4P-4

潜在的な意味を考慮した文書要約への取り組み

小倉由佳里　　小林一郎
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

1　はじめに

近年，自動要約の技術の必要性が高まり，様々な手法が提案されている．自動要約の代表的な手法として重要文抽出によるものがあり，要約における重要文抽出は，最適化問題に帰着させることができる．本研究では，重要文抽出において，遺伝的アルゴリズムによる多目的最適化手法を適用する．要約の生成においては，文の結束性や冗長性，内容の網羅性，重要度等，同時に考慮しなければならない要素が複数存在する．そのため，これらを目的関数として導入し，その多目的最適化を遺伝的アルゴリズムによる多目的最適化手法NSGAⅡ[1] を用いて，複数の条件を満たす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　Non-dominated Sorting Genetic AlgorithmⅡ(NSGAⅡ)

良い要約を生成するためには，要約長の制限や文の結束性，内容の網羅性，冗長性等のトレードオフな関係の複数の要素を同時に考慮することが求められる．これらの要素を考慮するため，要約生成を多目的最適化問題として定式化することができる．そこで本研究では，Deb[1] らによって開発された遺伝的アルゴリズムによる多目的最適化手法である NSGAⅡ を用いて最適な文の組み合わせを見つけ，要約生成を行う．図 1 はこのアルゴリズムにより生成される個体の例である．染色体の i 番目の要素は，文 s_i が要約に含まれている場合は 1，含まれていない場合には 0 となる．初期個体生成において，生成される要約に含まれる総単語数が，制限単語数を超えない個体のみを生成する．これにより，解が安定して収束しやすくなること，要約長の制約を満たす個体が得られやすくなることが考えられる．また子個体は，子個体は親個体のパターンに基づき生成される．しかし，要約生成においては，交叉により親個体と子個体で要素の構成が大きく変化することはあまり好ましくない．なぜならば，良い親同士の交叉であっても，目的関数の評価の低い子個体が多数生成されることが考えられるからである．そこで本研究における交叉，突然変異については Vahed ら [2] の手法を参考に行う．交叉では，まず交叉点をランダムで 1 つ選び，その場所より後ろを入れ替えることで行う．図 1 の場合，要素の白い部分が入れ替えられる．そして交叉後に子個体における 1 の数が親個体と異なる場合，子個体での 1 の数が親個体と等しくなるよう調整を行う．突然変異では，個体の要素において 1 と 0 が隣り合って存在している箇所を見つけ，それらの場所を入れ替えることにより行う．

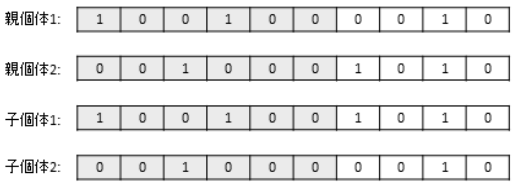

図 1: NSGAⅡ における生成される個体の例

3　提案手法

文書要約における重要文抽出の最適な組み合わせは，多目的最適化問題を解くことで得られる．本研究では，Vahed ら [2]，Huang ら [3] の手法を参考に (i) 文の結束性，(ii) トピックに関連する度合，(iii) 冗長性削減，これらに関する関数について考える．

3.1　文の結束性

良い要約においては，文同士が互いに高い類似度で結合していると考えられる．そのため，それぞれの文間の類似度が高い，文の組み合わせを抽出する必要がある．それを考慮するため，それぞれの文間の類似度の平均値を目的関数に導入する．文間の類似度は *tf-isf* から，コサイン類似度 (4) を用いて計算する．*tf-isf* は文ごとに計算され，s_j は j 番目の文，t_i は i 番目の単

Study on Multi-Document Summarization with Considering Latent Topics

†Yukari OGURA(ogura.yukari@is.ocha.ac.jp),
‡Ichiro KOBAYASHI(koba@is.ocha.ac.jp)
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University,2-1-1 Ohtsuka Bunkyo-ku Tokyo 112-8610

語を示している．

$$tf_{i,j} = \frac{t_{i,j}}{\sum_{i=1}^{n} t_{k,j}} \tag{1}$$

$$isf_i = \log \frac{N}{n_i} \tag{2}$$

$tf_{i,j}$ は，j 番目の文の i 番目の単語の単語の出現頻度であり，isf_i は，i 番目の単語が出現した文数の逆数である．ここで，N は総文数であり，n_i は単語 t_i を含む文数である．以上から単語 $w_{i,j}$ の重みは，式 (3) で計算される．

$$w_{i,j} = tf_{i,j} \times isf_i \tag{3}$$

それぞれの文間のコサイン類似度は式 (4) で表される．

$$sim(s_m, s_n) = \frac{\sum_{i=1}^{t} w_{i,m} \times w_{i,n}}{\sqrt{\sum_{i=1}^{t} w_{i,m}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,n}^2}} \tag{4}$$

文間の類似度の平均値を測る目的関数は式 (5) となる．

$$text\_coh_s = \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n-1} \big(1 - sim(s_i, s_j)\big)}{\frac{1}{2} n(n+1)} \tag{5}$$

ここで n は要約候補として抽出された文の総数である．

また，要約に出現する単語の共起を基に結束性を測る．単語間の結束性は，共起関係に基づく相互情報量 (MI) から得ることができる．

$$word\_coh_s(S) = \frac{\sum_{t_i,t_j \in S, i \neq j} \log\big(MI(t_i,t_j)\big)}{|(t_i,t_j) \in S| \cdot max\Big(\sum_{t_i,t_j \in S, i \neq j} \log\big(MI(t_i,t_j)\big)\Big)} \tag{6}$$

$$MI(t_i,t_j) = p(t_i,t_j) \cdot \log\Big(\frac{p(t_i,t_j)}{p(t_i) \cdot p(t_j)}\Big) \tag{7}$$

ここで，$p(t_i,t_j)$ はある文における単語 t_i と t_j の共起確率であり，$p(t_i)$ はある文における単語 t_i の出現確率である．

3.2 トピックに関連する度合

良い要約はその文書のタイトルに類似した文を含んでいる [4] ということが示されている．これは，タイトルは文書のトピックを端的に表現しているためと考えられる．そこで，生成された要約に含まれる各文とタイトルとの類似度 $TopicRelationFactor(TRF)$ を測る．方法として，文書のタイトルと各文との類似度の平均値を求める．要約 s における TRF は式 (8) で表される．

$$text\_TRF_s = \frac{\sum_{s_j \in summary} \big(1 - sim(s_j, q)\big)}{|S|} \tag{8}$$

ここで $|S|$ は生成された要約 s の文数であり，q は文書のタイトルである．

3.3 冗長性削減

文の結束性やタイトルとの関連度の高い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，文の含意関係を定式化した関数を目的関数に加える．高村ら [5] の先行研究では，文書要約を整数計画問題として定式化をして解く際に，内容的な観点から文 s_i が文 s_j を被覆している度合いを測ることにより，要約生成において文間の含意関係を活用している．

$$e_{ij} = \frac{|s_i \cap s_j|}{|s_i \cup s_j|} \tag{9}$$

ここで，s_i は，その文が含む単語の集合である．よって，$s_i \cap s_j$ は，文 s_i と文 s_j に共通して含まれる単語の集合を表す．文の結束性を考慮する際に，含意関係のある文の組み合わせを多く抽出することが考えられるため，冗長性削減のために e_{ij} を目的関数に導入する．

4 おわりに

本研究では，重要文選択の最適化問題において，トレードオフな関係である文の結束性，冗長性，内容の網羅性を同時に考慮し，最適な文の組み合わせを抽出するため，遺伝的アルゴリズムによる多目的最適化手法 NSGAⅡ[1] を用いる複数文書要約の提案を行った．また文の結束性，トピックに関連する度合，冗長性を最適化の目的関数として導入するための定式化を行った．今後の課題としては，DUC2002 を用いた実験により提案手法の有効性を確認したいと考えている．

参考文献

[1] Deb K, Agrawal S, Pratap A, and Meyarivan, T, : A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II, Lecture notes in computer science 1917, pp. 849–858, 2000.

[2] Qazvinian Vahed, L. Sharif, and Ramin Halavati, : Summarizing text with a genetic algorithm-based sentence extraction, IJKMS 4.2, pp. 426-444, 2008.

[3] Huang Lei, He Yanxiang, Wei Furu, and Li Wenjie, : Modeling document summarization as multi-objective optimization, In Intelligent Information Technology and Security Informatics, Third International Symposium, IEEE, pp. 382-386, 2010.

[4] Silla Jr, C. N, Pappa G. L, Freitas, A. A, and Kaestner. C. A, : Automatic text summarization with genetic algorithm-based attribute selection, In Advances in Artificial Intelligence-IBERAMIA, pp. 305-314, 2004.

[5] 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, 言語処理学会第 15 回年次大会発表論文集, pp. 60-63, 2009.