# 複数人議論における対話的役割分類モデルの比較

荻野 奈津美[1]　姫野 拓未[2]　嶋田 和孝[3]

**概要**：本研究では，複数人議論データにおける「質問」や「主張」などの発話の対話的役割分類を行う．具体的には AMI Corpus における Unit Label の分類問題に取り組む．分類には，Bag-of-Words などの発話の言語的素性に加え，時間情報などの対話的素性を導入し，これらの素性を入力とする分類モデルを構築する．分類モデルには Support Vector Machines (SVM) と近年自然言語処理の様々なタスクで高い精度を収めている事前学習モデルの BERT，その BERT と提案素性群を組み合わせて系列として学習する BERT-CRF の 3 つを比較する．実験の結果，対話的な特徴が Unit Label の分類に有効であることが確認された一方，全体的な精度では BERT が最も良いことが確認された．また，ラベルの分布に偏りがあるという問題点に対してリサンプリング手法を適用し，その効果について実験的に検証した．

# Comparison of unit label classification models in multi-party conversation

***Abstract:*** In this study, we deal with an unit label classification task of multi-party conversation. The unit labels were defined in the AMI Corpus with Twente Argument Schema (TAS). We propose not only linguistic features, such as bag-of-words, but also dialogical features, utterance time, for the task. In this paper, we compare three classification models; Support Vector Machines with our features, BERT, and CRF with BERT and our features. The experimental result shows the effectiveness of the dialogical features. The best performance was produced by the BERT model. In addition, to solve a problem caused by biased label distribution, we also discuss some resampling methods in the experiment.

## 1. はじめに

仕事や教育の現場における意志決定手段として，様々な議論が日々行われている．議論に直接参加していない人がその議論の内容を知るには議事録が不可欠である．議事録は議論の要約であるが，議論中での各話者の主張や質問など，発話の対話的な役割を明確にし，それを構造化できれば，議論をより深く理解することが可能になる．

Renals ら [1] は複数人議論コーパスにおける発話の対話的な役割分類を行っている．Renals らは，クエスチョンマークの有無や発話の単語数，N-gram などの言語的な特徴を素性に利用し，作成した素性を組み込んだ決定木で分類を行っている．実験の結果，高い分類精度が確認され，素性の中でも特に N-gram が有効であったと述べている．しかし N-gram などの言語的特徴だけでは，話者やトピックの変化などの議論の流れを考慮することができない．議論の場においては発話内容などの言語的な特徴だけでなく，やり取りの間や誰が発言したかなどの対話的な特徴も重要である．一方で，Himeno ら [2] は複数人議論における発話間の関係の有無の推定に際して，単語の分散表現だけでは複数人議論の特徴を掴めないという問題点を解消するために，対話や議論の特徴を捉えた素性を提案している．その追加素性を SVM と BiLSTM に適用し，対話的な特徴である時間情報と発話間の距離が特に有効であることを確認している．

本論文では，複数人議論コーパスである AMI Corpus [3] に対し，Twente Argument Schema (TAS) [4] と呼ばれる議論構造に関するスキーマによってアノテーションされたデータを対象とした対話的役割分類モデルを構築する．分類タスクは，TAS で定義された 5 つの Unit Label

[1] 九州工業大学 情報工学部 知能情報工学科
Department of Artificial Intelligence, Kyushu Institute of Technology, 680–4 Kawazu, Iizuka, Fukuoka, Japan
[2] 九州工業大学大学院 情報工学府
Department of Artificial Intelligence, Kyushu Institute of Technology, 680–4 Kawazu, Iizuka, Fukuoka, Japan
[3] 九州工業大学大学院 情報工学研究院
Department of Artificial Intelligence, Kyushu Institute of Technology, 680–4 Kawazu, Iizuka, Fukuoka, Japan

に「その他」を追加した 6 種のタグを分類する問題である．提案手法では，Support Vector Machines (SVM) を基盤とした手法と自然言語処理の分野で注目を集めている事前学習モデルである BERT [5] に基づく手法，さらにその BERT を系列学習のモデルである CRF [6] に組み込む手法（BERT-CRF）の 3 つを検証する．SVM と BERT-CRF に関しては言語的特徴だけではなく，Himeno らの手法に倣い，時間情報などの対話的な特徴を捉えた素性を作成し，発話の対話的役割分類を行う．また，6 つのタグの分布が不均衡であるという問題点を踏まえ，リサンプリングモデルを適用し，その有効性について議論する．

## 2. 関連研究

近年，議論コーパスを用いた議論分析は自然言語処理において注目を集めている研究分野の 1 つであり，議論検索，議事録の自動生成など様々な目的で行われている．Chernodub ら [7] は，議論検索のためのニューラル議論マイニングフレームワークである TARGER を提案している．TARGER では BiLSTM-CNN-CRF モデルによってエッセイなどウェブ上の 3 つのデータセットに自動タグ付けを行い，入力されたテキストに対してキーワードベースの議論検索を行うことができる．

Wachsmuth ら [8] はある議論に対する反論の自動検索に関する研究を行っている．反論は反対の立場を持っているという仮説のもと，トピックの類似性及び立場の非類似性を考慮するモデルを作成している．モデルの作成においては議論内の単語や埋め込み表現の類似度関数を作成し，ランキングアプローチとして 8 つのタスクで実験を行っており，その結果，3 分の 1 の確率で最良の反論を見つけ出すことに成功している．

Lugini ら [9] は，生徒たちによる教室での議論データを作成し，ある人物や状況，特徴的な語彙などの発話の特異性に注目した議論要素のマルチタスク分類を行っている．LSTM と CNN を用いて，発話を Claim，Warrant，Evidence の 3 つに分類している．実験結果から，特異性がマルチタスク分類に貢献していることを示している．

Shang ら [10] は話者の変化に注目した BiLSTM-CRF モデルによる談話行為タグ分類を行っている．実験の結果，CRF による系列学習に話者情報を用いることで，分類の正確率が向上することを確認している．

Schulz ら [11] は BiLSTM-CRF モデルによる複数データセットでの議論要素のマルチタスク分類を行っている．この研究では分類ラベルの依存性を考慮するために CRF を用い，発話を主張，前提，非議論的の 3 つに分類している．実験の結果，訓練データ数が少ないほどマルチタスク分類が有効であることを示している．

## 3. データセット

本研究で用いる英語の複数人議論データセットの AMI Corpus 及びアノテーションスキーマの Twente Argument Schema (TAS) について説明する．

### 3.1 AMI Corpus

本研究では，議論分析に必要な情報がアノテーションされた複数人議論コーパスである AMI Corpus を用いる．AMI Corpus は 4 人 1 組による議論コーパスであり，4 人は架空の家電企業の異なる役職の社員という設定である．議論は，新しいテレビリモコンのデザインをテーマに計 4 回行われている．AMI Corpus では発話時間，話者 ID，トピックのタグ，談話行為タグの計 4 つが各発話に付与されている．談話行為タグとは，言いよどみや相槌, 相手への行動の要求などの発話特徴を表すタグである．談話行為タグの詳細を**表 1** に示す．例えば “Stall” は言いよどみを表すタグであり，話者が “Well.....”，“Hmm.....” のように口ごもっていることを示している．また，本研究では先ほど述べた 4 つ以外に，発話時間の昇順に発話の順番，Stanford CoreNLP [12] を用いて算出した 5 種類の発話の極性値を各発話に付与している．

### 3.2 Twente Argument Schema (TAS)

本研究では，AMI Corpus に対して Twente Argument Schema (TAS) と呼ばれる議論構造に関するスキーマによってアノテーションされたデータを対象とする．TAS ではノードと呼ばれる単位で様々な情報が付与されている．ノードとはある話者の発言の全体，または言いかけて辞めたり他の話者に遮られるまでの発言の一部である．本節ではノード，次節以降では発話と表現する．ノードは複数文から成り立っている場合もある．本研究で扱うのは 92 対話中の 6368 ノードである．またノードをトピックごとにまとめたものをディスカッションと呼ぶ．

TAS では各ノードに Unit Label と呼ばれるものが付与されている．Unit Label とはノード（発話）の対話的な役割を示すラベルであり，「質問」と「主張」をさらに細かく分類した計 5 種類のラベルが存在する．本研究ではこの 5 種類の他に，Unit Label が付与されていないその他のノードを指す “Other” を加える．Unit Label の詳細を**表 2** に示す．“Open” は “How do you feel about that?” のように選択肢のない自由な質問，“A/B” は “What size batteries, double A, triple A?” のように選択肢を提示される質問，“Yes/No” は “Do we have a fancy look-and-feel?’ のように “Yes” か “No” で答えられる質問，“State” は “Oh, that would be nice.” のようなはっきりとした主張，“Weak” は “Maybe we should have also a digit button.” のような自

表 1　談話行為タグの詳細
**Table 1**　Type of dialog acts.

| 談話行為タグ | タグの詳細 |
|---|---|
| Backchannel | 相槌などの発話 |
| Stall | フィラーなどの話し始めの言いよどみ |
| Fragment | 言いかけてやめた発話 |
| Inform | 話者が聞き手に情報を与える発話 |
| Elicit-inform | 話者が聞き手に情報を求める発話 |
| Suggest | 話題の提案など<br>聞き手に働きかける発話 |
| Offer | 話者自身の行動を申し出る発話 |
| Elicit-Offer-or-suggestion | 話題の提案や<br>行動の申し出を求める発話 |
| Assess | 前の発話に対して<br>評価を与える発話 |
| Comment-about<br>-Understanding | 話者自身の理解,<br>不理解を示す発話 |
| Elicit-Assessment | 自分の発話を相手が<br>理解できたか確認する発話 |
| Elicit-Comment-about<br>-Understanding | 話者自身の発話を聞き手が<br>理解できたか確認する発話 |
| Be-Positive | 聞き手との信頼関係に<br>良い影響を与える発話 |
| Be-Negative | 聞き手の気分を害する<br>冗談や攻撃的な発話 |
| Other | 上記に該当しない<br>発話意図をもつ発話 |

表 2　Unit Label の詳細
**Table 2**　Type of the Unit Labels.

| Unit Label | ラベルの詳細 |
|---|---|
| Open issue(Open) | オープンクエスチョン |
| A/B issue (A/B) | 選択肢のある質問 |
| Yes/No issue (Yes/No) | Yes/No で答えられる質問 |
| Statement(State) | 主張 |
| Weak statement (Weak) | 弱い主張 |
| Other | その他の発話 |

信がない弱い主張，そして “Other” は “Mm-hmm.” のような議論に関係がないその他のノードを示している．本論文では，ノードが与えられた場合にその Unit Label を推定する問題を対話的役割分類として考え，分類モデルを構築する．

## 4. 手法

本論文では 3 つの手法（SVM，BERT，BERT-CRF）を比較する．そのうち SVM では，発話から得られる言語的素性（4.1.1 節）と対話的素性（4.1.2 節）の 2 種類の素性群を組み合わせて利用する．BERT-CRF では，BERT が出力する埋め込み表現と SVM でも利用する言語的素性と対話的素性を組み合わせて CRF で学習する．本節ではまず 2 つの素性群について述べ，その後に各分類モデルについて説明する．

### 4.1 素性群

本節では，SVM および BERT-CRF で利用する 2 種類の素性群について説明する．

#### 4.1.1 言語的素性

f1) 発話の分散表現： Google 社の word2vec [13][*1]を用いる．各単語の分散表現を足し合わせて平均を算出することにより発話の分散表現を得る．

f2) Bag-of-Words：品詞，ストップワードによる単語数削減のほか，出現文書数 5 以下または出現文書率 10%以上の単語を除外することで次元数を削減する．

f3) 単語 N-gram：bi-gram および tri-gram を対象とし，それぞれの出現頻度に基づき，Unit Label ごとに割合を算出する．対象発話内に出現する N-gram 表現の持つ割合の総和を素性とする．すなわち，特定の Unit Label で出現しやすい N-gram が対象発話で多く出現すれば，大きな値をとる素性となる．この素性により，構文やフレーズに関する Unit Label ごとの特徴を把握することができる．

f4) 品詞 N-gram：単語 N-gram と同様に素性値を算出する．

f5) 曖昧性を表す単語の有無： “maybe”, “likely”, “probably”, “might” の 4 つの単語の有無を素性とする．これは，自信がない主張や控えめな主張である Weak にはこれらの単語が多く含まれると考えられるからである．

f6) 接続語の数：文と文を繋ぐ際に一般的に利用される単語を人手で選定し，選定した単語 “because”, “so”, “though”, “whether”, “either”, “neither”, “therefore” の 7 単語の出現数を素性とする．これは文と文を繋ぐような語は長い主張文に使われることが多いと考えられるからである．

f7) 極性値：3 節で述べたように，発話には 5 種類の極性値を付与している．これを −2 から 2 の値に置き換えて利用する．発話に複数付与されている場合は極性値の和をとる．これは，“What do you think?” のように相手の意見を引き出す質問よりも，自分の意見を主張するときの方が感情が入りやすいと考えられるためである．

f8) 単語数：発話が複数文の場合は全単語数を文数で割ることで 1 発話文あたりの単語数を求める．これは，自分の意見とその根拠を論理的に伝えようとする主張は長くなる傾向があるためである．

f9) 疑問詞の出現数：自然言語処理ツールキット NLTK [14] を用いて得られた品詞情報から，疑問詞を示す 4 種類の品詞 “WDT”, “WP”, “WP$”, “WRB” の出現数を算出し，それを素性とする．これは，Open のように，選択肢を限定せず相手の意見を引き出すような疑問文では “What” や “Why” のような疑問詞が用いられる可能性が高いと考えられるためである．

[*1] https://code.google.com/archive/p/word2vec/

f10) 前後 2 発話との単語一致度： 前後 2 発話中の同じ単語数をカウントし，自身の単語数で割ることで単語一致度を計算し，それを素性とする．例えば “Blue or Yellow?” のように，A/B では選択肢が示されるため，次の発話には選択肢の単語が含まれる可能性が高いためである．

f11) パンクチュエーションの数：発話内のパンクチュエーション（“?”，“.”，“,”）の数を計算し，素性とする．これは，言いかけて途切れる発話文を含め，主張文と疑問文がどのくらい混ざった発話かを捉えるためである．例えば主張文+疑問文や主張文+疑問文+主張文のように発話には主張文と疑問文が複雑に混ざり合っているものがあるためである．

f12) クエスチョンマーク（“?”）の出現率： 発話内のクエスチョンマーク（“?”）の数を集計し，発話内の単語数で割った値を素性とする．これにより，主張文と疑問文が混在する発話の場合，どちらのニュアンスが強いかを確かめることができると考えられる．例えば，主張文+疑問文+主張文のような発話の場合，主張の方が割合が多いことが確認できる．

#### 4.1.2 対話的素性

f13) 発話位置：3 節で述べたように各発話には ID（出現順）を付与している．これを用いて，ディスカッション内での発話位置を素性とする．議論の序盤は，主張や他の人の意見を促す質問，終盤は意見をまとめる主張やそれに賛同する相槌が多くなると考えられる．発話の出現位置が分かれば，この対話的特徴をモデルに組み込むことができるようになる．

f14) 談話行為タグ：表 1 に示す談話行為タグを素性として利用する．これは，言いよどみを指す “Stall” や言いかけて辞めた発話を指す “Fragment” は，自信がない主張や控えめな主張である Weak に付与されやすく，相槌を指す “Backchannel” や前の発話への評価を示す “Assess” は State に付与されやすいと考えられるためである．

f15) 次の発話との時間差：ある発話の終了時間と次の発話の開始時間との差を素性とする．これは， 議論が白熱しているときは短時間で主張のやり取りが続く一方，相手の主張の理解ができず質問をするときや相手からの質問への応答を考えるときは時間差が生まれると考えられるためである．

f16) 話者 ID：AMI Corpus 中の話者 ID を素性として利用する．

f17) トピックタグ：AMI Corpus で付与されている 22 種類のトピックタグを素性として利用する．

### 4.2 モデル

本研究では，3 つの分類モデルを用いて対話的役割分類を行う．以下の節では各モデルの詳細を説明する．

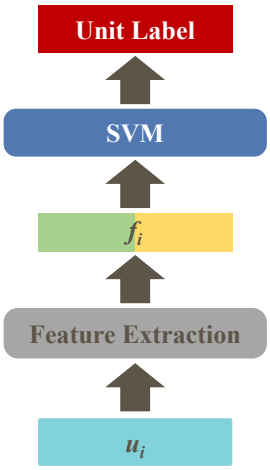


**図 1** SVM の概要図
**Fig. 1** SVM with all features.

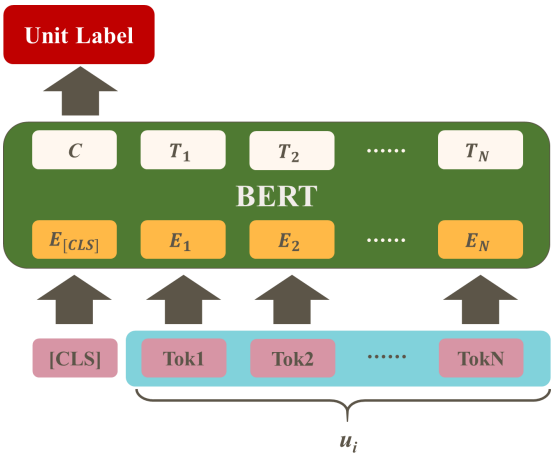


**図 2** BERT の概要図
**Fig. 2** BERT

#### 4.2.1 SVM

モデルの概要を**図 1** に示す．$u_i$ は発話であり，$u_i$ から言語的素性（4.1.1 節）と対話的素性（4.1.2 節）が抽出され $f_i$ が得られる．なお，前処理として，短縮形について “don’t” → “do not” のような書き換えを行う．

#### 4.2.2 BERT

近年自然言語処理の様々なタスクにおいて最高精度を達成している BERT をそのまま適用する．BERT の概要を**図 2** に示す．BERT は大規模データセットを用いた事前学習を行っており，本研究では学習済みのデータ[*2]を類似タスクであるシングルセンテンス分類タスク CoLA [15]，及び 6.1 節で示す訓練データで fine-tuning したものを用いる．入力は発話のみであり，4.1.1 節と 4.1.2 節で述べた素性は一切用いない．

#### 4.2.3 BERT-CRF

3 つ目は，BERT で得られる埋め込み表現と 4.1.1 節及

[*2] https://github.com/google-research/bert

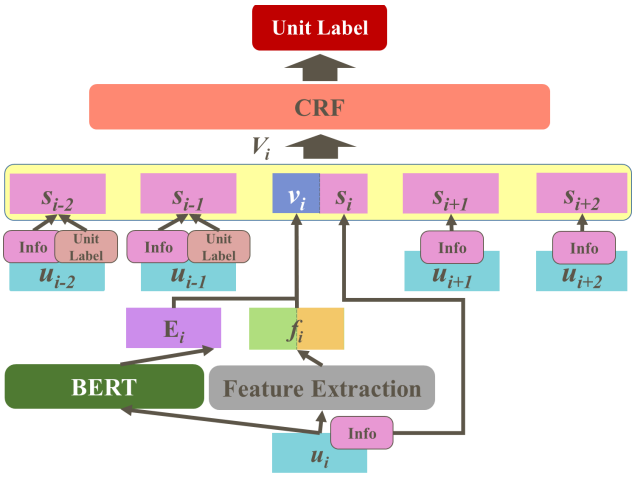


**図 3** BERT-CRF の概要図
**Fig. 3** BERT-CRF.

び 4.1.2 節の素性を統合し，系列学習器である CRF[*3]を用いるものである．モデルの概要を**図 3** に示す．このモデルでは，分類対象となる発話 $u_i$ に対して，その前後 2 発話（$u_{i-2}, u_{i-1}, u_{i+1}, u_{i+2}$）の情報も利用する．BERT からは，対象となる $u_i$ を BERT に入力した場合に得られる埋め込み表現 $E_i$ を利用する．具体的には，図 2 で分類の際に使用する [CLS] の埋め込み表現である．なお，実装は BERT における文章の埋め込み表現を獲得できるツール bert-as-service[*4]を用いる．この BERT の埋め込み表現とは別に，SVM のときと同様に $u_i$ に対して言語的素性（4.1.1 節）と対話的素性（4.1.2 節）のすべてを抽出する．これらを $E_i$ と結合させたものを $v_i$ とする．これとは別に，対象発話と前後 2 発話について，言語的素性の f7，対話的素性の f13 から f16 に相当する情報をデータセットから直接抽出し（図中の info），それぞれ $s_{i-2}$，$s_{i-1}$，$s_i$，$s_{i+1}$，$s_{i+2}$ とする．さらに $s_{i-2}$ と $s_{i-1}$ には前 2 発話の実際の Unit Label も素性として加える．この $V_i = \{s_{i-2}, s_{i-1}, v_i, s_i, s_{i+1}, s_{i+2}\}$ を CRF の入力とし，系列学習する．

## 5. データのリサンプリング

本論文でのタスクは各発話を 6 つのクラス（ラベル）に分類する問題である．それぞれのラベルに属する事例が均一であれば問題はないが，本タスクは議論に対する対話的な役割が分類対象であり，参加者同士の主張が続く場合など，議論の展開によって頻出するラベルとそうでないもの差が顕著になる．そのような場合，機械学習は多数派のラベルに寄せてしまう傾向がある．そこで，リサンプリングをし，データの偏りを軽減させる．

リサンプリング手法には，大きく分けると主に Under sampling と Over sampling がある．Under sampling は負例を減らすリサンプリング手法で，Over sampling は正例を増やす手法である．両者にはそれぞれデメリットがある．Under sampling は多数派データの削減により情報損失が生じうる．Over sampling は少数派データの拡張により過学習が起きる可能性がある．

これらの問題を解決するために新たなリサンプリング手法が提案されている．その中でも SMOTE [16] はこの問題を解決する手法の 1 つである．これは，一般的な Over sampling とは違い，今ある少数派データを複製するのではなく，K 近傍法により取ってきた少数派データ間に新たなデータを生成する手法であり，過学習を押さえることができる．

また，SMOTE の拡張版の研究が数多く行われている．Fernández ら [17] は数多くある SMOTE の拡張版モデルの概要や欠点をまとめ，マルチラベル学習や回帰などのより複雑な問題への応用の可能性について分析を行っている．さらに，ビッグデータにおいてデータのサンプリングや SMOTE 等の前処理のアプローチに注目することの重要性を説き，より新しいアプローチが課題であると言及している．

本論文では，Imbalanced-learn [18] で公開されているリサンプリングモデルのうち，SMOTE 及び SMOTE の拡張版モデル 2 種類を用いてリサンプリングを行う．用いるモデルを以下に示す．

- SMOTE
  K-近傍法で取ってきた少数派データとの間に新たな少数派データを生成する．
- ADASYN [19]
  SMOTE の拡張版の 1 つである．SMOTE より多数派データに近いデータを増やす．
- Borderline-SMOTE (B-SMOTE) [20]
  SMOTE の拡張版の 1 つである．SMOTE よりラベルの境界線付近にデータを増やす．

それぞれのモデルでリサンプリングを行ったのち，リサンプリングデータでモデルを学習する．

## 6. 実験

本節ではまず 3 つのモデルの精度比較を行う．次に，対話的素性の有効性について，SVM を対象に考察する．最後に，リサンプリングによる実験結果について議論する．

### 6.1 実験設定

3 節で述べたデータセットを，対話単位で 訓練データ：検証データ：テストデータ = 84：4：4 に分割した．各データにおける発話数を**表 3** に示す．表 3 より，State や Other のデータ数が多く，A/B のデータ数が少ないことが確認できる．

SVM のパラメータは，カーネルを rbf，カーネルパラメー

*3 https://github.com/scrapinghub/python-crfsuite.git
*4 https://github.com/hanxiao/bert-as-service

表 3　Unit Label ごとの発話数
**Table 3**　Distribution of each unit label in the dataset.

| Unit Label | 発話数 | | | |
|---|---|---|---|---|
| | Train | Develop | Test | All |
| Open | 205 | 17 | 7 | 229 |
| A/B | 61 | 4 | 2 | 67 |
| Y/N | 392 | 11 | 23 | 426 |
| State | 3538 | 102 | 206 | 3846 |
| Weak | 173 | 5 | 2 | 180 |
| Others | 1476 | 52 | 92 | 1620 |
| All | 5845 | 191 | 332 | 6368 |

タを 0.001，コストパラメータを 100 とした．BERT-CRF に用いている CRF のパラメータは L1 正則化係数を 1.0，L2 正則化係数を 0.001 とした．

### 6.2　実験：モデルの比較と素性の有効性

各モデルごとの分類結果の比較と，対話的素性の有効性の検証を行う．

#### 6.2.1　各分類モデルによる分類結果

各モデルにおける分類結果を**表 4** に示す．データ数はテストデータ中での各 Unit Label のデータ数を意味している．表 4 より，モデルごとに見ると BERT の分類結果が最も良いことが確認できる．このことから言語的な特徴はこの分類タスクにおいて有効であると推測できる．しかし，Unit Label ごとに見ると A/B 及び Weak はどのモデルでも分類できていない．全モデルでの誤分類の原因として，用いたデータセットにおいて，極めて事例数の少ないラベルがあったこと，意見の飛び交う議論の場において系列が汲み取りづらかったこと，話し言葉であり，言いかけの発話や関係のない発話も多く含まれていたことが考えられる．また，モデルごとの差が出たのは Open と Yes/No であった．そこでモデルごとにエラー分析を行う．

まず SVM のエラー分析を行う．SVM では State に関係する誤分類が多く見られ，特に State と Other の誤分類が多かった．誤分類例としては “Yeah.” や “Okay.” などの短文の発話が挙げられる．これらは特徴的な単語や N-gram を把握するのが困難な発話であり，議論に関係のない主張 Other と State の区別が人手でも付きづらい例である．また，SVM では State を Yes/No と誤分類する例は他のモデルよりも少なかったが，Yes/No を State と誤分類する例が確認された．Yes/No には主張文+ “did you?” のような付加疑問文から成り立つ例があり，SVM ではこのような文を State に分類してしまう傾向があった．さらに， A/B の誤分類の例としては “So lots of choices, what do we think?” のように，選択肢があることを暗示しているが Open と同じような特徴を持っている発話があり，データ数が少ないうえに分類が難しい発話であったと考えられる．他にも SVM における誤分類例には，主張文と疑問文が複雑に組み合わさった発話が多く見られた．本論文での提案素性群では，主張文と疑問文が組み合わさった発話を上手く分類するには不十分であったと考えられる．

次に BERT のエラー分析を行う．表 4 より，Yes/No における再現率は BERT が他のモデルを大きく上回っていることが確認できる．BERT では Yes/No を State と誤分類する数が他のモデルより少なかった．これは，他のモデルと比較すると BERT が長文の分類を得意としており，長文の出現頻度が高い State を正確に認識できていたためだと考えられる．また，SVM と同様に State と Other の誤分類が多く確認された．

最後に BERT-CRF のエラー分析を行う．表 4 より，BERT-CRF による Open の分類結果は他のモデルよりも悪いことが確認できる．BERT-CRF は State に誤分類する傾向が全モデルの中で最も強かった．CRF は系列学習を行うモデルであり，固有表現抽出などで良く用いられる．今回は対話の流れを系列であると見立て，固有表現抽出などと同様，CRF で精度が向上することを期待したが，良い結果を得ることができなかった．固有表現抽出で有効に機能する品詞などの系列と比較して，話者 ID，発話の極性，談話行為タグなどでは Unit Label の分類に効くほどの特徴がなかったものと考えられる．

#### 6.2.2　対話的素性の有無による結果

本節では対話的素性の有効性を検証する．SVM と BERT-CRF について，対話的素性をモデルに組み込んだ場合と組み込まなかった場合の結果を比較する．結果を**表 5**（SVM）と**表 6**（BERT-CRF）に示す．

表 5 より，SVM は対話的素性を用いることで精度・再現率ともに向上しており，その中でも Other の再現率が大きく向上していることが確認できる．このことから対話的素性は SVM を用いた Unit Label の分類において有効であることが確認された．SVM は State に分類する傾向があったが，対話的特徴を用いることでその傾向が緩和され，Other や Open など他のラベルにも分類していることが確認された．一方で，対話的特徴を加えることにより，相槌など短文の State を Other と誤分類する例が一部確認された．

次に BERT-CRF における結果を見る．表 6 より, BERT-CRF において対話的素性の有無にかかわらず Open や A/B，Weak の分類評価値に変化は見られなかった．しかし，Yes/No の精度を除くと対話的素性を加えることにより分類結果が改善する傾向が見られ，対話的素性は BERT-CRF を用いた Unit Label の分類においても有効であることが確認された．BERT-CRF では SVM と同じ提案素性群を組み込んでいたが，SVM も BERT-CRF も個々のラベルごとに見ると精度は 5%前後，再現率は 15%前後と同程度の評価値の向上が見られた．

以上の結果より，対話的素性が Unit Label（対話的役割）の分類で一定の効果があることが確認できた．一方

表 4　各モデルごとの分類結果
**Table 4**　Classification results by each model.

| Unit | SVM | | | BERT | | | BERT-CRF | | | データ |
|---|---|---|---|---|---|---|---|---|---|---|
| Label | 精度 | 再現率 | F 値 | 精度 | 再現率 | F 値 | 精度 | 再現率 | F 値 | 数 |
| Open | **0.80** | **0.57** | **0.67** | **0.80** | **0.57** | **0.67** | 0.00 | 0.00 | 0.00 | 7 |
| A/B | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 2 |
| Yes/No | **0.81** | 0.57 | 0.67 | 0.75 | **0.91** | **0.82** | 0.56 | 0.39 | 0.46 | 23 |
| State | 0.79 | 0.96 | 0.87 | **0.82** | **0.97** | **0.89** | 0.76 | 0.94 | 0.84 | 206 |
| Weak | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 2 |
| Other | 0.87 | **0.57** | 0.68 | **0.94** | 0.54 | **0.69** | 0.86 | 0.54 | 0.67 | 92 |
| All | 0.80 | 0.80 | 0.79 | **0.84** | **0.83** | **0.81** | 0.75 | 0.76 | 0.74 | 332 |

表 5　SVM における対話的特徴の有無による分類結果
**Table 5**　Classification results the presence or absence of interactive features in SVM.

| Unit | SVM（特徴有） | | | SVM（特徴無） | | | データ |
|---|---|---|---|---|---|---|---|
| Label | 精度 | 再現率 | F 値 | 精度 | 再現率 | F 値 | 数 |
| Open | **0.80** | **0.57** | **0.67** | 0.75 | 0.43 | 0.55 | 7 |
| A/B | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 2 |
| Yes/No | **0.81** | 0.57 | **0.67** | 0.76 | 0.57 | 0.65 | 23 |
| State | **0.79** | 0.96 | **0.87** | 0.75 | 0.96 | 0.84 | 206 |
| Weak | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 2 |
| Other | **0.87** | **0.57** | **0.68** | 0.83 | 0.42 | 0.56 | 92 |
| All | **0.80** | **0.80** | **0.79** | 0.76 | 0.76 | 0.73 | 332 |

表 6　BERT-CRF における対話的特徴の有無による分類結果の比較
**Table 6**　Classification results the presence or absence of interactive features in BERT-CRF.

| Unit | BERT-CRF（特徴有） | | | BERT-CRF（特徴無） | | | データ |
|---|---|---|---|---|---|---|---|
| Label | 精度 | 再現率 | F 値 | 精度 | 再現率 | F 値 | 数 |
| Open | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 7 |
| A/B | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 2 |
| Yes/No | 0.56 | **0.39** | **0.46** | **0.67** | 0.26 | 0.38 | 23 |
| State | **0.76** | 0.94 | **0.84** | 0.72 | **0.95** | 0.82 | 206 |
| Weak | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 2 |
| Other | **0.86** | **0.54** | **0.67** | 0.71 | 0.38 | 0.50 | 92 |
| All | **0.75** | **0.76** | **0.74** | 0.69 | 0.71 | 0.67 | 332 |

で，6.2.1 節で示したように，両モデルは BERT 単体と比較して精度が低い．より有効な対話的な特徴の分析や考察，BERT との統合による精度向上などについての議論が必要である．

### 6.3　実験：リサンプリングの有効性

6.2.1 節及び 6.2.2 節より，3 つのモデルはいずれにおいても Label ごとのデータ数の偏りの影響を受けている．A/B や Weak などの少数派データの分類精度はいずれも十分ではない．この問題は，言語的特徴だけでなく対話的特徴を用いた素性を取り入れた SVM と BERT-CRF においても同様であり，少数派データを分類することはできなかった．この問題の解決法として，本論文ではリサンプリングを導入した．リサンプリングの実験では SVM によるモデルでその有効性を検証した．

リサンプリングモデルによる結果を**表 7** に示す．なお，リサンプリングモデルのパラメータは全てデフォルト値とした．表 7 より，ほとんどすべてのラベルにおいてリサンプリング前の方が分類結果が良いことが分かる．少数派データの A/B や Weak の分類結果もリサンプリング前と変わらなかった．SVM は多数派ラベルである State に分類する傾向がある．しかし対話的素性を加えたときと同様に，リサンプリングを行うとその傾向が弱まり，他のラベルに分類する動きが見られた．そのため B-SMOTE では Other の再現率が大きく向上している．しかし少数派データに注目すると再現率に変わりはなく，予測できている数はほぼ変わらないことが確認できる．これは少数派データが少なすぎたため，K 近傍法で選択されるデータが毎回同じになってしまい，ほとんど同じようなデータが生成されていたこと，そもそも予測することが難しい発話の可能性があることが原因として考えられる．

## 7.　おわりに

本研究では，複数人議論データにおける質問と回答のペアの獲得や，話者ごとの質問及び主張の要約を目的とした議論構造把握のための発話の対話的役割分類を行った．分類対象は英語の複数人議論データである AMI Corpus の Unit Label である．

分類実験のために SVM，BERT，BERT-CRF の 3 つのモデルを構築し，分類精度を比較した．類似タスクの先行研究で示唆された対話的特徴の有効性を検証するために，言語的素性に加え，5 つの対話的素性を導入し，その有効性を検証した．類似タスクの先行研究で示唆された対話的特徴の有効性を検証するために，言語的素性に加え，5 つの対話的素性を導入し，その有効性を検証した．素性を組み込んだ SVM と BERT-CRF では，その対話的素性の有効性が実験から確認できた．一方で，全体的な精度では，近年自然言語処理分野で数々の最高精度を記録している事前学習モデルの BERT に敵わないことも確認された．また，今回扱ったようなタスクではデータの分布が不均一であることがままにある．実際，データの偏りが原因でどのモデルも分類ができていないラベルがあることが確認された．この問題を解決するため，リサンプリング手法を 3 つ

表 7　SVM におけるリサンプリングデータの分類結果
**Table 7**　Classification results of resampling data in SVM.

| Unit | No resampling | | | SMOTE | | | ADASYN | | | B-SMOTE | | | データ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Label | 精度 | 再現率 | F 値 | 精度 | 再現率 | F 値 | 精度 | 再現率 | F 値 | 精度 | 再現率 | F 値 | 数 |
| Open | **0.80** | **0.57** | **0.67** | 0.67 | **0.57** | 0.62 | 0.60 | 0.43 | 0.50 | 0.67 | **0.57** | 0.62 | 7 |
| A/B | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 2 |
| Yes/No | **0.81** | **0.57** | **0.67** | 0.67 | 0.52 | 0.59 | 0.67 | 0.52 | 0.59 | 0.71 | 0.52 | 0.60 | 23 |
| State | 0.79 | **0.96** | **0.87** | **0.83** | 0.83 | 0.83 | 0.82 | 0.70 | 0.76 | **0.83** | 0.69 | 0.76 | 206 |
| Weak | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 2 |
| Other | **0.87** | 0.57 | 0.68 | 0.72 | 0.67 | **0.70** | 0.57 | 0.72 | 0.63 | 0.56 | **0.75** | 0.64 | 92 |
| All | **0.80** | **0.80** | **0.79** | 0.77 | 0.75 | 0.76 | 0.72 | 0.68 | 0.70 | 0.73 | 0.69 | 0.70 | 332 |

導入し，リサンプリングなしのモデルと比較した．しかしながら，多くのラベルで F 値は向上せず，期待していた少数派事例の精度改善にも繋がらなかった．

今後は，今回最も分類結果の良かった BERT への効率的な対話的特徴の導入方法が精度改善の鍵になる．また，多数派データと少数派データを異なる弱分類器で分類するアンサンブルモデルを用いて，リサンプリングデータに対する分類を行い，精度向上を目指す．

**謝辞**　本研究は科研費 20K12110 の助成を受けたものです．

## 参考文献

[1] Steve, R.: AMI : Augmented Multiparty Interaction, *Proceedings of the NIST Meeting Transcription Workshop* (2004).
[2] Himeno, T. and Shimada, K.: Relation Identification Using Dialogical Features in Multi-Party Conversation, *In Proceedings of the 8th International Symposium on Applied Engineering and Sciences* (2020).
[3] McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V. et al.: The AMI Meeting Corpus, *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, p. 100 (2005).
[4] Rienks, R. and Verbree, D.: Twente Argument Schema Annotation Manual v 0.99 b, *University of Twente* (2005).
[5] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186 (2019).
[6] Lafferty, J., McCallum, A. and Pereira, F. C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data, pp. 633–723 (2001).
[7] Chernodub, A., Oliynyk, O., Heidenreich, P., Bondarenko, A., Hagen, M., Biemann, C. and Panchenko, A.: TARGER: Neural Argument Mining at Your Fingertips, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 195–200 (2019).
[8] Wachsmuth, H., Syed, S. and Stein, B.: Retrieval of the Best Counterargument without Prior Topic Knowledge, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 241–251 (2018).
[9] Lugini, L. and Litman, D.: Argument Component Classification for Classroom Discussions, *Proceedings of the 5th Workshop on Argument Mining*, pp. 57–67 (2018).
[10] Shang, G., Tixier, A., Vazirgiannis, M. and Lorré, J.-P.: Speaker-change Aware CRF for Dialogue Act Classification, *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 450–464 (2020).
[11] Schulz, C., Eger, S., Daxenberger, J., Kahse, T. and Gurevych, I.: Multi-Task Learning for Argumentation Mining in Low-Resource Settings, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 35–41 (2018).
[12] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S. and McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit, *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pp. 55–60 (2014).
[13] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* (2013).
[14] Bird, S.: NLTK: The Natural Language Toolkit, *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pp. 69–72 (2006).
[15] Warstadt, A., Singh, A. and Bowman, S. R.: Neural Network Acceptability Judgments, *Transactions of the Association for Computational Linguistics* (2019).
[16] Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P.: SMOTE: Synthetic Minority Oversampling Technique, *Journal of artificial intelligence research*, pp. 321–357 (2002).
[17] Fernández, A., Garcia, S., Herrera, F. and Chawla, N. V.: SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary, *Journal of artificial intelligence research*, pp. 863–905 (2018).
[18] Lemaître, G., Nogueira, F. and Aridas, C. K.: Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning, *The Journal of Machine Learning Research*, pp. 559–563 (2017).
[19] He, H., Bai, Y., Garcia, E. A. and Li, S.: ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning, *Institute of Electrical and Electronics Engineers international joint conference on neural networks*, pp. 1322–1328 (2008).
[20] Han, H., Wang, W.-Y. and Mao, B.-H.: Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning, *International conference on intelligent computing*, pp. 878–887 (2005).