英中韓から日本語への特許文向け統計翻訳
―高精度言語解析と事前並べ替えによる高精度な特許翻訳システムの構築―

English/Chinese/Korean-to-Japanese Statistical Machine Translation for Patent Documents

日本電信電話株式会社　コミュニケーション科学基礎研究所研究主任　須藤　克仁

PROFILE　2002 年京都大学大学院情報学研究科修士課程修了、同年 NTT 入社。統計的機械翻訳、音声言語処理の研究に従事。

日本電信電話株式会社　コミュニケーション科学基礎研究所主任研究員　鈴木　潤

PROFILE　2001 年慶應義塾大学大学院理工学研究科修士課程修了、同年 NTT 入社。博士（工学）。自然言語処理、機械学習の研究に従事。

日本電信電話株式会社　コミュニケーション科学基礎研究所／メディアインテリジェンス研究所主任研究員　秋葉　泰弘

PROFILE　1990 年早稲田大学大学院理工学研究科修士課程修了、同年 NTT 入社。博士（情報学）。機械学習、知識獲得、機械翻訳の研究に従事。

日本電信電話株式会社　コミュニケーション科学基礎研究所主任研究員　塚田　元

PROFILE　1989 年東京工業大学大学院理工学研究科修士課程修了、同年 NTT 入社。統計的機械翻訳、音声言語処理の研究に従事。

日本電信電話株式会社　コミュニケーション科学基礎研究所上席特別研究員　永田　昌明

PROFILE　1987 年京都大学大学院工学研究科修士課程修了、同年 NTT 入社。博士（工学）。統計的自然言語処理、統計的機械翻訳の研究に従事。

1　はじめに

外国の技術動向や特許抵触性の調査、あるいは外国への特許出願などにおいて、翻訳は不可欠である。調査等で大量の文献を高速に翻訳しようとする場合には機械翻訳が有望であり、すでに多くの機械翻訳システムが利用されている。こうしたシステムの多くは辞書や翻訳規則を人手で構築する「規則ベース翻訳」に基づくもので、長年にわたり整備されてきたものである。

機械翻訳の技術動向を見ると、辞書や翻訳規則に相当する翻訳知識を対訳文データから統計的に学習する「統計翻訳」がこの 10 年ほどの間に急速に発展しており、欧米を中心に広く実用化が進んだ。一方日本では、日本語が英語等と比較し語彙や文法の違いが大きく、特に特許等の長い文の翻訳において、統計翻訳では十分な翻訳精度が得られないことが多かった。しかし、近年の技術進展によって英日間の統計翻訳は大きく改善してきており、国際会議 NTCIR の英日特許翻訳タスク（2013 年開催）[1] においては、筆者らの統計翻訳システムが規則ベース翻訳システムを上回る翻訳精度を達成するに至った [2]。

本稿では、筆者らがこれまでの研究で得られた知見に基づいて構築した、英語・中国語・韓国語の特許文を日

本語に翻訳する統計翻訳システム[1]について述べる。本システムはまず入力文の言語解析（入力文の単語への分割、品詞推定、係り受け関係の推定）を行い、推定された係り受け関係に基づいて入力文を日本語の語順へ並べ替え（事前並べ替え。韓日翻訳では語順の違いが非常に小さいため不要）、統計翻訳によって日本語に翻訳する（図 1 に英日翻訳の例を示す）。統計翻訳に必要な対訳文データは、パテントファミリーから収集した。また、言語解析は特許データを追加して学習させ、特許向けに調整した。事前並べ替えは、係り受け関係の係り先を文後方に移動させる方法[4]を利用した。以下本稿ではこれらの技術について述べるとともに、特許文の翻訳実験の結果を示す。

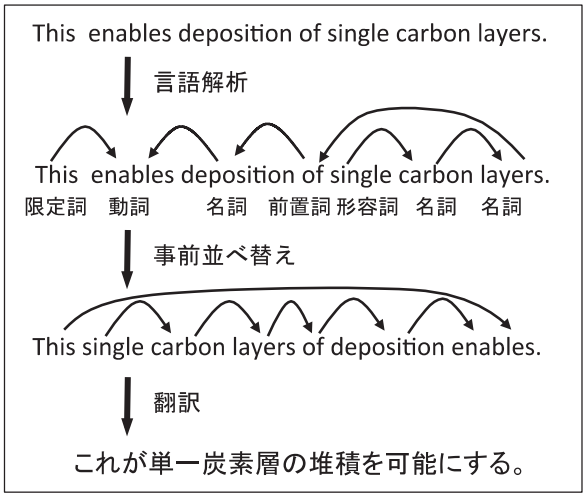

図 1　本システムの英日翻訳の流れ

2　対訳文データ収集

統計翻訳の精度向上には十分な量の対訳文データが必要である。先に述べた NTCIR[1] で提供された学習用対訳文データの規模は、約 320 万文であった。筆者らは同等規模以上の対訳文データを整備するため、各国への出願文書から自動的に対訳文を収集した。

2.1　文書単位対応付け

パテントファミリーに含まれる複数の国に出願された特許は、ほぼ対訳となっていると期待できる。NTCIR で提供された特許対訳データもパテントファミリーから対応する文を収集したものである [4]。本システムにおいても、パテントファミリーに含まれる特許の優先権主張に基づいて、対応する特許から対訳文を収集した。複数件の特許が対応する場合は個々の対訳文対応付けが困難であるため、1:1 に対応する出願文書対のみを用いた。

2.2　部分構造対応付け

文書対の中で対訳となっている文を探す際に、文書内の文のすべての組み合わせに対して対訳か否かを判断するのは膨大な計算時間を要する。そこで、近年の XML 形式の特許文書に付与されている文書構造情報（【発明の名称】、【発明を実施するための形態】等に相当）でまず部分的な構造を対応付け、その後で対訳文を探すことで、計算時間を削減するとともに精度向上を図った。[4] では XML 化されていない年代の文書を扱っていたこともあり、パターンマッチによって「発明の詳細な記述」「発明の背景」の箇所のみを取り出しているが、本システムでは XML でタグ付けされた部分構造すべてに対して対応付けを行った[2]。

2.3　段落・文対応付け

同等の部分構造の中で対訳文が存在する場合には、周辺の文も含めた段落の単位でも対訳となっていることが期待できる。[5] も同様の考え方に基づいて段落単位の対応付けと文単位の対応付けを交互に実行することで対訳文収集の精度向上を図っている。本システムでも英日対訳については [5] と同様の方法で段落・文対応付けを行った。中日・韓日については [6] と同様の文対応スコアに、段落内の文対応スコアの平均を乗算して最終的な文対応スコアとして用いた。文対応スコアの計算には、従来用いられてきた単語対訳辞書に加え、日本語の漢字と中国の簡体字の対応表や、統計翻訳の学習過程で得られる確率的な対訳辞書も利用した。

1　本システム構成技術のうち、英日翻訳手法については Japio YEAR BOOK 2013 への寄稿 [3] でも紹介している。

2　ただし、米国特許においては XML による構造化があまり活用されておらず、英日間では粗い対応付けを行うに留まった。

2.4　対訳文対応付けによる対訳文抽出結果

特許対訳文の収集のため、米国・中国・韓国の公開特許公報と日本国公開特許公報を用いて[3]上に述べた対訳文対応付けを行った。表 1 に対訳文収集に利用した 1:1 に対応する特許出願文書数と、得られた対訳文対の数を示す。英日の対訳文データについては NTCIR のものとほぼ同等の規模となった（対象とする出願年代は異なる）。

表 1　対訳文収集に用いた出願文書数と収集された対訳文数

言語対	1:1 対応文書数	対訳文対数
英日	11.8 万	364.2 万
中日	11.5 万	938.6 万
韓日	8.4 万	260.0 万

3　言語解析

本システムの言語解析部は、細分化すると文分割、単語分割、品詞付与、係り受け解析の 4 種類の処理を行う。これらは、文分割、単語分割、品詞付与といった、文字や単語の系列に対して文や単語の境界、品詞を表すラベルを付与する、系列ラベリングと呼ばれる種類の処理と、係り受け解析のように単語間の関係構造を予測する処理の 2 種類に分けられる。本システムでは、最初の 3 つの処理を [7] の同時解析法によって一括して行い、係り受け解析を [8] の依存係り受け解析手法によって行う。

3.1　新聞データと特許データの混合による学習

言語解析は、正解データと呼ばれる各処理の正解が付与された言語データに基づいて統計モデルを学習し、そのモデルに基づいて行う。本システムは特許文に対する言語解析を行うため、従来言語解析の研究に用いられてきた新聞記事の正解データに加え、新たに特許文の正解データを作成してモデル学習に利用した。さらに、特許文の正解データを大量に作成することは難しいため（本システムでは 1-2 万文）、各処理の正解が不明な平文データを大量（数千万〜数億文）に用い、半教師あり学習 [9] と呼ばれる方法でモデルを学習した（韓国語を除く）。

表 2　新聞記事と特許文に対する単語分割・品詞付与・係り受け解析精度（%）

	新聞			特許		
	単語	品詞	係受	単語	品詞	係受
英	99.9	97.0	91.8	99.3	94.7	86.7
中	95.1	88.4	82.0	92.7	85.5	81.2
韓	94.2	-	-	93.2	-	-

3.2　言語解析性能の評価結果

特許文に対する言語解析の性能を実験的に測定した結果を表 3 に示す。傾向として、特許の方が新聞記事と比べて解析精度がやや低くなっている。これは、特許の文が長いことと、未知語（正解データに出現しない単語）の割合が新聞記事と比較してやや多いことが原因の一つと考えられる。ただし、新聞記事に対する解析精度から大きく低下することはなく、次節の事前並べ替えに利用する観点で概ね問題がない程度の解析精度を達成していることが確認できた。

4　英語・中国語の事前並べ替え

正しい語順を得るための単語の並べ替えは、文が長くなるとともに計算時間が膨大になる上、モデル化も難しくなる。そのため、並べ替え距離を制限して計算量を抑え、比較的単純な並べ替えモデルによって解決することが多い。しかしながら、英日等の語順の違いが大きい言語間の翻訳においては長距離の並べ替えが避けられないため、様々な方法が試みられてきた。一つのアプローチが事前並べ替えと呼ばれる、原言語の文を目的言語の語順に近づけるように並べ替えるものである。事前並べ替えには規則に基づく方法と、自動単語対応付けに基づく統計的な方法があるが、本システムでは対訳データ量によらず安定して並べ替えが可能であるという面で前者を採用した。本システムにおいては、英日翻訳において顕著な性能向上を実現した主辞後置化（Head Finalization）[4]、及びその中日翻訳向け改良 [10] と同様の並べ替えを係り受け関係を利用して行った。

3　米国は 2004-2012 年、中国は 2007-2011 年、韓国は 2005-2011 年、日本は 2004-2011 年の公報を利用した。

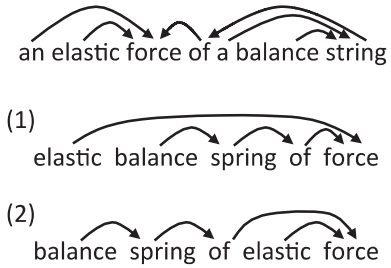

図２　係り受けを利用した主辞後置化の比較（冠詞 a,an は削除する）

主辞とは係り受け関係における係り先に相当するものであり、日本語は係り先が係り元よりも後方に置かれるという特徴を持つため、主辞をその係り元の後方へ移動させることで日本語に近い語順を得るのが主辞後置化の基本的な考え方である [3]。主辞後置化は二分木の構文木を利用していたが、係り受け解析の結果は二分木で通常表現できないため、係り先を後方へ移動するだけでは正しい並べ替えを行うことができない（図２（1））。主辞となる名詞句を後方から前置詞句が修飾するような構文構造において、元々の主辞後置化では名詞句ごと後置するのに対し、依存構造においては主辞となる単語のみを後置してしまったことに起因する。そのため、依存構造において主辞より前方にある修飾語は主辞と共に移動させるようにした [2]（図２（2））。また、句読点や引用符を跨ぐ並べ替えを行わないようにする例外規則を追加した。

また、中国語は図２のような名詞句における後置修飾がほとんどなく、単純な主辞後置並べ替えによって比較的日本語に近い語順が得られる。本システムではさらに [9] の知見を踏まえ、動態助詞（aspect particle）と副詞について主辞となる動詞の直後に移動するようにした。

5　翻訳実験

本システムの性能評価のため、収集した対訳文データを利用して翻訳実験を行った。英日・中日については事前並べ替えの有無による精度比較結果を示すが、韓日については事前並べ替えを行わなかったため、翻訳精度のみを示す。

表３　翻訳モデル用対訳データの諸元

言語対	文数	単語数（原）	単語数（日）
英日	283.7 万	8,004 万	9,435 万
中日	532.7 万	17,630 万	19,380 万
韓日	167.5 万	7,272 万	6,536 万

表４　言語モデル用対訳データの諸元

言語対	文数	単語数（日）
英日	346.3 万	14,950 万
中日	938.6 万	41,510 万
韓日	216.0 万	10,900 万

5.1　実験データ

本実験に利用した学習データについて、両言語の対応する語句の翻訳確率をモデル化した「翻訳モデル」のための対訳データについては表３に、翻訳結果の日本語らしさをモデル化した「言語モデル」のための日本語データについては表４に、それぞれ示す。対訳文データは２節で述べた手法により得られた対訳文のうち、一定の単語数以内（英日・中日は 64 単語、韓日は 80 単語）の文のみを抽出したものである。また、日本語データは得られた対訳文の日本語側のうち、非常に文が長い（文字数が 1,024 を超える）数文を除いたものである。また、パラメータの調整に利用する開発用データと、最終評価用データとして、英日・中日・韓日それぞれで約 1,000 文ずつを利用した。開発用データ・評価用データが含まれる出願は学習データには含まないようにした。

5.2　実装

各言語の言語解析処理は３節に示した通りに行った。日本語の単語分割についても他言語と同様の解析処理によって行った。統計翻訳の実装には Moses（ver. 1.0）を用いた。 各モデルの重みは開発用データを用いた誤り最小化学習（MERT）により最適化した。４節で触れた並べ替えの制限値は開発セットにおける RIBES の最大値に基づいて決定した。

5.3　評価尺度

翻訳結果の評価はテストセットの日本語側を用いた自動評価によって行った。特許翻訳の自動評価においては

語順を重視した評価尺度である RIBES が人手評価と非常に高い相関を持つことが知られており [1]、本実験でも RIBES を主要な評価尺度とする。また、統計翻訳の自動評価で最もよく用いられる BLEU も合わせて評価した。評価結果を表 5 に示す。

表 5: 翻訳の自動評価結果

	方式	RIBES（%）	BLEU（%）
英日	事前並べ替えあり	78.6	37.4
	事前並べ替えなし	72.1	34.8
中日	事前並べ替えあり	87.8	49.8
	事前並べ替えなし	85.6	47.9
韓日	事前並べ替えなし	94.3	70.4

5.4 英日翻訳

事前並べ替えによって、事前並べ替えを行わない方式から RIBES で 6.5% の性能向上を達成でき、本システムの事前並べ替えが有効に働くことが確認できた。

5.5 中日翻訳

事前並べ替えによる性能向上は RIBES、BLEU とも 2% ほどであり、英日の場合には及ばないものの、効果が確認できた。本実験のテストセットにおいては事前並べ替えなしでの性能が RIBES で 85.6%、BLEU で 47.0% と比較的高く、事前並べ替えによる性能向上の余地が限られていた可能性がある。事前並べ替えなしでの性能が高かった理由としては、対訳データ量が英日の 2 倍以上あること、特許において顕著な長い名詞句において中国語と日本語の間の語順の違いが小さいことが考えられる。

5.6 韓日翻訳

RIBES で 94.3%、BLEU で 70.4% と非常に高い翻訳性能を示しており、単語適合率が 86.3% にも達した。韓国語と日本語は文法的に非常に類似しており機械翻訳しやすいことはよく知られているが、本実験の結果により、改めて韓日翻訳における統計翻訳の有効性が確認できた。誤りの主因としては未知語、特に日本語でカタカナ語に相当する単語の翻訳漏れが多く見受けられた。

6 おわりに

本稿では、自動文対応付けによる大規模特許対訳コーパスを用い、高精度言語解析と統語的事前並べ替えに基づく英中韓 3 ヶ国語から日本語への特許翻訳システムについて報告した。特許文は専門用語が多く文も長いため、翻訳が難しいと考えられがちであるが、語彙や表現が定型的であって言語解析の対象として扱いやすく、大量の対訳データも利用できる点で統計翻訳が有効な対象であることが確認できた。

参考文献

[1] Goto et al. Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop. In Proc. NTCIR-10, 2013.

[2] K. Sudoh et al. NTT-NII Statistical Machine Translation for NTCIR-10 PatentMT. In Proc. NTCIR-10, 2013.

[3] 須藤他，語順の入れ替えに着目した特許の統計翻訳，Japio YEAR BOOK 2013, pp. 292-296, 2013.

[4] H. Isozaki et al. Head Finalization: A Simple Reordering Rule for SOV Languages. In Proc. WMT-MetricsMATR, pp. 244-251, 2010.

[5] M. Utiyama and H. Isahara. A Japanese-English Patent Parallel Corpus. In Proc. MT Summit XI, pp. 475-482, 2007.

[6] X. Ma. Champollion: A Robust Parallel Text Sentence Aligner. In Proc. LREC, pp. 489-492, 2006.

[7] 鈴木他，拡張ラグランジュ緩和を用いた同時自然言語解析法．言語処理学会第 18 回年次大会発表論文集，pp. 1284-1287, 2012.

[8] X. Carreras. Experiments with a Higher-Order Projective Dependency Parser. In Proc. EMNLP-CoNLL, pp. 957-961, 2007.

[9] J. Suzuki et al. Learning Condensed Feature Representations from Large Unsupervised Data Sets for Supervised Learning. In Proc. ACL-HLT, pp. 636-641, 2011.

[10] D. Han et al. Head Finalization Reordering for Chinese-to-Japanese Machine Translation. In Proc. SSST-6, pp. 57-66, 2012.