4-3　ベイジアンアライメントに基づく翻字システムと機械翻訳への応用

4-3 A Transliteration System Based on Bayesian Alignment and its Human Evaluation within a Machine Translation System

フィンチ アンドリュー　安田圭志
Andrew Finch and YASUDA Keiji

要旨

本稿では、翻字コーパスから、文字列アライメントに用いることができるベイジアンモデルについて説明する。本手法では Dirichlet process モデルを用い、ブロックギブスサンプリングをもとにベイジアン推定を行う。提案手法では、最尤推定における過学習の問題を解消することができる。翻字システムのモデルを構築するためにベイジアンアライメントを使用し、その有効性を実証する。提案手法を、従来法である GIZA＋＋、m2m バイリンガルアライナーと比較する。両方の場合で、ベイジアンアライメントのモデルは従来法よりもモデルサイズが小さく、翻字性能も高いことが分かった。次に、自動翻字システムと機械翻訳システムを統合し、翻訳システムの評価を実施した。評価においては、日本各地で実施された実証実験において収集されたデータを用いた。その結果、未知語対応のために翻字システムを使用すると機械翻訳の訳質が改善されることが確認された。

This paper reports on contributions in two areas. Firstly, we present a novel Bayesian model for unsupervised bilingual character sequence alignment of corpora for transliteration. The system is based on a Dirichlet process model trained using Bayesian inference through blocked Gibbs sampling implemented using an efficient forward filtering/backward sampling dynamic programming algorithm. The Bayesian approach is able to overcome the overfitting problem inherent in maximum likelihood training. We demonstrate the effectiveness of our Bayesian alignment by using it to build models for phrase-based statistical machine transliteration (SMT) systems. We compare our alignment technique to the commonly used GIZA++ word alignment process, and also to the state-of-the-art m2m bilingual aligner by using their alignments to train transliteration generation systems. In both cases the model resulting from our Bayesian alignment was considerably smaller than competitive technique, and in addition gave an increase in transliteration generation quality. Our second contribution is to conduct a large-scale real-world evaluation of the effectiveness of integrating an automatic transliteration system with a machine translation system. A human evaluation is usually preferable to an automatic evaluation, and in the case of this evaluation especially so, since the common machine translation evaluation methods are often being biassed towards translations in terms of their length rather than the information they convey. We evaluate our transliteration system on data collected in field experiments conducted all over Japan. Our results conclusively show that using a transliteration system can improve machine translation quality when translating unknown words.

[キーワード]
翻字，主観評価，機械翻訳，Dirichlet process モデル，ベイジアンアライメント
Transliteration, Human evaluation, Machine translation, Dirichlet process model, Bayesian alignment

1 はじめに

機械翻訳では、高い翻訳性能を実現するため、広範な固有名詞を取り扱う必要がある。しかしながら、実世界においては常に新しい固有名詞が作られており、固有名詞の数は日々増加している。限られた言語資源から、全ての固有名詞を網羅する対訳辞書を自動構築することは不可能であり、また、多言語辞書を網羅的に人手で構築する方法も現実的ではない。このようなことから、機械翻訳の分野においては、翻字技術の応用による固有名詞対応が有望である。本稿では、翻字モデルを構築するベイジアン手法を提案し、従来法で問題となっている過学習の問題を解決する。また、従来の翻字研究で用いられた評価だけではなく、実際の機械翻訳に、翻字システムを組み込んだ場合の評価についても述べる。統計的機械翻訳の研究では、BLEU スコア[1]や NIST スコア[2]などの翻訳自動評価が用いられることが多い。しかしながら、これらの評価手法においては、未知語が含まれる場合について、誤りが含まれうる方法で未知語の処理をするより、単に未知語を削除する方が、評価スコアが高くなることがある。従って、これらの評価手法では、未知語処理に関して適切な評価ができない。そのため、本稿では、主観評価により翻訳システムの評価を行っている。

本稿を通じ、各言語は、Unicode 文字の単位を用いて表現する。たとえば、英語では「a」、日本語では「ア」である。文字列は、これらの任意のシーケンスで、文字列のペアは2つの文字列で1つのペアであり、そのペアの各要素は各言語の文字列となる。たとえば、「hello」、「ハロー」などがある。

ここで、翻字モデルのトレーニングでベイジアンアライメント機構を使用した動機について述べる。

1.1 動機

翻字の問題は、順序の入れ替えが起こらない、文字列–文字列間の変換のタスクとして表現することができる。統計的機械翻訳技術[4]-[6]に基づくシステムおよびフレーズベースの joint source channel モデル[7]は近年、活発に研究されており、このタスクに対して高い性能を実現している。翻字タスクにおいては、変換が文字列–文字列間で直接的であるため、中間的音声表現を必要とせず、必要とされる学習データが対訳辞書のペアのみであるなどの利点もある。このため、本手法は、多くの言語ペアに適用可能である。本稿では、フレーズベース統計機械翻訳（PBSMT）手法を使用した翻字および joint source channel の基礎として文字列ペアを使用する翻訳の方法論について取り上げる。

PBSMT の中心的な構成要素はフレーズテーブルである。翻字タスクにおいては、これが、バイリンガル文字列ペアのセットとなる。PBSMT システムの典型的な学習手順において、フレーズセットテーブルの生成手順は次のとおりである。

1．GIZA＋＋[8]を使用した単語アライメント
2．ヒューリスティックを使用したフレーズペア抽出（たとえば、MOSES[9]ツールキットの *grow-diag-final-and*）

上記2ステップによるアプローチは実際に適切に機能するが、1ステップでバイリンガル文字列ペアのセット（文字列レベルのフレーズペアの類似物について説明するには、この用語を使用）を生成する方法の方がよりシンプルである。従来法である、尤度最大化のために EM アルゴリズムを使用すると、過学習が生じる可能性がある。極端な例として、シーケンスペアの数について一切制限のない場合を仮定すると、シーケンスペアに対して最も可能性のあるコーパスのアライメントは、1つのバイリンガルシーケンスペアとしてコーパス全体となる。

GIZA＋＋では、1対多で単語をアライメントすることにより、この問題が緩和される。アライメントの片側にある1つの単語は、バイリンガルペアのサイズに関して1つの制約として機能する。類似のアプローチを翻字においても適用できる。この場合、1つの言語の1つの文字をもう一方の言語の複数の単語にアライメントできるが、その逆は許可されない。英語–中国語の翻字[7][10]の場合では、1つの中国語の文字が複数の英語の文字に対応するため、このようなアプローチが有効である。

GIZA＋＋では、この1対多のアライメントが2回、ソースからターゲットとターゲットからソースの両方で実行される。単語間のアライメン

トに関するテーブルは、これらの両方のアライメント（通常は共通部分）から構築される。共通部分にない追加の単語アライメントはヒューリスティックに基づいて追加され、最後には、すべての可能なフェーズペアが抽出される。

文献[11][12]の手法では、多対多のアライメントは最尤トレーニングを用いて直接実行される。この方法では、モデルの過学習を回避するため、使用可能な学習データの一部を用いたヒューリスティックの適用が必要である。文献[13]では、類似のベイジアン手法を文法導入に適用することに成功しており、文献[14][15]では、SMTのバイリンガルフレーズペアの抽出における複雑なタスク向けのベイジアン法を開発した。文献[16]では、異なる思想であるにもかかわらず、我々のアプローチと多数の特性を共有する手法を用いて、leave-one-outによりフレーズアライメントの過学習の問題に取り組んでいる。また、文献[17]では、翻字のアライメントにおける、Adaptor Grammarを開発した。

本稿では、既存のモノリンガル単語分割モデル[18][19]をバイリンガルアライメントに拡張し、[20]の方法でベイジアンモデルを使用して、過学習することなく多対多でアライメントするため、シンプルかつ高性能な方法を提案する。

本稿は主に2つの要素から構成される。まず最初に、ベイジアンアライメント手法と、従来法の性能を評価するための実験について説明する。次に、提案手法を機械翻訳システムの未知語処理として利用する場合の有用性について検証する評価実験について述べる。具体的には、**2**で、ベイジアンモデルについて説明する。ここで、Dirichlet processモデル、Chinese Restaurant process（CRP）の概要を示し、我々のモデルとこれらの2つの表現との関連について説明する。**3**では、モデルの学習に使用したブロックギブスサンプリング手法について説明する。**4**では、機械翻訳のフレーズ抽出で使用したアライメント手法に関して我々のモデル構築のために実施した実験について説明する。**5**では、最新の多対多のシーケンスアライメントツールであるm2mアライナーと提案手法を比較するための実験および分析について示す。**6**では、主観評価により、機械翻訳の出力において未知語処理のために、翻字を使用した場合の評価実験について述べる。最後に**7**では、今後の研究の検討課題と結論について述べる。

2　方法

自然言語処理の分野では近年、ベイジアンモデルがよく利用されており、特に単語分割のタスクにおいて有効であることが示されている[18][19]。本稿で使用するモデルは、1-gram Dirichlet processモデルである。このアプローチを使用して機械翻訳の一般的な場合に単語アライメントを実行することは困難であるが、翻字のように、シーケンス（文字列）の長さが短く、並べ替えが起きない場合、特殊な最適化またはアニーリングを要することなく、アライメント問題に直接利用することができる。

2.1では、joint source channelモデルについて、**2.2**では、このモデルを使用した、Dirichlet processモデルの概要について説明する。

2.1　joint source channelモデル

原言語文字列 $\mathbf{s}_1^M = <s_1, s_2, \ldots, s_M>$ と目的言語文字列 $\mathbf{t}_1^N = <t_1, t_2, \ldots, t_N>$ で構成される対訳コーパスが与えられたとする。上線付きのボールド体フォントを使用して、単一文字の文字列を区別する。

基本となる生成モデルとして[7]のjoint source channelモデルを採用し、セグメントが互いに独立であることを前提とする（我々のアプローチは、これらの依存性をモデル化するために容易に拡張できる[19]）。このモデルでは、バイリンガルシーケンスペア（目的言語文字列と原言語文字列のペア）の連結によりコーパスが生成される。

バイリンガルシーケンスペアは、目的言語の文字列と原言語の文字列とで構成される一組の（s, t）である $(\mathbf{s}, \mathbf{t}) = (<s_1, s_2, \ldots s_i>, <t_1, t_2, \ldots, t_j>)$。

バイリンガルシーケンスペアの確率は、対訳コーパスを用いて、次式により計算される。

$$p(\mathbf{s}_1^M, \mathbf{t}_1^N) = p(s_1, s_2, \ldots, s_M, t_1, t_2, \ldots, t_N) = \sum_{\gamma \in \Gamma} p(\gamma) \quad (1)$$

ここで $\gamma = ((\mathbf{s}_1, \mathbf{t}_1), \ldots, (\mathbf{s}_k, \mathbf{t}_k), \ldots, (\mathbf{s}_K, \mathbf{t}_K))$ は、対訳コーパスから導出された、あるアライメントを表し、Γ は、すべてのアライメントのセットで

ある。

単一のアライメントにおける確率は、次式で計算される、各バイリンガルシーケンスペアの積で与えられる。

$$p(\gamma) = \prod_{k=1}^{K} p((\mathbf{s}_k, \mathbf{t}_k)) \tag{2}$$

実験で用いるコーパスは、バイリンガルの単語のペアで構成される。従って、我々のモデルは、コーパスにおける各バイリンガルシーケンスペアの原言語と目的言語の両方の文字列が単語の境界にまたがらないよう制約される。この点を考慮すると、式2は、コーパス内にある全ての単語ペアにおける積として計算できる。

$$p(\gamma) = \prod_{w \in \mathcal{W}} \prod_{(\mathbf{s}_k, \mathbf{t}_k) \in \gamma_w} p((\mathbf{s}_k, \mathbf{t}_k)) \tag{3}$$

ここで、γ_w はバイリンガルパートペア w からの導出である。

2.2 Dirichlet process モデル

Dirichlet process は、サンプルパスが S の確率分布であるセット S（我々の場合、すべての可能性のあるバイリンガルシーケンスペアのセット）で定義される確率過程である。

我々のアプローチで使用する Dirichlet process モデルは、言語モデリング [21] で使用されるキャッシュモデルと似ている。直感的には、2つの基本的なモデルからなり、これらは、少なくとも以前に1回生成されている結果を生成するためのモデルと、まだ一度も生成されていない結果に確率を割り当てるモデルである。理想的には、モデルパラメーターを再使用するため、一度も生成されていないバイリンガルシーケンスペアの生成確率は、以前に確認したシーケンスペアを生成する確率より大幅に小さくなければならない。これは、我々が使用する Dirichlet process モデルの特徴である。学習初期においては、新しいシーケンスペアの生成が優先されるが、学習後半では、生成される可能性が低くなる。このような方法では、新たなシーケンスペアを生成するよりも、信頼性が高くなった既存のシーケンスペアを用いる方が有利である。これらのバイリンガルシーケンスペア数（学習コーパスに出現しないペアも含む）の確率分布は、コーパスから直接学習することができる。このように、未観測のシーケンスペアに確率を割り当てることができるモデルであり、これを利用して学習データにおける候補のスコア付けを行うことができる。

バイリンガルシーケンスペアで構成されるコーパスの生成について基本となる確率過程は通常、次の形式で記述される。

$$\begin{aligned} G|_{\alpha, G_0} &\sim DP(\alpha, G_0) \\ (\mathbf{s}_k, \mathbf{t}_k)|G &\sim G \end{aligned} \tag{4}$$

G は、すべてのバイリンガルシーケンスペアに対する離散的な確率分布で、Dirichlet process prior と base measure G_0 および集中度パラメーター α により得られる。集中度パラメーター $\alpha > 0$ は、G の分散を制御し、値が大きいほど、G_0 は G に近づく。

2.2.1 Chinese Restaurant Process

可能性のあるバイリンガルシーケンスペアは無限にあるため、G を直接推定することはできない。これを実行するため、Chinese Restaurant Process（CRP）の考えを、バイリンガルシーケンスペアへ適用する [22]。CRP では、すべてのバイリンガルシーケンスペアは、中華料理店でテーブルに運ばれる皿に対応する。また、バイリンガルシーケンスペアの累積数は、各テーブルに座る客の数に対応する。店に来る新しい客は、テーブルの客数に比例する確率でそのテーブルに着席でき、着席できた場合は、すでに客がいるテーブルの皿の料理を食べる。また一定の確率で、空いているテーブルに着席でき、この場合は、シェフが選んだ皿の料理（バイリンガルシーケンスペア）を食べることになる（この例えでは、シェフの選択が、基底分布 G_0 に対応する）。

2.2.2 基底分布

新たなシーケンスペアの生成を制御する base measure においては、ジョイントスペリングモデルを用いる。このモデルでは、次式を用いて、G_0 を求める。

$$\begin{aligned} G_0((\mathbf{s}, \mathbf{t})) &= p(|\mathbf{s}|)p(\mathbf{s}||\mathbf{s}|) \times p(|\mathbf{t}|)p(\mathbf{t}||\mathbf{t}|) \\ &\simeq \frac{\lambda_s^{|\mathbf{s}|}}{|\mathbf{s}|!} e^{-\lambda_s} v_s^{-|\mathbf{s}|} \times \frac{\lambda_t^{|\mathbf{t}|}}{|\mathbf{t}|!} e^{-\lambda_t} v_t^{-|\mathbf{t}|} \end{aligned} \tag{5}$$

ここで、$|\mathbf{s}|$ および $|\mathbf{t}|$ はそれぞれバイリンガルシーケンスペアのソース側とターゲット側の文字の長さ、v_s および v_t はそれぞれ原言語と目的言

語の用語（アルファベット）のサイズ、λ_s および λ_t は原言語文字列と目的言語文字列の予測長である。

このモデルでは、原言語と目的言語の文字列は別々に生成される。いずれの場合も、文字列長はポアソン分布により与えられる。このモデルでは、原言語と目的言語の文字列において任意の長さのバイリンガルシーケンスペアに確率が割り当てられる。

基底分布を、より緻密にモデル化することも可能である。文献[14][15]では、IBM モデル 1 の尤度を利用し、対訳コーパスにおける、単語アライメントを考慮したバイリンガルペアを生成している。この手法は、文字レベルに変換し、我々の翻字タスクにも適用することができるが、今後の検討課題としたい。

文献[18]に従い、パラメーター λ_s、λ_t、および α は、それぞれ、2、2、0.3 とした。本来は、これらのパラメーターは、開発セット等を用いて決定されるべきであるが、我々の実験において、文献[18]の設定で十分な性能が得られた。また、予備実験により、これらのパラメータの変更は、最終結果に大きな差を与えないことが確認された。

2.2.3 生成モデル

生成モデルは以下の式 6 で与えられる。この式では、$(\mathbf{s}_{-k}, \mathbf{t}_{-k})$ が与えられた条件において、コーパスから導出された k 番目のシーケンスペア $(\mathbf{s}_k, \mathbf{t}_k)$ に確率が割り当てられる。ここで $-k$ により「k を含まない、k の 1 つ手前まで」を表す。

$$p((\mathbf{s}_k, \mathbf{t}_k))|(\mathbf{s}_{-k}, \mathbf{t}_{-k})) = \frac{N((\mathbf{s}_k, \mathbf{t}_k)) + \alpha G_0((\mathbf{s}_k, \mathbf{t}_k))}{N + \alpha} \quad (6)$$

ここで、N はこれまでに生成されたバイリンガルシーケンスペアの合計数（CRP の例では、これまでに来店した客の数）、$N((\mathbf{s}_k, \mathbf{t}_k))$ は、これまでに、シーケンスペア $(\mathbf{s}_k, \mathbf{t}_k)$ の発生した回数（CRP の例では、あるテーブルに座った客の数）をそれぞれ表す。G_0 および α は前述の基底分布および集中度パラメーターである。

3 ベイジアンインタフェイス

3.1 ギブスサンプリング

トレーニングには、ブロックドギブスサンプラーを使用した。文献[21]では、アニーリングを使用したサンプラー内の混合の問題が報告されており、文献[19]では、ダイナミックプログラミングアプローチと共にブロックサンプラーを使用することにより、この問題は解決されている。我々のアルゴリズムは文献[19]の手法に近いが、アニーリングを使用することなくサンプラーが急速に収束することが分かっている（図 5）。繰返し数は、予備実験において、収束動作を確認した後、30 回に設定した。

サンプリングアルゴリズムは図 1 に示されており、各ステップにおける処理は以下のように行われる。まず、バイリンガル単語ペアが学習コーパスから無作為に抽出される。

2 番目に、選択されたバイリンガル単語ペアの

```
Input: Random initial corpus alignment
Output: Unsupervised alignment of the corpus according to the model
foreach iter=1 to NumIterations do
    foreach bilingual word-pair w ∈ randperm(W) do
        foreach alignment γ_i of w do
            Compute probability p(γ_i|h)
            where h is the set of data (excluding w) and its hidden alignment
        end
        Sample an alignment γ_i from the distribution p(γ_i|h)
        Update counts
    end
end
```

図 1 ブロックギブスサンプリングアルゴリズム

すべての可能なアライメントの確率分布は、これ以前に得られたアライメント情報、それぞれの回数を用いて計算される。我々の翻字タスクに含まれているのは短いシーケンスのため、総当り的に確率分布を計算することができるが、効率化のため、文献[19]のForward Filtering /Backward sampling（FFBS）のDynamic programmingアルゴリズムを拡張し、バイリンガルアライメントを行った。この手法の詳細を図2に示す。

図2は、バイリンガルペアに対するすべての可能なアライメントを表している。各ノードは、部分的なアライメントの仮説を表し、各矢印はその矢印の末尾から先頭への遷移に使用されるバイリンガルフェーズペアを表す。図中の矢印は、このシーケンスペア（式6のモデルで与えられる）の対数確率が示されているため、完全なアライメント推定の対数確率は、開始ノード〈s〉から終わりのシンクノードabbaまでの各ノードの矢印の数値の和で与えられる。図では、最も確率の高いパスの矢印は太線で示されている。これは、アライメント「a-b-ba」に対応しており、これは「a/A」および「ba/BA」の両方が日本語の表音文字と関連付けられているために妥当であり、日本語の「TSU」はすぐ右の子音が繰り返されることを示す。図内の最も確率の低いアライメントは「abb/A a/TSU-BA」によって与えられる。図内の対数確率は学習時における3回目の繰返しから取得される値であり、ここでは最も確率の高いアライメントは、他に比べ圧倒的に値が大きい。

バイリンガルシーケンスペアのアライメントされていない部分が両方で同じ場合、2つのノードが組み合わせられ、コンパクトで効率の良い表現が生成される。

FFBSアルゴリズムはアライメントグラフ上で直接適用可能である。2つのステップがあり、まず、順方向フィルタリングステップでは、各ノードについての計算を行い、左側方向へサブグラフの確率計算を行い、開始ノードに戻る。この確率αは、そのノードに格納される（図2中のα）。このステップではシンクノードから開始され、再帰的に、深さ優先、後順走査によって処理が進む。既に確率が計算されたノードは適宜マークされるため、そのノードでの計算は、一度だけ行われる。

逆方向サンプリングステップでは、可能性のあるすべてのアライメントの確率分布に従い、バイリンガル単語ペアの導出がサンプリングされる。ここでは、順方向フィルタリングステップによって計算されたαの値を使用する。逆方向サンプリ

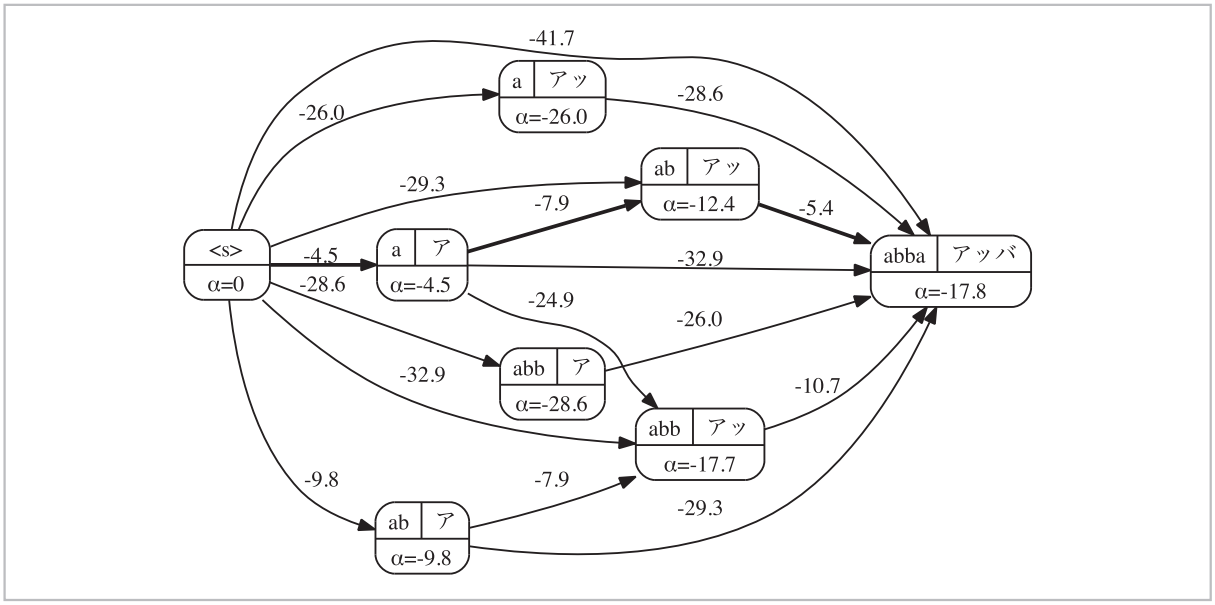

図2 英語の文字列「abba」と日本語の文字列「アッバ」に関してすべての可能なアライメントを表す

ノードのαは、そのノード自身を含むサブグラフの対数確率を表す。矢印上の値は、末尾から先頭への遷移する際のバイリンガルフレーズペアの対数確率で、式6のモデルによって与えられる。

ングもシンクノードから再帰的に計算される。入ってくる各矢印において、サンプル内のその矢印を含む確率は、矢印の確率と矢印の末尾のαの値との積で与えられる。この値は入ってくる各矢印について計算され、矢印サンプリングは、矢印の確率分布を用いて行われる。図の開始ノードに到達するまで、矢印の先頭部分で再帰的にサンプリングが行われる。

得られた矢印が、現繰返しにおいて、サンプリングにより導出されたバイリンガルペアである。このサンプルは、モデルにおける、すべての導出の確率分布に従う。

3.2 シーケンスペア抽出

典型的なフレーズベースのSMTシステムのフレーズテーブル生成中に、GIZA＋＋は、単語レベルのアライメントを生成するため、原言語から目的言語および、目的言語から原言語の2回実行される。この手順に続いて、grow-diag-final-and のオプションでは、2回のGIZA＋＋実行で生成される単語アライメントと一致するすべてのフレーズが抽出される。ギブスサンプリングにおける最後の繰返しで得られたアライメントからフレーズテーブルを構築する場合においても、同様の考え方によるヒューリスティックを用いる。実験では、これがシステムの性能を大幅に向上させることが確認された。

アライメントが実行されたコーパスからのフレーズ抽出に使用するアルゴリズムは次のとおりである。単一のバイリンガルワードペア内で、すべての隣接バイリンガルシーケンスペアをすべての可能な方法で生成するが、SMTシステムのトレーニングに使用する最大フレーズ長パラメーターに一致するよう、原言語と目的言語のフレーズのサイズを制限する（我々の実験では7に設定した）。これは厳密には必須ではないが、我々は、ベイジアンアライメントから生成されるフレーズテーブルがベースラインシステムによって生成されるものと同等になるように、この手順を実行した。このagglomerationアルゴリズムの概要については、図3で説明する。これにより抽出されるワードペアの例を図4に示す。

4 GIZA＋＋との比較

4.1 ベースラインシステム

実験では、文献[9]で紹介されているフレーズベースの機械翻訳手法を使用し、ログリニアのフレームワーク[23]を用いて、モデルを統合する。GIZA＋＋[8]を使用してワードアライメントを実行し、MOSES[9]ツールを使用してシーケンスペア抽出を行った。デコーダは、自作のフレーズベース機械翻訳のデコーダOCTAVIANを使用した。OCTAVIANは、MOSES[9]SMTデコー

```
INPUT:      a sequence of bilingual sequence pairs: ARRAY-of-sequence-pairs derivation
OUTPUT:     a set of all pairs formed by agglomerating the pairs in derivation

SET-of-sequence-pairs     agglomerations    = empty
ARRAY-of-sequence-pairs   chunk             = empty

FOR chunk in all subsequences of derivation
{
    sequence-pair pair = concatenation of sequence pairs in chunk

    IF (pair has source sequence length <= MAX_SOURCE_SEQUENCE_LENGTH AND
        pair has target sequence length  <= MAX_TARGET_SEQUENCE_LENGTH)
    {
        INSERT pair into agglomerations
    }
}

PRINT all_agglomerations
```

図3 シーケンスペアのagglomerationアルゴリズム

ダと同じ原理に従って動作し、これらの実験向けに構築されたデコーダである。

これらの実験では、Witten-Bell スムージングによって構築された 5-gram 言語モデルを用いた。システムは、ログリニアモデルの重みの最適化には、開発セットの BLEU スコアが最適になるよう、Minimum Error Rate Training（MERT）[24] を用いてチューニングしている。

Rama と Gali [25] は翻字のシーケンスペア抽出に関していくつかの手法を評価、比較しており、この結果、grow-diag-final-and が最も高性能であると報告されている。従って、本稿ではベースラインシステムにおいてこの方法を採用している。

4.1.1 デコードの設定

本稿における実験では、ビーム幅 100、スタック閾値なし、モノトーンデコードで実施された。

4.2 実験データ

学習データとして、NEWS2010 ワークショップの翻字シェアードタスクで使用した 27,993 単語対を用いた。パラメータチューニング用には 3,606 単語対からなる開発セットを用いた。また、評価セットとしては、前述の 2 つのデータセットに含まれない 1,935 単語対を用いた。これらの 3 つのデータの詳細を表 1 に示す。

我々はフレーズベースの SMT システムを学習するデータを使用して、英語から日本語への翻字を実行した。同じパラレルデータセットで Dirichlet process モデルをトレーニングし、繰返し終了時（繰返し回数　30）のコーパスのアライメントから翻字のフレーズテーブルを抽出した。

4.3 トレーニング手順

ギブスサンプリングでは、まず、コーパスをラ

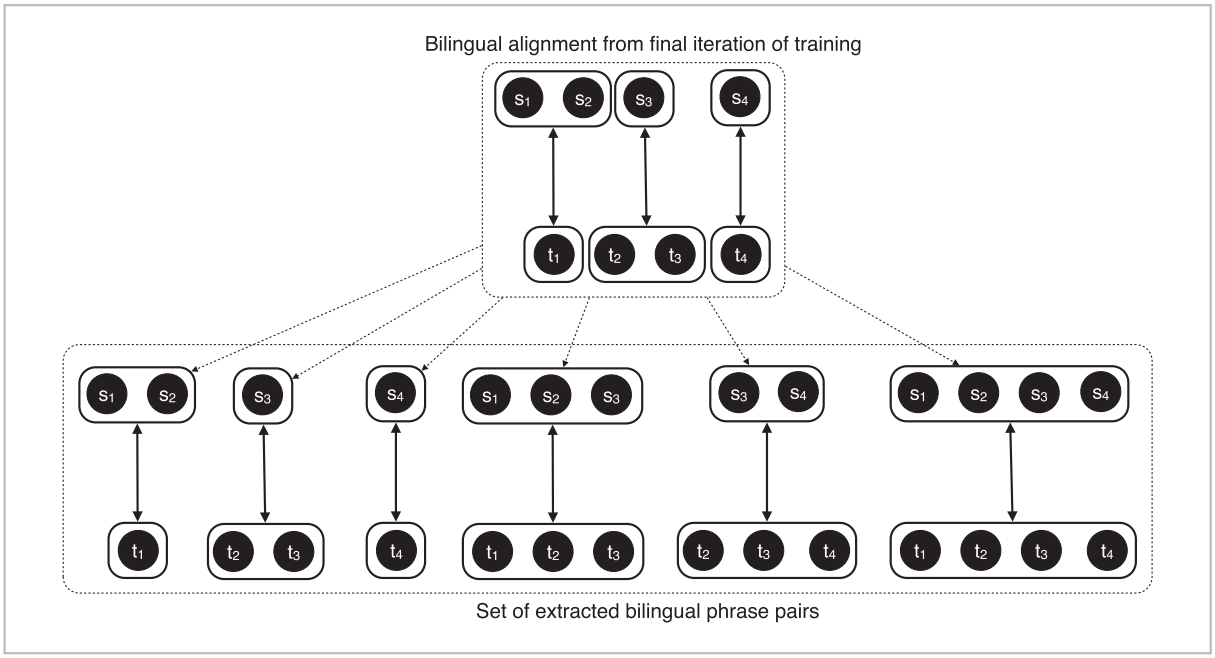

図 4 単一のバイリンガル単語ペアにおける、シーケンスペア抽出

表 1 英語-日本語の学習データの統計

Corpus	word-pairs	Characters		Avg. Word Len.	
		En	Ja	En	Ja
Training	27993	188941	131275	6.75	4.69
Development	3606	24066	16651	6.67	4.62
Evaluation	1935	11863	8199	6.13	4.24

ンダムにアライメントし、サンプリングを開始する。つまり、これは、コーパス内の各バイリンガル単語ペアについて、そのペアに関して可能性のあるアライメントの一様分布から単一のアライメントをサンプリングすることである。GIZA＋＋を前処理として用い、この結果を開始点としてサンプリングを行う方法が、有利かつ効率的であると考えられる。これは、現状の学習方法においても高い性能に到達しており、次に述べる実験結果によってその有用性を確認することができる。

4.4 評価手順

本稿の評価では、NEWS2010 翻字生成のシェアードタスク[26]で使用された評価基準を用いた。実験の結果では、ACC は 1 best の精度のスコアで、システム出力する 1 best が、正解に完全に一致する割合である。F-score は、システム出力と正解との距離により評価している。これらの評価法については、文献[26]を参照されたい。本稿では ACC と F-score の 2 つの評価指標を用いる。NEWS2010 においては、この 2 つに加え別の評価指標が用いられていたが、これらについても、ACC や F-score と似通った特性を持っていた。

4.5 結果

4.5.1 学習

学習時の繰返し回数と収束の関係を図 5 に示す。繰返し回数に対する学習コーパスを通じて各パスの最後にサンプリングされたバイリンガルシーケンスの対数確率をプロットしている。図 5 を見ると、品質の低い初期アライメントから急速に性能が向上し、その後は引き続き徐々に向上していることが確認できる。初期の無作為アライメントの対数確率は-1.5e06 であるが、図では省略されている。

4.5.2 自動評価結果

英語–日本語の翻字タスクの結果を表 2 に示す。学習の終了時におけるサンプルのみからなるシーケンスペアを使用すると、ベースラインシステムよりも性能が低くなっている。この方法で得られたフレーズテーブルのエントリは 3,372 であるのに対し、GIZA＋＋アライメントから抽出されたフレーズテーブルのエントリは 140,000 以上である。さらに、これらのシーケンスペアは、ベースラインシステムのフレーズテーブルのシーケンスペアと比較すると、非常に短い。ベースラインシステムでは 5 文字前後であったが、提案手法では、原言語と目的言語の両方で平均約 3 文字であった。

agglomeration アルゴリズムを適用した場合では、ベースラインのフレーズテーブルと同様、5

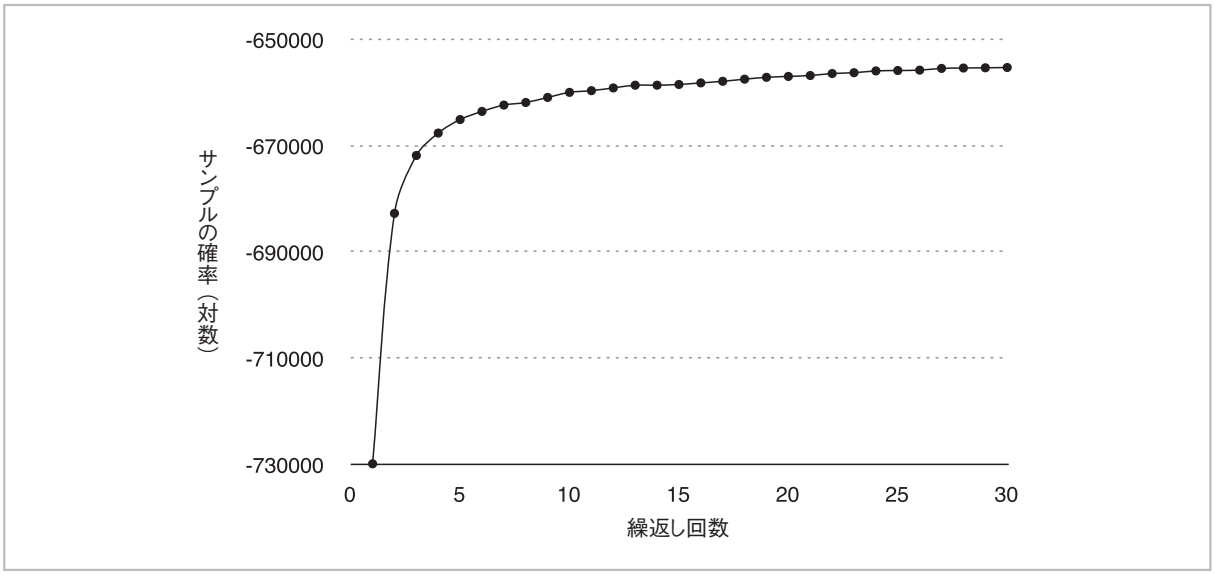

図 5 繰返し回数と学習の収束の関係

文字前後の長さのシーケンスペアとなった。また、フレーズテーブルのサイズは100,000エントリ程度で、ベースラインと比較し30％小さい。このように提案手法で得られるフレーズテーブルのサイズは小さいにも関わらず、ベースラインシステムのACCで約1％の改善がみられた。さらに、シーケンスペアは3,372個のコンポーネントシーケンスペアの連結であるため、必要に応じてこのモデルは非常にコンパクトに格納できる。また、ベースラインモデルと提案手法により得られるモデルを補間することにより、さらに性能を改善することができた。この改善はスムージングの効果によるものと考えられる。

我々の実験では、MERTによるチューニングを行うため、ベースラインモデルに有利になっている。提案手法において、得られたフレーズテーブルを用いてチューニングを行うことにより、高いスコアが得られる可能性がある。しかしながら、ここでは2つの異なったMERTの処理による影響を出さないため、このような方法は取っていない。ただし、agglomerationアルゴリズムの有効性を検証するため、agglomerationアルゴリズムを適用しない場合のフレーズテーブルに対して、チューニングを行っている。図2中のtuned on Bayesian phrase-tableがこれに該当する。レングスモデルに対する最適な重みが得られたために性能が向上していると思われるが、フレーズが短いので、agglomerationアルゴリズムを適用したフレーズテーブルの場合と比較して性能は劣っている。

興味深いのは、フレーズをより大きな単位にグルーピングするだけで、システムの性能が著しく向上したことである。これは、フレーズベース翻訳のアプローチの利点の1つである。agglomerationアルゴリズムにより得られるモデルは、より大きいシーケンスペアが目的言語文字列の構築に使用される場合に、その前後の文字列をコンテキストとして利用できている。agglomerationアルゴリズムを利用しない場合には、このようなコンテキストの情報は言語モデルのみが与えることになる。実験では5-gram言語モデルを使用したにもかかわらず、agglomerationアルゴリズムによる基本翻訳単位として長いシーケンスペアを含むモデルが性能改善に貢献している。表2のフレーズテーブルのオーバーラップの数に注目すると、agglomerationアルゴリズムによって、GIZA＋＋およびMOSESのフレーズ抽出を使用して生成されたフレーズテーブルと57％のオーバーラップがあることが分かる。

最終の実験では、学習過程における繰返し回数5回目から最後の繰返しまでに得られるシーケンスペアを集積していくことにより、37％大きいフレーズテーブルが得られることを確認したが、性能においては顕著な向上が得られなかった。

4.6 デコードにおける整合性

ここでは、システムの性能が向上した理由を精査する。仮説としては、ベイジアンシステムでは、デコード時において、より整合性の高いフレーズテーブルを生成できるという点が考えられる。Dirichlet processモデルでは、以前に検出されたバイリンガルシーケンスペアが再利用され

表2 フレーズテーブルの統計量と、3つのシステムの実験結果

フレーズ抽出モデル	ACC	F-score	フレーズテーブルエントリ	フレーズテーブルオーバーラップ（%）	平均フレーズ長	
					英語	日本語
GIZA＋＋およびgrow-diag-final-and	0.313	0.745	143382	100	5.41	4.80
ベイジアンアライナー（ベースラインフレーズテーブル）	0.278	0.726	3372	2	2.60	2.75
ベイジアンアライナー（ベイジアンフレーズテーブル）	0.283	0.732	3372	2	2.60	2.75
ベイジアンアライナー（+提案手法）	0.323	0.748	102507	57	5.54	4.83
ベイジアンアライナー（+統合）	0.329	0.752	164258	100	5.46	4.81

ここで、＋agglomeratedは、トレーニングの終了時に図3で説明した方法によりシーケンスペアを抽出した結果である。＋integratedでは、ベースラインシステムのフレーズテーブルと前述の方法を統合した結果で、等しい重みで線形補間している。提案手および統合結果のACCスコアを除くと、信頼係数0.05のt検定において、有意な差がある。

やすいという点からこのような仮説が考えられる。これにより、コンパクトなフレーズテーブルが生成され、コーパス内の類似した原言語の単語が同じ目的言語にデコードされる傾向が強くなる。仮説を検証するために、デコーダを修正し、評価データをデコードする際に使用されるバイリンガルシーケンスペアの種類の数をカウントした。提案手法であるベイジアンモデルから生成されたフレーズテーブルを使用したデコードにおいては、合計で3,496個のユニークなシーケンスペアを使用したのに対して、GIZA++および*grow-diag-final-and*で抽出したフレーズテーブルを使用したデコードでは、合計で3,970個のフレーズペアが必要であったことが分かった。ベイジアンモデルのフレーズテーブルの3,496個のシーケンスペアは、さらに1,289個のコンポーネントバイリンガルペアに分解でき、これらのペアは学習の最終段階に取得されたサンプル内のアライメントに存在していた。

4.7 学習時間

ベイジアン法はタスクによっては、学習時間の長さが問題になることがある。通常、長いシーケンスを処理できる最適化が必要になるが、翻字データの場合、シーケンス長が短いので、現実的な時間で処理することができる。例えば、日本語–英語のNEWS2010における全学習工程は、15分以内で完了することができる。データの各繰返し過程の所要時間は平均すると30秒前後であった。

5 m2mアライナーとの比較

この実験では、ベイジアンアライナーを、多対多のアライメントが可能なアライメントツール（EMアルゴリズムを使用してトレーニングされ、文献[28]で提案された手法に基づいて実装、公開されているm2mアライメントツール[*1][27]）と比較する。

NEWS2011シェアードタスク[29]に対するNICTエントリと同じ条件での比較を行うため、アライメント部分以外の実験条件を同一にしている。実験では、2009 NEWSワークショップのデータセットを使用し、前述のF-scoreにより評価を行った。

このシステムは、**4**の実験で使用した翻字システムをベースにしているが、joint source channelモデルがデコード過程に直接組み込まれるという点で異なる。アライナーは既定の設定と、原言語と目的言語のセグメントサイズ制限が同じ状態で実行された。アライナーのパラメーターを調整することで性能が向上する可能性があるが、ここでは対処していない。評価結果を表3に示す。すべての実験で、ベイジアンアライナーによって最高の性能が得られた。目的言語側の文字セットのサイズが大きい場合にとくに大きな改善が見られる。各言語対における、目的言語の文字セットのサイズは、表3の「Target Types」の列に示す。実験において、原言語は英語または日本語であり、原言語が日本語の場合もローマ字化した日本語であった。従って原言語の文字セットのサイズは、すべての実験において近い値で27前後で

＊1 http://code.google.com/p/m2m-aligner/

表3 提案手法とm2mアライナーの評価結果と統計量

言語対	Target Types	m2m F-score	ベイジアン F-score	m2m			ベイジアン		
				1-grams	2-grams	3-grams	1-grams	2-grams	3-grams
En-Ch	372	0.858	0.880	9379	44003	75513	4706	38647	72905
En-Hi	84	0.874	0.884	3114	15209	30195	1867	20218	34657
En-Ko	687	0.623	0.651	4337	11891	14112	2968	11233	14729
En-Ru	66	0.919	0.922	1638	6351	14869	1105	12607	23250
En-Ta	64	0.885	0.892	2852	14696	27869	1561	17195	30244
Jn-Jk	1514	0.669	0.767	7942	27286	38365	3532	22717	37560

あった。表3のN-gram統計に注目すると、サイズの大きい文字セットを持つ言語の場合、ユニグラムの数はm2mモデルで使用されるときの半分未満になる。コンパクトなモデルが得られるのは、ベイジアンモデルの特徴の1つである。新しいパラメーターをモデルに追加するには著しくコストが高くなるため、提案手法のように既存のパラメーターを再利用するという戦略が上手く働いている。

これらの2つのアプローチ間の性能の差が言語モデルのスパースネスに起因するという可能性についても検討したが、両モデルにおける2-gramsと3-gramsの数は極めて類似している。

コンパクトな1-gramにより、曖昧性の解消に貢献していることを確認するため、開発セットのパープレキシティによる評価も行った。システム間の差が最大であるJn-Jkタスクにおいて、ベイジアンアライメントで学習される結合言語モデルにおける1～3-gramの開発セットパープレキシティは、それぞれ218.3、88.4、87.5であった。一方、m2mのパープレキシティではそれぞれ321.8、120.5、および119.3であった。

コーパスのアライメントで使用されるセグメントの数は、この実験の両システムと同じであった。

表4に、両アライメント手法により得られた実例を示す。ベイジアンアライメントは高度な一貫性があり、原言語文字列「ara」はすべての場合の単一の単位として同様にアライメントされている。一方、m2mシステムでもある程度の一貫性を示すが、一部のシーケンスの始まり部分で一貫性が崩れている箇所がある。興味深いことに、ベイジアン法では漢字の正確な読みに従ってアライメントが行われている。このような現象は、文献[29]でも報告されている。

6 音声翻訳実証データを使用した主観評価

ここでは、我々の翻字システムの有効性を確認するため、翻字を機械翻訳システムの未知語処理として用いる事を想定した評価を行う。評価においては、音声翻訳実証実験により、携帯型端末上で稼動する音声翻訳アプリケーション上で収集された現利用データを用いる。

この評価のテストセットは、2009年度に行われた音声翻訳の実証実験における、ユーザーログデータから抽出されたものである[30]。図6に示すように、実証実験は、日本国中の5つの地方、関東、関西、九州、北海道、および中部で行われた。この実証実験は、総務省の「地域の観光振興に貢献する自動音声翻訳技術の実証実験」の一環

表4 提案手法とm2mで得られるアライメントの例

m2m			ベイジアン		
arad →荒	a →田		ara →荒	da →田	
ar →新	ae →江		ara →新	e →江	
ar →荒	ahori →堀		ara →荒	hori →堀	
ar →新	ai →井		ara →新	i →井	
ar →新	ai →居		ara →新	i →居	
ar →荒	ai →井		ara →荒	i →井	
ar →荒	ai →居		ara →荒	i →居	
araj →荒	ima →島		ara →荒	jima →島	
arak →新	i →木		ara →新	ki →木	
arak →荒	i →木		ara →荒	ki →木	
ar →荒	akid →木	a →田	ara →荒	ki →木	da →田
ar →荒	ao →尾		ara →荒	o →尾	
ar →荒	ao →生		ara →荒	o →生	
ar →荒	aoka →岡		ara →荒	oka →岡	
arasa →荒	wa →沢		ara →荒	sawa →沢	
ar →荒	aseki →関		ara →荒	seki →関	

として行われた。テストセット作成のため、まず、5地方のユーザーログデータの書き起こしから100文ずつサンプリングした。次に、この500文を実際に機械翻訳システムにより翻訳させ、出力に未知語が含まれる74文をテストセットとした。テストセット内の文には、ひらがなまたは漢字の文字で書かれた固有名詞が少なくとも1つ含まれている。

評価実験における翻訳方向は日英方向である。ベースラインシステムとして、我々は691,829文対からなる日英対訳コーパス、BTECを用い、標準的なフレーズベースの統計機械翻訳システムを使用した。また、機械翻訳の性能の上限を確認するため、単語のカテゴリとテストセットにおけるすべての固有名詞の英語翻訳で構成される固有名詞対訳辞書を手動で構築した。翻訳前処理において、入力文は翻訳前に、原言語文中の固有名詞はトレーニングコーパス内にある同じカテゴリの高頻度の単語と置き換えられる[31]。その後、機械翻訳により、置き換えられた原言語文が翻訳される。最後に、後処理として機械翻訳された目的言語内の該当する単語が前述の辞書を使って置き換えられる。このように、高頻度の単語に置き換える理由は、コーパス内の高頻度語は、モデルが学習されている、即ち、高頻度の単語はすでにフレーズテーブルでよく登場するため、十分な統計が提供されている可能性があるからである。主観評価においては、1名の評価者が5段階のスケールによって評価を行っている。

本評価で用いた翻字システム単体の性能は、NEWS2010ワークショップと同じ評価方法での翻字精度は19.14%であった。この値は、NEWS2010タスクのスコアより低いが、これは、本評価セット内に、翻字ではなく、対訳の関係になっているような名詞が含まれていることに起因すると考えられる。このように単語単位の翻字精度は低いがF-scoreは71.03であったことから、文字レベルの性能は極めて優れていることが分かる。

実際のデータを見ると、テストセットに含まれる単語における翻字誤りの大部分が、翻字ではなく翻訳されるか、あるいは部分的に翻字されて部分的に翻訳される必要があるような箇所で起きている。例えば、「伊丹空港」がその例で、これは「Itami Airport」と翻字（翻訳）されるべきである。我々の翻字システムは出力として「itami-kuukou」を生成する。完全な仮名への翻字であるが、翻訳システムの未知語処理としては不正確である。今後の研究では、翻字すべき場合と翻訳すべきでない場合とを特定する方法について取り組む必要がある。これらの誤りはテストセット中

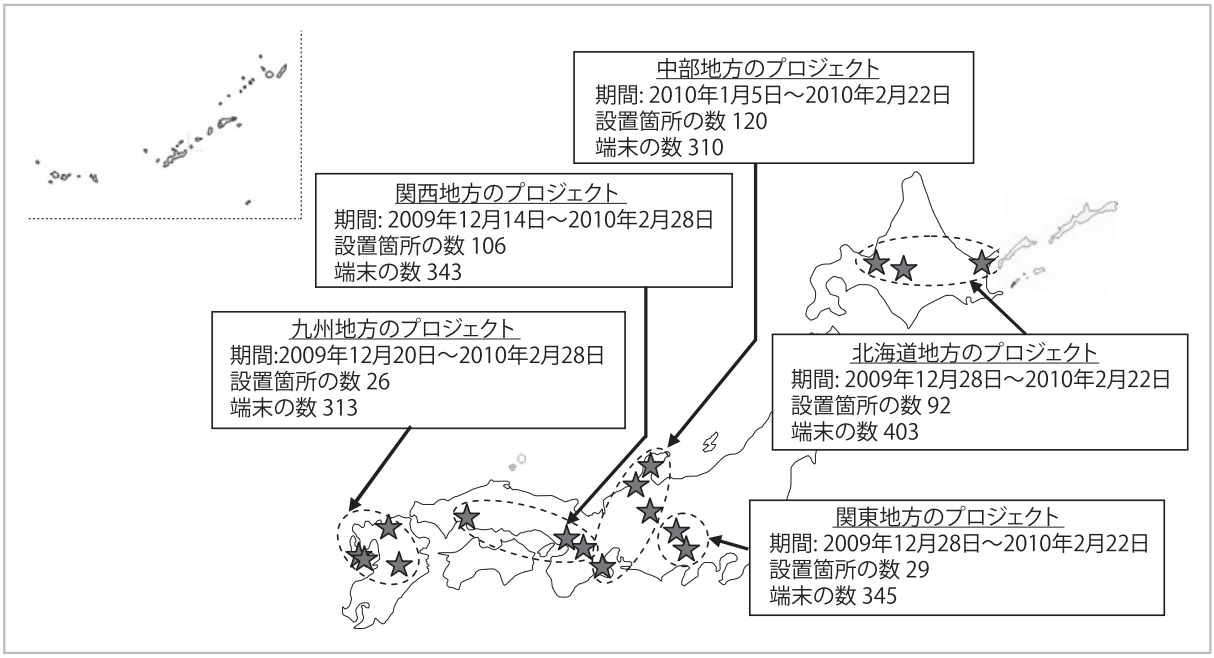

図6 5地方のプロジェクトの概要

の4%の文に対して生じた。並べ替えの問題も今後対応すべき課題である。たとえば「富士山」などの例がある。この場合のシステムの出力は「fujisan」であったが、正確な出力は「Mount Fuji」か「Mt. Fuji」である。この例では、翻字と翻訳の両方が必要であるが、さらに単語の順番が入れ替えられる。この並べ替え工程をモデル化することにより、システムの性能が更に向上すると考えられる。

図7に実証テストの翻訳評価結果を示す。縦軸は、全テスト文に対して、acceptable以上の評価結果が得られた翻訳の割合である。前述の固有名詞対訳辞書のサイズと翻訳性能の関係を確認するため、辞書サイズを制御した結果も示している。図中の横軸は辞書のカバレージを示し、これは評価セット内の固有名詞の合計数に対する辞書でカバーされている固有名詞の割合である。図中の○は従来の辞書利用[31]による機械翻訳の結果を表し、●のプロットは、固有名詞対訳辞書にエントリがない場合のみ、翻字結果を使用した時の結果を表している*2。

図から、機械翻訳システムで翻字を使用すると、翻訳品質が向上することが分かる。図7の横軸の0の値は、固有名詞対訳辞書が無い場合でのシステム性能である。このように固有名詞対訳辞書が無い場合でも、翻字システムを未知語処理に用いることにより、高い訳質が得られている。固有名詞対訳辞書によるカバレージが上がるにつれ、グラフの2本の線の差は徐々に減少する。当然のことながら、横軸の値が1になると、両手法とも固有名詞対訳辞書だけを用いるので2本の線は重なる。辞書のカバレージが小さければ、翻字によってシステム性能が大きく改善されることが分かる。“ベースライン＋辞書＋翻字”について、横軸の値が0と1の場合について注目すると、翻字誤りが翻訳文の品質に与える影響を知る事ができる。横軸の値が1の場合は、翻字誤りが全く無い場合に相当しており、現状の翻

*2 翻字の結果を機械翻訳と組み合わせるため、文献[31]で提案された従来の辞書ベースの手法と同じ枠組みを使用した。この方法では、単語のカテゴリが既知である必要がある。我々の実験では、翻字された単語であっても、単語のカテゴリが既知であるものとして処理を行った。

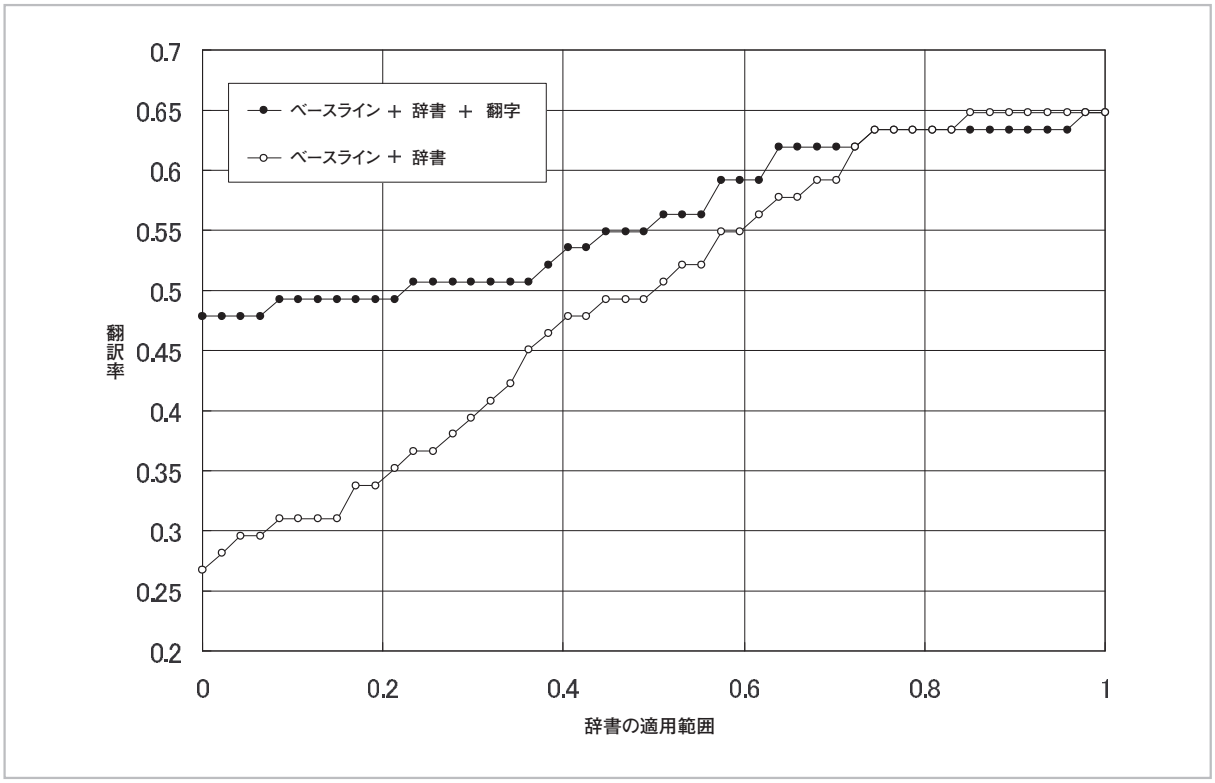

図7 機械翻訳の評価の結果

字性能（横軸が0の場合の性能）と比較し、性能が明らかに向上する。従って、今後の研究では、機械翻字システムの品質を向上させる必要がある。

7 結論

本稿では、ベイジアンバイリンガルアライメントの手法を紹介し、それをフレーズベースの統計機械翻訳による翻字向けに翻訳モデルを構築するタスクに適用した。さらに、主観評価を実行し、翻字システムを機械翻訳システムの未知語処理に用いる場合の有効性について示した。実験の結果から、翻字を機械翻訳システムの未知語処理に用いることで、機械翻訳の性能を大幅に改善することが確認できた。

従来法である、フレーズアライメントのモデルがEMアルゴリズムと組み合わせられた最尤推定により行われており、これが過学習の原因になっている。本稿で述べたバイリンガルアライメント用のベイジアンモデルは、この課題解決を目的とした。我々のアプローチにより、最尤推定に起因する課題が解決されただけでなく、コーパスから直接多対多のアライメントを抽出することが可能となった。

提案手法の評価として、標準的なGIZA++/grow-diag-final-andフレーズ抽出手順、およびm2mバイリンガルシーケンスアライナーを使用して構築されたモデルと直接比較する評価実験を行った。実験の結果、ベイジアンアプローチによって、これらの従来法よりも、コンパクトかつ整合性の高いモデルが得られ、翻字性能も優れていることが示された。

提案手法はDirichlet processモデルによって任意のバイリンガルワード単語ペアに確率を割り当てることができるという特徴も持っている。そのためモデルを利用することにより、翻字のデータのマイニングや、コーパスのフィルタリングなどにも利用することが可能である。文献[32]では、翻字のマイニングタスクにおいて提案手法が応用され、高い性能が得られている。

今後の研究課題として、基本となるDirichlet processモデルを改良した高次の階層的モデルを導入し、精度改善に取り組みたい。

参考文献

1 K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, pp. 311–318, Association for Computational Linguistics, 2001.

2 G. Doddington, “Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics,” Proceedings of the HLT Conference, San Diego, California, 2002.

3 X. Duan, D. Xiong, H. Zhang, M. Zhang, and H. Li, “I2r's machine translation system for iwslt 2009,” Proceedings of the International Workshop on Spoken Language Translation, pp. 50–54, 2009.

4 A. Finch and E. Sumita, “Phrase-based machine transliteration,” Proc. 3rd International Joint Conference on NLP, Hyderabad, India, 2008.

5 T. Rama and K. Gali, “Modeling machine transliteration as a phrase based statistical machine translation problem,” NEWS '09: Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration, Morristown, NJ, USA, pp. 124–127, Association for Computational Linguistics, 2009.

6 S. Noeman, “Language independent transliteration system using phrase based smt approach on substrings,” NEWS '09: Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration, Morristown, NJ, USA, pp. 112–115, Association for Computational Linguistics, 2009.

7 H. Li, M. Zhang, and J. Su, “A joint source-channel model for machine transliteration,” ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, p. 159, Association for Computational Linguistics, 2004.

8 F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," Computational Linguistics, Vol. 29, No. 1, pp. 19–51, 2003.

9 P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowa, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: open source toolkit for statistical machine translation," ACL 2007: proceedings of demo and poster sessions, Prague, Czeck Republic, pp. 177–180, June 2007.

10 D. Yang, P. Dixon, Y. C. Pan, T. Oonishi, M. Nakamura, and S. Furui, "Combining a two-step conditional random field model and a joint source channel model for machine transliteration," NEWS '09: Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration, Morristown, NJ, USA, pp. 72–75, Association for Computational Linguistics, 2009.

11 D. Marcu and W. Wong, "A phrase-based, joint probability model for statistical machine translation," In Proceedings of EMNLP, pp. 133–139, 2002.

12 M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," Speech Commun., Vol. 50, No. 5, pp. 434–451, 2008.

13 P. Blunsom, T. Cohn, C. Dyer, and M. Osborne, "A gibbs sampler for phrasal synchronous grammar induction," Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL, Suntec, Singapore, pp. 782–790, Association for Computational Linguistics, August 2009.

14 J. DeNero, A. Bouchard-Côté, and D. Klein, "Sampling alignment structure under a bayesian translation model," Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '08, Stroudsburg, PA, USA, pp. 314–323, Association for Computational Linguistics, 2008.

15 G. Neubig, T. Watanabe, E. Sumita, S. Mori, and T. Kawahara, "An unsupervised model for joint phrase alignment and extraction," ACL, pp. 632–641, 2011.

16 J. Wuebker, A. Mauser, and H. Ney, "Training phrase translation models with leaving-one-out," Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, pp. 475–484, Association for Computational Linguistics, July 2010.

17 Y. Huang, M. Zhang, and C. L. Tan, "Nonparametric Bayesian Machine Transliteration with Synchronous Adaptor Grammars," ACL (Short Papers), pp. 534–539, 2011.

18 J. Xu, J. Gao, K. Toutanova, and H. Ney, "Bayesian semi-supervised chinese word segmentation for statistical machine translation," COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics, Morristown, NJ, USA, pp. 1017–1024, Association for Computational Linguistics, 2008.

19 D. Mochihashi, T. Yamada, and N. Ueda, "Bayesian unsupervised word segmentation with nested pitman-yor language modeling," ACL-IJCNLP '09: 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: Vol. 1, Morristown, NJ, USA, pp. 100–108, Association for Computational Linguistics, 2009.

20 A. Finch and E. Sumita, "A Bayesian Model of Bilingual Segmentation for Transliteration," Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT), ed. M. Federico, I. Lane, M. Paul, and F. Yvon, pp. 259–266, 2010.

21 S. Goldwater, T. L. Griffiths, and M. Johnson, "Contextual dependencies in unsupervised word segmentation," ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, Morristown, NJ, USA, pp. 673–680, Association for Computational Linguistics, 2006.

22 D. J. Aldous, "Exchangeability and related topics," in École d'été de probabilités de Saint-Flour, XIII—1983, Lecture Notes in Math., Vol. 1117, pp. 1–198, Springer, Berlin, 1985.

23 F. J. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), pp. 295–302, 2002.

24 F. J. Och, "Minimum error rate training for statistical machine translation," Proceedings of the ACL, 2003.

25 T. Rama and K. Gali, "Modeling machine transliteration as a phrase based statistical machine translation problem," In Proc. ACL/IJCNLP Named Entities Workshop Shared Task, 2009.

26 M. Z. Haizhou Li, A Kumaran, and V. Pervouchine, "Whitepaper of news 2010 shared task on transliteration generation," In Proc. ACL Named Entities Workshop Shared Task, 2010.

27 S. Jiampojamarn, G. Kondrak, and T. Sherif, "Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion," Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, Rochester, New York, pp. 372–379, Association for Computational Linguistics, April 2007.

28 E. S. Ristad and P. N. Yianilos, "Learning string edit distance," IEEE Transactions on Pattern Recognition and Machine Intelligence, Vol. 20, No. 5, pp. 522–532, May 1998.

29 A. Finch, P. Dixon, and E. Sumita, "Integrating models derived from non-parametric bayesian cosegmentation into a statistical machine transliteration system," Proceedings of the Named Entities Workshop, Chiang Mai, Thailand, pp. 23–27, Asian Federation of Natural Language Processing, Nov 2011.

30 H. KAWAI, R. ISOTANI, K. YASUDA, E. SUMITA, U. Masao, S. MATSUDA, Y. ASHIKARI, and S. NAKAMURA, "An overview of a nation-wide field experiment of speech-to-speech translation in fiscal year 2009 (Japanese only)," Proceedings of 2010 autumn meeting of Acoustical Society of Japan, pp. 99–102, 2010.

31 H. Okuma, H. Yamamoto, and E. Sumita, "Introducint a translation dictionary into phrase-based smt," The IEICE Transactions on Information and Systems, Vol. 91-D, No. 7, pp. 2051–2057, 2008.

32 T. Fukunishi, A. Finch, S. Yamamoto, and E. Sumita, "Using features from a bilingual alignment model in transliteration mining," Proceedings of the 3rd Named Entities Workshop (NEWS 2011), pp. 49–57, 2011.

（平成 24 年 6 月 14 日 採録）

Andrew Finch
ユニバーサルコミュニケーション研究所
多言語翻訳研究室主任研究員
Ph. D. Computer Science
機械翻訳、自然言語処理
andrew.finch@nict.go.jp

安田圭志（やすだけいじ）
ユニバーサルコミュニケーション研究所
多言語翻訳研究室主任研究員
博士（工学）
機械翻訳、自然言語処理
keiji.yasuda@nict.go.jp