DEIM2021 D14-5

スタンス検出タスクにおける評価方法の選定

雨宮　佑基†　　酒井　哲也†

† 早稲田大学基幹理工学研究科情報理工・情報通信専攻　〒 169–8555 東京都新宿区大久保 3-4-1
E-mail: †yukiamemiya@fuji.waseda.jp, ††tetsuyasakai@acm.org

あらまし　スタンス検出（Stance Detection）は，ニュース記事や SNS 上での投稿を対象として特定のトピックや主張に対する書き手のスタンスを特定することを目的としており，フェイクニュースの検出や噂の検証などさまざまなアプリケーションにおいて重要な要素となっている．そのため，機械学習や自然言語処理技術を活用したスタンス検出システムを構築する研究が盛んに行われている．このようなスタンス検出システムの評価には，主に Accuracy や Macro F1 が評価方法として用いられているが，どのような評価方法がスタンス検出タスクに適しているかについて明確な議論は行われていない．さらに，スタンス検出タスクでは順序的なクラス（Ordinal Class）を扱うことが多いのにもかかわらず，Ordinal Class を区別できる評価方法が用いられた研究が少ないという課題もある．そこで本研究では，最適な評価方法を選定する際の基準としてシステムランキングの類似度と順位安定性，Ordinal Class の区別という 3 つの観点から 9 つの評価方法に対して比較実験を行い，スタンス検出タスクに最も適した評価方法を選定した．

キーワード　評価方法, 評価指標, スタンス検出, Fake News Challenge

1. はじめに

スタンス検出（Stance Detection）とは，ニュース記事や SNS 上での投稿といった様々な手段で発せられた主張における，その書き手のスタンスを自動的に特定することである．スタンスとは，ターゲット（特定の人物やトピック，出来事）に対する書き手の見解・立場を意味し，ターゲットに対して「賛成」か「反対」か「どちらでもない」かのいずれかを表明するのが一般的である．例えば，図 1 に示すように，インターネットを介して行われる遠隔授業である「オンライン授業」というターゲットに対し，「通学時間も交通費もかからないからオンライン授業って効率的だよね.」という主張が SNS 上で投稿されたケースを想定する．この例では発せられた主張がターゲットに対して同意を示しているため，スタンス検出によって特定される正しいスタンスは「賛成」となる．このようなスタンス検出は，フェイクニュースの検出や噂の検証などのアプリケーションにおける重要な要素となっている [1]．そのため，2016 年に開催された SemEval-2016 Task 6 でのツイートに対するスタンス検出タスク (注1)や，2017 年に開催された Fake News Challenge Stage 1 でのニュース記事に対するスタンス検出タスク (注2)などのように，機械学習や自然言語処理技術を活用しスタンス検出技術の向上を目的とした研究活動が世界中に広がっている [2]．

スタンス検出タスクは，主張とターゲットを入力として書き手のスタンスを表すクラスラベルを求める分類問題として設定されるため，タスクの参加者や研究者によって構築された分類システムは，分類問題の評価に一般的に用いられる評価方法である Accuracy や Macro F1 によって評価されることが多い．一方で，分類問題の評価に利用される評価方法は他にも多く存在するが，それらをスタンス検出タスクの評価に適用した研究はほとんどない．しかし，一般的にタスクにおける評価方法は，複数の評価方法を比較したうえで最も適切なものが選定されるべきである．さらに，スタンス検出タスクでは Ordinal Class，つまり順序関係があるクラスのラベルを扱うことが多いのにもかかわらず，一般的に利用される Accuracy や Macro F1 は評価の際に Ordinal Class を区別できていない．十分に性質が考慮されていない評価方法を用いて評価することは，目的のタスクに対する分類システムの最適化とはかけ離れた結果を生み出し，誤った結論が導かれる危険性があるため，評価方法の性質を考慮したうえで最も適切な評価方法を選定することが非常に重要である [3]．

そこで本研究では，スタンス検出タスクにおける最適な評価方法を選定するため，分類問題の評価で用いられる 9 つの評価方法に対して比較実験を行う．最適な評価方法を選定する際の基準は目的とするタスクによって異なるが，本研究ではシステムランキングの類似度と順位安定性，Ordinal Class の区別という 3 つの観点から評価方法を比較する．

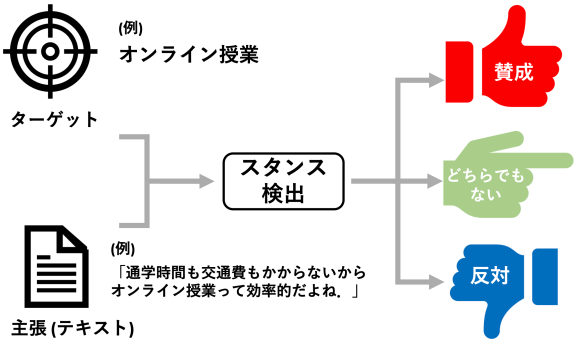

図 1　スタンス検出の例

(注1)：https://alt.qcri.org/semeval2016/task6/
(注2)：http://www.fakenewschallenge.org/

2. 関 連 研 究

Schiller ら [4] の研究で示されているように，既存のスタンス検出タスクにおける公式な評価方法としては Accuracy や Macro F1 などが用いられている．その中でもスタンス検出に焦点を当てた shared task として広く知られている SemEval-2016 Task 6，SemEval-2017 Task 8，Fake News Challenge Stage 1 という 3 つのタスクにおける評価方法に関する先行研究を以下に示す．

2016 年に開催された SemEval-2016 Task 6 では，ツイート本文とターゲットが与えられ，ツイートを投稿した人がターゲットに対して賛成（*Favor*）の立場なのか，反対（*Against*）の立場なのか，どちらでもないか（*Neither*）を分類するというタスクが課され，*Favor* クラスの F1 スコアと *Against* クラスの F1 スコアの平均によってスタンス検出システムが評価された [5]．また，Zarrella ら [6] や Wei ら [7] のように，このタスクに取り組んだ研究者のほとんどが上記の評価方法を利用していた．

2017 年に開催された SemEval-2017 Task 8 のサブタスク A [8] では，噂を発信したツイートが与えられたとき，その噂について議論している会話スレッド内のツイートが，発信された噂を支持しているか（*Support*），否定しているか（*Deny*），質問しているか（*Query*），コメントしているか（*Comment*）のいずれかに分類するというタスクが設定された．このタスクのために用意されたデータセットには大きなクラス分布の偏りが見られたが，評価方法としては Accuracy が用いられた．そのため，SemEval-2017 Task 8 のサブタスク A をさらに拡張したタスクである SemEval-2019 Task 7 のサブタスク A [9] ではクラス分布の偏りを考慮し，Macro F1 によってスタンス検出システムが評価された．

2017 年に開催された Fake News Challenge Stage 1（FNC-1） [10] では，与えられたニュース記事の見出しと本文のペアについて *Agree*, *Disagree*, *Discuss*, *Unrelated* の 4 つのスタンスから適切なものを分類するというタスクに対し，重み付けされた 2 段階のスコア付け手法が公式評価方法として用いられた．まず，見出しと本文の内容の関係が *Related*（*Agree*, *Disagree*, *Discuss*）か *Unrelated* かを正しく分類することができれば，分類システムのスコアに 0.25 ポイントが加算される．さらに，*Related* に分類されたペアについて *Agree*, *Disagree*, *Discuss* のいずれかに正しく分類することができれば，分類システムのスコアにさらに 0.75 ポイントが加算される．FNC-1 のスタンス検出タスクに対し，独自のモデルやアーキテクチャを組み込んだスタンス検出システムを構築し，精度改善を目的とした研究の多くは，前述の公式評価方法を用いてシステムを評価していた [11–18]．また，分類問題の評価指標として最も一般的である Accuracy を用いた先行研究も多く，それらはクラスごとの Accuracy やクラス全体の Accuracy による評価値を提示していた [14–19]．一方で，FNC-1 の公式評価方法や Accuracy による評価では偏ったデータ分布を考慮できないため，Hanselowski ら [13] や Slovikovskaya ら [18]，Umer ら [19] はクラスごとの F1 や Macro F1 を用いた評価方法を提案した．しかし，FNC-1 の公式評価方法や Accuracy，F1 は誤分類したときの深刻度を考慮することができない．例えば，正解が *Agree* であるペアを *Discuss* と誤分類したときと，*Agree* であるペアを *Disagree* と誤分類したときの評価値は同じになってしまう．しかし，本来は後者に大きいペナルティが与えられるべきであり，この問題に対処するために雨宮ら [20] は Weighted Cohen's Kappa による分類システムの評価を行った．

このように，既存のスタンス検出タスクでは評価方法として Accuracy や Macro F1 が選択されることが多かった．しかし，これらの評価方法がスタンス検出タスクに対して最も適しているかどうかについて検証した研究は，我々の知る限り存在しない．一般に，評価方法を評価するのは非常にチャレンジングな課題であり，評価の際に考慮すべき観点がタスクごとに異なるため，どの評価方法が良い評価方法であるかを明言することは難しい．しかし，Sakai ら [21] の研究のように，複数の評価方法を比較し各評価方法の特性を明らかにすることは，研究コミュニティ全体の方向性を定める評価方法の選定に極めて有用である．そこで，本研究ではスタンス検出タスクにおける最適な評価方法を選定するため，分類問題の評価で用いられる 9 つの評価方法に対して比較実験を行う．

3. 評価方法の定義

本研究では 9 つの評価方法を比較し，スタンス検出タスクにおいて最も適切な評価方法を選定することを目的としている．

表 1　各評価方法の定義に使われる変数

変数	説明
C	クラスラベルの集合 （例：{*Favor*, *Neither*, *Against*}）
i, j （$i, j \in C$）	クラスラベル（例：*Favor*）
c_{ij}	正解クラスが j であり，システムが クラス i と予測したアイテム数
N（$= \sum_j \sum_i c_{ij}$）	全アイテム数
$c_{i\bullet}$（$= \sum_j c_{ij}$）	システムがクラス i と予測したアイテムの総数
$c_{\bullet j}$（$= \sum_i c_{ij}$）	正解クラスが j であるアイテムの総数
$Prec_j$	適合率 (*Precision*) $c_{j\bullet} > 0$ のとき，$Prec_j = c_{jj}/c_{j\bullet}$
Rec_j	再現率 (*Recall*) $c_{j\bullet} > 0$ のとき，$Rec_j = c_{jj}/c_{\bullet j}$
$f1(p, r)$	F1 値 (*F1-score*) 適合率と再現率の調和平均で計算される
$Prec^M$	各クラスの適合率の平均 $Prec^M = \sum_{j \in C} Prec_j / \lvert C \rvert$ で計算される
Rec^M	各クラスの再現率の平均 $Rec^M = \sum_{j \in C} Rec_j / \lvert C \rvert$ で計算される
n_i	正解クラスが i であるアイテム数とシステムが クラス i と予測したアイテム数の総和
O_{ij}	クラス i，j に対する実際に観測された一致度 $O_{ij} = \sum_u n_i(u) n_j(u)$ で計算される
E_{ij}	クラス i，j に対する偶然による一致度 $E_{ij} = n_i n_j / (2N - 1)$ で計算される

そのため本節では Sakai [22] の定式化に従い，比較実験で使用する 9 つの評価方法を紹介する．まず，各評価方法の定義を表すうえで必要となる変数を表 1 にまとめる．

まず，既存のスタンス検出タスクで用いられることが多かった Accuracy や Macro F1 は以下のように定式化できる．ただし，Macro F1 には 2 つの異なる計算式が存在し，それらの計算式で算出された評価値によって分類システムを評価した際，2 つのシステムランキングが異なる可能性があると Opitz ら [23] は述べている．そのため，本研究では Macro F1 の 2 つの計算式を別の評価方法と捉え，各クラスの F1 値を計算した後にそれらをクラス数で平均する Macro F1 を $\mathcal{F}_1$，各クラスの適合率と再現率を平均した値を用いて調和平均を計算する Macro F1 を $\mathbb{F}_1$ と定義する．なお，$\mathcal{F}_1$ と $\mathbb{F}_1$ の表記の仕方は Opitz らの研究を参考にした．

$$Accuracy = \frac{\sum_{j \in C} c_{jj}}{N}$$

$$\mathcal{F}_1 = \frac{1}{|C|} \sum_{j \in C} f1\left(Prec_j, Rec_j\right)$$

$$\mathbb{F}_1 = f1\left(Prec^M, Rec^M\right)$$

また，分類システムの精度を評価する尺度として Kappa がよく知られている．Kappa は，たとえ Accuracy による評価値が同じである 2 つのシステムであっても異なる評価値を算出するため，その 2 つのシステムの精度を比較することができる．さらに Kappa は，分類システムによって予測されたクラスと正解クラスの偶然による一致を考慮することができるため，データセット内のデータサンプルに左右されずに複数のシステムを比較することができる [24]．なお，スタンス検出タスクでは順序的なクラスラベルを使用することが多く，予測されたクラスと正解クラスの不一致度合いを考慮する必要があるため，本研究では Kappa に各クラス対の重みを導入した Weighted Kappa を利用する．ここで，分類システムによって予測されたクラスと正解クラスが独立であるとき，それらの偶然による一致は $e_{ij} = c_{i\bullet} c_{\bullet j}/N$ と表され，誤分類したときのペナルティを事前に定義された重み $w_{ij} = |i-j|$ とすると，Weighted Kappa は以下のように定式化される．

$$\kappa = 1 - \frac{\sum_{j \in C} \sum_{i \in C} w_{ij} c_{ij}}{\sum_{j \in C} \sum_{i \in C} w_{ij} e_{ij}}$$

前述の通り，スタンス検出タスクの分類対象は順序的なクラス（Ordinal Class）であることが多く，このようなクラスを扱う際には Ordinal Class の性質を考慮した評価を行うことが必要である．例えば，SemEval-2016 Task6 では各ツイートに対し *Favor*，*Against*，*Neither* という 3 つの順序的なクラスのラベルが付与された．このとき，正解のラベルが *Favor* であるツイートに対して *Against* と誤分類することは，*Neither* と誤分類することよりも深刻であり，より多くの誤分類のペナルティが与えられるべきである．そこで，本研究では Ordinal Class を区別できる評価方法も研究対象とし，Ordinal Class を扱う分類問題における複数の評価方法を比較し各評価方法の性質を明らかにした Sakai [22] の研究で紹介された，Mean Absolute Error（MAE），Closeness Evaluation Measure（CEM），Krippendorff's α を使用する．MAE には，クラスラベルのデータ分布の偏りを考慮する MAE^M と，偏りを考慮しない MAE^μ の 2 種類があり，それらは以下のように定式化される．

$$MAE^M = \frac{1}{|C|} \sum_{j \in C} \frac{\sum_{i \in C} |i-j| c_{ij}}{c_{\bullet j}}$$

$$MAE^\mu = \frac{\sum_{j \in C} \sum_{i \in C} |i-j| c_{ij}}{N}$$

続いて，Ordinal Class を扱う分類問題で使用される評価指標が満たすべき 3 つの性質，すなわち *Ordinal Invariance*，*Ordinal Monotonicity*，*Imbalance* を満たしており，クラスラベル間の近さを考慮した評価指標である CEM^{ORD} [25] は以下のように定式化される．

$$CEM^{ORD} = \frac{\sum_{j \in C} \sum_{i \in C} prox_{ij} c_{ij}}{\sum_{j \in C} prox_{jj} c_\bullet j}$$

ここで，

$$K_{ij} = \begin{cases} c_{\bullet i}/2 + \sum_{l=i+1}^{j} c_{\bullet l} & (i \leq j) \\ c_{\bullet i}/2 + \sum_{l=j}^{i-1} c_{\bullet l} & (i > j) \end{cases}$$

とすると，Sakai に倣って CEM^{ORD} の定義中に現れる $prox_{ij}$ は以下のように表される．

$$prox_{ij} = -\log_2 \left(\max\left\{0.5, K_{ij}\right\}/N\right)$$

最後に，Krippendorff's α の定義を示す．本研究では，対象となるクラスラベルが順序的である場合（α^{ORD}）と間隔的である場合（α^{INT}）の 2 種類の Krippendorff's α を異なる評価方法とみなし，それぞれを以下のように定式化する．

$$\alpha^{ORD} = 1 - \frac{\sum_i \sum_{j>i} O_{ij} \left(\sum_{k=i}^{j} n_k - \frac{n_i + n_j}{2}\right)^2}{\sum_i \sum_{j>i} E_{ij} \left(\sum_{k=i}^{j} n_k - \frac{n_i + n_j}{2}\right)^2}$$

$$\alpha^{INT} = 1 - \frac{\sum_i \sum_{j>i} O_{ij} |i-j|^2}{\sum_i \sum_{j>i} E_{ij} |i-j|^2}$$

4. データセット・分類システム

本研究では，スタンス検出タスクの一例として Fake News Challenge Stage 1（FNC-1）のデータセットと，FNC-1 に対して構築された分類システムを用いて比較実験を行う．FNC-1 を選んだのは，FNC-1 のデータセットは十分な量のデータを含んでおり，FNC-1 はコンペティション開催期間中だけでなく開催後も多くの研究者によって分類システムの研究が進められる活発なタスクであるためである．さらに，他のスタンス検出タスクと比べて構築された分類システムのソースコードが公開されていることが多く，本来のタスクから条件を変えた実験ができるという利点も FNC-1 を選んだ大きな理由の一つである．

まず FNC-1 のデータセットについて説明する．本研究では実

際に FNC-1 で提供されたデータセットを使用し，訓練用データは（見出し，本文，スタンス）の三つ組，テスト用データは（見出し，本文）のペアから構成される．スタンス検出タスクの入力は（見出し，本文）であり，スタンスが出力となる．例えば，訓練用データは（*“Armed U.S. drones spotted flying over Syria in possible hunt for ISIS leader”*，*“American drones are being flown over Raqqa, Syria.”*，*Agree*）のような形になっている．スタンスは *Agree*，*Disagree*，*Discuss*，*Unrelated* という 4 種類のクラスラベルが用意された．ただし，本研究では各評価方法が Ordinal Class を区別できるかという点にも注目するため，クラスラベルは順序的である必要がある．しかし，この 4 種類のスタンスの中で *Unrelated* は他のスタンスと順序関係にあるとは言い難い．そこで，本研究では *Unrelated* クラスを除いた 3 クラスの分類問題を扱う．ここで，本研究で使用したデータセットのデータ分布を図 2 に示す．図 2 からわかるように，訓練用データセットとテスト用データセットはともにデータ分布の偏りが見られ，*Discuss* クラスのデータが最も多く，*Disagree* クラスのデータが最も少ない．また，訓練用データセットとテスト用データセット間のデータ分布が類似していることがわかる．

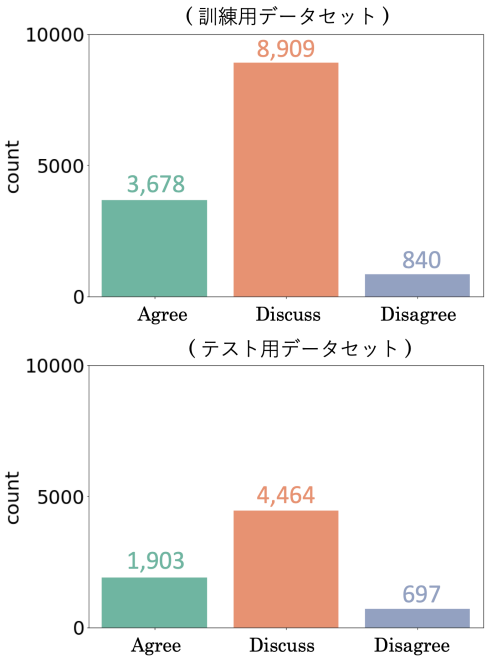

図 2　訓練用・テスト用データセットのデータ分布

続いて，本研究で使用した分類システムについて説明する．FNC-1 のスタンス検出タスクに対して，コンペティション参加者やコンペティション終了後にさらなる分類システムの精度向上に励んだ研究者によって様々な分類システムが提案されてきた．本研究では，それらの中でソースコードが公開され再現できる分類システムを 12 種類，実験で使用する分類システムの多様性を高める目的で自作した簡単な分類システムを 2 種類，合わせて 14 種類の分類システムを用意した．なお，再現した 12 種類の分類システムは本来 {*Agree*，*Disagree*，*Discuss*，*Unrelated*} の 4 クラス分類を目的として構築されているため，*Unrelated* クラスを除いた 3 クラス分類を行うように変更した．

まず，本研究では FNC-1 の主催者によって構築されたベースラインとなる分類システム (注3) を使用し，以後これを *baseline* と表記する．さらに，FNC-1 のコンペティションで 1 位を獲得した SOLAT in the SWEN チームによって公開された分類システム (注4) も使用する．このチームは Convolutional Neural Network（CNN）モデルと勾配ブースティング決定木モデルをアンサンブルしたモデルを最終成果物としているため，本研究ではこの CNN モデルを *TalosCNN*，勾配ブースティング決定木を *TalosTree*，これらのアンサンブルモデルを *TalosEnsemble* と表し，別の分類システムとして実験する．同様に，FNC-1 で 2 位を獲得した Athene チーム (注5)，3 位を獲得した UCL Machine Reading チーム (注6)，4 位を獲得した Chips Ahoy!チーム (注7) によって構築された分類システムはそれぞれ *Athene*，*UCLMR*，*Chips* と表される．

一方で，FNC-1 のコンペティション後に行われた研究では Hanselowski らや Slovikovskaya らにより構築された分類システムが公開されている (注8) (注9)．Hanselowski らは特徴量選択を修正することで *Athene* を改良した分類システム（*featMLP*）と，アブレーションテストで最も有効だった特徴量セットと Long Short Term Memory ネットワークを組み合わせた分類システム（*stackLSTM*）を考案した．また，Slovikovskaya らはラベルなしの大規模コーパスから事前学習した汎用言語モデルである BERT や XLNet，RoBERTa を FNC-1 のデータでファインチューニングした分類システム（それぞれ *BERT*，*XLNet*，*RoBERTa* と表す）を構築した．

さらに，訓練用データセットにおける最頻クラスラベル，つまり *Discuss* と常に予測する分類システム（*majority*）と {*Agree*，*Disagree*，*Discuss*} の中からランダムに 1 つ選んで予測する分類システム（*random*）の 2 種類の簡単な自作の分類システムを実装した．

5. 評価実験

本研究では，3. 節で定義した 9 つの評価方法をシステムランキングの類似度，順位安定性，Ordinal Class の区別という 3 つの観点から比較する．そこで，これら 3 つの比較実験の手法と実験結果，考察を以下に示す．なお，本実験では *Unrelated* クラスを除いた 3 クラス分類を扱うため，FNC-1 の公式評価方法は 3. 節で定義した Accuracy と同義になる．

5.1 システムランキングの類似度

どの評価方法同士が類似したシステムランキングを出力し，どの評価方法同士が異なったシステムランキングを出力するかを知ることは各評価方法の性質の理解につながると考えられる．そこで，本研究では評価方法同士のシステムランキングの類似

(注3)：https://github.com/FakeNewsChallenge/fnc-1-baseline
(注4)：https://github.com/Cisco-Talos/fnc-1
(注5)：https://github.com/hanselowski/athene_system
(注6)：https://github.com/uclnlp/fakenewschallenge
(注7)：https://github.com/shangjingbo1226/fnc-1
(注8)：https://github.com/AIPHES/coling2018-fake-news-challenge
(注9)：https://gist.github.com/vslovik

表 2　Kendall の τ による評価方法同士のシステムランキングの類似度．
類似度の高さを次のように色で表す．($\tau < 0.3$，$0.3 \leq \tau < 0.6$，$0.6 \leq \tau$)

	α^{INT}	MAE^{M}	MAE^{μ}	CEM^{ORD}	κ	$Accuracy$	$\mathcal{F}_1$	$\mathbb{F}_1$
α^{ORD}	0.5164	0.3406	0.3626	0.4945	0.6263	0.6043	0.7142	0.6923
α^{INT}	-	0.2527	0.6263	0.5384	0.5824	0.4285	0.4505	0.6483
MAE^{M}	-	-	0.5384	0.3186	0.1868	0.2087	0.2747	0.2527
MAE^{μ}	-	-	-	0.6483	0.3406	0.4965	0.2527	0.4065
CEM^{ORD}	-	-	-	-	0.6043	0.5824	0.3846	0.5824
κ	-	-	-	-	-	0.5384	0.6043	0.7142
$Accuracy$	-	-	-	-	-	-	0.4945	0.5604
$\mathcal{F}_1$	-	-	-	-	-	-	-	0.5384

図 3　*BERT*，*TalosTree*，*TalosCNN*，*baseline* による分類結果をまとめた Confusion Matrix
（縦軸が分類システムによる予測ラベル，横軸が正解ラベルを表す）

度を調査するため，Sakai [22] の研究を参考に以下のような手順で実験を行った．

（1）　評価方法ごとに，4. 節で述べたテスト用データセットに対して 14 種類の分類システムの評価値を求める．
（2）　評価方法ごとに，(1) の評価値をもとにシステムランキングを作成する．
（3）　(2) で作成された，9 つの評価方法に対応するシステムランキングの中から任意にペアを作成する．
（4）　すべてのペアに対して Kendall の順位相関係数（τ）を算出する．

なお，τ は $-1 \leq \tau \leq 1$ の範囲にあり，値が 1 に近いほど 2 つのシステムランキングが類似していることを意味する．

上記の実験を行った結果を表 2 に示す．表 2 から，システムランキングの類似度に従って 9 つの評価方法を 3 つのグループに分けることができる．まず，グループ A は α^{ORD} と κ，*Accuracy*，$\mathcal{F}_1$，$\mathbb{F}_1$ の 5 つから構成され，グループ B は MAE^{μ} と α^{INT} の 2 つから構成される．CEM^{ORD} はグループ A の評価方法とグループ B の評価方法の両方のシステムランキングに類似しているため，グループ A とグループ B の中間に位置づけられる．一方で，MAE^{M} はグループ A とグループ B のいずれの評価方法とも異なるシステムランキングを示すため，グループ C に分けられる．

上記 3 グループの特性について分析すると，グループ A に属する評価方法のシステムランキングとグループ B・C に属する評価方法のシステムランキングには以下のような 2 つの違いが見られた．例えば，グループ A に属する評価方法は *TalosTree* よりも *BERT* に対して高い評価値を出力する傾向にあり，グループ B・C に属する評価方法はその逆の傾向を示した．さらに，グループ A に属する評価方法は *baseline* よりも *TalosCNN* を高く評価する一方で，グループ B・C に属する評価方法は *TalosCNN* よりも *baseline* を高く評価していた．これらのシステムランキングの違いは，*Agree* クラスと *Disagree* クラスを誤分類した件数の違いによって生じたと考えられる．ここで，図 3 に表す Confusion Matrix からわかるように，*BERT* の *Agree* ⇄ *Disagree* の誤分類は $(397 + 103) = 500$ 件であるのに対し，*TalosTree* の *Agree* ⇄ *Disagree* の誤分類は $(187 + 0) = 187$ 件と少ない．また，*TalosCNN* の *Agree* ⇄ *Disagree* の誤分類は $(525 + 65) = 590$ 件であるのに対し，*baseline* の *Agree* ⇄ *Disagree* の誤分類はわずか $(78 + 40) = 118$ 件である．このように，グループ B・C に属する評価方法は *Agree* ⇄ *Disagree* を誤分類した件数が少ない分類システムに高い評価値を与える傾向にあり，これは評価の際に *Agree*，*Discuss*，*Disagree* の順序関係を考慮できるという α^{INT} や MAE^{M}，MAE^{μ} の性質と一致している．

5.2　順位安定性

角森ら [26] の研究では，最適な評価尺度を選定するための基準として順位安定性の観点から評価方法を比較している．ここで順位安定性とは，異なるデータセットを使用して分類システムを評価したときにシステムランキングがどれだけ変動するかということを意味し，適切な評価方法はデータセットが異なる場合でも同じシステムランキングを提示できると考えられる．さらに順位安定性について，System Ranking Consistency という形で Sakai [22] の研究においても同様の実験が行われている．本研究では角森らと Sakai の研究に倣い，以下のような手順で各評価方法の順位安定性を定量化する実験を行う．

（1）　4. 節で述べたテスト用データセットを，サブセット X とサブセット Y の 2 つにランダムに分割する．(なお，サブセット X とサブセット Y は同様のデータ分布を示した．)

表 3　順位安定性の実験結果

評価方法	Kendall の τ
κ	0.7811
$\mathcal{F}_1$	0.7388
CEM^{ORD}	0.6623
$\mathbb{F}_1$	0.6602
MAE^{M}	0.6217
$Accuracy$	0.6134
α^{INT}	0.5882
MAE^{μ}	0.5743
α^{ORD}	0.4727

（2） 9 つの評価方法の中から 1 つを選択し，その評価方法によってサブセット X に対する 14 種類の分類システムの評価値を求め，その評価値をもとにシステムランキングを作成する．

（3） （2）で選択した評価方法によってサブセット Y に対する 14 種類の分類システムの評価値を求め，その評価値をもとにシステムランキングを作成する．

（4） （2）と（3）で得られた 2 つのシステムランキングについて，Kendall の τ の値を算出する．

（5） （2）～（4）までの手順を 1,000 回繰り返し試行し，τ の平均値を計算する．

（6） （2）～（5）までの手順を，9 つの評価方法に対して実行する．

上記の実験を行った結果を表 3 に示す．表 3 から，κ や $\mathcal{F}_1$ は Kendall の τ の値が他の評価方法に比べて高く，異なるテスト用データセットを使用してもシステムランキングが大きく変わらないことがわかる．一方で，α^{ORD} や MAE^{μ} は τ の値が低く，テスト用データセットが異なるとシステムランキングも変動してしまうことがわかる．つまり，分類システムのランキングの安定性という観点では，評価方法として κ を利用することが最も適切であると言える．ここで，τ の値が最も高い κ と最も低い α^{ORD} によるサブセット X とサブセット Y のシステムランキング例をそれぞれ表 4 と表 5 に示す．この 2 つの評価方法によるシステムランキングの変動具合を分析することで，評価方法間でシステムランキングの安定性に違いが生じる要因を考察する．具体的には，14 種類の分類システムのランキング全体に対してではなく，システムランキングの上位（1 位～5 位），中位（6 位～10 位），下位（11 位～14 位）に分けてそれぞれの τ の値を算出した．その結果，κ における上位，中位，下位の τ の値はそれぞれ 0.8000，0.4000，1.0000 であり，α^{ORD} における上位，中位，下位の τ の値はそれぞれ 0.4000，0.2000，1.0000 であった．つまり，κ は α^{ORD} に比べて上位と中位のシステムランキングが安定しているため，全体として高い順位安定性を示すことができたと考えられる．

5.3　Ordinal Class の区別

3. 節で述べた通り，スタンス検出タスクでは Ordinal Class のラベルを扱うことが多く，本研究に関しては *Agree*，*Discuss*，*Disagree* の 3 クラスの順序を正しく区別できる評価方法を用いることが適切であると考えられる．そこで，本研究では角森ら [26] の研究から着想を得て，各評価方法がどの程度 Ordinal Class を区別できているかを定量化する実験を行う．角森らの研究では中間値のラベルが適切に扱われているかどうかを検討するため，複数のラベルを単一ラベルとみなして評価する評価尺度を使用しており，本研究ではこの手法に倣い 3 つのクラスラベルのうち 2 つのクラスを同一クラスとみなして各評価方法の評価値を算出する．つまり，すべての分類システムに対して *Agree* クラスと *Discuss* クラス，*Disagree* クラスと *Discuss* クラス，*Agree* クラスと *Disagree* クラスをそれぞれ単一ラベルとみなした 3 パターンの 2 クラス分類に対する評価値が出力され，その評価値によって分類システムのランキングが作成される．そして，この 2 クラス分類に対するシステムランキングと本来の 3 クラス分類に対するシステムランキングがどの程度類似しているかを Kendall の τ によって測定する．このとき，別のクラスとして扱われるべきである 2 つのクラスを単一クラスとみなしているため，Ordinal Class を正しく区別できる評価方法であれば，3 クラス分類に対するシステムランキングと 2 クラス分類に対するシステムランキングは大きく異なると

表 4　κ のシステムランキングの例（$\tau = 0.8242$）

サブセット X		サブセット Y	
システム	評価値	システム	評価値
RoBERTa	0.6304	*RoBERTa*	0.5926
XLNet	0.5542	*XLNet*	0.5455
BERT	0.3706	*BERT*	0.3392
Athene	0.3144	*TalosEnsemble*	0.3117
featMLP	0.3102	*UCLMR*	0.2854
UCLMR	0.2998	*Athene*	0.2848
TalosEnsemble	0.2973	*TalosTree*	0.2828
TalosTree	0.2615	*featMLP*	0.2771
Chips	0.2404	*Chips*	0.2127
stackLSTM	0.2248	*stackLSTM*	0.1982
TalosCNN	0.1687	*TalosCNN*	0.1853
baseline	0.0640	*baseline*	0.0399
majority	0.0000	*majority*	0.0000
random	-0.0118	*random*	-0.0164

表 5　α^{ORD} のシステムランキングの例（$\tau = 0.4945$）

サブセット X		サブセット Y	
システム	評価値	システム	評価値
RoBERTa	0.5903	*RoBERTa*	0.5760
XLNet	0.5364	*XLNet*	0.4948
BERT	0.2860	*TalosTree*	0.2594
Athene	0.2766	*Athene*	0.2512
TalosTree	0.2704	*UCLMR*	0.2507
featMLP	0.2702	*BERT*	0.2468
UCLMR	0.2637	*TalosEnsemble*	0.2448
TalosEnsemble	0.2632	*featMLP*	0.2279
Chips	0.1753	*stackLSTM*	0.1822
stackLSTM	0.1614	*Chips*	0.1612
TalosCNN	0.0701	*TalosCNN*	0.0414
baseline	0.0156	*baseline*	0.0204
random	-0.0273	*random*	-0.0325
majority	-0.0527	*majority*	-0.0434

表 6　Ordinal Class の実験結果

評価方法	*Agree*+*Discuss*	*Disagree*+*Discuss*	*Agree*+*Disagree*	平均
α^{ORD}	0.1428	0.7142	0.4065	0.4211
α^{INT}	0.5384	0.5384	0.1428	0.4065
MAE^{M}	0.1428	0.0989	0.1648	0.1721
MAE^{μ}	0.1648	0.6923	0.2967	0.3919
CEM^{ORD}	0.2747	0.4945	0.0989	0.2893
κ	0.4285	0.8901	0.3186	0.5457
Accuracy	0.2747	0.4505	0.5384	0.4212
$\mathcal{F}_1$	0.3846	0.7362	0.4725	0.5313
$\mathbb{F}_1$	0.2307	0.6263	0.4285	0.4285

考えられる．そのため，Kendall の τ の値が低い評価方法ほど Ordinal Class を区別して評価することができると言える．ここで，具体的な実験手順を以下に示す．

（1） 4. 節で述べたテスト用データセットの 3 つのクラスラベルのうち 2 つを単一クラスとみなした 2 クラス分類用のテスト用データセット，例えば *Agree* クラスと *Discuss* クラスを同一にしたデータセットを構築する．

（2） 9 つの評価方法のうち 1 つを選択し，(1) で構築したデータセットに対する 14 種類の分類システムの評価値を求め，システムランキングを作成する．

（3） (2) で選択した評価方法によって，本来の 3 クラス分類用のテスト用データセットに対する 14 種類の分類システムの評価値を求め，システムランキングを作成する．

（4） Kendall の τ を用いて，(2) と (3) で作成された 2 つのシステムランキングの類似度を測定する．

（5） (2) ～ (4) の操作を 9 つの評価方法に対して行う．

（6） (1) ～ (5) の操作を，3 つの条件下（*Agree* クラスと *Discuss* クラス，*Disagree* クラスと *Discuss* クラス，*Agree* クラスと *Disagree* クラスをそれぞれ単一ラベルとみなしたデータセット）で行う．

上記の実験の結果を表 6 に示す．*Agree* と *Discuss* を単一ラベルとみなしたとき，α^{ORD} が $\tau = 0.5384$ と最も高く，α^{INT} が $\tau = 0.1428$ と最も低かった．また，*Disagree* と *Discuss* を単一ラベルとみなしたとき，κ が $\tau = 0.8901$ と最も高く，MAE^{M} が $\tau = 0.0989$ と最も低い値となった．さらに，*Agree* と *Disagree* を単一ラベルとみなしたとき，*Accuracy* が $\tau = 0.5384$ と最も高かったのに対し，CEM^{ORD} が $\tau = 0.0989$ と最も低かった．これらの 3 パターンを平均すると $\tau = 0.1721$ と MAE^{M} が最も低い値を示し，これは MAE^{M} が 3 つの Ordinal Class を最も区別できていることを示唆している．このような結果から，Ordinal Class を区別できているかどうかという観点では，スタンス検出タスクの評価方法として MAE^{M} が最も適していると言える．

さらに，定義上では Ordinal Class を区別できる評価方法である α^{INT}，α^{ORD}，CEM^{ORD}，MAE^{M}，MAE^{μ} は 3 パターンの τ の平均値が低く，Ordinal Class を区別できるという予想通りの結果が得られた．一方で，定義上では Ordinal Class を区別できる評価方法であるのにもかかわらず，κ は τ の平均値が高く，予想とは異なる結果となった．特に *Disagree* と *Discuss* を単一ラベルとみなしたケースにおいて κ の τ の値が高かったが，これはテスト用データセットにおけるデータ分布の偏りが影響していると考えられる．本節の実験では *Disagree* と *Discuss* のクラスを単一ラベルとみなすことによって，*Disagree* $\rightleftarrows$ *Discuss* の誤分類は「正しく分類された」ものとして扱われる．そのため，誤分類に対してペナルティを課すことで分類結果を評価する κ は，*Disagree* $\rightleftarrows$ *Discuss* の誤分類を評価の際に考慮しなくなってしまう．これは各分類システムの評価値に大きく影響し，14 種類の分類システムによるシステムランキングが変動する可能性がある．しかし，4. 節で述べたように本研究で使用したテスト用データセットはデータ分布が極端に偏っているため，*Disagree* $\rightleftarrows$ *Discuss* の誤分類件数が少なく，分類システムの評価値に与える影響が小さくなり，その結果としてシステムランキングにあまり変化が起きなかったと考えられる．

上記の考察を検証するため，テスト用データセットのデータ分布を均一にして（4. 節のテスト用データセットの中から *Agree* クラスと *Discuss* クラスのデータをそれぞれダウンサンプリングし，*Agree* と *Discuss* と *Disagree* のテストデータ数を全て 697 件とした）同様の実験を行った．その結果，*Agree* と *Discuss* を単一ラベルとみなしたときは $\tau = 0.1209$，*Disagree* と *Discuss* を単一ラベルとみなしたときは $\tau = 0.4066$，*Agree* と *Disagree* を単一ラベルとみなしたときは $\tau = 0.1429$ となり，データ分布が偏ったテスト用データセットを使用したときと比較してすべてのパターンにおいて低い τ の値を示した．このように，テスト用データセットのデータ分布が均一であれば，κ は Ordinal Class を区別できる可能性がある．

5.4　スタンス検出タスクに対する最適な評価方法

5.1 節，5.2 節，5.3 節で得られた実験結果を表 7 にまとめる．なお，表 7 中の Similarity では 5.1 節で得られた結果をもとに分けられたグループ名が示され，Consistency と Ordinal Class ではそれぞれ 5.2 節と 5.3 節で得られた結果が優れている評価方法から順にランキングしたときの順位が示されている．表 7 に基づき総合的に判断すると，少なくとも FNC-1 のスタンス検出タスクに対する評価方法としては，順位安定性と Ordinal Class の区別という 2 つの観点でともに優れた結果を残した，CEM^{ORD} が最も適切であると考えられる．

表 7　5.1 節，5.2 節，5.3 節で得られた実験結果のまとめ

評価方法	Similarity	Consistency	Ordinal Class
α^{ORD}	A	9_{th}	5_{th}
α^{INT}	B	7_{th}	4_{th}
MAE^{M}	C	5_{th}	1_{st}
MAE^{μ}	B	8_{th}	3_{rd}
CEM^{ORD}	A/B	3_{rd}	2_{nd}
κ	A	1_{st}	9_{th}
Accuracy	A	6_{th}	6_{th}
$\mathcal{F}_1$	A	2_{nd}	8_{th}
$\mathbb{F}_1$	A	4_{th}	7_{th}

6. 結論と今後の課題

本研究ではスタンス検出タスクにおける最適な評価方法を選定するため，FNC-1 のデータセットと分類システムを用いて，システムランキングの類似度と順位安定性，Ordinal Class の区別という 3 つの観点から 9 つの評価方法に対して比較実験を行った．その結果，順位安定性と Ordinal Class の区別という観点ではそれぞれ κ と MAE^M が最も優れているということがわかった．そしてこれらの結果を総合的に考えると，少なくとも FNC-1 のスタンス検出タスクに対する評価方法としては，高い順位安定性を持ち Ordinal Class も区別できるという結果が得られた CEM^{ORD} が最も適切であると言える．また，本研究では 3 種類の実験を通して，評価方法同士のシステムランキングの類似度や各評価方法の利点・欠点を明らかにすることができた．例えば，κ は順位安定性の観点では最も優れているのに対して Ordinal Class の区別という観点では最も劣っていた．このように，本研究で得られた評価方法の性質に関する知見は，新たなスタンス検出タスクを設計する際の評価方法の選定に非常に有益であると考えられる．

一方，本研究ではスタンス検出タスクの一例として FNC-1 を扱ったが，他のスタンス検出タスクを対象に同様の実験を行った場合，本研究で得られた結果と同じような傾向が見られるのかどうかを検証することが今後の課題である．また，本研究で使用した 9 つの評価方法の他に，データ分布に偏りがあるデータセットに対して有効な評価方法や Cost-Sensitive な評価方法などを加え，より多様性を高めた比較実験についても検討していきたい．

文　献

[1] Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. Will-They-Won't-They: A Very Large Dataset for Stance Detection on Twitter. In *Proceedings of ACL 2020*, pp. 1715–1724, 2020.

[2] Dilek Küçük and Fazli Can. Stance Detection: A Survey. *ACM Computing Surveys (CSUR)*, Vol. 53, No. 1, pp. 1–37, 2020.

[3] Mehrdad Fatourechi, Rabab K Ward, Steven G Mason, Jane Huggins, Alois Schlögl, and Gary E Birch. Comparison of Evaluation Metrics in Classification Applications with Imbalanced Datasets. In *2008 Seventh International Conference on Machine Learning and Applications*, pp. 777–782, 2008.

[4] Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. Stance Detection Benchmark: How Robust Is Your Stance Detection? *arXiv preprint arXiv:2001.01565*, 2020.

[5] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of SemEval-2016*, pp. 31–41, 2016.

[6] Guido Zarrella and Amy Marsh. MITRE at SemEval-2016 Task 6: Transfer Learning for Stance Detection. *arXiv preprint arXiv:1606.03784*, 2016.

[7] Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. pkudblab at SemEval-2016 Task 6 : A Specific Convolutional Neural Network System for Effective Stance Detection. In *Proceedings of SemEval-2016*, pp. 384–388, 2016.

[8] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of SemEval-2017*, pp. 69–76, 2017.

[9] Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. SemEval-2019 Task 7: RumourEval, Determining Rumour Veracity and Support for Rumours. In *Proceedings of SemEval-2019*, pp. 845–854, 2019.

[10] Dean Pomerleau and Delip Rao. Fake News Challenge, 2017.

[11] Benjamin Riedel, Isabelle Augenstein, Georgios P Spithourakis, and Sebastian Riedel. A Simple but Tough-to-beat Baseline for the Fake News Challenge Stance Detection Task. *arXiv preprint arXiv:1707.03264*, pp. 1–6, 2017.

[12] James Thorne, Mingjie Chen, Giorgos Myrianthous, Jiashu Pu, Xiaoxuan Wang, and Andreas Vlachos. Fake news stance detection using stacked ensemble of classifiers. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pp. 80–83, 2017.

[13] Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. A Retrospective Analysis of the Fake News Challenge Stance-Detection Task. In *Proceedings of COLING 2018*, pp. 1859–1874, 2018.

[14] Gaurav Bhatt, Aman Sharma, Shivam Sharma, Ankush Nagpal, Balasubramanian Raman, and Ankush Mittal. Combining Neural, Statistical and External Features for Fake News Stance Identification. In *Companion Proceedings of the The Web Conference 2018*, pp. 1353–1357, 2018.

[15] Luís Borges, Bruno Martins, and Pável Calado. Combining Similarity Features and Deep Representation Learning for Stance Detection in the Context of Checking Fake News. *Journal of Data and Information Quality (JDIQ)*, Vol. 11, No. 3, pp. 1–26, 2019.

[16] Qiang Zhang, Shangsong Liang, Aldo Lipani, Zhaochun Ren, and Emine Yilmaz. From Stances' Imbalance to Their Hierarchical Representation and Detection. In *The World Wide Web Conference*, pp. 2323–2332, 2019.

[17] Chris Dulhanty, Jason L Deglint, Ibrahim Ben Daya, and Alexander Wong. Taking a Stance on Fake News: Towards Automatic Disinformation Assessment via Deep Bidirectional Transformer Language Models for Stance Detection. *arXiv preprint arXiv:1911.11951*, 2019.

[18] Valeriya Slovikovskaya and Giuseppe Attardi. Transfer Learning from Transformers to Fake News Challenge Stance Detection (FNC-1) Task. In *Proceedings of LREC 2020*, pp. 1211–1218, 2020.

[19] Muhammad Umer, Zainab Imtiaz, Saleem Ullah, Arif Mehmood, Gyu Sang Choi, and Byung-Won On. Fake News Stance Detection Using Deep Learning Architecture (CNN-LSTM). *IEEE Access*, Vol. 8, pp. 156695–156706, 2020.

[20] 雨宮佑基, 酒井哲也. Convolutional Neural Network を用いた Fake News Challenge の検討. *DEIM Forum 2019 A3-2*, 2019.

[21] Tetsuya Sakai and Zhaohao Zeng. Which Diversity Evaluation Measures Are "Good"? In *Proceedings of ACM SIGIR 2019*, p. 595–604, 2019.

[22] Tetsuya Sakai. Evaluating Evaluation Measures for Ordinal Classification and Ordinal Quantification. *in preparation*.

[23] Juri Opitz and Sebastian Burst. Macro F1 and Macro F1. *arXiv preprint arXiv:1911.03347*, 2019.

[24] Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for Multi-Class Classification: an Overview. *arXiv preprint arXiv:2008.05756*, 2020.

[25] Enrique Amigo, Julio Gonzalo, Stefano Mizzaro, and Jorge Carrillo-de Albornoz. An Effectiveness Metric for Ordinal Classification: Formal Properties and Experimental Results. In *Proceedings of ACL 2020*, pp. 3938–3949, 2020.

[26] 角森唯子, 東中竜一郎, 高橋哲朗, 稲葉通将. 対話破綻検出チャレンジ 3 における対話破綻検出の評価尺度の選定. 人工知能学会論文誌, Vol. 35, No. 1, pp. DSI–G_1, 2020.