中日・英日翻訳への定型利用翻訳技術の適用

富士 秀、鄭 育昌、角谷 昌剛、長瀬 友樹

(株) 富士通研究所

fuji.masaru@jp.fujitsu.com

1. 概要

長文に対する機械翻訳精度が低いという問題を解決するために、我々はこれまで、原文の定型性を利用して原文を構造部品に分割してから機械翻訳する「定型利用翻訳」の研究を行ってきた。定型利用翻訳の研究では、日本語文の定型性を利用して日英翻訳における長文に対する有効性を示すことができた。本研究では、同様の手法を英日翻訳および中日翻訳に適用できることを示し、翻訳精度向上における有効性を確認したので、その結果と言語対による特性について述べる。

2. 背景

ヨーロッパ言語間や日韓間のように、言語構造の類似した言語間の翻訳精度は、近年の統計翻訳の発達により飛躍的に進歩した。しかしながら、日英間や日中間のように、言語構造の大きく異なる言語間では、原文の文構造を正確にとらえてこの構造に対応する訳文構造に変換しなければ、適切な訳文を得ることができない。統計翻訳では、当初の単語や句レベルの処理に対して、構文的な処理を取り込むことによって、以前よりは大域的な構造をとらえる研究が行われるようになってきた[5]。しかしながら、ここで用いられる構文解析技術は従来のルールベースのものを超えるものではなく、長文の構造を正確にとらえることは、統計ベース翻訳とルールベース翻訳に共通の課題となっている。

3. 従来研究

定型性の高い文書では、長文であっても文の全体構造を比較的単純な定型パターンで書き表せるという特徴に着目し、筆者らは、日本語定型パターンをベースとした「定型利用翻訳」の研究を行ってきた[1][2][3][4]。日本語における定型性を利用して構造解析を行い、そこから英文を生成する構成の日英定型利用翻訳システムを開発し、特許明細書に現れる日本語長文の翻訳精度を向上できることを示した。

4. 本研究の目的

特許翻訳のもう一つの大きな需要としては、特許データベースに格納された大量の外国特許案件を日本語に自動翻訳して、日本人ユーザが簡便に斜め読みできるようにするというニーズがある。特に近年では中国語と英語で記述された外国特許案件が大量に出願されているため、本研究では、これまで日英方向で開発してきた定型利用翻訳技術を中日・英日翻訳に適用し、その効果を評価することを目的とする。

5. 本研究で構築したシステム

5.1. 対象文

外国語特許を日本語に自動翻訳して斜め読みする際にもっとも重要視されるのが、特許明細書の中の書誌情報である。本研究では、書誌情報の中で、テキストで記述される、「発明の名称」、「要約」、「請求項」を対象文として用いた。評価用として、英語案件と中国語案件を用意した。また、言語対の分析のために、パテントファミリー関係にある、英語案件と日本語案件の対、および中国語案件と日本語案件の対を用意した。

5.2. 言語対に固有の特性の分析

これまで日英翻訳を対象にして開発してきた定型利用翻訳技術を、英日翻訳および中日翻訳に適用するに当たって、各対象言語対の分析を行った。

文型のバリエーション

中日対訳案件および英日対訳案件それぞれ 200 件を一通り定型パターンとして記述しようとしたところ、①日英用に設計した構造部品のセット 6 種類をそのまま利用できること、②すべての文を定型パターンとして記述できることがわかった。ここで、連体・連用修飾の機能を持つ構造部品は、日英の場合と中日・英日では修飾関係が左右逆方向として扱う。

ちなみに、中国語構造と英語構造は、日本語構造と比較した場合に、文型のバリエーションが少なくなることもわかった。例えば、「X装置は、A部とB部を備える」という概念を表すのに、日本語だと以下にあげるような複数の文型のバリエーションがあるが、中国語や英語では、それぞれ一つのみである。これは、中国語も英語もともに、単語活用が少ないために、語順で文の構造を表す言語であるからだと思われる。

日本語：
a.「X装置は、A部とB部を備えることを特徴とする」
b.「A部とB部を備えるX装置である」
c.「X装置であって、A部とB部を備えることを特徴とするX装置」

中国語：
・「一种X，特征在于：包括A、B。」

英語：
・”X apparatus comprises: A part and B part.”

以上の分析結果から、今回の実験でも、中日翻訳および英日翻訳でも日英翻訳と同じ定型パターン記述の枠組みをそのまま利用し、定型パターンを記述することとした。

表現のバリエーション

日英定型翻訳における日本語構造解析では、入力文を定型パターンとマッチングさせるために、構造部品の表記バリエーションを記述する枠組みを用意した。例えば、「備える」と同じ概念を共有する動詞として「具備する」や「有する」等が使われており、これらの表記揺れを吸収しながら定型パターンとマッチングする枠組みを構築した。

今回、上記 200 件の中国語および英語の案件を対象に調査したところ、日本語と同様の表記揺れが起こっていることがわかった。中国語では「包括」に対する「由」や「它包括」等、また英語では”comprises”に対する”includes”や”has”等がある。このことから、中日翻訳と英日翻訳でも、日英定型利用翻訳と同じ枠組みを利用して、表記バリエーション規則を作成することとした。

5.3. システム構成

以上の言語対毎の分析結果に基づいて、中日・英日定型利用翻訳システムの構築を行った。図 1.にシステムの構成図を示す。基本的な構成は、中日・英日ともに共通である。

分析の結果から処理モジュール自体はこれまでの日英翻訳と同じ原理のものが使えることがわかっているため、図中の「定義ファイル」を今回の言語対（中日・英日）用および文種（特許抄録）用に新たに作成した。

なお、図中の「専用文法」は、我々のルールベース翻訳システムを利用して作成している。

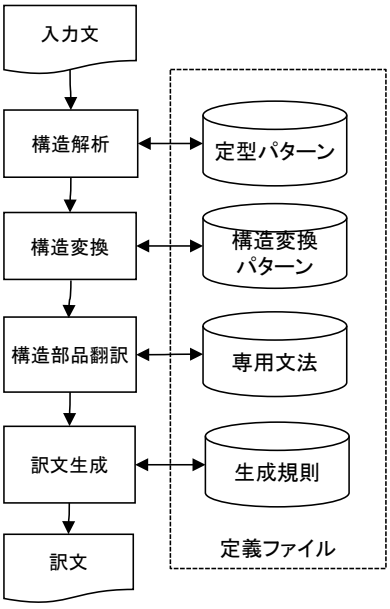

図 1. システム構成図

5.4. 処理フローの例

図 2.の中国語入力文を例にとって、構築した定型利用翻訳システムについて、処理フローを説明する。

一种透明荧光陶瓷封装的白光 LED 光源，由 LED 芯片、封装基座、支架、电极和透明荧光陶瓷封装材料构成，封装基座和支架组合安装。

図 2. 入力文（中国語）

構造解析

入力文が入力されると、図 3. の定型パターン（中国語）との照合が行われ、ヒットした定型パターンに沿って入力文が構造部品に分解される。ここでは、入力文と定型パターン P2 がヒットした結果、図 5.の構造解析結果が得られた。ヒットした定型パターン P2 に沿って、ラベルが付与されている。

ID	定型パターン（中国語）
P1	主題 主動詞 目的
P2	**主題 主動詞 目的* 述部** ⬅
P3	主題 主動詞 目的* 述部*

図 3. 定型パターンの例（中国語）

(一种)?([^，]*?)，其?特征(在于|是)?[，：]?(它?包括|由)：?([^。]+)。?

図 4. P2 に対応する表記揺れ正規表現（中国語）

ラベル	構造部品（中国語）
主題	透明荧光陶瓷封装的白光 LED 光源
主動詞	由
目的	LED 芯片
目的	封装基座
目的	支架
目的	电极和透明荧光陶瓷封装材料构成
述部	封装基座和支架组合安装

図 5. 構造解析結果（中国語）

構造変換

図 3.の定型パターンには、それぞれ、図 6.のように構造変換パターンが付与されている。構造変換では、この構造変換パターンを参照し、構造部品を対象言語の順番に並べ替える。パターンの左辺には、構造部品のラベルの並びを記述してある。「*」は繰り返しを表す。右辺は、変換後の構造部品の並びを表しており、左辺の各ラベルの左辺内の順番を表している。

図 5.の構造解析結果は、図 6.の構造変換パターン P2 とヒットして構造変換が動作し、図 7.の変換後構造部品が得られる。

ID	定型パターン（中）	並べ替え（日）
P1	主題 主動詞 目的	$1 $3 $2
P2	主題 主動詞 目的* 述部	$1 $3 $2 $4 ⬅
P3	主題 主動詞 目的* 述部*	$1 $3 $2 $4

図 6. 構造変換パターンの例（中⇒日）

ラベル	構造部品
主題	透明荧光陶瓷封装的白光 LED 光源
要素	LED 芯片
要素	封装基座
要素	支架
要素	电极和透明荧光陶瓷封装材料构成
主動詞	由
説明	封装基座和支架组合安装

図 7. 変換後の構造部品（日本語語順の中国語）

構造部品翻訳

各構造部品に対して、ラベルの内容に沿った適切な部品翻訳を行う。ここでは、構造部品のラベル毎に専用文法をあらかじめ用意しておく。

図 8.は、図 7.で得られた構造部品例に対する専用文法の適用結果である。図 7.のラベルに対応する専用文法が適用され、構造部品毎の訳文が得られる。

構造部品	構造部品訳文
透明荧光陶瓷封装的白光 LED 光源	透明蛍光セラミックパッケージ白色光 LED 光源
LED 芯片	LED チップ
封装基座	ベースパッケージ
支架	スタンド
电极和透明荧光陶瓷封装材料构成	電極と透明蛍光セラミックパッケージ材料構成
由	を備え、
封装基座和支架组合安装	ベースパッケージとスタンドは組み合わせて設置される。

図 8. 構造部品訳文

訳文生成

構造部品訳文を組合せて最終的な訳文を生成する。図 9.は生成された訳文の例である。

透明蛍光セラミックパッケージ白色光 LED 光源/であって、
LED チップ/と、
ベースパッケージ/と、
スタンド/と、
電極と透明蛍光セラミックパッケージ材料構成/と、を
備え、
ベースパッケージとスタンドは組み合わせて設置される。

図 9. 生成された訳文

6. 実験

6.1. 実験テキストの用意

実験では、中日・英日の特許明細書から、それぞれ 200 件の学習データと 50 件の評価データを分野が偏らないように抽出し使用した。明細書書誌の中の「発明の名称」、「要約」、「請求項」を対象に、中日・英日それぞれの定義ファイルを作成した。さらに今回は長文における翻訳精度向上を目的としたため、長文を多く含む「要約」と「請求項」を評価実験の対象とした。

6.2. 定義ファイルの作成

定義ファイルの作成段階では、学習セットの 200 件を見ながら定型パターンを人手で作成した。

中日翻訳のための定義ファイル

定型パターンおよび構造変換パターンに記述した文型数としては、「要約」用が 7 文型、「請求項」用が 12 文型と、少ない文型で学習セット全体をカバーすることができた。

定型パターンを記述する際の、表記揺れ吸収には、正規表現を用いたため、表記揺れバリエーションを定量的に表すことは難しいが、このような正規表現の例を図 4.に示す。正規表現の規模は、日本語構造解析の場合とほぼ同じである。

英日翻訳のための定義ファイル

定型パターンおよび構造変換パターンに記述した文型数としては、「要約」用が 5 文型、「請求項」用が 15 文型で全体をカバーすることができた。

表記揺れに対応するための正規表現は、日本語構造解析や中国語構造解析とほぼ同等であった。

6.3. 評価方法

文の全体構造の正確さを適切に評価するための統計的な評価手法がないため、各訳文に対して人手による評価を行った。

評価指標としては、acceptability 指標を用いている。以下の実験結果では、acceptability 指標の「AA」を「1」に、「A」を「2」に、「B」を「3」に、「C」を「4」に、「F」を「5」に読み替え、「1」～「5」を数値として扱って正解率を計算した。

なお、指標には「文法的に正しいか（Grammatically correct?）」という設問があるが、これは「文の全体構造が正しいか」に読み替えて評価を行っている。

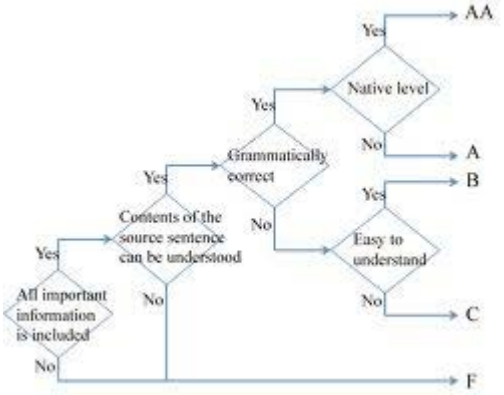

図 10. Acceptability 指標

7. 結果

図 11.および図 12.に実験の評価結果を示す。評価文は、特許明細書の「要約」と「請求項」である。

ここでは、縦軸に acceptability 指標の５段階評価値を示しており、最小値が１で最大値が５である。英日翻訳システムのほうが、中日翻訳システムと比べてもともとの翻訳品質が高いために、縦軸のプロット範囲を、中日と英日で変えてある。

比較のために、実用翻訳システムの評価値をあわせて記載した。これらシステムでは、「特許用」と銘打ってあるものについては「特許用」の欄に、特に何も記載のないものは左手の「一般用」のほうに記載した。「一般用」しか提供されていない場合には「一般用」のみに値を入れている。

なお、本手法の定型利用翻訳は、我々のルールベース翻訳（特許用）をもとに構築したものである。

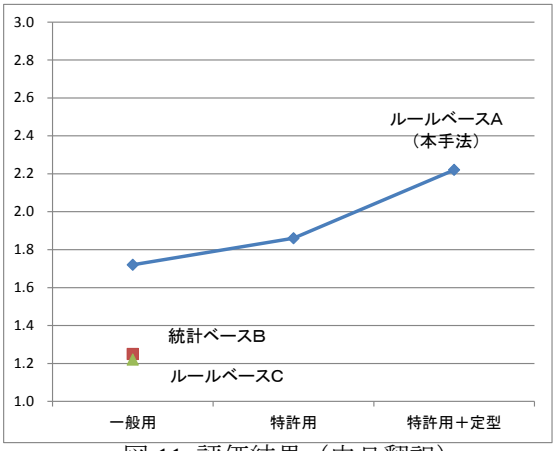

図 11. 評価結果（中日翻訳）

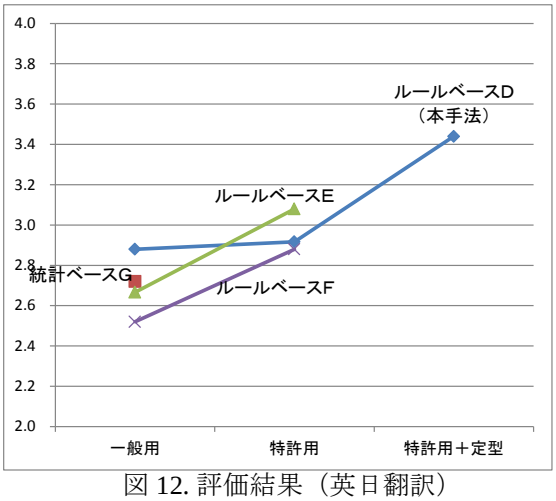

図 12. 評価結果（英日翻訳）

8. 考察

中日・英日翻訳に定型利用翻訳を適用することによって、いずれの言語対でも、従来翻訳システムと比べて翻訳精度が向上していることがわかる。実際の翻訳文を見てみると、全体の構造が正しく解析される文の比率が増えて、結果として、評価者にとっての読解性があがっていると考えられる。

定型パターンを登録した文種に限定されてはいるが、従来のルールベース翻訳における、単語登録のような対処と比べて、大きな改善幅が得られた。

9. まとめと今後

我々がこれまで開発してきた日英翻訳向けの定型利用翻訳技術を中日翻訳および英日翻訳に適用できることを示した。特に、長文における精度向上効果を評価するため、特許明細書における「要約」と「請求項」を対象に定義ファイルを作成して評価実験を行ったところ、従来では困難だった長文の全体構造の解析精度を向上させ、翻訳精度向上につなげることができた。

以上のようにして、対象を限定することによって長文の翻訳精度を向上させられることはわかったが、今後はその適用対象範囲を広げることを検討したい。今回扱った特許明細書の請求項等は定型性が非常に高い文の例であるが、そこまでの定型性はないものの、ある程度の定型性を持った文であれば世の中のコーパス中に多数存在する。このような文に対応できる、より適用範囲の広い処理の枠組みを検討していきたい。

また、文の全体構造のような大域的な特徴は人手でないと汎用化が難しい考えられるため、これまでは人手による定型パターンの作成を行ってきたが、今後は定型化のプロセスを部分的にでも自動化する手段を検討していきたい。

参考文献

[1]富士秀, 長瀬友樹, 潮田明, 増山顕成. 定型性の高い文章に対する日本語構造解析.言語処理学会第 14 回年次大会予稿集,2008.

[2]富士秀, 長瀬友樹, 潮田明, 増山顕成. 原文の定型性を活用した機械翻訳精度向上手法.言語処理学会第 15 回年次大会予稿集,2009.

[3]富士秀, 長瀬友樹, 潮田明, 増山顕成. 部品化された原文からの機械翻訳文生成.言語処理学会第 16 回年次大会予稿集,2010.

[4]富士秀, 潮田明. 定型利用翻訳における構造解析精度の評価.言語処理学会第 18 回年次大会予稿集,2012.

[5]星野翔, 宮尾祐介, 須藤克仁, 永田昌明(2013). 日英統計的機械翻訳のための述語項構造に基づく事前並べ替え. 言語処理学会第 19 回年次大会発表予稿集, 2013.