グラフを用いた時系列文書要約への取り組み

柏井香里　(指導教員：小林一郎)

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，近年では，Yan ら [6] 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 [7] が提案された．LexRank は，Erkan ら [2] によって提案された PageRank[1] 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．

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．

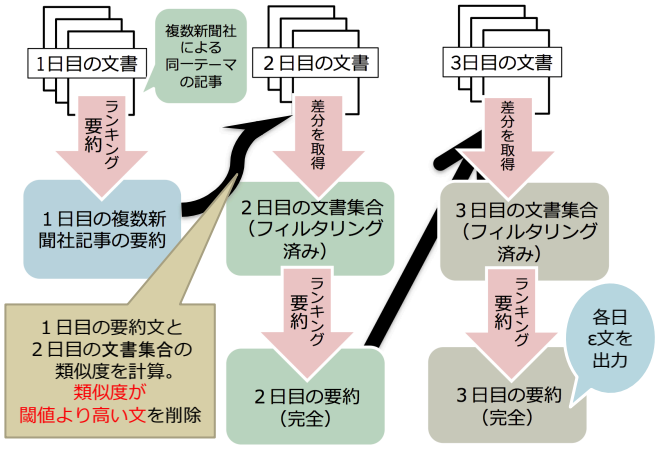

図 1: 提案手法の概要

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=$\{ 1,\cdots,T \}$ である．ここで，D_t は時刻 t に属する文書集合を表す．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 [2] で提案される LexRank アルゴリズムを用いた．

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

対象データには，Tran ら [5] 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．

生成する要約文の長さは，各日ランキング上位 10 文までとしたものと，元データの文数によって線形的に決めたものの 2 種類を生成した．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

表 2: 生成された時系列の要約文書 (BP Oil Spill)

2010-04-24
The Bureau of Ocean Energy Management , Regulation and Enforcement -LRB- BOEMRE -RRB- , formerly known as the · · ·
2010-04-25
Neither the Coast Guard nor their employers have released their names , though several of their families have come forward· · ·
2010-04-26
It is a big change from yesterday ... This is a very serious spill , absolutely , ” said Rear Adm Mary Landry .· · ·
2010-04-27
“ Eventually , things will return to normal . ” Forecasters say it could wash ashore within days near delicate wetlands · · ·

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	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	0.38	0.15	0.18	**0.80**	0.15	0.22	0.61	0.22	0.27
文長変化	0.64	0.21	0.29	0.65	**0.25**	**0.30**	0.72	0.13	0.22

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

グ処理を行った. ステミングには Porter の アルゴリズム [4] を用いる.

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[3] 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，闢値の値を 0.1，0.5，1.0 の 3 種類に設定し精度を確認した．

3.2 実験結果と考察

要約の出力結果を表 2 に示す．評価の結果を表 3 に示す.

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．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，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．0.5 より良い閾値がある可能性がある，

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闢値を，さらに細かく設定し実験を重ねて決定する必要がある．さらに，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．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，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] Sergey Brin and Lawrence Page, The Anatony of Large-scale Hypertextual Web Search Engine, Computer Networks and ISDN Systems, pp. 107-117, 1998

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

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

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

[5] 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.

[6] 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,2011a.

[7] 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.