The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

ディベート型人工知能によるサポート性推定に基づく反論生成

Debating AI: Counterargument Generation Based on Textual Supportiveness Recognition

佐藤 美沙 Misa Sato　柳井 孝介 Kohsuke Yanai　柳瀬 利彦 Toshihiko Yanase　是枝 祐太 Yuta Koreeda　黒土 健三 Kenzo Kurotsuchi

日立製作所 研究開発グループ
Research & Development Group, Hitachi, Ltd.

We will demonstrate a counterargument generation system in debating, which aims to utilize newswire articles for decision-making support. Users can specify a claim such as “We should legalize casinos because they promote economy.” and then the system outputs counterargument scripts against the claim.

1. はじめに

筆者らの研究グループでは，電子化されたテキストデータを意思決定支援に活用することを目的として，ディベートをするシステムを開発している [Sato 15]. このシステムを本稿ではディベート型人工知能と呼ぶ. 本発表ではディベート型人工知能による反論生成のデモンストレーション展示を行う. 聴講者が主張をテキスト文で与えると，このシステムは，反論する主張およびその理由や根拠を提示する. たとえば「カジノは経済を活性化させる」という主張が与えられると，反論として「カジノは経済を活性化させない」という反対側の主張を提示し，続いてその理由や根拠となる事例を新聞記事等から抽出して提示する.

反論においては，単に議題に対する肯定・否定の立場が相手と反対であれば良いわけではなく，相手の主張と噛み合った議論を行う必要がある. 本研究では，入力される主張と同じ観点に基づいて反対側の立場から論じることにより，話の噛み合った反論を行う.

本研究では，入力された主張の依拠する価値を認識して，それを踏まえた反論側の主張を作成し，続けて反論側の主張を支持する根拠を提示するシステムを作成する. 指定の価値についての反論側根拠を集めることを可能とするために，データソースから抽出したテキストが，ある価値に依拠する主張に対する根拠であるか否かを推定するサポート性推定 [佐藤 16b]の問題を組み込む. よって本稿では特にサポート性推定に着目し，従来のタスクで有用であった特徴量の有用性を調べることによりサポート性推定の特徴を調べる.

2. 関連研究

議論解析は近年 Argument Mining 分野として自然言語処理の観点から多くの研究がなされている. 反論に関するものとしては，議論中に出現する 2 つの文の関係が支持（support）であるか反論（attack）であるかを推定する Relation-based Argument Mining というタスクが提案されている[Carsteins15,Cocarascu 17]. また，[Bosc 16]ではツイート間が支持関係にあるか反論関係にあるかを識別している. これらは議論内の構造を明らかにするものであり，与えられた主張に対して反論を提示するものではない. また，どの観点で反論しているかは取り扱っていない.

反論の生成に関する研究としては，Bilu らは，与えられた主張文を否定する主張文を生成する手法を提案している[Bilu 15].

連絡先：佐藤美沙，株式会社 日立製作所 中央研究所，
〒185-8601 東京都国分寺市東恋ヶ窪一丁目 280 番地，
042-323-1111，misa.sato.mw@hitachi.com

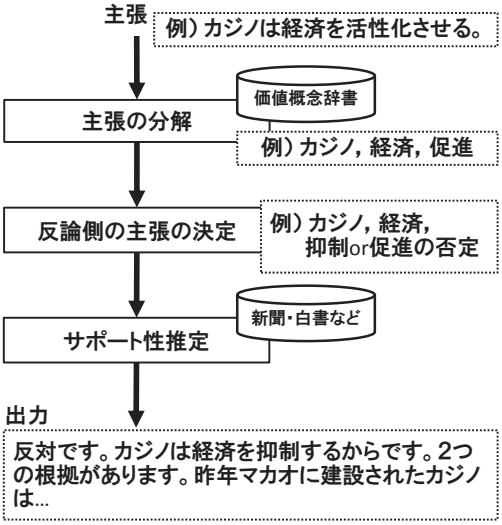

図 1. 反論生成システムの全体フロー

主張文に否定語“not”を加える，もしくは，感情語の極性を反転させるなどのルールベースの手法によって否定側の主張文を生成し，生成された否定側の主張文がディベートで使われた際にもっともらしいか否かを統計的な手法によって検証している. この研究では根拠の提示は行っていない.

本研究では，与えられた主張が問題としている議題および依拠する価値を推定し，議題から価値への影響が主張とは反対方向であることを示す根拠を探すことにより，反論を生成する. これにより，与えられた主張の観点を踏まえた反論を生成することができると考えられる.

3. ディベート型人工知能による反論生成

本研究では，[Sato 15]に説明しているディベート型人工知能の機能を一部組み替えて反論生成を行う. 入力される主張には，「カジノは経済を活性化させる」のような議論の対象と効果を示す文を想定している. 効果を論じる際に，価値というものを取り扱う. ここで言う価値とは，それ自体が人あるいは社会において良い存在（悪い存在）と捉えられるものであり，たとえば健康は良い価値，貧困は悪い価値，と考えることができる. つまり，ディベート型人工知能における肯定的な主張とは，議論の対象が良い価値を増やしていることへの指摘，あるいは，悪い価値を減らしていることの指摘であり，否定的な主張とは，議論の対象が悪い価値を増やしていることへの指摘あるいは良い価値を減らしていることの指摘である.

ディベート型人工知能における反論生成とは，与えられた主張に対して，それが肯定的な主張であれば否定的な主張およびその根拠を示し，否定的な主張であれば肯定的な主張およびその根拠を示すことである．図1に示す通り，全体フローは大きく以下の3ステップから成る．

1. 主張の分解
2. 反論側の主張の決定
3. 反論側の根拠の抽出

本章では各ステップの詳細を説明する．

3.1 主張の分解

主張の分解では，入力された主張文から議論の対象と価値を抜き出す．抜き出した後，議論の対象が価値に及ぼす影響の方向を推定する．影響の方向には，促進関係，抑制関係，促進関係の否定，抑制関係の否定，影響関係なし，の 5 通りを想定する．たとえば，「カジノは経済を活性化させる」という主張が入力された場合は，議論の対象が「カジノ」，価値が「経済」，影響の方向は「促進」となる．

議論の対象の抽出および価値の同定は，構文構造に対するルールベースの手法によって行う．優先度の付いた 5 つのルールを順番に適用し，合致した場合にそれを抽出する．最も優先度の高いルールは，影響関係に関する述語を同定し，その述語の主語を議論の対象，目的語を価値とするルールである [佐藤 16a]．影響関係に関する述語については約 350 語の辞書を用意している．

価値は，価値体系辞書によって分類を定義する [柳井 15]．価値体系辞書は 15 の大価値とそれにぶら下がる合計 69 の小価値からなる．それぞれの小価値には表現語が定義されている．表 1 に価値体系辞書の一部を示す．主張から価値を抜き出す際には，抽出した目的語がどの小価値にあたるかを表現語との単語列類似度により分類する．

次に影響の方向を推定する．ここでも構文構造に対するルールベースの手法を用いる．促進関係を 1，抑制関係を-1 とし，述語の極性と，述語に対する否定の係り受けの有無を掛け合わせることにより，方向を計算する．

表 2. データセットの内訳

分類	数
計	7955
影響関係なし	3784(48%)
影響関係あり	4171(52%)
促進関係	2861(68%)
抑制関係の否定	219(5%)
抑制関係	701(17%)
抑制関係の否定	390(9%)

3.2 反論側の主張の決定

反論側の主張においても，議論の対象が価値に影響を及ぼすという考え方を用いる．反論の依拠する価値を決定し，主張を決定する．入力された主張の価値を元に反論側の価値を決定するが，本研究では，元の主張の価値をそのまま反論側でも採用する．そして，影響の方向を反転させる．つまり，「カジノが経済を活性化させる」という主張に対しては，「カジノが経済を抑制する」という反論側の主張を行う．

3.3 反論側の根拠の抽出

前節で採用した反論側の主張を支持するような事例を新聞記事などのテキストから収集し，反論側の根拠として提示する．

サポート性推定では，入力されたテキストについて，議論の対象が価値に対して与える影響が，促進関係か，抑制関係か，あるいは影響関係が無いかを推定する． 根拠候補文のサポートする主張の影響関係の方向が反論側の主張と同一である場合，反論側の根拠であるとして抽出する．ここで反論側の根拠として抽出された文が最終的な反論文として表示されるため，サポート性推定の精度が高いほど，見逃しが少なく，誤りの少ない，良い反論生成システムになる．

最後に，信頼度の高い 2 文を選択し，反論側の根拠として提示する．

4. サポート性推定

4.1 手法

サポート性推定は，データソースから抽出したテキストが，主張を支持（否定）する根拠としてどの程度もっともらしいかを推定するタスクである [佐藤 16b]．サポート性推定の特徴を明らかにするため，関連タスクである根拠検知と賛否識別で使われている特徴量によってサポート性の推定がどの程度可能かを調べる．加えて，サポート性推定特有の特徴量の追加による推定精度への影響を調べる．

まず，根拠検知（Evidence Detection; ED）で有用な特徴量として，[Rinott 15]で使われている以下の 5 つを利用する．

・意味的類似度： 主張文と根拠候補文の間の類似度．Word2Vec のコサイン類似度[Mikolov 13]により計算する．

・センチメントの一致度： 根拠候補文のセンチメント値を[0,1]に正規化した値．センチメント値の導出には Stanford Core NLP [Manning 14] を用いた．

・語彙： 根拠文によく出現する語を集めた辞書を用意し，その辞書内の語がそれぞれ出現しているか否かを表す 1 か 0 の値．辞書は手作業で集めた 287 語からなる．専門家や公的機関の発言は説得力の元となるため，これらを表す語は根拠文に出現しやすいという考えの下に作成されている．

・固有表現クラス： 根拠候補文がそれぞれの固有表現クラスにあたる固有表現を含むか否かを表す 1 か 0 の値．固有表現およびクラスは Stanford Core NLP により抽出した．

表 1. 価値体系辞書の一部

大価値	小価値	極性	表現語
economy	economy	+1	economy, investment, development, …
〃	employment	+1	employment, job, …
〃	loan	-1	loan, debt, …
health	health	+1	health, healthy, …
〃	disease	-1	disease, complication, …
safety	crime	-1	crime, prostitution, …

・文パターン: 文が引用を含むか，人あるいは組織に寄る発言を含むか，人あるいは組織による行動を含むか，人あるいは組織の意見を含むか，を表すパターンを作成し，根拠候補文がそれぞれに当てはまるか否かを表す 1 か 0 の値.

次に，賛否識別（Stance Classification; SC）で有用性が知られている [Mandya 16]，以下の 1 つの特徴量を根拠候補文に対して適用する.

・依存関係: 根拠候補文において依存関係にある 2 語のペアを表す one-hot ベクトル.

最後に，サポート性推定（Supportiveness Recognition; SR）特有の特徴量として以下の 2 つを用いる. 議論対象と価値の関係に着目する問題であるため，候補文内のどの語句が議論対象および価値と対応するかを考慮した特徴量としている.

・機能語: 根拠候補文中の議論対象を表す語と価値を表す語の間に出現する動詞，名詞，副詞あるいは形容詞に対する one-hot ベクトル.

・否定語: 根拠候補文中の議論対象を表す語と価値を表す語の間に否定語が含まれるか否かを表す 1 か 0 の値. 否定語の辞書は，一般的な”not”などの否定語に加え，”ban”，“barrier”，“drop”，”hazard”などの極性を反転させるような語を含む 154 語からなる.

4.2 実験設定

データセットは，Annotated English Gigaword [Napoles 12] から候補となる文を取得し，人手でラベリングすることにより作成した. 候補文の取得は，31 通りの主張（議論の対象と価値のペア）について，議論の対象と価値の表現語の共起する文を抽出することにより行った. 表 2 にデータセットの統計値を示す.

反論側の根拠の候補には，(a)反論側主張と影響関係の方向が合致するもの，(b)影響関係の方向が反対であるもの，(c)影響関係がないもの，の 3 通りが存在する. 表 2 のラベルとの対応を考えると，反論側主張の影響関係の方向が促進関係であるとすると，(a)は促進関係あるいは抑制関係の否定，(b)は抑制関係あるいは促進関係の否定，(c)は影響関係なしが対応する. そこで，影響関係の有無を識別する（(a)と(b)あるいは(c)を識別する）2 値問題と，促進関係と抑制関係を識別する（(a)と(b)を識別する）2 値問題のそれぞれについて実験を行った.

前節で示した特徴量を用い，線形 SVM 識別器によって識別を行った. 3-fold の交差検定により精度を測定している. ただし，同じ主張に基づく根拠文が訓練データとテストデータの両方に含まれることはないように分割している. パラメータは，$C = \{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$とした. また，ベースラインの特徴量として Bag-of-Words（BoW）を用いた.

4.3 実験結果

表 3 に影響関係の有無の識別結果を示す. BoW，根拠同定および根拠分類の特徴量を用いた[BoW+ED+SR]が最も F 値が高く 57%となった. しかしながらいずれの特徴量の組み合わせも，BoW ベースライン特徴量の結果からさほどの向上は見られなかった. 類似のタスクの特徴量ならびに本稿で提案した特徴量は有用とはいえない.

表 4 に促進関係と抑制関係の識別結果を示す. [BoW]の結果に比べて[SC]は F 値が 5%高く，賛否分類の特徴量が有用であったことがみてとれる. 加えて，BoW 特徴量はうまく働かなかった. 賛否分類タスクにおいても BoW 特徴量はあまり機能しない [Mandya 16] ことから，この点も賛否分類との共通点といえる. しかしながら，賛否分類の特徴量にサポート性推定特有の特徴量を加えた[SC+SR]が最も良い精度となった. 議論の対象と価値に着目した特徴量によって精度の向上が見られた.

表 3. 影響関係の有無の識別実験結果

特徴量	適合率	再現率	F 値
[ED]	55%	55%	54%
[SC]	57%	56%	56%
[SC+SR]	56%	56%	56%
[BoW]	57%	56%	56%
[BoW+ED]	56%	55%	55%
[BoW+ED+SR]	**58%**	**57%**	**57%**

表 4. 促進関係と抑制関係の識別実験結果

特徴量	適合率	再現率	F 値
[ED]	58%	55%	56%
[SC]	62%	69%	64%
[SC+SR]	**68%**	**71%**	**69%**
[BoW]	61%	57%	59%
[BoW+ED]	62%	59%	60%
[BoW+ED+SR]	64%	63%	63%

5. おわりに

本稿では，反論生成システムの全体フローについて述べ，サポート性推定について実験を行った. サポート性推定というタスクを中心に据えることで，主張の依拠する価値を踏まえた反論を生成するシステムを開発した. 一方でサポート性推定については精度が未だ十分ではなく，特に影響関係の有無の識別については今回の実験では有効な特徴量を見つけることができなかった. エラー分析などにより問題の分析を進めていく.

今後は，精度向上と共に，立論システムと反論システムを統合して競技ディベートに出場可能なシステムとするため，立論から反駁まで一貫性のある議論を展開できるような手法の開発を進めていく.

参考文献

[Bilu 15] Yonatan Bilu, Daniel Hershcovich, and Noam Slonim: Automatic claim negation: why, how and when, *Proceedings of the 2nd Workshop on Argumentation Mining*, pp.84-93, Association for Computational Linguistics, 2015.

[Bosc 15] Tom Bosc, Elena Cabrio, and Serena Villata: Tweeties Squabbling: Positive and Negative Results in Applying Argument Mining on Social Media, *Computational Models of Argument (COMMA)*, pp.21-32, IOS Press, 2016.

[Carsteins 15] Lucas Carstens and Francesca Toni: Towards relation based argumentation mining, *Proceedings of the 2nd Workshop on Argumentation Mining*, pp.29-34, Association for Computational Linguistics, 2015.

[Cocarascu 17] Oana Cocarascu and Francesca Toni: Identifying attack and support argumentative relations using deep learning, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp.1374-1379, Association for Computational Linguistics, 2017.

[Mandya 16] Angrosh Mandya, Advaith Siddharthan, and Adam Wyner: Scrutable Feature Sets for Stance Classification, *Proceedings of the 3rd Workshop on Argumentation Mining*, pp.60-69, Association for Computational Linguistics, 2016.

[Manning 14] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard and David McClosky: The Stanford CoreNLP natural language processing toolkit, *Proceedings of the ACL 2014: System Demonstrations*, pp.55-60, Association for Computational Linguistics, 2014.

[Mikolov 13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado and Jeff Dean: Representations of Words and Phrases and their Compositionality, *Proceedings of the Third Workshop on Argument Mining*, pp.3111-3119, Curran Associates, Inc., 2013.

[Napoles 12] Napoles, Courtney, Matthew Gormley, and Benjamin Van Durme : Annotated English Gigaword LDC2012T21, *Philadelphia: Linguistic Data Consortium*, 2012.

[Rinott 15] Ruty Rinott, Lena Dankin, Carlos Alzate Perez, M. Mitesh Khapra, Ehud Aharoni and Noam Slonim: Show me your evidence - an automatic method for context dependent evidence detection, *Proceedings of the EMNLP 2015*, pp.440-450, Association for Computational Linguistics, 2015.

[Sato 15] Misa Sato, Kohsuke Yanai, Toshinori Miyoshi, Toshihiko Yanase, Makoto Iwayama, Qinghua Sun and Yoshiki Niwa: End-to-end Argument Generation System in Debating, *Proceedings of the ACL-IJCNLP 2015 System Demonstrations*, pp.109-114, Association for Computational Linguistics, 2015.

[佐藤 16a] 佐藤 美沙, 柳井 孝介, 柳瀬 利彦, 是枝 祐太, 丹羽 芳樹: ディベート人工知能における影響関係認識のためのテキスト内の論理構造に関する考察, 人工知能学会全国大会論文集, 第 30 巻, 4B1-3, pp.1-3, 人工知能学会, 2016.

[佐藤 16b] 佐藤 美沙, 柳井 孝介, 柳瀬 利彦, 三好 利昇, 是枝 祐太, 丹羽 芳樹: 意見文章自動生成のための組合せ構文特徴を用いたサポート性推定, 人工知能学会論文誌, 第 31 巻, no. 6, pp. AI30-L_1-12, 人工知能学会, 2016.

[柳井 15] 柳井 孝介, 三好 利昇, 柳瀬 利彦, 佐藤 美沙, 丹羽 芳樹, Reisert, P., 乾 健太郎:ディベート人工知能における意見生成,人工知能学会全国大会論文集, 第 29 巻, 3M3-2in, pp. 1-3, 人工知能学会, 2015.