最大被覆モデルを用いた電子掲示板の自動要約

田中 駿 †　　矢野 裕一朗 ††　　二宮 崇 ††　　高村 大也 †††

愛媛大学 工学部 情報工学科 †
愛媛大学 大学院理工学研究科 電子情報工学専攻 ††
東京工業大学 精密工学研究所 †††

{shun, yano}@ai.cs.ehime-u.ac.jp, ninomiya@cs.ehime-u.ac.jp
takamura@pi.titech.ac.jp

1　はじめに

近年，インターネット利用者が爆発的に増加し，インターネットを利用し情報を手に入れることが多くなってきている．特に電子掲示板 (BBS) では，誰でも情報を発信することが可能であり，多様な意見を交換することができるため，多くの人に利用されている．一方，誰でも情報を発信できるため，1つのトピックに対し多くの投稿が寄せられ，トピックに関する必要な情報のみを得ることは難しい．人手によって BBS 記事を要約している「まとめサイト」と呼ばれるウェブサイトが存在するが,「まとめサイト」の構築には時間的，人的コストがかかるため，BBS から必要な情報のみ抽出する自動要約の実現が期待されている．

本研究は BBS のための自動要約手法を提案する．「まとめサイト」は日々，人手によって作成されているため，要約の正解データを大量に入手することができ，機械学習により自動要約を実現する方法が考えられるが，単純に単語ベクトルなどを入力として学習する手法では高い精度を実現することは難しい．近年，文書要約 [2] を最大被覆問題として解く自動要約手法が研究されており，非常に高い要約精度を実現している [1, 4, 5, 3]．これらの自動要約手法を BBS 要約に適用することが考えられるが，これらの手法の多くは字数制限を最大被覆問題の制約とし，この制約の下でより多くの情報を再現する手法となっているため，字数制限がない BBS 要約にこれらの手法を直接適用することは難しい．本研究では要約対象の記事に対して投稿数と投稿番号に関する制約を与える手法を提案する．

1: 名前:A 2013/12/10 10:30:10 ID:xxxxxx 「mixiニュース」がスマホアプリに
2: 名前:B 2013/12/10 10:35:20 ID:aaaaaa いまどきmixiなんて使ってる人いるのか
3: 名前:C 2013/12/10 10:36:56 ID:bbbbbb 今なら無料 http://www.fkgame.com/
4: 名前:D 2013/12/10 11:14:10 ID:cccccc りんごたべたい
5: 名前:E 2013/12/10 11:45:50 ID:dddddd 情強はGoogleニュースだろ
6: 名前:F 2013/12/11 10:30:10 ID:eeeeee 20年間彼女いない俺はホグワーツに入学できる
7: 名前:G 2013/12/12 10:30:10 ID:ffffff おれちょっとmixiの株買ってくるわ

図 1: BBS 記事の例

2　BBS 要約

BBS には，たくさんの記事が含まれており，記事には1つの話題に対するたくさんの人の意見が含まれている．各記事は，複数の投稿から構成されており，投稿は話題提起であるトピックかそれに対する反応であるレスに分けられる (図 1)．トピックは記事の 1 番最初の投稿であり，レスは 2 番目以降の投稿である．

BBS 要約の目的は，不必要なレスを除去し，トピックに即したレスを抽出することである．図 1 中の黒地のレスは，トピックとは関係のないレスであり，このようなトピックとの関連性の低いレスを取り除き，図 1 の投稿番号 1，2，5，7 のようなトピックに即したレスのみを抜粋した記事を作成することで BBS 要約を行う．

本研究では，人手によって BBS 記事を要約している「まとめサイト」に掲載されている投稿を要約の正解と考え，元の BBS 記事 (元記事) における各投稿に対し,「まとめサイト」において採用されていれば採用タグを付与，採用されていなければ不採用タグを付与し，BBS 要約データを作成する.「まとめサイト」では，元記事のレス数の約 10 分の 1 程度までレスを圧

縮している．

要約率をどの程度要約対象データを圧縮したかを表す指標とし，式 (1) で定義する．

$$\text{要約率} = \frac{\text{採用投稿数}}{\text{元記事投稿数}} \tag{1}$$

3 最大被覆モデルを用いた研究

Filatova らは，最大被覆問題として文書要約を行う手法を提案した [1]．最大被覆モデルを利用するメリットとして，要約として採用された複数の文書に同じ内容を含むという冗長性へ対処することができるということが挙げられる [5]．Takamura らは，文書要約を最大被覆問題として定式化し，字数制限や被覆，関連性の制約の下で最適な文書を選び要約を行う手法を提案した [3]．Takamura らの研究で定式化された最大被覆モデルは式 (2) から式 (6) によって表される．

$$max. \quad \sum_j b_j z_j \tag{2}$$
$$s.t. \quad \sum_i x_i \leq K, \tag{3}$$
$$\sum_i a_{ij} x_i \geq z_j; \forall j, \tag{4}$$
$$x_i \in \{0,1\}; \forall i, \tag{5}$$
$$z_j \in \{0,1\}; \forall j. \tag{6}$$

x_i は i 番目の文が要約に採用される場合に $x_i = 1$ となる変数であり (式 (5))，また，z_j は単語 j が要約に採用される場合に $z_j = 1$ となる変数である (式 (6))．b_j は単語 j の重みを含む定数である．a_{ij} は i 番目の文中に含まれる単語 j の数であり，i 番目の文を採用する場合は単語 j が i 番目の文中に少なくとも一度は出現しなくてはならないという制約 (式 (4)) と，要約として採用する文の長さが K 以内であるという制約 (式 (3)) の下で，要約として採用された文に含まれる単語の重みを最大化する x_i と z_j を求める．

4 最大被覆モデルを用いた BBS 要約手法

BBS 要約に適用する最大被覆モデルと，最大被覆問題の制約式に追加する投稿番号コスト制約，可変要約長制約について述べる．

4.1 BBS 要約のための最大被覆モデル

3 節で述べた最大被覆問題を変更し，BBS 要約のための最大被覆問題を定義し，その最大被覆問題を解くことにより要約を行う．式 (7) から式 (11) に，本研究で使用する最大被覆問題を示す．

$$max. \quad \sum_j b_j z_j \tag{7}$$
$$s.t. \quad \sum_i d_i x_i \leq K(n), \tag{8}$$
$$\sum_i a_{ij} x_i \geq z_j; \forall j, \tag{9}$$
$$x_i \in \{0,1\}; \forall i, \tag{10}$$
$$z_j \in \{0,1\}; \forall j. \tag{11}$$

x_i は i 番目の投稿が要約に採用される場合に $x_i = 1$ となる変数であり，z_j，a_{ij}，b_j は式 (2) から式 (6) で定義される変数，定数と同義である．3.1 節の最大被覆モデルとの差異は，式 (3) と式 (8) にある．d_i は，投稿番号コスト制約によって決定される，投稿番号に対する要約採用コストである．$K(n)$ は，可変要約長制約によって決定される，要約として採用する投稿に対するコストの合計を決める関数であり，n は要約対象記事の投稿数である．これらの制約については以下で詳しく述べる．また，本研究では，単語ベクトルを素性ベクトルとする L2 正則化項付ロジスティック回帰 (L2LR) を学習し，L2LR を学習した結果得られる単語に対する重みを単語に対する重み (b_j) とした．

投稿番号コスト制約

図 2 は，要約として採用された投稿番号の分布である．「まとめサイト」では投稿番号の小さいレスを採用する傾向にあり，投稿番号 0-100 が 70% と，要約データの大部分を占めている．これは，たくさんの投稿が含まれる記事の場合，投稿番号が大きくなるにつれ，以前に投稿された内容と類似する投稿が出現することがあることや，「まとめサイト」の製作者が投稿番号の大きい投稿を要約対象としていないことなどが原因であると推測される．本研究でも，投稿番号に応じてコストを設け，投稿番号の大きい投稿を要約として採用しないように調整する制約をつけた．投稿番号 i に対するコスト d_i は，式 (12) で表される．

$$d_i = e^{0.08i} + 10 \tag{12}$$

可変要約長制約

本研究では，投稿数によって適切な要約長が決定されると仮定し，投稿数を n としたとき，要約として採用する投稿に対するコストの合計の上限を $K(n)$ とする．図 3 は，投稿数に対する要約率のグラフである．$K(n) = \beta n$ とすると，要約率を β で固定することに

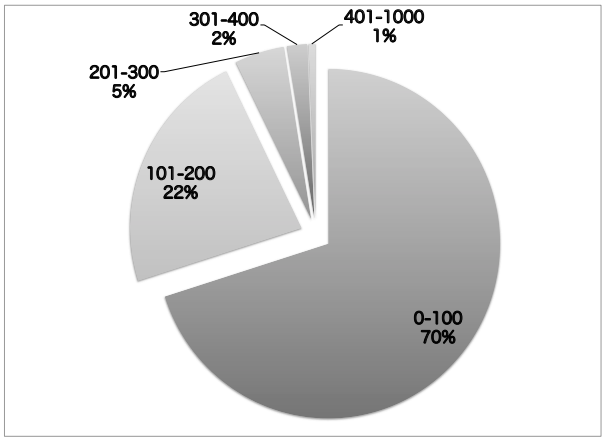

図 2: 要約データ中の投稿番号の内訳

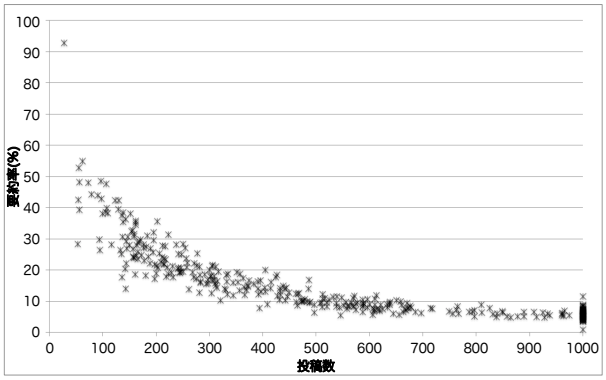

図 3: 投稿数と要約率の関係

なるが，例えば要約率を 10% に固定してしまうと，投稿数の少ない記事を要約する際は，要約で採用できる投稿数がかなり少なくなってしまい，十分に要約を行うことができない．そこで本研究では，n に対する 2 次式で $K(n)$ を定義する．

$$K(n) = \alpha n^2 + \beta n \tag{13}$$

式 (13) 中の α，β は，関数 $K(n)$ の出力値を調整するためのパラメータであり，開発データセットを使用し調整した結果，本研究では $\alpha = 0.0192$，$\beta = 24$ とした．

5 実験

5.1 大規模 BBS 要約データの作成

本研究で使用する BBS 要約データの作成方法について述べる．

1. データ取得
   ウェブサイトから BBS の記事データと，要約データを取得する．正解データとして，大手掲示板サイト [1] の「まとめサイト [2]」と呼ばれるウェブサイトの記事データを使用した．

2. HTML タグの除去
   取得したデータは HTML 形式なので，HTML タグの除去を行う．本研究では，改行コードである $\langle br \rangle$ タグは投稿中の文字の要素として扱うため，除去しない．

3. 各投稿に対し採用/不採用タグを付ける
   記事データと要約データを比較し，各投稿に対して要約として採用されている場合には 1，不採用の場合には 0 をタグ付けする．

5.2 実験環境

以下の 4 つの手法に対して精度の比較を行った．

1. L2LR
2. L2LR+リサンプリング
3. 最大被覆モデル (要約率固定)
4. 最大被覆モデル (提案手法)

L2LR は，L2 正則化項付ロジスティック回帰である．L2LR+リサンプリングは，正例と負例の割合が約 3:7 になるように要約データの正例の中からランダムに選択 (リサンプリング) し，正例の量を補正した L2LR である．要約データ中の正例と負例の割合は約 1:9 となっており，L2LR で要約を行うと負例にバイアスがかかる結果，要約率が著しく下がってしまうため，正例の量を補正した．また，最大被覆モデル (要約率固定) は，要約採用投稿数を要約対象記事の投稿数の 10 分の 1 とし，最大被覆モデルとして解いたものである．これは，投稿番号コスト制約を設けず，可変要約長制約については $\alpha = 0$，$\beta = \frac{1}{10}$ とした提案手法と同じである．最大被覆モデル (提案手法) は，投稿番号コスト制約と可変要約長制約を加えた最大被覆モデルである．

実験データとして用いる全データセットは 1,014 個の記事から成り，訓練データセットとして 812 記事，ハイパーパラメータ調整に使用する開発データセットとして 101 記事，テストデータセットとして 101 記事に分割して使用した．データセットの内訳を表 1 に示す．

[1] 2 ちゃんねる:http://www.2ch.net/
[2] 東アジア・政治経済ニュース:http://www.m9l-o-l.com/

表 1: データセットの内訳

	記事数	投稿数	要約採用投稿数	要約率
訓練データセット	812	457,226	38,626	8.45%
開発データセット	101	67,596	4,821	7.13%
テストデータセット	101	61,913	4,718	7.62%

表 2: 実験結果

	f-score	要約率
L2LR	5.3%	0.3%
L2LR+リサンプリング	14.5%	10.6%
最大被覆モデル (要約率固定)	6.2%	9.9%
最大被覆モデル (提案手法)	30.2%	11.2%

素性ベクトルとして用いる単語ベクトルは 200,129 個の単語から成り，L2LR によって各単語に対する重みを学習した．投稿内の各文の単語分割には MeCab Ver. 0.994 を用い，L2LR の学習には，LibLinear Ver. 1.93 を使用した．最大被覆問題の最適解を求める為の整数線形計画ソルバーとして，本実験では ILOG CPLEX Ver. 12.5.1 (IBM 社) を使用した．

本研究では，要約の精度を測る手法として f-score を使用した．f-score は Recall(再現率) と Precision(適合率) から導出され，それぞれ次式で定義される．

$$Recall = \frac{(|\text{ 正解データ中の投稿}\cap\text{システムが出力した投稿 }|)}{|\text{ 正解データ中の投稿 }|}$$

$$Precision = \frac{(|\text{ 正解データ中の投稿}\cap\text{システムが出力した投稿 }|)}{|\text{ システムが出力した投稿 }|}$$

$$f\text{-}score = \frac{2*Recall*Precision}{Recall+Precision}$$

5.3 実験結果

L2LR，L2LR+リサンプリング，最大被覆モデル (要約率固定)，最大被覆モデル (提案手法) の各手法を用いて BBS 要約の評価実験を行った．実験結果を表 2 に示す．

L2LR での要約精度 5.3%，L2LR+リサンプリングでの要約精度 14.5%，最大被覆モデル (要約率固定) での要約精度 6.2% と比べ，L2LR を用いた要約と同程度の要約率で精度が 30.2% と精度が向上した．

リサンプリングによって L2LR の精度が 5.3% から 14.5% まで向上した．最大被覆モデル (要約率固定) の精度が低いのは，投稿番号コスト制約がないためであると思われる．要約データでは，投稿番号の小さい投稿を多く採用しているが，最大被覆モデル (要約率固定) では投稿番号コスト制約がないため，投稿番号の大きい投稿も要約として採用する場合があった．

6 まとめと今後の課題

本研究では，BBS 要約に最大被覆モデルを適用し，要約精度の向上を実現した．今回の実験では，人手によって BBS 記事を要約した「まとめサイト」を正解データとし，データセットを作成した．訓練データセットに対し，L2 正則化項付ロジスティック回帰 (L2LR) を学習し，学習によって得られた単語重みを利用し最大被覆問題を作成した．整数線形計画ソルバーを用いてこれらの問題を解くことにより自動要約を実現した．既存の最大被覆モデルによる文書要約の多くの手法は字数制限を最大被覆問題の制約とするため，字数制限がない BBS 要約にこれらの手法を直接適用することは難しかった．本研究では，制約式中の要約長を投稿数に対する関数として与えることで BBS 要約を最大被覆問題として定式化した．また，投稿番号の小さい投稿ほど BBS 要約に採用され易いという傾向があるため，投稿番号コスト制約を最大被覆問題に追加した．

実験を行い，L2LR では要約精度 (f-score) が 5.3% であり，要約率を固定した最大被覆モデルによる手法では，要約精度 6.2% であった．それらと比較し，提案手法では 30.2% の要約精度を実現した．これらの実験により，最大被覆モデルは文書要約だけでなく，BBS 要約においても有効であることがわかった．また，BBS 要約全般において投稿番号コスト制約が有効であるとは一般には考えにくいが，本研究において「まとめサイト」を正解データとした場合には非常に有効であることがわかった．

今後の課題として，更に BBS 要約に特化した最大被覆問題への変更，素性ベクトルとして引用関係等を採用すること，国外の電子掲示板を正解データとして実験を行うことなどが考えられる．

参考文献

[1] E. Filatova and V. Hatzivassiloglou. A formal model for information selection in multi-sentence text extraction. In *Proc. of COLING 2004*, pp. 397–403, 2004.

[2] I. Mani. *Automatic Summarization*. John Benjamins Publisher, 2001.

[3] H. Takamura and M. Okumura. Text summarization model based on maximum coverage problem and its variant. In *Proc. of EACL 2009*, pp. 781–789, 2009.

[4] W.-T. Yih, J. Goodman, L. Vanderwende, and H. Suzuki. Multi-document summarization by maximizing informative content-words. In *Proc. of IJCAI-07*, pp. 1776–1782, 2007.

[5] 西川仁, 平尾努, 牧野俊朗, 松尾義博, 松本裕治. 冗長性制約付きナップサック問題に基づく複数文書要約モデル. 自然言語処理, Vol. 20, No. 4, pp. 585–612, sep 2013.