対訳パターンを用いた特許請求項文の翻訳手法

富士 秀　内山 将夫　藤田 篤　隅田 英一郎
情報通信研究機構
{fuji.masaru, mutiyama, atsushi.fujita, eiichro.sumita}@nict.go.jp

1. はじめに

特許明細書における請求項文は、特許発明が請求する権利範囲を規定するための重要なテキストであるが、現状の翻訳技術では高精度で翻訳することが難しい。請求項翻訳の難しさの要因は、長文であることと、独特の記述形式を持っていることにある。これらの問題を解決するための一つの手法として、請求項文の記述特性を利用したパターン翻訳が提案され、ルールベース翻訳における一定の有効性が示されている。本研究では、パターン翻訳の手法を統計的機械翻訳に適用し、訳質改善の効果を評価した。

2. 関連研究

日英中の特許文を対象とした機械翻訳研究では、NTCIR の特許翻訳タスク（"PatentMT"）[1]において共通的な訓練データ・テストデータの提供が行われ、これをきっかけとして業界をあげた研究開発が進み、翻訳精度が改善してきた。特許文書では、大規模な対訳データを構築することが比較的容易にできるため、特に統計的機械翻訳の精度向上につながってきた。さらに統計的機械翻訳では、構文情報の導入、また最近では構文並べ替え技術の進展により、PatentMT の英日特許翻訳タスクにおいて統計的機械翻訳がルールベース翻訳の精度を超えるまでに至った。

しかしながら PatentMT は、特許明細書の中でも比較的一般文に近い文体で記述される「実施例」を翻訳の対象としており、独特の記述形式かつ長文で記述される「請求項」の翻訳には未だに多くの課題が残されている。特許請求項文の平均文字数は 242 文字であるという分析結果もある[2]ほど、請求項は 1 文が長い。同分析によれば、新聞社会面記事の平均文長は 75 文字、政治記事は 56 文字とのことであり、請求項文が顕著に長いことがわかる。この長さの文を構文解析器や機械翻訳でそのまま処理しようとしても、そもそも計算コスト面から処理対象外となったり、処理できたとしても潜在的な曖昧性の膨大さゆえに処理精度が低くなったりする。

このような特徴を持つ特許請求項文の翻訳精度を向上させるための手法として、パターンを用いた翻訳の研究が行われてきた[3][4][5][6]。このうち定型利用翻訳[6]では、入力文を文節レベルで解析した結果に対してパターンを適用して構造部品に分解し、構造部品単位で機械翻訳を適用してからこれらを組み上げて訳文を生成する。請求項の独特の記述表現を手掛かりにパターンで分割を行い、短文の翻訳処理に落とし込むことで訳質向上をはかっている。この研究では従来、ルールベース翻訳を用いて構造部品の翻訳を行ってきたが、ルールベース翻訳では大量特許コーパスによるメリットを生かしにくいという課題があった。

3. 提案手法

本研究では、定型利用翻訳の翻訳エンジンとして統計的機械翻訳を利用することにより、大量のコーパスデータが存在する特許文データのメリットを生かした特許請求項翻訳の手法を提案する。

3.1 ベースライン

本研究では、上述の PatentMT を基本構成として想定している。ただし、PatentMT は、特許明細書の「実施例」が機械翻訳の対象であり、訓練・テストデータも実施例データを用いている。これに対して本研究では、NTCIR の実施例コーパスデータにさらに請求項データを足し合わせることによって請求項文にも対応できる混合コーパスを用意した。この混合コーパスを用いて統計的機械翻訳の訓練を行い、本研究のベースラインシステムとした。

請求項データとしては、文アラインメント手法[7]を用いて、米国特許明細書と日本特許明細書から自動的に対訳文を抽出し、ここから訓練データとテストデータを作成した。

3.2 提案手法　～対訳パターンによる翻訳～

本研究では、パターン翻訳における翻訳エンジンとして統計的機械翻訳を用いる構成とすることにより、統計的機械翻訳による流暢さと、パターン翻訳による大域的適切さの両立を目指す翻訳手法を提案する。①入力である英語請求項をあらかじめ人手で作成したパターンを用

いて分割し、②分割した構造部品に対してベースラインシステムで英日翻訳する。③そして対応する日本語パターンに沿って構造部品を組み上げて訳文を出力する。

図 1 は、提案手法の実行例である。ここではまず、(a)の英語原文に対して、正規表現による英語パターンが適用され、(b)のパターン処理結果が得られる。次に (b)英語表現に対応する日本語パターンが(c)のように生成される。次に、(b)の各構造部品に対してベースラインの翻訳エンジンを適用することによって、(d)の英日翻訳結果が得られる。最後に、(d)の各構造部品を組み合わせることによって、(e)の日本語訳文が得られる。

The actuator according to claim 1, wherein an even number of notches are formed in said body, and the displacement of said rod in the axial direction is extracted.

(a) 英語原文

PREA	the actuator according to claim 1
TRAP	wherein
PURP	an even number of notches are formed in said body, and the displacement of said rod in the axial direction is extracted

(b) 英語原文に対するパターン処理結果

PURP	an even number of notches are formed in said body, and the displacement of said rod in the axial direction is extracted
TRAP	wherein
PREA	the actuator according to claim 1

(c) 英語パターンに対する日本語パターンの出力

PURP	偶数個の切込みが形成されている前記本体であり、前記ロッドの変位には、軸方向 を抽出する
TRAP	ことを特徴とする
PREA	請求項 1 に記載のアクチュエータ

(d) 各構造部品に対する英日翻訳結果

偶数個の切込みが形成されている前記本体であり、前記ロッドの変位には、軸方向を抽出することを特徴とする請求項 1 に記載のアクチュエータ。

(e) 日本語訳文の出力

図 1. 提案手法の実行例

4. 実験設定

4.1 ベースラインの設定

実験に使用した統計的機械翻訳システムでは、言語モデル作成に KenLM [8]、フレーズ対の学習に SyMGIZA++ [9]、デコーダとして Moses 2.1.1 [10]を利用した。日本語のテキストデータは、MeCab 0.996 を用いて分かち書きしている。なお、モデルごとの重み付けの最適化処理は特に行っていない。

統計的機械翻訳システムの訓練用には、NTCIR 10 特許翻訳タスクでの実施例の英日対訳コーパス 1 万文対と、請求項の英日対訳コーパス 1 万文対を足し合わせて作成した、混合英日対訳コーパス 2 万文を用いた。

テストデータとしては、英日請求項対訳コーパスの対訳例文のうち、対応度の高さの順番で先頭から 100 文対を抽出し、これを用いた。

評価手法としては、BLEU-4 [11]と RIBES 1.03.1 [12]を用いた。

4.2 提案手法の設定

提案手法では最初に、正規表現で記述されたパターンを用いて入力文を分割するが、本実験では Perl 5.10.1 の正規表現を用いてパターンを記述した。提案手法によるシステムのためのパターン作成については次項で述べる。

4.3 パターンの作成

【英語パターン】

開発データの中の英語特許請求項文データを対象に分析を行い、人手によって英語パターンを作成した。分析結果に基づいて、英語請求項文は図 2 に示すパターン一覧で記述できることとした。

P_{e1}: SENT => PREA TRAP PURP+
P_{e2}: SENT => PREA TRAE ELEM+
P_{e3}: SENT => PREA TRAE ELEM+ TRAP PURP+

図 2. 英語パターン

ここで、PREA（preamble）は請求項の「前提部」であり、名詞句で構成される。ELEM（element）は請求項の「構成要素」であり、名詞句で構成される。PURP（purpose）は請求項の「目的」であり、動詞句で構成される。TRAE（transitional-for-element）は ELEM を導く「移行部」であり、TRAP（transitional-for-purpose）は PURP を導く「移行部」である。図中「+」は 1 回以上の繰り返しを表す。分析の結果得られた英語パターンを Perl の正規表現で記述した。入力英語文とのマッチングを行いマッチしたパ

ターンに沿って文が構造部品に分割されるようにした。

【日本語パターン】

上記開発データにおいて英語請求項に対応する日本語請求項を人手で分析し、同様に日本語パターンを抽出した。日本語パターンの一覧を図 3 に示す。

P_{j1}: SENT => PREA1 PURP+ TRAP PREA2
P_{j2}: SENT => PREA1 ELEM+ TRAE PURP+ TRAP PREA2
P_{j3}: SENT => ELEM+ TRAE PREA1
P_{j4}: SENT => PREA1 ELEM+ TRAE PREA2
P_{j5}: SENT => PURP+ TRAP PREA1
P_{j6}: SENT => ELEM+ TRAE PURP+ TRAP PREA1

図 3. 日本語パターン

日本語パターンは英語パターンと係り受けの方向が逆であるため、ELEM を導く移行部である TRAE は対応する ELEM の前にあり、PURP を導く移行部である TRAP は対応する PURP の前にある。また日本語では、「情報処理装置において、～することを特徴とする情報処理装置。」というように前提部を主題の位置に一旦提示するような記述方法がよく用いられる。この記述をカバーするために PREA1 と PREA2 を用意しており、例文では、一つ目の「情報処理装置」を PREA1、二つ目の「情報処理装置」を PREA2 としている。

【英語パターンと日本語パターンの対応付け】

前述のように日本語パターンでは前提部が２回提示される場合があるが、英語との対応を取るために文末の前提部１つにまとめたパターンを図 4.に示す。また、日本語パターンと構造部品の構成が同じであり、対応関係にある英語パターンを右辺に付記している。実験ではこれを用いて、英語パターンと日本語パターンの対応を求めて処理を行う。例えば（図 4 の 1 行目）、入力英語文が英語パターン P_{e1} にマッチしたとすると、対応日本語パターンは P_{j1} となりこの日本語パターンに沿って翻訳を行う。

P_{j1}: SENT => PURP* TRAP PREA :: P_{e1}
P_{j2}: SENT => ELEM* TRAE PURP* TRAP PREA :: P_{e3}
P_{j3}: SENT => ELEM* TRAE PREA :: P_{e2}
P_{j4}: SENT => ELEM* TRAE PREA :: P_{e2}
P_{j5}: SENT => PURP* TRAP PREA :: P_{e1}
P_{j6}: SENT => ELEM* TRAE PURP* TRAP PREA :: P_{e3}

図 4. 整理後の日本語パターンと、対応する英語パターン

5. 実験結果

表 1 に、ベースラインと提案手法における、翻訳精度の比較結果を示す。BLEU で比較するとベースラインと提案手法では 0.005 の差しかないが、RIBES で比較すると 0.395 から 0.765 と大幅な精度向上となっていた。

表 1. ベースラインとの比較結果

設定	BLEU	RIBES
ベースライン	0.360	0.395
提案手法	0.365	0.765

図 5 に、提案手法を導入したことによる改善例を示す。

この請求項文における英語前提部は”the actuator according to claim 1”であり、対応する日本語前提部は「請求項１に記載のアクチュエータ」である。日本語請求項文では一般的に、文末に前提部が配置されるべきだが、ベースラインシステムは、日本語前提部がバラバラに分かれ、また文の先頭に近い場所に訳出されていることにより、文全体として不適切な文を生成している。これに対して提案手法では、日本語前提部が一つにまとまって文末に訳出され、文の全体構造が正しく訳されるようになっている。

テスト文全体として、このように文の全体構造を正しく捉えて翻訳できた文が増えたことによって、RIBES 値の向上につながった。なお、ベースラインと提案手法では、個々の訳語や近距離の言語モデルはそれほど変わっていないため、BLEU 値の変化は小さいと考えられる。

The actuator according to claim 1, wherein an even number of notches are formed in said body, and the displacement of said rod in the axial direction is extracted.

英語原文

前記アクチュエータであることを特徴とする請求項１に記載の偶数個の切込みが形成されている前記本体であり、前記ロッドの変位には、軸方向を抽出する。

ベースラインの訳文

偶数個の切込みが形成されている前記本体であり、前記ロッドの変位には、軸方向を抽出することを特徴とする請求項１に記載のアクチュエータ。

提案手法の訳文

前記ボディに偶数個の切込みが形成され、前記ロッドの軸線方向の変位を取出すことを特徴とする請求 項1に記載のアクチュエータ。

参照訳

図 5. 提案手法による改善例

上述の提案手法を導入することによって大幅に訳質が向上したが、さらに残された誤り例を分析したところ、構造部品の種類（PREA, ELEM,…等）による訳し分けが十分にできていないことがわかった。この問題に対応するための改良手法として、ベースラインの訓練データにおいて、英語文と日本語文それぞれをパターンによって構造部品に分割し、対応する構造部品を対訳構造部品として切り出し、これをベースラインの訓練データに追加した。この改良手法による評価値を表 2 に示す。BLEU、RIBES ともに値がさらに向上していることがわかる。

表 2. 改良手法による評価値

設定	BLEU	RIBES
ベースライン	0.360	0.395
提案手法	0.365	0.765
改良手法	0.424	0.780

6. おわりに

請求項文の翻訳において、請求項文用パターンによって分割した構造部品を統計的機械翻訳で翻訳する翻訳手法を提案し、精度向上の度合いを測定した。

実験では、提案手法は RIBES 値においてベースラインを大きく上回る結果となった。このことから特許請求項の英日翻訳では、提案手法を用いることによってより適切な構造を持った訳文を得られるようになることがわかった。

今後は、各構造部品の特性を加味した訓練によって翻訳精度を向上させる方向で研究を進めたい。また、これまで人手で構築してきたパターンを自動的に取得する方向で研究を進めたい。

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 Proceedings of the 10th NTCIR Conference, pp 260-286, 2013.
2. 新森昭宏、奥村学、丸川雄三、岩山真、手がかり句を用いた特許請求項の構造解析､情報処理学会論文誌 Vol. 45 No. 3、pp 1-15、2004.
3. Jin’ichi Murakami, Isamu Fujiwara and Masato Tokuhisa, Pattern-Based Statistical Machine Translation for NTCIR-10 PatentMT. In Proceedings of the 10th NTCIR Conerence, pp 350-355, 2013.
4. 江原暉将、文パターンを用いた中日機械翻訳の精度改善、Japio Year Book 2014、pp241-251、2014.
5. Bart Mellebeek, Karolina Owczarzak, Declan Groves,Josef Van Genabith and Andy Way, A Syntactic Skeleton for Statistical Machine Translation. In Proceedings of EAMT 2006, pp xx-xx, 2006.
6. 富士秀、鄭育昌、角谷昌剛、長瀬友樹、中日・英日翻訳への定型利用翻訳技術の適用、言語処理学会 第 20 回年次大会、pp 380-383、2014.
7. Masao Utiyama and Hitoshi Isahara. A Japanese-English patent parallel corpus. In Proceedings of MT Summit XI, pages 475-482, 2007.
8. Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, Philipp Koehn, Scalable Modified Kneser-Ney Language Model Estimation. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp 13-21, 2013.
9. Marcin Junczys-Dowmunt and Arkadiusz Szał, SyMGiza++: A Tool for Parallel Computation of Symmetrized Word Alignment Models. In Proceedings of the International Multiconference on Computer Science and Information Technology, pp. 397–401, 2010.
10. Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the Association for Computational Linguistics Demo and Poster Sessions, pp 177–180, 2007.
11. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311-318, 2002.
12. Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, Hajime Tsukada, Automatic Evaluation of Translation Quality for Distant Language Pairs. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 944–952, 2010.