3O1-7in

潜在情報を考慮した時系列文書の要約への取り組み

A Study on Summarization of Time-series Documents using Latent Information

鈴木　聡子　　小林　一郎
Satoko Suzuki　　Ichiro Kobayashi

お茶の水女子大学大学院人間文化創成科学研究科理学専攻
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

In recent years, as the development of information technologies has enabled people to access enormous amount of documents, the necessity of automatic text summarization has also been increasing for helping people grasp the essential point of documents. Moreover, it is difficult to understand the whole figure of a particular news if it continues for a long period, therefore, we need a method to understand the evolution of time-series documents such as newspaper articles. Based on this, we aim to make a summary which summaries topic transition along time, employing latent information to estimate topic transition. In this paper, we propose sentence extraction method based on latent information as a part of making a summary and show the result of experiments.

1. はじめに

情報技術の発展に伴い大量のデータの蓄積・閲覧が可能となった近年では，ユーザが情報を取捨選択する必要がある．そのため，重要度の高い情報ほど選択されやすくするための情報検索や，膨大な情報の中から効率良く内容を把握するための自動要約において，より高精度な技術の必要性が高まっている．また新聞などにおいても，1つの話題に関して複数の記事が存在することや長期的に書かれていることから，情報量が膨大であるため内容の全体像を把握することは容易なことではない．そのため，そのような長期的な文書における話題の変化の軌跡を簡潔に理解したい，という欲求が生まれる．本研究では，統計的意味解析手法を用いて文書の潜在的意味に着目し，時間経過に伴う話題変遷の把握が可能な要約生成を目的とする．尚，今回は文の重要度を計算し，抽出するまでのプロセスに関する提案および実験を行った．

2. 関連研究

時系列文書を対象とした要約タスクにおいて，様々な研究が行われている．近年では，Yan らによって複数文書要約において使われている文章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 [2] 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を最適化することにとり要約を生成する手法 [3] が提案された．また Jiewi らは，トピックの進化パターンを考慮するために Evolutionary Hierarchical Dirichlet Process(EHDP) と呼ばれる新しいモデルの提案を行った [4]．ここでは，文章の選択には関連性，被覆率，結合性が考慮されている．

3. 提案手法

3.1 Latent Dirichlet Allocation

本研究では，潜在情報を反映するために統計的潜在意味解析手法である Latent Dirichlet Allocation(LDA)[1] を使用する．LDA とは，文書中には複数のトピックが存在し，トピックおよび単語の出現はそれぞれ Dirichlet 分布に従っていることを仮定したモデルである．図 1 に LDA のグラフィカルモデルを示す．仮定に基づき学習を行った結果，各文書におけるトピックの比率を表す $\boldsymbol{\theta}$ と各トピックにおける単語の出現確率を示す $\boldsymbol{\phi}$ を求めることが出来る．この結果を用いて，文の重みを決定する．

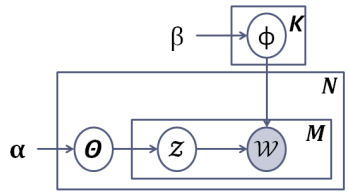

図 1: LDA のグラフィカルモデル

3.2 重要文の決定

以下の処理により重要文を決定する．

step1. 文書集合のトピックベクトルを生成

step2. 各時刻におけるトピックベクトルを生成

step3. 単語の重みを算出

step4. 文の重みを算出

step1 では，トピック比率 θ より式 (1) に従って，文書集合の持つトピックの特徴を表すベクトル $\boldsymbol{T}$ を生成する．

$$\boldsymbol{T} = (\hat{\theta_1}, \hat{\theta_2}, \cdots, \hat{\theta_K}) \tag{1}$$

$$\hat{\theta_k} = \frac{1}{N}\sum_{n=1}^{N}\theta_{n,k} \tag{2}$$

ここで，N は総文書数，K はトピック数を示し,$\theta_{n,k}$ は n 番目の文書における k 番目のトピックの割合を示す．つまり，ここではトピックごとに平均を求めたものを $\hat{\theta_k}$ とし，$\hat{\theta_k}$ を要素に持つベクトルを文書集合の平均的な特徴を表すベクトルとして生成する．また，対象とする文書は時系列文書であるため，時間情報を含む．ここで，各時刻を以下のようにおく．

$$e_t \in e_1, \cdots, e_T \tag{3}$$

連絡先: 鈴木聡子，お茶の水女子大学大学院人間文化創成科学研究科理学専攻情報科学コース小林研究室，〒 112-8610 東京都文京区大塚 2-1-1，suzuki.satoko@is.ocha.ac.jp

次に，step2 において各時刻における特徴を表すベクトルを求める．以下に，時刻 e_t における特徴ベクトルの計算を示す．

$$\boldsymbol{T}_{e_t} = (\hat{\theta_{1e_t}}, \hat{\theta_{2e_t}}, \cdots, \hat{\theta_{Ke_t}}) \tag{4}$$

$$\hat{\theta_{ke_t}} = \frac{1}{N_{e_t}} \sum_{n \in e_t} \theta_{n,k} \tag{5}$$

$$\boldsymbol{\Delta}_{e_t} = \boldsymbol{T}_{e_t} - \boldsymbol{T} \tag{6}$$

ここで，$\boldsymbol{T}_{e_t}$ は時刻 e_t におけるトピック比率の平均であり，$\boldsymbol{T}$ と $\boldsymbol{T}_{e_t}$ の差分を時刻 e_t での特徴ベクトルとする．次に，step2 で求めた Δ_{e_t} と LDA で求められるトピックにおける単語の出現確率を示す $\boldsymbol{\phi}$ より，時刻 e_t における単語 v の重み w_v を計算する．算出方法は以下の式に従う．

$$\boldsymbol{w}_v = (\boldsymbol{1} + \boldsymbol{\Delta}_{e_t}) \cdot \boldsymbol{\Phi}_{\cdot,v} \tag{7}$$

最後に，step4 では先ほど求めた単語の重みより文の重みの算出を行う．文 S_j の重みは以下の通りである．

$$S_j = \frac{1}{\sqrt{N_{S_j}}} \sum_{v \in S_j}^{V} w_v \tag{8}$$

N_{S_J} は文 S_j に含まれる語彙の総数であり，ここでは文長に左右されないために平方根を逆数として掛けている．

4. 実験

4.1 実験設定

使用するデータは New York Times など，10 個のニュース資源より集めた記事より，インフルエンザの流行に関するもの，全 428 記事を使用する．表 1 にニュース資源について示す．また，同じ日付に記述された文書を同じ時刻に含むものとした．LDA におけるパラメータの設定は，$\alpha = 0.1$，$\beta = 0.1$，トピック数 $k = 61$ とし，イテレーションの回数は 200 回とする．尚，トピック数はパープレキシティをもとに求めた．また，今回は無作為に選んだ特定の時刻（e_{17}:2009/04/28）に関して実験を行った．

表 1: ニュース資源

ニュース資源	国	ニュース資源	国
BCC	UK	New York Times	US
Guardian	UK	Washington Post	US
CNN	US	Fox News	US
ABC	US	MSNBC	US

4.2 実験結果

図 2 に文書集合全体の特徴ベクトル $\boldsymbol{T}$ と $\boldsymbol{T}_{e_{17}}$ を示す．ここで，文書集合は，2007 年 4 月 18 日から 2010 年 8 月 10 日までの全 428 記事であるに対し，時刻 e_{17} は 2009 年 4 月 28 日の全 7 記事である．次に，提案する単語の重みより重要度の高い 3 文を以下に示す．

- Navy Lt Sean Robertson said ill crew members had been treated with anti-viral medication and the remaining crew had been given prophylaxis .
- As of 19:15 GMT , 28 April 2009 , seven countries have officially reported cases of swine influenza infection .
- swine flu map click viru spread report canada april 26

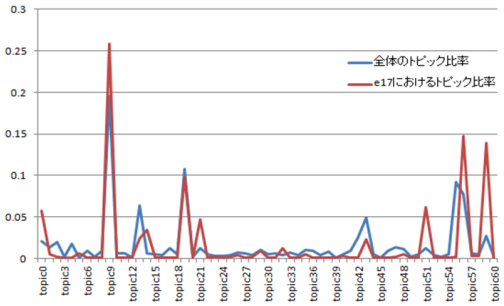

図 2: 全体と時刻 e_{17} の特徴ベクトル

5. 考察

図 2 より，全 428 記事の特徴ベクトルとその内の僅か 7 文書を対象としたトピック比率の平均を比較したところ，非常に類似した結果を示していることが分かった．このグラフより，差の大きい部分が時刻 e_{17} を特徴づけるトピックであることが考えられる．また，文の抽出では，いずれもインフルエンザに強く関連するものである．特に 2 番目に示す文は，この記事が記述された 2009 年 4 月 28 日の現状を示している．この結果より，時刻 e_{17} において重要な文の抽出をすることができた．

6. おわりに

本研究では，潜在情報を考慮した要約生成に向けて，LDA をもとに重要文の決定方法の提案を行った．また，提案手法に従って特定の時刻における実験を行った．実験では，定量的な評価や他手法との比較を行っていないため一概には言えないが，今回は話題や時刻と強く関連のある文を抽出することができた．今後は，他手法との比較や別のデータでの実験を行い，要約手法の提案へと進めていくつもりである．

参考文献

[1] David Blei, Andrew Ng and Micheal Jordan:Latent Dirichlet Allocation. Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.

[2] Rui Yan, Liang Kong, Congrui Huang, Xiaojun Wan, Xiaouming Li and Yan 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, 2011.

[3] Rui Yan, Liang Kong, Congrui Huang, Xiaojun Wan, Jahna Otterbacher, Xiaoming Li and Yan Zhang:Timeline Generation Evolutionary Trans-Temporal Summarization. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011.

[4] Jiwei Li and Sujian Li:Evolutionary Hierarchical Dirichlet Process for Timeline Summarization. In Proceedings of 51th Annual Meeting the Association for Computer Linguistics, pages 556-560, 2013.