F-024

アノテータごとの判断傾向を推定する破綻検出器の検討
Dialogue Breakdown Detection by Estimatting Tendency of Annotator

桑原 健太† 大村 英史† 桂田 浩一†
Kenta Kuwahara Hidehumi Ohmura Kturada Kouichi

1. はじめに

近年，非タスク指向の雑談対話システムが注目を集めている．しかし，現状の水準ではシステムが適切な応答を返答しつづけることは難しく,しばしば対話破綻が発生してしまう．この問題に対処するために対話破綻を検出する技術の研究が盛んに行われている．2015 年から開催されている対話破綻検出チャレンジ(Dialogue Breakdown Detection Challenge: DBDC) [1]では対話システムとユーザとの対話ログから破綻箇所を見つけるタスクのコンペティションが行われており，2019 年に実施された DBDC4 では Sugiyama の BERT を用いた検出器が日本語のデータに対し高い性能を示している[2].

対話破綻を検出するにあたって考慮すべき問題の一つに，対話が破綻しているかどうかの判断が人によって異なることが挙げられる．破綻しているかどうかは主観的な判断になるため，対話破綻検出用のデータにラベルを付与するアノテータごとに判断傾向の差が生じることになる．この問題に対して Takayama らは似通った判断傾向を持つアノテータをグループ化し，グループごとに検出器を学習することで高い性能を示した[3].

これに対して我々は，学習データ中のアノテータの判断傾向に影響を受けないラベル分布の学習によって破綻検出器を構築する方法を提案する．本研究では Tanno らの混同行列を利用した学習法[4]を用いて，アノテータの判断傾向の推定を行いながら目的のラベル分布を学習する．この手法を用いた検出器と用いない検出器との精度を比較し，アノテータの判断傾向の推定が本タスクにおいて有効かどうかを検証する．

2. アノテータごとの判断傾向を推定する破綻検出

2.1 作成した破綻検出器

対話破綻検出チャレンジで用いられる対話ログデータの各システム発話には，複数人のアノテータによって「破綻を引き起こしていない（O）」,「破綻を引き起こしていないが違和感がある（T）」,「破綻を引き起こしている（X）」のいずれかのラベルが付与されている．チャレンジの目的は，このラベルの確率分布を予測することである．本研究ではこのために Sugiyama がベースラインとして採用している BERT を利用した破綻検出器[2]をベースラインとして構築した．図 1 の内側がベースラインのモデルである．入力には判定するシステムと直前のユーザ発話，一つ前のシステム発話とその直前のユーザ発話を用いている．先頭に[CLS]トークンを入力し，続けて各発話を[SEP]トークンで区切って入力する．出力には[CLS]トークンに対する出力を用いて，入力xに対する各ラベルの確率分布$\mathrm{P}(x) = [p(\mathrm{O}), p(\mathrm{T}), p(\mathrm{X})]$を出力する．

† 東京理科大学大学院　理工学研究科,

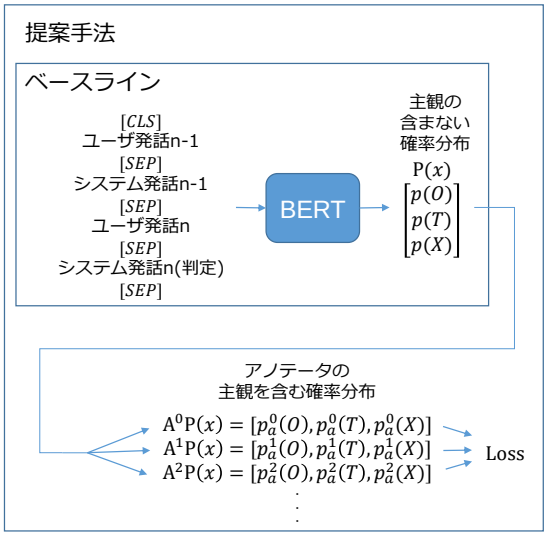

図 1　作成した検出器

既存の機械学習ベースの破綻検出器の多くは教師データとして「最も多くのアノテータが付与したラベル」か「ラベルの分布」をそのまま用いている．この方法では学習データ内のアノテータの判断傾向の影響を受けた分布を学習する可能性がある．そこで本研究では Tanno らの混同剛列を利用した学習手法を用いて，$\mathrm{P}(x)$をアノテータの判断傾向に影響されない分布にする方法を検討した．次節では具体的な学習方法について説明する．

2.2 判断傾向の推定を行う学習

Tanno らの手法は分類モデルの学習の際にアノテーションノイズを除いた分布を学習する手法である[4]．アノテーションノイズにはアノテータの判断傾向による真のラベル（最も多くのアノテータが付けたラベル）からのずれが含まれると考え，判断傾向を含めたモデル化を行うことを目指している．最初にデータセット中の各アノテータkに対し3×3の行列$\mathrm{A}^k = (a_{mn}^k)$を用意する．$m, n \in \{0,1,2\}$はラベルを表す値であり，それぞれ O,T,X を表す．a_{mn}^kは真のラベルがnのデータに対しアノテータkがmのラベルを付与する確率である．A^kはアノテータkの混同行列(CM)として扱う．BERT が出力する各ラベルの確率分布$\mathrm{P}(\mathrm{x})$に対し行列積をとることでアノテータkの判断傾向を加味した確率分布$\mathrm{P}_\mathrm{a}^k(x) = \mathrm{A}^k\mathrm{P}(x) = [p_a^k(\mathrm{O}), p_a^k(\mathrm{T}), p_a^k(\mathrm{X})]$へと変換する．BERT を fine-tuning する際にはA^kの学習を行う．入力データxに対する各$\mathrm{P}_\mathrm{a}^k(x)$が，アノテータ$k$が付与したラベルを予測するように学習させる．損失関数には次式を使用する．

$$\mathrm{Loss} = \sum_{k \in S(x)} \mathrm{CL}\left(y^k, \mathrm{P}_\mathrm{a}^k(x)\right) + \lambda * \mathrm{mean}_{k \in S(x)}(\mathrm{trace}(\mathrm{A}^k))$$

ここでCLは交差エントロピー誤差，mean は平均，traceは行列の対角成分の和である．$S(x)$は入力xにラベルを付与

表 1　破綻検出の結果

Model	Acc	F(X)	F(T+X)
提案手法($\lambda = 100$)	0.616	0.702	**0.842**
提案手法($\lambda = 1$)	**0.625**	**0.710**	0.844
ベースライン	0.608	0.694	0.834

Model	JS (O,T,X)	MSE (O,T,X)
提案手法($\lambda = 100$)	**0.0684**	**0.0376**
提案手法($\lambda = 1$)	0.0686	0.0378
ベースライン	0.0695	0.0381

したアノテータの集合，y^kはアノテータkが付与したラベル，λは係数である．各A^kの対角成分を最小化することで，A^kは対角成分以外の値が高くなるよう，すなわちP(x)が出力したラベルをアノテータの判断傾向に変換するよう学習される[4]．結果として BERT の出力であるP(x)はアノテータの判断傾向に非依存になるため，実際の予測時にはA^kによる変換を行わずにP(x)のみを用いることで，学習データ中のアノテータの判断傾向に影響されない予測を行う．

3. 評価実験

3.1 実験概要

図 1 の内側のみで構築したベースラインの検出器と，アノテータの判断傾向の推定を行う提案手法で構築した検出器の比較を行った．BERT の事前学習モデルには Kikuta が公開しているものを使用した[5]．使用した BERT は日本語 Wikipedia で学習されており，分かち書きには Sentencepiece が用いられている．ベースラインの検出器は学習データ中のラベルをそのまま正解データとして学習した．提案手法の検出器は Loss を$\lambda = 1$で学習した検出器と$\lambda = 100$で学習した検出器の 2 つを用意した．各A^kは単位行列で初期化した．学習データには，対話破綻検出チャレンジの開発用データと評価用データ，対話破綻検出チャレンジ 2 の開発用データと評価用データを用いた．評価データには対話破綻検出チャレンジ 3 の日本語の評価用データを用いた．評価尺度には，ラベルの一致率に関する指標である正答率(Acc)，ラベル X の F 値(F(X))，ラベル T を X とみなしたときの X の F 値(F(T+X))，確率分布の距離に関する指標である JS-Divergence(JS(O,T,X))，Mean Squared Error(MSE(O,T,X))を用いた．これらは対話破綻検出チャレンジで評価尺度として使用されている．

3.2 実験結果と考察

表 1 に結果を示す．提案手法はどちらもベースラインと比べて高い性能を示していることが分かる．特に Acc,F(X)では$\lambda = 1$のモデル，F(T+X)と JS(O,T,X),MSE(O,T,X)では$\lambda = 100$のモデルがより高い性能を示している．

学習した各A^kの値を調査してみたところ，$\lambda = 1$のモデルでは初期値から大きく変化せず，アノテータ間でA^kに差は見られなかった．λの値が小さかったために変化が少なく，A^kに差が表れなかったと考えられる．一方で$\lambda = 100$のモデルでは，λの値が大きくなったことでA^kに差が現れた．$\lambda = 1$のモデルと比べると Acc,F(X)に関しては悪化し，JS(O,T,X)，MSE(O,T,X)に関しては高い性能を示している．

表 2　A^kの例

アノテータのラベル ＼ 真のラベル	O	T	X
O	0.98605	0.00674	0.00721
T	0.00715	0.98564	0.00722
X	0.00650	0.00680	0.98670

この結果からアノテータの判断傾向の推定はラベルの正答率ではなく分布間距離の改善に繋がっていると考えられる．

表 2 に$\lambda = 100$のモデルで学習したA^kの一例を示す．真のラベルと異なるラベルを付与する確率のうち，X のデータに O，O のデータに T，X のデータに T を付与する確率が高いことが分かる．このアノテータが O を付与した対話の例を以下に示す．

ユーザ:　たけのこ派それともきのこ派？
システム:　江崎グリコのいちごポッキーに梨の味があることは知りませんでした。O:2 T:7 X:21

発話の隣の数字は，そのラベルを付与した人数である．約 2/3 が X のラベルを付与している発話に対し O をつけている．このことからも，このアノテータは X のラベルが多く付与されたデータに対し稀に O を付与するという傾向が確認できる．表 2 のA^kはその傾向をモデル化していることを確認できる．

今回アノテータの判断傾向による影響を除外するためにアノテータの CM の推定を行ったが，この CM は発話文に影響しない．そのため，例えば若者言葉への理解の違いからくる判断の差などの，発話文の解釈の差に基づくアノテータの判断傾向に関しては表現できていない．CM の推定に発話文の特徴量を加えることによって，発話文の特徴に基づく詳細な判断傾向の推定を行うことができると思われる．この詳細な推定は予測時の判断傾向による影響をより正確に除外できると考えられる．

4. おわりに

本研究ではアノテータの判断傾向を推定することで，アノテータの判断傾向に影響を受けない破綻検出器を提案した．この手法は従来の判断傾向の推定を行わない手法と比べて，高い性能を示すことが確認できた．また，判断傾向の推定を行う学習は検出器の分布間距離の改善に繋がることが判明した．今後は発話文の特徴量を用いた判断傾向の推定に取り組みたい．

参考文献

[1] Ryuichiro Higashinaka, Luis F. D'Haro, Bayan Abu Shawar, Overview of the Dialogue Breakdown Detection Challenge 4, International Workshop on Spoken Dialog System Technology(2019)
[2] Hiroaki Sugiyama, Dialogue breakdown detection using BERT with traditional dialogue features, International Workshop on Spoken Dialog System Technology(2019)
[3] Junya Takayama, Eriko Nomoto, Yuki,Arase, Dialogue Breakdown Detection Considering Annotation Biases, Dialog System Technology Challenges 6 Workshop(2017).
[4] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Learning From Noisy Labels By Regularized Estimation Of Annotator Confusion, CVPR(2019)
[5] Yohei Kikuta, BERT Pretrained model Trained On Japanese Wikipedia Articles, GitHub repository, https://github.com/yoheikikuta/bert-japanese (2019)