逐次変換方式による韓日翻訳ツールの評価

各務 宏昭　甲斐 郷子　中村 順一　　吉田 將
九州工業大学 情報工学部　　九州芸術工科大学

1　はじめに

韓国語は, 語順が日本語とほとんど同じであり, また日本語の助詞に相当するものが存在する. そのため, 構文・意味解析を行わなくても, 単語や助詞の単純な置換だけで, ある程度日本語への翻訳が可能である. そして, このような韓日両国語間の類似点に着目した機械翻訳システムの研究・発表が多く行われている [1] [2].

しかし, 両国語間の機械翻訳システムを実用化し, 高品質な翻訳結果を得るためには, 訳語, 助詞等の選択がやはり必要であるため, 深い意味解析を避けることができない. この点が実用システムを作成する上では問題となっている.

ここでは, 韓日両国語間の語順等の類似点を利用し, 訳語選択をユーザに任せる手法を考える. つまり, 翻訳過程でユーザとシステムが協力して翻訳するワードプロセッサ風の機械翻訳システムを作成することにした. 翻訳の際に, ユーザが行うことは, 入力と訳語選択であり, システムは単語切りや, 辞書引き等を行う [4, 5].

本システムで, 韓国語の2冊の本の各1000文節を翻訳した結果, 両本ともに, 約95%とういう高い割合で変換が可能であった. 本稿では, 作成したシステムの概要を示し, そのシステムを用いた翻訳実験結果について報告する.

2　逐次変換方式でのシステムの実現

韓国語には, 日本語の文節にあたるものを離して書く決まりがあり, この決まりを『分かち書き』と言う [3]. 本システムでは, この分かち書きを利用して対話的な翻訳を行う. 以下, 文節は分かち書きにより分けられたものとする. 本システムは, この文節を入力とし, 文節毎に入力・変換を行って翻訳を進める. これを, 逐次変換方式と呼ぶことにする.

逐次変換方式では, 意味マーカや細かな品詞情報等といった翻訳のための複雑な辞書情報を膨大に持つ必要がなくなる. そのため, ユーザは, 辞書を拡張したり, 文書の対象領域や自分の好みに合うよう辞書をチューニングすることが簡単にできる. そこで, 本システムでは, ユーザが翻訳を行いながら簡単に辞書を拡張できる辞書登録機能を備えた.

なお, 韓国語の中には, 後ろの単語と呼応してある意味をなす単語もあり, 逐次変換方式ではうまく処理できないものもある. このようなものを, 本システムでは, 熟語として処理を行うようにした.

3　システム構成

本システムは, 多国語エディタMule上でEmacs Lispを用いて実現した. 図 1にシステムの構成を示す. システムは, 入力文節を文節解析モジュールでテーブルを用いて, 語幹・補助語幹・語尾に分割し, 辞書登録形作成モジュールで辞書登録形になりうる語の候補を作成する. 次に, 辞書検索モジュールで辞書登録形の候補及び, 語尾を韓日対訳辞書で検索する. 検索できなかった場合や, 検索した日本語が妥当でなかった場合は, 辞書登録モジュールで新たに韓日対訳辞書, 日本語活用辞書に登録する. そして, 日本語形態素生成モジュールで, 日本語訳を生成し出力する.

なお本システムは, [4] で報告したものに対して, 辞書登録形作成モジュールで, 韓国語における語尾の縮約にも対応できるよう改良し, 辞書検索モジュールも, 複合語を処理することができるように改良したものである.

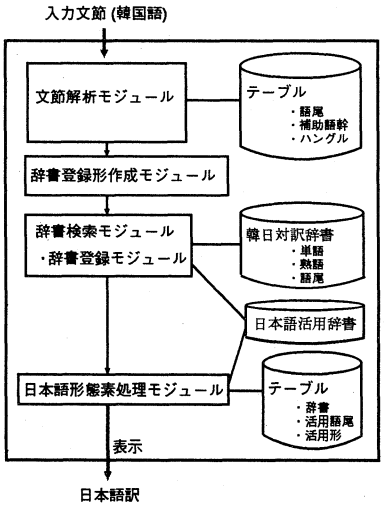

図 1: システム構成

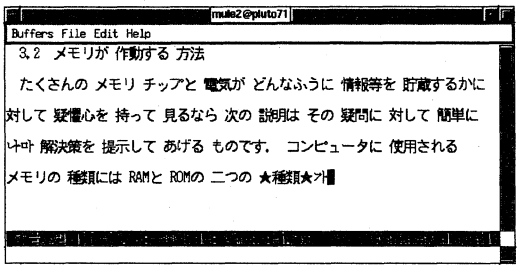

図 2: 翻訳画面

表 1: 変換結果

	MS-DOS	CIM 戦略
日本語に変換	956 文節	945 文節
〃 不変換	44 文節	55 文節

4 実験

4.1 翻訳実験

実験は, システムが持っている単語辞書に何も登録されていないものとして, 翻訳を行いながら, 未知語を辞書登録していくことにより行った. 実験は, 2 冊の異なるタイプの本の中の 1000 文節を対象に行った. 1 冊は, マニュアル的な本「MS-DOS 実践教室」 [6](以下, M とする), もう 1 冊は, ドキュメント的な本「CIM 戦略 ~ IBM 藤沢工場の挑戦~」 [7](以下, C とする) である. なお, C は, 日本語版の原本を翻訳したものである. 実際の翻訳中の画面例を図 2に示す.

変換結果は表 1 のようになった. 両本ともに, 約 95% というかなり高い割合で日本語への変換が可能であることがわかった. 以下に, 日本語へ変換できなかった文節のタイプとその例を述べる.

タイプ 1 (20 文節, M が 10 文節, C が 10 文節) : 韓国語における, 変則用言による語幹の変化.

例 : 따라(従って) ⇒ 따르(다)(従う) + 아(して)

⇒ 따라(다) + 아(して) (本システムの分析)

タイプ 2 (12 文節, M が 11 文節, C が 1 文節) : 韓国語における, 縮約による補助語幹の変化.

例 : 켜졌는카(ついたかどうか) ⇒ 켜지(다)(つく) + 었(過去)+는카(疑問)

⇒ 켜졌(다) + 는카(疑問) (本システムの分析)

タイプ 3 (15 文節, M が 6 文節, C が 9 文節) : 最長一致法による, 語尾の分割誤り.

例 : 시스템에(押す) ⇒ 시스템(システム) + 에(に)

⇒ 시스테 + ㅁ에 (本システムの分析)

タイプ 4 (7 文節, M が 3 文節, C が 4 文節) : 最長一致法による, 補助語幹の分割誤り.

例 : 임시로(臨時で) ⇒ 임시(臨時) + 로(で)

⇒ 임 + 시 + 로(本システムの分析)

タイプ 5 (28 文節, M が 6 文節, C が 22 文節) : 1 文節中に語尾を 2 語含むもの.

例 : 예를들면(例を挙げれば) ⇒ 예(例) + 를(を) + 들(다)(挙げる) + 면(れば)

タイプ 6 (4 文節, M が 0 文節, C が 4 文節) : 文節の中央に語尾を含むもの.

例 : 필요할때(必要な時) ⇒ 필요하(다)(必要だ) + ㄹ(である) + 때(時)

タイプ 7 (13 文節, M が 8 文節, C が 5 文節) : 一般辞書で見つからなかったもの. 例 : 다시말하여

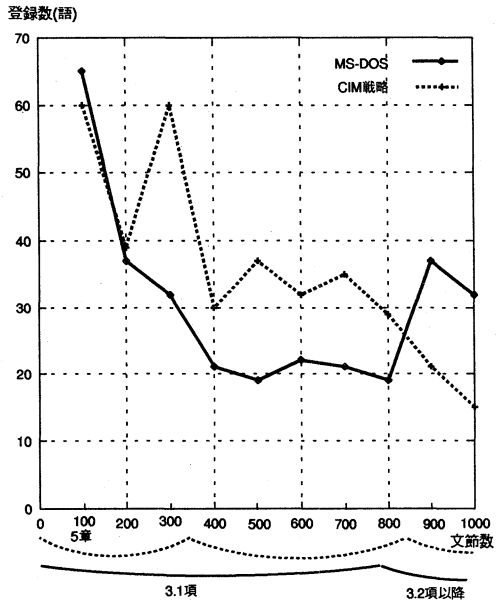

図 3: 100 文節毎の辞書登録語数グラフ

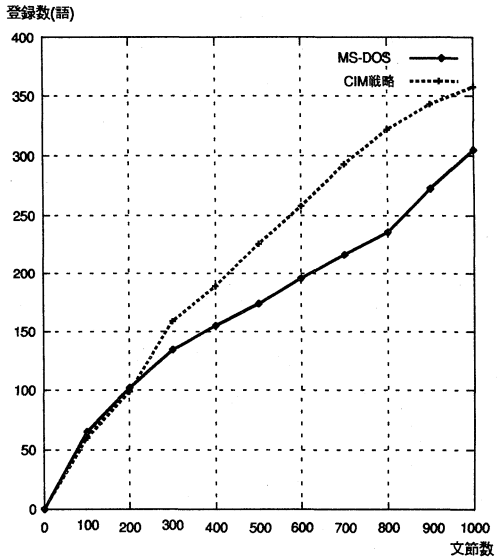

図 4: 延べ辞書登録語数グラフ

次に, 100 文節毎の辞書登録単語数の変化を図 3に示す. このグラフは, 横軸に文節数を取り, 縦軸に登録単語数をとった. 実線は, M の, 点線は, C の 100 文節毎の登録単語数の変化のグラフである. また, 横軸の下方に記した実線, 点線は, 各本の 3.1 項といった項目に対応している.

まず, M のグラフについて考えてみる. 実験の最初では, 辞書に単語が登録されていないため, 100 文節中, 単語登録をしたのは 65 語であった. しかし, 実験の終わりの段階では, 100 文節中, 単語登録をしたのは 32 語で, 最初に比べると, およそ半分の単語登録で済むようになった.

図 3のように, 実線で示した登録数は, 文節数の増加に従い減少しているが, 900 文節で顕著に増加しているのがわかる. これは, 800 文節付近で, 新しく 3.2 項に入ったためだと考えられる. また, M のグラフの減少の過程が, 割合に滑らかなのは, M はマニュアル的な本であるために, 比較的一定の割合で, 同じ単語が使われているためだと考えられる.

次に, C のグラフについて考えてみる. 実験の最初では, 辞書に単語が登録されていないため, 100 文節中, 単語登録をしたのは 65 語であった. しかし, 実験の終わりの段階では, 100 文節中, 単語登録をしたのは 15 語で, 最初に比べると, およそ 4 文の 1 の単語登録で済むようになった. また, C のグラフの減少の過程が, M のグラフに比べて, 変動が大きいのは, C はドキュメント的な本であるために,M に比べ, それほど一定の割合で, 同じ単語が使われていないためだと考えられる.

次に, 延べ辞書登録単語数の変化を図 4に示す. グラフの見方は, 図 3と同様である. 実線で示された M の登録単語数は, 最終的に, 305 語だったのに対し, 点線で示された C では, 358 語であった. M の方が, 登録単語数が多くなったのは, 先程記述したように, マニュアル的な本のため, 同じ単語がより多く使われているためだと考えられる.

4.2 複合語処理による辞書登録単語数の減少

現システムは, “開発管理” のような日本語訳を持つ韓国語の複合語を, “開発” と “管理” に分割して辞書検索を行うことが可能である. しかし, 以前に, Cの2000文節に対して翻訳実験を行ったときは, このような複合語処理にを行っておらず, 複合語により辞書登録数が多くなってしまった. そこで, ここでは, 以前に実験したCの2000文節に対して, 現システムで実験をしたら, どれくらい辞書登録単語数が減少するか調べてみた.

複合語処理による辞書登録単語数の変化を図5に示す. グラフの見方は図3と同様であるが, 実線は複合語処理を施した場合の100文節毎の登録単語数の変化のグラフを表わし, 点線は複合語処理を施さない場合の100文節毎の登録単語数の変化のグラフを表している.

最終的に, 複合語処理を施さない場合は, 登録単語数が605語であったのに対し, 複合語処理を施した場合は, 登録単語数が497語となり, 108語の登録単語数が減少したのが確認された.

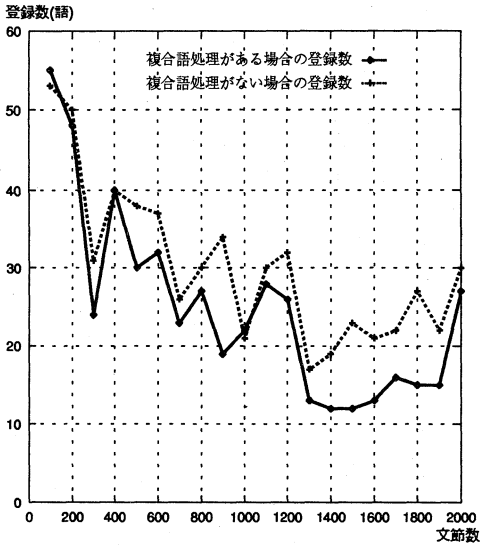

図 5: 複合語処理の辞書登録語数比較グラフ

図5のように, 文節数の増加に伴い, 両線の隔差が大きくなっている. これは, 文節数の増加に伴い, 複合語処理の効果が次第に大きくなっていることを意味すると考えられる.

5 おわりに

本稿では, 逐次変換方式による韓日機械翻訳システムの実験結果について述べた. システムの問題点としては以下のものがある. 一つは, 逐次変換で人間が訳語選を行いながら翻訳を行うために, 文節が文章の最初, 文脈から何の情報も得られない場合には, 訳語選択が困難になる場合があり, これは, 特に助詞の選択に見られる. 他には, 3文節以上にわたる, 韓国語独特の言い回しによる表現等は, 本システムは, うまく翻訳できていないことである. また, 現システムは, 処理できない文節があり, この多くの部分は, 韓国語における変則用言等の文法的なものである.

今後, これらの問題点を解決しつつ, システムの実験, 評価を重ね, 韓日翻訳システムの実現に向けて, さらに改良, 拡張を図る予定である.

参考文献

[1] 李 義東他 : “助述表現の意味対応による日韓述部機械翻訳システム”, 情報処理学会論文誌, Vol.31, No.6, pp.801-809 (1990).

[2] 李 秀炫他 : “韓日機械翻訳のための音韻表現形式による用言の活用処理”, 情報処理学会論文誌, Vol.33, No.12, pp.1565-1577 (1992).

[3] 油谷 幸利 : ハングルの基礎, 大修館書店 (1988).

[4] 各務 宏昭他: “逐次変換方式による韓日翻訳ツールの試作”, 電子情報通信学会技術研究報告, Vol.94, No.292, pp.47-54 (1994).

[5] Hiroaki Kagami, et al. : “Interactive Korean to Japanese Translation Tool Using Sequential Selection Method”, PACLING95 (to be appeared).

[6] C. C. Soo : MS-DOS 実践教室, PC ワールド出版部 (1992)

[7] CIM 開発研究会: CIM 戦略— 藤沢工場의 挑戦 —, 하이테크정보 (1989) in Korean.