統計翻訳における英辞郎を利用したパラレルコーパスの効果

日野聡子 村上仁一 徳久雅人 村田真樹
鳥取大学 工学部 知能情報工学科
{s072040, murakami, tokuhisa, murata} @ ike.tottori-u.ac.jp

1　はじめに

統計翻訳は翻訳の際に翻訳モデルと言語モデルを用いる．翻訳モデルは日本語と英語の句対応を管理するフレーズテーブルを用いる．通常フレーズテーブルは，学習データの日本語文と英語文を自動的に対応付けをして作成する．そのため作成されたフレーズテーブルのカバー率は高いが信頼性は低い．

そこで，本実験ではフレーズテーブルの信頼性を高めるために人手で作成された対訳フレーズ辞書を用いる．人手で作成された対訳フレーズ辞書は信頼性が高く，統計翻訳に組み込むことにより翻訳精度が向上するのではないのかと考えられる．具体的には人手で作成された対訳フレーズ辞書を用い，学習データ中の対訳フレーズを1つのまとまりとした新しい学習データを作成する．本実験ではこれを対訳フレーズデータと呼ぶ．

本実験では対訳フレーズデータを用いて作成した信頼性の高いフレーズテーブルにより翻訳精度を向上させることを目的とする．

類似した先行研究として東江ら [1] は英辞郎 [2] の対訳フレーズ対をフレーズテーブルに追加し，翻訳精度が向上したことを報告している．

2　提案手法

本実験では人手で作成された対訳フレーズ辞書を用いて対訳フレーズデータを作成し，統計翻訳を行う．日英統計翻訳の場合の提案手法の流れを図1に示す．

手順1　対訳フレーズ辞書のフレーズ対と，学習データ(日本語文 (1) と英語文 (1)) のマッチングを行う．一致した場合は手順2を行う．例を表1に示す．
表1の例では，学習データの英語文に “old story” 日本語文に “ありふれ た 話”が含まれる文がマッチングする．

表1　マッチングの例

対訳フレーズ辞書	: old story ||| ありふれ た 話
学習データ 日本語文 (2)	: それ は ありふれた 話 だ 。
学習データ 英語文 (2)	: It is an old story .

手順2　学習データの一致した句を一つのまとまりとするために，単語間のスペースを“-”に置き換えた対訳フレーズデータの日本語文 (2) と英語文 (2) を作成する．例を表2に示す．

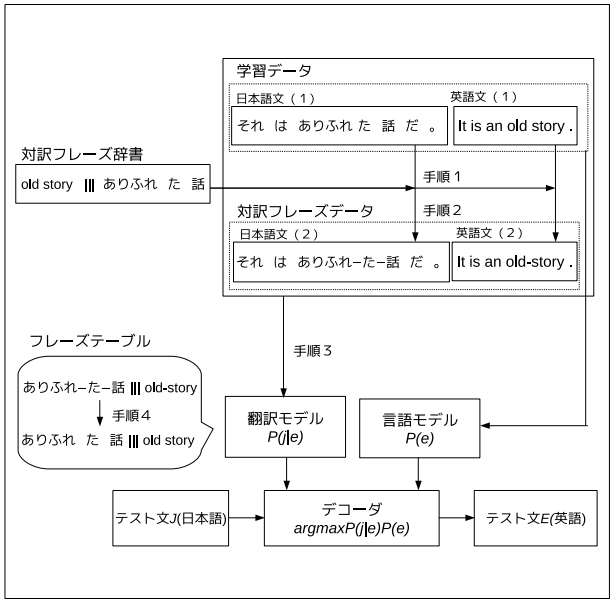

図1　提案手法の流れ

表2　対訳フレーズデータの例

対訳フレーズデータ 日本語文 (2)	: それ は ありふれ-た-話 だ 。
対訳フレーズデータ 英語文 (2)	: it is an old-story .

手順3　学習データと手順2で作成した対訳フレーズデータを用い，フレーズテーブルを作成する．

手順4　フレーズテーブルに含まれる “-”を取り除く．例を表3と表4に示す．

表3　“-” を取り除く前のフレーズテーブル

ありふれ-た-話 ||| old-story

表4　“-” を取り除いた後のフレーズテーブル

ありふれ た 話 ||| old story

手順5　手順4で作成したフレーズテーブルを用いて日英統計翻訳を行う．

3　実験環境

3.1　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，人手で作成された対訳フレーズ辞書として英辞郎 [2] と鳥バンク [3] の2種類を用いる．また，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．

3.2 学習データ

単文コーパスの実験には辞書の例文より抽出した単文コーパス [4] 182,899 文から学習データとして 100,000 文，テストデータとして 10,000 文用いる．また，重文複文コーパスの実験では重文複文コーパス [5] 121,719 文から学習データとして 100,000 文，テストデータとして 10,000 文用いる．

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.perl[7]”を用いて分かち書きを行う．前処理を行った対訳文の例を表 5 に示す．

表 5 対訳文の例

石油 の 発見 で その 国 は 裕福 に なっ た 。
The discovery of oil enriched the country .
梅雨 が 始まっ た 。
The rainy season has set in .
彼 は 食料 品 店 を 営ん で いる 。
He runs a grocery store .

3.3 統計翻訳

3.4 学習データの数

以下にベースラインと提案手法の言語モデルと翻訳モデルを作成する学習データの数を示す．

ベースライン
言語モデル: 100,000 文対
翻訳モデル: 100,000 文対

提案手法
言語モデル: 100,000 文対
翻訳モデル: 100,000 文対＋対訳フレーズデータの文数

対訳フレーズデータの数は 4 章で記述している．

3.4.1 翻訳モデルの学習

翻訳モデルはフレーズテーブルで管理されている．フレーズテーブルの学習には，多くの方法がある．本実験では，moses の付録である“train-factored-phrase-model.perl”を用いる．

3.4.2 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 モデルの学習には “SRILM”の ngram-count を用いる．またスムージングに “-kndiscount” を用いる．

3.4.3 デコーダのパラメータ

デコーダは “Moses[7]” を使用する．“weight-t” は “0.5 0 0.5 0 0” “distortion weight”は 0.2 とする．また，本実験ではパラメータの最適化は行わない．

3.5 対訳フレーズ辞書

本実験では人手で作成された対訳フレーズ辞書として英辞郎と鳥バンクを用いる．

3.5.1 英辞郎

英辞郎 [2] 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フレーズ対の他に翻訳例や注釈であったり，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フレーズ対のみの形にした 1,366,575 フレーズ対を用いる．英辞郎を用いて統計翻訳を行う実験を提案手法 (英辞郎) とよぶ．表 6 にクリーニング後の英辞郎のフレーズ対の例を示す．

表 6 クリーニング後の英辞郎のフレーズ対の例

come out from ||| から 出 て くる
come out from ||| の 結果 として 生じる
obtain information on ||| に 関する 情報 を 得る

3.5.2 鳥バンク

鳥バンク [3] 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 フレーズ対 [8] 用いる．本実験では鳥バンクを用いて統計翻訳を行う実験を提案手法 (鳥バンク) とする．フレーズ対の例を表 7 に示す．

表 7 鳥バンクフレーズ対の例

コート の すそ
the edge of my coat
偉大 な 学者
become a great scholar
カメラ を 買う
buy a camera

3.6 評価実験

本実験では，出力の評価として自動評価法 “BLEU[9]”, “NIST[10]”, “METEOR[11]”を用いる．また，翻訳後の文からランダムに 100 文抽出し，人手で対比較評価も行う．

4 翻訳実験

4.1 対訳フレーズデータ

提案手法 (英辞郎) で作成された対訳フレーズデータの数を表 8，提案手法 (鳥バンク) で作成された対訳フレーズデータの数を表 9 に示す．

表 8 対訳フレーズデータの数 (英辞郎)

単文	重文複文
83,017	84,782

表 9 対訳フレーズデータの数 (鳥バンク)

単文	重文複文
97,649	99,902

鳥バンクを用いて作成した対訳フレーズデータは，英辞郎を用いて作成した対訳フレーズデータより多くなった．これは鳥バンクの対訳フレーズ辞書が単文コーパス及び重文複文コーパスと分野が同じであるため，各コーパスにより多くマッチングしたためと考えている．

5 実験結果

5.1 自動評価

日英統計翻訳の単文の結果を表 10 に，重文複文の結果を表 11 に示す．英日統計翻訳の単文の結果を表 12 に，重文複文の結果を表 13 に示す．

表 10 自動評価結果 (日英翻訳 単文)

	BLEU	NIST	METEOR
ベースライン	0.1091	3.9823	0.4597
提案手法 (英辞郎)	0.1114	3.9891	0.4625
提案手法 (鳥バンク)	0.1068	3.8390	0.4566

表 11 自動評価結果 (日英翻訳 重文複文)

	BLEU	NIST	METEOR
ベースライン	0.0916	3.6270	0.2645
提案手法 (英辞郎)	0.0944	3.6269	0.2667
提案手法 (鳥バンク)	0.0894	3.3828	0.2526

表 12 自動評価結果 (英日翻訳 単文)

	BLEU	NIST
ベースライン	0.1519	4.1148
提案手法 (英辞郎)	0.1527	4.1090
提案手法 (鳥バンク)	0.1522	4.0884

表 13 自動評価結果 (英日翻訳 重文複文)

	BLEU	NIST
ベースライン	0.1132	3.4877
提案手法 (英辞郎)	0.1157	3.5117
提案手法 (鳥バンク)	0.1102	3.3543

結果より，提案手法 (英辞郎) はベースラインと比べて BLEU，METEOR のいずれの自動評価においてもスコアが向上し，提案手法の有効性を確認することができた．しかし，提案手法 (鳥バンク) は日英翻訳の単文の BLEU のみベースラインの評価値より優れているという結果になった．

5.2 人手評価

英日統計翻訳，日英統計翻訳において人手評価を行った．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優れていることを示す．提案手法 (英辞郎) の人手評価結果を表 14 から表 17，提案手法 (鳥バンク) の人手評価結果を表 18 から表 21 に示す．

表 14 人手評価結果 (英辞郎 日英 単文)

ベースライン○	提案手法○	差無し	同一出力
8	1	52	39

表 15 人手評価結果 (英辞郎 日英 重文複文)

ベースライン○	提案手法○	差無し	同一出力
1	5	83	11

表 16 人手評価結果 (英辞郎 英日 単文)

ベースライン○	提案手法○	差無し	同一出力
3	3	58	36

表 17 人手評価結果 (英辞郎 英日 重文複文)

ベースライン○	提案手法○	差無し	同一出力
4	2	73	21

表 18 人手評価結果 (鳥バンク 日英 単文)

ベースライン○	提案手法○	差無し	同一出力
5	0	70	25

表 19 人手評価結果 (鳥バンク 日英 重文複文)

ベースライン○	提案手法○	差無し	同一出力
4	2	90	4

表 20 人手評価結果 (鳥バンク 英日 単文)

ベースライン○	提案手法○	差無し	同一出力
3	8	66	23

表 21 人手評価結果 (鳥バンク 英日 重文複文)

ベースライン○	提案手法○	差無し	同一出力
2	4	88	6

人手翻訳結果において提案手法 (英辞郎) の日英統計翻訳はベースラインより優れているという結果になった．また，提案手法 (鳥バンク) の英日統計翻訳はベースラインより優れているという結果になった．

5.3 翻訳例

提案手法 (鳥バンク) の英日翻訳においての翻訳例を示す．単文の翻訳例を表 22 と表 23 に，重文複文の翻訳例を表 24 と表 25 に示す．

表 22 ベースライン○の翻訳例 (英日 単文)

入力文	He was walking down the street .
正解文	彼 は 通り を 歩い て い た 。
ベースライン	彼 は 通り を 歩い て い た 。
提案手法 (鳥バンク)	通り を 歩い て い た 。
入力文	This town was not prepared for an earthquake .
正解文	この 町 は 地震 に対して 無 防備 だっ た 。
ベースライン	この 町 は 地震 の 用意 し なかっ た 。
提案手法 (鳥バンク)	この 町 は 地震 で い た 。

表 23 提案手法○の翻訳例 (英日 単文)

入力文	I shall never forget your kindness .
正解文	ご 親切 は 決して 忘れ ませ ん 。
ベースライン	私 は 後々 まで 忘れ ませ ん 。
提案手法 (鳥バンク)	ご 恩 は 決して 忘れ ませ ん 。
入力文	I moved to the suburbs last year .
正解文	去年 、 郊外 に 引っ越し まし た 。
ベースライン	私 は 郊外 へ 移転 し た 。
提案手法 (鳥バンク)	私 は 去年 郊外 へ 移動 し た 。

表 24 ベースライン○の翻訳例 (英日 重文複文)

入力文	Ensure that the disk is not exposed to dust , dirt , fingerprints , etc .
正解文	ディスク に ほこり や 指紋 、 汚れ が 付か ない よう に 注意 し て ください 。
ベースライン	それ は 、 ディスク は 、 指紋 が 汚れ ない よう に し て ください 。
提案手法 (鳥バンク)	その こと は あり ませ ん が 、 など 、 土 に 指紋 が ない 。
入力文	Osaka is one of the industrial cities representing Japan .
正解文	大阪 は 日本 の 代表 的 な 工業 都市 で ある 。
ベースライン	大阪 は 日本 代表 都市 の 1つ で ある 。
提案手法 (鳥バンク)	大阪 は 日本 の 産業 を 代表 し て いる 。

表 25 提案手法○の翻訳例 (英日 重文複文)

入力文	He did not give up and stuck it out to the end .
正解文	彼 は あきらめ ず に 最後 まで 粘っ た 。
ベースライン	彼 は あきらめ て は なら ない 。
提案手法 (鳥バンク)	彼 は あきらめ ず に 最後 まで 粘っ た 。
入力文	When the window is opened , the mountain could be seen clearly .
正解文	窓 を 開ける と 山 が よく 見え た 。
ベースライン	窓 を 開ける と が はっきり 見え た 。
提案手法 (鳥バンク)	窓 を 開け て 、 山 が はっきり 見え た 。

6 考察

6.1 提案手法 (英辞郎) と提案手法 (鳥バンク)

人手評価において提案手法 (鳥バンク) の英日翻訳結果はベースラインより優れていた．一方，自動評価においてスコアが低かった．

人手評価において提案手法 (鳥バンク) が優れていた原因として，鳥バンクの対訳フレーズ対は単文コーパス及び重文複文コーパスと分野が同じであるため対訳フレーズデータを多く作成できた．その結果，信頼性の高いフレーズテーブルを自動的に作成することが出来たためであると考えている．

また，人手評価において提案手法 (英辞郎) はベースラインの翻訳品質とさほど変わらなかった．しかし，自動評価においてスコアが向上した．この原因として，人手で作られた対訳フレーズ辞書を用いて作成した信頼性の高いフレーズテーブルは，ベースラインで作成したフレーズテーブルに既に存在する場合が多く，そのため殆どの人手評価で提案手法の翻訳品質が向上しなかったと考えている．

6.2 人手評価と自動評価

提案手法 (英辞郎) は人手評価においてベースラインの翻訳品質と比べてさほど変わらなかったが，自動評価において翻訳精度の向上が確認できた．一方，提案手法 (鳥バンク) は人手評価において日英統計翻訳はベースラインより優れていたが，自動評価においてベースラインよりスコアが低くかった．自動評価で優れていると判断した実験結果は人手評価において優れている場合もあった．

今後自動評価の問題点もふまえて提案手法 (英辞郎) と提案手法 (鳥バンク) の更なる調査を考えている．

7 おわりに

本実験では人手で作成された辞書として英辞郎を用いた提案手法 (英辞郎) と鳥バンクを用いた提案手法 (鳥バンク) の翻訳実験を行った．また，単文コーパスと重文複文コーパスを用い，それぞれ日英統計翻訳と英日統計翻訳を行った．したがって，合計 8 種類の翻訳実験を行った．

その結果，人手評価において提案手法 (英辞郎) の日英翻訳の重文複文と提案手法 (鳥バンク) の単文と重複文の英日翻訳の評価結果はベースラインより優れており，それ以外の翻訳実験はベースラインより劣っているという結果になった．

一方，自動評価において提案手法 (英辞郎) の全ての実験においては BLEU，METEOR の値がベースラインより向上した．提案手法 (鳥バンク) はほとんどの自動評価においてベースラインよりスコアが低くなった．

これは鳥バンクの対訳フレーズ辞書がフレーズ対が単文コーパス及び重文複文コーパスと分野が同じであるため，効果的な対訳フレーズデータを作成し，良い翻訳結果が得ることができたからと考えている．

今後，提案手法 (英辞郎) と提案手法 (鳥バンク) の更なる調査を行っていきたい．

参考文献

[1] 東江恵介, 村上仁一, 徳久雅人, 池原悟, “日英統計翻訳における英辞郎の効果” , 言語処理学会第 1 6 回年次大会発表論文集,pp.641-644,2010.

[2] 英辞郎 http://www.alc.co.jp/

[3] 鳥バンク http://unicorn.ike.tottori-u.ac.jp/toribank/. 2007.

[4] 西山七絵, 村上仁一, 徳久雅人, 池原悟, “単文文型パターン辞書の構築” , 言語処理学会第 1 1 回年次大会,pp.372-375,2005.

[5] 村上仁一, 池原悟, 徳久雅人, “日本語英語の文対応の対訳データベースの作成” ,「言語, 認識, 表現」第 7 回年次研究会,2002.

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

[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] 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] 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.

[10] NIST,Automatic Evaluation of Machine Translation Quality Using n-gram Co-Occurence Statistics http://www.itl.nist.gov/

[11] Lavie, Alon and Denkowski, Michael. “The METEOR Metric for Automatic Evaluation of Machine Translation” , 2009.