修辞構造と係り受け構造を制約とした単一文書要約手法

菊池 悠太 † 平尾 努 § 高村 大也 ‡ 奥村 学 ‡ 永田 昌明 §
† 東京工業大学 総合理工学研究科, ‡ 東京工業大学 精密工学研究所
§NTT コミュニケーション科学基礎研究所
†kikuchi@lr.pi.titech.ac.jp, ‡{takamura,oku}@pi.titech.ac.jp,
§{hirao.tsutomu,nagata.masaaki}@lab.ntt.co.jp

1 はじめに

抽出型要約は，原文書をある言語単位の集合とみなしそこから長さ制約を満たす部分集合を抽出することで要約を生成する．近年，部分集合の選択に対する組合せ最適化の導入により要約システムの ROUGE スコアは大きく向上した [9, 2, 12]．しかし，これらの手法は原文書を抽出単位の集合とみなすので原文書が本来持つ構造を考慮できるとは限らない．この問題を解決するため Hirao らは，修辞構造理論 (Rhetorical Structure Theory: RST)[7] に基づき，文書をおおよそ節に相当する Elementary Discourse Unit(EDU) をノードとした修辞関係の係り受け木として表現し，木制約のもとでナップサック問題を解くことで原文書の大域的構造を保持した要約の生成法を提案した [5].

抽出型要約を考える上では，その抽出単位の粒度も重要となる．Hirao らの手法は EDU という，文よりも小さな抽出単位を扱うため，修辞関係に基づく大域的構造を考慮しているものの，組合せ最適化問題として解いた場合は，目的関数が最大になるよう多くの文から細かな断片を集める形になってしまい，情報の過度な断片化が生じてしまう．一方，文を抽出単位とすると文レベルの文法性が担保されるが，不要な情報も含むことになるため高圧縮な要約生成には不向きである．

本稿ではこれらの問題を解決するため，文書を文の間の係り受け木，文を単語の間の係り受け木で表現した木の入れ子構造として表現し，その木制約のもと，文書から単語を選択する単一文書要約手法を提案する．生成される要約は，文の根付き部分木，また文の中では単語の任意の部分木が抽出されたものになっている．これにより，修辞構造を保持することで原文書の構造を利用しつつ，単語の任意の部分木を抽出することで文書中から重要な箇所のみを残した要約が生成できる．

単語の係り受け木からその部分木を抽出するアプローチは文圧縮の一つとして，文抽出と併せる形で文書要約の手法に盛んに取り込まれている [13, 11, 10, 4]. このような文圧縮を要約に取り込んだ従来の研究では，抽出する部分木は根付き部分木に限定されていた．しかし根付き部分木の抽出のみでは，that 節など，文の中に内包された他の完全文のみを抽出することができない．そこで提案手法では文中の任意の動詞を根とした部分木を抽出することで，より柔軟な要約生成を可能とする．また，従来の文抽出と文圧縮を同時に解く手法は，冒頭で記した通り文抽出の際に修辞構造のような文間の大域的な構造は利用していない.

Filippova らは，文圧縮のタスクにおいて，根付き部分木に限らない部分木の抽出が可能な手法を提案しており，提案手法の文圧縮の部分と最も関連した手法であるといえる [3]．彼女らは単語の係り受け木を規則で変換することで根だけに限らずそれ以外の動詞も根の候補として，そのいずれかを根とした部分木を抽出できるようにした．ただし，文圧縮というタスクは文書要約とは異なり単一の文からそれよりも短い文を，文毎に個別の圧縮率に基づき圧縮を行う．文書要約において文書中の各文に個別の圧縮率を設定することは困難であるため，文圧縮タスクにおける彼女らの手法をそのまま要約に適用することは困難である.

2 木の入れ子構造からの要約生成

2.1 木の入れ子構造による文書の表現

本稿では，文書を文の係り受け木，各文の中では単語の係り受け木によって成る木の入れ子構造で表現する．その後，文の係り受け木からその根付き部分木，根付き部分木の各ノードの中では単語の任意の部分木が抽出されるような制約の下で，目的関数を最大化するように単語を選択する.

RST では，文書を EDU に分割し，隣り合う EDU 間に階層的に修辞関係[1]を割り当てることで，EDU を葉ノード，それらの関係を節ノードとする RST Discourse Tree(RST-DT) を構築する．Hirao らは，文書要約に RST-DT をそのまま用いることの問題点を指摘し，RST-DT を EDU 間の係り受け関係を直接表現した Dependency-based Discourse Tree(DEP-DT) へ変換した．具体的な変換方法は [5] を参照されたい．

我々は Hirao らの DEP-DT をベースにしているが，文単位の係り受け関係を用いるために DEP-DT を変換する．具体的にはある文の DEP-DT の根となっている EDU の係り先の EDU が属する文を，その文の係り先の文とみなす．これにより，EDU 単位の係り受け木から文単位の係り受け木へと変換することができる．さらに各文に対して構文解析器による単語の係り受け木を考えることで，文書を，文の係り受け関係，文の中では単語の係り受け関係を表した木の入れ子構造として表現する．図 1 に木の入れ子構造の例を示す．文レベルの係り受け木のそれぞれのノードの中に単語の係り受け木が存在している．

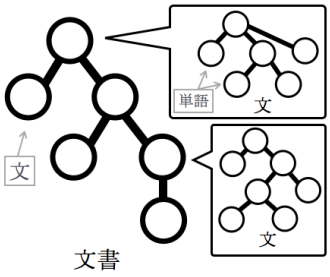

図 1: 木の入れ子構造の例

$$
\begin{array}{lrll}
\text{maximize} & \sum_{i,j} \frac{log(1+tf_{ij})}{depth(i)^2} z_{ij} & & \\
\text{subject to} & \sum_{i,j} z_{ij} \leq L; & \forall i,j & (1) \\
& x_{\text{parent}(i)} \geq x_i; & \forall i & (2) \\
& z_{\text{parent}(i,j)} - z_{ij} + r_{ij} \geq 0; & \forall i,j & (3) \\
& x_i \geq z_{ij}; & \forall i,j & (4) \\
& \sum_j z_{ij} \geq \min(\theta, \text{len}(i)) x_i; & \forall i & (5) \\
& \sum_j r_{ij} = x_i; & \forall i & (6) \\
& \sum_{j \notin R_c(i)} r_{ij} = 0; & \forall i & (7) \\
& r_{ij} \leq z_{ij}; & \forall i,j & (8) \\
& r_{ij} + z_{\text{parent}(i,j)} \leq 1; & \forall i,j & (9) \\
& r_{\text{iroot}(i)} = z_{\text{iroot}(i)}; & \forall i & (10) \\
& \sum_{j \in \text{sub}(i)} z_{ij} \geq x_i; & \forall i & (11) \\
& \sum_{j \in \text{obj}(i)} z_{ij} \geq x_i; & \forall i & (12)
\end{array}
$$

図 2: 提案手法の定式化 ($x_i, z_{ij}, r_{ij} \in \{0, 1\}$)

2.2 整数計画問題としての定式化

本節では，2.1 節で定義した文書の表現を用いた単一文書要約手法の定式化について述べる．与えられた文書から文書レベルでは文の根付き部分木，文レベルでは単語の任意の部分木になるような木制約のもとでの単語選択による要約生成を，整数計画問題により定式化する．

提案手法の定式化を図 2 に示す．決定変数 x_i, z_{ij} は，それぞれ文 i，文 i 内の単語 j(以下，単語 ij) が要約に含まれるとき 1 となる．r_{ij} は，単語 ij が，抽出される部分木の根として選択されたとき 1 となる．(1) は，選択される単語の数が L 以下になることを保証する制約式である．(2)，(3) はそれぞれ修辞構造，係り受け構造の親子関係を保持するための制約である．ある子ノードが選択された時は，その親ノードも必ず選択されることを保証する．ただし，単語の係り受けに関しては r_{ij} の導入により，単語 ij が部分木の根となる場合のみ，その親ノードを要約に選択しないことを許容する．制約式 (4)，(5) は文と単語の整合性を保つための制約で，文を選択せずに単語を選択することや，逆に文を選択しているにも関わらずその文中の単語が選択されていないという状態を防ぐ．また (5) の右辺により，ある文は高々θ 単語までしか圧縮しないことを保証する．ここで $len(i)$ は文 i の単語数を返す関数である．これには，細かい部分木を多く抽出することで情報の過剰な断片化を防ぐ狙いがある．実験時は θ を 8 とした．

制約式 (6)-(9) により，文から任意のノードを根とする単語の部分木を選択可能とする．(6) により各文からは高々一つの部分木を抽出することを保証し，(7) により，根となりうる単語を制限する．ここで $R_c(i)$ は文 i の中で根の候補となる単語集合を返す関数であり，今回は文中の動詞を根の候補とする．(8) により，ある単語を根とした場合は必ずその単語を要約に含めることを保証し，(9) により，根として選んだ単語の親の単語は要約に含めないことを保証する．また (10) により，係り受け解析器の出力した根の単語を候補として選ばない場合は，その単語を要約に含めないことを保証する．ここで $\text{root}(i)$ は，文 i において係り受け解析器が根と出力した単語のインデックスを返す関数である．(11)，(12) は，それぞれ原文に主語 (SUB)，目的語 (OBJ) があった場合，それらのうち一つ以上を圧縮後の文に含めるという制約である．ここで，$\text{sub}(i), \text{obj}(i)$ は，それぞれ引数に与えられた文 i のうち，係り受けタグが “SUB”, “OBJ” である単語インデックスの集合を返す関数である．

[1]EDU は核とその補助内容である衛星のいずれかの属性を持ち，EDU 間には 78 種類の修辞関係が定義されている．

2.3 言語的制約

本節では，圧縮後の文の文法性を担保するため，大別して二種類の追加的な制約を用意する：

$$z_{i,k} = z_{i,l}, \quad (13)$$

$$\sum_{k \in s(i,j)} z_{ik} = |s(i,j)| x_i. \quad (14)$$

(13), (14) はそれぞれある単語対，単語列が同時に選択されることを保証する制約式である．(13) に該当する単語対には，以下のものがある: 係り受けタグが “PMOD” である単語とその係り先の単語，係り受けタグが “VC” である単語とその係り先の単語，否定詞とその係り先の単語, 動詞にタグ “SUB” で係っている名詞とその動詞, 動詞にタグ “OBJ” で係っている名詞とその動詞，比較級 (JJR) または最上級 (JJS) とその係り先の単語, to とその係り先の動詞, 冠詞とそれに係る単語, %とそれに係る単語．また，(14) に該当する単語列には，以下のものがある: 固有名詞列 (品詞タグが PRP$，WP$，あるいは POS である単語列)，所有格とその係り先の名詞と，その間に含まれる単語列.

3 評価実験

3.1 実験設定

実験には RST Discourse Treebank[1] を用いた．これは，Penn Treebank のうち 385 記事に対し，RST-DT を人手で付与したコーパスである．385 記事のうち 30 記事に対して，各記事に一つずつ人間により作成された要約 (参照要約) が割り当てられている．これらはそれぞれ原文書の 10%ほどの単語数である．なお，これらの文書ははじめに Marcu らによって要約の評価に用いられた [8]．これら 30 文書について，参照要約の単語数を L として，各手法により要約を生成する．評価指標には，文書要約において広く用いられている ROUGE を用いた [6]．提案手法 (**任意部分木**) との比較として Hirao らによる edu 選択手法 [5] (**EDU 選択**) を用いる．また，提案手法の中でも，根付き部分木選択手法 (**根付き部分木**)，文選択手法 (**文選択**) との比較を行う．根付き部分木選択は，制約式 (6) の $R_c(i)$ が係り受け解析器が根とした単語 id のみを返すよう変更することで実現する．また文選択は，制約式 (5) における右辺を $len(i)x_i$ とすることで実現する．なお，本実験における文の係り受けは，全てコーパスにアノテートされた RST-DT から得た DEP-DT に基づいている．

表 1: 実験結果

	ROUGE-1
任意部分木	**0.354**
根付き部分木	0.352
文選択	0.254
edu 選択 [5]	0.321

3.2 結果と考察

3.2.1 ROUGE による比較

ROUGE-1 スコアを表 1 に示す．表を見ると，文選択手法では十分な ROUGE スコアが得られないことが分かる．ここに文圧縮を加えることで大幅な向上が見られた．そのスコアは，本データセットで最も良い ROUGE スコアを得ていた EDU 選択手法よりも高い．なお，Holm 法による多重検定の結果，提案手法である任意部分木が EDU 選択手法，文選択手法を有意に上回っていることがわかった．任意部分木と根付き部分木との差はわずかであるが，これについて 3.2.3 節で実例を示し，定性的な分析を行う．なお，ニュース記事の単一文書要約におけるベースライン手法で，しばしば高い ROUGE スコアを持つ傾向にある LEAD[2]の ROUGE スコアは 0.240 であった．

3.2.2 情報の断片化に関する考察

本節では EDU 選択において問題であった情報の過剰な断片化について分析を行う．ここで，情報の過剰な断片化とは，多くの文から小さな部分を抽出している状態を指す．すなわち，生成した要約中の文の数は，情報の断片化を測る指標とみなすことができる．要約を生成する際に利用した原文の文の数について集計すると．提案手法の中央値と平均値はそれぞれ 4 文，4.73 文であった．これに対し EDU 選択手法はそれぞれ 5 文，5.77 文であり，提案手法の方が EDU 選択よりも，要約生成に利用した文の数が有意に少ない[3]．つまり，提案手法は情報の過剰な断片化を緩和しつつ，高い ROUGE スコアを保つことができていることがわかる．要約に利用した文の数の分布を示すため，箱ひげ図を図 4 に示す.

[2]要約長に達するまで記事の先頭から文を抽出する.
[3]ウィルコクソンの符号順位検定 (有意水準 5%)

原文	:	John Kriz , a Moody ’ s vice president , {said} Boston Safe Deposit ’ s performance has been hurt this year by a mismatch in the maturities of its assets and liabilities .
根付き部分木	:	John Kriz a Moody ’ s vice president [{said}] Boston Safe Deposit ’ s performance has been hurt this year
任意部分木	:	Boston Safe Deposit ’ s performance has [been] hurt this year
原文	:	Recent surveys by Leo J . Shapiro & Associates , a market research firm in Chicago , {suggest} that Sears is having a tough time attracting shoppers because it has n ’ t yet done enough to improve service or its selection of merchandise .
根付き部分木	:	surveys [{suggest}] that Sears is having a time
任意部分木	:	Sears [is] having a tough time attracting shoppers

図 3: 二つのシステムが共通して要約に選択した文と，それぞれが抽出した部分木の例.

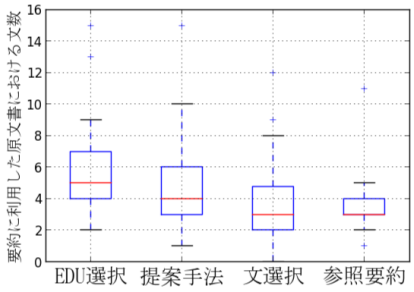

図 4: 各手法で抽出した文数の分布

3.2.3 任意の部分木による文圧縮の定性評価

単語の係り受け木から任意の部分木を抽出することの有用性を定性的に考察する．図 3 に，二つの手法が共通に選択した原文と，それぞれが出力した圧縮後の文を示す．{·} は係り受け木の根となっている単語であり，[·] は，抽出する部分木の根として提案手法が選択した単語である．これらは提案手法が根付き部分木以外の部分木を選択した例である．いずれの文も，提案手法が原文の目的節や that 節など重要かつ根付きでない部分木を抽出している．このように，文から任意の部分木を抽出することは，生成できる要約の長さが制限されている文書要約において重要な性質となる．

4 おわりに

本稿では文の修辞構造と単語の係り受け構造を利用した単一文書要約手法を提案した．提案手法は，EDU 選択手法よりも有意に少ない文数でそれよりも有意に高い ROUGE スコアを持つ要約を生成することができた．これは情報が過剰に断片化されることを緩和していると捉えることができる．ただし，これが実際に要約の可読性に与える影響については人手評価によって確かめる必要がある．また，文の係り受け木から単語の任意の部分木を抽出することについて定性的な考察を行った．今後の課題としては，前述の通り要約の可読性などは ROUGE スコアだけでは評価ができないため，今後は人手による主観的な評価も交えて提案手法の有用性を確かめたい．

参考文献

[1] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *SIGDIAL*, pages 1–10, 2001.

[2] Elena Filatova and Vasileios Hatzivassiloglou. A formal model for information selection in multi-sentence text extraction. In *COLING*, 2004.

[3] Katja Filippova and Michael Strube. Dependency tree based sentence compression. In *INLG*, pages 25–32, 2008.

[4] Dan Gillick and Benoit Favre. A scalable global model for summarization. In *ILP*, pages 10–18, 2009.

[5] Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. Single-document summarization as a tree knapsack problem. In *EMNLP*, pages 1515–1520, 2013.

[6] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, page 10, 2004.

[7] William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, pages 243–281, 1988.

[8] Daniel Marcu. Improving summarization through rhetorical parsing tuning, 1998.

[9] Ryan T. McDonald. A study of global inference algorithms in multi-document summarization. In *ECIR*, pages 557–564, 2007.

[10] Hajime Morita, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. Subtree extractive summarization via submodular maximization. In *ACL*, pages 1023–1032, 2013.

[11] Xian Qian and Yang Liu. Fast joint compression and summarization via graph cuts. In *EMNLP*, pages 1492–1502, 2013.

[12] Hiroya Takamura and Manabu Okumura. Text summarization model based on the budgeted median problem. In *CIKM*, pages 1589–1592, 2009.

[13] 奥村 学 富田 紘平, 高村 大也. 重要文抽出と文圧縮を組み合わせた新たな抽出的要約手法. *IPSJ SIG Technical Report 2009-NL-189*, pages 13–20, 2009.