トピックを考慮したグラフによる複数文書要約への一考察

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，情報技術の発展に伴って大量のテキストデータが蓄積されるようになり，適した情報を効率よく選択することが重要になってきている．そのため，人々が必要としている情報を選択するために自動要約の技術の必要性がますます高まっている．自動要約技術においては，様々な手法が提案されている一方で，文の関係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，グラフベースの文書要約技術が提案されており，その有用性が知られている．特に，LexRank [1] はリード手法 [2] や文ベクトルの重心に基づいた手法 [3] のようなベンチマーク手法として用いられる様々な手法よりも良い結果を示すことが知られている．この手法は文間の類似度を計算するのに表層情報に対するコサイン類似度を用いている．本研究では，潜在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 [1]を用いた文書要約実験を通して，LexRank との性能の比較および考察を行う．

2　関連研究

自動要約技術としては，多くの手法がこれまでに提案されてきている一方で，文の類似度をグラフ表現したものを用いる手法が高い精度で文書要約を行えることが知られている [1, 4, 5]．Erkan ら [1] および Mihalcea ら [4] は，対象となる文書の概要をまとめた要約生成を行っており，前者では複数文書を，後者では単一文書を対象としている．加えて，Otterbacher ら [5] は，クエリに特化した要約生成を行っている．これらの研究は，リード手法 [2] や重心法を用いた手法 [3] などのベースラインとなる手法よりも高い精度を示している．要約技術において PageRank アルゴリズムを適用している代表的かつ初期の研究としては，LexRank [1] がある．LexRank は，Erkan ら [1] により提案された，PageRank [6] の概念に基づいた複数文書要約手法である．LexRank は，対象文書内の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，その分に比例して対象ノードを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．

[1]http://www-nlpir.nist.gov/projects/duc/guidelines/2004.html

3　提案手法

3.1　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．文内のトピック分布を推定するための手法として，Latent Dirichlet Allocation (LDA) [7] を用いる．類似度グラフは文間の類似度を要素とした接続行列に基づいて生成される．

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 (1) 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，$adj[u]$ 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 [6]．制動係数 d の値は，Brin ら [6] の結果を参考に $d = 0.85$ とした．類似度 $sim(u, v)$ の計算については，3.2 節に示す．

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

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．

3.2 文間類似度

文間の類似度として，LexRank では表層的な類似度のみを用いている一方で，提案手法である TopicRank では、対象文書群の表層的な類似度と潜在的な類似度の両方を用いる．式 (2) は，TopicRank の枠組みにおいて定義されている文 S と文 T の間の類似度を示している．P と Q は，それぞれ文 S と文 T のもつトピック分布である．式 (3) は，トピック分布に基づいた類似度を示している．文の重要度は，式 (2) を用いて式 (1) によって計算される．トピック分布の類似度指標には，Jensen-Shannon ダイバージェンスを用いる．

$$sim(S,T) = \alpha * sim(P,Q) + (1-\alpha) * cosine(tfidf(S), tfidf(T)) \quad (2)$$

$$sim(P,Q) = 1 - D_{JS}(P,Q) \quad (3)$$

3.3 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) [8] 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 (4) 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，λ は重みパラメータを表わす．

$$MMR \equiv argmax_{v_i \in D \setminus D'} [\lambda TopicRank(v_i) - (1-\lambda) max_{v_j \in D'} Sim_{cosine}(v_i, v_j)] \quad (4)$$

4 要約精度に基づく性能評価実験

4.1 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．評価指標としては，それぞれの手法によって生成された要約に対して ROUGE [9] を適用する．特に，人間の評価と相関していることが示されている，ROUGE-1 値を用いる [9]．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，まず，文間の類似度を計算するための式 (2) における重みパラメータ α，および，冗長性を削減するための式 (4) における重みパラメータ λ の適切な値について考察する．また，LDA によりトピック推定を行う際に，カテゴリの近い文書群からは類似したトピック分布が得られると仮説を立て，要約対象とする文書の潜在トピックを大量の文書を用いてより正確に推定することを考え，要約対象以外の文書を追加した場合についての実験を行う．追加する文書群については表 1 に示した．なお，文書を追加する前の語彙数は 10091 である．最後に，提案手法である TopicRank および TopicRank(+MMR) を，従来の手法である LexRank と比較する．なお，本実験においてトピック数は 50 とする．LDA において潜在トピックの推定手法としては，ギブスサンプリングを用い，その反復回数は 200 とする．各手法につき 10 回実験を行いその平均を示す．

表 1: 追加する文書データ

コーパス	文書数	追加後の語彙数
DUC2003 Task1	615	15100
DUC2003 Task3	326	13341
DUC2004 Task5	994	15483
上記全て	1609	19121

4.2 実験結果

図 1 に，TopicRank における α の変化に伴う ROUGE-1 値の変化を示す．標準偏差の平均は with, without で 0.0014，0.0013 である．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 λ の変化に伴う ROUGE-1 値の変化を示す．標準偏差の平均は with，without で 0.0011，0.0016 である．λ 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 ともに $\lambda = 0.5$ のときである．図 3 に，LDA によるトピック推定の際に文書データを追加した場合の結果を示す．標準偏差の平均は，with，without で 0.0018，0.0030 である．手法は，前の実験結果より，$\alpha = 1.0$，$d = 0.95$ のときの TopicRank および $\alpha = 1.0$，$d = 0.95$，$\lambda = 0.5$ のときの TopicRank(+MMR) である．いずれにおいても，データを追加せずにトピック推定した場合よりも ROUGE-1 値が高くなっており，特に全ての文書群を

追加した場合に最も高い値を示している．

表 2 に，各手法間の ROUGE-1 値の比較を示す．前の実験結果より，各パラメータに最適な値をあてはめた TopicRank，TopicRank (+MMR) 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．

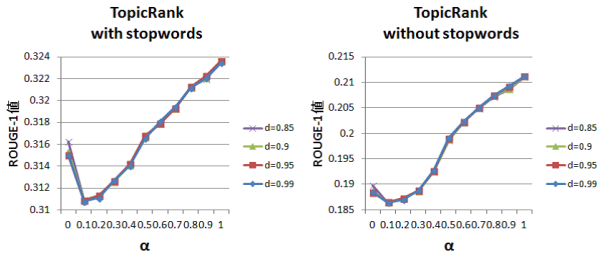

図 1: TopicRank における ROUGE-1 値の変化

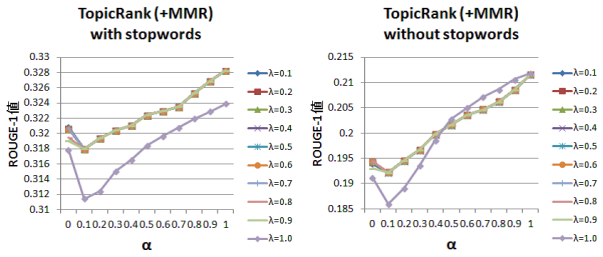

図 2: MMR 導入後の ROUGE-1 値の変化

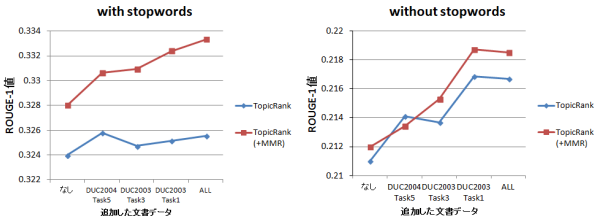

図 3: 文書データ追加後の ROUGE-1 値の変化

4.3 考察

実験結果から，$tfidf$ 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，$d = 0.95$ の場合に ROUGE-1 値が高くなったことに関しては，PageRank や LexRank における $d = 0.85$ と比較すると値が大きく，これは TopicRank における類似度グラフが PageRank で用いられるグラフよりも密であることから，非隣接ノードを考慮する割合である $(1 - d)$ の値が小さい方が良かったのではないかと考える．

表 2: ROUGE-1 値の比較

method	with	without
LexRank	0.222	0.035
TopicRank	0.326	0.217
TopicRank (+MMR)	0.333	0.219

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 α 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．α 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 を導入した後でも，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．

5 要約課題に着目した各手法の比較

5.1 実験設定

提案手法に対するより深い考察を行うために，50 セットの各要約課題について生成された要約の評価を行う．そして，提案手法が特に有効であった文書群や，LexRank の方が高い精度を示した文書群の傾向を調べることにより，提案手法の性質について考察する．手法は LexRank，TopicRank，および TopicRank(+MMR) を対象とし，各手法におけるパラメータの値は 4 章の結果を適用する．各手法，要約課題につき 10 回実験を行い，その平均を示す．

5.2 実験結果

図 4 に，要約課題ごとの各手法の比較を示す．横軸は，DUC2004 の Task2 で与えられた要約課題の番号である．要約課題によって，ROUGE-1 値の高い手法が異なっていることが分かる．標準偏差の平均は LexRank，TopicRank，TopicRank(+MMR) それぞれに対して，0.1006，0.0994，0.0998 である．特に，要約課題 1，3，44，55 では，LexRank よりも TopicRank の方が精度が高く，要約課題 22，1009 では，TopicRank よりも LexRank の方が精度が高くなっている．また，要約課題 40，48，1038 では，TopicRank(+MMR) の精度が高く，冗長性削減が効いていることが分かる．

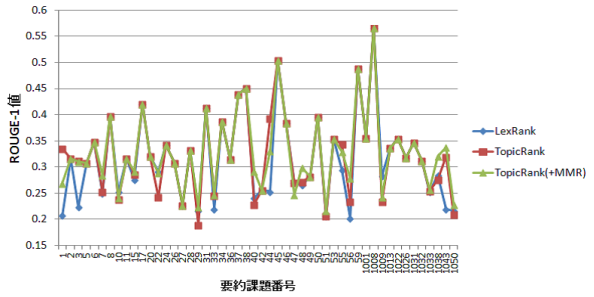

図 4: 要約課題ごとの各手法の比較

5.3 考察

実験結果より特徴の見られた要約課題に対する文書群について調べることにより，提案手法の考察を行う．ここで，LexRank の方が高い精度を示した文書群のセットを D_L で表わし，同様に TopicRank，TopicRank(+MMR) の場合を D_T，D_M で表わす．文書群の特徴として，ここでは各語彙の出現回数の平均を表わす平均単語頻度と，文書群内の文書がどのくらい類似しているかを表わす指標を用いる．後者は式 (5) のように定義し，Ave_{cos} で表わすことにする．D は対象文書群，d は D 内の文書である．

$$Ave_{cos} = \frac{1}{|D|} \sum_{d_1,d_2 \in D \wedge (d_1 \neq d_2)} sim_{cosine}(d_1, d_2) \quad (5)$$

表 3: 各文書群の特徴

文書群	語彙数	平均単語頻度	Ave_{cos}
D_L	630	2.882	0.080
D_T	651	2.864	0.082
D_M	707	3.140	0.095

表 3 に，各文書群の特徴を示した．いずれも政治に関連した新聞記事であり，テーマの傾向の差は見られなかった．表より，D_M では平均単語頻度と Ave_{cos} において高い値を示していることが分かる．これより，文書群に含まれる単語が偏っていて各単語の出現回数が高く，また，類似した文書が多く含まれるような文書群を対象に要約生成を行うような場合には，冗長性削減を考慮した TopicRank(+MMR) が確かに有効に働いたと考えられる．また，D_L と D_T を比較すると，僅差ではあるが，D_T の方が両指標において値が大きくなっていることが分かる．これより，類似した文書群を要約する場合や，ある限定された単語が頻出するような細かいテーマに関する文書群を要約する場合では，表層的な情報よりもトピックに基づいて要約生成を行った方が適していると考える．今回のデータセットでは手法によって精度に差が出たものが少なかったため，顕著な差を発見することはできなかったが，対象とするデータが様々なジャンルを含む場合には，手法による精度の差が現れやすくなり，より深い考察ができるのではないかと考える．

6 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．また，一考察として，類似度の高い文書が含まれる文書群において提案手法が特に効いてくるのではないかという知見を得た．今後の課題としては，要約対象データとしてレビュー等の異なる種類のものを適用し，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 Salience in Text Summarization, Journal of Artificial Intelligence Research, 22, pp. 457–479, 2004.
[2] Ronald Brandow, Karl Mitze and Lisa F. Rau, Automatic Condensation of Electronic Publications by Sentence Selections, Information Processing and Management: an International Journal – Special issue: summarizing text, pp. 675–685, 1995.
[3] Dragomir R. Radev, Hongyan Jing and Malgorzata Budzikowska, Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-based Evaluation, and User Studies, ANLP/NAACL Workshop on Summarization, 2000.
[4] Rada Mihalcea and Paul Tarau, TextRank: Bringing Order into Texts, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 401–411, 2004.
[5] Jahna Otterbacher, Gunes Erkan and Dragomir R. Radev, Using Random Walks for Question-focused Sentence Retrieval, Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 915–922, 2005.
[6] Sergey Brin and Lawrence Page. The Anatony of a Large-scale Hypertextual Web Search Engine, Computer Networks and ISDN Systems, pp. 107–117, 1998.
[7] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet Allocation. Journal of Machine Learning Research, Vol. 3, pp. 993–1022, 2003.
[8] 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.
[9] C. Lin, ROUGE: a Package for Automatic Evaluation of Summaries, In Proc. of the Workshop on Text Summarization Branches Out, pp. 74–81, 2004.