文パターンを用いた中日機械翻訳の精度改善

Improvement of Chinese to Japanese machine translation using sentence patterns

山梨英和大学人間文化学部人間文化学科教授 **江原　暉将**

PROFILE 1967 年早稲田大学理工学部卒。同年 NHK 入局。2003 年、諏訪東京理科大学教授。2009 年より現職。アジア太平洋機械翻訳協会 (AAMT) ／ Japio 特許翻訳研究会委員。

1　はじめに

機械翻訳において正確な翻訳を行うためには、適切な訳語を出力することと入力言語の構文構造を出力言語の構文構造に適切に変換することが必要であり、機械翻訳の 2 大課題とされている [1]。特許文書では、専門用語が多く使われ、かつ文長が長く複雑な構文構造を持つという特徴がある。これらの特徴は、上記 2 大課題を共に困難にさせ、機械翻訳の精度を低くする要因となっている。

筆者らは、上記課題のうち、特に構文構造の変換に着目して研究を進めている [2]。複雑な構文構造を正確に捉え、出力言語の構文構造に適切に変換する方法の一つとしての文パターンの利用がある。特許文は構文が複雑であるが、類似した構文が繰り返して出現するという特徴がある。このような場合、頻出する構文を文パターンとして用意しておき、翻訳に利用することで構文変換の精度を向上できる可能性がある。文パターンを利用するというアイデアは古くからあり [3, 4]、近年、特許文書にも適用されている [5, 6, 7, 8, 9]。特に、文献 [10] においては、特許文書を対象にした中日機械翻訳において、文パターンを用いることで acceptability 指標 [1] が 0.4 程度向上できることが示されている。

このように機械翻訳の精度向上に有効な文パターンの利用であるが、まず問題になるのが、頻出する文パターンをどのようにして収集し、翻訳規則としてルール化するのかということである。人手で収集・規則化する方法はコストがかかるという難点がある。一方、構文解析器を利用して自動または半自動で収集する方法が考えられるが、そもそも構文解析の精度が低いから文パターンを利用するのであり、この方法では本末転倒になってしまう。

日本語の文パターンを抽出する一方法として、文末表現に着目する方法がある [12]。日本語は、文の重要部分が文末にくる主要素後置型言語であるため、文末に着目することで文パターンが抽出できる。しかし、この方法は、主要素後置型言語ではない英語や中国語に適用することはできない。筆者らは中国語の頻出する文パターンを抽出する一方法として、2 段階の方法を提案している [2]。まず、日本語と中国語の文対応コーパスを用意し、その日本語部分の文パターンを文末表現を手掛かりにして求める。次に、同一の文末表現を持つ日本語文に対応する中国語文に共通する表現に着目して中国語の文パターンを求める。しかしこの方法では、日本語文パターンは自動的に求まるものの、対応する中国語文パターンは手動で求めなければならない。本報告では、統計的機械翻訳の中で用いられている単語対応機能を用いて、対応する中国語の文パターンを半自動的に求める手法を提案する。

以下、2 章で手法の詳細、3 章で評価実験について述べ、4 章でまとめを行う。

1　文献 [11] で提案された指標であり、精度を 5 段階に分類して評価するものである。文献 [10] では、5 段階を 1 ～ 5 に数量化して用いている。

2 中国語文パターンの抽出

以下の手順で中国語文パターンを抽出する。

a） 日中文対応コーパスを用意する。

b） 上記コーパスの日本語部分から文末表現に着目して日本語文パターンを抽出する。

c） 統計的機械翻訳ツールの Moses[13] を用いて、日本語の文パターンに対応する中国語文パターンを抽出する。

d） 得られた中国語文パターンと日本語文パターンを用いてパターン翻訳規則を作成する。

上記のうち a）と b）については昨年度の報告と同様の手法を用いた。その結果、パテントファミリーの要約部分から 43,404 文対の中日対訳コーパスが得られ、その日本語部分から日本語文パターンが得られた。頻度上位の 56 種類の日本語文パターンで 5,773 文（約 13%）をカバーしている。

c）の手順は、さらに下記にように細分される。

c1）Moses を利用して日本語文パターンに対応する中国語表現を求める。

c2）類似する中国語表現をクラスタリングによってまとめる。

c3）得られた各クラスターの中から代表的なパターンを中国語文パターンとして抽出する。

以下、日本語文パターン「・・・を提供することを目的とする。」を例として上記の手順を説明する。上記日本語文パターンの頻度は 425 件（カバー率は約 1%）である。

2.1 Moses を利用した中国語表現の抽出

Moses を利用して「提供することを目的とする。」に対応する中国語表現を求める方法には 2 種類がある。第 1 の方法は、フレーズテーブルから求める方法であり、「提供 する こと を 目的 と する」[2] の対訳としてフレーズテーブルに存在する中国語表現を求める。その結果、表 1 に示すような中国語表現が求まった。

表 1 フレーズテーブルから得られた中国語表現例

的 目的 在于 提供 一 种
的 目的 在于 提供
的 目的 是 提供 一 种
提供 一 种
本 发明 提供 一 种
发明 的 目的 在于 提供 一 种
目的 在于 提供
本 发明 的 目的 在于 提供
发明 的 目的 在于 提供
一 种
目的 是 提供 一 种
提供
目的 是 提供
的 目的 是 提供
目的 在于 提供 一 种

しかしながら、フレーズテーブルからだけでは、分離した文パターン（例えば「本发明以提供・・・为目的。」など）を求めることができない。そこで今回は、Moses の中で用いられている単語対応ツール Giza++[14] の結果を利用することにした。Giza++ での単語対応から得られた中国語表現の例を表 2 に示す。表中（?:.+）は任意表現を表す。任意表現部分を挟んで分離した文パターンが得られている。例えば「一 种（?:.+）的（?:.+）。」は「一 种 具有 较 高 的 记录效率 的 信息 文件 记录方法 及 装置 。」に適合する。第 1 の任意表現部分は「具有 较 高 的 记录效率」に対応し、第 2 の任意表現部分は「信息 文件 记录方法 及 装置」に対応する。この方法で、表 2 に示すような中国語表現が 150 表現求まった。

表 2 Giza++ から得られた中国語表現例

本 发明 的 目的 在于 提供 一 种 （?:.+） 的
本 发明 的 目的 是 提供 一 种 （?:.+） 的 （?:.+） 的
本 发明 提供 一 种 （?:.+） 的 （?:.+） 的
本 发明 的 目的 在于 提供 （?:.+） 的 （?:.+） 的
本 发明 的 目的 是 提供 （?:.+） 的
目的 在于 提供 （?:.+） 的 （?:.+） 。
本 发明 的 目的 是 （?:.+） 提供 （?:.+） 的
本 发明 的 目的 是 提供 一 种 （?:.+） 和 （?:.+） 的
本 发明 的 目的 是 提供 一 种 （?:.+） 的
本 发明 提供 （?:.+） 的
一 种 （?:.+） 的 （?:.+） 。

2.2 クラスタリングの実施

表 2 に示すような中国語表現の中で類似する表現をまとめるためにクラスタリングを行った。表現間の距離は以下のように求めた。表現 W_i と W_j の単語対応を動的計画法で求め、その結果、対応単語が異なる単語数を n_{ij} とし、W_i の単語数を n_i、W_j の単語数を n_j とすると

2 形態素解析器 (ChaSen) を用いて単語分割を行っている。また、句点 (。) は対応から除外している。

き距離 $d(i,j)$ は

$$d\ (i,j) = {}^{n_{ij}}/_{(n_i+n_j)}$$

で求める。クラスタリングは、最遠隣法で行い、26 のクラスターにまとめた。同一クラスターに含まれる表現の例を表 3 に示す。W_i=「可 提供 (?:.+) 一边 (?:.+)。」、W_j=「目的 在于 提供 (?:.+) 的 (?:.+) 。」に対して、n_{ij}=5、n_i=6、n_j=7 であるから、$d(i,j)$ =5/13 となる。

表 3　クラスターの例

可 提供 (?:.+) 的 (?:.+) ,
可 提供 (?:.+) 一边 (?:.+) 。
目的 在于 提供 (?:.+) 的 (?:.+) 。
提供 (?:.+) 防止电源 之间 (?:.+) 的 (?:.+) 。
提供 (?:.+) 的 目的 。

2.3　中国語文パターンの抽出

得られたクラスターの中から代表的な中国語表現を取り出して、中国語文パターンとする。現在、この部分は手作業になっている。その結果、表 4 に示す 12 個の中国語文パターンが得られた。

表 4　中国語文パターン

本发明以提供(?:.+)为目的。
本发明的目的在于提供(?:.+)的(?:.+)。
本发明的目的是提供(?:.+)的(?:.+)。
本发明提供(?:.+)的(?:.+)。
本发明的目的是(?:.+)提供(?:.+)的(?:.+)。
一种(?:.+)的(?:.+)。
本发明目的在于提供(?:.+)的(?:.+)。
本发明的目的在于(?:.+)提供(?:.+)的(?:.+)。
提供(?:.+)。
目的在于提供(?:.+)的(?:.+)。
提供(?:.+)的(?:.+)。
本发明的目的旨在提供(?:.+)的(?:.+)。

2.4　パターン翻訳規則の作成

表 4 に示す中国語文パターンと抽出に用いた日本語文パターンをもとにパターン翻訳規則を作成した。表 4 に対応する日本語文パターンを表 5 に示す。表中 <$1> や <$2> は中国語側の第 1 の任意部分や第 2 の任意部分に対応する日本語訳を挿入することを示す。現在、日本語文パターン作成部分も手作業となっている。

表 5　日本語文パターン（表 4 に対応）

本発明は、<$1>を提供することを目的とする。
本発明は、<$1><$2>を提供することを目的とする。
本発明は、<$1><$2>を提供することを目的とする。
本発明は、<$1><$2>を提供する。
本発明は、<$1><$2><$3>を提供することを目的とする。
<$1><$2>を提供する。
本発明は、<$1><$2>を提供することを目的とする。
本発明は、<$1><$2><$3>を提供することを目的とする。
<$1>を提供する。
<$1><$2>を提供することを目的とする。
<$1><$2>を提供する。
本発明は、<$1><$2>を提供することを目的とする。

3　文パターンを用いた翻訳実験

表 4 と表 5 の中日文パターン対を用いて、市販の規則方式中日機械翻訳システムで翻訳実験を行った。表 4 に適合する文を、中日対訳データ（43,404 文対）の中国語部分から抽出したところ 7,018 文あった。それらの中国語文を機械翻訳させたところ、文パターンを利用することで出力が変化した文は、2,564 文であった[3]。これらの文をテストセットとして機械翻訳結果を考察する。テストセットのうち、日本語部分の文末がパターン抽出に用いた「提供することを目的とする。」に一致する文は 133 文、それ以外の文末を持つ文が 2,431 文であった。前者をテストセット A、後者をテストセット B と呼ぶ。

これらの 2 種類のテストセットについて、文パターンを利用する場合と、利用しない場合の機械翻訳結果を、自動評価基準である RIBES[15] と IMPACT[16] を用いて評価した。評価値の平均を表 6 に示す。テストセット A および B ともに、文パターンを利用した場合のほうが評価値の平均値が向上している。

表 6　機械翻訳出力の自動評価結果

テストセット	文パターン	RIBES	IMPACT
A	非利用	0.653	0.304
	利用	0.806	0.423
B	非利用	0.699	0.333
	利用	0.766	0.375

3　機械翻訳システムの機能上、中国語文パターンがマッチしても必ずしも出力が変わるわけではない。

文パターン利用時の RIBES の値から非利用時の RIBES の値を引いた差の分布を図 1 に示す。IMPACT に対する同様の値を図 2 に示す。いずれもテストセット A のほうが B よりも正の方向によっているのが分かる。また、負の部分も存在し文パターンを利用することで評価値が下がる場合があることが分かる。

付録 1 に翻訳例を示す。テストセット A では多くが改善例であったが、テストセット B では改悪例も見られた。翻訳例（3）では、主動詞が「指示」であるが、文パターンを利用することで「提供」と誤って認識してしまい、その結果、翻訳が改悪されている。

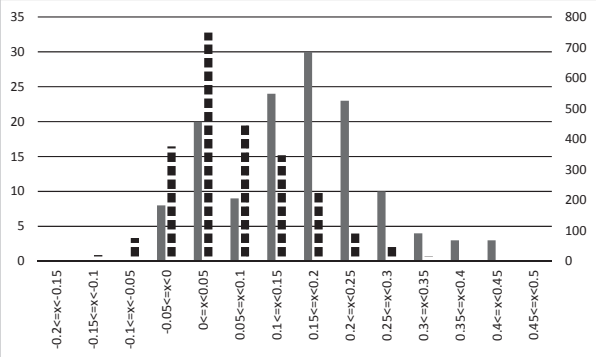

図 1 RIBES の差の分布（パターン利用－非利用）
テストセット A：塗潰、左側軸
テストセット B：横縞、右側軸

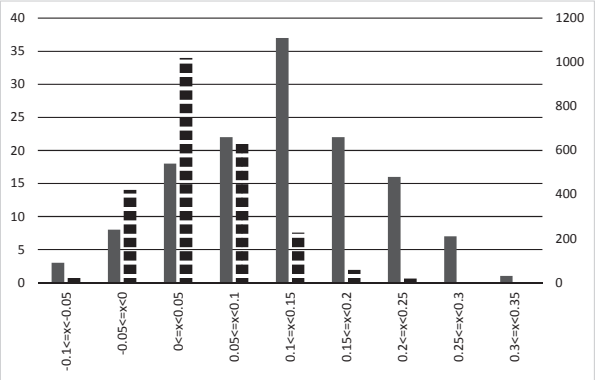

図 2 IMPACT の差の分布（パターン利用－非利用）
テストセット A：塗潰、左側軸
テストセット B：横縞、右側軸

4 あとがき

文パターンを利用した翻訳システムにおいて、パターン翻訳規則を半自動的に収集する手法について述べ、翻訳実験によって有効性を確認した。

今後の課題としては以下のことが考えられる。

- パターン翻訳規則収集の最終段階が手作業になっているので、この部分を自動化する。
- 今回は日本語文パターンとして、「提供することを目的とする。」のみを考察したが、他の文パターンについてもパターン翻訳規則を収集する。
- 文末以外の日本語文パターンについても自動収集する方法を考案する。

参考文献

[1] 江原暉将、田中穂積：機械翻訳における自然言語処理、*情報処理、自然言語処理技術の応用特集号*、Vol.34, No.10, pp.1266-1273, Oct. 1993.

[2] 江原暉将：中国語特許文書から文パターンを抽出する一方法、*Japio YEAR BOOK 2013*, pp.270-275, Nov. 2013.

[3] Hiroyuki Kaji, Yuuko Kida, Yasutsugu Morimoto : Learning Translation Templates from Bilingual Text, *Proceedings of COLING-92*, pp.672-678, Aug. 1992.

[4] 加藤直人：定型パターンを含む文の機械翻訳手法、*情報処理学会論文誌*、Vol.36, No.9, pp. 2081 – 2090, Sept. 1995.

[5] 船守茉美：中国公開特許公報の日本語への機械翻訳、*特技懇 262 号*、pp.3-10, Aug. 2011.

[6] Minah Kim : Current Status of Korea’s Machine Translation for Patent Domain Users, *第 1 回特許情報シンポジウム資料* , Dec. 2010.

[7] Wang Dan : Making Effective Use of Machine Translation for Patent Documents: Practice of CPIC, *第2回特許情報シンポジウム資料* , Nov. 2012.

[8] 张冬梅，刘小蝶，晋耀红：基于模板的汉英专利机器翻译研究，*计算机应用研究* , Vol.30, No.7, pp.2044-2047, July 2013.

[9] Jin’ichi Murakami et. al : Pattern-Based Statistical Machine Translation for NTCIR-10 PatentMT, *Proceedings of the 10th NTCIR Conference*, pp.350-355, June 2013.

[10] 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 回年次大会論文集*、pp.380-383, March 2014.

[11] Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita and Benjamin K. Tsou : Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop, *Proceedings of NTCIR-9 Workshop Meeting*, pp.559-578, Dec. 2011.

[12] 特許庁：日本語特許出願書類の中国語への機械翻訳に関する調査報告書、Feb. 2011.

[13] Philipp Koehn, Franz J. Och, Daniel Marcu : Statistical Phrase-Based Translation, *Proceedings of HLT-NAACL 2003*, pp.48-54, May-June 2003.

[14] Franz Josef Och, Hermann Ney. : A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, Vol. 29, No. 1, pp. 19-51, March 2003.

[15] Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, Kevin Duh : Head finalization: A simple reordering rule for SOV languages. *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pp. 244-251, July 2010.

[16] Hiroshi Echizen-ya, Kenji Araki : Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum, *Proceedings of the Eleventh Machine Translation Summit* (*MT SUMMIT XI*) , pp.151-158, Sept. 2007.

付録　翻訳例

(1) テストセット A (改善例)

【原文】本发明的目的是提供改善了数据信号检测特性的接收机、通信系统和接收方法。

【基準翻訳文】データ信号検出特性を改善した受信機及び通信システム、並びに受信方法を提供することを目的とする。

【機械翻訳文（文パターン非利用）】当発明の目的は改善にデータ信号の検出特性の受信機、通信システムと受け渡し方法を提供したのだ。

RIBES=0.650

IMPACT=0.338

【機械翻訳文（文パターン利用）】本発明は、データ信号の検出特性を改善した受信機、通信システムと受け渡し方法を提供することを目的とする。

RIBES=0.925
IMPACT=0.646

(2) テストセットB（改善例）
【原文】本发明提供一种可用以实现缩小设置空间的逆变器模块。
【基準翻訳文】設置スペースの縮小を図ることができるインバータモジュールを提供する。
【機械翻訳文（文パターン非利用）】当発明が1種提供してそれによって設置空間のインバータのモジュールを縮小するのを実現することができる。
RIBES=0.433
IMPACT=0.262
【機械翻訳文（文パターン利用）】本発明は、1種はそれによって設置空間を縮小するのを実現することができるインバータのモジュールを提供する。
RIBES=0.791
IMPACT=0.386

(3) テストセットB（改悪例）
【原文】提供有关服务的信息的服务器按照所请求的服务内容给终端指示测位精度。
【基準翻訳文】サービスに関する情報を提供するサーバは、要求されたサービスの内容に従って端末に測位精度を指示する。
【機械翻訳文（文パターン非利用）】関係するサービスの情報のサーバを提供して願い出たサービス内容によって端末に精度を測りを指示する。
RIBES=0.807
IMPACT=0.411
【機械翻訳文（文パターン利用）】サービスの情報のサーバに関して願い出たによってサービス内容は端末に精度を測りを指示するを提供する。
RIBES=0.623
IMPACT=0.370