

機械翻訳システムにおける文書の形式構造の利用

3F-8

伊吹潤

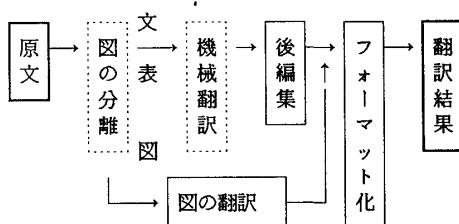
(株)富士通研究所

1.はじめに

現在の機械翻訳システムは基本的には一文ごとの処理を行うため、文脈に依存した処理を後編集や前編集の形で人手に委ねている。こうした処理は一般に深い意味処理を必要とするため、様々な文章を扱う機械翻訳において実現するには困難である。我々は、文章の形式的な構造に関する情報を利用することによって信頼性の高い文脈処理を行う機械翻訳の枠組みについての実験を行った。

2. 機械翻訳と文脈処理2.1 機械翻訳を利用した翻訳作業

計算機マニュアルの翻訳について実際に最終的な結果が得られるまでの過程を下図に示す。



機械翻訳を利用した翻訳作業例
(…で囲まれた部分は自動化されたもの)

ここでは翻訳結果の内容の手直し（後編集）以外にも、図の翻訳や翻訳結果のフォーマット化（フォント選択、段付け等の整形処理）の処理が人手で行われていることがわかる。翻訳処理の効率化を図る際には機械翻訳部自体の効率化と合わせてこれら後編集やフォーマット化の処理を自動化することも大きな目標となる。

2.2 後編集作業

現在、マニュアルの翻訳で行なわれている後編集の作業項目を示す。

- 語句の置き換え
- 冠詞、所有形容詞の付加、変換
- 名詞の単複の変換
- 文型の書き換え

これらの中では文型の書き換えがタイプ量が多く、最も大きな労力を必要とする。こうした作業が必要とされる原因としては特に複文中の文同士のかかり受け関係や、格の共有関係の解析誤りが挙げられるが、解析が成功している場合にもしばしば書き換えが必要になっている。後者の原因について考える。

a. 必須格の違い

日本語の文において述語の格要素（述語を修飾する要素）は省略可能であり、文脈等により特定できる場合はよく省略される。これに対し英語の場合は主語、目的語（他動詞の場合）は文の不可欠な要素である。このため翻訳の際に自然な英文とするために元の日本語文にはない格を補う必要が生じる。

現在の機械翻訳では省略された主語に対応するために受動態に変換する等の処理を行っているが、これは文の意味を変化させ、訳文として不適切なものにしてしまう可能性がある。（下図参照）

Utilization of formal structure

at machine translation system

Jun IBUKI

FUJITSU Laboratories, LTD.

WRITE: データを表示する。

機械翻訳 → WRITE: data is displayed.
後編集 → WRITE: WRITE displays data.

主語の省略された文の翻訳例

b. 文を名詞句にする

全く同一の文でもテキスト中のどこに置かれているかによって、翻訳結果の文体が異なることがある。例えばタイトルの場合は全体を名詞句とすることが多い。

例 同一文の訳し分けが必要な場合

データを変換する	←→ Data is converted (本文)
翻訳	converts data (本文)
結果	Conversion of data (タイトル)
	Convert data (アルゴリズムの一部)

このような処理はテキストの構造についての知識をもたない現在の機械翻訳の枠組みでは不可能であり、後編集の段階で人手で対応することになる。

2.3 文脈処理の問題点

省略された格の補完を考えてみよう。日本語では、一般に主語や目的語等の省略が行われるのは、あらかじめ主題化がされている場合に限ると言われており、主題から簡単に省略語句の補完ができるようになる。ところが実際に補完処理を行おうとすると、主題の解析自体や、複数の主題からの補完対象の選択などに大きな負担がかかってしまう〔1, 2〕。このような深い意味的な処理は様々なテキストを扱う機械翻訳システムに対して処理コスト（辞書、文法の整備等）、信頼性の点で問題があった。

3. 文書構造を利用した翻訳システム

上述の問題点に対し、本システムでは、形式的な構造化されたテキスト（章節の枠組みをもつもの）を対象とし、さらに補完の対象と補完の候補を信頼性の高いものに限定することにより対処を図っている。以後、このシステムの詳細について述べる。

3.1 処理内容

本システムは文書の形式的な構造情報をを利用して次に述べるような処理を行う。

i) 省略格の補完

本システムでは、翻訳に大きく影響するものとして、主語の補完のみに目的を絞っている。また、省略された主語の探索範囲をタイトル内の固有名詞に限ることによって処理誤りの減少を図っている。

ii) 生成スタイルの指定

文書中の各部分（タイトル等）に対しそれぞれに対応した生成スタイルを指定することにより、より読みやすい文を生成させる。

iii) フォーマット情報の自動処理

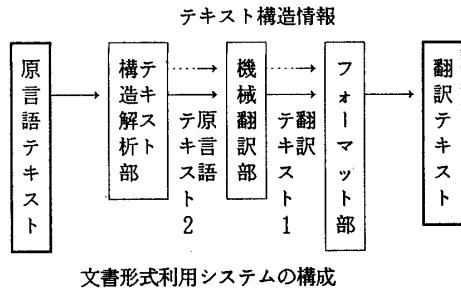
テキストの章節の構造を抽象的なレベルで保存して、それを元にして自動的な段付け、ヘッダの付加を行う。

3.2 システムの構成

システムの設計にあたり、機械翻訳システムとの独立性を保つために、本システムを前編集を主体としたシステムとして実現することとした。このために次のような基本方針を定めた。

- 翻訳前に予め文書構造に関する情報を各文に付加し、必要なら文自体を書き換える。これによって文脈情報を機械翻訳システム本体に伝達する。
- 翻訳処理本体の文法、辞書の変更を変更することによって文書構造に対応した処理を行なう。

これに基づいたシステムの構成を次に示し、各部について説明する。



a. テキスト構造解析部

テキスト構造解析部は内部にテキスト構造とヘッダとの対応表を持ち、これによってテキストの構造の解析を行なう。出力された各文には次の情報が付加される。

- i) 自分の直接属するテキスト構造のマーク
- ii) 自分を含むテキスト構造のもつタイトルのもつ固有名詞に関する情報

これは自分が属するテキスト構造に近いところのタイトルから順に記述される。

下に入力とその処理結果の例を示す。

第2節 WRITEコマンド
〔機能〕
データをスクリーン上に表示する。

テキスト構造解析部の入力例

/SECTION/TITLE/ WRITE コマンド
/SUBSECT/TITLE/ 機能
/PROP/ WRITE/SENT/ データをスクリーン上に表示する。

テキスト構造解析部の出力例

b. 機械翻訳部

ここでは従来の文毎に処理を行う機械翻訳の枠組みをそのまま用いている。但しテキスト構造解析部で付加された情報を扱うために、文法、辞書の内容を調整している。次に示すような付加的な処理を行う。

i) 省略された主語を補完する

文の主語がなく、かつ特別なアスペクトに関する情報がなければ主語の補完を行う。この場合、自らのもつ固有名詞のリストの中から順に主語の意味的制約を満たすか調べ、最初に条件を満たすものを選ぶ。

ii) テキスト中の位置に応じた生成スタイルを選択する

タイトル部、アルゴリズムのステップ部、本文部等のテキスト中の位置に応じた生成を行う。

下に出力例を示す。

/SECT/TITLE/WRITE command
/SUBSECT/TITLE/function
/SENT/WRITE displays data on the screen.

機械翻訳本体の出力例

c. フォーマット部

訳文に保存されているテキスト構造情報をを利用してレイアウト処理を行う部分であり、段付け、ヘッダの自動生成等を行う。

本システムによる処理の例を下に示す。

1.2 WRITE command
1.2.1 function
WRITE displays data on the screen.

フォーマット部の出力例

4. 評価

ソフトウェアの操作マニュアルを実験の対象に選び、評価を行った。詳細を下に示す。

〔評価の対象〕

題名 : FACOM PFD マニュアル
(ソフトウェアの操作マニュアル)
総文数 : 707 (タイトルを含む。図表は除く)

〔評価結果〕

ここでは格の補完作業に限って処理結果の評価を行った。

○後編集で対処が必要な文数

(括弧の中の数字は本システムでの処理が成功したもの)

主語の補完 95/103(79)
それ以外の格の補完 8/103(0)

〔考察〕

- i) 格の補完処理の中では、主格の省略への対応が大部分を占めている。これはコマンドの機能の説明文がほとんどであり、タイトルの直後の文であることが多い。
- ii) 但し単純に主語を補うのではなく、補完対象の名詞の複雑さ、タイトルの形式に応じて代名詞化する、不定詞句として主語なしのままにする等の選択が必要である。
- iii) 生成スタイルをテキスト中の位置に応じて変更する作業はなかった。これはあらかじめタイトルを名詞句とするような文体のチェック等がきちんと行われているせいだと思われる。

5. 終わりに

本試作システムの簡略化された処理でも、計算機マニュアルの翻訳に対し、ある程度の有効性が確認された。ただし計算機マニュアルの文章はあらかじめ文体の統制がされている、タイトルに示されるのはもっぱら自分で自律的に動作を行うもの（例えばプログラムのモジュール）となる、等の理由でかなり特殊な種類の文章であることも事実である。

同じマニュアルでも、自動車の整備マニュアルの場合はタイトルに示されるのは動作の対象となるような事物であることが多い。こうした文章に対処する際には、目的格を対象とした格の補完も行う必要がある。また論文等の場合、タイトルに現れるのは内容の全体をまとめる抽象的な主題が多く、補完対象はむしろ先行する文中に存在するようになる。

処理対象とする格の種類や補完する対象物の範囲が広ぐるに従って、処理の信頼性は低下する。このため処理内容の改善とともに、複数の候補をユーザーに示し、変換結果の確認を行いながら処理を行うようなインタラクティブな処理の導入が今後必要となると思われる。

参考文献

[1] 吉本 啓：日本語談話中のゼロ代名詞の同定
情報処理第36回全国大会3T-8(1988)

[2] 野垣内出、飯田仁：名詞句の同一性の理解と応用
情報処理第38回全国大会6D-5(1989)