

3E-1

過去の翻訳結果を利用した翻訳支援システム

伊藤悦雄 長谷部浩一 武田公人 天野真家

(株)東芝 総合研究所

1. はじめに

現状では機械翻訳は自動翻訳ではなく人間の翻訳作業の支援を行うものである。そのため、機械翻訳が出力した結果に修正(ポストエディット)を加えたものが最終結果として利用されることが多い。従ってポストエディット作業の軽減は機械翻訳システムの性能向上の重要な要因の一つである。

本稿では文書のバージョン管理を行うことによって過去の翻訳結果の利用を可能とし、ポストエディットの軽減をはかる翻訳支援方式について述べる。

2. 翻訳支援機能への要求

従来の翻訳システムでは、入力文書はそれぞれ独立した文書として扱われる。すなわち、過去に翻訳した文書(既翻訳文書)の一部に修正を加えられた文書(修正文書)を翻訳する場合、この文書は新規文書として扱われ全体が翻訳される。このため、修正文書の翻訳結果に対しては全体に渡ってポストエディットが必要となる。従って、以前に翻訳を行った文と同一文の翻訳を行う際には、既に存在するポストエディット後のデータを利用し、翻訳効率を向上させたいという要求が生じる。

3. バージョン管理機能

上記の要求に従来のシステムで応じるためには、ユーザ自身が修正情報を管理し、修正された部分のみを翻訳し直し、その結果を以前の翻訳結果と差替えるという作業が必要である。しかし、細かい修正が広範囲に渡って行なわれた場合や、文書作成者と翻訳者が異なる場合などでは修正情報のユーザによる管理は困難である。

そこで、我々は過去の翻訳結果を再利用するために、英日機械翻訳における支援機能として次の三機能を導入した。

第一の機能は二つの文書(既翻訳文書と修正文書)の差、すなわち修正された箇所を求める修正箇所抽出機能である。第二の機能は第一の機能で求めた情報を

基に、過去の翻訳結果を利用して翻訳を実施する機能である。第三の機能は修正された部分をユーザに知らせる修正情報の可視化機能である。

なお、これらの機能では我々が開発した機械翻訳システム AS-TRANSAC(TRANSLation ACcelerator)による翻訳結果を用いる。このデータは、第一に文単位に分割されていること、第二に原文と訳文とが対応して記録されているため、原文を指定することにより訳文を得ることが容易であること等の特徴を有する。

3.1. 修正箇所抽出機能

この機能は二つの文書を一文ずつ比較し、文書の変更点を求める機能である。

通常、文書からの文の抽出には形態素解析を必要とする。しかし、修正箇所抽出機能の実現に関しては、下記のような形態素解析を必要としない文抽出方式を採用している。その理由は、第一に二つの文書は同一ではないため文の完全な抽出がされてもすべての文の一致は不可能であるためであり、第二に処理速度の向上のためである。

- ・ 文末には文の終りを示す記号(?!?;:等)が存在し、その次の文字が空白または改行である。
- ・ 前置詞や冠詞以外の単語の先頭文字が大文字で、かつ、次の行が空行であるか、次の行の先頭単語が大文字で始まる場合その行末で文は終了する。

このような条件によって抽出された修正文書の文を既翻訳文書と先頭から順次比較する。この比較によって一致しない場合には、次の理由が考えられるため、既翻訳文書の他の部分と比較を行う必要がある。

- a. 文の一部が修正された
→ 一致する文はないが、次文は一致する
- b. 複数の文に渡って修正された
→ 一致する文はないが、
数文先では一致する
- c. 挿入された文である
→ 一致する文がない

d. 移動により挿入された文である

→ 別の位置に一致する文がある

e. 複写によって挿入された文である

→ 別の位置に一致する文がある

これらの比較により一致する文が存在する場合は修正文書の文が既翻訳文書のどの文と等しいかという情報をバージョン管理ファイルへ出力する。一方、一致する文が存在しない場合には翻訳処理が必要な文であるという情報を出力する。

上記の不一致理由の内、b~eの場合には一致する文の位置の発見が困難となり、最終的には全文に対する検索が必要となる場合がある。この検索を高速に行うために文末の二単語をキーとして一次検索を行う方式を採用した。これは、これらの単語は文中からの抽出が容易である上、これらの単語によって文の限定が可能であるからである。また、この検索方式は「同一単語で終了する英文は少ない」という英語の性質に基づいている。

第一表に示すように文末の二単語の比較により約80%の文が一意に決定される。このため、この検索方式を用いることにより約1000文からなる既翻訳文書とそれに約一割の修正を加えた修正文書との比較が数秒で実現された。

3.2. 修正情報を用いた翻訳機能

バージョン管理機能を用いて得られた修正場所の情報をを用いて翻訳処理を行う。

第一表 英文における
同一単語の文末出現頻度

(単位:単語)

文数	一単語	二単語
1	444	764
2	69	66
3	29	27
4	20	7
5	13	1
6	8	1
7	3	1
8	4	0
9	4	0
10	4	0
11	2	0
12	0	0
13	1	0

注:文書中のある文末の一単語あるいは二単語により限定される文の数を示す。(総数996文)

既翻訳文書中に一致する文がないと判断された修正文書中の文は従来通り、機械翻訳が行われる。一方、一致する文に関しては既翻訳文書から翻訳結果を得る。AS-TRANSACで翻訳を行った文書の原文・訳文の対応が取られており、既にポストエディットされた翻訳結果を容易に得ることが可能である。

3.3. 修正情報の可視化機能

修正文書のうち新たに翻訳が行われた部分に対してはポストエディットを加える必要がある。このときの支援として、修正情報の可視化が上げられる。

対応する原文と訳文を同一画面上に表示し、対応をとりながら翻訳/修正を行う機能(対話翻訳機能)をAS-TRANSACは有しているが、この画面上で新たに翻訳が行われた部分呼びだす機能を加えポストエディットの支援を行う。さらに、修正文とその翻訳結果の位置を英文/和文ワードプロセッサ上で反転表示する機能による修正情報の可視化も実現している。

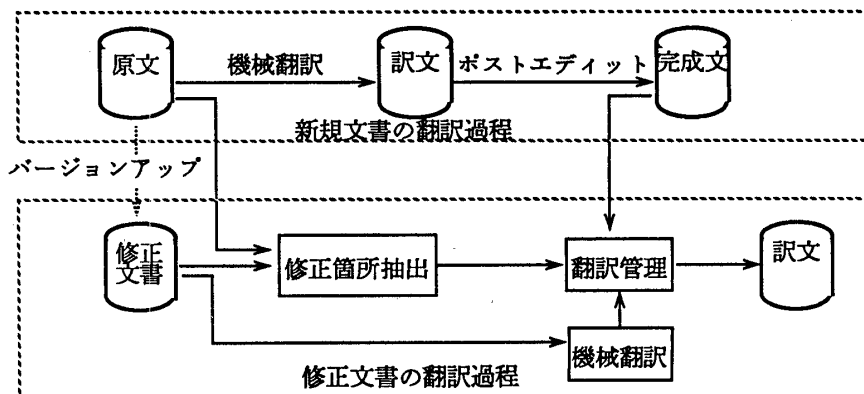
4. おわりに

本稿で述べた翻訳支援機能はマニュアルなど文書のバージョンアップが頻繁に行われる文書に対して特に有効である。

最後に本システムに関する今後の機能拡張について述べる。第一の拡張はAS-TRANSACによって翻訳された文書以外の文書を扱えるようにすることである。第二の拡張は翻訳する文に類似した文(例えば文中の数字・名詞のみ異なるような文)もそのポストエディットされた結果を翻訳時に利用可能とすることである。これらの機能拡張により一層効率のよい翻訳システムが得られる。

参考文献

安達他、日英相互翻訳システムのエディタ方式、情報処理学会34回全国大会、1988



第一図 過去の翻訳結果を利用した機械翻訳処理の流れ