議論システムにおける賛成／反対意見の生成のための発話のベクトル化手法の検討*

☆古舞千暁, 滝口哲也, 有木康雄 (神戸大)

1　はじめに

情報網・Web2.0 の発展や放送のデジタル化により，情報整理が困難なメディア，映像，画像，音響などの普及が，情報の氾濫を招いている．情報量の爆発とプ ラットフォームの多様化により，ユーザーがほしい情 報を入手できない状況にあり，効率的にユーザーが欲しい情報だけを入手できる環境が，必要とされてきている．そこで，我々人同士のコミュニケーション手 段である音声を介したインタフェースが注目を集めており，音声インタフェースによる乗換案内システム [1] や情報案内システム [2, 3] が研究されてきた．NetTv[4] は，ネットニュース動画において，動画 インデキシングと音声インタフェースを用いた検索 システムである．すなわち，ユーザーが快適に動画 を視聴でき，視聴中に生じた疑問をその場で解決できるシステムを構築し，ユーザーの検索負担軽減と ニュースに関する知識の向上を目的としている．

本研究では，ユーザーが NetTv を使用するにあたり，ニュースに関してユーザーが議論できるシステムを構築し，ユーザーのニュースに関する理解を深めることを目指す．議論に関連する研究 [5, 6] はいくつか行われているが，我々が構築する議論システムでは，ユーザーの主張に対して類似する反論意見をシステムが提示することで，ユーザーと議論を行う．その際のユーザーの主張や根拠の推定については検討を行った．[7] 本研究では，類似度をどのような文書のベクトル化手法を用いて計算するべきかについての検討を行う．

2　NetTv の概要

NetTv では，日々変化するダイナミックな環境において，音声認識により情報検索することを目的としている．そのため，インターネットニュースの映像メディアに対して，以下のような機能を円滑に行えることに重点を置き，システムを構築している．

- ネットニュースにおける動画インデキシング
- 音声インタフェースによる検索と質問応答機能

本研究では，ユーザの幅広い質問に対して，回答できるような QA(質問応答) モジュールを構築した．

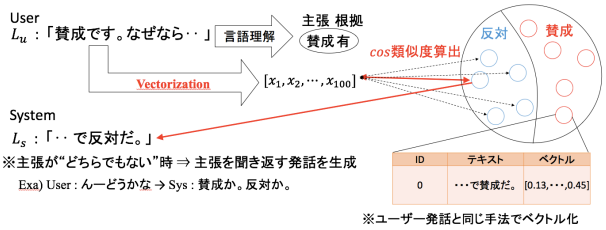

Fig. 1　Debate management

3　議論システム

ユーザーがニュースに関する情報を得るだけでなく，対話的な行為を通して理解を深めることができれば，見落としていた観点などを得ることができる．そこで，QA(質問応答) モジュールの機能の 1 つとして，議論システムを構築した．議論とは，互いに対立する意見を述べ，論じ合うことである．そのため，本研究では，Fig. 1 に示すように，言語理解部でユーザーの主張 (賛成/反対/どちらとも言えない) と根拠 (有/無) を推定し，推定したユーザーの主張に対して反論する意見 (ユーザーの主張が賛成なら反対の意見) をデータベースから決定し，ユーザーに提示することで議論を行う．ユーザーに意見を提示するために，あらかじめ，WWW 上で議題に対する賛成意見と反対意見を自動収集し，ディベートデータベースを構築し，システム発話として利用する．

3.1　ディベートデータベース

井上らの手法 [8] に基づいてディベートデータベースを構築する．井上らは任意の議題に対して，賛成意見と反対意見を検索 API を用いて WWW 上から自動で収集している．その方法は，初期検索と再検索の 2 段階にわたる検索手法である．初期検索では，具体的な検索クエリを検索 API に入力することにより，高精度の検索を行う．例えば， X を議題とするとき，検索 API に「X に賛成 (反対) です」と入力し，「X に賛成 (反対) です」という表現が含まれる Web ページを取得する．取得された Web ページからテキストを抽出し，検索クエリの表現を中心に前後の文を取得し，意見候補とする．ここで，初期検索で使用された検索クエリとして，例に挙げた表現の他に ”X(に or は) 賛成 (です or だ or で ある or します) ”

*Investigation of vectorizing sentence for generating supporting/opposing opinion in debate system, by Kazuaki Furumai, Tetsuya Takiguchi, Yasuo Ariki (Kobe univ.)

の同義表現を用いる．再検索では，初期検索で取得した意見の集合から，頻出する単語を関連語として抽出する．抽出した関連語の集合を「X 賛成 (反対)」の後ろに追加し，検索 API の入力として，Web ページを取得する．取得した Web ページからテキストを抽出し，検索クエリに使用した単語のうち，3 語以上を含むパッセージを意見候補として取得する．上記のような初期検索と再検索は，賛成と反対に関してそれぞれ行い，意見候補を収集する．しかし，再検索として取得した意見は質の悪い文になるため，フィルタをかけることで，精度が高い意見のみを取得する．初めに，初期検索で取得した高精度の意見候補を学習データとして，賛成，反対を分類する 2 クラスの SVM 分類器を構築する．構築した分類器に初期検索と再検索で取得した各意見候補を入力し，スコアを出力させる．スコアが賛成と反対の中間に近い値の意見候補を除外することで，高精度な意見のみをデータベースに格納する．また，取得された賛成意見と反対意見は，個別に LDA(Latent Dirichlet Allocation) を用いてクラスタリングし，各意見のクラスをデータベースに格納する．

3.2 議論制御部

議論制御部では，ユーザーの発話に対して，システムの発話を決定する．例えば，推定されたユーザーの主張が賛成の場合はディベートデータベースに格納されている反対意見群からシステム発話を選択する．その際，ユーザー発話と cos 類似度が最も高い反対意見を選択する．また，過去に選択した意見のトピック (クラス) とは異なるトピックの意見をユーザーに提示することで，ユーザーに幅広い意見を提示することを可能にしている．cos 類似度の計算に用いるベクトル化手法については以下の 4 つの手法を比較・検討した．

- TF-IDF に基づいた Bog-of-Words によるベクトル化
- Word2vec を用いた特徴ベクトル平均によるベクトル化
- Doc2vec によるベクトル化
- Sparse Composite Document Vectors (SCDV) を用いた特徴ベクトル平均によるベクトル化

4 文書のベクトル化手法

4.1 TF-IDF に基づいた Bog-of-Words

Bag-of-Words は，文書データを表現するために最も広く用いられている手法である．BoW では文書中に含まれている単語の出現頻度でベクトル化を行う．本研究では，$TF\text{-}IDF$ に基づいて，学習に不必要な単語を除いている．$TF\text{-}IDF$ は，文書中の単語に関する重要度を示す指標であり，TF と IDF の 2 つの指標の積で計算される．文書 d 内で任意の単語 t の出現回数を n_{td} とし，d 内すべての単語の出現回数の和を $\sum_s n_{sd}$ としたとき，$TF = n_{td}/\sum_s n_{sd}$ で計算される．したがって，TF の値が大きい単語すなわち，出現頻度が高い単語は重要である．また，任意の単語 t を含む文書数を df_t とし，全文書数を D としたとき，$IDF = \log(D/df_t)$ で計算され，IDF が高い単語は重要である．これは，多数の文書に出現する単語は重要でないことを意味する．本研究では，$TF\text{-}IDF$ が低い単語，すなわち出現頻度が 3 以下で，全体の 7 割以上の文書で出現するような単語は TF が低く，IDF が低いため削除する．そうして得られたベクトルデータに対して，Latent Semantic Indexing(LSI) を用いて次元削減を行い，200 次元の特徴量を分類に使用している．

4.2 Word2vec を用いた特徴ベクトル平均

Word2vec は，Mikolov ら [9] によって提案された，分布仮説に基づいた単語の分散表現を獲得する手法である．収集したディベートデータベースの意見群を用いて，Skip-gram モデルによる学習を行い，各単語について 200 次元の特徴ベクトルを学習した．学習した word2vec を用い，文章中の各単語の特徴ベクトルの平均をその文書のベクトルとして用いた．ただし，句読点や感嘆符などといった約物や数字は除去している．

4.3 Doc2vec によるベクトル化

Doc2vec[10] とは，Word2vec を文書単位に拡張したものであり，パラグラフや文書全体の意味表現を獲得するものである．収集した意見群を用いて学習を行い，各文書に対して 150 次元の特徴ベクトルを作成した．

4.4 SCDV を用いた特徴ベクトル平均

SCDV[11] は，一度学習した Word2vec による単語空間上でクラスタリングを行い，その後，各単語がどのようなトピックに属しているかを考慮して，特徴ベクトルを拡張させる手法である．クラスタリングには Gaussian Mixture Model (GMM) が用いられ，ディベートデータベースに格納した意見群を用いて作成した word2vec による 200 次元の特徴ベクトルを，10 個のトピックにクラスタリングした．結果として，SCDV による単語の特徴ベクトルは 2000 次元となった．

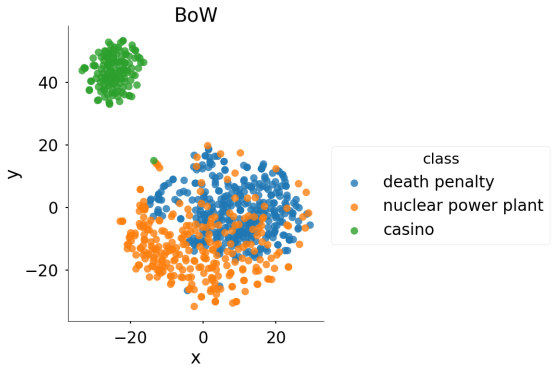

Fig. 2　Visualization of opinion with BoW

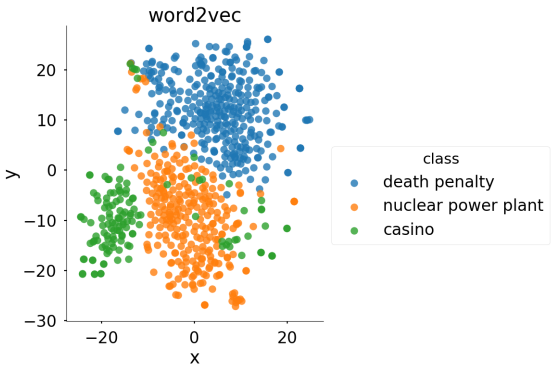

Fig. 3　Visualization of opinion with Word2vec

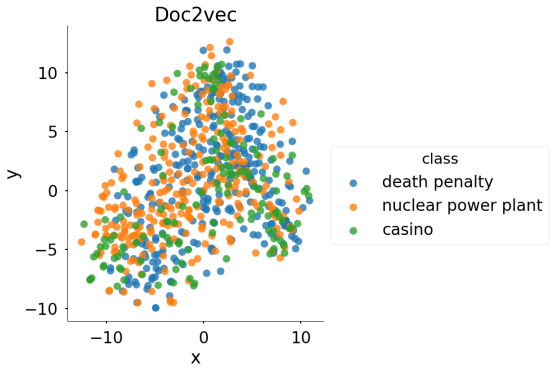

Fig. 4　Visualization of opinion with Doc2vec

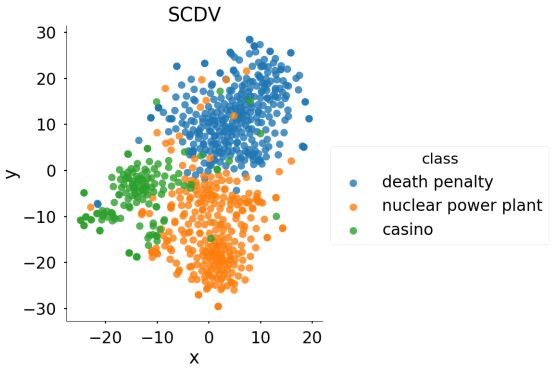

Fig. 5　Visualization of opinion with SCDV

5　実験と考察

取り扱う議題として，”カジノ法案”, ”死刑制度”, ”原発再稼働問題”の 3 つを設定し，それぞれについて WWW 上から自動で意見収集を行い，データベースを作成した．各議題で収集した意見数は Table 1 のとおりである．

Table 1　Number of opinions

	賛成	反対
カジノ法案	38	68
死刑制度	158	183
原発再稼動	54	220

獲得した各文書の特徴ベクトルに対して，t-SNE を用いて次元削減を行い可視化をおこなった．それぞれの手法による結果を，Fig. 2~5 に示す．

また，獲得した各意見の特徴ベクトルを用いて，文書の分類タスクによる比較も行った．LightGBM[12] を用いて，”死刑制度 (death penalty)”，”原発再稼働問題 (nuclear power plant)”，”カジノ法案 (casino)” への分類実験を行った．721 個の意見のうち，574 個で LightGBM の学習を行い，174 個の意見を分類した．結果を Table 2 に示す．

Fig. 2 より，BoW によるベクトル化では，カジノ法案に関する意見は他 2 つの意見から分離されているが，死刑制度と原発再稼働の意見は，Fig. 3, 5 の Word2vec, SCDV によるベクトル化に比べ分離されていない．Fig. 4 より，Doc2vec によるベクトル化では，3 つの議題が混ざり合っているように見え，Table 2 の結果を見ても精度が低いことが分かる．これは，文書単位を 1 つのデータとして扱う Doc2vec において，本研究で収集した意見数が学習データとして少なかったとからであると考えられる．対して，SCDV によるベクトル化は最も良い結果を示した．議論制御部における類似度計算に SCDV を用いた場合のユーザーとの議論の例を以下に示す．

Table 2　Document classification with LightGBM

	Accuracy	Precision	recall	f1
BoW	0.867	0.864	0.864	0.863
Word2vec	0.878	0.874	0.878	0.874
Doc2vec	0.626	0.623	0.626	0.615
SCDV	**0.952**	**0.953**	**0.952**	**0.952**

Sys　：原発についてどう思いますか．賛成ですか．反対ですか．
User：賛成です．
Sys　：何故ですか．詳しく教えてください．
User：手放しで賛成というわけではありませんが，

将来的には低いコストで安定した電力を供給できなければ様々な発展の妨げになると思われ，技術としては捨てるべきではないと思うからです．

Sys ：どんなに厳しい安全対策をしたところで，自然には勝てません．また，一旦事故が起きれば，県内広範囲にわたって人が住めない地域となることから，島根原発再稼働には反対します．

User：前回の事故を受けて，新しい基準が設けられていますが，その基準がまだ信用できないということでしょうか．

Sys ：原発再稼働に向けた原子力規制委員会の安全審査で，最も早く再稼働する可能性が高いとみられている九州電力川内原発（鹿児島県）．しかし再稼働審査では，大規模な火山噴火を“想定外“にしていることが判明しました．福島の...(長文のため省略)

User：そうだったんですね．

Sys ：またこんな意見もあります．センター側と県との間で議論になったのは，原子力発電所の監視体制など．運転状況に関してこれまで事業所はウソをついてきた．専門的な知見を持った職員を配置するなど監視体制を抜本的に強化すべきだ．(省略)

User：新基準を設けても，第三者による監視体制が整っていないということですね．

Sys ：原発輸出を戦略として掲げている以上，政府金融機関による支援は間違いではない．ただし，公金が...(省略)

システムの発話内で，島根原発の話になっていたり，敬体と常体が統一されていないなど課題が残るが，ユーザーが知らなかった情報を提示しつつ議論を行えている部分も見受けられる．最後のシステム発話に関しては，公金がどのようなどのようなことに使われているか議論するべきだという意見であったが，「第三者による監視体制」という文言に関わっていると思われるものの，論点としてはずれた発話となっている．また，意見数が少ないことも課題であり，今後より大規模なデータベースを作成できる手法の検討を行う必要がある．

謝辞　本研究の一部は，JSPS 科研費 JP17K00236，JP17H01995 の支援を受けたものである．

参考文献

[1] Lori Lamel *et al.*, ”User evaluation of the mask kiosk”, Speech Communication, pp. 131-139, 2002.

[2] Koichiro Yoshino, Tatsuya Kawahara, ”News Navigation System based on Proactive Dialogue Strategy”, Int’l Workshop Spoken Dialogue Systems, 2015.

[3] 木田 学　他, ”音声情報案内システムにおける質問応答データベース最適化手法の検討”, 情報処理学会, pp. 81-86, 2006.

[4] 田中 克幸，滝口 哲也，有木 康雄，”NetTv：NetNews とテレビ放送のクロスプラットホームにおける音声検索”, 日本音響学会誌，pp. 179-180，2007

[5] 佐藤 美沙　他，”国会会議録を用いたディベート人工知能による意見生成”, 人工知能学会全国大会論文集, 2017

[6] Peter Potash, Anna Rumshisky, ”Towards Debate Automation: a Recurrent Model for Predicting Debate Winners”, Empirical Methods on Natural Language Processing (EMNLP), 2017

[7] Rikito Marumoto, Katsuyuki Tanaka, Tetsuya Takiguchi, Yasuo Ariki , ”Debate Dialog for News Question Answering System ‘NetTv’ -Debate Based on Claim and Reason Estimation- ”, IWSDS, May 2018.

[8] 井上 結衣，藤井 淳，”Web 世論からの意見抽出と賛否に基づく分類”, 言語処理学会論文集, pp. 364-367

[9] Tomas Mikolov *et al.*, ”Distributed Representations of Words and Phrases and their Compositionality”, Proc. of NIPS, pp. 3111-3119, 2013

[10] Quoc Le, Tomas Mikolov, ”Distributed representations of sentences and documents”, Proc. of ICML, pp. 1188-1196, 2014

[11] Dheeraj Mekala *et al.*, ”SCDV : Sparse Composite Document Vectors using soft clustering over distributional representations”, Proc of EMNLP, 2017

[12] Guolin Ke *et al.*, ”LightGBM: A Highly Efficient Gradient Boosting Decision Tree”, Proc of NIPS, pp. 3146-3154 2017