日本語を原言語とする機械翻訳における日本語等価変換機構

吉田雄太　　宮崎正弘

新潟大学大学院自然科学研究科
{yuta, miyazaki}@nlp.ie.niigata-u.ac.jp

1　はじめに

日本語から他の様々な言語への翻訳を行う際、障害となるものの一つに日本語特有の表現のもつ曖昧性をどうするかという問題が存在する。複数の語が連なって一つの意味を持つ連語や、本来の意味とは全く異なる意味を持つ慣用句、みかけの重文・複文などはそういった曖昧性を生む要因である。そのような表現を検出し、文構造が簡単で意味的曖昧さがより少ない意味的に等価とみなせる表現に変換することにより、変換部に負荷をかけずにより自然な翻訳が可能になる。本稿では日本語特有の表現をより翻訳に適した中立的な表現に変換するための日本語の意味的等価変換について論じる。

2　研究の背景

「努力によって成功。」という文を構文解析する場合、「努力/に/よ/っ/て/成功/。」というようにまず形態素解析によって文を最小構成要素である単語に分割し、同時に各単語の品詞を決定する必要がある。しかし、このような単語列をそのまま構文解析の入力として解析をすると、「によって」が、'格助詞「に」+本動詞「よっ」+助動詞「て」'として扱われてしまい、「努力」という「場所」に「寄って」、「成功」という誤った構造を導き出してしまう。また、「悪事から足を洗う。」などの文を解析した場合では、「悪事/から/足/を/洗う/。」と形態素解析されるが、この構造から、慣用句的用法の意味である「(悪事・悪い稼業などを)やめる」という意味を取り出す事は出来ず、きれいにするという意味での「足を洗う」と解析してしまう。こういった1：2などの対応づけがなされた非線形変換は、対訳対のパターンを用いて行うと、翻訳する言語の数だけそのパターンを用意せねばならず、パターン変換の負荷も大きくなる。

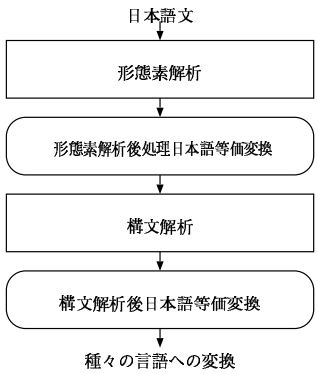

図1:日本語等価変換機構の流れ

そこで、図1に示すように、まず形態素解析の結果から構文解析における曖昧性の原因となる部分を判別、抽出し先の例であれば「努力/によって/成功/。」という単語列に書き換える。その上で構文・意味解析を行うことによって、「努力」という「原因」で、「成功」という正しい構造を導き出すことが可能になる。また、「悪事から足を洗う。」という文に対しては、構文解析後に文構造を利用しながら「悪事をやめる」という文に変換し、それを変換部にまわすことにより、「足を洗

Japanese Equivalent Rewriting Mechanism
In Machine Translation From Japanese
Yuta Yoshida, Masahiro Miyazaki
Niigata University

う」の意味の判断を行う必要無く線形変換が可能になる。このような処理を実現するために日本語等価変換機構を提案する。

3　日本語特有の曖昧性の種類

日本語特有の表現が持つ機械翻訳における曖昧性の種類としては以下のようなものがある。

表 1 . 日本語特有の曖昧性の主な例

連語	同形語	慣用句
として	行った	油を売る
という	一度	足を洗う
について	平野	水と油

3.1　連語の持つ曖昧性

連語とは、いくつかの単語が連鎖し、ある特定の機能を持つものである。

例 :「によって」
車が右によって いる。(動詞「寄る」のかな表記)
食生活 によって 健康は左右される。(格助詞相当の連語「によって」)

上記の例では、どちらも形態素解析の結果は「に/よ/っ/て」であるのだが、後者の文では、格助詞相当として扱われる連語の意味を持っている。これを動詞の「寄る」ととらえると後者の文は重文として扱われ、誤った構造を導き出してしまう。したがって構文解析以前にによってがどちらであるかを決定し、格助詞相当であるならばその形態素をひとまとめにしておく必要がある。

3.2　同形語の持つ曖昧性

同形語とは、みかけは同じ形であるが、語の読みが異なる語のことをいう。

例 :「行った」
山道を 行った。(読み「いった」)
運動会を 行った。(読み「おこなった」)

上記の例では、どちらもみかけは「行った」であるが、「いった」と読むか「おこなった」と読むかによって意味が大きく異なる。また、どちらの例文でも五段動詞音便形「行っ」＋助動詞「た」と解釈されるため、動詞の格パターンなど意味情報を利用してその読みを決定しておく必要がある。

3.3　慣用句の持つ曖昧性

ここでいう慣用句とは、二つ以上の語が結合することで、全体として一つの意味を持つようになったものの事をいう。

例 :「足を洗う」
悪事から 足を洗う。(慣用句的意味「(悪事などを) やめる」)
家に帰ったら汚い 足を洗う。(動詞「洗う」)

上記の例では前者が慣用句的意味の「足を洗う」で、後者が動詞「洗う」の本来の意味である。このようなものに対しては格情報を利用した対訳パターンを用意し、変換部で処理しようとすると多言語翻訳では、こういった対訳パターンを翻訳対象の言語の数用意して実装しなければならず、変換部に多大な負荷を生じることとなる。よって構文解析後に意味的に等価でより文構造の簡単な日本語文に変換を行うことが望ましい。

4　日本語等価変換機構

日本語等価変換機構は 3 で述べた日本語特有の表現を機械翻訳で処理しやすいように意味的にほぼ等しい範囲での書き換えを行う機構である。

4.1　連語処理

ある種の連語は構文解析において曖昧性の原因となるため、形態素解析後に連語書き換え規則を適用することにより行う。

連語書き換え規則の例
＊(ダ型静詞、「に」に接続する副詞、場所を表す名詞を除く), ［に (格助詞), よ (五段動詞語幹),っ (五段動詞語尾), て (既定の助動詞)］, ＊（形式動詞を除く）
→によって (格助詞)

字面と品詞の組み合わせが一つの単語に対応し、その単語の連鎖列によって [ ] の単語列を矢印の右辺に示す格助詞相当の連語に書き換えるようになっている。また、[ ] の前後の単語列は連語書き換え規則の適用の可否やその条件を表す。なお、これらの条件は省略も可能であり、その場合、無条件に連語書き換えを行う。

4.2 同形語処理

同形語は読みを付加するのみで構造の変換を必要としないため、同形語処理は同形語書き換え規則を適用することによって形態素解析後に行う。

同形語書き換え規則の例
＊（名詞）＜人間活動＞, を, ＊,[行（五段動詞語幹）,っ（五段動詞語尾）]
→行（五段動詞語幹）＜読み「おこな」＞,っ(五段動詞語尾)
＊（名詞）＜場所＞, を, ＊,[行（五段動詞語幹）,っ（五段動詞語尾）]
→行（五段動詞語幹）＜読み「い」＞,っ(五段動詞語尾)

基本となる形式は連語と同一であるが、単語の意味および読みの情報を表せる＜＞部が拡張されている。これを付加することにより同形語の処理をほぼ同じ枠組で行えるようにしてある。

4.3 慣用句処理

慣用句はその処理上文構造を必要とするため、慣用句処理は構文解析後に慣用句書き換え規則を適用することにより行う。

慣用句書き換え規則の例
(N1 から) 足を洗う
→ (N1 を) やめる
非修飾属性:足, 非関連語：洗う,N1 属性:人間活動, 職業

この規則は、格の変換を行うため、適用の条件は厳しいものとなっている。「洗う」の意で用いられた文を「やめる」と変換しないためである。その他に、「足」に「きれいな」などの修飾が行われない、「石鹸」などの「洗う」の関連語が文中にない、格変換を行う場合 N1 は人間活動あるいは職業という適用条件が付加されている。また、慣用句としての用法でしか解釈されないもの、例えば「道草を食う（寄り道する）」や「拍車をかける（速める）」「根掘り葉掘り（しつこく）」や、「うなぎ登り（急激に）」などについては、一意に変換するものとする。

5 実験

機械翻訳における日本語の意味的等価変換の有効性を確認するために小規模な実験を行った。

実験の方法は、規則に記述された連語・同形語・慣用句を含む文をインターネット上から無作為に抽出し、その文が意味的に正しいかの調査を行った。

5.1 評価

実装したルールのうちのいくつかについて、その語を含む文 100 文を抽出した結果を表 2 , 表 3 , 表 4 に示す。

表 2 :「という」の評価実験の結果

	一般用法の文	連語の文
総数	15	85
解析正解数	11	85

表３：「行った」の評価実験の結果

	おこなった	いった
総数	18	82
解析正解数	18	75

表４：「足を洗う」の評価実験の結果

	慣用句用法の文	一般用法の文
総数	56	44
解析正解数	38	44

5.2 考察

変換を誤った文の中から、いくつか注目すべきものをあげておく。

- 勝利という声が響き渡った。

  この文に関しては「という」は同格の格助詞相当とも取れるし、「と言う」と実際に言ったとも取れる本質的曖昧な文であるためである。こういった文に対しては、文脈から判断する必要がある。

- 植林を山へ行った。(「いった」と読み付け)

  この文では、規則の優先順位が直近の名詞句になっているため、「山へ行った（いった)」を優先したために失敗している。本来意味的には「植林を山に行った。」の表現が正しいと考えられるが、文として「植林を山へ行った」も誤った文ではない。こういったいくつかの文に対しては、係り受け構造を利用して読み付けをすることも必要である。

- レースから足を洗う→レースをやめる

  この文では、「レース（の世界）をやめる」という意味で用いられているものを、そのまま「やめる」に変換したため、「(一つの）レースをやめる」という別の意味の文となってしまっている。この場合「レース」に「レース(の世界)」という職業的な意味と、「(その時行われた一つの）レース」という瞬間的な事象の二つの意味が存在するためこのような誤りが生じていると考えられる。このような、省略された意味や多義語の場合にどの程度まで判断できるかというのも大きな問題である。一つの対処法としては、「N1から足を洗う」のN1の意味制約（この場合は職業、人間活動）を見て、N1がその意味制約と異なる語義を持っていた場合、適切な形に変換してから（この場合は「レースの世界」)、補完するという方法が考えられるが、世界認識なども必要となり容易ではない。

このようにまだまだ問題点は多いものの元来、翻訳は近似であるので、大きくニュアンスがはずれなければその言語内であらかじめ構造が簡素で意味的に解釈しやすい表現に変換することは機械翻訳において有効であるといえる。しかし、「足を洗う」を「やめる」に変換した時、そこには日本語において感じる印象の違いが存在することもまたあきらかな事実である。どこまでが翻訳に影響を与えない日本語等価変換で、どこからがそうでないのかと言う問題は、今後の大きな課題であるといえる。

6 おわりに

本稿では日本語を原言語とする機械翻訳における日本語等価変換機構を用いることにより、解析における様々な曖昧性を持つ連語や同形語、慣用句等に対し効果的に書き換える機構について論じた。この枠組みを利用して、中間表現を利用した多言語翻訳への応用や、異なった形態素解析システムの出力結果への変換など様々な応用が期待される。

以下のような課題が残されている。

- 慣用句の誤用に対応した拡張
- 文脈に依存した意味決定への対応
- 主格の省略などへの対応

参考文献

[1] 尾島、宮崎:日本語形態素解析システムにおける部分的再試行機構の導入とその効果、情報処理学会第 58 回全国大会 1E-4(1999)