グラフを用いた時系列文書要約への取組み

柏井香里
お茶の水女子大学 理学部 情報科学科

小林 一郎
お茶の水女子大学 基幹研究院 自然科学系

{g1220515, koba}@is.ocha.ac.jp

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，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 時系列複数文書の要約

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 [10] 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 [11] 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 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．

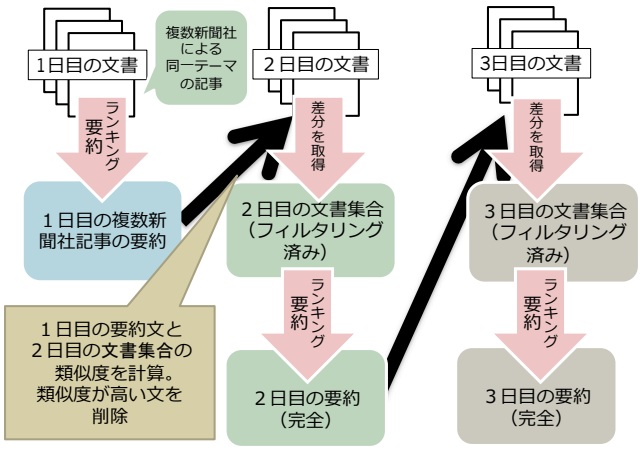

図 1: 提案手法の概要

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．

入力として，$\boldsymbol{D}$，$\boldsymbol{S}$，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．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．

Algorithm 1 要約のプロセス

Input: $\boldsymbol{D}$, $\boldsymbol{S}$, ϵ, α, l
$\boldsymbol{S}$ = { }
$\epsilon \leftarrow$ 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
    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 $\boldsymbol{S}$

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 [8][9] 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．このデータセットは以下の論文で使用されている．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．

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．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 [7] を用いる．

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

3.2 評価手法

各新聞社の人手で作成された正解要約を１つに集約し，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．

3.3 実験結果

トピック BP Oil Spill について生成した要約文の一例と正解データを表 2 に示す．また，表 3 は閾値=0.5 における要約の評価結果である．

表 3:閾値=0.5 における結果

要約対象/手法	精度	再現率	F 値
BP Oil Spill 全日	0.428	0.084	0.135
BP Oil Spill 期間限定	0.447	0.083	0.136
H1N1 全日	0.374	0.094	0.139
H1N1 期間限定	0.302	0.120	0.146
haiti 全日	0.342	0.054	0.078
EgyptianProtest 全日	0.317	0.073	0.120
Financial crisis 全日	0.293	0.040	0.067
IraqWar 全日	0.311	0.065	0.105

表 2:生成された時系列の要約文書 (BP Oil Spill)

生成された要約文	正解文	精度	再現率	F 値
2010-05-30 He said he did not know why it failed to stop the gusher . It is not tolerable . " Experts say it will be difficult to create a watertight seal on a high-pressure gushing pipe at a depth of 1,500 metres -LRB- 5,000 ft -RRB- . But ultimately the pressure forcing the oil upwards proved greater than the force of the mud , which was delivered at a pressure of 6,800 pounds per square inch . Photograph : Win McnameeGetty Images An uncontrollable fountain of oil could gush into the Gulf of Mexico until August , **the Obama administration warned today , as BP conceded it was moving to a containment strategy after failing to plug the well at the center of the most environmentally disastrous spill in US history .** Louisiana , the nearest state to BP 's gushing undersea well that is 42 miles -LRB- 67 km -RRB- out in the Gulf of Mexico , has been the most impacted by the spill so far . " We 're moving to a containment operation . " Whoever can clean it up the quickest , BP gets their bill too . After that , the company could place another blowout preventer on top of the existing one . Ms Browner said BP had been told to drill another relief well in case the first did not work .	**2010-05-30 BBC:**Carol Browner , President Barack Obama 's adviser on energy policy , says the spill is the worst environmental disaster in US history , worse even than the 1989 Exxon Valdez spill in Alaska . **Reuter:**Hayward , who is British , shocks Gulf residents when he says " I 'd like my life back . " He also disputes scientists ' claims that there are large plumes of oil under the surface of the Gulf . **Guardian:**Hayward causes outrage after telling reporters , " There 's no one who wants this over more than I do . I would like my life back . " BP 's clumsy response to oil spill threatens to make a bad situation worse	0.302	0.117	0.168
2010-08-04 This discussion is now closed . High winds make coastal protection efforts difficult . News , features , and opinions on environmental policy , the science of climate change , and tools to live a green life . Meanwhile , the 100-ton box meant to capture the leak is not working . " The well is now being monitored , per the procedure , to ensure the well remains static . " We 've pretty much made this well not a threat , but we need to finish this from the bottom , " said Thad Allen , the official appointed by Barack Obama to lead the federal response to the disaster . **BP said it had completed a process known as static kill , in which heavy mud was pumped in to plug the stricken well , producing a " textbook " result . In the Gulf : Crews prepared to pump mud into the blown-out well , provided a test on the process is successful .** The campaign 's Web site features dozens of images of the burning rig , oil-smeared birds and other environmental devastation from the spill . We have to reverse the damage that 's been done .	**2010-08-04 BBC:**The US government says three-quarters of the oil spilled in the Gulf has been cleaned up or broken down by natural forces . Meanwhile , **BP reports " encouraging " progress with the " static kill " operation to plug the well with mud and seal it with cement . Guardian:BP says the ' static kill ' attempt to stop the oil leak has been successful , though more mud may still have to be pumped into the well to close it permanently .** BP says ' static kill ' has successfully plugged oil well Legislation introduced into the Senate by Democrats to cap oil spill compensation claims at 75m has been stopped because there was n't enough support from within the party . Oil spill damages legislation thwarted in Senate by Democrats BP says the ' static kill ' attempt to stop the oil leak has been successful , though more mud may still have to be pumped into the well to close it permanently . BP oil spill mostly cleaned up , says US The US government announces that the majority of oil from the BP spill has been cleaned up . BP oil spill mostly cleaned up , says US	0.524	0.186	0.275

表 4:閾値毎の性能評価

要約対象/閾値	0.1			0.5			1.0		
指標	精度	再現率	F 値	精度	再現率	F 値	精度	再現率	F 値
BP Oil Spill 全日	0.425	**0.083**	**0.134**	**0.432**	0.077	0.128	0.425	0.083	0.134
BP Oil Spill 期間限定	0.447	0.083	0.136	0.447	0.083	0.136	0.447	0.083	0.136
H1N1 全日	**0.404**	0.093	**0.140**	0.374	**0.094**	0.139	0.369	0.093	0.137
H1N1 期間限定	**0.306**	0.080	0.140	0.302	**0.120**	**0.146**	0.283	0.116	0.140
IraqWar 全日	0.310	0.065	0.105	**0.311**	0.065	0.105	0.310	0.054	0.096
haiti 全日	0.343	0.058	0.098	0.342	0.054	0.078	**0.353**	**0.060**	**0.100**
EgyptianProtest 全日	**0.361**	**0.077**	**0.124**	0.317	0.073	0.120	0.311	0.067	0.106

全期間を通した要約を生成する他，BP Oil Spill については，連続している 2010 年 6 月 14 日から 23 日までの 10 日間, H1N1 については，連続している 2009 年 4 月 29 日から 5 月 1 日までの 3 日間の，限定された期間内で要約を作成し精度を比較した．

また，闢値の値を 0.1，0.5，1.0 の 3 種類に設定し，それらについても精度をそれぞれ確認した．前日の要約と比較し類似度が閾値を超える文を要約対象から除いていたが，閾値=1.0 としたもの, 類似度は 1.0 を超えることは無いのでつまり全ての文を要約対象とするものも用意した．これらを比較することによって, この手法の有用性を確認した．

3.4 考察

BP Oil Spill について，約 1400 文書を各日 10 文までの合計 1700 文程度と，総文数と比べてかなり短く要約をすることができた．H1N1 についても同様に，約 400 文書を各日 10 文までの合計 120 文程度に要約できた．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，上記の要約結果のように出力された要約文に冗長性はあまり見られず，新しく追加された情報を把握しやすくなった．また，表 2 にあるように，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．

ほとんどのものが前日と類似しているものを要約対象から外した方が，外していないものよりも精度は同じまたは高くなったことから，前日と重複を避けてその日の要約を作ることはある程度有用だと考えられる．しかし，期間を限定した BP Oil Spill では 精度に全く違いがなかったことから，この場合前日の要約と比較したときの類似度が 0.1 以上のものは要約対象には含まれていなかったということがわかる．前日との単語の類似度を用いているので，内容が同じでも言い回しや使われている単語が異なると類似度は低くなるため，前日との類似度が低いものが多くなったのではないかと考えられる．各日ごとの精度を見ると，また，日付が連続していない，前後に間がある部分は，前日との関連が薄くなるため類似度を取る必要性は薄いと考え，連続した期間に限定した方が前日との関連がある文書群になるため精度は高くなると予想されたが，文書によっては精度が低くなるものもあった．これは日付が離れていても，同じ様な文が繰り返し含まれていたからだと考えられる．

4 おわりに

LexRank による重要文抽出と，前日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することができた．これにより時系列に沿った要約文生成を行った．しかし実験では各日 10 文と，長期になれば要約結果も長くなってしまい読みにくくなるので，出力する文の数の上限などを要約文がより見やすくなるように設定し直す必要がある．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，どのくらいの値にするかを様々な実験を重ねて決定する必要がある．また，現段階では前日との類似のみを見ているが，直前の日だけではなく，数日前までさかのぼって比較することも考えられる．さらに，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，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．

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] Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz，Multi-document Summarization by Sentence Extraction，In Proceedings of the 2000 NAALP-ANLP Workshop on Automatic Summarization, pp.40-48，2000.

[5] J. Li and S. Li，Evolutionary hierachical dirichlet process for timeline summarization，In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL'13, pages 556-560. Association for Computational Linguistics，2013.

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

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.