Vol.2021-GN-114 No.7
Vol.2021-SPT-42 No.7
2021/5/10

大規模合意形成支援システムD-Agreeにおける BERTを用いた不適切発言フィルタリングの精度向上の検証

佐藤 拓実[1,a)]　長谷川 拓也[1,b)]　伊藤 孝行[2,c)]

概要：自然言語処理を用いた不適切発言フィルタリングの研究を，実オンライン議論での利用における精度向上を目指した手法の詳細と先行研究手法との比較検討について示す．近年，SNS 等の発展に併せて誹謗中傷を含めた不適切発言の増加が大きな社会問題となっている．そのため，ユーザー同士が交流を行うようなオンラインのサイトでは，不適切発言を自動的に判定し取り除く仕組みであるフィルタリングが重要視されてきている．我々はその中でも，ユーザー同士がテキストで議論を行うオンラインテキスト議論に注目をした．中でも，次世代の民主主義プラットフォームとして注目を集めているオンライン上のクラウド（Crowd）スケールの議論支援プラットフォームの一つである大規模合意形成支援システム D-Agree を利用する．自然言語処理を用いた不適切発言フィルタリングの先行研究では既に高い精度が報告されている．本稿では先行研究に対して学習データを追加するとともに，自然言語処理モデルにより精度が良いとされる BERT を用いた．実利用における有効性を評価する評価実験を行なった結果と，それらを先行研究手法の結果と比較をし考察を行なった．

Verification of Accuracy Improvement of Inappropriate Remarks Filtering Using BERT in the Crowd Discussion Support System, D-Agree

Abstract: This paper presents a research on inappropriate remark filtering using natural language processing, including details of the method aimed at improving accuracy in real online discussions and a comparison with previous research methods. In recent years, with the development of social networking services (SNS), the increase in inappropriate remarks, including slander, has become a major social problem. For this reason, filtering, which is a system that automatically detects and removes inappropriate remarks, is becoming more and more important for online sites where users interact with each other. In this paper, we focus on online text discussion, in which users discuss with each other in text. We use D-Agree, the Crowd Discussion Support System, which is one of the online crowd-scale discussion support platforms that are attracting attention as a next-generation democracy platform. Previous research on inappropriate remark filtering using natural language processing has already reported high accuracy. In this paper, we add training data to the previous studies and use BERT, which is considered to be a more accurate natural language processing model. This paper discusses the results of the accuracy evaluation in terms of actual use, and compares them with the results of the previous research methods.

1. はじめに

近年，オンラインでユーザー同士がコミュニケーションを行えるコミュニケーションツールが発展しており．誰もがウェブ上で気軽に情報を発信したり入手することができる．ウェブ上のコミュニケーションにはこれらのような大きなメリットがある一方，コミュニケーションツールにおける誹謗中傷的な発言による被害は大きな社会問題に発展している．そのため，SNS をはじめとしたユーザー同士が交流を行うようなウェブサイトでは，誹謗中傷等の不適切な発言を自動的に判定し取り除く仕組みであるフィルタリングが重要視されてきている．

本研究の目的は，不適切発言を自動的に判定し取り除く

[1] 名古屋工業大学
Nagoya Institute of Technology
[2] 京都大学
Kyoto University
[a)] sato.takumi@itolab.nitech.ac.jp
[b)] hasegawa.takuya@itolab.nitech.ac.jp
[c)] ito@i.kyoto-u.ac.jp

仕組みを実現し，誹謗中傷を含めた不適切発言に関する社会問題に対しての解決方法を示すことである．そのために本論文では，コミュニケーションツールの中でも特にオンライン上でのテキスト議論に注目し，議論中の不適切な発言を自動的に判定し取り除けるような，実オンライン議論での利用のための精度向上を目指す．第 3 章で述べるような不適切文書フィルタリングの先行研究ではすでに高い精度が示されている．しかし，一般的な評価方法である F 値でしか判定しておらず実フィールドへの応用という観点では評価が不十分である．そのため我々は実議論での利用という観点からの評価実験設定を提案すると同時に，F 値ではなくその評価実験における精度の向上を目指した．

本稿では，第 2 章で本論文にて注目するオンライン上でのテキスト議論プラットフォームである大規模合意形成支援システム D-Agree について紹介する．第 3 章で不適切発言フィルタリングの先行研究について紹介する．第 4 章で不適切発言フィルタリングの精度を向上させるために行った手法を説明し，第 5 章では議論プラットフォームでの実利用における精度を評価する評価実験設定について述べるとともに，実験の結果・考察について述べている．最後に第 6 章で本論文をまとめる．

2. 大規模合意形成支援システム D-Agree

ウェブ上のコミュニケーションの一つに，テキストでのオンライン議論がある．一般的にオンライン議論では，ユーザー同士があるテーマについて議論を交わし合意形成を目指す．中でも大規模合意形成支援システム D-Agree [1], [2]，そしてその前身である COLLAGREE [3], [4] では多くの社会実験で一般的なオンライン議論プラットフォームよりも議論支援に関する高い有効性を示しており [5], [6], [7], [8], [9], [10], [11], [12], [13]，D-Agree のようなオンライン上のクラウド（Crowd）スケールの議論支援プラットフォームは次世代の民主主義プラットフォームとして注目を集めている [14].

D-Agree の合意形成システムの概要を図 1 に示す．合意最適化エージェントによる様々な機能を通して合意形成を支援しているのが特徴である．中でも D-Agree の最も大きな特徴として自動ファシリテーションエージェントがある．合意形成支援として自動ファシリテーションエージェントがユーザーの議論に自動的に介入し，ファシリテートを行う．自動ファシリテーションエージェントは図 1 のように議論の合意構造に従って，ユーザーの投稿を課題・アイデア・長所・短所に分類しながら適切なファシリテートを行う．議論の合意構造は IBIS (Issue-Based Information System) 構造 [15] と呼ばれる創造的かつ建設的な議論を進めるための構造に従っている．自動ファシリテーションエージェントの自動ファシリテート（問いかけ介入）の実際の例を図 2 に示す．自動ファシリテーションエージェントがリアルタイムで行っている議論構造の抽出は，ノード分類とリンク抽出の 2 つのサブタスクからなっておりそれぞれ研究が進められている [16], [17]．なお，議論における意見のことをノードと呼び，意見間の関係のことをリンクと呼ぶ．

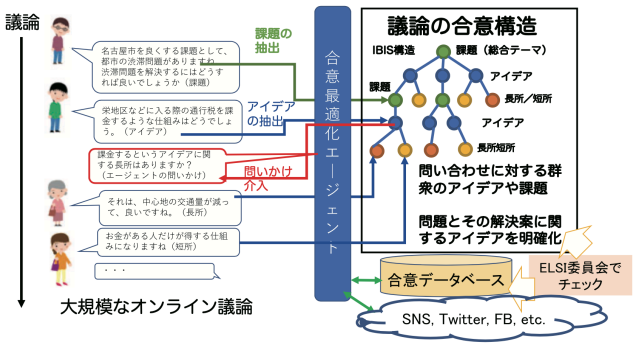

図 1　大規模合意形成システムの概要（[18] より引用）

図 2　自動ファシリテーションの成功事例（[18] より引用）

D-Agree 等オンライン議論の課題として，議論中に故意に攻撃的，差別的，侮辱的な投稿を行い炎上を引き起こす投稿がある．炎上を引き起こすような不適切な投稿があると議論の進行が妨げられるのは勿論，現在社会問題となっているような精神的な被害も起こり得る．そのため我々は，議論支援の一環として有効性の高いフィルタリングを実装し炎上を引き起こすような発言を予め防ぐことで，議論における合意形成を支援することができる．

3. フィルタリング手法の先行研究

3.1 Filtering of Impertinent Remarks using distributed expression

関連研究として，平石ら [19], [20] の不適切文書判定の研究がある．平石らの研究では，分散表現化した文書を文書類似度計算とディープニューラルネットワーク（DNN）を用いて，暴力的文書・性的文書・一般的な文書の 3 クラス分類を行い不適切発言フィルタリングを目指した．平石らの手法の概要を図 3 に示す．以下図 3 に沿った手法の詳細を [20] より引用する．

ステップ 1

学習データは Web 上からクローラーを用いて学習データ

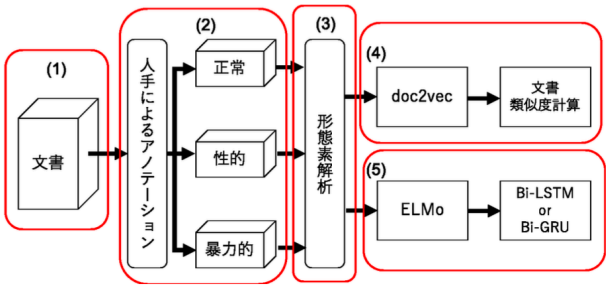

図 3 平石らの提案手法の概要図（[20] より引用）

を自動で収集し，収集に用いるサイトは 5 ちゃんねる掲示板，および本研究室で運用を行なっている COLLAGREE，D-Agree とした．今回の実験では約 1 万件のデータを収集した．

ステップ 2

収集した学習データを正常な投稿，性的（卑猥な発言や，出会い系の発言など）な投稿，暴力的（攻撃的，差別的，侮辱的など）な投稿に分類する．ラベル付けは，人の目視により行なった．

ステップ 3

形態素解析を用いて学習データを単語分割する．本研究では，MeCab [21] を形態素解析機に用いて，ステップ 4 もしくはステップ 5 に入力した．

ステップ 4

形態素ごとに解析された文書データを doc2vec モデル [22] に入力し，分散表現にする．フィルタの構築では分散表現になった各ラベルの文書を学習し，代表ベクトルとする．生成された代表ベクトルと新規投稿の間で類似度計算を行い，最も値が高くなったクラスに分類する．

ステップ 5

形態素ごとに分割された文書を形態素ごとに ELMo [23] で分散表現にする．分散表現となったものを Bidirectional Long Short-Term Memory Unit (Bi-LSTM) [24], [25], Bidirectional Gated Recurrent Unit (Bi-GRU) [26] のレイヤーに入力として，それぞれの出力を出力層に渡す．出力層の活性化関数は softmax を用いて，各クラスの確率を求め，最も確率の高いクラスに分類する．

ステップ 1 で述べているように，教師データは平石らがウェブ上からクローラーを用いて収集したものを利用している．収集に用いたサイトは 5 ちゃんねる掲示板，および COLLAGREE，D-Agree である．性的である文書 2,400 件，暴力的である文書 1,920 件，一般の発言 3,200 件を集めた．これらの定義として，性的である文書とは性器を連想させるなどといった嫌がらせを目的とするセクハラ発言とし，暴力的である文書とは「死ね」や「殺す」などといった明確に害悪を加える意思を示すものや相手の名誉を毀損させるような他人を恐れさせることを目的とする発言としており，それらのラベル付けは平石らによって人手で行われた．平石らの手法の精度を表 1，表 2 に示す．表 2 はテストデータとなる文書を性的・暴力的・一般の文書の 3 クラスに分類した際の F 値である．表 2 は文書を不適切か・適切かの 2 クラスに分類し，先行研究と比較した際の F 値である．これらより平石らの提案手法は先行研究を上回る精度を示しており，高い精度で不適切文書を分類できていることが分かる．

表 1 平石らの 3 クラス分類精度の平均 F 値

Method	F 値
doc2vec Ensemble Classifier	0.9360
Bi-LSTM	0.9190
Bi-GRU	0.9164

表 2 平石らの 2 クラス分類の先行研究との精度比較

Method	F 値
doc2vec Ensemble Classifier（平石らの提案手法）	0.962
BI-LSTM （平石らの提案手法）	0.944
Bi-GRU（平石らの提案手法）	0.941
PV-CBOW [27]	0.943
Naive Bayes - Bayesian FIlter [28]	0.93
Gray Robinson - Bayesian Filter [29]	0.884

本研究での提案手法との違いは，文書の分散表現化モデルに言語処理分野の多数のベンチマークタスクで SOTA (State of the Art) を記録した BERT (Bidirectional Encoder Representations from Transformer) [30] を用いたこと，また D-Agree 上での議論における一般性の高い実議論データを適切な発言データとして追加し学習させ，更に実利用を想定した評価実験を行い精度検証した点である．

4. 精度向上手法の提案

4.1 doc2vec から BERT への置換

BERT (Bidirectional Encoder Representations from Transformer) [30] は 2018 年登場当時，言語処理分野の多くのベンチマークタスクで SOTA (State of the Art) を記録した自然言語処理モデルである．双方向 transformer を用い文章を双方向（文頭から文末・文末から文頭）から学習しているのが特徴である．言語処理の多数の分野で高い精度を誇る BERT への置換によって汎用性のある精度向上を期待する．我々は，平石らの手法の doc2vec 部分を BERT に置き換えた．それに伴って MeCab 部分は SentencePiece [31] に，文書類似度計算部分は全結合層による計算へと置き換えた．つまり，手法の概要は図 4 のようになる．基本的には平石らの提案手法をベースとしており，文書を入力すると性的な文書・暴力的な文書・一般の文書の 3 クラスに分類される．今回は不適切な文書（性的な文書・暴力的な文書）と適切な文書（一般の文書）の 2 クラス分類として本手法を利用する．これは，ウェブ上のコミュニケーションツールでの実利用を考えた際，投稿を削

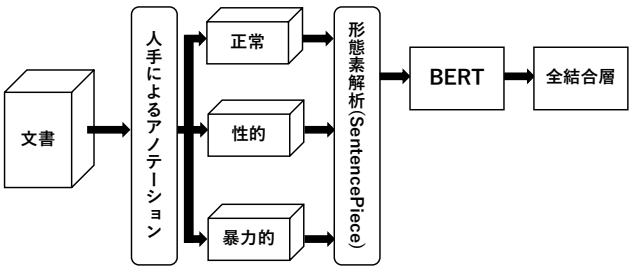

図 4　BERT を用いた新しいフィルタリング手法の概要図

除する・削除しないの 2 クラスで利用されることを想定しているからである．平石らの手法と異なるのは形態素解析以降である．形態素解析では，SentencePiece を用いて文書をトークン化する．トークン化された文書を BERT モデルに入力し分散表現にする．BERT では日本語 wikipedia によって事前学習されたモデルを用い，fine-tuning によって不適切発言の分類に対応させる．入力した発言は全結合層へと入力され，性的な文書・暴力的な文書・一般の文書の 3 クラスに分類がされる．性的な文書・暴力的な文書として分類されたものは不適切な発言，一般の文書として分類されたものは適切な発言とみなす．

4.2　学習に利用するデータの追加

先程の BERT に置換した手法にさらに，学習データを追加する．追加したデータは，COLLAGREE・D-Agree によって 2018 年に行われた名古屋市次期総合計画議論，2019 年に行われた浜岡原発移転に対する県民投票の可否に関する議論，2020 年に行われた日本の社会問題に関する議論，の全て合わせて 1,400 件である．目視で確認したところ不適切な発言は見当たらなかったため，これらを全て適切な文書として追加した．元々平石らが利用していた適切な文書のデータと合わせ，適切文書のデータは 4,600 件となった．学習させるデータ数を増やすことで，手法により汎用性を持たせ一般の議論における精度向上を期待する．更に学習データの増減による精度の違いも比較検討する．

5.　実利用を想定した評価実験と考察

5.1　実験設定

フィルタリング評価実験には，D-Agree 上で行われた実際の議論の議論データを用いた．議論は 2021 年 1 月に行われ，一般募集した 30 代の男女 16 名を 4 名ずつの 4 グループに分け行なったものである．議論されたテーマは 4 種あり，日常生活の差別について，仕事に対するモチベーションを保つ方法について，子ども達の創造性を向上させるためには，新型コロナウイルスによる私たちの仕事・学習・生活の変化について，である．本議論データは一般性が高いと判断しテストデータに用いた．今回の評価実験では各グループから 1 テーマずつを利用し，計 307 投稿を利用する．それらの各投稿を研究室の学生 1 名が「議論から削除すべき・しないべき」という尺度でアノテーションを行い，アノテーションされたテストデータに対して各種手法を用い正答率を求める．アノテーションを担当する学生には，「議論から削除すべき・しないべき」は有害か無害か，炎上を引き起こすか引き起こさないか，と言い換えることもできると具体的に伝えた上でアノテーションを行なってもらった．本実験によって，不適切発言フィルタリングを実議論に用いた場合における有効性を評価することができ，実際の議論に用いる際の問題点や不適切発言フィルタリングの必要性まで考察することができる．これらを，先行研究である doc2vec Ensemble，ベースラインとして doc2vec，BERT を用いた新しい手法，の計 3 手法について，4.2 節で述べた学習データを追加する前と後の計 6 パターンについて実験を行う．4.2 節で述べた学習データを追加する前の学習データは平石らが利用していた学習データと同じものを利用する．doc2vec Ensemble は平石らの手法に基づいて再現している．また，doc2vec では python のライブラリである gensim の doc2vec モデルを用いている．

表 3　評価実験の結果

Method	正答率
doc2vec	0.560
doc2vec + 追加学習	0.834
doc2vec Ensemble（平石らの提案手法 [19], [20]）	0.879
doc2vec Ensemble + 追加学習	0.896
BERT	0.531
BERT + 追加学習	0.906

表 4　doc2vec Ensemble+追加学習の評価実験結果の混同行列

	適切	不適切
適切	275	26
不適切	5	1

5.2　実験結果

各手法を用いた評価実験の結果を表 3，表 4，表 5 に示す．学習データを追加する前では，先行研究の doc2vec Ensemble が最も高い正答率を示し，BERT を用いた手法が最も低い正答率を示した．また学習データを追加した後では，BERT が最も高い正答率を示し，doc2vec 単体を用いた手法が最も低い正答率を示した．ただし，どれも正答率としては 0.8 を超えており，先行研究の doc2vec Ensemble と BERT を比較すると正答率に大きな差はないことが分かる．

更に，実験結果についていくつか実例を示す．これらはどちらも追加データ有の状態で比較している．

表 5　BERT+追加学習の評価実験結果の混同行列

	適切	不適切
適切	276	25
不適切	3	3

適切な投稿に対して，doc2vec Ensemble が正しく判定し BERT が正しく判定できなかった例

- テレワークを始めるようになってから毎朝、朝のルーティンとして豆からコーヒーを淹れています。濃い目の熱いコーヒーを一気に飲み干すと、仕事モードに入れる気がします。気分を切り替えたい時や集中力が続かない時のために何でも良いので仕事モードに入る為のルーティンを決めておくと、仕事のモチベーションを維持するのに役に立つかもしれません。
- 私もエイジさんのご意見に近いです。
  在宅勤務とはいえ「仕事モード」に入るために、私は寝間着やジャージで仕事はしないようにしています。あとは、愛犬の「あそぼ」攻撃に負けない強い心をもって仕事をしています (笑)

適切な投稿に対して，BERT が正しく判定し doc2vec Ensemble が正しく判定できなかった例

- 全く知識が無ければ創造性というステップに行けませんから、知識を付けさせることが創造性への第一歩になるというのは賛成です。
  私が書いたのは、知識を付けることへの固執が問題だということです。
- そのまま失敗することを素直にうけとめられずに大人になっていくのが多いように見受けられます。

不適切な投稿に対して，BERT が正しく判定し doc2vec Ensemble が正しく判定できなかった例

- マイコさんの気持ちわかります。
  黒人の方は体格もいいですし、勝手なイメージで襲われたらまず敵わないなと思ってしまいます。
  メディアなどで見聞きする情報から少し怖いイメージを私ももってしまいます。
  個人個人はもちろんそんなことはないことはわかっていてもです…

適切な投稿に対して，doc2vec Ensemble も BERT も正しく判定できなかった例

- 工事の音を曜日によって調整している新たな発見も、コロナがもたらしてくれた気付きですね (笑)
  私はコロナは色んな気付きや、当たり前が当たり前でないと気付かせてくれるチャンスだと思っています。
- 疲れて帰ってきても、待っててくれる人（猫ちゃん）がいると思えば、毎日頑張れますね！わたしもです＾＾
- 私は現在、コロナの影響もあり在宅勤務をしていますが、幸いこういった仕事のやり方の方が自分には合っていたようで、またコロナ前のように、毎朝毎晩、電車に揺られて仕事場まで出向く生活に戻れと言われたら厳しいというのが正直なところなのですが、コロナ前は仕事をしながら、よく家族と飼っている猫の事を思い出しては家族はともかく猫のために仕事をしなければと自分のモチベーションを高めていた事を思い出します。自分ひとりですと最悪のたれ死んでもその時はその時だとついつい考えてしまうのですが、自分の事を必要としてくれている人（猫？）がいたり、誰かが自分を待っていてくれているというのはありがたい事だと感じます。最も猫の方はお前なんか別にいなくてもいいよと思っているかもしれませんが。
- レイナさんのご意見についてどう思われますか？

5.3　考察

学習データを追加する前の結果に注目すると，BERT よりも doc2vec Ensemble の正答率の方が圧倒的に高い．また学習データを追加した後も，正答率は BERT の方が高くはあるが，doc2vec Ensemble との差は僅かである．これらのことから，BERT は学習データが十分にある際は有効であり，doc2vec Ensemble は学習データが十分でない際も安定的な精度を出すことができることが考えらえれる．また，学習データを追加する前は BERT は doc2vec よりも正答率が低くなっている．これは，BERT の性質上，読み込むトークン数の上限を 256 としているため長文の投稿が最後まで読み込まれないことが精度に影響していることが考えられる．一方表 4，表 5 で示したように，追加学習をした後の結果に対する混同行列に注目すると，doc2vec Ensemble では不適切発言を適切と判断してしまう割合が BERT に比べて高い．不適切発言フィルタリングにおいては不適切な発言を不適切と判定する精度が重要であるので，BERT の方が優れていると言える．

課題として，適切な投稿に対して doc2vec Ensemble も BERT も正しく判定できなかった例として「レイナさんのご意見についてどう思われますか？」という投稿がある．これは自動ファシリテーターの投稿である．ファシリテーターの発言はルールベースでフィルタリングされないようにするなど，実利用に向けて対策が必要である．

6.　結論

本論文では，オンライン上でのテキスト議論プラットフォームにおいて，不適切発言フィルタリングの実利用を目指した精度向上の検証を行った．我々は BERT を用いた新しい手法を提案し，先行研究と比較を行なった．評価実験では実議論データを用いて，実利用の観点から評価実験を行なった．結果として，我々の BERT を用いた新しい手法が最も精度が高いことが示されたが，先行研究の doc2vec Ensemble との精度の差は僅かであった．ただ，不

適切発言を不適切発言と正しく判定する割合は BERT の方が優れていたため，実用的な観点では BERT が優れていると考えた．また，学習データが少ない場合においては BERT を用いた我々の手法は有効ではなく，先行研究で用いられている doc2vec Ensemble の方が優れていた．

現在は，最も高い有効性を示した BERT の手法を用いて，不適切発言の自動フィルタリングシステムの D-Agree への実装を進めている．ただし，フィルタリングを全自動で行ってしまうと一般の投稿まで消されてしまいユーザーの不満を溜めてしまう一因になりかねないことから，人力とのハイブリットとして実装することを考えている．具体的には，自動的にフィルタリングされた投稿を議論テーマの管理者が確認でき，万が一一般の投稿が自動的に削除されていた場合は取り消せるようにする．我々の不適切発言フィルタリングを実際のシステムに組み込むことによって，議論における合意形成支援となることを期待する．

謝辞 本研究は JST CREST 研究課題番号 JPMJCR20D1 の助成を受けている．

参考文献

[1] Ito, T.: Discussion and Negotiation Support for Crowd-Scale Consensus, *Handbook of Group Decision and Negotiation* (Kilgour, Marc, D., Eden and Colin, eds.), Springer (2021).

[2] Ito, T., Suzuki, S., Yamaguchi, N., Nishida, T., Hiraishi, K. and Yoshino, K.: D-agree: Crowd Discussion Support System based on Automated Facilitation Agent, *Proceedings of 35th AAAI conference*, Vol. 2020 (2020).

[3] Ito, T., Imi, Y., Sato, M., Ito, T. and Hideshima, E.: Incentive mechanism for managing large-scale internet-based discussions on collagree, *Collective Intelligence*, Vol. 2015 (2015).

[4] Ito, T., Imi, Y., Ito, T. and Hideshima, E.: COLLAGREE: A faciliator-mediated large-scale consensus support system, *Collective Intelligence 2014* (2014).

[5] Hadfi, R., Haqbeen, J., Sahab, S. and Ito, T.: Argumentative Conversational Agents for Online Discussions, *Journal of Systems Science and Systems Engineering. Special Issue on AI-enabled System Simulation and Modelling* (2020).

[6] Hadfi, R. and Ito, T.: Exploring Interaction Hierarchies in Collaborative Editing using Integrated Information, *The 8th ACM Collective Intelligence 2020, Boston-Copenhagen* (2020).

[7] Haqbeen, J., Ito, T., Hadifi, R., Nishida, T., Sahab, Z., Sahab, S., Roghmal, S. and Amiryar, R.: Promoting Discussion with AI-based Facilitation: Urban Dialogue with Kabul City, *The 8th ACM Collective Intelligence 2020, Boston-Copenhagen* (2020).

[8] Haqbeen, J., Ito, T., Hadifi, R., Nishida, T., Sahab, Z., Sahab, S., Roghmal, S. and Amiryar, R.: Agent that Facilitates Crowd Discussion, *Proceedings of ACM Collective Intelligence*, Vol. 2019 (2019).

[9] Haqbeen, J., Ito, T., Sahab, S., Hadfi, R., Okuhara, S., Saba, N., Hofaini, M. and Baregzai, U.: A Contribution to COVID-19 Prevention through Crowd Collaboration using Conversational AI & Social Platforms, *IJCAI 2019 Workshop on AI for Social Good* (2020).

[10] Ito, T.: Towards Agent-based Large-scale Decision Support System: The Effect of Facilitator, *The 51st Hawaii International Conference on System Sciences (HICSS2018)* (2018).

[11] Nishida, T., Ito, T. and Ito, T.: Verification of Effects Using Consensus-Building Support System in Continuous Workshops for City Development, *Journal of the Science of Design* (2018).

[12] Kawase, S., Ito, T., Otsuka, T., Sengoku, A., Shiramatsu, S., Matsuo, T., Oishi, T., Fujita, R., Fukuta, N. and Fujita, K.: Cyber-physical hybrid environment using a largescale discussion system enhances audiences' participation and satisfaction in the panel discussion, *The IEICE Transactions on Information and Systems*, Vol. E101.D, No. 4, pp. 847–855 (2018).

[13] Nishida, T., Ito, T., Ito, T., Hideshima, E., Fukamachi, S., Sengoku, A. and Sugiyama, Y.: Core Time Mechanism for Managing Large-Scale Internet-based Discussions on COLLAGREE, *the Proceedings of the 2nd IEEE International Conference on Agents (IEEE ICA2017)* (2017).

[14] Malone, T. W.: *Superminds: The surprising power of people and computers thinking together*, Little, Brown Spark (2018).

[15] Kunz, W. and Rittel, H. W.: *Issues as elements of information systems*, Vol. 131, Citeseer (1970).

[16] Suzuki, S., Yamaguchi, N., Nishida, T., Moustafa, A., Shibata, D., Yoshino, K., Hiraishi, K. and Ito, T.: Extraction of Online Discussion Structures for Automated Facilitation Agent, *Annual Conference of the Japanese Society for Artificial Intelligence*, Springer, pp. 150–161 (2019).

[17] Suzuki, S., Ito, T., Moustafa, A. and Hadfi, R.: A Node Classification Approach for Dynamically Extracting the Structures of Online Discussions, 人工知能学会全国大会論文集， Vol. JSAI2020, pp. 2G5ES302–2G5ES302 (2020).

[18] 伊藤孝行，柴田大地，鈴木祥太，山口直子，西田智裕，平石健太郎，芳野 魁：エージェント技術に基づく大規模合意形成支援システムの創成：自動ファシリテーションエージェントを用いた大規模社会実験，人工知能学会全国大会論文集，Vol. JSAI2019, pp. 2F3OS5a01–2F3OS5a01 (2019).

[19] 平石健太郎，柴田大地，西田智裕，山口直子，鈴木祥太，芳野 魁，Moustafa, A.，伊藤孝行：Filtering of Impertinent Remarks using distributed expression, 人工知能学会全国大会論文集， Vol. JSAI2019, No. 0, pp. 2F4OS5b04–2F4OS5b04 (2019).

[20] 平石健太郎，柴田大地，西田智裕，山口直子，鈴木祥太，芳野 魁，伊藤孝行：分散表現を用いた不適切文書判定，研究報告知能システム（ICS），No. 14 (2019).

[21] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying conditional random fields to Japanese morphological analysis, *Proceedings of the 2004 conference on empirical methods in natural language processing*, pp. 230–237 (2004).

[22] Le, Q. and Mikolov, T.: Distributed Representations of Sentences and Documents, *Proceedings of the 31st International Conference on Machine Learning* (Xing, E. P. and Jebara, T., eds.), Proceedings of Machine Learning Research, Vol. 32, No. 2, Bejing, China, PMLR, pp. 1188–1196 (online), available from ⟨http://proceedings.mlr.press/v32/le14.html⟩ (2014).

[23] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L.: Deep Contextualized Word Representations, *Proceedings of the 2018 Con-*

ference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, Louisiana, Association for Computational Linguistics, pp. 2227–2237 (online), DOI: 10.18653/v1/N18-1202 (2018).

[24] Hochreiter, S. and Schmidhuber, J.: Long short-term memory, *Neural computation*, Vol. 9, No. 8, pp. 1735–1780 (1997).

[25] Schuster, M. and Paliwal, K. K.: Bidirectional recurrent neural networks, *IEEE transactions on Signal Processing*, Vol. 45, No. 11, pp. 2673–2681 (1997).

[26] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation, *arXiv preprint arXiv:1406.1078* (2014).

[27] 佐藤元紀，伊藤孝行：Paragraph Vector と多層パーセプトロンを用いた有害文書分類手法の提案，第 77 回全国大会講演論文集，Vol. 2015, No. 1, pp. 165–166 (2015).

[28] 大塚孝信，Deyue, D.，伊藤孝行：3 単語共起フィルタリングによる有害文書分類手法と大規模データ処理，電気学会論文誌. C， Vol. 134, No. 1, pp. 168–175 (2014).

[29] 吉村卓也，藤井雄太郎，伊藤孝行：RF-007 Robinson 型判定手法を用いた単語共起フィルタの検証 (FIT 論文賞受賞論文, 情報推薦,F 分野:人工知能・ゲーム)，情報科学技術フォーラム講演論文集，Vol. 10, No. 2, pp. 85–90 (2011).

[30] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019).

[31] Kudo, T. and Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing (2018).