5J－10

帰納的学習を用いた文脈に依存した訳語推定手法*

笹岡久行　荒木健治　桃内佳雄　　栃内香次†
北海学園大学工学部‡　　北海道大学工学部§

1. はじめに

従来の機械翻訳システムにおいて、辞書に未登録である単語 (辞書未登録語) を翻訳することはできない。我々は、この問題の解決を目指して研究 [1] を進めている。従来の多くのシステム [2] のように辞書の登録語数を増やすだけではこの問題を解決することは難しいと考えている。なぜなら、存在する全ての単語を辞書に登録することは非常に困難であり、また造語等の新出語により辞書未登録語は無くならないからである。

既に、我々は辞書未登録語の訳語を推定する手法とその訳語推定に利用する単位の抽出手法を提案した [1]。しかし、機械翻訳システムでは、翻訳対象の文脈に合った訳語が一つだけ必要とされる。しかし、先に提案した手法ではそのような訳語を選択できるまでには至っていない。そこで、本稿では先の訳語推定手法に対する評価実験から得られた結果に対する考察について述べ、さらに我々の訳語推定手法を利用し文脈に適した訳語を推定する手法を実現するために有効であると考えられる情報について考察する。

2. 訳語推定に利用する単位

我々は、帰納的学習により２つの異なる文字列から、字面における共通な文字列と異なる文字列の抽出により得られた文字列を一つの単位と見なし、単語片 (a Piece of Word:PW) と呼んでいる。さらに、共通部分として抽出された文字列に対して、変数を付与する。この変数は抽出された文字列において、差異部分である文字列が存在した位置に付与される。また、この位置が訳語推定の際に他の単語片を組み合わせる位置となる。さらに、２つの異なる言語の単語片の組みを単語片対 (a Pair of Pieces of Word:PPW) と呼んでいる。図 1に、「meaningful, 意味深長な」と「meaningless, 意味のない」の２組みから抽出した単語片対の例を示す。

3. 処理概要

図 2に我々の手法を基にして作成したシステムの概略を示す。まず、訳語を推定する辞書未登録語を入力する。システムが、訳語推定部で単語の訳語を推定する。推定結果が、正しいものであった場合にはその結果を利用して次の学習部へ進み、誤ったものであった場合には校正

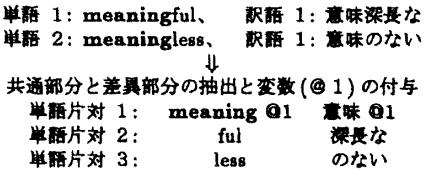

図 1: 単語片対の例

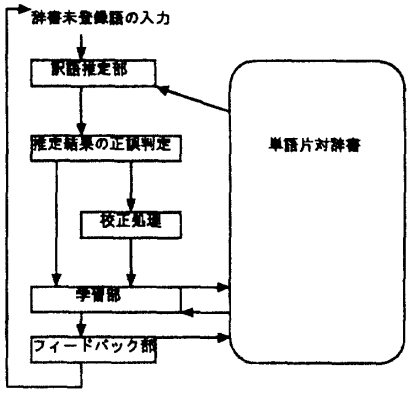

図 2: 実験システムの概略

処理を行い正しい訳語をシステムに与えた後に学習部での処理を行う。学習部では、入力された単語とその訳語の組みとそれ以前に獲得された単語片対の間から新たな単語片対の獲得を行う。さらに、推定された推定結果の正誤判定結果に応じて、フィードバック部において推定処理に利用した単語片対の持つ訳語推定処理の際に確実性を表す数値に対して処理を行う。このフィードバック処理により、それ以降の訳語推定処理において、訳語推定能力を改善する。

4. 評価実験

4. 1 実験方法

先に提案した訳語推定手法を基に作成した実験システムに対して評価実験を行った。本実験では、電子化された英和辞典 "gene" [3] に見出し語として登録されていない語を辞書未登録語とした。また、英文コーパス "Susanne Corpus" [4] には４つのカテゴリー「１. "Press Reportage", ２. "Belles Letters, Biography, Memories", ３. "Learned Writing", ４. "Adventure and Western Fiction"」があり、各々異なり語数 4,425 語、4,986 語、

*"Prediction Method of Adaptable Word for Translation Using Inductive Learning

†Hisayuki Sasaoka, Kenji Araki, Yoshio Momouchi, Koji Tochinai
‡Fac. of Engineering, Hokkai-Gakuen University
§Fac. of Engineering, Hokaido University

3,851 語、3,896 語が存在する。その中の辞書未登録語である派生語と複合語は、それぞれ 179 語、305 語、369 語、202 語であり、それらを実験データとした。評価実験は、各カテゴリー毎に行った。

本実験では、上述した辞書 "gene" の中から英単語と訳語の組み 102,156 組みを抽出した。さらに、文献 [5] を参照して、同じ接辞を持つ英単語とその訳語から接辞とその訳語からなる 351 組みを抽出した。合計 102,507 組みを初期状態の単語片対辞書に与えて実験を開始した。

本実験において、推定結果は英和辞典 [6] を利用して、分類 a「辞典 [6] の訳語と一致するあるいは一単語のみが同じ意味の単語と置き換わったもの」、分類 b「辞典 [6] の訳語と語幹の部分までは一致しているもの」、分類 c：「分類 1 と 2 に属さないもの」と分類した。

また、本実験における有効推定率と無効推定率を以下のように定める。

$$\text{有効推定率}(\%) = \frac{(\text{分類 a}) + (\text{分類 b})}{(\text{実験データ数})} \times 100.0$$

$$\text{無効推定率}(\%) = \frac{(\text{分類 c})}{(\text{実験データ数})} \times 100.0$$

4．2 実験結果

表 1: 実験結果

カテゴリー	有効な推定			無効な推定		合計
	a	b	推定率	c	推定率	
1	26	87	63.1	66	36.9	179
2	41	143	63.0	108	37.0	292
3	50	180	62.3	139	37.7	369
4	20	114	66.3	68	33.7	202

表 1は、この実験における各分類の数と有効推定率および無効推定率を示す。この表から、各カテゴリーで有効推定率が約 60%であることが分かる。従来の機械翻訳手法に基づくシステムでは、辞書未登録語の訳語を知ることは全く不可能であった。しかし、本手法を利用することにより、辞書未登録語の派生語と複合語の約 60%の訳語あるいは訳語の語幹の部分を推定可能であり、本手法は有効であると考えられる。

4．3 考察

表 1での分類 a を、構成している単語片対によって、分類 a－1：「初期単語片対辞書中に存在した単語片対のみによって構成されているもの」、分類 a－2：「初期単語片対辞書中に存在した単語片対と学習処理により獲得された単語片対により構成されたもの」分類 a－3：「学習処理により獲得された単語片対のみによって構成されたもの」に分類した。

また、上の分類では推定された結果が文脈に合っているのかを判定していない。そこで辞書 [6] に記載されている訳語の中で、それぞれの辞書未登録語が存在していた英文の文脈に最も合うものを人手により選び出した。これに一致するものと一致しないものの数を表 2 に示す。

表 2: 分類 a における文脈に一致した推定結果の数

カテゴリー	a－1		a－2		a－3		合計
	一致	不一致	一致	不一致	一致	不一致	
1	15	5	3	2	1	0	26
2	16	8	6	5	5	1	41
3	15	6	11	4	12	1	50
4	9	5	2	2	2	0	20

この表 2より、分類 a では学習処理により獲得された単語片対のみにより構成された推定結果が文脈に合った訳語となる割合が高いことがわかる。このことから、訳語推定処理の際に、訳語推定に利用する単語片対が初期状態から与えられた単語片対であるのかあるいは学習処理により獲得された単語片対であるのかを表す情報を利用することは、複数ある推定結果の中から文脈に合った訳語を選択するために有効な情報の一つであると考えられる。

5．おわりに

本稿では、単語片対を利用した訳語推定手法を基に作成したシステムに対して行った評価実験の結果とその考察について述べた。さらに、この実験結果に対する考察から、我々の手法により辞書未登録語の訳語推定をする際、推定に利用した単語片対に関する情報が文脈に適した訳語を複数の推定結果の中から選択するために有効な情報の一つであると考えられる。今後は、訳語を選択する手法の詳細について検討し、それが有効であることを確認するための評価実験を行う予定である。

参考文献

[1] 笹岡久行，荒木健治，桃内佳雄，栃内香次，"語基および接辞の接続情報を用いた辞書未登録語の訳語推定手法，" 信学技報，NLC96-64，Mar. 1997.

[2] 長尾真 編，"自然言語処理，" 岩波書店，東京，1996.

[3] 久保正治，"英和・和英電策辞典 *gene*，"，技術評論社，東京，1995.

[4] Geoffery, S, "ENGLISH FOR THE COMPUTER, " Oxford University Press, Oxford, 1995.

[5] 前田健三，"強くなる英単語，" 有精堂，東京，1994.

[6] 小稲義男 他，"新英和大辞典，" 研究社，東京，1980.