対訳文書を用いた訳語の学習

7Q－6

中山圭介 木下聡 平川秀樹
(株) 東芝 研究開発センター

1　はじめに

最近対訳コーパスから翻訳に関する知識を抽出する方法が注目されている。対訳コーパスから翻訳テンプレートを抽出する方法 [1]、翻訳規則を抽出する方法 [2]、動詞の格フレームを獲得する方法 [3] などが提案されている。

本稿では日本語と英語の対訳例文を照合し、その結果から翻訳システムに必要な訳語学習データを抽出する。日本語と英語の照合に際しては、日本語と英語という構造が大きく異なる言語間の照合を柔軟に行なうために、2言語間の照合を、単語の訳語情報を元にしてボトムアップに行なうとともに統語解析結果に対してではなく依存構造に対して行なう。

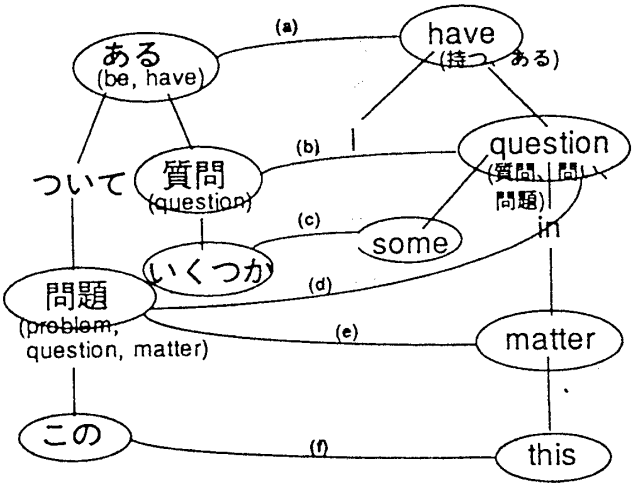
2　訳語学習の方式

本稿で使用する対訳文書はあらかじめ文対応が付けられているものとする。対訳文書の英語側を目標訳とし、日本語側を機械翻訳システムで翻訳した結果と比較する。そして目標訳と機械翻訳出力で訳語が異なっているものを取りだし、その語について日本語の語と目標訳の語の対を訳語学習データとして出力する。

次に1つ1つの対訳文についての訳語学習処理の流れを説明する。まず対訳文の日本語側を構文・意味解析し図1の左側のような依存構造を作成する。同様に対訳文の英語側 (目標訳) を構文・意味解析し図1の右側のような依存構造を作成する。構文・意味解析には機械翻訳システムの解析ルーチンを使用する。統語解析結果ではなく依存構造を用いることによって日本語と英語という構造が大きく異なる言語間の対応が取りやすくなる。

2つの依存構造ができあがると、次にそれらの間でどの単語とどの単語が対応するか、あるいはどの熟語とどの熟語が対応するかの候補を求める。これには機械翻訳システムの辞書の訳語情報を用いる。対応する語の候補が複数個ある場合には後述する方法によって対応関係を一意に決定する。単語の訳語情報を元にボトムアップに対応関係を決定することで、2つの言語の依存構造の構造上の違いに対してロバスト性を持つことができる。

日本語と英語の間の単語や熟語の対応関係を定めた後、英語側の単語、熟語のうち機械翻訳システムが訳出したものと異なるものについて、その日本語側の単語、熟語と英語側の単語、熟語のペアを訳語学習データとして出力する。



日本語: この問題についていくつかの質問があります。
英語: I have some questions in this matter.

図1: 両言語の依存構造

3　対訳例文の対応付け

訳語学習の前提として、まず2つの言語の依存構造のノードの間の対応関係を定める。

3.1　対応候補の作成

概念依存構造間のノードの対応付けには翻訳システムの辞書の訳語情報を基本的情報として用いる。日本語側のノード、英語側のノードの両方から対応する単語、熟語の候補を求めると図1の (a) ～ (f) のようになる。以下対応する単語、熟語の候補を対応候補と呼ぶ。対応候補を作成する段階では翻訳システムの辞書の訳語情報を参照して対応する可能性のある全ての単語、熟語が組み合わされる。

3.2　対応セット

図1の (a) ～ (f) に示す対応候補の内、(b) と (d) は "question" が日本語の違ったノードと対応するという

Acquisition of Translation Equivalents from Bilingual Text
Keisuke NAKAYAMA, Satoshi KINOSHITA and Hideki HIRAKAWA
Research and Development Center, Toshiba Corp.
keisuke@isl.rdc.toshiba.co.jp

矛盾した情報を持っており、両立しないものである。同様に (d) と (e) も両立しない。両言語の依存構造のノード間の対応候補の中で矛盾しない対応関係のみを集めた集合を対応セットと呼ぶことにする。図 1 の例では対応セットは 2 つ求まる。(図 2)

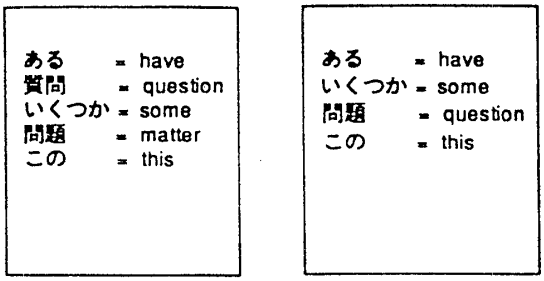


図 2: 図 1 の例に対する 2 つの対応セット

3.3 最適対応セットの抽出

全ての対応セットの中で、どの対応セットが最も確からしいかを決定するために、それぞれの対応セットに対してスコア付けを行なう。対応セットのスコアは、その対応セットに含まれる対応候補のスコアを合計することにより得られる。対応候補のスコアを計算する際には翻訳システムの辞書の訳語情報だけでなく周囲のノードとの整合性も考慮される。対応セットのスコア付けの結果、最もスコアの高かったものを最適対応セットとする。そして最適対応セットに含まれる対応候補を 2 つの依存構造間のノードの対応関係を表すものとする。

4 訳語学習データの抽出

最適対応セットに含まれる対応候補は日本語文と英語文の間の対応する語の対であると見なせる。対応付けされた単語、熟語の内訳語学習の対象となるものは、翻訳システムが訳出した英語の単語と対応付けされた英語の単語が異なるものである。今回は名詞、合成名詞を対象として、対応付けの結果と翻訳システムの訳が異なるものについて訳語学習データを抽出した。

5 実験

同一分野の文対応が付けられた対訳例文 248 文を対象に、以下の手順で訳語学習データの抽出の実験を行なった。対訳文の日本語側 1 文に含まれる自立語の数は平均 10.2 個であった。

1. 対訳文の日本語側を機械翻訳システムで翻訳する。
2. 翻訳システムの翻訳結果と対訳文の英語側を人間が比較し、両者の間で訳語が異なっている名詞と合成名詞に注目して、学習目標データ (正解データ) を作成する。
3. 訳語学習プログラムが出力した結果と学習目標データを比較する。

実験結果は以下の表のようになった。

用意した正解データ	114個
学習した名詞、合成名詞	83個
正解データと一致	52個(63%)
正解データと不一致	31個(37%)
準正解	22個(26%)
学習の誤り	9個(11%)

この結果は全自動で学習するには精度が不足しているが、人間が介在した学習には利用可能な精度であると考えられる。学習の誤りの多くは、日本語と英語の間の表現の違いにより誤った対応付けが行なわれてしまったことが原因であった。また準正解は、正解データが「指サック = fingerstall」と日本語側が合成名詞になっているのに「サック = fingerstall」と学習した等、誤った学習とは言えないが正解データとは一致しなかったものである。

6 結論

対訳例文から単語や熟語の対応付けを元にして翻訳システムの訳語学習データを抽出する方法を提案した。今後は対応付けアルゴリズムの拡張により、学習の精度を向上させいく予定である。

参考文献

[1] H.Kaji, Y.Kida and Y.Morimoto : Learning Translation Templates from Bilingual Text, Proc. of COLING-92,(1992),pp.672-678

[2] H.Watanabe : A Method for Extracting Translation Patterns from Translation Examples, Proc. of TMI-93,(1993),pp.292-301

[3] T.Utsuro, Y.Matsumoto and M.Nagao : Lexical Knowledge Aquisition from Bilingual Corpora, Proc. of COLING-92,(1992),pp.581-587