特許文書向け韓国語翻訳システム

Korean Translation System for Patent Documents

東芝ソリューション株式会社　プラットフォームソリューション事業部参事　**熊野　明**

PROFILE: 1982 年東京工業大学卒業。同年東京芝浦電気（株）（現（株）東芝）入社。2010 年から東芝ソリューション（株）プラットフォームソリューション事業部参事。自然言語処理システムの研究開発に従事。アジア・太平洋機械翻訳協会理事。AAMT/Japio 特許翻訳研究会メンバー。2007 年度から Japio 特許版・産業日本語委員会委員。

1　はじめに

近年、日本と中国・韓国の双方から企業進出や投資が増加し、それに伴い多くの技術情報が交換されている。その結果、中国語・韓国語と日本語との間の翻訳の必要性がますます高まっている。例えば、中国の発明特許出願件数は 2011 年に世界で最多となり、また、韓国の発明特許出願は、中国、米国、日本に続いて４番目の件数を占めている[1]。最新の技術情報を理解するためには、これらの情報を活用することが必須である。そのため、中日・日中および韓日・日韓翻訳の翻訳負荷の低減による業務効率化やコスト削減が求められている。

これを解決するために、われわれはこれまで中国語翻訳システムを開発し、精度向上を図ってきた[2][3]。本稿では、新たに開発した韓国語翻訳システム[4]について述べる。なおこのシステムは、The 翻訳シリーズの韓日・日韓機械翻訳製品[5]として提供している。

2　韓日機械翻訳の処理方式

韓国語から日本語への機械翻訳システムについて、その方式を説明する。処理の概要を図１に示す。

まず韓国語原文に対して単語辞書を参照して形態素解析を行い、文を構成する単語を認識する。次に、この単語列に対し、原文の語順を利用して日本語としての語順を決める。続いて、各韓国語単語に対して、適切な日本語の訳語を決定する。最後に、原文の形態素解析で得られた各単語の属性をもとに、日本語の訳語に対して形態素生成を行う。このようにして得られた表現を接続して、日本語訳文を完成する。

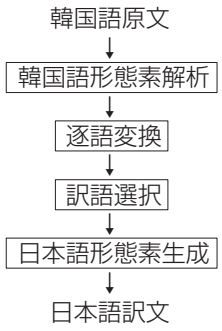

図１ 韓日機械翻訳の処理

以下詳しく紹介する韓国語から日本語への翻訳方式は、英語や中国語から日本語への翻訳方式と異なる点がある。それは、韓国語と日本語の特性によるものである。

2.1　韓国語形態素解析

韓国語の形態素解析では、中国語の形態素解析同様、統計情報を利用した解析を行う。

単語辞書と接続テーブルを参照して、形態素ラティスを構築する。この形態素ラティスには複数の形態素系列が含まれているので、その中からもっともらしい単語・品詞系列を選択する必要がある。

一般には、接続テーブルだけでは完全に曖昧性を解消することができない。確からしい結果を得るために、形態素の間の接続コストを最小にする系列を選択する。この接続コストは、韓国語のタグ付きコーパスから統計的に学習することができる。

2.2 逐次変換

英語や中国語では、精密な構文解析処理により、原文中の語の間の係り受け関係、修飾関係を明らかにし、原文の構文構造を出力する。この構文構造を語彙知識、構造知識を使って日本語の構文構造に変換する。その後、日本語の文法に基づいて語順を決定し、必要な語尾変化を施して訳文を出力する。精密な構文解析処理と構造変換が必要な理由は、英語や中国語の表層構造は、日本語の構造と大きく異なるからである。

これに対して韓国語は、その語順が日本語の語順と極めて類似している。したがって、韓国語の語順をできるだけ利用し、日本語訳文を出力する。形態素解析を行って得られた単語列に対して、明示的な構文解析や構造変換は行わず、直接日本語訳文となる単語列に逐次変換する。

2.3 訳語選択

韓国語を表記するハングルは、日本語や中国語の漢字（表意文字）と異なり、表音文字である。このため、一つのハングル表記の単語に対して、複数の異なった意味の可能性がある。また名詞の多くは漢語であり、日本語と発音が類似しているものがあることも特徴である。

表 1 に、同音異義語の例を示す。

表 1 韓国語同音異義語の例

韓国語	品詞	日本語（例）
기구	名詞	機構、器具、気球
검사	名詞	検査、検事、剣士
독자	名詞	独自、読者
쓰다	動詞	使う、かぶる、書く
들다	動詞	入る、要る、持つ

日本語に翻訳する際には、これらの多義を解消しなければならない。この曖昧性解消には、いくつかの方法が考えられる。

ここでは、「이름을 쓴다」の翻訳を例にあげて説明する。前半の「이름을」は、「名前を」を意味する表現であり、ほぼ曖昧性はない。これに対して、「쓴다」(基本形「쓰다」)には複数の意味の可能性があり、その中から正しい意味を選択することが必要である。

2.3.1 意味的制約

「쓴다」には、「使う」「かぶる」「書く」の多義がある。この語と直接係り受け関係のある「이름을」（名前を）との意味的制約を利用することで、正しい意味を選択することができる。

日本語の構造に関する十分な意味情報を利用すれば、「名前を使う」、「名前を書く」は意味的にもっともらしいが、「名前をかぶる」は意味的に可能性が低いと判断できる。このことから、「名前を使う」または「名前を書く」が正しい解釈の候補と考えることができる。

しかしこの判断には、（1）韓国語の正確な係り受け解析、（2）日本語の十分な意味知識が必要である。係り受け解析には、日本語の構文解析に相当する高精度の韓国語構文解析技術が必要であり、開発コストが大きい。また日本語の十分な意味知識を構築するにも大きなコストがかかる。

2.3.2 統計的言語モデル

「이름을 쓴다」の訳文には、「名前を / 使う」、「名前を / かぶる」、「名前を / 書く」の可能性がある。これらの表現のうち、日本語の表現として通常使われないものは正しい解釈である可能性が低い。つまり、日本語で最も多く使われているものが、もっともらしい解釈であると考える。

この判断のためには、大規模な日本語テキストを形態素解析して作成するタグ付きコーパスが必要である。現在日本語形態素解析の精度は十分に高く、ここで必要なタグ付きコーパスを構築することは困難でない。

われわれのシステムでは、この統計的言語モデルで、訳語選択を行う。大規模な日本語コーパスをもとに、単語の並びから統計情報を学習し、言語モデルとして構築した。訳語選択では、この言語モデルを参照することにより、複数の訳語候補から適切なものを選択することが可能になった。

図 2 に、ここで述べた統計的訳語選択の処理の概要を示す。

2.4 日本語形態素生成

訳語選択で決まった日本語訳語のうち、動詞や形容詞など語尾変化のあるものに対して、形態素生成を行う。

韓国語形態素解析の際に得られた各単語の属性を利用

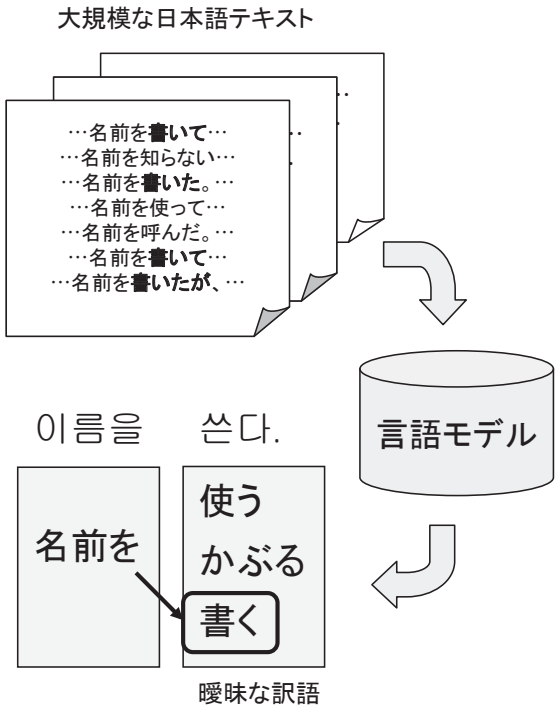

図 2 統計的訳語選択の処理

して、日本語の活用語に対して適切な語尾変化や助動詞の付与などにより、最終的な訳語文字列を決定する。

ここで得られた各訳語の文字列を連結し、最終的な日本語訳文を出力する。

3 特許文書向け翻訳システム

特許文書を対象とした韓日翻訳システムは、韓国語の特許文書の特徴と、日本語の特許文書の特徴を考慮しなければならない。

韓国語に限らず、特許文書には専門用語や新語が多く記述されている。また、特許文書特有の表現も含まれている。特許文書向け韓日翻訳システムは、これまでに述べたものに、特許文書に向けた言語知識の強化、言語モデルの適応を加えたものである。

3.1 言語知識の強化

特許文書には、分野に応じた専門用語が多く出現する。これらの専門用語を正しく翻訳することは、機械翻訳精度を大きく左右する。特に複数の単語が連続してなる複合語を辞書に登録することで、構成語の訳語の曖昧性が減り、確実な訳語を出力することができる。

今回、基本語辞書に加えて、12 分野の専門用語辞書と特許専用の用語辞書を整備した。

3.2 言語モデルの適応

統計的訳語選択で利用する言語モデルは、学習対象とする日本語タグ付きコーパスを変更することによって調整が可能である。

特許文書に適した訳語選択を実現するために、大量の日本語特許文書を学習対象テキストに追加してタグ付きコーパスを作成し、新たな言語モデルを構築した。これを利用することにより、特許文書に適応した日本語訳語・表現を優先して出力することができる。

これらの言語知識強化、言語モデル適応によって、特許文書向け韓日翻訳システムを実現することができた。このシステムは、多くの専門用語と特有の表現を含む特許文書に対応した、高い精度の翻訳を提供することができる。

4 おわりに

韓国語と日本語の特性を利用した韓日翻訳システムを開発した。さらに、特許文書の翻訳に必要な言語知識、言語モデルを追加構築した。この結果、韓国語の特許文書を高精度で日本語に翻訳するシステムを実現することができた。今後さらに増加する韓国語翻訳のニーズに対して、翻訳業務の効率化やコスト削減に貢献していく。

なお本システムは、東芝ソリューション（株）の機械翻訳システム“The 翻訳エンタープライズ™”として製品化している。

参考文献

[1] WIPO: World Intellectual Property Indicators 2012（2012）
[2] 出羽，熊野：中日・日中機械翻訳システム，東芝レビュー Vol.64 No.7（2007）
[3] 熊野：中国語特許翻訳を支援する機械翻訳技術，Japio YEAR BOOK 2011（2011）
[4] 韓日・日韓翻訳技術，東芝レビュー Vol.68 No.3（2013）
[5] 東芝ソリューション（株）：The 翻訳エンタープライズ
http://mt-server.toshiba-sol.co.jp/（2013）