

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に関する一考察[†]

5R-1

笹岡久行 荒木健治 桃内佳雄
北海道学園大学工学部[‡]

栢内香次[†]
北海道大学大学院工学研究科[¶]

1. はじめに

近年、我々には母国語以外の言語を扱う機会が増えており、そのために機械翻訳に対する需要が高まりつつある。しかし、実際の機械翻訳システムには種々の問題がある。そこで、機械翻訳システムにおける辞書未登録語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 [1],[2]。そして、評価実験において4つの異なる分野から抽出した辞書未登録語での手法の有効性を確認した。しかし、この評価実験において推定結果の優先順位を決定する方法が十分ではないこと等の幾つかの問題点が考察された。そこで、本稿では、この訳語推定結果の優先順位決定方法の問題を解決するために既出の原言語と目標言語の文字列の組を利用し、システムが自動的に訳語推定に利用する単位の確からしさを評価する方法を提案する。

2. 基本的な考え方

我々の帰納的学習を用いた訳語推定手法は、人間が持つ言語獲得能力の工学的な実現という目的のもとで行われている研究の一つである。従来の手法において、推定結果が複数存在する場合、その推定に利用している単位の確からしさに応じての優先順位は決定されている。我々が定義した単語片対という単位は、それ以前の推定処理における使用結果に応じて評価されてきた。この従来の単位の評価方法は、正しい訳語推定処理に数多く利用された単位はより確からしく、逆に誤った訳語推定処理に数多く利用された単位は確からしくないものと見なしている。しかし、この方法では獲得された多くの単位が訳語推定に使用されなかったり、使用されてもその回数が少ないために単語片対の単位の間に確からしさの差が生じ難く、その結果として訳語推定結果の優先順位決定方法が有効となっていなかったことがこれまでの評価実験において考察されている [2]。さらに、これまでの評価実験から文脈に適した訳語を推定するためには、推定に必要な単位を学習する対象も文脈に適したものであった方が良いことが考察されている [2]。本研究では、「既出の用例を利用し、獲得した規則の確からしさを評価する」という基本的な考え方に基いて規則の評価を行い、その

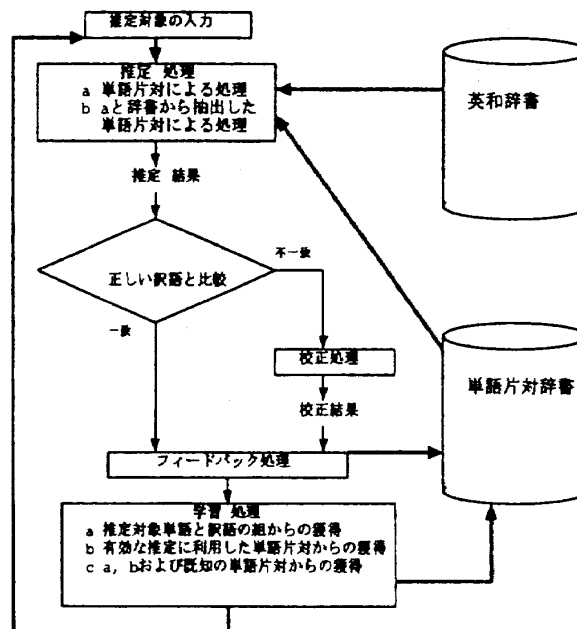


図1: 実験システム

評価結果は推定結果の優先順位の決定のために利用する。

3. 処理過程

図1に、実験システムの概要を示す。

システムでは、推定対象単語が入力されると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、辞書の見出し語とその訳語の文字列の組から新たな単語片対を抽出し、その単位を利用して訳語推定処理を進める。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、正推定

[†]Consideration of the Effectiveness of Selection of Learning Target on Prediction Method of Target Word Using Inductive Learning

[‡]Hisayuki Sasaoka, Kenji Araki, Yoshio Momouchi and Koji Tochinnai

[§]Faculty of Engineering, Hokkai-Gakuen University

[¶]Graduate School of Engineering, Hokkaido University

- 対象単語より前に出現している文字列の組

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

- 対象データ

単語	正しい訳語
electron device engineering:	電子デバイス工学講座

1. 単語片対辞書

英語の単語片	日本語の単語片
electron @1:	電子 @1
electron @1:	電子の @1
engineering:	工学講座
	他

2. 英和辞書の見出し語

対象単語	正しい訳語
engineering:	エンジニアリング
engineering:	工学技術
device:	装置
	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
electron @0:	電 @0
device @1:	デバイス @1
	他

↓

- 推定結果

電 装置 工学講座
電子 装置 工学講座
電子の 装置 工学講座
電 デバイス 工学講座
電子 デバイス 工学講座
電子の デバイス 工学講座
他

図 2: 推定処理例

度数あるいは誤推定度数を操作する。そして、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、新たに追加された文字列の組とそれ以前に存在していた文字列の組の間から、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文字列とその訳語の文字列の組と部分一致するもののみとした。これは、原言語と目標言語の誤った対応を持つ単語片対の爆発的な増加を抑えるためである。また、文脈に出現した文字列のみに制限しなかったのは、その制限により訳語推定の再現率が低下するのを防ぐためである。

4. 推定処理例

今回は、大学の工学部の講座名(10個)を実験データ

に利用し、訳語推定実験を行った。この実験において英和辞書は「gene」[3]を利用した。その結果、8個について推定が完了し、そのうち4個が優先順位10位以内に正しい訳語と一致する訳語を得ることができた。

図2に、訳語推定に成功した例を示す。この中で、「electron @1:電子 @1」は「electron affinity:電子親和力」と「electron gun:電子銃」の間の共通部分として抽出された単語片対であり、「electron @1:電子の @1」は「electron affinity:電子親和力」と「electron gun:電子銃」の間の共通部分として抽出した。また、従来の優先順位決定方法では、「engineering:工学講座」、「engineering:エンジニアリング」および「engineering:工学技術」の確からしさは各々が訳語推定に利用されるまで知ることが不可能であった。しかし、今回提案している手法では、「対象単語より前に出現している文字列の組」を参照することにより、「engineering:工学講座」の確からしさが他の2つよりも大きく、その結果、この単位を利用した推定結果の優先順位が高くなる。このように、提案した方法により正しい訳語と一致する推定例が優先順位10位以内に得られている。なお、図中の単語片対の中の「@1」は変数を表し、他の文字列の接続位置を表している。

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いて、システムが既出の原言語と目標言語の文字列の組を利用し、訳語推定に利用する単位の確からしさを決定し、推定結果の優先順位を決定する手法を提案し、その推定例を示した。今後は、大量の実験データを用いて評価実験を行い、本手法の有効性を確認し、考察を進める予定である。

謝辞

本研究の一部は文部省科学研究費補助金(課題番号09878070)により行われた。

参考文献

- [1] 笹岡久行、荒木健治、桃内佳雄、柄内香次: 帰納的学習を用いた訳語推定手法の固有名詞における有効性の評価、言語処理学会第4回年次大会論文集、pp.344-347,1998.
- [2] Hisayuki Sasaoka, Kenji Araki, Yoshio Momouchi and Koji Tochinnai, "Evaluation of Prediction Method of Target Words Using Inductive Learning," In *Proceedings of Artificial Intelligence and Soft Computing*, pp. 183 - 186, Cancun, Mexico, 1998.
- [3] 久保正治: 英和・和英電索辞典 gene, 技術評論社, 1995(東京).