GAによる構文木枝刈りを用いた単一文書要約

A Parse-Tree Pruning Approach to GA-Based Summarization

立林 裕太朗* 藤田 充洋** 古谷 克司* 三輪 誠* 佐々木 裕*

Yutaro Tatebayashi* Mitsuhiro Fujita** Katsushi Furutani* Makoto Miwa* Yutaka Sasaki*

*豊田工業大学

*Toyota Technological Institute

** (株)豊田中央研究所

**Toyota Central R&D Labs., Inc.

1 はじめに

自動要約は, 最適化問題と見なせるため, 組み合わせ最適化手法の一つである遺伝的アルゴリズム(以下, GA) が, 有用だと考えられる. しかし, これまでGAによる良好な要約結果は報告されていない. そこで, 本稿ではGAによる要約が, どの程度の品質を実現可能か探る. また, 要約例となる教師データの不足により, 新聞記事や学術論文の一部しか要約システムが適用できないケースもある. そこで本研究では, これまで他の研究で自動要約の対象とされていなかった機械加工報告書も対象に, 教師なし手法であるGAによる自動要約利用の可能性を検証する.

2 関連研究

GA を用いた要約の研究はいくつか報告されているが, 補助的な役割で利用しているものが多い. Yeh ら [1] は, 整数計画問題の重み付け調整を行うためにGAを用いている. またSillaら [2] は, 分類アルゴリズムの性能を向上させるために GA に基づいた特徴選択の有効性を調べている. GAを直接用いた要約手法として, 小倉ら [3] による複数文書要約を対象にした要約手法がある. 小倉らは, 重要文抽出を多目的最適化問題と見なし, 最適な文の組合せを選択する手法を採用した. また筆者らは, 参照要約を用いたフィットネス関数により, GA を使った要約手法の性能について予備的検討を行った [5] . しかし, 正解を用いたにも関わらず GA による要約の品質は良くなかった. その理由として, N-gram ベースの手法であること, 要約文の構文構造に関する特徴を取り込むことが課題となっていた.

3 単一文書要約におけるGAの利用

2節で示した問題を解決するため, これまでの筆者らの研究[4,5]に従って, 前処理および重要文抽出を行ったあと, 構文木枝刈りを用いたGAによる文短縮を行う方法を提案する.

3.1 実験データの前処理

重要文抽出の精度向上を目的に SVM を用いた前処理を行う. 機械加工報告書には項目を並べた箇条書きが存在するが, 箇条書きは例示や詳細化のために利用され, 重要な内容を含まない傾向がある. そこで, 文毎に句読点の有無や長さなどの素性をSVMに学習させ, 前処理として重要文候補に適さない表現を除去する. 表 1 に用いた素性の一覧を示す.

3.2 重要文抽出

重要文抽出を行うために, TF-IDF値を用いた抽出および 重要な情報が含まれやすい文を重点的に抽出するリード法を用いる. TF-IDF値は, 単語の出現頻度TF, その単語が文書群の中で出現した文書の数の逆数IDFの2つから計算する. リード法は, 「最初の文が重要な文になりやすい」などを考慮し, 文頭や文末, 見出し語の次の文のスコアに重みを与える手法である.

表1. 実験データの前処理で用いた素性

素性	次元	内容
文の長さ	150	文Sの文字数
単語ID	1900	文中に出現した単語のID
文の位置	30	文書からみた文の位置番号
句読点	1	文中に句読点が存在するか
漢字の数	80	文に含まれる漢字の割合
文を占める漢字の割合	10	文に含まれる漢字の割合
カタカナの数	80	文に含まれるカタカナの数
文を占めるカタカナの割合	10	文に含まれるカタカナの割合
ひらがなの数	80	文に含まれる単語の数
文を占めるひらがなの割合	10	文に含まれる単語の割合
半角文字数	80	文に含まれる英語や数字の数
文を占める半角英数字の割合	10	文に含まれる英語や数字の割合
文頭の文字	5	文頭の文字の種類

3.3 文短縮

非文の生成防止を目的に構文木の枝刈りを用いた教師あり学習，教師なし学習による文短縮を行う．教師あり学習としてSVM，教師なし学習としてGAを用いる．SVMでは，先行研究 [4] に，構文木に関する素性を与え，文節毎に要否を判別するよう拡張する．用いた素性を表2, 3に示す．一方，GAでは，単語の出現頻度と構文木の深さを考慮した値TF-IDEPをフィットネス関数として文短縮を行うことを新たに提案する．

$$TF-IDEP = \frac{\log(1+TF_{ij})}{depth(i)^2}$$

ここで，TF_{ij} は文書 j における単語 i の出現頻度，depth (i)は，構文木の根からの深さである．TF_{ij} に対数，depth (i)に２乗の重みを付与したのは，重みを付与しない場合との比較実験による結果を考慮したためである．構文木枝刈りは特定の文節を除去する場合，係り元となる文節も除去することである．構文木枝刈りの例を図 1 に示す．図中の「非文が生成される例」のように構文木の枝刈りを考慮しない場合は，「③切断する」を除去した場合，「②材料を」が残り，文の一貫性を損ねてしまうが，枝刈りを考慮する事で②も除去する事が可能となる．

表 2. SVM 文短縮における文節に関する素性一覧

素性	次元	内容
文末表現	1	文節Pは文末かどうかを示す二値
係る文節数	10	文節Pに直接係る文節の数
文節番号(文毎)	30	文中の文節番号
文節番号(文全体)	50	文全体の文節番号
文位置	100	文全体からみた文節頭の単語番号
係り受ける文節数	30	文節Pが直接係る文節の数
構文木の深さ	10	構文木の根からの距離

表 3. SVM 文短縮における単語に関する素性一覧

素性	次元	内容
単語の数	50	一文節に含まれる単語数
単語ID	1900	単語ID
品詞ID	69	品詞ID

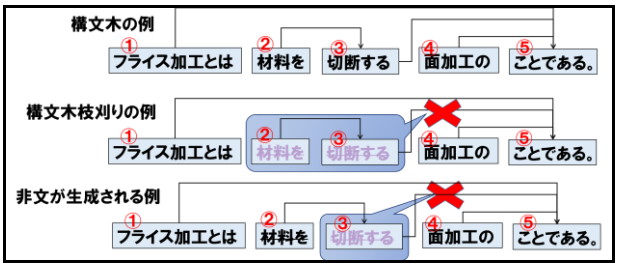

図 1. 構文木枝刈りによる要約手法

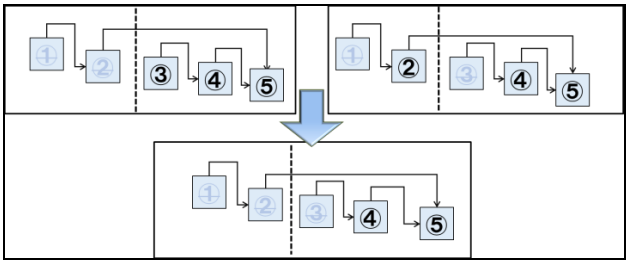

図 2.　2 つの遺伝子を交叉させる様子

3.4 GA を用いた解法

a) 染色体表現

GA に基づく解法では，染色体上の遺伝子への問題の表現方法により，解法の性能が大きく異なる．また遺伝子型から表現型への変換方法も解法の性能に影響する．一般的には，遺伝子型と表現型が 1 対 1 に対応し，表現型において総ての遺伝子型が実行可能解となることが望ましい．本研究では，構文構造を考慮した文短縮を行うために，木構造を遺伝子型とする事で，表現型が持つ構文情報を保持した．符号化については，遺伝子配列を 0 と 1 の 2 進数からなるものにしている．遺伝子が 1 ならばその文節は必要な情報として，要約文に配置する．一方で，0 ならばその文節は不要であるとして，要約文に配置しないという符号化手法を用いる．図 1 で振られている数字は，染色体の配列の順番である．

b) 交叉の方式

ルーレット・モデルまたはモンテカルロ・モデルとも呼ばれる適応度比例戦略を用いた．この戦略は，各個体の適応度に比例した確率で子孫を残せる可能性がある戦略である．また，任意の構文木の枝同士を無作為に交叉させることで，不要な末端の枝を切った組み合わせを新たに生成する木構造コーディングを採用した．

c) 突然変異の方式

突然変異のオペレータとして，枝刈りもしくは枝の復活をランダムに行う．例を図 3 に示す．この例では，②から⑤に係っている枝を刈る操作を行っている．この

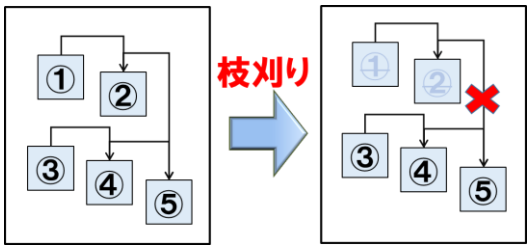

図 3 .突然変異による枝刈り

操作によって実行不能解は生成されない. すなわち生成される個体，順列は必ず実行可能解となる.

d) その他のパラメータ

表 4 に，GA のパラメータについて示す.

4 実験

4. 1 評価指標

自動要約の評価法として用いられている ROUGE-*N* [8] を用いた．ROUGE-*N* は，参照要約と生成要約で一致する *N*-gram の割合を計算した値である． ここではスペースの都合により，ROUGE-1 の結果のみ報告する.

4. 2 対象コーパス

NTCIR3-TSC2 [10] による 329 文書，および機械加工の現場で作成された機械加工報告書 96 文書をそれぞれ用いた. 機械加工報告書のデータについて表 5 に示す. 機械加工報告書には，人手でつけられた要約部と加工内容に関する説明文が与えられている．本研究では，要約部を参照要約として扱う．要約部は，タイトルのような簡潔な文であるため，本研究の機械加工報告書に対する要約は，ヘッドライン要約ととらえることができる．機械加工報告書が持つ特徴として，原文から重要な個所を抜粋する方法で正解要約を作成していないこと，箇条書きなどの表現が多く存在することがある. 形態素解析には MeCab [8]，係り受け解析には CaboCha [9] を用いた．専門用語の解析のため，機械加工のための MeCab ユーザ辞書 [10] を用いた.

表 4. GA のパラメータ

	集団数	交叉	突然変異
実験条件	200 個	80%	2%

表 5. 機械加工報告書の概要

データ種類	文字数	行数
原文書	30,938	1,032
参照要約	4,570	191

4. 3 実験手法

・実験データの前処理

箇条書きを分類するため，対象である機械加工報告書，TSC2 の全文書に対してタグ付けを行った. 箇条書きの数は機械加工報告書では 769 であり，TSC2 では 23 だった. 機械学習には LIBSVM [11] を用いた. 学習の際，データを 8 分割して交差確認を行った． 実験の結果，重要文として適さないとした表現は，次項の重要文抽出の対象外としている.

・重要文抽出

TF-IDF スコアの算出のため，対象である機械加工報告書・TSC2 文書に対し，形態素解析を行った後，名詞であった単語のみを抽出した. なお，形態素解析を行う際，機械加工の専門用語を追加したユーザ辞書を用いた. TF-IDF を用いて，対象の文書毎に，文書に含まれる全ての名詞群に対し，TF-IDF を算出する．その後，各文書の任意の文中に出現する単語の TF-IDF の総和を文の重要度として計算し，文の重要度が高い文を抽出した.

・文解析

重要文抽出にて生成した重要文集合の全ての文に対して，形態素解析，係り受け解析を行った．係り受け解析の結果から，文節毎に分割した.

・比較対象手法

- [SVM]：3.3 節で説明した SVM による文短縮
- [GA]： 3.1 節の前処理を使わない GA
- [GA+Prep]：提案手法. 3.1 節の前処理結果を用いた GA
- [Greedy]：参考のため，正解要約を基に評価値の高い順に文節を取り込む Greedy 法を用いて，構文木枝刈りにより生成された要約が取りうる評価値の上限値（近似値）を算出したもの.

なお，SVM と GA は「枝刈りあり」と「枝刈りなし」で比較実験を行った. また，比較のために，GA のフィットネス関数として TF-IDEP に加えて，TF-IDF を用いた場合についても実験を行った.

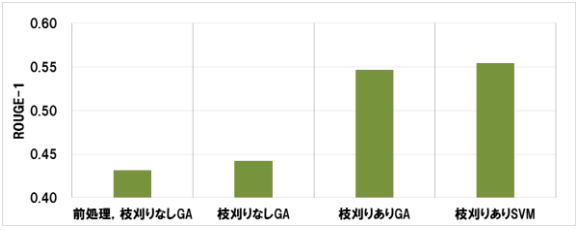

図 4. 前処理, 枝刈りの有無による比較 (機械加工技術文書)

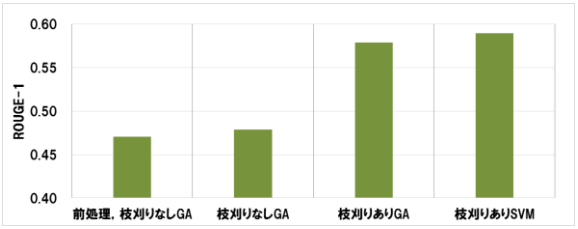

図 5. 前処理, 枝刈りの有無による比較 (TSC2)

4.4 結果と考察

ROUGE-1 を用いて, NTCIR3-TSC2 と機械加工報告書についてそれぞれ生成要約と正解要約の類似度を評価した. 視覚的な比較のため, 機械加工報告書に関する主要結果を図4に, NTCIR3-TSC2 の結果を図5に示す. また, 詳細な結果について表 6, 7 に示す. t 検定による有意差検定を行った結果、構文木枝刈りの有無による有意差が確認でき, GA と SVM による手法の違いには有意差が見られなかった. また, 機械加工報告書を用いた場合, 事前に不要な個所を削る「前処理」がある程度の効果を持つことが分かった。このことから正解要約を用いずとも教師あり手法に匹敵する要約を生成できること, 特定の文書においては予め不要な個所を同定することの有用性が示された. また, 生成された要約を実際に読んでみたところ, 機械加工報告書の一部の文書では, 抽出された重要文集合が談話構造を保持しておらず, 原文書の意図が把握しにくい問題が存在していた.

5 おわりに

機械加工報告書を対象とした構文木枝刈りを用いた GA による要約手法を提案した. 実験の結果, GA によりSVMを用いた要約と同等の品質で要約を生成することが出来た. 今後の課題としては, 文書の一貫性や情報の網羅性に着目した要約手法が挙げられる. 重要文の抽出を行う際に, 文と文の依存関係, 指示語などの単語間の依存関係を考慮することで, 原文書の情報をより網羅した自動要約が実現できるのではないかと考えている.

表 6. ROUGE-1 のスコア(機械加工報告書)

	GA		GA+Prep		SVM	Greedy
	TF-IDF	TF-IDEP	TF-IDF	TF-IDEP		
枝刈り有り	0.4589	0.4756	**0.5335**	**0.5464**	**0.5543**	**0.5895**
枝刈り無し	0.4319	0.4451	0.4422	0.4814	0.4860[4]	

表 7. ROUGE-1 のスコア(TSC2)

	GA		GA+Prep		SVM	Greedy
	TF-IDF	TF-IDEP	TF-IDF	TF-IDEP		
枝刈り有り	0.5501	0.5654	**0.5554**	**0.5788**	**0.5896**	**0.6832**
枝刈り無し	0.4706	0.4816	0.4788	0.4912	0.4996	

参考文献

[1] C. N. Silla et.al, Automatic text summarization with genetic algorithm-based attribute selection, in Proceedings of the IBERAMIA , pp. 305–314, Springer, 2004.

[2] J.Y. Yeh, et al., Text summarization using a trainable summarizer and latent semantic analysis, IP&M, pp. 75–95, 2005.

[3] 小倉由佳里, 小林一郎, 多目的GAを用いた複数文書要約への取り組み. 人工知能学会全国大会論文集, 28, pp.1–3, 2014.

[4] 立林裕太朗, 藤田充洋, 古谷克司, 佐々木裕, 機械加工技術文書の自動要約, 言語処理学会第20回年次大会, pp.638-641, 2014.

[5] 立林裕太朗, 藤田充洋, 古谷克司, 佐々木裕, GAによる機械加工メモの自動要約に関する予備検討, 第21回言語処理学会年次大会, pp.581-584, 2015.

[6] C.Y. Lin, E. Hovy, Automatic evaluation of summaries using n-gram co-occurrence statistics, ACL, 2003.

[7] 難波英嗣, 奥村学, 第 2 回 NTCIR ワークショップ自動要約タスク (TSC) の結果および評価法の分析, 電子情報通信学会技術研究報告, pp.45-52, 2001.

[8] Mecab: Yet another part-of-speech and morphological analyzer, http://mecab. sourceforge. net/, 2005.

[9] CaboCha: Yet Another Japanese Dependency Structure Analyzer, https://code.google.com/p/cabocha/, 2005.

[10] 増田和浩, 寺本一成, 古谷克司, 佐々木裕, SVMを用いた機械加工文書からの直接的因果関係の抽出, 第20回年次大会発表論文集, pp.4-12, 2014.

[11] C.C. Chang and C.J. Lin. LIBSVM: a library for support vector machines, 2001.