統計情報を用いた対訳単語辞書の作成

4L－2

大森 久美子　堤 純也　中西 正和
慶應義塾大学大学院 理工学研究科 計算機科学専攻 修士課程1年

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．

対訳辞書作成には確立した手法がないため，現在は人間の手作業で作成する場合が多い．そこで，本研究では統計情報を用いたP. F. Brown [1] の手法をもとに対訳単語辞書の自動作成を試みる．

2. Brownの手法

Brownは，用いる対訳コーパスに対し1. 各対訳単語間の一対一対応. 2. 各文間の一対一対応. という2点を前提としている．

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．

$$MI(e_j, f) = \log_2 (P(e_j|f) / P(e_j)) \qquad (1)$$

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．

$$P(e_j|f) = C(e_j, f) / M(f) \qquad (2)$$

また，$P(e_j)$ は，式(3)のように表される．

$$P(e_j) = \sum_f (P(e_j|f) / P(f)) = \sum_f (P(e_j|f)\ M(f) / M_f) \qquad (3)$$

3. 本システムの構成

本システムではBrown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* [2] (総単語数39,983語)，及び *From the Earth to the Moon* [3] (総単語数56,463語) のうち805文(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) をコーパスとして用いた．

3.1 対訳単語辞書作成の手順

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，Brownの手法を用いてその仏文の対訳文に出現する，英単語との相互情報量を計算する．
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．

Extraction of bilingual dictionary from bilingual corpus based on statistical information
Kumiko OHMORI　Junya TSUTSUMI
Masakazu NAKANISHI
Department of Mathematics, Faculty of Science and Technology, Keio University 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223, Japan

3.2 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. そこで, 出現回数が2回以上の英単語に対し, 以下の2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.

4. 実験結果及び評価

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, 各順位までに「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	2,000	23.2 %	58.0 %	73.7 %

4.2 Brownの手法の改良

3.2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を用いて再度相互情報量を計算をした. 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ので出現回数が2回以上の英単語についてのみ, 各手法を用いて結果を比較したところ表2のようになった. 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 2: 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 %

実験結果から1. 出現回数が2回以上の英単語についてはBrownの手法は有効ではない. 2. 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. ということがわかった.

5. 今後の展望及び予定

- 熟語の扱い...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熟語を一単語と同等に切り出すシステムの必要がある.
- 対訳コーパス...相互情報量は出現する単語の頻度に依存するが, 各単語が均等に出現して必ず同じ単語は同じ訳になっているコーパスは存在しない. 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.
- 他言語間への応用...今後, 日英対訳コーパスに本システムを適用し, 日英間の対訳単語辞書の作成を試みる予定である.

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

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