特許明細書の日英機械翻訳における課題

株式会社東芝　研究開発センター
熊野　明

PROFILE 1982年東京工業大学工学部卒業。同年東京芝浦電気(株)(現､(株)東芝)入社。以来、機械翻訳、電子化辞書などの自然言語処理技術の研究開発に従事。現在、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。情報処理学会、人工知能学会、言語処理学会会員。2007年度から特許版産業日本語委員会委員

akira.kumano@toshiba.co.jp　044-549-2239

1 背景

日本の特許文書を海外の技術者が情報検索する際には、間違いなく英訳が必須である。IPDL [1] やAIPN [2] では、機械翻訳を利用した海外向けの日本特許検索サービスが既に提供されている。一方、国内の技術者にとっても、自分の特許を外国に出願する場合は英訳する必要が生ずる。最終的に出願する英語明細書は人手でチェックしたものであるが、その下訳作成には機械翻訳が利用されることが多い。

このように日本語の特許文書に対する日英機械翻訳の利用は広まっているが、機能・精度ともに改良の余地がある。そこで今回、現状の機械翻訳ソフトの機能や精度をもとにその課題を洗い出したので、今後の機能強化・翻訳精度向上の材料とするために紹介する。

2 現行ソフトの分析

現在の機械翻訳ソフト“The翻訳2008プレミアム特許エディション [3]”を使って日本語特許明細書を英訳する過程を通して、機械翻訳ソフトの改良すべき点を分析した。社内で外国出願特許を担当する知的財産担当者に加え、機械翻訳研究開発の技術者から集まった問題点や改良希望項目を、その対策案とともに表1に示す。

3 機能・精度に対する課題の整理

前章の分析結果をもとに、現行翻訳ソフトの問題点やユーザからの改良要求を整理し、技術的実現性を考慮して課題を抽出した。それらの課題を、（A）確実な効果が期待でき、実現も比較的容易なものと、（B）対策に試行錯誤が必要で、実現コストの高いものに分類した。

（A）確実な効果が期待でき、実現も比較的容易なもの

（A-1）用語抽出機能

日本語明細書（原文）中に出現する用語を収集し、その訳語とともに表示する機能である。現行の機械翻訳ソフトでも実現されており [4]、製品搭載以降も特許文書に向けた技術改良を行っている [5] [6]。現行ソフトでは機械翻訳辞書に登録されていない未知語と複合語だけが対象である。既登録語でも、翻訳前に訳語をチェックすることは、望ましい訳文を作成するには有効である。この機能は、現行の用語抽出機能を拡張することで実現可能である。

表1　現行機械翻訳ソフトの特許翻訳における問題点

分類	問題点	重要度	対策	困難度
辞書	専門用語の訳語が正しくない 関連語の訳語が統一されていない	中	訳語を調査して修正	易
辞書	一般語の訳語が特許文書に適していない	中	頻度の高い例文を調査	易
前処理	請求項の文に前処理規則が適用されない	中	記述パターンを洗い出して前処理規則追加	中
前処理	サブクレームに適切な前処理が適用されない	中	記述パターンを洗い出して前処理規則追加	中
文分割	フローチャートに合った文分割ができていない	軽	長文分割規則を追加	中
解析	原文が正しくなくて翻訳できなくても、無理に訳文を出力する	重	翻訳を中断する機構を新設	難
変換	英訳を指定した原文に対して、訳語が重複する	重	変換規則を追加	中
変換	名詞文が直訳されると英語として不自然 （例：解析時、置換を実行）	中	語彙変換規則の追加	中
生成	請求項（sub-claimを含む）らしい表現ができていない	中	生成文法の追加	中
文脈処理	既出語に定冠詞 the が付かない	中	文脈処理の枠組みを新設	難
文書構造	英語の明細書の構造（文の順序）が実現されていない	中	構造変換の枠組みを新設	中

（A-2）特許固有の表現の訳し分け

特許文書特有の訳し分けの知識は、その多くが現行機械翻訳ソフトに収録されているが、まだ完全なものではない。多くの特許文書の翻訳結果から、追加登録すべき訳し分け知識を選定し、辞書に登録することが可能である。

（A-3）訳語重複チェック機能

原文では意図的に使い分けた用語でも、機械翻訳の英訳では同じになる場合がある。例えば原文中で「半導体装置」と「半導体デバイス」を使い分けた明細書を機械翻訳する場合、使用する専門用語辞書の組合せによると、どちらも「semiconductor device」と訳出されてしまう場合がある。このような現象を事前に検出すれば、訳語の変更が必要な場合に辞書登録できる。（A-1）の用語抽出機能で抽出された訳語をチェックすることで実現可能である。

（A-4）訳語統一チェック機能

（A-3）とは逆に、本来統一されるべき用語の訳語が、訳文中で統一されない場合がある。例えば、「音声認識装置」と「音声認識部」が含まれた明細書を機械翻訳する場合、使用する専門用語辞書の組合せによると、前者は「voice recognition equipment」、後者は「speech recognizing unit」と訳出されてしまう場合がある。ここでは、voice recognitionとspeech recognizingを訳し分ける必要はない。これも、（A-1）の用語抽出機能で抽出された訳語をチェックすることで実現可能である。

（B）対策に試行錯誤が必要で、実現コストの高いもの

（B-1）長文自動分割の精度

（B-2）翻訳メモリのヒット率

（B-3）特許文前処理の再現率

いずれも現行の機械翻訳ソフトに実装されている技術であるが、満足できる精度が得られていないものである。精度向上によって特許明細書の翻訳作業で効率向上が期待できるものである。

4 課題解決のための検討

前章で、「(B) 対策に試行錯誤が必要で、実現コストの高いもの」として分類した技術に関して、現状精度を評価し、今後考えられる対策を検討した。

4.1 長文自動分割

長文を正しく翻訳することは、構文解析おける曖昧性が増大するため、非常に困難である。このような長文を短く分割して、分割単位ごとに翻訳する機能である。長文は特許文書だけの問題ではないが、特許明細書には一般の技術文書にない長文が多く見られる。

翻訳エンジンは、文中の読点や特徴的な表現を手掛かりに分割位置を決定するが、適切な分割位置を決定することは必ずしも容易ではない。不適切な位置で分割した結果、下訳として全く役に立たない翻訳結果になることがある。

(例)

これを解消する手法の1つとして、担体に固定した捕捉プローブと、▼

電気化学的に活性である物質を標識した標識プローブと、遺伝子サンプルとをハイブリダイズさせた後、▽電圧を印加して、遺伝子サンプルに結合した標識プローブからの電気化学的な信号を検出することにより、▼

目的とする遺伝子の存在を検出する手法がある。

上の例では、▼で示す2箇所で長文分割が行われるが、最適な位置とは言えない。特に最初の分割は1行目の▼でなく▽で示す位置で行いたい。

そこで、現行の翻訳ソフトの長文分割の精度を評価した。

(1) 精度の現状

- 評価対象の全18,675文のうち、1,005文(5.4%)が長文分割された
- このうち、68%(全評価文の3.7%)は不適切な分割であった

長文は一文当たりの文字数が多いため、文数の割合である3.7%という数字以上に、文書全体に対する影響は大きい

(2) 今後可能な対策

- 並列構造(特に名詞の列挙)を認識し、その途中では分割しない
- 時を表す連用修飾句や主格の後の読点では分割してもよい
- 連用中止や従属節で優先的に分割する

4.2 翻訳メモリ

過去の翻訳例を蓄積・検索して利用する機能である。翻訳メモリの検索機能を利用すれば、全く同じ、あるいは、非常に類似した翻訳例が見つかった場合に、訳文の修正をほとんど必要がないという利点がある。

特許文書には、類似の表現を使った文が繰り返し用いられているので、通常の技術文書より翻訳メモリの効果が大きいと考えられる。

(例) 入力文

図1は、本発明によるレイアウト方法の処理の流れを示すフローチャートである。

ヒットした対訳用例 (類似度70%)

[原文] 図3は、本発明による図形変形処理の概要を示すフローチャートである。

[訳文] FIG. 3 is a flow chart showing the out-

line of the graphic transformation in accordance with the present invention.

下線を施した部分は、入力文と対訳用例原文とが異なる部分である。ここでヒットした対訳用例を利用することで、下記の訳文を生成することが可能である。

FIG. 1 is a flow chart showing the flow of the layout process in accordance with the present invention.

翻訳メモリがどの程度検索できるのか、実際の特許文書に対して評価を行った。

（1）精度の現状

- 2年分のパテントファミリ（全分野）から自動抽出して作成した対訳用例300万対のヒット率を調査した
- IPCがG06F17の明細書から抽出した26,503文のうち、類似度60%を閾値として検索した結果、1,948文（7.4%）がヒットした
- ヒットしたのには定型表現に近いものが多く、翻訳作業の負荷軽減に寄与しそうなものは少なかった

（2）今後可能な対策

- 一文単位だけではなく、より小さい節や句の単位で用例を利用する
- 翻訳メモリを大規模化することで、ヒット率を向上させる

4.3 特許文前処理

特許文書独特の長文に対して，機械翻訳システムが翻訳しやすいよう自動的に前処理する機能である[7]。パターンマッチングによる言い換え規則にとって実現している。

（例）

原文（オリジナル）

【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。

言い換え規則

（NP1）において、（NP2）と、（NP3）と、（NP4）と、（NP5）とからなる構成とした。

↓

（NP1）に以下を備えて構成する。

（NP2）

（NP3）

（NP4）

（NP5）

原文（前処理後）

【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

言い換え規則における（NP1）、（NP2）などは、単独の名詞句を表している。

特許請求の範囲や、構成要素を列挙する文では、このような前処理が有効であると考えられる。そこで、現行の翻訳ソフトの長文分割の精度を評価した。

（1）精度の現状

3分野の特許のクレーム（特許請求の範囲）の文に対して、前処理規則がどのくらい適用されるかを分析した。IPCのC12Q1、G06F17、H01L21の各分野から10件ずつ、合計30件の特許を無作為に選び、322のクレーム文を取り出して選択して評価を行った。その結果を表2に示す。

表2の分析結果から、以下の問題が明らかになった。

- 分野によってばらつきがあるが、適用すべきクレーム文の約1割にしか前処理規則が適用されない
- 前処理規則が適用されても，正しく適用されないことも多い

（2）今後可能な対策

- 特許文書から構成要素列挙型の長文表現パターンを洗い出し、そのパターンに対応する前処理規則を増強する
- 特に、サブクレームに対する規則を増強する
- 表層の文字列の特徴を利用するだけでなく、句の単位を認識した上で、構造的な言い換えを行う

5 まとめ

特許明細書の機械翻訳作業を効率化するために有効な、機械翻訳ソフトの機能強化、精度向上の課題を検討した。用語抽出機能や訳語チェック機能は、技術文書に共通の課題であるが、比較的実現が容易であり、確実な効果が期待できる。長文自動分割、翻訳メモリ、特許文前処理は、現行ソフトで実現されている技術ではあるが、知識や機構の改良には試行錯誤が必要で、実現コストが高いものである。

これらの課題の中には、今年度Japioで策定が進めら

表2　現行機械翻訳ソフトの特許文前処理精度

IPC		C12Q1	G06F17	H01L21	合計
特許文書数		10	10	10	30
クレーム総文数		113	103	106	322
適用すべき文	(a) 正しく適用された（○）	4	4	3	11
	(b) 適用が不正確（×）	3	17	14	34
	(c) 適用されない（×）	15	43	17	75
適用の不要な文	(d) 適用された（×）	0	0	0	0
	(e) 適用されない（○）	91	39	72	202

れている日英機械翻訳向き産業日本語の仕様や、産業日本語オーサリングシステムによって解決するものもある。しかしその解決は、人間と機械が互いに歩み寄って成立するものであり、そのためには、機械翻訳そのものの性能向上も不可欠である。今後は、これらの課題解決のための技術開発、知識蓄積を行い、より実用的な特許明細書用日英機械翻訳を実現していきたい。

参考文献

[1] 工業所有権情報・研修館、特許電子図書館、http://www.inpit.go.jp/info/ipdl/service/index.html

[2] 守屋：最近の特許行政の動き、2005特許・情報フェア&コンファレンス、http://www.japio.or.jp/fair/files/2005kouen01.pdf

[3] 東芝ソリューション、英日/日英翻訳ソフト The翻訳シリーズ、http://hon-yaku.toshiba-sol.co.jp/

[4] 熊野、平川：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巻11号 （1994）

[5] 熊野：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（1995）

[6] 熊野：特許文書の特徴を利用した対訳知識抽出、Japio 2006 Year Book （2006）

[7] 鈴木、熊野：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（2001）