統計情報を用いた対訳単語辞書の作成

大森 久美子　堤 純也　中西 正和
慶應義塾大学理工学部 数理科学科

1 はじめに

機械翻訳システムの質は，そのシステムが用いる対訳辞書に大きく依存する．

これまで対訳辞書は，既存の辞書を計算機上にコピーしたものや，人間の手により作成したものが用いられてきた．しかし，人間の手による対訳辞書の作成は莫大な労力を要する上，作業ミスも含まれやすい．

そこで，本研究では統計情報を用いた P. F. Brown [1] の手法，及び北村 [2] の手法をもとに，仏英間の対訳単語辞書の自動作成を行い，両者の問題点を解決する手法を提案する．その際，対訳コーパスから得られる単語間の相互情報量，及び単語間の類似度を用いる．

2 本システムの背景

以下の二つの研究では，用いる対訳コーパスに対し次の二つを仮定している．

- 各対訳単語間に一対一対応がついている．
- 各文の間に一対一対応がついている．

2.1 Brown の手法

Brown は仏英単語間の相互情報量を用いた対訳辞書作成を提案している．

任意の仏単語 f と英単語 e_j の結び付きの強さを表す相互情報量は，任意の仏単語 f が英単語 e_j に訳される確率 $P(e_j|f)$ と，あるランダムに選び出された仏単語 f が英単語 e_j に訳される確率 $P(e_j)$ を用いて，式 (1) のように定義される．

$$MI(e_j, f) = \log_2 \frac{P(e_j|f)}{P(e_j)} \tag{1}$$

この $MI(e_j, f)$ の値を最大にする e_j が f の訳語と考えられる．ここで式 (1) に現れる $P(e_j|f)$, 及び $P(e_j)$ を求める手順を以下に述べる．

まず始めに，任意の仏単語 f が出現する仏文の対訳文に，英単語 e_j が出現する回数 $C(e_j, f)$ を以下の手順に従って求める．

1. 仏単語 f と英語のすべての語彙 e_j に対し，$C(e_j, f) = 0$ とセットする．
2. 仏単語 f が出現する仏文に対応する英文が，n 個の単語から成る $E = e_{j_1}\ e_{j_2}\ \cdots\ e_{j_n}$ である時，$C(e_{j_1}, f), C(e_{j_2}, f), \cdots, C(e_{j_n}, f)$ を $1/n$ 増やす．
3. すべての仏単語に対し，2 を繰り返し $C(e_j, f)$ を求める．

この手順で求めた $C(e_j, f)$ を用いて，$P(e_j|f)$ は，仏文コーパスの全単語数を M_f, 仏文コーパスに仏単語 f が出現する回数を $M(f)$ とすると，式 (2) のように表される．

$$P(e_j|f) = \frac{C(e_j, f)}{M(f)} \tag{2}$$

また，$P(e_j)$ は，式 (3) のように表される．

$$\begin{aligned} P(e_j) &= \sum_f \frac{P(e_j|f)}{P(f)} \\ &= \sum_f \frac{P(e_j|f)\ M(f)}{M_f} \end{aligned} \tag{3}$$

ここで，$C(e_j, f)$ を上記のように 1 回出現するごとに

1/(その英文の単語数)

ずつ増加させているのは，対訳文の単語数が多いほどそれを構成している英単語の一つが，仏単語 f の対訳になる確率は低いためである．

しかし，必ず仏英間で一対一に単語が対応しているとは限らないことから，この方法では熟語を構成しているそれ自身の意味を表さない単語や，訳されない単語について良い結果が得られないと考察している．

2.2 北村の手法

北村は，式 (4) から算出される単語間の類似度を用いた対訳単語対抽出を提案している．式 (4) は，仏英両コーパスの各単語に対して，その単語が出現する文の対訳文に出現する，仏単語，英単語の組合せが全コーパス

中にどれくらいの頻度で出現するかを示している．C_e は英文コーパスに英単語 w_e が出現する回数，C_f は仏文コーパスに仏単語 w_f が出現する回数，C_{ef} は仏単語 w_f が出現する仏文の対訳文に，英単語 w_e が出現する回数を表す．

$$\begin{cases} sim1(w_e, w_f) = \dfrac{2C_{ef}}{C_e + C_f} \\ sim2(w_e, w_f) = \dfrac{C_{ef}}{C_f} \\ sim3(w_e, w_f) = \dfrac{C_{ef}}{C_e} \end{cases} \quad (4)$$

この単語間の類似度を用いた手法では，類似性の高い仏英単語対のみに限定するため，式 (4) の $sim1$，及び C_{ef} の値が，

$$C_{ef} \geq 2 \text{ かつ } sim1(w_e, w_f) \geq 0.15$$

の条件を満たす単語対のみをデータとして採用する．また，採用する単語間の類似度 $sim(w_e, w_f)$ の値は $sim1$, $sim2$, $sim3$ の最大値とする．この理由は，単語間の類似度を $sim1$ と定義した場合，片方の言語にしか多義性がない場合でも，その多義性が類似度の値に反映されてしまい類似度の値が低下してしまうためである．

3 本システムの構成

本システムでは Brown の手法，及び北村の手法をもとに，仏英対訳コーパスから得られる統計情報のみを用いて仏英間の対訳単語辞書の自動作成を行う．

本研究においては，約 50,00 文から成る *De la terre à la lune* [3] (総単語数 39,983 語)，及び *From the Earth to the Moon* [4] (総単語数 56,463 語) のうち 805 文をコーパスとして用いた．

805 文は，

- 仏文…総単語数 6,733 語，語彙数 2,137 語
- 英文…総単語数 6,772 語，語彙数 2,000 語

から成る．

3.1 対訳単語辞書作成の手順

まず始めに，辞書作成の手順を以下に述べる．

1. 2 節で述べた対訳コーパスに対する前提条件より，対訳単語対抽出のための前準備として，仏英両コーパスを文単位で単語に切り分ける．
2. 1 の結果，得られるコーパスのすべての仏単語に対し，その仏文の対訳文に出現する，英単語との統計量を計算する．
3. 英単語それぞれに対し結び付きの強い仏単語を序列化する．
4. 英単語に対し，用いた対訳コーパスをもとに「正しい」と思われる対訳仏単語が何番目に現れるかを評価する．

辞書作成の手順を図 1に示す．

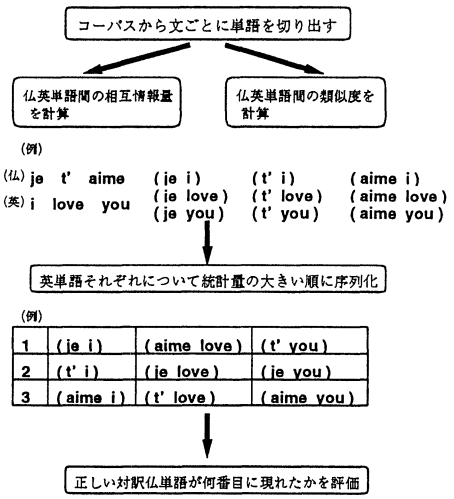

図 1: 対訳辞書作成の手順

3.2 Brown の手法

2.1 節において述べた Brown の手法を用いて仏英対訳単語対を抽出する．

3.3 Brown の手法の改良

Brown の手法において，式 (1) の分子 $P(e_j|f)$ が大きい時ほど $MI(e_j, f)$ の値は大きくなる．さらに式 (2) により，この $P(e_j|f)$ の値は，仏単語の出現回数 $M(f)$ が小さい時ほど大きくなることがわかる．つまり，式 (1) を用いて求めた相互情報量は出現回数の少ない，特に出現回数が 1 回の仏単語との結び付きが上位にきてしまう．

そこで，以下の 3 通りの計算方法で相互情報量を再度計算する．

A. 出現回数が1回の英単語についてはBrownの手法を用いる．2回以上出現する英単語については，1回しか出現しない仏単語と結び付くことはないと仮定して，対訳文中の出現回数が1回の仏単語との相互情報量は計算しない．

B. 出現回数が1回の英単語については，手法Aと同様にBrownの手法を適用する．2回以上出現する英単語については，「正しい」仏単語との組合わせは2回以上出現すると仮定して，1回しか出現しない仏英単語対の相互情報量は計算しない．

C. 出現回数が10回未満の単語はデータとして与える影響が少ないと思われるので，仏英ともに出現回数が10回未満の単語についてはそれらの対訳単語は抽出せずに，出現回数が10回以上の単語に対してのみBrownの手法を用いる．

3.4 北村の手法

2.2節において述べた北村の手法を用いて，仏英対訳単語対を抽出する．

4 実験結果

仏文コーパスに出現するすべての仏単語に対し，その仏単語が出現する仏文の対訳文に出現する，英単語との相互情報量，及び類似度を計算し，英単語それぞれに対し結び付きの強い仏単語を序列化した．その後，それぞれの英単語に対し，用いた対訳コーパスをもとに「正しい」と思われる対訳仏単語が何番目に現れるかを評価した．

4.1 Brownの手法

英単語2,000語それぞれに対し，その英単語との相互情報量の値が大きい仏単語を順に並べ，1位，3位，5位までに「正しい」対訳仏単語が現れた英単語がどれくらいあったかを評価したところ表1のような結果になった．表1の各データは，英単語2,000単語を全体とした場合の正解の割合を示す．

表1: Brownの手法による実行結果

	単語数	1位	3位まで	5位まで
Brownの手法	2,000	23.2 %	58.0 %	73.7 %

4.2 Brownの手法の改良

3.3節において述べたように，2回以上出現した英単語については手法A, B, Cを用いて再度相互情報量を計算をした．

Brownの手法と比較してみると表2のような結果になった．

表2: 「正しい」対訳仏単語の出現順位

	出現回数	Brown	手法A	手法B	手法C
with	46	46位	3位	1位	1位
the	601	101位	98位	89位	3位

すべての英単語に対して評価したところ，表3のような結果になった．表中の単語数はデータを採取することが出来た英単語の数を表す．従って，各手法の正解率は表中の単語数を全体とした場合，「正しい」対訳仏単語をその順位までに持つ英単語の割合を示す．

表3: Brownの手法との比較

	単語数	1位	3位まで	5位まで
Brown	2,000	23.2 %	58.0 %	73.7 %
手法A	2,000	33.5 %	51.6 %	63.3 %
手法B	1,946	35.6 %	63.0 %	73.5 %
手法C	79	39.2 %	49.4 %	54.4 %

ここで手法A, Bは出現回数が1回の英単語についてはBrownの手法を用いている．そこで出現回数が2回以上の英単語についてのみ，各手法を用いて相互情報量を計算したところ表4のような結果が得られた．表2と同様に，表中の単語数はデータを採取することが出来た英単語数を表し，各手法の正解率は表中の単語数を全体とした場合の正解の割合である．

表4: 出現回数が2回以上の英単語についての結果

	単語数	1位	3位まで	5位まで
Brown	697	18.5 %	47.2 %	63.1 %
手法A	697	43.6 %	60.6 %	65.4 %
手法B	643	50.2 %	59.5 %	61.7 %

図2は，2回以上出現する英単語に対するBrownの手法と，手法A, Bの結果の比較をグラフに表したものである．図2において棒グラフはその順位に「正しい」

の対訳仏単語が現れた英単語の個数を，折れ線は1位からその順位までに「正しい」対訳仏単語が現れた英単語の個数を表している．

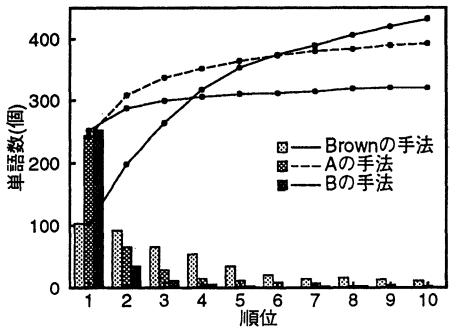

図 2: 出現回数が2回以上の英単語についての結果

4.3 北村の手法

北村の手法を用いて，仏英単語間の類似度を計算し，Brown の手法と同様に評価したところ，表5のような結果が得られた．この手法においてデータを採取することが出来た英単語は，全体の 28.9 % (2,000 単語中 578 単語) であった．表5の各データは，英単語 578 単語を全体とした場合の正解の割合を示す．

表 5: 北村の手法による実行結果

	単語数	1位	3位まで	5位まで
北村の手法	578	46.2 %	65.0 %	66.5 %

5 実験結果の評価

- 出現回数が2回以上の英単語については Brown の手法は適さない．
- Brown の手法，及び手法 A はすべての英単語について対訳仏単語を抽出することが可能である．
- 出現頻度が非常に高い英単語については，表2より手法 C が有効である．しかし，手法 C は出現回数が 10 回以上の単語のデータしか採ることが出来ない．
- 出現回数 10 回未満の英単語については，Brown の手法に比べ手法 A, B の方が有効である．
- 北村の手法は出現頻度に依存することなく対訳単語対を抽出することが可能であるが，すべての英単語について対訳仏単語を抽出することは出来ない．

6 今後の展望

6.1 仏語動詞の活用

仏語の動詞の過去形には継続を表す半過去，一般の過去形に相当する複合過去，過去の完了を表す大過去の3種類がある．従って英語動詞に対する対訳仏単語が一意に定まらないという問題がある．

6.2 熟語の扱い

仏英間で単語が一対一に対応しない場合がある．訳されない単語，あるいは互いに単独では意味を持たないが熟語として一つの意味を成す単語もあるので，単語を切り出す際に，仏英の両コーパスから熟語を一単語と同等に切り出すシステムが必要である．

6.3 対訳コーパスについて

一般に，文の対応が成されているコーパスは少ない．本研究においても手作業で文の対応付けを行った．また，相互情報量は出現する単語の頻度に依存するという結果も得られた．

以上のことから，対訳辞書作成には各単語が均等に出現して必ず同じ仏単語は同じ英訳になっているコーパスが理想といえが，実際にはそのようなコーパスは存在しないことから，今後，どのようなコーパスに対しても対応出来るようなシステムに改善していく必要がある．

参考文献

[1] Peter F. Brown, *A Statistical Approach to Language Translation*,
International Conference On Computational Linguistics, v1, P. 71-76, 1988

[2] 北村美穂子 松本裕治,「二言語対訳コーパスからの翻訳知識の自動獲得」,
電子情報通信学会, 言語理解とコミュニケーション研究会, 1994.

[3] Jules Verne, *De la terre à la lune*,
Hetzel, Paris, 1865.

[4] Jules Verne, *From the Earth to the Moon*,
Project Gutenberg, 1993.