階層的バックオフモデルに基づく同期文脈自由文法の学習

上垣外 英剛 †　渡辺 太郎 ‡　高村大也 ††　奥村学 ††
† 東京工業大学 大学院総合理工学研究科　‡ 情報通信研究機構
†† 東京工業大学 精密工学研究所
†kamigaito@lr.pi.titech.ac.jp ††{takamura,oku}@pi.titech.ac.jp

1 はじめに

階層型統計的機械翻訳では, フレーズベースの統計的機械翻訳と同様, 対訳コーパスに対し, 単語アライメントを学習し, ヒューリスティックな手法により同期文脈自由文法 (SCFG) に基づく同期ルールの抽出を行う. このヒューリスティックな抽出手法によって生成されたルールテーブルには, 翻訳時にほとんど用いられない, もしくは誤った対応となっているルールを多く含んでいる. ルールテーブルに含まれるルール数は膨大なものであり, 速度の低下や翻訳誤りをもたらす.

フィッシャーの検定値 [4] やエントロピー [6] などの指標を用いて, ルールの数を削減する手法が提案されているが, これらの手法は, 閾値を適切に設定する必要がある. ヒューリスティックな手法に対し, ディリクレ過程や Pitman-Yor 過程にもとづいて, 過学習を防ぎつつ, コンパクトな同期文脈自由文法を学習し, その結果に基づいて網羅的にフレーズ対応を決定する手法が提案されている [2, 5]. ルールテーブルを直接学習するためには, サンプリングの際に同期文脈自由文法に従った構文解析を行うのが望ましいが, $O(n^6)$ の計算量が必要なため, 繰り返しサンプリングを行うベイズ的な学習モデルにおいては, 膨大な時間を要するため, 相性が良くない. Blunsom, Levenberg らは共に, 初期化以外では構文解析を必要としない手法を提案する事でこの問題に対処しているが, この場合は, 初期化の際に用いた導出木への依存が強くなるという問題が生じる.

本研究では, ノンパラメトリックベイズ法に基づき, 複数の粒度のフレーズ対を網羅的に学習しつつ, 同期文脈自由文法の一つである, 自由度の高い Hiero 文法を同時に学習する手法を提案する. 具体的には, Neubig らの階層的なバックオフモデル [7] を, Hiero 文法へと拡張し, バックオフ時に任意のルールを用いる. また, 高速な学習のため, Xiao らが提唱した $O(n^3)$ の計算量で構文解析が可能な, 二段階法 [10] を用いる. ビーム幅に基づく枝刈りでは, 枝刈りされたモデルに, スパンが採用される保証がないため, 詳細釣り合い条件を満たさない. そこで, Blunsom ら [1] と同様, スライスサンプリング [9] を導入し, 詳細釣り合い条件を満たしながら, 実用的な時間でのサンプリングを可能とした.

提案手法は, WMT2010 の news-commentary コーパスのドイツ語-英語対において, BLEU を大きく減少させる事無く, ルールテーブルを大幅に削減する事に成功した.

2 従来モデル

従来モデルでは, Pitman-Yor 過程に基づく Levenberg らの様に, コーパス中の各文に対応する同期文脈自由文法の導出木から, ノンパラメトリックベイズ法に基づいたサンプリングを行う事で, モデルにおける, 同期文脈自由文法の各同期ルールの確率を学習している. 本稿では, その中でも階層型 Pitman-Yor 過程に基づく Levenberg らのモデルを説明する. 例文 “日本語を英語に翻訳する事は難しい。/Japanese is difficult to translate into English.” の同期文脈自由文法における導出の例を次に示す.

S → X_1 英語 X_2 難しい。 / X_1 difficult X_2 English.
X_1 → X_3 を / X_3 is
X_2 → X_4 翻訳 する X_5 は / X_4 translate X_5
X_3 → 日本語 / Japanese
X_4 → に / into
X_5 → 事 / to

また, 対応する導出木を図 1 に示す. Levenberg らは各文における導出木 G_X, 割引値 d, スムージング値 θ,

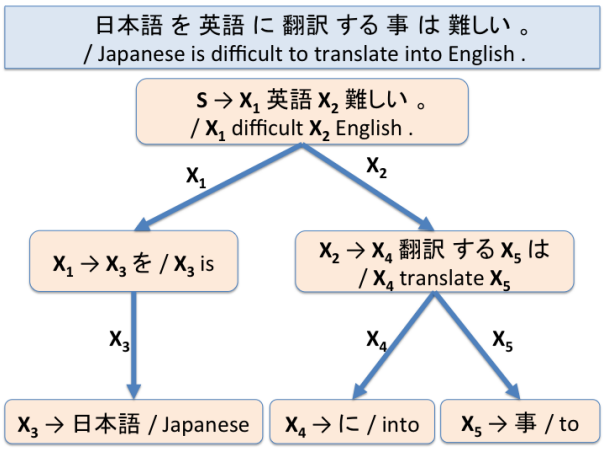

図 1: 従来モデルの導出木

基底測度 G_0 に対して次のような生成過程

$$G_X \sim PY(d, \theta, G_0), \quad (1)$$

$$X \to \langle \mathbf{s}, \mathbf{t}, \mathbf{a} \rangle \sim G_X, \quad (2)$$

を提案している．$PY(d, \theta, G_0)$ は Dirichlet 過程をより一般化し，割引値 d を明示的にモデル化した Pitman-Yor 過程 [8] に従い，同期ルール k の客の数 c_k，テーブルの数 $|\varphi_k|$，客の総数 n，テーブルの総数 $|\varphi|$ を用いて

$$PY(d, \theta, G_0) = \frac{c_k - d \cdot |\varphi_k|}{\theta + n} + \frac{\theta + d \cdot |\varphi|}{\theta + n} \cdot G_0, \quad (3)$$

のような式で表される．式 3 の右辺の第一項は，同期ルール k を表すテーブルが存在する場合の確率を，第二項はテーブルが新しく生成される場合の確率をそれぞれ表している．基底測度 G_0 は同期文脈自由文法の各同期ルールに対して，次のような生成過程によって構成される．

- 原言語側の記号数
  $|s| \sim Poisson(1)$
- 原言語側に対する目的言語側の終端記号の数
  $|T_t| \sim Poisson(|s| - NT_s + \lambda_0)$
  λ_0 は，ポアソン分布のパラメータが 0 になる事を避けるための，微小な値である．
- 原言語の記号 s_i の種類
  $type_i \sim Bernoulli(\phi^{|s|})$
  ϕ は $0 < \phi < 1$ を満たすハイパーパラメータである．
- 原言語側と目的言語側の終端記号のアライメント
  IBM Model1 の双方向の確率の算術平均に従う．

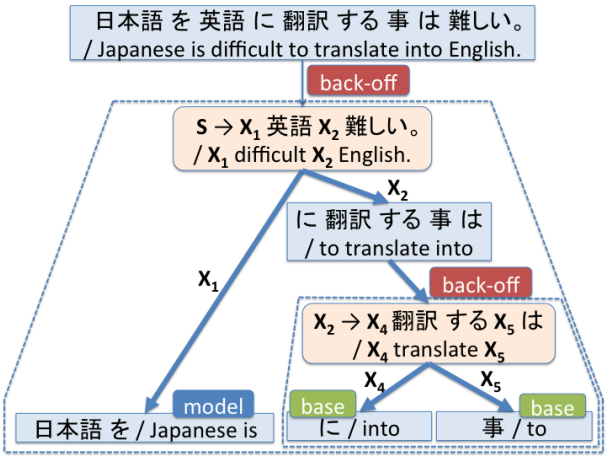

図 2: 提案モデルの導出木

モデルを学習する際に，Levenberg らは計算量の問題から，各文の構文木の一部の同期ルールに対してギブスサンプリングを行っている．しかし，この方法では，初期化木を作る際以外には構文解析を行わないため，結果が初期化木に依存してしまう．

3 階層的バックオフモデル

従来法では，どのようなフレーズもルールによって必ず分解されるため，フレーズ対により明示的な二言語間の対応付けが不可能である事から，学習された結果を，単語アライメントへと変換してから，網羅的に Hiero 文法の学習を行っていた．そこで，Neubig らの階層的なバックオフモデルを，任意の同期文脈自由文法へと拡張し，網羅的にフレーズ対を学習すると同時に，Hiero 文法を学習するモデルを提案する．提案モデルでは，図 2 のように，まずフレーズ対の生成を試み，もし生成できない場合，同期ルールを生成する．さらに，生成された同期ルールに含まれる非終端記号から，再帰的にフレーズ対の生成をおこなう．より具体的に，生成過程は，

$$G_r \sim PY(d_p, \theta_p, G_p), \quad (4)$$

$$G_p \sim PY(d_r, \theta_r, G_r), \quad (5)$$

のように表される．G_p は model, back-off, base の三つの状態を取る．それぞれの状態における G_p の出力は次のようになっている．

1. model
   現在モデルに存在するフレーズ対について，式 3 の右辺の第一項と同じくモデル確率を出力する．

2. back-off
   現在のフレーズから生成された同期文脈自由文法の同期ルールの確率と, その非終端記号に含まれる全てのフレーズ確率の積を, フレーズ確率として出力する.

3. base
   フレーズ対について式 3 の右辺の第二項と同じく基底測度から確率を出力する.

フレーズ対の基底測度は次のような生成過程によって構成される.

- 原言語側の単語数 w_s
  $|w_s| \sim Poisson(1)$
- 原言語側に対する目的言語側の単語数
  $|w_t| \sim Poisson(|w_s| + \lambda_0)$
- 原言語と目的言語の単語アライメント
  IBM Model1 の双方向の確率の相乗平均に従う.

G_r については従来モデルと同じである. 提案モデルは同期文脈自由文法の同期ルールに対して model, base を, フレーズ対に対して model, back-off, base を扱う. このため, 従来モデルと比較して, 各スパンに存在する導出候補の数が多くなり, 計算時間が増大する. 我々は, 二段階法とスライスサンプリングを併用する事で, 構文解析の高速化を行った. 両言語を扱う, 通常の構文解析手法の計算量 $O(n^6)$ とは異なり, 二段階法では, 原言語側のみでスパンを扱うため, 計算量については, 単言語に対する構文解析と同じく $O(n^3)$ で解析を行う事が出来る. ただし, 二段階法を用いる場合は, 目的言語側のスパンの組み合わせを全て考慮する事が出来ないため, 構文解析が失敗する可能性が存在する. 今回対象としたコーパスでは, 各文に対する解析の成功率は 99%を超えているため, 構文解析の失敗については問題にはならない水準であると判断した. スライスサンプリングは, ビームによる枝刈りとは異なり, 詳細釣り合い条件を満たしながら, 枝刈りを行う事が出来る. スライスサンプリングは動的計画法との相性がよく, 特に iHMM における状態遷移の枝刈り [9] において使用されている. 我々は, Blunsom ら [1] と同じく, 前回のイテレーションで採用された導出木に含まれる同期ルールに対しては, そのスパンに対する基底測度を基準とした一様分布からのサンプルを用いて枝刈りを行い, 前回のイテレーションで採用されなかった同期ルールに関しては, ベータ分布からのサンプルを基準にした枝刈りを行っている. またこの手法を用いる際には, 初回のイテレーションでは, 前回のイテレーションで採用された導出木が存在しないことから, 初期化のための導出木が必要である. この問題については, 二段階法でビーム幅を 30 に設定して導出された導出木を, 初期化木として用いることで解決した.

4 実験

WMT2010 のドイツ語-英語対を対象に, 原言語はドイツ語, 目的言語は英語として比較実験を行った. 比較対象は, ヒューリステックな同期ルールの抽出手法と, 従来モデルである. 翻訳モデルの訓練データとしては, news-commentary コーパスを用い, 言語モデルの訓練データとしては, 全ての news-commentary コーパスと europarl-v7 を合わせて用いた. 翻訳モデルを訓練する際には, news-commentary コーパスの先頭 10 万文を使用し, それらに対して, 小文字へと変換した後に, 原言語および目的言語の双方で, 単語数が 40 以内の対訳文を用いた. 使用する言語モデルは 5-gram とし, IRSTLM で学習を行い, スムージングには Kneser-Ney を用いた. デコーダには cdec[3] を使用した. デコーダのパラメータには cdec の初期値を, ペナルティには WordPenalty と ArityPenalty を用いた. 翻訳モデル, 言語モデルのパラメータチューニングには mira を使用した. mira によるパラメータチューニングには結果にゆらぎがあるため, 今回の実験では, 三回の結果の平均を比較に使用している. 各回のイテレーション数は 25 とした. 従来モデルと提案モデルにおける同期ルールの学習には, 20 イテレーションのサンプリングを行い, 最後のイテレーションにおいてサンプルされた, 同期ルールとフレーズを翻訳モデルに用いた. ルールテーブルでは, 従来モデル, 提案モデル共に, モデルから計算される条件付き確率を翻訳モデルの素性として用いた. 従来モデル, 提案モデル共に, ハイパーパラメータ d, θ は各イテレーションの最後に, スライスサンプリングを行うことで推定した. また基底測度に含まれるハイパーパラメータ ϕ についても, d, θ を推定した後で, メトロポリスヘイスティング法を用いて推定した. 提案モデルにおいて, ルールテーブルにフレーズを入れる際は, 単語長が 7 以下のものに限定した. ヒューリステックな抽出手法に関しては, GIZA++によって生成された単語アライメントから, grow-diag-final によってフレーズ対応を抽出し, それらに対し, mosesdecoder を用いて Hiero 文法に拡張している. 結果を表 1 に示す. 表中のサイズはルールテーブルの行数を示している.

表 1: 提案モデルとの比較

手法	BLEU	テーブルサイズ
ヒューリステック	**16.83**	7.07M
従来モデル	15.20	**516k**
提案モデル	15.50	904k

BLEU スコアはヒューリスティックな手法が最も高い. しかし, ヒューリスティックな手法は, ルールテーブルのサイズも大きい事が分かる. 一方, 従来モデルは, BLEU スコアが最も低いが, ルールテーブルのサイズは最も小さい. 提案モデルは, BLEU スコア, ルールテーブルのサイズ共に, ヒューリスティックな手法と従来モデルとの間に位置している. この実験結果からは, 提案モデルと比較手法との間に, 明確な優位性はなく, BLEU スコアと, ルールテーブルのサイズとの間に, トレードオフの関係が存在している, ということが導ける.

5 結論

提案モデルの BLEU スコアは, 従来モデルを上回り, またルールテーブルのサイズは, ヒューリステックな抽出手法よりも小さいものとなった. これらのことから, 提案モデルは比較手法に比して, いくつかの点で優れていると言える. しかし, BLEU スコアに関しては, ヒューリスティックな手法よりも低く, Neubig らのフレーズベースのモデルでは, Pitman-Yor 過程を用いて学習を行った結果が, ヒューリステックな手法よりも高い BLEU スコアが出ている事を考えると, 改善の余地がある. 今後は, 使用する素性を増やす事等で, BLEU スコアの向上を目指したい. また, 対象とする言語対を増やして, 各言語に対する特徴を, より詳細に分析する必要もあると考えている. さらに, スライスサンプリングに更なる改良を加える事や, 構文解析により高速なアルゴリズムを用いる事で, 大規模なコーパスにおいても, 提案手法を用いて, 実験を行う事を検討したい.

参考文献

[1] Phil Blunsom and Trevor Cohn. Inducing synchronous grammars with slice sampling. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 238–241. Association for Computational Linguistics, 2010.

[2] Phil Blunsom, Trevor Cohn, Chris Dyer, and Miles Osborne. A gibbs sampler for phrasal synchronous grammar induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pp. 782–790. Association for Computational Linguistics, 2009.

[3] Chris Dyer, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of ACL*, 2010.

[4] John Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. Improving translation quality by discarding most of the phrasetable. 2007.

[5] Abby Levenberg, Chris Dyer, and Phil Blunsom. A bayesian model for learning scfgs with discontiguous rules. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pp. 223–232. Association for Computational Linguistics, 2012.

[6] Wang Ling, Joao Graça, Isabel Trancoso, and Alan Black. Entropy-based pruning for phrase-based machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 962–971. Association for Computational Linguistics, 2012.

[7] Graham Neubig, Taro Watanabe, Eiichiro Sumita, Shinsuke Mori, and Tatsuya Kawahara. An unsupervised model for joint phrase alignment and extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 632–641. Association for Computational Linguistics, 2011.

[8] Yee Whye Teh. A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 985–992. Association for Computational Linguistics, 2006.

[9] Jurgen Van Gael, Yunus Saatci, Yee Whye Teh, and Zoubin Ghahramani. Beam sampling for the infinite hidden markov model. In *Proceedings of the 25th international conference on Machine learning*, pp. 1088–1095. ACM, 2008.

[10] Xinyan Xiao, Deyi Xiong, Yang Liu, Qun Liu, and Shouxun Lin. Unsupervised discriminative induction of synchronous grammar for machine translation. In *COLING*, pp. 2883–2898, 2012.