潜在トピックの重要度を考慮した要約文の生成

重松遥　(指導教員：小林一郎)

1　はじめに

文書要約の分野では，各々の文書に特有の情報や，共通の基本的な情報などをまとめて要約することができる複数文書要約に注目が集まっている．また，文書要約の手法としては，重要文抽出によるものが主流であり，この手法では，文の重要度を測る基準をどのように設定するかが重要となる．

本研究では，トピックモデルと呼ばれる文書の確率的生成モデルを用いて，文書セット内に潜在しているトピックを抽出し，個々の文に含まれるトピックの比率に基づいて文の重要度を定義する．そして，整数計画法を用いて，総重要度が高くなるような文の組み合わせを要約として出力する複数文書要約手法を提案する．

2　潜在トピックに基づく重要文決定

文書群中の潜在トピックを確率的に求めるトピックモデルとして Latent Dirichlet Allocation(LDA)[1] がある．このモデルでは，文書中に存在している単語は，文書に潜在的に存在しているいくつかのトピック（話題）に基づいて生成されるとする．各トピックは単語出現確率として表され，文書群中に存在している総単語に対して，総和が 1 となる出現確率が割り当てられる．また，トピック自身にも文書内において出現確率の総和が 1 となるトピック比率として確率が付与される．

文 i の重要度 b_i は，トピック t に対する文 i の重要度 b_{ti} の総和として定義する．b_{ti} は，文 i を構成している単語の重みを総和したもの W_{ti} に，文長を考慮した係数として単語の総数 N_i の平方根の逆数を掛けている．

$$b_{ti} = \frac{W_{ti}}{\sqrt{N_i}} \quad , \; b_i = \sum_t b_{ti}$$

複数文書要約は単一文書要約とは異なり，文書セット中に内容が重複した文が出てくる場合が多い．どの文書にも書かれているような文が重要文だとみなされた場合，ただ重要文抽出をするだけでは，生成された要約は類似する文の集合で構成されてしまう．そのため，文間の冗長をなるべく避けた重要文抽出が必要となる．ここでは高村ら [2] を参考に，文間の被覆関係を考慮し，生成される要約文の冗長性を回避する (3.3 節に詳述)．

3　実験

3.1　実験仕様

本実験では，評価型ワークショップ TSC3 で用いられたテストセット [3] を利用する．テストセットには，話題の異なる 30 の文書セットが用意されており，1 文書セットあたり約 10 記事から成っている．各文書セットには，短い要約と長い要約の正解要約が複数示されており，これに基づいて生成した要約を評価する．

生成された要約の評価方法としては，Precision(精度) と Coverage(被覆度)[3] を用いる．Precision は生成した文集合の内，正解要約集合に含まれる文の割合であり，Coverage は生成した文集合の冗長度合いを考慮しつつ，その文集合がどれだけ正解要約の内容に類似しているかを測る指標である．30 文書セット全てに対し実験を行い，Precision と Coverage の平均を求める．抽出する文数は，最小の文数で最大の情報を伝えることが望ましいとの考え [3] より，正解要約例の中で最も少ない要約文数を指定する．トピック数はパープレキシティによって調べ，トピックの推定にはギブスサンプリングを用い，反復回数は 200 回とした．

3.2　要約生成におけるトピック比率の考慮

30 文書セット中，パープレキシティによりトピック数が 10 個と推定された 5 つの文書セットに対してトピックを抽出し (トピック比率が高い順に topic0>topic1>…>topic9 とする)，そこで抽出されたトピックが正解要約中にどのような比率で入っているのかをグラフに表した．図 1 に 5 つの文書セットの平均を示す．

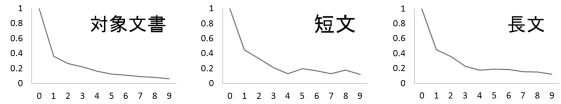

図 1: 5 文書セットの平均トピック比率

図の左から順に，要約対象文書セットのトピック比率，短い正解要約のトピック比率，長い正解要約のトピック比率を表したグラフとなっている．このグラフより，文書セット中で比率が高いトピックほど正解要約中にも多く含まれる傾向があることが分かる．そこで，対象文書セットと同じような割合で正解要約にもトピックが割り当てられていると推測し，各トピックの単語出現確率にそのトピックの比率を掛けることで，トピックの重要度に応じた単語の重み付けをする．

3.3　整数計画法を用いた要約文生成

上述した内容を重要文抽出の条件に反映させ，整数計画法を用いて，文書セット中の全文から以下の制約条件を満たす文の組み合わせを探す．

制約条件：

1. 指定した文数だけ文を選択する

$$S_i \in \{0, 1\}; \; \forall i \tag{1}$$

$$\sum_i S_i = d \tag{2}$$

S_i は，文 i が要約文として選択されているときは 1，そうでないときは 0 となるような決定変数とする．S_i の総和を，指定された文数 d にするよう制約を与える．

2. 冗長文の考慮

$$Z_{ij} \in \{0, 1\}; \; \forall i, j \tag{3}$$

$$Z_{ij} \leq S_i; \; \forall i, j \tag{4}$$

$$\sum_i Z_{ij} = 1;\ \forall j \tag{5}$$

$$Z_{ii} = S_i;\ \forall i \tag{6}$$

できるかぎり文書セット全体を被覆しているような文集合を選択するために，文書セット中の全文を，要約文として選択された文のいずれか一つに被覆させる．Z_{ij} は，文 j が文 i に被覆されているとき 1，そうでないとき 0 の決定変数とする．よって，$Z_{ij} = 1$ のときは文 i が要約文として選択されている必要があり，これは制約式 (4) で表される．また，制約式 (5) より，全文がいずれか 1 つの文に被覆されることが保証される．さらに，制約式 (6) は，選択された文はその文自身に被覆されることを意味する．

目的関数：

- 重みと被覆度が大きい文集合を選択する

$$\sum_{i,j}(e_{ij}b_j)\cdot Z_{ij} \rightarrow max \tag{7}$$

e_{ij} は文 i が文 j を被覆する度合いとし，

$$e_{asy,ij} = \frac{|W_i \cap W_j|}{|W_j|} \quad , \quad e_{sym,ij} = \frac{|W_i \cap W_j|}{|W_i \cup W_j|}$$

という式のどちらかを適用する．$e_{asy,ij}$ は，文 i と文 j について非対称の被覆度で，W_i は文 i を構成する単語の集合である．また，$e_{sym,ij}$ は被覆を対称化したものである．

この被覆度に，文 j の重要度を表す b_j を掛けることによって，冗長性を考慮した重要文抽出ができる．また，文の重要度を決める際に，各トピックの重要単語上位 30 個のみを用いても総単語を用いた際とほぼ重要度の値が変わらないことが予備実験で確認できたことから，30 単語のみを用いて計算時間の削減を実現した．

3.4 実験結果と考察

上記で提案した冗長性を考慮した要約生成手法と共に，冗長性を考慮しない場合についても実験を行った．冗長性を考慮しない要約文生成については，上で示した制約式の式 (3)(4)(5)(6) を外し，$\sum_i S_i b_i$ が最大となるような目的関数を設定した．また，各手法とも 30 単語を考慮した場合と総単語を考慮した場合の 2 種類について結果を求めた．

30 文書セットの Precision と Coverage，計算時間 (秒) を求め，平均した結果を表 1 に示す．

表 1 の結果から，冗長性を考慮しない手法については Precision の値に対して Coverage が低い値となっており，抽出した文集合が冗長であることを示している．一方，冗長性を考慮した手法では Coverage の値が高くなり，冗長な文を削減できていることが分かる．

次に，文間の被覆度が非対称なモデル $e_{asy,ij}$ と対称なモデル $e_{sym,ij}$ について比較する．表 1 より，非対称なモデルの方が対称なモデルよりも Coverage が上回っており，より冗長性が削減できていることが分かる．これは，非対称なモデルの方が，被覆の方向性を明確に捉えているためと考えられる．

表 1: 提案手法の要約精度

手法	長さ	Prec	Cov	計算時間 (秒)
提案手法 (asy，30)	Short	.455	.408	193
	Long	.587	.480	194
提案手法 (sym，30)	Short	.476	.398	193
	Long	.551	.421	194
提案手法 (asy，総)	Short	.460	.425	560
	Long	.588	.483	567
提案手法 (sym，総)	Short	.481	.418	554
	Long	.552	.432	562
提案手法 (冗長考慮無，30)	Short	.563	.318	14
	Long	.591	.340	14
提案手法 (冗長考慮無，総)	Short	.545	.329	14
	Long	.588	.341	14
eventLDA	Short	.418	.340	-
	Long	-	-	-
Lead	Short	.426	.212	-
	Long	.539	.326	-
TF-IDF	Short	.497	.292	-
	Long	.604	.325	-

また，考慮する単語数を比較すると，30 単語の場合は総単語の場合より，計算時間が約 3 分の 1 に短縮でき，かつ，Precision と Coverage の値にはあまり差が見られなかった．この結果より，考慮する単語を 30 個に限っても精度を落とさず計算時間を削減することができると分かった．

更に，各文書の先頭から順に 1 文ずつ重要文としてとってくる Lead 手法，および TF-IDF によって求めた単語重要度の和で文のスコアを定義する TF-IDF 手法による実験結果 [3]，北島らによって提案された eventLDA[4] による実験結果との比較を行った．Precision の値はどの手法ともあまり差が見られないが，Coverage の値は本研究で提案した手法が一番高くなっていることを確認した．

4 おわりに

本研究では，LDA により文書セット中の潜在的トピックを抽出し，抽出されたトピックに基づく複数文書要約の提案を行った．文書セット中のトピック比率と類似したバランスで要約文中にもトピックが含まれていることが分かり，トピック比率を考慮した要約の妥当性を示した．実験により，非対称の被覆度を採用し，単語数を 30 として生成した要約の精度が，計算時間も含めて相対的に良い結果になることが分かった．これにより，考慮する単語数の削減や冗長性回避の効果が分かり，提案手法が他の手法より高い性能をもつことを示すことができた．今後は，重要文抽出と冗長性削減のバランスを考えて，再度，制約条件を検討し，さらに性能が高い要約生成手法の開発を目指す．

参考文献

[1] David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, p. 2003, 2003.

[2] 高村大也, 奥村学. 施設配置問題による文書要約のモデル化. 人工知能学会論文誌 25(1), pp.174-182, 2010.

[3] 平尾努, 奥村学, 福島孝博, 難波英嗣. Tsc3 コーパスの構築と評価. 言語処理学会年次大会発表論文集, pp. A10B5–02, 2004.

[4] 北島理沙, 小林一郎. 文書内の事象を対象にした潜在的ディリクレ配分法による要約. *DEIM Forum 2011*.