句に基づく対訳文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

江木孝史 村上仁一 徳久雅人
鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
{s082008, murakami, tokuhisa} @ ike.tottori-u.ac.jp

1　はじめに

パターン翻訳は古典的な方法であり，古くから研究が行われている．しかし，パターン翻訳には多くの問題点がある．本研究ではコストとカバー率と翻訳精度の問題を取り上げる．

パターン翻訳は対訳単語辞書と対訳文パターンを用いて翻訳文を出力する [1] 方法である．この対訳単語辞書と対訳文パターン辞書は人手で作成する．このため開発にコストと時間がかかる [2].

また，パターン翻訳は一般的にカバー率が低くなる傾向がある．このため，カバー率向上のために対訳文パターンを汎化させる必要がある．しかし対訳文パターンを汎化させた場合は翻訳精度が低下する傾向がある．

本研究ではこれらの問題点を解決するために，統計的手法を用いて大量の対訳文パターンを作成し，英日パターン翻訳を行う．そして提案手法の有効性を調査する．

2　パターン翻訳

2.1　パターン翻訳の概略

パターン翻訳 [1] は 1960 年代に提案された翻訳方法であり，古典的な方法である．パターン翻訳には対訳単語辞書と対訳文パターンが必要であり，通常は人手で作成する．パターン翻訳は入力文が適切な文パターンに適合した場合，精度の高い翻訳文が得られる．以下に一般的な英日パターン翻訳 [3] の手順を示す．

手順 1　対訳単語辞書と対訳文パターンを用意する．
手順 2　入力文と原文パターンを照合する．
手順 3　目的語文パターンと対訳単語辞書を用いて翻訳文を生成する．

図 1 に英日パターン翻訳の例を示す．

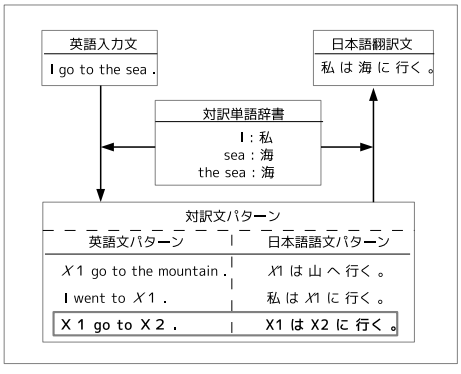

図 1　英日パターン翻訳

2.2　パターン翻訳の問題点

パターン翻訳には多くの問題点があるが，本論文ではコストとカバー率と翻訳精度の問題を取り上げる．

2.2.1　コスト

対訳単語辞書と対訳文パターン辞書は通常，人手で作成する．そのため開発にコストと時間がかかる．

2.2.2　カバー率

パターン翻訳は入力文に文パターンが適合しなければ翻訳が不可能である．そのため，従来のパターン翻訳ではカバー率向上のために，汎化させた対訳文パターンを用いる．しかし，通常汎化させた対訳文パターンを用いると，翻訳精度が低下する傾向にある．

2.2.3　翻訳精度

汎化させた対訳文パターンを用いると，しばしば文法的に誤った対訳文パターンが適合する．そのため，品質の低い翻訳文が出力される．

3　GIZA++

現在，機械翻訳において統計翻訳の研究が行われている．統計翻訳が提案された当初は単語に基づく統計翻訳であった．単語に基づく統計翻訳は IBM モデル [4] を基にしている．GIZA++ は IBM モデルを用いて，原言語と目的言語の対訳文から対訳単語と単語翻訳確率を自動的に得るツールである [5].

4　提案手法

4.1　提案手法の概略

従来のパターン翻訳は 2.2 節に示す問題点がある．以下に各問題点に対する具体的な解決策を示す．

4.1.1　コスト

開発コスト削減のために，GIZA++ を用いて対訳学習文から対訳単語辞書と対訳文パターンを自動的に作成する．

4.1.2　カバー率

カバー率の向上を図るために，大量の対訳単語辞書と対訳文パターンを生成する．

4.1.3　翻訳精度

翻訳精度を向上させるために，翻訳時に英語入力文と文パターンの字面を比較する．そして最も多く字面が一致する文パターンを優先して選択する．

本研究では 5 つのステップを用いて英日パターン翻訳を行う．以下に手順を示す．

手順 1 対訳単語辞書
GIZA++ を用いて，対訳単語辞書を作成する．
手順 2 単語に基づく対訳文パターン辞書
対訳単語辞書を用いて，単語に基づく対訳文パターン辞書を作成する．
手順 3 対訳フレーズ辞書
単語に基づく対訳文パターン辞書を用いて，対訳フレーズ辞書を作成する．
手順 4 句に基づく対訳文パターン辞書
対訳フレーズ辞書を用いて，句に基づく対訳文パターン辞書を作成する．
手順 5 英日パターン翻訳
対訳フレーズ辞書と句に基づく対訳文パターン辞書を用いて，英日パターン翻訳を行う．

各手順の詳細は 4.2 節で説明する．なお，単語に基づく英日パターン翻訳の実験結果は [7] で報告している．

4.2　提案手法の詳細

手順 1 対訳単語辞書

対訳単語辞書を作成するために，対訳学習文と GIZA++ [5] を用いる．本研究では GIZA++ で得た確率を対訳単語確率と呼ぶ．また，対訳単語辞書は対訳

Copyright(C) 2014 The Association for Natural Language Processing.
All Rights Reserved.

単語確率から得る．図 2 に対訳単語辞書の作成手順と例を示す．

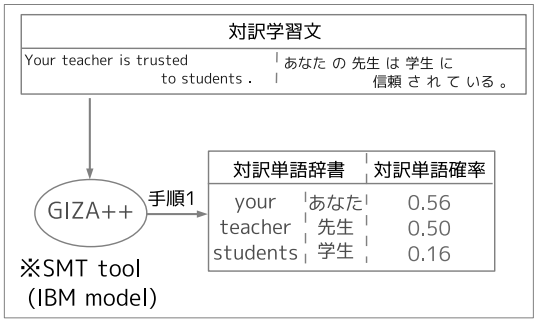

図 2　対訳単語辞書の作成手順

手順 2 単語に基づく対訳文パターン辞書

単語に基づく対訳文パターン辞書を作成するために，対訳単語辞書 (手順 1) と対訳学習文を用いる．

なお，大量の句に基づく対訳文パターンを生成させるため，可能な限り単語に基づく対訳文パターンを生成する．具体的には変数化するときに，変数の組み合わせを考慮し，可能な限り多くの単語に基づく対訳文パターンを生成する．

図 3 に単語に基づく対訳文パターン辞書の作成手順と例を示す．

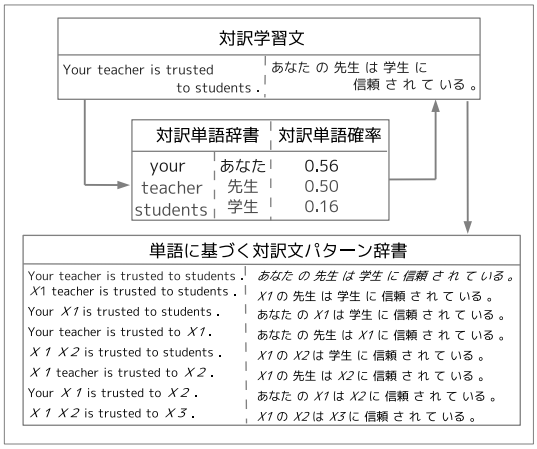

図 3　単語に基づく対訳文パターン辞書の作成手順

図 3 において変数化される対訳単語は“your | あなた”，“teacher | 先生”，“students | 学生”である．これら 3 つの対訳単語が変数化される場合とされない場合の組み合わせを全て考慮し，2^3=8 通りの単語に基づく対訳文パターンを生成する．

手順 3 対訳フレーズ辞書

対訳フレーズを抽出するために，単語に基づく対訳文パターン辞書 (手順 2) と対訳学習文を用いる．

1. パターン照合
   対訳学習文と単語に基づく対訳文パターン辞書を照合する．
2. 対訳フレーズの抽出
   対訳学習文が単語に基づく対訳文パターンに適合した場合，単語に基づく対訳文パターンの変数部に対応する対訳フレーズを抽出する．また，本研究では対訳フレーズの英語側を英語フレーズ，日本語側を日本語フレーズと呼ぶ．

   図 4 に対訳フレーズの抽出手順と例を示す．

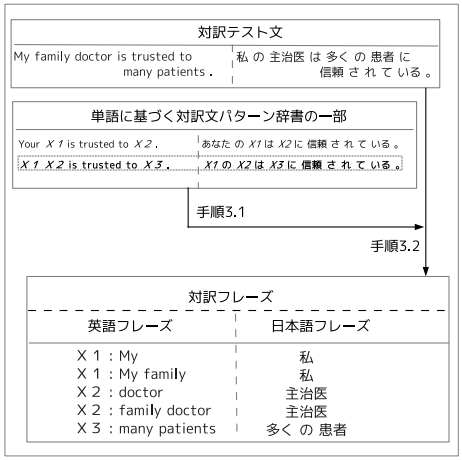

図 4　対訳フレーズの抽出手順

3. 対訳フレーズの翻訳確率の計算
   対訳フレーズの翻訳確率を計算する．以下に手順を示す．

   A. 単語の組み合わせの取得
   対訳フレーズにおいて，英語フレーズの単語と日本語フレーズの単語の全ての組み合わせを得る．

   B. 翻訳確率の計算
   各英単語に対応する日本語単語の中で，GIZA++ の単語翻訳確率が最大となる対訳単語確率を得る．

   C. 翻訳確率の付与
   得られた対訳単語確率に対して対数を取り，総和を求める．総和を対訳フレーズの翻訳確率として付与する．本研究では計算した確率を対訳フレーズ確率と呼ぶ．

図 5 に翻訳確率の計算の方法と例を示す．

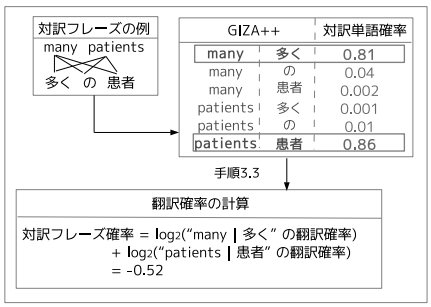

図 5　対訳フレーズ確率の付与手順

図 5 に対訳フレーズの例として“many patients | 多くの 患者” を示す．まず，英語フレーズの単語と日本語フレーズの単語の全ての組み合わせを得る．“many”には“多く”，“の”，“患者”が対応する．GIZA++ の対訳単語確率を用いて，各組み合わせの中から最大となる対訳単語確率を得る．図 5 では “many | 多く” に付与された確率“0.81”が最も高いため，0.81 に対して対数を取る．“patients”も同様に単語翻訳確率に対数を取り総和を求める．

手順 4 句に基づく対訳文パターン辞書

句に基づく対訳文パターン辞書を作成するために，対訳フレーズ辞書 (手順 3) と対訳学習文を用いる．

1. 句に基づく対訳文パターン辞書の作成

Copyright(C) 2014 The Association for Natural Language Processing.
All Rights Reserved.

対訳フレーズが照合に成功した場合，該当箇所を変数化し，対訳文パターンを生成する．変数化するときに，変数の組み合わせを考慮し，可能な限り多くの句に基づく対訳文パターンを生成する．本研究では，句に基づく対訳文パターンの英語側を英語フレーズ，日本語側を日本語フレーズと呼ぶ．

図 6 に句に基づく対訳文パターン辞書の作成手順と例を示す．なお，プログラムの実装には，CYK アルゴリズムを用いる．

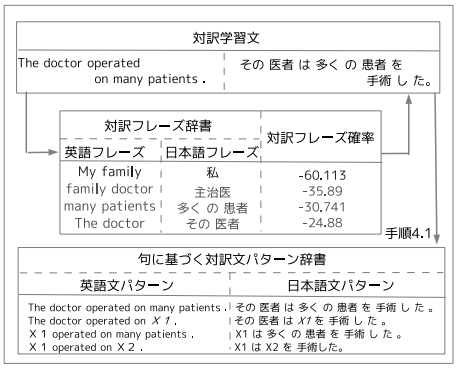

図 6 句に基づく対訳文パターン辞書の作成手順

図 6 において変数化される対訳単語は“The doctor | その 医者”，“many patients | 多く の 患者”である．この 2 つの対訳フレーズが変数化される場合とされない場合の組み合わせを全て考慮し，2^2=4 通りの句に基づく対訳文パターンを生成する．

2. 句に基づく対訳文パターンの翻訳確率の付与
対訳文パターンの字面と GIZA++ の単語翻訳確率を用いて，対訳文パターンに翻訳確率を付与する．翻訳確率の付与は，手順 3.3 で説明した対訳フレーズ確率の付与と同じ手法を用いる．本研究では計算した確率を対訳文パターン確率と呼ぶ．

図 7 に，対訳文パターン確率の付与手順と例を示す．

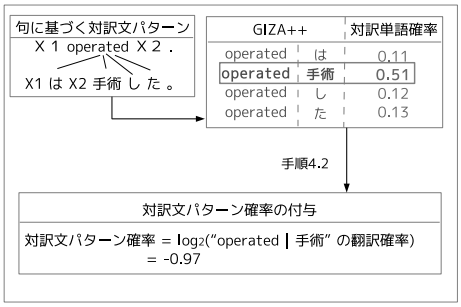

図 7 対訳文パターン確率の付与手順

手順 5 英日パターン翻訳

日本語翻訳文を出力するために，対訳フレーズ辞書 (手順 3) と句に基づく対訳文パターン辞書 (手順 4) を用いる．

翻訳精度を向上させるために，翻訳時に英語入力文と英語文パターンの字面を比較する．そして最も多く字面が一致する英語文パターンを優先して選択する．

日本語翻訳文の絞り込みには対訳フレーズ確率と対訳文パターン確率と日本語翻訳文の tri-gram スコアを用いる．総和を取り，確率が最大となる日本語翻訳文を出力する．以下に英日パターン翻訳の手順を示す．

1. 英語文パターンの選択
英語入力文を読み込み，英語入力文と英語文パターンの字面を比較する．そして最も多く字面が一致する英語文パターンを優先して選択する．
2. 英語フレーズの取得
一致する英語文パターンの変数部に対応する英語フレーズを得る．
3. 日本語文パターンの取得
英語文パターンに対応する日本語文パターンと対訳文パターン確率を得る．
4. 日本語フレーズの取得
日本語文パターンの変数部に対応する日本語フレーズと対訳フレーズ確率を得る．
5. 日本語翻訳文の生成
日本語文パターンの変数部を手順 5.4 の日本語フレーズに置き換える．そして，日本語翻訳文として出力する．
6. tri-gram スコアの算出
手順 5.5 の日本語翻訳文に対して tri-gram スコアを計算する．
7. 日本語翻訳文の選択
対訳フレーズ確率と対訳文パターン確率と日本語翻訳文の tri-gram スコアの総和を求め，日本語翻訳文に付与する．最後に総和が最大となる日本語翻訳文を出力する．

図 8 に日本語翻訳文を出力するまでの手順を示す．なお，実装したプログラムは，高速化を図るために viterbi アルゴリズムと CYK アルゴリズムを用いる．

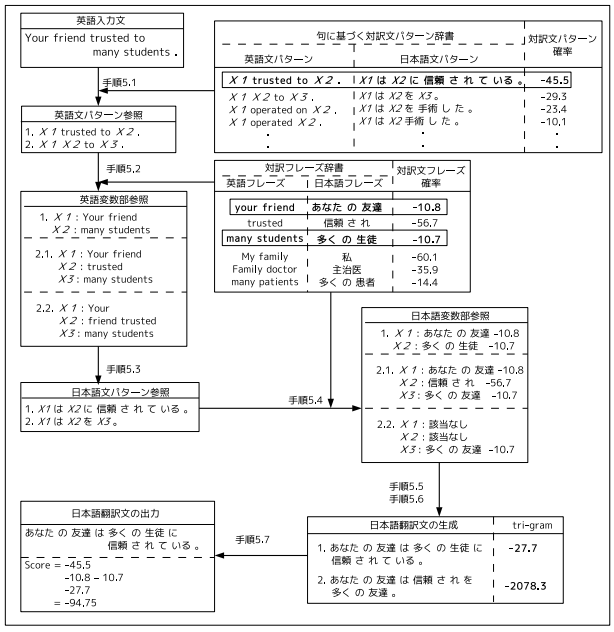

図 8 日本語翻訳文の生成手順

5 実験

実験には電子辞書から抽出した単文データベースを用いる [8]．なお，実験に使用した単文データは日本語文が単文であるが，英語文は単文とは限らず，重文・複文が含まれる．

5.1 実験条件

以下に実験条件を示す．

手順 1 対訳単語辞書

- 対訳学習文 100,000 文を用いる．
- 単語に基づく対訳文パターンの出力数を抑制するため，閾値を 0.1 とする．

手順 2 単語に基づく対訳文パターン辞書

- 対訳学習文 100,000 文を用いる．

Copyright(C) 2014 The Association for Natural Language Processing.
All Rights Reserved.

手順 3 対訳フレーズ辞書

- 対訳学習文 100,000 文を用いる．
- 句に基づく対訳文パターンの出力数を抑制するため，対訳フレーズの選別には以下の条件を用いる．
  - 閾値を −100.0 とする．
  - 英語フレーズの単語数を基準とし，対応する日本語フレーズの単語数が± 5 単語以内の対訳フレーズのみを選別する．

手順 4 句に基づく対訳文パターン辞書

- 対訳学習文 1 文に対し，対訳文パターンの出力数は最大 100,000 文対までとする．
- 句に基づく対訳文パターンの作成には，手順 3 で作成した対訳フレーズ辞書を用いる．ただし，1 つの英語フレーズに対して，付与された対訳フレーズ確率が高い上位 2 つの日本語フレーズを抽出して利用する．

手順 5 英日パターン翻訳

- 英日パターン翻訳は入力文として対訳テスト文 100 文を用いる．
- 句に基づく対訳文パターンの作成には，手順 3 で作成した対訳フレーズ辞書を用いる．ただし，1 つの英語フレーズに対して，付与された対訳フレーズ確率が高い上位 512 の日本語フレーズを抽出して利用する．
- 英語文パターンの選択
  - 英語文パターンを選択する際は，英語入力文と英語文パターンの字面を比較し，最も多く字面が一致する英語文パターンを優先して選択する．
  - 英語入力文 1 文に対し，英語文パターンの選択数は 1,000 文までとする．
- tri-gram スコアの計算
  - 日本語翻訳文に対して tri-gram スコアを計算する．
  - tri-gram は対訳学習文の日本語文 100,000 文を用いる．
- 日本語翻訳文
  - 翻訳精度の低い出力文を除外するために，閾値を −1,000.0 とする．

5.2 実験結果

表 1 に 4.2 節の各手順で得た単語数，対訳文パターン数，フレーズ数，日本語翻訳文数を示す．

表 1 4.2 節で得たデータ数

対訳単語辞書	22,574 単語
単語に基づく対訳文パターン辞書	1,464,042 文対
対訳フレーズ辞書	4,128,831 フレーズ
句に基づく対訳文パターン辞書	175,087,300 文対
英日パターン翻訳	24 文

6 評価結果

24 文の日本語翻訳文に対し，人手による対比較評価を行った．ベースラインシステムには Moses [9] を用いる．表 2 に提案手法とベースラインの人手較評価結果を示す．

表 2 提案手法とベースラインの人手評価結果

提案手法○	提案手法×	差なし	同一出力
6	3	10	5

6.1 人手評価における提案手法○の例

提案手法○の例を表 3 に示す．

表 3 提案手法○の例

英語入力文	The signal changed from green to red .
英語文パターン	The $X00$ $X01$ from $X02$ to $X03$.
日本語文パターン	$X00$ が $X02$ から $X03$ に $X01$ 。
提案手法	信号 が 青 から 赤 に なっ た 。
ベースライン	信号 が 赤 に 緑 を 変え た 。
正解文	信号 が 青 より 赤 に 変わっ た 。

6.2 人手評価における提案手法× の例

提案手法×の例を表 4 に示す．

表 4 提案手法×の例

英語入力文	The quarrel lasted for years .
英語文パターン	The $X00$ lasted for $X01$.
日本語文パターン	その $X00$ $X01$ 続い た 。
提案手法	その けんか で も 何 年 も 続い た 。
ベースライン	その けんか は 何 年間 も 続い た 。
正解文	その けんか は 何 年 も 糸 を 引い た 。

6.3 評価結果のまとめ

表 2 の結果から，提案手法はベースラインと比較して優れている文が多いことがわかる．よって，評価結果から提案手法の有効性が確認された．

7 考察

6 章より，提案手法はベースラインよりも優れていることがわかった．しかし，人手評価結果を解析したところ，提案手法×の日本語翻訳文を 3 文得た．この 3 文に対して誤り解析を行ったところ，3 文中 2 文が不適切な英語文パターンの選択による誤訳であった．

以下に表 4 における誤り解析の結果を述べる．表 4 で誤った翻訳文が出力された原因として，不適切な英語文パターンの選択が挙げられる．仮に対訳文パターンとして “$X00$ lasted for $X01$. | その $X00$ $X01$ 続い た 。” が選択されたとすると，適切な日本語翻訳文が出力される．さらに上記の対訳文パターンは，本実験で用いた句に基づく対訳文パターン辞書に含まれている．

以上から英語文パターンの選択方法の改善により，正しい日本語翻訳文を得られる可能性がある．具体的には各辞書に用いる閾値の調整や対訳文パターン確率の計算方法，英語文パターン選択数の見直しが必要である．

8 まとめ

本研究では，句に基づく対訳文パターンをプログラムで自動的に作成し，統計的手法を用いて英日パターン翻訳を行った．実験の結果，英語入力文 100 文から日本語翻訳文 24 文を得た．提案手法を評価するために，Moses をベースラインとし，提案手法との対比較評価を行った．対比較評価の結果，提案手法○が 6 文，提案手法×が 3 文であり，提案手法の有効性が認められた．

また，提案手法×の日本語翻訳文に対して誤り解析を行った．誤り解析の結果，不適切な英語文パターンの選択が原因であることがわかった．今後は閾値の調整や対訳文パターン確率の計算方法，英語文パターン選択数の見直しにより，さらに翻訳精度が向上すると考えている．

参考文献

[1] Hiroshi Maruyama: “Pattern-Based Translation: Context-Free Transducer and Its Applications to Practical NLP”, n Proc.of Natural Language Pacific Rim Symposium, pp.232-237, 1993.
[2] 池原 悟, 他: “日本語語彙大系”, 岩波書店, 1997.
[3] 池原悟, 他: “透過的類推思考の原理による機械翻訳方式”, 電子情報通信学会技術研究報告, pp.7-12, 2002.
[4] Peter F. Brown, et al.,: “The Mathematics of Statistical Machine Translation: Parameter Estimation”, Computational Linguistics, pp.263−311, 1993.
[5] Franz Josef Och, Hermann Ney: “A Systematic Comparison of Various Statistical Alignment Models”, Computational Linguistics, pp.19-51, 2003.
[6] 春野瑞季, 他: “文パターンを用いた句の抽出方法の検討”, 言語処理学会第 19 回年次大会, pp.741-744, 2013.
[7] 江木孝史, 他: “統計的手法を用いた英日パターン翻訳”, 言語処理学会第 18 回年次大会, pp.263-266, 2012.
[8] 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.
[9] Philipp Koehn, et al.,:“Moses: Open Source Toolkit for Statistical Machine Translation”, Proceedings of the ACL 2007 Demo and Poster Sessions, pp.177-180, 2007.

Copyright(C) 2014 The Association for Natural Language Processing.
All Rights Reserved.