4I1-R-9-1

潜在トピックの比率に基づく文書要約手法の提案

Text Summarization based on Latent Topic Distribution

重松 遥　　小林　一郎
Haruka Shigematsu　Ichiro Kobayashi

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
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

In recent years, analyzing texts with probabilistic topic models has become an active area of research. Latent Dirichlet Allocation (LDA) is one of the well-known topic models. This study describes multi-document summarization based on latent topics extracted from multiple documents by means of LDA. By the result of a preliminary experiment to investigate the relation between target documents to be summarized and their correct summaries from a viewpoint of topic distribution in documents, we confirmed that the topic distribution of target documents is roughly the same as that of correct summaries of the documents, so we define the important degree of a sentence by taking account of the topic ratio in the sentence. In the case of summarizing multiple documents, there are many cases where some sentences share their contents with other sentences. We therefore generate a summary based on sentence extraction considering the coverage among sentences by means of integer programming. We confirmed the validity of our proposal technique by comparison with the previous researches.

1. はじめに

文書要約には，大別すると単一文書要約と複数文書要約の二種類が挙げられる．前者は，個々の文書に対しそれぞれ要約するもので，後者は同じ話題について書かれた複数文書をまとめて要約するものである．特に近年，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，文書要約の代表的な手法として、重要文抽出によるものがある．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，近年盛んに研究が行われている，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，複数文書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算出する．そして，整数計画問題を用いて，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. 関連研究

トピックモデルを用いた要約文手法は，Blei ら [4] による LDA の提案以降，数多くの研究がなされてきている．そのアプローチも様々であり，トピックを抽出する手法に工夫を加えたものとして，Chang ら [5] や Wang ら [6] は，LDA の確率割り当て対象を単語から文に変え，トピックに基づく複数文自動要約手法を提案している．北島ら [11] は，確率割り当て対象を文内の係り受けに基づく単語の組に変更した要約手法を提案している．また，トピック抽出のモデルに対する工夫として，Barzilay ら [2] や Einsensteing ら [7] は，文書内容を表現する構造的なトピックモデルの提案に基づく文書要約を実現している．他にも，Haghighi ら [1] は，文書セット内の複数のサブトピックを発見するため，階層型 LDA[3] に基づくトピックモデルを利用した要約文生成手法を提案している．

連絡先: 重松遥，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小林研究室，
〒 112-8610　東京都文京区大塚 2-1-1，03-5978-5708，
shigematsu.haruka@is.ocha.ac.jp

要約文生成には，重要文抽出に基づく手法を採用する研究が現在主流を占めており，重要文抽出に最適化の手法が多く適用されている．

要約文生成における最適化手法の適用は様々な試みがなされており，高村ら [8] は，文書の内容をより含意するような文の組み合わせを最適な要約と定義し，整数計画法を利用することで，含意度が最も高くなるような要約を生成した．この際，文書の始めの部分に重要な文が出現しやすいという考えにより文の重みを定義し、含意度と文の重みを掛け合わせたものが最大となる文集合を要約とした．また，平尾ら [9] は，文抽出と文短縮を同時に行い，スコアを最大にする組み合わせを選択する最適化問題として要約問題を捉え，動的計画法を用いてそれを解決する手法を提案している．

本研究においては，複数文書内に潜在しているトピックに基づき文の重みを定義する．そして，要約生成を整数計画問題として捉えることで，重要箇所をおさえつつ，内容の重複を避けた重要文抽出を行う．

3. 潜在的トピックによる重要文決定

3.1 潜在的トピック抽出

複数文書内の潜在的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)[4] がある．このモデルでは，文書中に存在している単語は，独立に出現しているのではなく，文書中に潜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．つまり，文書はトピックの集まりによって構成されており，トピックごとに出現しやすい単語があると考える．各トピックは単語出現確率として表され，複数文書内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セット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．

3.2 重要文決定

単語出現確率に基づき，トピックごとに文の重みを決定する．そして，全てのトピックに対する文の重みの総和をその文の最終的な重みとし，この重みが大きくなるような文を重要文

とする．文 i の重み b_i を求める場合，まずトピック t に対する文 i 重み b_{ti} を調べる．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，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．この係数は，文長に左右されない文の重み付けを行うためのものである．そして，b_{ti} の総和を最終的な文 i の重みとする．

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

3.3 冗長文の考慮

複数文書要約は単一文書要約とは異なり，文書群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した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 [8] を参考に，文間の含意関係を考慮し，生成される要約文の冗長性を回避する (4.4 節に詳述)．

4. 実験

4.1 実験仕様

本実験では，評価型ワークショップである TSC3 で用いられたテストセット [10] 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 文書セットあたり毎日新聞と読売新聞がほぼ同数入った約 10 記事から成っている．各文書セットには，長い要約と短い要約の正解要約例が複数示されており，これに基づいて生成された要約を評価する．

生成された要約の評価方法としては，Precision(精度) と Coverage(被覆度)[10] 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例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 [10] 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．

4.2 要約生成におけるトピック比率の考慮

実験を行う前に，文書セット中の潜在的トピックが，正解要約中にどのような比率で入っているのか調べる．

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 {0460，0520，0560，0610，0650} 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それぞれの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 1-5 に上記 5 つの文書セットの結果，図 6 に 5 つの文書セットの平均を示す．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．このグラフを見ると，多少の凹凸はあるものの，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．さらに，図 6 の 5 つの文書セットを平均したグラフを見ても，対象文書セットのトピック形状と正解要約のトピック形状が似ていることが分かる．このことから，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できる．この結果より，トピック比率を考慮した要約生成を考える．

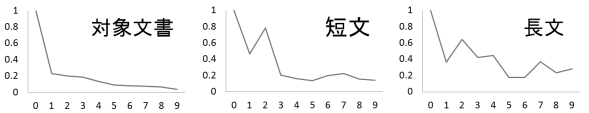

図 1: 文書セット 0460 のトピック比率

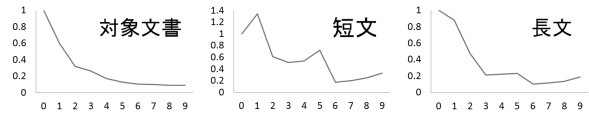

図 2: 文書セット 0520 のトピック比率

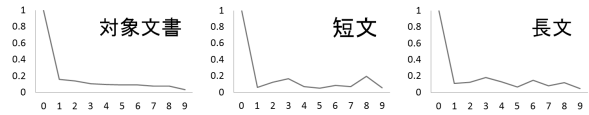

図 3: 文書セット 0560 のトピック比率

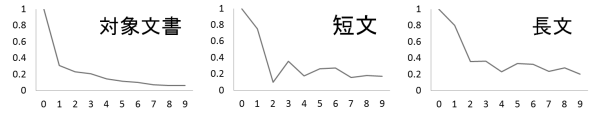

図 4: 文書セット 0610 のトピック比率

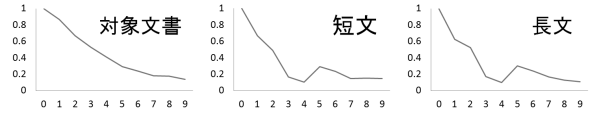

図 5: 文書セット 0650 のトピック比率

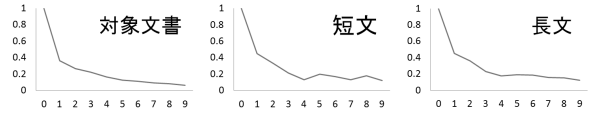

図 6: 5 文書セット平均のトピック比率

4.3 制約条件に反映される仕様の決定

4.2 節で示したとおり，文書セットと正解要約のトピック比率が類似していることより，要約文中のトピックの重要度は文書セット内のトピック比率に応じて決まると仮定する．そこで，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．

また，トピックを代表させる単語の個数を減らすことにより計算時間の無駄を省くことができると考え，文の重みを決定する際に重要単語の上位何単語を考慮するか決める．そこで，上位 10 個を考慮した場合と，20 個，30 個，40 個，50 個，100 個，総単語数の 1/3，1/2，総単語を考慮した場合の正解要約中のトピック形状を見る．

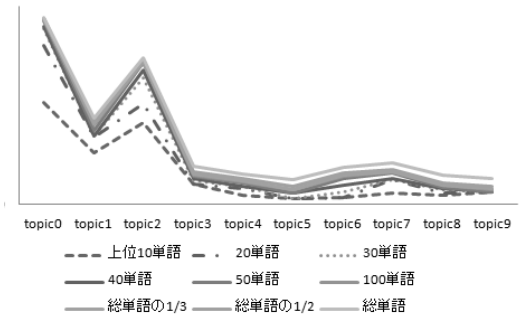

図 7: トピックを代表する重要単語の調査

文書セット 0460 の短い正解要約について分析した結果を図 7 に示す．各トピック 20 単語以下を考慮している場合は他の場合と比べて差が見られるが，30 単語以上を考慮した場合は

ほとんど値に大きな差が見られないことが分かる．つまり，正解要約の多くは，各トピックの重要単語上位 30 個で構成されていることが推測される．よって，本実験では，トピックを代表する 30 個の単語のみを考慮し，それ以外の単語は重みを 0 として考慮しない．なお，他の文書セットに対しても同様の分析をしたところ，どれも 30 単語以上を考慮したときの値に大きな差は見受けられなかった．

4.4 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の組み合わせを探すために，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，

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

という式で表す．$e_{asy,ij}$ 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，$e_{asy,ij}$ を対称化した

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

も検討する．

更に，単語の表層的な一致ではなく，潜在トピックの観点で被覆度を測ることにし，

$$e_{asy,ijt} = \sum_t \frac{|b_{ti} \cap b_{tj}|}{|b_{tj}|}$$

という文間のトピックの被覆度を考慮した式も試す．

これらの被覆度に，文 j の重みを表す b_j を掛けることによって，冗長性を考慮した重要文選択ができる．

4.5 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．

表 1: 提案手法の要約精度

手法	長さ	Prec	Cov	時間 (秒)
提案手法 ($e_{asy,ij}$，30)	Short	.455	.408	193
	Long	.587	.480	194
提案手法 ($e_{sym,ij}$，30)	Short	.476	.398	193
	Long	.551	.421	194
提案手法 ($e_{asy,ij}$，総)	Short	.460	.425	560
	Long	.588	.483	567
提案手法 ($e_{sym,ij}$，総)	Short	.481	.418	554
	Long	.552	.432	562
提案手法 ($e_{asy,ijt}$，30)	Short	.442	.335	425
	Long	.593	.422	430
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
TF-IDF(整数計画法)	Short	.398	.376	169
	Long	.516	.460	169
TF-IDF(クラスタリング) [平尾ら]	Short	.417	.299	-
	Long	.528	.358	-
TF-IDF(MMR) [平尾ら]	Short	.454	.305	-
	Long	.553	.374	-
TF-IDF	Short	.497	.292	-
	Long	.604	.325	-
Lead	Short	.426	.212	-
	Long	.539	.326	-
eventLDA [北島ら]	Short	.418	.340	-
	Long	-	-	-

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．また，文間のトピックの被覆度を考慮した $e_{asy,ijt}$ についての結果を見てみると，文を表層的に捉えた被覆度である $e_{asy,ij}$ と比べて Coverage の値が下がり，要約の精度が落ちている．よって，要約生成の際は，文書の潜在的な面，表層的な面を上手く組み合わせることが重要だと推測できる．

次に，文の重みを測る際の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

生成要約文 (冗長考慮無，30)

$\{S_1, S_4, S_7, S_{10}, S_{55}, S_{62}, S_{73}\}$　Precision 1.000 / Coverage 0.333

【ロンドン３日共同】「そこに山があるから」――の名文句で知られ、エベレスト（チョモランマ）でなぞの死を遂げた英国の登山家ジョージ・マロリー氏の遺体が７５年ぶりにエベレスト山中で見つかったことが３日、米国の「マロリー＝アービン捜索隊」のホームページで明らかになった。マロリー氏は１９２４年６月、英国登山隊の一員としてエベレスト初登頂に挑戦。【ロンドン３日共同】「そこに山があるから」――の名文句で知られ、エベレスト（チョモランマ）でなぞの死を遂げた英国の登山家ジョージ・マロリー氏の遺体が７５年ぶりにエベレスト山中で見つかったことが３日、米シアトルに拠点を置く「マロリー＝アービン捜索隊」のホームページで明らかになった。マロリー氏は１９２４年６月、英国登山隊の一員としてエベレスト初登頂に挑戦。【カトマンズ２５日ビナヤ・グルアチャリャ】世界最高峰のエベレスト（中国名チョモランマ、８８４８メートル）で１９２４年、消息を絶ち、先ごろ頂上の下約６００メートルで遺体が見つかった英国人登山家、ジョージ・マロリー氏（当時３８歳）＝写真＝の捜索隊が２５日、カトマンズで記者会見し、マロリー氏らはエベレスト登頂に成功していない可能性が大きいと語った。【カトマンズ２５日ビナヤ・グルアチャリャ】世界最高峰のエベレスト（中国名チョモランマ、８８４８メートル）で１９２４年、消息を絶ち、先ごろ頂上の下約６００メートルで遺体が見つかった英国人登山家、ジョージ・マロリー氏（当時３８歳）の捜索隊が２５日、カトマンズで記者会見し、マロリー氏らはエベレスト登頂に成功していない可能性が大きいと語った。マロリー氏は三十八歳だった一九二四年、英国登山隊の一員としてエベレスト初登頂に挑んだ。

生成要約文 (asy，30)

$\{\mathbf{S_7}, \mathbf{S_{55}}, \mathbf{S_{58}}, \mathbf{S_{71}}, \mathbf{S_{72}}, S_{135}, \mathbf{S_{136}}\}$　Precision 0.857 / Coverage 0.750

【ロンドン３日共同】「そこに山があるから」――の名文句で知られ、エベレスト（チョモランマ）でなぞの死を遂げた英国の登山家ジョージ・マロリー氏の遺体が７５年ぶりにエベレスト山中で見つかったことが３日、米シアトルに拠点を置く「マロリー＝アービン捜索隊」のホームページで明らかになった。【カトマンズ２５日ビナヤ・グルアチャリャ】世界最高峰のエベレスト（中国名チョモランマ、８８４８メートル）で１９２４年、消息を絶ち、先ごろ頂上の下約６００メートルで遺体が見つかった英国人登山家、ジョージ・マロリー氏（当時３８歳）＝写真＝の捜索隊が２５日、カトマンズで記者会見し、マロリー氏らはエベレスト登頂に成功していない可能性が大きいと語った。遺体の最初の発見者、アンカー隊員も「個人的には、登頂できなかったと思う」と述べ、同隊員が２週間後、同じ場所から頂上に着くまで約１０時間を要したことを明らかにした。四日の英紙などによると、マロリー氏の遺体は、米シアトルに拠点を置く「マロリー＆アービン捜索隊」によって一日、標高８２９０メートルの地点で発見された。乾燥した空気と氷点下の温度のため、遺体の状態は「きわめて良好」で、腰にロープを巻き付けていたという。遺体を発見した「マロリー＆アービン捜索隊」隊長のエリック・シモンソン氏が、マロリー氏の遺体のそばにあった登山用ゴーグルや酸素ボンベ、高度計などを公開したもので、シモンソン氏は、これらの遺品を世界各地の博物館に展示することを検討しているという。**マロリー氏は三十八歳だった一九二四年、英国登山隊の一員としてエベレスト初登頂を目指している途中、アンドルー・アービン隊員（当時二十二歳）と共に消息を絶った。**

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．

また，文の重みを TF-IDF 値で求め，冗長性回避を本手法で利用した整数計画法で行った手法も試してみた．その結果，潜在トピックによって文の重みを求めた提案手法の方が Coverage の値が良く，TF-IDF 値よりも潜在トピックの方が文書の内容を捉える指標として優れているのではないかと推測された．

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 [10]，北島らによって提案された eventLDA[11]，平尾らによって提案されたクラスタリングおよび MMR に基づく要約手法 [10] 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．

例として，文書セット 0590 の提案手法による要約生成結果を以下に示す．

文間の冗長性を考慮しない (冗長考慮無) 場合は，正解データから，要約として抽出された 7 文全てが重要文と分かり，Precision=1.0 となる．しかし，内容には多くの冗長が見られ，Coverage=0.333 と低い値になっている．一方，冗長性を考慮した手法では，太字となっている 6 つの文が重要文となり，冗長考慮無の場合と比べて Precision の値は低い．しかし，冗長はあまり見られず，Coverage は 0.75 と大幅に上がり，多くの内容を網羅した要約生成ができていることが分かる．

5. 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．

謝辞

TSC3 のデータ利用に関しまして，NTT コミュニケーション科学基礎研究所の平尾努氏，広島市立大学の難波英嗣氏からご支援を頂きました．この場を借りて御礼申し上げます．

参考文献

[1] A.Highighi and Vanderwende. Exploring content models for multi-document summarization. *Proc. of NAACL HLT-09*, 2009.

[2] Regina Barzilay and Lillian Lee. Catching the drift:probabilistic content models, with applications to generation and summarization. *Proc. of HLT-NAACL*, 2004.

[3] David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems*, p. 2003. MIT Press, 2004.

[4] 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.

[5] Ying-Lang Chang and Jen-Tzung Chien. Latent dirichlet learning for document summarization. *ICASSP*, pp. 1689–1692, 2009.

[6] Tao Li Dingding Wang, Shenghuo Zhu and Yihong Gong. Multi-document summarization using sentence-based topic models. *Proc. of the ACL-IJCNLP 2009*, pp. 297–300, 2009.

[7] Jacob Einsenstein and Regina Barzilay. Bayesian unsupervised topic segmentation. *Proc. of EMNLP-SIGDAT*, 2008.

[8] 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.

[9] 平尾努, 鈴木潤, 磯崎秀樹. 最適化問題としての文書要約. 人工知能学会論文誌, Vol. 24, No. 2, pp. 223–231, 2009.

[10] 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.

[11] 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, F4-2*, 2011.