断片数制御を取り入れた部分文抽出型要約

安田宜仁　　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
{yasuda.n,nishino.masaaki,hirao.tsutomu,nagata.masaaki}@lab.ntt.co.jp

概 要

本稿では, 組合せ最適化による抽出型要約を単語単位の抽出に対して適用することを検討する. 従来, 抽出型要約では文が抽出単位とされてきたが, 要約の長さに対して厳しい制限がある場合, より短い単位での抽出が望まれる. しかし, 単純に既存の重要部抽出を小さな抽出単位に適用した場合，過剰に断片化された要約が生成されてしまうおそれがある．本稿では, 短い抽出単位の抽出型要約において, 過剰な断片化を防ぐために, 組合せ最適化による定式化に対して, 断片の連なりを考慮した制約を導入する. これにより元の文内での連なりを保った部分単語列 (部分文) の抽出が可能となる. 実験では, 提案法が既存手法に比べ高い ROUGE スコアを達成することを示す.

1　はじめに

テキスト自動要約手法のうち, 抽出による要約生成 (重要部抽出) は, 要約対象の文書あるいは文書群から重要であると考えられる部分を取り出したものを結合することによって要約を生成する技術である. 重要部抽出は, 事前に定めた抽出の単位の集合であるところの文書群から, 何らかの尺度を最大化するような部分集合を取り出す問題と捉えることができ, 近年，組合せ最適化を用いたを定式化によって大幅な精度の向上が見られる [McDonald 07, Gillick 09b, Takamura 09].

重要部抽出における抽出の単位に制限はないものの, 従来, 文が抽出単位として用いられてきた. 文はそれ自体で意味をなすため, 任意の箇所から抽出した文を結合したとしても言語的に破綻することがないとう利点があり, 文を抽出単位とすることは理に適っている. しかし, 要約の長さに対して厳しい制限がある場合, たとえば携帯端末への提示を前提とした場合に, 文単位では選択の余地が限られてしまい, 適切な要約を生成するには単位が大きすぎる場合がある. たとえば, Google のスマートフォン向けの検索での検索結果スニペットの長さは現在 60 文字程度に設定されている. これに対し, 平均的な日本語の一文の長さは 40 文字程度であるため, 文抽出では 1, 2 文しか選択できないことになってしまう.

本稿では, 近年の組合せ最適化に基づく重要部抽出の進展を, 単語単位での抽出に適用することによって, 高い圧縮率が求められるような場合においても高精度な要約を生成することを検討する. しかし, 単純に従来の重要部抽出を単語に適用した場合, 過度に断片化された要約が出力されてしまうおそれがある. なぜなら, 従来の定式化ではスコアが高く, 冗長性のないような断片をばらばらに選択することが, もっとも目的関数の値を高くするからである. このような過度の断片化は, 言語的な並びや構造を保存しないため, 意味を伝えるのが困難になってしまう. 本稿ではそのような過剰に断片化された要約を避けるため, 断片数の制御を組合せ最適化問題に取り入れることを提案する. 断片数の制御は, 目的関数への断片数によるディスカウント項の導入と, 制約としての最大断片数によって行う. また, このような制御を実現するために, 整数計画問題による定式化において, 連続した単語列を取り扱う仕組みを導入する.

2　部分文抽出による要約

本節では手法の定式化について述べる. 組合せ最適化による重要部抽出の定式化の方法としてはこれまでに幾通りか提案されている．これらのうち，本稿では，高い性能が報告されているため [Takamura 09]，最大被覆問題に基づいた定式化について検討する. 最大被覆問題による定式化においては，被覆する単位としての「概念単位 (conceputal unit)」[Filatova 04] を定義する必要がある．概念単位の定義はさまざま考えられるが, 特別な辞書や解析器を必要としない一方で, 単語よりも大きな意味のまとまりを取り扱えるため, 本稿では，概念単位としてバイグラムを採用する.

最大被覆問題としての重要部抽出は, 結果の要約に含まれる概念単位を重要度を加味した上で, 制限された要約の長さ内でできるだけ多く含むような問題とみなすことができる. k 種類目のバイグラムが要約結果に含まれるかどうかを示す二値変数を b_k, その重要度に応じた重みを v_k で表すとすれば, 以下の目的関数の b_k への二値割当を最大化する問題とすることができる:

$$\sum_k v_k b_k$$

これに本稿の提案である, 選択された部分文の総数によるディスカウント項を目的関数に加える:

$$\sum_k v_k b_k - \alpha \frac{n}{|D|}$$

ここで, α は部分文の総数によるディスカウントの重みを与えるためのパラメータである. また, $|D|$ は要約対象文書群中の文の総数を, n は要約結果に含まれる部分文の数を示す.

要約に含まれる単語を示すために, $x_{j,i}$ を j 番目の文の i 番目の語に対応する二値変数とし, この語が要約に含まれる場合に $x_{j,i} = 1$, そうでない場合に 0 であるとする. 以下では上記の目的関数に対する制約について述べる.

まず, 抽出単位としての単語と, 目的関数に含まれている概念単位としてのバイグラムとを結びつける. このために 2 種類の単位を「つなぎ」のための変数 $d_{k,p}$ を用いた以下の制約を導入する:

$$\forall k : \sum_p d_{k,p} \geq b_k \quad (1)$$

$$\forall k, p : x_{\sigma(k,p)} + x_{\sigma(k,p)+1} \geq 2 d_{k,p} \quad (2)$$

ここで, $\sigma(k,p)$ は, b_k に対応するバイグラムが文書中で p 番目に出現した位置を (文番号, 文内位置) の対として返す関数である.

次に, 本稿の提案である, 部分文としての連続した単語の選択を表現する. このためにまず, 文中での部分文の開始位置と終了位置を示す二値変数 $s_{j,i}$, $e_{j,i}$ を導入する. 開始位置については, $s_{j,i} = 1$ のとき j 番目の文の i 番目の語からの部分単語列を要約として選択することを示す. 終了位置については, $e_{j,i+1} = 1$ のとき j 番目の文の選択範囲は i 番目の語で終わることを意味する. これらを用いて, 以下の 4 つの制約によって, 部分文の選択を表現する:

$$\forall j : x_{j,0} = s_{j,0} \quad (3)$$

$$\forall i, j : x_{j,i+1} - x_{j,i} + e_{j,i+1} - s_{j,i+1} = 0 \quad (4)$$

$$\forall j : e_{j,l(j+1)} = x_{j,l(j)} \quad (5)$$

maximize $\sum_k v_k b_k - \alpha \frac{n}{|D|}$,
subject to
$\forall k : \sum_p d_{k,p} \geq b_k$ (1),
$\forall k, p : x_{\sigma(k,p)} + x_{\sigma(k,p)+1} \geq 2 d_{k,p}$ (2),
$\forall j : x_{j,0} = s_{j,0}$ (3),
$\forall i, j : x_{j,i+1} - x_{j,i} + e_{j,i+1} - s_{j,i+1} = 0$ (4),
$\forall j : e_{j,l(j+1)} = x_{j,l(j)}$ (5),
$\forall j : \sum_i s_{j,i} = \sum_i e_{j,i}$ (6),
$\sum_{j,i} x_{j,i} \leq L$ (7),
$\sum_j \sum_i s_{j,i} = n \leq F$ (8),
$\forall j : \sum_i s_{j,i} \leq 1$ (9).

図 1: 提案法による組合せ最適化問題

$$\forall j : \sum_i s_{j,i} = \sum_i e_{j,i} \quad (6)$$

ここで, $l(j)$ は j 番目の文の長さを与える関数である. これにより, $x_{j,l(j)}$ は, j 番目の文の最後の単語を指す.

一般的な組合せ最適化に基づく要約同様, 要約の長さ制限に対する制約を導入する. 要約の長さ制限が単語数 L で与えられているとして以下の制約となる:

$$\sum_{j,i} x_{j,i} \leq L. \quad (7)$$

また, 本稿では長さ制限同様に最大の断片数 (部分文数) F も与えられているとし, 部分文数が F を越えないことを保証するための制約として以下を導入する:

$$\sum_j \sum_i s_{j,i} = n \leq F. \quad (8)$$

ところで, 最大被覆問題は一般には NP 困難なクラスに分類される問題であるため, 問題のサイズが大きくなった場合に解を得ることが困難となる. そこで, 現実的な時間内に解を得やすくするために解候補の制限を導入する. 具体的には, 以下の式により, 各文から選択できる部分文を 1 文につき高々1 に制限する:

$$\forall j : \sum_i s_{j,i} \leq 1. \quad (9)$$

以上をまとめると, 提案法による組合せ最適化問題は図 1 となる.

なお, 提案法は, 式 (9) までを含めた場合, 文を選択し, 選択した文について前後を切取るという操作を同時に行っているという見方もできる.

3 評価

実験条件

要約対象文書群と参照要約は，自動要約の評価型会議である Document Understanding Conference (DUC)[1]の 2005 年から 2007 年までのデータセットを用いた．評価指標には要約の自動評価指標として広く用いられている ROUGE[Lin 04] を用い，ユニグラムを計数の単位とする ROUGE-1 とバイグラムを計数の単位とする ROUGE-2 の 2 種類での再現率と F スコアで評価した．

長さ制約 L は，DUC の評価の設定に従い，250 単語とした．一方，最大部分文数の制約 F については，提案法でのみ利用されるパラメータであり，標準的な値は存在しないが，本稿のねらいである過剰な断片化を避けるという観点と，連続した単語を選択することによって意味的なまとまりを保証するという観点から，通常の文抽出での出力数を参考に設定する．DUC2005-2007 に参加した各システムの平均出力文数は約 10 であったことから，この値を参考に，$F = 10$, $F = 20$ の 2 種類とした．目的関数中の，部分文数によるディスカウントのためのパラメータ α の値は 1 とした．本実験の目的は部分文抽出の ROUGE スコアへの効果を明らかにすることであるため，パラメータの探索は行わず，上記のようにできる限り単純なパラメータ設定とした．また，本実験で用いたデータセットは質問に応じた要約生成のためのデータセットだが，上記と同様の理由から，提案法においては質問文は利用していない．

概念単位としたバイグラムの重要度は [Lin 00] に従って算出した χ^2 値の対数値と要約対象文書内での当該バイグラムの頻度との積を用いた．文分割とトークナイズには，`splitta`[Gillick 09a] を用い，ステミングには Porter アルゴリズムを用いた．

比較に用いた手法は，過去の DUC において最もスコアが高かった各手法と，現時点で最も高い ROUGE スコアが報告されている Lin らの手法 (以後 Lin11)[Lin 11] である．なお，これらはいずれも文抽出による手法である．

3.1 評価結果

評価結果を表 1 に示す．括弧内の値は各比較手法の論文内で報告されていた値である．値がやや異なるのは，参照要約に対しての前処理の差のためだと考えられる．下線のついた値は，Wilcoxon 符合順位検定 (有意水準 0.05) により有意差が認められたことを示す．

表より，提案法は $F = 20$ の場合に，DUC-05, DUC-06 それぞれの最高スコアの手法よりも高い ROUGE スコアであることが分かる．また，DUC-05, DUC-06 データセットにおいて Lin11 よりも高い ROUGE スコアであることが分かる．DUC-07 データセットの結果については，提案法が有意に上回っているとは言えない．今回提案法では何らかのトレーニングやチューニング，外部リソースの利用は行っていないのに対し，DUC-07 の peer 15 は検索エンジン由来の外部データを使った上での結果であり [Pingali 07]，また，Lin11 は DUC-05 と DUC-06 データセットをトレーニングデータとして使った上での結果であることを踏まえると，提案法は今後の余地を含めて有効であると言える．

提案法で $F = 10$ の場合の結果は，$F = 20$ の場合と比べて若干スコアが低い．これは，より多くの断片を許容することによって，高い ROUGE スコアが得られることを示唆している．ただし，高い ROUGE スコアが得られるからといって，手放しに大きな F を設定することは好ましくないと考える．なぜなら，F を大きな値に設定することによって，場合によっては過剰に断片化され，意味が通りにくい要約になってしまう可能性があるからである．

4 関連研究

重要部抽出では冗長さを避けつつ関連性の高い断片を要約に含める必要がある．しかし，各断片を逐次的に選んでいった場合，要約全体としての冗長性の排除は難しい．これに対して近年要約全体での大域的な最適化を行う組合せ最適化による定式化によって大幅な精度の向上が見られる．McDonald は重要文抽出をナップサック問題として定式化し，整数計画法 (ILP) を用いた厳密解を得る手法提案した [McDonald 07]．文を概念単位の集合として表現し，概念単位の最大被覆問題として解くモデルが Filatova らによって提案され [Filatova 04]，Takamura らによって分枝限定法による厳密解を得る方法が示された [Takamura 09]．厳密解を得る手法は計算量が大きいという問題に対して，Lin らは単調非減少劣モジュラ関数による近似を提案した [Lin 11]．ここでは劣モジュラ関数による近似は効率的に計算でき，精度についても最適解に対する誤差の保証ができることが示され，実際高い性能が得られている．これらの手法はいずれも文抽出を対象として実験が行われてきた．提案手法はより小さな抽出単位に

[1] http://duc.nist.gov/

表 1: 実験結果: ROUGE スコア (再現率と F スコア)

	DUC-05		DUC-06		DUC-07	
	再現率	F スコア	再現率	F スコア	再現率	F スコア
	ROUGE-2 スコア					
提案法 (F=20)	**8.67**	**8.51**	**10.83**	**10.71**	**12.83**	**12.59**
提案法 (F=10)	7.70	7.55	10.22	10.10	11.93	11.70
Lin11	(7.82)	(7.72)	(9.75)	(9.77)	(12.38)	(12.45)
各 DUC での最高手法	7.57(7.44)	7.47(7.43)	9.77(9.51)	9.72(9.51)	12.45(12.45)	12.24(12.29)
	システム番号: peer 15		システム番号: peer 24		システム番号: peer 15	
	ROUGE-1 scores					
提案法 (F=20)	**39.70**	**39.00**	**43.66**	**43.16**	**46.35**	**45.42**
提案法 (F=10)	38.03	37.34	42.40	41.90	45.17	44.29
各 DUC での最高手法	38.99	38.38	42.14	42.00	45.36	44.59

おいても, 上記のような組合せ最適化による進展の恩恵をより活かそうとするものである.

提案手法は見方を変えれば, 文抽出と, 抽出した文の前後を切る方法による文短縮を組合せた手法と捉えることもできる. 文抽出と文短縮を同時に行う試みとしては, [富田 09] や [Gillick 09b] によって, ひとつの整数計画問題に文短縮と重要文抽出を組み込んだモデルが提案されている. これらの手法は構文木を必要とするため外部の構文解析器を必要とするのに対し, 提案手法では何ら外部のモジュールを必要としないことが特徴である.

5 おわりに

本稿では, 組合せ最適化に基づく単語抽出による要約生成を提案した. 従来重要部抽出で使われていた文ではなく, 単語を抽出単位とすることで限られた長さ制限の中でも冗長さの少ない高精度な要約を生成することを狙っている. 単純に単語単位での抽出を行った場合の過剰な断片化という問題に対処するため, 目的関数への断片化数によるディスカウント項の導入と, 連続した単語列としての部分文の選択を導入することで, 断片数の制御を組合せ最適化問題に取り入れた.

複数年分の DUC データセットを用いた評価実験により, 提案法は特別なパラメータの調整なしに DUC-05, DUC-06 の最高スコアの手法よりも高い ROUGE-1,2 スコアを得ることができ, DUC-07 の最高スコアの手法や現在報告されている最も高いスコアの手法とも同等なスコアを得ることができた.

今後の課題を述べる. まず, 本稿では計算時間についての議論は行わなかったが, 提案法は整数計画問題の厳密解を計算しているため, 潜在的な計算コストはとても大きい. 計算時間の削減のため, 厳密解を必要としない解法について検討したい. また, 提案法は連続した単語列を選択しているとはいえ, 元の文の前後を取り除いている. このため文が保持していた言語的な構造を破壊している可能性があり, 単純に ROUGE スコアが高いからといって良い要約とはいえない. これについては人手による主観評価を含めた他の指標による評価を是非行いたい.

参考文献

[Filatova 04] Filatova, E. and Hatzivassiloglou, V.: A Formal Model for Information Selection in Multi-Sentence Sentece Extraction, in *Proceedings of the 20th COLING* (2004)

[Gillick 09a] Gillick, D.: Sentence boundary detection and the problem with the U.S., in *Proceedings of the HLT-NAACL* (2009)

[Gillick 09b] Gillick, D. and Favre, B.: A scalable global model for summarization, in *Proceedings of the Workshop on ILP for NLP* (2009)

[Lin 00] Lin, C.-y. and Hovy, E.: The Automated Acquisition of Topic Signatures for Text Summarization, in *Proc. Of the COLING Conference*, pp. 495–501 (2000)

[Lin 04] Lin, C.-y.: ROUGE: A Package for Automatic Evaluation of Summaries, in *Proc. of Workshop on Text Summarization Branches Out*, pp. 74–81 (2004)

[Lin 11] Lin, H. and Bilmes, J.: A Class of Submodular Functions for Document Summarization, in *Proceedings of the ACL/HLT* (2011)

[McDonald 07] McDonald, R.: A Study of Grobal Inference Algorithm in Multi-document Summarization, in *Proceedings of the 29th ECIR* (2007)

[Pingali 07] Pingali, P., K, R., and Varma, V.: IIIT Hyderabad at DUC 2007, in *Proceedings of DUC* (2007)

[Takamura 09] Takamura, H. and Okumura, M.: Text Summarization Model based on Maximum Coverage Problem and its Variant, in *Proceedings of the 14th EACL* (2009)

[富田 09] 富田 紘平, 高村 大也, 奥村 学：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, 情報処理学会研究報告 自然言語処理研究会報告, Vol. 2009, No. 2, pp. 13–20 (2009)