ARG WI2 No.12, 2018
WI2-2018-05

Argument Mining のための Web市民議論データのアノテーション

森尾 学[a]　　藤田 桂英[b]

東京農工大学大学院 工学府 情報工学専攻

a) *morio@katfuji.lab.tuat.ac.jp* b) *katfuji@cc.tuat.ac.jp*

概要　本論文では，Argument Mining における Argument Component や Relation を用いた自動構造化の試みをスレッド構造に適用する手法を提案する．本論文の貢献は以下の通りである．(1) スレッド構造を持つオンライン市民議論データに対して投稿内と投稿間の関係に着目したスキームを適用した．(2) そのスキームを用いてアノテーションを行い，アノテータ間一致度を評価した．最終的に，Argument Mining の分野において最大規模のコーパスを作成する事ができた．(3) SVM や End-to-End なニューラルモデル等を用いて，文のタグ付け，投稿内関係および投稿間インタラクションを分類する識別モデルを提案した．識別モデルの評価実験の結果，我々のアノテーションデータを用いて，スレッド構造に対して比較的安定した精度で End-to-End に識別することが可能であることが分かった．

キーワード　Argument Mining，市民議論，アノテーション，SVM，ニューラルネットワーク

1　はじめに

オンラインでの市民議論により，膨大な人数の市民が場所や時間に問わず，意見を投稿し，議論を通じて結論（合意）を導くようなことが可能になっている．しかし，膨大な数の投稿の整理や議論の流れの理解，結論への導出（合意）を支援するために，意見の理解につとめる Opinion Mining [1] や，Claim と Premise，Evidence [2] の関係の理解につとめる Argument Mining [4, 3] の適用が求められている．

Argument Mining (AM) とは，議論的言説から Claim（主張）や Premise（前提）となる部分を抽出し，それらの関係を推論して構造化する研究領域である．AM に関する多くの既存研究はアーティクルやエッセイなどのフォーマルな文書を主な対象としている．例えば，草分け的な AM のデータセットには Persuasive Essays [7, 8] がある．彼らのデータでは，学生のエッセイ文書に対して Claim や Premise となる部分を綿密にアノテーションしたため，多くの研究で用いられている [9, 10, 11]．しかし，Persuasive Essays のようなコーパスは単一の議論的言説にフォーカスしており，スレッド構造のような議論的言説間の関係は考慮されない（エッセイは単体で完結しているため）．そこで我々は，スレッド構造を持つオンラインフォーラムに対して，AM を適用させる方法論を提案する．

本研究の主要なポイントは次の 3 つである．(1) AM をスレッド構造に適用させるために新たなスキームを提案する．具体的には文章のタイプ (Premise / Claim / Non-Argumentative(NA)) と，投稿内および投稿間の文章の関係を定義した．(2) スレッド構造を持つ市民議論データに対して，提案スキームを用いたアノテーションを行った．具体的には，過去に行われた大規模なオンライン市民議論での議論データに対して 2 段階のアノテーションを行い，アノテータ間一致度を評価した．最後に，(3) SVM や End-to-End なニューラルモデル等による推論手法を提案した．具体的にはスレッドに含まれる文章のタイプ，投稿内と投稿間それぞれで文章同士の関係を識別するモデルを実現した．

2　関連研究

AM はコーパス駆動の研究領域であるため，データは特に重要である．ニュース記事や Web 議論などを含み，議論関係がアノテーションされた草分け的なコーパスである AraucariaDB [13] や，Web 上のフォーラムに焦点を当てたアノテーションが存在する [14]．他には，Reddit の 2016 US Presidential Debates に対するアノテーション [12] や，*ChangeMyView*[1] の解析 [40, 18] がある．とは言え，これらのフォーラムやエッセイにおけるデータは極めて議論的かつ明確な構造を扱うケースが多い．すなわち，我々が目的とするような，多人数によるより砕けた市民議論とは性質が異なっている．また，スレッド構造に対する AM スキームを用いたアノテーション済みデータは我々の知る限りではほとんど存在しない．

AM においては議論構造を自動的に識別する研究が盛んである．例えば，投稿に対する議論要素の特定（ACI）[27, 26] やリンク抽出（LE）[25, 17] の AM タスクの研究が存在する．

[1] https://reddit.com/r/changemyview

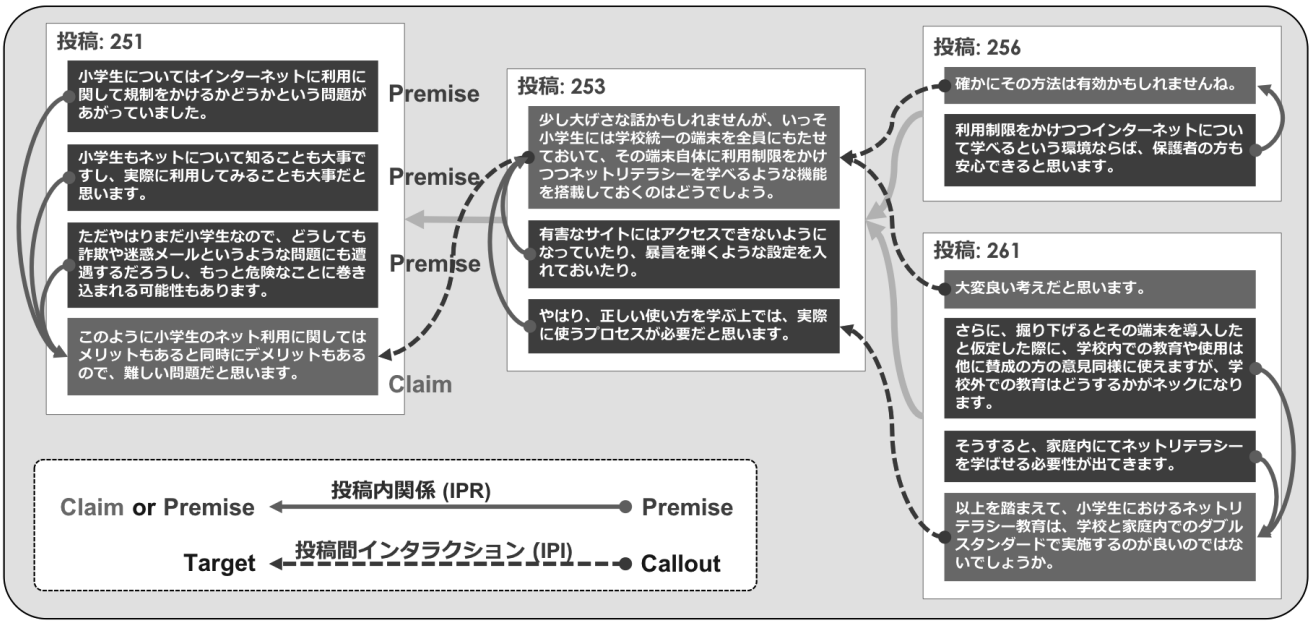

図 1　スレッドのアノテーション例．白色のボックスは投稿を表す．投稿 251 に対して 253 が返信しており，同様に 253 に対して 256 と 261 が返信している．青色のボックスが Premise，赤色のボックスが Claim を表す．投稿内のリンクは投稿内関係（IPR）を示している．投稿間の赤色のリンクは投稿間インタラクション（IPI）を示している．

近年では End-to-End (E2E) な識別モデルが目下進行中の研究対象である．例えば [17] は LSTM を用いて AM に対するマルチタスク学習 [23, 24] を行う手法や，LSTM-ER [39] を適用する手法を提案した．

3　スレッド構造のためのスキーム

本節では，どのようなアノテーションスキームを導入するかについて説明する．AM に対する堅実なスキームは数多く存在している．そのため既存のスキームを組み合わせ，それを発展させることがスレッド構造に AM スキームを適用させるための鍵であると言える．我々は，一つの投稿をそれ単体で談話であると見なした．さらに投稿と投稿の間のインタラクションに着目した．そこで「投稿内」と「投稿間」の 2 つのスキームモデルを組み合わせることにした．投稿内における AM（Claim/Premise/NA のタグ付けやリンクの制約付け）は [22] のスキームを導入した．投稿間を結びつけるインタラクションに関しては [14] の Target/Callout のスキームモデルを導入することにした．図 1 は我々の組み合わせスキームを用いたアノテーション例である．

3.1　データ

スレッド構造に対する AM のコーパスを作成するために，独自に収集したオンライン市民議論データに対してアノテーションを行うことにした．今回ターゲットとしたデータは，スレッド構造を持つ *COLLAGREE* [6, 5, 16][2]と呼ばれるオンラインフォーラムで実施された，市民議論データである[3]．このデータは名古屋市と共同で 2016 年 12 月から 2017 年 1 月にかけて行われた Web 議論をベースとしている．399 のスレッドと 1327 件の投稿が蓄積されている[4]．

3.2　アノテーションの過程

[28] によれば，AM におけるタスクは次の 3 つのプロセスに分けられる: (1) Segmentation，(2) Segment classification，(3) Relationship identification．初めに，Segmentation については，Argument Component Idenfication (ACI) や Argument Component Detection (ACD) と呼ばれ，このタスク単体においても重要な研究分野である [41, 29]．しかし，ACI はトークンレベルでのアノテーションを行わなければならず，時間や金銭的コストの面で負担が多い．そこで我々はルールによって自動で Segmentation を行うことにした．具体的には 1 つの文章を 1 つの Component として見なし [19]，句読点をベースに Argument sentence (AS) に分割する [15]．次に，(2) の Segment classification については，(1) で分割した AS を Claim，Premise および NA にアノテーションする．最後に，(3) の Relationship identification については投稿内の AS 同士の関係をアノテーションする．また，投稿間の AS 同士のインタラクション (Target/Callout) のアノテーションも行う．アノテーション

[2] http://collagree.com/

[3] 市民の意見や個人情報を含むため，データセットは現在非公開．

[4] 1 スレッドあたりの平均投稿数 3.33（標準偏差 3.29），スレッドの平均階層深さ 1.09（標準偏差 1.19），1 投稿あたりの平均文章数 4.19（標準偏差 3.33，総文章数 5559），1 文章あたりの平均単語数 21.63（標準偏差 19.92）．

対象となる AS タイプおよび関係の定義は次のとおりである．

- **Claim** とは議論の的となる言説文である．投稿内には関係先を持たない．
- **Premise** は Claim や他の Premise に対して理由を与えたり支持や攻撃する言説文のことである．
- **投稿内関係 (IPR)** は投稿の中での Claim や Premise の関係である．関係を $(parent \leftarrow child)$ と表す時，$child$ は $parent$ に対して理由を与える．そのため，あり得るパターンは $(Claim \leftarrow Premise)$ もしくは $(Premise \leftarrow Premise)$ の 2 通りしか存在し得ない．注意点として，Premise のみでループを作ることは禁止である．すなわち Claim をルートノードとする木構造とならなければならない．
- **Target** は投稿間の AS と AS のインタラクションにおいて，後に呼び出される対象となる言説文である[5]．
- **Callout** は投稿間の AS と AS のインタラクションにおいて，*Target* に対してコメントを与える言説文である．データの都合上，明示的な返信関係にある 2 つの投稿間のみを考慮する．本研究では議論に焦点を当てるため，Callout は Claim に限定する[6]．
- **投稿間インタラクション (IPI)** は 2 つの返信関係にある投稿同士の関係である．関係を $(parent \leftarrow child)$ と表す時，$child$ は Callout，$parent$ は Target である．

複数のサブタスクが存在する場合に，複数段階に分けてアノテーションを行うことが一般的である [20, 22]．我々は次の 2 段階に分けてアノテーションを行った．初めに，AS タイプ (Claim/Premise/NA) のアノテーションと投稿内関係 (**IPR**) のアノテーションを行い，ゴールドスタンダードを作る．次にそのゴールドスタンダードを用いて投稿間インタラクション (**IPI**) (Target/Callout) のアノテーションを行う．すなわち，全アノテータは同じ IPR のアノテーション結果を用いて，IPI のアノテーションを行うことになる．

ゴールドスタンダードを決定する際には，アノテータの多数決投票で決定することにした．なお今回，1 つのタスクを担当するアノテータは 3 人とした．投稿内 AM のゴールドスタンダードを決定するプロセスは次の通りである．

A1: AS のタグ (Premise/Claim/NA) を過半数 (2/3) の一致で採用．多数決で決まらないものは NA に分類する．
A2: IPR を，過半数 (2/3) の一致で仮採用する．
A3: A1 と A2 の結果を統合することで，文章タグが Claim である文章をルートノードにした木を得る（Claim の数だけ木ができる）．
A4: どの木にも含まれない Premise は定義を満たさないので，NA に分類し直し，関係も削除する．

[5]Target を含む投稿が別の投稿に対して返信していれば，同時に Callout にもなりうる．

[6]ファシリテーションの観点では，Claim がユーザ間でどのように評価されているかを知ることが望ましい．また，Callout を Claim に限定することによって，AS から出る関係先が最大一つになり，問題がシンプルになるという利点があるため，このような制約条件をつけた．

表 1 AS タイプと IPR，IPI の IAA．

コーパス	タイプ	サイズ	κ
COLLAGREE (我々のコーパス)	Claim	1449	.531
	Premise	2762	.554
	NA	1348	.529
	IPR	2762	.466
	IPI	745	.430
Persuasive Essays	Claim	1506	.635
	Premise	3832	.833
	Relation (Supp)	3613	.708
	Relation (Att)	219	.737

図 2 は，A1〜A4 のプロセスの具体例である．また，IPI のゴールドスタンダードについても，投稿内のゴールドスタンダードを決定した後に，同様に関係の多数決を取ることで決定した（A2 のプロセスの IPR を IPI に置き換えて実施する）．

3.3 アノテータ間一致度の評価

我々が提案したスキームにおいて，アノテータ間一致度の評価，すなわち Inter-annotator Agreement (IAA) の対象となるのは，AS タイプと，IPR と，IPI の 3 つである．IAA の手法は Fleiss's κ [21] を採用する．しかし，予備実験として IPR のゴールドスタンダードを評価したところ $\kappa = 0.420$ と比較的低い値となってしまった．この理由として,「前提の前提」を許していることが考えられる．すなわち投稿内の前提が多い場合に，どのように前提を組み合わせるかでアノテータが一致しにくいことが原因であると仮定した[7]．そこで，A1〜A4 のプロセスの前に，前提を親に持つ前提があれば，直接主張へと関係を修正するプロセス A0 を導入した[8]．

表 1 に各 AS タイプや IPR，IPI のアノテーション数と Fleiss's κ を示す[9] [10] [11]．比較のために Persuasive Essays [7, 8] のデータも併記した．この表から，我々のデータが AM において最大規模のコーパスに比較的近い規模でアノテーションを行うことが出来たと言える．また，エッセイなどのフォーマルな文章 [8] と異なり，我々のデータは砕けた表現が多いので一致度が比較的低い結果になったが， [30] によれば，0.41〜0.60 の一致度は

[7]エッセイのデータセット [8] と異なり，市民の文章はよく構造化されていないことが多いため，必要以上に複雑な構造化をする意味がないと考えている．

[8]例えば，$\{(claim1 \leftarrow premise1), (premise1 \leftarrow premise2)\}$ という関係が存在した時に，プロセス A0 によって $\{(claim1 \leftarrow premise1), (claim1 \leftarrow premise2)\}$ に修正する．

[9]実際には，時間と予算の都合でデータセットを 4 分割し，それぞれの分割において 5 人ずつ人員を割り当てた．よって κ は 4 つの平均値を算出したものである．

[10]IPI における，Target となる AS タイプ別の数は Claim,Premise,NA がそれぞれ 574,109,62 であった．本研究では Callout の AS タイプが Claim のみであることに留意すると，「Claim → Claim」が全体の 77%を占めていることが分かる．すなわち，IPI は極めて議論的であると言える．

[11]支持関係/攻撃関係 [31] のアノテーションも行っており，支持関係が 86%，攻撃関係が 7%と，圧倒的に同調する Claim が多い結果となった．

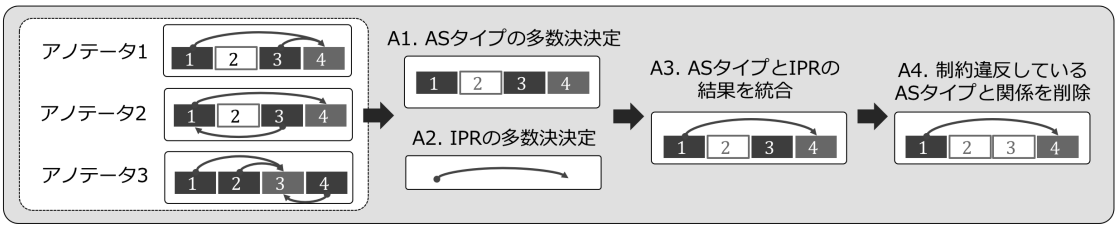

図 2 投稿内アノテーションにおけるゴールドスタンダードの決定の例.

「Moderate agreement」であるため，分類学習が可能であると考えられる．さらに，プロセス A0 を導入したことにより IPR の一致度が上昇した．議論要素の数について言えば，Premise，Claim がそれぞれ全体の 50%，26%であり，非議論的な要素（挨拶など）も非常に多くアノテーションされる結果となった．

3.4 スレッド表現の入力系列への変換

スレッドに含まれる投稿は，ある特定のテーマに対しての議論に対して言及しているため，E2E な識別モデルの入力系列はスレッド単位が望ましい．すなわち，入力をスレッドとして，各投稿に含まれる文章の AS タイプ，IPR と IPI を同時に全て出力するモデルを実現したい．そこで，本節ではスレッド構造を入力系列として表現するための，モデリング方法について述べる．

我々は，既存研究の系列モデリング手法 [17] を拡張することにした．初めに入力の単位を単語ではなく文章単位とする．このようにすることで，LSTM [32] への入力系列長を短くすることが出来る．続いて，スレッドでの投稿を階層の深さ順にならべ，各階層で時系列順に並べる．また各階層の区切りと各投稿の区切りにそれぞれ区切り表現を挿入し，学習させやすくする．このようにすることで，階層構造や返信関係を入力系列の中で表現できる．図 1 を入力系列としてモデリングした例を図 3 に示す.

4 評価実験

アノテーションの先の最終的な目標は，文章の AS タイプと IPR と IPI を自動的に識別することである．本節では，我々のスキームを用いてアノテーションされたデータが，機械学習によって学習可能であることを示すための試験的な実験を行う．

4.1 識別モデル

本論文では，比較のために 4 つの識別モデルを評価することにした．1 つ目は，[17] で提案された Multi Task Learning (MTL) モデル [23] (**STagBLSTM**) を拡張して，スレッドに対する E2E 学習を提案する．このモデルを導入する理由は，スレッドのように比較的大きな入力単位においても E2E な識別が可能であることを示すためである．STagBLSTM は AS タイプの識別器と関係識別器の 2 つの出力層が，同じ双方向 LSTM [34] の中間層を共有するモデルである[12] [13] [14]．今回は，STagBLSTM の，双方向 LSTM を一つだけ用いる ($\mathcal{Y}$-3) を再現実装して実験を行った[15]．

STagBLSTM の再現実装は Chainer [35] を用いて行った．ハイパーパラメータは次のとおりである: 単語埋め込み層の次元数 512，双方向 LSTM の隠れ層の次元数 256，ドロップアウトレート [36, 37]0.9，オプティマイザに Adam [38]，ミニバッチサイズ 16．さらに，時間の都合で，50 epochs で実験を打ち切り，その時点でのモデルを用いてテストを行った．学習データの内訳は，訓練:テスト=8:2 である．

2 つ目の識別モデルには特徴ベースの SVM [8] を導入する（**SVM - T**）．この識別手法を用意した理由は，タスクに特化した特徴ベースの手法が効果的かどうかを検証するためである．T はそれぞれのタスク（Claim, Premise, IPR, IPI）別になっており，独立した分類器である．特徴素として，頻度が上位 500 件の単語の BoW を用いた[16]．

3 つ目と 4 つ目は，SVM - T と同様の目的で，ランダムフォレスト（**RF - T**）と，ロジスティック回帰 [42] によるモデルを用意した（**Simple - T**）.

4.2 評価方法

正当な評価を行うために，本節では説得性のあるメトリクスを導入する．本論文では，文章の AS タイプを正

[12] ただし我々の場合，入力の単位が文章単位であるため，トークンレベルの代わりに文章レベルでの入力を行う．

[13] AS タイプの出力層では，3 クラス（Claim, Premise, NA）分類のクロスエントロピー誤差関数によって学習する．IPR と IPI の出力層では，出力時点のタイムステップにおいて，何個先（もしくは何個前）に関係させるか，という実数値（その結果をさらに整数値に丸めたものが推定結果）を出力し，2 乗誤差関数を用いて学習する．

[14] 文章レベルでの入力には，文章に含まれる全単語の Word Embedding [33] や Bag-of-words (BoW) を合計したものを用いることが多い．今回は単純な入力でも学習が可能であることを示すために，入力を BoW とし，埋め込みベクトル $\boldsymbol{E}$ をモデルのパラメータとして学習させた．

[15] 予備実験において ($\mathcal{Y}$-3) だけでなく，($\mathcal{Y}$-3, C-2) など複数の MTL を試したが，どれも精度向上には繋がらなかった．

[16] [8] では，feature-rich な入力を用いていたが，我々のデータセットの性質とは大きく異なるため，比較のために BoW のみを用いた．学習は 2fold のクロスバリデーションを用いてベストなハイパーパラメータを選択した．

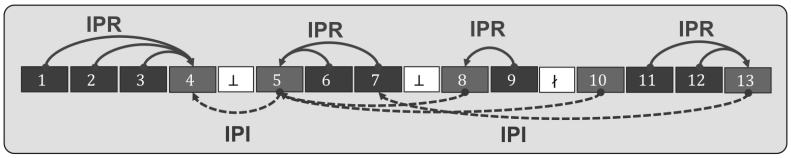

図 3　スレッド構造を入力系列としてモデリングする例．最優先順序は階層で，その次の優先順位は時系列となっている．⊥ は階層の区切り表現，⫽ は投稿の区切り表現である．矢印は便宜のために表示していることに留意．

表 2 各モデルの最良識別性能 (F1 スコア (%))．各列の最大値を太字で表示した．

モデル種別	モデル名	AS タイプの識別			関係の識別	
		Claim - F1	Premise - F1	NA - F1	IPR - F1	IPI - F1
End-to-End (E2E)	STagBLSTM	**54.2**	65.6	**56.9**	14.9	**12.6**
	STagBLSTM - 区切り表現無し	51.8	66.1	55.2	14.5	10.8
Task Specific	SVM - T	53.3	64.4	52.3	**22.4**	11.5
	RF - T	41.0	**66.8**	38.3	0	0
	Simple - T	41.1	66.1	38.3	0	0

しく分類できたかの指標として適合率，再現率，F 値を導入した．また IPR の識別と IPI の識別についても適合率，再現率，F 値を算出できるようにした．IPR については，self-loop を除く全ての関係のペアを作る [8] ことで評価が可能である[17]．IPI についても同様にペアを作ることで評価する[18]．注意点として，返信元から返信先への向きのみを考慮することに留意されたい．なぜならば，明らかに返信元の文章のほうが生成時刻が早いためである．

4.3　実験結果

表 2 に，各識別モデルの F1 スコアを示す．それぞれのモデルにおいて最良モデルのスコアを記載した．なお，スペースの都合で適合率や再現率の記載は省略した．驚くことに，我々のデータは κ が低いにも関わらず，比較的良好な F 値を算出している．特に IPR と IPI はネガティブケースが圧倒的に多い不均衡データであるため，単純な素性と SVM で IPR において 22.4%となったことは注目に値する．全体として，AS タイプの識別性能は E2E なモデルの STagBLSTM が効果的であることが分かる．また，IPR においても STagBLSTM が最も高いパフォーマンスを示した．特徴ベースなモデルのうち RF と Simple は IPR と IPI を全く識別できなかった．これは，ネガティブケースが多い不均衡データであることに起因すると考えられる．

[17]例えば，ある投稿内の文章が (S_1, S_2, S_3) であったとして，ゴールドスタンダードの IPR が $(S_1 \leftarrow S_2)$，$(S_1 \leftarrow S_3)$ となっている場合には，ポジティブケースは明らかに $\{(S_1 \leftarrow S_2), (S_1 \leftarrow S_3)\}$ の 2 つである．ネガティブケースは self-loop を含まない全てのペア，すなわち $\{(S_2 \leftarrow S_1), (S_2 \leftarrow S_3), (S_3 \leftarrow S_1), (S_3 \leftarrow S_2)\}$ の 4 つである（ここでの self-loop は $\{(S_1 \leftarrow S_1), (S_2 \leftarrow S_2), (S_3 \leftarrow S_3)\}$ の 3 つ）．

[18]例えば，返信元の投稿の文章が (S_1, S_2, S_3)，返信している投稿の文章が (S_4, S_5) とする．このとき，ゴールドスタンダードのインタラクション $(S_2 \leftarrow S_5)$ が存在するとする．明らかに $\{(S_2 \leftarrow S_5)\}$ がポジティブケースとなる．ネガティブケースは self-loop を除く全ての組み合わせなので $\{(S_1 \leftarrow S_4), (S_1 \leftarrow S_5), (S_2 \leftarrow S_4), (S_3 \leftarrow S_4), (S_3 \leftarrow S_5)\}$ の 5 つである．

表 2 では，区切り表現の有効性を示すための実験結果も示している．「STagBLSTM - 区切り表現無し」と記載してあるモデルは区切り表現を省略して STagBLSTM を学習した時のスコアを示している．Premise 以外の識別性能において STagBLSTM の性能を下回る結果となったため，本論文で提案した区切り表現が効果的であることが分かった．この理由について我々は，区切り部分の情報が LSTM でエンコードされることで，構造的な情報を保持できているのではないかと考える．

5　結論

本論文では，スレッド構造を持つオンライン市民議論に対して，Argument Mining (AM) のスキームがどのように適用されるかを示した．初めに Claim, Premise, Target, Callout の概念を取り入れ，スレッド構造のための新たな AM スキームを提案した．続いて実際の大規模オンライン市民議論データに対して提案スキームに基づくアノテーションを行った．アノテーションの結果，AM の分野において最高レベルの規模のコーパスが得られた．また，アノテーションされたデータを用いて機械学習が可能であるかどうかを示すために，自動構造化のための識別モデルを提案した．識別モデルとして我々は，SVM や End-to-End (E2E) なニューラルモデルなどの手法を提案した．評価実験の結果，我々の提案した E2E なモデルが全体的に高いパフォーマンスを示した．すなわち，スレッド構造を入力として，文章のタイプや投稿内関係，投稿間のインタラクションを一気に判別することが出来るようになった．今後はよりスレッド構造に特化した学習モデルの研究を進めていく予定である．

謝辞

本研究は，JST，CREST の支援を受けたものである．また，COLLAGREE による市民議論データを提供いた

だいた，名古屋工業大学伊藤孝行教授，秀島栄三教授，伊藤孝紀准教授，白松俊准教授に感謝する．

参考文献

[1] Pang, B. and Lee, L.: Opinion Mining and Sentiment Analysis, Found. Trends Inf. Retr., pp. 1–135, 135, 2008.

[2] Rinott, R., Dankin, L., Perez, C. A., et al.: Show Me Your Evidence - an Automatic Method for Context Dependent Evidence Detection, Proc. of the 2015 Conference on EMNLP, pp. 440–450, 2015.

[3] Palau, R. M. and Moens, M-F.: Argumentation Mining: The Detection, Classification and Structure of Arguments in Text, Proc. of ICAIL '09, pp. 98–107, 2009.

[4] Lippi, M. and Torroni, P.: Argumentation Mining: State of the Art and Emerging Trends, ACM Trans. Internet Technol, pp. 10:1–10:25, 2016.

[5] Nishida, T., Ito, T., Ito, T., et al.: Core time mechanism for managing large-scale internet-based discussions on COLLAGREE, Proc. of IEEE International Conference on Agents, pp. 46–49, 2017.

[6] T., Ito., Imi, Y., Ito, T., et al.: COLLAGREE: A Faciliator-mediated Large-scale Consensus Support System, Proc. of the 2nd International Conference of Collective Intelligence, 2014.

[7] Persing, I. and Ng, V.: Modeling Argument Strength in Student Essays, Proc. of the 53rd Annual Meeting of the ACL and the 7th IJCNLP, pp. 543–552, 2015.

[8] Stab, C. and Gurevych, I.: Parsing Argumentation Structures in Persuasive Essays, Computational Linguistics, pp. 619-659, 2017.

[9] Nguyen, H. V. and Litmann, D. J.: Contextaware argumentative relation mining, Proc. of the 54th Annual Meeting of the ACL, pp. 1127–1137, 2016.

[10] Taghipour, K. and Ng, H. T.: A Neural Approach to Automated Essay Scoring, Proc. of the 2016 Conference on EMNLP, pp. 1882–1891, 2016.

[11] Ghosh, D., Khanam, A., Han, Y., et al.: Coarse-grained Argumentation Features for Scoring Persuasive Essays, Proc. of the 54th Annual Meeting of the ACL, pp. 549–554, 2016.

[12] Lawrence, J. and Reed, C.: Using Complex Argumentative Interactions to Reconstruct the Argumentative Structure of Large-Scale Debates, Proc. of the 4th Workshop on Argument Mining, pp. 108–117, 2017.

[13] Reed, C. and Rowe, G.: ARAUCARIA: SOFTWARE FOR ARGUMENT ANALYSIS, DIAGRAMMING AND REPRESENTATION, Artificial Intelligence Tools, pp. 961–979, 2004.

[14] Ghosh, D., Muresan, S., Wacholder, N., et al.: Analyzing Argumentative Discourse Units in Online Interactions, Proc. of the First Workshop on Argument Mining, hosted by the 52nd Annual Meeting of the ACL, pp. 39–48, 2014.

[15] Kitagawa, R. and Fujita, K.: Automatic Summarization Considering Time Series and Thread Structure in Electronic Bulletin Board System for Discussion, Proc. of the 5th IIAI International Congress on Advanced Applied Informatics, pp. 681–686, 2016.

[16] Morio, G. and Fujita, K.: Predicting Argumentative Influence Probabilities in Large-Scale Online Civic Engagement, Companion Proceedings of the The Web Conference, pp. 1427–1434, 2018.

[17] Eger, S., Daxenberger, J. and Gurevych, I.: Neural End-to-End Learning for Computational Argumentation Mining, Proc. of the 55th Annual Meeting of the ACL, pp. 11–22, 2017.

[18] Hidey, C., Musi, E., Hwang, A., et al.: Analyzing the Semantic Types of Claims and Premises in an Online Persuasive Forum, Proc. of the 4th Workshop on Argument Mining, pp. 11–21, 2017.

[19] Rocha, G. and Lopes Cardoso, H.: Towards a Relation-Based Argument Extraction Model for Argumentation Mining, Proc. of the 5th International Conference, SLSP, pp. 94–105, 2017.

[20] Meyers, R. A. and Brashers, D.: Extending the Conversational Argument Coding Scheme: Argument Categories, Units, and Coding Procedures, Communication Methods and Measures, pp. 27-45, 2010.

[21] Fleiss, J. L.: Measuring nominal scale agreement among many raters, Psychological Bulletin, pp. 378–382, 1971.

[22] Stab, C. and Gurevych, I.: Annotating Argument Components and Relations in Persuasive Essays, Proc. of the 25th International Conference on Computational Linguistics, SLSP, pp. 1501–1510, 2014.

[23] Søgaard, A and Goldberg, Y.: Deep multi-task learning with low level tasks supervised at lower layers, Proc. of the 54th Annual Meeting of the ACL, pp. 231–235, 2016.

[24] Martínez Alonso, H. and Plank, B.: When is multitask learning effective? Semantic sequence prediction under varying data conditions, Proc. of the 15th Conference of the EACL, pp. 44–53, 2017.

[25] Persing, I. and Ng, V.: End-to-End Argumentation Mining in Student Essays, Proc. of Human Language Technologies: The 2016 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 1384–1394, 2016.

[26] Eckle-Kohler, J. and Kluge, R. and Gurevych, I: On the Role of Discourse Markers for Discriminating Claims and Premises in Argumentative Discourse, Proc. of the 2015 Conference on EMNLP, pp. 2236–2242, 2015.

[27] Lippi, M. and Torroni, P.: Context-independent Claim Detection for Argument Mining, Proc. of the 24th IJCAI, pp. 185–191, 2015.

[28] Peldszus, A. and Stede, M.: From Argument Diagrams to Argumentation Mining in Texts: A Survey, Int. J. Cogn. Inform. Nat. Intell., pp. 1–31, 31, 2013.

[29] Gao, Y., Wang, H., Zhang, C., et al.: Reinforcement Learning Based Argument Component Detection, arXiv, abs/1702.06239, 2017.

[30] Landis, J. R. and Koch, G. G.: The Measurement of Observer Agreement for Categorical Data, Biometrics, 33, 1, 1977.

[31] Cocarascu, O. and Toni, F.: Identifying attack and support argumentative relations using deep learnin, Proc. of the 2017 Conference on EMNLP, pp. 1374–1379, 2017.

[32] Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, Neural Comput., pp. 1735–1780, 48, 1997.

[33] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., et al.: Distributed Representations of Words and Phrases and their Compositionality, Advances in Neural Information Processing Systems 26, pp. 3111–3119, 2013.

[34] Graves, A. and Schmidhuber J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures, NEURAL NETWORKS, pp. 5–6, 2005.

[35] Tokui, S., Oono, K., Hido, S., et al.: Chainer: a Next-Generation Open Source Framework for Deep Learning, Proc. of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on NIPS, 2015.

[36] Srivastava, N., Hinton, G., Krizhevsky, A., et al.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research, pp. 1929–1958, 15, 2014.

[37] Zarrella, G. and Marsh, A.: MITRE at SemEval-2016 Task 6: Transfer Learning for Stance Detection, Proc. of the 10th International Workshop on Semantic Evaluation, pp. 458–463, 2016.

[38] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, arXiv, abs/1412.6980, 2014.

[39] Miwa, M. and Bansal, M.: End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures, Proc. of the 54th Annual Meeting of the ACL, pp. 1105–1116, 2016.

[40] Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., et al.: Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions, Proc. of WWW, pp. 613-624, 2016.

[41] Petasis, G. and Karkaletsis, V.: Identifying Argument Components through TextRank, Proc. of the Third Workshop on Argument Mining, pp. 94–102, 2016.

[42] Peldszus, A. and Stede, M.: Joint prediction in MST-style discourse parsing for argumentation mining, Proc. of the 2015 Conference on EMNLP, pp. 938–948, 2015.