教師あり学習モデルの適用範囲及びグラフに基づく要約手法の拡張

鈴木聡子　(指導教員：小林一郎)

1　はじめに

情報技術の発達により，膨大なデータの蓄積・閲覧が可能となり，情報検索や自動要約の必要性が高まっている．これらの精度を向上させるために，潜在情報に着目した研究が盛んに行われている．トピックモデルの代表的な手法である Latent Dirichlet Allocation(LDA)[1] に基づいて提案された Labeled LDA(L-LDA)[2] では，テキストに付与されているタグを利用し，LDA を上回る性能が得られる事が報告されている．本研究では，テキストから得られる情報より疑似ラベルを生成し，タグ付きデータでなくても適用可能な手法を提案し，文書分類を通して性能評価を行う．また自動要約においては，対象データとして新聞記事が用いられることが多い．しかし新聞記事には短期的な話題と長期的な話題が存在し，後者においては，話題の概要が分かるだけでなく，時間に伴う内容の変化も把握できる要約が望ましい．よって，長期的な話題に対して前述のような欲求を満たす要約生成を目的とした要約手法を提案し，実験および評価を行う．

2　疑似ラベルによる潜在ディリクレ配分法

2.1　提案手法

L-LDA では，トピック分布の推定において，文書に付与されたタグの情報を教師情報とし，射影行列を生成することによってハイパーパラメータ α を制限する．本研究では，文書の 2 つの表層的な情報から教師情報となる疑似ラベルを生成する．1 つは，単語の共起情報である．文書の潜在意味の一貫性は単語の共起と関係があるということから，共起性の強い単語より疑似ラベルを生成することを考える．この方法では，TF-IDF 値の高い単語の自己相互情報量（PMI）を共起性の指標とし，閾値以上の PMI を持つ単語のグループを疑似ラベルとする．また，PMI 値は低いが出現頻度の高い単語もラベルとして採用する．そして，疑似ラベルを構成する単語が抽出されている文書に同じラベルを与える．ただし，複数の単語で構成されている場合は，対象単語が複数抽出された場合のみ，ラベルを与える．もう 1 つは，文書の類似度である．類似性の高い文書に同じ疑似ラベルを与える．ここでは，Leader-Follower 法と Crouch 法の 2 つの方法において疑似ラベルを生成する．両分類法は，1 文書に対し複数のラベルを与えることが可能である．

2.2　実験

文書分類課題を通じて各手法と LDA との比較を行う，対象文書として，20Newsgroups の 20 カテゴリの内，10 個のカテゴリから成る文書集合を用意した．単語の共起情報に基づく手法（パターン 1）では，抽出する単語数を各文書ごとに TF-IDF の値が上位 30 単語と 50 単語と設定して実験を行った．また，1 単語から構成されるラベル数は，出現回数が上位 5 位までの単語を選んだ．文書の類似度に基づく手法（パターン 2）では，類似度の閾値は，[0.1,0.9]（2a）と 0.1 以下（2b）において実験を行った．ハイパーパラメータの値は，α=0.1，η=0.1 とした．文書のトピック分布 θ から，k-means 法を用いて分類を行った精度により提案手法を評価する．評価には，正規化した相互情報量 $\widehat{MI}$ を用いる．

2.3　結果と考察

実験結果を図 1〜3 に示す．単語の共起情報に基づく手法では，全ての閾値で LDA と比べ低い精度を示した．抽出単語数による結果の変化は見られない．この手法は，評価結果が文書集合の影響を大きく受けやすく，生成した疑似ラベルにはトピックの情報が反映されにくいと考えられる．また，文書数に偏りがあった場合には，良質な疑似ラベルの生成が，さらに困難になることが考えられる．文書の類似度に基づく手法では，Crouch 法を用いた方が Leader-Follower 法を用いた場合よりも，全体的に安定した評価値を得ている．最も良い精度は，Leader-Follower 法を用い，閾値を小さく設定した場合に得られている．実験より，文書の類似関係によるラベル生成の方が単語の出現による精度の依存が小さく，良い精度が得られることが分かった．

3　時系列文書を対象とした要約

3.1　提案手法

本研究では，各文の重要度を決定するためにグラフ構造を用いる．まず，文書集合 $D_t \in \boldsymbol{D}$ について考える．t は時刻単位を表し，t=$\{$ 1,$\cdots$,T $\}$ である．ここで，D_t は時刻 t に属する文書集合を表す．本研究では，時間が経過するとともに新しく文書が追加されることを想定する．Algorithm1 に要約を生成する手順を示す．

Algorithm 1 要約のプロセス

Input: $\boldsymbol{D}$, $\boldsymbol{S}$, ϵ, l
$\boldsymbol{S} = \{\ \}$
$\epsilon \leftarrow$ threshold
for $t = 0$ to T **do**
　$\boldsymbol{S}' \leftarrow \boldsymbol{S} + D_t$
　ranking $\boldsymbol{S}'$ with LexRank
　if length of $\boldsymbol{S}' > \epsilon$ **then**
　　$\boldsymbol{S} \leftarrow$ top ϵ sentences of $\boldsymbol{S}'$
　else
　　$\boldsymbol{S} \leftarrow \boldsymbol{S}'$
　end if
end for
return top l sentences of $\boldsymbol{S}$

入力として，$\boldsymbol{D}$, $\boldsymbol{S}$, ϵ, l を与える．ここで，$\boldsymbol{S}$ は出力する要約の候補となる文集合，ϵ は閾値であり，l は要約として出力する文の数である．文集合 $\boldsymbol{S}'$ に含まれる文で構成されるグラフを考える．グラフの各頂点は各文を表し，エッジは文同士の関係を表している．文のランキングアルゴリズムには [3] で提案される LexRank アルゴリズムを用いた．

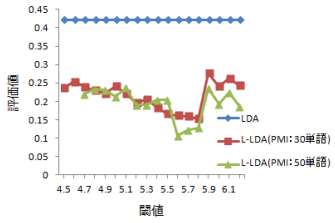

図 1: $\widehat{MI}$ パターン 1

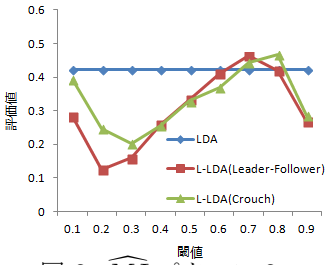

図 2: $\widehat{MI}$ パターン 2a

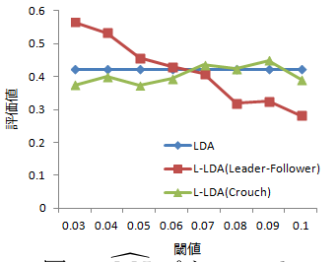

図 3: $\widehat{MI}$ パターン 2b

提案手法では，グラフの大きさを制限する．グラフの大きさが設定した閾値を超えた場合には，閾値以下の大きさにグラフを縮小する．常に閾値以下のグラフサイズを保った状態で文のランキングを行い，要約が必要なタイミングでスコアの高い文を抽出する．

3.2 実験

対象データには，Tran[4] らが提供しているタイムライン要約のためのデータセットを用いる．比較のためのベースラインとして，ランダムに抽出したものと LexRank を用意する．なお，正解要約は最終時におけるものとする．全てのシステムにおいて，生成する要約の長さは，各トピックにおける正解要約の文の長さと等しいものとした．また，前処理として ‘a’ や ‘the’ といったありきたりな語であるストップワードの除去と，語尾の異なるものを同一とみなすためのステミング処理を全てのシステムにおいて行った．評価には，ROUGE-1 の再現率と F 値を用いる．また，評価の際にはストップワードの有無の両パターンにおいて実験を行った．

3.3 結果と考察

グラフサイズを固定値とした結果を表 1 に示す．

表 1: グラフサイズを固定値とした結果

	with		without	
	Recall	F 値	Recall	F 値
random	0.509	0.464	0.314	0.304
Lexrank	0.620	0.434	0.430	0.328
提案手法 A	0.607	0.473	0.424	0.340
提案手法 B	0.658	0.490	0.495	0.377

提案手法 A は，グラフサイズが全体のグラフサイズの 1/2 程度のものであり，提案手法 B は 2/3〜3/4 程度の大きさのものである．提案手法 B において高い精度を得られたことから，不必要な情報の削除による効果を確認できた．

次に，グラフサイズを比率とした実験結果を図 4〜7 に示す．この実験では，グラフを更新する期間を 1，3，7 日と変化させた．結果から，グラフを比率に設定した場合には LexRank と比べ精度が良くないことが分かる．これは，グラフサイズが小さい状態でのランキング結果においてもグラフを制限してしまうため，精度が落ちていると考えられる．また，ストップワードの有無による LexRank との精度の差が大きいことから，提案手法では，ストップワードを多く含む文を抽出してしまっている可能性が考えられる．グラフの更新期間による精度の差は，比率が小さい程顕著である．

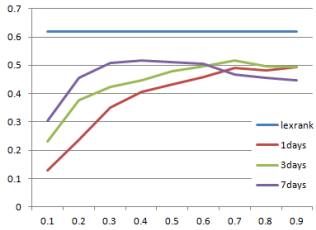

図 4: Recall/with

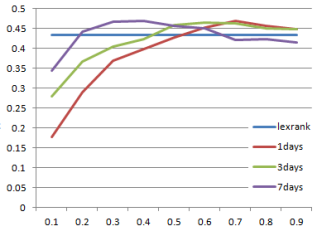

図 5: F 値/with

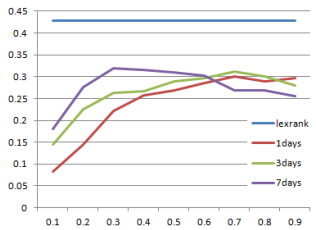

図 6: Recall/without

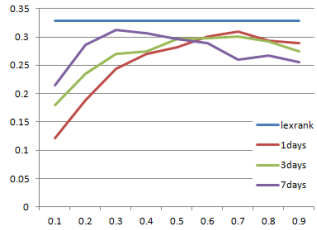

図 7: F 値/without

4 おわりに

本研究では，教師あり学習モデルである L-LDA に対して，疑似ラベルを与えることにより，適用範囲を拡張した．その結果，教師なしモデルである LDA を上回る精度により，疑似ラベルの有用性が確認できた．また，時系列的な文書に対してグラフを用いた要約手法を提案し，実験を行った．結果，グラフサイズが不十分な状態でのノードの制限は，後の要約に悪い影響を与えるが，ノードを制限することによる精度向上を確認することができた．グラフサイズの設定に関して，更なる考察が必要であると同時に，他の手法との比較や途中時間における要約の評価を今後の課題とする．

参考文献

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

[2] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning: Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora, EMNLP2009, pp. 248-256, 2009.

[3] G. Erkan and D. R. Radev, LexRank: Graph-based Lexical Centrality as Salience in Text Summarization, Journal of Artificial Intelligence Research, pp. 457-479, 2003.

[4] G. B. Tran, M. Alrifai, and D. Q. Nguyen, Predicting Relevant News Events for Timeline Summaries, In Proceedings of the 22nd international conference on World Wide Web Companion, pages 91-92, 2013.