

パターンベース翻訳システム PalmTree の訳語選択

3N-6

宮平知博、神山淑朗、羽鳥洋美

日本アイ・ビー・エム株式会社

1. はじめに

パターンベース翻訳システム PalmTree [1,2,3]は、同期文法[4]の考え方に基づいて翻訳の解析から生成までを翻訳パターンという一つの枠組みで処理する英日翻訳システムであり、翻訳ソフト「インターネット翻訳の王様」の翻訳エンジンとして使用されている。機械翻訳システムが今後ますます広く使用されていくためには訳質向上が不可欠であるが、そのためには原文の正確な解析と共に適切な訳語の選択が重要な鍵となる。本稿では、PalmTree の訳語選択機構と今回新たに導入した分野別辞書の自動切り換え機能について述べる。

2. 訳語選択

PalmTree では、以下の3つの方法の組み合わせで訳語選択を行なっている。

- ・ 格フレーム
- ・ パターン中の変数の素性
- ・ 共起処理・文脈処理

2.1 格フレーム

辞書中の各語には統語情報以外に、その単語の意味的な情報を表すために素性が振られている。動詞と形容詞の訳語選択の際には、主語や目的語の要素の素性によって動詞や形容詞の訳語選択を行なっている。

たとえば、動詞 have は30種類以上格フレームを持っている。以下にその中のいくつかを示す。

主語：動物、目的語：食べ物、訳語：食べる
 主語：有意味体、目的語：動物、訳語：飼う
 主語：有意味体、目的語：病気、訳語：かかる

これによって、以下のような訳し分けが可能になっている。

She had a cake.

彼女はケーキを食べました。

I have cats.

私はネコを飼っています。

He had lung cancer.

彼は肺癌にかかりました。

2.2 パターン中の変数の素性

パターン中には変数を記述することができ、その変数部分に素性を指定することで、その素性にマッチする訳語が選択される。

たとえば、

occupation of ~ : 場所 = ~の占領

というパターンによって、

fortification 抽象物 武装化

場所 砦

科学 アルコール添加

という複数の訳語候補から

occupation of fortification → 砦の占領

という訳が得られる。

2.3 共起処理・文脈処理

一文内の共起処理だけでなく、複数文にわたる文脈処理を行なうために、共起辞書と文脈メモリを備えている[5]。複数文にまたがる共起語の訳語選択を行なうと共に、名詞の訳として複合名詞の訳語として選択されたパターンを訳語選択に用いる処理を行なっている。

3. 辞書の自動切り換え

PalmTree ではシステムの基本辞書以外に複数の分野別辞書を使用することが可能で、翻訳の王様V3では以下の6つの専門分野辞書が用意されている。

インターネット一般、アート、ビジネス、スポーツ、政治、エンターテイメント

3.1 辞書自動切り換えの問題点

原文の大意を掴んだうえで適切な分野の辞書を選択することが理想であるが、大意の把握は容易ではなく、また、話題が転換していく文章においてどの部分を辞書を切り換えるべき単位とするかの決定も難しい。そのため、辞書の自動切り換えを行なう現在の一般的な方法は、各辞書を選択するためのキーワードをあらかじめ選んでおき、それが原文中に現れたら、そのキーワードに対応する辞書を選択するというものである。しかし、これには以下のような問題点がある。

- ・ キーワードの選定が容易ではない
- ・ 辞書を更新するたびにキーワードも更新する必要がある
- ・ 原文のどの範囲を辞書の切り換える単位とすべきかの決定が難しい

3.2 自動切り換えの仕組み

前記のような問題点を解決するために、今回我々は以下のような方法を取ることにした。

- ・ 辞書は常にすべて使用する。その優先度を自動的に切り換える
- ・ 複合語と単語は別に考え、複合語に関しては優先度の切り換えは行なわない
- ・ 単語の優先度を上げるトリガーに複合語を用いる
- ・ 複合語がヒットしなくなったら、その分野の話題は終わったものと考え、優先度を下げる

下記にその動きを詳しく述べる。

3.3 自動切り換えの詳細な動き

まず、初期状態ではシステムの辞書は以下のような優先順位となっている。

- 1 ユーザー辞書
- 2 ユーザーが選択した分野別辞書
- 3 システム基本辞書
- 4 ユーザーが選択しなかった分野別辞書

この状況から入力文を翻訳し始める。複合語に関しては辞書の優先度はこのまま固定されることになるが、他の辞書中に同一字面の複合語が含まれることは少ないので、実際には最低優先度の4の分野別辞書中の複合語も、それが入力文に現れれば使用される可能性が高い。

入力文を翻訳する内に4の分野別辞書に含まれる複合語が実際に翻訳の訳語として使用された場合に、その辞書中の単語の優先度をシステム基本辞書より

上にする。たとえば、「Tiger Woods = タイガー・ウッズ」が入力文中に現れて、スポーツ分野辞書中のこの複合語が翻訳に使用された場合には、スポーツ分野辞書中の単語の優先度が上がる。結果として、以下のような優先度となる。

- 1 ユーザー辞書
- 2 ユーザーが選択した分野別辞書
- 3 スポーツ分野辞書の単語
- 4 システム基本辞書
- 5 ユーザーが選択しなかった分野別辞書 (スポーツ分野辞書の複合語も含む)

この結果、birdie という単語が出てきても「小鳥」という訳にはならず、スポーツ分野辞書の「バーディー」という訳語が選択される。複合語と違い分野別辞書の単語はその多くがシステム基本辞書にも含まれているため、このようにシステム基本辞書より優先度を上げないと訳語として使用されない。一度優先度を上げた分野別辞書 (単語のみ) は、話題が変わったら優先度を元にもどす必要がある。そのため、最後の複合語のヒットから10文経ったら優先度を元にもどすことで話題の転換に対応している。

また、自動切り換えによりシステム基本辞書より優先する分野別辞書 (単語のみ) は複数 (現在は3つ) 保持しており、他の分野の人名などが一回だけ現れるような場合にも対処している。

4. まとめ

PalmTree の訳語選択と辞書の自動切り換えについて述べた。格フレーム、パターン中の変数の素性、共起・文脈処理による訳語の柔軟な選択機構を備えている。また、分野別辞書の自動切り換えによって、より適切な訳語選択が可能となっている。

[参考文献]

- [1] Takeda, K., "Pattern-Based Context-Free Grammar for Machine Translation," Proc. of 34th ACL, pp. 144-151, 1996
- [2] Takeda, K., "Pattern-Based Machine Translation," Proc. of 16th Coling, Vol. 2, pp. 1155-1158, 1996
- [3] 渡辺, 武田, "パターンベース翻訳システム: PalmTree", 情報処理学会第55回全国大会, 1997
- [4] Shieber, S., and Shabes, Y., "Synchronous Tree Adjoining Grammars," Proc. Of COLING 90, 1990
- [5] 宮平, 渡辺, 武田, 那須川, "パターンベース翻訳システム PalmTree の文脈処理", 情報処理学会第57回全国大会, 1998