An Empirical Study of Span Representations in Argumentation Structure Parsing

Tatsuki Kuribayashi††††, Hiroki Ouchi†††††, Naoya Inoue††††††††, Jun Suzuki†††††, Paul Reisert††††, Toshinori Miyoshi††††† and Kentaro Inui†††††

Argumentation Structure Parsing (ASP) is the task of predicting the roles of argumentative units (e.g., claim, premise) and the relations between the units (e.g., support, attack) in an argumentative text. ASP has received a great deal of attention due to its usefulness for applications such as automatic assessment of argumentative
texts. As textual spans (i.e., argumentative units) are basic units of ASP, it is important to explore an effective design for representing them. Inspired by the current span representation design in other natural language processing tasks, we propose a method to obtain effective span representations of argumentative units in ASP. Our proposed method leverages multiple levels of global contextual information, such as argumentative markers in surrounding contexts, for obtaining each span representation. We show that using our span representation improves performance on several benchmark datasets—especially when parsing complex argumentative texts, which have been difficult to parse with existing methods. Furthermore, we report the effectiveness of our span representations when using word representations obtained from existing, powerful language models such as BERT.

Key Words: Argument Mining, Discourse Parsing, Span Representation

1 はじめに

議論は人間にとって主要な言語活動のひとつである。議論の参加者は、前提・根拠に基づきながら深く考え、筋道を立てて自身の意見を伝える。例えば、国の方向を決定したり、親を説得したり、価格を交渉したり、問題解決や合意形成において議論は欠かせないものである。近年、自然言語処理の分野では、特に作文や意見文といった独話的な論述文を対象とし、議論の解析を行う議論マイニング (Argument Mining) が進展を遂げてきた (Cabrio and Villata 2018; Lawrence and Reed 2019)。議論マイニングにおける知見やシステムは、人々の意見の集約 (Stab, Miller, Schiller, Rai, and Gurevych 2018; Reimers, Schiller, Beck, Daxenberger, Stab, and Gurevych 2019) や、議論の質の自動評価・フィードバック (Stab and Gurevych 2016; Wachsmuth, Naderi, Hou, Bilu, Prabhakaran, Thijum, Hirst, and Stein 2017; Reisert, Vallejo, Inoue, Gurevych, and Imii 2019) などへ応用が期待されている。議論マイニングにおける中心的なゴールとして、論述文における言語構造の解析（以降, 論述構造解析）が挙げられる。本研究では、論述構造解析のためのベンチマークデータセット (Peldszus and Stede 2016; Stab and Gurevych 2017) 上で、高性能な論述構造解析モデルの開発を行う。

論述構造解析モデルは、与えられた論述文について、話題単位間の論述関係とその種類 (Support や Attack)、話題単位の機能 (Premise や Claim, MajorClaim) などの構造を予測する。図 1 に論述文とその論述構造の例を示す。この例では、In addition, I believe that city provides more work opportunities than the countryside. という主張（話題単位 1）について、主張を支持する事例（話題単位 2）や、主張と対立する意見（話題単位 3）などが述べられている。論述構造はグラフで表現され (Peldszus and Stede 2015; Stab and Gurevych 2017)。グラフの頂点は話題単位を、グラフの辺は話題単位間の論述関係を表す。

話題単位のように一単語以上からなる意味的関連のあるまとまりをスパンと呼ぶ。論述構造
解釈は論文中の談話単位スパンの役割を解釈するタスクである。スパンに対する特徴ベクトル（スパン分散表現）をどのように計算するかはモデル設計において重要な点である。論述構造解析における既存研究（Potash, Romanov, and Rumshisky 2017）では、ニューラルネットワークベースのスパン分散表現を用いることで、高い解釈性能が実現されてきた。同様に、統語解析や意味解析などの自然言語処理においても、ニューラルネットワークベースのスパン分散表現は注目を集めている。より効果的なスパン分散表現抽出方法について知見が得られてきた（Wang and Chang 2016; Stern, Andreas, and Klein 2017; He, Lee, Levy, and Zettlemoyer 2018; Ouchi, Shindo, and Matsumoto 2018）。これらの知見と論述文特有の言語的性質を踏まえ、本研究では論述構造解析において効果的なスパン分散表現抽出方法の提案を行う。談話単位の機能や役割は文脈に大きく依存するため、各談話単位のスパン分散表現に論述の文脈情報をうまく取り入れることが解釈精度向上の鍵である（Nguyen and Litman 2016; Lawrence and Reed 2019）。

論述構造を予測する上で、重要となる文脈情報は大きく二つあると考える。一つは、ある談話単位の周辺にどのような論述関係が存在するかという高次の情報であり、文章中の接続表現の配置からある程度推測することができる。例えば，“of course”→“but”といった接続表現の系列からは、一旦譲歩した上で反論し返すという攻撃関係の連鎖など、典型的な部分構造が捉えられる（図1中、談話単位3と4）。もう一つは、談話単位間の語彙的な結びや話題の変化である。近い話題について議論している談話単位同士は、論述関係を共有している可能性が高いと考える。例えば、図1中、談話単位1と2では“more work”→“more jobs”という非常に意味の近い内容語が含まれており、これらの談話単位間の関係は支持の関係で結ばれている。

そこで本研究では、各談話単位を機能的な表現（接続表現）と内容（命題）に分離し（図1中、下線付き部分と下線なし部分）、談話単位、接続表現、命題という様々な観点における文脈情報を考慮しながら、談話単位のスパン分散表現を獲得する（3節）。接続表現の系列は論述の
テーマ非依存な文章の型に関する手がかりであるのに対し、命題は論述のテーマに大きく依存した内容であるなど、両者が異なる性質を持つことも、接続表現と命題を区別して扱う動機の一つである。

実験から、本研究で提案したスパン分散表現獲得方法を用いることで解析性能が向上することを示した。また、BERTなどの強力な言語モデルから得られる単語分散表現を用いた際も、既存のスパン分散表現獲得方法を用いた場合では十分な性能が得られないが、本研究で提案したスパン表現獲得方法を用いることで、大幅な性能向上が得られることを示す。分析から、特に複雑な構造を対象とした論述文において、スパン表現の工夫による性能向上が得られること分かった。

本研究の貢献は以下の通りである。

- 論述構造解析において、言語処理における他タスクで有効とされていたスパン分散表現と、本タスクのために拡張したスパン分散表現の有効性を調査した。
- 実験結果から、既存のスパン分散表現、およびスパン分散表現の拡張が本タスクにおいて有効であることを示し、複数のベンチマークデータで最高性能を達成した。
- 分析から、複雑な論述構造（深いグラフ）を持つ文章において、特に我々のスパン分散表現の獲得方法が有効であることが分かった。

2 関連研究

論述構造解析。論述構造解析は、論述文から議論の構造を抽出するタスクである。議論学の知見（Toulmin 1958; Freeman 2011）に基づき、論述構造アノテーションの仕様検討から、論述構造解析モデルの開発に至るまで様々な研究が行われてきた（Reed 2006; Mochales and Moens 2011; Peldszus and Stede 2013, 2015, 2016; Stab and Gurevych 2014, 2017; Niculae, Park, and Cardie 2017; Potash et al. 2017; Eger, Daxenberger, and Gurevych 2017; Habernal and Gurevych 2017）。近年は、論述構造を考慮したエッセイコアリングや論述文の評価も注目を集めている（Ke, Carlile, Gurrapadi, and Ng 2018; Nguyen and Litman 2018）。

近年、論述構造解析のための比較的高品質なデータセット（Stab and Gurevych 2017; Peldszus and Stede 2016）が公開され、これらのベンチマークデータセットの上で、論述構造解析のための様々なモデルが提案されてきた。論述構造解析モデルの設計は、(i) 人手で設計された索引に基づくモデル（Peldszus and Stede 2015; Stab and Gurevych 2017; Afantenos, Peldszus, and Stede 2018）と、(ii) ニューラルネットワークベースのモデル（Potash et al. 2017; Eger et al. 2017）に分けられる。人手による索引に基づくモデルでは、論述文の言語的な性質を踏まえ、解析精度向上のため論述文や論話に関連した索引が設計されてきた。これらの研究では、論述関係を予測する際の広い文脈を考慮する重要性や、周辺文脈に出現する接続表現が手がかりとなること
が報告されてきた。一方，ニューラルネットワークベースのモデルでは，言語的に動機付けされた素性はほとんど用いられていないものの，高性能な解析が実現されてきた。したがって，それぞれのアプローチで得られた知見が，もう一方のアプローチでは活かされてこなかった。本研究では，素性ベースの研究で得られた知見に着目し，ニューラルベースモデルに論述文の性質を考慮した改良を加えることで，両アプローチにおけるこれまでの知見を統合し，解析性能の向上を図る。

論述構造分析

論述構造解析と関連の強いタスクとして，修辞構造理論（Mann and Thompson 1987）に基づく論述構造解析や，Penn Discourse TreeBank (PDTB) (Prasad, Dinesh, Lee, Miltasakaki, Robaldo, Joshi, and Webber 2008) タイプの論述関係認識がされられる。修辞構造理論やPDTBタイプの論述解析では一般的なドメインの話題を解析対象とし，話題単位間で生じている原因，結果，対比，並列などの関係を予測する。一方，論述構造解析では論述的，記述的な文章に焦点が当たり，相手を説得するという目標における談話単位の機能（根拠・主張）や談話単位間の関係（支持・反論）を予測する。

修辞構造理論や論述構造といった，異なる言語構造アノテーション間の対照的な性質についても，近年分析が行われている。既存研究 (Stede, Afantenos, Peldszus, Asher, and Perret 2016) では，同じテキストに論述構造，修辞構造理論に基づく談話構造など複数の談話構造を重ねて付与し，付与された辺（どの談話単位間に関係があるか）全体のうち，両構造で重複する辺は高々60%程度であるなどの報告があり，既存の論述解析タスクとの性質の違いが示唆されている。また，PDTBタイプの論述関係認識は，二つの談話単位が与えられ談話単位間の関係の種類を予測するタスクであるにせよ，論述構造解析や修辞構造理論に基づく談話構造解析では，2つ以上の談話単位が与えられ，どの単位間に関係が存在するかも予測する点でタスク設定が異なる。

接続表現と命題

人による効性設計を行った論述構造解析の研究 (Stab and Gurevych 2017) では，典型的な接続表現のリストを作成し，各談話単位の周辺文脈に存在する接続表現を効性として用いることで，有意な精度向上を実現している。例えば，ある談話単位が Therefore，で始まる場合，その談話単位は結論であると考えられる上に，その直前の談話単位は根拠である可能性が高い。接続表現は，一般的な談話構造解析（Mann and Thompson 1987; Prasad et al. 2008）でも，非常に有用な手がかりとされている (Marcu 2000; Braud and Denis 2016)。接続表現の活用は，談話構造解析の枠を超えて，文の表現学習などの分野でも注目を集めている (Sileo, Van-De-Cruys, Pradel, and Muller 2019; Pan, Yang, Zhao, Zhuang, Cai, and He 2018)。

また，テキスト言語学の文脈 (De Beaugrande and Dressler 1981; Halliday and Hasan 2014) では，節や文の繋がりについて，結論性 (cohesion)・一貫性 (coherence) という二つの観点から論じられてきた。結論性は文章の表層に現れる繋がりを指し，接続詞・指示代名詞などによって実現される文法的結論性と，同じ語の繋り返しなどで実現される語彙的結論性に更に分類される。
自然言語処理 Vol. 27 No. 4 December 2020

また、一貫性は節、文間の意味的な繋がり（出費が多い ⇒ 収入が必要など）を指し、ニューラルネットワーク内部で接続詞と命題を区別して文章の流れを捉える本研究の試みは、文章の文法的結び性（接続表現がどう配置されているか）と、語彙的結び性（命題の語彙・意味的な繋がり）の観点それぞれに特化した機構を、ネットワーク構造に導入する試みと考えることもできる。

スパン分散表現

統語解析 (Wang and Chang 2016; Stern et al. 2017; Kitaev and Klein 2018)、意味役割付与 (He et al. 2018; Ouchi et al. 2018)、共参照解析 (Lee, He, Lewis, and Zettlemoyer 2017; Lee, He, and Zettlemoyer 2018)、話分析解析 (Li, Li, and Chang 2016) などの言語処理タスクにおいて、スパン分散表現の設計は大きな注目を受けています。ここで、スパンとは複数の単語からなる言語単位（句やフレーズなど）を指す。近年、LSTM-minus と呼ばれるスパン特徴表現抽出方法が提案され (Wang and Chang 2016)。様々なタスクで適用されている。この手法では、スパン(i,j)に対してスパン両端のトークンに対応する BiLSTM の隠れ層の差 (h_j - h_{i-1}) を用い、文脈情報を取り入れたスパン分散表現を計算する。論文文中の該当単位の解釈において文脈情報は重要であり (Nguyen and Litman 2016; Lawrence and Reed 2019)。例えば、「りんごには栄養がある」という発話の機能は、「毎朝りんごを食べるべきだ」と続く発話であれば、主張を支持していることになるが、「果物は体に悪い」とする発話であれば反論になる。本研究でも、文脈を考慮したスパン分散表現獲得法である LSTM-minus をベースとし、より効果的なスパン分散表現について研究を行う。

3 モデル

論述構造解析モデルの概要、及び提案手法について説明する。

3.1 モデルの概要

論述構造解析は、(i) 論述単位分割、(ii) 論述関係認識、(iii) 論述関係タイプ分類、(iv) 論述単位タイプ分類の4つのサブタスクによって構成される。既存研究 (Potash et al. 2017; Niculae et al. 2017; Peldszus and Stede 2015) に従い、本研究では論述単位は分割済みであると仮定する。論述構造解析モデルは論文を入力とし、各論述単位のスパン分散表現を計算する (3.2.3.3.4節)。その後、各論述単位間の論述関係の有無、各論述関係のタイプ分類、各論述単位のタイプ分類を行う (3.5節)。

3.2 スパンの種類

論述文は T 個の単語 \( w_1, \ldots, w_T \) で構成され、K 個の論述単位スパン \( s_{1,K}^{ADU} = (s_{1}^{ADU}, s_{2}^{ADU}, \ldots, s_{K}^{ADU}) \) をもつ。ここでスパン \( s_k^{ADU} \) は、単語インデックス \( i_k, j_k \) を用いて \( (i_k, j_k) \)
図 2 論述文の例と、接続表現と命題の区別をする／しないモデルの概要。ADU は論述単位、AM は接続表現、AC は命題を指す。

と表す (1 ≤ i_k ≤ j_k ≤ T)。本研究では、各論述単位 s_k^{ADU} をさらに接続表現 s_k^{AM} と命題 s_k^{AC} に分離する。本研究で扱う英語の論述文では、接続表現は論述単位の先頭に存在し、命題は論述単位中の接続表現よりも後ろとする（図 2 右）。すなわち、各論述単位 s_k^{ADU} = (i_k, j_k) は、単語インデックス i_k (i_k ≤ j_k < j_k) を用いて、接続表現 s_k^{AM} = (i_k, j_k) と命題 s_k^{AC} = (j_k + 1, j_k) に分割できる。論述文中に論述単位が K 個存在する場合、接続表現および命題も K 個ずつ存在する。以下、論述単位を ADU (argumentative discourse unit)、命題を AC (argument component)、接続表現を AM (argumentative marker) と表す。

3.3 LSTM-minus 基づくスパン分散表現

はじめに、言語処理他タスクに有効なスパン分散表現である LSTM-minus (Wang and Chang 2016) 基づく、論述構造解析モデルを設計する（図 2 左）。本モデルでは、AC と AM を区別する。
せず、ADU を一つのスパンとして扱う。ADU スパン $s_{k}^{ADU} = (i_k, j_k)$ のスパン分散表現 $h_{k}^{ADU}$ を以下のように得る。

$$w_{1:T} = f_{emb}(w_{1:T}),$$
$$h_{1:T} = BiLSTM(w_{1:T}),$$
$$h_{k}^{ADU} = \left[ \hat{h}_{j_k}, \hat{h}_{i_k-1}; \hat{h}_{i_k}, \hat{h}_{j_k+1}; \hat{h}_{i_k-1}; \hat{h}_{j_k+1}; \phi(w_{i_k:j_k}) \right].$$

ここで、単語埋め込み層 $f_{emb}$ は、入力系列 $w_{1:T}$ を単語分散表現の系列 $w_{1:T}$ に変換する。BiLSTM レイヤは $w_{1:T}$ を入力とし、隠れ層の系列 $h_{1:T}$ を出力する。$h_{1:T}$ から、各 ADU スパン $s_{k}^{ADU}$ について LSTM-minus ベースのスパン分散表現 $h_{k}^{ADU}$ を計算する。なお、既存研究 (Potash et al. 2017) に従い、各スパン分散表現に特徴ベクトル $\phi(w_{i:j})$ を追加している。

本スパン分散表現を用いたモデルを LSTM モデルと呼ぶ。

### 3.4 接続表現と命題を区別したスパン分散表現

LSTM-minus に基づくスパン分散表現の拡張として、AM と AC を区別したスパン分散表現を用いる (図 2 右)。AC は論述の論理的な筋を構成するか否か、AM は However などの論の流れを汲み取る上で標識となる接続表現に対応する。AC の内容は論述のテーマに依存するのに対し、AM は論述のテーマに非依存であるなど、AC と AM は異なる性質を持つ。

各論述単位 $s_{k}^{ADU}$ を接続表現 $s_{k}^{AM}$ と命題 $s_{k}^{AC}$ に分離する。LSTM スパンモデルと同様に、各 AM スパン $s_{k}^{AM}$ をスパン分散表現 $h_{k}^{AM}$ に、各 AC スパン $s_{k}^{AC}$ をスパン分散表現 $h_{k}^{AC}$ にエンコーデする。$K$ 個の AM/AC スパン分散表現の系列をそれぞれ、$H_{1}^{AM} = (h_{1}^{AM}, h_{2}^{AM}, \ldots, h_{K}^{AM})$。$H_{1}^{AC} = (h_{1}^{AC}, h_{2}^{AC}, \ldots, h_{K}^{AC})$ と表す。

$H_{1}^{AM}$ と $H_{1}^{AC}$ について、異なる BiLSTM を用いてそれぞれ文脈情報を取り入れる。

$$H_{1}^{AM,txt} = BiLSTM^{AM}(H_{1}^{AM}),$$
$$H_{1}^{AC,txt} = BiLSTM^{AC}(H_{1}^{AC}).$$

ここで $H_{1}^{AM,txt} = (h_{1}^{AM,txt}, h_{2}^{AM,txt}, \ldots, h_{K}^{AM,txt})$ と $H_{1}^{AC,txt} = (h_{1}^{AC,txt}, h_{2}^{AC,txt}, \ldots, h_{K}^{AC,txt})$ は、それぞれ文脈情報を取り込んだ $K$ 個の AM/AC スパン分散表現の系列を表す。

AM と AC を別々に処理することで、複数の観点における論の主な流れを捉える。例えば、直前の談話単位が議論 (of course...) をしていた場合、その直後の談話単位では議論の内容に反論し返す (ATTACK 関係をもつ) のが自然である。議論している節では対立する見解が述べられていると考えられ、自分の主張に対して不利な見解には直ちに反論しなければ、論述文の説得力が低下する恐れがあるからだ。このような特徴量は BiLSTM^{AM} によって捉える。ま
た。一つのテーマについて論じられている文章であっても、しばしば論文の中では様々な論点が挙げられ、テーマに対して多角的に批判・分析が行われる。近い話題について議論している談話単位同士は、論述関係を共有している可能性が高いと考える。このような話題、結論性の変化は BiLSTMACによって捉える。

各 ADU スパン $s_k^{ADU}$ を、$h_k^{AM,ctx}$ と $h_k^{AC,ctx}$、LSTM モデルと同様の特徴ベクトル $\phi(w_{i:k})$ を用いて以下のように表す。

$$h_k^{ADU, dist} = [h_k^{AM,ctx}; h_k^{AC,ctx}; \phi(w_{i:k})].$$

本スパン分散表現を用いたモデルを LSTM+dist モデルと呼ぶ。

### 3.5 出力層

以降は LSTM モデル、LSTM+dist モデルで共通のレイヤーとなる。3.3 節または 3.4 節で得られた、ADU スパン $s_k^{ADU}$ に対する分散表現を $h_k^{ADU}$ と表記し、$K$ 個のスパン分散表現の系列を $H_1^{ADU} = (h_1^{ADU}, h_2^{ADU}, \ldots, h_K^{ADU})$ とする、$H_1^{ADU}$ を BiLSTM に入力し、ADU のレベルでスパン分散表現に文脈情報を取り入れる。

$$H_1^{ADU,ctx} = \text{BiLSTM}(H_1^{ADU}). \quad (1)$$

LSTM モデルと LSTM+dist モデルを対比するために、LSTM+dist モデルでは、式 1 における BiLSTM の層数を減らし、LSTM モデルとパラメータ数を抑えている。得られた ADU スパン分散表現 $H_1^{ADU,ctx} = (h_1^{ADU,ctx}, h_2^{ADU,ctx}, \ldots, h_K^{ADU,ctx})$ をもとに、それぞれのサブタスクにおけるクラスの確率分布を求める。サブタスクは、(i) 論述関係認識、(ii) 論述関係タイプの分類、(iii) 論述単位のタイプの分類の 3 つである。いずれのタスクも、各 ADU に対する多クラス分類問題として定義される。以降簡単化のため、論述単位 $s_n^{ADU}$ のスパン分散表現 $h_k^{ADU,ctx}$ を $h_k$ と表記する。

**論述関係認識** 論述関係認識レイヤでは、スパン $s_m^{ADU}$ がスパン $s_n^{ADU}$ に対して有向辺を張る確率を以下のように計算する。

$$\text{score}_{m,n}^{link} = w_{\text{link}} \cdot [h_m; h_n; h_m \odot h_n; \phi(m, n)].$$

$$P(n|s_m^{ADU}) = \frac{\exp(\text{score}_{m,n}^{link})}{\sum_{n'=1}^{K} \exp(\text{score}_{m,n'}^{link})},$$

ここで $w_{\text{link}}$ は学習対象となるパラメータである。$\phi(m, n)$ は、$m$ 番目のスパンと $n$ 番目のスパンの間の相対的な距離を示す one-hot vector である。本研究で扱うデータセットでは、論述構

---

3 LSTM モデルの場合、3.3 節における $h_k^{ADU}$、LSTM+dist モデルの場合、3.4 節における $h_k^{ADU,dist}$ に対応する。
論述関係タイプの分類。スパン $s^{ADU}_o$ が論述関係タイプ $r$ に分類される確率を、スパン分散表現 $h_o$ をもとに以下のように計算する。

$$\text{score}_{o,r}^{\text{link-type}} = w_r^{\text{link-type}} \cdot h_o + b_r^{\text{link-type}}, \quad (2)$$

$$P(r|s^{ADU}_o) = \frac{\exp(\text{score}_{o,r}^{\text{link-type}})}{\sum_{r' \in R} \exp(\text{score}_{o,r'}^{\text{link-type}})}, \quad (3)$$

ここで、$w_r^{\text{link-type}}$, $b_r^{\text{link-type}}$ は学習対象となるパラメータである。また、$R = \{\text{Support}, \text{Attack}\}$ である。既存研究に従い、各 ADU スパンを係り元とする論述関係のタイプの分類という形で各 ADU スパン分散表現をもとに予測を行う。また、本研究で用いるデータセットの一つである Persuasive essay corpus (PEC) (Stab and Gurevych 2017) では、CLAIM に筆者が同じ立場か異なる立場かというスタンスのラベルが付与されている。既存研究に従い、PEC では本スタンスの分類も論述関係タイプ分類問題として扱う。

論述単位タイプの分類。式 2、3 と同様に、パラメータ $w_r^{\text{ac-type}}$, $b_r^{\text{ac-type}}$ を用いて、スパン $o$ が論述単位タイプに分類される確率 $P(r|s^{ADU}_o)$ を、スパン分散表現 $h_o$ をもとに求める。

### 3.6 学習

以下の訓練データセット $D$ を用いる。

$$D = \{(X, Y^{\text{link}}, Y^{\text{link-type}}, Y^{\text{ac-type}})_{d=1}^{|D|} ,$$

$$X = \{w_{1,1}, s^{ADU}_{1,K}, s^{AM}_{1,K}, s^{AC}_{1,K}\},$$

$$Y^{\text{link}} = \{h_1, \ldots, h_K\},$$

$$Y^{\text{link-type}} = \{t_1, \ldots, t_K\},$$

$$Y^{\text{ac-type}} = \{r_1, \ldots, r_K\}$$

ここで、$h_k \in \{\text{root}, 1, 2, \ldots, K\}$, $t_k \in \{\text{Support}, \text{Attack}\}$, $r_k \in \{\text{MajorClaim}, \text{Claim}, \text{Premise}\}$である。マルチタスク学習が有効であるという既存研究 (Potash et al. 2017; Stab and Gurevych 2017) の知見に従い、本研究ではこれら 3 つのサブタスクをマルチタスクで解く。以下の式より、モデルパラメータ $\hat{\theta}$ を得る。

$$\hat{\theta} = \arg \min_{\theta} \mathcal{L}(\theta),$$

4 microtext corpus (Peldszus and Stede 2016) では、MajorClaim クラスは存在しない。
$L(\theta) = - \sum_{(X,Y^{\text{link}},Y^{\text{link-type}},Y^{\text{ac-type}}) \in D} \left( \alpha \ell_\theta^{\text{link}}(X,Y^{\text{link}}) + \beta \ell_\theta^{\text{link-type}}(X,Y^{\text{link-type}}) + (1 - \alpha - \beta) \ell_\theta^{\text{ac-type}}(X,Y^{\text{ac-type}}) \right),$ 

ここで、$\alpha$ と $\beta$ は各サブタスクに対応する損失関数の重みを決める係数であり、人手により決定するハイパラメータである。ただし、$\alpha \geq 0$, $\beta \geq 0$, $\alpha + \beta \leq 1$ とする。また、各サブタスクの損失関数は以下のように表される。

\[
\ell_\theta^{\text{link}}(X,Y^{\text{link}}) = \sum_{k \in \{1,2,\ldots,K\}} \log P_\theta(h_k | s_k^{\text{ADU}}), \\
\ell_\theta^{\text{link-type}}(X,Y^{\text{link-type}}) = \sum_{k \in \{1,2,\ldots,K\}} \log P_\theta(t_k | s_k^{\text{ADU}}), \\
\ell_\theta^{\text{ac-type}}(X,Y^{\text{ac-type}}) = \sum_{k \in \{1,2,\ldots,K\}} \log P_\theta(r_k | s_k^{\text{ADU}}).
\]

学習時のハイパラメータは、付録表 11 に記載する。

4 実験

4.1 実験設定

データセット。本研究では、Persuasive essay corpus (PEC) (Stab and Gurevych 2017) と arg-microtext corpus (MTC) (Peldszus and Stede 2016) の 2 つのパーコープデータセットを用いる。PEC はオンラインフォーラムに投稿された 402 作文（1,833 段落）からなる。PEC における論述構造は各段落内に固定されているため、既存研究に従い、各段落にアルゴリズムを適用した。データセット作成者 (Stab and Gurevych 2017) が定めた訓練／評価データの分割を用い、学習データから無作為に選ばれた 10%のデータを開発データとした。また、既存研究 (Potash et al. 2017) に従い、PEC におけるスコアについては、異なるシードを用いて行なった 3 回の試行の平均を報告した。MTC は、様々な年齢や教育課程のドイツ語話者によって書かれた 112 テキストを、英語に翻訳したもので構成される (Peldszus and Stede 2016)。本データセットは规模が小さいため、既存研究 (Peldszus and Stede 2015) の分割に従い、5 分割交差検証を分割の仕方を変えた 10 回行った際の平均スコアを報告する。表 1 に PEC と MTC の統計を示す。PEC は MTC のおよそ 10 倍大きいデータセットである。表中の論述構造の深さは、それぞれの段落における論述構造の深さの平均値であり、ルートオブジェクトは深さの計算に含まれていない。例えば CLAIM とそれに直接接する PREMISE のみが存在する場合は、段落の深さは 1 とする。また、PEC における導入の段落など論述関係が存在しない段落は計算から外した。MTC の方が比較的深い構造を有しているが、これは、MTC データの収集時に「なるべく反論を含めるこ

5 https://essayforum.com/
表1 PECとMTCの統計

|                | PEC | MTC |
|----------------|-----|-----|
| 論述文         |     |     |
| 文章数         | 402 | 112 |
| 段落数         | 1,833 | 112 |
| 文数           | 7,116 | -   |
| トーカン数       | 147,271 | -   |
| 論述単位数     |     |     |
| MajorClaim     | 751  | -   |
| Claim          | 1,506 | 112 |
| Premise        | 3,832 | 464 |
| 論述関係数     |     |     |
| SUPPORT        | 3,613 | 290 |
| ATTACK         | 219  | 174 |
| 論述構造の深さ | 1.46 | 2.06 |

と」など論述構造が複雑になる方向の指示がされていることが原因であると考えている。

接続表現の抽出。PECとMTCで付与されているスパンが異なる。PECでは、ACスパンは付与されているものの、AMスパンは付与されていない。本研究では、ルールに基づいてAMスパンを抽出した。各ACが属する文のうち、ACに先行する部分をAMスパンとする。ただし、AMスパンは他のACスパンと重ならない。また、AMにおいて主節・従属節の区別を明確にするため、AMスパン前の文の終了記号（i.e., ., !, and ?）やテキストの始まりを示す特殊記号を含めた、例として、以下の二つの文における節BのAMを考える。

(1) a. A. Because B, C.
   b. A because B, C.

例 (1)-aにおいてBはCの従属節であり、例 (1)-bにおいてBはAの従属節である。節BのAMを、例 (1)-aでは“Because”，例 (1)-bでは“because”として抽出した。ADUがAMを持たない場合は、前の文の終了記号やテキストの始まりを示す特殊記号のみがAMとなる。PECデータセットでは、63%のADUが終了記号以外の単語を含むAMを持っていた。ACに対応するAMを合わせて、ADUスパンとした。

MTCでは、ADUが付与されており、ACとAMの区別はついていない。そこで、PECで同定したAMとPenn Discourse TreeBank (PDTB) (Prasad et al. 2008)における談話標識のリストをもとにAMのリストを作成し、各ADUからAMを抽出した6（表2）。AMのリストには

6 AMリスト作成段階では、前の文の終了記号（i.e., ., !, and ?）やテキストの始まりを示す特殊記号は含めず、ADUからAMを抽出する段階で、それらの記号をAMスパンの一部として追加した。
1,131の表現（平均5.38トークン）が含まれ、うちPDTBから収集した表現が173種類。PECから収集した表現が958種類である。各ADUについて、AMリスト中に存在する表現がADUの先頭に存在する場合、最長一致するフレーズをそのADUのAMとした。ADU中のAM以外の箇所はACとした。MTC中のADUのうち、およそ48%にAMが付与された。

AMが文頭以外に出現した場合（Others, however, think that these children may disrupt their school work and should be allowed to leave school early to find a job.），AMとACを区別するとACが複数のスパンに分かれてしまう。従ってこのようなケースでは、AMに対応する表現は存在しないとみなした。AMがADUの途中に挿入される事例は稀であり、例えば““，however,”が存在するADUは、PEC中の全ADUのうち1%程度である。

MTC上で、本アプローチによるAMの同定性能を調べた。MTCからランダムに抽出した100個のADUについて、人手でAMスパンのアノテーションを行い、本アプローチによるAMの抽出結果を評価した（表3）。表中のスコアは、AMスパンの完全一致を正解とみなして計算している。抽出された結果を観察すると、“but”などの典型的なAMは概ね抽出できているものの、例えば“This would mean that”などの長い表現については抽出されず目立ち、改善の余地があった。実験では、本アプローチで同定したAMの情報を活用し、MTC上で性能が向上した。

| 表2 | 接続表現の例 |
|---|---|
| the other reason is that  |  |
| consequently, |  |
| in conclusion, from the above views, although |  |
| first, as you can see that |  |
| in this essay, the reasons for why i agree that |  |
| it is a debatable subject that |  |
| unfortunately |  |
| although some argue that |  |
| furthermore, it’s undeniable that the |  |
| in short, |  |
| another thing that put big cities in front of small towns is |  |
| in conclusion, despite the contribution of it to the society, |  |
| however, some say that |  |

| 表3 | 接続表現の同定性能 |
|---|---|
| 適合率 | 再現率 | F1値 |
| 79.2 | 74.5 | 76.8 |

注7: https://github.com/davidsbatista/NER-Evaluationによって、exactスコアを求めた。
していることから，ある程度妥当な同定が実現できていると解釈している。

単語表現． 単語表現の獲得手段として，GloVe (Pennington, Socher, and Manning 2014), ELMo (Peters, Neumann, Iyyer, Gardner, Clark, Lee, and Zettlemoyer 2018), BERT (Devlin, Chang, Lee, and Toutanova 2019), RoBERTa (Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer, and Stoyanov 2019), XLNET (Yang, Dai, Yang, Carbonell, Salakhutdinov, and Le 2019)を用いた．ELMoを用いた実験では，1層目から3層目の表現の平均9を用いた．BERT, RoBERTa, XLNETを用いた単語表現については，Transformers (Wolf, Debut, Sanh, Chaumond, Delangue, Moi, Cistac, Rault, Louf, Funtowicz, et al. 2019)で公開されている事前学習済みモデル9に文章を入力し，各トークンに対応する中間表現を獲得した．各層から獲得した中間表現の重みつき和を対応するトークンの表現とし，重みはタスク依存で学習した．トークンが複数のサブワードに分割される場合は，それらのサブワードに対応する表現の平均をトークンの表現とした．

ベースライン． 既存の論述構造解析モデル (Potash et al. 2017)で活用されているスパン分散表現を用い，LSTM-Minusに基づくスパン分散表現と比較した．Potashらのスパン分散表現では，スパン中の単語ベクトルの平均などが用いられており，以降このスパン分散表現を用いたモデルを BoW モデル (Bag-of-Words) と呼ぶ．詳細な実装は付録に記載する．

4.2 結果

既存研究に従い，3タスクそれぞれにおけるF1スコアと，モデルの総合的な性能として3タスクのF1スコアの平均を報告する．各タスクのMacroF1スコアについて，プートストラップ法 (Koehn 2004)で検定を行った (Dror, Baumer, Shlomov, and Reichart 2018)．

表4にPEC上での実験結果を，表5にMTC上での実験結果を示す．結果から，(i) BoWモデルに比べて，LSTMモデルの方が良い性能を示していること，(ii) ACとAMの区別を行うモデル（LSTM+dist）が，区別を行わないモデル（LSTM）よりも良い性能を示していること，(iii)これらの傾向は，異なるデータセット，異なる単語表現を用いた場合も概ね一貫していることがわかる．XLNETやRoBERTaを用いた実験結果においてもベースライン (BoW) の性能が低いことから，強力な言語モデルから得られた単語表現を活用するだけでは十分な性能は引き出せず，これらの単語表現を用いてどのようにスパン分散表現を獲得するかが更なる精度向上の鍵であると分かる．

LSTM-minusベースのスパン分散表現を用いたことによる性能向上（LSTM > BoW）は特にPEC上で顕著である（表4）。MTC上では性能向上の傾向が比較的小さが（表5）、MTCが小規模なデータセットであるため，LSTM-minus スパン分散表現の抽出に用いるパラメータを

---

8 事前実験において，重みつき和をとるよりも平均をとった方が性能が高かった。
9 bert-large-cased，roberta-large，xlnet-large-casedを使用した。
表4 PEC上でのLSTM+dist, LSTM, BoWモデルの性能比較

| 順位 | モデル   | 3タスク  | 論述関係認定 | 論述関係タイプ分類 | 論述単位タイプ分類 |
|------|----------|----------|--------------|-------------------|-------------------|
|      | LSTM+dist | 84.3     | 82.2   | 75.8  | 96.2  | 60.7  | 89.4  | 94.5  | 79.8  | 94.1  |
|      | LSTM     | 82.5     | 80.6   | 77.8  | 96.6  | 59.1  | 89.1  | 94.5  | 79.5  | 94.2  |
|      | BoW      | 77.4     | 76.4   | 71.6  | 95.7  | 47.6  | 84.3  | 91.4  | 71.0  | 90.4  |
|      | LSTM+dist | 82.9     | 80.9   | 79.4  | 96.9  | 61.9  | 88.4  | 94.7  | 77.7  | 92.8  |
|      | LSTM     | 82.9     | 80.6   | 77.9  | 96.5  | 59.3  | 89.1  | 94.3  | 79.6  | 93.3  |
|      | BoW      | 75.6     | 73.1   | 71.9  | 95.4  | 48.4  | 81.9  | 89.0  | 67.3  | 89.3  |
|      | LSTM+dist | 81.8     | 80.9   | 78.0  | 96.5  | 59.6  | 86.4  | 92.5  | 74.9  | 92.0  |
|      | LSTM     | 80.6     | 80.4   | 74.9  | 96.0  | 53.8  | 86.6  | 92.2  | 75.4  | 92.2  |
|      | BoW      | 73.9     | 71.8   | 69.4  | 95.6  | 43.1  | 79.9  | 87.8  | 63.8  | 88.1  |
|      | LSTM+dist | 81.8     | 80.7   | 79.0  | 96.8  | 61.1  | 85.7  | 91.6  | 73.3  | 92.1  |
|      | LSTM     | 81.8     | 80.4   | 78.2  | 96.7  | 59.8  | 86.9  | 92.4  | 76.4  | 92.0  |
|      | BoW      | 77.1     | 76.2   | 72.3  | 96.2  | 48.3  | 82.9  | 90.4  | 68.6  | 89.6  |
|      | GloVe    | 79.7     | 78.8   | 76.5  | 96.5  | 56.6  | 83.9  | 91.2  | 72.1  | 88.4  |

表5 MTC上でのLSTM+dist, LSTM, BoWモデルの性能比較

| 順位 | モデル   | 3タスク  | 論述関係認定 | 論述関係タイプ分類 | 論述単位タイプ分類 |
|------|----------|----------|--------------|-------------------|-------------------|
|      | LSTM+dist | 79.3     | 75.8   | 73.7  | 82.1  | 65.3  | 88.3  | 81.1  | 95.6  |
|      | LSTM     | 76.0     | 73.6   | 69.3  | 77.1  | 61.6  | 85.2  | 75.8  | 94.6  |
|      | BoW      | 76.3     | 73.3   | 71.1  | 79.1  | 63.1  | 84.6  | 74.6  | 94.6  |
|      | LSTM+dist | 77.1     | 74.0   | 71.6  | 80.5  | 62.6  | 85.9  | 7.1   | 94.6  |
|      | LSTM     | 73.1     | 71.8   | 66.4  | 75.2  | 57.7  | 80.9  | 68.6  | 93.3  |
|      | BoW      | 73.4     | 71.2   | 66.7  | 76.2  | 57.2  | 82.2  | 70.8  | 93.8  |
|      | LSTM+dist | 76.1     | 73.0   | 71.1  | 79.7  | 62.6  | 84.0  | 74.2  | 93.9  |
|      | LSTM     | 70.7     | 69.9   | 64.2  | 73.5  | 54.9  | 77.9  | 63.2  | 92.6  |
|      | BoW      | 71.9     | 70.0   | 65.2  | 75.6  | 54.9  | 80.4  | 67.6  | 93.2  |
|      | LSTM+dist | 78.2     | 73.9   | 77.2  | 84.2  | 70.3  | 83.4  | 72.9  | 94.0  |
|      | LSTM     | 75.0     | 73.1   | 71.3  | 78.7  | 64.0  | 80.5  | 68.1  | 93.0  |
|      | BoW      | 73.3     | 71.2   | 67.5  | 76.7  | 58.5  | 81.2  | 69.0  | 93.5  |
|      | GloVe    | 76.5     | 72.6   | 75.4  | 82.3  | 68.4  | 81.5  | 69.9  | 93.1  |

十分に学習できなかったことなどが原因として考えられる。また，ACとAMの区別による性能向上（LSTM+dist > LSTM）は，特にMTC上で大きい（表5）。MTCデータセットは，PECに比べてSUPPORTに対するATTACK関係の比率が高いことから，対立意見を踏まえて議論し，反論し返すなどの複雑な構造を持つ文章が多いことが考えられる。AC・AMごとの流れ（例え
表 6  PEC における既存研究との性能比較

| モデル | 3 タスク | 論述関係認識 | 論述関係タイプ分類 | 論述単位タイプ分類 |
|--------|----------|-------------|------------------|------------------|
| LSTM+dist (XLNET) | 83.4 | 82.2 | 70.3 | 94.2 |
| BoW (XLNET) | 77.4 | 76.4 | 60.4 | 92.3 |
| LSTM+dist (GloVe) | 79.7 | 78.8 | 64.6 | 93.0 |
| Potash+ 2017 | — | — | 60.1 | — |
| Niculae+ 2017 | — | — | — | — |
| Stab+ 2017 | 75.2 | 75.1 | 58.5 | 91.8 |

MC は MAJORCLAIM の略である。Potash+ 2017 の結果は、再実装によるものである。

表 7  MTC における既存研究との性能比較

| モデル | 3 タスク | 論述関係認識 | 論述関係タイプ分類 | 論述単位タイプ分類 |
|--------|----------|-------------|------------------|------------------|
| LSTM+dist (XLNET) | 79.3 | 75.8 | 60.7 | 90.8 |
| BoW (XLNET) | 76.3 | 73.3 | 56.8 | 89.8 |
| LSTM+dist (GloVe) | 76.5 | 72.6 | 55.4 | 89.8 |
| Potash+ 2017 | — | — | 60.1 | — |
| Afantenos+ 2018 | 78.5 | 72.2 | — | — |
| Stab+ 2017 | 76.2 | 68.3 | 48.6 | 88.1 |

Potash+ 2017 の結果は、再実装によるものである。

ば、of course→but などの接続表現の系列）を追うことにより、このような複雑な文章においても顕著に解析できたのではないかと考える。また、PEC データセットは非常に多様なトピックに関する作文で構成されるのに対し、MTC データセットには同じトピックに関する論文が複数含まれている。そのため MTC では、トピックの影響が大きい AC のレベルにおいても、同一トピックの他の論文から学習したパターンを、新たな論文の解析に転用しやすかった可能性を考えられる。例えば模擬試験における作文の自動評価への論述構造解析の応用を考えると、特定のテーマに関する論文が複数存在する MTC の状況は自然である。

LSTM+dist モデルと既存研究における結果との比較を表 6 と表 7 に示す。スパン分散表現の工夫と最先端のモデルから得られた単語表現を組み合わせることで、PEC において論述関係認識における Link 有りクラスの F1 値を 10 ポイント近く向上させる (60.8→70.3) など、解析性能を大きく更新した。強力な言語モデルによる単語ベクトルを用いた場合も、BoW(XLNET) と Potash+ 2017 の比較から、スパン表現の獲得方法次第では既存のモデルと大きく変わらない
性能を示すことが分かり、LSTM+dist(XLNET) と BoW(XLNET) の比較から、提案法によるスパン表現によって性能が大きく向上することが分かる。また、スパン分散表現獲得法により、言語レベルの文脈を積極的に取り入れることで、論述構造解析における性能が大きく向上することから、単語の穴埋めや次の文の予測といった言語モデルの目的関数によって得られた単語分散表現のものには、間接的の流れといった話題的な文脈情報は十分に含まれていない可能性があることも示唆される。

5 分析

論述構造解析において主要に取り組まれてきた論述関係認識・分類タスクに着目し、予測の難しいインスタンスや既存のモデルとの出力傾向の違いについて分析する。本分析では、比較的規模の大きい PEC の開発データを用いた。

5.1 論述構造の深さ

既存研究 (Stab and Gurevych 2017) では、論述構造解析モデルが正解の木よりも浅い木を予測しがちであることが指摘されていた。この指摘を踏まえ、木の深さと予測性能の観点から、既存のモデルと本研究における提案モデルを比較する。図 3 に、ルートオブジェクトの認識精度（深さ 0）、深さ 1 の ADU から出る関係の認識精度、深さ 2 以上の ADU から出る関係の認識精度を示す。各数値は正しいキューを当てることができた正解率を示す。全体的な傾向として、正解の構造において深い位置にある論述関係ほど、予測が難しいことが分かる。深さ 2 以上の関係の予測精度に着目すると、既存のモデル (Potash+ 2017) の予測精度は大きく低下しているのに対し、スパン分散表現に多粒度の文脈情報を取込んでいる LSTM+dist モデルは、比較的
頑健に予測できていることが分かる。これらの結果から、深い位置の関係を予測する際には文脈情報が重要であり、文脈情報のスパン表現において重要であることが示唆される。

論述構造が深くなる典型的なケースとして、一旦想定される対立意見を述べてそれに対して反論し返すという、Attack 関係の連鎖が挙げられる（Peldszus and Stede 2013; Freeman 2011）。図 4 では、Attack が連鎖する部分構造（論述単位 4→論述単位 3→論述単位 1）を持つ論述構造の例と、既存モデルの予測および提案モデルの予測結果である。既存のモデルは予測に失敗しているが、LSTM+dist モデルは正しく予測できている。表 8 に、Attack 関係が連鎖している部分構造における論述関係の予測精度を示す。LSTM+dist モデルが良い性能を示していることから、AC と AM の区別が予測に良い影響を与えていることが分かる。異なる立場の意見を挙げて議論を進める文章では、何か自分の主張で何かが対立意見なのかを読み手が追いやすいように、接続表現 (AM) が頻繁に用いられると考えられる。従って、of course→but といった AM の流れが予測の手がかりになったと考えられる。また、このような部分構造は特に MTC で頻繁に観察され、MTC において LSTM モデルと LSTM+dist モデルの性能差が大きい理由の一つであると考えられる。

木の深さと論述関係分類性能についても分析を行う。正解の木における深さごとの論述関係分類性能を求めた（表 9）。深さ 0 における論述関係分類は CLAIM が作文の筆者が自身の立場のものであるか、対立する立場のものであるかという問題に対応し、特に深さ 0 における予測性能が LSTM+dist を用いた場合に一貫して向上していることが分かる。この結果は、論述関係認識において深い位置の ADU の係り先の予測精度が向上している傾向と対照的である。

CLAIM の立場を推測するためには、その主張が周囲の ADU から反論されているのか、補強されているのかといった明確な論の流れが手がかりとなると考えられる。様々な粒度で大域的
表 8  ATTACK 関係が連鎖している部分構造における論述関係の予測精度

| 埋め込み | モデル     | 精度 |
|---------|------------|------|
| XLNET   | LSTM+dist  | 71.6 |
|         | LSTM       | 67.8 |
|         | BoW        | 60.7 |
| RoBERTa | LSTM+dist  | 69.9 |
|         | LSTM       | 67.2 |
|         | BoW        | 56.8 |
| BERT    | LSTM+dist  | 69.3 |
|         | LSTM       | 68.8 |
|         | BoW        | 57.4 |
| ELMo    | LSTM+dist  | 71.0 |
|         | LSTM       | 68.9 |
|         | BoW        | 60.7 |
| GloVe   | LSTM+dist  | 63.9 |
|         | LSTM       | 59.0 |
|         | BoW        | 57.9 |

表 9 深さごとの論述関係分類性能

| 埋め込み | モデル     | 深さ   |
|---------|------------|--------|
|         | 0          | 1      | 2  以上 |
| XLNET   | LSTM+dist  | 78.2   | 76.5 | 77.3 |
|         | LSTM       | 77.1   | **77.4** | **78.0** |
|         | BoW        | 73.2   | 67.3 | 69.5 |
| RoBERTa | LSTM+dist  | 81.6   | 76.4 | 74.2 |
|         | LSTM       | 76.2   | **77.1** | **79.6** |
|         | BoW        | 73.2   | 66.4 | 72.5 |
| BERT    | LSTM+dist  | 79.5   | **74.4** | **73.9** |
|         | LSTM       | 76.8   | 70.8 | 71.2 |
|         | BoW        | 72.6   | 60.4 | 72.1 |
| ELMo    | LSTM+dist  | 79.1   | 75.4 | **78.8** |
|         | LSTM       | 79.0   | **76.0** | 75.9 |
|         | BoW        | 73.5   | 68.3 | 68.9 |
| GloVe   | LSTM+dist  | **77.4** | 69.7 | 75.1 |
|         | LSTM       | 73.9   | **71.6** | **76.9** |
|         | BoW        | 70.0   | 59.4 | 60.2 |

値は Macro F1 値である。

771
表 10 接続表現の有無と論述関係認識性能

| 埋め込み | モデル     | 接続表現なし | 接続表現あり |
|----------|------------|-------------|-------------|
| XLNET    | LSTM+dist  | 73.2        | 78.8        |
|          | LSTM       | 69.8        | 77.6        |
|          | BoW        | 62.9        | 72.6        |
| RoBERTa  | LSTM+dist  | 69.6        | 78.2        |
|          | LSTM       | 70.8        | 79.1        |
|          | BoW        | 59.8        | 67.8        |
| BERT     | LSTM+dist  | 70.3        | 77.2        |
|          | LSTM       | 68.9        | 76.7        |
|          | BoW        | 59.5        | 67.5        |
| ELMo     | LSTM+dist  | 71.8        | 76.0        |
|          | LSTM       | 70.3        | 76.4        |
|          | BoW        | 63.1        | 71.5        |
| GloVe    | LSTM+dist  | 66.6        | 72.7        |
|          | LSTM       | 65.4        | 73.2        |
|          | BoW        | 57.6        | 67.1        |
| Potash+  | LSTM       | 67.0        | 70.9        |

な文献を挿取る LSTM+dist モデルが、深さ 0 における論述関係分類性能の向上に有効であったと考えられる。また、表 8 における ATTACK 関係の連鎖の分析では、筆者に対して立場の CLAIM とそれに反論し返す PREMISE といった部分構造の予測性能も対象となっている。対立的な CLAIM は周囲の PREMISE によって再反論されやすいという高次の傾向が存在し（Kuribayashi, Reisert, Inoue, and Inui 2017）。LSTM+dist モデルが対立する CLAIM をうまく同定できることで、付近に存在する PREMISE の係り先（反論先）を予測する手がかかりとなり、表 8 にみられるようなマクロな構造の解析性能の向上や、図 3 でみられるような深い位置の PREMISE の係り先の予測性能の向上に影響を与えていると考えている。

5.2 接続表現の有無

本研究では、論述文における接続表現 (AM) に焦点を当ててモデルの改良を行ったため、分析においても接続表現に焦点を当て、接続表現をもつ／もたない ADU から出る論述関係の認識精度について調査する。表 10 に分析結果を示す。各数値は正しい親を当てることができたかの正解率を示す。全体の傾向として、接続表現のある ADU よりも、接続表現のない ADU の方が論述関係の予測精度が低いことが観察される。この傾向は、PDTB スタイルの論述関係認識などにおいて、話作標識が表示されていない非明示的な論述関係の認識性能が難しいこと
6 まとめ

本研究では、論述構造解析において効果的なスパン分散表現獲得手法を提案し、実験結果から、接続表現と命題をネットワーク内で明示的に区別し、スパン分散表現に多粒度の文脈を取り込むことが重要であることが示唆される。今後の展望として、論述単位の分割も含めて全てのサブタスクを一貫通貫学習で解くことや、修辞構造理論に基づく談話構造の解析や、論述構造に関連した他タスク（Trautmann, Daxenberger, Stab, Schütze, and Gurevych 2020）における本手法の有効性の検証を行うことの重要性を示唆する。

謝辞

本研究の一部は The 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019) で発表したものです (Kuribayashi, Ouchi, Inoue, Reisert, Miyoshi, Suzuki, and Inui 2019)。また、本研究の一部は JST CREST（課題番号：JPMJCR1513）の支援を受けて行いました。

参考文献

Afantenos, S., Peldszus, A., and Stede, M. (2018). “Comparing Decoding Mechanisms for Parsing Argumentative Structures.” Argument & Computation, 9 (3), pp. 177–192.

Braud, C. and Denis, P. (2016). “Learning Connective-based Word Representations for Implicit Discourse Relation Identification.” In Proceedings of the EMNLP 2016, pp. 203–213.

Cabrio, E. and Villata, S. (2018). “Five Years of Argument Mining: A Data-driven Analysis.” In Proceedings of the IJCAI 2018, pp. 5427–5433.

De Beaugrande, R. and Dressler, W. U. (1981). Introduction to Text Linguistics. Routledge.
Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). “BERT: Pre-training of Deep
Bidirectional Transformers for Language Understanding.” In Proceedings of the NAACL
2019, pp. 4171–4186.

Dror, R., Baumer, G., Shlomov, S., and Reichart, R. (2018). “The Hitchhiker’s Guide to Testing
Statistical Significance in Natural Language Processing.” In Proceedings of the ACL 2018,
pp. 1383–1392.

Eger, S., Daxenberger, J., and Gurevych, I. (2017). “Neural End-to-End Learning for Computa-
tional Argumentation Mining.” In Proceedings of the ACL 2017, pp. 11–22.

Freeman, J. B. (2011). Dialectics and the Macrostructure of Arguments: A Theory of Argument
Structure, Vol. 10. Walter de Gruyter.

Habernal, I. and Gurevych, I. (2017). “Argumentation Mining in User-generated Web Discourse.”
Computational Linguistics, 43 (1), pp. 125–179.

Halliday, M. A. K. and Hasan, R. (2014). Cohesion in English. Routledge.

He, L., Lee, K., Levy, O., and Zettlemoyer, L. (2018). “Jointly Predicting Predicates and Argu-
ments in Neural Semantic Role Labeling.” In Proceedings of the ACL 2018, pp. 364–369.

Ke, Z., Carlile, W., Gurrapadi, N., and Ng, V. (2018). “Learning to Give Feedback: Modeling
Attributes Affecting Argument Persuasiveness in Student Essays.” In Proceedings of the 27th
IJCAI, pp. 4130–4136.

Kitaev, N. and Klein, D. (2018). “Constituency Parsing with a Self-Attentive Encoder.” In
Proceedings of the ACL 2018, pp. 2676–2686.

Koehn, P. (2004). “Statistical Significance Tests for Machine Translation Evaluation.” In Pro-
cceedings of the EMNLP 2004, pp. 388–395.

Kuribayashi, T., Ouchi, H., Inoue, N., Reisert, P., Miyoshi, T., Suzuki, J., and Inui, K. (2019).
“An Empirical Study of Span Representations in Argumentation Structure Parsing.” In
Proceedings of ACL2019, pp. 4691–4698, Florence, Italy. Association for Computational Lin-
guistics.

Kuribayashi, T., Reisert, P., Inoue, N., and Imi, K. (2017). “Examining Macro-level Argument-
ative Structure Features for Argumentative Relation Identification.” IPSJ SIG Technical
Report, 2017-NL-234, pp. 1–6.

Lawrence, J. and Reed, C. (2019). “Argument Mining: A Survey.” Computational Linguistics,
45, pp. 765–818.

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). “End-to-end Neural Coreference Resol-
tion.” In Proceedings of the EMNLP 2017, pp. 188–197.

Lee, K., He, L., and Zettlemoyer, L. (2018). “Higher-Order Coreference Resolution with Coarse-
to-Fine Inference.” In Proceedings of the NAACL-HLT 2018, pp. 687–692.
Li, Q., Li, T., and Chang, B. (2016). “Discourse Parsing with Attention-based Hierarchical Neural Networks.” In Proceedings of the EMNLP 2016, pp. 362–371.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). “RoBERTa: A Robustly Optimized BERT Pretraining Approach.”
Mann, W. C. and Thompson, S. A. (1987). “Rhetorical Structure Theory: Description and Construction of Text Structures.” In Natural Language Generation, pp. 85–95.
Marcu, D. (2000). The Theory and Practice of Discourse Parsing and Summarization. MIT press.
Mochales, R. and Moens, M.-F. (2011). “Argumentation Mining.” Artificial Intelligence and Law, 19 (1), pp. 1–22.
Nguyen, H. and Litman, D. (2016). “Context-aware Argumentative Relation Mining.” In Proceedings of the ACL 2016, pp. 1127–1137.
Nguyen, H. V. and Litman, D. J. (2018). “Argument Mining for Improving the Automated Scoring of Persuasive Essays.” In Proceedings of the AAAI 2018, pp. 5892–5899.
Niculae, V., Park, J., and Cardie, C. (2017). “Argument Mining with Structured SVMs and RNNs.” In Proceedings of the ACL 2017, pp. 985–995.
Ouchi, H., Shindo, H., and Matsumoto, Y. (2018). “A Span Selection Model for Semantic Role Labeling.” In Proceedings of the EMNLP 2018, pp. 1630–1642.
Pan, B., Yang, Y., Zhao, Z., Zhuang, Y., Cai, D., and He, X. (2018). “Discourse Marker Augmented Network with Reinforcement Learning for Natural Language Inference.” In Proceedings of the ACL 2018, pp. 989–999.
Peldszus, A. and Stede, M. (2013). “From Argument Diagrams to Argumentation Mining in Texts: A Survey.” International Journal of Cognitive Informatics and Natural Intelligence, 7 (1), pp. 1–31.
Peldszus, A. and Stede, M. (2015). “Joint Prediction in MST-style Discourse Parsing for Argumentation Mining.” In Proceedings of the EMNLP 2015, pp. 938–948.
Peldszus, A. and Stede, M. (2016). “An Annotated Corpus of Argumentative Microtexts.” In Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, pp. 801–815.
Pennington, J., Socher, R., and Manning, C. D. (2014). “GloVe: Global Vectors for Word Representation.” In Proceedings of the EMNLP 2014, pp. 1532–1543.
Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). “Deep Contextualized Word Representations.” In Proceedings of the NAACL-HLT
Potash, P., Romanov, A., and Rumshisky, A. (2017). “Here’s My Point: Joint Pointer Architecture for Argument Mining.” In Proceedings of the EMNLP 2017, pp. 1375–1384.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., and Webber, B. L. (2008). “The Penn Discourse TreeBank 2.0.” In Proceedings of the 6th Conference on Language Resources and Evaluation, pp. 2961–2968.

Reed, C. (2006). “Preliminary Results from an Argument Corpus.” In Bermúdez, E. M. and Miyares, L. R. (Eds.), Linguistics in the Twenty-first Century. Citeseer.

Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., and Gurevych, I. (2019). “Classification and Clustering of Arguments with Contextualized Word Embeddings.” In Proceedings of the ACL 2019, pp. 567–578.

Reisert, P., Vallejo, G., Inoue, N., Gurevych, I., and Inui, K. (2019). “An Annotation Protocol for Collecting User-Generated Counter-Arguments Using Crowdsourcing.” In Proceedings of the AIED 2019, pp. 232–236.

Sileo, D., Van-De-Cruys, T., Pradel, C., and Muller, P. (2019). “Mining Discourse Markers for Unsupervised Sentence Representation Learning.” In Proceedings of the NAACL-HLT 2019, pp. 3477–3486.

Stab, C. and Gurevych, I. (2014). “Identifying Argumentative Discourse Structures in Persuasive Essays.” In Proceedings of the EMNLP 2014, pp. 46–56.

Stab, C. and Gurevych, I. (2016). “Recognizing Insufficiently Supported Arguments in Argumentative Essays.” In Proceedings of the EACL 2016, pp. 980–990.

Stab, C. and Gurevych, I. (2017). “Parsing Argumentation Structures in Persuasive Essays.” In Proceedings of CL, Vol. 43, pp. 619–659. MIT Press.

Stab, C., Miller, T., Schiller, B., Rai, P., and Gurevych, I. (2018). “Cross-topic Argument Mining from Heterogeneous Sources.” In Proceedings of the EMNLP2018, pp. 3664–3674.

Stede, M., Afantenos, S., Peldszus, A., Asher, N., and Perret, J. (2016). “Parallel Discourse Annotations on a Corpus of Short Texts.” In Proceedings of the LREC 2016, pp. 1051–1058.

Stern, M., Andreas, J., and Klein, D. (2017). “A Minimal Span-Based Neural Constituency Parser.” In Proceedings of the ACL 2017, pp. 818–827.

Toulmin, S. E. (1958). The Uses of Argument. Cambridge Univ. Press (Cambridge).

Trautmann, D., Daxenberger, J., Stab, C., Schütze, H., and Gurevych, I. (2020). “Fine-Grained Argument Unit Recognition and Classification.” In Proceedings of the AAAI 2020, pp. 9048–9056.

Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijmm, T. A., Hirst, G., and
Stein, B. (2017). “Computational Argumentation Quality Assessment in Natural Language.” Proceedings of the EACL 2017, pp. 176–187.

Wang, W. and Chang, B. (2016). “Graph-based Dependency Parsing with Bidirectional LSTM.” In Proceedings of the ACL 2016, pp. 2306–2315.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). “Transformers: State-of-the-art Natural Language Processing.” arXiv preprint arXiv:1910.03771.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). “Xlnet: Generalized Autoregressive Pretraining for Language Understanding.” In Proceedings of the NIPS 2019, pp. 5754–5764.

付録

A 追加索引

既存研究 (Potash et al. 2017) に従い、ADU に関する以下の離散的な特徴量を用いた
- ADU 中に出現する単語の集合（出現単語に対応する次元が 1, それ以外が 0 である語彙数長のベクトル）
- ADU 中に含まれる単語の分散表現について、average, max, min プーリングを施したもの
- 論述全体の中で何番目の ADU か (one-hot-vector)
- 段落内で何番目の ADU か (one-hot-vector)
- ADU が論述文の中で、最初の段落、最後の段落、それ以外の段落のいずれかに属するか (one-hot-vector)

Potash らに従い、これらの特徴から得られた離散特徴ベクトルを全結合層によって、連続的な特徴表現に変換した。特徴表現の次元数は、LSTM, LSTM+dist モデルでは 512 次元。BoW モデルでは LSTM-Minus ベースのスパン表現の次元数と合わせるため 1536 次元にしている。また、単語の分散表現を用いた索引については、各実験設定で用いた分散表現と同じもの（例えば、XLNET の各層から得られた分散表現の重みつき和）を用いている。
表11 ハイパーパラメータの一覧

| Name                  | Value                                      |
|-----------------------|--------------------------------------------|
| 単語埋め込み表現       |                                            |
| - Glove               | 300 次元                                    |
| - ELMo                | 1024 次元                                   |
| - BERT                | 1024 次元                                   |
| - RoBERTa             | 1024 次元                                   |
| - XLNET               | 1024 次元                                   |
| BiLSTMs               | 256 次元                                    |
|                       | (LSTM モデルでは 300 次元)                   |
| ミニバッチサイズ       | 16                                         |
| 最適化手法             | Adam                                       |
| 学習率                 | 0.001                                      |
| エポック数             | 500 (MTC では 1000)                         |
| 損失関数の重み          |                                            |
| - α                   | 0.5                                        |
| - β                   | 0.25                                       |
| ドロップアウト           |                                            |
| - 出力層               | 0.5 (MTC では 0.9)                          |
| - BiLSTMs             | 0.1 (MTC では 0.9)                          |
| - 単語埋め込み層       | 0.1                                        |

B ハイパーパラメータ

表11にハイパーパラメータを示す。LSTMは1層のものを用いた。BoW, LSTM, and LSTM-distモデルをパラメータ数の観点で正当に比較するため、LSTM モデルではBiLSTMの隠れ層の次元を大きくし、ADU レベルの LSTM の層の数を 2 層にしている。BoW モデルでは、パラメータ数を増やしても性能がほとんど変わらなかったため表11の通りのハイパーパラメータを用いている。

略歴

栗林 樹生：2020年東北大学大学院情報科学研究科博士前期課程修了。現在、東北大学大学院情報科学研究科にて博士課程取得に向けて研究を進めている。2020年より日本学術振興会特別研究員（DC1）。Langsmith株式会社共同創業者。

大内 啓樹：奈良先端科学技术大学院情報科学研究科にて、2015年博士前期課程修了。2018年博士後期課程修了。2018年より理化学研究所革新知能統合研究センター特別研究員。
井之上直也：2013年東北大学大学院情報科学研究科博士後期課程修了。株式会社デンソー基礎研究所研究員、東北大学大学院情報科学研究科助教、理化学研究所客員研究員を経て、2020年よりStony Brook University研究員。

鈴木 潤：2001年から2018年まで日本電信電話株式会社コミュニケーション科学基礎研究所研究員（主任研究員／特別研究員）。2005年奈良先端科学技术大学院情報科学研究科博士後期課程修了。現在、東北大学データ駆動科学・AI教育研究センター教授。

Paul Reisert：2017年東北大学大学院情報科学研究科博士後期課程修了。2017年より理化学研究所革新戦略研究センター特別研究員。

三好 利昇：2002年京都大学理学部卒業。2007年同大学院情報科学研究科博士課程修了。博士（情報学）。同年、（株）日立製作所中央研究所入社。現在、日立製作所研究開発グループテクノロジーイノベーションセンター勤務。自然言語処理などの研究に従事。

乾 健太郎：1995年東京工業大学大学院情報理工学研究科博士課程修了。2010年より東北大学大学院情報科学研究科教授。2016年より理化学研究所AIPセンター自然言語理解チームリーダー兼任。

（2020年4月1日 受付）
（2020年7月1日 再受付）
（2020年8月19日採録）