1J3-5in2

逐次的な差分に基づく複数時系列文書要約への取り組み

An Approach to Summarizing Multiple Time-series Documents based on the Sequential Difference of Events

柏井香里 *1　小林一郎 *2
Kaori Kashiwai　Ichiro Kobayashi

*1お茶の水女子大学理学部情報科学科
Dept. of Information Sciences, Faculty of Sciences, Ochanomizu University

*2お茶の水女子大学基幹研究院自然科学系
Graduate School of Humanities and Sciences, Ochanomizu University

Like TV news or newspaper articles, much new information is provided time to time. It is not realistic for readers to read and understand all of such new resources because of being time-consuming. So, a summarizing method, which we can comprehend the contents of such time-series documents provided by multiple information sources, has been required. Based on this, in this research we propose a method to make a summary of long-term news articles provided by multiple newspaper publishing company, focusing on new additional information.

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

2.1 先行研究

時系列文書を対象とした要約として，Allan らは temporal summarization を定義した [1]．近年では，Yan ら [8] により文のランキングアルゴリズムをベースとしたグラフの拡張を行い，異なる時間から 1 つの平面に文章を射影することによって要約を生成する手法や，関連性・被覆率・結合性・多様性のような異なる側面の組み合わせを考慮した関数の最適化により要約を生成する手法 [9] が提案された．LexRank は，Erkan ら [3] によって提案された PageRank[2] に基づいた複数文書要約手法である．この手法では，対象文書中の各文をノードとし，ノードをつなぐエッジを文同士の類似性としてグラフを生成する．多くの文と類似している文は重要度が高いという概念のもと，グラフにおける固有ベクトルの中心性の概念に基づいて文の重要度を計算している．Erkan らは，グラフを生成する際に，類似度の値からエッジの重みを利用する重み付きグラフと，閾値を用いて枝刈りを行う重みなしグラフを提案している．

2.2 提案手法

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 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，お茶の水女子大学理学部情報科学科小林研究室，〒 112-8610 東京都文京区大塚 2-1-1，g1220515@is.ocha.ac.jp

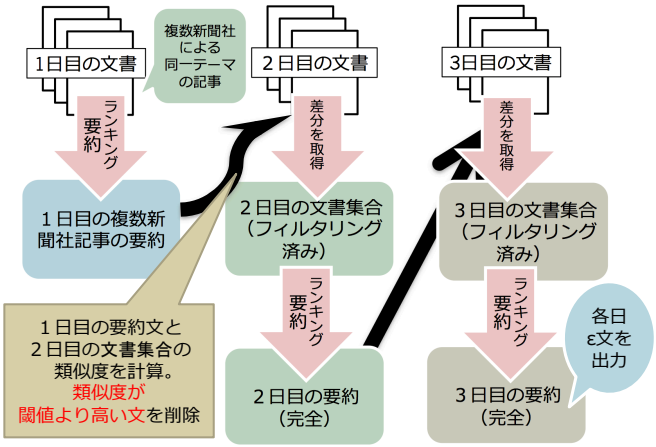

図 1: 提案手法の概要

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

2.3 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,⋯,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．Algorithm1 に要約を生成する手順を示す．

表 2: 閾値 0.5，文長変化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 .

Algorithm 1 要約のプロセス

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
    if length of $\boldsymbol{S}_t > \epsilon$ **then**
      $\boldsymbol{S}'_t \leftarrow$ top ϵ sentences of $\boldsymbol{S}_t$
    else
      $\boldsymbol{S}'_t \leftarrow \boldsymbol{S}_t$
    end if
  end if
  $\boldsymbol{S} \leftarrow \boldsymbol{S}'_t$
end for
return $\boldsymbol{S}$

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 [3] で提案される LexRank アルゴリズムを用いた．

3. 実験

3.1 実験設定

対象データには，Tran ら [6] 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．類似度の閾値によって差分をとる際，何日前まで遡って計算するかを、１日前、２日前、３日前の三種類において実験を行った．これらの提案手法と精度を比較するため，ランダムに文を取得し要約文を作る方法も行った，生成する要約文の長さは，各日ランキング上位 10 文までとしたもの (文数固定) と，元データの文数によって線形的に決めたもの (文数可変) の 2 種類を生成した．文数可変の場合は，元データのそう文数が 100 文以下の場合は出力文数は 2 文，500 文以下の場合は 4 文，500〜1000 文の場合は総文数÷ 100 文，1000 文以上の場合は 10 文とした．そのため，文数可変で作成したものは文数固定のものよりも短い要約となる．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 [5] を用いる．

3.2 評価手法

各新聞社の人手で作成された正解要約と，提案手法によって作成した要約文とを比較した．評価には ROUGE[4] を用い，今回はとくに ROUGE-1 における精度と再現率と F 値を評価に用いる．各日 10 文とする場合と，元データによって文長を定める場合それぞれにおいて，閾値の値を 0.1，0.5，1.0 の 3 種類に設定し精度を確認した．閾値 1.0 の場合は，類似度によるフィルタリングを全く行わないという事である．

3.3 実験結果と考察

要約の出力結果を表 2 に，評価の結果を表 3 にそれぞれ示す.

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

表 3: ROUGE-1 による閾値毎の性能評価

出力文数/閾値		0.1			0.5			1.0		
指標	遡る日数	精度	再現率	F 値	精度	再現率	F 値	精度	再現率	F 値
文長固定	ランダム	-	-	-	-	-	-	0.30	0.06	0.09
	1 日前	0.38	0.15	0.18	**0.80**	0.15	0.22	0.61	0.22	0.27
文長変化	ランダム	-	-	-	-	-	-	0.19	0.06	0.08
	1 日前	0.64	0.21	0.29	0.65	0.25	**0.30**	0.72	0.13	0.22
	2 日前	0.69	0.16	0.25	0.68	0.25	0.30	0.72	0.13	0.22
	3 日前	0.68	0.24	0.30	0.62	**0.26**	0.30	0.72	0.13	0.22

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

LexRank による重要文抽出と，数日前の要約との冗長性を避ける文抽出により，各日毎の重要となる情報を含む文から要約文を生成した．出力する文数は 10 文だと多いが，減らしすぎても正解を取りこぼしてしまうので，より少ない文数で正確に正解を導けるような手法を追求したい．そして，どの程度前日の要約文と似ている文を要約対象から除外するかを決める閾値を，さらに細かく設定し実験を重ねて決定する必要がある．また，前日との比較を文同士の単語の類似度によって計算しているが，これでは同じ単語として使用していても異なる意味を表現している文を区別することは難しいので，トピックトラッキングなどを用いて内容によるより性能の良い類似を発見する手法を模索したい．また，正解データの見直しなども行いたい．

参考文献

[1] James Allan, Rahul Gupta, and Vikas Khandelwal，Temporal Summaries of News Topics，In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval，2001.

[2] Sergey Brin and Lawrence Page, The Anatony of Large-scale Hypertextual Web Search Engine, Computer Networks and ISDN Systems, pp. 107-117, 1998

[3] 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.

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

[6] 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.

[7] G. B. Tran, M. Alrifai, and D. Q. Nguyen，Predicting Relevant News Events for Timeline Summaries，In Proceedings of the 22nd international conference on World Wide Web Companion，pages 91-92.International World Wide Web Conferences Steering Committee，2013.

[8] R. Yan, L. Kong, C. Huang, X. Wan, X. Li, and Y. Zhang, Evolutionary Timeline Summarization: a Balanced Optimization Framework via Iterative Substitution，In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval，2011a.

[9] R. Yan, C. Huang, X. Wan, J. Otterbacher, X. Li, and Y. Zhang，Timeline Generation Evolutionary Trans-Temporal Summarization，In Proceedings of the Conference on Empirical Method in Natural Language Processing，2011b.