自然言語処理 128−14
（1998. 11. 6）

帰納的学習を用いた訳語推定手法における学習対象選択の有効性の評価

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

あらまし
我々は，機械翻訳手法における問題の一つである辞書未登録語処理の問題の解決を目指して，帰納的学習を用いた訳語推定手法を提案した。この手法の有効性は確認しているが，翻訳対象に適合した訳語を推定した割合は低かった。しかし，これまでの研究から，同一分野の単語と訳語の組から抽出された訳語推定に利用する単位は，我々の訳語推定手法において有効に働いていることが分かった。そこで，有効な単位を獲得するために，この単位の抽出元を選択することが必要であると我々は考えた。本稿では，帰納的学習を用いた訳語推定手法において訳語推定単位の抽出元をシステムが選択する方法とそ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る。

Evaluation for Selection of Learning Data on Prediction Method of Target Words Using Inductive Learning

Hisayuki Sasaoka†, Araki Kenji‡, Yoshio Momouchi† and Koji Tochinai‡
†Faculty of Engineering, Hokkai-Gakuen University
‡Graduate School of Engineering, Hokkaido University

Abstract
We have studied the prediction method of target words using inductive learning. We have confirmed this method but the system on our method could not predict many target words which adjust to the context. In our research, the units for prediction method, which are extracted from text in the same field, are effective for the prediction process on our method. Our system needs to select the learning data to acquire the effective units. This paper presents the selection method of learning data on the system and the results of evaluation experiment used on our method.

1 はじめに

近年，機械翻訳に対する需要の増加に伴い，様々な機械翻訳手法が提案されている。しかし，提案されている手法には，翻訳精度の点においても翻訳結果の品質の点においても不十分であること等の問題がある [1]。このような問題の中の一つに辞書未登録語の処理がある。従来提案されている多くの機械翻訳手法では，システムの辞書に登録されていない単語を翻訳することはできない。そのために，このような辞書未登録語が翻訳対象に出現する場合，翻訳精度の低下を招いている。そこで，この辞書未登録語の処理の問題の解決を目指し，我々は帰納的学習を用いた訳語推定手法を提案した [2], [3], [4]。

機械翻訳において利用する辞書のカバレッジの向上を目指し，原言語と目的言語の対訳表現を抽出する手法の研究が提案されている [5],[6],[7]。機械翻訳で利用する辞書の改善は，翻訳精度の向上のための一つの手法であると考えられるが，対訳表現を抽出するためには大量の良質である同一分野の対訳コーパスが必要となる。また，依然として翻訳対象に適合した訳語を選択する訳語選択の問題が存在する。

実例を用いた翻訳手法として佐藤らにより MBT3 が提案されている [8]。MBT3 では，原言語と目的言語の間の対応関係を与えた複合名詞と翻訳例の集合の中から，翻訳対象と類似している翻訳例を抽出し，これを利用して翻訳を行っている。しかし，翻訳例中の原言語と目的言語の各要素の対応を付けることは大変困難であり，大きな労力が必要となる。佐藤らの手法では翻訳例においてこの原言語と目的言語の対応付けを利用するため，システム作成者が原言語と目的言語の要素間に対応を付けた大量の翻訳例を用意する必要があり，しかも翻訳例が用意された分野でしか利用できない。これに対して我々の提案する手法では，単語片対という原言語と目的言語の文字列の組を定義した。そして，単語と訳語の組あるいは既に獲得された単語片対の間から帰

- 単語と訳語の組

単語	訳語
diamagnetic	反磁性体
ferromagnetic	強磁性体
	他

- 抽出される単語片対

共通部分	@1 magnetic	@1 磁性体
差異部分 1	dia	反
差異部分 2	ferro	強

図 1: 単語片対抽出例

納的学習を用いてシステムが自動的にこの単位を獲得している。そのために，我々の手法では対象分野に自動的に適応することが可能である。

しかし，これまでに我々が提案した手法では訳語の推定は行われているが，翻訳対象に適合した訳語を推定できる割合は低かった [4]。さらに，推定対象となる分野に出現した単語と翻訳対象に適合した訳語の組から抽出した単語片対は翻訳対象に適合した訳語推定に有効であることが確認された。そのために，翻訳対象に適合した訳語を推定するには単語片対の抽出元の単語と訳語の組を選択する必要があると我々は考えた。本稿では，帰納的学習を用いた訳語推定手法において訳語推定に利用する単位である単語片対の抽出元となる文字列の組の選択処理方法およびその有効性について述べる。

2 基本的な考え方

我々がこれまで進めてきた帰納的学習を用いた訳語推定手法の研究において，「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」を単語片対と呼んでいる [2],[3],[4]。

図 1に単語片対の抽出例を示す。この中で，「＠１」は変数部分を表している。変数は共通部分として抽出された単語片対に対して，抽

出元の文字列におい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の位置に他の単語片対を組み合わせ，新たな文字列の組を生成する。

上述したようにこれまでの研究から，同一分野に出現している単語と訳語の組から抽出された単語片対は有効な単位であることが明らかになっている [4]。しかし，単語片対を抽出する抽出元を文脈に出現する単語と訳語の組のみとするとデータのスパースネスのため，必要な単語片対の獲得は困難になる [2]。また，英和辞書の全ての見出し語と訳語の組を単語片対の抽出元とすることは，その処理量が大きくなり，実現が難しい。

そこで，我々は単語片対の抽出元となる単語と訳語の組を訳語を推定する単語と関連する組とした。訳語を推定する単語と関連があると判断する基準は，「推定対象単語と一致する文字数」とした。処理対象単語は訳語さえもわからない未知語であるため，利用可能な情報はプリミティブな情報に限られる。そこで、字面から得られる情報のみを利用して単語片対の抽出元の選択を行うことにした。また、これは人間が未知語の訳語推定を行う場合，利用できる情報が制限されている場面では，訳語を推定する単語に関連する幾つかの単語を想起し，それらの単語から何らかの対応関係を発見し，訳語推定を行うことに対応すると考えられる。

3 処理過程の概要

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を利用して訳語推定を試みる。もし，訳語推定処理が完了すれば，推定結果の正誤判定に処理を進める。また，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。そして，既に獲得している単語片対と英和辞書から獲得された単位を利用して訳語推

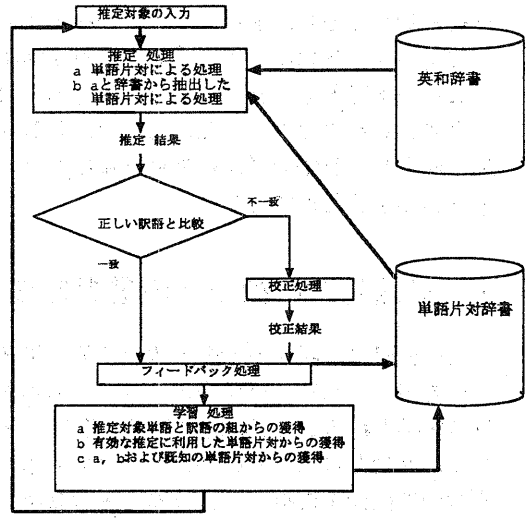

図 2: 実験システム

定処理を進める。また，訳語推定処理において複数の推定結果が現れた場合，各推定結果を構成している単語片対と既出の原言語と目標言語の文字列の組が部分的に一致する回数を計算し，その数値が大きいものを優先する。これにより優先順位がつけられない場合，各推定結果を構成している各単語片対の訳語推定処理における過去の利用状況を示す数値である正推定度数あるいは誤推定度数を参照し，推定結果の優先順位を決定する。その後，推定結果の正誤判定に処理を進める。もし，推定結果が正しかった場合，フィードバック処理に進み，推定結果が誤っていた場合には，推定結果に校正処理を行い正しい推定結果とした後にフィードバック処理に進む。フィードバック処理では，推定結果を構成する各単位にその正誤判定結果に応じて，各単位の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，推定対象となっている単語とその訳語の組，有効な訳語推定に利用した単語片対および有効な訳語推定に利用した単語片対の抽出元となった辞書の見出し語と訳語の組を単語片対辞書に追加する。そして，新たに追加された単語片対とそれ以前に存在していた単語片対の間から，さらに，新たな

単語片対を抽出し，追加する。この処理において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と一致あるいは包含されるもののみとした。これは，翻訳対象に適合した訳語の推定に有効な単語片対のみを追加するためである。

4 単語片対の抽出元の選択方法

同一分野に出現する単語と訳語の組に単語片対を限定すれば，訳語推定に有効な単語片対が獲得されることが確認されている [4]。そこで，有効な単語片対を獲得するために単語片対の抽出元を選択する。英和辞書の見出し語と訳語の組を利用して単語片対を抽出する際にこの選択を行う。選択の方法は，訳語を推定する単語と英和辞書の各見出し語と一致する文字数の最大値を求め，その一致文字数が多いものを選択する。

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。我々の一致文字数を基にした方法では，'*electrical materials*' との間では，'*material*' の方が '*mate*' よりも関連が深いと判断し，選択される順位が高くなる。そして，実際の単語片対の抽出は予め定めた順位以上の組を利用して行う。

5 評価実験

5.1 実験方法

帰納的学習を用いた訳語推定手法において，単語片対の抽出元を制限することの有効性を確認するために，上述した実験システムを用いて評価実験を行った。実験データは大学の工学部の講座名および講座で履修されている科目名の英語と日本語での表記を用いた。データ数は 100 組とした。

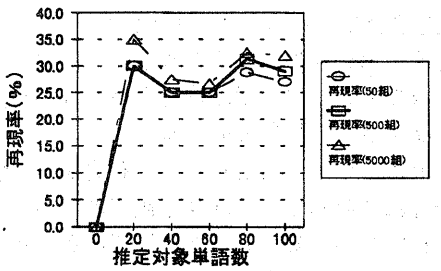

図 3: 再現率の推移

実験システムにおいて利用した英和辞書は，電子化された辞書 [9] を利用した。この辞書は，一般的な英語文書を読むことを目的として見出し語が選ばれており，英語の見出し語数としては 49,336 語登録されている。そして，1 つの単語が複数の訳語を持つ場合もあるために，この辞書中には単語の訳語の組としては 102,156 組ある。

推定結果の評価基準は，「推定結果の優先順位 10 位以内に正しい訳語が存在する推定」を正推定とし，「推定結果を出力したが，10 位以内に正しい訳語がない推定」を誤推定と見なした。そして，以下の式のように適合率および再現率を計算した。

$$\text{再現率 (\%)} = \frac{\text{推定を完了した個数}}{\text{推定対象単語数}} \times 100.0$$

$$\text{適合率 (\%)} = \frac{\text{正推定の個数}}{\text{推定を完了した個数}} \times 100.0$$

表 1: 実験終了時の実験結果

選択した組数	再現率 (%)	適合率 (%)
50	27.0	29.6
500	29.0	44.8
5000	32.0	34.4

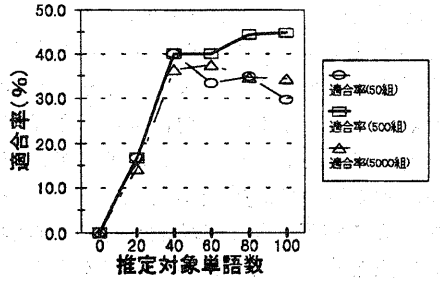

図 4: 適合率の推移

5.2 選択する組数

本評価実験では，上述したように英和辞書の見出し語と訳語の組から単語片対を抽出する際に抽出元の組を選択した。そして，選択した組数は，50 組，500 組，5000 組とした。

5.3 実験結果

再現率と適合率の推移は，それぞれ図 3，4 のようになった。実験終了時の再現率及び適合率は表 1となり，再現率は単語片対の抽出元の組み数が 5000 組の場合が一番高く 44.8 %であり，適合率は 500 組の場合で 34.4 %であった。

5.4 考察

5.4.1 単語片対の抽出元の選択の有効性

これまでに行った訳語推定実験 [4] の有効な推定を行った推定結果の中での，翻訳対象に適合した訳語が 10 位以内に出現した割合を計算してみると，4 つの異なる分野における平均で 12.3%であった。この数値と比較すると本評価実験の適合率は最大の時で 44.8%あり、32.5 ポイントの向上があり，適合率が最低の時でも 29.6%あり 17.3 ポイントの向上があった。しかし、再現率は最大の時でも 32.0%であり、十分に高いとは言えない。

次に，英和辞書の見出し語と訳語の組から抽出された単語片対を利用して訳語の推定に成功した実験結果の例を図 5に示す。「electronic system engineering:電子システム工学講座」の訳語推定において，単語片対「electronic @1，電子 @1」および「system engineering，システム工学講座」を利用して正しい訳語と一致する訳語を生成した。単語片対「electronic @1，電子 @1」は，英和辞書の見出し語と訳語の組「electronic book，電子ブック」と「electronic mail，電子郵便」の間の共通部分の文字列の組として抽出されている。そして，「system engineering，システム工学講座」は，推定対象単語と同一の文脈中の単語と訳語の

- 対象単語より前に出現した文字列の組

対象単語	正しい訳語
electrical system engineering:	電気システム工学講座

- 対象データ

単語	正しい訳語
electronic system engineering:	電子システム工学講座

1. 単語片対辞書

英語の単語片	日本語の単語片
system engineering:	気システム工学講座
system engineering:	システム工学講座

他

2. 英和辞書

対象単語	正しい訳語
electronic book:	電子ブック
electronic mail:	電子郵便

他

3. 英和辞書から抽出したもの

英語の文字列	日本語の文字列
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 5: 推定処理例

表 2: 実質的な再現率

選択した組数	再現率 (%)
50	49.0
500	52.7
5000	58.2

組「electrical system engineering, 電気システム工学講座」および英和辞書の見出し語と訳語の組「electrical conductivity, 電気伝導率」の間の差異部分の文字列の組として抽出されている。

5.4.2 推定未了の推定対象単語

訳語の推定が完了しなかったものの原因の考察を行った。推定対象データにおいて単語と訳語の組が英和辞書あるいは他の推定対象データ中に存在するのかしないのかを調査した。本手法は英和辞書と他の推定対象データの字面情報を元にして訳語を生成している。このために，推定対象データとその訳語が訳語と英和辞書あるいは他の推定対象データ中に存在しないものについては本手法では推定できないものになる。このようなデータは全データ 100 組中に 45 組存在した。つまり，本手法を用いて推定できる可能性があるものは 55 組であった。この数値をもとにして，実質的な再現率を計算し直すと表 2のようになった。この表から，本手法により訳語が推定可能な推定対象データの中で，約 60%～ 50%が訳語推定を完了していることが確認できる。

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システムが単語片対の抽出元を選択する方法とその有効性を確認するための評価実験の結果について述べた。単語片対を抽出する元となる文字列の組の数を変化させて実験を行ったところ，抽出元の組数が 5000 組の時に再現率が 32.0%で最高になり，適合率は抽出元の組数が 500 組の時に 44.8%で最高となった。本実験において，従来の手法 [2],[3],[4] よりも翻訳対象に適した訳語を推定した割合は高くなっていることにより本手法の有効性を確認した。しかし，十分に高い再現率および適合率は得られてはいない。そこで，実験結果に対する考察を更に進め，ヒューリスティクスによる制限に対する検討を行い，再現率および適合率の向上を図る予定である。

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] 長尾真 編,“自然言語処理”, 岩波書店,1996.

[2] 笹岡久行, 荒木健治, 桃内佳雄, 栃内香次, “帰納的学習による単語片対抽出を用いた未登録語の訳語推定”, 1996 信学ソ大, D-53,Sep., 1996.

[3] Hisayuki Sasaoka, Kenji Araki, Yoshio Momouchi and Koji Tochinai, “Prediction Method of Word for Transaltion of Unknown Word”, In *Proceedings of Artificial Intelligence and Soft Comuting*, pp.228–231, Banff, Canada.

[4] 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, vol.J81-D-II, No.9, pp.2146 – 2158, 1998.

[5] 山本由紀雄, 坂本仁, “対訳コーパスを用いた専門用語対訳辞書の作成”, 情処自然言語処理研究報告, NL94-12, pp.85–92, 19 March 1993.

[6] 熊野明, 平川秀樹, "対訳文書からの機械翻訳専門用語辞書作成", 情処学論, vol.35, no.11, pp.2283–2290, 1994.

[7] 北村美穂子, 松本裕治, "対訳コーパスを利用した対訳表現の自動抽出", 情処学論, vol.38, no.4, pp.727–736, 1997.

[8] 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.

[9] 久保正治, 英和・和英電索辞典 gene, 技術評論社,1995.