3G2-3

係り受け構造と情報利得に着目した複数の小規模コーパスによるWebタウンミーティングにおける投稿の自動分類

Automatic Classification of Posts considering Dependency Structures and Information Gain in Web Town Meetings using Multiple Small Corpus

岩佐幸翠 *1　　藤田桂英 *1
Kosui IWASA　　Katsuhide FUJITA

*1東京農工大学大学院 工学府 情報工学専攻
Department of Computer and Information Sciences, Graduate School of Engineering, Tokyo University of Agriculture and Technology

It is hard for participants to understand the outline of discussions on web town meetings because of many posts. Most existing works related to automatic classification of documents by supervised learning use word's contexts of posts as a part of the feature, therefore, a large corpus is necessary. However, it is difficult to prepare a large-scale corpus of web town meetings because of much time and costs. We propose the dimension reduction method for supervised learning with multiple small corpuses using Japanese dependency structure and information gain. In addition, our proposed method is evaluated by classifying the posts in web town meeting to six labels; agreeing, question, opinion, information, experience, suggestion.

1. はじめに

地方自治体において，地域住民の意見を集約した上で政策に反映することは重要である．そのため，タウンミーティングは各住民の持つ立場・価値観・考え方に基いて意見集約・合意形成を行うことができるために多くの自治体が活用している．特に，Web 上での非対面式タウンミーティングは参加者を時間的・物理的に拘束しないことから，近年では注目が集まっている．しかし，参加者が Web タウンミーティングにおける議論の流れを把握することは，大量の投稿を読まなければならないことから大きな精神的負担を伴う．そのため，Web タウンミーティングにおける各投稿を自動でラベリングすることで，参加者による議論の概観の把握を支援することができると考えられる．本論文では，複数の小規模なコーパスを用いた教師あり学習によって，Web タウンミーティングにおける投稿を自動で同意・質問・提案・意見・情報・経験の 6 つのラベル区分に基いてマルチラベル分類することを目指す．

議論における発言の教師あり自動分類に関する既存研究として，Web 上での政治的なディベートにおける意見を賛成・反対に分類する手法 [Walker 12]，Twitter における投稿をある議論の主題に対して肯定的・否定的のいずれかへ分類する手法 [Mohammad 13] がある．しかし，これらの手法は対立的な議論を対象とし，各投稿を肯定的・否定的のいずれかに分類するが，タウンミーティングのようなブレインストーミング形式で進行される議論では投稿された意見への明確な否定はブレインストーミングのルール上行われることが少ないことから，これらの手法の適用が難しい．また，随筆における論争的な談話のそれぞれの文を前提，主張，主要な主張へ分類する手法が提案されている [Stab 14] が，タウンミーティングのような対話的な発言が多く含まれる議論を対象とする場合，Stab らの分類区分では質問や同意などを分類できない．

また，Bag-of-ngrams により素性を得る手法 [Zhang 03] や，分類対象の投稿と返信先の投稿との単語アライメントを素性として用いる手法 [Deepak 14] など，既存手法の多くは主に投稿文中の出現単語や投稿文と議論の主題との単語アライメントを素性として自動分類を行う．しかし，タウンミーティングでは各実施地域や議題について固有な単語が頻出することから，さまざまな地域・議題に関するタウンミーティングに対して汎用的に既存手法を適用する場合は，大規模なコーパスが必要となるが，意見の閲覧や投稿は各参加者が自らの余暇時間に合わせて行うため実施に時間がかかるため，大規模なコーパスを構築するには長い時間が必要となるため，これらの手法を適用することは難しい．

本論文では，文頭および文末を中心に単語を抽出することで，同意ラベルにおける「私も～だと思います.」や提案ラベルにおける「～するのはいかがでしょうか.」などの，各ラベルにおいて文頭や文末に現れやすい言い回しに着目し，実施地域や議題によらない素性抽出法を提案する．さらに，複数の小規模コーパスから学習するに当たり，特定のコーパスに固有な言い回しを学習することを防ぐために，情報利得を用いた特徴空間の次元削減法を提案する．評価実験において，提案手法が Bag-of-Words や Bag-of-ngrams, word2vec による既存手法に対して，提案手法が既存手法を F 値に関して上回ることを示す．

以下に本論文の構成を示す．まず，対象とする分類問題に関して説明する．次に，係り受け構造と情報利得に着目した複数の小規模コーパスによる自動分類手法を提案する．その後，評価実験の結果を示し，最後に，本論文のまとめと今後の課題を示す．

2. Webタウンミーティングにおける投稿の自動分類

本論文では，返信構造を備えた大規模議論掲示板における投稿文を対象として，6 種類のラベル (表 1) に基づくマルチラベル分類を教師あり学習によって行う．なお，本論文では議論掲示板においてある主題を持った投稿群を「議論コーパス」，議論における主題を「議論テーマ」，各投稿の文章を「発言」，発言を構成するそれぞれの文を「投稿文」と呼ぶ．

連絡先: 氏名: 岩佐 幸翠
所属: 東京農工大学
住所: 〒 184-8588 東京都小金井市中町 2-24-16
メールアドレス: iwasa@katfuji.lab.tuat.ac.jp

表 1: 対話的議論における投稿文の分類区分

ラベル名	ラベルの定義とその例
同意	返信先の意見に対して肯定する文. 例) そうですね，わたしも反対です. 例) 僕もそう思います.
意見	ある事柄について考えを示している文. 例) インターネットは自由であるべきです.
情報	ある事柄について事実を述べている文. 例) 渋谷にはハチ公があるらしいですね. 例) おじいさんとおばあさんがおったそうな.
質問	参加者に質問をしている文. 例) 渋谷にはハチ公がありますか？
提案	参加者に提案している文. 例) 具体的な解決策について話し合いましょう！
経験	ある事柄について経験談を述べている文. 例) 渋谷で人ごみに遭遇した経験があります.

3. 係り受け構造と情報利得に着目した自動分類手法

以下に，提案手法のおおまかな流れを示す.

- 学習
  1. 各議論掲示板における発言群を取得
  2. 各発言に対して前処理を実施し，文ごとに分割
  3. 文ごとに人手によるラベル付けを実施
  4. 文の特徴ベクトル化
  5. 特定の議論コーパスに依存した特徴素の削除
  6. 各ラベルに対応する分類器への学習
- 分類
  1. 各議論掲示板における発言群を取得
  2. 各発言に対して前処理を実施し，文ごとに分割
  3. 文の特徴ベクトル化
  4. 特定の議論コーパスに依存した特徴素の削除
  5. 各ラベルに対応する分類器による文の分類

各発言に対する前処理と文分割

前処理として，投稿文に対して URL および文分割に用いない記号を削除する．さらに，括弧や括弧に括られていない部分のうち，「！」「!」「？」「?」「.」「。」「…」が一つまたは連続している部分を区切りとして文分割を行う.

素性の抽出

以下に示す素性抽出法から得られた素性群 α と β，およびそれぞれの文の文字長を分類器への素性として与える.

素性群 α: 係り受け木の枝刈り込みにより短縮された文からの Bag-of-words

提案手法に対して用いる議論コーパスは，第一章でも述べたとおり小規模なコーパスを想定している．そのため，例えば「愛知の産業」に関する議論では，意見ラベル，情報ラベルの投稿に「名古屋」や「豊田」などの主要な都市名が頻出するなど，小規模な対話的議論コーパスにはそれぞれの議論テーマに強く関係がある単語が多く含まれる．このようなコーパスからそのまま単語を素性として抽出した上で分類器に学習させた場合，異なる議論テーマを扱う投稿に対する汎化性能が下がってしまう．例えば，名古屋に関するコーパスを用いて，「東京における犯罪防止」に関する議論の投稿文に自動分類を実行しても，各投稿に「名古屋」や「豊田」などの単語が含まれていないため，意見ラベル，情報ラベルの分類が十分に行えなず，再現率が低下する．そのため，教師データから特定の議論テーマに関する単語を取り除く必要がある.

図 1: 係り受け構造を利用した文短縮 ($d = 1$ の場合)

文書分類や文書要約の手法において，文末表現や係り受け構造に着目する手法は多く存在する．例えば，松本らはウェブページの主観，客観度の判定手法を提案するに当たって，文末表現に着目することでそれぞれのページにおけるトピックに依存しにくくなるという手法を提案している [松本 09]．また，石井らは文を要約する手法において，係り受け木における各ノード (文節) の深さが，文における各ノード (文節) に存在する単語の重要度に対して負の相関性があると仮定し，これを利用することを提案している [石井 01]．これらの研究成果を参考に，特定の議論テーマの単語が多く含まれた少量のコーパスを教師データとして用いて分類を行うために，係り受け構造を用いた文短縮によって文頭・文末表現のみに着目する．上から，各ラベルに属する文が含んでいる特徴的な言い回しを素性として抽出することで，教師データのトピックに対する過学習を防ぐことが可能となる.

以下に，提案する素性抽出手法を示す.

1. 文から係り受け構造における根ノードからの距離がしきい値 d を超えている文節を削除（図 1）.
2. 文中の各形態素を基本形に変換.
3. 文における各形態素 N-gram の出現回数をベクトル化. (Bag-of-ngrams)

素性群 β: 各形態素に係り受け木の深さを付与した文からの Bag-of-words

これまで提案されている Bag-of-Words や Bag-of-ngrams による素性の抽出方法では，文における各形態素の位置が考慮されていないため，同意ラベルが付与される「その/意見/に/賛成/し/ます/.」という文における「賛成/し」という形態素 2-gram と，同意ラベルが付与されない「多く/の/人/が/その/法案/に/賛成/し/てい/ます/が/，/私/は/より/慎重/に/なる/必要/が/ある/と/考え/ます/．/」という文における「賛成/し」という形態素 2-gram を区別して扱うことができない．そのため，「賛成/し」という形態素 2-gram は，各文に対して同意ラベルが付与されるべきか判断する上で十分な情報量を持たない．そこで，提案する素性抽出手法では，以下の手法で文から素性を得る.

1. それぞれの形態素に対して，各形態素が属する分節の係り受け構造における根ノードからの距離を付与.
2. 文における各形態素 N-gram の出現回数をベクトル化.

情報利得を利用した特徴空間の次元削減法

素性群 α 及び及び β では，各ラベルに属する文の文頭・文末に特徴的な言い回しを素性として抽出することで，特定の議論コーパスに含まれる単語を学習することによる汎化性能の低下を防ぐことを企図した．しかし，各素性における形態素 N-gram の出現回数をそのまま特徴ベクトルとして用いた場合，特定のユーザのみが用いる言い回しや，特定の議論テーマにおいて頻出する言い回しなどの特定の議論コーパスのみに多数出現するような言い回しを学習することにより，過学習を引き起こし汎化性能が低下する可能性が考えられる．そこで，教師データの特定の議論コーパスに偏って出現する素性を削除することを考える．教師データの議論群への各素性の出現確率の偏り程度が高いほど，属する議論コーパスおよび発言が未知の投稿 p に関して，p 中にある素性 f が出現するか否かの情報が取得できた場合の，p が教師データの議論コーパス群のいずれに属するかについての情報利得 $G(f)$ が高くなる．ここで，教師データは N 個の議論 $d_1, d_2, ...d_{N-1}, d_N$ から構成されているとし，教師データから得られた素性群 $\alpha \cdot \beta$ が属する M 個の素性を $f_1, f_2, ...f_{M-1}, f_M$，$p$ が教師データを構成する議論の一つ d_i に含まれている確率を $P(p, d_i)$ とすると，$G(f)$ は数式 1 のように表すことができる．

$$G(f) = \sum_{k=1}^{N} P(p, d_k) \log P(p, d_k) - \sum_{k=1}^{N} P(p, d_k|f) \log P(p, d_k|f) \quad (1)$$

提案手法では，$f_1, f_2, ...f_{M-1}, f_M$ から $G(f_h)$ が高い $f_h (1 \leq h \leq M)$ の上位 τ% を削除する．

Random Forest による学習と自動分類

提案手法では分類器として Random Forest を用いる．Random Forest とは弱学習器を決定木としたアンサンブル学習の一つであり，分類，回帰のために用いる機械学習アルゴリズムである [Breiman 01].

4. 評価実験

提案手法の効果を評価するために，表 3 に示す 11 つの大規模議論データセットについて学生 5 名によるアノテーションによって作成した分類ラベル付きコーパスを用いて，議論テーマを分割単位とした leave-one-out 交差検証を行った上でベースラインおよび提案手法のそれぞれの適合率，再現率，F 値を比較する．ベースラインについては，word2vec(Wikipedia 日本語版をコーパスとしてモデル生成) および Bag-of-ngrams を用いる．また，Bag-of-ngrams および提案手法におけるハイパーパラメータについては，$N = 2, 3, 4$，提案手法におけるハイパーパラメータについては $\tau = 99$ とする．

表 2 にベースラインおよび提案手法のそれぞれの適合率，再現率，F 値を示す．提案，情報，経験ラベルでは，ベースラインである Bag-of-words による素性抽出と比較して提案手法は F 値において優れている．一方，それ以外のラベルでは提案手法がベースラインと並ぶ結果になった．

同意ラベルにおいて，提案手法がベースラインに対して F 値において上回ることができなかった理由は，同意ラベルには「賛成です.」「全くそのとおりだと思います。」「私もその考えに賛成です。」などの短く簡潔な文が多く，議論テーマに固有な単語が含まれる場合が少ないためであると考えられる．提案手法における素性 α で用いた係り受け木の枝刈り込みによる文短縮のみでは，期待したような議論テーマに固有な単語を削除することによる汎化性能の向上効果を得ることは難しい．また，「すごく良いアイデアですね！」「良いアイデアだと思います！」などの文章における形態素「良い」や形態素 2-gram「良い/アイデア」など，係り受け構造によらず特定のラベルに出現しやすい形態素 (または形態素列) が存在するが，提案手法では係り受け構造によらない素性について着目することができないことも理由として考えられる．

表 2: 評価実験の結果 (ラベル別)

手法	ラベル	適合率	再現率	F 値
word2vec	同意	**0.88**	0.32	0.47
	質問	0.74	0.07	0.13
	提案	0.67	0.01	0.02
	意見	0.78	0.90	0.84
	情報	0.63	0.06	0.10
	経験	0.00	0.00	0.00
Bag-of-ngrams	同意	0.82	0.43	0.56
	質問	0.83	0.68	0.75
	提案	0.78	0.29	0.42
	意見	0.81	0.91	0.86
	情報	0.67	0.13	0.21
	経験	0.23	0.02	0.04
提案手法	同意	0.79	0.44	0.56
	質問	0.83	0.69	0.75
	提案	0.76	**0.36**	**0.49**
	意見	0.81	0.91	0.86
	情報	0.69	**0.21**	**0.32**
	経験	**0.47**	0.05	**0.10**

意見・質問ラベルでは，提案手法がベースラインに対して F 値において上回ることができなかった．この結果の理由として，意見・質問ラベルは他のラベルと比較して十分な投稿数があるため，ベースラインにおいて各議論コーパスに含まれる議論テーマに固有ないし強い関連のある単語による汎化性能の低下が起こらなかったことによると考えられる．

ベースライン及び提案手法では，同じ単語を含んでいても文脈によって意味が変わるような文へ対応することが出来ないという課題がある．例えば，同じ「そうですね.」という文でも，これは相手の意見に対する同意としての「そうですね.」と，相槌としての「そうですね」が存在する．そのため，対象となる文が同意ラベルを付与されるべきかは，対象となる文自体から得られる情報では判定することが難しい．今後，対象の前後にある文の形態素や評価極性を素性として与えることが考えられる．

ベースライン及び提案手法はともに隣り合う形態素に着目する一方で，必ずしも各ラベルにおいて出現しやすい形態素のペアが隣り合うとは限らないため，十分に分類のための素性を抽出することができていない可能性がある．例えば，「すごく良いアイデアですね！」「良いアイデアだと思います！」という同意ラベルに分類される文について考える．「良い」という形態素と「！」という形態素が，同意ラベルに分類されるような文に出現しやすいと仮定した場合，ベースラインおよび提案手法はともにこの情報を利用しない．そのため，提案手法の今後の展望として，共起する形態素に着目するための方法を

表 3: 評価のためのデータセットに用いた議論一覧

議論番号	議論テーマ	データ引用元	投稿者数	投稿数	文の数
1	ネットリテラシーの問題点	COLLAGREE[*1]	59	283	926
2	名古屋における災害	COLLAGREE	21	332	1162
3	名古屋における魅力	COLLAGREE	34	392	1422
4	名古屋における環境	COLLAGREE	20	261	944
5	名古屋における人権	COLLAGREE	21	238	857
6	名古屋を活気づけるために	COLLAGREE	34	244	903
7	日本の国民食	SYNCLON[*2]	5	118	306
8	ロボット/人工知能の未来について	SYNCLON	8	149	575
9	3D プリンタでの拳銃製造について	SYNCLON	7	75	298
10	国際社会は世界共通語を導入するべきか	SYNCLON	4	30	109
11	夫婦別姓制度について	SYNCLON	7	108	307

Bag-of-ngrams から何らかの方法へ変更する事が考えられる．

5. おわりに

本論文では，大規模議論掲示板における投稿の構造化を目的として，文ごとに同意，質問，提案，意見，情報，経験の分類区分に対してラベリングする手法を提案した．提案手法では，まず，投稿を文ごとに分割し，素性を抽出した上で，情報利得によりトピックへの依存が強い素性を削除する．その後，ランダムフォレストによって学習および分類を行う．素性を選択するに当たって，議論テーマに固有な単語を排除することで汎化性能を向上させるために，係り受け木の枝刈り込みによって短縮された文，および形態素に係り受け木の深さを付与された文のそれぞれから Bag-of-ngrams によって素性を抽出した．提案手法は，対話的議論データを構造化する上で，小規模なコーパスから学習しなければならない場面において有効であった．また，評価実験の結果，特定のラベルに対してベースラインを上回る高い F 値であった．

今後の展望としては，提案手法の性能をさらに向上させるため，返信構造や前後文を利用した素性を用いることにより議論の流れを学習させる必要がある．また，係り受け構造に依存しない素性を加えて導入することにより，各ラベルに特徴的な言い回しのうちの係り受け位置によらないものについても利用することができると考えられる．さらに，Bag-of-ngrams の拡張である提案手法では隣り合う形態素しか見ることができないため，共起する形態素をより広く探索できる方法の検討，導入が挙げられる．

謝辞

本研究は，JST，CREST の支援を受けたものである．

参考文献

[Breiman 01] Breiman, L.: Random Forests, *Machine Learning*, Vol. 45, No. 1, pp. 5–32 (2001)

[Deepak 14] Deepak, P. and Visweswariah, K.: Unsupervised Solution Post Identification from Discussion Forums., in *ACL (1)*, pp. 155–164 (2014)

[Mohammad 13] Mohammad, S. M., Kiritchenko, S., and Zhu, X.: NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets, *CoRR*, Vol. abs/1308.6242, (2013)

[Stab 14] Stab, C. and Gurevych, I.: Identifying Argumentative Discourse Structures in Persuasive Essays., in *EMNLP*, pp. 46–56 (2014)

[Walker 12] Walker, M. A., Anand, P., Abbott, R., Tree, J. E. F., Martell, C., and King, J.: That is your evidence?: Classifying stance in online political debate, *Decision Support Systems*, Vol. 53, No. 4, pp. 719–729 (2012)

[Zhang 03] Zhang, D. and Lee, W. S.: Question classification using support vector machines, in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 26–32ACM (2003)

[松本 09] 松本 章代, 小西 達裕, 高木 朗, 小山 照夫, 三宅 芳雄, 伊東 幸宏 : 文末表現を利用したウェブページの主観・客観度の判定, 第 1 回データ工学と情報マネジメントに関するフォーラム (DEIM), A5-4 (2009)

[石井 01] 石井 弘志, 林日華, 古郡 廷治 : 単語の中心性に基づくテキスト自動要約システム, 情報処理学会研究報告自然言語処理（NL）, Vol. 2001, No. 20, pp. 83–90 (2001)

*1 http://collagree.com/
*2 http://synclon3.com/