

表層的な文脈情報を用いた自然な文生成の試み

6 B-6

荻野 紫穂 那須川 哲哉

日本アイ・ビー・エム株式会社 東京基礎研究所

1. 背景

英日機械翻訳において、代名詞の翻訳は解決が非常に困難なものの一つである。ゼロ代名詞の補完 [2] や代名詞の照応 [3]、あるいは特定分野における代名詞の訳語選択 [1] などと言及されることが多いが、自然な日本語を生成するために、英語で代名詞になっているものをそのまま代名詞で翻訳するか、あるいは照応先の名詞で置き換えるかなどの翻訳方法の差異の要因は、まだ明らかでない。代名詞処理は補完あるいは照応解析に主眼が置かれており、特に生成の際の省略に関しては、「日本語は省略が可能な言語だから重複する情報はどんどん省略すればよい」などと言われることがある割には、その妥当性についてはあまり触れられることがない。また、一般の機械翻訳システムでは、省略による誤解や誤った情報欠落を防ぐために、重複する情報であっても重複したまま出力することが多い。

人間の翻訳者による翻訳はある程度自然な文生成の一例と言える。本稿では、自然な日本語文を生成するための一つの試みとして、代名詞の翻訳方法に着目し、英文と人間の翻訳者によるその日本語への翻訳文を対象に代名詞翻訳方法を分類して、自然な日本語を生成する際の代名詞の生成処理方法について考察する。また、その生成処理の中で、表層の文脈情報をどれだけ適用することができるかを調査し、代名詞訳出における表層的な文脈情報の有効性を検証する。

2. 方法

今回の調査では、英語の代名詞 “you,” “it,” “this,” “that,” “they,” “these,” “those” を対象として、それらが翻訳日本語文の中でどのように訳出あるいは省略されているかを調査した。これら以外の人称代名詞は、データ中には出現しなかった。また、限定詞的な用法の “this,” “that,” “these,” “those” は、調査対象から外した。更に、我々の文脈処理 [3] から得られる表層的な文脈情報を適用することによって、このような翻訳がどの

	a	b	c	d	e	f	total
“you”	0	2	49	0	0	13	64
他代名詞	8	-	23	5	7	2	45
total	8	2	72	5	7	15	109

- a そのまま訳出
- b 翻訳慣例使用
- c 省略
- d 照応先を訳出
- e 限定詞+照応先
- f その他

表 1: 代名詞訳出方法

くらい可能になるかどうかを調査した。調査には、代名詞とその翻訳表現の対応が取りやすいことから、英語のコンピュータマニュアル約1,000文と、その日本語版における翻訳文をデータとして使用した。

3. 調査結果

3.1. 概論

表 3.1. に調査結果を示す。

“you” の訳出については、出現した 64 例中 49 例で省略されている。他研究でも言及される通り、マニュアル文では “you” は省略されるかあるいは “ユーザ” などの訳語が使用される慣例があると言われるが [1]、この調査結果もそれを裏付ける。“you” の 76.5% が省略されているのに対し、それ以外の代名詞は約半分が省略されているだけである。3.2. で “you” 以外の訳出方法に関する考察を述べる。以下、“you” 以外の代名詞を一般代名詞と仮に呼ぶ。

3.2. “you” 以外の代名詞訳出

翻訳の際に省略された一般代名詞 23 例のうち、16 例がコマンドなどの定義リスト中の定義文の主語として現れている。このような代名詞は定義リストの見出し語が照応先であるが、これらをそれぞれ日本語の代名詞や照応先である見出し語として訳出すると日本語としてしつこくなりやすいため、特にマニュアル翻訳において

Journal Number.

This is the number of the CICS journal that the system uses to store messages.

Operator ID.

This is the user ID of the person who last modified the row.

Sequence Number Seed.

This is the starting number that is to be used by the scanner when assigning sequence numbers to documents being scanned.

ジャーナル ID

システムがメッセージを記録するために使用する CICS ジャーナルの番号。

操作員 ID

最後に該当行を修正した操作員のユーザ ID。

順序番号シード

スキャンする文書に順序番号を割り当てるとき、スキヤナによって使用される開始番号。

図 1: 定義リスト

は定義リストの定義文の主語として使用される一般代名詞は、省略されやすいと言える。定義リストの例を図 1 に示す。英語原文中斜体で示された代名詞がここで該当する代名詞である。

日本語においては、異なる述部に対して同じ格役割を持つ要素は、一つにまとめて表現しやすく、これが日英機械翻訳などにおけるゼロ代名詞補完研究の一要因となっている。省略された一般代名詞 7 例中、これに該当する例が 4 例、“It - that” 構文など訳出しづらい例が 2 例で、これらに当てはまらない代名詞省略は 1 例であった。この例を下に示す。英語原文中斜体が省略された代名詞である。

x distinguishes the data sets.

It is A for the primary data set and B for the secondary data set.

x によりデータセットを区別します。

1 次データセットの場合は A、2 次データセットの場合は B です。

省略されずに翻訳された 22 例のうち、12 例が照応先名詞で置き換えられている。このうち 5 例について同文中に照応先名詞が存在しており、残り 7 例は同文脈内の他の文に照応先名詞が存在している。この意味で、22 例中 30% は表層的な文脈情報を使用することによって訳出が可能であると言える。また、照応先で置き換える際には限定詞を伴っていることも多い。

22 例中 2 例については、明確な照応先名詞を文章内に発見できなかった [4]。

4. まとめ

自然な日本語文生成の試みの一環として、代名詞の生成に着目した。人間の翻訳者による翻訳を自然な文生成の一例と見て、人手による英日翻訳において、英語の代名詞がどのように翻訳されているかを調査することにより、代名詞訳出処理の方法を考察した。

今回の調査は、データの絶対量が少なく、また、分野もコンピュータマニュアルに限定されているため、マニュアル翻訳における傾向と可能性を示すにとどまった。今後は更に深い調査を進めるとともに、その結果を実際の自然な日本語文の生成に生かしていく必要がある。また、適用分野や対象も翻訳や代名詞に限らず、重複する情報を以下に変形・省略して自然な日本語文を生成するかを考えなければならない。

文献

- [1] 森辰則, 瀧野弘幸, 中川裕志 (1995) 日本語マニュアル文における条件表現「と」「れば」「たら」「なら」から導かれる制約, 『自然言語処理』vol. 2, no. 4, pp. 19 - 35.
- [2] 吉本啓 (1986) 談話処理における日本語ゼロ代名詞の扱いについて自然言語処理研究会報告 56-4.
- [3] Tetsuya Nasukawa (1994) “Robust Method of Pronoun Resolution Using Full-Text Information,” *Proceedings of COLING-94*, pp. 1157 - 1163.
- [4] Eugene Charniak (1973) “Jack and Janet in Search of a Theory of Knowledge,” *Proceedings of IJCAI-73*, pp 337 - 343, 1973.