知的財産のグローバル化を加速する機械翻訳技術

Machine Translation Technology to Accelerate Globalization of Intellectual Property

熊野　明
■ KUMANO Akira

わが国には，特許情報など多くの知的財産が存在する。これらの多くは，海外からアクセスする価値のあるものであるが，必ずしも英語で著されているものではない。機械翻訳技術は，これらの技術情報を英訳するのに欠かせない手段である。特許文書を機械翻訳するには，特許特有の課題がある。請求範囲などに書かれる不自然な長文は，一般には機械翻訳が困難であるが，前処理することで英訳が容易になる。また，翻訳精度を上げるための専門用語辞書の蓄積には，対訳文書からの辞書構築技術が有用である。

東芝の機械翻訳技術は，これらの技術の蓄積により高精度翻訳を実現し，多くの製品やサービスに活用されている。

There is a large volume of intellectual property documentation, including patent documents, in Japan. Although these documents are worth accessing from other countries, very few of them are written in English. Machine translation is essential as a means of translating them into English. However, specific problems are encountered in the machine translation of patent documents, particularly the difficulty of translating the long sentences in claims. Pre-editing is of assistance in this area. Dictionary building technology using a parallel corpus is also useful for the compilation of technical terms.

Toshiba has developed a machine translation technology as an accumulation of these technologies. This machine translation technology makes it possible to realize high-quality translations for widespread use in commercial products and Internet services.

1　まえがき

グローバル化が進む今日の社会では，日本国内で発明された技術を世界に伝えることは非常に重要である。最新の科学技術を表す代表的なものとして論文や特許があり，多くの科学技術論文は英語で書かれ出版される。しかし，国内で年間30万～40万件出願される特許は，国内で出願されるかぎり日本語で十分であり英語には翻訳されていない。発明は世界共通の財産であり，欧米から日本で出願されている特許の情報にアクセスしたいという要望が非常に強い。

東芝は，このような要望に応えるために，高精度な機械翻訳技術を開発した。ここでは，特許文書の特徴を利用した翻訳技術と，専門用語辞書を自動構築する技術の概要及び特長について述べる。

2　特許文書の特徴を利用した翻訳技術

2.1　特許文書の前処理機能

特許出願には，特許請求の範囲，明細書，及び要約書の3種類の文書が必要である。

特許請求の範囲は，一つの請求項を1文で記述する例が多く，非常に長い文であることが知られている。明細書にも，類似の長文が記述される。以下に，明細書の一例を示す。

［原文1］

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。

すべての構成要素を含めて1文にした，特許文書特有の表現である。この文をこのまま機械翻訳すると，例えば，次のような訳文が出力される。

［訳文1］

A support means 20 to support a turntable 18, enabling free rotation while receiving the weight of the thing 19 on a turntable 18 and this turntable 18 as load and changing this load into load torque in weight detection equipment 1,

It considered as the composition which consists of a motor 3 which carries out the rotation drive of the above-mentioned turntable 18 with the drive torque corresponding to load torque, and load measurement

equipment 2 which detects the load torque of this motor 3 and measures the weight of the thing 19 on a turntable 18 indirectly from the correlation of load torque and load.

原文が長いために，係り受けのあいまい性が多くなった結果，そのすべてを正しく解析することが困難になり，1文としての解析に失敗して2文に分割して翻訳したものである。

当社は，逆にこの翻訳困難な文の特徴を利用することを考えた。前述したように，このような表現は多くの特許に共通のものである。書かれている構成要素はそれぞれの特許で異なっていても，文全体の構成はほとんど共通である。その骨組みを次に示す。

［NP1において，NP2と，NP3と，NP4と，NP5とからなる構成とした。］

ここで，NP1～NP5は任意の名詞句を表す。

この構造を，前処理として機械的に次のように書き換える。

［NP1に，以下を備えて構成する。▽ NP2 ▽ NP3 ▽ NP4 ▽ NP5］

▽は文の分割位置を表し，一つの長文を5文に分割することを示す。

書き換えた後の文とその翻訳結果は，次のとおりである。

［書換え後の原文1］

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19の重量を間接的に測定する荷重測定装置2

［書換え後の訳文1］

Weight sensing device 1 comprises :

Turntable 18.

Support means 20 to support turntable 18 enabling free rotation while receiving weight of object 19 on this turntable 19 as load and changing this load into load torque.

Motor 3 which rotates said turntable 18 with drive torque corresponding to load torque.

Load measuring device 2 which detects load torque of this motor 3 and measures weight of object 19 on turntable 18 indirectly from correlation of load torque and load.

この前処理の結果，訳文の精度が向上した。少なくとも，構成要素の切れ目が明確になった結果，英文のクレーム記述の様式に似た表現が実現できた。東芝ソリューション（株）製の機械翻訳パッケージソフト“The翻訳TM2008プレミアム[1]”では，このように，特許文書専用の書換え機能を使うことにより，英訳の精度が向上している[2]。

2.2 翻訳メモリ機能

出願特許の審査過程では，次のような文が記述される。

［原文2］

この出願の下記の請求項に係る発明は，その出願前日本国内又は外国において頒布された下記の刊行物に記載された発明又は電気通信回線を通じて公衆に利用可能となった発明に基づいて，その出願前にその発明の属する技術の分野における通常の知識を有する者が容易に発明をすることができたものであるから，特許法第29条第2項の規定により特許を受けることができない。

拒絶理由通知書でよく見る，定型文の一例である。このまま機械翻訳すると，例えば，非常に不自然な訳文が出力される。このままでは審査過程の英訳としては使えない。

しかし，この定型文に対しては固定の英訳がある。これも，いわゆる定型文である。

［固定訳文2］

The inventions in the claims listed below of the subject application should not be granted a patent under the provision of Section 29 (2) of the Patent Law, since they could have easily been made by persons who have common knowledge in the technical field to which the inventions pertain, on the basis of the inventions which were described in the distributed publications listed below or made available to the public through electric telecommunication lines in Japan or foreign countries prior to the filing of the subject application.

原文2全体に対して，固定訳文2全体を登録することが，翻訳メモリ機能と呼ばれるものである。この翻訳メモリを利用すれば，原文2がまちがいなく固定訳文2に訳されるので，英訳をチェックする手間が省ける。

3 専門用語辞書の自動構築技術

機械翻訳ソフトの辞書には，基本語と専門用語を合わせると，数百万の用語が収録されている。これらの用語はソフトウェア開発時点での使用状況を考慮して集められたものであるが，あらゆる文書を翻訳する場合に十分というものではない。特に特許文書には，様々な分野で日々生まれる最新の専門用

語とその訳語を収めた専門用語辞書が必要であるが，そのような辞書をあらかじめ準備することは事実上不可能である。

そこで，人が翻訳した訳文とその原文を利用して専門用語の対訳辞書を作成する，専門用語辞書の自動構築技術を開発した。The翻訳TM2008プレミアムに搭載されている辞書構築支援機能は，その研究成果を実用化したものである。

3.1　辞書構築支援機能

機械翻訳辞書に収録されていない用語を正しく翻訳するためには，用語とその訳語をユーザー辞書に登録する必要がある。通常ユーザー辞書登録は，対象文書を機械翻訳ソフトで翻訳した結果から正しく訳されていない用語を探し出し，その訳語をほかの文献で調べて入力する方法と，あらかじめ訳語の不備な用語リストを準備し，個々の用語に正しい訳語を付与した後，そのリストを使って一括登録処理する方法が一般的であった。

1990年代半ば以降，コーパス（corpus）からの知識抽出技術に関する研究が盛んに行われてきた。コーパスとは，テキストデータを大量に集めて整理したものである。現在，日本語文書とその訳文の英語文書のペアを準備することが，ある程度可能になっている。日本語特許の抄録とそれを人手で翻訳したPAJ（Patent Abstracts of Japan）は，その好例である。このような対訳データがあれば，その原文中からユーザー登録すべき用語を抽出し，その訳語を訳文中から推定して提示することが可能である。

The翻訳TM2008プレミアムの辞書構築支援機能は，このような目的で開発された。現在リリースされている支援ツールの抽出・表示例を**図1**に示す。画面の左が日本語文（原文）で，中央が英語文（訳文）である。原文ウィンドウに特許抄録文の一例を，訳文ウィンドウにPAJを入力した状態を表示している。ここで，用語抽出及び訳語推定の指示をすると，原文と訳文の分析を開始し，右のウィンドウに，抽出した専門用語とその推定英訳語を，原文での頻度とともに表示する。

3.2　専門用語の抽出

専門用語として抽出するのは，機械翻訳ソフトの辞書に収録されていない用語であり，ユーザー辞書登録することで，翻訳品質の向上が見込まれる用語である。その主なものは未知語であり，かたかな語であることが多い。

しかし，未知語のほか，既知語を組み合わせた複合語も抽出する。例えば，“ロータリー”と“耕耘（うん）機”は機械翻訳辞書に収録されていても，原文に出現する“ロータリー耕耘機”として収録されていなければ専門用語として抽出する。図1に示した特許文書でも，ロータリー耕耘機の英訳語はrotary tillerであるが，これは，ロータリーの標準的な英訳語rotaryと耕耘機の標準的な英訳語cultivatorを単純に連結したものと異なる。訳語がrotary cultivatorなら，ロータリー耕耘機をユーザー辞書に登録する必要はないが，rotary tillerと訳出するためには，ロータリー耕耘機＝ rotary tillerという対訳情報を辞書登録する必要があるからである。

3.3　英訳語の推定

複合語の場合は，その構成語それぞれについて機械翻訳辞書を参照することで，訳語候補を複数個得ることができる。それらの訳語候補の組合せと訳文中の英単語の表現を比較して，複合語に対する英訳語を推定する。

例えば，“オープンビット線方式”という複合語を辞書に登録すべき専門用語として抽出した場合を考える。この複合語は，“オープン”，“ビット”，“線”，“方式”の4語に分割できる。

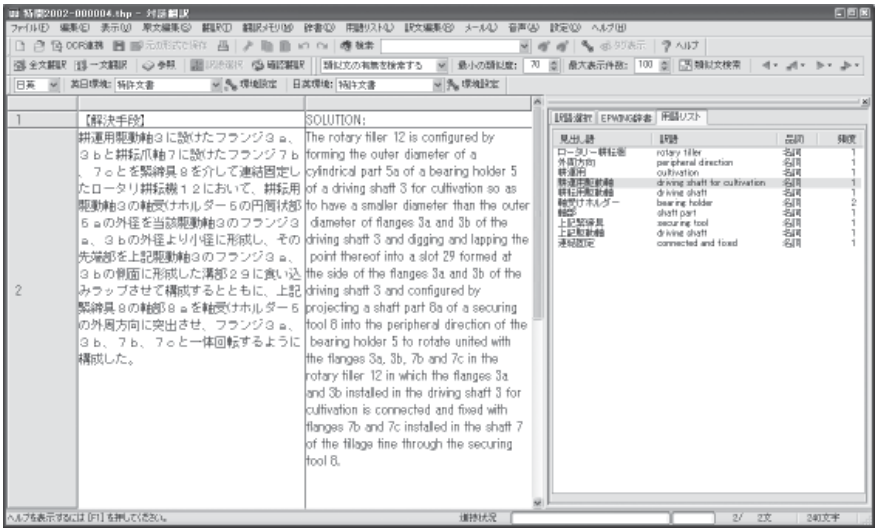

図1．対訳文書からの専門用語抽出 — 特許の日英対訳文書から，辞書登録すべき用語を抽出し，英訳語を推定して表示する。
Extraction of technical terms from parallel text

英訳語候補は，対応する英訳文中から，任意の連続した英単語を取り出すことで得られる。例えば，英訳語候補として，訳文中から次のような英単語列が得られたとする。

(1) **open bit line**
(2) **open bit line** configuration
(3) **bit line** configuration

太字で示した英単語は，open=オープン，bit=ビット，line=線のように，日本語の構成語の訳語候補と一致するものである。(1)は，単語列のすべてが日本語の構成語と一致するが，構成語数が1語足りない。(2)は，構成語数は同じだが，configurationと方式との対訳関係が機械翻訳辞書では得られない。(3)は，構成単語数が1語足りないうえ，configurationが方式と対応しない。どれも，機械翻訳辞書を利用しただけでは完全な一致が得られないが，構成語数と個々の対訳関係の条件を考慮して，この例では(2)を英訳語の第1候補として出力する[3]。

未知語の場合は，機械翻訳辞書による対訳情報がないため，それを利用した訳語推定はできない。しかし，かたかな語の場合は，英語のつづりとの類似性を利用することができる。例えば，当社AVノートパソコンのブランド名“コスミオ”という用語が未知語として抽出されたとき，英文中のQosmioがその訳語であることを推定する。

まず，かたかな語から発音要素候補を作成する。促音“ッ”，長音“ー”を除き，拗音(ようおん)を含めて1音節ずつに分解し，それぞれ一つ以上の候補を出力する。具体的には，“コ”はko，“ス”はsu/s，“ミ”はmi，“オ”はoの候補を出力する。スは，スーパー(super)のような母音を伴った場合にも，ストア(store)のように子音だけの場合にも使われるので，2種類の候補を出す。次に，訳語候補の英単語も発音要素候補を作成する。英語のつづりを各子音の直前で分割し，同様に発音要素候補を作成する。“Qo”はkou/ko/ka，“s”はs/z，“mio”はmio/maioの候補を出力する。sは，books[buks]のような清音の場合にも，pens[penz]のように濁音の場合にも使われるので，2種類の候補を出す。最後に，かたかな語から生成可能な候補列と英単語から生成可能な候補列を比較する。この場合は，コスミオから生成可能な“ko-s-mi-o”と，Qosmioから生成可能な“ko-s-mio”が一致するので，Qosmioはコスミオの訳であると判断できる[4]。

4　研究成果を利用したソフトウェアやサービス

当社の特許翻訳に関する技術は，パッケージソフトや機械翻訳サーバなどの製品として市場に出ている。

特許電子図書館[5]では，海外のユーザーに対して，日本の特許を英語に自動翻訳して表示するサービスを行っているが，ここにもこの機械翻訳サーバが用いられている。また，日本の特許庁は，海外の特許庁におけるサーチ・審査負担を軽減するために，高度産業財産ネットワーク(AIPN：Advanced Industrial Property Network)を構築し，審査関連情報を英語に翻訳して提供するサービスを行っている。ここにもこの機械翻訳エンジンが用いられている。

2008年，国立情報学研究所(NII：National Institute of Informatics)が主催する評価ワークショップNTCIR-7[6]では特許翻訳タスクを設定し，応募した機械翻訳の出力に対し，自動評価と人手評価を行った。当社は日英翻訳と英日翻訳の両タスクに応募し，自動評価及び人手評価の両方でトップの成績を収め，特許文書に対する機械翻訳精度の高さを確認することができた。

5　あとがき

特許文書の共有ニーズは，日本語と英語の間にとどまらない。近年急激に出願件数が伸びている中国の特許も同様の課題を持っている。当社は，日英・英日翻訳だけではなく，日中・中日翻訳技術も開発し，2007年に機械翻訳サーバとして製品化した。また，東芝中国社が北京に持つ研究所では，当社の技術をベースにして，英中・中英翻訳技術を開発している。このように，日本語，英語，及び中国語を相互に機械翻訳できる環境が整いつつある。

特許情報などの知的財産は，国内だけでなく全世界で流通することでその価値がより高まる。機械翻訳技術はその流通を加速するものであり，今後も，より高い精度の実現を目指して研究開発を行っていく。

文　献

(1) 東芝ソリューション(株)．“英日/日英翻訳ソフトThe翻訳シリーズ”．<http://hon-yaku.toshiba-sol.co.jp/>，(参照 2008-10-20).
(2) 鈴木博和，ほか．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, 2-255－2-256.
(3) 熊野　明，ほか．対訳文書からの機械翻訳専門用語辞書作成．情報処理学会論文誌．35，11，1994，p.2283－2290.
(4) 熊野　明．カタカナ表記からの英訳推定による専門用語辞書作成．言語処理学会第1回年次大会，東京，1995-03, p.221－224.
(5) 独立行政法人 工業所有権情報・研修館．“特許電子図書館(IPDL)の運営”．<http://www.inpit.go.jp/info/ipdl/index.html>，(参照2008-10-20).
(6) NTCIR-7 Information Retrieval Evaluation Site. "NTCIR-7 Workshop". available from <http://ntcir.nii.ac.jp/>，(accessed 2008-10-20).

熊野　　明　KUMANO Akira
研究開発センター　知識メディアラボラトリー主任研究員。自然言語処理技術の研究・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。
Knowledge Media Lab.