

潜在的トピックの類似度グラフに基づく要約文生成

北島 理沙 小林 一郎

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

{kitajima.risa, koba}@is.ocha.ac.jp

1 はじめに

近年、情報技術の発達により大量のテキストデータが蓄積され、必要な情報を取捨選択することが重要となっている。そのことから、必要な情報をユーザに容易に判断させるために、テキスト自動要約の必要性が高まっている。テキスト自動要約には様々な手法が存在するが、その一つである、文の類似度に基づいたグラフを利用した手法が高い精度で要約文生成を行えることが知られている。通常、この手法で要約文生成に利用しているのは文の表層的な情報であり、その潜在的トピックについては考慮されていない。そこで本研究では、文の持つ潜在的トピックを考慮した類似度グラフを用いた要約文生成手法を提案する。そして、実験により生成された要約文について、従来の手法と提案手法との比較および考察を行う。

2 関連研究

文書要約については様々な手法が提案されており、グラフを用いた文書要約手法が高い精度を示すことが知られている [1, 2, 3]。Erkan ら [1]、Mihalcea ら [2] は、文書の全体像をつかむ包括的な要約生成を行っており、特に、前者は複数文書、後者は単一文書を対象としている。一方、Otterbacher ら [3] は、ある視点に特化したクエリに基づく要約生成を行っている。いずれにおいても、重心法 [4] などのベースラインとされる手法と比較して、高い精度を示している。

また、潜在的トピック推定を文書要約に応用した研究としては、Hennig [5] の研究や、Tang ら [6] の研究がある。どちらも、クエリに基づく要約生成を行っており、前者は、Probabilistic Latent Semantic Analysis を、また、後者は、Latent Dirichlet Allocation を要約手法に適用した手法を提案している。

本研究では、グラフを用いた要約手法に対して、潜在的トピック推定を導入することにより、より内容を考慮した文書要約を生成することを目指す。

3 LexRank

LexRank は、Erkan ら [1] により提案された、PageRank [7] を応用した複数文書要約手法である。要約手法には、文書の全体像をまとめる要約と、ある視点に特化した内容をまとめる要約とがあるが、LexRank は前者の要約を対象としている。LexRank は、文のグラフ表現における固有ベクトル中心性の概念に基づいて文の重要度を計算する手法である。これは、単に次数の多いノードを評価するだけでなく、次数の多いノードと隣接しているノードの重要度についても考慮し、その分に比例して高く評価することができる。この手法では、文間のコサイン類似度に基づいた連結性行列が文のグラフ表現の隣接行列として使われており、その隣接行列の第 1 固有ベクトルの成分を各ノードの中心性を表すスコアと考える。

Erkan らは、類似度グラフを生成する際に、上で述べたように枝の重みを利用した重みつきグラフとして表わす手法の他に、その枝の重みに対して閾値 t を用いて枝刈りを行い、重みなしグラフとして表わす手法を提案している。前者の手法は Cont. LexRank、後者の手法は LexRank と呼ばれている。LexRank および Cont. LexRank は、実際には上述の処理のみで要約文を生成するのではなく、Radev ら [8] の提案した要約システムである MEAD¹ の内部に組み込み、冗長性削減のための指標などと組み合わせることで要約文を生成することを前提としている。本研究では、MEAD に組み込む前の LexRank との比較により、提案手法の評価を行う。

4 提案手法

4.1 TopicRank

LexRank では、文間の類似度として $tfidf$ 値を要素とする文ベクトルのコサイン類似度を用いているのに対して、文のもつトピック分布の類似度を文間の類似

¹<http://www.summarization.com>

度として用いる手法を提案し、これを TopicRank と呼ぶことにする。

潜在的意味解析手法としては、Latent Dirichlet Allocation（以下、LDA）を用いる。LDA とは、一つの文書に対して複数のトピックが存在すると想定した確率的トピックモデルであり、それぞれのトピックがある確率を持って文書上に生起するという考えの下、そのトピックの確率分布を導き出す手法である [9]。トピック分布の類似度判定指標には、LDA において精度が高いと報告されている [5]、Jensen-Shannon ダイバージェンスを用いる。また、距離から類似度への変換は、式 (1) を用いる。

$$\text{sim}(P, Q) = 1 - D_{JS}(P, Q) \quad (1)$$

文間のトピック分布類似度に基づいた連結性行列は、対象文書群に含まれる文数が 7 のとき、例えば表 1 のように表される。ただし、s0 は、対象文書群に含まれる文のうち 0 番目の文を表す。

表 1: 文間のトピック分布類似度

| | s0 | s1 | s2 | s3 | s4 | s5 | s6 |
|----|------|------|------|------|------|------|------|
| s0 | 1.00 | 0.03 | 0.02 | 0.01 | 0.07 | 0.45 | 0.08 |
| s1 | 0.03 | 1.00 | 0.37 | 0.23 | 0.02 | 0.30 | 0.24 |
| s2 | 0.02 | 0.37 | 1.00 | 0.09 | 0.12 | 0.18 | 0.21 |
| s3 | 0.01 | 0.23 | 0.09 | 1.00 | 0.14 | 0.03 | 0.19 |
| s4 | 0.07 | 0.02 | 0.12 | 0.14 | 1.00 | 0.16 | 0.11 |
| s5 | 0.45 | 0.30 | 0.18 | 0.03 | 0.16 | 1.00 | 0.27 |
| s6 | 0.08 | 0.24 | 0.21 | 0.19 | 0.11 | 0.27 | 1.00 |

次に、この連結性行列を隣接行列とみなし、類似度グラフを生成する。例として、表 1 に対する類似度グラフは、図 1 のように示される。各節点は文、枝は文間の類似度を示す。ここでは、トピック分布類似度の値を枝の重みとした重みつきグラフとして示した。

次に、生成された類似度グラフに対して、固有ベクトル中心性に基づいた各文の重要度を計算する。文 u の重要度は、Erkan ら [1] の手法を参考にして、式 (2) で求められる。ここで、 N は対象としている文書群の総文数、 $\text{adj}[u]$ は文 u の隣接ノード集合、 d はある一定の割合で非隣接ノードとの類似度を考慮するための制動係数（damping factor）である。制動係数 d の値は、Brin ら [7] の結果を参考に $d = 0.15$ とした。

$$p(u) = \frac{d}{N} + (1-d) \sum_{v \in \text{adj}[u]} \frac{\text{sim}(u,v)}{\sum_{z \in \text{adj}[v]} \text{sim}(z,v)} p(u) \quad (2)$$

次に、重要度を要素とした行列に対してべき乗法を用いて第 1 固有ベクトルを計算する。これにより、中心性の高い文と類似していることがその文の重要度を高める、という概念に基づいた文の重要度を求めるこ

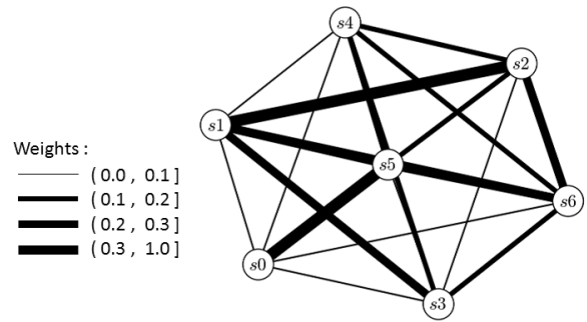


図 1: 類似度グラフ

とができる。最後に、計算された重要度に基づいて文をランク付けし、上位から文を選択していくことで、要約文が生成される。

類似度グラフを生成する際、上で述べたように枝の重みを利用した重みつきグラフとして表わす手法の他に、その枝の重みに対して閾値 t を用いて枝刈りを行い、重みなしグラフとして表わす手法も考える。本研究では、前者を Cont. TopicRank、後者を TopicRank と呼ぶ。なお、後者によって生成された類似度グラフにおける文の重要度は、Erkan ら [1] の手法を参考にして、式 (3) によって求められる。ここで、 $\text{deg}(v)$ はノード v の次数を表わす。

$$p(u) = \frac{d}{N} + (1-d) \sum_{v \in \text{adj}[u]} \frac{p(v)}{\text{deg}(v)} \quad (3)$$

4.2 TopicLexRank

文間の類似度として、LexRank では表層的類似度のみを、また、TopicRank ではトピック分布の類似度のみに用いた。ここでは、表層的類似度とトピック分布の類似度の両者を考慮した類似度を文間の類似度とした手法を提案し、これを TopicLexRank と呼ぶことにする。TopicLexRank における文 S 、文 T 間の類似度 $\text{sim}2$ は、それぞれのトピック分布を P, Q で表わすとき、式 (1) で求められるトピック分布の類似度 sim を用いて、式 (4) で表わされる。また、文 u の重要度は、式 (4) を用いて式 (5) で表わされる。

$$\begin{aligned} \text{sim}2(S, T) &= \alpha * \text{sim}(P, Q) \\ &+ (1 - \alpha) * \text{cosine}(\text{tfidf}(S), \text{tfidf}(T)) \end{aligned} \quad (4)$$

$$p(u) = \frac{d}{N} + (1-d) \sum_{v \in \text{adj}[u]} \frac{\text{sim}2(u,v)}{\sum_{z \in \text{adj}[v]} \text{sim}2(z,v)} p(u) \quad (5)$$

本手法に対しても、類似度グラフを生成する際に重みつきグラフとして表わす手法と、重みなしグラフとして表わす手法を考える。本研究では、前者を Cont. TopicLexRank、後者を TopicLexRank と呼ぶ。なお、後者によって生成された類似度グラフにおける文の重要度は、式 (3) によって求められる。

5 実験

5.1 実験設定

対象データには，DUC2004 の Task2 で使われた文書データを用いた．約 10 件の新聞記事からなる文書群が 50 セット用意されており，それらを用いて複数文書要約を行う．ここでは，LexRank，TopicRank および TopicLexRank の比較，また，Cont. LexRank，Cont. TopicRank および Cont. TopicLexRank の比較を行うこととし，各手法によって生成された要約文に対して ROUGE を用いて評価する．特に，人間の評価と関連していることが示されている [10]，ROUGE-1 値を用いる．また，ストップワードを含めた値とストップワードを除いた値を求めることにし，前者を with，後者を without として示す．本実験においては，TopicRank における枝刈りのための閾値 t を変化させながら実験を行い，TopicRank における適切な閾値について調べる．TopicLexRank についても適切な閾値の判定が必要であるが，まずは Cont. TopicLexRank において潜在的トピックと表層的情報の重みを表すパラメータ α を変化させ，適切なパラメータ値を調べる．そして，求められたパラメータ値を TopicLexRank に適用した後に，適切な閾値について調べる．各手法の比較は，これらの予備実験によって適切な閾値およびパラメータ値を求め適用した上で行うこととする．

なお，LexRank において枝刈りをするためのコサイン類似度の閾値は先行研究 [1] の結果から 0.1 とする．また，LDA において指定するトピック数による精度の違いについては本実験では考慮しないこととし，その数を 20 と設定することにする．

5.2 実験結果

図 2 に，TopicRank における閾値 t の変化に伴う ROUGE-1 値の変化を示す．閾値 t の値が大きくなるにつれて，ROUGE-1 値も高くなっている．しかし， $t = 0.5$ を超えると，ROUGE-1 値は安定し，あまり変化が見られなくなることが分かる．この結果から，TopicRank における適切な閾値 t の値は $t = 0.5$ とした．

図 3 に，Cont. TopicLexRank における重みパラメータ α の値の変化に伴う ROUGE-1 値の変化を示す．重みパラメータ α の値は，ROUGE-1 値にあまり大きな影響を与えていないことが分かる．ROUGE-1 値が最も高くなるのは， $\alpha = 0.7$ のときであるため，Cont. TopicLexRank における適切な重みパラメータ α の値は， $\alpha = 0.7$ とする．

図 4 に，TopicLexRank における閾値 t の変化に伴う ROUGE-1 値の変化を示す．なお，ここでの重みパ

ラメータ α の値は，前の実験結果から， $\alpha = 0.7$ としている． $t = 0.5$ のとき，ROUGE-1 値は最も大きくなり，それ以上閾値 t が大きくなっても ROUGE-1 値は変わらないことが分かる．したがって，TopicLexRank における適切な閾値 t は， $t = 0.5$ とする．

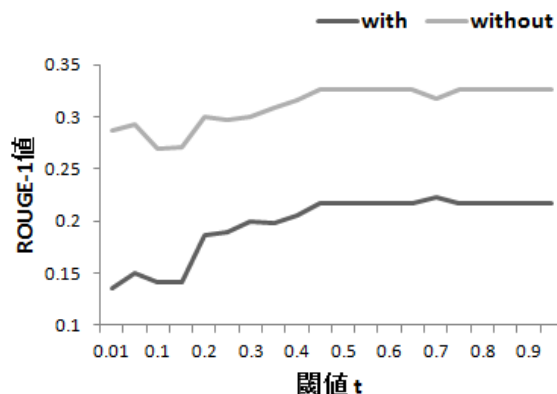


図 2: TopicRank における ROUGE-1 値の変化

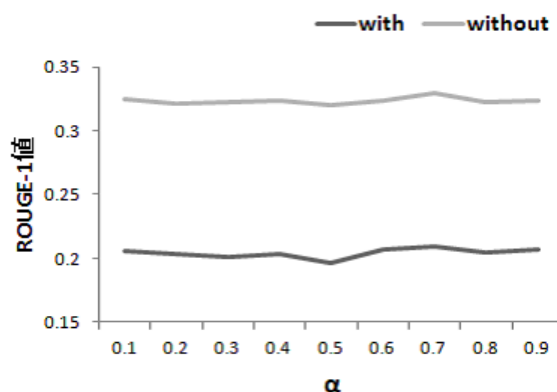


図 3: Cont. TopicLexRank における ROUGE-1 値の変化

TopicRank，Cont. TopicRank，および TopicLexRank について適切な閾値 t ，重みパラメータ α の値が求められたので，最後に，手法の比較を行う．表 2 に，類似度グラフに対して枝刈りを行う場合の ROUGE-1 値の比較を示す．枝刈りを行う場合は，提案手法は従来の LexRank よりも低い ROUGE-1 値を示している．また，潜在的トピックのみを考慮した手法である TopicRank よりも，表層的情報と潜在的トピックの両方を考慮した手法である TopicLexRank の方が，ROUGE-1 値は低くなる事が分かる．

表 2: 枝刈りを行う場合の ROUGE-1 値の比較

| 手法 | with | without |
|--------------|-------|---------|
| LexRank | 0.246 | 0.340 |
| TopicRank | 0.223 | 0.326 |
| TopicLexRank | 0.216 | 0.326 |

表 3 に，類似度グラフに対して枝刈りを行わない場合の ROUGE-1 値の比較を示す．枝刈りを行わない場

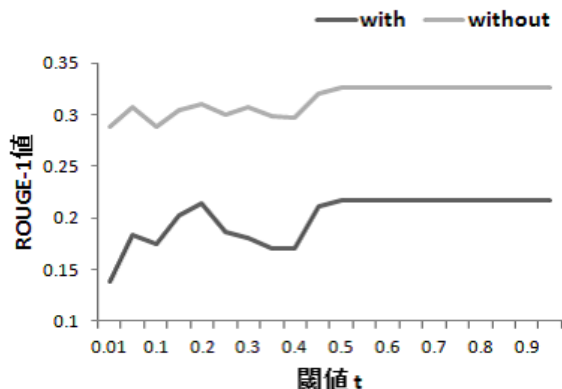


図 4: TopicLexRank における ROUGE-1 値の変化

合は，提案手法が従来の手法よりも高い ROUGE-1 値を示している．また，潜在的トピックのみを考慮した手法である TopicRank よりも，表層的情報と潜在的トピックの両方を考慮した手法である TopicLexRank の方が，ROUGE-1 値は高くなることが分かる．

表 3: 枝刈りを行わない場合の ROUGE-1 値の比較

| 手法 | with | without |
|--------------------|-------|---------|
| Cont. LexRank | 0.142 | 0.285 |
| Cont. TopicRank | 0.170 | 0.313 |
| Cont. TopicLexRank | 0.209 | 0.329 |

5.3 考察

実験結果より，枝刈りをする場合には表層的情報を用いた手法の方が精度が良く，一方で，枝刈りをしない場合には潜在的トピックを考慮した手法の方が精度が良くなるということが分かった．前者については，トピック分布の Jensen-Shannon ダイバージェンスが，表層的情報のコサイン類似度に比べて閾値による影響が反映されていないことが予想される．これについては，同程度のトピック分布の類似度に対して差が付きやすくなるように類似度判定指標を選択することで，閾値の設定がより効果的に枝刈りの結果に反映されるのではないかと考える．一方，後者については，潜在的トピック分布の類似度が表層的情報の類似度に比べて文のもつ意味を捉えられているからであると考えられる．このことから，枝刈りをする際に閾値や類似度判定指標を工夫することで，TopicRank が LexRank よりも高い精度を示す可能性があると考えられる．

また，Cont. LexRank，Cont. TopicRank，Cont. TopicLexRank を比較すると，他の 2 つの手法を組み合わせた手法である Cont. TopicLexRank は最も良い精度となり，表層的情報と潜在的なトピック分布の両方を考慮することが重要であることが分かった．

6 おわりに

本研究では，類似度グラフを用いた要約文生成に対して，文の持つ潜在的トピックを考慮した手法を提案した．具体的には，トピック分布の類似度のみを考慮した TopicRank と，表層的情報とトピック分布の類似度の両者を考慮した TopicLexRank を提案し，また，それらにおいて閾値による枝刈りを行わない手法についても比較対象として用いた．実験により，閾値を設けて枝刈りをしない場合には提案手法が従来の手法を上回る精度を示し，潜在的トピックを考慮した提案手法が要約文生成において有効であることが分かった．

今後の課題としては，今回得られた考察を踏まえて，冗長性削減のための手法などを提案手法に組み込み，より精度の高い手法へ拡張していきたいと考える．また，潜在的トピック推定については，与えるトピック数を変化させるなどしてトピック数の違いによる生成された要約文の精度についても調べることや，トピック分布の類似度判定において他の指標を用いた場合の枝刈りの影響を調べることを予定している．

参考文献

- [1] G. Erkan and D. R. Radev, LexRank: Graph-based Lexical Centrality as Saliency in Text Summarization, *Journal of Artificial Intelligence Research*, 22, pp. 457–479, 2004.
- [2] R. Mihalcea and P. Tarau, TextRank: Bringing order into texts, In *Proc. of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 401–411, 2004.
- [3] J. Otterbacher, G. Erkan and D. R. Radev, Using random walks for question-focused sentence retrieval, In *Proc. of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 915–922, 2005.
- [4] D. R. Radev, H. Jing, Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies, In *ANLP/NAACL Workshop on Summarization* Seattle, WA, 2000.
- [5] L. Hennig, Topic-based Multi-Document Summarization with Probabilistic Latent Semantic Analysis, *International Conference RANLP 2009-Borovars, Bulgaria*, pp. 144–149, 2009.
- [6] J. Tang, L. Yao and D. Chen, Multi-topic Based Query-Oriented Summarization, In *Proc. of SIAM International Conference on Data Mining*, pp. 1147–1158, 2009.
- [7] S. Brin and L. Page, The anatomy of a large-scale hypertextual Web search engine, *Computer Networks and ISDN Systems*, 30, pp. 107–117, 1998.
- [8] D. R. Radev, S. Blair-Goldensohn and Z. Zhang, Experiments in single and multi-document summarization using MEAD, In *First Document Understanding Conference* New Orleans, LA, 2001.
- [9] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.
- [10] C. Lin, ROUGE: a Package for Automatic Evaluation of Summaries, In *Proc. of the Workshop on Text Summarization Branches Out*, pp. 74–81, 2004.