施設配置問題による文書要約のモデル化

高村 大也　奥村 学
東京工業大学 精密工学研究所
{takamura,oku}@pi.titech.ac.jp

1 序論

文書要約とは，与えられた単数あるいは複数の文書 (文書クラスタ) から，その内容を簡潔に表した短い文書 (要約) を生成する研究課題である (Mani, 2001)．文書要約の代表的な手法として，文選択によるものがある．これは，与えられた文書から必要な文を選択することにより要約を生成する手法である．この手法は，出力となる要約において，少なくとも文レベルでの文法性は保証されるという特長がある．本稿ではこの手法を取り扱う．

さて，文書要約について再考してみよう．要約は，文書クラスタの内容を表している必要がある．文選択手法ならば，“選択された文集合が，文書クラスタ内のすべての文の内容を表現している” ことが望ましい．これを簡単化し，“文書クラスタ内のすべての文が，選択された文集合内のいずれかの文によってできる限り表現される” ような要約を作るモデルを考える．このようなモデルは，施設配置問題と呼ばれる最適化問題の変種で定式化することができる．提案モデルは，多くの既存手法と異なり，文書要約での要請を直接的にモデル化したものである．また，文間の含意関係 (textual entailment) などの非対称な関係が，自然な形で取り込めるという利点もある．

本稿では，施設配置問題による文書要約モデルを提案すると同時に，提案モデルで使われる係数の算出方法の例を述べる．また，実験を行い，実際に提案モデルが高い要約性能を持つことを示す．提案モデルのさらなる改良方法についても述べる．

2 関連研究

これまで数多くの文選択要約手法が提案されてきた．ここでは代表的なもの，本研究と関連が深いものを取り上げる．

Goldstein (2000) は，文の逐次選択による文書要約を実現した．彼らは，既に選択された文と似た文に対してペナルティを与えることで要約における冗長性を排除した．

Nomoto ら (2001) は，k-means 法で文クラスタリングを行い，クラスタを用いて要約を作成した．最小記述原理によりクラスタ数を決定している．Bhandari ら (2008) は，確率的潜在意味解析を用いて文をクラスタリングし，クラスタを要約作成に用いている．

また，Filatova ら (2004) は，文を概念単位の集合で表現することにより，なるべく多くの概念単位を被覆することを目的とする最大被覆問題で文選択による文書要約を定式化した．

McDonald(2007) は，各文にスコアを与え，それらの文をナップサックに入れるか入れないかを考えるナップサック問題として文書要約を定式化し，大域解の厳密解及び近似解を求めた．

これらの既存手法と提案手法の違いについての議論は，3.4 節で行う．

3 施設配置問題による文書要約モデル

提案モデルの説明を行う．まず，要約を整数計画問題として定式化する．次に，整数計画問題の係数の計算方法について説明する．

3.1 整数計画問題としての定式化

“文書クラスタ内のすべての文が，選択された文集合内のいずれかの文によってできる限り表現される” ような要約を作るモデルは，次の整数計画問題で定式化できる．

$$
\begin{aligned}
maximize \quad & \textstyle\sum_{i,j\in I} e_{ij} z_{ij} & & \\
s.t. \quad & z_{ij} \le x_i; & \forall i,j \in I & \quad (1)\\
& \textstyle\sum_{i\in I} c_i x_i \le K; & \forall j \in I & \quad (2)\\
& \textstyle\sum_{i\in I} z_{ij} = 1; & \forall j \in I & \quad (3)\\
& z_{ii} = x_i; & \forall i \in I & \quad (4)\\
& x_i \in \{0,1\}; & \forall i \in I & \quad (5)\\
& z_{ij} \in \{0,1\}; & \forall i,j \in I & \quad (6)
\end{aligned}
$$

x_i は，文 s_i が選択されるとき 1 となり，そうでないとき 0 となるような決定変数である．また，z_{ij} は，選択された文 s_i に，文 s_j が割り当てられるときに 1 となり，そうでないとき 0 となるような決定変数である．I は，文書クラスタ内の文のインデックス集合である．$z_{ij} = 1$ のときは，文 s_i は選択されている必要があり，これは制約式 (1) により保証されている．内容的な観点から，文 s_i が文 s_j を被覆している度合いを e_{ij} で表し，これを文間係数とよぶことにする．よって，目的関数 $\sum_{i,j\in I} e_{ij} z_{ij}$ は，文書クラスタ内のすべての文が要約によって表現されている度合いを表していることになり，これを最大化することが我々の目的となる．また，c_i は文 s_i の長さを表すので，制約式 (2) は，要約長 $\sum_{i\in I} c_i x_i$ が与えられた値 K 以内であることを保証する．文の長さは，単語数あるいはバイト数で測られることが多い．また，制約式 (3) は，すべての文がいずれかの文に割り当てられることを保証する．さらに，制約式 (4) は，選択された文はその文自身に割り当てられることを意味している．

これは，組み合わせ最適化の分野で研究されている施設配置問題 (Korte and Vygen, 2002) の変種とみなすことができる．開設施設数の上限が与えられた施設配置問題は k-施設配置問題とよばれ盛んに研究されているが，本稿で扱う問題はその一般化であり，開設施設の量がナップサック型制約で与えられている施設配置問題であるといえる．ただし，乾ら (2008) が論じているように，施設の配置位置候補と利用者の位置の集合が異なりうる一般の施設配置問題と異なり，本稿で扱う問題では配置位置候補は利用者の位置の集合と同一になる．

3.2 文間係数 e_{ij} の計算方法

3.2.1 含意関係

文間係数 e_{ij} は内容的な観点から文 s_i が文 s_j を被覆している度合いを表している．内容的な観点から文 s_i が文 s_j を被覆するということは，文 s_i が文 s_j を含意していることであると換言できる．よって，提案モデルでは，文間の含意関係 (textual entailment) での成果 (Bar-Haim et al., 2006) を用いることが可能である．ここでは，Rus ら (2005) が含意関係認識においてベースラインとして用いた次の量を，文間係数として用いることにする:

$$e_{\mathrm{asy},ij} = \frac{|s_i \cap s_j|}{|s_j|}. \quad (7)$$

ここでは，s_i などを，その文が含む単語の集合とみなしている．よって，$s_i \cap s_j$ は，文 s_i と文 s_j に共通してい含まれる単語の集合を表わす．$e_{\mathrm{asy},ij}$ は，i と j について非対称である．asy は asymmetric の略であり，非対称であることを表わしている．

また，$e_{\mathrm{asy},ij}$ を対称化した

$$e_{\mathrm{sym},ij} = \frac{|s_i \cap s_j|}{|s_i \cup s_j|} \quad (8)$$

も試す．sym は symmetric の略であり，対称であることを表わしている．

より洗練された含意関係認識手法を活用することにより，さらなる性能改善が期待される (Bar-Haim et al., 2006).

3.2.2 利得

提案手法では，文書クラスタ内の各文を選択された文のいずれかに割り当てるという形で，文書クラスタ全体を被覆することを試みる．さて，文書クラスタ内の文には，被覆することが重要であると思われる文とそうでない文が存在することが考えられる．よって，各文 s_j が与える利得 b_j を考え，これを文間係数に組み込むことにより，重要なものをより強く被覆するように手法を改良する．具体的には，次のような文間係数を考える:

$$e'_{\mathrm{asy},ij} = b_j^{\beta} e_{\mathrm{asy},ij}^{1-\beta}. \quad (9)$$

同様に $e'_{\mathrm{sym},ij}$ も定義する．ただし，$e_{\mathrm{sym},ij}$ は対称だが，$e'_{\mathrm{sym},ij}$ は対称ではないことを強調しておく． によって，利得と含意関係の重みを調整する．

ここでは，利得 b_j は次のようにして計算する:

$$b_j = p\frac{1}{pos(s_j)} + (1-p)\cos(\overrightarrow{s_j}, \sum_{k\in I}\overrightarrow{s_k}). \quad (10)$$

ここで，$pos(s_j)$ は，文 s_j の文書内での出現位置 (何文目に出現するか) を表わす．出現位置の逆数を利用したのは，新聞記事要約では文書の始めの部分に重要な文が出現しやすいからである．また，$\overrightarrow{s_j}$ は，文 s_j を bag-of-words で表現したベクトルであり，第二項の $\cos(\overrightarrow{s_j}, \sum_{k\in I}\overrightarrow{s_k})$ は，$\overrightarrow{s_j}$ と $\sum_{k\in I}\overrightarrow{s_k}$ の正弦値である．これは，文書クラスタ全体に類似した文ほど重要であろうという考えを，文 s_j の単語出現分布と，文書クラスタ全体の単語出現分布との近さを用いることで表わしたものである．パラメータ p によって，出現位置の逆数と正弦値の重みを調整する．$p = 0.5$ のときは両者を同等に扱うことになり，定数倍を除いて McDonald(2007) が文のスコアとして使用した値と一致する．

3.3 部分緩和

施設配置問題においては，決定変数 x_i 及び z_{ij} に対して 0-1 整数制約が課せられている．しかし，z_{ij} は文間の割り当てを司る決定変数であり，施設配置問題の解から要約を生成にあたっては，解における z_{ij} の値は必要でない．そこで，我々は式 (6) で表わされた z_{ij} に関する制約 $z_{ij} \in \{0, 1\}$ を次のように緩和したモデルも試してみる:

$$z_{ij} \in [0, 1]; \quad \forall i, j \in I. \quad (11)$$

一般に，変数に整数制約が課せられている問題よりその緩和問題の方が高速に解けることが知られている (Hromkovič, 2003)．ここでは，一部の決定変数に対してのみ緩和が施されているので，線形計画問題にまで問題を簡単化できるわけではないが，解が高速に求められることが期待できる．

3.4 議論

文クラスタリングを用いた既存の要約手法 (Nomoto and Matsumoto, 2001; Bhandari et al., 2008) と提案手法は，どちらも文の集合を形成するという点で関連している．しかし，提案手法は，似ている文の集合を形成しているわけでなく，各選択文についてその選択文が含意する文の集合を形成しているのである．非対称である含意関係を自然な形で取り込むことができるのは，提案手法の特長である．

また，Filatova ら (2004) は，文を概念単位の集合で表現することにより，文選択による文書要約をなるべく多くの概念単位を被覆することを目的とする最大被覆問題で定式化した．Filatova らのモデルは，人間の直感にあったモデルであり，適切な概念単位を用いることができれば高い性能を発揮することが期待できる．しかし，文を概念単位の集合に分解しなくてはならないという制約は，モデルの柔軟性を損なう可能性もある．提案モデルには，そのような制約は存在しない．含意関係やカーネル関数など，文間の関係を測る既存手法を容易に取り入れることができる．ただし，Filatova らのモデルでは決定変数の数は $O(\max\{|I|, N\})$ であるが，提案手法では $O(|I|^2)$ である．ここで N は概念単位の数を表わす．よって，Filatova らのモデルの計算量は，比較的小さいという利点がある．また，2 つの文により 1 つの文の内容を被覆しているような状況は，Filatova らのモデルは適切に表現できるが，上で説明した提案モデルではこのような状況は表現できない．しかし，提案手法に拡張を施すことにより可能になる．拡張方法については，結論で触れる．

McDonald(McDonald, 2007) は，各文にスコアを与え，それらの文をナップサックに入れるか入れないかを考えるナップサック問題として文書要約を定式化し，大域解の厳密解あるいは近似解を求めた．似ている文が選ばれることを防ぐために，選択文同士の類似度の和に -1 をかけて目的関数に加えた．McDonald のモデルは文間の様々な類似度を取り入れることが可能であるが，非対称な関係を McDonald のモデルにどのように取り込むかについては明らかでない．

表 1: 施設配置問題による要約手法の ROUGE-1 値と計算時間．without はストップワードを除いた ROUGE-1 値であり，with はストップワードを含めた ROUGE-1 値である．

文間係数	p	ROUGE-1 値 without	ROUGE-1 値 with	計算時間
$e_{\mathrm{asy},ij}$	−	0.303	0.390	206.46
$e_{\mathrm{sym},ij}$	−	0.249	0.352	194.82
$e'_{\mathrm{asy},ij}$	0	**0.340**	**0.419**	396.59
$e'_{\mathrm{sym},ij}$	0	0.303	0.392	225.99
$e'_{\mathrm{asy},ij}$	0.5	**0.343**	**0.420**	316.49
$e'_{\mathrm{sym},ij}$	0.5	0.300	0.387	352.61
$e'_{\mathrm{asy},ij}$	1	0.312	0.387	665.70
$e'_{\mathrm{sym},ij}$	1	0.301	0.393	294.77

表 2: 部分緩和を行った場合の ROUGE-1 値と計算時間．without はストップワードを除いた ROUGE-1 値であり，with はストップワードを含めた ROUGE-1 値である．

文間係数	p	ROUGE-1 値 without	ROUGE-1 値 with	計算時間
$e_{\mathrm{asy},ij}$	−	0.301	0.387	47.26
$e_{\mathrm{sym},ij}$	−	0.244	0.346	31.45
$e'_{\mathrm{asy},ij}$	0	0.316	0.392	87.41
$e'_{\mathrm{sym},ij}$	0	0.290	0.374	44.91
$e'_{\mathrm{asy},ij}$	0.5	0.319	0.392	214.16
$e'_{\mathrm{sym},ij}$	0.5	0.293	0.377	83.12
$e'_{\mathrm{asy},ij}$	1	0.315	0.391	104.54
$e'_{\mathrm{sym},ij}$	1	0.288	0.376	59.65

4 実験

4.1 実験設定

実験には，DUC'04(2004) のデータを用い，DUC'04 の task 2 と同じ設定で実験を行った．これは複数文書要約タスクであり，それぞれ 10 個程度の文書から成る 50 個の文書クラスタが与えられる．各文書クラスタに対して一つの要約を作成することが求められる，要約の長さは 665 バイト以内とした．文間係数における利得と含意関係の重みを調整する　は 0.5 とし，ここでは利得と含意関係を等しく重み付けした．

提案手法において e_{ij} を計算する際に，文を単語の集合とみなす．このとき，内容語 (名詞，形容詞，動詞) であり，かつ ROUGE version 1.5.5(Lin, 2004) のストップワードリストに入っていない単語のみを用いた．また各単語は Porter(1980) の語幹抽出アルゴリズムを用い，語幹に変換した．文を bag-of-words でベクトル表現する際も，語幹に変換した内容語をベクトルの各要素とした．評価には ROUGE version 1.5.5 を用いた．特に ROUGE-1 値を用いて結果の分析を行った[1]．また，ILOG CPLEX version 11.1 を用いて整数計画問題を解いた．

実験を行ったモデルは，$e_{\mathrm{sym},ij}$，$e_{\mathrm{asy},ij}$，$e'_{\mathrm{sym},ij}$，$e'_{\mathrm{asy},ij}$ の 4 つである．$e'_{\mathrm{sym},ij}$ と $e'_{\mathrm{asy},ij}$ に関しては，p を 0，0.5，1 と変化させて実験を行った．p は利得における出現位置情報の重みであるので，$p = 0$ ならば利得は正弦値のみになり，$p = 0.5$ ならば出現位置情報と正弦値とが同等に扱われ，$p = 1$ ならば出現位置情報のみとなる．

4.2 実験結果

実験結果を表 1 に示す．この表は，各モデルについて，ストップワードを除いた ROUGE-1 値，ストップワードを含めた ROUGE-1 値，さらに一文書クラスタあたりの平均計算時間 (秒) を示す．まず，文間係数が非対称なモデルと対称なモデルを比較すると，いずれにおいても文間係数が非対称なモデルが上回っている．これは，非対称な文間関係を取り入れられる提案手法の良さを示す結果である．

また，利得を考慮していない $e_{\mathrm{asy},ij}$ や $e_{\mathrm{sym},ij}$ と利得を考慮している $e'_{\mathrm{asy},ij}$ や $e'_{\mathrm{sym},ij}$ との比較では，いずれも利得を考慮している方が良い．特に，$p = 0$ や $p = 0.5$ のときの $e'_{\mathrm{asy},ij}$ は (表 1 ではボールド体で示した)，非常に高い ROUGE-1 値を得ている．文書要約の共通タスクである DUC'04 において最も高い ROUGE-1 値を出した peer65(Conroy et al., 2004) では，ストップワードを除いた評価で 0.309 であり，ストップワードを含めた評価で 0.382 となっている．この peer65 の値と比較すると，$p = 0$ や $p = 0.5$ のときの $e'_{\mathrm{asy},ij}$ が非常に高い性能を発揮していることがわかる．$p = 0$ でも高い評価値を出していることから，文の出現位置情報無しでも提案手法はうまく機能することがわかる．

計算時間を見ると，例えば，$p = 0.5$ のときの $e'_{\mathrm{asy},ij}$ では一文書クラスタあたり 5 分程度の時間がかかっている．それ以外のモデルでも 3 分から 11 分程度の時間がかかっている．高速な文書要約が必要とされる応用においては，計算時間の短縮する技術の導入が今後求められるであろう (Shmoys, 2000)．

次に，部分緩和を行った場合の結果を表 2 に示す．この場合も，$e'_{\mathrm{asy},ij}$ は peer65 を上回る結果を出しており，提案手法が高性能であることがわかる．計算時間については，表 1 よりは全体的に短くなっており，予想通り高速化することができた．しかし，$p = 0.5$ の $e'_{\mathrm{asy},ij}$ では 3 分以上かかるなど，依然として長い時間がかかっている．ROUGE-1 値による評価は表 1 と比較すると大きく低下しているので，性能低下に見合うだけの高速化が実現できたとは言い難い．やはり，高速化を考える場合は，施設配置問題の高速化に関する既存研究を参考にし，理論的な基盤の上で手法を改良することが必要であろう．

5 結論

文選択による文書要約を施設配置問題でモデル化した．提案モデルは，すべての文が選択された文集合によりできる限り表現されるという文書要約における要請を直接的にモデル化したものである．提案モデルは，文間の含意関係などの非対称な関係を自然な形で取り込むことができる．実験により，提案モデルは非常に高い要約性能を有することを示した．

提案モデルのさらなる改良について述べる．まず，今回はベースライン的な手法を用いて文間の含意関係を表わした．しかし，より洗練された手法が数多く存在しているので (Bar-Haim et al., 2006)，それらを提案手法に組み込むことを考えている．

本稿では “文書クラスタ内のすべての文が，選択さ

[1] 用いたオプションは，ストップワードを除いた評価では，
-n 4 -x -m -2 4 -u -f A -p 0.5 -l 100 -t 0 -d -s,
ストップワードを含めた評価では，
-n 4 -x -m -2 4 -u -f A -p 0.5 -l 100 -t 0 -d .

れた文集合内のいずれかの文によってできる限り表現される” として定式化したが，例えば，選択された文集合内の 2 つの文で表現されていても構わない．これは，文 s_i と文 s_j の対に文 s_k が割り当てられるとき 1 となり，そうでないとき 0 となるような決定変数 z'_{ijk} を導入し，目的関数を $\sum_{i,j,k \in I} e'_{ijk} z'_{ijk}$ として適切に制約を設定することで実現できる．ここで e'_{ijk} は，内容的な観点から文 s_i と文 s_j を合わせた文対が文 s_k を被覆している度合いを表す．このようにして割り当て先を 1 文から 2 文に拡張できるが，割り当て先はさらに，選択された文集合の部分集合にまで拡張できる．

また，実験において述べたように，提案手法は最適化に時間がかかっている．Shmoys ら (2000) は施設配置問題の近似手法について記述しているので，これらを参考に高速化を試みるのがよいであろう．また，前述の割り当て先の拡張を行った場合，計算量が増えるので高速化は不可欠になると考えられる．

また，今回は文間係数に含意関係を考えた．しかし，これ以外にも例えば，文間のカーネル関数を用いることが考えられる．部分文字列カーネル (Bunescu and Mooney, 2006) や木構造カーネル (Collins and Duffy, 2002; Zelenko et al., 2003) など，文の間の関係を緻密に測ることができる手法が存在するので，これらを提案手法に組み込むことも考えている．

参考文献

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, pages 1–9.

Harendra Bhandari, Masashi Shimbo, Takahiko Ito, and Yuji Matsumoto. 2008. Generic text summarization using probabilistic latent semantic indexing. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP)*, pages 133–140.

Razvan C. Bunescu and Raymond J. Mooney. 2006. Subsequence kernels for relation extraction. In *Advances in Neural Information Processing Systems 18 (NIPS 2005)*, pages 171–178.

Michael Collins and Nigel Duffy. 2002. Convolution kernels for natural language. In *Advances in Neural Information Processing Systems 14 (NIPS 2001)*, volume 1, pages 625–632.

John M. Conroy, Judith D. Schlesinger, John Goldstein, and Dianne P. O’Leary. 2004. Left-brain/right-brain multi-document summarization. In *Proceedings of the Document Understanding Conference (DUC)*.

DUC. 2004. Document Understanding Conference. In *HLT/NAACL Workshop on Text Summarization*.

Elena Filatova and Vasileios Hatzivassiloglou. 2004. A formal model for information selection in multi-sentence text extraction. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pages 397–403.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *Proceedings of ANLP/NAACL Workshop on Automatic Summarization*, pages 40–48.

Juraj Hromkovič. 2003. *Algorithmics for Hard Problems*. Springer.

Bernhard Korte and Jens Vygen. 2002. *Combinatorial Optimization: Theory and Algorithms*. Springer.

Chin-Yew Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, pages 74–81.

Inderjeet Mani. 2001. *Automatic Summarization*. John Benjamins Publisher.

Ryan McDonald. 2007. A study of global inference algorithms in multi-document summarization. In *Proceedings of the 29th European Conference on Information Retrieval (ECIR)*, pages 557–564.

Tadashi Nomoto and Yuji Matsumoto. 2001. A new approach to unsupervised text summarization. In *Proceedings of the 24th ACM International Conference Research and Development in Information Retrieval (SIGIR2001)*, pages 26–34.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Vasile Rus, Art Graesser, Philip M. McCarthy, and King-Ip Lin. 2005. A study on textual entailment. In *Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI’05)*, pages 326–333, Washington, DC, USA. IEEE Computer Society.

David B. Shmoys. 2000. Approximation algorithms for facility location problems. In *Approximation Algorithms for Combinatorial Optimization (Lecture Notes In Computer Science; Vol. 1913)*, pages 369–378.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of Machine Learning Research*, 3:1083–1106.

乾孝司, 橋本泰一, 高村大也, 内海和夫, 石川正道. 2008. キーワード抽出の整数計画問題としての定式化. 情報処理学会自然言語処理研究会 (NL-188-5), pages 29–36.