文単位のパターンを用いた統計翻訳

西村拓哉 村上仁一 徳久雅人 池原悟
鳥取大学 工学部 知能情報工学科
{s062044, murakami, tokuhisa, ikehara} @ ike.tottori-u.ac.jp

1 はじめに

現在，機械翻訳において統計翻訳が注目され，研究が盛んに行われている [1]．統計翻訳では一般に N-gram モデルを用いる．N-gram モデルは局所的な文法情報であるため，特異な翻訳文が生成される場合がある．そこで本研究では，まず日英文パターン辞書を用いて日英パターン翻訳を行い，文法構造を英語に近づける．そして出力文に対し，統計翻訳でさらに英英翻訳を行う．この日英文パターンが持つ大局的な文法情報を用いることで N-gram モデルにおける局所的な構文問題が解消でき，翻訳精度の向上が可能であると考えられる．

2 翻訳システム

2.1 パターン翻訳の基本概念

日英パターン翻訳は，まず日本語入力文 j が与えられたとき，日英文パターン辞書と日英単語辞書を参照する．次に，日本語入力文に適合する日本語文パターンと英語文パターンの変数部を単語に置き換えることで翻訳を行う．尚，従来のパターン翻訳では日英文パターン辞書を人手で作成する．そのため，開発に時間がかかるが，日英文パターンが適合した場合に翻訳精度の高い翻訳文が得られる．図 1 に日英パターン翻訳の手順を示す．

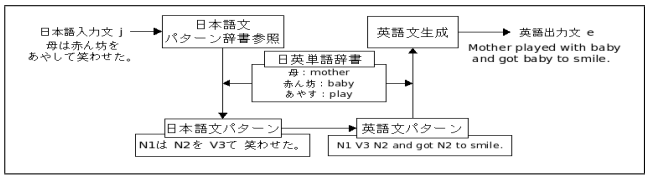

図 1 日英パターン翻訳の手順

2.2 統計翻訳の基本概念

日英統計翻訳は，日本語入力文 j が与えられたとき，全ての組合せから確率が最大値となる英語文 $\hat{e}$ を探索して翻訳を行う．

$$\hat{e} = \arg\max_{e} P(e|j)$$
$$\approx \arg\max_{e} P(j|e)P(e)$$

$P(j|e)$ は翻訳モデル，$P(e)$ は言語モデルと呼ぶ．図 2 に統計翻訳の手順を示す．

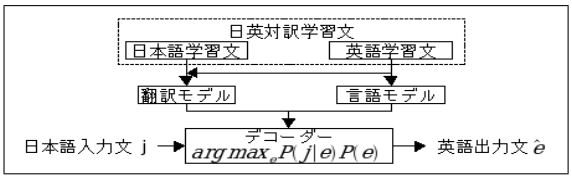

図 2 日英統計翻訳の手順

2.2.1 翻訳モデル

翻訳モデルは英語から日本語の単語列へ確率的に翻訳を行うためのモデルである．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例を表 1 に示す．

左から，日本語フレーズ，英語フレーズ，日英フレーズ内単語対応，英日のフレーズ内単語対応，フレーズの日英方向の翻訳確率，日英の単語翻訳確率の積，フレーズの英日方向の翻訳確率，英日の単語翻訳確率の積である．

表 1 フレーズテーブルの例

1 0 パーセント ||| a ten percent ||| (0,1) (1) (2) ||| (0) (0,1) (2) ||| 1 0.02 0.25 0.005
あなた と 私 ||| you and I ||| (0) (1) (2) ||| (0) (1) (2) ||| 0.5 0.01 0.5 0.03

2.2.2 言語モデル

言語モデルは単語列の生じる確率を与えるモデルである．日英翻訳では翻訳モデルで生成された翻訳候補から英語として自然な文を選出する．統計翻訳では一般に，N-gram モデルを用いる．

3 翻訳システム

3.1 本研究の翻訳システムの概要

統計翻訳では言語モデルとして，一般に N-gram モデルを用いる．しかし N-gram モデルは局所的な文法情報をモデル化するので，大局的な文法情報がない．したがって，文法構造が大きく異なる言語間では特異な文章が生成される可能性がある．

そこで本研究では，まず統計翻訳を行う前にパターン翻訳を行う．この処理により，日英文パターンが有する大局的な文法情報を用いるので N-gram モデルの局所的な構文問題が解消出来ると考えている．次に，パターン翻訳の出力文に対し統計翻訳を行う．この処理により，局所的な修正を行うことで翻訳精度が向上すると考えている．

また本研究では，日英文パターン辞書を人手で作成するのではなく，プログラムで自動的に作成する．この手法により，日英文パターン辞書の開発時間が短縮出来る．図 3 に提案手法の翻訳手順を示す．

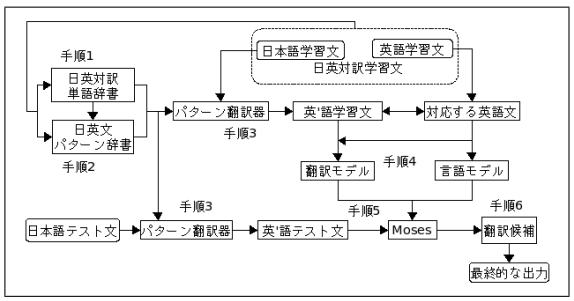

図 3 提案手法の翻訳手順

3.2 本研究の翻訳システムの手順

本研究の翻訳システムの手順を以下に示す．

手順 1 日英対訳単語辞書の作成

まず日英対訳学習文から GIZA++[2] を用いて日英単語辞書と英日単語辞書を作成する．次に両辞書における各単語の確率を掛け合わせる．そして閾値以上の確率を持つ単語を用いて日英対訳単語辞書を作成する．表 2 に日英対訳単語辞書の例を示す．

手順 2 日英文パターン辞書の作成

日英対訳単語辞書を用いて日英対訳学習文から日英文パターン辞書を自動的に作成する．図 4 に日英文パターン辞書の作成手順を示す．

表 2 日英対訳単語辞書の例

日本語	英語	確率
彼	He	0.4
彼女	She	0.5
ハワイ	Hawaii	0.4

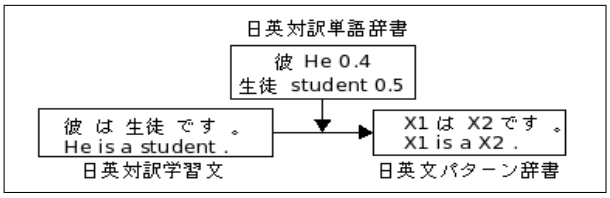

図 4 日英文パターン辞書の作成手順

日英文パターン辞書は日英対訳単語辞書を参照し，日英対訳学習文中で適合する単語を変数化して作成する．図 4 において，日英対訳学習文中にある単語 “彼” と “He” が日英対訳単語辞書中にある．したがって，両者を変数 “X1” に置換する．同様に “生徒” と “student” も変数 “X2” として置換する．以上の処理を日英対訳学習文全てに対して行い，日英文パターン辞書を作成する．

手順 3 パターン翻訳

手順 1 の日英対訳単語辞書と手順 2 の日英文パターン辞書を用いて，日本語テスト文と日本語学習文に対してパターン翻訳を行う．パターン翻訳を行う際，日本語入力文中の単語と日英対訳単語辞書中の単語で，対応する単語が複数ある場合には，全ての組合せを翻訳候補として出力する．次に各翻訳候補で使用した単語の確率を掛け合わせ，翻訳候補の中で確率が最も高い候補文を選択する．以後，選択した文を英’ 語出力文とする．また入力文 1 文に対して複数のパターンに適合する場合，各文パターンにつき 1 文を出力する．図 5 にパターン翻訳の例を示す．

図 5 の例では，日本語入力文に対して 2 つの日英文パターンが適合する．まず，日本語入力文の “彼女” に対応する単語が日英対訳単語辞書に 2 つあり，“先生” に対応する単語が 1 つあるので，文パターン 1，文パターン 2 においてそれぞれ 2 文が翻訳候補として出力される．次に，翻訳候補が使用した単語の確率を掛け合わせる．そして掛け合わせた確率を使用し，最も確率が高い翻訳候補を選択する．文パターン 1 では “出力文 1a” を，文パターン 2 では “出力文 2a” を選択し，2 つの出力文を英’ 語出力文とする．以上の処理と同様にして，入力文全てに対してパターン翻訳を行う．

尚，提案手法のパターン翻訳において，次の場合にはパターン翻訳の出力をしない．

- 日本語入力文が日英文パターンに適合しない
- 日英対訳単語辞書を参照する時に適合する単語が日英対訳単語辞書に登録されていない

また日英文パターンに適合しない日本語テスト文に関しては，3.3 章で述べるベースラインシステムと同様の翻訳を行う．

手順 4 統計翻訳の翻訳モデルと言語モデルの学習

学習データには，日本語学習文のパターン翻訳で出力された英’ 語学習文と，その英’ 語学習文に対応する英語学習文を用いる．この英’ 語学習文と英語学習文を用いて翻訳モデルを，英語学習文を用いて言語モデルを学習する．

手順 5 統計翻訳における英語文生成

本研究における統計翻訳のデコーダには Moses[3] を用いる．手順 4 で学習した翻訳モデルと言語モデルを用いて，手順 3 の英’ 語テスト文に英’ 英統計翻訳を行う．図 6 に統計翻訳の例を示す．

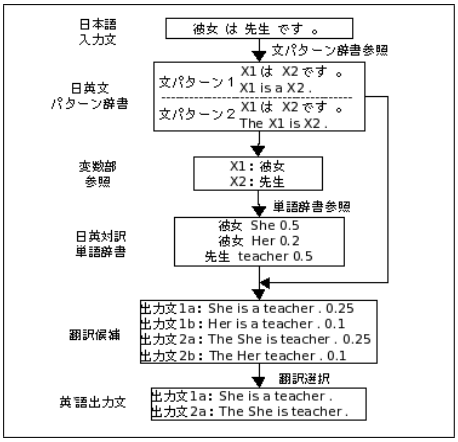

図 5 パターン翻訳の例

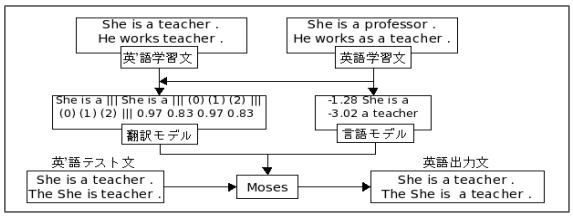

図 6 統計翻訳の例

手順 6 翻訳候補の選択

手順 3 で得られた英’ 語テスト文には，入力文 1 文に対して複数の翻訳候補がある．そこで，複数の翻訳候補の中から手順 5 の統計翻訳の出力時における確率が最も高い文を最終的な出力として 1 文選択する．

3.3 ベースラインシステム

Moses を用いた日本語テスト文から英語出力文への翻訳を行うシステムをベースラインシステムと呼ぶ．翻訳モデルの学習には日英対訳学習文を，言語モデルの学習には英語学習文を用いる．

4 実験環境

実験データには，辞書の例文から抽出した日英対訳文の単文のみを用いる．学習文には 100,000 文，テスト文には 10,000 文を用いる．また前処理として，日本語文に対しては chasen[4] を用いて形態素解析を行い，形態素と句読点の間にスペースを入れる．英語文に対してはコンマ，ピリオドの前後にスペースを入れる．表 3 に日英対訳文の例を示す．

表 3 日英対訳文の例

日本語文	もっと 右 へ 寄っ て ください 。
英語文	Please move over more to the right .
日本語文	この 傷 は 至急 手当 を せ ね ば なら ない 。
英語文	The wound requires prompt treatment .

4.1 言語モデルの学習

統計翻訳に用いる言語モデルには N-gram モデルを用いる．本研究では，SRILM[5] の “ngram-count” を用いて 5-gram の言語モデルを学習する．尚，スムージングに “-ndiscount” を用いる．

4.2 デコーダのパラメータ

過去の研究 [6] から，翻訳モデルには日英翻訳確率と英日翻訳確率の共起確率を用いる．したがって，フレーズテーブルの各種の重み “weight-t” を “0.5 0.0 0.5 0.0 0.0” とする．また翻訳時にフレーズの位置の変化を柔軟に対応するために “distortion-weight” を “0.2” とする．尚，本研究ではパラメータチューニング [7] は行わない．

5 翻訳実験

5.1 パターン翻訳部の実験

日本語テスト文 10,000 文と日本語学習文 100,000 文のパターン翻訳には，同じ日英対訳単語辞書と日英文パターン辞書を用いて行う．提案手法で用いる日英対訳単語辞書と日英文パターン辞書の作成手順を以下に示す．

手順 1 GIZA++ を用いて日英対訳学習文 100,000 文から日英単語辞書と英日単語辞書を作成する．

手順 2 手順 1 の日英単語辞書と英日単語辞書の各単語の確率を掛け合わせ，閾値以上の確率を持つ単語を用いて日英対訳単語辞書を作成する．

手順 3 日英対訳学習文 100,000 文と手順 2 の日英対訳単語辞書を用いて日英文パターン辞書を作成する．

日英対訳単語辞書を作成する時の閾値について，今回の実験では 0.25，0.02，0.01 の 3 つの場合で日英対訳単語辞書を作成する．表 4 に各閾値における日英対訳単語辞書の登録単語数を示す．

表 4 日英対訳単語辞書の登録単語数

閾値	登録単語数
0.25	3,033
0.02	34,587
0.01	53,185

5.2 パターン翻訳の実験結果

実験は以下の 2 つの実験条件で行う．

実験条件 1

閾値が 0.25 以上の日英対訳単語辞書を用いて日英文パターン辞書を作成し，その日英文パターン辞書と閾値が 0.01 以上の日英対訳単語辞書を用いてパターン翻訳を行う

実験条件 2

閾値が 0.01 以上の日英対訳単語辞書を用いて日英文パターン辞書を作成し，その日英文パターン辞書と閾値が 0.02 以上の日英対訳単語辞書を用いてパターン翻訳を行う

日本語学習文での出力文数を表 5 に，日本語テスト文での出力文数を表 6 に示す．尚，表中における “適合した文数” は 10,000 文中でパターン翻訳が行われた日本語入力文の数を示し，“出力文数” は適合した文数から得られた英’ 語出力文の数を表している．

表 5 日本語学習文での適合した文数と出力文数

	適合した文数	出力文数
実験条件 1	47,053	51,441
実験条件 2	80,988	728,819

表 6 日本語テスト文での適合した文数と出力文数

	適合した文数	出力文数
実験条件 1	362	566
実験条件 2	2,103	61,467

表 6 の結果から，実験条件 1 では日本語テスト文の適合した文数が 362 文であるのに対し，実験条件 2 では 2,103 文と大きく増加している．

5.3 提案手法における統計翻訳部の実験

翻訳モデルの学習には表 5 の英’ 語学習文と英’ 語学習文に対応する英語学習文を，言語モデルの学習には英’ 語学習文に対応する英語学習文を用いる．この言語モデルと翻訳モデルを用いて，表 6 の英’ 語テスト文に対して英’ 英統計翻訳を行う．

5.4 ベースラインシステムにおける実験

ベースラインシステムの翻訳モデルの学習には日英対訳学習文 100,000 文を，言語モデルの学習には英語学習文 100,000 文を用いる．この翻訳モデルと言語モデルを用いて，日本語テスト文 10,000 文に対して日英統計翻訳を行う．

6 実験結果

日本語テスト文 10,000 文での結果と，その 10,000 文における表 6 の適合した文数での結果を各実験条件ごとに示す．尚，実験条件 1 と実験条件 2 では実験環境が異なるので，ベースラインの BLEU と NIST の値も異なっている．

6.1 自動評価

出力文の評価は自動評価法の BLEU[8] と NIST[9] と METEOR[10] を用いる．尚，本研究では入力文 1 文に対して正解文 1 文を用いて評価を行う．

6.1.1 実験条件 1 での評価結果

表 7 に 10,000 文での評価結果を示す．また表 8 に，表 6 における適合した文数 362 文での評価結果を示す．

表 7 10,000 文での結果

	BLEU	NIST	METEOR
ベースライン	0.102	4.034	0.347
提案手法	0.107	4.150	0.355

表 8 362 文での結果

	BLEU	NIST	METEOR
ベースライン	0.355	5.450	0.564
提案手法	0.376	5.520	0.568

表 7 において，ベースラインと比較して提案手法では BLEU 値が 0.5% 向上していることから，提案手法の有効性が分かる．また表 8 において，ベースラインと比較して提案手法では 2.1% 向上していることから，適合した文での翻訳精度の向上が表 7 の結果に影響し，提案手法での翻訳精度が向上した．

6.1.2 実験条件 2 での評価結果

表 9 に 10,000 文での評価結果を示す．また表 10 に，表 6 における適合した文数 2,103 文での評価結果を示す．

表 9 10,000 文での結果

	BLEU	NIST
ベースライン	0.107	4.157
提案手法	0.103	4.062

表 10 2,103 文での結果

	BLEU	NIST
ベースライン	0.206	5.133
提案手法	0.171	4.499

表 9 において，ベースラインと比較して提案手法では BLEU 値が 0.4% 低下していることから，提案手法が有効でないことがわかる．また表 10 において，ベースラインと比べて提案手法では 3.5% 低下していることから，適合した文での翻訳精度の低下が表 9 の結果に影響し，提案手法での翻訳精度が低下した．

6.2 人手による評価

表 8 と表 10 のそれぞれの出力に対してベースラインとの対比較実験を行う．

6.2.1 判断基準

人手による 4 つの判断基準に基づいて評価を行う．評価基準と評価例を以下に示す．尚，未知語はローマ字変換して評価する．

評価 1(>) 提案手法の翻訳結果がベースラインよりも優れている

入力文 私 は 彼女 に 結婚 を 申し込ん だ 。
正解文 I proposed to her .
ベースライン I He asked her for her hand .
提案手法 I made a proposal of marriage to her .

評価 2(<) 提案手法の翻訳結果がベースラインよりも劣っている

入力文　仕事 は 山場 に 入っ た 。
正解文　Work has reached the critical point .
ベースライン　The work is appear to have entered the final stage .
提案手法　work went into the labor-management .

評価 3(≈)　どちらも似たような文である，またはどちらも入力文で伝えたい情報が理解できない

入力文　豊作 に なり そう だ 。
正解文　The harvest looks promising .
ベースライン　Hopes looks like .
提案手法　It looks like rejoicing .

評価 4(=)　提案手法とベースラインの翻訳結果が全く同じである

入力文　彼 は 故郷 を 恋し がっ て いる 。
正解文　He is sick for home .
ベースライン　He is homesick .
提案手法　He is homesick .

6.2.2　評価結果

各翻訳結果からランダムに抽出した 100 文を対象に評価を行う．評価結果を表 11 に示す．

表 11　対比較実験の結果

	実験条件 1	実験条件 2
評価 1(>)	24 / 100	26 / 100
評価 2(<)	18 / 100	24 / 100
評価 3(≈)	18 / 100	40 / 100
評価 4(=)	40 / 100	10 / 100

実験条件 1 では提案手法の出力文がベースラインよりも優れている文が多い．しかし英' 語テスト文の多い実験条件 2 では実験条件 1 に比べて提案手法がベースラインよりも劣っている文が多い．

7　考察

7.1　日英文パターン辞書における問題点

表 6 の結果から実験条件 1 と比べて実験条件 2 では日本語テスト文の適合数が多い．しかし表 8 の結果，実験条件 1 では提案手法が有効であるが，表 10 の結果から，実験条件 2 では提案手法が翻訳精度の低下を引き起こしている．この原因として，日英文パターン辞書を作成するときに用いた日英対訳単語辞書の閾値が影響している．実験条件 1 の日英文パターン辞書の作成には，閾値が 0.25 以上の登録単語数が少ない日英対訳単語辞書を使用している．したがって，日英文パターン中の変数が少なく，文法情報が多く残され，翻訳精度の高い文が出力されたと考えられる．しかし実験条件 2 の日英文パターン辞書では閾値が 0.01 以上の登録単語数が多い単語を使用している．したがって，日本語テスト文の適合数は増加するが日英文パターン中の変数が多く，複雑になり文法情報が損なわれ，翻訳精度の低い文が出力されたと考えられる．表 12 に日英文パターンの例を示す．

表 12　日英文パターン辞書の例

実験条件 1 の 日英文パターン	私 は X1 を 見 に 行く 。
	I go to see a X1 .
実験条件 2 の 日英文パターン	X1 X2 X3 を 見 X4 X5 。
	X1 X5 X4 see X2 X3 .

実験条件 1 の英語文パターンでは “I” が主語であり “go to see” が複合動詞であることが分かる．しかし実験条件 2 では “see” が動詞であると考えられる．したがって，実験条件 1 では翻訳精度の高い翻訳文が出力されるが実験条件 2 では翻訳精度の低い文が出力されると考えられる．

この問題に対し，今後日英文パターンが文法情報を十分に保持し，なおかつ日本語テスト文の適合数を増やすための方法を考える必要がある．

7.2　パターン翻訳の出力における解析

提案手法での出力の成功例を表 13 に，失敗例を表 14 に示す．尚，表中の “パターン” はパターン翻訳のみの出力文を表す．

表 13　出力の成功例

入力文	彼 は じっと 横 に なっ て い た 。
正解文	He lay without movement .
ベースライン	He intently at the side .
パターン	His lay still .
提案手法	He lay still .

表 14　出力の失敗例

入力文	彼女 は 家庭 の 事情 で 高校 を 中退 した 。
正解文	She quit high school for family reasons .
ベースライン	She 中退 a high school because of family circumstances .
パターン	The Her midyear family a of high in his circumstances .
提案手法	She midyear of the high in his family circumstances of that time .

表 13 の例では，ベースラインの出力文には動詞がなく入力文の意味が理解出来ない．パターン翻訳のみの出力文では動詞はあるが主語が “His” となっている．提案手法では，“His” を “He” に翻訳され入力文の意味が理解できる．

表 14 の例では，ベースラインの出力文には未知語が存在しているが，文全体として入力文の意味が理解できる文となっている．しかしパターン翻訳のみの出力文では動詞が存在せず，提案手法の出力文でも動詞が存在しないので，入力文の意味が理解出来ない．

したがって，失敗例の結果からパターン翻訳における翻訳精度の改善を行う必要があると考えられる．この問題に対し，品詞タグを用いることで動詞の入る場所には動詞のみを置換するルールを用いるなどの方法が考えられる．

8　おわりに

本研究では日本語-英語間における大きく異なる文法構造に着目し，まずパターン翻訳を行うことで文法構造を英語に近づけ，次に統計翻訳を行うことで文法構造に対し，局所的な修正を行う手法を提案した．実験の結果，日英文パターン辞書における変数の個数が少ない場合には，翻訳精度が向上した．しかし変数の個数が多い場合には，翻訳精度が低下する結果となった．この問題に対し，今後日英対訳単語辞書の閾値を考慮する必要がある．またパターン翻訳における翻訳精度を向上するために，パターン翻訳時に品詞を考慮したルールを用いるなどの処理を行い，翻訳精度の高い翻訳候補を得る必要がある．

参考文献

[1] Richard Zens, Franz Josef Och, Hermann Ney “Phrase-based Statistical Machine Translation”, KI 2002, pp35-56, 2002
[2] GIZA++, Training of statistical translation models http://www.fjoch.com/GIZA++.html
[3] Moses, statistical machine translation system http://www.statmt.org/moses/
[4] chasen, 日本語形態素解析器 http://chasen-legacy.sourceforge.jp/
[5] SRILM, The SRI Language Modeling Toolkit http://www.speech.sri.com/projects/srilm/
[6] Jin’ichi Murakami, Masato Tokuhisa, Satoru Ikehara, “Statistical Machine Translation using Large J/E Parallel Corpus and Long Phrase Tables”, International Workshop on Spoken Language Translation 2007, pp.151-155, 2007
[7] Franz Josef och, “Minimum Error Rate Training in Statistical Machine Translation”, Association for Computational Linguistics 2003, pp160-167, 2003
[8] BLEU, NIST Open MT Scoring http://www.itl.nist.gov/iad/894.01/tests/mt/2008/scoring.html
[9] NIST, Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics http://www.itl.nist.gov/iad/mig/tests/mt/2008/scoring.html
[10] METEOR, The METEOR Automatic Machine Translation Evaluation System http://www-2.cs.cmu.edu/ alavie/METEOR/