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

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　　北海道大学工学部

1 はじめに

機械翻訳システムはその有用性から高い需要があり、種々の手法が提案されているが、引き続き考察が必要な課題が残されている [1]。その一つに固有名詞の処理がある。

従来の自然言語処理分野での固有名詞に関する研究において、Wakaoらの研究 [2] は辞書的な知識や文法規則等を利用し、固有名詞をテキストから抽出し分類する手法を提案している。さらに、島津らの研究 [3] では機械翻訳における固有名詞処理を扱い、ヒューリスティクスおよびキーワードを利用し文中の固有名詞を抽出する手法を提案し、機械翻訳システムにおいてその有効性を確認している。Wakaoらの研究 [2] および島津らの研究 [3] では、固有名詞の抽出に予め用意された手がかりを用いている。しかし、固有名詞の中の組織名等では新出語が出現するために、用意された有限個の手がかりだけで処理するのは困難である。また、島津らの研究 [3] では固有名詞の訳語を生成してはおらず、アルファベットの綴りのまま訳出している。

これまでの我々の研究では、帰納的学習を用いた訳語推定手法における派生語と複合語における有効性を確認した [4]。本稿では帰納的学習を用いた訳語推定手法について述べ、さらに、タグ付けされたコーパスから人手により抽出した固有名詞を用いて行った評価実験の結果とその考察について述べる。

2 訳語推定に利用する単位

本手法では、「要素合成原理」 [1] に基づき原言語および目的言語の文字列を組み合わせて訳語の生成を行っている。

本手法では、訳語推定に必要となる基本的な単位は予め用意する。そして、用意された以外で必要となるものは帰納的学習を用いてシステムが獲得する。用意された基本的な単位ならびに獲得された単位を利用して訳語を生成する。

ここで、帰納的学習を用いて獲得するもので、「2つの原言語の文字列あるいは2つの目的言語の文字列における共通部分あるいは差異部分として抽出された文字列」を単語片と呼ぶ。また、「2組の原言語の文字列と目的言語の文字列の組から抽出された原言語の単語片と目的言語の単語片の組」を単語片対と呼んでいる。

抽出元 1:

単語 1: Hokkaido University
訳語 1: 北海道 大学

抽出元 2:

単語 2: Hokkai-Gakuen University
訳語 2: 北海学園 大学

⇓

共通部分と差異部分の抽出

共通部分:	University	大学
差異部分 1:	Hokkaido	北海道
差異部分 2:	Hokkai-Gakuen	北海学園

⇓

変数 ("@1") の付与

単語片対 1:	@1 University	@1 大学
単語片対 2:	Hokkaido	北海道
単語片対 3:	Hokkai-Gakuen	北海学園

図 1: 単語片対抽出例

従来の研究 [2, 3] 等で行われているように、固有名詞を処理するための規則を人手により与えることは可能であるが、実際に与えられる規則の数は有限個である。上述したように、自然言語における固有名詞を有限個の規則だけで処理するのは困難である。そこで、本手法では訳語推定において人手により与えられた規則以外に必要となるものを帰納的学習を用いて獲得する。

図 1に、単語片対抽出の例を示す。図中で、"@1" は変数を表している。この変数は、訳語を生成する際に他の文字列を代入する位置となる。仮に、図 1において抽出された単語片対「@1 University, @1 大学」以外に「Kyushu, 九州」という単語片対が存在し、「Kyushu University」の訳語を推定する場合、推定結果は「九州 大学」となる。

3 処理過程

3.1 システムの概要

本手法を基にして作成した実験システムの概要を図 2に示す。

辞書未登録語がシステムに対して入力されると、最初に訳語推定部において訳語推定処理が行われる。この訳語推定部では、システムが持つ訳語推定単位辞書の中に存在する訳語推定単位を利用して訳語推定処理を行う。推定結果と正しい訳語が一致する場合には、学習部に処

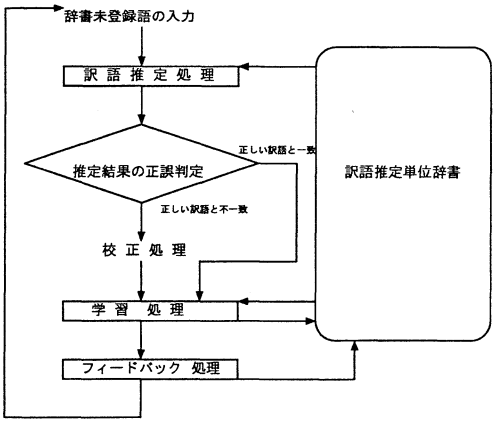

図 2: 実験システム

理を進める。しかし、推定結果と正しい訳語が一致しない場合には、人手により校正処理を行い、正しい訳語をシステムに与えた後に学習部に処理を進める。学習部ではこの辞書未登録語と正しい訳語の組および既に獲得されている訳語推定単位を利用して新しい単語片対を獲得する。最後に、フィードバック部ではシステムの訳語推定能力の向上のためのフィードバック処理を施す。この処理では、推定結果の正誤判定に応じて、訳語推定のための単位が持つ確からしさを表す数値を操作することにより訳語推定能力の向上を図る。

3.2 推定処理

訳語推定処理の手順は、以下の通りである。

1. 最初に原言語の辞書未登録語の字面に一致する単位を持つ訳語推定単位を辞書の中から検索
2. 検索された原言語の文字列を組み合わせ、その目的言語側の文字列も合成
3. 生成された原言語の文字列が辞書未登録語の綴りを再現可能な場合、それと対を成す目的言語の文字列を推定結果として生成

このような字面情報による推定では、複数の組み合わせが存在する場合がある。そのような場合には、以下の数値を参照して推定結果に対する優先順位を決定する。

（1） 推定結果に利用されている単位の数

（2） 推定結果に利用されている単位の出現度数の平均

（3） 推定結果に利用されている単位の正推定度数の平均

（4） 推定結果に利用されている単位の誤推定度数の平均

（5） 推定結果に利用されている単位が持つ目的言語側の文字列の正接続度数の平均

（6） 推定結果に利用されている単位が持つ目的言語側の文字列の誤接続度数の平均

ここで、訳語推定のための単位が正しい訳語と一致する推定結果に利用された回数を「正推定度数」と呼び、訳語推定のための単位が正しい訳語と一致しない推定結果に利用された回数を「誤推定度数」と呼んでいる。また、訳語推定のための単位が持つ目的言語側の文字列ともう一方の単位が持つ文字列の間の接続が正しい訳語と一致する推定結果に利用された回数を「正接続度数」と呼び、このような接続が正しい訳語と一致しない推定結果に利用された回数を「誤接続度数」と呼んでいる。

（1）は数値の大きい推定結果を優先する。これは、推定結果においてはより長い文字列を利用した組み合わせの方が確からしいというヒューリスティクスに基づいている。（2）は数値の大きい推定結果を優先する。この理由は、過去の推定処理においてより多く利用された訳語推定単位の方が確からしいというヒューリスティクスに基づいている。（3）と（5）は数値の大きい推定結果を、（4）と（6）は数値の小さい推定結果を優先する。これは、過去の推定処理において正しい訳語と一致する推定結果に利用された回数が多いものほど確からしく、また、正しい訳語と一致しない推定結果に利用された回数が少ないものほど確からしいというヒューリスティクスに基づくものである。

```
単語 1:    gentleman
訳語 1:    紳士
単語 2:    gentlemanly
訳語 2:    紳士的な
              ⇓
単語片対 1:  gentleman @1    紳士 @1
単語片対 2:       ly          的な
```

図 3: 条件 1 に合う単語片対の抽出例

```
単語 1:    thermometer
訳語 1:    温度計
単語 2:    voltmeter
訳語 2:    電圧計
              ⇓
単語片対 1:  @1 meter    @1 計
単語片対 2:   thermo      温度
単語片対 3:    volt       電圧
```

図 4: 条件 2 に合う単語片対の抽出例

3.3 学習処理

学習部では異なる2組の訳語推定単位から新たな単語片対を獲得する。学習部における学習処理は、訳語推定単位を構成している原言語の文字列と目的言語の文字列の各々の字面における共通部分と差異部分の抽出により新たな単語片を抽出し、その単語片の組を単語片対とする。

しかし、この字面情報だけを利用した学習処理では原言語の単語片と目的言語の単語片の間の意味的な対応関係が誤った単語片対を数多く抽出することが既に行った実験 [5] から明らかになっている。そこで、新たな単位抽出のために幾つかの抽出条件を設定した。

まず、本研究での原言語である英語の単語片の抽出では、「共通部分として抽出される文字数は2文字以上」とした。これは、英語を表記する文字の種類は目的言語である日本語を表記するための文字の種類に比べて少ない。また、英語を表記する文字は表音文字であり、1文字で意味を表わす場合は稀である。このような理由から、共通部分における文字数に関する英語の単語片の抽出条件を設定した。

さらに、2つの単語片対の抽出に関する条件を以下のように設定した。

（条件1）　2組ある抽出元は、片方の組は1組の共通部分である文字列の組のみで構成され、もう一方は1組の共通部分である文字列の組と1組の差異部分である文字列の組により構成される場合

（条件2）　2組ある抽出元は、双方とも1組の共通部分である文字列の組と1組の差異部分である文字列の組で構成される場合

（条件1）に当てはまる抽出例を図 3に示し、（条件2）に当てはまる抽出例を図 4に示す。

4 評価実験

4.1 実験方法

本手法の固有名詞における有効性を確認するために評価実験を行った。本評価実験では、“Susanne Corpus” [6] の “Press Reportage(報道記事)” における6個のテキストから実験データを抽出した。このコーパスは予めタグ付けされており、本実験では固有名詞としてタグ付けされた 424 個を実験データとした。

訳語推定に必要となる基本的な単位は、電子化された辞書 “Gene” [7] および “名辞郎” [8] に記載されている英単語とその訳語の組とした。その組の総数は、205,118組であり、これを訳語推定のための基本的な単位として単語片対辞書に与えた状態から実験を行った。

実験データに対する正しい訳語は辞書 [9] を参照し、人手により作成した。そして、正しい訳語と一致し、推定結果における優先順位が5位以内であるものが存在する場合を有効な推定とし、有効な推定以外を無効な推定とした。また、訳語推定における有効な推定の占める割合を表す有効推定率を以下のように定義した。

$$\text{有効推定率 (\%)} = \frac{\text{(有効な推定数)}}{\text{(実験データ数)}} \times 100.0$$

図 5: 有効推定率

上述した実験データに対し、本手法を基にした実験システムを用いて訳語推定実験を行い、評価した。

4.2 実験結果と考察

表 1: 実験結果

有効な推定数	率 (%)	無効な推定数	率 (%)	総数	率 (%)
152	35.8	272	64.2	324	100.0

表 1に実験結果を示す。有効推定率は、35.8%であった。

対象データ: Texas supreme court

単語片対 1:　“Texas @1,　テキサス @1 ”
単語片対 2:　“Texas @1,　テキサス州 @1 ”
単語片対 3:　“supreme court,　最高裁判所 ”
単語片対 4:　“court,　裁判所 ”

他

⇓

推定結果 1:　テキサス　最高裁判所
推定結果 2:　テキサス州　最高裁判所

正しい訳語　テキサス州最高裁判所

図 6: 有効な推定例

図 6に、有効な推定結果の例を示す。この推定例において、単語片対「Texas @1, テキサス州 @1」は「Texas research league, テキサス州研究団体」と「Texas bankers, テキサス州の銀行家」の間の共通部分として抽出されている。一方、「supreme court, 最高裁判所」は基本的な単位として与えられた単位であった。

図 7に、カタカナ表記された訳語の推定に成功した例を示す。この推定例において、単語片対「Sa @1, サ @1」は「Sam Caldwell, サム・コールドウェル」と「Sa'gya, サギャ」の間の共通部分として抽出されていた。また、単語片対「ndman, ンドマン」は「Hyannis Port, ハイアニス・ポート」と「Hyndman, ハインドマン」の差異部分として抽出されていた。

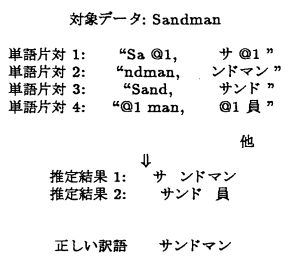

図 7: カタカナ表記の訳語を持つ単語の有効な推定例

次に、無効な推定に対する考察を行う。無効な推定272個を、無効な推定となった原因に基づいて分類すると以下のようになる。

1. 必要な文字列の単位が不足したもの 145 個
2. 必要な文字列の単位は存在するが、単位中に変数が存在しないので文字列の接続が不可能であったもの 113 個
3. 訳語が略称等のために推定が不可能だったもの 9 個
4. 正しい訳語を生成したが、その優先順位が 5 位以下だったもの 5 個

ここで、1に属するものを不足している単位の種類を基にして分類すると以下のようになる。

a. カタカナ表記される人名や地名などの主要な部分のみが不足したもの 43 個
b. 人物の肩書などの補助的な部分のみが不足したもの 37 個
c. 主要な部分および補助的な部分が不足したもの 37 個
d. 人名におけるミドルネーム (アルファベット表記) や Jr. 等の人名に関わるもののみが不足したもの 28 個

このような推定を有効な推定に改善するには、まず、b および d のような文字列は固有名詞の訳語推定のための基本的な単位と見なし、予め用意しておくことが考えられる。一方、単語片対の学習量を増加させるにつれ有効な単語片対は増加する。そのため、a および c のような無効な単語片対は単語片対の学習量を増加させるにつれ減少させることが可能である。

また、2に属するものの改善には、単語片対の接続に関する学習手法の検討等が必要である。そして、3に属するものに関しては、本手法だけでは訳語を生成することは困難であるので、人手により正しい訳語を与えなくてはならない。最後に、4に属するものの改善には、優先順位決定手法に対する検討が必要である。

5 おわりに

本稿では、帰納的学習を用いた訳語推定手法を提案し、本手法の固有名詞における有効性を調べるために行った評価実験の結果について述べた。本実験において、有効推定率は 35.8% と十分に高い数値は得られなかった。しかし、実験結果に対する考察から、幾つかの改良方法により改善が可能であることが明らかになった。今後は、これらの方法を用いて本手法を改良する予定である。

参考文献

[1] 長尾真 編, "自然言語処理," 岩波書店, 東京, 1996.

[2] Takahiro Wakao, Robert Gaizauskas and Yorick Wolks, "Evaluation of an Algorithm for the Recognition and Classification of Proper Names," In *Proceedings of Coling '96*, , pp. 418 – 423, Copenhagen, Denmark, 1996.

[3] 島津美和子、吉村裕美子、"機械翻訳における固有名詞処理," 言語処理学会第 3 回年次大会論文集, pp. 35 – 38, 1997.

[4] Hisayuki Sasaoka, Kenji Araki, Yoshio Momouchi, Koji Tochinai, "Prediction Method of Word for Translation of Unknown Word," In *Proceedings of Artificial Intelligence and Soft Computing*, pp. 228 – 231, Banff, Canada, 1997.

[5] 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抽出を用いた未登録語の訳語推定," 電子通信学会全国大会, D-53, Sep. 1996.

[6] Geoffery Sampson, "English for the Computer," Oxford University Press, Oxford, 1995.

[7] 久保正治, "英和・和英電策辞典 *Gene*," 技術評論社, 東京, 1995.

[8] EDP, "ハイパー英語辞典, " 技術評論社, 東京, 1997.

[9] 小稲義男 他, "新英和大辞典" 研究社, 東京, 1980.