組合せ最適化アルゴリズムを用いた潜在的意味に基づく複数文書要約

理学専攻 情報科学コース
重松遥

1 はじめに

自動要約は，大量の文書データを効率よく把握する手段として盛んに研究されており，最適化問題を解くことで実現されることが多い．文書要約における最適化問題とは，要約対象となる文書群の中から，要約文数などの制約条件の下で目的関数を最大にするような文の組合せを見つけることであり，目的関数は文書の解析手法によって多様に考えられる．近年では，解析手法として，文書群中に潜在する話題を確率的に求めるトピックモデルの有用性が示されている．代表的なトピックモデルには pLSA や LDA[1] などがあり，各文書 d はトピック比率ベクトル θ_d，各トピック t は単語分布ベクトル ϕ_t で表される．本研究では，文書群中の潜在トピックに着目し，トピックの観点から適した要約を生成することを目的とする．第 2 章では，最適化に整数計画法を利用し，文書群中のトピック比率を考慮した要約手法を，第 3 章では，NP 困難を避けるため最適化に収束効率の良い差分進化を利用し，各トピックの単語分布 ϕ_t を文書の内容を代表する文として着目する要約手法を提案する．

2 トピック比率を考慮した複数文書要約

2.1 概要

内容の重要度が高く，なおかつ冗長が少ない組合せを高く評価するため，目的関数は，文の重要度ともに，文の組合せが文書群全体の内容を被覆する度合も考慮して設定する．文の重要度は，LDA によって推定したトピック比率ベクトル θ_d と単語分布ベクトル ϕ_t に基づいて定義する．要約評価型ワークショップ TSC3 のデータを用いた予備実験により，要約対象となる文書群中のトピック比率と，正解要約中のトピック比率がほぼ同じになることを確認したことから，トピックの重要度がトピック比率に応じて決まると推測し，トピック比率を考慮した要約手法を提案する．

2.2 手法

目的関数は，文の組合せの重要度と被覆度の積で定義する．組合せの重要度は，組合せを構成している文の重要度の総和で測る．文 i の重要度 b_i は，各トピック t における文 i の重要度 b_{ti} の総和とした．b_{ti} は，トピック t における文 i を構成している単語の重要度を総和したものに，トピック比率を掛けることで求める．トピック t における単語の重要度には，単語分布ベクトル ϕ_t の値を利用した．また，組合せの被覆度は，文書群中の全文を組合せを構成している文のどれかひとつに被覆させ，それぞれの被覆度の総和で測る．文 i が文 j を被覆する度合い e_{ij} は，文間に共通する単語の数を，文 j を構成する単語の数で割ることで求める．制約条件としては要約長の制限を与え，要約長に見合う中で，最も目的関数を最大にする文の組合せ選択する最適化問題を考える．なお，最適化手法には整数計画法を用いる．

2.3 実験

2.3.1 実験仕様

TSC3 のデータで実験を行う．約 10 件の日本語ニュース記事からなる文書群が 30 セット用意されており，各文書群に予め設定された要約長に従い，短文要約と長文要約を生成する．短文要約 “short”，長文要約 “long” を Coverage 値によって評価する．また，DUC’04 で提供された英語文書データによる実験も行う．約 10 件の英文記事からなる文書群が 50 セット用意されており，各文書群に対し，665byte 以内で要約を生成する．評価には ROUGE を用い，ストップワードを含めた ROUGE-1 値 “with” とストップワードを除いた ROUGE-1 値 “without” を求める．また，LDA におけるトピック数はパープレキシティによって決定した．

2.3.2 結果と考察

表 1: Coverage 値の比較

method	short	long
提案手法	0.425	0.483
Lead	0.212	0.326
TF-IDF	0.292	0.325

表 2: ROUGE-1 値の比較

method	with	without
提案手法	0.389	0.326
CLASSY	0.382	0.309
takamura	0.399	0.326

TSC3 による結果を表 1 に示す．Coverage は生成要約中の冗長度合を考慮しつつ，正解要約とどれだけ近いかを測る指標である．提案手法は，自動要約において基本的な手法である Lead 法，TF-IDF 法と比較して，Coverage 値に優れており，提案手法の妥当性が示された．また，DUC’04 による結果を表 2 に示す．DUC’04 において ROUGE-1 値が最も高かった手法である CLASSY と比較すると，提案手法は CLASSY を上回っている．また，高村らによって提案された手法 [2] は，被覆度の考え方が提案手法と同様であり，文の重要度は，文書群全体に類似した文ほど重要だという表層的な情報に基づき求めている．提案手法は高村らの手法と比べ同等の精度が出せており，要約生成において，文書の潜在的意味 (トピック) は表層情報と遜色なく重要な情報であると分かった．

3 トピックを捉え差分進化を用いた複数文書要約

3.1 概要

最適化問題による要約手法の多くは，最適化手法として整数計画法などの厳密解法が使用される．しかし，厳密解法には解の探索空間が大きいほど解を求めるコストが大きくなり NP 困難となる問題がある．そこで，本手法では，厳密解を求めるのではなく実時間の中で近似解を求める手法として効率の良い収束性が示されている差分進化 (DE)[3] を用いた要約生成を行う．DE は進化的アルゴリズムの一種であり，初期個体群を，淘汰 (選択)・交叉・突然変異によって適合度の高い個体へと進化させていくアルゴリズムである．要約生成に

おいては，個体を文の組合せと捉え，予め設定した世代数の中で適合度が高くなるようなものに文の組合せを進化させていくことで，適合度が高い要約を生成する．なお，適合度は，第 2 章で文書要約における潜在的意味の効果が示されたことより，引き続きトピックを考慮して定義する．

3.2 手法

文書群中には k 個のトピックが潜在し，各トピック t の代表文 O_t に類似した文ほど重要であるという仮定の下，適合度関数を定義する．

$$f(X) = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\max_{t=1,2,\cdots,k}\{sim(s_{ti},O_t)\}+\max_{t=1,2,\cdots,k}\{sim(s_{tj},O_t)\}\right)\cdot x_i\cdot x_j}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} sim(s_i,s_j)\cdot x_i\cdot x_j}$$

x_i は，文 i が組合せ X に含まれるとき 1，そうでないとき 0 となる二値変数である．$\max_{t=1,2,\cdots,k}\{sim(s_{ti},O_t)\}$ は，文 i と一番近いトピック代表ベクトルとのコサイン類似度で，文 i の重要度と考える．$sim(s_i,s_j)$ は，文 i と文 j の冗長度を表し，tf-isf 値のコサイン類似度で求める．代表文 O_t は，LDA によって推定された各トピック t の単語分布ベクトル ϕ_t を当てはめる．

個体 X_i の次世代候補個体 Z_i は，個体 X_i に突然変異個体 Y_i を交叉率 CR で交叉して生成する．突然変異個体 Y_i は以下の式で求める．

$$Y_i = X_i + F\cdot(X_{best} - X_a) + F\cdot(X_{best} - X_b)$$

X_{best} は個体群の中で最も適合度が大きい個体 (ベスト個体)，X_a と X_b は個体群からランダムに選んだ 2 個体を表す．個体 X_i にベスト個体とランダム個体の重み付き差分を足す事で，ベスト個体の情報を取り入れた突然変異個体を作る．

次世代個体は，各個体と次世代候補個体を比較し，より要約に適した方を選ぶ．ここで，要約生成には要約長の制約があるため，選択の際に適合度だけでなく要約長も考慮することで制約を加味した最適化を行う．選択のルールは，(i) どちらも制約を満たす場合，適合度が大きい方，(ii) どちらかが制約を満たさない場合，制約を満たす方，(iii) どちらも制約を満たさない場合，制約の逸脱度が低い方をそれぞれ選択することにする．

3.3 実験

3.3.1 実験仕様

DUC'04 によって提案手法 ($TopicDE$) を評価する．差分進化は，10000 世代目のベスト個体を生成要約とし，個体数，交叉率 CR，差分の重み F は経験的に 50，0.7，0.45 と設定する．初期個体群は通常ランダムな確率で数値を設定することが多いが，要約生成においては，ランダムな初期個体を設定すると初期個体が 665byte を大きく上回る文の組合せになってしまう．すると，いくら進化させても制約を満たす個体が現れず，適合度を考慮できないまま世代が終了してしまうという問題がみられた (図 1 左)．そこで，予め初期個体が 665byte を下回るように数値の発生確率を操作することで問題を解決した (図 1 右)．要約は各文書セットあたり 20 個生成し，20 個のうち，最も適合度が高い個体 $TopicDE_{best}$，最も低い個体 $TopicDE_{worst}$，20 個の平均 $TopicDE_{ave}$ の ROUGE-1 値を求める．

(左軸：byte 数，右軸：適合度)

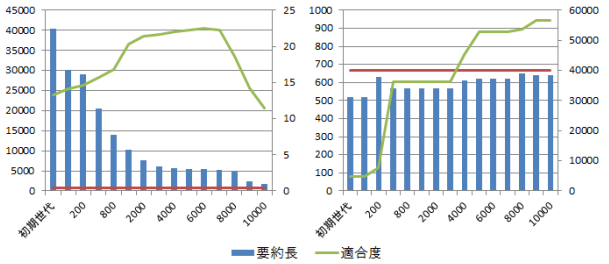

図 1: 初期個体群の操作による収束性の変化

3.3.2 結果と考察

50 文書セットの平均 ROUGE-1 値を表 3 に示す．$TopicDE_{best}$，$TopicDE_{ave}$，$TopicDE_{worst}$ は ROUGE-1 値の分散が小さく，差分進化によって安定した近似解を求められていることが分かった．また，$TopicDE_{best}$ が $TopicDE_{worst}$ よりも精度が高いことより，適合度の高さが要約の精度に関係していることが分かる．しかし，$TopicDE$ は，表 2 に示される他手法と比較して半分の精度も出せておらず，適合度関数が不十分であったと推測できる．現在，組合せの重要度を組合せの冗長度で割ることで適合度を求めているが，これだと重要度と冗長度のバランスを考慮できず，あまり重要ではないが冗長性は低い文の組合せを高く評価してしまう問題が出てくる．この問題によって $TopicDE$ の正確な精度が得られないと推測し，今後は適合度関数の再設定を課題とする．

表 3: 提案手法 TopicDE の評価

method	with	without
$TopicDE_{best}$	0.284	0.150
$TopicDE_{ave}$	0.283	0.142
$TopicDE_{worst}$	0.281	0.138

4 おわりに

本研究では，文書中の潜在トピックを捉えた二種類の要約手法を提案した．前者では，トピックは文書中の含有率に応じて重要であると想定して，トピック比率を考慮した要約手法を提案し，実験により文書要約においてトピックは有効な潜在情報であると分かった．後者では，最適化手法に差分進化を用いることでコストを抑えた要約手法を提案し，実験によって差分進化の有効性を示すことができたが，適合度関数が不十分であった．今後は，適合度関数を重要度と冗長性のバランスを考えて定義しなおすことを課題とする．

参考文献

[1] David M. Blei, Andrew Y. Ng, Michael I. Jordan: Latent Dirichlet Allocation, Journal of Machine Learning Research 3, pp. 993-1022, 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 巻, pp. 174-182, 2010.

[3] Storn R, Price K: Minimizing the Real Functions of the ICEC96 Contest by Differential Evolution, in Proc. of the Intemational Conference on Evolutionary Computation, pp. 842-844, 1996.