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

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問題の解決を目指して，帰納的学習 [1] を用いた訳語推定手法を提案し，既にその有効性を確認した [2]. 我々が提案した手法では，単語と訳語の文字列の組の字面から得られる情報のみを利用して訳語推定を行う．単語と訳語の組の間から獲得される単位を単語片対と呼び，この単語片対を組み合わせることにより訳語を生成している．

複合名詞の翻訳手法として，原言語と目的言語の間の対応関係を付与した複合名詞と訳語の組を利用する MBT 3 [3] を佐藤らは提案している．この手法では，大量の良質な複合名詞と訳語の翻訳例を必要とする．そして，これらの翻訳例に対して原言語と目的言語の対応関係を付与するには大きな労力が必要となる．また，異なる分野の複合語を翻訳するためには分野毎に人手により翻訳例の原言語と目的言語の間の対応付けを行わなくてはならない．これに対して我々の手法では，単語片対という単位はシステムが自動的に獲得するので，原言語と目的言語の対応付けには大きな労力を必要としないという利点がある．

ところで，我々の従来の訳語推定手法では文脈に適合した訳語が推定される割合は十分ではない [2]. ここでの，文脈に適合した訳語とは翻訳対象が含まれる問題領域に適合した訳語のことである．我々の従来手法において文脈に適合した訳語が推定される割合が低くなる原因は，訳語推定にとって有効な単語片対が十分獲得されていないためであった．しかし，獲得される単語片対の増加のために，無制限に単語と訳語の間から単語片対の抽出を行うことは，理論的には可能であるが，実際には組み合わせの爆発が起こり処理することは不可能である．また，そのような抽出によって獲得される単語片対の中には原言語と目的言語の対応関係が誤った単語片対が含まれ，そのような単語片対は推定される訳語の精度および質の低下を生む原因になる．本研究では，単語内や訳語内の字面情報のみを用いて文脈に適合した訳語の推定を行うことを目指す．そのために，訳語推定に利用する単語片対を抽出する元となる単語と訳語の組をシステムが制約し，その結果，選択した単語と訳語の組の間から単語片対を抽出し，訳語を推定する．本稿では，文脈に適合した訳語を推定するために単語と訳語の組を制約する方法とその有効であることを示すために行った評価実験の結果およびその考察について述べる．

2. 基本的な考え方

我々が提案した帰納的学習を用いた訳語推定手法において，「単語あるいは訳語の字面の共通部分と差異部分の抽出結果から得られる文字列の並び」を単語片 (PW:a Piece of Word) と呼び，「単語と訳語の二つの異なる文字列の組から抽出される原言語と目的言語の単語片の対」を単語片対 (a Pair of Pieces of Words) と呼んでいる [2].

単語と訳語の組
(diamagnetic, 反磁性体)
(ferromagnetic, 強磁性体)

⇓

抽出された単語片対
単語片対1 (@1 magnetic, @1 磁性体)
単語片対2 (dia, 反)
単語片対3 (ferro, 強)

図 1: 単語片対抽出例

図 1に単語片対の抽出例を示す．この中で，「@1」は変数を表しており，共通部分として抽出された単語片対に対して差異部分が存在していた位置に置かれる．訳語推定処理の際に，変数を持つ単語片対における原言語と目的言語の変数を他の単語片対に置き換え，新たな文字列の組を生成する．

単語片対の抽出元の制約には，人手により作成されたシソーラスのような静的な情報を用いることも考えられる．しかし，そのような情報を利用する場合はシソーラス自体の質が問題となる．さらには，シソーラスに未登録であった場合の処理も問題である．従来から利用されているような静的な情報のみを用いるには検討すべき問題が残されている．そのために，学習を用いて知識を獲得していくことにより，種々の問題の解決を図って行くことを我々は目指している．知識の獲得方法等については，今後解決しなければならない問題とし，研究を進める予定である．本研究では単語内あるいは訳語内の字面から得ることができる情報のみを利用して，単語片対の抽出元となる単語と訳語の組を選択する．

3. 処理過程

3.1 概要

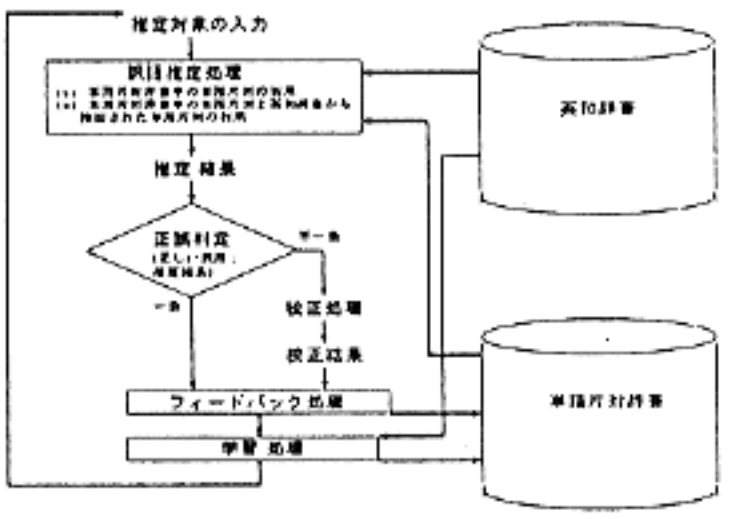

図 2: 実験システム

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る。もし，処理が完了しない場合には，英和辞書の見出し語とその訳語の組から新たな単語片対を抽出する。この際に，単語片対の抽出元となる単語と訳語の組を選択する．そして，既に獲得している単語片対と英和辞書から新たに獲得された単語片対を利用して訳語推定処理を進める．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には，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．この処理において単語片対辞書に追加される単語片対は，既出の推定対象となった単語とその訳語の組に包含されるもののみとした．これは，既出の単語と訳語の組を利用して訳語推定処理において有効な単語片対のみを単語片対辞書に追加するためである．

3.2 単語片対の抽出元の制約方法

単語片対の抽出元の単語と訳語の組を制約する基準は，「推定対象単語と連続して一致する文字列の文字数」とした．

例えば，'*electrical materials*' と '*material*' の２つの文字列の間には 'ri'(2 文字)，'al'(2 文字) および 'material'(8 文字) の共通な文字列が存在する．その中で，この２つの文字列の一致文字数の最大値は 8 となる．また，'*electrical materials*' と '*mate*' の間の一致文字数の最大値は 4 となる．本手法では，'*electrical materials*' との間では，'*material*' の方が '*mate*' よりも関連が深いと判断し，選択される選択順位を高くする．このように，単語片対の抽出元の単語と訳語の組を制約する基準を，「推定対象単語と連続して一致する文字列の文字数」としたことに対する理論的な根拠は存在しないが，形態素解析における「最長一致法」 [4] と同様に，できるだけ長い一致文字列には関連性があると考え，この基準を用いている．

そして，実際の単語片対の抽出元は予め定められた選択順位以上の組を利用する．また，学習処理において抽出されたものの中で単語片対辞書に追加されるのは既出の推定対象となった単語とその訳語の組に包含されるもののみとしている．本手法では，このような制約により文脈に適合した訳語の推定を行っている．

4. 評価実験

4.1 実験方法

帰納的学習を用いた訳語推定手法において，単語片対の抽出元となる単語と訳語の組を制約することの有効性を確認するために評価実験を行った．

実験データは，大学の講座名，研究室名，履修される科目名および専門分野名の英語と日本語の表現 100 組とした．実験では，英語から日本語の訳語を推定するが，実験データの文脈に適した訳語はこの英語と対になっている日本語の表現を使用した．

また，実験システムにおける単語片対の抽出元の母集団となる英和辞書は，電子化された英和辞書 [5] の見出し語と訳語の組 102,156 組と文献 [6] を参照し取り出した英語の接辞とその日本語の訳語の組 351 組を合わせた，102,507 組を利用した．

単語片対の抽出元は上述した方法により選択するが，本実験ではその選択数を 100 組とした．

推定結果は，「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」を正推定と判定した．また，「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」を誤推定と判定した．そして，実験の推定結果に対する評価には以下の式で計算される再現率および適合率を用いた．この式における推定完了数とは，本手法を用いることにより推定結果を得ることができた数である．

$$\text{再現率 [\%]} = \frac{\text{正推定の個数}}{\text{推定対象単語数}} \times 100.0$$

$$\text{適合率 [\%]} = \frac{\text{正推定の個数}}{\text{推定完了数}} \times 100.0$$

4.2 実験結果

実験結果は，表 1のようになった．

我々は，単語片対の抽出元制約の有効性を確認するために，本手法による実験結果と我々の従来手法による訳語推定結果とを比較した．従来手法による実験結果を表 2に示し，両実験における再現率と適合率の推移はそれぞれ図 4, 図 5に示す．両実験の結果では，適合率の差はほとんどないが再現率は本手法の方が約 2 倍高くなっているおり，表 1と表 2により，本手法の正推定の数が従来手法に比べ約 2 倍になっている．これらにより，本手法を用いることは文脈に適合した訳語を推定するために有効であると考えられる．

表 1: 実験結果

データ数	再現率 (%)	適合率 (%)
0	0.0	0.0
25	8.0	25.0
50	14.0	31.8
75	12.0	26.5
100	14.0	26.4

実験データ
(electronic system engineering, 電子システム工学講座)
⇓
既出の実験データ
electrical system engineering: 電気システム工学講座
他

単語片対辞書中の単語片対
system engineering: 気システム工学講座
system engineering: システム工学講座
他

選択された単語片対の抽出元
electronic book: 電子ブック
electronic mail: 電子郵便
他

抽出された単語片対
electronic @0: 電 @0
electronic @0: 電子 @0
他
⇓
推定結果
電 システム工学講座
電子 システム工学講座
他

図 3: 正推定処理例

5. 考察

本実験における正推定の 1 つの例を図 3に示す．

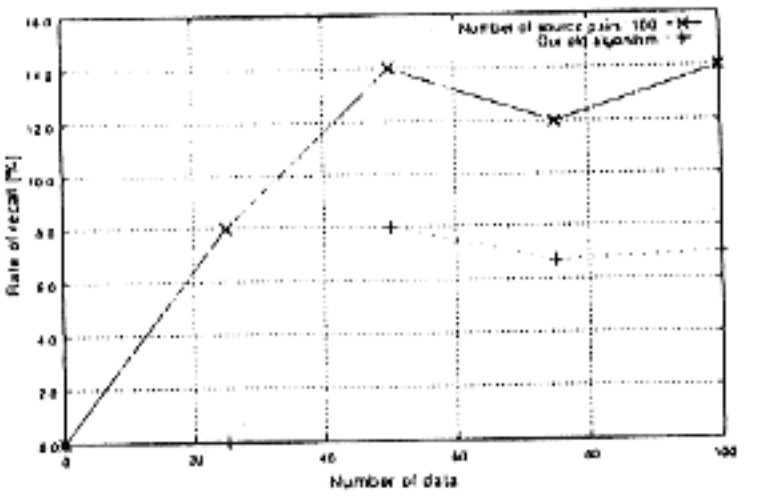

図 4: 従来手法と本手法の再現率の比較

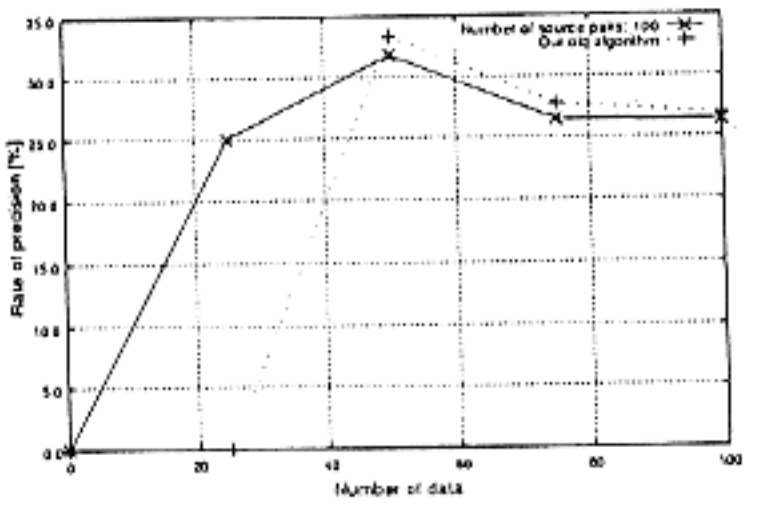

図 5: 従来手法と本手法の適合率の比較

また，誤推定結果と判定された原因を分類すると

1. 原言語と目的言語の対応に誤りのある単語片対を使用したため
2. 語基の訳語に揺れが生じているため
3. 文脈に適合する単語と一致する結果が優先順位 11 位以下に出現したため

となる．誤推定の原因に基づく分類は表 3 のようになる．

原因 (1) により誤推定となったものに関しては，単語片対の抽出の際に原言語側あるいは目的言語側の分割位置を誤り，誤推定の原因となる単語片対を抽出したと考えられる．本手法では，上述したように字面情報のみを用いて，共通部分あるいは差異部分

表 2: 従来手法を用いた実験結果

データ数	再現率 (%)	適合率 (%)
0	0.0	0.0
25	0.0	0.0
50	8.0	33.3
75	6.7	27.8
100	7.0	26.9

表 3: 誤推定となった原因に基づく分類

原因	個数	割合 [%]
(1)	24	68.6
(2)	7	20.0
(3)	4	11.4
合計	35	100.0

として獲得される組を単語片対としている．このような誤推定を招く単語片対を減少させるには，字面情報以外の情報を利用することを検討しなくてはならない．例えば，語基の分類を行い，その分割規則および接続規則を利用することである．しかし，上述したように静的な情報のみを利用するには様々な問題が伴うために，その獲得方法等も含めて今後検討していく予定である．

原因 (2) により誤推定となったものには「gaseous electronics」等があった．この単語に対する文脈に適合した訳語は「気体電子工学」であったが，システムは「気体エレクトロニクス」という訳語を推定した．この推定が行われる以前に実験データの中に「gaseous electronics laboratory, 気体エレクトロニクス研究室」や「power electronics, パワーエレクトロニクス」等が出現していた．そのために「electronics」を「エレクトロニクス」と対応付けるためにこのような推定を行った．しかし，原因 (2) により誤推定となったものは本実験の評価基準では誤推定と判定されるが，出現した単語と訳語の組に忠実な訳語を推定したと見なすことができる．また，限定された分野において特定の訳者により訳された翻訳対象ではこのような語基の訳語の揺れは生じにくいと考えられる．そのために，本手法はそのような場面での利用が適切な利用方法であり，利用分野毎あるいは利用者毎にシステムが動的に適応することによりこのような文脈に適合した訳語を推定することが可能になる．

原因 (3) により誤推定となったものには「laser engineering, レーザ工学」等があった．この推定では，文脈に適合した訳語「レーザ工学」を推定したが，優先順位を 14 位と判定した．また，原因 (3) に関しても原因 (1) と同様な解決方法が必要になると考えられる．本手法では，3 章で述べたように推定結果が複数存在する場合には，字面から得られる情報を利用して推定結果に優先順位を付けている．しかし，本手法で利用している単語内あるいは訳語内の字面情報から得られる情報では情報が不十分であるために，このような誤推定を生んだと考えられる．このような誤推定を減らす方法の一つとしては，単語片対間の接続規則を獲得し，それらにより推定結果の優先順位を決定することが考えられる．

6. 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いて単語内あるいは訳語内の字面情報のみによる訳語推定単位の抽出元の制約の方法を提案した．そして，本手法の有効性を確認するために行った評価実験の結果について述べた．本手法と従来手法による訳語推定結果とを比較すると，再現率および正推定数は約 2 倍に増えており，このことより本手法の有効性を確認した．しかし，本手法による訳語推定結果の再現率および適合率は十分に高いものではなく，解決すべき問題が依然として残されていることが明かになった．そこで，再現率および適合率を向上させるために語基，形態素あるいは単語の分類規則の獲得手法，そしてそれらの接続規則の獲得手法，さらにはそれらの意味情報等の獲得手法について検討する予定である．

謝辞

本研究の一部は文部省科学研究費 (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] 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.

[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.

[3] 佐藤理史, "アナロジーによる機械翻訳", 認知科学モノグラフ 4, 共立出版, 1997.

[4] 長尾 真 (編), "自然言語処理", 岩波講座ソフトウェア科学 15, 岩波書店, 東京, 1996.

[5] 久保正治, 英和・和英電気辞典 gene, 技術評論社,1995.

[6] 前田健三, "強くなる英単語," 有精堂, 東京, 1994.

[7] Nifty Serve 英会話フォーラム, "ハイパー英和辞書," 技術評論社, 東京, 1997.