部品化された原文からの機械翻訳文生成

富士 秀[1]、長瀬友樹[1]、潮田明[1]、増山顕成[2]

[1]富士通研究所、[2]富士通

fuji.masaru@jp.fujitsu.com

1. 概要

本報告では、部品化された原文を入力として、そこから機械翻訳文を得るために必要となるシステム構成について説明する。原文を人手分割して各部品に機械翻訳を適用すれば翻訳支援を実現でき、他方、原文を自動分割して各部品に機械翻訳を実行すれば機械翻訳の高精度化を目指すことになる。
部品化された原文では、個々の部品が完全な文を構成していないため、通常の翻訳処理をそのまま適用するのではなく、部品に合わせた翻訳処理を行う必要がある。また、原文と訳文で部品の配置が異なるため、これを構造変換パターンとして持ち、訳文において自然な語順で出力する必要がある。
本研究では対象文書の調査を行うことにより、必要となる構造変換機構と、各部品に対する構造部品翻訳機構を割り出し、英文を生成するために必要なシステムを一通り完成させた。また、この日英対向け構成を、他の言語対で必要となる構成と対比して考察した。

2. 背景

長文に対する翻訳は、人間にとっても機械処理にとっても難易度が高い。このため、翻訳に先立って入力原文を部品化し、各部品に対して翻訳を行い、最後に各部品の翻訳を統合することによって、高品質な訳文を得る試みがなされてきた。

2.1. 入力文の部品化

ここでは、入力文の部品化について、日本語特許明細書を例にとって説明する。図 1.は翻訳対象の入力文であり、特許明細書特有の定型性に沿って記述されている。

用紙を収納するフィルム状の本体部と、密封加工によって形成された把持部とから構成され、把持部の中央には手掛け部が形成されると共に、下縁中央には切込部が形成されている。

図 1. 入力文

図 2.は図 1.の入力文に対する部品化の結果である。原文に現れる、「～と、」や「～から構成される」等の当該分野特有の表現を手掛かりとして、構造部品列への部品化を行う。ここで「ラベル」とは、切り出された各構造部品の原文中における役割を表しており、後段の処理ではこれら「ラベル」を参照しながら訳文生成が行われる。

ラベル	構造部品
要素	用紙を収納するフィルム状の本体部/と、
要素	密封加工によって形成された把持部/と
主動詞	から/構成/され、
説明	把持部の中央には手掛け部が形成される/と共に、
説明	下縁中央には切込部が形成されている/。

図 2. 部品化された入力文

2.2. 部品化の手段

実際の翻訳作業に照らして考えると、入力文の部品化の手段としては、主に以下の 3 種類が考えられる。

人手による部品化

翻訳者の作業効率化を支援するための翻訳者ワークベンチでは、翻訳者が原文を構造部品に分割し、これら構造部品に対して機械翻訳を実行することによって翻訳文を得る。
従来研究[1]では、”divide-and-conquer”という操作性・効率性の高い方式を採用することによって、人手分割および構造部品からの機械翻訳生成を支援し、翻訳作業の効率化を実現してきた。

自動処理による部品化

部品化のもう一つの用途は、機械翻訳のための前処理である。日英間のように構造の大きく異なる言語対の機械翻訳では、主に、ルールベース手法と、統計・用例ベース手法が試みられているが、いずれの手法も長文に対する翻訳品質は不十分である。
この問題に対して、筆者らはこれまでに、日本語に対する自動部品化の研究を行ってきた[2]。原文の定型性に着目した処理を行うことによって、ある程度の精度で入力文を構造部品に自動分割することが可能となった。翻訳文書の大半を占める産業分野文書（特許・法令・契約書、等）では特に定型性が高く、このような手法が有効である。
なお、日本語入力文に対する構造化では、複数の候補が出力される場合があるが、この中でもっとも評価値の高い候補を一つ抽出し、各構造部品に対して機械翻訳処理を行えば、全体として人手を介すことなく自動的に訳文を得ることができるようになる。

半自動処理による部品化

定型性が高い文書であっても、日本語入力文に対する構造化では複数の候補が出力され、人間でなければ正解を判断できない場合も多い。また、一つの構

造から生成される機械翻訳文が複数存在する場合も多い。このため、各段階で人間の選択が入るような利用形態も考えられる。
このような構成とすることで、人間による正確な判断と、機械処理による効率化の両方を狙うことができる。

3. 構築した英文生成システム

本研究では、部品化された日本語原文を入力として、そこから英訳文を生成するシステムを構築した。システムの構成図を図 3 に示す。
ここで、①、②、④、⑤は機械翻訳システムとはほぼ独立に構築でき、③は機械翻訳システムにある程度依存した内容となる。

①部品化された日本語入力文がシステムへの入力となる。人手分割によって部品化された入力文、もしくは自動分割によって部品化された入力文が渡される。
②構造変換処理では、あらかじめ用意した構造変換パターンを参照しながら、目標言語の構造になるように構造部品の並べ替えを行う。なお、ここで必要となる付随機能について後述する。
③構造部品翻訳では、各構造部品に対して、分野の定型性を活かした適切な専用文法を適用することによって構造部品の翻訳を得る。
④候補ソートでは、構造部品の候補を評価値順にソートする。
⑤最後に訳文生成では、構造部品単位で作成された翻訳を統合し、最終的に⑥英語訳文を出力する。

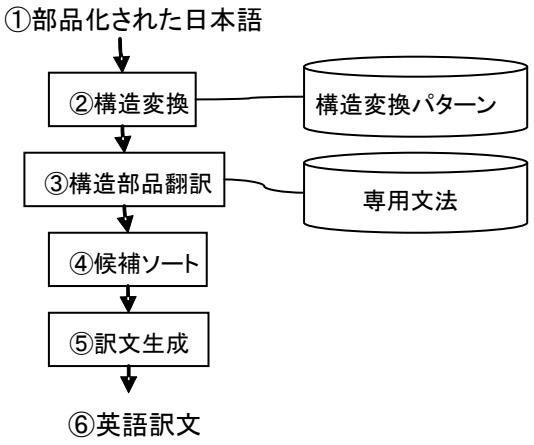

図 3. システム構成

以下に、システムの各構成要素について説明する。

3.1. 構造変換

構造変換では、対象分野に合わせてあらかじめ用意した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。図 4.は、構造変換パターンの例である。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。

ID	左辺		右辺
P1	要素* 主動詞 説明*	⇒	主題 $2 $1 $3
P2	主題 要素* 主動詞 説明*	⇒	$1 $3 $2 $4

図 4. 構造変換パターンの例

図 5.は変換前の構造の例であり、図 6.は変換後の構造である。図 5.の変換前の構造において、ラベルの並びを構造変換パターンの左辺と照合し、一致したら、右辺の並びに並べ替える。ここでは、図 6.の P1 の構造変換パターンがヒットして構造変換がなされている。

ここで、1 つの入力構造に対して複数の構造変換パターンがヒットする場合は、全ての変換を行って複数候補を生成する。このようにして、構造の異なる複数の候補が作成される。

ラベル	構造部品
要素	用紙を収納するフィルム状の本体部/と、
要素	密封加工によって形成された把持部/と
主動詞	から/構成/され、
説明	把持部の中央には手掛け部が形成される/と共に、
説明	下縁中央には切込部が形成されている/。

図 5. 変換前の構造

ラベル	構造部品
主題	
主動詞	構成
要素	用紙を収納するフィルム状の本体部
要素	密封加工によって形成された把持部
説明	把持部の中央には手掛け部が形成される
説明	下縁中央には切込部が形成されている

図 6. 変換後の構造

3.2. 構造変換に付随する諸機能

前項では、目的言語の文構造に合うように、構造部品を並べ替える手順を説明したが、目的言語の全体の構造を生成するには、以下の機能も必要になる。

訳語文字列の生成

目的言語の文を生成するには、原文を直訳しても得られない文字列を生成する必要がある場合がある。この機能では、訳文中の指定された位置に必要な文字列を生成する。
図 7.では、訳文において、「要素」の連続と「説明」の連続の間に必要となる”wherein”という訳語文字列を生成している。

ラベル	構造部品
主題	
主動詞	構成
要素	用紙を収納するフィ
要素	密封加工によって形成された把持部
	wherein
説明	把持部の中央には手掛け部が形成される
説明	下縁中央には切込部が形成されている

訳語文字列生成

図 7. 訳語文字列の生成

主題の補完

日本語では主語が省略されることがよくあるが、特に主題となる主語が存在しないと、英語の訳文として成り立たない場合が多い。
定型性の高い文書では、文章内で前方を参照すれば、省略された主題をほぼ確実に見つけることができる。特に特許文書では、「発明の名称」が全体の主題を表わすことが多く、そこから主要部分を切り出して持ってくることによって省略された主題を補完することができる。
図 8.は、主語の補完の様子を表す。英語訳文では必ず主語が必要になるが、日本語原文には存在しないため、文書中から主語を推定して補完している。

ラベル	構造部品
主題	***包装袋***
主動詞	構成
要素	用紙を収納 部
要素	密封加工によって形成された把持部
	“wherein”
説明	把持部の中央には手掛け部が形成される
説明	下縁中央には切込部が形成されている

補完された主題

図 8. 主題の補完

主語の結合

日本語の特許文書では、「○○装置において、～することを特徴とする○○装置である。」のように主語が重複して現れることがある。このため、構造変換に当たって、同じ事物を指し表わす主語を一つにまとめておく機能が必要となる。
ここで、同じ事物を指し現す複数の主語がまったく同じ表記で出現する場合は、単純にどちらを採用すればよい。しかし、表記が異なっている場合に、①同じ事物を指し示すかどうかの判断と、②統合する場合にどの表記を採用するかの判断は、難しい場合がある。特に②ではさらに、どちらかの表記をそのまま採用する場合と、二つの表記を融合してから翻訳に回す場合とがある。これらの事象に対応するためには、あらかじめ大量の対象文書を分析しておく。

文分割・文結合

翻訳を行う際、原文の 1 文を複数文として訳出したり、逆に複数の原文を 1 文の訳文として訳出したりすることがある。特に定型性の高い文書では、この文分割と文統合がある程度規則性を持って行われるため、構造変換の過程においてこれらの操作が自由に指定できる必要がある。
図 9.は、文分割機能の適用例である。分割した 2 文目には主題機能によって主題が補完されている。

ラベル	構造部品
文1	
主題	包装袋
主動詞	構成
要素	用紙を収納するフィルム状の本体部
要素	密封加工によって形成された把持部
文2	
主題	包装袋
説明	把持部の中央には手掛け部が形成される
説明	下縁中央には切込部が形成されている

図 9. 文分割の例

3.3. 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訳文を生成する。このために、以下の 2 段階の処理が必要になる。

専用文法の作成

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 10.は、「説明」ラベルを翻訳するための専用文法の例である。

日本語構造部品		英訳構造部品
主語が～される	⇒	主語 is 動詞+ed
主語が～される	⇒	主語 is to be 動詞+ed

図 10. 「説明」ラベル用専用文法の例

専用文法の適用

各構造部品に対して専用文法を適用する。
図 11.は、図 8.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 8.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
包装袋	wrapping bag
構成	comprises
用紙を収納するフィルム状の本体部	main part that stores the paper
密封加工によって形成された把持部	gripper formed by sealing process
“wherein”	wherein
把持部の中央には手掛け部が形成される	handling part is formed in the central part of the gripper
下縁中央には切込部が形成されている	cutout part is formed at the centre of the inferior edge

図 11. 構造部品訳文

3.4. 訳文生成

構造部品訳文を組合せ、最終的な訳文を生成する。
英文生成として、ここでは、先頭単語の先頭文字の大文字化、文末ピリオドの挿入、並列表現におけるセミコロンや”and”の挿入等がある。
図 12.は生成された訳文の例である。途中の処理段階で複数の候補があった場合でも、ここでは、それぞれの段階で最も評価値の高い候補を採用し、これらをつなぎあわせている。
細部の問題はあるものの、全体としては、元の機械翻訳の訳文より品質が向上している。

A wrapping bag
comprises:
a main part that stores the paper; and
a gripper formed by sealing process;
wherein:
the handling part is formed in the central part of the gripper: and
cutout part is formed at the centre of the inferior edge.

図 12. 生成された訳文

4. 考察

以上のようにして本研究では、日本語を構造部品に分解し、そこから機械翻訳で英文生成するための処理について調査を行ってシステム構築を行った。これによって、日英方向の翻訳の高精度化に向けて必要となるシステム構成を割り出すことができた。

これまでの一連の研究では、日英方向の他にも、英日方向および中日方向で同様の研究を行ってきたが、以下では今回の日英方向との比較を行い、各言語対の特性を考察する。

日英方向

日英方向では、日本語の解析における曖昧性が大きいことが最大の特徴としてあげられる。日本語文では、定型性の高い文章であっても、英語や中国語と比較すると多くの多様性がある。例えば、日本語では、主題が、文頭／文末／文頭および文末、に現れることがある。
日英方向の処理では、構造変換において、これらバリエーションをまとめあげる機構が必要になる。また、日英対に依存した構造部品翻訳が必要となる。

英日方向

定型性の高い英語文書では、日本語と比べて格段に構造が固定化されている。このため、英文の部品化は、パターンマッチ的な手法である程度の精度が得られる。また、英日方向で部品から訳文を生成する場合も、一意な対応の構造変換で処理できる場合が多い。
構成としては、一意の変換ができるシンプルな構造変換があれば十分で、あとは英日対に依存した構造部品翻訳が必要となる。

中日方向

定型性の高い中国語文書は、英語文書と同様に、構造が固定されている。英日方向と同様の手法によって処理することが可能である。
なお、日中方向の処理を行うためには、日英方向と同様の道具立てが必要になることが想定される。

5. まとめ

本研究では、日本語構造部品から英文を生成するために必要な機能の洗い出しを行い、これに沿ったシステム構築を行った。
今後は、構築したシステムに対する性能評価を行うと同時に、さらなる問題点の抽出と、改良による性能向上を進めることが求められる。

参考文献

[1] 潮田明, 富士秀, 大倉清司, 山下達雄. 機械翻訳と訳例検索を統合した翻訳支援システム.言語処理学会第 9 回年次大会予稿集,2003.

[2]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.

[3]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.