240　帰納的学習を用いた訳語推定手法における形態素情報の利用

笹岡久行† 荒木健治† 桃内佳雄† 栃内香次‡
† 北学園大工　　‡ 北大工

1　はじめに

我々は、機械翻訳システムにおける辞書未登録語問題の解決を目指し、帰納的学習を用いた訳語推定手法 [1] を提案し、その有効性を確認した。我々は、訳語推定処理に利用する単位である単語片対を定義している。この手法では、単語と訳語の組を抽出元として、字面情報を基に帰納的学習 [2] を用いて単語片対を獲得する。しかし、従来手法では単語片対の過分割あるいは抽出の見逃しなどの問題があり、十分な量の単語片対を獲得することは困難であった。そこで、このような問題の解決を目指して、従来の字面情報に加え、形態素分割位置や品詞のような形態素解析結果を学習に利用した単語片対の抽出手法を本稿では提案する。さらに、獲得された単語片対を用いて行われた訳語推定実験の結果を報告する。

2　基本的な考え方

2.1　従来手法の問題

従来手法では抽出元の文字列の共通部分と差異部分を抽出する。そして、原言語と目的言語の各々の抽出結果を単語片対として獲得する。しかし、字面情報のみに基づいた単語片対の抽出には単語片対の過分割などの問題がある。例えば、単語片対の抽出元が「electrical system engineering, 電気システム工学講座」と「electrical field, 電場」の場合には、「system engineering, 気システム工学講座」等が抽出される。ここで、単語片対として「electronic ◎1, 電子 ◎1」[1] が獲得されており、「electronic system engineering」を処理する例について考える。正しい訳語は「電子システム工学講座」であるが、従来手法では単語片対を組み合わせると、「電子 気システム工学講座」という誤った訳語を生成する。

2.2　形態素解析結果を利用した単語片対の獲得

従来手法の問題を解決するために、形態素の区切り位置情報や品詞情報等の解析的な知識に基づいた単語片対の獲得手法の提案を行う。上述したように字面情報に基づく抽出では、字面の共通な並び及び異なる並びを抽出する。この抽出を原言語と目的言語において各々行い、単語片対として獲得する。次に、形態素区切り位置情報に基づく抽出では、形態素に分割された抽出元において、形態素毎に共通な並びあるいは異なる並びを抽出する。この抽出を原言語と目的言語において各々行い、単語片対として獲得する。そして、品詞情報に基づく抽出では、抽出元に形態素解析結果から品詞を付与する。品詞が付与された列において、共通な品詞の並びあるいは異なる並びを抽出する。この抽出を原言語と目的言語において各々行い、単語片対として獲得する。

字面情報に基づく抽出処理では、既存の形態素に影響されることなく単語片対を抽出する。つまり、既存の単位では未登録となっているものを単語片対として抽出することが可能である。一方、形態素区切り位置情報および品詞情報に基づく抽出処理では、形態素毎に単語片対を抽出する。これにより、単語片対の過剰な分割を防ぐことができる。

さらに、各抽出結果を基にして、より確からしい単語片対とそうではないものを区別し、優先順位を決定する。優先順位の付け方は、「各抽出処理結果が同一の抽出元の組から同一の抽出結果を得た場合にはより確からしい単位と見なす。」というものである。これは単一の情報のみによる学習結果よりも、数多くの情報による学習結果が一致した学習結果の方がより確からしいというヒューリスティクスに基づいている。

Fig. 1にその例を示す。図中の "‖" は形態素の区切り位置を示す。この抽出例では、各抽出結果が一致した。本手法では、このように各情報に基づく抽出結果が一致するものを一致しないものよりもより確からしいものと判断する。

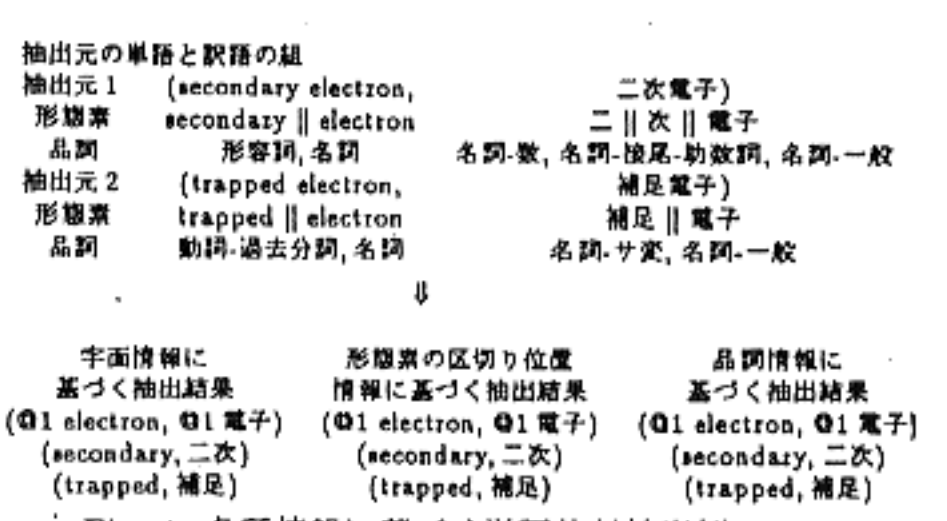

Fig. 1　各種情報に基づく単語片対抽出例

3　処理過程

Fig. 2に、実験システムの概要を示す。システムは、推定対象単語が入力されると、既に獲得している単語片対のみを利用して訳語推定を試みる。もし、獲得さ

[1] ただし、単語片対において、「◎1」は変数を表す。変数の位置に他の文字列を代入することにより新たな文字列の生成を行う。

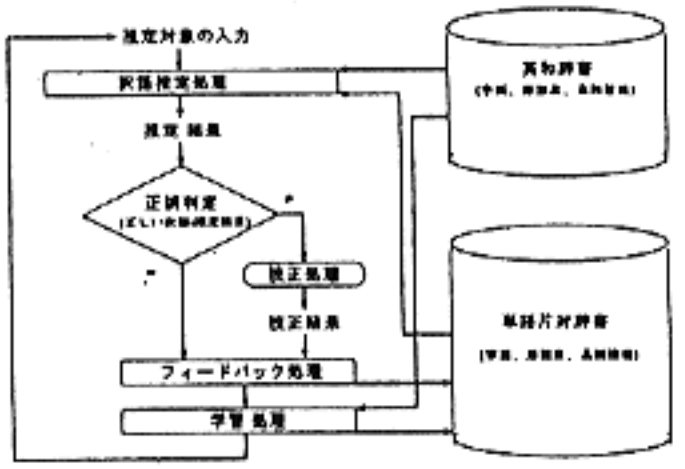

Fig. 2 実験システム

れた単語片対を用いて推定が完了しない場合には、さらに英和辞書からも単語片対を抽出する。そして、訳語推定処理において複数の推定結果が生成された場合、各推定結果を構成している単語片対が既出の単語と訳語の組に包含される回数、過去の利用状況を示す数値である出現度数、正推定度数および誤推定度数を参照し優先順位を決定する。その後、推定結果の正誤判定を行い、推定結果が誤ったものであった場合だけ、推定結果に校正処理を施す。次のフィードバック処理では、その正誤判定結果に応じて、推定結果を構成する各単位の出現度数と正推定度数あるいは誤推定度数を操作する。そして、学習処理では新たな単語片対の抽出を行う。学習処理では、上述したように各種の情報に基づいて単語片対の獲得を行う。

4 予備実験

4.1 実験方法

本手法を用いて行った予備実験の結果について報告する。実験データとしては、大学の講座名や専門分野名等の英語と日本語の組 50 組を用いた。また、形態素解析ツールは英語では「Tagger」 [3] を、日本語では「茶筅」 [4] を、初期状態の文法設定のまま利用した。また、システムが単語片対抽出のために利用する英和辞書は「gene」を形態素解析結果を付与した上で利用した。推定結果は，推定が完了したものと未了であるものに分類した。さらに、推定が完了したものの中で、「優先順位 10 位以内に文脈に適合した訳語と一致する推定結果が存在するもの」を正推定、「推定を完了したが，優先順位 10 位以内に文脈に適合した訳語と一致する推定結果が存在しないもの」を誤推定と分類した.

4.2 実験結果

Table 1に、本手法における推定完了と推定未了のものの数と割合を示し、さらに、比較のために従来手法 [1] における結果も示した。また、Table 2では、本手法と従来手法における正推定と誤推定の数と割合を示した。

この結果において、本手法の方が従来手法より高い精度で訳語推定が行われていることが確認される。これにより、本手法が従来手法に比べて有効に働くと考えられる。

Table 1 本手法と従来手法の推定完了と推定未了

	推定完了数 ([%])	推定未了数 ([%])	データ数 ([%])
本手法	26 (52.0)	24 (48.0)	50 (100.0)
従来手法	22 (44.0)	28 (56.0)	50 (100.0)

Table 2 本手法と従来手法の正推定と誤推定

	正推定数 ([%])	誤推定数 ([%])	推定完了数 ([%])
本手法	14(53.8)	12(46.2)	26 (100.0)
従来手法	6(27.3)	16(72.7)	22 (100.0)

5 おわりに

本稿では、帰納的学習を用いた訳語推定における単語片対獲得の問題を解決するために、形態素解析結果を利用した単語片対獲得手法を提案した。さらに少量の実験データを用いて行った予備実験の結果を報告した。この実験での本手法と従来手法の訳語推定結果の比較から本手法が有効に働くことが確認された。

今後は、本手法を基にしたシステムを作成し、大量のデータを用いた評価実験を行い、本手法の有効性を確認する予定である。

謝辞

本研究の一部は科学研究費 (No. 09878070, No.10680367) および北海学園大学ハイテク・リサーチ・センター研究費による補助のもとに行われた。

参考文献

[1] 笹岡久行, 荒木健治, 桃内佳雄, 栃内香次, "帰納的学習を用いた訳語推定手法における単語片対の抽出元の選択数に関する性能評価", 言処学会, 第 5 回年次大会, pp357–360,March 1999.

[2] 荒木健治, 高橋祐治, 桃内佳雄, 栃内香次, "帰納的学習を用いたべた書き文のかな漢字変換," 信学論 (D-II), vol.J79-D-II, No.3, pp.391 – 402, March 1996.

[3] E. Brill. "Some advances in rule-based part of speech tagging." Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), Seattle, Wa., 1994.

[4] 松本裕治, 北村啓, 山下達雄, 今一修, 今村友明. "日本語形態素解析システム『茶筅』version2.0 使用説明書." Technical Report NAIST-IS-TR99008, 奈良先端科学技術大学院大学,1999.

[5] 久保正治, 英和・和英電索辞典 gene, 技術評論社,1995.