

外97-19

早稲田大学大学院理工学研究科

博 士 論 文 概 要

論 文 題 目

言語特性を利用した
日一韓機械翻訳システムに関する研究

申 請 者

朴	哲 済
Chul-Jae	PARK

1997 年 10 月

機械翻訳システムにおいては、処理する言語に関する情報をどれほど豊かにそなえているかが、そのシステムの性能に大きな影響を与える。とくに分かち書きをしない日本語では、その形態素解析だけのためにも膨大な量の辞書データをそろえる必要がある。しかし、辞書データの蓄積は、自動的に行うことが困難であり、人手による膨大な時間と労力を必要とする。幸い、最近では公開の辞書データの入手も可能となってきたが、それでもなお、新しい文法体系を試みるような場合には、その辞書を用意するのに手間がかかりすぎて、本題の研究にかかれないことがおきる。

また、直接翻訳方式を利用して機械翻訳システムを構築するとき、直面する問題点として、語彙の曖昧性と、部分的な語順の調整問題があげられる。この問題点の中で高品質の翻訳結果を得るためには、語彙の曖昧性の問題を解決することが重要である。語彙の曖昧性の問題は大きくカテゴリの曖昧性、多義性、および、多訳性の3つに分けられる。この中で、カテゴリの曖昧性は形態素解析の段階でヒューリスティックスを用いてほとんど解決できる。反面、多義性や多訳性の問題は構文と意味情報を基に解決しなければならない。語彙の曖昧性の問題は日-韓機械翻訳においても一番大きな問題点として残っている。

このような観点から本論文では、日本語と韓国語が極めて類似していることから日-韓機械翻訳システムが軽装かつ高精度に構成できることを実証した。具体的には形態素レベルでの変換を核とし、そこに統計モデルに基づく形態素解析と連語パターン知識に基づく変換生成の新手法を開発実装して、難文コーパスを対象に在来システムの約2倍の成功率を収めた。また、機械翻訳システムでの性能を左右する辞書データの自動蓄積に関しても新手法を提案・実現してその有効性を示した。

本論文は5章より構成される。以下にその概要を示す。

第1章では、機械翻訳システムの現状とその技術的課題を明らかにするとともに、機械翻訳システムに関する研究の現状を概観し、本研究の背景と目的を明らかにした。

第2章では、辞書データがほとんどない状態から始めても、大量の日本語テキストを与えることで、形態素に関する辞書データを自動的に蓄積する方法について論じた。具体的には、形態素に関する種々の規則と、統計的知識を利用して、未知の形態素の切出しとその品詞、活用種類、活用形などの推定を行う。推定するたびにその信頼性を評価し、大量のテキストを走査するうちに十分高い信頼性を得るに至ったものを、正しい形態素として辞書に登録する。

現在までに、計算機によって自動的に辞書情報を獲得するいくつかの研究が行われてきている。しかし、初期に用いる辞書データを用意することでも困難であり、人手による膨大な時間と努力を必要とする。本研究では、形態素の接続関係に着目し、ほとんど辞書が整備されていない環境でも形態素とその属性を自動獲

得していける方法を提供した。実際に、システムの最低要件である助詞・助動詞と、約100語の活用をもたない、かながき形態素だけに初期設定し実験システムを構成して、比較的簡易な機構によって目的が達成できた。

実験システムは、朝日新聞の社説6カ月分241,573形態素中228,450形態素について正しくその形態素属性を推定した(成功率94.6%)。とくに活用品詞類に対しての成功率は90.5%であり、その他の品詞に対する成功率は95.2%であった。実験の結果は、ほとんど辞書が整備されていない環境でも形態素とその属性を自動獲得できるシステムを提供する、という目的からすると、十分に満足いくものであった。とくに、これだけ簡易なシステム構成であっても、助詞・助動詞に着目することで多くの情報が自動的に獲得でき、その有効性を明らかにした。

第3章では、改訂CYK法に単語接続に関する統計モデルを利用した日本語の形態素解析手法について論じた。

形態素解析の代表的手法である幅優先探索法として三角行列を用いるCYK(Cocke-Younger-Kasami)法がある。CYK法は、まず、辞書検索によって分割可能なすべての形態素候補を求めたのち、接続情報モデルを利用して隣合う形態素間の結合妥当性を検査して正しい解析結果のみ出力する。したがって、形態素間の結合妥当性検査のために用いる接続情報の正確性によって高い形態素解析成功率が得られる。既存の研究では、主に開発者の言語知識に依存して2つの形態素間の結合可能性を判断して、接続情報表を作成している。しかし、数多い形態素間の結合関係を辞書の中に入れるのは不可能である。また、単純な形態素間の接続有無のみ表現する既存の方式では解析成功率を上げることに限界がある。

本研究では、動的プログラミング技法であるCYK法を改訂した形態素解析を日-韓機械翻訳システムの解析部として構築した。これはまず、接続情報を検査し接続が可能な形態素解析結果をすべて得たのち、ヒューリスティックスを利用して優先順位を決める。本稿では、接続情報表の値を確率として用いることにより接続の強度を表現し、その強度により形態素候補の優先順位を決めた。このとき用いる確率情報は、形態素解析の対象言語に関する確率モデルとして統計情報抽出機構から得られる。このような形態素解析手法を日本語を対象に約24万形態素のコーパスから接続情報を抽出し実験した結果95.2%の解析成功率を得た。また、ビット形式の接続表と確率形式の接続表に分けて比較評価した結果、ビット形式の接続表を用いたより確率接続表を用いた方が1.9%正解率が向上された。全体の形態素解析において第一順位の結果では統計モデルを用いることによって、より自然な解析結果を得ることができ、本確率接続情報を利用した形態素解析手法の有効性を確認した。

第4章では、直接翻訳方式において連語パターンによる日-韓機械翻訳システムの実現と評価について論じた。

本稿では、日本語での連語パターンを作成し、変換規則として用いることによ

り語彙の曖昧性の問題を解決した。また、意味素性をベースにした選択制限として、連語パターンの中で最もパターン類似度が高いものを選択する類似度のスコア計算の方法を提案した。連語パターンでは語と語の接続関係をすべての単語間(動詞-名詞, 名詞-助詞など)にも選択制限として用いた。本システムでは、連語パターンを人が簡単に理解できるように記号で定義し、それを直接辞書に記述して曖昧性の問題を解決した。また、意味素性をベースにした選択制限として連語パターンの中で一番パターン類似度が高いものを選択する類似度計算の方法を提案した。言語構造が似ている日-韓翻訳においては、連語パターンで処理するだけで、本格的な構文解析や意味解析を行わず、実用的には高品質の翻訳が可能であることを確認した。

韓国語生成は、意味素を利用して述部の生成における順序関係の問題を解決した。日本語述部と韓国語述部との順序の違いは、様相類意味素テーブルに表現されている順位関係を利用することで解決した。一つの日本語形態素がいろいろな意味を表現している場合は、様相類意味素テーブルに多くの意味素を活性化して解決した。また、日本語用言に否定の助動詞が使われそれに対応する韓国語に否定的訳語が存在する場合は、否定的訳語として生成することによりもっと自然な韓国語表現が得られた。韓国語用言の不規則処理は、韓国語形態素の接続情報に現われた不規則情報を用いて処理した。助詞と語尾の異形態処理は、韓国語の接続可能性を調査し処理した。

評価実験では、個々のレベル(形態素, 単語, 熟語)での成功率から翻訳システム全体を評価する方法として、荷重により総合評価点数を求めた。システムは、8,319 日本語形態素中 7,831 形態素について正しく韓国語に翻訳した(成功率 94.13%)。本システムで付加した荷重による総合評価点数は 93.46 であった。これはシステムをトレーニングする前の翻訳結果としてはかなり高い。他のシステムとの翻訳性能の比較は、難しいのが現状である。しかし、翻訳性能は 2 つの観点からチェックすることができる。(1) 実際の文章でどの程度の翻訳率なのか。(2) 翻訳するのに難しい文章をテストコーパスとして使って、どの程度の翻訳率なのか。現在(1)の方法では、我々のシステムと他のシステムはほぼ同じ精度であった。(2)の方法では、我々のシステムの精度が最も高い。実際、2 つのシステムを比較し、本システムの問題点を見つけるため比較評価を行った。その結果、我々のシステムは在来システムの約 2 倍の成功率を収めた。

最後に、第 5 章では、本研究を総括し、今後の研究の展望について論じた。本研究で達成された成果は、日-韓機械翻訳システムにおいては両国語の言語特性を利用することで効果的なシステムが実現できたことである。また、機械翻訳システムの性能に大きな影響を与える辞書データを自動的に蓄積する手法に関してもその有効性が明らかになった。