている。我々もかな漢字変換に似たインターフェイスから機械翻訳システムの利用を可能にすることにより、ユーザの日本語入力による英文作成を支援する英文作成支援システムを開発している[1, 2, 3]。本システムにおいて日本語から英語への変換(以降、日英変換と呼ぶ)は、機械翻訳で生じる曖昧性を解消するためにユーザの知識を利用しようとするものである。

しかし、日英変換のために要求されるキー操作が増えることにより、ユーザはその思考を中断され、文書作成に集中できなくなるという問題がある。これを解決するため、本稿ではユーザの選択、確定操作を減少させる辞書学習機能を提案する。

従来、機械翻訳のための辞書学習機能には後編集の一機能として提案されたものがある[4]。しかし、これらは完成した文を対象とした翻訳結果の修正操作のみを手がかりとしたため、名詞の学習しかできなかった。これに対し、本システムでは文章を作成しながらの翻訳、修正といった操作の全てをモニターできるため、より細かな辞書学習が可能となる。また、かな漢字変換においても辞書学習機能が存在するが、日英変換では変換対象が異言語であり、その語彙や語順が異なる。そこで、日英変換では文章構造を考慮して辞書学習を行う。これにより、ユーザの日英変換の履歴を反映した第一候補が出力され、選択操作を減少させることができる。また、本システムでは英単語に対する直接の修正も受け付けるため、これを自動登録することによって未登録語などを処理することが可能になる。

本稿では、まず入力文での品詞ごとに、辞書学習の優先順位を学習する方法について述べ、その後に自動登録によって実現できる機能について記す。

2 日英変換における辞書学習

2.1 優先順位更新によるのユーザ適応

本英文作成支援システムでは、日英変換で得たい結果が出力されなかった場合、ユーザはシステムが提示する他の候補から正しい翻訳結果を選択することができる。ここで、頻度情報を用いて候補の順番をユーザに適応させると、ユーザのキーストロークが少なくなるということが、かな漢字変換の辞書学習機能によって知られている。ただし、かな漢字変換が日本語内での変換であるのに対し、日英変換では語彙や文法が変換前後で変化するため、単語単位の対応を登録するだけでは正しい出力が得られる保証はない。このため、日英変換においては登録する語は文章構造の情報を含めて登録する必要がある。

本辞書学習機能では原言語の品詞ごとに登録すべき情報が異なっていることを考慮し、それぞれについて表1に示す構造ごとに頻度を計算する。以下に各品詞について説明する。

2.1.1 名詞の辞書学習

名詞についてはかな漢字変換で行われる辞書登録と同様の方法で優先順位の更新が行われる。すなわち、日本語と英語の一つ以上の単語からなる語の対に対し、その頻度情報を更新する。例えば、日本語入力「大学」に対して「college」を選択すると、この組合せの頻度情報が更新され、次回の変換では「大学」の第一候補として「college」が出力されるようになる。これにより、機械翻訳を利用するメリットである用語の統一が、ユーザの行った選択の履歴に応じてなされるようになる。

2.1.2 用言の辞書学習

動詞、形容詞、形容動詞といった品詞で構成される用言は、日本語と英語の間の語彙や文法の違いのため、かな漢字変換で用いられているように単純な単語対で辞書学習しても正しく結果が反映するとは限らない。

例えば、「張る」は「幕を張る」と「網を張る」では学習すべき訳語が異なる。そこで本システムの辞書学習機能では用言に関しては、必須格とその直前の名詞を含めた日本語文章構造パターンと英語動詞を対応づけて学習する。例えば、「池の周りに網を張った。」といった文章を翻訳する場合には、本辞書学習機能ではユーザがこの文章の用言として「stretch」を選択することによって「縄を / 張る」と「stretch」を対応づけて登録する。これにより、「幕を / 張る」−「spread」の組み合わせには影響を与えることなしに、「網を / 張る」−「stretch」の組を学習することができる。

2.1.3 長文パターンの辞書学習

本英文作成支援システムは、複文や重文を単文の並びに分割する接続助詞といった長文パターンを有している[5]。このような長文パターンは文全体の構造を考慮した

表1: 各品詞の頻度計算の単位構造

品詞	頻度計算の単位となる構造(日本語−英語)
名詞	1つ以上の単語−1つ以上の単語
用言	必須格(名詞+格助詞)+用言終止形−動詞原形
長文パターン	長文パターン
その他	変換前選択範囲−日英変換結果

An English documentation through input in Japanese
— Dictionary Learning —
Shinichi Ando, Susumu Akamine, Kazunori Muraki
NEC Corp.

対応関係であるため、従来のかな漢字変換様のシステムでは辞書学習の対象外のものであった。しかし、本システムでは翻訳機能の一部として長文パターンを利用しているため、これを学習することによってユーザに適した長文パターンを出力することができる。例えば、長文パターン「(単文)が、(単文)」を考えると、

「彼はその本を読み始めたが、」

という日本語文が入力されたとき、システムの自動変換機能によって翻訳が始められる。しかし、この入力の後に

「〜が、すぐに止めてしまった。」

などのように逆接となる文が続くのか、

「〜が、その本は私にとって難しかった。」

などのように制限用法として「が」が使われるのかは分からないため、「but」「which」のどちらを翻訳結果として出力すべきかは決めることができない。本辞書学習ではこれらの長文パターンの使用頻度を学習し、より頻繁に利用されている長文パターンを第一候補として出力する。これによって、長文パターンをユーザの使用する長文パターンの癖に適応させることができる。また長文パターン辞書の初期状態は任意に設定できるため、辞書開発のコストも削減することができる。

2.1.4 その他の品詞の辞書学習

上記以外の品詞については現在のところ、名詞と同様の辞書学習を行っている。ただし、助詞については辞書学習の対象としていない。また、節、句などの単位で変換された場合も、名詞と同様に辞書学習を行う。このときには翻訳処理の前に選択された変換範囲と日英変換の結果として決定された部分を対応させて辞書に登録する。

2.2 自動登録による機能

かな漢字変換では直接変換できない漢字が存在した場合、新規にその漢字文字列を登録することができる。しかし、通常、漢字による直接入力が行えないため、新規登録の際には陽なキー操作で登録を行う必要がある。すなわち、かな漢字変換で直接変換できない漢字列が存在した場合、ユーザは間接的に変換を行って目的とする漢字を得た後に、編集することによって漢字列を得る必要がある。このため、システムは得られた新規漢字列の読み(ひらがな)を得ることはできない。しかし、日英変換はその出力である英語(アルファベット)による入力が可能であり、直接的な修正を行うことができる。このため、本英文作成支援システムではアルファベットによる修正を検出することによって、さらに次の2つの機能が備えられている。

2.2.1 未登録語の新規登録

本システムは未登録語を発見した場合、入力分の翻訳結果を得るためにユーザに対して修正を要求する。ここでは日本語文の修正の他に、アルファベットによる対訳の入力も受け付ける。これによって未登録語の日英対訳を得ることができ、以降の変換においてはこの対訳を利用することができる。

本辞書学習では未登録語に対する修正があったとき、その未登録語の日英の表記と入力された日本語文から推定された未登録語の品詞、そして品詞ごとに定義された表1の構造情報を辞書に新規登録する。ユーザにとって、これらの修正操作は単に入力した文の翻訳結果を得るための後編集的な操作であるが、これを自動的に辞書登録することによって以後の日英変換に反映させることができる。

2.2.2 0代名詞の自動登録

日本語文では必須格の省略が頻繁に起こることが知られており、機械翻訳においてはこの省略を補完する機能が必要とされる。しかし、省略の補完は文脈処理を必要とし、実際にこの機能を備えた実用的なシステムはまだ存在しない。このため通常、0代名詞の補完として特定の代名詞が補われている。

しかし、例えば、日記のように自分について書いた文書では「私」が、手紙などでは「あなた」が、そしてマニュアルなどでは「それ」や「それら」が省略されやすいというように、一つの主題で書かれたテキストでは補完すべき代名詞に傾向があると考えられる。そこで本システムでは、補完すべき代名詞の傾向をテキストごとに登録し、利用する。すなわち、ユーザがアルファベット入力によってシステムの行った補完を修正した場合、これをテキスト毎に設定された辞書に自動登録する。これにより、テキストごとにそのテキストの傾向に沿った代名詞が次回の日英変換では優先される。もちろん、以上の手法では完全な翻訳結果を出力することは不可能であるが、ユーザのキーストローク数の減少が期待できる。これについての評価は以後行う予定である。

3 おわりに

本稿では現在開発中の英文作成支援システムにおける辞書登録機能について提案した。特にここでは日本語と英語の日英変換という、かな漢字変換とは本質的に異なる変換方法を用いているため、文書構造までを含めた辞書学習が必要となる。今後は辞書学習機能の有無によるキーストローク数の違いを基準として、本機能の評価を行う予定である。

参考文献

[1] 赤峯他「日本語入力による英文作成支援」, 情処第43回全国大会 No.3, pp.205-206, 1991.

[2] 赤峯他「日本語入力による英文作成支援インターフェース」, 情処第44回全国大会 No.3, pp.261-262, 1992.

[3] 赤峯他「日本語入力による英文作成支援システム」, 本大会予稿集, 1993.

[4] 堤「簡単な学習機能を備えた機械翻訳のためのエディタ」, 情処第33回全国大会 No., pp.1771-1772, 1986

[5] 佐藤他「日本語入力による英文作成支援システム −長文パターンによる翻訳−」, 本大会予稿集, 1993.