日韓機械翻訳における活用語処理のための拡張翻訳テーブルの改善

金 政仁 *　　権 泰光*　　大駒 誠一*
kim@ae.keio.ac.jp　　kwon@ae.keio.ac.jp　　okoma@ae.keio.ac.jp

1　はじめに

日本語と韓国語は文法的に似ている言語である．両国語の類似性をうまく用いることにより，日韓機械翻訳では構文解析や意味解析のかなりの部分を省略することができる．特に，品詞の分類や語順がほとんど同じであることは1対1の直接対応による翻訳処理を可能とする．しかし，両国語には活用ルールが異なること，多義性を持っている単語の下位区分が一致していないこと，助詞の使い分けが一致しないことなどの相違点があり，直接翻訳方式を用いた日韓機械翻訳システムを実用化するためには，もう少しの活用語の処理，多義語の処理などを必要とする．活用語の処理に対しては，両国語の述部表現の意味対応による活用語処理 [李　義東，中嶋正之，安居院猛 1990]，韓日機械翻訳に適用した音韻表現形式による述部の活用処理 [李　秀絃，小沢慎治 1992] などが提案されたが，韓国語の用言の不規則的な活用をうまく表現しきれなかった．用言の不規則的な活用をルール化することを避ける方策として，訳語をあらかじめ翻訳テーブルに用意する活用語の処理 [金　泰錫，浦　昭二 1992] が提案され，その有効性が確認された．また，著者らは翻訳テーブルを改良した拡張翻訳テーブルを用いた翻訳方式を提案した [金　政仁，金　泰錫，大駒誠一 1995]．多義語の処理としては，語と語の関係を用いた意味解析による単文の多義語処理 [李　義東，中嶋正之，安居院猛 1989]，体言の意味素性を羅列した格形式パターンを用いた動詞の多義性処理 [金　政仁，大駒誠一 1992] が提案された．本論文は，日韓直接翻訳方式で，活用語処理に有効であると確認された翻訳テーブル方式から，更に隣接2単語までの接続関係を記述することができるように拡張翻訳テーブルの構造を新たにし，隣接2単語までの接続ルールに従った日韓直接翻訳方式について述べる．

*慶應義塾大学 理工学部 管理工学科，横浜市港北区日吉3-14-1

2　類似性を用いた日韓機械翻訳

類似性を用いた日韓機械翻訳は，構文解析や意味解析のかなりの部分が省略できる長所の代わりに，翻訳の際に利用できる情報が少ないという短所がある．直接翻訳方式で利用する翻訳情報としては，品詞，体言の意味素性，用言の活用型，助詞・助動詞・記号の一連番号である．これらの情報を用いて，前後に接続された単語との接続関係を考慮しながら翻訳処理を行なう．体言の意味素性は，１９種類に分けられているIPAL動詞辞書の分類法を引用した．接続関係の正確な記述のため，活用型の分類は学校文法より細かく分類した．分類基準は形であり，例えば，連用形が連用１、連用２、連用３とあるのは，形容動詞の連用形が３種類あって，それを区別するためである．(「重要だ」の連用形:重要だっ，重要に，重要で)．助詞および助動詞はその数も少ないし，各々が接続されることによって訳語が変わることが多いので，一連番号を与えて接続関係を詳しく記述するために用いた．

- 品詞:名詞 (サ変名詞)，準体言 (の)，動詞 (上１，下１，カ行変格，サ行変格，５段)(可能，進行，補助)，形容詞，形容動詞，助動詞，助詞 (格，接続，副，終)，副詞，連体詞，接続詞，感動詞，記号など

- 体言の意味素性:具体名詞 (動物，人間，組織及び機関，植物，生物の部分，自然物，生産物及び道具)，現象名詞，抽象名詞 (動作及び作用，精神，言語作品，性質，関係，空間及び方角，時間，数量)，その他

- 用言の活用型:未然1，未然2，未然3，連用1，連用2，連用3，終止，連体，仮定，命令1，命令2，語幹

- 助詞，助動詞および記号の一連番号:助詞58個，助動詞19個，記号30個

筆者等が先に提案した、従来の翻訳テーブル方式は，品詞別に6種類に分けられていたが，拡張翻訳テーブル方式では，機能別に2種類の翻訳テーブルに分けた．図1に従来の翻訳テーブル方式の概念を，図2に拡張翻訳テーブル方式の概念を示す．従来の翻訳テーブル方式では，翻訳対象単語の品詞によって，まず翻訳テーブルの種類を決定し，その翻訳テーブルから適切な訳語を選択するため，接続ルールを翻訳システムの中に記述した．拡張翻訳テーブルを用いた日韓直接翻訳方式は，単語別の1対1対応による翻訳処理を基本とするため，品詞に関わらず，全ての日本語の単語の代表訳語をエントリテーブルに用意する．活用語に対しては，活用型別の代表訳語をエントリテーブルに記述する．そして，1対nの対応関係を用いた単語には，候補単語らおよび接続ルールを接続情報テーブルに用意する．拡張翻訳テーブルの主な改善点は，次のようである．

- 翻訳テーブル方式では品詞別の接続ルールを翻訳システムに記述し，複数の候補単語だけを翻訳辞書に記述したが，拡張翻訳テーブル方式では，代表訳語はエントリテーブルに，候補訳語と接続ルールは拡張翻訳テーブルの中に各々記述した．品詞別ではなく，単語別の接続ルールが記述できる．

- 翻訳テーブル方式では品詞別に異なる構造の翻訳テーブルを用いたが，拡張翻訳テーブル方式では，全ての品詞の翻訳テーブルの構造を統一させ，翻訳処理の一貫性を持たした．

- 従来の翻訳テーブルはその単語の原型が翻訳テーブルのエントリになっていたが，拡張翻訳テーブルでは，活用型をエントリとした．即ち，翻訳システムの中で原型から活用語に変換する処理を省略させ，各単語別の形による1:1対応処理を基本とした．

図1: 従来の翻訳テーブル方式の概念

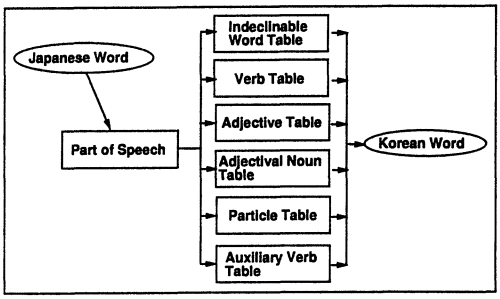
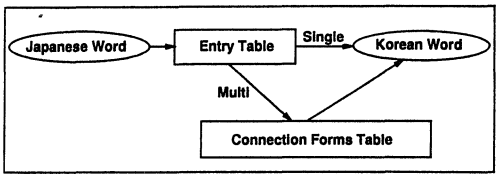


図2: 拡張翻訳テーブル方式の概念



3 接続関係を用いた拡張翻訳テーブル方式の改良

日本語の用言，助詞，助動詞を適切な韓国語に翻訳させるために提案した従来の翻訳テーブル方式は，前後に接続されている複数の単語らとの意味接続関係を用いた．しかし，ほとんどの場合，翻訳対象単語は前後に隣接した単語の情報によって，適切な訳語を選ぶことができる．そして，前接単語より前に接続された単語(以下，前前接単語)との接続関係，後接単語より後に接続された単語(以下，後後接単語)との接続関係など，いわゆる，隣接した2単語までの接続関係を用いることによって，ほとんどの日本語の単語が適切な韓国語に翻訳される．拡張翻訳テーブルに隣接2単語との接続ルールを記述できるように改良する．

3.1 隣接2単語との接続関係の記述

用言，助詞，助動詞などを翻訳するためには，隣接した単語の情報が重要な役割を果たす．しかし，隣接した単語の情報だけでは適切な韓国語を選択するのに充分な情報にならない場合がある．接続助詞「て(で)」は，用言の後ろに接続され，表1のように後接関係を

表 1: 後後接単語との接続関係

活用語	後接単語	後後接単語	訳語
行っ	て	,(記号)	GaSeo
行っ	て	から(助詞)	GaGo
行っ	て	も(助詞)	Ga
行っ	て	は(助詞)	GaSeo
行っ	て	見る(補助動詞)	Ga
行っ	て	しまう(補助動詞)	Ga
行っ	て	いる(補助(進行)動詞)	GaGo
行っ	て	来る(動詞)	Gass

表 2: 隣接 2 単語との接続関係の形態

区分	前前接	前接	翻訳対象	後接	後後接
$CF0$			W		
$CF1$		α	W		
$CF2$			W	β	
$CF3$		α	W	β	
$CF4$	α'	α	W		
$CF5$			W	β	β'

考慮した用言の訳語を選択するのに同じ情報を与えてしまう．この現状は過去，完了，状態などを表す助動詞「た(だ)」，丁寧を表す助動詞「ます」などにも現われる．隣接する単語だけで接続関係を表すことができない場合，1) 2 単語の複合語化，2) 前前接単語および後後接単語 (隣接 2 単語) との接続関係を記述する方法が考えられる．ここで 1) 単語の複合語化は，形態素解析の結果から複合語を作るための作業が必要であるか，複合語を考慮した形態素解析をする必要が出てくるので，よい解決策とは言えない．2) 隣接 2 単語との接続関係を用いる方法は，形態素解析の部分を触らずに翻訳辞書の構造を新たにすることで実現可能である．実際に翻訳対象単語の前後に接続されている隣接 2 単語を次のように表示すると，

$$\alpha' \ \alpha \ W \ \beta \ \beta'$$

(α':前前接単語，α:前接単語，W:翻訳対象単語，β:後接単語，β':後後接単語)

翻訳対象単語 Wに対しては，表2のように接続形態別に分けることができる．

ここで前前接情報α'，前接情報α，後接情報β，後後接情報β' を CF(Connection Form) の形式として翻訳辞書の中に記述し，翻訳処理のとき，隣接 2 単語との接続情報と比較することによって適切な訳語が選択できる．

3.2 翻訳処理の例

適切な訳語を選択するために接続関係を評価し，最も高い数値が得られた訳語を選択するが，評価値が同じである場合は先に評価された訳語を優先させる．すなわち，評価値が同じであるときは，辞書に登録した訳語の順番が優先順番になる．表??は翻訳過程の例を表している．日本語の文「学校に行きませんでした.」の中で活用語「行き」を翻訳するために用いられる隣接 2 単語の接続情報は

$$\alpha'(\text{学校})+\alpha(\text{に})+\underline{W(\text{行き})+\beta(\text{ませ})}\ +\beta'(\text{ん})$$
$$\text{HakGyoE } \underline{\text{GaJi}} \text{ AnhSeupNiDa}$$

であり，接続形態 $CF2$ に従い，後に「ませ」が接続された状況から「GaJi」に翻訳される．

しかし,「でし」を翻訳する場合は

$$\alpha'(\text{ませ})+\alpha(\text{ん})+\underline{W(\text{でし})+\beta(\text{た})}\ +\beta'(.\ )$$
$$\text{Anh}\underline{(\text{A/Eo})\text{ssSeupNiDa}}$$

の接続関係から後接した「た」の接続関係で「(A/Eo)ssSeupNi」に翻訳するが，日本語の文が「学校に行きませんでしたので」に変わった場合は対応できない．この場合は，$CF5$ の接続形態からβ'(ので) の情報を接続関係に加えて「でし」は「(A/Eo)ss」に翻訳し,「た」は $CF3$ の接続関係を用いて韓国語の「Gi」に翻訳させる．

$$\alpha'(\text{ませ})+\alpha(\text{ん})+\underline{W(\text{でし})+\beta(\text{た})\ +\beta'(\text{ので})}$$
$$\text{Anh}\underline{(\text{A/Eo})\text{ssGi}}\ \text{TtaeMunE}$$

また,「ませ」+「ん」+「でし」+「た」の 4 回繰り返しの助動詞同士は,「丁寧+否定+丁寧+過去」の順序であるのに対し，韓国語は「否定+過去+丁寧」の順序である．この時,「ませ」は後接単語が「ん」であれば，その訳語を「NULL」にする．そして，助動詞「ん」を翻訳する時，後接単語が終了の記号であって前接単語が「ませ」であれば，韓国語の打ち消し語「AnhSeubNiDa」に翻訳する接続ルールを助動詞「ん」のところに記述する．即ち，丁寧の意味を持った「ませ」の意味は「ん」に移して翻訳することができる．文の終了でない場合は,「ん」を「Anh」に翻訳し，次の「でし」を「(A/Eo)ssSeubNi」に,「た」を「Da」に翻訳できるように接続ルールを記述する．このように，隣接単語に自分の意味を移した接続ルールを記述することができるので，助動詞の連結順序が異なる場

表 3: 助動詞「です」のエントリテーブル

原型	翻訳対象単語	品詞	活用型	番号	代表訳語	マルチ	例
です	でしょ	助動詞	未然1形	19	(I)GessJi	N	行くでしょう
	でし	助動詞	連用1形	19	(I/Y)EossSupNi	Y	本でした
	です	助動詞	終止形	19	IpNiDa	Y	本です

表 4: 助動詞「です」の接続関係テーブル

翻訳対象単語					第1接続関係			第2接続関係				
単語	品詞	活用型	番号	接続関係	品詞	活用型	番号	品詞	活用型	番号	候補訳語	例文
でし	助動	連用1形	19	$CF3$	助動	終止形	12	助動	終止形	13	(A/Eo)ssSeupNi	行きませんでした.
	助動	連用1形	19	$CF5$	助動	終止形	12	助詞		15	(A/Eo)ss	行きませんでしたので
です	助動	終止形	19	$CF1$	助動						SeupNiDa	行かないです
	助動	終止形	19	$CF1$	形動						HapNiDa	静かです
	助動	終止形	19	$CF2$				助詞		15	(I)Gi	東京ですので
	助動	終止形	19	$CF2$				助詞		15	(I)Gi	東京ですから

合も自然な訳文の生成が可能である. 表3と表4は助動詞「です」のエントリテーブルと接続情報テーブルの一部である.

4 むすび

日韓機械翻訳は, 両国語の類似性を用いた直接翻訳方式がよく使われている. 類似性をうまく用いることにより構文解析や意味解析のかなりの部分の省略ができ, シンプルな翻訳システムの構築が可能である. しかし, 活用語については, 活用語幹と語尾を分離して翻訳処理をすることでは, 自然な訳語の生成ができなかった. そこで、著者らは前後単語の意味接続関係を考慮してあらかじめ韓国語を用意する翻訳テーブル方式を提案し, その有効性を確認した. しかし, 翻訳テーブルを用いた翻訳処理方式は品詞に大きく依存するシステムになり, 品詞別の特殊な処理が多かった. 訳語の選択ルールも翻訳システムの中に品詞別に記述した.

本論文は, 品詞に依存しない拡張翻訳テーブルに, 隣接2単語との接続ルールを単語別に記述することを提案した. このことで, 隣接2単語の属性を用いて, 辞書に書かれている候補単語別の接続ルールとの比較処理だけで, 適切な訳語の生成ができる. 今後, 用言や助動詞などでよく現われる多義性の問題などが処理できるように改良を続けて, 両国語の類似性を用いたシンプルな日韓機械翻訳システムの実用化を向けて, さらに改良していくつもりである.

参考文献

[Choi.k.s, ほか] 日本語翻訳システム環境下での韓国語翻訳システム開発のための一考察, 情報処理学会, 自然言語処理研究報, Vol 86,No 4,1988

[李 義東, 中嶋正之, 安居院猛 1989] 語と語の関係を用いた意味解析による日韓単文機械翻訳システム, 電子通信学会論文誌 (D-II) Vol 72,No 10,1989

[村田賢一, ほか 1987] 計算機用日本語動詞辞書 IPAL(Basic Verbs)- 解説編, 情報処理振興事業協会技術センター

[李 義東, 中嶋正之, 安居院猛 1990] 助詞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,pp 801-809,1990

[李 秀絃, 小沢慎治 1992] 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. 33,1565-1577

[金 泰錫, 浦 昭二 1992] 日韓機械翻訳における意味接続関係を用いた韓国語の生成方法, 情報処理学会論文誌,Vol. 33,1578-1588

[金 政仁, 大駒誠一 1992] 日韓機械翻訳における動詞の多訳性処理, 情報処理学会４５回全大集,Vol. 3,97-98

[金 政仁, 金 泰錫, 大駒誠一 1995] 拡張翻訳テーブルを用いた日韓機械翻訳, 情報処理学会５１回全大集,Vol. 3,89-90