句の依存関係確率モデルを用いた統計的対訳文アライメント

中澤 敏明　　黒橋 禎夫
京都大学大学院情報学研究科
nakazawa@nlp.kuee.kyoto-u.ac.jp　kuro@i.kyoto-u.ac.jp

1　はじめに

既存の統計的対訳文アライメント手法の多くは、文を単純な単語列として見た単語アライメント ([1, 4]) である。英語とフランス語のように語順変化が局所的であり、大きな構造の違いがない言語対においてはこのような手法でも十分な精度を達成できるが、日本語と英語などのように構造が大きく異なる言語対に対しては脆弱であり、良い精度は得られない。単純な単語アライメント手法による結果の上に構文情報を追加する手法 ([7, 3]) もいくつか提案されているが、ベースとなるアライメントそのものの精度が高くない状況では大きな精度の向上を達成することは難しい。

そこで本稿では、初めから言語構造を利用した対訳文アライメントモデルを提案する。本モデルは依存構造解析された二つの木構造において、句の対応のペアの原言語側と目的言語側での依存関係をモデル化したものであり、単語列アライメントで扱うのが困難な距離の大きな語順変化にも対応することができる。言い替えれば、本モデルは木構造上での reordering モデルである。

もちろん木構造を利用したアライメントモデルはすでにいくつか提案されている。[9] は構文解析結果をアライメントで利用しているが、木構造に対して挿入、削除、複製などの操作をする必要があり、モデルが複雑であるし、ロバストではない。[2] は提案手法と同様に依存構造解析結果を利用しているが、単語単位の手法であり、句などの大きな単位が扱えない上に、1 対 1 対応のみという制約があり、やはりロバスト性に欠ける。提案モデルでは多対多の対応を生成することができるし、アライメントの単位は単語よりもう少し大きい、意味的なまとまりのある単位である (2.1 章)。

我々が以前提案した手法 [5] では内容語は内容語のみに、付属語は付属語のみにしか対応しないという制約があり、これがしばしば誤った対応を生成していた。本手法ではこの制約を取り除き、より柔軟なアライメントを可能とした。また以前のモデルでは各方向でのアライメント結果を得た後に、それらをヒューリスティックな方法で対称化することにより多対多の対応を獲得しているが、本モデルでは推定時に双方向の確率を同時に用いており、多対多の対応も推定時に自然と学習される。

2　依存関係確率モデル

2.1　アライメントの単位

以後の説明では原言語に日本語、目的言語に英語を仮定する。ただし、提案手法は言語対によらない汎用的なものであることに注意されたい。

まず対訳文を両言語とも依存構造解析する。日本語文に対しては JUMAN および KNP を用い、英語文に対しては Charniak のパーサを用いて句構造に変換し、これにフレーズの head を定義するルールを適用することにより依存構造に変換する。これらの処理により、文は 1 つの内容語と 0 個以上の機能語とをひとまとまりとした句をノードとする依存構造木に変換される。

次に機能語を、内容語同士をつなぐものと考える。つまり機能語は、内容語とその係り先の内容語との間に位置しており、それらをつなぐ枝として働くと考える。なお機能語がない句については、その句の係り受け情報を機能語の代わりとする。係り受け情報としては各パーサによって定義される “係り受けタイプ” と “係る方向” を利用する。係り受けタイプは日本語では “文節内” や “連用”、“連体” などが、英語では “NP”、“VP”、“PP” などがある。係る方向はその句が係り先の前から係るか、後ろから係るかの情報である。これは英語で同士の前から係る名詞は主語 (日本語での “ガ格”) である可能性が高く、後ろから係る名詞は目的語 (“ヲ格”) である可能性が高いといった現象を扱うのに有効である。

図 1 に依存構造木の例を示す。提案モデルではこれら内容語ノード、付属語ノードそれぞれをアライメントの最小単位として扱う。なお、モデル推定時には連

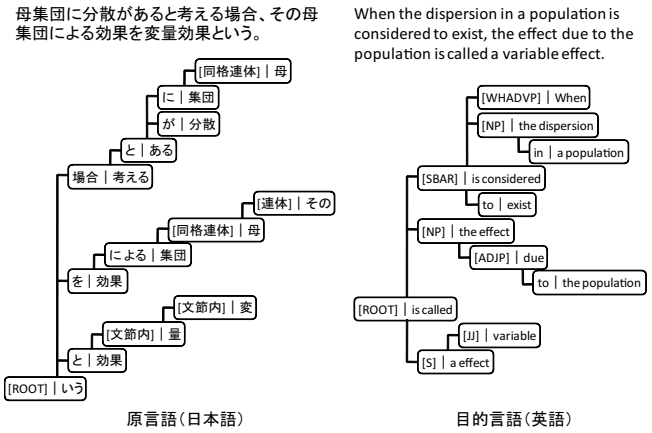

図 1: 依存構造木の例

続する複数のノードのかたまりもアライメントの単位として考慮することにより、多対多対応を実現する (3.2.3 章参照)。

2.2 モデルの定義

提案するモデルでは、IBM Model[1] のような方向性のある確率モデルを双方向同時に利用する。そこでまず方向性のある確率モデルを定義する。

IBM Model では与えられた原言語文 f と目的言語文 e に対して最も良いアラインメント $\hat{a}$ を

$$\hat{a} = \mathrm{argmax}\, p(f|e,a) \cdot p(a|e) \tag{1}$$

により求める。ここで $p(f|e,a)$ は lexicon probability、$p(a|e)$ は alignment probability と呼ばれている。

提案モデルでは木構造を利用しているため、以下のような確率を考える。与えられた原言語の依存構造木 T_f と目的言語の依存構造木 T_e に対して、IBM Model と同様に以下のように書ける。

$$\hat{a} = \mathrm{argmax}\, p(T_f|T_e,a) \cdot p(a|T_e) \tag{2}$$

さらに T_f が J 個のノード $f_1, ..., f_J$ から、T_e が I 個のノード $e_1, ..., e_J$ から構成されているとし、各ノードの係り先 (親) のノードを、添字にダッシュを付けることにより表し (例えばノード f_j の係り先のノードは $f_{j'}$)、それぞれの確率を以下のように定義する。

$$p(T_f|T_e,a) \approx \prod_{j=1}^{J} p(f_j|e_{a_j}) \tag{3}$$

$$p(a|T_e) \approx \prod_{j=1}^{J} p(rel(e_{a_j} \leftarrow e_{a_{j'}})|rel(f_j \leftarrow f_{j'})) \tag{4}$$

a_j は f_j の対応先のノード番号であり、e_{a_j} は対応ノードを表す。

表 1: カテゴリ確率の有無による精度比較

	適合率	再現率	F 値
カテゴリ確率なし	58.25	42.77	49.33
カテゴリ確率あり	**76.52**	**50.23**	**60.65**

$p(f_j|e_{a_j})$ 単語翻訳確率[1]であり、IBM Model 1 と同様の定義である。$rel(f_j \leftarrow f_{j'})$ は $f_{j'}$ からみた f_j の依存関係であり、$rel(e_{a_j} \leftarrow e_{a_{j'}})$ も同様である。$p(rel(e_{a_j} \leftarrow e_{a_{j'}})|rel(f_j \leftarrow f_{j'}))$ は依存関係確率であり、原言語側で直接の親子関係になっている二つのノード (f_j と $f_{j'}$) の対応する目的言語側のノード (e_{a_j} と $e_{a_{j'}}$) の依存関係がどうなっているかを表す確率である。

提案モデルでは対訳文のアライメント確率 $p(a|f,e)$ を lexicon probability と alignment probability を双方向分すべて掛け合わせたものとして以下のように定義し、この確率が最大となるアライメントを最も良いアライメントとする。

$$p(T_f|T_e,a) \cdot p(a|T_e) \cdot p(T_e|T_f,a) \cdot p(a|T_f) \tag{5}$$

3 トレーニング

モデルのトレーニングではまず単語翻訳確率のみを推定 (Model 1) し、これを初期値として依存関係確率と単語翻訳確率を同時に推定 (Model 2) する。推定には EM アルゴリズムを用いる。

3.1 Model 1

基本的には IBM Model 1 と全く同じ方法で、各方向独立に推定を行う。ただし提案モデルでは内容語、機能語が内容語、機能語、NULL のどれに対応するかといった確率 (カテゴリ確率) も利用する点が異なる。予備実験ではこの確率の導入により、Model 1 のみのアライメントの精度が表 1 のように向上することがわかった。カテゴリ確率は出現頻度が極端に少ない単語を正しくアライメントするの有効である。これは出現頻度が少ない語はほぼ確実に内容語であるからである。なおカテゴリ確率は Model 1 の推定にのみ利用し、Model 2 では利用しない。

Model 1 の推定の際には対応の単位は各ノード単体のみであり、複数ノードのかたまりは考慮しない。つまり、この時点では 1 対 1 対応のみしか扱えない。多対多対応は Model 2 の推定から考慮し、多ノード候補を動的に作り出すことにより実現する。これは Model 1 の段階で多ノード候補全てを考慮すると、アライメント候補数が爆発し、扱えなくなるためである。

[1]我々は複数の単語のかたまりや係り受け情報もノードとしているため正確には単語ではないが、便宜上こう呼ぶ。

3.2 Model 2

単語翻訳確率と依存関係確率を同時に推定する。Model 2 では双方向の確率を両方用い、一つのアライメント結果を得る。

しかし Model 1 とは違い近似なしでモデルを完全に推定することはできないため、単語翻訳確率のみを用いて初期アライメントを生成し、山登り法により依存関係確率を考慮した n-best アライメントを探索する。

依存関係確率は、原言語側で直接の親子関係にあるノード (f_j と $f_{j'}$) の対応する目的言語側のノード (e_{a_j} と $e_{a_{j'}}$) の関係 ($rel(e_{a_j} \leftarrow e_{a_{j'}})$) に対して与えられる確率 $p(rel(e_{a_j} \leftarrow e_{a_{j'}})|rel(f_j \leftarrow f_{j'}))$ である。$rel(P_2 \leftarrow P_1)$ はあるノード (P_1) から別のあるノード (P_2) までの経路で表し、離れているノード数分だけ以下の表記を並べることにより表現する。

- P_2 が P_1 の (前の子、後ろの子): (c-、c+)
- 同じ句内で P_2 が内容語、P_1 が機能語: c
- P_2 が P_1 の (前の親、後ろの親): (p-、p+)
- 同じ句内で P_2 が機能語、P_1 が内容語: p

例えば図 1 において “いう” から見た “場合” の依存関係は “c-”、“考える” は “c-;c”、“考える” は “c-;c”、“ある” は “c-;c;c-;c” などとなる。

3.2.1 初期アライメントの生成

原言語側、目的言語側の全てのノード及び多ノード候補間の単語翻訳確率を計算し、単語翻訳確率の最も高いペアから順に採用する。ただし、すべてのノードは 1 度しかアライメントされない。つまり対応の採用は排他的に行われるが、そもそも多ノード候補を考慮しているため自然と多対多対応が実現される。多ノード候補の生成については 3.2.3 章で述べる。なお、単語翻訳確率は両方向の単語翻訳確率を単純に掛け合わせることにより計算される。

3.2.2 山登り法

初期アライメントの状態から、依存関係確率を考慮しながらアライメントを修正していき、徐々に確率の高いアライメントを探索していく。修正手段としては以下の 2 種類を考える。

対応の入れ替え: 任意の 2 つの対応に注目し、それらの対応を入れ替える。

大きな対応への拡大: 任意の 1 つの対応に注目し、その原言語側または目的言語側いずれかの対応を、親または子方向に 1 ノード分だけ拡大する。拡大先のノードにすでに対応が存在する場合、その対応は棄却し、反対言語側のノードは NULL 対応となる。

修正後のアライメント確率が修正前よりも高くなる場合にのみ修正を実行し、修正された状態から再度修正を行っていく。確率が高くなる修正箇所がなくなるまで修正を繰り返し行う。最終的に得られたアライメントが、最も確率の高いアライメントとなる。

3.2.3 多ノード候補の生成

獲得された最も確率の高いアライメントの中に、NULL に対応付けされたノードがあった場合、それらを親、または子の NULL 対応でないノードに併合したものを新たな多ノード候補として生成し、次のイタレーションから探索に入れる。この新たな多ノード候補はその対訳文から 1 回出現したものとして数える。

4 実験と考察

JST 日英抄録コーパスを用いてアライメント実験を行った。このコーパスは、科学技術振興機構所有の約 200 万件の日英抄録から、内山・井佐原の方法 [8] により、情報通信研究機構が作成したものであり、100 万対訳文からなる。このうち 100 対訳文についてアライメントの正解を人手で付与し、アライメント精度を適合率、再現率、F 値により測定した。実験は全ての語を原形に戻して行った。また提案モデルでは英語の冠詞は無視している。実験結果を表 2 に示す。

Model 1 は両方向とも 5 回ずつイタレーションを行い、得られたパラメータを用いて 3.2.1 章で説明した初期アライメントを生成し、これを評価したものである。Model 2-1 は Model 1 のアライメント結果から依存関係確率を推定して Model 1 のパラメータに追加した状態でのアライメント精度である。つまり Model 1 と Model 2-1 とを比較することにより、依存関係確率の効果が見て取れる。以後 Model 2 のイタレーションを 5 回行い、その都度アライメント精度を算出した。

また比較として GIZA++による双方向のアライメント結果を様々な対称化アルゴリズムでマージした結果も示す。これらの結果と比較すると、提案モデルの方がより精度の高いアライメントが実現できていることが分かる。さらに、対訳辞書などの言語リソースを用い、アライメントの整合性尺度を用いた手法 [6] による結果も示した。この手法と比較すると精度は若干劣っているが、対訳辞書などを全く用いていなくてもある程度の精度が達成できるということが分かった。また、提案手法に置いても数字の汎化や同じアルファベットの出現などに注目するなどの枠組を入れることにより、[6] と同等かそれ以上の精度を実現することができると考えられる。

表 2: アライメント実験結果

	適合率	再現率	F 値
Model 1	76.52	50.23	60.65
Model 2-1	81.58	54.95	65.67
Model 2-2	82.79	60.34	69.81
Model 2-3	83.12	62.27	71.20
Model 2-4	81.70	62.78	71.00
intersection	88.14	40.18	55.20
grow-final-and	78.00	52.93	63.06
grow-diag-final-and	74.95	54.26	62.95
整合性尺度 [6]	64.24	84.19	72.88

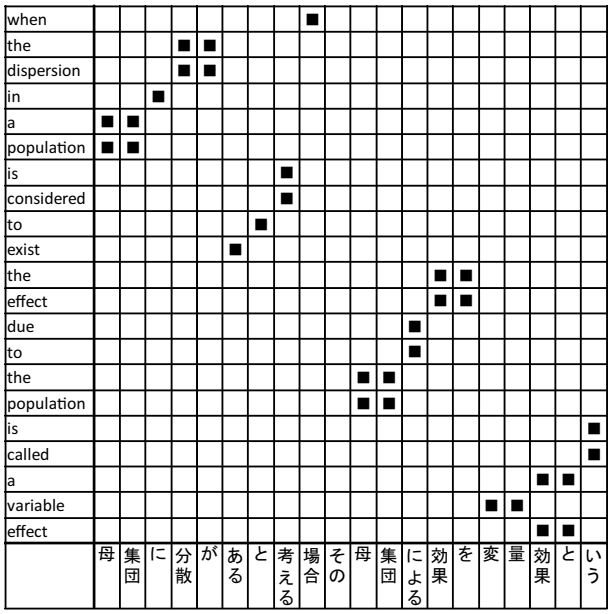

図 2: アライメント結果例

図 2 にアライメントの例を示す。例を見ると、“母 集団 ↔ population”、“変 量 ↔ variable”、“に よる ↔ due to” など、期待する多対多の対応が得られていることが分かる。またアライメント誤りの原因として最も大きいのはパース誤りであり、次いで多いのが初期アライメントのミスによるものであった。一方の言語で省略が起こっている場合や、同じ語が複数回出現する場合など、翻訳確率だけでは正しいものを判断できず、山登り法でも修正できなかったケースが多かった。これは初期アライメント生成時に曖昧性がある場合などは、多少依存関係確率を考慮するなどすれば解決すると思われる。

5 結論

本稿では句の依存関係確率モデルを用いた統計的アライメント手法を提案した。この確率モデルは木構造上での reordering モデルと言うことができ、シンプルなモデルながらも言語構造の違いを柔軟に吸収し、精度の高いアライメントを実現できた。今回はアライメントの精度のみを評価したが、この結果が翻訳の精度にどのように影響するかを調査する必要がある。

提案手法は依存構造解析に大きく依存しており、依存構造解析誤りが容易にアライメントの誤りにつながってしまう。各言語独立に依存構造解析の精度向上を期待することはできるが、対訳がある状況では両言語の解析結果を照らしあわせて、文構造を修正しつつアライメントすることも可能なはずであり、現在検討中である。これが実現できれば、依存構造解析とアライメント双方の精度向上が可能となると考える。

参考文献

[1] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Association for Computational Linguistics*, 19(2):263–312, 1993.

[2] Colin Cherry and Dekang Lin. A probability model to improve word alignment. In *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics*, pages 88–95, 2003.

[3] Brooke Cowan, Ivona Kučerová, and Michael Collins. A discriminative model for tree-to-tree translation. In *Proceedings of the 2006 Conference on EMNLP*, pages 232–241, Sydney, Australia, July 2006. Association for Computational Linguistics.

[4] Philipp Koehn, Franz J. Och, and Daniel Marcu. Statistical phrase-based translation. In *HLT-NAACL 2003: Main Proceedings*, pages 127–133, 2003.

[5] Toshiaki Nakazawa and Sadao Kurohashi. Linguistically-motivated tree-based probabilistic phrase alignment. In *In Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA2008)*, 2008.

[6] Toshiaki Nakazawa, Kun Yu, and Sadao Kurohashi. Structural phrase alignment based on consistency criteria. In *In Proceedings of Machine Translation Summit XI (MT-Summit XI)*, pages 337–344, 2007.

[7] Chris Quirk, Arul Menezes, and Colin Cherry. Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 271–279, 2005.

[8] Masao Utiyama and Hitoshi Isahara. A japanese-english patent parallel corpus. In *MT summit XI*, pages 475–482, 2007.

[9] Kenji Yamada and Kevin Knight. A syntax-based statistical translation model. In *Proceedings of 39th Annual Meeting of the ACL*, pages 523–530, 2001.