列生成法を用いた高速なアラインメント

西野正彬 † 鈴木潤 † 梅谷俊治 * 平尾努 † 永田昌明 †

† 日本電信電話株式会社　NTT コミュニケーション科学基礎研究所
* 大阪大学 大学院情報科学研究科
{nishino.masaaki,suzuki.jun,hirao.tsutomu,nagata.masaaki}@lab.ntt.co.jp
umetani@ist.osaka-u.ac.jp

概要

本稿では，2 つの系列が与えられたときに，系列のアラインメントを求める方法について述べる．特に，それぞれの系列の要素がいくつかのグループに分割され，かつそれらのグループ間の対応付けの順序が大きく変動するようなデータに対して，高速にアラインメントを求める方法を述べる．著者らが以前に提案した集合分割問題に基づくアラインメント法は，このようなアラインメントを可能とするが，系列の要素数が増加すると整数線形計画法で扱わなければならない変数の数が急激に増加し計算が困難になるという問題があった．本稿では大規模な整数線形計画問題を解くために用いられる列生成法を用いることで，高速にアラインメントを求める手法を提案する．

1　はじめに

本稿では自然言語処理におけるアラインメント問題を解く手法について述べる．アラインメントとは，2 つの系列が与えられたときに片方の系列のどの要素がもう片方の系列のどの要素と対応するかを求める問題であり，文アラインメントや単語アラインメント等，自然言語処理においては様々な単位でアラインメントを求める必要がある．

本稿では様々なアラインメントのうち，特に文を求めるためのアルゴリズムを例にとって議論する [1]．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（例えば [Moore 02, Ma 06] など），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．

上記のように文の順序が大きく変動する対訳文書を想定した文アラインメント法として，著者らは以前に組合せ最適化問題の一種である集合分割問題として文アラインメントを定式化して解く方法を提案した [西野 13]．集合分割問題に基づく手法では，F, E の文全体に対するアラインメントを，いくつかの文のまとまりごとの部分的な文アラインメントが集まったものであるとして定式化して解く．このとき，部分的な文アラインメントは既存の動的計画法による手法と同様に文の交差が発生しないものとして解き，文のまとまり単位の対応付けでは文の交差を考慮しないことによって，文のまとまりの順序が変動するアラインメントを実現した．

集合分割問題に基づく文アラインメント法の課題として，E, F に含まれる文の数が増加するに伴い，急激に処理に時間がかかる点が挙げられる．これは，それぞれの文集合に含まれる文の総数を $|E|$, $|F|$ とすると，集合分割問題における変数の数が $O(|E|^2|F|^2)$ となるため，整数線形計画法ソルバを用いた求解が困難になるからである．本稿ではこの課題に対処するために，大規模な（整数）線形計画問題を解く際に用いられる，**列生成法** [Lübbecke 05] を用いることで高速な文アラインメントを実現する方法を提案する．列生成法は大規模な問題を一度に解く代わりに，出現する変数の個数を制限した小さな問題を，変数を逐次追加しながら解く方法である．列生成法を用いることで，よ

[1] なお，提案手法は文アラインメントに特化した手法ではなく，他のアラインメント問題にも適応可能である．

$$\text{minimize} - \sum_{ijkl}(w_{ijkl} + \log(\lambda))x_{ijkl}$$

$$\begin{array}{ll} \text{subject to} & \sum_{i \le y \le j} \sum_{kl} x_{ijkl} = 1 \quad \forall y : 1 \le y \le n_f \\ & \sum_{ij} \sum_{k \le y \le l} x_{ijkl} = 1 \quad \forall y : 1 \le y \le n_e \\ & x_{ijkl} \in \{0, 1\} \end{array}$$

図 1: 集合分割問題による文アラインメント法の整数線形計画問題としての定式化.

り少ない数の変数からなる問題を解くことで解を得ることが期待できる．きなお，列生成法を用いることで線形計画問題の最適解を得られることは保証されているが整数線形計画問題については，解の下限は求まるものの最適解を得られることは必ずしも保証されていない．そこで本稿では最適解と列生成法で得られた整数解とを実験によって比較し，よい近似解が得られていることを確認する．

2 集合分割問題による文アラインメントの定式化

はじめに集合分割問題に基づく文アラインメント法について簡単に説明する．詳細は [西野 13] を参照されたい．文アラインメント問題は，可能なアラインメントにスコアを定めることで，スコアを最大（最小）化する最適化問題として定式化して解くことができる．対応付けを行う文の集合を $F = \{f_1, \ldots, f_{|F|}\}$, $E = \{e_1, \ldots, e_{|E|}\}$ とし，$|F|, |E|$ はそれぞれの文の集合に含まれる文の総数とする．F の i 文目から j 文目までの文のまとまりと，E の k 文目から l 文目までの文のまとまりに対し，文の交差を許さない既存の文アラインメント法（例えば [Moore 02] など）を適用すると，最適なそれらの文のまとまり間での最適アラインメントスコアを求めることができる．そうして得られたスコアを w_{ijkl} とすると，集合分割問題に基づく文アラインメント法は図 1 に示す整数線形計画問題として定式化することができる [2]．ここで x_{ijkl} は，最終的に得られる文アラインメントに文の集合 $f_i, \ldots, f_j$ と $e_k, \ldots, e_l$ を文の交差を許さない文アラインメント法で対応付けた結果が含まれることを示す二値変数である．また，λ は解に含まれる変数 x_{ijkl} の個数に対するペナルティであり，λ が大きいほどできるだけ少ない個数の x_{ijkl} からなる文アラインメントが得られやすくなる．最適化問題の 2 つの制約式は，それぞれ F, E に含まれるすべての文が文アラインメントに出現するいずれかの x に正確に一度だけ含まれることを要求している．

[2] この定式化は [西野 13] とは異なるが，等価なものである．

3 列生成法

前節で述べた手法を実行するためには，すべての可能な文のまとまりのペアについて交差を許さない文アラインメントスコアを求めたうえで，それらの文のペアの個数の変数が含まれる整数線形計画問題を解く必要がある．しかし文のまとまりのペアの個数は $|F||E|(|F|+1)(|E|+1)/4$ 個存在するため，文の数が増加するとそれぞれの問題を解くのに多大な時間を要するようになる．この問題を解決するため，本稿では列生成法を用いて最適化問題を解く．列生成法は大規模な線形計画問題を一括で解く代わりに，逐次的に変数を追加しながら問題を解くことで解を求めるアプローチである．扱う必要がある変数の数を減らすことで解を高速に求められることが期待できる．

列生成法を導入するにあたり，いくつかの概念を定義する．まず図 1 の整数線形計画問題を線形緩和した問題，つまり制約 $x_{ijkl} \in \{0, 1\}$ を $0 \le x_{ijkl} \le 1$ へと緩和した問題を主問題（Master problem: MP）とよぶ．主問題に含まれるすべての変数の集合を X とする．MP からいくつかの変数を取り除いた問題を制限された問題を制限された主問題（Restricted master problem: RMP）とよぶ．RMP に出現する変数の集合を $X' \subseteq X$ と表す．

列生成法は RMP の求解と RMP に追加する変数を求める問題とを繰り返し解くことで MP を解く．追加する変数を求める問題は列生成部分問題とよばれ，具体的には $x_{ijkl} \in X \setminus X'$ であるような x_{ijkl} のうち，

$$\overline{w}_{ijkl} = w_{ijkl} - \sum_{n=i}^{j} u_n - \sum_{m=k}^{l} v_m \qquad (1)$$

を最大とするものを一つ求める問題である．ここで，u_n は RMP の文 f_n に関する制約に対応する双対変数であり，v_m は文 e_m に関する制約に対応する双対変数である．現在の RMP の解が求まっているなら双対性によって u_n も v_m も容易に求めることができる．以下では $\overline{w}_{ijkl}$ のことを**被約費用**とよぶ．

前述のとおり，変数の個数は $|F||E|(|F|+1)(|E|+1)/4$ 個あるため，その全てについて被約費用を求めるのは困難である．ここで，スコア w_{ijkl} が動的計画法によって求められること，および被約費用の式 (1) において u_n, v_m が対応する文ごとにそれぞれ独立に作用していることに着目すると，最大の被約費用が Smith-Waterman 法 [Smith 81] によって求められることが分かる．Smith-Waterman 法は配列のローカル

1: $w_{1|F|1|E|}$ を計算し変数 $x_{1|F|,1|E|}$ を RMP に追加する．
2: **loop**
3: RMP を解く．
4: 列生成部分問題を解き，被約費用を最大とするアイテム x_{ijkl} を選択する．
5: **if** $\overline{w}_{ijkl}$ の費約費用が負 **then**
6: **break**
7: **end if**
8: RMP に x_{ijkl} を追加．
9: **end loop**
10: RMP に整数制約を追加し，整数線形計画問題として解く．

図 2: 列生成法

アラインメントを求めるためのアルゴリズムであり，動的計画法に基づいて長さ N, M の二本の配列に対するスコア最大の局所アラインメントを $O(NM)$ 時間で求めることができる．具体的には，$q[j,l]$ をその末尾の要素がそれぞれ f_j, e_l であるような文のまとまりのペアの被約費用 $\overline{w}_{ijkl}$ $(1 \leq i \leq j, 1 \leq k \leq l)$ の最大値とすると，$q[j,l]$ は

$$q[j,l] = \min \begin{cases} -\log \lambda \\ q[j-1,l-1] - w(f_j, e_l) - u_j - v_l \\ q[j-1,l] - w(f_j) - u_j \\ q[j,l-1] - w(e_l) - v_l \end{cases} \tag{2}$$

として再帰的に計算することができる．ここで $w(f_j, e_l)$, $w(f_j)$, $w(e_l)$ は，既存の動的計画法（例えば [Moore 02] など）による文アラインメントにおいて利用されるスコアであり，それぞれ文 f_j 番目の文と e_l 番目の文とを対応付けたときのスコア，f_j を E のどの文とも対応させなかったときのスコア，e_l を F のどの文とも対応させなかったときのスコアである．すべての $1 \leq j \leq |F|$, $1 \leq l \leq |E|$ について $q[j,l]$ を計算したのちに，それらのうち最大値をとることで被約費用の最大値を求めることができる．さらに，動的計画法によって $q[j,l]$ を計算する際に (1) のどの式をもとに計算したかを記憶しておけば，$q[j,l]$ からバックトラッキングを実行するすることによって被約費用を最大とする x_{ijkl} を求めることができる．バックトラッキングは高々$O(|F|+|E|)$ 時間で実行できるため，Smith-Waterman 法によって被約費用を最大とするアイテムを効率的に選択できる．

最後に列生成法の手順を図 2 に示す．まず RMP に含まれる変数の集合を $X^* = \{x_{1|F|1|E|}\}$ として初期化する (line 1)．$x_{1|F|1|E|}$ はすべての文からなる文のまとまりであり，実行可能解であることから，以降の RMP は必ず実行可能解をもつことが保証される．以降，RMP の求解 (line 3) と被約費用最大の変数を Smith-Waterman 法によって決定する問題 (line 4) を解くことを繰り返す．もしすべての変数で被約費用が負となったら (line 5)，その時点で MP の最適解が得られていることになるので，最後に RMP に整数制約を追加したうえで整数線形計画問題を解いて得られた解を出力する (line 8)．ここで，RMP に整数制約を追加して得られた解が必ずしも元の MP に整数制約を追加して得られた解と一致するわけではないことに注意する必要がある．すなわち，提案法はヒューリスティクスであり，必ずしも厳密な最適解を得られるわけではない．そこで検証によって厳密解との差を評価する．

4 検証

提案手法の有効性を検証するため，列生成法で文アラインメントを求めた場合と整数線形計画法として元の問題を解いた時の結果を比較した．検証のためのデータは，[西野 13] と同様に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．部分的な文アラインメントとそのスコアは Moore の手法 [Moore 02] を用いて求めた．

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 $K = 3, 6$ 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 $K = 3, 6$ 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．

評価尺度として，文の対応付けの再現率 (recall)，適合率 (precision)，F 値 (F-measure)，実行時間，目的関数値の増加率，そして変数の数を用いた．なお，増加率は

$$\text{増加率} = \frac{\text{列生成法の目的関数値} - \text{厳密解の目的関数値}}{\text{厳密解の目的関数値}}$$

表 1: 検証結果（対称データ, $K = 3$）

	再現率	適合率	F 値	実行時間 (秒)	増加率 (%)	変数の数
CPLEX	0.986	0.917	0.949	5.57×10^2		3.35×10^6
列生成法	0.986	0.923	0.953	2.35	0.211	1.02×10^3

表 2: 検証結果（対称データ, $K = 6$）

	再現率	適合率	F 値	実行時間 (秒)	増加率 (%)	変数の数
CPLEX	1.000	0.822	0.901	6.08×10^2		3.35×10^6
列生成法	1.000	0.814	0.897	1.31	0.268	8.30×10^2

表 3: 検証結果（非対称データ, $K = 3$）

	再現率	適合率	F 値	実行時間 (秒)	増加率 (%)	変数の数
CPLEX	0.989	0.876	0.928	8.80×10		1.50×10^6
列生成法	0.989	0.876	0.928	2.01	0.215	7.14×10^2

表 4: 検証結果（非対称データ, $K = 6$）

	再現率	適合率	F 値	実行時間 (秒)	増加率 (%)	変数の数
CPLEX	0.990	0.869	0.925	9.04×10		1.50×10^6
列生成法	0.990	0.843	0.910	1.09	0.240	7.18×10^2

として定義した．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．整数計画問題，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 $\lambda = 0.1$ とした．

4.1 結果

実験結果を表 1 から表 4 に示す．表より，CPLEX で数十秒から数百秒かかっていた問題が列生成法によって数秒で解けていることが確認できる．すべてのデータで 40 倍から 400 倍程度の高速化が達成できたが，特に変数の総数が多い対称データでは 200 倍以上の高速化が確認できた．これは，利用した変数の数が列生成法ではいずれの問題でも 10^3 個程度であり，10^6 個以上の変数からなる元の問題と比較して非常に小さな規模の問題とできたことに起因すると考えられる．

厳密解と比べたときの目的関数の値の増加率の平均値は，いずれのデータセットにおいても 0.3% 以内におさまった．この結果は，提案法はヒューリスティクスではあるものの文アラインメント問題ではよい解を出力できることを示している．

再現率，適合率，F 値については，列生成法で解くことによる精度の低下は最大で 3 ポイントにおさまっていることが確認できた．なお，表 1 では列生成法のほうが適合率，F 値が大きくなる結果が得られているが，これは目的関数を最大とする文アラインメントが必ずしもこれらの指標を最大化するものではないことに起因する．

5 おわりに

本稿では，集合分割問題に基づくアラインメント法を高速に実行するための手法を示した．数理計画法で利用される列生成法を適用することによって，最適化問題を解くときに扱わなければならない変数の数および各変数のスコアの計算に必要となる動的計画法の実行回数を劇的に減らすことができ，結果として高速な求解を可能とした．

提案手法は厳密解を求めることができないが，列生成法を用いた厳密解の求解アルゴリズムとして，分枝価格法 (Branch and price method) [Barnhart 98] をはじめ，様々な手法が提案されている．今後はこれら厳密解法の適用も検討したい．

参考文献

[Barnhart 98] Barnhart, C., Johnson, E. L., Nemhauser, G. L., Savelsbergh, M. W., and Vance, P. H.: Branch-and-price: Column generation for solving huge integer programs, *Operations research*, Vol. 46, No. 3 (1998)

[Lübbecke 05] Lübbecke, M. E. and Desrosiers, J.: Selected Topics in Column Generation, *Operations Research*, Vol. 53, No. 6 (2005)

[Ma 06] Ma, X.: Champollion: A robust parallel text sentence aligner, in *Proceedings of LREC* (2006)

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

[Smith 81] Smith, T. F. and Waterman, M. S.: Identification of Common Molecular Subsequences, *Journal of Molecular Biology*, Vol. 147, (1981)

[西野 13] 西野 正彬, 平尾 努, 永田 昌明：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, 言語処理学会年次大会 (2013)