

## トピックを考慮したグラフ表現に基づく複数文書要約

北島理沙<sup>†</sup>小林一郎<sup>‡</sup>

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

## 1 はじめに

近年、自動要約の技術の必要性が高まり、様々な手法が提案されている一方で、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た手法が提案され、特に、LexRank [1] はその有用性が知られている。これは、PageRank の概念に基づいた要約手法であり、文をノード、文間のコサイン類似度をエッジとした類似度グラフに基づいて文の重要度を計算する。しかし、LexRank では文の表層的な情報のみを用いており、文の持つ潜在的な情報であるトピックについては考慮されていない。本研究では、トピックに基づいた文の類似度グラフを用いる複数文書要約手法を提案し、DUC2004 \* を用いた文書要約を通して LexRank との性能の比較および考察を行う。

## 2 提案手法

## 2.1 TopicRank

LexRank ではグラフを構成する文間の類似度として *tfidf* ベクトルのコサイン類似度に基づき文の中心性を算出するのに対して、我々はトピックに基づいた文間類似度を扱い、各々の文に割り当てられるトピック分布のベクトルに基づき中心性を算出する手法を提案し、これを TopicRank と呼ぶことにする。なお、文のトピック分布推定には、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る、Latent Dirichlet Allocation (LDA) [2] を用いる。式 (1) に、TopicRank における文  $S, T$  間の類似度を示す。 $P, Q$  は、それぞれ文  $S, T$  に対応するトピック分布である。トピック分布の類似度指標には Jensen-Shannon 距離を用いる。最終的な文間の類似度は  $\alpha(0 \leq \alpha \leq 1)$  によって潜在的類似度と表層的類似度を混合して設定することが可能となっている。次に、計算された文間類

似度を重みとした類似度グラフを生成する。文  $u$  の重要度は、Erkan ら [1] の手法を参考にして、式 (2) で求める。ここで、 $N$  は対象文書群の総文数、 $adj[u]$  は文  $u$  の隣接ノード集合、 $d$  は制動係数 (damping factor) である。文の重要度は反復的に計算されるため、これらを要素とした行列に対し、べき乗法を用いて第 1 固有ベクトルを計算する。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から選択していくことで要約文を生成する。

$$sim(S, T) = \alpha * sim_{JS}(P, Q) + (1 - \alpha) * sim_{cosine}(tfidf(S), tfidf(T)) \quad (1)$$

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

## 2.2 冗長性削減

TopicRank に従って文を抽出していくと冗長性のある要約文が生成される可能性がある。これに対し、MMR(Maximal Marginal Relevance) [3] を応用した指標を提案する。MMR は、抽出済の文との類似度に対応するペナルティ値を与えることで類似文の抽出を防ぐ指標であり、クエリに特化した要約においてしばしば使用される。提案手法では、高い TopicRank をもち、かつ、抽出済の文と表層的に類似していない文を抽出したいと考え、式 (3) のように応用する。なお、 $v_i$  は対象文書群内の文、 $D$  は対象文書群、 $D'$  は要約文として既に選ばれた  $D$  内の文集合、 $\lambda$  は重みパラメータを表わす。

$$MMR \equiv argmax_{v_i \in D \setminus D'} [\lambda TopicRank(v_i) - (1 - \lambda) max_{v_j \in D'} Sim_{cosine}(v_i, v_j)] \quad (3)$$

## 3 実験

## 3.1 実験設定

DUC2004 の Task2 で使われた、10 件の新聞記事群 50 セットからなるデータを用いる。評価指標には、ROUGE-1 値 [4] を採用する。まず、TopicRank において類似度の重みを制御する  $\alpha$  と制動係数  $d$  の値を変化させ、次に、MMR を導入した TopicRank において、

Multi-document Summarization with a Graph of Latent Topics in Sentences

<sup>†</sup>Risa KITAJIMA(kitajima.risa@is.ocha.ac.jp),

<sup>‡</sup>Ichiro KOBAYASHI(koba@is.ocha.ac.jp)

Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University, 2-1-1 Ohtsuka Bunkyo-ku Tokyo 112-8610

\*<http://www-nlpir.nist.gov/projects/duc/guidelines/2004.html>

類似度の重みを制御する  $\alpha$  と冗長性削減の重みを制御する  $\lambda$  の値を変化させ、最後にデータにストップワードを含めた場合の “with” と含めない場合の “without” を条件において手法間の比較を行う。LDA におけるトピック数は 50，潜在変数推定にはギブスサンプリングを用いる。各手法につき 10 回実験を行いその平均を示す。

### 3.2 実験結果

図 1 に、TopicRank における  $\alpha$  の変化に伴う ROUGE-1 値の変化を示す。  $d$  に関わらず  $\alpha = 1.0$  の場合に値が高く、特に  $d = 0.95$  の場合が最も高い値となっている。図 2 に、MMR を考慮した TopicRank における  $\lambda$  の変化に伴う ROUGE-1 値の変化を示す。  $\lambda$  に関わらず、 $\alpha = 1.0$  の場合に値が高く、 $\alpha = 1.0$  で比較した際に最も精度が高いのは、with, without とも  $\lambda = 0.5$  のときである。表 1 に、各手法間の ROUGE-1 値の比較を示す。前の実験結果より、TopicRank は  $\alpha = 1.0, d = 0.95$  の場合、TopicRank (+MMR) では  $\alpha = 1.0, \lambda = 0.5$  の場合を示した。提案手法である TopicRank は、LexRank よりも高い ROUGE-1 値を示していることが分かる。一方で、MMR を導入したことでの精度の差はあまり大きく見られず、TopicRank に対する冗長性削減の効果は小さいことが分かる。

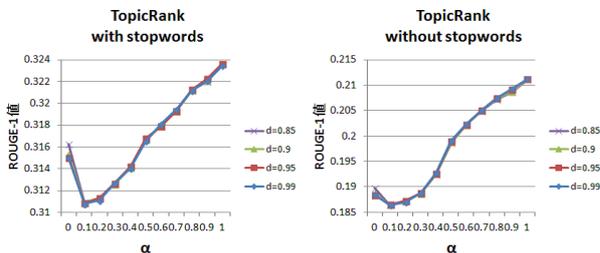


図 1: TopicRank における ROUGE-1 値の変化

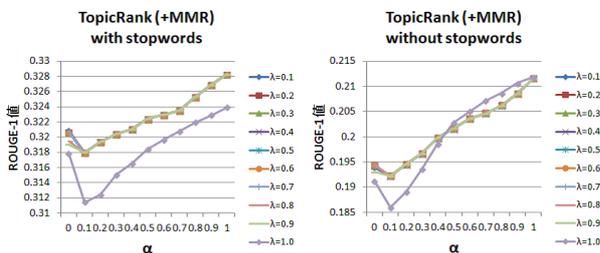


図 2: MMR 導入後の ROUGE-1 値の変化

### 3.3 考察

トピックに基づいた類似度を考慮する割合について考察した結果として、トピック分布の類似度のみを用いたときに精度が高かったことから、*tfidf* ベクトルのような表層的な情報よりもトピックに基づいた類似度

表 1: ROUGE-1 値の比較

method	with	without
LexRank	0.222	0.035
TopicRank	0.324	0.211
TopicRank (+MMR)	0.328	0.212

の方がグラフに基づく複数文書要約において役立つことが分かった。また、MMR 導入後の精度の差が小さかったことに関しては、以下のように考える。 $\lambda = 1.0$  の場合、つまり、冗長性削減を考慮しない場合に着目したときに、with では  $\alpha$  が大きくなるにつれて冗長性削減を考慮した場合 ( $\lambda \neq 1.0$ ) との差が小さくなり、without においても  $\alpha > 0.4$  のときに冗長性削減を考慮した場合よりも高い ROUGE-1 値を示している。 $\alpha$  が大きいことは、トピック分布の類似度をより考慮することを意味するため、トピックに基づいて文の重要度を計算することで、冗長性の少ない要約生成を行えたといえる。このことから、冗長性削減のための手法である MMR を導入した後も、TopicRank の精度があまり変わらなかったのではないかと考える。

## 4 おわりに

本研究では、トピックを考慮したグラフに基づく複数文書要約手法である TopicRank を提案し、DUC2004 を用いた実験を通して提案手法の考察を行った。結果として、グラフに基づいた要約においてトピックが有用であり、その特性として冗長性の少ない要約生成が行えることが分かった。今後の課題としては、対象データの種類の違いと精度との関係を調査することにより、提案手法に対するより深い考察を行いたいと考えている。

## 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, pp. 457–479, 2004.
- [2] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet Allocation. Journal of Machine Learning Research, Vol. 3, pp. 993–1022, 2003.
- [3] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz, : Multi-document Summarization by Sentence Extraction, Proceedings of the 2000 NAALP-ANLP Workshop on Automatic Summarization, pp. 40–48, 2000.
- [4] C. Lin, : ROUGE: a Package for Automatic Evaluation of Summaries, In Proc. of the Workshop on Text Summarization Branches Out, pp. 74–81, 2004.