簡単な学習機能を備えた機械翻訳のためのエディタ

7J-3

堤　豊

日本アイ・ビー・エム株式会社サイエンス・インスティチュート

１．はじめに

近年、機械翻訳が注目を浴ているが、自然言語を対象としているため実用システムとして稼働するにはかなりの困難が考えられる。。その原因として、翻訳の品質の問題、プリエディトの繁雑さ、ポストエディト機能の問題、辞書整備の労力などが考えられる。

従来よく議論されてきた翻訳の品質については、今や文脈処理を本格的に取り入れないと飛躍的な向上は望めないと思われる。今後は翻訳文の品質だけではなく、ユーザー・インターフェースを含めたシステム全体の構成が非常に重要になるっであろう。

本稿では、ＩＢＭの計算機マニュアルを対象とした英日機械翻訳システムＳＨＡＬＴ(System for Human-Assited Language Translation)[1][2]用に試作した翻訳用後編集プログラム（翻訳エディタ）について報告する。この翻訳エディタの最大の特徴は学習機能を翻訳部とリンクして実現しているため使えば使うほど翻訳の品質を向上させることにある。

２．翻訳エディタの概略

機械翻訳システムを実際に使用する場合に労力を必要とするのは、

（１）翻訳システムの仕様に合わせたプリエディト作業
（２）ユーザー別、分野別の辞書の整備
（３）誤訳に対するポストエディト作業

である。本稿では、このうち（３）に着目して翻訳作業の効率を上げることを考える。本エディタの特徴は、

- 曖昧さを持つ単語の表示および選択機能
- 辞書引き機能
- 単語挿入および置換機能

である。

機械翻訳では単語の訳し分けは、重要な問題の１つである。これは、１つの辞書見出し語に対して複数の訳語が登録されていた場合に、どれを訳出するかという問題である。機械翻訳システムが、曖昧さのある単語をすべて正しく訳し分けることは、現状では困難であるので、誤訳あるいはニュアンスが違うものは訳し直すことが必要である。

ＳＨＡＬＴでは曖昧さのある単語のうち８割強はそのまま使用でき、修正する必要がない。従ってすべての訳語をあらかじめ表示する方法では正しく訳されていても選択の作業が必要であり、修正に余分な手間がかかる。

本エディタでは最も確からしい訳語を１つ選出し、画面上では前後の単語と色を変えて表示することにより曖昧さがあることを示している。また原文の単語と訳文の単語の対応がとられており、必要に応じて辞書中の内容を表示し、任意の訳語を選択できる。これにより修正が必要な訳語のみ選択作業を行えばよく、後編集が効率良く行える。

本エディタでは、出来るかぎりキーボードをたたく回数を減らすために、辞書参照された訳語を直接訳文中に挿入したり、あるいは置換することができる。この場合、翻訳エディタは、自動的に訳文を単語単位に分割するので、カーソルを動かすだけで挿入、置換ができる。

３．学習機能

従来の機械翻訳システムでは、情報は計算機から人間への一方通行であった。本エディタでは、後編集を行うことによって、システムの性能をあげることができるような学習機能を備えている。例えば、

ｏｐｅｒａｔｏｒ　ｃｏｍｍａｎｄ

では、「操作員指令」「演算子指令」のいずれなのかは、システムの内部で判断するものがほとんどである。この判断を間違えると誤訳となる。これを修正するには、一般に次の３通りの方法がある。

（１）ユーザー辞書にｏｐｅｒａｔｏｒを「操作員」として登録する
（２）用語辞書に「ｏｐｅｒａｔｏｒ　ｃｏｍｍａｎｄ」を「操作員指令」として登録する。
（３）システム内部の訳し分けの情報を変更する

このうち（１）の方法ではほとんどのシステムではユーザー辞書を優先して参照するため「操作員」と「演算子」を訳し分けることができなくなるという欠点がある。（２）の方法は最も有効でかつ簡単な方法であるが、用語辞書が膨大になってしまう。（３）の方法は普通はユーザーに開放されているわけではなく、仮にユーザーがいじれるとしても、設定方法が非常にやっかいで副作用の心配もある。これを解決するために、本エディタでは、学習機能を導入することにした。

３．１．学習対象

一口に学習といっても何を学習させるかが一番の問題点である。翻訳システムに限っていえば、曖昧さのあるところということになる。つまり、訳語の選択、係り受けの選択、接続詞の範囲などが学習の対象として好適と思われる。本エディタでは、このうちもっとも簡単でかなりの効果が期待できる、名詞の訳語選択の部分に学習機能をつけた。

３．２．学習のタイミング

学習のさせ方にはいろいろあるが、人間が意識せずに機械が学習することが望ましい。また繁雑な手順で学習

The Editor with Simple Learning Feature for Machine Translation
Yutaka TSUTSUMI
Science Institute, IBM Japan, Ltd.

させることは、システム全体の効率を低下させることになってしまう。システムが誤って訳出した文はどのみち修正しなければならないので、そのときに同時に学習させるのがもっとも効率がよさそうである。そうすればユーザーは学習を意識することなく、後編集するだけで翻訳システムの能力を高めることができる。また、こうすることにより何度も同じ修正をしたり、修正してさらに辞書に入れるといった2度手間を避けることもできる。

3．3．エディタと翻訳システムのリンケージ

図1に翻訳システム全体の流れを示す。翻訳システムは、最も確からしい訳文を1つ出力し、付随情報として、曖昧さのある単語の位置、および名詞句をエディタに渡す。ユーザーは2章で述べた方法により訳語を修正する。エディタは、このときの修正情報を翻訳システムにフィードバックする。これにより翻訳システムは、訳語選択について学習をおこなう。実際の訳語選択についての学習のアルゴリズムを次節で述べる。

3．4．学習のアルゴリズム

実際の学習過程について述べる前にＳＨＡＬＴ内部の名詞句の翻訳の手順について簡単に触れておく。ＳＨＡＬＴでは名詞句の翻訳だけを特別に処理する部分を設けており、単純名詞句処理部と呼んでいる。これにより、文のレベルの翻訳と名詞句の翻訳を独立して行うことができる。本来、文のレベルの構造と名詞句レベルの構造はかなり異なるので、単純名詞句処理部を設けることにより、名詞句の翻訳をきめ細かく行うことができる。

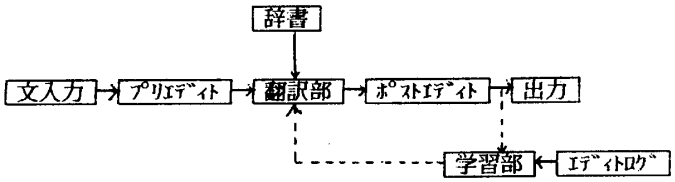

図1　翻訳システム全体図

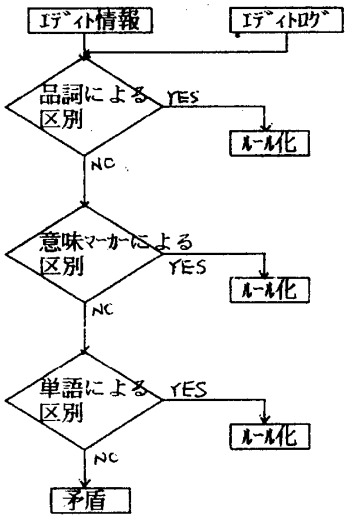

図2　学習の手順

単純名詞句処理部では、各単語の品詞を認定し、単語と単語の共起頻度により単語間の係り受け関係をきめ、次に訳語が2つ以上ある場合には、訳語およびその抽象概念である意味マーカーの共起頻度を用いて訳語を決定している。ただし意味マーカーは、名詞にのみつけられている。

学習に使えるデータとしては、

（1）各単語の品詞名
（2）名詞についている意味マーカー
（3）各単語間の共起頻度データ

である。このうち、共起頻度のデータについては、本システムでは、扱わない。

図2に本システムの学習の手順を示す。このように、本システムでは、学習は品詞から、意味マーカー、単語というように広い範囲から狭い範囲へと条件を増やす方向に動く。学習部はエディト情報が入ると今までのエディトログを参照して品詞で区別できるかどうかを調べる。もし区別できれば、それをｉｆ－ｔｈｅｎ－ｅｌｓｅのルールにし、翻訳部に渡す。そうでなければ、意味マーカーで区別、単語で区別と進んでいき、それでも区別できない場合には、ユーザーに矛盾であることを知らせる。

ここで判別の順序が品詞－意味マーカー－単語の順になっているのは、本エディタが条件の細分化を行なうように学習をするためである。学習には一般に本システムのように細分化するものと逆に一般化するものとがあるが、ここで対象としている名詞の訳し分けの場合には、あまりデータ量が増えないことや、細分化する手順がかなり明確であることから、細分化の方式が適当であると思われる。

4．まとめ

本稿では、翻訳エディタの特徴と簡単な学習機能について述べた。現在、学習のアルゴリズム自体は条件の細分化を行っているだけで、何ら特徴がないが、今後はデータが増えた場合には一般化等を導入して、情報量を減らすようにする予定である。また、学習対象も今は名詞の訳し分けについてのみ行っているが、これを発展させて動詞の訳し分けにも使いたいと考えている。ワード・プロセッサが出荷時点ではそれほど高い変換率ではないのにかなり使いやすくなっているのは、学習機能のおかげである。同様に機械翻訳システムにもこのような機能が要求されるようになるであろう。

参考文献

［1］堤泰治郎他：
情報処理用語の英日機械翻訳について
情報処理学会第２７回全国大会、１９８３
［2］堤　豊他：
英日機械翻訳システムＳＨＡＬＴにおける単純名詞句の翻訳手法について
自然言語研究会５６－１、１９８６