オンライン会議の雰囲気予測における困難な事例のサンプリングによる精度向上

新美翔太朗[1]　後藤啓介[1]　西田典起[2]　松本裕治[2]　廣島雅人[1]
[1] 京セラ株式会社
[2] 理化学研究所革新知能統合研究センター
{shotaro.niimi.gt, keisuke.goto.fj, masahito.hiroshima.zs}@kyocera.jp
{noriki.nishida, yuji.matsumoto}@riken.jp

概要

オンライン会議は参加者の内容の理解度によらず会議が進行してしまい，意思疎通が難しくなることがこれまでの研究で示唆されている．本研究ではオンライン会議の雰囲気をシステムに予測させることで非対面下においても雰囲気を共有することを目指し，特に検出が難しい険悪な事例の分類精度の向上を目的とする．具体的には，まず険悪な雰囲気の事例 (正例) の近傍にある良好な雰囲気の事例 (負例) を候補集合として収集し，そこから正例と同数になるまで，正例との距離に基づいて負例をサンプリングする手法 (HES_s) と，特徴が偏ることを回避するためにクラスタリングに基づいて負例をサンプリングする手法 (HES_v) を提案する．評価データを用いた実験の結果から，提案手法が Focal Loss やランダムなアンダーサンプリングよりも有効であることを確認した．

1　はじめに

近年，ビデオ通話アプリの技術的進歩と社会情勢の変化により，デバイスを用いてオンライン会議を行う機会が増加している．オンライン会議はオフライン会議と比較して時間や場所等の制約が少ないため，参加・開催するハードルは非常に低い．

しかし一方で，オンライン会議はオフライン会議とくらべて一人あたりの発言量が減り，意思疎通度も低くなることが宮内ら [1] によって確認されている．宮内らはその原因として参加者が話すタイミングを伺う必要があることや，理解度によらず会議が進行してしまう点を挙げている．著者らは，オンライン会議における理解の促進を図り発言を活発化させる研究について取り組みを行っている [2, 3]．中でも雰囲気を共有することは，オンライン会議における重要な課題である．

本稿では，オンライン会議の雰囲気を「良好」と「険悪」の 2 種類に分類することを考える．しかし，一般的にほとんどの会議は良好な雰囲気で進行されるため，険悪な雰囲気の事例を大量に収集することは難しい．特定のクラスのデータ数が極端に少ない，いわゆるラベル不均衡下では，マイナーラベルの予測精度が低下することが知られている．

そこで本研究では，険悪な雰囲気 (マイナーラベル) の検出精度を向上させるために，良好な雰囲気 (メジャーラベル) に対する 2 つのアンダーサンプリング手法を提案する．提案手法では，雰囲気予測が特に困難と考えられる事例 (hard example) をサンプリングの対象とする．具体的には，険悪な雰囲気の事例 (正例) の近傍にある良好な雰囲気の事例 (負例) は hard example であると考える．そして，各正例に対して発話系列上の距離が最も近い負例を一つずつサンプリングする手法 (HES_s) を提案する．また，サンプリングされたデータ集合の特徴が偏ることを回避するために負例集合をクラスタリングし，各クラスタの中心に最も近い負例を 1 つずつサンプリングする手法 (HES_v) を提案する．

本研究では UUDB コーパス [4] に収録されている対話データに対して雰囲気ラベルを付与し，Multilogue-Net [5] をベースモデルとして用いて実験を行った．ランダムサンプリングとの比較を行ったところ，HES_s は Micro F1 において 0.01 ポイント上回り，HES_v は同じく Micro F1 において 0.001 ポイント上回った．これらの結果は，オンライン会議の雰囲気分類におけるラベル不均衡問題に対して，分類困難な事例を積極的にデータサンプリングするアプローチが有効であり，また発話の並びの近さに基

づく困難性の推定とクラスタリングによる偏りの是正が有効であることを示している．

2 関連研究

分類モデルの訓練において，データセット中のラベル数の偏りから生じる問題はラベル不均衡問題と呼ばれ，自然言語処理を含む様々な機械学習タスクで多くの取り組みが行われてきた [6, 7, 8, 9, 10, 11]．

これらのうち，損失関数を工夫することでラベル不均衡性の解決を試みた代表的な手法として Focal Loss が提案されている [12]．Focal Loss ではモデルの予測確率の大きさをデータの難易度と見なし，その大きさに基づいて損失の重みを決定する．

一方，データ集合の加工によって解決しようとする手法はデータサンプリングと呼ばれる．特にマイナークラスのデータ数に合わせてメジャークラスのデータを抽出する手法はアンダーサンプリングと呼ばれる．メジャークラスのデータの抽出手法として，山口ら [13] は Hard Example Sampling (HES) を提案している．HES は，学習が困難な事例 (hard example) は学習が容易な事例よりも重要性が高いと考えて，メジャークラスのうち hard example を積極的に抽出することを目的とする．

本研究では，オンライン会議のマルチモーダル雰囲気予測におけるラベル不均衡問題を改善するために，HES を拡張する．山口らの HES では hard example からランダムにメジャークラスのデータを抽出していたが，本研究ではこの手法を特徴量の偏りを避けて抽出を行うように拡張する．具体的には，発話系列上の距離を考慮して抽出を行う拡張，クラスタリングの結果に基づいて均等に事例をサンプリングする拡張の 2 つを提案する．

3 タスクの定義

本研究ではオンライン会議を対象に，その発話ごとに会議の雰囲気が険悪であるか良好であるかを分類する．すなわち，会議 x は n 個の発話の系列 $x = x_1, \cdots, x_n$ として表され，さらに各発話 x_i はテキスト t_i と音声 a_i から構成されるものとする: $x_i = (t_i, a_i)$．各発話 x_i に対して，その時点までの発話系列 $x_{i-w}, \ldots, x_i$ から，その時点の会議の雰囲気 $y_i \in \{$ 険悪, 良好 $\}$ を予測する．本稿では $d_i = ([x_{i-w}, \cdots, x_i], y_i)$ を 1 つの「事例」として扱う．ウィンドウ幅 w は固定であり，本研究の実験では $w = 10$ とした．

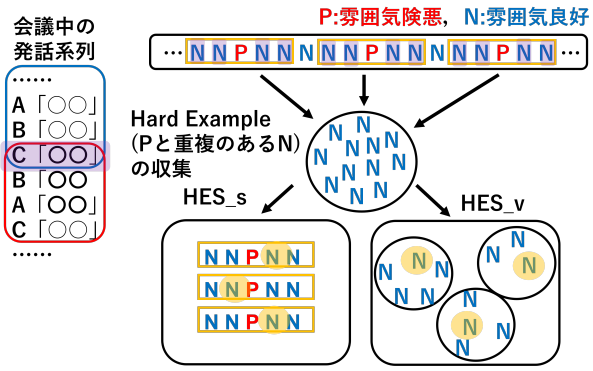

図 1 提案手法の概要図．提案手法では各正例に対して発話区間が部分重複する負例 (Hard Negative) をサンプリング候補として収集する．次に，HES_s では候補集合から各正例に対して系列上の距離が最も近い負例を 1 つずつ抽出する．HES_v では候補集合をクラスタリングし，各クラスタの中心に最も近い負例を 1 つずつ抽出する．

4 提案手法

1 章で述べた通り，一般的に会議は良好な雰囲気で行われる時間が長く，険悪な雰囲気の事例を大量に収集することは難しいため，ラベル不均衡問題が顕在化する．そこで本研究では，険悪な雰囲気を正例，良好な雰囲気を負例として，正例と同数の負例をサンプリングすることを考える．特に，負例の中でも雰囲気分類が難しいと予想される事例を重要視し，積極的にサンプリングすることでラベル不均衡の改善を試みる．具体的には，HES [13] を拡張し，正例の発話区間とのオーバーラップの有無に基づいて負例の難しさを決定し，負例サンプリングの候補集合を構築する (4.1 節)．そして，その候補集合から正例と同数までサンプリングを行う．候補集合からのサンプリングの方法として，正例との発話系列上の距離の近さに基づいてサンプリングする HES_s と，サンプリング結果の特徴の偏りを回避するために候補集合をクラスタリングし，各クラスタからサンプリングする HES_v を提案する (4.2 節)．HES_s, HES_v のサンプリング過程については図 1 に示す．

4.1 負例サンプリングの候補集合の構築

まず，負例サンプリングの対象となる候補を収集する．元になるデータセットの正例集合，負例集合をそれぞれ P，N ($|P| \ll |N|$) とする．各正例 $d_i \in P$ に対して，その近傍 (ウィンドウ r) 内にありかつ負例である事例 $N'_i = \{d_j \mid j \in [i-r, i+r] \wedge d_j \in N\}$ を見つけ，すべての正例について統合したものを負例サンプリングの候補集合とする: $\tilde{N} = \bigcup_{i=1}^{|P|} N'_i$．

4.2 HES_s と HES_v

負例サンプリングの候補集合 $\tilde{N}$ が構築できたら，次はそこから正例と同数になるまでデータをサンプリングする．本研究では，各正例 $d_i \in P$ について，発話系列上の距離 $|i-j|$ が最も近い負例候補 $d_j \in \tilde{N}$ をサンプリングする．以上のサンプリング方法を，本稿では文 (sentence) の並びの距離に基づいた HES として，HES_s と呼称する．

HES_s は発話系列上の距離のみに基づいてサンプリングするため，サンプリング後の最終的な負例集合の特徴が偏る可能性がある．そこで，本研究では候補集合 $\tilde{N}$ を負例候補 $d_i \in \tilde{N}$ の特徴ベクトルに基づいてクラスタリングし，各クラスタから均等に負例候補をサンプリングする方法を提案する．具体的には，候補集合 $\tilde{N}$ を K-means 法によってクラスタ数 $K = |P|$ でクラスタリングし，各クラスタの中心に最も近い負例候補を選択する．クラスタリングと分類タスクで用いる特徴ベクトルが異なると，折角クラスタリングを通しても，サンプリング後に分類タスクで用いられる負例集合の特徴が偏ってしまう可能性がある．そこで，本研究では後述するベースモデルをデータサンプリングを行わずに一度訓練し，各負例候補をそのモデルによってエンコードし，中間層の値をその特徴ベクトルとした．以上のサンプリング方法を，本稿では各負例候補の特徴ベクトルに基づいた HES として，HES_v と呼称する．

5 実験

5.1 データ

オンライン会議の雰囲気の変化についてアノテーションされたデータセットは存在しない．そこで，本研究では対人関係や態度に関して複数の項目によりアノテーションされたデータセットである UUDB [4] を利用し，新たに雰囲気のラベルを付与した．UUDB は，大学生の友人同士がバラバラにされた 4 コマ漫画の 2 コマずつを与えられ，自身が持つコマを相手に見せず会話のみで，協同して 4 コマ漫画を復元する過程を収録したデータセットである．UUDB には合計で 27 個の対話が収録されており，各発話に対して書き起こしテキストと音声情報が収録されている．また，各対話は平均 179.26 個の発話から構成されており，合計で 4,840 個の発話が収録されている．各発話に対しては 6 つの観点についてそれぞれ 3 人の評価者による 0〜6 の 7 段階スコアの平均 (e.g., 4.7) が付与されており，本研究ではその中で「快・不快」の観点のラベルを利用した．

UUDB では各発話 x_i に対して独立にラベルが付与されているが，雰囲気は文脈にも依存するため，発話 x_i のみを見てその時点での雰囲気ラベル y_i を決定することは難しい．そこで本研究では，入力の発話系列 $[x_{i-w}, \cdots, x_i]$ に基づいて雰囲気ラベル y_i を決定した．具体的には，UUDB に既にラベル付けされている「快・不快」のスコアを各入力系列ごとに平均化し，それが 3.5 未満ならば雰囲気が険悪，3.5 以上ならば雰囲気が良好とした．結果として，正例の数 $|P|$ が 324 件，負例の数 $|N|$ が 4,516 件，負例サンプリングの候補数 $|\tilde{N}|$ が 2,401 件であった．

5.2 評価方法

UUDB では 27 個の各対話に対して話者に関するメタ情報が付与されている．そこで，同一話者による対話が訓練セットとテストセットで重複して現れないように UUDB を 7 つのサブセットに分割し，各サブセットをテストセット，残りを訓練セットとして 7 分割交差検証を行った．評価指標には Micro 評価を採用し，各サブセットに含まれる事例への予測から Precision と Recall，F1 を求めた．

5.3 ベースライン

提案手法の有効性を評価するために，下記の 4 つのベースラインと提案手法との比較を行った．

ALL data with Focal loss (以下 ALL w/ FL) データサンプリングを行わず，Focal Loss [12] を用いて困難な事例の損失に対してより大きな重みを割り当てるようにした．

Over Sampling (以下 OS) 負例と同数になるように正例をランダムにコピーする方法でオーバーサンプリング [10] を行った．

Under Sampling (以下 US) 負例候補集合 $\tilde{N}$ を考慮せず，正例と同数の負例をランダムにアンダーサンプリングを行った．

Random HES (以下 HES_r) 負例候補集合 $\tilde{N}$ を考慮して，正例と同数の負例をランダムにアンダーサンプリングを行った．

5.4 ベースモデル

提案手法の有効性を評価するためのベースモデルとして，Shenoy ら [5] を参考に，ニューラルネッ

表 1　UUDB における 7 分割交差検証の結果.

Method	Precsion	Recall	F1
ALL w/ FL	**0.128**	0.086	0.103
OS	0.055	0.170	0.083
US	0.082	0.336	0.132
HES_r	0.074	0.380	0.124
HES_s	0.083	**0.503**	**0.142**
HES_v	0.079	0.420	0.133

トワークモデルを構築した．具体的には，まず各発話 x_i を独立に特徴ベクトル $f(x_i)$ にエンコードし，次に特徴ベクトルの系列 $f(x_{i-w}), \cdots, f(x_i)$ を GRU に入力し，各発話の隠れ状態ベクトルを計算した．次に，それらの隠れ状態ベクトルに対して Multi-Head Attention [14] を適用し，入力系列全体の特徴ベクトルを求めた．最後にこの特徴ベクトルを線形層に入力し，発話系列全体に対する雰囲気ラベルを予測した．UUDB の各発話 x_i はテキスト t_i と音声 a_i から構成されており，本実験では $f(x_i) = [f_{\text{text}}(t_i); f_{\text{audio}}(a_i)]$ とした．$f_{\text{text}}(t_i)$ は BERT [15] の事前学習済モデルにテキスト t_i を入力したときの [CLS] トークンに対応する最終埋め込みとした．$f_{\text{audio}}(a_i)$ は音声 a_i のメルスペクトログラム画像を事前学習済み ResNet [16] に入力したときの最終隠れ層とした．

5.5　結果と考察

実験の結果を表 1 に示す．ALL w/ FL とその他の手法を比べると，ALL w/ FL の Precision は最も高いが，Recall は最も低い．これは，ALL w/ FL が正例の出力に消極的過ぎであり，Focal Loss がラベル不均衡の影響を改善できていないことを示唆している．また，アンダーサンプリング (US, HES_r, HES_s，HES_v) の F 値はオーバーサンプリング (OS) よりも 0.041 ポイント以上高く，Recall も 0.166 ポイント以上高い．以上から，ラベル不均衡問題に対して Focal Loss やオーバーサンプリングよりもアンダーサンプリングのほうが効果的であることが示唆される．

US と HES_r を比べると，F 値は US がわずかに上回っている．これは，発話系列スパンが正例と重複する負例を hard example としてサンプリングするだけでは必ずしも精度向上につながらないことを示唆している．一方で，US と HES_s，HES_v を比べると，HES_s が 0.01 ポイント，HES_v が 0.001 ポイントそれぞれ F 値において上回っている．以上から，負例候補集合 $\tilde{N}$ に対象を絞り，さらに効果的な事例にしぼってサンプリングすることが，他のランダムなアンダーサンプリング手法 (US, HES_r) と比較して効果的であることを示唆している．具体的には，各正例について最も近い負例を選択する HES_s と，特徴空間が偏ることを回避するためにクラスタリングしてからサンプリングする HES_v の有効性が確認できた．

最後に HES_s と HES_v を比較すると，Precision, Recall, F 値すべてにおいて HES_s が上回っている．HES_s と HES_v の違いは負例抽出の際に注目した観点である．HES_v ではクラスタリングされた負例候補集合から，各クラスタの中心に最も近い負例を 1 つずつ抽出する．しかし，特徴空間において多様な負例集合が，常に正例集合に対して効果的であるとは限らない．このことが HES_v が HES_s を下回った要因の 1 つだと考えられる．また，HES_v がクラスタリングを行う際に特徴抽出で用いたモデルである ALL w/ FL は今回の結果において最も F 値が低く，クラスタリングが失敗している可能性も考えられる．大規模な事前学習済みモデルによって生成された特徴ベクトルを用いて作成されたクラスターからサンプリングするデータを決定することで，HES_v の精度の向上が期待出来る．

6　おわりに

本稿では，オンライン会議の雰囲気分類におけるラベル不均衡問題を改善するために，2 つのアンダーサンプリング手法を提案した．具体的には，まず雰囲気の予測が特に困難である可能性が高い，険悪な雰囲気の事例 (正例) の近傍にある良好な雰囲気の事例 (負例) を候補集合として収集した．さらにそこから正例と同数になるように，各正例に対して最も系列上の距離が近い負例を 1 つずつ抽出する HES_s と，負例候補集合をクラスタリングして，各クラスタの中心に最も近い負例を 1 つずつ抽出する HES_v を提案した．UUDB を用いた実験の結果から，提案手法が Focal Loss やランダムなアンダーサンプリング (US)，候補集合からのランダムなサンプリング (HES_r) よりも有効であることを確認した．

今後は正例集合に対する最適な負例集合の構築方法についての検討や，抽出する負例の数を徐々に増やした場合の実験について取り組んでいく．

参考文献

[1] 宮内佑実, 遠藤正之. オンライン会議とオフライン会議の意思疎通の比較. 経営情報学会, pp. 144 – 147, 2020.

[2] 後藤啓介, 新美翔太朗, 荒川智也, 西田典紀, 松本裕治, 廣島雅人. オンライン会議における議論の要点と対話の雰囲気の認識技術の開発. 情報処理学会研究報告, Vol. 2022-NL-253, No. 15, 2022.

[3] 新美翔太朗, 後藤啓介, 荒川智也, 西田典紀, 松本裕治, 廣島雅人. オンライン会議での自動要約のためのマルチモーダル情報を考慮した重要発話抽出に関する検討. 情報処理学会研究報告, Vol. 2022-NL-253, No. 16, 2022.

[4] 森大毅, 相澤宏, 粕谷英樹. 対話音声のパラ言語情報ラベリングの安定性. 日本音響学会誌, Vol. 61, No. 12, pp. 609–697, 2005.

[5] Aman Shenoy and Ashish Sardana. Multilogue-net: A context aware rnn for multi-modal emotion detection and sentiment analysis in conversation. In **Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 19–28, 2020.

[6] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks. In **Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 1–13, 2020.

[7] PiCkLe at SemEval-2022 Task 4: Boosting Pre-trained Language Models with Task Specific Metadata and Cost Sensitive Learning. Manan suri. In **Semantic Evaluation**, pp. 464–472, 2022.

[8] Harish Tayyar Madabushi, Elena Kochkina, and Michael Castelle. Cost-sensitive bert for generalisable sentence classification on imbalanced data. In **Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda**, pp. 125–134, 2019.

[9] Jinfen Li and Lu Xiao. syrapropa at semeval-2020 task 11: Bert-based models design for propagandistic technique and span detection. In **Semantic Evaluation**, pp. 1808–1816, 2020.

[10] Nathalie Japkowicz. The class imbalance problem: Significance and strategies. In **International Conference on Artificial Intelligence (IC-AI)**, pp. 111–117, 2000.

[11] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. Vol. 16, pp. 321–357, 2002.

[12] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In **International Conference on Computer Vision (ICCV)**, pp. 2980–2988, 2017.

[13] 山口泰弘, 進藤裕之, 渡辺太郎. ラベルの不均衡を考慮した end-to-end 情報抽出モデルの学習. 言語処理学会 第 27 回年次大会, pp. 400–404, 2021.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Advances in Neural Information Processing Systems**, pp. 6000–6010, 2017.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In **The North American Chapter of the Association for Computational Linguistics (NAACL)**, pp. 4171—-4186, 2019.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In **Computer Vision and Pattern Recognition (CVPR)**, pp. 770–778, 2016.