形態素解析とHMMを用いたフレーズアラインメント

揚石　亮平†　　三浦　孝夫†

† 法政大学大学院 工学研究科 電気工学専攻　〒184–8584 東京都小金井市梶野町 3-7-2
E-mail: †ryohei.ageishi.08@gs-eng.hosei.ac.jp, ††miurat@k.hosei.ac.jp

あらまし　本稿では、統計機械翻訳の問題の１つとして知られているフレーズアラインメント問題を論じる。フレーズアラインメントでは、異なる言語表現間のフレーズを対応させることによって翻訳精度を向上させる。しかし、日本語と英語では表現形式が大きく異なり、フレーズ間の対応を取ることは容易ではない。ここでは、隠れマルコフモデルと形態素解析を用いた対応づけを自動的に行い、その有用性を検証する。
キーワード　統計翻訳，アラインメント，隠れマルコフモデル

Phrase Alignment using by Hidden Markov Model and Morphological Analysis

Ryohei AGEISHI† and Takao MIURA†

† Dept.of Elect.& Elect. Engr., HOSEI University　3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
E-mail: †ryohei.ageishi.08@gs-eng.hosei.ac.jp, ††miurat@k.hosei.ac.jp

Abstract　In this investigation we discuss Phrase Alignment Problem as known one of Statistical Machine Translation problems. In Phrase Alignment, we can improve the translation accuracy by making the phrase between different language expressions correspond. However, Japanese expressive form is greatly different from English and it is not easy to take the phrase alignment. Here to align the phrase automatically using by Hidden Markov Model and Morphological Analysis. We show some experimental results to see the effectiveness.
Key words　Statistical Machine Translation，Alignment, Hidden Markov Model

1. ま え が き

近年，インターネットの普及により，他言語による情報に接することが容易になってきた．それに伴い，機械翻訳に注目が集まるようになり，様々な研究が行われ，その学習目的のための大規模対訳コーパスが利用可能となった．そして，現在では，対訳コーパスからの学習方法として，統計翻訳 [1](Statistical Machine Translation，SMT) が著しい発展を遂げている．

SMT の 1 つの問題として，アラインメント問題が挙げられる。ここで，アラインメントとは，ある単語とその語と等価な意味を持つ翻訳語との対応を取ることである．例えば，*I like apples* と**私はりんごが好き**という 2 つの文の単語アラインメントを考えた場合，*I(***私は***)*，*like(***好き***)*，*apples(***りんごが***)* となる．そして，アラインメントの精度が上昇することにより，翻訳精度が上昇すると考えられる．

しかし，アラインメントを取ることは容易なことではない．例えば，辞書を用いた場合，*book* という単語が**本**, **予約する**など複数の意味を持つように，多くの英単語は一義的に意味を決定することができない．これは，形態素解析の問題と関連している．

また，たとえ形態素解析を行ったとしても，日本語と英語のような表現形式が大きく異なる言語では，単語間の対応を取ることは容易ではない．例えば，*I was practicing tennis in the park* と**私は公園でテニスを練習していました**という 2 つの文の場合，英単語 *was* に対応する日本語は**た**であるが，これは**練習する**という語の助詞として扱われてしまっている．また，英単語 *in*, *the* に対応するような日本語は存在していない．したがって，本研究では，意味を持つ単位としてフレーズを用い，フレーズをベースとしたアラインメントを考える．すなわち，2 つの対訳文を *I/was practicing/tennis/in the park* と**私は/公園で/テニスを/練習していました**というようにフレーズに分割し，その後，対応付けを行う．

本研究では，隠れマルコフモデルと形態素解析を用いた手法により，英語と日本語のフレーズアラインメントを獲得する手法を提案する．その際，形態素解析と辞書を用いた発見的方法により，フレーズを生成し，フレーズ間の対応付け行う．そして，提案手法が効果的であることを実験で検証する．

第 2 章で，フレーズアラインメントのために用いた隠れマル

コフモデルの適用方法を示す．第 3 章では，形態素解析と辞書によるフレーズの対応付け方法を述べる．第 4 章では，実験結果を述べ，本手法の有効性を示す．

2. 隠れマルコフモデル

2.1 隠れマルコフモデル

隠れマルコフモデル (Hidden Markov Model，HMM) [7] とは，確率的な状態遷移と記号出力を備えたオートマトンである．次の 5 項組 $M = (X, Y, A, B, \pi)$ で定義される．

（1） 出力記号系列 $X = \{x_1, ..., x_n\}$ であり，観測可能である．

（2） 状態遷移系列 $Y = \{y_1, ..., y_n\}$ であり，観測不可能である．

（3） 状態遷移確率分布 $A = \{a_{ij}\}$ であり，状態 y_i から状態 y_j への遷移確率である．単純マルコフを仮定している．

（4） 記号出力確率分布 $B = \{b_i(x_t)\}$ であり，状態 y_i で記号 x_t を出力する確率である．出力は現在の状態にのみ依存すると仮定している．

（5） 初期状態確率分布 $\pi = \{\pi_i\}$ であり，状態 y_i が初期状態である確率である．

すなわち，図 1 のようなモデルが構成される．

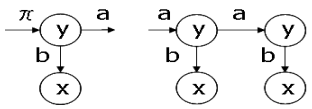

図 1 隠れマルコフモデル

本稿では，対訳コーパスからモデルを生成し，これを基に，最適な状態遷移系列の推定を行う教師あり機械学習の手法を用いることとする．

また，その推定には，Viterbi アルゴリズムを用いる.Viterbi アルゴリズムは以下のようなステップを持つ．

（1） 各状態 i=1,⋯,N に対して，変数の初期化を行う．

$$\delta_1(i) = \pi_i b_i(o_1)$$
$$\psi_1(i) = 0$$

（2） 各状態の遷移ステップ t=1,⋯,T-1, 各状態 j=1,⋯,N について，再帰的に計算を行う．

$$\delta_{t+1}(j) = max_i[\delta_t(i)a_{ij}]b_j(o_{t+1})$$
$$\psi_{t+1}(j) = argmax_i[\delta_t(i)a_{ij}]$$

（3） 再起計算の終了

$$\hat{P} = max\delta_T(i)$$
$$\hat{q}_T = argmax\delta_T(i)$$

（4） バックトラックによる最適状態遷移系列を復元する.t=T-1,...,1 に対して行う．

$$\hat{q}_t = \psi_{t+1}(\hat{q}_{t+1})$$

Viterbi アルゴリズムにより，確率値を最大にする状態遷移系列を得ることができる．

2.2 HMM フレーズアラインメント

この節では，フレーズアラインメントを HMM に対応させる方法を述べる．フレーズ化の方法については，次章で説明する．

英語文は，観測できない状態遷移系列として，日本語文を持っていると考える．例えば，*I was practicing tennis in the park* という文章の場合，その観測できない (隠れ) 状態列として，**私は公園でテニスを練習していました**を持っていると考えられる．そして，これらの文章をそれぞれフレーズに分割し，フレーズ間の対応を取ると，図 2 のようになる．

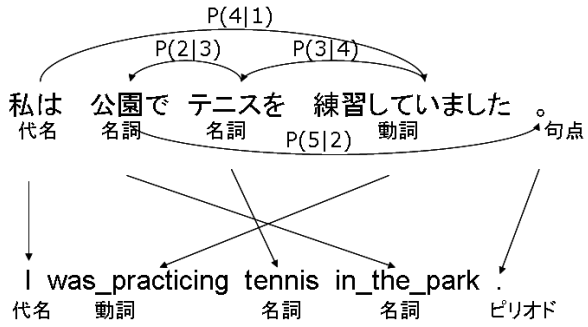

図 2 HMM の対応図

この場合，状態遷移として，**私は**，**練習していました**，**テニスを**，**公園で**の順に状態が遷移していったと考え，その各状態での記号出力として，**私は**/*I*，**練習していました**/*was practicing*，**テニスを**/*tennis*，**公園で**/*in the park* をそれぞれの状態が出力したと考える．

したがって，本稿で扱う品詞推定は，隠れマルコフモデルにおいて，状態遷移系列を日本語フレーズ，出力記号系列を英語フレーズとし，学習データから，(A,B,π) を以下で示した式のように学習することにより，モデルを生成する．ただし，$P(y|x)$ は条件付き確率であり，条件 x の下で y が生起する確率を示したものである．また $C(x)$ は生起頻度であり，語 x が当該文書で出現する頻度を表す．$BOS(Beginning\ Of\ Stream)$ は初期状態であり，文の先頭であることを表す．

- $\pi_i = P(y_i|BOS) = \dfrac{C(y_i)}{C(BOS)}$
- $a_{ij} = P(y_j|y_i) = \dfrac{C(y_i, y_j)}{C(y_i)}$
- $b_i(x_t) = P(x_t|y_i) = \dfrac{C(y_i, x_t)}{C(y_i)}$

3. HMM のためのフレージング

この章では，前章に述べた HMM フレーズアラインメントモデルを適用するために，対訳コーパスを，フレーズに分割し，対応付けを行う手法を述べる．ここで，HMM は観測列 (英語文) と隠れ状態列 (日本語文) のそれぞれの語の対応が明確でないと学習できないことに注意されたい．

3.1 フレーズ生成

まず，対訳コーパスをそれぞれ形態素解析する．英文形態素解析には，揚石ら [5] の手法を，日本語形態素解析には，MeCab [9] を用いる．

揚石らの手法では，HMM によって英文に品詞 (タグ) を付与した後，自動的に抽出したルールにより固有名詞の再推定を行っている．固有名詞を正しく認識できることは，翻訳にとって重要であると考えられる．例えば，*While House* が，**白い家**

か**ホワイトハウス**なのかでは大きく意味が異なる．したがって，本稿ではこの手法を採用する．

この手法により，*I was practicing tennis in the park* という文の形態素解析を行うと，*I_PPSS was_BEDZ practicing_VBG tennis_NN in_IN the_AT park_NN* となる．ここで，*PPSS* は一人称代名詞を，*BEDZ* は be 動詞過去形を，*VBG* は動名詞を，*IN* は前置詞を，*AT* は冠詞を，*NN* は名詞を意味している．

このようにして得られた形態素をもとに，発見的規則によりフレーズ化を行う．例えば，冠詞 (*the*) に続く名詞 (*park*) は，フレーズ (*the_park*) とし，その品詞を名詞とする，といった形である．

また，日本語形態素解析に用いた Mecab は，京都大学情報学研究科－日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクトを通じて開発されたオープンソース形態素解析エンジンであり，CRF によってパラメータの推定を行っている．

MeCab により，**私は公園でテニスを練習していました**という文の形態素解析を行うと，表 1 のような結果となる．これは，左から順に，表層形，読み，原型，品詞，細品詞 1，細品詞 2，となっている．

《私》《ワタシ》《私》《名詞》《一般》《代名詞》
《は》《ハ》《は》《助詞》《*》《係助詞》
《公園》《コウエン》《公園》《名詞》《*》《一般》
《で》《デ》《で》《助詞》《一般》《格助詞》
《テニス》《テニス》《テニス》《名詞》《*》《一般》
《を》《ヲ》《を》《助詞》《一般》《格助詞》
《練習》《レンシュウ》《練習》《名詞》《*》《サ変接続》
《し》《シ》《する》《動詞》《*》《自立》
《て》《テ》《て》《助詞》《*》《接続助詞》
《い》《イ》《いる》《動詞》《*》《非自立》
《まし》《マシ》《ます》《助動詞》《*》《*》
《た》《タ》《た》《助動詞》《*》《*》
《。》《。》《。》《記号》《*》《句点》

表 1　MeCab による形態素解析

この結果をもとに，英語文の場合と同様にして，形態素によるフレーズ化を行う．そして，これらの英語と日本語のフレーズ化により，図 2 のようなフレーズを獲得することができる．

3.2　フレーズの対応付け

次に，以下の得られたフレーズ間の対応付けを行う方法について説明する．

- *I*/**代名詞** *was_practicing*/**動詞** *tennis*/**名詞** *in_the_park*/**名詞**
- **私は/代名詞 公園で/名詞 テニスを/名詞 練習していました/動詞**

まず，品詞によってユニークに決定できる場合，それらの語を対応付ける．すなわち，*I* と**私は**が，*was_practicing* と**練習していました**が対応していると考える．

次に，品詞によって決定することができなかった残りの候補を辞書を引くことによって対応付けをする．ここで，辞書には英辞郎 [8] を用いた．英辞郎の構成は表 2 のようになっている．

park 他動-1：～を駐車｛ちゅうしゃ｝させる、駐車｛ちゅうしゃ｝する
park 他動-2：～を置いておく
park 名-1：公園｛こうえん｝■・Karl took a walk in the park after lunch. カールは昼食の後、公園を散歩した。
park 名-2：遊園地｛ゆうえんち｝
park 名-3：大庭園｛だい ていえん｝
park 名-4：競技場｛きょうぎじょう｝
park 名-5：駐車場｛ちゅうしゃじょう｝
park 名-6：〈米話〉野球場｛やきゅう じょう｝
park 自動-1：駐車｛ちゅうしゃ｝する
park 自動-2：《野球》（ホームランを）打つ
park：【レベル】1、【発音】p ɑ':(r)k、【@】パーク、【変化】《動》parks — parking — parked
park a bicycle：（自転車｛じてんしゃ｝を）駐輪｛ちゅうりん｝する
park a car in a garage：駐車場｛ちゅうしゃじょう｝に車を止める
park a car in a narrow space：狭い場所｛ばしょ｝に駐車｛ちゅうしゃ｝する

表 2　英辞郎の例

ここで，{}は品詞を，その前は単語を，:より後が翻訳語を表している．また，■より後ろは用例を表している．

本研究では，:と■の間の語に着目し，品詞が一致し，かつ翻訳語が一致するフレーズを対応付けを行う．そして，最後にもう一度残った候補の中から，品詞によってユニークに決定できるフレーズを対応付ける．

これにより，対応付けされたフレーズ，**私は**/*I*，**練習していました**/*was practicing*，**テニスを**/*tennis*，**公園で**/*in the park*，を得ることができる．

4.　実　　験

4.1　準　　備

本研究では，対訳コーパスとして読売新聞と The Daily Yomiuri の対訳文 [3] を用いる．これは，内山らによって作成された対訳コーパスで，1989 年から 2001 年までの読売新聞と The Daily Yomiuri からなっており、文対応ペアは 1 対 1 対応が約 15 万ペアある．このうち，無作為に選択した 2,000 行をテストデータとし，残りを学習データに当てる．対訳コーパスの例を表 3 に挙げる．

<J> 欧州は、エディンバラにおいて合意され、コペンハーゲンにおいて強化された成長イニシアチブを精力的に実行しつつある。</J>
<E>Europe is carrying out vigorously the Growth Initiative agreed in Edinburgh and strengthened in Copenhagen. </E>
<J> 我々は、ロシアの経済発展にとって、改善された市場アクセスが重要であることを認識する。/<J>
<E>We recognize the importance of improved market access for economic progress in Russia./<E>

表 3　対訳コーパス:読売新聞–The Daily Yomiuri

4.2　評 価 方 法

Word Error Rate (WER) [4] は語順を考慮した不一致率であり，正解との編集距離によって計算される．

$$WER = \frac{\sum_i (\text{挿入語数 } i + \text{削除語数 } i + \text{置換語数 } i)}{\sum_i \text{参照訳 } i \text{ の語数}}$$

また，Position independent word Error Rate(PER) は語順を無視し、文を単語集合とした不一致率である．これは，WER では翻訳が正しくても，正解と語順が著しく異なる場合に結果が悪くなってしまうため，語順を考慮せず，語のレベルでその翻訳が正しいかを判断するための基準である．

$$PER = 1 - \frac{\sum_i (\text{翻訳文 } i \text{・参照訳 } i \text{ 間の一致語数})}{\sum_i \text{参照訳 } i \text{ の語数}}$$

どちらも，誤り率であるので，数値が低いほど良い結果であるといえる．本研究では，語を並び替える操作は考えていないため，PER の結果が良いことが望ましい．

4.3 実験結果

テストデータ 2,000 行に対し，Viterbi アルゴリズムを用いて，日本語の推定を行い，WER，PER の値を測定した．実験結果を表 4 に示す．

	WER	PER
読売新聞	1.13	0.64

表 4　実験結果

PER の値が 0.74 であるので，およそ 4 語に 1 語の割合で正しい日本語を推定できていると考えられる．また，本実験では，並び替えは考えていないため，WER の値が 1 を超えてしまっているこれは，翻訳語と正解語が違うことに加えて，翻訳語が，正解語よりも多いため，削除を多くしてしまっているためである．

4.4 考　　察

PER の値が，低い理由として，学習不足が考えられる．実際に学習データとして 148,000 行与えているが，対応付けを自動的に行い，学習に使用することができたデータ数は，1315 行であった．これにより，多くの未知語 (学習データに出現しない語) が増えてしまったため，精度が悪化していると考えられる．学習がうまくいく例としては，*We/PPSS recognize/VB the-importance/NN of-improved/VBN market-access/NN for-economic-progress/NN in-Russia/NP ./.* と **我々は/代名詞 ロシアの/固有名詞 経済発展にとって/名詞 改善された/動詞 市場アクセス-が/名詞 重要であることを/名詞 認識する/動詞 。/句点**の対訳文が挙げられる．この例のように，英語文と日本語文のそれぞれのフレーズが，順序が異なるだけであれば学習はうまくいく．この 2 つの文の対応付けを行うと，*We/***我々は** *recognize/***認識する** *the-importance/***重要であることを** *of-improved/***改善された** *market-access/***市場アクセスが** *for-economic-progress/***経済発展にとって** *in-Russia/***ロシアの** *./*。となる．しかしながら，*International-terrorism/NP is/BEZ a-grave-threat/NN to-world-peace/NN and/CC security/NN ./.* と**国際テロは/名詞 世界の/名詞 平和と/名詞 安全に対する/名詞 重大な脅威-だ/名詞 。/句点**，のような場合は対応付けを行うことができない．これはまず，英語文の動詞 *is* に対応する日本語が存在していない．このように英語と日本語の文法構造が異なる文では，うまく学習することができない．また，英語の *to-world-peace/NN and/CC security/NN* と日本語の**世界の/名詞 平和と/名詞 安全に対する/名詞**のフレーズ構造が異なってしまっている．これは，英文では，*to-world-peace/NN(***世界平和***)* と security/NN(安全) と解釈したが，日本語文では，**世界の** *(***平和と安全***)* と解釈してしまっているためである．このようにフレーズ化した際の違いによっても，学習が困難となっている．

さらに，テストデータを調査してみたところ，テスト語 32606 語のうち，19471 語が未知語と判定されていた．未知語を除いて PER の値を計算すると，PER=0.41 であったので，未知語を減らすことによって，精度が上昇すると考えられる．したがって，発見的ルールを見直し，より多くの学習データを使用することができれば，それに伴い，精度を上げられる可能性があると考えられる．

5. 結　　論

本研究では，英語と日本語のフレーズアラインメントを生成する方法として，形態素解析を隠れマルコフモデルを用いた方法を提案した．実際に実験を行い，その結果から提案手法がフレーズアラインメントを生成する方法として期待できることを確認した．

文　　献

[1] Adam Lopez, Statistical Machine Translation, ACM Computing Surveys, Vol. 40, No. 3, Article 8, Publication date: August 2008.

[2] Bernard Merialdo: Tagging English Text with a Probabilistic Model. *Association for Computational Linguistics* pp. 155-171, 1994

[3] Masao Utiyama and Hitoshi Isahara. Reliable Measures for Aligning Japanese-English News Articles and Sentences. ACL-2003, pp. 72–79.

[4] OCH, F. J., TILLMAN, C., AND NEY, H. 1999. Improved alignment models for statistical machine translation. In Proceedings of EMNLP-VLC. 20.28.

[5] 揚石 亮平, 三浦 孝夫. 2008. Named Entity Recognition Based On A Hidden Markov Model in Part-Of-Speech Tagging, ICADIWT 2008.

[6] 浅原 正幸:系列ラベリング問題に関するメモ，奈良先端科学技術大学院大学, 2006

[7] 北 研二:確率的言語モデル，東京大学出版会, 1999

[8] 英辞郎:100 万語収録付録 CD-ROM, 株式会社アルク

[9] MeCab: http://Mecab.sourceforge.net/