1J1-2in2

文書の潜在情報と表層情報を考慮したタイムライン要約への取り組み

A Study on Time-line Summarization Using the Latent and Surface Information of Documents

柏井 香里 *1　小林 一郎 *2
Kaori Kashiwai　Ichiro Kobayashi

*1お茶の水女子大学大学院　*2お茶の水女子大学基幹研究院
Ochanomizu University　Ochanomizu University

The time-series documents such as periodicals, newspapers and so on catch up new information and take it to the articles day by day. It will take much time for readers to grasp the contents of the documents because of quite a few information in themselves. So, a method to summarize the documents from multiple news resources along the timeline should be necessary. In this study, we propose a method to generate a summary of time-series documents provided from multiple news resources by extracting essential sentences from the documents along the timeline, focusing on new information sequentially added along the timeline.

1. はじめに

ニュースや新聞記事といった時系列文書は時々刻々と新しい情報が追加されていく．そのような文書の全てを読んで理解することは膨大な時間がかかってしまい現実的ではない．複数の情報源からの文書を要約し，時間の経過とともにその内容を把握できる要約手法が望まれる．本研究ではそのことを踏まえて，複数の新聞社による長期にわたる記事を一つにまとめながら，数日前には無かった新しく追加された情報に重きを置いた要約文を時系列順に生成する手法を提案する．

2. 時系列文書要約

本研究では，上述した時系列文書要約とグラフを用いた文書要約のそれぞれの手法を踏まえた時系列複数文書要約手法を提案する．提案手法の概要を図 1 に示す．

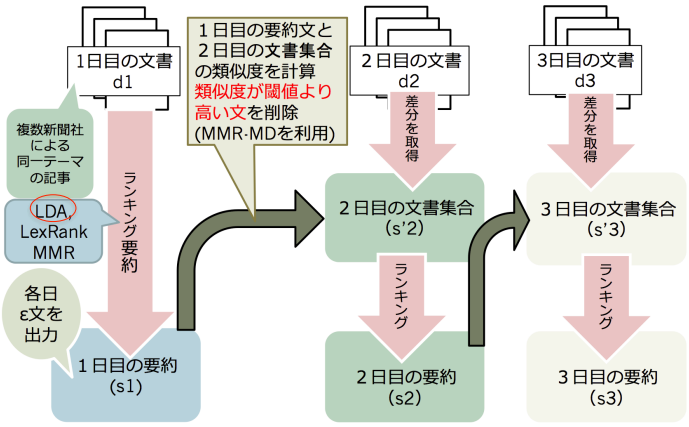

図 1: 提案手法の概要

図 1 には 1 日前まで遡った時の，3 日目までの要約の流れを示してある．複数の新聞社による記事を入力とし，各日毎の要約文を出力する．

連絡先: 柏井香里，お茶の水女子大学，〒 112-8610 東京都文京区大塚 2-1-1 理学部 3 号館 506 小林研学生室，g1220515@is.ocha.ac.jp

2.1 要約の流れ

本研究では，各文の重要度を決定するためにグラフ構造を用いる．まず，文書集合 $D_t \in \boldsymbol{D}$ について考える．t は時刻単位を表し，t=$\{ 1,\cdots,T \}$ である．ここで，D_t は時刻 t に属する文書集合を表す．本研究では，時間が経過するとともに新しく文書が追加されることを想定する．入力として，$\boldsymbol{D}$，$\boldsymbol{S}$，n，ϵ，α を与える．ここで，$\boldsymbol{S}$ は出力する要約の候補となる文集合，α は前日の要約文と当日の文との類似度の閾値，n は遡る日数であり，ϵ は要約として出力する文の数である．文集合 $\boldsymbol{S}_t$ に含まれる文で構成されるグラフを考える．文のランキングアルゴリズムに [5] で提案される LexRank アルゴリズムを用いた．本研究では，閾値による枝刈りを行わない重みなしグラフを適用した．

2.2 潜在情報を用いた要約

文の類似度を決定する際に，単語の表層的一致と潜在的意味の一致を考え，潜在的意味の抽出には Latent Dirichlet Allocation(LDA)[1] を用いる．LDA は，Blei ら [1] によって提案された文書中の単語のトピックを確率的に求める手法であり，文書中の単語は潜在的なトピックを持ち，同一トピックの単語は同一文書に出現しやすいと考え，そのトピックを教師なしで推定することができる．LDA を使用する際にはこれを応用し，本来なら文書単位で確率を求めているものを文単位でトピックを推定し，文の潜在的意味の類似度を測る事を可能にしている．また，文の潜在的意味と表層的意味どちらも考慮するために，潜在的意味と表層的意味の割合を 0 から 1 までの間の値で変化させる．グラフを用いた文書要約に潜在情報を取り入れたアルゴリズム（手法 A）を Algorithm1 に示す．

2.3 表層情報を用いた要約

この手法では文の重要度を計算する際に，単語の特徴量を考える．単語の特徴量は tf-idf によって計算し，tf-idf のスコアの上位 n 単語を文書を特徴付ける重要単語とし，重要単語が多く含まれる文ほど重要とした．本手法（手法 B）での手順を Algorithm2 に示す．入力は手法 A と同じものに加え，上位 n 単語を決定する n を与える．

Algorithm 1 要約のプロセス（手法 A：LexRank+潜在情報）

Input: $\boldsymbol{D}$, n, $\boldsymbol{S}$, ϵ, α, l
$\boldsymbol{S}$ = { }
$n \leftarrow$ past n days
$\epsilon \leftarrow$ threshold1
$\alpha \leftarrow$ threshold2
for t = 0 to T **do**
  if t=0**then**
    $\boldsymbol{S}_t \leftarrow D_t$
  else
    $\boldsymbol{S}_t$ = [ ]
    for d to $|D_t|$ **do**
      for k to n **do**
        for s to $|S_{t-k}|$ **do**
          if similarity(d, s) $< \alpha$ **then**
            $\boldsymbol{S}_t \leftarrow d$
          end if
        end for
      end for
    end for
    ranking $\boldsymbol{S}_t$ with LexRank and topic (latent) information
    if length of $\boldsymbol{S}_t > \epsilon$ **then**
      $\boldsymbol{S}'_t \leftarrow$ top ϵ sentences of $\boldsymbol{S}_t$ through MMR'
    else
      $\boldsymbol{S}'_t \leftarrow \boldsymbol{S}_t$
    end if
  end if
  $\boldsymbol{S} \leftarrow \boldsymbol{S}'_t$
end for
return $\boldsymbol{S}$

Algorithm 2 要約のプロセス（手法 B：tf-idf による表層情報）

Input: $\boldsymbol{D}$, $\boldsymbol{S}$, ϵ, α, n, l
$\boldsymbol{S}$ = { }
top_w = { }
$\epsilon \leftarrow$ threshold1
$\alpha \leftarrow$ threshold2
for t = 0 to T **do**
  if t=0**then**
    $\boldsymbol{S}_t \leftarrow D_t$
  else
    $\boldsymbol{S}_t$ = { }
    for d to $|D_t|$ **do**
      for s to $|S_{t-1}|$ **do**
        if similarity(d, s) $< \alpha$ **then**
          $\boldsymbol{S}_t \leftarrow d$
        end if
      end for
    end for
    $top_w \leftarrow$ top n word by tf-idf $\boldsymbol{S}_t$
    for s to $|S_t|$ **do**
      if top_w in s
        $\boldsymbol{score}_s$ += $\boldsymbol{score}_{topw}$
      end if
    ranking $\boldsymbol{S}_t$ by $\boldsymbol{score}$
    $\boldsymbol{S}'_t \leftarrow$ top ϵ sentences of $\boldsymbol{S}_t$
  $\boldsymbol{S} \leftarrow \boldsymbol{S}'_t$
end for
return $\boldsymbol{S}$

3. 実験

3.1 実験設定

使用したデータ，正解データなど実験に関する設定を記載する．対象データには，Tran ら [2] が提供しているタイムライン要約のためのデータセットを用いた．これらは，複数のニュース源から集められた 9 つのトピックに属している新聞記事である．本研究では 9 つのうち 6 つのトピックに関する記事を用いた．表 1 に用いたデータセットの詳細を示す．

表 1: ニュース資源

トピック	ニュース源	文書数	正解の文数
BP Oil Spill	BBC	293	98
BP Oil Spill	Foxnews	286	52
BP Oil Spill	Guardian	288	307
BP Oil Spill	Reuters	298	30
BP Oil Spill	Washingtonpost	296	19
H1N1 Influenza	BBC	122	40
H1N1 Influenza	Guardian	76	34
H1N1 Influenza	Reuters	207	23
Finiancial Crisis	WP	298	520
Haiti Earthquake	BBC	296	86
Iraq War	Guardian	344	410
Egyptian Protest	CNN	273	55

手法 A を用いた実験 1∼5 と，手法 B を用いた実験 6∼8 を行う．各実験の出力文数の詳しい設定は表 2 に示す．

実験 1： 単語の一致による表層的意味のみを利用し，LexRank によりランキングをし (手法 A)，出力文数は総文数に比例する．

実験 2： 表層的意味と LDA[1] による潜在的意味を半分ずつ利用し，LexRank によりランキングをし (手法 A)，出力文数は総文数に比例する．

実験 3： 表層的意味 0.2，潜在的意味 0.8 の割合で利用，LexRank によりランキングをし (手法 A)，出力文数は総文数に比例する．

実験 4： 表層的意味と LDA による潜在的意味を半分ずつ利用し，LexRank によりランキングをし (手法 A)，出力文数は総単語数に比例する．

実験 5： 表層的意味と LDA による潜在的意味を半分ずつ合わせて利用し，LexRank によりランキングをし (手法 A)，出力文数は実験 4 の半分とする．

実験 6： tf-idf による上位 1 単語の特徴量を利用し (手法 B)，出力文数は総文数に比例する．

実験 7： tf-idf による上位 3 単語の特徴量を利用し (手法 B)，出力文数は総文数に比例する．

実験 8： tf-idf による上位 5 単語の特徴量を利用し (手法 B)，出力文数は総文数に比例する．

実験 9： 手法 A において，表層的意味と潜在的意味の割合を 0.1∼1.0 まで 0.1 単位で変化させる，

また, 前処理として 'a' や 'the' といったストップワードの除去と, ステミング処理を行った. ステミングには Porter の アルゴリズム [3] を用いる.

3.2 評価手法

評価には ROUGE[4] を用い，各新聞社の人手で作成された正解要約をすべて正解データとし，その単語の種類を作成した要約文と比較し単語の一致を見ることで精度と再現率と F 値を計算する．各日毎にそれらの指標とする値を計算し，平均を取ることで全体の要約の性能とした．

表 2: 出力文数

実験 1 ～ 3，6~8		実験 4		実験 5	
元データの総文数	出力文数	元データの総文数	出力文数	元データの総文数	出力文数
1 ～ 100	2 文	1~1000	2 文	1~1000	1 文
101 ～ 500	4 文	1001 ~2000	4 文	1001 ～ 2000	2 文
501 ～ 1000	総文数÷ 100	2001 ～ 5000	総単語数÷ 500	2001 ～ 5000	総単語数÷ 1000
それ以上	10 文	それ以上	10 文	それ以上	10 文

表 3: 閾値 0.5 で生成された時系列の要約文書 (Haiti Earthquake)

2010-01-16
Did you complete the donation via another method ? “ Emergency stocks were distributed pretty much straight away . Please note that if your comments are published , your name and location may also be published . A special televised appeal for the DEC was shown on Friday night on BBC One and ITV1 .
2010-01-17
“ Nearly every house was destroyed here . The Pan American Health Organization put the death toll at 50,000-100 ,000 , while Haitian Prime Minister Jean-Max Bellerive said 100,000 “ would seem a minimum ” . We were tossed around incredibly violently , buildings were falling down around us . The US Southern Command ’s Lt-Gen Ken Keen said that while streets were largely calm there had been an increase in violence .
2010-01-18
That ’s a good thing . No.40 , Jean Carlos . I ’d also recommend you look into Lindblom ’s work , who explains rather lucidely the fact that business has a disproportionate influence in public decisionmaking . More than that , however , this is money that is Haiti ’s own . Surely M. Kouchner ’s past makes it all the more understandable that he should feel exasperated if he perceives the US are not letting aid through ? Do you think Africa ’s response is adequate ? At the moment though there simply is n’t really anywhere to go . I suspect a dilemma here for the EU . “ I suspect a dilemma here for the EU . : -RRB- “ Petty remark , does n’t actually disprove my point . ”
2010-01-19
Thus , earthquakes , tsunamis . Ben Brown reports from Port-au-Prince . Haiti was fully within his control . And the leading US general in Haiti , Lt Gen Ken Keen , said there was currently less violence in Port-au-Prince - already a troubled city - than there had been before the earthquake . “ We are worried sick . The Italians are supporting two medical non-governmental organizations and 70 volunteers who are fast running out of medical supplies .

3.3 実験結果と考察

表 4: 実験結果

	精度	再現率	F 値
LexRank	0.72	0.13	0.22
実験 1	0.65	0.29	0.30
実験 2	0.73	**0.31**	**0.38**
実験 3	0.73	**0.31**	0.37
実験 4	**0.83**	0.22	0.31
実験 5	0.73	**0.31**	**0.38**
実験 6	0.75	0.24	0.33
実験 7	0.77	0.20	0.30
実験 8	0.78	0.20	0.28

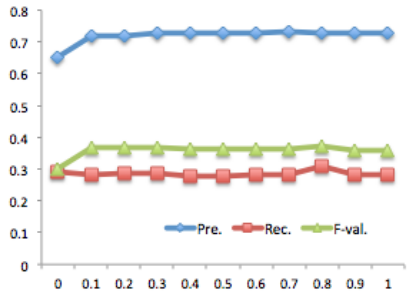

図 2: 表層情報と潜在情報の比率に対する精度

表 3 に実験 1 で出力された HaitiEarthquake の要約結果を示す．実験 1~8 の結果を表 4 に示す．既存の手法である LexRank のみを使った場合と比較して，実験 1~9 はすべて性能が上回った．精度が高く再現率が低くなったのは，異なる新聞社の複数の要約文を正解データとして利用したので，正解要約とするものが本来あるべき要約文よりもサイズが大きくなっているためだと考えられる．表層的意味と潜在的意味の割合を比較すると，実験 1~3 の中では 2 が最も F 値が高かったことから，表層的意味と潜在的意味どちらとも使う手法が有効だと分かった．また実験 9 の図 2 より，潜在的意味を利用することで性能は向上するが，潜在的意味と表層的意味の割合を変えることでは性能の違いは得られなかった．
また，各実験による出力文数の決定手法を比較すると，実験 2，4，5 では 2 と 5 はほぼ同じ結果となった．4 と 5 では 4 の方が精度が高く，5 の方が再現率が高くなったのは，4 の方が出力文数が多く正解文を多く含んだがその分不正解分も多く含んだからだと考えられる．これらから，この実験の結果のみからでは出力文数を元データの総文数と総単語数どちらから決定するのが良いかは判断できないので，さらに多様な設定を考えて実験する必要があると思われる．

実験 2，3(手法 A) と実験 6~8(手法 B) を比較すると，手法 A の方が性能は良かったことより，文の類似度を求める際に，tf-idf による単語の特徴量のみにより文のスコアを決定している手法 B よりも，手法 A の表層情報に加え潜在情報を用いることは有効であると言える．

3.4 考察

入力された文書の各文と前日の要約文との類似度を計算し，前日の要約で既に登場した情報を含む文を取り除くことによって，冗長性のない，新しく追加された情報を把握しやすい要約を生成した．また，複数の新聞社の記事に共通する内容を含んでいるため，要約文は複数の新聞社にも同じ内容が載っている重要で信憑性の高いものになった．ランダムに文をとってきた場合よりも、提案手法の方が精度と再現率ともに高くなっているので，この手法は有用だと考えられる，また，文長を 10 文に固定した時よりも元データの文数によって文長を決めたときの方が，精度は下がることもあったが再現率は上がっていた．これは，10 文だと多くの文をとってくるので正解も含まれや

すいが同時に不正解の分もとってきやすく，元データによって適度に文長を短くすると含まれる正解データの量は減るが不正解も減り正解の割合が多くなるからである．

再現率が低いと，ユーザが正解を得る為に多くの文を読まなくてはいけなくなり負担になる．よって，precision の値を上げる為に出力文数を適度に減らし，少ない量で内容を理解できるようにする事がより重要だと考えられる．また，F 値は全ての場合で閾値が 0.5 のときが最も良い結果となった．0.1 では正解まで要約対象から外されてしまうからである．しかし 0.5 より良い閾値がある可能性がある，さらに，遡る日数を変化させても結果にあまり違いは見られなかった．理由として，元データにおいて数日前と類似している文がなく閾値によってフィルタリングされないからだと考えられる．使用する正解データは人手によって作成された要約文なので，今回の手法である数日前との差分をとるというコンセプトとは異なっており，それが全体を通して精度が上昇しない原因だと考えられる．

4. おわりに

実験結果から，今回提案したグラフと潜在的情報を用いた手法は既存の Erkan らが提案した LexRank[5] よりも性能が良いことが分かった．しかし，出力文数の設定方法はさらに細かく値などを変え実験を行いよりより出力文数を探すことで更なる性能の向上が期待できる．また，単語の特徴量による重要文抽出手法は潜在的意味と表層的意味を用いた場合よりも性能が低い事より，潜在的情報を取り入れることは有用だと分かった．今後の課題として，トピック追跡などを用いて内容を考慮しつつ，より少ない文数で正確に正解要約を導ける手法を追求したい．

参考文献

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan “Latent Dirichlet Allocation”, Journal of Machine Learning Research, 3:993-1022, 2003.

[2] G. B. Tran, Tuan A. Tran, N. Tran, M. Alrifai, and N. Kanhabua, Leveraging Learning To Rank in an Optimization Framework for Timeline Summarization, SIGIR, 2013.

[3] M.F. Porter, An algorithm for suffix Stripping, Program, Vol. 14 No.3,pp.130-137, 1980.

[4] C. Lin, ROUGE: a Package for Automatic Evaluation of Summaries, In Proceedings of the Workshop on Text Summarization Branches Out, pp. 74-81, 2004.

[5] Gunes Erkan and Dragomir R. Radev, LexRank: Graph-based Lexical Centrality as Salience in Text Summarization, Journal of Artificial Intelligence Research, pp. 457-479, 2003.