パターンベース翻訳システムPalmTreeの文脈処理

5R－4

宮平知博　渡辺日出雄　武田浩一　那須川哲哉

日本アイ・ビー・エム株式会社

1. はじめに

PalmTree[2,3,4]は同期文法[1]の考え方に基づ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。パターンベース翻訳システムは訳質向上のための翻訳パターンの追加が容易であり、また、システム構成がシンプルであるという利点があるが、さらなる訳質の向上、特に適切な訳語の選択をするためには、文脈に依存する処理が必要となる。本稿では、PalmTreeの拡張として共起辞書と文脈メモリによる文脈を利用した訳語選択処理について述べる。

2. 共起辞書

一つの英単語が複数の異なる意味の訳語を持つ場合が少なくないが、その訳語の選択は必ずしも容易ではない。たとえば、

I went to the bank to get money.

I went to the bank to get fish.

という2つの英文が与えられた場合、bank の訳語として前者では「銀行」、後者では「土手」を選択するためには、本来は文章の意味を適切に理解する必要があるが、現在の機械翻訳システムではまだ文章を理解して翻訳することは不可能である。

そのため、今回我々はPalmTreeに共起辞書を導入した。共起辞書では、同一の文章内に単語Aと単語Bが現れる場合、それぞれにどのような訳語を選択すべきかを指定する。

共起辞書中の各エントリは、以下のような論理フォーマットである。

単語1;語語2;...;単語n

ここで、単語iは、

英語文字列 品詞 日本語訳語

の組であり、日本語訳語は省略可能である。同一文中に単語iと単語jが現れる場合に、それぞれ指定した訳語を選択する。訳語が省略されている場合には、その単語の訳語選択では特に何も特別な処理はしない。

上記の bank の場合のエントリは以下のようになる。

("-" は省略を示している)

bank NOUN 銀行;money NOUN -

bank NOUN 土手;fish NOUN -;swim VERB -

money や fish 訳語は "-" になっているので訳語選択の際には特別な処理はしない。

また、この例では swim も「土手」の共起語としているので、

I went to the bank to swim.

でも「土手」が訳語として選択される。

以下に共起辞書を使用した場合としない場合の翻訳結果をいくつか挙げる。それぞれ、初めが共起辞書を使用した場合の訳、2番目が使用しない場合である。

I went to the bank to get fish.

私は、魚を手に入れるために、土手に行きました。

私は、魚を手に入れるために、銀行に行きました。

Tom belongs to the house and John belongs to the senate.

トムは下院に属します、そして、ジョンは上院に属します。

トムは家に属します、そして、ジョンは上院に属します。

She bathed her foot in the spring.

下彼女は泉に彼女の足を浸しました。

下彼女は春に彼女の足を浸しました。

3. 文脈メモリ

同一文中に共起する語が存在する場合にしか共起辞書を使用することができないと、適用できる範囲が限られてしまう。これを解決するために、複数文にまたがる

Context Processing of Pattern Based Translation System PalmTree
Tomohiro Miyahira, Hideo Watanabe, Kohichi Takeda, Tetsuya Nasukawa
IBM Japan, Ltd.

文脈情報を保持する文脈メモリを用意している。文脈メモリは文脈単語メモリと文脈訳語メモリから成っており、適当なタイミングで初期化する必要がある。

4．1 文脈単語メモリ

共起処理の対象となる単語が原文中に現れた場合、それが実際に共起語として処理されたかどうかにかかわらず、その英語文字列と品詞を文脈単語メモリに保持する。その文中では共起する語が現れなくても、後の文中で現れた場合に共起処理を行なうのが目的であり、そのため、共起辞書に出現しない単語は記憶しておく必要がない。

前にも出した以下の bank の例の場合、

bank NOUN 土手; fish NOUN －; swim VERB －

fish(NOUN)や swim(VERB) が文中に現れると文脈単語メモリに保持され、同一文中の単語だけでなく、文脈単語メモリの語も共起語の候補として使用される。その結果、以下のような複数文にまたがる場合も適切に共起処理がなされる。

I wanted to swim. I went to the bank.

私は泳ぎたかったです。私は土手に行きました。

She bathed her foot in the water. The spring was cold.

彼女は、水で彼女の足を浸しました。泉は寒かったです。

4．2 文脈訳語メモリ

PalmTreeでは、英語の複合語や言い回しなどをパターンとして扱っている。この内、複合名詞はその一部が後続の他の文で現れることがあり、その場合の訳語は複合名詞の訳語として選択されたものを考慮する方が適切な訳になる[5]。

たとえば、civil trial が「民事裁判」と訳された場合、以降の文脈では trial という単語は「裁判」と訳すほうがよさそうだと推測できる。このため、複合名詞の訳語を記憶しておき、その後の単語の訳を以前の複合名詞の訳から推定するのがこの文脈訳語メモリの機能である。

文脈訳語メモリには、最終的に解として使われた複合名詞パターンが登録される。複合名詞パターンが

EW1 ... EWn -> NP NP <- JW1 ... JWm

である場合、それぞれの英単語 EWi に関して文脈訳語メモリに以下のレコードを追加する。

EWi = JW1 ... JWm

ある単語の訳語を選択する際には、その単語の訳語候補を文脈訳語メモリ中の訳語とのマッチング度によってソートする。マッチング度の計算は現在のところ単純に単語の訳語文字列が複合語の訳語文字列に含まれるかどうかのチェックで行なっている。

この文脈訳語メモリによって以下のような翻訳が可能となる。

It was a civil trial about an accident. The trial was scheduled at Monday morning.

それは、事故についての民事裁判でした。裁判は、月曜日午前に予定されました。

The defense secretary stood up. The secretary started a speech.

国防長官は立ち上がりました。長官はスピーチを開始しました。

5. 処理の順序

以上述べたように3つの文脈を利用した訳語選択処理をPalmTreeでは行なっている。

一文内の共起処理

文脈単語メモリによる複数文にまたがる共起処理

文脈訳語メモリによる複合語訳語からの訳語選択

距離の近い語の方が密接に関係していると思われるので、現在のところ、この順に処理を行ない、ある処理が成功したら以下の処理は行なわないようにしている。

6. まとめ

PalmTreeにおける文脈を利用した訳語選択処理について述べた。共起辞書と文脈メモリによって、複数文にまたがる文脈情報の利用を実現している。現在はまだ実験段階であり小さな共起辞書しか構築できていないが、今後共起辞書の拡充を行なう予定である。

また、文脈訳語メモリのマッチング度の計算方法も現在の単純な文字列の包含関係ではなく、一致する文字数の多少など他の計算方法も検討していきたい。

[参考文献]

[1] Shieber, S., and Shabes, Y., "Synchronous Tree Adjoining Grammars," Proc. Of COLING 90, 1990

[2] Taked, K., "Pattern-Based Context-Free Grammar for Machine Translation," Proc. of 34th ACL, pp. 144-151, 1996

[3] Takeda, K., "Pattern-Based Machine Translation," Proc. of 16th Coling, Vol. 2, pp. 1155-1158, 1996

[4] 渡辺日出雄、武田浩一, "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

[5] 那須川哲哉, "文脈辞書を用いた頑健な多義性解消", 言語処理学会第4回年次大会, 1998