2K−2　帰納的学習を用いた訳語推定手法における同一分野から抽出した単語片対の利用の有効性

笹岡久行[†]　荒木健治[†]　桃内佳雄[‡]　栃内香次[‡]
[†]北海学園大学大学院工学研究科　[‡]北海道大学大学院工学研究科

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処理の問題の解決を目指し、研究を進めている。辞書未登録語が翻訳対象に出現した場合、その単語の翻訳は困難であり、大きな問題となる。そこで我々は、帰納的学習を用いた訳語推定手法を提案し、その有効性を確認した [1]。提案手法では、帰納的学習を用いて、単語と訳語の組の中から、訳語推定に利用する単位を獲得する。本研究ではこの単位を単語片対と呼ぶ。提案手法では、字面に基づく単語片対の抽出 [1] に加え、形態素解析処理の結果に基づいて単語片対を獲得する。そして、それらを用いて訳語推定を行う手法を提案し、有効性を確認した [2]。本稿では、この手法において、同一分野に出現する単語と訳語の組の中から獲得した単語片対の利用の有効性について述べる。そのために、獲得した単語片対を利用するシステムによる訳語推定結果と獲得した単語片対を利用しないシステムによる訳語推定結果とを比較した。

2. 基本的な考え方

帰納的学習を用いた訳語推定手法の基本的な考え方は、単語と訳語の中に存在する単位を組み合わせて、翻訳を行うというものである。そして、この単位を単語片対と呼び、帰納的学習 [3] を用いて単語と訳語の組から獲得する。単語と訳語の字面情報、区切り位置の情報および品詞情報の3つの情報に基づいて、共通部分と差異部分をそれぞれ抽出する。そして、抽出した単語片対を組み合わせることにより、新たな単語と訳語の組を生成する。この単語片対とは、単語と訳語を構成する単位となりうるものであり、単語と訳語の組を抽出元とし、その抽出元から帰納的学習を用いて獲得される。

図1は字面情報に基づく単語片対の抽出例を示す。この例では、抽出元1「electrochemistry, 電気化学」と抽出元2「electrolytic, 電気分解の」の間から、字面の共通部分と差異部分を抽出する。ここで、'@1'は変数を表す。この変数は、共通部分として抽出された単語片対の原言語および目的言語の各々に対して付与され、そ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われる。

抽出元 1　(electrochemistry,　電気化学)
抽出元 2　(electrolytic,　電気分解の)

↓

PPW1　(electro@1,　電気@1)
PPW2　(chemistry,　化学)
PPW3　(lytic,　分解の)

図 1: 単語片対の抽出例

3. 実験システム

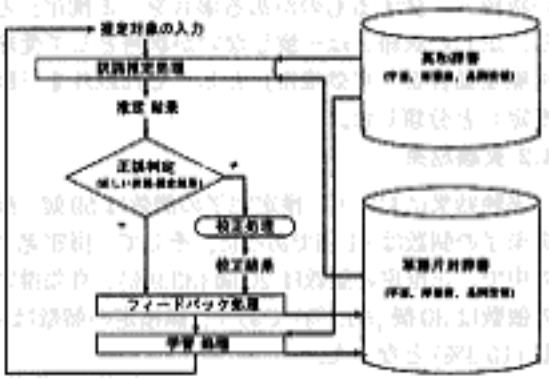

図 2: 実験システム

図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、もし、獲得された単語片対を用いて推定が完了しない場合には、英和辞書の単語と訳語の組の間から新たな単語片対を抽出し、それらを用いて訳語推定を試みる。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、誤った推定結果に対して人手により校正処理を施す。次

にフィードバック処理を行なう。この処理により、これ以降のシステムが持つ訳語推定能力を向上させる。最後に、学習処理で新たな単語片対の抽出を行う。

4. 評価実験

4.1 実験方法

本実験では、大学の講座名や専門分野名等の英語の単語とその日本語の訳語の組 100 組を実験データとした。単語片対抽出に利用する形態素解析結果を得るために、形態素解析システムとして、英語では「Brill Tagger」 [4]、日本語では「茶筌」 [5] を利用した。そして、実験システムにおいて利用する英和辞書としては、「gene」 [6] を利用した。そして、上述した実験システムを用いて、実験データである英語の単語に対する日本語への訳語推定を繰り返した。

実験結果の評価方法としては、推定結果を、推定が完了した「推定完了」と推定が完了しない「推定未了」に分類した。さらに、「推定完了」を、推定結果の優先順位 10 位以内に、用意された正しい訳語と一致するものがある場合を「正推定」とし、正しい訳語とは一致しないが訳語として受理可能な場合を「有効推定」とし、それ以外を「誤推定」と分類した。

4.2 実験結果

実験結果において、推定完了の個数は 59 個、推定未了の個数は 41 個であった。そして、推定完了の中で、正推定の個数は 20 個 (33.9%)、有効推定の個数は 30 個 (50.8%) であり、誤推定の個数は 9 個 (15.3%) となった。

そして、我々の実験システムにおいて単語片対辞書を持たないシステム、つまり、帰納的学習を用いて同一分野に出現する単語と訳語の組の間から獲得した単語片対を利用しないシステムを用意し、上述の実験データを用いて、訳語推定実験を行った。その結果、推定完了の個数は 34 個であり、推定未了の個数は 66 個であった。さらに、推定完了の中で、正推定の個数は 8 個 (23.5%)、有効推定は 18 個 (52.9%) であり、誤推定の個数は 8 個 (23.5%) であった。

2 つの実験を比較すると、正推定となったものの個数の差は 12 個、正推定率の差は 10.4 ポイントであった。この結果から同一の分野に出現する単語と訳語の組から獲得した単語片対を、訳語推定において利用することの有効性を確認した。

5. おわりに

本稿では、帰納的学習を用いた訳語推定手法における、システムが獲得する単語片対の利用の有効性について考察するために、獲得した単語片対を利用した訳語推定結果と利用しない訳語推定結果を比較した。その結果、10.4 ポイントの正推定率の向上があった。この結果から、帰納的学習を用いた訳語推定手法における、システムが同一分野に出現する単語と訳語の組の間から獲得した単語片対の利用の有効性を確認した。今後は、機械翻訳システムにおける本手法の汎用的な有効性を確認するための評価実験を行う予定である。そのために、幾つかの機械翻訳システムにおいて辞書未登録語となったものについて本手法を用いて訳語推定を行い、その結果から有効性を考察する。

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の派生語および複合語における有効性の評価", 信学論 (D-II), vol.J81-D-II, No.9, pp.2146–2158, 1998.

[2] 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解析的知識の有効性について", 情処学会 第 134 回 NL 研究会 情処研報 99-NL-134, pp147–154,Nov. 1999.

[3] 荒木健治, 栃内香次, "帰納的学習による語の学習および確実性を用いた語の認識", 信学論 (D-II), vol.J75-D-II, No.7, pp.1213 – 1221, 1992.

[4] 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.

[5] 松本裕治, 北村静, 山下達雄, 今一修, 今村友明. "日本語形態素解析システム『茶筌』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.

[6] 久保正治, 英和・和英電索辞典 gene, 技術評論社,1995.