原文の定型性を活用した機械翻訳精度向上手法

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

一般的に、長文に対する機械翻訳精度は低い。本研究では、入力された長文が高い定型性を持っている場合に、その定型性を活用した処理を導入することによって翻訳精度向上を図った。
長文は、意味的な単位である「構成部品」から成り立つとみなすことができる。定型性の高い長文では、各構成部品が分野に特徴的な文字列表現によって区切られているため、これらを比較的容易に切り出すことができる。そして、切り出した各構造部品に対して部品の特性に即した翻訳処理を行い、これらをつなぎ合わせることによって高品質な最終訳文を得ることができる。
本研究では、今回このような構造単位の翻訳処理を導入した翻訳システムのプロトタイプを作成したのでこれについて述べる。

2. 背景

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。
ルールベース手法は、構文解析を中心とした手法であり、ある程度の長さの文であれば、正確な文構造で訳出が可能であるが、それより長い文になる曖昧性が増加して翻訳精度は低くなる。一方、統計・用例ベース手法は単語の連接を中心とした手法であり、フレーズレベルではルールベースよりも自然な訳が生成できるものの、構文的な構造の認識に難があり、特に長文での翻訳品質は低い。

長文の解析精度を上げるための取り組みとして、本研究ではこれまで、定型性の高い文章に対する日本語構造解析[1]の研究を行ってきた。これは、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）には長文が多いが、これらの文が定型性を持っており、複数の意味的ブロック「構造部品」から構成されるという特徴を活用して、原文の解析精度向上を図るものである。

3. 目的

本論文では、これまで研究開発してきた日本語構造解析によって得られた構造部品列を入力として、ここから訳文を生成する機械翻訳システムの構築を目指す。これによって、対象文書の定型性を活かし、機械翻訳の品質向上を図る。なお現段階では、プロトタイプ作成までを目標とする。

4. 従来技術

4.1. 従来の機械翻訳

従来の機械翻訳では、入力文に定型性がある場合でもその定型性を生かすことができなかった。ルールベース手法の構文解析にせよ、統計・用例ベース手法の文字列の合成にせよ、処理は基本的にボトムアップに行われるため、文全体に関わる定型性を反映させることは困難である。

以下本稿では、特許要約文を例にとって説明する。図 1 は翻訳対象の入力文であり、文全体としては、『「要素」および「要素」を「説明」した「主題」を「目的」する。』のように、定型性を持っている。

接合用鋼管を有する解体容易な柱と梁の接合構造およびその接合体を用いた接合方法を提供すること。

図 1. 入力文

図 2 は従来の機械翻訳の結果だが、入力文に存在する定型性を翻訳処理に生かすことができていない。ボトムアップ的な解析の結果、たまたま採用された構文解釈は誤ったものであり、誤った解析結果から生成された訳文も必然的に誤っている。さらに、『「提供する」は”to provide”に翻訳されやすい』というような、生成における定型性も活かすことができていない。

Offer an easy pillar that has the steel pipe for the joint to dismantle, the junction structure of the beam, and the joint method of using the zygotic.

図 2. 従来の機械翻訳の出力文（誤り）

4.2. 入力文の構造解析

ここで、本研究でこれまでに行ってきた、定型性の高い入力文に対する構造解析について触れておく。本構造解析手法では、入力文の定型性を生かした解析を行うことによって、入力文を「構造部品」に分割し、各構造部品に対してその役割を表す「ラベル」を付与する。

構造解析システムは、まず、局所的な処理である文節処理と、大域的な処理であるパターンマッチを融合した枠組みにより、分割候補を作成する。そして、作成された複数の候補に対して、定型性に鑑みた全体的なバランスの良さから評価値付けおよびランキングを行って、最終的に正解構造を導き出す。

図 3 と図 4 を用いて処理の流れを説明する。

図 3 は通常の文節処理である。なお、本構造解析システムでは、用いる言語処理は文節処理までとし、文全体としての構文解析は行わない。これは、長文の解析では、そもそも構文解析が正常に動作しない可能性が高いためである。

接合用鋼管を/ 有する/ 解体容易な/ 柱と/ 梁の/ 接合構造および/ その/ 接合体を/ 用いた/ 接合方法を/ 提供する/ こと/ 。/

図 3. 文節処理結果

図 4 は、パターン処理の例である。
各候補は文節を単位として候補が作成されている。例えば、「説明」は、主題部分を修飾する構造部品であり、複数の文節から構成されている。
なお、大域的なパターンとしては、この分野に頻繁に現れる構造パターンをあらかじめ用意しておき、これにマッチする文節の組合せを構造候補として出力する。

構造候補Ａ（正解）
構造パターン：　説明* 主題 説明* 主題 目的

ラベル	構造部品
説明	接合用鋼管を/ 有する/
説明	解体容易な/
主題	柱と/ 梁の/ 接合構造および/
説明	その/ 接合体を/ 用いた/
主題	接合方法を/
目的	提供する/ こと/ 。/

構造候補Ｂ（不正解）
構造パターン：　説明* 主題 目的

ラベル	構造部品
説明	接合用鋼管を/ 有する/
説明	解体容易な/
説明	柱と/ 梁の/ 接合構造および/ その/ 接合体を/ 用いた/
主題	接合方法を/
目的	提供する/ こと/ 。/

構造候補Ｃ（不正解）
構造パターン：　説明* 主題 目的

ラベル	構造部品
説明	接合用鋼管を/ 有する/
説明	解体容易な/ 柱と/ 梁の/ 接合構造および/ その接合体を/ 用いた/
主題	接合方法を/
目的	提供すること。/

図 4. 構造候補の例

最後に、全ての構造候補に評価値を付与し、評価値順に複数候補を出力する。評価値は、対象分野における定型性に鑑み、各構造の出現しやすさを表すように設定したものである。例えば、構造候補Ａの構造パターンは、ＢやＣのパターンよりも多くの加点を得るようになっている。また、例えば、構造中に並列要素がある場合、各並列要素の類似性等も加点として加味している。

5. 構造部品からの訳文生成

ここからは、本論文の目的である、構造化された入力文に対する訳文生成について説明する。

5.1. システム構成

図 5 は、本研究で作成したプロトタイプのシステム構成である。

システムへの入力は、構造化された入力文である。複数の構造候補の中から、もっとも評価値の高い構造候補が入力として渡される。

構造変換処理では、最初に、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。

次の構造部品翻訳では、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。

最後に訳文生成では、構造部品単位で作成された翻訳を組合せ、最終的な訳文を出力する。

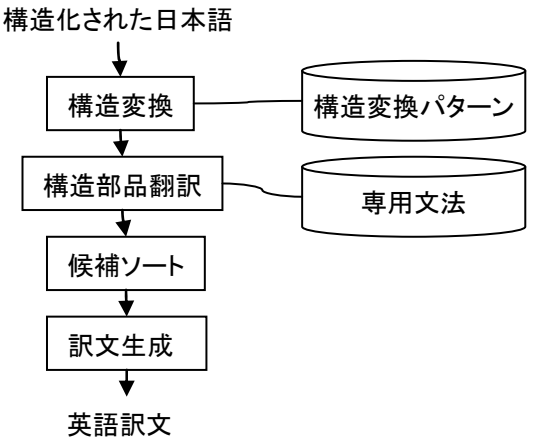

図 5. システム構成

5.2. 処理フロー

以下、図 1 の例文を使って処理フローを説明する。なお、処理の途中で、何箇所かで複数候補が出力されるが、これらの箇所について「★」印を付加している。

構造変換

構造解析では、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 6 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	説明* 主題 説明* 主題 目的	⇒	$5 $2 $1 $4 $3
P2	説明* 主題 目的	⇒	$3 $2 $1

図 6. 構造変換パターンの例

図 7 は変換前の構造の例であり、図 8 は変換後の構造である。図 7 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ことになる。（★複数候補）

ラベル	構造部品
説明	接合用鋼管を有する
説明	解体容易な
主題	柱と梁の接合構造および
説明	その接合体を用いた
主題	接合方法を
目的	提供すること。

図 7. 変換前の構造

ラベル	構造部品
目的	提供すること
主題	柱と梁の接合構造
説明	接合用鋼管を有する
説明	解体容易な
主題	接合方法
説明	その接合体を用いた

図 8. 変換後の構造

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、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 9 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
…連体形動詞～	⇒	～, 動詞 ing …
…連体形動詞～	⇒	～、which 動詞 …

図 9.「説明」ラベル用専用文法の例

1 つの構造部品に対して複数の専用文法が適用可能な場合には、これら複数の翻訳が得られることになる。例えば、図 10 では、「接合用鋼管を有する」という構造部品に対して、二つの構造部品訳文の候補が出力されている。（★複数候補） これは、図 9 の二つの専用文法が両方とも適用された結果である。

構造部品	構造部品訳文
提供する	to provide
柱と梁の接合構造	a junction structure of pillar and beam
接合用鋼管を有する	having the steel pipe for the joint
	which has the steel pipe for the joint
解体容易な	where there is the easy dismantlement pat
接合方法	a joint method
その接合体を用いた	wherein the zygotic was used

図 10. 構造部品訳文

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、並列表現におけるカンマや”and”の挿入等がある。
図 11.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。

To provide
a junction structure of pillar and beam,
having the steel pipe for the joint,
where there is the easy dismantlement pat, **and**
a joint method
wherein the zygotic was used**.**

図 11. 生成された訳文

6. プロトタイプ構築結果

以上の構成に基づいてプロトタイプを作成した。

6.1. システム出力の表示

本研究で作成したプロトタイプシステムは、各段階で内部的に複数候補が作成され、これが最後まで保持されるようになっている。
最終結果としてこれをどう出力するかは、実際の利用シーンや用途に合わせて変えることが可能である。

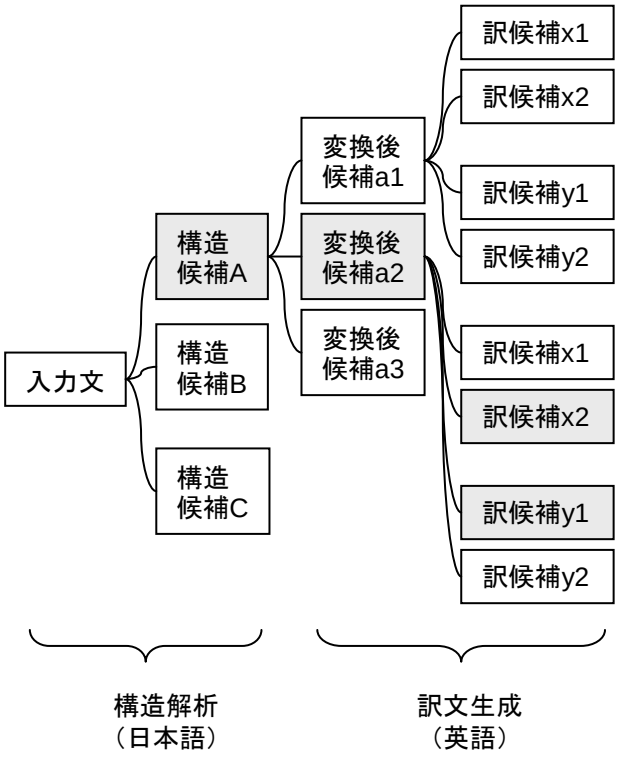

図 12. 各段階での複数候補

インタラクティブ翻訳

機械翻訳の途中に人間が介在することができる、または介在が望まれる場合には、内部的に生成される複数候補を、評価値順に見られるようにする。画面インタフェース上で、ユーザが自由に複数候補から選択したり、別の候補を選びなおしたりできるようにすることにより、簡単な操作で適切な訳文が得られるようになる。

自動翻訳

各段階での複数候補について、最も評価値の高いもの一つを選択していくことによって、通常の自動翻訳で使われるような、唯一の訳文を得ることができる。各段階で、対象分野の定型性に即した最適化が行われるため、結果として得られた訳文も、定型性を反映したものとなる。

7. 課題

現時点では、出力結果が一通り得られるようになった段階であり、目視上では期待された結果が得られるようになってきた。しかし、細部のチューニングはこれからである。
現時点でわかっている課題をあげる。

訳文全体としての整合性

今回試作したプロトタイプでは、それぞれの構造部品に対して、複数の可能な翻訳候補を出力することができる。人間の翻訳では、訳文全体のバランスを見ながら各パーツの訳を調整していくというプロセスがあるので、このプロセスを取り入れる必要がある。人間であれば、例えば、並列関係にある複数のフレーズは、同じような表現を使って訳出しようとするが、このようなプロセスをシミュレートする必要がある。

対象言語側の属性の導入

現在の設計では、入力文の構造部品には、入力側言語の特性に合わせたラベルが付与されている。しかし、より柔軟な翻訳を行うためには、出力側言語の特性も導入する必要がある。例えば、生成文法において、出力側言語の訳し分けを実現するような分類が必要になる。

第 2 階層以下の階層の処理

今回のプロトタイプでは、原文の構造候補は、意味的な構造の第 1 階層のみを扱うようになっている。（「階層」については[1]を参照されたい）しかし、実際には、第 1 階層の構造部品は、さらにその中に第 2 階層以下の構造を持っていることが多い。現状の、第 1 階層のみの分割では、構造部品が十分に短くならない場合には、第 2 階層以下の処理を行わなければ十分な翻訳精度が得られない場合がある。

8. まとめと今後

定型性の高い入力分に対して構造解析を行った結果をもとに、訳文を得るような機械翻訳システムのプロトタイプを作成した。
対象分野の定型性を利用した処理を行うことにより、解析の精度をあげ、また、訳文の質を高める構造を考案することができた。

今後は、より実用に近い訳文を得られるような枠組みを検討する。また定量評価を実施する。

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.