# ハイブリッド翻訳のためのフレーズアラインメント

潮田明

（株）富士通研究所

*ushioda@jp.fujitsu.com*

## 1. はじめに

日英・英日機械翻訳のように言語構造の大きく異なる２言語間の自動翻訳においては、従来より構文解析を含むある程度深い言語的解析が重要だと考えられ、実用システムは主にルールベース翻訳を中心に開発が行われて来た。一方仏英間などヨーロッパ言語間での適用から始まった統計翻訳(Statistical Machine Translation,以下 SMT)は、近年著しい発展をとげ、最近ではアラビア語英語間や中英間などの翻訳においても実用化を目指した開発が進められている。SMT では現在フレーズベースの統計翻訳[1]が最も有望視されているが、フレーズベースの統計翻訳における「フレーズ」とは一般に言語学的フレーズ、あるいは constituent とは無縁の場合が多い。そのため、翻訳の単位として汎用性に欠ける、対訳コーパスの分野に過適合するため対訳フレーズの分野間移植性が低い、などの本質的問題を抱えている。

本発表では、日英翻訳を念頭に、従来のルールベース翻訳および用例翻訳[2]と、フレーズベースの統計翻訳とを融合するハイブリッド翻訳について展望し、融合に際してのカギとなるハイブリッド翻訳のためのフレーズアラインメント方式を提案する。本方式では、統計情報と辞書情報を組み合わせながらボトムアップにバイリンガルパージングを進めることにより統計的最適化の枠組みの中に言語学的制約を組み込むことが可能となる。

## 2. ハイブリッド翻訳

ルールベース翻訳は汎化が比較的容易で、ルールによる細かい細工が可能であるなどの長所がある一方、辞書や文法には大量の人手コストが必要な上、ユーザによる調整が難しいなどの問題を抱えている。更には、最近多くの言語間で整備が進みつつある大量の対訳コーパスから、直接自動で（そのまま既存の翻訳エンジンに組み込める形で）文法規則を抽出する手立てが現時点では得られていない。用例翻訳はコーパスから自動で翻訳知識が構築できる点で優れているが、不特定分野の翻訳においては、大量の用例収集が必要であり、対訳コーパスを効率的に活用するためには、用例の汎用化を自動で行うメカニズムの開発が不可欠である。統計翻訳は対訳データさえあれば人手による辞書作成やルール作成が不要というメリットがある反面、学習用対訳データとは全く違う分野や文種の翻訳は苦手であるという欠点もあり、汎用的な翻訳システムとして見たときに従来手法に比べて実際どのくらいの翻訳品質が単独で得られるかはまだ未知な部分が多い。

そこで、少なくとも日英・英日翻訳においては、従来より商用システムとしても幅広く使われてきたルールベース翻訳および用例翻訳と、大量の対訳コーパスから翻訳に有用な情報を定量的に抽出できる統計翻訳の長所を組合わせたハイブリッド翻訳が次世代自動翻訳の有力候補として期待できる。しかし、単にハイブリッドといっても、組合わせ方には様々な形態が考えられる。最も疎な組合わせ方としては、それぞれの翻訳結果からベストと思われる結果を抽出する方法、すなわち投票方式がある。またそれぞれの翻訳システムの中間結果を相互利用する方法も考えられる。たとえば、フレーズベース統計翻訳において、抽出されたフレーズテーブルの中から、パーサの出力と一致するフレーズ（すなわち constituent）のみを選択、あるいは一致するフレーズを優先して使う方式などが提案され、その有効性ついても報告されている[3]。更に踏み込んだ融合型のハイブリッド方式では、統計情報と言語解析情報を結合しながら翻訳を進めて行く方法などが考えられる。

いずれのアプローチにせよ、従来のルールベース翻訳や用例翻訳の資産を最大限に活用しようと考えたときに必ず問題となる重要なポイントは、ルールベース翻訳におけるフレーズとフレーズベース統計翻訳におけるフレーズの間に整合性が全くないことである。前者は構文木における constituent をフレーズの単位として解析を進めるのに対して、後者のフレーズは、言語学的意味とは無縁に、統計的にある意味で有意な単語の連なりをフレーズとして活用している。SMT の枠組みの中で constituent を用いることの是非についてはここでは深くは論じないが、少なくとも SMT の最大の問題の１つである異分野間移植性、すなわち、ある分野の対訳コーパスから学習した SMT を全く異なる分野での翻訳へ適用しようとしたときの適用可能性、を考えた場合、人間の言語知識を基に築かれた汎化性に裏打ちされた constituent の概念が重要な役割を果たすことは十分考えられる。ましてや、ハイブリッド方式の中で従来のルールベース翻訳の開発過程で築かれた文法規則あるいは文法記述の枠組みや、言い換え可能性を基準に構築された用例翻訳用の用例の蓄積を有効に活用しようと考えたときには、効率よく constituent あるいはそれに準拠した対訳フレーズと統計翻訳の枠組みとを組み合わせる手立てを考える必要がある。本研究では、フレーズベース SMT の枠組みを基盤として、その中に syntax ベースの要素をいかに取り入れていくかという観点から,フレーズアラインメントの手法を検討する。

## 3. フレーズアラインメント方法と実験

constituent を基準にしたフレーズアラインメントを抽出すると言っても、実際の対訳コーパスの中で意味的に過不足なく対応の付く部分（チャンク）を括り出して行ったときに、結果的に得られたチャンクが実際 constituent になっているかと言うと、その補償はない。実際には一方の言語において constituent であっても、対応するもう一方の言語側のチャンクは constituent でない、ということは多々ある。従って両言語側とも constituent であるという制約は強すぎる可能性はある。またもちろんそもそも何が constituent であるかも、もとになる文法のルールに依存するため、両言語側の文法の相性と言ったものも関係してくる。

そこで今回は、日本語側のフレーズが constituent になるべく近づくことを優先するアプローチを選択した。実験には、GIZA++[4]、日英・英日機械翻訳ソフト[5]用の日本語パーサ、英語パーサ、日英対訳辞書（８６万訳対）の各種リソースを用いた。対訳コーパスは第３回 NTCIR ワークショップ[6]特許検索タスクのテストコレクションの中から１５万文対の日英対訳特許抄録の課題文を抽出して用いた。

提案手法ではまず対訳コーパスに対して GIZA++ により日英・英日双方向の単語アラインメント A1, A2 を同定した後インターセクション $A = A1 \cap A2$ を求める。次に A の中での英語単語 e の出現頻度 $N(e)$, および e と j が対応付けられた頻度 $N(e,j)$、さらに対訳辞書中で e と j が互いに訳語として登録されているか否かを示す判別関数 $\delta(e,j)$ をもとに単語ベースの翻訳確率 $Pc(j|e)$ を以下のように求める。

$$Pc(j|e)=(N(e,j)+\delta(e,j))/(N(e)+\Sigma_t \delta(e,t))$$

ここで $\delta(e,j)$ は辞書に登録されていれば１、いなければ０を取る関数である。

図１に本手法に基づくシステムの構成図を示す。提案手法では、統計情報と辞書情報および文法規則を組み合わせながらボトムアップに隣接する２つの単語あるいは単語列の接合（マージ）を進めることによりフレーズアラインメントを行う。まず簡単のために、言語学的制約のない場合のマージングの進め方について説明する。この場合、図１の「統計評価値」は「総合評価値」に等しい。

まず入力対訳文$(\mathbf{J}, \mathbf{E})$の日本語側の文 $\mathbf{J}$ は、shallow parser によってベースフレーズ(minimum phrase)の抽出が施され、$\mathbf{J} = \mathbf{j}_1, \mathbf{j}_2, \cdots, \mathbf{j}_M$ なるチャンクの列に分割される。これらのチャンクはそれ自体が constituent をなすが、最終的に得られるフレーズアラインメントのフレーズの最小単位でもある。英語文 $\mathbf{E}$ は N 個の単語よりなるとして、$\mathbf{E} = w_1, w_2, \cdots, w_N$ と表すことにする。

ここで、対訳コーパスと対訳辞書から得られた単語の翻訳確率 $Pc(j|w_i)$ を用いて、単語 $w_i$ の訳語がチャンク $\mathbf{j}_j$ 中に含まれる確率を求めることを考える。まず、$w_i$ の訳語が必ず英語文中に出現するという仮定より、

$$\Sigma_j \Sigma_t P(t|w_i) P(t \text{ appears in } \mathbf{j}_j) = 1 \quad ①$$

ここで、t は $w_i$ の翻訳候補、$P(t|w_i)$ は与えられた対訳文内において $w_i$ が t に翻訳される翻訳確率、$P(t \text{ appears in } \mathbf{j}_j)$ は t が $\mathbf{j}_j$ 中に出現する確率を表す。ここでは、対訳文は既知であるから、$\mathbf{j}_j$ が文字列として t を含まなければ $P(t \text{ appears in } \mathbf{j}_j)$ は 0 であり、含めば１であると認定できる。正確には $\mathbf{j}_j$ が文字列として t を含んでいても t が $w_i$ の訳語として存在しているのではない可能性もあるが、ここでは上記の通り認定するものとする。また、与えられた対訳文内での翻訳確率 $P(t|w_i)$ は対訳コーパスおよび辞書から文脈に依存しない形で得られる上記翻訳確率 $Pc(t|w_i)$ に比例すると仮定する。すなわち、すべての t と $w_i$ について

$$P(t|w_i)=\lambda Pc(t|w_i) \quad ②$$

ここでλはある定数。

①と②より、

$$\lambda \Sigma_j \Sigma_t Pc(t|w_i) P(t \text{ appears in } \mathbf{j}_j) = 1$$

ここで

$$C_{ij} = \Sigma_t Pc(t|w_i) P(t \text{ appears in } \mathbf{j}_j)$$

と定義し、C を対訳フレーズマトリックスと呼ぶことにする。すると、

$$\lambda = 1 / \Sigma_j C_{ij}$$

$$P(t|w_i) = Pc(t|w_i) / \Sigma_j C_{ij}$$

となる。

これより、単語 $w_i$ の訳語がフレーズ $\mathbf{j}_j$ 中に出現する確率 $P(\mathbf{j}_j|w_i)$ は、

$$P(\mathbf{j}_j|w_i) = \Sigma_t P(t|w_i) P(t \text{ appears in } \mathbf{j}_j)$$
$$= \Sigma_t P(t \text{ appears in } \mathbf{j}_j) * Pc(t|w_i) / \Sigma_j C_{ij}$$
$$= C_{ij} / \Sigma_j C_{ij} \quad ③$$

と求まる。従って、単語 $w_i$ の訳語がフレーズ $\mathbf{j}_j$ 中に出現する確率 $P(\mathbf{j}_j|w_i)$ は、$C_{ij}$ の値を行マージンで割った値（行内相対値）として求まる。また $C_{ij}$ の値は定義より Pc から直接求めることができる。

同様に、日本語側フレーズ $\mathbf{j}_j$ について、$\mathbf{j}_j$ の訳語が $w_i$ として表された確率 $P(w_i|\mathbf{j}_j)$ を C の列内相対値として以下のよう求めることが可能である。もちろん他の指標をもとにした求め方も可能であるが、本実験では以下のように仮定する。

$$P(w_i|\mathbf{j}_j) = C_{ij} / \Sigma_i C_{ij} \quad ④$$

さてここで、対訳フレーズマトリックスの i 行に着目し、$w_i$ の訳語（翻訳表現）が日本語文のどのチャンクに出現したかを判定する際の確からしさを考える。もし $P(\mathbf{j}_j|w_i) = 1$ ならば、単語 $w_i$ の訳語がチャンク $\mathbf{j}_j$ 中に出現したことは１００％確かであり、情報論的に言えば、判定のエントロピー

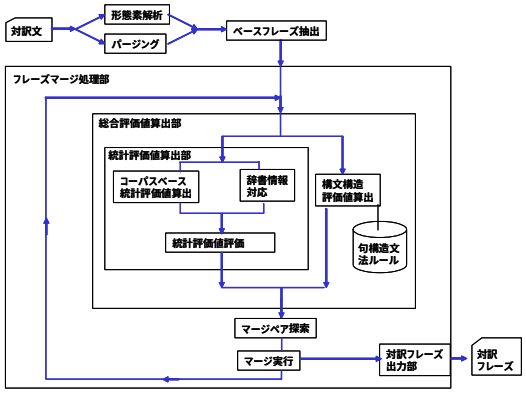


図 1. フレーズアラインメントの構成図

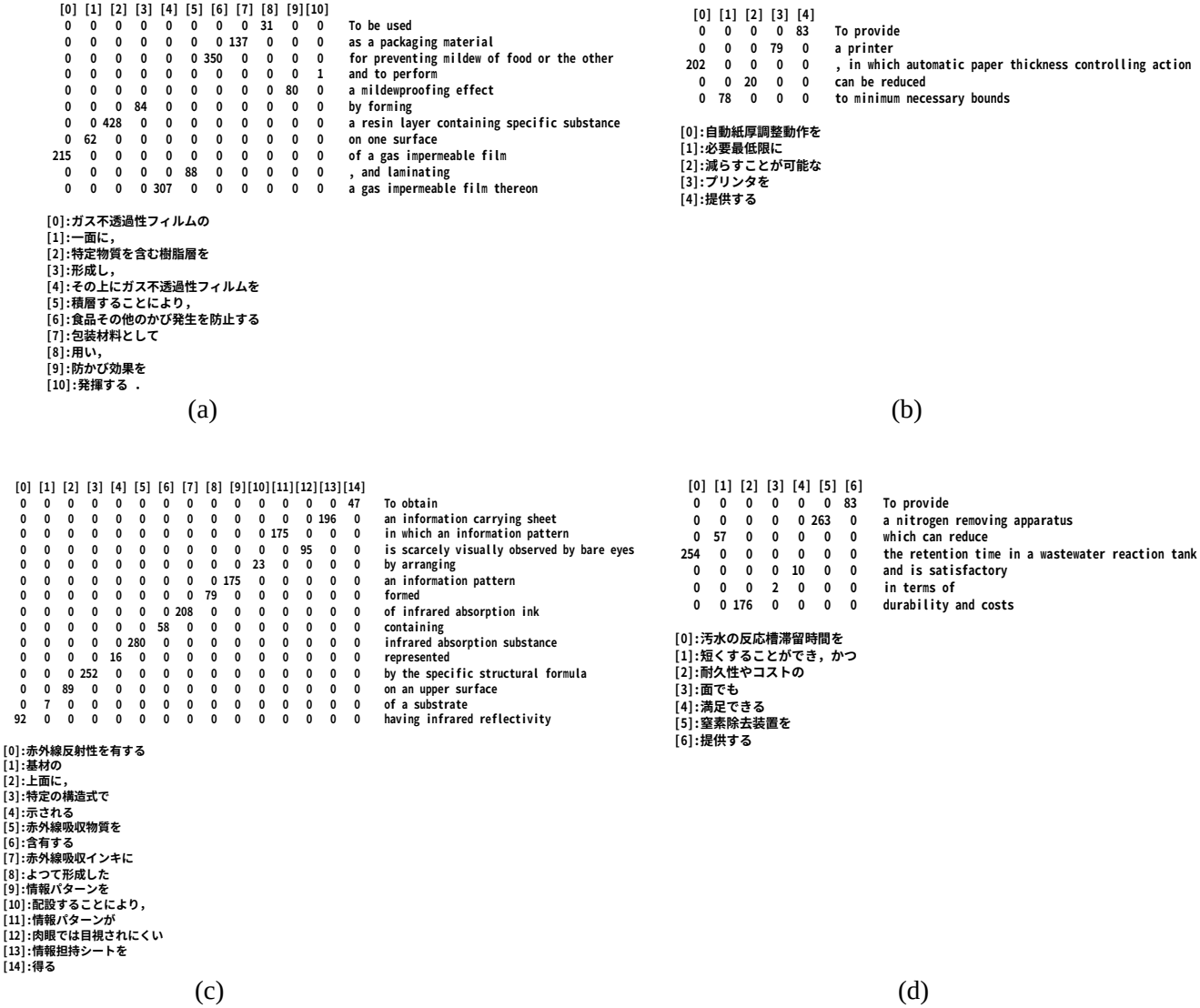


図 2. 獲得されたフレーズアラインメントの例

は0である。ここで、エントロピー $H(i)$ は

$$H(i) = -\Sigma_j P(\mathbf{j}_j \mid w_i) \log_2 P(\mathbf{j}_j \mid w_i) \quad ⑤$$

で与えられる。但し $\lim_{X \to 0} X \log_2 X = 0$ の関係を利用して、すべての $j$ に対して $P(\mathbf{j}_j \mid w_i) = 0$ のケースでは $H(i) = 0$ と解釈することにする。

本提案手法では、以下に示すエントロピーを基準にした総合評価値を用いて、対訳フレーズマトリックスの隣接する行同士あるいは隣接する列同士を順次マージして行き、評価値が極小値を取ったところでマージを終了し、対訳フレーズマトリックスの正値の各要素 $C_{ij}$ に対して $i$ 行目の英語単語列と $j$ 列目の日本語単語列の対を対訳フレーズとして抽出する。マージによる評価値の変化量の算出方法は、決定木において、サンプル属性の値に応じてサンプル集合を分割して行く際のエントロピー変化量（あるいは情報量のゲイン）の算出方法と等しい。ここでは、各行を決定木における分割されたサンプル集合に見立て、サンプル集合の大きさは対訳フレーズマトリックスの行マージン( $\Sigma_j C_{ij}$ )で表されると仮定する。但し決定木では順次データ集合を部分集合に分割して行くのに対して本手法では逆にフレーズ同士をマージして行く。

上述の考え方に従って、各英語フレーズを主体に見たときに、それぞれが各日本語フレーズに対応付けられる不確実性を表す指標（エントロピー）を以下のように定義する。

$$H = \Sigma_i [\Sigma_j C_{ij}] H(i) / \Sigma_i \Sigma_j C_{ij}$$

同様に、各日本語フレーズを主体に見たときの指標を

$$H_t = \Sigma_j [\Sigma_i C_{ij}] H(j) / \Sigma_i \Sigma_j C_{ij}$$

として、総合評価値を両者の平均値として定義する。

$$H_{tot} = (H + H_t)/2$$

なおここで「評価値」という表現を用いているが、ここでは評価値は小さい方が優位である。各マージングステップにおいて、行をマージするか列をマージするかは、総合評価値によって判断される。またマージングは greedy に行うと local minimum に陥るため、beam search により絶えず複数の候補を保ちながら探索を行う。今回の実験ではビーム幅は 300 から 1000 の値を用いた。また今回は日本語文の各フレーズと英語文の各フレーズが１対１に対応するケースのみを扱うため、総合評価値ゼロの地点をマージの終点と見なす。１対１に対応しないケースには、日本語単語の同一訳語が英文中の離れた場所に当該日本語単語の訳として２回以上出現する場合などがある。$Pc(t|w_i)$ の値が２つ以上の異なる t に対して正値の場合は、そのまま使うことも可能であるが、その場合総合評価値はゼロには到達しないケースが多いため、今回は $w_i$ はどれか１つの列にのみ対応すると仮定してそれぞれ場合分けを行い探索を行った。$w_1, w_2, \cdots w_n$ のｎ個の単語がそれぞれ $k_1, k_2, \cdots, k_n$ 個の対応候補を持つ場合全体の組合わせの数は $k_1 k_2 \cdots k_n$ となる。

以上は言語学的制約のない場合のマージングの進め方であるが、実際には各マージ判定の過程で、syntax 情報が制約あるいは選好として導入される。たとえばマージ候補としてトップ２組のペア(i-行, i+1-行), (k-列, k+1-列)がありそれぞれのマージの総合評価値が H1, H2 とした場合、言語学的制約がない場合は単に評価値の小さい方のマージペアを選択するが、言語情報を制約として導入した場合は、制約条件を満たしたペアのみが選択される。選好として導入する場合は、評価値に選好の度合いを係数として掛け合わせて評価する。constituent をなすフレーズを単位としたアラインメントのみを求める場合には、マージしようとしているペアが constituent の境界をまたぐ表現かどうかを判断基準に制約をかけることになる。今回の実験に用いた制約および選好の例を以下に示す。

[日本語マージの選好]
評価値が等しい場合は constituent をなすマージ候補を優先する。

[日本語マージの制約]
接続詞、句読点は前方と接続する。

[英語マージの制約]
ベースフレーズの境界をまたがない。接続詞、前置詞、句読点は後方と接合する。

本手法により得られたフレーズアラインメントの例を図２．に示す。図中のマトリックス中の数字は対訳フレーズマトリックスの要素の値であるが、見易くするために 100 倍してある。まだ定量評価は行っていないが、日本語側のフレーズは概ね constituent と見なせるものが獲得できている。

## 4. 関連研究

従来のフレーズベース統計翻訳の代表的なものは Och らの提案した手法[1]である。この手法では両言語方向のアラインメントのインターセクションを出発点として、片方のアラインメントにしか現れない単語対応の中で一定の条件を満たす「有望な」対応を順次追加して行く。このようにして得られたフレーズは、一般に言語学的には必ずしも意味を持たない単なる単語列となる。

フレーズベース統計翻訳に（言語学的）構文解析処理を用いる手法もこれまでに提案されているが[7]、従来の手法においては構文解析処理は,対訳コーパス間の単語対応や単語翻訳確率などとは独立に、単言語コーパスに対して既存の構文解析器を用いて行われていたため、パーサが誤認識をした場合、エラーを修復する手立てがなかった。

対訳コーパスから自動的に Synchronous CFG を学習し、フレーズテーブルの汎用化を図る試み[8]も行われているが、ここで言う文法規則は言語学的なものではなく、また Synchronous CFG 生成の出発点となるフレーズテーブルは従来のフレーズベース SMT の手法によるものである。

## 5. まとめ

統計情報と辞書情報を組み合ながらボトムアップにバイリンガルパージングを進めることにより、統計的最適化の枠組みの中に言語学的制約を組み込んだ新しいフレーズアラインメント方式について提案した。本手法では、統計をベースにしたフレーズアラインメントと部分的構文解析処理(syntax による制約と選好)とを同時に行うことにより両者の欠点を補う効果が期待できる。今回は syntax 情報はマージの条件に一方的に用いたが、マージの過程で得られた信頼度の高いフレーズの情報をパーサにフィードバックすることによりパーサの精度を高めたり、また相互に信頼度の高い情報を交換しながら連携してマージとパージングを進めて行く方法も考えられる。

## 参考文献


[1]Franz-Josef Och and Hermann Ney (2004) “The alignment template approach to statistical machine translation.” *Computational Linguistics*, 30(4),pp.417-450.

[2] Makoto Nagao (1984) “A Framework of a Mechanical Translation between Japanese and English by Analogy Principle.” In Alick Elithorn & Ranan Banerji, eds., Artificial and Human Intelligence, Elsevier Science Publishers, chap. 11, pp. 173-180.

[3]Andreas Zollmann, et. al.(2006) “The CMU-UKA augumented machine translation system for IWSLT-06.” *Proceedings of International Workshop on Spoken Language Translation*, pp138-144.

[4]Franz-Josef Och and Hermann Ney (2000) “Improved statistical alignment models.” *Proceedings of the 38th Annual Meeting of the ACL*, pp.440-447.

[5]富士通.英日・日英翻訳ソフト ATLAS. http://software.fujitsu.com/jp/atlas/.

[6] http://research.nii.ac.jp/ntcir/ntcir-ws3/ws-ja.html

[7]Kenji Yamada and Kevin Knight (2001) “A syntax-based statistical translation model.” *Proceedings of the 39th Annual Meeting of the ACL*, pp.523-530

[8] David Chiang (2005) “A hierarchical phrase-based model for statistical machine translation.” *Proceedings of the 43rd Annual Meeting of the ACL*, pp263-270.