多目的遺伝的アルゴリズムを用いた組合せ最適化による要約生成

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
{ogura.yukari, koba}@is.ocha.ac.jp

1 はじめに

文抽出による複数文書要約に関する研究では，要約生成を文の組合せ最適化問題として解く方法が主流である．しかしながら，考えうる組合せが増えるほど，厳密解法による組合せ最適化では時間のコストがかかる．また要約生成では，考慮すべき要因が複数あり，多目的の最適化が必要であるといえる．そこで本研究では，組合せ最適化に Non-dominated Sorting Genetic Algorithm-II(NSGA-II)[1] という多目的遺伝的アルゴリズムを用いる．適合度関数には，文の出現位置や単語の重要度，Jensen-Shannon 距離などを用いた．実験は DUC2004 を用い，パレート最適解に関する実験や他のシステムとの比較を行った．パレート最適解から生成された要約から，どの要因が要約生成において重要であるかということに関する考察を行った．

2 関連研究

遺伝的アルゴリズム (Genetic Algorithm:GA) は，代表的な最適化手法の 1 つである．GA は，自動文書要約 [4, 5, 8, 9, 13, 15] においても用いられており，文の組合せ最適化 [4, 8, 9, 13, 15]，文のランク付け [15]，素性の選択や重みの決定 [3, 14]，文の分類 [2] など，複数の操作で用いられる．要約生成において，GA を用いて組合せ最適化を行うことの利点は，厳密解法と比較して計算が高速であることである．要約対象の文書集合が大きくなる程，考えうる文の組合せは増えるため，これを厳密解法で解を求めようとすると，膨大な計算時間が必要となる．また，要約生成においては，複数のトレードオフな関係の要因を考慮しなければならず，一方の要因を満たそうとすると，もう一方の要因が改悪されてしまうような関係を持つ要因間の最適化を行わなければならない．そこで本研究では，文の組合せ最適化問題に対し，GA による多目的最適化を用いることで要約生成を試みる．

3 NSGA-II を用いた要約生成

NSGA-II[1] は，多目的最適化を行い，パレート最適解を得ることで複数の解を得られることが特徴である．GA による文の組合せ最適化では，1 つの個体を解の候補とし，適合度の高い遺伝子を持つ個体を次世代に残すことにより，より適合度の高い解を探索し，要約生成を行う．本研究では，要約生成において重要であると考えられる複数の素性を考慮し，それらを適合度関数として用いる．

3.1 個体と表現方法

各個体は，解の候補を示し，出力する要約の文の組合せを表現している．図 1 は，個体の例である．個体における 1 つのマスが遺伝子座であり，各遺伝子座は，要約対象の文書群が含む各文に対応する．i 番目の遺伝子座の持つ値が “1” である時，文 s_i は要約に含まれることを示し，遺伝子座の持つ値が “0” である時，要約

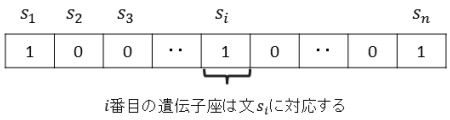


図 1: 個体表現の例

3.2 初期集団の生成

初期集団の生成においては，要約長の制約に基づき各個体の生成を行う．つまり，要約長の制約に関する式 (1) を満たす個体のみを生成する．

$$\sum_{s_i \in S} |\{w_i | w_i \in s_i\}| \leq |L| \tag{1}$$

ここで，S は要約候補に含まれる文集合，w_i は文 s_i に含まれる単語，$|L|$ は予め設定された要約長である．

3.3 交叉

交叉は，2 点交叉を行う．交叉後に，個体の示す要約候補が持つ文字数に応じてランダムに選択した文を個体に追加，または削除する操作を行う．この操作を行う目的は，制約文字数を大幅に超えた個体が生成されることを防ぐことと，交叉により全ての遺伝子座の持つ値が “0” である個体が生成されることを防ぐことである．特に，多くの個体がスパースであることから，実験により後者のような個体が多数生成されることが観測された．

3.4 突然変異

本手法では，遺伝子をランダムに 1 つ選択し，その遺伝子座の持つ値を反転するという操作を行う．つまり，i 番目の遺伝子座が選択されたとき，遺伝子座の持つ値が “1” であった場合は “0” に更新され，遺伝子座の持つ値が “0” であった場合は “1” に更新される．

3.5 適合度関数

複数の文書から要約を生成する場合，システムが最適化すべき要因は複数ある．最適化では，各適合度関数値の最小化を行う．以下に本研究で用いた適合度関数について説明を述べる．

3.5.1 単語出現頻度による文のスコア

単語の出現頻度から，要約がどれだけ重要な文を含んでいるかを定義する．i 番目の文 s_i のスコアを c_i とする．要約 x に対する適合度関数 $f_0(x)$ は，

$$f_0(x) = \frac{1}{\sum_{i=1}^{N} \frac{c_i}{c_{max}} x_i} \tag{2}$$

と書ける．ここで，x_i は，i 番目の文 s_i の状態を表す変数 $x_i \in \{0, 1\}$ であり，文 s_i が要約に含まれるならば $x_i = 1$ となり，そうでないなら $x_i = 0$ となる．c_{max} は，すべての c_i のうちで最大値をもつものである．人が作る要約に含まれる単語は，要約対象の文書で出現頻度が高い傾向が見られる [10] という報告がなされていることから，Nenkova ら [10] の定義に従い，c_i を式 (3) とする．

$$c_i = \sum_{w_j \in s_i} \frac{p(w_j)}{|\{w_j | w_j \in s_i\}|} \tag{3}$$

ここで，w_j は，文 s_i が含む単語を示している．また $p(w_j) = \frac{t}{n}$ であり，t は単語 w_j が要約対象文書群で出現した回数，n は要約対象文書群の総単語数である．

3.5.2 文の位置による文のスコア

文書において，文がどの位置に出現するかによりスコアを与える．i 番目の文 s_i の出現位置のスコアを pos_i とする．要約 x に対する適合度関数 $f_1(x)$ は，

$$f_1(x) = \frac{1}{\sum_{i=1}^{N} pos(s_i) x_i} \tag{4}$$

$$pos(s_i) = max\Big(\frac{1}{i}, \frac{1}{|S_d| - i + 1}\Big) \tag{5}$$

と書ける．ここで，$|S_d|$ は文 s_i が出現する文書 d の総文数である．

3.5.3 Jensen-Shannon 距離による文の組合せに関するスコア

Kullback-Leibler 距離を用いた要約生成手法はいくつか存在する．代表的な手法として，KLSUM[6] では，要約対象文書群と生成要約それぞれの言語モデルから推定した単語分布の Kullback-Leibler 距離を最小化することにより，要約の生成を行っている．また，Hong ら [7] は，Kullback-Leibler 距離を用いて新聞記事とその要約の単語分布の比較から考察を行っていることから，Kullback-Leibler 距離を用いることは要約生成において有用であると考えられる．本手法では，彼らの手法を応用し，要約対象の文書群 D と生成した要約 S での単語出現頻度の比較に，Kullback-Leibler 距離を基に正規化を加えた Jensen-Shannon 距離による式 (10) を適合度関数として用いる．

$$KL(S||D)(w) = P_S(w) \cdot log \frac{P_S(w)}{P_D(w)} \tag{6}$$

$$KL(D||S)(w) = P_D(w) \cdot log \frac{P_D(w)}{P_S(w)} \tag{7}$$

$$JS(S||D)(w) = \frac{KL(S||M)(w) + KL(D||M)(w)}{2} \tag{8}$$

$$M(w) = \frac{P_S(w) + P_D(w)}{2} \tag{9}$$

ここで，$P_S(w)$，$P_D(w)$ はそれぞれ要約 S，要約対象の文書群 D での単語 w の出現確率である．要約 x に対する適合度関数は式 (10) となる．

$$f_2(x) = \sum_{i=1} JS(S||D)(w_i) \tag{10}$$

3.5.4 要約の文字数の評価

要約には，要約長の制約がある．GA を用いて要約を生成する際には，この制約を満たす個体で，適合度の高い個体を次世代へ残す必要がある．しかし，交叉，突然変異などの操作により，制約を超える個体も多数生成されてしまう．そこで，個体が示す要約候補の文字数に関する適合度関数を用いる．

$$f_3(x) = \frac{Length}{L_{const.}} \tag{11}$$

$$Length = \sum_{i=1}^{N} |\{w_i | w_i \in s_i\}| x_i \tag{12}$$

ここで，$L_{const.}$ は，予め設定された要約長の制約である．

4 実験

4.1 実験設定

データセットには DUC2004 の task2 を用い，各文書セットに対して 665byte 以内の要約を生成する．評価指標には ROUGE-1，ROUGE-2 を用いる．実験は，他の要約システムとの性能の比較と，最終的に得られたパレート最適解の個体が示す文の組合せからなる要約に関して，ROUGE-1 値と適合度関数との関係を調べることである．交叉率は 1.0，突然変異率は 0.005，初期個体数は 50，世代数は 100 とする．また，DUC2004 の全ての文書セットに対し実験を行ったが，結果を全て載せると冗長になるため，結果を掲載する文書セットはランダムに選択した．

4.2 実験結果

表 1 は，他のシステムとの比較結果である．図 2 は各適合度関数と ROUGE-1 値との関係を示している．図と対応する適合度関数はそれぞれ，単語出現頻度による文のスコア (式 (2))，文の位置による文のスコア (式 (4))，Jensen-Shannon 距離による文の組合せに関するスコア (式 (10)) である．図の横軸は各適合度関数値，縦軸は ROUGE-1 値となっており，各点は個体を示している．単語出現頻度による文のスコア (式 (2)) と ROUGE-1 値の関係からは，強い相関は見られなかった．一方で，文の位置による文のスコア (式 (4)) では，文書セットにより異なる傾向が見られた．また，Jensen-Shannon 距離による文の組合せに関するスコア (式 (10)) では，このスコアが良いと ROUGE-1 値も良くなる傾向が見られた．

表 1: 他のシステムとの ROUGE-1，ROUGE-2 の比較

システム	ROUGE-1	ROUGE-2
提案手法	37.98	9.48
FreqSum[11]	35.30	8.11
Greedy-KL[6]	37.98	8.53
LR[12]	39.00	9.60

4.3 考察

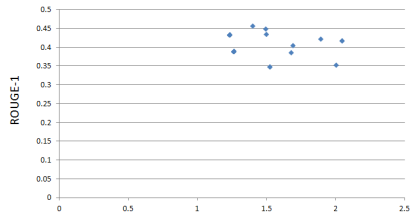
表 1 の他手法との比較より，提案手法は，1 つの要因の目的関数に関して貪欲法で最適化を行う手法よりも高い ROUGE 値を示していることから，要約生成に関して考慮すべき要因は複数必要であることが考えられる．図 2 では，各適合度関数と ROUGE-1 値の関係を調べた．その結果，この 3 つの関数のうち，Jensen-Shannon 距離による文の組合せに関するスコア (式 (10)) の値が良い場合に，ROUGE-1 値が最も高くなる傾向が見られた．しかしながら，この値が最も良い個体であったとしても，最も高い ROUGE-1 値になるとは限らず，少しこの値が悪くても，他の適合度関数値が良い個体の方が，ROUGE-1 値が高い事例も観測された．表 2 は，その一例である．表中の F_0，F_1，F_2 はそれぞれ適合度関数の f_0，f_1，f_2 に対応する．この結果より，1 つの適合度関数値のみを最適化しても良い要約の生成は難しく，各適合度関数のバランスをとることが重要であるといえる．

表 2: 適合度と ROUGE-1 値の関係

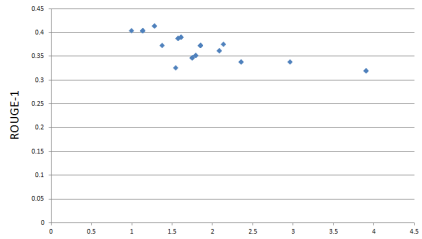
	F_0	F_1	F_2	ROUGE-1
事例 1	2.1475	**1.0**	**0.14346**	0.4038
事例 2	**2.0287**	1.2857	0.1543	**0.4132**

5 おわりに

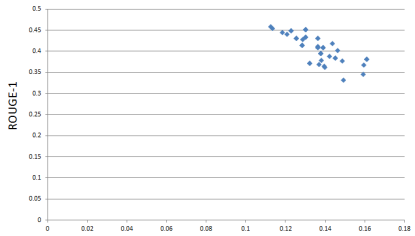
本研究では，文の組合せ最適化問題に対し，多目的遺伝的アルゴリズムである NSGA-II を用いることで実時間で質の良い要約を生成することを目的として，要約生成システムの提案を行った．適合度関数には，文の素性や文の組合せに関するスコアを用いた．実験には DUC2004 を用い，評価は ROUGE で行った．実験では，他のシステムとの比較，パレート最適解と各適合度関数との ROUGE-1 値との関係について考察を



(a) 単語頻度による文のスコア (式 (2))



(b) 文の位置による文のスコア (式 (4))



(c) Jensen-Shannon 距離による文の組合せに関するスコア (式 (10))

図 2: 各適合度関数と ROUGE-1 の関係を図示した例

行った．適合度関数によって，ある適合度関数値が良くなると ROUGE-1 値が改善する傾向が見られた．しかし，1 つの適合度関数を最適化するだけでは不十分であり，各適合度関数とのバランスが重要であるとの考察を得た．また，他の要約生成システムとの比較では，ROUGE-1，ROUGE-2 の値において，他のシステムと並ぶ性能を示した．今後の課題としては，ニュース記事以外のデータを用いた実験や，システムの評価において，ROUGE 以外の評価指標を用いることが考えられる．

参考文献

[1] K. Deb, A. Pratap, S. Agarwal and T. Meyarivan. 2002. A fast and elitist multi-objective genetic algorithm: NSGA2. IEEE Transactionon Evolutionary Computation 6.2:149–172.

[2] R. Alguliev, and R. Aliguliyev. 2009. Evolutionary Algorithm for Extractive Text Summarization. Intelligent Information Management:128–138.

[3] Aristoteles, Y. Herdiyeni, A. Ridha, and J. Adisantoso. 2012. Text Feature Weighting for Summarization of Documents in Bahasa Indonesia Using Genetic Algorithm. International Journal of Computer Science Issues.9(3).

[4] R. Arnulfo, G. Hernandez, and Y. Ledeneva. 2013. Single extractive text summarization based on a genetic algorithm. Pattern Recognition:374–383.

[5] N. Chatterjee, A. Mittal, and S. Goyal. 2012. Single document extractive text summarization using Genetic Algorithms. Emerging Applications of Information Technology.

[6] A. Haghighi, and L. Vanderwende. 2009. Exploring content models for multi-document summarization. In Proc. of NAACL.

[7] K. Hong, and A. Nenkova. 2014. Improving the Estimation of Word Importance for News Multi-Document Summarization. In Proc. of EACL:712–721.

[8] D. Liu, Y. He, D. Ji, and H. Yang. 2006. Genetic algorithm based multi-document summarization. PRICAI 2006: Trends in Artificial Intelligence:1140–1144.

[9] K. Nandhini, and S. R. Balasundaram. 2013. Use of genetic algorithm for cohesive summary extraction to assist reading difficulties. Applied Computational Intelligence and Soft Computing 2013:8.

[10] A. Nenkova, and L. Vanderwende. 2005. The impact of frequency on summarization. Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101.

[11] Ani Nenkova, Lucy Vanderwende, and Kathleen McKeown 2006. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In Proc. of ACM SIGIR:573–580.

[12] M. Nishino, N. Yasuda, T. Hirao, J. Suzuki, and M. Nagata. 2013. Lagrangian Relaxation for Scalable Text Summarization while Maximizing Multiple Objectives. Information and Media Technologies 8.4:1017–1025.

[13] V. Qazvinian, L. S. Hassanabadi, and R. Halavati. 2008. Summarising text with a genetic algorithm-based sentence extraction. International Journal of Knowledge Management Studies 2.4:426–444.

[14] C. N. Silla Jr., G. L. Pappa, A. A. Freitas, and C. A. A. Kaestner. 2004. Automatic text summarization with genetic algorithm-based attribute selection. IBERAMIA 2004:305–314.

[15] Z. Xie, X. Li, B. D. Eugenio, P. C. Nelson, W. Xiao, and T. M. Tirpak. 2004. Using gene expression programming to construct sentence ranking functions for text summarization. In Proc. of Coling.