中国特許の調査や特許出願を効率化する機械翻訳技術

Machine Translation Technologies for Chinese Patent Searching and Application

熊野　明
■KUMANO Akira

複雑な技術表現を含む特許文書を翻訳するためには，特許翻訳専用の機能と知識が必要である。
東芝ソリューション（株）は，中国特許固有の機械翻訳技術を開発し，中国特許の調査や，特許出願の効率化を実現した。中日翻訳エンジンには，発明の名称や発明者の名前を正しく翻訳するために，文書構造を利用した中国語解析方式を開発した。また，日中翻訳エンジンには，日本語の長文表現を構成要素に分割する自動前編集機能を，日英翻訳エンジンから導入した。更に，特許明細書特有の定型表現を翻訳するために，両方向のエンジンに特許専用の翻訳メモリを開発した。

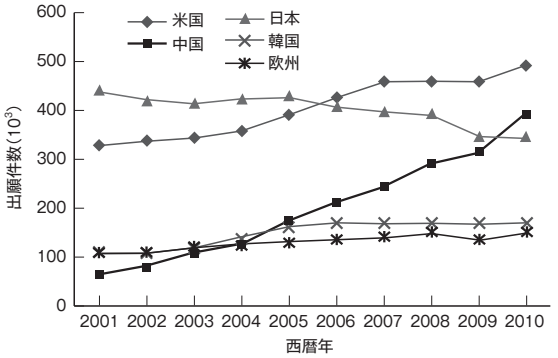
In order to translate patent documents containing complex technical expressions, dedicated functions and knowledge specific to patent documents are essential.

Toshiba Solutions Corporation has developed machine translation technologies for Chinese patents to improve the efficiency of patent searches and applications, and released the Hon-Yaku Enterprise V15 translator capable of Japanese-to-Chinese (J-to-C) and Chinese-to-Japanese (C-to-J) translation. To translate invention titles and inventors' names correctly, we have developed a sentence analysis technique using Chinese document structure in the C-to-J translation engine. We have also introduced an automatic pre-editing function to divide long Japanese patent sentences in the J-to-C translation engine. Furthermore, we have developed translation memories for fixed phrases specific to patent specifications in both translation engines.

1　まえがき

近年，東アジア諸国は急速な経済発展を続けており，様々な産業において，諸外国からの技術進出，製品販売，及びビジネス展開が急激に進んでいる。なかでも巨大市場であり，安価な労働力を提供できる中国では，2000年頃からわが国の企業を含む先進国企業によるビジネス展開が加速した。その結果，多くの労働者の雇用を生み出し，都市への人口流入を加速している。また，先進国からの技術やノウハウを導入して，これまでにない技術革新により飛躍的な成長を遂げた企業がある。現在では，中国市場に参入した外資系企業よりも大きなビジネスを行うまでに成長した企業も多い。これらの経済成長は，低成長時代を迎えた欧米やわが国に代わって，世界の経済成長を推進する原動力になっている。

中国の経済が発展するに従って，ビジネスを左右する知的財産への注目が高まっている。これらの事実は，特許出願件数のデータにも表れており，2000年以降，中国特許庁への特許出願が急増している。主要5大特許庁での2001年から2010年の特許出願件数の推移を**図1**に示す[1]。中国での特許出願は，2001年にはわが国の15％にも満たない63,450件であったが，年平均20％以上増加し，2010年には391,177件まで伸びた。一方，増加を続けていたわが国の特許出願は2001年の440,248件がピークで，2005年以降減少を続け，2010年には344,598件までに減った。その結果，米国に次ぐ出願件数世界第2位の座を中国に譲るに至った。



出典：World Intellectual Property Organization
「World Intellectual Property Indicator 2011」[1]

図1．主要5大特許庁の特許出願件数の推移 ― 中国での特許出願数が2010年にわが国での出願数を上回った。
Trends in number of patent applications in top five countries

中国に出願される特許は，中国ビジネスを続ける企業やこれから始める企業にとって，重要な技術情報である。しかし日本人にとって，中国特許の調査や出願の業務は大きな課題である。言うまでもなく中国語という言語と，特許という複雑な技術表現に起因する課題である。

課題の場面の一つは，中国特許の調査における内容理解である。キーワードや分野情報で関連特許を絞り込むことはで

きるが，絞り込まれた中国特許を十分に理解することは容易でない。日本語への翻訳を翻訳会社に依頼することが多いが，様々な専門分野に対して十分な翻訳結果を得るためには，大きなコストがかかる。

もう一つの場面は，中国への特許出願である。中国に特許出願する場合は，中国語の特許明細書が必要である。中国語で明細書を作成できるわが国の技術者は，英語に比べてはるかに少ない。既に日本語の明細書が作成済みの場合，中国語への翻訳を翻訳会社に依頼することもできるが，やはりコストが課題である。

これらの課題に対して，特許庁では機械翻訳の活用を検討しており，実際の翻訳結果を評価して問題点を整理している[2]。

東芝ソリューション（株）は，従来，英日・日英機械翻訳システムに，特許文書用の機能を開発し，提供してきた[3]。更に，前述の中国特許の課題を軽減できるように，中日・日中機械翻訳システムに中国の特許文書専用の機能と知識を強化した。

ここでは，中国特許文書の翻訳に関わる技術について述べる。これらの技術は，The翻訳エンタープライズ$_{TM}$V15[4]で提供している。

2　中日・日中機械翻訳システムの構成

機械翻訳の方式には，規則ベース機械翻訳や，例文ベース機械翻訳，統計的機械翻訳などがある。The翻訳$_{TM}$シリーズの中日・日中機械翻訳エンジンの基本部は規則ベース機械翻訳であり，次の三つの処理を経て翻訳する[5]。

(1)　原文の意味を理解して構造化する原文解析
(2)　原文の構造を出力言語の構造に変換する構造変換
(3)　出力言語の構造から訳文を出力する訳文生成

これに加え，定型表現の翻訳のために，翻訳メモリを使った例文ベース機械翻訳も採用している。翻訳メモリとは，過去の翻訳結果をもとに，原文に対する直接の訳文を集めたデータの集合であり，入力文が原文データと一致すれば，訳文データをそのまま出力する。

新語や専門用語を多く含む特許文書を翻訳する場合，専門用語辞書の拡充は不可欠である。しかし，辞書の拡充だけでは解決できない問題も多い。中国特許の調査や出願をするうえで，高精度の機械翻訳を実現するためにはいくつかの課題がある。

3　中国特許調査を支援する中日機械翻訳

中国特許の調査業務では，中国特許文書を中日機械翻訳して内容の理解を支援することで，効率化できる。既存の中日機械翻訳システムをもとに，特許翻訳用に新たに開発した技術とその効果について述べる。

3.1　文書構造を利用した中国語解析

中日機械翻訳の精度を左右するいちばんの課題は，原文解析での単語分割と品詞判定である。中国語は全て漢字で表記され，単語の間に空白は存在しない。したがって，単語分割の曖昧性が問題になることが多い。更に，同じ表記の単語でも品詞の曖昧性を持つものがある。例えば，名詞の多くは動詞としても機能する。動詞の一部には，機能語や副詞になるものもある。

特許明細書での発明の名称は，特許明細書の主旨を的確に表現したものである。技術の高度化や細分化に伴い，構成要素や特徴を記述した長い表現になる傾向がある。長い表現になるほど，原文解析の曖昧性は増大する。名詞句として記述された表現でも，機械的には動詞句や介詞（英語の前置詞相当）句として解釈できる場合もある。

中国特許明細書での発明の名称と発明者名の記述例を**図2**に示す。この記述例を，発明の名称や発明者の記述であるという情報を利用することなく，通常に翻訳した結果を**図3**に示す。

発明の名称は名詞句として記述されているが，介詞句として解釈されているため，訳文としては正しくない。また，発明者名は，動詞を含む文として解釈されている。

このような誤った解釈を軽減するため，言語外情報である文書構造を利用して中国語解析を行う方式を開発した。

発明の名称や発明者は，特許明細書の中でHTML（Hypertext Markup Language）やXML（Extensible Markup Language）によって文書構造が明確に示されている。**図4**は，図2を例にHTML記述の構造を示すものである。

このような文書構造を利用することで，正しい解析ができる。“**发明名称**”（発明の名称）である“**基于加密算法技术的商品防伪方法**”の解析では，名詞句としての解釈を優先することで，介詞句などの誤った構造解釈を避けることができる。

发明名称	基于加密算法技术的商品防伪方法
发明人	龙传红

図2．中国特許明細書の記述例（一部）— 上側が発明の名称で，下側が発明者の情報である。
Example of invention title and inventor's name in Chinese patent

発明の名称	暗号化アルゴリズム技術の商品偽造防止方法に基づいて
発明者	龍は伝わって赤くなる

図3．通常に中日の機械翻訳をした結果 — 発明の名称と発明者が正しく翻訳されていない。
Result of normal C-to-J machine translation

一般論文

```
<TR>
  <TD>发明名称</TD>
  <TD>基于加密算法技术的商品防伪方法</TD>
</TR>
<TR>
  <TD>发明人</TD>
  <TD>龙传红</TD>
</TR>
```

図4．中国特許明細書のHTML記述（一部）— 発明の名称と発明者が文章構造的に明確である。

HTML (HyperText Markup Language) description in part of Chinese patent

発明の名称	暗号化アルゴリズム技術に基づく商品偽造防止方法
発明者	竜伝紅

図5．文書構造を利用した中日翻訳結果 — 発明の名称と発明者が正しく翻訳されている。

Result of C-to-J machine translation using document structure

表1．中日特許翻訳用の翻訳メモリの効果

Effect of translation memory for C-to-J machine translation

原　文	標準辞書での翻訳	翻訳メモリを使った翻訳
其特征在于：	その特徴は次の点にある：	その特徴は以下のとおりである：
以下参照附图说明本发明的实施例。	以下に図面説明 本発明の実施例を参照する。	以下に図を参照して本発明の 実施例を説明する。

また発明者の記述も，“龙传红”が“发明人”（発明者）に対応することがわかるので，人名としての解釈を優先して，固有名詞辞書の見出し語を使って解析できる。図2の情報を，文書構造を利用して中日翻訳した結果を**図5**に示す。

3.2　中国特許特有の表現に対応した翻訳メモリ

中国特許明細書には，特許特有の技術表現が頻繁に使われる。文法的な表現であるかぎり，原文解析，構造変換，及び訳文生成の処理を経て日本語に翻訳できる。しかし，原文の表現が特許特有で，通常の翻訳知識では日本特許に適した訳出ができない場合もある。

このような表現に対して，特許専用の翻訳メモリを開発した。原文の意味を伝える訳文として日本語特許明細書特有の表現を出力することで，中国語特許明細書の理解を高めることができた。

中日特許翻訳用の翻訳メモリの効果例を**表1**に示す。

4　中国特許明細書作成のための日中機械翻訳

中国への特許出願業務は，日本語の特許明細書に対して日中機械翻訳を活用することによって効率化できる。既存の日中機械翻訳システムを基に，特許翻訳用に開発した技術と効果について述べる。

4.1　自動前編集による長文表現の分割

中国への特許出願では，日本語の特許明細書を中国語に翻訳する場合が多い。機械翻訳を利用して日中翻訳する場合，原文が長いほど曖昧性が増加し，訳文精度の低下を招きやすい。これは日中機械翻訳だけの問題ではなく，日英機械翻訳にも共通する問題である。

日本語の特許明細書に現れる長文表現の一例を**図6**に示す。これは，四つの構成要素の並列表現を示したものである。この長文を通常の方法で日中翻訳すると，1文としての解釈に失敗し，解析できる部分に分割する。その結果，**図7**に示すような断片的な訳文を出力し，原文の構造を反映できない。

このような構成要素を列挙した長文に対して，構成要素に自動分割し，分割要素ごとに翻訳する自動前編集機能を開発した。これは，日英機械翻訳用の特許文書向け自動前編集機能[6]を導入して開発したものである。

長文分割には，日本特許文書の表層的特徴を利用した前編集規則を適用する。この例で適用した規則を**図8**に示す。<NPn>は，文中に現れる任意の名詞句を表す。原文パターンにマッチした長文は，前編集パターンに従って再構成されて出力される。この規則を適用して，図6の原文を自動前編集し

【構成】
重量検出装置１において、ターンテーブル 18 と、このターンテーブル 18 上の物体 19 の重量を荷重として受け、この荷重を負荷トルクに変換しながらターンテーブル 18 を回転自在に支持する支持手段 20 と、前記ターンテーブル 18 を負荷トルクに対応した駆動トルクで回転駆動するモータ３と、このモータ３の負荷トルクを検出して負荷トルクと荷重との相関関係からターンテーブル 18 上の物体 19 の重量を間接的に測定する荷重測定装置２とからなる構成とした。

図6．日本語の特許明細書の長文表現例 — 四つの構成要素を並列表記した長い文である。

Example of long sentence in Japanese patent

放重量查出装置 1，作为负荷接受转盘 18 和这个转盘 18 上的物体 19 的重量，把这个负荷一边转换为负荷转动力矩一边转自由自在的支持转盘 18 的支撑手段 20，由在对应负荷转动力矩的驱动转动力矩旋转驱动上述转盘 18 的电动机 3 和查出这台电动机 3 的负荷转动力矩从负荷转动力矩和负荷的相互关系间接地测量转盘 18 上的物体 18 的重量的负荷测量装置 2 变成的构成和下面。

図7．長文に対する通常の日中翻訳結果 — 長文が自動的に分割され断片的な訳文となり，正しい構造を表現できない。

Result of normal J-to-C machine translation of long sentence

[原文パターン]
<NP1>において、<NP2>と、<NP3>と、<NP4>と、<NP5>とからなる構成とした。
↓
[前編集パターン]
<NP1>に、以下を備えて構成する。
<NP2>
<NP3>
<NP4>
<NP5>

図8. 長文分割に適用する前編集規則（例）— 構成要素の併記による長文を，構成要素に自動分割し，分割要素ごとに翻訳する。
Pre-editing rule for long patent sentence

【構成】
重量検出装置 1 に、以下を備えて構成する。
ターンテーブル 18
このターンテーブル 18 上の物体 19 の重量を荷重として受け、この荷重を負荷トルクに変換しながらターンテーブル 18 を回転自在に支持する支持手段 20
前記ターンテーブル 18 を負荷トルクに対応した駆動トルクで回転駆動するモータ 3
このモータ 3 の負荷トルクを検出して負荷トルクと荷重との相関関係からターンテーブル 18 上の物体 19 の重量を間接的に測定する荷重測定装置 2

図9. 日本特許文書の長文を自動前編集した結果 — 構成要素がそれぞれ独立した文になる。
Result of automatic pre-editing of long patent sentence

【构成】
重量查出装置 1，具有以下构成：
转盘 18，
作为负荷接受这个转盘 18 上的物 19 的重量，把这个负荷一边转换为负荷转动力矩一边转自由自在的支持转盘 18 的支持手段 20
在对应负荷转动力矩了的驱动转动力矩旋转驱动上述转盘 18 的电动机 3
查出这台电动机 3 的负荷转动力矩从负荷转动力矩和负荷的相互关系间接地测量转盘 18 上的物 19 的重量的负荷测量装置 2

図10. 自動前編集出力に対する日中翻訳結果 — 自動前編集による分割出力は，機械翻訳の曖昧性が少ない簡潔な表現であり，構成要素ごとに正しく翻訳されている。
Result of J-to-C machine translation of pre-edited sentence

た結果を**図9**に示す。

自動前編集による分割出力は，機械翻訳の曖昧性が少ない簡潔な表現である。これらを日中翻訳した結果を**図10**に示す。分割部分ごとに翻訳することで，原文の表す特許の構造を正しく伝える中国語表現が出力できる。

4.2 日本特許特有表現に対応した翻訳メモリ

日本語の特許明細書においても，特許特有の技術表現が頻繁に使われる。原文の表現が特徴的で，通常の翻訳処理では特許特有の中国語訳が出力できない表現に対して，特許専用の翻訳メモリを開発した。

日中特許翻訳用の翻訳メモリの効果例を**表2**に示す。中国語特許特有の表現を出力することで，中国語特許明細書に適した訳出ができた。

表2. 日中特許翻訳用の翻訳メモリの効果
Effect of translation memory for J-to-C machine translation

原　文	標準辞書での翻訳	翻訳メモリを使った翻訳
その特徴は以下を含むことである。	那个特征是含有以下。	其特征在于包括。
本発明の技術的課題を解決する手段は以下のとおりである。	解决本个发明的技术性题目的手段如下。	本发明解决其技术问题的技术方案是。

5 あとがき

中国特許の調査や出願を効率化する機械翻訳技術について述べた。この技術は，The 翻訳エンタープライズ$_{TM}$V15の中日・日中機械翻訳エンジンに実装した。

中国特許の翻訳を支援するための機械翻訳には，専門用語辞書の拡充などによる翻訳精度向上とともに，複雑な技術表現に対応した特許文書専用機能と知識が必要である。今後も，中国語の特許調査や特許出願に必要な翻訳業務を効率化するため，翻訳精度の向上と機能強化を進める。

文　献

(1) World Intellectual Property Organization. "World Intellectual Property Indicators 2011". <http://www.wipo.int/ipstats/en/wipi/>, (accessed 2012-01-10).
(2) 船守茉美. “中国公開特許公報の日本語への機械翻訳”. 特許庁技術懇話会. <http://www.tokugikon.jp/gikonshi/262/262tokusyu1.pdf>, (参照 2012-03-30).
(3) 熊野　明. 知的財産のグローバル化を加速する機械翻訳技術. 東芝レビュー. **64**, 2, 2009, p.10－13.
(4) 東芝ソリューション. “The 翻訳エンタープライズ$_{TM}$”. <http://mt-server.toshiba-sol.co.jp/pro/hon_yaku/seihin/server/index_j.htm>, (参照 2012-03-30).
(5) 出羽達也 他. 中日・日中機械翻訳システム. 東芝レビュー. **62**, 4, 2007, p.30－33.
(6) 鈴木博和 他. “特許文書用前編集機能を備えた機械翻訳システム”. 情報処理学会第63回全国大会. 山口, 2001-09, 情報処理学会. 2001, p.255－256.

熊野　　明　KUMANO Akira
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ソフトウェア開発部参事。自然言語処理の研究・開発に従事。情報処理学会，人工知能学会，言語処理学会会員。
Toshiba Solutions Corp.

一般論文