フレーズテーブルを用いた教師なし用語対訳抽出手法の比較

井手上 雅迪[†]　山本 和英[†]　内山 将夫[‡]　隅田 英一郎[‡]
長岡技術科学大学 電気系[†]　情報通信研究機構 MASTAR プロジェクト[‡]

1　はじめに

用語の対訳は、機械翻訳などの自然言語処理の分野に留まらず、翻訳支援など広い分野で必要とされている。これらの用語対訳を人手で整備すると大変なコストがかかるため、用語対訳を自動抽出する研究が行われてきた。

用語対訳の自動抽出として、フレーズテーブルから分類器によって正しい用語対訳を得る手法がある [2]。フレーズテーブルとは単語列の対訳関係とその翻訳確率などの集合であり、フレーズベース統計的機械翻訳の訓練過程で生成される。

他にも、既存の対訳辞書を利用する手法等がある [5]。これらの手法では人手による学習データを用意したり、予め対訳辞書を用意しておかなければならない。

本稿では、フレーズテーブルから正しい用語対訳を教師なしで抽出する。精度の高い用語対訳を教師なしで抽出するために、C-value[1] と Fisher's exact test での p 値 [3] を尺度として利用する。これらの尺度に加えて、用語対訳の語対応を考慮した対数尤度比を提案し、各尺度で抽出した用語対訳の比較を行った。

2　関連研究

フレーズテーブルを用語対訳抽出に用いた手法に Itagaki et al.[2] の手法がある。この手法では対訳コーパスからフレーズテーブルを獲得し、対訳リストを作成する。次に、対訳コーパスの原言語側から複合名詞リストを作成する。この 2 つのリストで原言語側のフレーズが重複しているものを用語対訳候補とする。フレーズテーブルを用いて得られる用語対訳候補には信頼度の低いものも含まれている。Itagaki et al. は正しい用語対訳であるか判別するために、用語対訳の正解データを予め用意して分類器を学習させている。

既存の対訳辞書を用いたものとして、外池ら [5] の要素合成法がある。この手法は用語を構成する単語・形態素毎に既存辞書から訳語を獲得し、これらを結合することで用語対訳を獲得する。これらの手法では、用語対訳の正解データや対訳辞書が必要になり、人手のコストがかかる。本稿では、学習データや対訳辞書の事前知識を必要としない 3 つの尺度を用いてフレーズテーブルから用語対訳を抽出する。

[†]{ideue,yamamoto}@jnlp.org
[‡]{mutiyama,eiichiro.sumita}@nict.go.jp

3　用語対訳抽出手法

本稿では日英対訳コーパスから用語対訳を抽出する。用語対訳抽出手法の概要を以下に述べる。

(1) 対訳コーパスから単名詞や複合名詞等のパターンマッチにより、両言語の用語候補を抽出する。
(2) フレーズベースの統計的機械翻訳ツールキットである Moses[4] を用いて、対訳コーパスからフレーズテーブルを作成する。
(3) フレーズテーブルから両言語とも用語候補であるフレーズ対を抽出し、用語対訳の候補を作成する。
(4) 候補から正しい用語対訳を獲得するために、各候補に対して尺度を計算する。
(5) 候補を尺度で順位付けし、上位の候補を用語対訳として抽出する。

尺度が等しくなった用語対訳候補に対しては、両言語の用語候補を連結し、文字コードの値で並び替える。

3.1　尺度

フレーズテーブルから抽出した用語対訳候補の全てに対して尺度を計算し、上位の候補を正しい用語対訳とする。用語対訳候補の順位付けのために Fisher's exact test の p 値、用語対訳の語対応を考慮した対数尤度比、C-value のそれぞれに基づいた 3 つの尺度を定義する。

3.1.1　Fisher's exact test

Howard et al.[3] は Fisher's exact test による有意性検定により、フレーズテーブルから信頼性の低いフレーズ対を削除している。本稿では信頼性の高い用語対訳候補である程、正しい用語対訳である可能性が高いと仮定する。

ある用語対訳候補 $T_{J,E}$ における日本語の用語候補を J、英語の用語候補を E とすると、対訳コーパスから表 1 のような分割表が得られる。ここで、

$\boldsymbol{N}$　対訳文の総数
$\boldsymbol{C(J)}$　日本語側に J を含む対訳文数
$\boldsymbol{C(E)}$　英語側に E を含む対訳文数
$\boldsymbol{C(J,E)}$　J と E を含む対訳文数

である。

J と E が独立に出現する場合、表 1 のような分割表が得られる確率は組み合わせの数より (1) 式で定義さ

れる。$C(J,E)$ 以上の各値について確率を計算し、それらの総和が p 値となる。p 値は (2) 式で定義される。

$$P_h(C(J,E)) = \frac{\begin{matrix} C(J) & N-C(J) \\ C(J,E) & C(E)-C(J,E) \end{matrix}}{\begin{matrix} N \\ C(E) \end{matrix}} \quad (1)$$

$$p\text{-}value(C(J,E)) = \sum_{k=C(J,E)}^{\infty} P_h(k) \quad (2)$$

p 値が 0 に近づく程 J と E は従属であるとする。最終的な尺度は以下の式で表される。

$$Score_F = -\log(p\text{-}value) \quad (3)$$

表 1: 分割表

$C(J,E)$	$C(J)-C(J,E)$
$C(E)-C(J,E)$	$N-C(J)-C(E)+C(J,E)$

3.1.2 対数尤度比

本稿ではフレーズテーブルから正しい用語対訳を獲得するための尺度として、用語対訳の語対応を考慮した対数尤度比 (Log-likelihood Ratio, LLR) を提案する。

LLR ではフレーズベース統計的機械翻訳で生成される語対応を構成要素の対応とし、用語対訳内の語対応の強さを尺度とする†。

LLR の計算には用語対訳候補 $T_{J,E}$ 内の語対応と、対訳コーパス内の各対訳文に対する語対応の情報が必要となる。これらの情報は Moses が翻訳モデル学習の際に出力するものを用いる。

$T_{J,E}$ の日本語用語 J は $j_1, j_2, \ldots, j_k$ で構成されるとし、英語側を $e_1, e_2, \ldots, e_l$ とすると、$T_{J,E}$ の尺度は以下の式で定義される。

$$Score_L(T_{J,E}) = \sum_{(j_k,e_l)\in A_{j,e}} LLR_{j,e}(j_k, e_l \mid T_{J,E}) + \sum_{c\in A_c} LLR_{c,\varphi}(c, \varphi \mid T_{J,E}) \quad (4)$$

ここで、$T_{J,E}$ 内での語対応の集合を $A_{j,e}$、$T_{J,E}$ 内で対応先を持たない構成要素の集合を A_c とする。日本語用語 $J = \{j_1, j_2, j_3\}$, 英語用語 $E = \{e_1, e_2, e_3\}$ で構成される $T_{J,E}$ の語対応を例として図 1 に示す。

$LLR_{j,e}(j_k, e_l \mid T_{J,E})$ は語対応の強さを表し、$\alpha \geq 0$ として以下の式で定義される。

$$LLR_{j,e}(j_k, e_l \mid T_{j,e}) = \log \frac{P(+1 \mid j_k, e_l)}{P(-1 \mid j_k, e_l)} \quad (5)$$

$$P(+1 \mid j_k, e_l) = \frac{j_k \text{と } e_l \text{に対応がある文数} + \alpha}{j_k \text{と } e_l \text{が共に出現した文数} + 2\alpha} \quad (6)$$

$$P(-1 \mid j_k, e_l) = 1 - P(+1 \mid j_k, e_l) \quad (7)$$

†外池ら [5] は用語対訳抽出において、構成要素の対応を考慮することは有効である可能性が高いとし、要素合成法によって高精度の用語対訳を獲得している。

(5) 式は $T_{J,E}$ の日本語用語 J と英語用語 E が対訳文に出現したとき、対訳文内で j_k と e_l が対応する場合と対応しない場合の対数尤度比である。α はスムージングのために設けており、本稿では $\alpha = 1$ として進める。

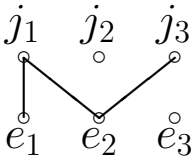
$T_{J,E}$ が対訳文に出現したときに対応を持たない語は空の構成要素 φ と対応していると考え、$P(+1 \mid c, \varphi)$ は以下の式で定義される。

$$P(+1 \mid c, \varphi) = \frac{c \text{ の対応先がない文数} + \alpha}{c \text{ が出現した文数} + 2\alpha} \quad (8)$$

$LLR_{c,\varphi}(c, \varphi \mid T_{j,e})$ は (5) 式と同様に求める。

図 1 の語対応例では、$Score_L(T_{J,E})$ を以下のように計算する。

$$Score_L(T_{J,E}) = LLR_{j,e}(j_1, e_1|T_{J,E}) + LLR_{j,e}(j_1, e_2|T_{J,E}) + LLR_{j,e}(j_3, e_2|T_{J,E}) + LLR_{c,\phi}(j_2, \phi|T_{J,E}) + LLR_{c,\phi}(e_3, \phi|T_{J,E})$$



$A_{j,e} = \{(j_1, e_1), (j_1, e_2), (j_3, e_2)\}$
$A_c = \{j_2, e_3\}$

図 1: $T_{j,e}$ の語対応例

3.1.3 C-value

C-value[1] は (9) 式で定義され、用語 T が安定して使用される度合いを表す。

$$C\text{-}value(T) = (|T| - 1)\left(n(T) - \frac{t(T)}{c(T)}\right) \quad (9)$$

ここで、

$\boldsymbol{|T|}$ T の構成要素数
$\boldsymbol{n(T)}$ コーパスにおける T の出現頻度
$\boldsymbol{t(T)}$ T を部分文字列として含む用語の延べ語数
$\boldsymbol{c(T)}$ T を部分文字列として含む用語の異なり語数

である。

$(|T| - 1)$ より、構成要素数が 1 の用語は C-value が 0 になり、構成要素数に比例して高い値をとる。

本稿では用語対訳候補について、両言語側とも C-value が高いという条件を満たせば、正しい用語対訳である可能性が高いと仮定する。C-value による用語対訳候補の順位付けは、(1) 両言語の用語に対して C-value を計算し、各言語で独立に順位付けを行う。(2) 用語対訳候補を構成する両言語の用語に対して、順位の平均を計算する。(3) 順位の平均を尺度 ($Score_C$) として、用語対訳候補を順位付けする。という手順で行う。

3.2 部分文字列を考慮した計数

本稿では対訳コーパスから両言語の用語候補を抽出し、各尺度の計算に必要な用語候補の出現頻度を数え

る。ある用語候補 T の部分文字列として文章に出現した用語候補 T' は、T' 単体で使用されていない。用語候補の適切な出現頻度を数えるために、用語候補が単体で使用されていない場合は出現頻度を数えないという制限を設ける。

C-value は既に部分文字列を考慮した尺度なので、Fisher's exact test と LLR に対してこの制限を設けて用語対訳抽出を行い、制限を設けない場合と比較する。

4 実験

アパレル分野の約 6 万文対を日英対訳コーパスとして用い、3 章で述べた手順によって用語対訳を抽出した。フレーズテーブルから抽出した用語対訳候補は 22,543 対であった。これらの用語対訳候補に対して各尺度を計算した。各尺度において上位 1,000 対を正しい用語対訳として抽出し、抽出した用語対訳の比較を行った。

4.1 評価方法

用語対訳候補を各尺度で順位付けし、上位 1,000 対から無作為に 100 対を抽出して対訳精度を評価した。評価結果を表 2 に示す。評価は人手で行い、対訳として正しいものを A、対訳として正しいが、文脈に依存するものを A'、部分的に正しいものを B、対訳として正しくないものを C とした。評価結果を表 2 に示す。

表 2: 各尺度で獲得した用語対訳の精度 (100 対中)

尺度＼評価 (対)	A	A'	B	C
$Score_F$	43	25	24	8
$Score_L$	77	5	18	0
$Score_C$	78	6	14	2
$Score'_F$	71	18	8	3
$Score'_L$	79	4	17	0
$Score_{FLC}$	87	2	11	0

4.2 各尺度での対訳精度

表 2 より各尺度での A 評価の数を比較すると、$Score_F$ が最も対訳精度が低かった。$Score_L$ と $Score_C$ では 1 対の差で $Score_C$ が最も良い精度となったが、C 評価の数は $Score_L$ が最も少なく、$Score_L$ と $Score_C$ は同程度の精度で用語対訳を抽出することが分かった。

$Score'_F$ と $Score'_L$ は部分文字列を考慮した場合の評価結果である。$Score'_F$ での評価 A の数は 71 対となり、制限を設けない場合よりも 28 対多く抽出した。$Score'_L$ でも精度が向上しており、用語対訳の部分文字列を考慮することは有効であることが分かった。

4.3 各尺度で抽出した用語対訳の比較

$Score'_F$,$Score'_L$,$Score_C$ の各尺度において、どのような用語対訳を抽出しているのか調査した。図 2 に用語対訳候補の上位 5,000 に対する日本語側の構成要素数と出現頻度の変化を示す。各点は 1,000 位毎の平均値である。英語側の構成要素数は日本語側の性質とほぼ同じ性質を示したため省略する。

図 2 より、各尺度で性質の異なる用語対訳が上位に集中していることが分かった。$Score'_F$ は高頻度で構成要素数が少ない用語対訳を抽出した。$Score'_L$ によって抽出した用語対訳も高頻度であるが、構成要素数は多いものを抽出した。$Score_L$ は対応の強さを語対応の数だけ加算するため、構成要素数が多く、それぞれの対応が強い用語対訳が上位に集中する。$Score_C$ では低頻度で構成要素数が多い用語対訳を抽出した。

3.1.3 節で述べたように構成要素数が 1 の用語は C-value が 0 となるため、$Score_C$ による用語対訳候補の順位では、構成要素数が 2 以上のものが集中している。$Score_C$ は両言語で独立に C-value を計算した結果を統合しているので、日本語用語がどの英語用語に翻訳されるかについて考慮していない。しかし、対訳精度は高いことから、フレーズテーブルから用語対訳を抽出する本手法において、両言語側の用語が安定して出現し、構成要素数が 2 以上であるという制限が対訳精度に影響を与えることが分かった。

4.4 部分文字列を考慮した計数の影響

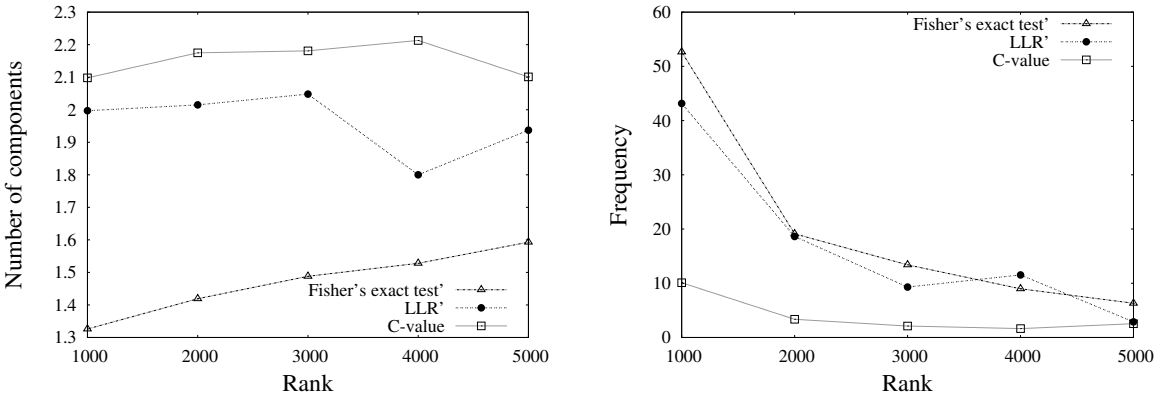
表 3 に $Score_F$ で抽出され $Score'_F$ では抽出されなかった用語対訳の例と、それぞれの尺度における順位を示す。括弧内は用語対訳の正誤判定である。

1 番目と 2 番目の例は、間違った用語対訳を $Score_F$ によって抽出した例である。「リング/coloring」の正しい用語対訳は「カラー リング/coloring」である。このような間違った用語対訳に高い $Score_F$ が与えているのは、用語対訳の部分文字列を考慮せずに出現頻度を数えているためである。「カラー リング/coloring」の例では、日本語用語が「カラー」と「リング」に分割された形態素解析結果となっている。「coloring」に対して「リング」が単体で出現することは少ないにもかかわらず、部分文字列として出現した場合を考慮していないので「リング/coloring」の出現頻度が高くなる。

2 番目の例については「side pocket」中の「side」と「サイド ポケット」が同時に出現する頻度を数えているので、間違った用語対訳が抽出された。

表 3: $Score_F$ で抽出され $Score'_F$ で抽出されなかった用語対訳例と各尺度における順位

用語対訳	$Score_F$	$Score'_F$
リング/coloring (X)	35	21,676
サイド ポケット/side (X)	837	15,222
スリーブ/sleeve (O)	749	5,433



(a) 構成要素数 (日本語側)　　(b) 出現頻度

図 2: 各尺度で抽出した用語対訳の性質 (上位 5,000)

3 番目の例は $Score_F$ で正しく抽出されているが、$Score'_F$ では抽出できなかった例である。抽出失敗の理由は「スリーブ」と「sleeve」がそれぞれ部分文字列として多く出現し、十分な頻度が得られなかったためである。用語対訳候補内には「スリーブ」を部分文字列として含む日本語用語候補が 38 個、「sleeve」を含む英語用語候補が 49 個あった。「スリーブ/sleeve」の例のように、他の用語の部分文字列となりやすい用語で構成されている用語対訳は、用語対訳単体で出現する頻度が低くなる。

$Score'_L$ についても $Score_L$ より精度が向上しているが、$Score'_F$ 程大きな向上ではない。$Score_L$ の上位には構成要素数の多い用語対訳が集中するため、部分文字列となっている用語対訳が抽出されにくい。

4.5 各尺度の統合

4.3 節より、各尺度によって抽出する用語対訳の性質が異なることが分かった。そこで各尺度の性質を持った用語対訳を抽出することで、高品質の用語対訳が獲得できると考えた。そのために各用語対訳候補について $Score'_F$,$Score'_L$,$Score_C$ での順位の平均値を $Score_{FLC}$ と定義し、これを各尺度を統合した尺度として用語対訳の抽出を行った。

評価は 4.1 節と同様に行い、評価結果は表 2 に示す。

各尺度を統合したところ、尺度を統合しない場合で最も精度が高かった $Score'_L$ より A 評価が 8 対増え、各尺度を統合することで精度の高い用語対訳を抽出できることが分かった。

5 おわりに

本稿では、フレーズテーブルから教師なしで用語対訳を抽出するために、Fisher's exact test の p 値、用語対訳の語対応を考慮した対数尤度比、C-value のそれぞれを利用した尺度と用語対訳に適した計数方法を考え、各尺度の比較を行った。

各尺度について対訳精度を評価したところ、C-value と対数尤度比のそれぞれを用いた尺度で高精度な用語対訳が抽出できた。Fisher's exact test を用いた尺度については、用語の部分文字列を考慮して出現頻度を数えることで、高精度な用語対訳が抽出できることを確認した。各尺度で抽出できた用語対訳は、用語対訳の構成要素数と出現頻度の 2 点で性質が異なっていた。

最後に、各尺度で抽出できる用語対訳の性質が異なるため、尺度を組み合わせて抽出を行った。尺度を組み合わせない場合よりも対訳精度が向上し、教師なしの手法でも高精度の用語対訳を抽出できた。

参考文献

[1] Katerina T. Frantzi and Sophia Ananiadou. Extracting nested collocations. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING 96)*, pp. 41–46, 1996.

[2] Masaki Itagaki, Takako Aikawa, and Xiaodong He. Automatic validation of terminology translation consistency with statistical method. In *Proceedings of MT summit XI*, pp. 269–274, 2007.

[3] J Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 967–975, 2007.

[4] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics(ACL)*, pp. 177–180, 2007.

[5] 外池昌嗣, 宇津呂武仁, 佐藤理史. ウェブから収集した専門分野コーパスと要素合成法を用いた専門用語訳語推定. 自然言語処理, Vol. 14, No. 2, pp. 33–68, 2007.