形態素の意味情報を用いた訳語推定手法

4B－3

笹岡久行　荒木健治　桃内佳雄
北海学園大学工学部

1. はじめに

インターネットの普及などにより我々には母国語以外の言語を扱う機会が増えてきている。その影響により機械翻訳の需要が高まりつつあり、また機械翻訳に関する研究も盛んになっている [1]。しかし、実際に機械翻訳システムを利用するとシステムの辞書に未登録である単語が英文に出現することがある。機械翻訳手法の研究におけるこの辞書未登録語の処理はこれまで大きな問題となっていた。そこでこの問題の解決を目指し、我々は帰納的学習による未登録語の訳語推定手法 [2] を提案している。この手法では訳語推定に利用する単位を獲得するために英単語とその訳語に対する帰納的学習が必要であった。この学習に大量の英単語と訳語の組みおよび膨大な処理時間を必要とする問題が存在した。この問題を解決するために本手法では、訳語推定に必要となる知識は人間が作成し、システムはその知識が与えられた状態で未登録語の訳語推定を行なう。つまり、本研究では知識を持った人間の問題解決能力の実現とその工学的応用を目的としている。本手法では、訳語推定を単語の語基および接辞を基にして行なう。本稿では、単語の語基および接辞を利用した英日の訳語推定手法および本手法に対する評価実験の結果と考察について述べる。

2. 未登録語

辞書 [4] に見出し語として登録されていない単語を未登録語とした。文献 [3] の 1478 種類の英単語の中から 239 種類の未登録語を抽出した。これらを分類すると以下のようになった。

a. 固有名詞：Brian など
b. 派生語：internally など
c. 複合語：general-purpose など
d. その他：ANSI(略語),1983(数値) など

本手法は単語の語基あるいは接辞を基に訳語推定を行なうので、b と c に分類できる単語に対して有効であると我々は考えている。239 種類の単語の中で b あるいは c に分類できる単語は 54 種類であった。また a や d の単語については別の手法により対応する予定である。

3. 処理手法

本手法を基に作成したシステムの概要を図 1 に示す。

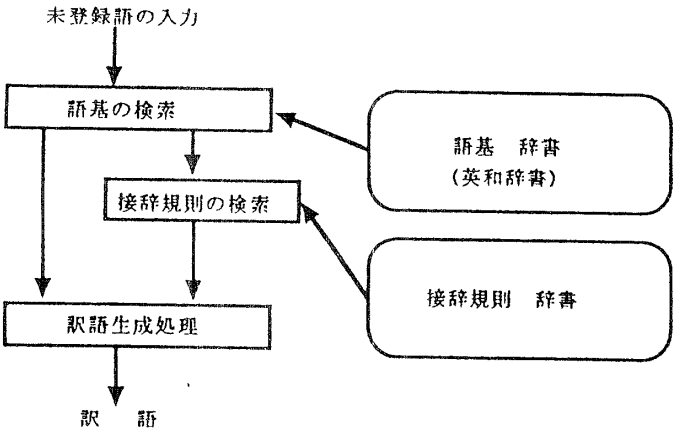

図 1：システムの概要

単語の語基とその訳語の組みは英和辞書 [4] の見出し語とその訳語の組みとした。また、文献 [5] における単語の分類を参考にし、それらの単語と訳語の組みから人手により抽出した接辞と訳語の組みを接辞規則とした。

本手法では、最初に推定対象単語の語基とその訳語を検索する。次に、その語基以外の部分に適応可能な接辞規則あるいは語基と訳語の組みを検索する。語基の訳語あるいは接辞規則の日本語側を組み合わせて未登録語の訳語推定を行なう。

図 2 において派生語 vocalist の訳語推定を例に本手法の処理過程を説明する。まず、単語の語基とその語基の訳語を辞書中のから検索する。これは、未登録語の文字列と辞書の見出し語との左側からの最長一致法により検索する。次に、語基以外の部分において適応可能な接辞規則を検索する。ここで @0 は変数を表し、文字列が代入される。また、語基と接辞規則の英語側の組み合わせにより未登録語の綴りを構成可能な場合は、接辞規則の日本語側の変数部分に語基の訳語を代入して推

Prediction Method of Word for Translation Using Information of Morpheme
Hisayuki Sasaoka, Kenji Araki, Yoshio Momouchi
Fac. of Engineering, Hokkai-Gakuen University

1. 語基の検索
   語基：vocal「声楽」
2. 接辞規則の検索
   接辞規則：@0 ist「@0 家」
3. 訳語の組み合わせ
   推定結果：「声楽家」

図２：派生語 vocalist の処理例

1. 語基の検索
   語基：small「小さい」
2. １の語基以外の部分での語基の検索
   語基：scale「規模」
3. 訳語の組み合わせ
   推定結果：「小さい規模」

図３：複合語 small-scale の処理例

英単語：typeless
語基：type「型」
接辞規則：@0 less「@0 のない」
推定結果：「型のない」、正しい訳語：「型のない」

図４：有効な推定結果の例

英単語：end-of-file
語基：end「終り」,of「の」,file「ファイル」
推定結果：「終りのファイル」、
正しい訳語：「ファイルの終り」

図５：誤った推定結果の例

定結果を得る。

次に、図３において複合語 small-scale の訳語推定を例に本手法の処理過程を説明する。派生語の場合と同様に、左側からの最長一致法で語基を検索する。次に、最初に検索された以外の部分に適応可能な語基を検索する。語基の英語側の組み合わせにより未登録語が構成可能ならば、英語の語基の語順通りに語基の訳語側を並べて訳語を完成させる。

4. 評価実験

4.1 実験方法

上述した 54 種類の未登録語に対して本手法による訳語推定を行なった。本手法では語基あるいは接辞規則に複数の訳語が存在する場合には全ての可能な組合せを推定結果とした。また、推定結果の中に正しい訳語が含まれている場合を有効な推定結果とした。

4.2 実験結果

54 個中の 38 個 (70.4%) が有効な結果であった。図４に正しい推定結果の例、図５に誤った推定結果の例をそれぞれ示す。

4.3 考察

評価実験の誤った推定結果に対する考察から本手法だけでは正しく訳語を推定できない複合語が存在することが明らかになった。本手法による複合語の訳語推定では、英語側の語基の語順通りにその訳語を組み合わせている。このために英語の語基とその訳語の語順に対応関係がない複合語では正しく訳語を推定できない。そのために、シソーラスを利用した意味分類に基づく語順規則を作り、それを利用した訳語推定手法を検討する予定である。

また、本手法では語基あるいは接辞規則に複数の訳語が存在する場合は可能な全ての組合せを推定結果として生成しているため、訳語選択の負担をユーザーにかけるという問題が存在する。適切な訳語の選択には、対象英文の文脈情報が必要である。そこで、文脈情報を利用した手法も検討する予定である。

5. おわりに

対象とした未登録語の 70.4%について有効な推定結果を得たことから、本手法の有効性は確認された。しかし、これは全未登録単語 1478 語の中の 15.9%に対し有効な推定を行なったに過ぎない。そこで、他の種類の未登録語に対処する手法を検討する予定である。また、考察から明らかになった問題を解決する手法についても研究を進める予定である。

参考文献

[1] 長尾真：自然言語処理、岩波書店、1996(東京).

[2] 笹岡久行、荒木健治、桃内佳雄：英文からの帰納的学習による訳語推定手法へのヒューリスティクスの適応、平成 8 年度電気関係学会北海道支部連合大会講演論文集、No.294.

[3] Brian,W. K. and Dennis,M.R.: THE C PROGRAMMING LANGUAGE, PRENTICE HALL, 1988,(New Jersey).

[4] 久保正治：英和・和英電索辞典 gene, 技術評論社、1995(東京).

[5] 前田健三：強くなる英単語、有精堂、1994(東京).