英中韓から日本語への特許文向け統計翻訳システム

須藤克仁　鈴木潤　秋葉泰弘　塚田元　永田昌明
NTT コミュニケーション科学基礎研究所
sudoh.katsuhito@lab.ntt.co.jp

概要

本稿では、英語・中国語・韓国語の特許文を日本語に翻訳するシステムについて報告する。本システムは、独自に収集した大規模特許対訳コーパスを用いた統計翻訳システムであり、特許分野に適応させた多言語形態素解析・依存構造解析とそれに基づく統語的事前並べ替えを行う（韓国語を除く）ことを特徴とする。

1　はじめに

各種産業における国際的競争は厳しさを増しており、外国における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となりつつある。特に調査目的で大量の文献を扱うような場合においては費用や時間等の面で機械翻訳が有望であり、すでに多くの翻訳システム・ソフトウェアが販売されている。こうしたもの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、この 10 年ほどの間で統計翻訳技術が急速に発展し、欧米を中心に広く実用化が進んでいる。一方、日本においては、日本語が英語等の他主要言語との語彙的・統語的差異が大きいことに起因して統計翻訳の性能向上が進んでいなかった。しかし近年の技術進展によって、昨年の国際会議 NTCIR の英日特許翻訳タスク [1] の人手評価において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結果を達成するに至った [2]。

筆者らはこれまでの言語解析や統計翻訳の知見に基づき、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を構築した。本システムはまず入力文の言語解析（単語分割・品詞付与・依存構造解析）を行い、解析結果に基づく統語的事前並べ替えを施した後（韓国語を除く）、統計翻訳によって日本語に翻訳する（図1）。統計翻訳に必要な対訳コーパスは、日本と各国の双方に出願された特許から収集した。また、特許文の言語解析に対応するため、既存の言語解析正解コーパスに加え、特許文を対象とした正解コーパスを新たに作成し、言語解析を特許分野に適応させた。統語的事前並べ替えについては主辞後置並べ替え [3] を依存構造に対して適用するための拡張を行った。以下本稿ではこれらの技術について述べるとともに、各国語から日本語への特許文翻訳実験の結果を示す。

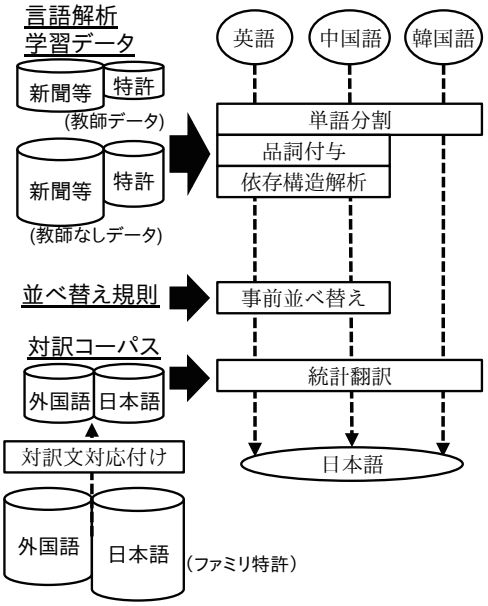

図1　本システムの大まかな構成

2　対訳文対応付け

統計翻訳の精度向上には十分な量の対訳コーパスが必要である。NTCIR 特許翻訳タスク [1] での学習用対訳コーパスの規模は、英日で約 320 万文、中英で約 100 万文であった。これと同等程度以上の対訳コーパスの作成のため、各国に出願された特許出願文書を用いて対訳文の対応付けと抽出を行った。

2.1　文書単位対応付け

特許は同等の内容で複数の国に出願されることがあり（パテントファミリー）、それらはほぼ対訳となっていると期待できる。NTCIR の特許対訳コーパスもパテントファミリーに基づいて対応付けられたものである [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情報（他国で先に出願した特許への参照）に基づき、一方が他方を優先権主張している出願、もしくは第三国の同一の出願に対して優先権主張をしている出願を対応付けた。複数の特許同士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 に対応する特許出願文書対のみを用いた。

2.2　部分構造対応付け

文書対の中で対訳文候補を探す際に、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ことは計算量・精度の両面で容易ではない。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、文対応付けの計算量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い[*1]、その他の箇所においても対訳文対応付けを行っている。

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。[4] 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文対応付け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 [4] 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 [5] 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、統計翻訳モデルで利用する自動単語対応付けに基づく確率的単語対訳辞書も利用した。

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[*2]上に述べた対訳文対応付けを行った。表1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。英日の対訳コーパスについては NTCIR 学習コーパス (約 320 万文) とほぼ同等の規模となったが、以下の点で NTCIR コーパスとは異なっている。

- 文献年代の違い: NTCIR コーパスは 1993-2005 年の公開特許公報に基づいている。
- パテントファミリーの扱いの違い: NTCIR コーパスは日本へ出願された特許に対して優先権主張している米国出願特許のみを用いている。
- 文対応付け対象の違い: NTCIR コーパスは「発明の詳細な記述」「発明の背景」に限って行っている。

表1 対訳文対応付けに用いた 1:1 対応の特許出願文書数と対応付けされた対訳文対数

言語対	1:1 対応文書数	対訳文対数
英日	117,620	3,642,583
中日	115,634	9,385,767
韓日	84,476	2,159,821

表2 言語解析モデル用訓練データの文数、単語数、および、ファイルサイズ (FS, bytes)

	教師あり						教師なし	
	新聞			特許			新聞	特許
言語	文数	単語数	FS	文数	単語数	FS	FS	FS
英	35K	1.0M	6M	10K	0.2M	1M	24G	0.2T
中	31K	0.7M	5M	35K	1.2M	7M	9G	0.1T
韓	52K	1.7M	19M	12K	0.6M	6M	—	—

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、依存構造解析の 4 つのサブタスクで構成されている。これは、生テキストが入力された際に、事前並べ替え型翻訳を行うために必要な言語解析の最小構成要素と捉えることができる。

これらの 4 つのサブタスクは、文分割、単語分割、品詞付与といった、いわゆる系列ラベリング問題として定式化できるサブタスクと、木構造予測問題として定式化される依存構造解析サブタスクの 2 種類の問題と考えることができる。そこで、本システムでは、系列ラベリング問題とみなせる 3 つのサブタスクは、文献 [6] の同時解析法を用いて一括処理を行う。また、依存構造解析は、文献 [7] の 2 次依存係り受け解析モデルを用いて依存木を予測する。

3.1 新聞データと特許データの混合による学習

特許用言語解析モデルを構築するのに利用した訓練データを表2に示す。訓練データには、正解の解析結果が付与された教師ありデータと、正解の解析結果が付与されていない教師なしデータを利用して、文献 [8] に基づいた半教師あり学習を行った。教師あり、教師なしデータともに、新聞記事と特許文書の 2 つのドメインから構成される。なお、教師なしデータに関しては、明確な単語、文区切りが不明なため、ファイルサイズのみを記載した。また、韓国語に関しては、通常の教師あり学習のみを行ったため、教師なしデータは用いていない。

3.2 言語解析性能の評価結果

特許文書から作成した正解データを用いて簡単な特許用言語解析実験を行った。表3に解析結果を示す。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（教師ありデータに出現しない語）に相当する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。

[*1] 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] 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。

表3　言語解析実験の評価結果 (単語、品詞、文はセグメンテーション F 値 (%) による評価、依存構造解析はラベルありの二項関係の正解率 (%) による評価)

	新聞				特許			
言語	単語	品詞	文	依存	単語	品詞	文	依存
英	99.9	97.0	97.4	91.8	99.3	94.7	99.5	86.7
中	95.1	88.4	81.4	82.0	92.7	85.5	97.4	81.2
韓	94.2	—	98.3	—	93.2	—	90.3	—

4　英語・中国語の事前並べ替え

統計翻訳における語句の並べ替えは、距離が大きくなると解探索の計算量が膨大となるとともにモデル化が難しくなる。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。

その中の一つのアプローチが事前並べ替え (pre-ordering あるいは pre-reordering) 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。事前並べ替えには統語的な並べ替え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句構造木における主辞後置化 (Head Finalization)[3]、及びその中日翻訳向け改良 [9] と同様の並べ替えを依存構造上で行うための拡張を行った。

4.1　英語依存構造に対する主辞後置並べ替え

依存構造は二分木でないため、二分木上の主辞後置化 [3] と同じように主辞となる単語を後方へ移動しても望ましい並べ替え結果が得られない (図2に例を示す)。これは、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右分岐型の構造において、主辞後置化によって名詞句ごと後置されるのに対し、依存構造において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]。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。その他、主格・目的格の格助詞に相当する仮想単語の挿入や冠詞の削除、並列句を並べ替えない例外規則については [3] と同様だが、複数形名詞の原形への変換は行わなかった。

4.2　中国語依存構造に対する主辞後置並べ替え

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、依存構造上の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。

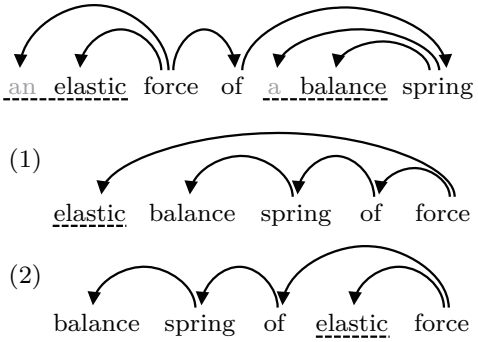

図2　依存構造における主辞後置並べ替えの比較。(1) のように主辞を後方へ移動するだけでは修飾語 “elastic” が離れてしまうが、本システムでは (2) のように前方の修飾語を主辞と合わせて移動することで正しく並べ替えられる (冠詞 a,an は削除される)。

表4　翻訳モデル用対訳データの文数及び単語数

言語対	文数	単語数（原）	単語数（日）
英日	2,836,988	80,038,847	94,349,535
中日	5,326,800	176,295,945	193,783,729
韓日	1,674,526	72,721,721	65,355,417

表5　言語モデル用日本語データの文数及び単語数

言語対	文数	単語数（日）
英日	3,642,563	149,542,488
中日	9,385,763	415,066,042
韓日	2,159,818	109,020,902

5　翻訳実験

本システムの性能評価のため、収集した対訳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。

5.1　実験データ

本実験に利用した学習データについて、翻訳モデル用対訳データについては表4に、言語モデル用日本語データについては表5、それぞれ示す。対訳データは 2 節で述べた文対応付け手法により得られた対訳文のうち、対訳文対応付けスコアが閾値以上、かつ両言語とも一定の単語数以内（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、開発データ・テストデータとして、英日・中日・韓日それぞれでおよそ 1,000 文ずつを利用した。開発データ・テストデータが含まれる出願は学習データには含まれていない。

表6　翻訳の自動評価結果 (DL は distortion limit)

(a) 英日翻訳

英日	DL	RIBES(%)	BLEU(%)
事前並べ替え	6	78.6	37.4
ベースライン	18	72.1	34.8

(b) 中日翻訳

中日	DL	RIBES(%)	BLEU(%)
事前並べ替え	3	87.8	49.8
ベースライン	9	85.6	47.9

(c) 韓日翻訳

韓日	DL	RIBES(%)	BLEU(%)
ベースライン	3	94.3	70.4

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。単語対応付けは MGIZA++ を用い、単語対応の重ね合わせは grow-diag-final-and、並べ替えモデルは wbe-msd-bidirectional-fe とした。また、言語モデルは modified Kneser-Ney 平滑化を行って学習した単語 6-gram モデルを用いた。各モデルの重みは開発セットを用いた誤り最小化学習 (MERT) により最適化した。並べ替えの制限値である distortion limit 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 についても合わせて示す。

5.4　英日翻訳

英日翻訳の評価結果を表6(a)に示す。事前並べ替えによって、事前並べ替えを行わないベースラインから RIBES で 6.5% の性能向上を達成した。また、bootstrap resampling により RIBES、BLEU とも統計的に有意な差を確認した。この結果は NTCIR における結果と概ね同等であり、異なるデータセットを用いた場合であっても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

中日翻訳の評価結果を表6(b)に示す。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統計的に有意な差が確認できた。本実験のテストセットにおいてはベースラインの性能が 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。ベースライン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

韓日翻訳の評価結果を表6(c)に示す。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。韓国語と日本語は統語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] Isao Goto, Ka Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K. Tsou. Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop. In *Proc. NTCIR-10*, 2013.

[2] Katsuhito Sudoh, Jun Suzuki, Hajime Tsukada, Masaaki Nagata, Sho Hoshino, and Yusuke Miyao. NTT-NII Statistical Machine Translation for NTCIR-10 PatentMT. In *Proc. NTCIR-10*, 2013.

[3] Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. Head Finalization: A Simple Reordering Rule for SOV Languages. In *Proc. WMT-MetricsMATR*, pp. 244–251, 2010.

[4] Masao Utiyama and Hitoshi Isahara. A japanese-english patent parallel corpus. In *Proc. MT Summit XI*, pp. 475–482, 2007.

[5] Xiaoyi Ma. Champollion: A robust parallel text sentence aligner. In *Proc. LREC*, pp. 489–492, 2006.

[6] 鈴木潤, Kevin Duh, and 永田昌明. 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, March 2012.

[7] Xavier Carreras. Experiments with a Higher-Order Projective Dependency Parser. In *Proc. EMNLP-CoNLL*, pp. 957–961, 2007.

[8] Jun Suzuki, Hideki Isozaki, and Masaaki Nagata. Learning condensed feature representations from large unsupervised data sets for supervised learning. In *Proc. ACL-HLT*, pp. 636–641, 2011.

[9] Dan Han, Katsuhito Sudoh, Xianchao Wu, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. Head Finalization Reordering for Chinese-to-Japanese Machine Translation. In *Proc. SSST-6*, pp. 57–66, 2012.