大規模因果関係テキストを用いた反論生成手法

鎌倉 まな [1,2]　飯田 龍 [1,2]　呉 鍾勲 [1]　田仲 正弘 [1]
Julien Kloetzer[1]　淺尾 仁彦 [1]　鳥澤 健太郎 [1,2]
[1] 国立研究開発法人 情報通信研究機構 (NICT)
データ駆動知能システム研究センター (DIRECT)
[2] 奈良先端科学技術大学院大学 先端科学技術研究科
{kamana,ryu.iida,rovellia,mtnk,julien,asao,torisawa}@nict.go.jp

概要

議論は他者と意見交換を行い意思決定を行う重要な手段であり、ユーザと議論する対話システム [1, 2, 3] が研究されてきたが、データ構築の困難さにより適用できる議論は限定的であった。本研究では、ブレインストーミングが可能な音声対話システムの実現を目指し、Web から 7.9 億件の大規模な因果関係を表すテキストを、テキスト中の省略を補完しつつ収集し、それを活用することで、ユーザの多様な主張に反論を出力する手法を提案する。また、提案手法が出力した反論を人手で評価した結果を報告し、今後対処すべき課題について論じる。

1　はじめに

議論は日常生活と社会生活において、他者との間でそれぞれの意見を交換し、集団内でより良い意思決定を行うため重要な手段である。特に、他者の意見に対して反論することで、集団による不適切な意思決定を回避したり、より精密な意思決定を行うことが可能になる。例えば、ある人が「リニアが奈良を通ったらいいよね」と言ったとき、他の人が「でも、リニアが奈良を通ると京都の経済力が落ちるかもしれないよ」と、潜在的リスクを挙げて反論を行えば、より適切な意見、意思決定につながることが期待できる。本研究では、Walton [4] が論証スキームの一つとして挙げた、ネガティブな帰結からの論証 (Argument from Negative Consequences、ANC) 「もし A を実行すると、B が起きる。B は悪い結果である。ゆえに、A を実行すべきでない」をベースとし、大量の Web データから反論となるテキストを見つけ出す手法を開発した。上述のリニアに関する反論は、A を「奈良にリニアを通す」、B を「京都の経済力が落ちる」とすれば、まさにこの ANC の典型例と言える。反論となるテキスト候補としては、Web から抽出した因果関係知識（例えば、原因文「奈良にリニアを通す」、帰結文「京都の経済力が落ちる」のように、因果関係を持つ原因文、帰結文の対）を 7.9 億件用いており、多岐にわたる話題に関して反論テキストを発見することが可能である。

NICT DIRECT では、現在、雑談対話システム WEKDA [5] や介護支援用対話システム MICSUS[1)] [6, 7] の開発を行なっている。これらのシステムは Web 情報を用いて雑談を行うが、本研究で開発した手法は、そうした雑談に反論を導入し、多様な意思決定に関するブレインストーミングを可能にし、これらの対話システムの適用範囲をより広げる狙いがある。

提案手法では、ユーザの主張が入力されると、因果関係知識のデータベースから、BERT ベースのモデルを複数個使い、ユーザの主張を含意する／に含意される原因文を持つ因果関係で、帰結文がなんらかのトラブルを表しているものを反論として発見する。手法の評価のため、「リニアが奈良を通る」のような何らかの出来事、行為、施策を表す主張を 500 件作成し、そのうち提案手法が反論候補を出力できた 452 件について、著者以外の 3 人の作業者の多数決で、それらの反論候補が本当に反論として成立するかどうか判定した。各々の主張に対して、トラブル認識のスコアの最上位の出力が反論として成立した割合は 50.0%、3 位まで出力した場合、そのうちのどれかが反論として成立した割合は 79.0%であった。この精度は高くはないが、反論生成に特化したデータを使っていない初の試みとしてはまずまずの

1) けいはんな R&D フェア 2021「高齢者介護支援マルチモーダル音声対話システム MICSUS」 https://keihanna-fair.jp/exhibition/ai/899 、YouTube NICTchannel「高齢者向けマルチモーダル音声対話システム“MICSUS”」 https://www.youtube.com/watch?v=gCUrC3f9-Go

結果であると考えている。本論文の最後で、誤りの原因も分析し、今後の改善策についても言及する。

2 関連研究

これまでもユーザと議論する手法がいくつか研究されてきた。Sakai ら [8, 9] は、約 2 千件の主張や根拠を表す文を支持・不支持関係で結びつけた論証構造を手動で作成し、論証構造に基づいてユーザと議論する対話システムを提案した。また、Reddit 等のオンラインディベートフォーラムのデータを用いて、議論のためのテキストを発見したり、生成したりする研究もある [2, 10, 11] 。これらの研究の課題は、まず、議論可能なトピックや反論がディベートフォーラムのデータや人手で作成したデータによって強く制約されてしまうことである。また、テキスト生成を行う手法では、根拠のない議論を行ってしまう可能性がある。加えて、英語ではディベートフォーラムが盛んであり、中には支持・不支持等のタグづけがユーザによってなされているものもある一方、日本語ではそこまでデータが集まっておらず、日本語を対象とするシステムでの活用は難しいという問題がある。一方、我々の手法では、対象が反論に限定されてはいるが、ディベートフォーラムであるか否かを問わず、非常に大量の Web ページから取得した因果関係テキストを使うため、極めて広範な話題に関して反論を出力でき、また、反論は実際に誰かが Web に書いたテキストを元にしているため、全く根拠のない反論が出力されるリスクはきわめて小さく、優位性があるものと考えている。

3 反論を出力するアルゴリズム

提案手法は、ユーザの主張を表す入力文に対して反論を出力するが、4 節で述べる大規模な因果関係に関する知識ベースと、入力文に関して知識ベースを検索した結果をフィルタリングするための各種分類モデル（4 節で述べる含意等認識器、トラブル認識器）を利用する。因果関係に関する知識ベースは、原因と帰結のペアがそれぞれ文の形式で（例えば、原因文「これから導入されるマイナンバーは、国だけではなく、一般企業も給与の支払いなどで使用する」と帰結文「マイナンバーで情報漏えいが起きるリスクが高い」のように）保持している。Hashimoto らの研究 [12] と異なり、原因文、帰結文にはそれぞれ名詞 1 個、述語 1 個といった制限はなく、任意の形態の文が含まれる。

手続きとしては、入力文に含まれる内容語（「プロ野球に FA 制度がある」の「プロ野球」や「FA 制度」）を原因文に含む因果関係を、Lucene (https://lucene.apache.org/) を用いて検索し、次に得られた原因文・帰結文ペアの各々に関して、含意等認識器、トラブル認識器を適用し、1) 入力文と含意（含意の方向は問わない）もしくは等価であると分類された原因文と 2) トラブルを含むと分類された帰結文からなるペアのみを選択し、その帰結文を反論として出力する。なお、評価ではトラブル認識器の分類スコアが最も高い帰結文を最大で 3 個出力した。なお、後述するように含意等認識器、トラブル分類器ともに、BERT ベースの二値分類器であり、入力が含意等の性質を持つ確率を分類スコアとして出力しているが、その確率が 0.5 を上回るものをその性質を持つ入力であるとみなした。本手法が出力した反論の例を表 1 に示す。

4 因果関係獲得と含意等文間関係・トラブル認識処理

3 節のアルゴリズムで利用する知識の獲得は下記の手順で行う。まず、Web から収集した 7 文からなるテキストパッセージと、そのパッセージに Oh ら [13] の CRF ベースの因果関係抽出手法を適用して得られる因果関係の候補を表すテキストの原因部分と帰結部分の単語列を入力とし、さらに高精度な因果関係を抽出するため、BERT で重ねて分類を行った。この BERT モデルの学習のためにアノテータ 3 名の多数決を実施し、学習用、開発用、評価用データとして、それぞれ 127,240 事例、 20,937 事例、 20,954 事例を得た。実際には、呉らの因果関係をさらに、解決法、「〜が起きると、〜の可能性が高くなる」等いくつかのタイプの因果関係に再分類しており、上記学習データはそれらの再分類された因果関係のタイプに関するアノテーションを含んでいるが、本研究では、先述の「可能性が高くなる」因果関係のみを利用するので他の詳細は割愛する。NICT が収集した大規模 Web テキスト 350GB で事前学習した $\text{BERT}_{\text{large}}$ でファインチューニング[2)]した結果、評価用データに関する上述の「可能性が高くなる」因果関係に関する自動分類の性能は平均精度で 91.7% となった。

上述の手法で抽出された因果関係は、ゼロ照応等による省略が含まれており、原因文、帰結文それぞれを単独で読んでも意味が通らないことが多い。こ

2) 他の関係判別の学習データと multi-task で学習した。

表 1　入力と反論の例

入力	反論
マイナンバーを導入する	マイナンバーで情報漏えいが起きるリスクが高い
墓じまいを進める	都市部では墓じまい後の墓石の放置問題が発生する
プロ野球にＦＡ制度がある	一部の球団にスター選手が集中してしまい、特にパリーグの戦力が著しく低下した
所得税を上げる	働き盛りの会社員の負担が増える
物事の結果よりも過程を重視する	ビジネスの世界では苦労する
イギリスのＥＵからの脱退が決まった	世界の３大マーケットのマーケットバランスが崩れる恐れがある
小学校でプログラミングを必修にする	家庭でのプログラミングの教育に不安を感じる
統合型リゾートを誘致する	外国人向けの宿泊施設の不足が懸念されている

のため、省略等が補完され原因文、帰結文それぞれの単体を読んで内容が理解できるような形式で整形された原因と帰結の文の対を生成する課題を設計し、その処理に必要な判定、生成の学習・評価用データを作成した。先述の BERT 学習用データの原因部分、帰結部分に各事例 1 名のアノテータが元パッセージからの省略等の補完や文の整形などを行う編集作業を実施した。また、省略が補完された原因文、帰結文アノテーション学習用、開発用、評価用データの件数はそれぞれ他の関係と合わせて 113,486 事例、17,216 事例、17,364 事例である[3)]。これをもちいて、上述の $\text{BERT}_{\text{large}}$ をベースとして、さらに事前学習を行った Transformer Encoder-Decoder モデルをファインチューニングして生成器を作成し、この性能は評価データの ROUGE-1 F 値で 80.0 %である。これらの因果関係判定器、生成器を Web 約 200 億ページから得た約 35 億パッセージに適用し、「可能性が高くなる」関係に関しては約 7.9 億件の知識を獲得した。

また、含意の認識処理についても、類似する文対をヒューリスティックに獲得し、それらに対して意味的に等価、含意（含意の方向性として、２つの可能性があるので、２種類の関係）、（本研究では使用していないが）矛盾の 4 つの関係をアノテーションした。5 万件の事例に対し 3 名のアノテータが独立に判定を行い、多数決でラベルを決定した。この事例を学習用、開発用、評価用データとして 35,000 事例、5,000 事例、10,000 事例に分割し、上記の関係判別と同様に $\text{BERT}_{\text{large}}$ をファインチューニングして実験を行った。この結果、等価、含意（2 種類）、矛盾の判定性能は評価データの平均精度でそれぞれ 88.0%、87.6%、88.8%、55.4%という結果を得ている。

また、トラブル認識のためにも学習・評価用データを作成した。この作成時には、負担・トラブル表現リスト[4)] [14] と活性・不活性辞書 [15] を組み合わせ、「結核を患う」のようなトラブル名詞と活性述語テンプレートの組み合わせ、また、「出生率が低下する」のような非トラブル名詞と不活性述語テンプレートの組み合わせをトラブル認識の対象となる核表現とし、その核表現を含む文をその前後文とともに Web 文書から 4 万件サンプリングした。3 名のアノテータが文中の核表現がトラブルを表す内容として記述されているかを判定し、多数決でラベルを決定した。ただし、3 節で述べた提案手法では帰結文全体を入力してトラブルを含むか否かを判定する必要があるため、核表現に対するアノテーション結果をその核表現が含まれる文に対するアノテーション結果とみなして学習・評価に利用した[5)]。上述の 4 万事例を学習用、開発用、評価用としてそれぞれ 28,000 事例、4,000 事例、8,000 事例に分割し、上記の $\text{BERT}_{\text{large}}$ でファインチューニングを行った。この結果得られた分類器の性能は、評価用の平均精度で約 91.0% である。

5　実験

対話システムにユーザが話しかけるとき「マイナンバーで役所の手続きが簡単になるといいね」「小学校でプログラミングが必修になるんだって」のようにユーザの主張が明示されないか、間接的な言い方で示されることが想定される。ここでは問題を単純化するために、「マイナンバーを導入する」「小学校でプログラミングを必修にする」のように、ユー

3）　上述の因果関係の判定アノテーションでは複数の因果関係のタイプに関して判定を行うため 1 事例に対して複数の正例が得られる場合もあるが、生成のアノテーションではその個別の正例を 1 事例としてアノテーションを行っているため、上記の判定の事例数とはアノテーションの件数が異なっている。

4）　https://alaginrc.nict.go.jp/li-outline.html#A-3

5）　アノテーション時には前後文を含む 3 文を見て判断を行ったが、学習・評価時には問題の核表現を含む 1 文のみを入力して利用した。

表 2　評価用主張文 500 件に対する反論候補の評価結果

反論が出力できた事例	P@1	P@3
452	226 (50%)	357 (79.0%)

ザ発話中の出来事を取り出し、賛成と反対の両方の立場が仮定できるシンプルな主張に書き換え、ユーザが主張に「賛成」「そう願っている」と仮定する。以下では書き換え後の文をユーザの主張として扱い、提案手法が出力した反論候補が反論として成立するかどうか評価する。

より具体的には、雑談のための入力発話として、任意のトピックに関してアノテータが作成したユーザ入力発話の中から、賛成と反対の両方の立場が仮定できるシンプルな主張に書き換え可能なものを著者の一人が選び、上記のようにシンプルな主張への書き換えを行い、ユーザの主張文とした。ユーザの主張は、開発用、評価用としてそれぞれ 50 事例、500 事例を用意した。提案手法の適用に際しては、含意等認識を適用する因果関係の候補として、Lucene の検索結果上位 5,000 件を利用した。また、反論候補が複数得られた場合にはトラブル認識の分類スコアが最も高い反論候補を最大 3 つ出力することとした。評価としては、主張文と反論候補の組を提示し、主張文の内容に対して「賛成」「そう願う」「よい選択だ」のような立場をある人がとったと仮定した場合に、反論候補がその人に対する反論となるかどうかを、3 名の作業者がそれぞれ独立に判定し、多数決でラベルを決定する。

5.1　実験結果

入力となる主張文、開発用 50 事例、評価用 500 事例のうち、反論が出力できたのはそれぞれ 41 事例、452 事例であった。評価様主張文の人手評価結果を表 2 に示す。トラブル認識の分類スコアスコア最上位の反論候補が反論になると判定された場合 (P@1) が 50.0%、トラブル認識の分類スコア 3 位までの反論候補に、反論になると判定されたものが含まれる割合 (P@3) が 79.0%となった。Fleiss の κ 値は 0.605 であった。特に分類スコア最上位の場合は高精度とは言えないが、特に反論生成に特化した学習データを作成していないことを考えると、まずまずの結果であると考えている。

誤りの原因を探るため、評価用データに対して出力された反論について、トラブル認識の分類が最大の反論候補のうち、反論ではないと判定された事例の中から 100 件サンプリングし、反論が成り立たない要因を分析した。その要因の分類と事例数を付録表 3 に示す。主な要因は因果関係が迂遠で分かりづらいことであった。例えば、主張「ホームドアを設置する」に対する反論「事故が起きる可能性がある」では、「ホームドアが設置されている駅とされていない駅が混在することで視覚障害者が混乱してしまい事故につながる」といった説明があれば反論として成立すると考えられる。また、含意等判定器、因果関係判定器、生成器にも改善の余地があることがわかる。今後は、反論に特化した学習データを作成するとともに、こうした課題に対処していく予定である。

表 3　反論として成立しなかった要因（合計 100 件）

要因	件数
因果関係が迂遠で普通の人には分かりづらい	36
原因文と帰結文の間に因果関係がない・帰結文の生成に失敗している	25
入力文と原因文の間の含意等認識に失敗している	17
帰結文の問題検出に失敗している	8
その他	14

6　おわりに

本論文では、ユーザのブレインストーミングを助けより良い意思決定を支援する手法の開発を目指し、因果関係知識を大規模に獲得し、ユーザに対して帰結に問題を含む因果関係を用いて、反論として出力する手法を提案した。提案手法による反論は情報源が Web であるため、信頼性に難があったり意見が偏っていたりするものもあり得るが、今後は、DISAANA [16] 等のデマ検出技術で培ったテキスト間の矛盾検知技術 [17] 等も活用し、よりバランスの取れた反論や、因果関係 [12] の連鎖による推論等も用いてよりクリエイティブな反論や、ユーザや文脈により適合した反論を提供することを目指す。加えて、深層学習自動並列化ミドルウェア RaNNC[18] を用いて事前学習したより巨大な言語モデルや、そもそもアーキテクチャの異なる言語モデル、例えば [19] も用いてさらなる高精度化を図る。また、反論だけではなく、支持のような議論の他の構成要素も同様の手法で出力が可能だと考えており、これらも用いて多様な意思決定に関するブレインストーミングを可能にする対話システムの実現を目指す。

参考文献

[1] Ryuichiro Higashinaka, Kazuki Sakai, Hiroaki Sugiyama, Hiromi Narimatsu, Tsunehiro Arimoto, Kiyoaki Matsui, Takaaki Fukutomi, Yusuke Ijima, Hiroaki Ito, Shoko Araki, Yuichiro Yoshikawa, Hiroshi Ishiguro, and Yoshihiro Matsuo. Argumentative dialogue system based on argumentation structures. In **Proceedings of the 21st Workshop on the Semantics and Pragmatics of Dialogue**, 2017.

[2] Dieu Thu Le, Cam-Tu Nguyen, and Kim Anh Nguyen. Dave the debater: a retrieval-based and generative argumentative dialogue agent. In **Proceedings of the 5th Workshop on Argument Mining**, pp. 121–130, 2018.

[3] Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, Ron Hoory, Yufang Hou, Shay Hummel, Michal Jacovi, Charles Jochim, Yoav Kantor, Yoav Katz, David Konopnicki, Zvi Kons, Lili Kotlerman, Dalia Krieger, Dan Lahav, Tamar Lavee, Ran Levy, Naftali Liberman, Yosi Mass, Amir Menczel, Shachar Mirkin, Guy Moshkowich, Shila Ofek-Koifman, Matan Orbach, Ella Rabinovich, Ruty Rinott, Slava Shechtman, Dafna Sheinwald, Eyal Shnarch, Ilya Shnayderman, Aya Soffer, Artem Spector, Benjamin Sznajder, Assaf Toledo, Orith Toledo-Ronen, Elad Venezian, and Ranit Aharonov. An autonomous debating system. **Nature**, Vol. 591, No. 7850, pp. 379–384, 2021.

[4] Douglas Walton, Chris Reed, and Fabrizio Macagno. **Argumentation Schemes**. Cambridge University Press, 2008.

[5] 水野淳太, クロエツェージュリアン, 田仲正弘, 飯田龍, 呉鍾勲, 石田諒, 淺尾仁彦, 福原裕一, 藤原一毅, 大西可奈子, 阿部憲幸, 大竹清敬, 鳥澤健太郎. WEKDA: Web 40 億ページを知識源とする質問応答システムを用いた博学対話システム. 人工知能学会第 84 回言語・音声理解と対話処理研究会, pp. 135–142, 2018.

[6] Yoshihiko Asao, Julien Kloetzer, Junta Mizuno, Dai Saiki, Kazuma Kadowaki, and Kentaro Torisawa. Understanding user utterances in a dialog system for caregiving. In **Proceedings of the 12th Language Resources and Evaluation Conference**, pp. 653–661, 2020.

[7] 淺尾仁彦, 水野淳太, 呉鍾勲, Julien Kloetzer, 大竹清敬, 福原裕一, 鎌倉まな, 緒形桂, 鳥澤健太郎. 介護支援対話システムのための意味解釈モジュール. 言語処理学会 発表論文集, 2022.

[8] Kazuki Sakai, Akari Inago, Ryuichiro Higashinaka, Yuichiro Yoshikawa, Hiroshi Ishiguro, and Junji Tomita. Creating large-scale argumentation structures for dialogue systems. In **Proceedings of the 11th International Conference on Language Resources and Evaluation**, 2018.

[9] Kazuki Sakai, Ryuichiro Higashinaka, Yuichiro Yoshikawa, Hiroshi Ishiguro, and Junji Tomita. Hierarchical argumentation structure for persuasive argumentative dialogue generation. **IEICE Transactions on Information and Systems**, Vol. E103.D, No. 2, pp. 424–434, 2020.

[10] Xinyu Hua, Zhe Hu, and Lu Wang. Argument generation with retrieval, planning, and realization. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 2661–2672, 2019.

[11] Khalid Al Khatib, Lukas Trautner, Henning Wachsmuth, Yufang Hou, and Benno Stein. Employing argumentation knowledge graphs for neural argument generation. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 4744–4754, 2021.

[12] Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics**, pp. 987–997, 2014.

[13] Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. Why-question answering using intra- and inter-sentential causal relations. In **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics**, pp. 1733–1743, 2013.

[14] Stijn De Saeger, Kentaro Torisawa, and Jun'ichi Kazama. Looking for trouble. In **Proceedings of the 22nd International Conference on Computational Linguistics**, pp. 185–192, 2008.

[15] Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong-Hoon Oh, and Jun'ichi Kazama. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In **Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning**, pp. 619–630, 2012.

[16] Junta Mizuno, Masahiro Tanaka, Kiyonori Ohtake, Jong-Hoon Oh, Julien Kloetzer, Chikara Hashimoto, and Kentaro Torisawa. WISDOM X, DISAANA and D-SUMM: Large-scale NLP systems for analyzing textual big data. In **Proceedings of COLING 2016: System Demonstrations**, pp. 263–267, 2016.

[17] Julien Kloetzer, Stijn De Saeger, Kentaro Torisawa, Chikara Hashimoto, Jong-Hoon Oh, Motoki Sano, and Kiyonori Ohtake. Two-stage method for large-scale acquisition of contradiction pattern pairs using entailment. In **Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing**, pp. 693–703, 2013.

[18] 田仲正弘, 田浦健次朗, 塙敏博, 鳥澤健太郎. 自動並列化深層学習ミドルウェア RaNNC. 自然言語処理, Vol. 28, No. 4, pp. 1299–1306, 2021.

[19] Jong-Hoon Oh, Ryu Iida, Julien Kloetzer, and Kentaro Torisawa. BERTAC: Enhancing transformer-based language models with adversarially pretrained convolutional neural networks. In **Proceedings of the 59th ACL and the 11th IJCNLP**, pp. 2103–2115, 2021.