集合パッキング問題に基づく文アラインメントのモデル化

西野正彬　　平尾努　　永田昌明

NTT コミュニケーション科学基礎研究所
{nishino.masaaki,hirao.tsutomu,nagata.masaaki}@lab.ntt.co.jp

概要

文アラインメントは統計的機械翻訳に必要な文単位での対訳コーパスを用意するために不可欠な処理である．これまでに既存の文アラインメント法は，対訳関係にある二つの文書において，文の順序が大きく入れ替わらないことを前提としていた．しかし，文書の種類によってはパラグラフの順序が入れ替わることも想定される．本稿では，ある大きさの文のまとまりを単位として文の順序が大きく順序が変動する対訳コーパスに対して文アラインメントをとるための方法を示す．手法のポイントは，重み付き集合パッキング問題として定式化して解くことで，文のまとまりの発見と対応付けとを同時に行えるようにした点にある．

1　はじめに

統計的機械翻訳では，対訳コーパスにおいてどの文がどの文を翻訳したものであるかという文対文での対応付けが与えられているという前提のもとで処理が適用される．しかし実際の対訳コーパスでは，文書対文書での対応付けは行われていても，文対文の対応付けはとられていないものも多い．そのため，対訳文書間での文同士の正しい対応付け（文アラインメント）を求めることは，精度のよい統計的機械翻訳を実現するための重要な前処理として位置づけられる．

これまでに多くの文アラインメント手法が提案されてきているが，いずれも対訳関係にある二つの文章における対応する文の出現順序が大きく違わないことを前提としていた．すなわち，対訳文書のペア F，E があったとき，F の i 番目の文に E の j 番目の文が対応するとしたら，F の $i+1$ 番目の文に対応する E の文は，(存在するならば) j の近傍にあるという前提のもとで文同士の対応付けを行っていた．この前提は，例えば小説のように文の順序が大きく変動すると内容が損なわれてしまうような文書に対しては妥当なものである．一方で，例えば百科事典の記事のように，一つの文書が独立な複数の文のまとまりからなる場合には，文のまとまりの出現順序が大きく変動しても内容が損なわれないことがある．このような文書においては，文の順序が大きく変動しないという前提は必ずしも正しいものではないため，既存手法では正しい文アラインメントが行えない可能性が高い．

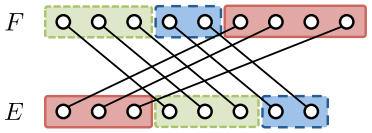

図 1: 提案法による文アラインメントの概観

本稿では，文の順序が，ある大きさの文のまとまりを単位として大きく変動する対訳文書において，文アラインメントを行う方法を提案する．仮に文書 F の文が E の任意の文と対応してもよいとすれば，文間の類似度を設定することによって，問題は二部グラフにおける最大重みマッチング問題として定式化して解くことができる．しかし，任意の文と対応してもよいという前提では，近傍の文間のつながりを無視して対応付けを行うことになるため，ある程度の大きさの文のまとまりを単位として順序が変動する状況では，正しい対応付けが行えない可能性が高い．そこで，提案手法では文アラインメントを組合せ最適化の問題の一つである集合パッキング問題として定式化して解く．これにより，前後の部のつながりを加味しつつ，文の順序の変更を許容する文アラインメントを実現することができる．

2　集合パッキング問題に基づく文アラインメントのモデル化

はじめに本稿で提案する文アラインメント法の概観を示す．まず，文アラインメントを求める対象である

文書 F とその対訳である文書 E のそれぞれを，同じ数 (K とする) の連続する文のまとまりに分割する．次にそれら K 個ずつの文のまとまりについて，一対一の対応付けを求める．最後に，対応付けられた文のまとまりのペアに対し，文の並び方が大きく違わないことを前提とする既存の文アラインメント法を適用することでそれぞれの文アラインメントを行う．このように文アラインメントを行うことによって，段落などの大きなまとまり単位で出現位置が大きく異なっているような文書に対しても，正しい文アラインメントを求めることができる．$K = 3$ としたときの文アラインメントの様子を図 1 に示す．図中の黒い丸がひとつの文に対応している．また，複数の丸を囲む四角が文のまとまりを表す．四角に囲まれた文同士では，既存の文アラインメント手法が適用されるため，対応付けの順序が交差することはない．

次に実際にアラインメントを求める方法を述べる．上記の説明ではまず文のまとまりを求めるとしていたが，前もって K および個々の文のまとまりを適切に定めることは難しい．そこで，実際には任意の文のまとまり同士を対応づけたときのスコアをまず求めた後に，文全体のアラインメントスコアが最大となるように文のまとまりの個数 K とその対応付けを集合パッキング問題を解くことで求める．

以下，各手順の詳細を順に説明するが，その前に以下で用いる記法について述べる．対応付けをとる対象の 2 つの文書を F, E とし，それぞれ n_f，n_e 個の文からなるとする．f_i を F に含まれる i 番目の文，e_k を E に含まれる k 番目の文とする．F の i 番目から j 番目までの連続する文の集合を $f_{ij} \subseteq F$ とする．ただし $1 \leq i \leq j \leq n_f$ である．$e_{kl} \subseteq E$ は同様に E の k 番目から l 番目までの連続する文の集合とする．ただし $1 \leq k \leq l \leq n_e$ である．また，a を文のまとまりのペア (f_{ij}, e_{kl}) を表現するために用いる．$f(a) = f_{ij}$, $e(a) = e_{kl}$ とする．

2.1 系列マッチングによる文のまとまり間のアラインメント

まず文のまとまり間のマッチングスコアを計算する手順について説明する．この手順では既存の文アラインメント法を適用することで，文の順序が入れ替わらない前提のもとで文のまとまり間の文アラインメントを求める．提案手法では任意の系列マッチングに基づいた既存の文アラインメント法を用いることができるが，本稿では代表的な文アラインメント法である Moore による手法を用いた [Moore 02]．Moore の手法は，文の長さと文中の語の翻訳確率とを用いることでペアのスコアを定義し，文アラインメントに含まれるペアのスコアが最大となるようにアラインメントを求める．具体的には，$s \in F$ である文 s と $t \in E$ である文 t とのペアのスコア $S(s,t)$ を

$$S(s,t) = \frac{P(m_s, m_t)}{(m_s+1)^{m_t}} \left(\prod_{j=1}^{m_t} \sum_{i=1}^{m_s} tr(t_j|s_i)\right) \left(\prod_{i=1}^{m_s} u(s_i)\right) \tag{1}$$

として定める．ここで，m_s, m_t はそれぞれ文 s, t に含まれる単語の総数である．また，s_i, t_j はそれぞれ s の i 番目の語，t の j 番目の語を表す．$tr(t_j|s_i)$ は語 s_i が t_j に翻訳される確率である．$u(s_i)$ は語 s_i の文書中での相対頻度を表す．$P(m_s, m_t)$ は，文の長さ（語の数）に応じてスコアを定める関数であり，ポアソン分布を用いて

$$P(m_s, m_t) = \frac{\exp(-m_s r)(m_s r)^{m_t}}{m_t!} \tag{2}$$

として定義される．r はパラメータである．各確率分布は [Brown 93] にある手法によってデータから推定できる．Moore の手法を適用することによって，任意の文のまとまりのペア f_{ij}, e_{kl} について，系列マッチングを行った際の最大のスコアを得ることができる．このスコアを $\mathrm{seqMatch}(f_{ij}, e_{kl})$ と表現する．

2.2 集合パッキング問題に基づく定式化

前節で求めた $\mathrm{seqMatch}(f_{ij}, e_{kl})$ を用いて，文のまとまり同士の一対一の対応付けを求める．文のまとまり同士の対応付けを求めることができれば，文同士の対応付けは既に前節で求めてあるので，文アラインメントが得られる．前節で求めた可能なすべての文のまとまりのペアの集合を $\mathcal{M}$ とすると，ある文同士の対応付け A は $A \subseteq \mathcal{M}$ である．ただし，A に含まれる任意の $a, a' \in A$ について $f(a) \cap f(a') = \emptyset$ かつ $e(a) \cap e(a') = \emptyset$ であり，$\cup_{a \in A} f(a) = F$ かつ $\cup_{a \in A} e(a) = E$ を満たすものとする．上記の条件を満たす A の集合を $\mathcal{A}$ とすると，文アラインメントを求める問題は，

$$A^* = \underset{A \in \mathcal{A}}{\mathrm{argmax}} \{\mathrm{score}(A)\} \tag{3}$$

として，マッチングのスコアを最大とする A^* を求める問題として定式化することができる．ここで $\mathrm{score}(A)$ は，F と E に対して，対応付け A を定めたときのス

$$
\begin{aligned}
\text{maximize} \quad & \sum_{ijkl}(w_{ijkl} + \log(\lambda))a_{ijkl} \\
\text{subject to} \quad & \sum_{i \le x \le j} f_{ij} = 1 \qquad \forall x : 1 \le x \le n_f \\
& \sum_{k \le x \le l} e_{kl} = 1 \qquad \forall x : 1 \le x \le n_e \\
& f_{ij} = \sum_{k,l} a_{ijkl} \qquad \forall i, j \\
& e_{kl} = \sum_{i,j} a_{ijkl} \qquad \forall k, l \\
& f_{ij}, e_{kl}, a_{ijkl} \in \{0, 1\} \\
& 1 \le i \le j \le n_f, \quad 1 \le k \le l \le n_e
\end{aligned}
$$

図 2: 整数線形計画問題としての定式化

コアであり，以下のように定義する．

$$\text{score}(A) = \lambda^K \prod_{a \in A} \text{seqMatch}(f(a), e(a)) \quad (4)$$

ここで λ はペアの個数に応じて課されるペナルティを表すパラメタであり，K は A 中に含まれるペアの総数である．$0 < \lambda \le 1$ を満たす．小さい λ を設定することによって，文書が多くの小さなまとまりに分割されてしまうことを防ぐことができる．

A^* を求める問題は (3) の対数をとったものを最大化する整数線形計画問題 (ILP) として定式化することができる．ILP による定式化を図 2 に示す．ここで，w_{ijkl} は $\log \text{seqMatch}(f_{ij}, e_{kl})$ の値である．a_{ijkl} はペア (f_{ij}, e_{kl}) を表す変数であり，$a_{ijkl} = 1$ のときはペア (f_{ij}, e_{kl}) が文アラインメントに含まれるとする．また，f_{ij} は，対応付けにおいて F の i 番目の文から j 番目の文までが一つのまとまりとして利用されることを示す変数である．e_{kl} についても同様である．

制約は，F と E に含まれる各文が最終的に得られた文のまとまり同士の対応付けのいずれか一つに必ず含まれることを保証するものである．上の二つの制約は，F, E の各文が必ず 1 つの文のまとまりに含まれることを保証している．次の二つの制約は，ある文のまとまりが二つ以上のペアで同時に利用されないことを保証している．

今回用いた定式化は，任意のペア (f_{ij}, e_{kl}) に対応する集合の集まりである集合族に対する集合パッキング問題となっている．集合パッキング問題は，ある集合族 $\mathcal{S}$ が与えられたときに，含まれる集合同士が互いに素であるような $C \subseteq \mathcal{S}$ を求める問題である．集合パッキング問題は NP 完全であり厳密な多項式時間のアルゴリズムは知られていない．そこで検証では ILP ソルバを用いて最適解を求めた．

3 検証

提案手法の有効性を検証する．検証のためのデータとして，文対応のついた日本語と英語の対訳文書から生成した人工データを用いた．対訳文書はそれぞれ約 25,000 文からなる．この文書から取り出した 2,500 文から文の長さが一定以上に長いものと短いものとを除いたものをテストデータを生成する元データ，残りを翻訳確率等を推定するための訓練データとして用いた．

テストデータの生成手順は以下のとおりとする．まず，元データのそれぞれの文書集合から，K 個の対応関係にある連続する文のまとまりをランダムに取り出す．そしてその文のまとまりをランダムに並べなおしたしたのちに，各まとまりに含まれる文を順に並べることで文のまとまり単位での移動があるデータセットを作成した．テストデータの文の数は日本語，英語ともに 60 文とし，まとまりの数は $K = 3, 6, 12$ とした．また，日本語と英語の文の数が異なる非対称データセットも同様に作成した．こちらでは日本語の文数を 60 文，英語の文数を 40 文とし，日本語の 20 文は対応する文が存在しないようにした．日本語のまとまりの数は $K = 3, 6, 9$ とし，英語のまとまりの数は日本語のまとまりの数の 2/3 とした．

比較対象として，Moore らによる系列マッチングに基づく手法 (Moore) と，二部グラフの重み最大マッチングとして解いた方法 (BM) とを用いた．なお，重み最大マッチングにおけるリンクの重みは (1) を用いた．評価は Moore[Moore 02] にならって，文の対応付けのの再現率 (recall)，適合率 (precision)，F 値 (F-measure) を算出した．対称，非対称の各データセットについて，異なる K ごとに 5 つのデータセットを生成し，その平均値を最終的な評価値とした．翻訳確率の算出には GIZA++[Och 03] を用いた．整数線形計画問題のソルバとして ILOG CPLEX を用いた．文のまとまりの個数に対するペナルティ λ は，$\lambda = 0.1$ と $\lambda = 0.01$ の 2 種類を試した．

3.1 結果

実験結果を表 1, 2 に示す．表より，いずれのデータセット，および K の値においても，$\lambda = 0.1$ としたときの提案手法が高い F 値を示していることが分かる．

表 1: 検証結果 (対称データ)

	$K=3$			$K=6$			$K=12$		
	再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値
提案法 ($\lambda=0.1$)	**0.986**	0.917	0.949	**1.000**	0.822	**0.900**	**0.990**	0.764	**0.859**
提案法 ($\lambda=0.01$)	**0.986**	**0.923**	**0.953**	0.804	**0.831**	0.814	0.609	**0.776**	0.680
BM	**0.986**	0.853	0.912	**1.000**	0.730	0.843	**0.990**	0.694	0.814
Moore	0.585	0.872	0.690	0.524	0.743	0.613	0.506	**0.776**	0.603

表 2: 検証結果 (非対称データ)

	$K=3$			$K=6$			$K=12$		
	再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値
提案法 ($\lambda=0.1$)	**0.989**	0.876	0.928	**0.990**	0.869	**0.925**	**0.988**	0.758	**0.856**
提案法 ($\lambda=0.01$)	**0.989**	0.883	**0.932**	0.879	**0.896**	0.880	0.721	**0.790**	0.750
BM	**0.989**	0.772	0.865	**0.990**	0.794	0.879	**0.988**	0.674	0.796
Moore	0.638	**0.907**	0.739	0.707	0.827	0.758	0.486	0.719	0.564

λ の値の違いによる影響として，いずれのデータセットにおいても $K=3$ のときは $\lambda=0.01$ の方がよい F 値を示していることが分かる．これは，λ が文のまとまりの個数に対するペナルティであり，λ が小さいほど大きなペナルティを与えていることによって説明できる．つまり，K が小さいときは文のまとまりの個数が小さくなりがちな $\lambda=0.01$ の方がよい結果を出力し，K が大きいときはより多くのまとまりが出現することを許容する $\lambda=0.1$ の方がよい結果を出力していると考えられる．Moore による系列マッチングに基づいた対応付け手法は，今回用いたデータセットのように文のまとまり単位で文の順序が大きく変動するデータは想定していないため，他の方法と比べると極端に F 値が悪くなっているのが確認できる．

4 関連研究

これまで，文の長さを対応付けに利用する方法 [Gale 93]，語の翻訳確率と文の長さを利用する方法 [Moore 02, Braune 10], などが提案されてきているが，ほとんどの方法で系列マッチングによって文アラインメントができることを前提としている．例外として，Deng らは系列マッチングとクラスタリングをあわせて利用することで，文の順序が入れ替わる場合でも文アラインメントを行う手法を提案している [Deng 07]．しかし，Deng らの手法は，ある隣接する二つの文の順序が入れ替わるなど，順序の入れ替わりが小さい範囲で起きることを想定した手法となっている．一方，提案手法では文のまとまり単位で大きく順序が入れ替わることを想定している．

5 おわりに

本稿では，文のまとまりの出現順序が変動する対訳文書間で正しい文アラインメントを求めるための方法を提案した．最大集合パッキング問題として定式化し，整数線形計画法を用いて解くことによって，既存手法では対応付けをとるのが難しい状況でも対応付けができることを示した．今後の予定として，文書集合が巨大な場合でも動作するよう該当の問題をより効率的に解く近似解法について検討する予定である．

参考文献

[Braune 10] Braune, F. and Fraser, A.: Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora, in *Proceedings of COLING 2010* (2010)

[Brown 93] Brown, P. F., Petra, S. A. D., Pietra, V. J. D., and Mercer, R. L.: The mathematics of statistical machine translation: parameter estimation, *Computational Linguistics*, Vol. 19, No. 2 (1993)

[Deng 07] Deng, Y., Kumar, S., and Byrne, W.: Segmentation and alignment of parallel text for statistical machine translation, *Natural Language Engineering*, Vol. 13, No. 3 (2007)

[Gale 93] Gale, W. A. and Church, K. W.: A program for aligning sentences in bilingual corpora, *Computational Linguistics*, Vol. 19, No. 1 (1993)

[Moore 02] Moore, R. C.: Fast and accurate sentence alignment of bilingual corpora, in *Proceedings of AMTA'02*, pp. 135–144 (2002)

[Och 03] Och, F. J. and Ney, H.: A systematic comparison of various statistical alignment models, *Computational Linguistics*, Vol. 29, No. 1 (2003)