対訳単語辞書の精度調査

中村友哉 *1　村上仁一 *2
*1 鳥取大学 工学部 知能情報工学科
{s122034,murakami}@ike.tottori-u.ac.jp

1 はじめに

現在のパターンベース統計翻訳の翻訳精度は低い．そしてパターンベース統計翻訳には対訳単語を起点としている．翻訳精度が低い原因としてこの対訳単語の精度が低いことが問題と考える．そこで本研究では対訳単語の日本語と英語の対応の精度を調査した．

2 パターンベース統計翻訳

パターンベース統計翻訳の翻訳手順を以下に示す [1].

手順 1　対訳単語
対訳学習文と GIZA++ を用いて対訳単語を作成する．

手順 2　単語レベル文パターン
対訳単語辞書 (手順 1) と対訳学習文を用いて単語レベル文パターンを作成する．

手順 3　対訳句の作成
単語レベル文パターン (手順 2) と対訳学習文を用いて対訳句を作成する．

手順 4　句レベル文パターンの作成
対訳句 (手順 3) と対訳学習文を用いて句レベル文パターンを作成する．

手順 5　翻訳
対訳句 (手順 3) と句に基づく対訳文パターン辞書 (手順 4) を用いて翻訳を行う．

パターンベース統計翻訳の流れ図を図 1 に示す.

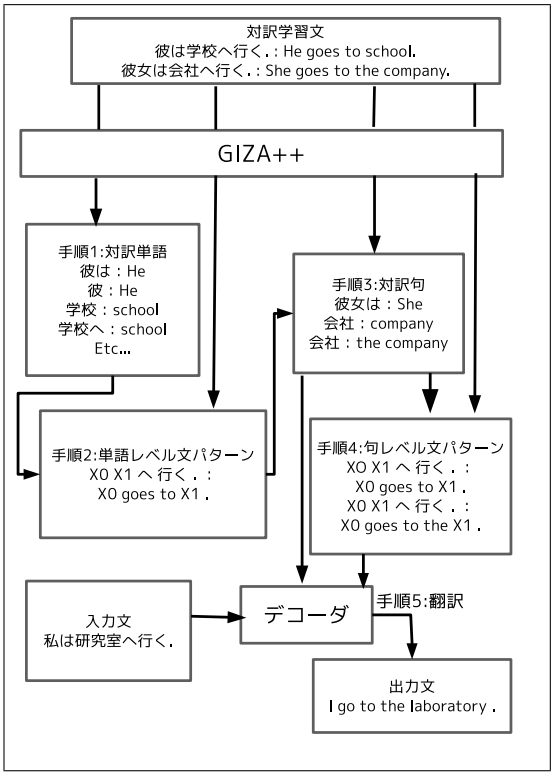

図 1　パターンベース統計翻訳の流れ図

3 対訳単語の問題

現在のパターンベース統計翻訳には対訳単語を起点としている．その対訳単語の精度の低さが翻訳精度の低下に繋がっていると考える．翻訳精度低下の原因を以下に示す.

1. 対訳単語に誤りが含まれることにより単語レベル文パターンに誤りが含まれる．
2. 単語レベル文パターンに誤りが含まれることで対訳句と句レベル文パターンに誤りが含まれる．
3. 対訳句と句レベル文パターンに誤りが含まれることで翻訳において翻訳精度が低下する．

4 研究の目的

対訳単語辞書は対訳単語確率を基にして作成する．対訳単語確率は対訳文と IBM 翻訳モデルにより計算される．本研究では対訳単語に関して日本語単語と英語単語の適切な対応がとられているかの調査を行う.

5 対訳単語辞書の調査条件

学習文 100,000 文と IBM 翻訳モデルで作成した対訳単語 327,604 単語において，日英の対応の精度を調査する．対訳単語の計算に用いる GIZA++ のパラメータを以下に示す.

- m1=4，m2=0，mh=4，m3=0，m4=0，t1=4

6 対訳単語辞書の調査

評価は全てランダム 100 単語を取り出して行う．
評価基準を表 1 に示す．

表 1　評価基準

○	適切な対訳単語
×	不適切な対訳単語

また，表中の $P(E/J)$ は日本語単語が英語単語に訳される GIZA++ の対訳単語確率．$P(J/E)$ は英語単語が日本語単語に訳される GIZA++ の対訳単語確率である．

6.1 全対訳単語

全対訳単語を調査した結果を以下に示す．

表 2　全対訳単語の評価

評価対象数	○	×
327,604	17	83

表 3　全対訳単語の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
ご馳走	feast	-3.47	-4.18	○
ら	Association	-21.43	-17.70	×
’	’hour’	-3.16	-1.70	×
1	in	-5.33	-16.21	×

6.2 記号及びひらがなの評価

6.2.1 記号

全対訳単語中の記号を調査した結果を以下に示す．

表 4　記号の評価

評価対象数	○	×
317	0	100

表 5　記号の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
°	triangle	-1.76	-19.60	×
「	brand	-23.12	-4.58	×

6.2.2 ひらがな 1 文字

全対訳単語中のひらがな 1 文字を調査した結果を以下に示す．

表 6　ひらがな 1 文字の評価

評価対象数	○	×
630	0	100

表 7　ひらがな 1 文字の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
き	blow	-20.44	-17.98	×
は	This	-14.61	-3.84	×

表 3，5 より記号やひらがな 1 文字の対訳単語は全て不適切である．

6.3 頻度における調査

6．3 節において頻度とは当該単語を含む対訳学習文の数である．

6.3.1 日本語単語の頻度 1

全対訳単語中の日本語単語の頻度 1 の対訳単語を調査した結果を以下に示す．

表 8　日本語単語の頻度 1 の単語の評価

評価対象数	○	×
30,468	16	84

表 9　日本語単語の頻度 1 の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
ぜん息	asthma	-2.32	-10.81	○
希少	stamps	-8.18	-18.64	×

6.3.2 英語単語の頻度 1

全対訳単語中の英語単語の頻度 1 の対訳単語を調査した結果を以下に示す．

表 10　英語単語の頻度 1 の単語の評価

評価対象数	○	×
43,001	14	86

表 11　英語単語の頻度 1 の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
むっつり	sullenly	-1.56	-1.87	○
自称	professedly	-2.16	-1.59	×

日本語単語の頻度 1 の対訳単語の評価 (6．3．1 節) と英語単語の頻度 1 の対訳単語の評価 (6．3．2 節) は全対訳単語の評価 (6．1 節) と差が小さい．

6.3.3 日本語単語と英語単語が両方同時に含まれる対訳文の頻度 1

全対訳単語中の日本語単語と英語単語が両方同時に含まれる対訳文の頻度 1 の対訳単語を調査した結果を以下に示す．

表 12　頻度 1 の対訳単語の評価

評価対象数	○	×
246,217	14	86

表 13　頻度 1 の対訳単語の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
こい	carp	-14.47	-16.19	○
どんどん	better	-13.29	-8.45	×

この種の対訳単語の評価は全対訳単語の評価 (6．1 節) と差が小さい．

6.3.4 日本語単語の頻度 1 かつ英語単語の頻度１

全対訳単語中の日本語単語の頻度１かつ英語単語の頻度 1 の対訳単語を調査した結果を以下に示す.

表 14 日本語単語の頻度 1 かつ英語単語の頻度 1 の対訳単語の評価

評価対象数	○	×
7,083	37	63

表 15 日本語単語の頻度 1 かつ英語単語の頻度 1 の対訳単語の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
マストドン	Mastodons	-1.07	-1.59	○
ダンツァス	Hamburg	-14.61	-3.87	×

なおこの種の対訳単語は固有名詞が大半であった.

6.3.5 日本語単語と英語単語が両方同時に含まれる頻度 ≧ 2

全対訳単語中の日本語単語と英語単語が両方同時に含まれる文の数が頻度 2 以上の対訳単語を調査した結果を以下に示す.

表 16 頻度 2 以上の対訳単語の評価

評価対象数	○	×
83,017	40	60

表 17 頻度 2 以上の対訳単語の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
船底	bottom	-1.04	-5.86	○
川	crossed	-14.70	-16.14	×

日本語単語の頻度 1 かつ英語単語の頻度 1 の対訳単語 (6．3．4 節) と頻度 2 以上の対訳単語 (6．3．5 節) は精度が高い.

6.4 数字及びアルファベットの評価

6.4.1 数字

全対訳単語中の対訳単語において数字を調査した結果を以下に示す.

表 18 数字の評価

評価対象数	○	×
569	8	92

表 19 数字の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
6	6	-3.29	-0.64	○
1	to	-9.33	-22.33	×

6.4.2 アルファベット大文字 (日本語単語)

全対訳単語中の対訳単語において日本語単語がアルファベット大文字の対訳単語を調査した結果を以下に示す.

表 20 アルファベット大文字 (日本語単語)

評価対象数	○	×
238	12	88

表 21 アルファベット大文字 (日本語単語) の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
A	A	-0.30	-7.73	○
A	Company	-6.55	-3.02	×

6.4.3 アルファベット小文字 (日本語単語)

全対訳単語忠の日本語単語がアルファベット小文字の対訳単語を調査した結果を以下に示す.

表 22 アルファベット小文字 (日本語単語)

評価対象数	○	×
92	11	81

表 23 アルファベット小文字 (日本語単語) の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
x	x	-1.77	-1.33	○
z	Hz	-3.68	-1.28	×

6.4.4 アルファベット大文字 (英語単語)

全対訳単語中の英語単語がアルファベット大文字の対訳単語を調査した結果を以下に示す.

表 24 アルファベット大文字 (英語単語)

評価対象数	○	×
271	11	89

表 25 アルファベット大文字 (英語単語) の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
私	I	-0.91	-0.75	○
から	I	-14.03	-18.84	×

6.4.5 アルファベット小文字 (英語単語)

全対訳単語中の英語単語がアルファベット小文字の対訳単語を調査した結果を以下に示す.

表 26 アルファベット小文字 (英語単語)

評価対象数	○	×
99	5	94

表 27 アルファベット小文字 (英語単語) の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
x	x	-1.77	-1.33	○
です	a	-7.31	-20.85	×

アルファベットや数字では６と 6 のように対応を取るもの以外は不適切な対訳単語である.

6.5 翻訳に用いる対訳単語の評価

翻訳に用いる対訳単語は以下の条件で枝刈りを行って作成する．

- 付与した対訳単語確率を用いて作成した対訳単語の順位が日本語・英語ともに 8 位以上の単語
- 対訳単語確率が $log2(0.20)$ より高い単語
- 日本語単語と英語単語が両方同時に含まれる文の数 (頻度) が 2 以上の単語

枝刈りした後の対訳単語辞書を調査した結果を以下に示す．

表 28　翻訳に用いる対訳単語の評価結果

評価対象数	○	×
4340	90	10

表 29　翻訳に用いる対訳単語の評価例

日本語	英語	$P(E/J)$	$P(J/E)$	評価
興味	interested	-2.24	-1.72	○
十郎	Danjuro	-1.22	-1.46	×

7 まとめ

本研究では対訳単語の対応の精度調査を行った．
6 節の実験より以下のことがわかった．

1. ひらがな 1 文字や記号の対訳単語は全て不適切である．
2. 日本語単語の頻度 1 の対訳単語の評価 (6．3．1 節) と英語単語の頻度 1 の対訳単語の評価 (6．3．2 節) は全対訳単語の評価 (6．1 節) と差が小さい．日本語単語と英語単語が両方同時に含まれる対訳文の頻度 1 の対訳単語の評価 (6．3．3 節) も同様である．
3. 日本語単語の頻度 1 かつ英語単語の頻度 1 の対訳単語 (6．3．4 節) と頻度 2 以上の対訳単語 (6．3．5 節) は精度が高い．
4. 数字とアルファベットに関する調査では６と 6 やＡと A など対応を取るもの以外は対訳単語が不適切である．

参考文献

[1] 江木孝史，句に基づく対訳句パターンの自動作成と統計的手法を用いた英日パターン翻訳 言語処理学会第 20 回年次大会，A6-2，pp.951-954，2014.

[2] Moses：http://www.statmt.org/moses/.

[3] Franz Josef Och，Hermann Ney，A Systematic Comparison of Various Statistical Alignment Models，Computational Linguistics，pp．19-51，2003.