

階層型統計翻訳における辞書登録の効果

池淵堅斗 *1 村上仁一 *2 坂田純 *2 徳久雅人 *2

*1 鳥取大学 工学部 知能情報工学科

*2 鳥取大学大学院 工学研究科 情報エレクトロニクス専攻

*1,*2 {s102002, murakami, d112004, tokuhisa} @ ike.tottori-u.ac.jp

1 はじめに

近年、機械翻訳の分野で原言語から目的言語に翻訳する統計翻訳が注目されている。統計翻訳は対訳文を用いてフレーズごとの翻訳確率や単語確率を学習する。統計翻訳において、対訳文数が多ければ翻訳精度は高くなるが、対訳文数の数を増やすだけでは精度の伸びに限界がある。

この問題に対し、Maja Popovićらは、セルビア語英語間の翻訳において、小規模のコーパスに辞書データを追加し句に基づく統計翻訳(PSMT)を行った。結果、自動評価結果が向上した[1]。また日野らの研究においても、PSMTにおける学習データへの対訳句の追加の効果が報告されている[2]。

ところで、階層型統計翻訳(HSMT)は文の構造を考慮して翻訳を行うため、PSMTよりも翻訳精度が高いことが知られている。

そこで本研究では、HSMTでも同様に、辞書のデータから抽出した対訳句を学習データに追加し、翻訳精度の調査を行う。

2 提案手法

本研究では、辞書のデータから抽出した対訳句を学習データに追加し、HSMTにおける翻訳精度の調査を行う。

図1に日英統計翻訳の場合の提案手法の流れを示す。

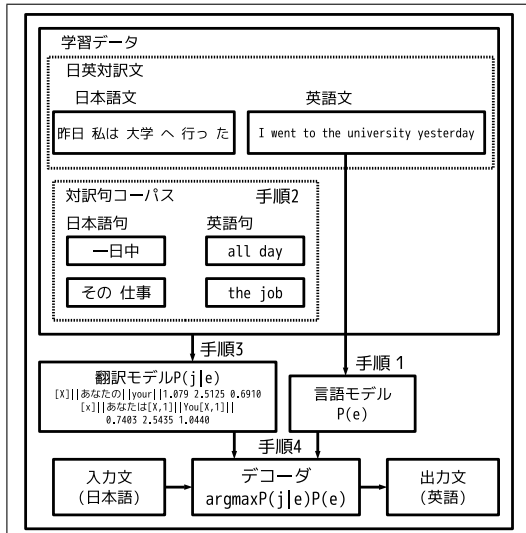


図1 日英統計翻訳の場合の提案手法の流れ

提案手法の手順を以下に示す。

- 手順1 英語文を学習データとして言語モデルを作成する
- 手順2 日英対訳文に対訳句コーパスを追加する
- 手順3 手順2で作成したコーパスを学習データとして翻訳モデルを作成する
- 手順4 手順1と手順3で作成したモデルを用いて統計翻訳を行う

3 実験環境

3.1 日英対訳文

本研究では、日英対訳文として、単文コーパスと重文複文コーパス[3]を用いる。統計翻訳の前処理として、各コーパスの日本語文に対して、MeCab[4]を用いて形態素解析を行う。また、英語文に対して“tokenizer.sed[5]”を用いて分かち書きを行う。前処理を行った単文コーパスの例を表1に、重文複文コーパスの例を表2に示す。

表1 単文コーパスの例

彼の考え方は極端すぎる。
His way of thinking goes too far.

表2 重文複文コーパスの例

その知らせを聞いて彼女の顔は明るくなった。
Her face lit up when she heard the news.

3.2 対訳句コーパス

本研究では対訳句コーパスとして鳥バンク[6]と英辞郎[7]を用いる。本研究で対訳句コーパスとして用いる鳥バンク、英辞郎の対訳句数を表3に示す。

表3 対訳句対数

鳥バンク	548,694 句
英辞郎	1,366,458 句

3.2.1 鳥バンク

鳥バンクは自然言語処理のための言語知識ベースを収録したデータバンクであり、日本語の重文と複文を対象とする「意味類型パターン辞書」が収録されている。本研究では、このパターン辞書から抽出した対訳句を用いる。対訳句の例を表4に示す。

表4 鳥バンクから抽出した対訳句の例

彼のお母さん
his mother
息子の話
son's story
社会通念
commonly accepted ideas

3.2.2 英辞郎

英辞郎は、EDP(Electronic Dictionary Project)がアップデートし続けている英和・和英辞書である。英辞郎のデータには対訳句の他に翻訳例や注釈、本来の文に出てこない“~”等の記号が含まれる。表5に英辞郎の対訳句の例を示す。

表5 英辞郎から抽出した対訳句の例

に急落する
plummet to
から出てくる
come out from
の結果として生じる
come out from

3.3 実験データ

本研究では、日英対訳文として辞書の例文から抽出した単文コーパス [3]、重文複文コーパス [3] を表 6 の内訳で用いる。また、本研究では鳥バンクの対訳句として表 6 中の重文複文の学習データ 100,000 文から抽出した、548,694 句を用いる。

表 6 日英対訳文対数

	単文	重文複文
学習データ	100,000 文	100,000 文
ディベロップメントデータ	1,000 文	1,000 文
テストデータ	10,000 文	10,000 文

3.4 デコーダ

本研究ではデコーダとして“Moses[9]”を用いる。

3.5 実験内容

本研究では、以下に示す合計 16 種類の翻訳実験を行う。

- 日英対訳文として、単文と重文複文を用いる。
- 翻訳実験として、日英統計翻訳と英日統計翻訳を行う。
- 統計翻訳の手法として、HSMT と PSMT を用いる。
- 追加する対訳句として、鳥バンクと英辞郎を用いる。

3.6 評価方法

本研究では、出力文の評価として自動評価と人手評価を行う。自動評価は自動評価法“BLEU[10]”、“NIST[11]”、“METEOR[12]”、“RIBES[13]”を用いる。人手評価は出力文からランダムに 100 文抽出し、対比較評価を行う。

4 実験結果

自動評価を 4.1 節、人手評価を 4.2 節に示す。表記方法を以下に示す。

- 対訳句の追加を行っていない HSMT の翻訳結果を“HSMT”と表記する。
- 対訳句の追加を行っていない PSMT の翻訳結果を“PSMT”と表記する。
- 学習データに対して鳥バンクの対訳句の追加を行った HSMT の翻訳結果を“HSMT+ 鳥”と表記する。
- 学習データに対して鳥バンクの対訳句の追加を行った PSMT の翻訳結果を“PSMT+ 鳥”と表記する。
- 学習データに対して鳥バンクの対訳句の追加を行った HSMT の翻訳結果を“HSMT+ 英”と表記する。
- 学習データに対して鳥バンクの対訳句の追加を行った PSMT の翻訳結果を“PSMT+ 英”と表記する。
- 対比較評価において翻訳後の出力が完全に同一な場合は、“一致”と表記する。
- 対比較評価において翻訳精度にかかわらず同程度である場合は、“差なし”と表記する。

4.1 自動評価

自動評価における、単文の日英翻訳の結果を表 7、単文の英日翻訳の結果を表 8、重文複文の日英翻訳の結果を表 9、重文複文の英日翻訳の結果を表 10 に示す。

表 7 単文 日英翻訳

	BLEU	NIST	METEOR	RIBES
HSMT	0.1397	4.8430	0.4590	0.7143
HSMT+ 鳥	0.1544	5.2274	0.4909	0.7268
HSMT+ 英	0.1527	5.2628	0.4919	0.7315
PSMT	0.1388	4.9161	0.4608	0.7152
PSMT+ 鳥	0.1561	5.2203	0.4897	0.7308
PSMT+ 英	0.1595	5.3405	0.4964	0.7355

表 8 単文 英日翻訳

	BLEU	NIST	METEOR	RIBES
HSMT	0.1827	4.7136	0.4581	0.7718
HSMT+ 鳥	0.1998	5.0070	0.4824	0.7848
HSMT+ 英	0.2064	5.1298	0.4868	0.7871
PSMT	0.1811	4.6653	0.4563	0.7699
PSMT+ 鳥	0.1904	4.9788	0.4665	0.7765
PSMT+ 英	0.1963	4.9880	0.4762	0.7760

表 9 重文複文 日英翻訳

	BLEU	NIST	METEOR	RIBES
HSMT	0.1263	4.5297	0.4183	0.6770
HSMT+ 鳥	0.1400	4.9040	0.4486	0.6915
HSMT+ 英	0.1412	4.9925	0.4546	0.6998
PSMT	0.1223	4.3717	0.4051	0.6694
PSMT+ 鳥	0.1429	4.9310	0.4569	0.6925
PSMT+ 英	0.1426	4.9136	0.4525	0.6961

表 10 重文複文 英日翻訳

	BLEU	NIST	METEOR	RIBES
HSMT	0.1502	4.4037	0.4050	0.7220
HSMT+ 鳥	0.1580	4.4772	0.4154	0.7234
HSMT+ 英	0.1648	4.6284	0.4271	0.7311
PSMT	0.1498	4.2698	0.3983	0.7140
PSMT+ 鳥	0.1573	4.5137	0.4116	0.7155
PSMT+ 英	0.1635	4.6480	0.4257	0.7230

4.2 人手評価

人手評価における、単文の日英翻訳の結果を表 11、単文の英日翻訳の結果を表 12、重文複文の日英翻訳の結果を表 13、重文複文の英日翻訳の結果を表 14 に示す。

具体的には、表 11 の一番上の表は、対訳句を追加していない HSMT における翻訳結果と、学習データに鳥バンクの対訳句を追加した HSMT の翻訳結果との対比較を行い、HSMT における翻訳結果が優れていると判断したのが 6 文、HSMT において学習データに鳥バンクの対訳句を追加して行った翻訳結果が優れていると判断したのが 11 文、差なしと判断したのが 79 文、一致と判断したのが 4 文であることを表している。

表 11 単文 日英翻訳

HSMT	HSMT+ 鳥	差なし	一致
6	11	79	4
HSMT	HSMT+ 英	差なし	一致
4	10	79	7
PSMT	PSMT+ 鳥	差なし	一致
4	16	73	7
PSMT	PSMT+ 英	差なし	一致
4	14	72	10

表 12 単文 英日翻訳

HSMT	HSMT+ 鳥	差なし	一致
2	10	83	5
HSMT	HSMT+ 英	差なし	一致
1	12	76	11
PSMT	PSMT+ 鳥	差なし	一致
4	9	80	7
PSMT	PSMT+ 英	差なし	一致
5	14	72	9

表 13 重文複文 日英翻訳

HSMT	HSMT+ 鳥	差なし	一致
6	11	78	5
HSMT	HSMT+ 英	差なし	一致
5	7	84	4
PSMT	PSMT+ 鳥	差なし	一致
3	14	78	5
PSMT	PSMT+ 英	差なし	一致
3	16	81	0

表 14 重文複文 英日翻訳

HSMT	HSMT+ 鳥	差なし	一致
4	7	83	6
HSMT	HSMT+ 英	差なし	一致
4	9	84	3
PSMT	PSMT+ 鳥	差なし	一致
2	11	83	4
PSMT	PSMT+ 英	差なし	一致
4	17	77	2

4.3 対比較評価の出力例

対比較評価の出力例を表 15, 表 16 に示す. また, 出力例において誤りであると判断した箇所を下線で示す.

表 15 “HSMT”と“HSMT+ 鳥”の出力例 (単文 日英翻訳)

HSMT の例	
入力文	我が社には優秀な技術者が居ります。
正解文	Our company has some talented technicians .
HSMT	There is excellent enginerrrs in our company .
HSMT+ 鳥	Our company has excellent engineers is 居り .
HSMT+ 鳥 の例	
入力文	彼はその作品を新聞で褒めた。
正解文	He praised the work of art in the newspaper .
HSMT	He was the work in the newspaper .
HSMT+ 鳥	He praised the work in the newspaper .

表 16 “HSMT”と“HSMT+ 鳥”の出力例 (重文複文 英日翻訳)

HSMT の例	
入力文	I think he made a good choice .
正解文	彼はよい選び方をしたと思う。
HSMT	彼がよい選択したと思う。
HSMT+ 鳥	彼を信用したのがいいと思います。
HSMT+ 英 の例	
入力文	they demolished the old factory and built a condominium .
正解文	彼らはその古い工場を取りこわしてマンションを建てた。
HSMT	彼らはその古い工場 demolished ているマンションを建てた。
HSMT+ 鳥	彼らはその古い工場を壊してマンションを建てた。

4.4 実験結果のまとめ

- HSMT においても全ての実験結果において, 対訳句の追加における翻訳精度の向上が見られた. これは PSMT と同様の結果である.
- HSMT と PSMT のそれぞれにおける対訳句の追加の効果の方が高かった. ただし, 単文の英日翻訳においては HSMT における対訳句の追加の効果の方が高かった.
- 鳥バンクと英辞郎の対訳句の追加における翻訳精度を比べると, 差は見られなかった.

5 考察

5.1 HSMT と PSMT の比較

HSMT と PSMT の翻訳精度の比較を行うため, 以下に示す人手評価を行った. 結果を表 17 から表 20 に示す. また, 対比較評価の出力例を表 21, 表 22 に示す.

表 17 単文 日英翻訳

PSMT	HSMT	差なし	一致
3	5	70	22
PSMT+ 鳥	HSMT+ 鳥	差なし	一致
6	5	64	25

表 18 単文 英日翻訳

PSMT	HSMT	差なし	一致
2	6	71	21
PSMT+ 鳥	HSMT+ 鳥	差なし	一致
3	8	72	17

表 19 重文複文 日英翻訳

PSMT	HSMT	差なし	一致
5	8	77	10
PSMT+ 鳥	HSMT+ 鳥	差なし	一致
8	5	71	16

表 20 重文複文 英日翻訳

PSMT	HSMT	差なし	一致
5	7	75	13
PSMT+ 鳥	HSMT+ 鳥	差なし	一致
7	3	82	8

表 21 “PSMT+ 鳥”と“HSMT+ 鳥”の出力例 (重文複文 日英翻訳)

PSMT+ 鳥 の例	
入力文	我々は彼は無罪だと推定した。
正解文	We presumed his innocence .
PSMT+ 鳥	We presumed that he is innocent .
HSMT+ 鳥	We are that he is innocent .
HSMT+ 鳥 の例	
入力文	女が二人寄ればきっと着物の品定めだ。
正解文	When you see two women together , you may be sure they are discussing their own clothes .
PSMT+ 鳥	When two women get together , they will be in kimono .
HSMT+ 鳥	When two women get together , I am sure they discuss their clothes .

表 22 “PSMT+ 鳥”と“HSMT+ 鳥”の出力例 (単文 英日翻訳)

PSMT+ 鳥 の例	
入力文	The bill failed to reach the Senate floor .
正解文	その法案は上院の審議に回されなかった。
PSMT+ 鳥	その法案は上院届かなかった。
HSMT+ 鳥	その法案は 上院を床に失敗した。
HSMT+ 鳥 の例	
入力文	This cold medicine is effective .
正解文	この風邪薬はよく効く。
PSMT+ 鳥	この寒さに効く薬はない。
HSMT+ 鳥	この薬は風邪に効果的だ。

表 17 から表 20 の結果より，対訳句の追加を行った場合は，PSMT が HSMT よりも翻訳精度が高くなった。ただし，単文の英日翻訳においては，HSMT に対訳句を追加した翻訳の方が高くなった。この結果より以下のことが言える。

- HSMT における学習データへの対訳句の追加の効果よりも，PSMT における学習データへの対訳句の追加の効果の方が高い。
- PSMT において学習データに対訳句を追加して行った翻訳の結果が，すべての翻訳結果の中で一番精度が高い。

5.2 鳥バンクと英辞郎の比較

鳥バンクと英辞郎の対訳句の追加の効果を比較するために以下に示す人手評価を行った。結果を表 23 から表 26 に示す。

表 23 単文 日英翻訳

HSMT+ 鳥	HSMT+ 英	差なし	一致
9	4	82	5
PSMT+ 鳥	PSMT+ 英	差なし	一致
7	10	73	10

表 24 単文 英日翻訳

HSMT+ 鳥	HSMT+ 英	差なし	一致
8	8	81	3
PSMT+ 鳥	PSMT+ 英	差なし	一致
7	7	74	12

表 25 重文複文 日英翻訳

HSMT+ 鳥	HSMT+ 英	差なし	一致
6	6	85	3
PSMT+ 鳥	PSMT+ 英	差なし	一致
10	8	80	2

表 26 重文複文 英日翻訳

HSMT+ 鳥	HSMT+ 英	差なし	一致
6	6	86	2
PSMT+ 鳥	PSMT+ 英	差なし	一致
7	9	82	2

表 23 から表 26 より，英辞郎と鳥バンクの対訳句の追加における翻訳精度の差は見られなかった。

しかし，対訳句数に着目すると，英辞郎は鳥バンクに対して 2 倍以上も対訳句数が多い。それにも関わらず，翻訳精度に差が生じなかった理由として，鳥バンクの対訳句が重文複文コーパスから抽出されたものであるため，実験で用いたテスト文との親和性が高かったことが要因のひとつではないかと考える。

6 分野外実験

鳥バンクは重文複文コーパスから抽出した対訳句であるので，重文複文コーパスとの親和性が高い。そこで，重文複文の実験において，新たに分野外のオープンなデータである 3,950 文の重文複文コーパスをテスト文として用いて，分野外の実験を行った。

6.1 分野外実験の内容

分野外実験では，3.5 節の実験内容で示した項目の日英対訳文として重文複文のみを扱い実験を行う。したがって，合計 8 種類の翻訳実験を行う。

6.2 分野外実験のデータ

学習データとディベロップメントデータとして 3.3 節で示したものと同じものを使用し，テストデータとして新たに用意した重文複文 3,950 文を用いる。

6.3 分野外実験結果

実験結果として人手評価を示す。日英翻訳の結果を表 27，英日翻訳の結果を表 28 に示す。

表 27 重文複文 日英翻訳

HSMT	HSMT+ 鳥	差なし	一致
4	6	88	2
HSMT	HSMT+ 英	差なし	一致
4	8	88	0
PSMT	PSMT+ 鳥	差なし	一致
5	10	81	4
PSMT	PSMT+ 英	差なし	一致
5	9	84	2

表 28 重文複文 英日翻訳

HSMT	HSMT+ 鳥	差なし	一致
4	12	82	2
HSMT	HSMT+ 英	差なし	一致
6	12	79	3
PSMT	PSMT+ 鳥	差なし	一致
2	11	87	0
PSMT	PSMT+ 英	差なし	一致
1	7	91	1

6.4 分野外実験結果のまとめ

実験結果より，HSMT においても対訳句の追加における翻訳精度の向上が見られた。また，HSMT と PSMT のそれぞれにおける学習データへの対訳句の追加の効果を比べた場合，PSMT における学習データへの対訳句の追加の効果の方が高かった。これは，4 章の結果と同様の傾向である。

7 おわりに

本研究では，HSMT に対して，対訳句コーパスを追加し翻訳精度の調査を行った。実験結果より，提案手法の有効性について確認することができた。また，PSMT への対訳句の追加が HSMT への対訳句の追加よりも，高い効果を示していた。さらに，PSMT において学習データに対して対訳句を追加した翻訳の結果が，すべての実験結果の中で一番高い精度を示した。また，鳥バンクと英辞郎の対訳句の追加における翻訳精度の差は見られなかった。

今後は，今回の実験で得られた結果の誤り分析を行い，さらなる精度の向上を目指し調査を行いたい。

参考文献

- [1] Popović Maja, and Ney Hermann “Statistical Machine Translation with a small amount of bilingual training data”, 5th LREC SALTML Workshop on Minority Languages, pp.25-29. 2006.
- [2] 日野聡子：“日英統計翻訳における対訳句コーパスの効果”，2013 年修論
- [3] 村上仁一，藤波達 “日本語と英語の対訳文対の収集と著作権の考察”，第一回コーパス日本語ワークショップ，pp.119-130. 2012.
- [4] MeCab <http://mecab.sourceforge.net/>
- [5] tokenizer.sed <http://www.cis.upenn.edu/~treebank/tokenizer.sed>
- [6] 鳥バンク <http://unicorn.ike.tottori-u.ac.jp/toribank/>
- [7] 英辞郎 <http://www.alc.co.jp/>
- [8] 藤味良太，村上仁一，徳久雅人，池原信，“統計翻訳における人手で作成された大規模フレーズテーブルの効果”，言語処理学会第 14 回年次大会，pp.224-227. 2008.
- [9] Moses: “Open Source Toolkit for Statistical Machine Translation”, Proceedings of the ACL 2007 Demo and Poster Sessions, pp.177-180. 2007.
- [10] BLEU: “A Method for Automatic Evaluation of Machine Translation”, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp.311-318. 2002.
- [11] NIST, “Automatic Evaluation of Machine Translation Quality Using n-gram Co-Occurrence Statistics” Proceedings of the Human Language Technology Conference (HLT), pp.128-132. 2002.
- [12] Meteor: Lavie Alon, and Denkowski Michael “An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments”, Proceedings of the Second Workshop on Statistical Machine Translation, pp.228-231. 2007.
- [13] Hideki Isozaki, “Automatic Evaluation of Translation Quality for Distant Language Pairs”, Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp.944-952. 2010.