潜在的意味の抽出に基づく文書処理手法の提案とその応用

理学専攻 情報科学コース　北島 理沙　（指導教員：小林 一郎）

1　はじめに

近年，文書の潜在トピックを扱う機会が増え，LSI, pLSI，LDA などの手法が利用されるようになってきた．また，大量の文書データを効率よく処理するための一手法として自動要約の必要性が高まっている．本研究では，潜在的意味に基づいて文書の内容をより捉えた処理を行うことを目的とし，まず，トピックの割り当て対象を単語対に変更した手法を提案し，レビュー記事を用いた実験を通して性能評価を行う．また，潜在的意味に基づいた自動要約手法として，トピックに基づいた文のグラフ表現を用いる複数文書要約手法を提案し，DUC2004 [1] を用いた文書要約を通して先行研究との性能の比較および考察を行う．

2　文書上の単語対を素性とした潜在トピック推定

2.1　提案手法

文書上の各事象を *Event* と呼び，単語の組として表現する．その構成条件としては，次の 4 種類を用意した．同文中で共起する 2 つの語は関連性が高いと考え，1 文中で共起する 2 つの語を組とする．この素性タイプを *Event0* と定義する．また，係り受け関係にある 2 つの自立語の共起を組とする．この素性タイプを *Event1* と定義する．*Event1* の中には，重要でない組み合わせも存在すると考えられる．したがって，必要と考えられる組み合わせを経験的に定めることとし，(主語，述語) , (述語 1，述語 2) の条件を満たす係り受け関係にある組を抽出する．この素性タイプを *Event2* と定義する．(名詞，名詞) , (名詞，形容詞) , (動詞，名詞) , (動詞，副詞) のいずれかを満たす係り受け関係にある語の組み合わせをイベントとして抽出する．この素性タイプを *Event3* と定義する．上記，各 *Event* を構成する際の構文解析には構文解析器 CaboCha を使用した．イベント－文書行列の作成後，LDA [1] によるトピック推定を行う．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る．

2.2　実験

2.2.1　実験設定

共通の文書検索課題を通じて素性タイプを評価する．トピック分布の類似度指標には，Jensen-Shannon 距離を用いる．対象データは楽天トラベル[2]のホテル・施設に関するレビュー・評価データとし，その中から無作為に選んだ 4000 件を対象文書群とする．トピック数はパープレキシティの値によって決定することとし，予備実験を行った．予備実験の結果を表 1 に示す．使用するクエリは「サービスが良かった」「料理が良かった」「部屋が良かった」「立地が良かった」とし，その平均を調べる．評価指標には 11 点平均適合率を使用する．

表 1: パープレキシティの比較

素性タイプ	語彙数	$K=5$	$K=10$	$K=50$	$K=100$	$K=150$	$K=200$
Word	7620	1201	1083	975	1112	1258	1382
Event0	211136	58233	50887	45238	51631	57402	60707
Event1	31459	9103	7748	7940	9621	10824	11871
Event2	14704	5070	4099	4666	6286	7480	8346
Event3	23644	8052	6647	7088	9197	10809	12225

2.2.2　実験結果および考察

表 2 に，各素性タイプによる 11 点平均適合率の変化を示す．文内の自立語の共起を用いた *Event0* で，最も良い性能となっていることが分かる．また，経験則 1 を満たす係り受け関係を用いた *Event2* では性能が低く，従来の単語を単位とした LDA を下回っていることが分かる．

表 2: 11 点平均適合率の比較

素性タイプ	11 点平均適合率
Word	0.7453
Event0	0.7967
Event1	0.7657
Event2	0.7345
Event3	0.7661

実験結果から，文内の自立語の共起関係を素性に使用することが，潜在的トピック推定において有効であることが分かる．これは，文書を単語集合として扱った場合に無視されてしまった，同文中における単語の依存関係が，潜在的トピックに大いに寄与していることを意味する．一方，係り受け関係を用いた場合は少し精度が下がることから，単語対の構成条件が厳しすぎたのではないかと考える．

3　トピックを考慮したグラフ表現に基づく複数文書要約

3.1　提案手法

要約手法の一つとして，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ており，特に LexRank [2] 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グラフに基づいて文の重要度を計算する．これに対して我々は，文のトピック分布に基づいて計算される文間類似度によって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．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 のトピック分布である．トピック分布推定には LDA [1] を用い，類似度指標には Jensen-Shannon 距離を用いる．α は，潜在的類似度と表層的類似度の重みパラメータである．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．ここで，文 u の重要度は式 (2) で求める [2]．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 は制動係数である．文の重要度は反復的に計算されるため，これ

[1] http://www-nlpir.nist.gov/projects/duc/guidelines/2004.html
[2] http://travel.rakuten.co.jp/

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．

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，λ は重みパラメータを表わす．

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

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

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

3.2 実験

3.2.1 実験設定

対象データには DUC2004 で使われた 10 件の新聞記事群 50 セットを用い，評価指標には ROUGE-1 値を採用する．まず，TopicRank においてパラメータ α と d の値を変化させ，次に，MMR を導入した後のパラメータ α と λ の値を変化させる．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 とし，各手法につき 10 回実験を行いその平均を示す．

3.2.2 実験結果および考察

図 1 に，TopicRank における α 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 導入後の λ の変化に伴う ROUGE-1 値の変化を示す．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 $\lambda = 0.5$ のときである．表 3 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．TopicRank は LexRank よりも高い ROUGE-1 値を示したが，一方で，MMR を導入したことでの精度の差はあまり見られず，TopicRank に対する冗長性削減の効果は小さいことが分かる．

実験結果より，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，表層的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，MMR 導入後の結果に着目すると，$\lambda = 1.0$ の場合には，with では α が大きくなるにつれて冗長性削減を考慮したときの差が小さくなり，without においても $\alpha > 0.4$ のときに高い値を示している．α が大きいことはトピック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．

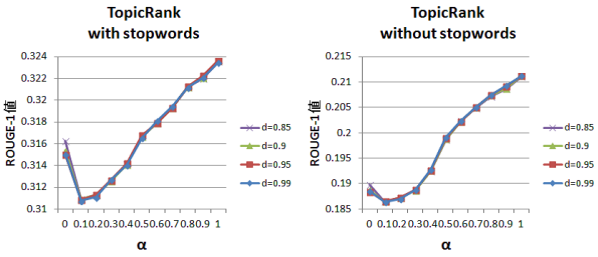

図 1: TopicRank における ROUGE-1 値の変化

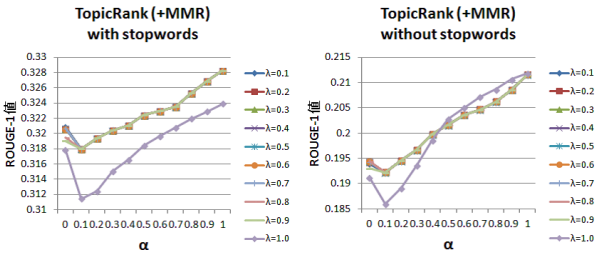

図 2: MMR 導入後の ROUGE-1 値の変化

表 3: 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

4 おわりに

本研究では，潜在的意味に基づいた文書処理手法として，トピックの割り当て対象を単語対に変更した手法を提案し，文書の内容をより捉えた文書検索が行えることを示した．また，潜在的意味に基づいた自動要約手法として，トピックに基づいた文のグラフ表現を用いる複数文書要約手法を提案し，先行手法よりも精度が高く，冗長性の低い要約生成が行えることを示した．今後の課題としては，今回用いたデータと異なる種類の文書データを対象に実験を行って提案手法へのより深い考察を行いたいと考えている．

参考文献

[1] D.M. Blei, A.Y. Ng, and M.I. Jordan : Latent Dirichlet Allocation, Journal of Machine Learning Research, vol.3, pp.993–1022, 2003.

[2] G. Erkan and D. R. Radev, : LexRank: Graph-based Lexical Centrality as Salience in Text Summarization, Journal of Artificial Intelligence Research, pp. 457–479, 2004.

[3] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz : Multi-document summarization by sentence extraction, Proc. of ANLP/NAACL Workshop on Automatic Summarization, vol.4, pp.40–48, 2000.