目的言語の単言語コーパスを利用した訳語学習方式

鈴木 博和　　熊野 明

(株)東芝 研究開発センター 知識メディアラボラトリー

hirokaz.suzuki@toshiba.co.jp

1. はじめに

機械翻訳システムにおいて高品質な翻訳を実現するためには、一文内で訳語が正しいだけでなく、訳語がユーザの嗜好や分野、使用目的に依存するような単語に対しても適切な訳語を出力できることが不可欠である。訳文に不適切な訳語が使用されている場合は、訳語候補の中から適切な訳語を選択しなおしシステムに指示する作業(学習作業)を行い、以降の翻訳ではその訳語を優先的に選択することによって適切な訳語を出力することができるようになる。この一連の流れを訳語学習と呼ぶ。

従来の機械翻訳システムの訳語学習では、ユーザが文書ごとに逐次的に学習作業を行わなければならず非常に煩雑であった。従って、学習作業を自動化することによる効率的な訳語学習が期待される。

自動的な学習作業に関しては、2言語コーパスを用いる手法 [8, 11, 12] と単言語コーパスを用いる手法 [10, 14] が提案されている。前者はユーザの使用分野・目的に合った2言語コーパスの入手が難しいという問題があり、後者では如何にして訳語選択の精度を上げるかという点が問題になる。

本稿では目的言語の単言語コーパスを用いて高精度の訳語学習を実現するためのアルゴリズムを提案する。また、本方式を基に実験モデルを構築し、そのモデルに対しての実験を通して有効性の検証を行う。

2. 訳語学習

2.1 概要

ユーザはあらかじめ目的言語の任意の文書を単言語コーパスとして保持しておく。これを利用することによって、ユーザの嗜好や分野・使用目的に合わせた訳語学習が可能となる。本稿では翻訳方向を「英→日」とし、訳語学習の対象単語の品詞を「名詞」に限定する。

本方式による訳語学習を以下のように定義する:

> **訳文中の名詞に対して複数の訳語候補が存在する場合、各訳語候補に対して単言語コーパス中での統計的情報に基づく評価基準を満たす訳語候補を優先して選択し、ユーザに適した訳語として採用する。**

例えば、次の文の翻訳において宇宙分野の単言語コーパスを用いた訳語学習を行う場合を考える:

> *In 1978 the United States launched the Pioneer Venus mission.*

まずこれを通常に機械翻訳すると以下の訳文が得られる:

> 1978年には、アメリカが初期のビーナス・ミッションを打ち上げました。

原文中の "Venus" には「ビーナス」「金星」「美人」「ウェヌス」「ヴィーナス」の訳語候補が存在する。これらの訳語候補の中で最適な訳語は、その文中での意味だけでなくユーザの嗜好や分野、使用目的に依存して決まるものである。本稿ではこれらの訳語候補について、単言語コーパス中での統計的情報に基づく評価基準で判定し、最適と思われる訳語を決定する。例えば「金星」という訳語が他の訳語候補よりも極めて多く出現していたり、その他の訳語との共起強度が大きい場合は、"Venus" の訳語を「金星」に決定する。従って訳語学習により訳語が自動的に変更され、以下のような結果になる:

> 1978年には、アメリカが初期の金星ミッションを打ち上げました。

太字部分が修正された箇所である。

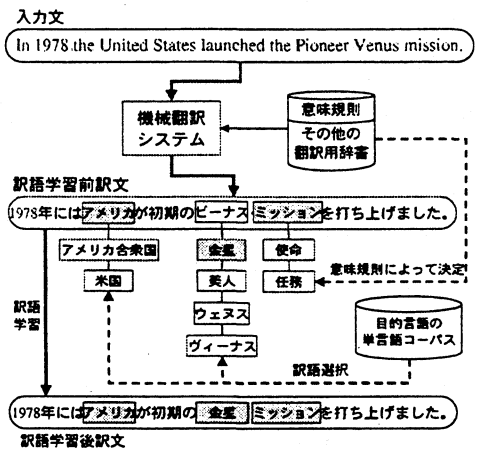

図 1:単言語コーパスを用いた訳語学習

このように、ユーザがわざわざ "Venus" の訳語を「金星」に変更して訳語の学習作業をすることなく、単言語コーパスを用いて最適な訳語が自動的に選択される。

上記の例では、"mission" の訳語「ミッション」は意味規則によって決められたものであるのに対し、"Venus" の訳語は意味規則によって決定されず標準的な訳語「ビーナス」が当てられただけであった。従って、厳密に決定された訳語「ミッション」の方が「ビーナス」よりも確実な訳語であり、"mission" に比べ "Venus" の訳語は選択し直す余地があると考えられる。本稿では、この例の "mission" のように訳語が確定的に決定する名詞を訳語確定名詞、その訳語を確定訳語とし、それ以外の、即ち上記の例での "Venus" のように訳語を選択し直す余地がある名詞を訳語選択対象名詞とする。そして訳語選択対象名詞の訳語候補の中から、各訳語候補のコーパスでの統計的情報を元に、最適な訳語を選択する問題を考える。

2.2 訳語学習アルゴリズム

ここではまず訳語学習の具体的な流れについて説明する。

訳語学習アルゴリズムは機械翻訳の後処理として機能する。従って、機械翻訳処理が一旦終了している状態を考える。この状態から以下の流れで訳語学習を行う:

(1) 訳文中の全ての名詞を抽出する。

(2) 訳語が意味的規則によって決定しているものは、その訳語を確定訳語リストに格納する。

(3) (2) 以外の名詞について、その訳語 (key) と全訳語候補 (value) をセットにして訳語選択対象アレイに格納する。

(4) 訳語選択対象アレイのある key に対して value として登録してある全訳語候補の中から、訳語選択

アルゴリズムによって最もふさわしい訳語を選択する。これをすべての key に対して繰り返す。

(5) 選択された訳語は、そのときの訳語選択対象アレイの key から変更された「新たな訳語」として翻訳結果の文に反映される (変更されない場合もある)。

(4) の訳語選択アルゴリズムとは、確定訳語リストと訳語選択対象アレイが与えられたときに各訳語選択対象名詞の最適な訳語を選択するアルゴリズムである。次章では、この訳語選択アルゴリズムで用いる訳語選択基準について説明する。

3. 訳語選択基準

本稿では、コーパスからの統計的情報の一つとして共起情報を用いる。そこで、共起している文のカウント法と共起の強度の評価基準について説明する。

3.1 重み付き共起文数の定義

コーパスから共起する単語対を取得する場合、「どのような場合に共起する単語対と見なすか」を始めに決定しなければならない。この共起取得法として、一文内に同時に出現する単語全てを共起単語とする一文内共起を採用する。一文内共起を採用した理由は、

- アルゴリズム設計が容易であり、実装に有利であるため。
- 依存単語対を取りこぼすことがないため。

などによる。今回の手法では学習対象の単語の品詞を「名詞」に限定しているので、依存関係のある単語対は互いに離れている可能性があり、特に後者の理由は重要である。

ここではまず、上記の定義を基に本稿で用いる共起のカウント法について述べる。

原文書中の文 T_S の翻訳に対して、T_S 中の訳語選択対象名詞の集合を $\{W_{S1}, W_{S2}, \cdots, W_{Sr}\}$ で表す。$W_{Si}(1 \leq i \leq r)$ の訳語候補の数を n_i とし、訳語候補をそれぞれ $w_{ij}(1 \leq j \leq n_i)$ とする。W_{Si} の訳語として最適なものを W_{Ti} とする。一文中での確定訳語を $W_C = \{\tilde{w}_1, \tilde{w}_2, \cdots, \tilde{w}_L\}$ とおく。ここでの「確定訳語」は、実際にはルールによって決定された訳語や他の評価基準によって既に決定された訳語などに相当する。本稿では訳語選択対象名詞の各訳語候補に対して、これらの名詞との共起を調べることで最適な訳語を決定する。

また、コーパス中の文を T_m とし、コーパスを $D_M = \{T_1, T_2, \cdots, T_M\}$ で表す。M はコーパス中の全文数を表す。

訳語候補 w_{ij} と確定訳語 $\tilde{w}_l(1 \leq l \leq L)$ が共起する文数を $C_{w_{ij},\tilde{w}_l}$ とする。また、文 T_m 中に存在する単語対 $(w_{ij}, \tilde{w}_l)$ の数を $P_{T_m,w_{ij},\tilde{w}_l}$、コーパス D_M 内での単語対 $(w_{ij}, \tilde{w}_l)$ の数を $P_{w_{ij},\tilde{w}_l}$ とする。即ち:

$$P_{T_m,w_{ij},\tilde{w}_l} \equiv \min\{P_{T_m,w_{ij}}, P_{T_m,\tilde{w}_l}\} \quad (1)$$

$$P_{w_{ij},\tilde{w}_l} \equiv \sum_{m=1}^{M} P_{T_m,w_{ij},\tilde{w}_l} \quad (2)$$

$P_{T_m,w_{ij}}$、$P_{T_m,\tilde{w}_l}$ はそれぞれ文 T_m における w_{ij} と $\tilde{w}_l$ の出現数を表す。

定義により、一文中に単語対が存在する場合にそれらの単語は共起していると認定するが、文によってはその単語対が複数ある場合も存在する。単語対が 1 つしかない文と複数ある文とを比較すると、複数存在する文の方がその単語対がより強く共起しているということができる。従って、本稿では共起している文の中での単語対の出現数に応じて重みをつけて共起を数えることにする。

まず w_{ij} と $\tilde{w}_l$ とが共起している文を抽出し、それらを $T_{k,w_{ij},\tilde{w}_l}(1 \leq k \leq C_{w_{ij},\tilde{w}_l})$ とおく。ここで、文 $T_{k,w_{ij},\tilde{w}_l}$ の中での単語対 $(w_{ij}, \tilde{w}_l)$ の出現数に応じて重み $\Gamma_{T_{k,w_{ij},\tilde{w}_l}} \geq 1$ を設け、重み付き共起文数 $S_{w_{ij},\tilde{w}_l}$ を以下のように定義する:

$$S_{w_{ij},\tilde{w}_l} \equiv \sum_{k=1}^{C_{w_{ij},\tilde{w}_l}} \Gamma_{T_{k,w_{ij},\tilde{w}_l}} \quad (3)$$

ここで $1 \leq \Gamma_{T_{k,w_{ij},\tilde{w}_l}} \leq P_{T_{k,w_{ij},\tilde{w}_l},w_{ij},\tilde{w}_l}$ となるように $\Gamma_{T_{k,w_{ij},\tilde{w}_l}}$ を決めると、$C_{w_{ij},\tilde{w}_l} \leq S_{w_{ij},\tilde{w}_l} \leq P_{w_{ij},\tilde{w}_l}$ という関係が成立する。本稿では、重み $\Gamma_{T_{k,w_{ij},\tilde{w}_l}}$ を $P_{T_{k,w_{ij},\tilde{w}_l},w_{ij},\tilde{w}_l} = 1$ のとき 1、$P_{T_{k,w_{ij},\tilde{w}_l},w_{ij},\tilde{w}_l} = 2$ のとき 1.2、$P_{T_{k,w_{ij},\tilde{w}_l},w_{ij},\tilde{w}_l} \geq 3$ のとき 1.5 に設定した。

3.2 共起強度

次に共起強度の評価基準について述べる。これは総重み付き相互情報量と平均相互情報量の線形結合の関数を用いる。

3.2.1 総重み付き相互情報量

始めに、相互情報量 $I(w_{ij} : \tilde{w}_l)$ による共起強度の評価基準を定義する:

$$I(w_{ij} : \tilde{w}_l) = \log_2 \frac{P(w_{ij}, \tilde{w}_l)}{P(w_{ij}) \cdot P(\tilde{w}_l)} \quad (4)$$

式 (4) の $P(w_{ij}, \tilde{w}_l)$、$P(w_{ij})$、$P(\tilde{w}_l)$ はそれぞれ w_{ij} と $\tilde{w}_l$ が共起する確率、w_{ij} の出現確率、$\tilde{w}_l$ の出現確率を表している。これらの確率は一般に未知であるので、式 (3) の重み付き共起文数を用いた次の推定値を用いることにする:

$$\hat{P}(w_{ij}, \tilde{w}_l) = \frac{S_{w_{ij},\tilde{w}_l}}{M} \quad (5)$$

$$\hat{P}(w_{ij}) = \frac{\sum_{m=1}^{M} P_{T_m,w_{ij}}}{M} \quad (6)$$

$$\hat{P}(\tilde{w}_l) = \frac{\sum_{m=1}^{M} P_{T_m,\tilde{w}_l}}{M} \quad (7)$$

このとき、式 (4) の相互情報量 $I(w_{ij} : \tilde{w}_l)$ の推定値 $\hat{I}(w_{ij} : \tilde{w}_l)$ は

$$\hat{I}(w_{ij} : \tilde{w}_l) = \log_2 \frac{S_{w_{ij},\tilde{w}_l} \cdot M}{\sum_{m=1}^{M} P_{T_m,w_{ij}} \cdot \sum_{m=1}^{M} P_{T_m,\tilde{w}_l}} \quad (8)$$

となる。この相互情報量を重み付き相互情報量と呼ぶことにする。また、重み付き相互情報量 $\hat{I}(w_{ij} : \tilde{w}_l)$ の W_C に関する和、即ち $\sum_{l=1}^{L} \hat{I}(w_{ij} : \tilde{w}_l)$ を総重み付き相互情報量と呼ぶことにし、$WI(w_{ij} : W_C)$ で表すことにする:

$$WI(w_{ij} : W_C) = \sum_{l=1}^{L} \hat{I}(w_{ij} : \tilde{w}_l) \quad (9)$$

この総重み付き相互情報量は w_{ij} と W_C の依存関係の大きさを表している。

3.2.2 平均相互情報量

次にエントロピーを用いた共起強度の評価基準について述べる。確定訳語 W_C を知った上での訳語候補 w_{ij} に対する平均情報量 (条件付エントロピー) $H(w_{ij}|W_C)$ は、

$$H(w_{ij}|W_C) = -\sum_{l=1}^{L} P(w_{ij}, \tilde{w}_l) \log_2 P(w_{ij}|\tilde{w}_l) \quad (10)$$

ここで、$P(w_{ij},\tilde{w}_l)$ の推定値 $\hat{P}(w_{ij},\tilde{w}_l)$ は式 (5) で与える。また、条件付確率 $P(w_{ij}|\tilde{w}_l)$ の推定値 $\hat{P}(w_{ij}|\tilde{w}_l)$ は式 (5)、(7) を用いて、次式で与える:

$$\hat{P}(w_{ij}|\tilde{w}_l) = \frac{\hat{P}(w_{ij},\tilde{w}_l)}{\hat{P}(\tilde{w}_l)} = \frac{S_{w_{ij},\tilde{w}_l}}{\sum_{m=1}^{M} P_{T_m,\tilde{w}_l}} \quad (11)$$

以上から、平均情報量 $H(w_{ij}|W_C)$ の推定値 $\hat{H}(w_{ij}|W_C)$ は、式 (10)、(5)、(11) より

$$\hat{H}(w_{ij}|W_C) = -\sum_{l=1}^{L} \frac{S_{w_{ij},\tilde{w}_l}}{M} \log_2 \frac{S_{w_{ij},\tilde{w}_l}}{\sum_{m=1}^{M} P_{T_m,\tilde{w}_l}} \quad (12)$$

となる。また、w_{ij} の平均情報量 (エントロピー)$H(w_{ij})$ は

$$H(w_{ij}) = -P(w_{ij}) \log_2 P(w_{ij}) \quad (13)$$

であり、$H(w_{ij})$ の推定値 $\hat{H}(w_{ij})$h は式 (6) を用いて

$$\hat{H}(w_{ij}) = -\frac{\sum_{m=1}^{M} P_{T_m,w_{ij}}}{M} \log_2 \frac{\sum_{m=1}^{M} P_{T_m,w_{ij}}}{M} \quad (14)$$

とする。以上より、平均相互情報量 $EI(w_{ij} : W_C)$ は、式 (12) と (14) を用いて

$$EI(w_{ij} : W_C) = \hat{H}(w_{ij}) - \hat{H}(w_{ij}|W_C) \quad (15)$$

で与えられる。この平均相互情報量は W_C を知っている上での w_{ij} の曖昧さの減少量を表している。

3.2.3 共起強度による訳語選択基準

式 (9)、(15) を用いて次のような評価式 $\epsilon[w_{ij}]$ を定義する:

$$\epsilon[w_{ij},W_C] \equiv \alpha WI(w_{ij} : W_C) + \beta EI(w_{ij} : W_C) \quad (16)$$

この $\epsilon[w_{ij},W_C]$ を用いて訳語を決定する場合、

$$W_{T_i} = arg\max_{w_{ij}} \epsilon[w_{ij},W_C] \quad (17)$$

で決定する。評価基準 $\epsilon[w_{ij}]$ の第 1 項目は総重み付き相互情報量に関する評価基準であり、第 2 項は平均相互情報量に関する評価基準になっている。このとき係数 α と β の値を変えることによってどちらの評価基準を重視するかを決めることができる。係数 α や β は経験的に決められ、今回は $\alpha = 0.5$、$\beta = 600$ とした。

3.3 出現頻度による訳語選択基準

さらに、別の訳語決定法として区間推定を用いた手法も用いる。W_{S_i} の訳語候補 $w_{ij}(1 \leq j \leq n_i)$ のコーパス中での出現数が最も大きい名詞を w'、その出現数を n' とし、2 番目に大きい名詞を w''、その出現数を n'' とする。ここで、

$$\ln \frac{n'+0.5}{n''+0.5} \geq \theta_{confidence} + Z_{1-\alpha}\sqrt{\frac{1}{n'+0.5} + \frac{1}{n''+0.5}} \quad (18)$$

が成り立つとき、またそのときに限り W_{S_i} の訳語 W_{T_i} を w' に決定する [1]。本稿では閾値 $\theta_{confidence} = 0.2$ とし、$Z_{1-\alpha} = 1.04$(信頼度は 85.083%) に設定した。

4. 訳語選択アルゴリズム

2.2 節の (4) での訳語選択アルゴリズムは以下のようにした:

原文書中の文 T_S に対して、訳語選択対象アレイの key の集合を $W_U = \{W_{S_1},\cdots,W_{S_r}\}$ とし、W_{S_i} の value を $\{w_{i1},\cdots,w_{in_i}\}$ とする。意味規則によって決定した確定訳語を $W_C = \{\tilde{w}_1,\cdots,\tilde{w}_L\}$ とおく。W_{S_i} の訳語として選択されるものを W_{T_i} とする。

Step1: [一意に決定する訳語を確定訳語に追加] 訳語候補が一つしかないものを W_C に追加し、これを $W_C{}'$ とする。

Step2: $W_U = \phi$ ならば終了。

Step3: [エントロピー上位 2 語を確定訳語に追加] 次のいずれかの条件:

(a) $W_C{}' = \phi$

(b) $W_C{}' \neq \phi$ かつ $\forall\tilde{w}_l \in W_C{}', \forall T_m \in D_M, P_{T_m,\tilde{w}_l} = 0$

(c) $W_C{}' \neq \phi$ かつ $\exists\tilde{w}_l \in W_C{}', \exists T_m \in D_M, P_{T_m,\tilde{w}_l} \neq 0$ かつ $|W_C{}'| < 6$

が成立するならば、式 (14) よりエントロピー $\hat{H}(w_{ij})$ が

$$\hat{H}(w_{ij}) > \theta_{entropy}$$

を満たす w_{ij} の中で、$\hat{H}(w_{ij})$ を最大にするものを $\tilde{w}_{L+1}$、2 番目に大きくするものを $\tilde{w}_{L+2}$ として $W_C{}'$ に追加し、訳語選択対象アレイから対応する key のエントリを削除する。ここで $\theta_{entropy} = 0.002$ とする。修正された確定訳語リストを $W_C{}''$ とする。

Step4: [確定訳語を含む *Step1,Step3* での確定訳語との共起による訳語決定] 修正された訳語選択対象アレイの key の集合を $W_U{}' = \{W_S{}'_1,\cdots,W_S{}'_{r'}\}$ とし、$W_S{}'_h$ の value を $\{w'_{h1},\cdots,w'_{hn'_h}\}$ とする。また、$W_S{}'_h$ の訳語として選択されるものを $W_T{}'_h$ としたとき

$$W_T{}'_h = arg\max_{w'_{hk}} \epsilon[w'_{hk},W_C{}'']$$

によって訳語を決定する。

Step5: [信頼区間を利用した訳語決定] *Step4* で決定しなかった訳語選択対象名詞を $\{W_S{}''_1,\cdots,W_S{}''_{r''}\}$ とする。$W_S{}''_u(1 \leq u \leq r'')$ に対して、コーパス中での出現数が多い上位 2 語の訳語候補を w'、w'' とし、そのときの出現数をそれぞれ n'、n'' とする。$W_S{}''_u$ の訳語として選択されるものを $W_T{}''_u$ とする。このとき

$$W_T{}''_u = w' \quad \text{if} \quad \text{式 (18) が成立}$$

によって訳語を決定する。

Step6: [確定訳語を含む全確定訳語との共起による訳語決定] *Step4*、*Step5* で決定した訳語を全て $W_C{}''$ に追加し、これを $W_C{}'''$ とする。

まだ決定していない訳語選択対象名詞を $\{W_S{}'''_1,\cdots,W_S{}'''_{r'''}\}$ とし、$W_S{}'''_v$ の訳語候補を $\{w'''_{v1},\cdots,w'''_{vn'''_v}\}$ とする。また、$W_S{}'''_v$ の訳語として選択されるものを $W_T{}'''_v$ としたとき

$$W_T{}'''_v = arg\max_{w'''_{vy}} \epsilon[w'''_{vy},W_C{}''']$$

によって決定する。

Step7: 以上で決まらなかったときは元の訳語のままにする。(終了)

5. 実験と評価

学習用の目的言語の単言語コーパスは、主に NASDA(宇宙開発事業団) の HP [13] で収集した『宇宙』に関する文書 (総文数 8224 文) から作成した品詞タグ付きコーパスと、インターネットで収集した『F1』に関する文書 (総文数 10018 文) から作成した品詞タグ付きコーパスの 2 つを用いる。

またそれぞれのコーパスを用いたときのテスト用文書には、主に NASA やチャンドラ X 線探査、SEC の HP [4, 7, 9] などから収集した宇宙に関する文書 (全 201 文) と、主に F1 のニュースや技術的な説明の多い HP [3, 6] などから収集した F1 に関する文書 (全 198 文) を用いて訳語学習の実験を行った。

ここでは訳語学習を行う前後で全ての名詞に対して表 1 の 6 段階で評価を行う。表 1 中の"Right"は「正しい、あるいは妥当な訳語が選択された」ことを表し、"Wrong"は「不適切な訳語が選択された」ことを表す。

客観的に評価するために次の指標を与える:

$$\text{(品質向上率)} = \frac{\text{(正解名詞数)} - \text{(不正解名詞数)}}{\text{名詞数}}$$

$$\text{(改善率)} = \frac{\text{(訳語が変更され正解になった名詞数)}}{\text{名詞数}}$$

評価	訳語学習前	訳語学習後	訳語の変更あり
Unchanged_Right	Right	Right	×
Wrong → Right	Wrong	Right	○
Right → Right	Right	Right	○
Right → Wrong	Right	Wrong	○
Wrong → Wrong	Wrong	Wrong	○
Unchanged_Wrong	Wrong	Wrong	×

表 1: 訳語学習前後の名詞の評価カテゴリ

分野	宇宙		F1	
モード	Normal	Collo	Normal	Collo
文数	201		198	
名詞数	2047		1678	
変更された名詞数	385	387	485	488
Unchanged_Right	1580	1586	1145	1142
Wrong → Right	118	125	173	189
Right → Right	181	180	209	216
Right → Wrong	66	62	76	64
Wrong → Wrong	20	20	27	19
Unchanged_Wrong	82	74	48	48

表 2: 名詞毎の評価結果

分野	宇宙		F1	
モード	Normal	Collo	Normal	Collo
品質向上率	0.8359	0.8476	0.8200	0.8439
改善率	0.1460	0.1490	0.1990	0.2414
Applicability	0.9179	0.9238	0.9100	0.9219
Precision	0.7766	0.7881	0.7876	0.8299
有意確率 P_0	$7.75E^{-5}$	$2.38E^{-6}$	$3.61E^{-10}$	$8.29E^{-16}$

表 3: 宇宙分野と F1 分野での各指標値

$$\textbf{(Applicability)} = \frac{\text{(正解名詞数)}}{\text{名詞数}}$$

$$\textbf{(Precision)} = \frac{\text{(訳語が変更され正解になった名詞数)}}{\text{(変更された名詞数)}}$$

また、あらかじめ学習用コーパスから複合語と思われるものを抽出し、それぞれの分野における複合語候補リストを作成しておく。本実験では今まで述べてきた訳語学習アルゴリズム (Normal モード) の *step1* において、翻訳結果の文において名詞が 2 つ以上並んでいる場合、各訳語候補の組み合わせの内複合語候補リストに登録されているものを優先的に選択し、それらを確定訳語に追加するようにしたアルゴリズム (Collo モード) についても評価した。結果を表 2、3 に示す。いずれも Precision は 77%以上であり、高い訳語学習精度であるといえる。特に Collo モードの場合 Precision は大きく上昇していることから、本稿のアルゴリズムにコーパスから抽出した複合語に関する情報を取り入れるのは大きな効果があるといえる。

さらにそれぞれのモードにおける符号検定の有意確率は表 3 に示す通りである。いずれの場合も信頼度 **99.9999%**以上で翻訳品質に差が出たといえるため、本訳語学習方式は有効であると考えられる。

6. まとめ

本稿では、訳語候補が複数存在する場合に、ユーザの嗜好や分野・使用目的に合った訳語選択を行う訳語学習方式を提案した。本方式では、各訳語候補と確定訳語の目的言語の単言語コーパス中での共起の強度を、総重み付き相互情報量と平均相互情報量を組み合わせた関数 ϵ を用いた評価基準で判定し訳語を決定する。また、宇宙分野と F1 分野の学習用単言語コーパスとそれぞれの分野のテスト文書を用いて訳語学習の実験を行った。その結果、宇宙分野では複合語候補リストを用いない場合、Applicability は 91.8%、Precision は 77.7%であった。また名詞改善率は 14.6%であることから、本アルゴリズムでは実験用のテスト文書中の全名詞の約 15% において、正しい訳語への変更を行うということが分かった。さらに複合語候補リストを用いた場合、Applicability は 92.4%、Precision は 78.8%に上昇した。F1 分野では Applicability は 91.0%、Precision は 78.8%であった。また名詞改善率は 19.9%であることから、本アルゴリズムでは実験用のテスト文書中の全名詞の約 20% において、正しい訳語への変更を行うということが分かった。さらに複合語候補リストを用いた場合、Applicability は 92.2%、Precision は 83.0%に上昇した。さらに符号検定を行い本訳語学習方式の有効性を検証した。

7. 今後の課題

実用化に関しては、Precision の目標を 80%以上と考えている。そのためには、以下を考慮に入れる必要がある:

- n-gram による共起認定——今回は一文中に単語対が存在した場合に共起していると認定しているが、この手法では依存関係のない単語対を多くとる傾向があり精度が高くない。そこで n-gram で共起単語を抽出すれば、依存関係のない単語対を少なくすることができ、共起認定の精度が向上すると考えられる。
- 訳語選択対象名詞の訳語どうしの共起——今回の手法では、共起は「確定訳語」と「訳語選択対象名詞の各訳語候補」との間の共起のみを考え、訳語を決定していた。しかし訳語を決定する際にはその前後の訳語との共起についても考慮して訳語を決定した方がよい。これは複合語を翻訳する場合等に非常に有効であると考えられる。従って、確定訳語との共起だけでなく訳語選択対象名詞の訳語どうしの共起情報も考慮して訳語選択を行う。

参考文献

[1] Ido Dagan and Alon Itai. Word Sense Disambiguation Using a Second Language Monolingual Corpus. *Association for Computational Linguistics*, pp. 563–596, 1994.

[2] Shinichi Doi and Kazunori Muraki. Evaluation of DMAX Criteria for Selecting Equivalent Translation based on Dual Corpora Statistics. *TMI'93*, pp. 302–311, 1993.

[3] F1MECH.COM. http://www.f1mech.com/.

[4] NASA ホームページ. http://www.nasa.gov/.

[5] Fung Pascale. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. *Proceedings of the 33rd Annual Conference of the Association for Computational Linguistics*, 1995.

[6] Planet-F1. http://www.planet-f1.com/.

[7] The Chandra X ray Observatory Center. http://chandra.harvard.edu/.

[8] H. Ney S. Niessen, S. Vogel and C. Tillman. A dp based search algorithm for statistical machine translation. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pp. 960–966, 1998.

[9] SEC ホームページ. http://www.sec.noaa.gov/.

[10] Kumiko TANAKA and Hideya IWASAKI. Extraction of Lexical Translations from Non-Aligned Corpora. *Proceedings of the 16th International Conference on Computational Linguistics(COLING)*, pp. 580–585, 1996.

[11] Ye-Yi Wang and Alex Waibel. Decoding algorithm in statistical machine translation. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pp. 366–372, 1997.

[12] Gale William and Kenneth Church. A Program of Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, No. 19(1), pp. 75–102, 1993.

[13] 宇宙開発事業団ホームページ. http://www.nasda.go.jp/.

[14] 野上宏康, 熊野明ほか. 既存目的言語文書からの訳語の自動学習方式. 情報処理学会第 42 回全国大会講演論文集, Vol. 3, pp. 29–30, 1991.