オンラインデータベース翻訳サービスにおける機械翻訳の利用

2C−12

渡辺敏彦　山本真理　加藤晴嗣
(株)富士通静岡エンジニアリング

1．はじめに

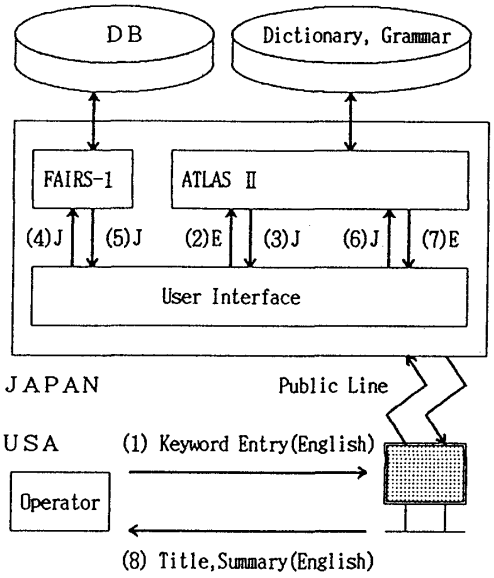
現在、富士通は機械翻訳システムATLAS の応用として、日本語データベースを海外から英語でオンライン検索することができるシステムを検討している。

このシステムでは英語キーワードを日本語キーワードに機械翻訳し、その日本語キーワードを使って日本語データベースの検索を行い、検索された日本語データは機械翻訳されて英語として出力されるというものである。このため、日本語を全く知らない外国人が英語で簡単に日本語データベースを検索することができる。

本稿では、このオンラインデータベース機械翻訳システムの概要を説明すると共に、機械翻訳の利用可能性、問題点、その解決策について述べる。

2．概要

オンラインデータベース機械翻訳システムの処理概要を図−1に示す。



図−1　処理概要

3．機械翻訳の利用可能性

現在の機械翻訳システムはマニュアル文、説明文を翻訳対象とし、その範囲内で実用化されている。この分野での機械翻訳システムの応用では、人手による手直し作業をできるかぎり軽減するため、自然で読みやすい訳文が要求されている。一方、データベース検索に機械翻訳を利用する場合、その目的は即時翻訳であり、人手による作業は前提としない。そのため、訳文の読みやすさに難点が出てくる場合もある。しかし、訳文についてある程度の読解性があり概要把握が可能であれば検索用として十分利用可能であると思われる。

そこで、概要把握を目的とした機械翻訳システムの要件を調査するため、実際にデータベース化されたデータを対象とし機械翻訳をおこなった。

4．検索用翻訳の要件

データベース化された新聞文約1,500 文をATLAS を用いて翻訳し、問題となる文の原因分析を行った。結果を表−1に示す。

表−1　問題がある文の原因

問題点	%	内容
未翻訳文	37.0	原文が長すぎ、訳文が出力されない。
辞書	32.3	辞書の未登録語
原文解析	19.5	係り先の誤り
述語省略	3.8	動詞の省略
口語的表現	3.0	「ずいぶん〜わけだ」等
並立	3.0	並立句の解析誤り
多義語	0.7	多義語の解析誤り
生成	0.7	訳文生成の誤り

この結果、原文が長すぎ訳文が出力できない場合が問題原因の3割以上を占めていた。

この未翻訳文の存在は原文の概要把握を妨げる原因であり、可能なかぎり意味の把握が可能な訳文を出力する必要がある。

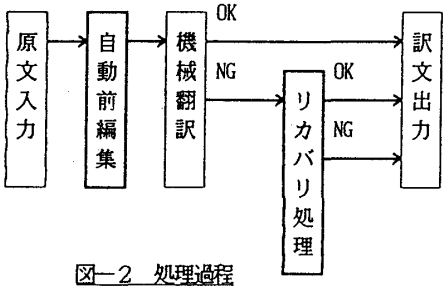
また、新聞文には概要把握にとってあまり重要でない補助的情報も多く含まれている。このような補助的情報により原文が長く複雑になる場合もあり、原文解析の障害となっている。

以上の分析から概要把握を行うためには未翻訳文の解決と補助的情報への対応が必要であると考えられる。

そこで、概要把握にあまり重要でない部分については原文を自動的に加工する処理をおこない、訳文が出力されない場合は、句単位に再度機械翻訳しリカバリを行う機械翻訳システムの作成を試みた。

5．自動前編集とリカバリ処理を持つ機械翻訳システム

最終結果が得られるまでの過程を図−2に示す。



図−2　処理過程

The Utilization of Machine Translation in Online Database Translation Service
Toshihiko Watanabe,Mari Yamamoto,Harushi Kato
Fujitsu Shizuoka Engineering,Ltd.

5．1自動前編集

自動前編集では、原文の意味を損なわずに、翻訳結果が読みやすく簡潔に訳出できるように、原文に対して加工を行っている。以下に主な機能を示す。

(1) 補助的情報の処理

補助的情報は一文の概要把握を行う場合、あまり主要な役割を担っていない。また、補助的情報の訳出は訳文を長く複雑にする原因となっており、読みやすさという点からも翻訳対象外とする処理をおこなった。

例) 同社の一階に「ショールーム=写真右上」を開設
↓自動前編集
同社の一階に「ショールーム」を開設

(2) 引用記述の処理

一文中に引用記述がある場合、引用の中身を別文として翻訳できるように前編集する。

例) 同社は「ショールームを開設する。」という。
↓自動前編集
同社は次のようにいう。
ショールームを開設する。

(3) 項番の処理

項番により意味が区切れる場合、項番ごとに別文として翻訳できるように前編集する。

例) 特徴は①同社一階に開設すること②同時にサービスを開始することである。
↓自動前編集
特徴は
①同社一階に開設すること
②同時にサービスを開始することである。

以上に示す自動前編集を行うことにより、長く複雑な訳文が出力されることがなくなり、読みやすく概要把握のしやすい訳出が可能となった。

5．2リカバリ処理

訳文が出力されない場合のリカバリ処理として、句ダイレクト方式を利用した。この方法は一文中で意味のまとまった句を解析し、その句ごとに機械翻訳を適用し翻訳する方法である。生成された句は原文の句の並び順にダイレクトに並べることにより訳文が出力される（図−3）。句単位の翻訳ができなかった場合、階層化した句認識ルールに従いさらに細分化した句で翻訳を行い、やはり原文の並び順に出力する（図−4）。

同社はショールームを開設、サービスを開始する。
↓
［同社はショールームを開設］［サービスを開始する。］ ↓
[This company establishes the showroom].
[The service is started] . ←

図−3 句ごとのダイレクト翻訳

同社はショールームを開設、サービスを開始する。
↓
［同社は］［ショールームを開設］［サービスを開始する］
↓
[This company] , [establishes the showroom].
[The service is started] . ←

図−4 さらに細分化した句で翻訳

句認識ルールは原文を形態素解析し、その形態素の文法的役割、または単語表記から句の区切れ部分を認識する。また句の区切れ可能性の強弱から図−5のように区切れ位置を階層化し、句ダイレクト翻訳を行う際、この階層の上位のある句の区切れ位置から適応を行っていく。

同社は、ショールームを開設、サービスを開始する。
↑ ↑ ↑
第二階層位置 第一階層位置
第三階層位置 第三階層位置

図−5 句の区切れ一の階層化

また、新聞、雑誌等の文書約7,500 文に対し分析を行い、句の認識ルールを作成した。この句認識ルールを簡単に説明したものの一部を以下に示す。

第一階層ルール
以下の条件が原文中にある時、句の区切れとみなす。
訳文生成時には句の切れ目はピリオドで区切る。
・格助詞「を」「が」「で」「に」＋サ変名詞＋読点
・接続助詞＋読点

第二階層ルール
訳文生成時には句の切れ目はカンマで区切る。
・英数字以外の表記＋読点＋英数字以外の表記
・表記「〜を用いて」
・表記「〜に関して」

第三階層ルール
訳文生成時には句の切れ目はカンマで区切る。
・格助詞「を」「が」「で」「に」
・接続助詞

6．評価

以上に示した諸方式を利用して読解性の可否という点から翻訳評価を行った。この結果、形態素解析に成功していればほぼ１００%近い訳文の出力が可能であった。また訳文の読解性については１５〜２０%の向上が認められた。

このことから、当翻訳システムを利用することにより、即時翻訳を目的とし、読解性、概要把握を重視した検索用機械翻訳が十分に利用可能であると思われる。

今後、大量のデータを調査することにより、階層的句認識ルールと自動前編集仕様のさらなる改善をおこなっていく予定である。

参考文献

〔１〕信国：自然言語における長文分割方式
情報処理学会第３９回全国大会
4U-7(1989)