識別的隠れ半マルコフモデルによるテキスト一貫性を考慮した単一文書要約

西川 仁[1] 有田一穂[2] 田中 克己[2] 平尾 努[3] 牧野 俊朗[1] 松尾 義博[1]
[1]NTT メディアインテリジェンス研究所 [2]NTT サービスエボリューション研究所
[3]NTT コミュニケーション科学基礎研究所
{ nishikawa.hitoshi, arita.kazuho, tanaka.katsumi
hirao.tsutomu, makino.toshiro, matsuo.yoshihiro } @lab.ntt.co.jp

1 はじめに

単一文書要約はナップサック問題として定式化できることが知られている [11]．したがって個々の文に要約を構成する文としての重要度を与えることができれば，動的計画ナップサックアルゴリズム [7] で最適な解が効率的に求まる．ナップサック問題を利用した要約モデルの 1 つの問題点は，文間の関係をモデルに組み込むことができない点である．一方，単一文書要約に文間の関係を組み込む際には，隠れマルコフモデル (HMM) や条件付き確率場 (CRFs) が有効であることが知られている [2, 16] ．ただしこれらの手法は，文書中の文を要約に含めるか否かという 2 値分類を系列ラベリングに基づいて行っているに過ぎず，その結果がナップサック制約を満たすとは限らない．すなわち HMM や CRFs で推定されたパラメータはナップサック制約を考慮して最適化されておらず，あくまで単なる文の系列ラベリングとして最適化されている．

本論文ではこれらの問題を解決するため，まず HMM がナップサック制約を考慮できるようにこれを拡張し，ナップサック制約を伴う HMM として自然に隠れ半マルコフモデル (HSMM) [18] が現れることをみる．さらに HSMM を識別的に訓練し，種々の特徴量を盛り込むことで，良好な要約が生成できることを示す．

2 関連研究

単一文書要約は，要約としてふさわしい文（重要文）を選択する問題として定式化できる [13] ．

McDonald は要約長内で重要文を選択する問題をナップサック問題とみなせることを示した [11] ．ナップサック問題に基づいて単一文書要約を行う場合は個々の文に対して重要度を与えればよいが，他の文との関係まで考慮して文を要約に含めるべきか否かの決定を行うことができない．そのため，一貫性に欠けた要約ができる恐れがある．

テキスト一貫性を考慮した単一文書要約の方法は大きくわけて 2 種類ある．1 つは修辞構造理論に基づく談話構造解析によって得られる，文書の木構造表現を利用するものである [10, 4] ．もう 1 つは文同士の局所的な一貫性を利用するものである [2, 16] ．前者は文書の大域的な情報を利用して要約を行うことができるという長所があるが，談話構造解析器が必要であり，その精度に要約の精度が強く依存してしまう．後者は局所的な情報を利用して要約を行うものの，談話構造解析器は不要であり，その分頑健であると期待できる．そのため，本稿では後者に焦点を当てる．

特に本稿と関連するものとして，Barzilay らによる HMM を利用した単一文書要約 [2] と Shen らによる CRFs を利用した単一文書要約がある [16]．これらの手法は HMM や CRFs によって文が要約に含まれるべきか否かを決定するものであるが，要約長を直接制御することができないという問題がある．本論文ではこの問題を克服するため，隠れ半マルコフモデルを用いて単一文書要約を行う．

3 要約モデル

3.1 隠れ半マルコフモデル

ナップサック制約が文数として与えられていれば，通常の隠れマルコフモデルでも最適な要約が可能である．10 文から 5 文を選択する問題であれば，5 個の異なる状態，すなわち文を遷移すればよい．しかし，自動要約においてはナップサック制約は単語数や文字数として与えられることが多い．例えば，100 文字の要約を作成することを考える．入力される文はそれぞれ異なる文字数であると考えられるから，文を状態としたときに個々の状態が異なる時間継続することを許す必要がある．これを許すのが隠れ半マルコフモデルであり，時間は文字数によるナップサック制約とみなすことができる．個々の状態，すなわち文は，それぞれ異なる文字数を要約の中で占めることができる．隠れ半マルコフモデルは隠れマルコフモデルと同様に出力確率，遷移確率を持つから，状態を文と考えれば，これらをそれぞれ文の重要度，ある 2 つの文同士の局所的一貫性とみなすことができる．

3.2 要約モデルの定式化

n 個の入力文 $s_1, s_2, ..., s_n$ があるとする．これらの文は長さ $\ell_1, \ell_2, ..., \ell_n$ と重要度 $w_1, w_2, ..., w_n$ を持つ．要約に含まれるべき文ほど高い重要度を持つものとする．いくつかの文は，何らかの文短縮器や言い換え器などを利用して生成できる，元の文 s_i の m_i 種類の亜種 $s_{i,1}, s_{i,2}, ..., s_{i,m_i}$ を持つものとする．これらの亜種も元の文と同様に長さ $\ell_{i,1}, \ell_{i,2}, ..., \ell_{i,m_i}$ と重要度 $w_{i,1}, w_{i,2}, ..., w_{i,m_i}$ を持つものとする．以降，記法を単純にするために，元の文 $s_1, s_2, ..., s_n$ を

$s_{1,0}, s_{2,0}, ..., s_{n,0}$ と書く．$s_{0,0}$ と $s_{n+1,0}$ をそれぞれ文書の先頭と末尾を表す記号とする．文 $s_{g,h}$ と文 $s_{i,j}$ の間に定義される局所的な一貫性のよさを $c_{g,h,i,j}$ として表す．要約は，入力文書に含まれる文の系列 g として表現される．入力文書に含まれる文の集合から作られうる系列の全体を G とする．すなわち， $g \in G$ である．最後に，求められる要約の最大長を K とする．これらの記法により，提案モデルは以下のように記述される：

$$g^* = \arg\max_{g \in G} \sum_{s_{i,j} \in sent(g)} w_{i,j} + \sum_{(s_{g,h}, s_{i,j}) \in adj(g)} c_{g,h,i,j}, \quad (1)$$

$$s.t. \sum_{s_{i,j} \in sent(g)} \ell_{i,j} \leq K. \quad (2)$$

ここで， $sent(g)$ と $adj(g)$ はそれぞれ，系列 g に含まれる，文の集合を返す関数，隣接する 2 つの文の組の集合を返す関数とする．すなわち，我々のモデルは入力された文とその亜種のなかから，長さ K のナップサック制約の下で，要約が含む文の重要度と一貫性において最良となる系列を選択するものである．

3.3 パラメータ最適化

$w_{i,j}$ と $c_{g,h,i,j}$ の値は後述する訓練事例を用いて Passive-Aggressive Algorithm [3] にて定めた．このとき，損失関数には ROUGE [9] を用いた．

文の特徴量には，文が含む内容語[1]の文書内頻度の合計，単語の表記と品詞，固有表現およびそのクラス，文の長さ，文の絶対位置と相対位置を用いた．

テキスト一貫性のための文間の特徴量には，Lapata によって提案された 2 つの隣接する文に含まれる単語の対 [8]，Pitler らによって提案された 2 つの文の類似度 [15]， Barzilay らによって提案された Entity Grid [1] を用いた．

3.4 文の亜種の生成

元の文を書き換えた亜種を生成し，これを要約に利用する．以下の 2 種類の方法で亜種の生成を行う．

- 文中の括弧とその中身の除去する．
- 係り受け木の枝刈りによる文短縮を行う．これは野本による方法 [14] と基本的には同一だが，必須格の脱落を防ぐために，動詞とその必須格について格納した辞書を用いた．

今回用いるコーパスに含まれる参照要約を分析したところ，参照要約に含まれる文の約半分は元の入力文書に含まれるそれと同一であった．また，書き換えが含まれている文の多くも，元の文との編集距離を見る限り書き換えはわずかであった．したがって上述したような比較的単純な方法で原文に対する亜種を生成しておけば，参照要約に近い要約を生成できると期待した．

4 デコード

デコードは動的計画ナップサックアルゴリズム [7] を拡張することで行うことができる．デコードは 2 段階にわかれる．最初の段階では漸化式を繰り返し解くことである時点における最適解を求める．次の段階では最後の時点の最適解からポインタをたどって計算過程をバック・トレースすることで，最良の文の組み合わせを得る．

[1]名詞，動詞および形容詞．

アルゴリズム 1 漸化式を解く

```
 1: x = ⟨x_0, ..., x_{n+1}⟩
 2: for i = 0 to n + 1 do
 3:     x_i = −1
 4:     V[0][i] ← −1
 5:     P[0][i] ← −1
 6:     S[0][i] ← 0
 7: V[0][0] = 0
 8: for k = 1 to K do
 9:     for i = 1 to n do
10:         V[k][i] ← V[k − 1][i]
11:         P[k][i] ← P[k − 1][i]
12:         S[k][i] ← S[k − 1][i]
13:         for v = 0 to m_i do
14:             if ℓ_{i,v} ≤ k then
15:                 for h = 0 to i − 1 do
16:                     u = V[k − ℓ_{i,v}][h]
17:                     if u ≠ −1 ∧ S[k−ℓ_{i,v}][h]+w_{i,v}+c_{h,u,i,v} ≥
                        S[k][i] then
18:                         V[k][i] ← v
19:                         P[k][i] ← h
20:                         S[k][i] ← S[k − ℓ_{i,v}][h] + w_{i,v} + c_{h,u,i,v}
21: V[K + 1][n + 1] ← 0
22: P[K + 1][n + 1] ← 0
23: S[K + 1][n + 1] ← 0
24: for h = 1 to n do
25:     u = V[K][h]
26:     if S[K][h] + c_{h,u,n+1,0} ≥ S[K + 1][n + 1] then
27:         P[K + 1][n + 1] ← h
28:         S[K + 1][n + 1] ← S[K][h] + c_{h,u,n+1,0}
```

最初の段階の擬似コードをアルゴリズム 1 に示す．1 行目から 7 行目では変数の初期化を行う．ベクトル $\mathbf{x} = \langle x_0, ..., x_{n+1} \rangle$ は文とその亜種に関する決定変数を格納する．$x_i = j$ であれば， $s_{i,j}$ が要約に含まれ，$x_i = -1$ であれば s_i およびその亜種は要約には含まれない．2 次元の配列 V ， P および S は計算の途中過程を記録するのに用いられる．V はある時点でどの亜種が要約に使われたか，あるいは使われなかったかを記録する．P はある文が要約に含まれたときどの文と接続されるかを記録する．S はある時点での解の目的関数値を記録する．

8 行目から 36 行目は，以下の漸化式を段階的に解くものである：

$$S[k][i] = \begin{cases} \max_{h=0...i-1, v=0...m_i} S[k - \ell_{i,v}][h] \\ \qquad + w_{i,v} + c_{h,V[k-\ell_{i,v}][h],i,v} & \text{(A)}, \\ S[k-1][i] & \text{(B)}. \end{cases} \quad (3)$$

ここで，(A) は $V[k - \ell_{i,v}][h] \neq -1 \wedge \ell_{i,v} \leq k \wedge S[k - 1][i] \leq S[k - \ell_{i,v}][h] + w_{i,v} + c_{h,V[k-\ell_{i,v}][h],i,v}$ の場合で， (B) はそれ以外の場合である．この漸化式は，ある時点 k, i において，文 $s_{i,j}$ を挿入するだけの文字数が要約にまだ残されており，かつ挿入しなかった場合より要約の目的関数値が増加する場合のみ， $s_{i,j}$ を要約に挿入するということを意味している．

次の段階は，最後の時点の最適解からバック・トレースを行うことで，最良の文の組み合わせを得るものである．擬似コードをアルゴリズム 2 に示す．

アルゴリズム 2 バック・トレース

$k \leftarrow K + 1$
$i \leftarrow n + 1$
while $i \neq 0$ **do**
  $v \leftarrow V[k][i]$
  $x_i \leftarrow v$
  $j \leftarrow k$
  $k \leftarrow k - \ell_{i,v}$
  $i \leftarrow P[j][i]$
$x_0 \leftarrow 0$
return x

	Document	Reference
Avg. # of characters	476.2	142.0
Avg. # of words	298.6	88.3
Avg. # of sentences	9.7	2.9

表 1: コーパスの統計量.

5 実験

5.1 データ

実験のため，新聞記事とその要約を 12,748 対用意した．2001 年および 2002 年の Document Understanding Conference (DUC) [2] で利用された単一文書要約データは合わせて 60 対であり，第 1 回と第 2 回の Text Summarization Challenge (TSC) [3] で利用された単一文書要約データはそれぞれ 180 対と 60 対であった．これらのデータの規模と比べて，本実験のために用意されたデータの規模は 2 桁以上大きく，そのためより信頼のおける定量的評価が可能である．全ての参照要約は 150 文字以内で書かれている．コーパスの統計量を表 1 に示す．表の示すように，本タスクは概ね元の記事を 3 分の 1 の長さに要約するタスクとなる．

5.2 評価尺度

要約の内容性の評価には ROUGE [9] を用いた．合わせて，参照要約をどれくらい模倣できたかを評価するために，参照要約と完全に同一の要約が生成された割合を調査した．

言語的品質の評価には米国国立標準技術研究所による評価尺度 [12] を用い，人手によって評価した．単一文書の要約を対象とするため，文法性，照応関係の明瞭さ，構造・一貫性，全体の 4 つの観点から評価を行った．評価のため，各手法によって生成された要約からそれぞれ 100 文章を抽出し，7 人の評価者でこれを評価した．評価者に著者らは含まれていない．

5.3 比較手法

我々は以下の 8 手法を比較した : RANDOM（ランダムに文を選択），LEAD（先頭から 150 文字を切り出す），KP（ナップサック問題に基づく方法，文重要度は tf-idf で計算），KP(S)（KP の文重要度を教師で推定），CRFs（KP の文重要度を CRFs で推定），HSMM（提案手法），HSMM(C)（文の亜種も含めた HSMM）.

学習の際には 10 分割交差検定を実施した．検定にはウィルコクソンの符号付き順位検定 [17] を用いた．

[2]http://duc.nist.gov/
[3]http://lr-www.pi.titech.ac.jp/tsc/

Method	R-1	R-2	Idt.
RANDOM	0.417	0.291	1.2%
LEAD	$0.779^{\mathrm{C,S,U,R}}$	$0.727^{\mathrm{C,S,U,R}}$	4.4%
KP	0.704^{R}	0.611^{R}	9.3%
KP(S)	$0.729^{\mathrm{U,R}}$	$0.647^{\mathrm{U,R}}$	10.4%
CRFs	$0.741^{\mathrm{U,R}}$	$0.675^{\mathrm{S,U,R}}$	11.3%
HSMM	$0.769^{\mathrm{C,S,U,R}}$	$0.703^{\mathrm{C,S,U,R}}$	15.2%
HSMM(C)	$0.785^{\mathrm{C,S,U,R}}$	$0.722^{\mathrm{C,S,U,R}}$	20.4%

表 2: ROUGE による評価の結果．R-1 と R-2 はそれぞれ ROUGE-1 と ROUGE-2 の結果である．Idt. は参照要約と完全に同一であった要約の比率である．表中の $^{\mathrm{C,S,U,L,R}}$ はそれぞれ CRFs ， KP(S), KP, LEAD, RANDOM に対する統計的有意差をあらわす.

Method	Gram.	Ref.	S./C.	Overall
LEAD	1.9	3.9	2.5	2.1
KP	4.1^{L}	3.7	3.4	3.5
KP(S)	4.2^{L}	3.6	3.5	3.6^{L}
CRFs	4.1^{L}	3.9	3.7^{L}	3.6^{L}
HSMM	4.3^{L}	4.0	4.1^{L}	4.0^{L}
HSMM(C)	4.0^{L}	3.9	4.0^{L}	3.9^{L}
HUMAN	4.7^{L}	4.5	4.7^{L}	4.8^{L}

表 3: 言語的品質の評価の結果．Gram. は文法性， Ref. は照応関係の明瞭さ， S./C. が構造・一貫性， Overall が全体の評価をそれぞれ示す．値域は 1 から 5 までであり，1 がもっとも悪く，5 がもっともよい．統計的有意差については表 2 を用いて示した.

多重比較となるため，有意水準は $\alpha = 0.05$ とし，ホルム法 [6] にてそれを調整した.

6 結果と考察

まず ROUGE による内容性評価の結果について述べる．結果を表 2 に示す．我々の提案する手法は LEAD を除く他の全ての手法に有意に勝っている．また，文の亜種も考慮した手法 HSMM(C) が出力した要約の約 20%は人間が作成した要約と全く同一であり，提案手法が人間の要約をうまく模倣できていることを示している．LEAD が非常に良好な性能を示している理由は，新聞記事においては重要な情報はまず冒頭に記述されるため，先頭から 150 文字を切り出すという単純な手法が有効に働くためである．一方で，人間と同等の要約を作成する能力は低く， LEAD の出力のうち，人間が作成した要約と同一の要約は 4%程度でしかなかった．CRFs に対しても我々の手法は有意に優越しており，これは学習の際にナップサック制約を考慮してパラメータを最適化することの有効性を示している．HSMM と HSMM(C) をみると，文の亜種を要約に組み入れることで，参照要約と全く同一の要約の割合が 5 ポイント程度増加しており，書き換えが参照要約を模倣するためにうまく働いていることを示している．

訓練事例数が要約の ROUGE 値に及ぼす影響を調べるため，原文書と参照要約 2,748 対を固定し，訓練事例を 100 対，250 対，500 対，750 対，1,000 対，2,500 対，5,000 対，7,500 対，10,000 対と変化させた．図 1 に HSMM の学習曲線を示す．図 1 の学習曲線を見る

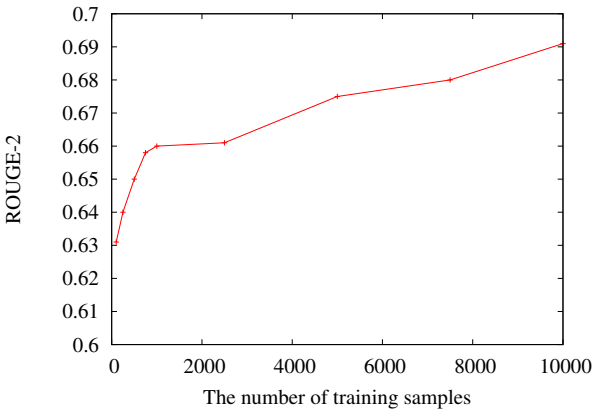

図 1: HSMM の学習曲線.

限り，訓練事例の追加が ROUGE 値の改善につながっている．学習曲線はまだ収束していないように観察されるため，さらに訓練事例を用意することができればROUGE 値の更なる改善が期待できる．Filippova は文短縮において大規模な訓練事例の効果を示した [5]．本稿は大規模な単一文書要約コーパスを訓練に利用した結果を報告する初めてのものであり，その結果は，単一文書要約においても大規模な訓練事例が出力を改善することを示したといえる．

次に，言語的品質の評価について述べる．言語的品質の評価においては人間によって作成された要約も評価の対象とし，これを HUMAN として示す．内容性と同様に，提案手法が最もよい評価を得た．ROUGE では文の亜種を含む要約が最も高い評価を得ていたが，言語的品質では亜種のないものが最もよい．これは，文短縮によって文法的でない文が生成される場合があるからであり，実際に HSMM(C) の文法性は HSMM に比べ 0.3 ポイント低い．提案手法の強みは構造・一貫性にあり，他の手法を比べこの点で高い評価を得ている．これは一貫性を考慮した特徴量がうまく働いたことを示している．一方，ROUGE で高い性能を示した LEAD は言語的品質では劣った評価を得ている．これは先頭から 150 文字だけ切り出すという方法は往々にして文の断片を要約の末尾に残してしまい，これが言語的品質の評価において問題となるからである．

総合すると，既存の単一文書要約手法に対して提案手法は内容性，言語的品質のいずれの点においても優越しており，単一文書要約手法として高い性能を持つ．

7 おわりに

本稿では，隠れ半マルコフモデルに基づく新しい要約モデルを提案した．隠れ半マルコフモデルはナップサック問題に基づく要約モデルの自然な拡張とみなすことができる．我々のモデルは文間の一貫性を自然に捉えることができ，この性質は高い内容性，言語的品質を両立させるために重要である．我々の提案した動的計画法に基づくアルゴリズムは効率的に最適解を求めることができる．大規模単一文書要約コーパスを用いた実験では，我々の手法がベースラインを上回る性能を持つことを示した．また，大規模単一文書要約コーパスが要約の品質の改善に寄与することも示した．

今後，より厳しい要約率を求められるタスクに本稿で提案した手法を用いることを予定している．そのためにはより洗練された文の書き換え，例えば言い換えなどが必要になるだろう．

謝辞

本稿で用いたコーパスは株式会社毎日新聞社の所有物であり，本稿で述べた実験のために NTT メディアインテリジェンス研究所に貸与されたものである．ご厚意に衷心より御礼申し上げる．

参考文献

[1] Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, Vol. 34, No. 1, pp. 1–34, 2008.

[2] Regina Barzilay and Lillian Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *HLT-NAACL 2004: Main Proceedings*, pp. 113–120, 2004.

[3] Koby Crammer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, Vol. 7, No. Mar, pp. 551–585, 2006.

[4] Hal Daume, III and Daniel Marcu. A noisy-channel model for document compression. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 449–456, 2002.

[5] Katja Filippova. Overcoming the lack of parallel data in sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1481–1491, 2013.

[6] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, Vol. 6, No. 2, pp. 65–70, 1979.

[7] Bernhard Korte and Jens Vygen. *Combinatorial Optimization*. Springer-Verlag, third edition, 2008.

[8] Mirella Lapata. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 545–552, 2003.

[9] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of ACL Workshop Text Summarization Branches Out*, pp. 74–81, 2004.

[10] Daniel Marcu. From discourse structure to text summaries. In *Proceedings of ACL/EACL 1997 Summarization Workshop*, pp. 82–88, 1997.

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

[12] National Institute of Standards and Technology. The linguistic quality questions, 2007. `http://www-nlpir.nist.gov/projects/duc/duc2007/quality-questions.txt`.

[13] Ani Nenkova and Kathleen McKeown. *Automatic Summarization*. Now Publishers, 2011.

[14] Tadashi Nomoto. A generic sentence trimmer with crfs. In *Proceedings of the 46th Annual Conference of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pp. 299–307, 2008.

[15] Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 186–195, 2008.

[16] Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. Document summarization using conditional random fields. In *Proceedings of the 20th international joint conference on Artifical intelligence (IJCAI)*, pp. 2862–2867, 2007.

[17] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, Vol. 1, No. 6, pp. 80–83, 1945.

[18] Shun-Zheng Yu. Hidden semi-markov models. *Artificial Intelligence*, Vol. 174, No. 2, pp. 215–243, 2010.