機械加工技術文書の自動要約

Automatic Summarization of Machining Technical Documents

立林 裕太朗*　藤田 充洋**　古谷 克司*　佐々木 裕*
Yutaro Tatebayashi*　Mitsuhiro Fujita**　Katsushi Furutani*　Yutaka Sasaki*
*豊田工業大学　**㈱豊田中央研究所
*Toyota Technological Institute　**Toyota Central R&D Labs., Inc.

1　研究背景

近年，コンピュータを始めとした情報技術の発展は目まぐるしく，企業や研究機関などの様々な場所で，情報技術が利用されている．その一つとして，報告書などの文書資料の電子化が挙げられる．文書資料を電子化することで，従来と比較して，大量の資料を容易に管理することが可能となった．しかし，管理できる文書数は飛躍的に向上しても，文書に含まれる情報量は変わっていないため，これらの電子文書から必要な情報を探し出すことが難しくなっている．つまり，大量の電子文書を管理することが出来るようになっても有効活用できていない．今後も，電子化された文書の量は増加すると考えられるため，ユーザが必要な情報を効率よくアクセスするための支援技術の開発が早急に求められている．

本研究は，企業や研究機関で利用できる，機械加工文書を対象にした実用レベルの自動要約システムを構築することを目的としている．具体的には，自動要約でよく行われている手法のひとつである重要個所抽出手法による機械加工文書の自動要約の評価を行った．

2　重要個所抽出手法の現状

重要個所抽出に関する既存研究の多くは，最初に重要文抽出を行い，次に抽出された文の非重要個所を削除する手法を採用している．例えば，非重要個所の削除手法として，構文木や動詞連体修飾節に基づいて削除個所を決定する方法[1][2]，平尾らの要約コーパスに基づいてSVMによる機械学習を行う手法[3]，Knightらの決定木や確率モデルを用いて非重要個所の判断知識を獲得する手法[4]があげられる．これらの手法の共通点として，要約に含まれるべき情報を選択するのは文抽出処理に任せておき，非重要個所の削除は要約をより短縮するために行うという考え方である．重要個所抽出の利点として，文全体の構成に影響のないような個所を削除することによって文法的な正しさを保存しやすいことが挙げられる．

3　本要約生成システム

本要約生成システムは，大きく分けると二つのフェーズに分けられる．それぞれ重要文抽出と文短縮と呼ばれ，それぞれのフェーズで構成要素や研究手法が異なる．この節では，それぞれのフェーズについて解説を述べる．

3.1 重要文抽出

重要文抽出とは，入力された原文書に対して，ある基準を用いて評価を行い，より重要だと判断した文を抽出して要約を作成する方法である．重要文を選択する評価尺度には，文の出現位置や節見出しなどといった様々なものが利用されているが，本研究では，文書の関心の中心となる事柄について言及する単語の出現頻度や重要度などを利用して，文の重要度を算出する手法を採用した．単語の重要度の算出には，TF-IDF法を用いた．

図１に重要文抽出フェーズの概要を示し，以下に対応する処理のプロセスについて示す．

① 名詞の抽出

TF-IDF スコアを算出するために，実験対象である文書から名詞を収集する必要がある．まず，これらの文書に対して，Mecab を用いて形態素解析を行い，品詞の同定を

行った．次に，品詞の同定結果から名詞のみを抽出した．なお，Mecab による形態素解析を行う際，機械加工特有の用語を追加したユーザ辞書を用いた．

② TF-IDF スコア算出

TF-IDF 値を計算するために，(1)実験対象の文書に含まれる全ての名詞リストと(2)実験対象文書毎に含まれる名詞リストの二種類のファイルを作成した．これらの二種類のファイルを用いて，実験対象の文書毎に含まれている名詞の全てについて TF-IDF 値を算出した．

③ 重要文判定および抽出

各文書の任意の文中に出現する名詞の TF-IDF スコアの総和を文の重要度として計算し，文書毎に文の重要度が高かった上位 2 文を重要文集合として抽出した．

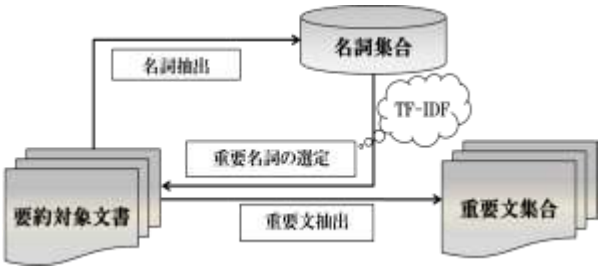


図 1 重要文抽出フェーズの概要

3.2 文短縮

文短縮は，一文毎に重要でない個所を削り，主要な情報を減らすことなく，テキストを短く表現し直す要約手法である．この手法は，段落，文を対象としていた重要文抽出とは違い，句や文字列を単位とした重要個所抽出である．文短縮の主な研究として，短い文（要約文）にノイズが挿入され長い文（原文）が観測されたと考えるモデルを仮定する方法[5]が挙げられる．この方法では，スコアリング手法を用いていたが，本研究では機械学習器として Support Vector Machine(SVM)[6]を用いた．また，多クラス識別問題に適用するために one-against-one[7]の方法を採用した． 図2に文短縮の概要を示し，以下に対応する処理のプロセスについて示す．

①準備

まず，3.1 節の文書の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，Cabocha を用いて形態素解析，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．また，形態素解析，係り受け解析を用いて，単語毎の品詞および係り受け関係の同定を行った．

② ID 付け

重要文集合および正解要約文に含まれる全ての単語およびその品詞について，3.3 節の素性 ID を割り振った．

③ 機械学習による判別

LIBSVM[8]を用いて，重要文集合に対し，文節毎にどのくらい重要かの判定を行った．

④ 文短縮

結果をもとに，短縮率70%で短縮した重要文集合を生成した．

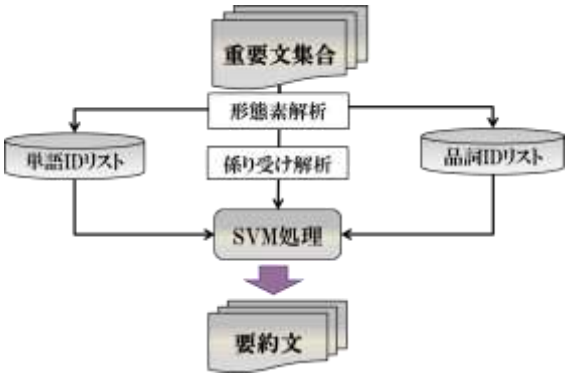


図 2 文短縮フェーズの概要

3.3 素性選択

文節中の単語に対して，実験で用いた素性を表 1 にまとめた．

表1 文節に関する素性一覧

素性	次元	内容
文末表現	1	文節は文末かどうかを示す二値
係る文節数	10	文節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
文節の TF-IDF 値	30	係り元の数

表2 文節に含まれる単語に関する素性

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

本実験では，一定の要約率を維持するために，3.1節にて出力された重要文集合について，文節に含まれる名詞および動詞が正解要約文に含むか否か，係り受け元の文節は存在するか否かについて，多値分類を通じて学習を行っている．

作成および使用した学習モデルの精度を表3に示す．

表 3 学習モデルの精度

	recall	precision	F-score
ALL	0.6094	0.6792	0.6424

4 実験

4.1 実験条件

本研究において，実験と評価の際に用いた実験データおよび正解データについては，機械加工の現場で作成された，要約付きの文書データセットを用いた．

それぞれのデータセットの仕様は以下の通りである．

表 4 実験データと正解データの仕様

データ種類	文字数	行数	大きさ(KB)
実験データ	30938	1032	82.5
正解データ	4570	191	12.3

4.2 システム評価

本研究では，BLEUスコア[9]を用いて自動評価を行った．BLEUは入力に対して，あらかじめ用意した正解要約文とシステムが出力した要約文を比較し，正解にどれだけ近い文が得られたかを評価することによってシステムの優劣を測るものである．BLEUは，1-gramからN-gramまでの適合率の重み付き和で定義されており，N-gramのマッチ率に基づく手法を用いている．

$$\mathrm{BLEU\text{-}N} = \exp\left(\sum_{1}^{N} \frac{1}{n} \times \log P_n\right)$$

$$P_n = \frac{\text{比較対象文書中と正解文書中で一致した n-gram の数}}{\text{比較対象文書中の n-gram の総数}}$$

BLEUの評価は適合率であるため，システムが出力した文が正解文に対してあまりにも短いと評価を不当に上げてしまう恐れがある．そのためペナルティが導入されることがある．しかし，要約文の評価では，短い出力文である方が高圧縮なので一般的に良いとされており，このペナルティは課さない場合が多く，本研究でも長さによるペナルティは考えないものとして評価を行った．

また，BLEUの式にある変数Nの値は一般的にはN=4であることが多いが，本研究では対象が日本語であるため，BLEUスコアが必要以上に小さくなってしまう恐れがある．そこで，N=1,2,3,4の四通りについて，それぞれ評価を行った．

以下の表5および図3に，96個の実験データ(要約対象文)，抽出した重要文(重要文集合)，文短縮された重要文集合(出力要約文)の場合について，BLEUスコアの平均を示した．

表 5 本手法を用いたことによる BLEU の変化

	原文	重要文集合	出力要約文
BLEU-1	0.1090	0.2328	0.3622
BLEU-2	0.0805	0.1585	0.2520
BLEU-3	0.0628	0.1194	0.1908
BLEU-4	0.0498	0.0944	0.1490

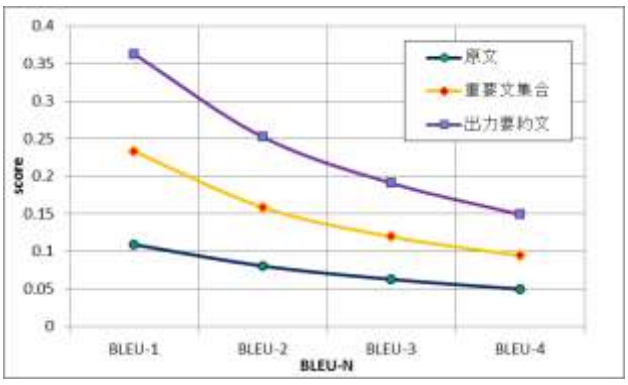
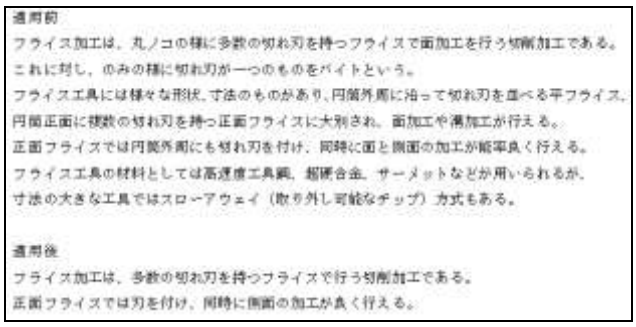


図3 本手法を用いたことによるBLEUの変化

次に，本要約システムの適用前と適用後の文書の比較例を以下に示す.


適用前
フライス加工は，丸ノコの様に多数の切れ刃を持つフライスで面加工を行う切削加工である。
これに対し，のみの様に切れ刃が一つのものをバイトという。
フライス工具には様々な形状，寸法のものがあり，円筒外周に沿って切れ刃を並べる平フライス，
円筒正面に複数の切れ刃を持つ正面フライスに大別され，面加工や溝加工が行える。
正面フライスでは円筒外周にも切れ刃を付け，同時に面と側面の加工が能率良く行える。
フライス工具の材料としては高速度工具鋼，超硬合金，サーメットなどが用いられるが，
寸法の大きな工具ではスローアウェイ（取り外し可能なチップ）方式もある。

適用後
フライス加工は，多数の切れ刃を持つフライスで行う切削加工である。
正面フライスでは刃を付け，同時に側面の加工が良く行える。

図 4 本システムによる処理結果の例

表5より，重要文抽出，文短縮という処理を経る毎に，BLEU-Nのスコアが高くなっていることが分かる．このことから本手法を用いることで，ある程度の冗長表現の除去が期待できることが分かる．図2は，フライス加工について解説した文書であるが、フライス加工を端的に何かと示した最初の文を重要文として抽出できていることが分かる．他にも，「能率」や「丸鋸の様に」などの修飾表現を除去できていることが分かる．しかし、「円筒外周」を始めとして，多くの文節を除去したために文章の本意が分からなくなってしまった事例も存在した.

5 結論

本研究を通して，TF-IDFを用いた重要文抽出，単語や係り受け関係をもとに文短縮を行う重要個所抽出方法について，機械加工文書でも一定の効果があることを確認できた．しかし，鈴木ら[10]が指摘しているように，文抽出処理で重要文と認識されなかった文中に重要な情報を持つ箇所がある場合，それらの情報は要約から欠落してしまうことや，則本ら[11]が指摘した係り受け関係と正解文に出現したキーワードだけを考慮した場合は，助詞等の文末の語を削除する手法を入れてないので、不要な部分を正しく除去できないといった従来法の問題点もみられた.

今後の課題として，さらに高品質な要約を生成するために，要約長の制約の中で，文書中の句や文節が最も自然にかつ多くの内容を保存するような組み合わせを最適化手法により生成する研究があげられる.

参考文献

[1] H. Jing: Sentence reduction for automatic text summarization, in Proc. of the 6th Applied Natural Language Processing Conference (ANLP'2000), pp. 310-315, 2000.

[2] 酒井浩之，増山繁：動詞連体修復節の省略可能性に関するコーパスからの知識獲得 (自然言語処理). 電子情報通信学会論文誌. 87(8), pp.1641-1652, 2004.

[3] 平尾努ら：Support Vector Machine を用いた重要文抽出法，情報処理学会論文誌， pp. 2230-2243, 2003.

[4] K. Knight and D. Marcu: Summarization beyond Sentence Extraction: A Probabilistic Approach to Sentence Compression, Artificial Intelligence, Vol. 139, No. 1, pp. 91–107, 2002 .

[5] K. Knight and D. Marcu: Statistics-based Summarization - Step One: Sentence Compression. in Proc. of AAAI/IAAI'00, pp. 703-710, AAAI Press / The MIT Press, 2000.

[6] V.Vapnik: Statistical Learning Theory, Wiley, New York. 1988.

[7] C.-W. Hsu and C.- J. Lin: A Comparison of Methods for Multiclass Support Vector Machines, Neural Networks, IEEE Transactions on, 13(2), pp. 415-425, 2002.

[8] LIBSVM : A Library for Support Vector Machines (http://www.csie.ntu.edu.tw/~cjlin/libsvm/)

[9] K. Papineni et al. "BLEU: A Method for Automatic Evaluation of Machine Translation, in Proc. of the 40th Annual Meeting on Association for Computational Linguistics, 2002.

[10] 鈴木大介，内海彰：Support Vector Machine を用いた文書の重要文節抽出—要約文生成に向けて—，人工知能学会論文誌，21(4)，pp. 330-339, 2006.

[11] 則本 達哉，小山 登：VOD講義のための字幕強調や短縮表示法，情報処理学会研究報告．データベース・システム研究会報告，2010-DBS-151(6), pp. 1-7, 2010.