鳥バンクと英辞郎を日英対訳文に追加した統計翻訳の調査

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ことが知られている．しかし，利用できる対訳文数には限りがある．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との報告がある [1].

そこで，本実験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句単位の対訳対を日英対訳文に追加し，翻訳精度の調査を行う．句単位の対訳対として鳥バンク [2] と英辞郎 [3] を用いる．

2　提案手法

本実験では辞書のデータから抽出した句単位の対訳対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．

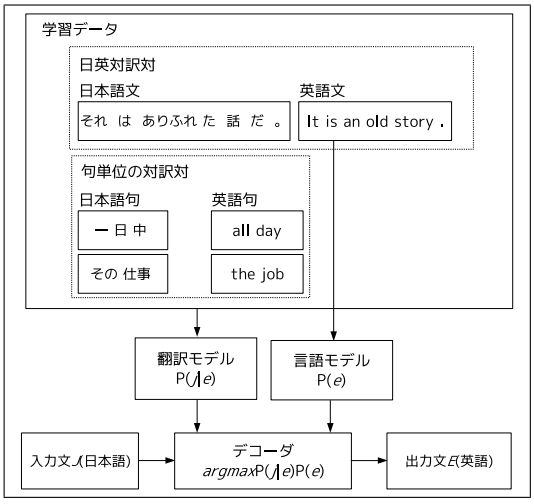

図 1　日英統計翻訳の場合の提案手法の流れ

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　日英対訳文に句単位の対訳対に追加する

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　コーパス

本実験では単文コーパスと重文複文コーパスを用いる．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対訳文の例を表 1 に，重文複文対訳文の例を表 2 に示す．

表 1　前処理後の単文対訳文の例

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

表 2　前処理後の重文複文対訳文の例

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 .

なお，本実験で使用する単文コーパスは日本語文が単文の対訳対である．そのため，英語文は重文複文である場合もある．また，単文コーパスは日本語文が重文複文の対訳対である．そのため，英語文は単文である場合もある．

3.2　句単位の対訳対

本実験では句単位の対訳対として鳥バンク [2] と英辞郎 [3] 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．本実験では鳥バンクから抽出した 698,472 対訳対 [6] を用いる．対訳対の例を表 3 に示す．

表 3　鳥バンク対訳対の例

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や，本来の文に出てこない “〜” 等の記号が含まれる．本実験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 1,366,458 対訳対を用いる．表 4 にクリーニング後の英辞郎の対訳対の例を示す．

表 4　クリーニング後の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

本実験ではデコーダとして “moses[7]” を用いる．

3.4 翻訳モデルの学習

本実験では翻訳モデルの作成に moses の付録である “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[8]”の ngram-count を用いる．またスムージングに “-kndiscount” を用いる．

3.6 パラメータチューニング

パラメータチューニングはディベロップメントデータを翻訳したとき BLEU スコアが高くなるようにパラメータを最適化する．本実験ではパラメータの最適化のために moses の付録の“mert-moses.pl[9]”を用いる．

4 実験内容

本実験では単文コーパスと重文複文コーパスを用いた 2 種類の実験を行う．また，句単位の対訳対として鳥バンク [2] と英辞郎 [3] 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．いずれの翻訳実験もパラメータチューニングを行う．

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，句単位の対訳対の数を表 6 に示す．

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
ディベロップメントデータ	1,000	1,000

表 6 句単位の対訳対の数

鳥バンク	698,472
英辞郎	1,366,458

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句単位の対訳対を用いる．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

言語モデル：日英対訳文
翻訳モデル：日英対訳文＋句単位の対訳対

本実験において句単位の対訳対として，鳥バンクを用いる翻訳実験を提案手法 (鳥バンク) と呼び，英辞郎を用いる翻訳実験を提案手法 (英辞郎) とよぶ．

5 評価実験

本実験では，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]”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

日英統計翻訳の単文コーパスの結果を表 7 に，重文複文コーパスの結果を表 8 に示す．英日統計翻訳の単文コーパスの結果を表 9 に，重文複文コーパスの結果を表 10 に示す．

表 7 自動評価結果 (日英翻訳 単文)

	BLEU	NIST	METEOR
ベースライン	0.1277	4.665	0.5000
提案手法 (鳥バンク)	0.1528	5.220	0.5321
提案手法 (英辞郎)	0.1555	5.203	0.5357

表 8 自動評価結果 (日英翻訳 重文複文)

	BLEU	NIST	METEOR
ベースライン	0.0929	4.030	0.4436
提案手法 (鳥バンク)	0.2559	6.819	0.5863
提案手法 (英辞郎)	0.1178	4.700	0.4810

表 9 自動評価結果 (英日翻訳 単文)

	BLEU	NIST
ベースライン	0.1728	4.612
提案手法 (鳥バンク)	0.1865	4.982
提案手法 (英辞郎)	0.1955	4.984

表 10 自動評価結果 (英日翻訳 重文複文)

	BLEU	NIST
ベースライン	0.1370	4.161
提案手法 (鳥バンク)	0.2461	6.121
提案手法 (英辞郎)	0.1573	4.554

表 7 から表 10 の結果より，全ての自動評価法において，句単位の対訳対を用いた提案手法はベースラインよりも精度が向上した．また，重文複文コーパスの翻訳実験において提案手法 (鳥バンク) は提案手法 (英辞郎) よりも大幅に精度が向上した．

6.2 人手評価

英日統計翻訳，日英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文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文質が優れていることを示す．また，差無しは文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 11 から表 14 に示す．

表 11 人手評価結果 (日英翻訳 単文)

	ベースライン○	提案手法○	差無し	同一出力
提案手法 (鳥バンク)	3	14	74	9
提案手法 (英辞郎)	2	13	77	8

表 12 人手評価結果 (日英翻訳 重文複文)

	ベースライン○	提案手法○	差無し	同一出力
提案手法 (鳥バンク)	2	11	84	3
提案手法 (英辞郎)	0	10	90	0

表 13 人手評価結果 (英日翻訳 単文)

	ベースライン○	提案手法○	差無し	同一出力
提案手法 (鳥バンク)	3	13	63	11
提案手法 (英辞郎)	4	14	71	11

表 14 人手評価結果 (英日翻訳 重文複文)

	ベースライン○	提案手法○	差無し	同一出力
提案手法 (鳥バンク)	6	21	67	6
提案手法 (英辞郎)	4	14	76	6

表 11 から表 14 の結果より，全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．

6.3 翻訳例

表 12 においてベースライン○と判断した文を表 15 に，提案手法 (鳥バンク) ○と判断した文を表 16 に示す．また，表 14 においてベースライン○と判断した文を表 17 に，提案手法 (鳥バンク) ○と判断した文を表 18 に示す．

表 15 ベースライン○の例 (日英翻訳 重文複文)

入力文	あなた に 会い たく て たまらない 。
正解文	I am dying to see you .
ベースライン	I am dying to see you .
提案手法 (鳥バンク)	dying to see you .

表 16 提案手法○の例 (日英翻訳 重文複文)

入力文	大学 を 出 たら 一 人 で 独立 する よう に し なさい 。
正解文	You have to be independent after graduating from college .
ベースライン	Try to make a college , I left alone .
提案手法 (鳥バンク)	Try to independence alone after graduating from college .

表 17 ベースライン○の例 (英日翻訳 重文複文)

入力文	There is no point in worrying a lot about the future .
正解文	先 の 事 を あまり 心配 し て も 無駄 だ 。
ベースライン	将来 の こと が 心配 で も しかた が ない 。
提案手法 (鳥バンク)	将来 の こと を 心配 し ない で いろいろ な 点 で は ない 。

表 18 提案手法○の例 (英日翻訳 重文複文)

入力文	I have decided to vote for him .
正解文	私 は 彼 に 投票 しよ う と 決め た 。
ベースライン	彼 に する こと に 決め た 。
提案手法 (鳥バンク)	私 は 彼 に 投票 する こと に 決め た 。

6.4 実験結果まとめ

6.1 章自動評価と 6.2 章人手評価の結果より，提案手法 (鳥バンク) と提案手法 (英辞郎) はベースラインより翻訳精度が向上し，提案手法の有用性を示すことができた．提案手法 (鳥バンク) の重文複文コーパスを用いた日英統計翻訳は BLEU スコアでベースラインより 0.1630，提案手法 (英辞郎) の単文コーパスを用いた日英統計翻訳は 0.0278 ベースラインより向上した．特に鳥バンクを用いた重文複文コーパスの翻訳実験で BLEU スコアが大幅に向上していることがわかる．

7 考察

7.1 提案手法での向上の原因

ベースラインと提案手法の出力文に用いたフレーズテーブルを調査した．表 18 において，ベースラインと提案手法 (鳥バンク) の出力文に用いたフレーズテーブルと日英フレーズの対応を表 19 に示す．

提案手法 (鳥バンク) の出力文はベースラインのフレーズテーブルに存在しない “vote 投票 する”が用いられたことにより翻訳品質が向上した．これは鳥バンクの対訳対 “vote 投票 する”を学習データに追加したことにより，ベースラインに存在しないフレーズテーブルが提案手法で作成されたためであると考えられる．このように出力文に用いる有効なフレーズテーブルの作成に句単位の対訳対が役に立った．そのため提案手法の翻訳精度が向上したと考える．

表 19 フレーズテーブルと日英フレーズの対応

入力文	I have decided to vote for him .
ベースライン出力文	彼 に する こと に 決め た 。
I have decided	こと に 決め た
to vote	する
for him	彼 に
.	。
提案手法 (鳥バンク) 出力文	私 は 彼 に 投票 する こと に 決め た 。
I have	私 は
decided to	こと に 決め
vote	投票 する
for him	彼 に
.	た 。

7.2 未知語を含む文の割合調査

ベースラインよりも提案手法の精度が向上した原因として，出力文における未知語の減少が考えられる．そこで，各翻訳実験の出力文で未知語を含む文の割合を調査した．日英翻訳実験の結果を表 20 に，英日翻訳実験の結果を表 21 に示す．

表 20 出力文 1 万文中に未知語を含む文数 (日英翻訳)

	単文	重文複文
ベースライン	4,760	5,580
提案手法 (鳥バンク)	1,629	568
提案手法 (英辞郎)	1,261	1,796

表 21 出力文 1 万文中に未知語を含む文数 (英日翻訳)

	単文	重文複文
ベースライン	3,940	4,318
提案手法 (鳥バンク)	1,690	285
提案手法 (英辞郎)	1,371	1,186

表 20 と表 21 から，どちらの提案手法もベースラインの出力文と比べて未知語を含む文数の減少が確認できた．この結果は単文コーパスを用いた翻訳実験での “ベースライン＜提案手法 (英辞郎) ＜提案手法 (鳥バンク)”の傾向と，重文複文コーパスを用いた翻訳実験での “ベースライン＜提案手法 (鳥バンク) ＜提案手法 (英辞郎)”というの傾向と同じであるため，提案手法の翻訳精度の向上の原因であると考えられる．

7.3 提案手法 (鳥バンク) と提案手法 (英辞郎) の比較

重文複文コーパスの英日統計翻訳実験において，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) より文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) より文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 22 に示す．

表 22 人手評価結果 (英日翻訳 重文複文)

提案手法 (鳥バンク) ○	提案手法 (英辞郎) ○	差無し	同一出力
12	4	78	6

表 22 より，提案手法 (鳥バンク) は提案手法 (英辞郎) よりも優れている．

7.4 鳥バンクと英辞郎の差

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．

7.4.1 正しいフレーズテーブルの選択

提案手法 (鳥バンク) ○と判断した文を表 23 に示す．

表 23 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 23 において，提案手法 (鳥バンク) と提案手法 (英辞郎) 出力文に用いたフレーズテーブルと日英フレーズ対応を表 24 に示す．

表 24 フレーズテーブルと日英フレーズの対応

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 24 において，提案手法 (鳥バンク) の出力文は鳥バンクに “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．このように，句単位の対訳対に入力文に対して有効な対訳対が存在したとき，文質は良くなる傾向がある．

7.4.2 未知語の減少

表 22 の提案手法 (鳥バンク) ○において，未知語の減少によって文質が向上した文は 12 文中 3 文であった．表 25 に例を示す．

表 23 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ため，出力文のフレーズテーブルに用いられた．英辞郎には “quip 皮肉 を 言う”の対訳対が含まれているが，英語句が進行形の “quipping”である対訳対は含まれていないため，未知語となった．

表 25 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 。

未知語の減少によって文質が向上した他の 2 文において，名詞の複数形が未知語として出力された．英辞郎には未知語の単数形の名詞は存在したが，複数形の対訳対は含まれていなかったため，未知語となった．このように，英辞郎には多数の動詞や名詞の対訳対が存在するが，動詞の活用が原型であったり，名詞が単数形である場合が多い．

英辞郎に動詞を過去形や現在進行形に変化させた対訳対を追加すれば，翻訳精度は向上すると考える．

8 おわりに

本実験では，提案手法として句単位の対訳対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．句単位の対訳対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，人手評価法においてもベースラインより提案手法の性能が良いという傾向を示すことができ，提案手法の有効性を示すことができた．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．

今後は英辞郎に固有名詞や動詞を過去形や現在進行形に変化させた対訳対の追加を考え，さらなる精度向上を目指す．また，鳥バンクを用いた重文複文コーパスの翻訳実験において大幅に精度が向上した原因のさらなる調査を行いたい．

参考文献

[1] Popović Maja, and Ney Hermann “Statistical Machine Translation with a small amount of bilingual training data”, 5th LREC SALTMIL Workshop on Minority Languages, 2006.
[2] 鳥バンク http://unicorn.ike.tottori-u.ac.jp/toribank/ 2007.
[3] 英辞郎 http://www.alc.co.jp/
[4] MeCab http://mecab.sourceforge.net/
[5] tokenizer.sed http://www.cis.upenn.edu/ treebank/tokenizer.sed
[6] 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.
[7] Moses: Open Source Toolkit for Statistical Machine Translation, Proceedings of the ACL 2007 Demo and Poster Sessions, pages 177-180, Prague, June 2007.
[8] SRILM : The SRI Language Modeling Toolkit, http://www.speech.sri.com/projects/srilm
[9] Franz Josef och, “Minimum Error Rate Training in Statistical Machine Translation”, Association for Computational Linguistics 2003, pp.160-167, 2003.
[10] BLEU: a Method for Automatic Evaluation of Machine Translation, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 311-318, July 2002.
[11] NIST,Automatic Evaluation of Machine Translation Quality Using n-gram Co-Occurence Statistics http://www.itl.nist.gov/
[12] Lavie, Alon and Denkowski, Michael. “The METEOR Metric for Automatic Evaluation of Machine Translation” , 2009.