特許を対象とする高精度な自動翻訳技術の異言語検索における実用化
—異言語特許の検索に有用な統計翻訳技術の最新成果—

High-quality automatic translation technologies of patent document and their practical use in cross-lingual retrieval

独立行政法人情報通信研究機構　ユニバーサルコミュニケーション研究所多言語翻訳研究室室長　**隅田　英一郎**

PROFILE: 博士 (工学)。日本 IBM、国際電気通信基礎技術研究所 (ATR) を経て、情報通信研究機構 (NICT) の多言語翻訳研究室室長。言語処理学会副会長。

eiichiro.sumita@nict.go.jp　　TEL 0774-98-6350

1 はじめに

アイデアは、特許制度がなければ、他人に簡単に盗まれてしまう。特許制度は、こういったことが起こらないよう、発明者には一定期間、独占的な権利を与えて保護を図る[1]ものであり、日本の特許法第 1 条には、「この法律は、発明の保護及び利用をはかることにより、発明を奨励し、もって産業の発達に寄与することを目的とする」とある。逆に、製品を作り販売するためには、他人の特許を侵害しないように、予め調査をしておく必要がある。これを怠ると、膨大な補償金を払うことになりかねない。

特許制度は国ごとに定められており、日本では日本語、中国では中国語で、各国政府に申請することになっている。一方、経済はグローバル化しているから、例えば、日本の企業が中国に製品を輸出するためには、中国の特許の調査が不可欠になる。中国は今や世界第 2 位の経済大国であるので、日本企業も中国市場への進出が今後の発展の要になる。一方で、最近の中国では特許の出願数も急速に伸びており、今や世界第 2 位である。実際に侵害・訴訟事案が増加している[1]。困ったことに、中国語を日本語に翻訳ができる翻訳者の数は限られるし、人間による翻訳はコストと時間がかさむ。そこで、中国語特許文書の高精度の自動翻訳システムの開発が焦眉の課題となっていた。英語でも、従来の特許用の自動翻訳システムの精度は十分でない場合が多く、高精度の自動翻訳システムが求められていた。

2 節では、特許の翻訳の難しさと従来技術の翻訳品質を確認し、3 節では、特許を対象とした場合について、統計翻訳技術とその高度化について述べ、4 節で、外国語から日本語への特許の統計翻訳技術の事業化について紹介する。

1　第 16 代アメリカ合衆国大統領リンカーンの言葉で「特許制度は、天才の火に利益という油を注いだ」が残っている。

2 特許の翻訳の難しさと従来技術の翻訳品質

特許文を翻訳することは大変難しい。実際、翻訳会社の翻訳費用の単価も高度な技能が必要とされることから、他の分野の文書より大幅に高額になっている。

理由の一つ目は 1 文の長さが非常に長いことによって、解釈が困難になり翻訳誤りが増えること、二つ目は専門用語が膨大で、これを十分カバーする対訳辞書が存在しないし、専門用語はどんどん新規に作られるため追補が追い付かないこと、がある。

表 1　従来技術の翻訳品質

中国語の原文	图一是表示应用本发明的车用发动机的传感器设置结构的发动机一的整体结构的图
模範訳	図 1 は、本発明に係る車用エンジンのセンサ配設構造を応用したエンジン 1 の全体構成図を示している
従来技術 A の訳	図はちょっと本発明した車を応用することを示してエンジンでのセンサーであり構造のエンジンの 1 の全体構造の図を設置します
従来技術 B の訳	図は 1 つは応用の当発明の自動車用エンジンのセンサーが構造のエンジン 1 の全体の構造の図を設けると表しています

さらに、中国語・英語と日本語は文法が全く異なることから、従来の自動翻訳技術では翻訳精度が低い状況[2]だった。表 1 の３～４行に示したように、意味不明な翻訳が出力されることが少なくない。次節で、この状況を打開する新しい手法である統計翻訳（SMT）について述べる。

3 統計翻訳技術による特許翻訳

3.1 統計翻訳（SMT）の基本

コンピュータのハードウエアの処理速度や記憶容量が格段に進歩したこと、文章や辞書が大量にコンピュータ上に集積されるようになったこと、などを受けて、自動翻訳の研究において、対訳コーパス（同じ意味の原文と訳文の文レベルの対を集めたもの）から，翻訳に必要な知識を自動的に構築する技術が興り、現在、主流の研究パラダイムとなっている。

1990 年前後に興った統計翻訳 (SMT) と呼ばれる技術[2]はその一つである。SMT では、対訳コーパスから二言語間の単語や句の対応関係を抽出した翻訳モデル（確率付きの対訳辞書と確率付きの語順変換表）と訳文の言語らしさを表現する言語モデル（日英翻訳であれば、並びの自然さを表す確率付き英語の単語連鎖データ）を導出し、これらの確率の積を最大化する訳文候補を出力する。図 1 は、対訳コーパスから、確率付きの対訳辞書が学習されることを示している。「どこですか」は

図 1 特許翻訳の基本

2 また、このように、特許文の翻訳は大変困難なため、いきなり、完璧な翻訳を目指すのでなく、まず、多少の不自然さを許容しても通じる翻訳を目指して、段階的に進める必要がある。

「Where is」に確率 60% で対応するなど、確率付きの対訳辞書の一部が例示されている。

3.2 従来技術と統計翻訳（SMT）の対比

NTCIR という国際的共同研究のフォーラムでの特許の自動翻訳に関する評価を軸にしながら、SMT による自動翻訳の急速な進展に関して紹介する。

3.2.1 フォーラム NTCIR と自動翻訳

NTCIR[3]という、情報アクセスシステムの評価を国内外の多数の研究者が共同実施するオープンイノベーションのためのフォーラムがあり、1998 年の立ち上げ以降 1 年半ごとに、多様なタスクの評価結果について議論するために国際会議を開催している。

2008 年 12 月開催の第 7 回目の NTCIR の会議からは、特許を対象とした自動翻訳に関する評価も行われている。この中で、従来技術と SMT の比較が重要なテーマの一つになっている。

3.2.2 フォーラム NTCIR の知見その 1（検索精度と翻訳品質の関係）

第 7 回目の NTCIR で、はじめて、特許の自動翻訳の従来技術と SMT の比較が行われた。

まず、翻訳品質に関して比較された。英日翻訳において、従来技術と SMT の訳文を、意味が通じるか（adequacy）と流暢か（fluency）の 2 つの観点[4]で 5 段階評価したところ、平均を取った尺度で、従来技術を使ったシステムが SMT の全てのシステムより、品質が良かった。

次に、異言語検索の精度と翻訳品質の関係が調べられた。異言語検索とは、日本語の特許文献を英語の検索キーワードで検索するなど、検索対象と検索キーワードの言語が異なった検索のことである。この言語の差を対訳辞書や自動翻訳で解消する。

検索の精度には MAP[5]と呼ばれる標準的な評価尺度があり、翻訳の品質には BLEU[4]と呼ばれる標準的な評価尺度がある。実験では、BLEU と MAP には強い相関（相関係数 0.936）があり、BLEU が高ければ高いほど、MAP が高くなることが分かった。SMT はこの BLEU を目的関数にして最適化するため、研究の進捗に伴い年々 BLEU が改善されており、同時に MAP

も改善されていることになる。

一方、従来技術は、BLEU も MAP も低かった。さらに、理屈の上からも、経験的にも、従来技術を改良しても、BLEU は大きく変化しない。従って、MAP も大きく改善しないと想定しても間違う可能性は低い。

まとめると、検索の側面から考えると従来技術は有効性が低く、改良の期待も薄いが、逆に、SMT は既に従来技術を圧倒しており、その後の進展も急速であり、性能が飽和する兆しも、まだ見えていないので、軍配は明らかである。

一方で、翻訳品質の面では、当時は、従来技術が SMT に上回っていたが、後述のように、その後の SMT のアルゴリズムの進展と、対訳データの増量で逆転することになり、SMT が翻訳品質でも検索性能でも最終的には優位にたつことになった。

3.2.3　SMT の訳語選択は高精細

今述べた異言語検索における SMT の優位について、ここでさらに考察する。

例えば、「communicate」には、「通信」、「連通」などの多様な訳語がある。従来技術では、訳し分け条件を分野等で指定し、対訳辞書中にある単語を出力に指定する規則を人間が作成する。一方、SMT では、①「通信」に翻訳される場合と②「連通」に翻訳される場合を識別する条件と出力単語が、実在の特許の対訳データから、自動学習されるので SMT の訳語選択は、特許表現の現実を反映して、高精度である [6]。また、特許用 SMT の対訳辞書の規模は約 10 億に上り、従来の自動翻訳ソフトの専門用語の収録数の約百～千倍に相当し、精緻な訳語選択が可能になっている。このことが異言語検索での高性能を可能にしている。

表 2　「communicate」訳語のサンプル

allow computers all over the world to **communicate** with one another	世界中のコンピューターが互いに通信できるようにする
The end portion has an outlet which **communicates** with the interior of the wand.	端部 62 は、ワンドの内部と連通する流出口を有している。

3.2.4　フォーラム NTCIR の知見その2（従来技術と統計翻訳（SMT）の翻訳品質の関係）

言語対によって翻訳の難しさは違い、自動翻訳、特に、統計翻訳の研究開発や翻訳性能の状況は異なる。英日のように語順が大きく異なる言語では翻訳品質が低く、年々、精度の改善を示しつつも長い間、従来技術の翻訳品質を上回ることができなかった [3]。

2010 年前後に提唱された一連の語順変換と訳語選択を分離する技術（3.2.6 節）が奏功し、英日の場合、2013 年 6 月開催の第 10 回 NTCIR で [7]、統計翻訳の翻訳品質が従来法の品質を上回ることが確認された。

日英の場合は、毎回、性能が改善しているものの、現時点では、まだ統計翻訳の翻訳品質は従来法の品質を下回っている。

3.2.5　大規模実験での翻訳品質

ただし、これは NTCIR で利用された 300 万文 という限定された対訳を使った評価の結果である。一方、SMT は、対訳コーパスの量が多ければ多いほど、翻訳品質が良くなることが分かっているので、対訳コーパスの量は増やすことによって品質向上が確実に可能であり、SMT の優位性に疑いを差し挟む余地は少ない。実際、独自に、NICT は自動文対応技術 [8] を駆使して 2700 万文の日英対訳コーパスを構築し、これを使った日英・英日の SMT は従来技術によるシステムを品質で圧倒していることを確認した。

3.2.6　SMT での構文を使った語順変換

同じデータ量でもアルゴリズムによる性能差が大きく変わるので , 与えられたデータでより高精度を実現する良いアルゴリズムの研究が重要になる。

日本語と英語のように、文法が異なる言語間の翻訳を高精度化するための新しい技術が盛んに研究開発されている。

たとえば、英日翻訳では、英語文の構文を計算し、主辞が必ず拘置されるという日本語の特性を利用して英語の構文構造を日本語のそれに変換した後、自動学習による確率付き対訳辞書などを用いて訳語を選択する手法

3　英仏など語順が似た言語対の場合は、2000 年前後には実用レベルに達している。

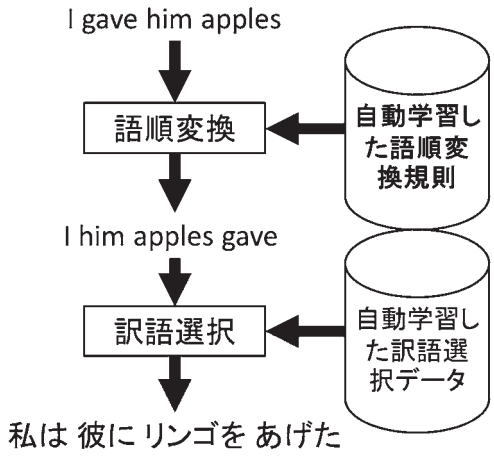

図 2　NICT の新方式の統計翻訳

で、翻訳の精度を大きく改善できることが示されている[9]。さらに、NICT は、英日翻訳に限定することなく適用可能となるように、語順制御を自動的に学習する手法を提案し、語順制御する新技術を提案し（図 2）、これによって、特許文書のような長文の文書においても高精度な翻訳を可能とした。

平均約 25 語のテストセットで、中国語特許文書を翻訳率 80%、英語特許文書を翻訳率 85% という高精度で日本語に翻訳できた。表 1 のサンプル入力に対する提案システムの訳文を表 3 に示す。

表 3　提案手法の翻訳品質

中国語の原文	**图**一是表示**应**用本**发**明的**车**用**发动**机的**传**感器**设置结构**的**发动**机一的整体**结构**的**图**
模範訳	図 1 は、本発明に係る車用エンジンのセンサ配設構造を応用したエンジン 1 の全体構成図を示している
提案技術での訳	図一は本発明に係る車両用エンジンのセンサ配設構造のエンジン一の全体構成を示す図

4　特許の自動翻訳の事業化

NICT は 3 節の最後に述べた技術をただちに事業化した。

日本特許情報機構（Japio）と共同で開発した、「中日自動翻訳ソフトウェア」の翻訳者が判定した精度は、従来技術の 3 倍以上の値を達成している。図 3 のように、この「中日自動翻訳ソフトウェア」によって、Japio は中国の特許文献を日本語に翻訳し、データベース化し、特許検索事業としてサービスを開始した[10]。図 4 に「センサ配設構造」に対する検索結果の中国語特許とその自動翻訳の一部を例示する。

また、ニッパツと共同で開発した「英日自動翻訳ソフトウェア」では、（特許要約 1 件あたりの）訳語誤り数を従来技術と比べて、約 12 分の 1 に削減するという高い品質を実現した。ニッパツは、今回開発した「英日自動翻訳ソフトウェア」によって、英語特許を対象にして同様のサービスを事業化した[11]。

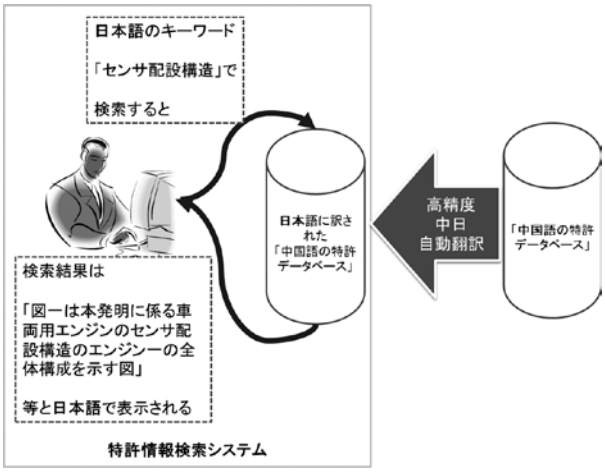

図 3　外国語特許の日本語による検索

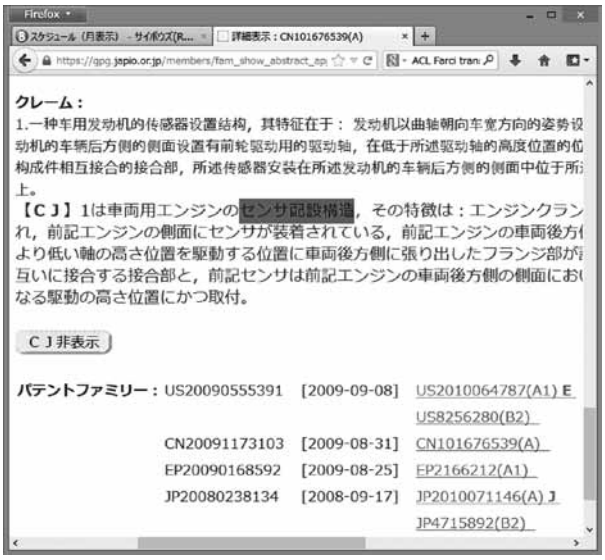

図 4　自動翻訳を利用した検索結果

5 おわりに

本稿では、SMT 技術の改良によって、従来技術を大きく上回る高精度の特許用の自動翻訳システムが実用化されたことを述べた。この新技術は、検索品質と翻訳品質の両面において、従来技術を凌駕している。

この自動翻訳システムは企業の知財部や弁理士や審査官の調査のための特許検索で役だつだろう。また、利用者からのフィードバックによって同システムは改良されていくだろう。

今後は、要素技術である構文解析の特許向け精度の改善とこれによって翻訳精度をさらに改善すること、より長文となるクレームの正確な翻訳を実現するための技術や文書全体での訳語の一貫性を実現する技術の創出などが課題になる。

また、一文が長く専門用語が多く、非常に翻訳が困難な特許での SMT の成功は、他の分野への応用研究を加速するだろう。

参考文献

[1] 産業構造審議会　平成 24 年 6 月 25 日配布資料「知財立国に向けた新たな課題と対応 http://www.jpo.go.jp/cgi/link.cgi?url=/shiryou/toushin/shingikai/sangyou_kouzou.htm

[2]「統計に基づく翻訳」、p.266-269、言語処理学辞典、共立出版、2009

[3] http://ntcir.nii.ac.jp/jp/about/

[4] 安田圭志 , 隅田英一郎 , " 機械翻訳の研究・開発における翻訳自動評価技術とその応用 ," 人工知能学会誌 , 23 巻 1 号 , pp.2-9, 2008.

[5] 岸田 和明 , 情報検索における評価方法の変遷とその課題 , 情報管理
Vol. 54 (2011) No. 8 P 439-448

[6] 特許と自動翻訳の新たなトレンド , 隅田 英一郎 , JAPIO YEARBOOK 2007 年版 , pp.262-265.

[7] Patent Machine Translation Task at NTCIR-10, http://ntcir.nii.ac.jp/PatentMT-2/

[8] Masao Utiyama and Hitoshi Isahara. (2003) Reliable Measures for Aligning Japanese-English News Articles and Sentences. ACL-2003, pp. 72--79.

[9] Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh: Head Finalization: A Simple Reordering Rule for SOV Languages, Proceedings of WMT-2010 (ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR), pp.244--251, 2010.

[10] NICT の高精度な中日自動翻訳ソフトウェアが Japio のサービスに
http://www.nict.go.jp/press/2013/03/28-1.html

[11]“英語特許文”の高精度「自動翻訳ソフトウェア」を開発
http://www.nict.go.jp/press/2013/03/21-1.html