日英統計翻訳における対訳句コーパスの効果

日野聡子 村上仁一 徳久雅人 村田真樹
鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
{s072040, murakami, tokuhisa, murata} @ 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目的言語らしさを学習する．統計翻訳において，対訳文数が多ければ多いほど翻訳精度は高くなる．しかし，利用できる対訳文数には限りがある．

そこで，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翻訳を行った．その結果，自動評価結果が向上した [1]．また，ブルトン語フランス語間の翻訳においても自動評価結果が向上した [2]．

そこで，本研究では日本語英語間の翻訳において，辞書のデータから抽出した対訳句コーパスを日英対訳文に追加し，翻訳精度の調査を行う．対訳句コーパスとして鳥バンク [3] と英辞郎 [4] を用いる．

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コーパスを日英対訳文に追加し，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．

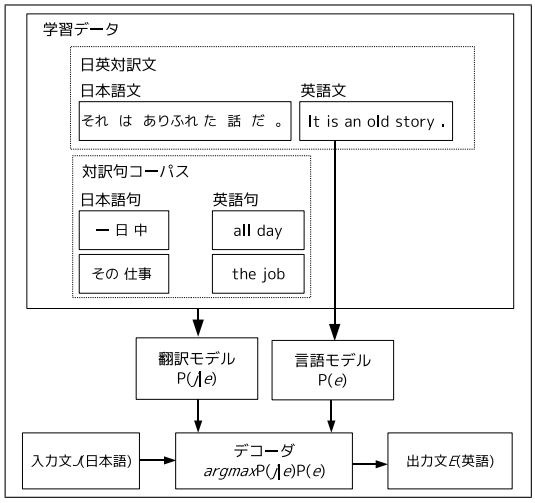

図 1　日英統計翻訳の場合の提案手法の流れ

提案手法の手順を以下に示す．

手順 1　日英対訳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に追加する
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 [5] 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[6] を用いて形態素解析を行う．また，英語文に対して “tokenizer.sed [7]” 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　前処理後の単文コーパスの例

魚 が たくさん 取れ た 。
Many fish were caught .

表 2　前処理後の重文複文コーパスの例

勉強 を し て いる 間 は ラジオ を 切っ て おき なさい 。
While studying , turn off the radio .

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 [3] と英辞郎 [4] を用いる．

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．本研究では鳥バンクから抽出した 698,472 対訳対 [8] を用いる．鳥バンクは重文複文コーパスから抽出したため，重文複文コーパスと非常に親和性が高い．対訳対の例を表 3 に示す．

表 3　鳥バンクの対訳対の例

発酵 槽 に 移し
transferred to the fermenter
摩擦 熱
Friction accounts
コート の すそ
the edge of my coat

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対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．本研究では英辞郎のクリーニングを行い，必要な英語と日本語の対訳対のみの形にした 1,350,299 対訳対を用いる．英辞郎は辞書のデータであるので，重文複文コーパスと親和性が低い．表 4 に英辞郎の対訳対の例を示す．

表 4　英辞郎の対訳対の例

の 姿 に 変装 する
be disguised as
から 出 て くる
come out from
の 結果 として 生じる
come out from

3.3　デコーダ

本研究ではデコーダとして “moses[9]” を用いる．

3.4　翻訳モデルの学習

本研究では翻訳モデルの作成に “train-model.perl” を用いる．

3.5 言語モデルの学習

言語モデルは, N-gram モデルを用いる．N-gram モデルの作成には “SRILM[10]”の ngram-count を用いる．またスムージングに “-kndiscount” を用いる．

4 実験内容

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 [5] を用いた 2 種類の実験を行う．また，対訳句コーパスとして鳥バンク [3] と英辞郎 [4] の 2 種類を用いる．翻訳実験は日英統計翻訳と英日統計翻訳を行う．したがって，合計 8 種類の翻訳実験を行う．

単文コーパスの実験は単文コーパス 182,988 文から，重文複文コーパスの実験では重文複文コーパス 102,712 文から表 5 の内訳で用いる．また，対訳句コーパスの数を表 6 に示す．なお，全ての翻訳実験において，パラメータチューニングは行わない．

表 5 使用した実験データ

	単文	重文複文
日英対訳文	100,000	91,712
テストデータ	10,000	10,000

表 6 対訳句コーパスの数

鳥バンク	698,472
英辞郎	1,350,299

4.1 ベースラインと提案手法

ベースラインでは言語モデルと翻訳モデルの作成に日英対訳文を用いる．提案手法では言語モデルの作成に日英対訳文を，翻訳モデルの作成に日英対訳文と対訳句コーパスを用いる．対訳句コーパスとして，鳥バンクを用いる翻訳実験を提案手法 (鳥バンク) とよび，英辞郎を用いる翻訳実験を提案手法 (英辞郎) とよぶ．以下に言語モデルと翻訳モデルの作成に使用するコーパスについてまとめる．

ベースライン

言語モデル：日英対訳文
翻訳モデル：日英対訳文

提案手法 (鳥バンク)

言語モデル：日英対訳文
翻訳モデル：日英対訳文＋鳥バンク

提案手法 (英辞郎)

言語モデル：日英対訳文
翻訳モデル：日英対訳文＋英辞郎

5 評価実験

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[11]”, “NIST[12]”, “RIBES[13]”, “METEOR[14]”を用いる．人手評価はベースラインと提案手法の出力文からランダムに 100 文抽出し，対比較評価を行う．

6 実験結果

6.1 自動評価結果

自動評価結果を表 7 に示す．表 7 より，すべての自動評価法において，提案手法はベースラインより精度が向上した．また，重文複文コーパスにおいて提案手法 (鳥バンク) は提案手法 (英辞郎) より大幅に精度が向上した．

表 7 自動評価結果

	BLEU	NIST	RIBES	METEOR
日英翻訳 単文				
ベースライン	0.1195	4.521	0.6999	0.4923
提案手法 (鳥バンク)	0.1401	4.964	0.7144	0.5189
提案手法 (英辞郎)	0.1379	5.017	0.7173	0.5199
英日翻訳 単文				
ベースライン	0.1620	4.384	0.6241	-
提案手法 (鳥バンク)	0.1701	4.759	0.6371	-
提案手法 (英辞郎)	0.1813	4.795	0.6503	-
日英翻訳 重文複文				
ベースライン	0.0852	3.743	0.6420	0.4310
提案手法 (鳥バンク)	0.2193	6.086	0.7101	0.5533
提案手法 (英辞郎)	0.1068	4.318	0.6627	0.4609
英日翻訳 重文複文				
ベースライン	0.1239	3.953	0.5618	-
提案手法 (鳥バンク)	0.2275	5.895	0.6459	-
提案手法 (英辞郎)	0.1383	4.276	0.5782	-

6.2 人手評価結果

日英統計翻訳，英日統計翻訳においてベースラインと提案手法の人手評価を行った．ベースライン○は提案手法がベースラインより翻訳品質が劣っていることを示し，提案手法○は提案手法がベースラインより翻訳品質が優れていることを示す．また，差無しは翻訳品質に差が無いことを示し，同一出力は出力文が完全に同一であることを示す．人手評価結果を表 8 に示す．

表 8 人手評価結果

	ベースライン○	提案手法○	差無し	同一出力
日英翻訳 単文				
提案手法 (鳥バンク)	3	14	74	9
提案手法 (英辞郎)	2	13	77	8
英日翻訳 単文				
提案手法 (鳥バンク)	3	13	63	11
提案手法 (英辞郎)	4	14	71	11
日英翻訳 重文複文				
提案手法 (鳥バンク)	2	11	84	3
提案手法 (英辞郎)	0	10	90	0
英日翻訳 重文複文				
提案手法 (鳥バンク)	6	21	67	6
提案手法 (英辞郎)	4	14	76	6

表 8 より，すべての人手評価結果において，提案手法はベースラインよりも優れている．

7 考察

7.1 提案手法 (鳥バンク) と提案手法 (英辞郎) の比較

提案手法 (鳥バンク) と提案手法 (英辞郎) の翻訳品質を比較するために，人手評価を行う．提案手法 (鳥バンク) と提案手法 (英辞郎) の出力文からランダムに 100 文抽出し，人手評価を行った．提案手法 (鳥バンク) ○は提案手法 (鳥バンク) が提案手法 (英辞郎) より翻訳品質が優れていることを示し，提案手法 (英辞郎) ○は提案手法 (鳥バンク) が提案手法 (英辞郎) より翻訳品質が劣っていることを示す．差無しと同一出力は 6.2 章と同じである．人手評価結果を表 9 に示す．

表 7 と表 9 の結果より，以下のことが示された．

1 人手評価結果では，提案手法 (鳥バンク) は全ての翻訳実験において優れている．

2 自動評価結果では，単文コーパスを用いた英日翻訳実験において，提案手法 (英辞郎) が優れているが，その他の翻訳実験において，提案手法 (鳥バンク) が優れている．

表 9 人手評価結果

提案手法 (鳥バンク) ○	提案手法 (英辞郎) ○	差無し	同一出力
日英翻訳 単文			
12	7	75	6
英日翻訳 単文			
7	3	83	7
日英翻訳 重文複文			
5	4	88	3
英日翻訳 重文複文			
12	4	78	6

7.2 鳥バンクと英辞郎の比較

提案手法 (鳥バンク) が提案手法 (英辞郎) よりも精度が良い原因として，“正しいフレーズテーブルの選択”と“未知語の減少”が考えられる．

7.2.1 正しいフレーズテーブルの選択

表 9 の重文複文コーパスを用いた英日翻訳において，提案手法 (鳥バンク) ○と判断した文を表 10 に示す．

表 10 提案手法 (鳥バンク) ○の例 (英日翻訳 重文複文)

入力文	I am still a member of the rank and file , though I joined this company ten years ago .
正解文	入社 し て 1 0 年 たっ て も ぼく は まだ 平 だ 。
提案手法 (鳥バンク)	私 は 1 0 年 前 に この 会社 に 入っ た のに 、 まだ 平 だ 。
提案手法 (英辞郎)	私 は 1 0 年 前 に は 一般 の メンバー を し て も 、 この 会社 に 加わっ た 。

表 10 において，提案手法 (鳥バンク) と提案手法 (英辞郎) の出力文で利用されたフレーズテーブルを表 11 に示す．

表 11 利用されたフレーズテーブル

入力文	I am still a member of the rank and file , though I joined this company ten years ago .
提案手法 (鳥バンク) 出力文	私 は 1 0 年 前 に この 会社 に 入っ た のに 、 まだ 平 だ 。
I	私 は
am still a member of the rank and file	まだ 平 だ
, though	のに 、
I	た
joined	に 入っ
this company	この 会社
ten years ago	1 0 年 前 に
.	。
提案手法 (英辞郎) 出力文	私 は 1 0 年 前 に は 一般 の メンバー を し て も 、 この 会社 に 加わっ た 。
I am	私 は
still a	を し
member of the	の メンバー
rank and file	一般
, though	て も 、
I	は
joined	に 加わっ た
this company	この 会社
ten years ago	1 0 年 前 に
.	。

表 11 において，提案手法 (鳥バンク) は鳥バンクに“am still a member of the rank and file まだ 平 だ”の対訳対が存在し，フレーズテーブルが作成された．このフレーズテーブルが出力文に利用されたため，翻訳品質は向上した．このように，対訳句コーパスに入力文に対して有効な句の対訳対が存在したとき，翻訳品質は良くなる傾向がある．

7.2.2 未知語の減少

表 9 の重文複文コーパスを用いた英日翻訳の提案手法 (鳥バンク) ○において，未知語の減少によって翻訳品質が向上した文は 12 文中 3 文であった．表 12 に例を示す．

表 12 提案手法 (鳥バンク) ○の例 (英日翻訳 重文複文)

入力文	Stop quipping about the boss .
正解文	ボス を 皮肉る の は やめろ 。
提案手法 (鳥バンク)	上司 を 皮肉る の は やめ なさい 。
提案手法 (英辞郎)	上司 について quipping の は やめ なさい 。

表 12 において，鳥バンクに日英対訳文には存在しない“quipping 皮肉る”の対訳対が含まれていた．そのため，フレーズテーブルが作成され，出力文に利用された．英辞郎には“quip 皮肉 を 言う”の対訳対が存在するが，英語句が進行形の“quipping”である対訳対は存在しなかったため，未知語となった．

他の 2 文においては名詞の複数形が未知語として出力された．英辞郎では未知語の単数形の名詞は存在したが，複数形の対訳対は存在しなかったため，未知語となった．

このように，英辞郎には多数の対訳対が存在するが，動詞の活用が原形であったり，名詞が単数形である場合が多い．

8 追加実験

7.2 章の結果から，鳥バンクは重文複文コーパスから抽出した対訳対であるので，重文複文コーパスとの親和性が高く，翻訳精度が向上しやすいのではないかと考える．一方，英辞郎は辞書のデータであるので，重文複文コーパスとの親和性が低く，翻訳精度の向上の幅が小さいのではないかと考える．

ところで，対訳句コーパスに含まれる動詞句に着目すると，鳥バンクの日本語句は終止形で終わるだけでなく，活用しているものも含まれる．例を表 13 に示す．

表 13 鳥バンク動詞の例

やっと そこ に 着い
I finally arrived
とても 緊張 し
am very nervous

一方，英辞郎の日本語句は終止形で終わる動詞が多い．そのため，精度に差が出た可能性がある．そこで，英辞郎の動詞句を活用させ，翻訳実験を行う．

8.1 実験方法

英辞郎の動詞句を活用させ，翻訳実験を行う．活用させた動詞句を用いた翻訳実験を英辞郎 (動詞活用) とよぶ．

手順を以下に示す．

手順 1 日英対訳文を学習データとして言語モデルを作成する

手順 2 英辞郎の日本語句を MeCab を用いて形態素解析をする

手順 3 手順 2 の結果から，日本語句の最後の単語が動詞である動詞句を抽出する

手順 4 動詞を語幹のみ（なし），未然 1，未然 2，連用 1，連用 2，連用 3，終止，連体，仮定，命令，可能の 11 種類にそれぞれに活用させる

手順 5　日英対訳文に英辞郎と手順 4 で作成した動詞句を追加する

手順 6　手順 5 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 7　手順 1 と手順 6 で作成したモデルを用いて統計翻訳を行う

手順 3 で抽出できた動詞句は 286,828 句であった．表 14 に手順 4 における動詞の活用例を示し，表 15 に手順 4 で作成した対訳句コーパス数を示す．

表 14　英辞郎の動詞活用例

活用形	動詞句	活用形	動詞句
基本形	外へ伸ばす	連用 3	外へ伸ば
なし	外へ伸ば	終止	外へ伸ばす
未然 1	外へ伸ばそ	連体	外へ伸ばす
未然 2	外へ伸ばさ	仮定	外へ伸ばせ
連用 1	外へ伸ばし	命令	外へ伸ばせ
連用 2	外へ伸ばし	可能	外へ伸ばせる

表 15　動詞活用後の対訳句コーパスの数

英辞郎	1,350,299
英辞郎 (動詞活用)	3,155,108

8.2　実験結果

8.1 章で作成した英辞郎 (動詞活用) を用いて翻訳実験を行った．

8.2.1　自動評価結果

自動評価結果を表 16 に示す．

表 16　自動評価結果

	BLEU	NIST	RIBES	METEOR
日英翻訳 単文				
提案手法 (英辞郎)	0.1379	5.017	0.7173	0.5199
英辞郎 (動詞活用)	0.1280	4.856	0.6986	0.5004
英日翻訳 単文				
提案手法 (英辞郎)	0.1813	4.795	0.6503	-
英辞郎 (動詞活用)	0.1686	4.551	0.6347	-
日英翻訳 重文複文				
提案手法 (英辞郎)	0.1068	4.318	0.6627	0.4609
英辞郎 (動詞活用)	0.1038	4.282	0.6560	0.4527
英日翻訳 重文複文				
提案手法 (英辞郎)	0.1383	4.276	0.5782	-
英辞郎 (動詞活用)	0.1317	4.150	0.5654	-

表 16 の結果より，全ての翻訳実験において英辞郎 (動詞活用) は提案手法 (英辞郎) よりも翻訳精度が低下した．

8.2.2　人手評価結果

人手評価結果を表 17 に示す．

表 17　人手評価結果

提案手法 (英辞郎) ○	英辞郎 (動詞活用) ○	差無し	同一出力
日英翻訳 単文			
12	7	75	6
英日翻訳 単文			
7	3	83	7
日英翻訳 重文複文			
5	4	88	3
英日翻訳 重文複文			
12	4	78	6

表 17 の結果より，英辞郎 (動詞活用) は提案手法 (英辞郎) よりも翻訳品質が低いことがわかる．

8.3　追加実験のまとめ

鳥バンクを用いた提案手法の精度向上の原因として，英辞郎には無い動詞の活用が考えられた．そこで英辞郎において，末尾単語が動詞である日本語句を 11 種類の活用形に活用させ，学習データに追加し翻訳実験を行った．しかし，自動評価結果と人手評価結果から精度の向上は見られなかった．他の手法として，英辞郎の英語句において，名詞を複数形にする．動詞を過去形，過去完了形，進行形，三人称単数現在形にする．日本語句において，動詞を過去形，進行形にするなどの方法が考えられる．

なお，鳥バンクは重文複文コーパスから抽出された対訳対なので，重文複文コーパスとの親和性が高く，翻訳精度が向上した可能性がある．

9　おわりに

本研究では，提案手法として対訳句コーパスを日英対訳文に追加したコーパスを学習データとして使用し，統計翻訳を行った．対訳句コーパスとして鳥バンクと英辞郎を用い，単文コーパスと重文複文コーパスに対して日英統計翻訳と英日統計翻訳をそれぞれ行った．したがって合計 8 種類の翻訳実験を行った．

その結果，全ての自動評価結果において，提案手法はベースラインよりも精度が向上した．また，人手評価結果においてもベースラインより提案手法の翻訳精度が良いことがわかった．特に重文複文コーパスの翻訳実験において，鳥バンクを用いた提案手法はベースラインや英辞郎を用いた提案手法と比較して大幅に精度が向上した．また，全ての人手評価結果においても，鳥バンクを用いた提案手法は英辞郎を用いた提案手法よりも優れていることがわかった．

鳥バンクには英辞郎にはない動詞の活用が多くあった．そこで，英辞郎の日本語句の動詞を 11 種類の活用形に活用させ，翻訳実験を行ったが，精度は向上しなかった．他の手法として動詞活用以外の手法も考えられる．今後はそれらを検討していく予定である．

参考文献

[1] Popović Maja, and Ney Hermann “Statistical Machine Translation with a small amount of bilingual training data”, 5th LREC SALTMIL Workshop on Minority Languages, pp.25-29. 2006.

[2] Francis M Tyers “Rule-based augmentation of training data in Breton-French stastical machine translation”, 13th Annual Conference of the European Association for Machine Translation, pp.213-217. 2009.

[3] 鳥バンク http://unicorn.ike.tottori-u.ac.jp/toribank/

[4] 英辞郎 : http://www.alc.co.jp/

[5] 村上仁一, 藤波進 “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学ワークショップ, pp.119-130. 2012.

[6] MeCab http://mecab.sourceforge.net/

[7] tokenizer.sed http://www.cis.upenn.edu/ treebank/tokenizer.sed

[8] 鏡味良太, 村上仁一, 徳久雅人, 池原悟, “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] SRILM : The SRI Language Modeling Toolkit, http://www.speech.sri.com/projects/srilm

[11] 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.

[12] NIST, “Automatic Evaluation of Machine Translation Quality Using n-gram Co-Occurence Statistics” Proceedings of the Human Language Technology Conference (HLT), pp.128-132. 2002.

[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.

[14] 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.