階層的モデルを用いた機械翻訳のためのフレーズアライメント

Graham Neubig†‡*　渡辺 太郎‡　隅田 英一郎‡　森 信介†　河原 達也†
†京都大学 情報学研究科　‡情報通信研究機構　*日本学術振興会 特別研究員

1　はじめに

フレーズベース統計的機械翻訳 (SMT,[11]) の学習は単語アライメントされていない対訳コーパスを入力とし、スコア付きのフレーズテーブルを出力する。従来法では、フレーズテーブルを2段階で構築する。まず、単語や最小フレーズを対応付ける**単語アライメント**を行なってから、これらを複数の粒度で網羅的に組み合わせる**フレーズ抽出**を行う。長いフレーズで語彙的曖昧性を解消しながら、短いフレーズでスパースなデータに対応することができるため、フレーズベースSMTの強みはこの複数のフレーズ粒度が利用できる枠組みにあると言える。

しかし、このような2段階法では単語アライメントとフレーズ抽出を独立に行うため、翻訳に最適なフレーズテーブルが得られない。DeNeroら [8] はこの問題に対して、**教師あり**モデルを用いて単語アライメントとフレーズ抽出を同時に行い、翻訳結果の精度向上を実現した。

本稿では、複数の粒度でフレーズアライメントを行う**教師なし**モデルを提案する。具体的には、Inversion Transduction Grammar (ITG,[15]) を用いた階層的なモデルを実現した。先行研究 [16] と同様に、簡潔なフレーズテーブルを学習するためにノンパラメトリックベイズ法に基づく確率過程を利用する。この枠組みで、フレーズ分布が自分自身を基底測度に含む再帰的なモデルで複数の粒度のフレーズを学習する。この学習されたフレーズを直接フレーズテーブル構築に利用するため、ヒューリスティックなフレーズ抽出を行わなずに高い翻訳精度が実現できる。

仏英・日英翻訳における評価実験では、提案手法は2段階法と同程度の翻訳精度を実現しながらフレーズテーブルのサイズを大幅に削減できた。

2　フレーズ抽出の確率モデル

統計的機械翻訳は学習コーパス $\langle \mathcal{E}, \mathcal{F} \rangle$ と翻訳したい原言語文 $\boldsymbol{f}$ が与えられた場合、最も確率の高い目的言語文 $\boldsymbol{e}$ を探索する。

$$\hat{\boldsymbol{e}} = \underset{\boldsymbol{e}}{\operatorname{argmax}} P(\boldsymbol{e}|\boldsymbol{f}, \langle \mathcal{E}, \mathcal{F} \rangle)$$

未観測のパラメータ集合 θ があり、θ が与えられた場合、$\boldsymbol{e}$ は学習コーパスと条件付き独立であると仮定し、目的言語文の確率は以下のようになる。

$$P(\boldsymbol{e}|\boldsymbol{f}, \langle \mathcal{E}, \mathcal{F} \rangle) = \int_\theta P(\boldsymbol{e}|\boldsymbol{f}, \theta) P(\theta|\langle \mathcal{E}, \mathcal{F} \rangle) \quad (1)$$

θ がスコア付きのフレーズテーブルであるならば、従来のフレーズベースSMTを利用して $P(\boldsymbol{e}|\boldsymbol{f}, \theta)$ を探索することができるため、本稿ではパラメータの事後確率 $P(\theta|\langle \mathcal{E}, \mathcal{F} \rangle)$ の求め方に着目する。ベイズ則を用いて、事後確率をコーパス尤度とパラメータの事前確率に分解し、

$$P(\theta|\langle \mathcal{E}, \mathcal{F} \rangle) \propto P(\langle \mathcal{E}, \mathcal{F} \rangle|\theta) P(\theta)$$

右側の2つの確率をモデル化する。第3節で従来モデルについて述べ、第4節で提案手法について述べる。

3　従来のITGモデル

近年、フレーズアライメントは研究されており、特にInversion Transduction Grammar（ITG）を利用する先行研究が多い [16, 1]。ITGは同期文脈自由文法の一種で、非終端記号を生成する時に単語の並べ換えを行うことが特徴である [15]。ITG制限を利用することにより計算量を減らし、多項式時間でアライメントの最尤解や周辺確率が計算できる [7]。

あるフレーズペアの生成確率を $P_{flat}(\langle e, f \rangle; \theta_x, \theta_t)$ とし、フレーズペア確率 θ_t と記号確率 θ_x でパラメータ化する。従来のITGモデルは以下の生成過程を利用する：

1. シンボル x を多項式分布 $P_x(x; \theta_x)$ に従って生成する。x が取り得る値は TERM,REG,INV である。
2. x の値に従って：
   (a) $x =$ TERM（終端記号）の場合、フレーズペア確率 $P_t(\langle e, f \rangle; \theta_t)$ に従ってフレーズペアを生成する。
   (b) $x =$ REG（普通非終端記号）の場合、P_{flat} に従ってフレーズペア $\langle e_1, f_1 \rangle$ と $\langle e_2, f_2 \rangle$ を生成し、$\langle e_1 e_2, f_1 f_2 \rangle$ で1つのフレーズペアに融合する。
   (c) $x =$ INV（倒置非終端記号）の場合、(b) と同じように2つのフレーズペアを生成するが、f_1 と f_2 を逆順に並べる：$\langle e_1 e_2, f_2 f_1 \rangle$。

各文に対する P_{flat} の積を取り、コーパス尤度が計算できる。

$$P(\langle \mathcal{E}, \mathcal{F} \rangle|\theta) = \prod_{\langle e,f \rangle \in \langle \mathcal{E}, \mathcal{F} \rangle} P_{flat}(\langle e, f \rangle; \theta).$$

従来のITGモデルを FLAT と呼ぶ。

3.1 ベイズ学習によるモデル化

前節のモデルはそのまま最尤推定で学習できるが、最尤解では非常に長いフレーズペア（1 文 1 フレーズ）が得られてしまう。Zhang ら [16] は簡潔なフレーズ辞書に高い確率を与える事前確率 $P(\theta) = P(\theta_x, \theta_t)$ を利用することで、長いフレーズの問題を解決する。

ここでは、θ_x の事前確率に Dirichlet 分布を利用し、θ_t にはノンパラメトリックベイズ法に基づく Pitman-Yor 過程 [14] を利用する。

$$\theta_t \sim PY(d, s, P_{base}) \tag{2}$$
$$\theta_x \sim Dirichlet(\alpha)$$

Pitman-Yor 過程の割引パラメータ d と強さパラメータ s を Teh[14] と同様に推定する。P_{base} は次節で述べる基底測度（base measure) である。

Pitman-Yor 過程による事前分布を用いる利点は、生成されたフレーズペアを記憶するという確率過程の性質にある。分布から頻繁に生成されるフレーズペアの確率が高くなり、さらに生成されやすくなる（いわゆる「rich-gets-richer 効果」）。Pitman-Yor 過程を用いた学習によって、より少ない、より役に立つフレーズから構成されるフレーズテーブルが構築できる。また、P_t から生成されるフレーズのみが記憶されることに注意する必要がある。FLAT では、終端記号の最小フレーズペアのみが P_t から生成されるため、記憶されるのも最小フレーズペアのみである。

3.2 基底測度

式 (2) の P_{base} はモデルにおけるフレーズペアの事前確率であり、適切に決めることでフレーズのアライメントしやすさに関する事前知識をモデルに組み込める。ここで、P_{base} は対応なしのフレーズ（$|e| = 0$ または $|f| = 0$）を生成するかどうかを一定の確率 P_u で選び[1]、対応なしのフレーズを P_{bu} から生成し、対応ありのフレーズペアを P_{ba} から生成する。

P_{ba} は DeNero ら [5] と同じく以下の形とする。

$$P_{ba}(\langle e, f\rangle) = M_0(\langle e, f\rangle) P_{pois}(|e|; \lambda) P_{pois}(|f|; \lambda)$$
$$M_0(\langle e, f\rangle) = (P_{m1}(f|e) P_{uni}(e) P_{m1}(e|f) P_{uni}(f))^{\frac{1}{2}}$$

P_{pois} は平均長パラメータ λ を持つポアソン分布である。長いフレーズを避けるために、λ に小さい値を利用する[2]。P_{m1} は単語確率に基づく IBM モデル 1 確率である [2]。これを利用することで、フレーズを構成する単語の翻訳確率が高ければフレーズの確率も高くなる。両方向の条件付き確率の相乗平均を利用することで、両モデルが一致するフレーズを優先的にアライメントする [12]。

P_{bu} では、e と f の中から空でない単語列を g とし、確率を以下のように定義する。

$$P_{bu}(\langle e, f\rangle) = P_{uni}(g) P_{pois}(|g|; \lambda)/2.$$

e と f を両方考慮するため、P_{bu} を 2 で割っている。

4 階層的 ITG モデル

FLAT では、最小フレーズのみが記憶されるが、複数の粒度のフレーズを利用した機械翻訳に比べて最小フレーズのみを利用した機械翻訳の精度が低いと知られている [5]。このため、先行研究は FLAT で最小フレーズをアライメントしてから、ヒューリスティックスに基づいて網羅的に長いフレーズを抽出する。提案手法では、階層的なモデルを利用することで、複数の粒度のフレーズを直接確率モデルで表現する。このため、ヒューリスティックスに基づくフレーズ抽出を行わずに高い翻訳精度を実現する。この階層的モデルを HIER と呼ぶ。

FLAT と同様、HIER のフレーズペア確率 $P_{hier}(\langle e, f\rangle; \theta_x, \theta_t)$ を定義する。モデルの違いは生成過程の順番にある。FLAT はまず導出木の分岐点を P_x から生成してからフレーズペアを P_t から生成するのに対して、HIER はまず P_t からフレーズペア $\langle e, f\rangle$ を生成しようとする。P_t からフレーズペアが生成できなかった場合、基底測度で再帰的により小さいフレーズペアを 2 つ生成し、組み合わせることで新たなフレーズペアを生成する。正式には、式 (2) で利用した基底測度 P_{base} の代わりに、新しい基底測度 P_{dac}（“divide-and-conquer”）を定義し、θ_t の式は以下のようになる。

$$\theta_t \sim PY(d, s, P_{dac}) \tag{3}$$

P_{dac} の生成過程は P_{flat} と類似しており、以下のような ITG に基づく生成過程となる。

1. 記号 x を $P_x(x; \theta_x)$ に従って生成する。x は BASE、REG、INV の値を取り得る。
2. x の値に従って：
   (a) x = BASE の場合、新しいフレーズペアを第 3.2 節の P_{base} から直接生成する。
   (b) x = REG の場合、$\langle e_1, f_1\rangle$ と $\langle e_2, f_2\rangle$ を P_{hier} から生成し 1 つのフレーズペア $\langle e_1 e_2, f_1 f_2\rangle$ を作成する。
   (c) x = INV の場合、(b) と同じように 2 つのフレーズペアを生成するが f_1 と f_2 を逆順に並べる $\langle e_1 e_2, f_2 f_1\rangle$。

FLAT と HIER の導出木の比較を図 1 に示す。図 1 の通り、FLAT の P_t は最小フレーズのみを生成するが、HIER では複数の粒度のフレーズを P_t から生成し、記憶する。

4.1 実装

フレーズアライメントの先行研究の多くはサンプリングを用いてモデルを学習する [5, 1]。本研究では Blunsom ら [1] に従い、文ごとのブロックサンプリングを利用す

[1] P_u は 10^{-2}、10^{-3}、10^{-10} の中からヘルドアウトデータで精度が最もよくなるように選択する。

[2] λ を 1、0.1、0.01 の中からヘルドアウトデータで精度が最も高くなるように選択する。

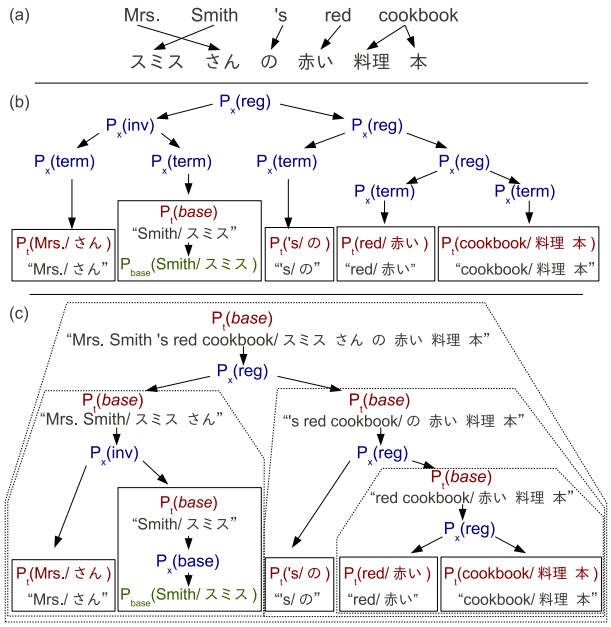

図 1: 単語アライメント (a)、FLAT の導出木 (b)、HIER の導出木 (c)。実線と点線はそれぞれ最小フレーズペアとその多のフレーズペアであり、モデルに記憶されるフレーズはそのフレーズを生成した P_t の下に書いてある。「Smith/スミス」は P_{base} によって生成された

る。ITG の導出木候補を現実的な時間で探索するために Saers ら [13] のビームサーチに基づくチャート法を採用し、確率ビームを $P > 10^{-10}$ とする。

従来のモデルに比べて、フレーズペアの頻度管理に注意する必要がある。あるフレーズペア t_a が t_b と t_c から構成される場合、t_a を含むサンプルが削除される時に t_b と t_c の頻度を減らさなければならない場合がある。P_t を中華料理店過程 (CRP,[14]) で表現する場合、フレーズの管理は容易となる。t_a を生成するテーブルに対して、客の数だけではなく、テーブルを生成した時に利用したフレーズペア t_b と t_c も記録しておく。t_a の客数が 0 になった場合、t_b と t_c の客数も 1 人減らす。

5 フレーズ抽出

本節では、従来のフレーズ抽出と提案手法のフレーズ抽出について述べる。

5.1 ヒューリスティクスに基づくフレーズ抽出

従来のフレーズ抽出は単語アライメントに従ってフレーズを網羅的に抽出する [11]。フレーズペアに対して、最尤推定による両方向の条件付き確率 $P_{ml}(f|e)$ と $P_{ml}(f|e)$、単語の翻訳確率を用いる両方向の lexical weighting 確率 [11]、各フレーズに対する定数のペナルティという 5 つの素性を計算する。このフレーズ抽出法を HEUR-W と呼ぶ。HEUR-W に必要な単語アライメントは IBM モデル [2] で得ることができ、これを 1 つ目のベースラインとして利用する。

提案手法で得られるアライメントもヒューリスティクスに基づくフレーズ抽出と組み合わせることができるため、これを 2 つ目のベースラインとして利用する。しかし、提案手法は長いフレーズを取ることもあり、最小フレーズでもデータがスパースになることもある。このため、さらに細かいアライメントを得るために、モデルがあるフレーズを生成した場合、そのまま使う（HEUR-P）だけではなく、1 対多アライメントになるまでを分解する（HEUR-B）手法や、1 対 1（または 0）のアライメントまで分解する（HEUR-W）手法も試みる（図 2）。

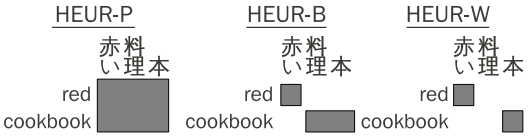

図 2: フレーズ・ブロック・単語のアライメント

5.2 モデル確率に基づくフレーズ抽出

ITG モデルの生成確率 $P_t(\langle e, f\rangle)$ に基づくフレーズテーブル構築法も提案する。フレーズテーブルの素性として、条件付き確率 $P_t(f|e)$ と $P_t(e|f)$ や、lexical weighting 確率、フレーズペナルティなどを利用する。前節の条件付き確率は最尤推定によるものであったが、ここではモデル確率 P_t を使って条件付き確率を計算する：

$$P_t(f|e) = P_t(\langle e, f\rangle) / \sum_{\{\tilde{f}: c(\langle e, \tilde{f}\rangle) \geq 1\}} P_t(\langle e, \tilde{f}\rangle)$$

$$P_t(e|f) = P_t(\langle e, f\rangle) / \sum_{\{\tilde{e}: c(\langle \tilde{e}, f\rangle) \geq 1\}} P_t(\langle \tilde{e}, f\rangle).$$

なお、サンプルに 1 回以上現れるフレーズペアのみをフレーズテーブルに入れる。

さらに、2 つの素性を加える。1 つ目はモデルによるフレーズペアの同時確率 $P_t(\langle e, f\rangle)$ である。2 つ目は inside-outside アルゴリズムで計算されたスパンの事後確率に基づいて、あるフレーズペア $\langle e, f\rangle$ が入っているスパンの平均事後確率を素性とする。スパン確率は頻繁に起こるフレーズペア、または頻繁に起こるフレーズペアを元に構成されるフレーズペアで高くなるため、フレーズペアがどの程度信頼できるかを判定するのに有用である。このモデル確率に基づくフレーズ抽出を MOD と呼ぶ。

6 実験評価

提案手法を仏英翻訳と日英翻訳のタスクで評価した。仏英翻訳において Workshop on Statistical Machine Translation (WMT)[3] のデータを用い、翻訳モデル学習に news commentary のコーパス、言語モデル学習に news commentary と Europarl のコーパスを利用した。日英翻訳は NTCIR の特許翻訳タスク [9] のデータを用い、翻訳モデルにパラレルコーパスの最初の 10 万文、言語モデルにパラレルコーパス全体を利用した。コーパスの諸元を表 1 に示す。データの前処理として単語分割（トークン化）と小文字化を行い、翻訳モデルの学習に 40 単語以下の文のみを利用する。デコーダとして

表 1: 各コーパスの単語数

	WMT		特許	
	fr	en	ja	en
翻訳モデル	1.56M	1.35M	2.78M	2.38M
言語モデル	-	52.7M	-	44.7M
重み学習	55.4k	49.8k	80.4k	68.9k
テスト	72.6k	65.6k	48.7k	40.4k

表 2: BLEU スコアとフレーズテーブルのサイズ。太字は最も高い精度のシステムに比べて統計的に有意な差ではない（$p < 0.05$ のサインテストにより [4]）

抽出法		fr-en BLEU	fr-en サイズ	ja-en BLEU	ja-en サイズ
GIZA	HEUR-W	21.35	4.01M	**23.20**	4.22M
FLAT	MOD	19.09	271k	21.07	263k
HIER	MOD	**21.50**	751k	**23.23**	723k

Moses[10] を利用する。フレーズの最大長を 7 とし、言語モデルは Kneser-Ney 平滑化を用いた 5-gram モデルである。評価基準は 4-gram までの BLEU スコアとする。

最初の実験では、FLAT と HIER のモデル確率を利用したフレーズ抽出（MOD）と、GIZA++から得られたアライメント（GIZA）とヒューリスティックスに基づくフレーズ抽出の精度を比べる。

GIZA の場合は Model 4 までの標準的な学習設定を用いて、`grow-diag-final-and` で両方向のアライメント結果で組み合わせる。提案手法では 100 イタレーションの学習を行い、最後のサンプルを利用する。実際には 100 イタレーション目まで尤度が単調増加したが、翻訳精度は 5〜10 イタレーション目以降ほぼ同等となった。1 イタレーションは 1 コアで約 1.3 時間かかったため、良い翻訳精度は 6.5〜13 時間で実現することができた。

実験結果を表 2 に示す。仏英・日英ともに、階層的モデルの確率を利用したフレーズテーブルは GIZA++とヒューリスティックスに基づくフレーズ抽出の精度をわずかに上回った。完全な確率モデルがヒューリスティックスに基づくフレーズ抽出を上回ったのは本稿で初めてである。さらに、提案手法で得られたフレーズテーブルのサイズも従来法の 20%弱に収まった。また、モデル確率を用いた場合、HIER は FLAT を大きく上回った。これは、先行研究が報告する通り [6]、最小フレーズのみを利用すると高い精度が得られないからである。

最後に、モデル確率に基づくフレーズ抽出と従来法の比較を表 3 に示す。HIER や FLAT のアライメントを利用し、モデル確率を用いる提案手法 MOD に加えて、第 5 節で説明したフレーズ HEUR-P、ブロック HEUR-B、単語 HEUR-W を最小単位とするヒューリスティック抽出を比較した。HIER と MOD の組み合わせはヒューリスティック抽出とほぼ同等、またはより高い精度を示しながら、フレーズテーブルのサイズを大幅に削減した。

表 3: 様々なフレーズ抽出法による翻訳精度とフレーズテーブルサイズ（仏英）

	FLAT		HIER	
MOD	19.09	271k	21.50	751k
HEUR-W	21.16	6.05M	21.68	5.39M
HEUR-B	21.16	3.39M	21.41	2.61M
HEUR-P	19.14	1.12M	21.47	1.62M

7 おわりに

本稿はベイズ学習と ITG に基づく階層的モデルを用いて、機械翻訳のためのフレーズアライメントと抽出を同時に行う手法を提案する。提案手法を使った評価実験では、従来の 2 段階法とほぼ同等の精度を保ちながらフレーズテーブルのサイズを大幅に削減できた。

参考文献

[1] P. Blunsom and T. Cohn. Inducing synchronous grammars with slice sampling. In *Proc. NAACL*, 2010.

[2] P. F. Brown, V. J. Pietra, S. A. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–311, 1993.

[3] C. Callison-Burch, et al. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proc. WMT/MetricsMATR*, pp. 17–53, 2010.

[4] M. Collins, P. Koehn, and I. Kučerová. Clause restructuring for statistical machine translation. In *Proc. ACL*, pp. 531–540, 2005.

[5] J. DeNero, A. Bouchard-Côté, and D. Klein. Sampling alignment structure under a Bayesian translation model. In *Proc. EMNLP*, pp. 314–323, 2008.

[6] J. DeNero, D. Gillick, J. Zhang, and D. Klein. Why generative phrase models underperform surface heuristics. In *Proc. WMT*, pp. 31–38, 2006.

[7] J. DeNero and D. Klein. The complexity of phrase alignment problems. In *Proc. ACL*, pp. 25–28, 2008.

[8] J. DeNero and D. Klein. Discriminative modeling of extraction sets for machine translation. In *Proc. ACL*, pp. 1453–1463, 2010.

[9] A. Fujii, M. Utiyama, M. Yamamoto, and T. Utsuro. Overview of the patent translation task at the NTCIR-7 workshop. In *Proc. NTCIR-7*, pp. 389–400, 2008.

[10] P. Koehn, et al. Moses: Open source toolkit for statistical machine translation. In *Proc. ACL*, 2007.

[11] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proc. NAACL*, pp. 48–54, 2003.

[12] P. Liang, B. Taskar, and D. Klein. Alignment by agreement. In *Proc. NAACL*, pp. 104–111, 2006.

[13] M. Saers, J. Nivre, and D. Wu. Learning stochastic bracketing inversion transduction grammars with a cubic time biparsing algorithm. In *Proc. IWPT*, 2009.

[14] Y. W. Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proc. ACL*, 2006.

[15] D. Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403, 1997.

[16] H. Zhang, C. Quirk, R. C. Moore, and D. Gildea. Bayesian learning of non-compositional phrases with synchronous parsing. *Proc. ACL*, pp. 97–105, 2008.