Vol.2018-DBS-167 No.28
Vol.2018-IFAT-132 No.28
2018/9/14

対話破綻検出コーパスに対する学習データ選別の検討

河東 宗祐[1,a)]　酒井 哲也[1,b)]

概要：対話システムなどの自然言語処理システムの評価においては人間の主観的な判断を扱う必要があり，このためのひとつのアプローチとして，正解を 1 つのラベルではなく複数の判定者によるラベルの分布として扱う評価方法がある．この代表的なものに，与えられたシステム・ユーザ間対話の各システム発話の破綻ラベル上の判定者分布を予測する対話破綻検出チャレンジがある．ここで，破綻ラベルは NB (対話を継続することができないあきらかにおかしい発話)，PB (違和感を感じる発話)，B (破綻ではない) の三種類からなり、複数の種類の対話システムを用いて収集された対話集合が与えられる。本研究では，対話集合に付与された破綻ラベルの確率分布群の均衡性，均一性に注目し，学習データの選別を試みた．提案する均一性に注目した学習データ選別手法は，学習データを一切選別しない手法に比べ統計的に有意な性能の向上が見られた．

キーワード：均一性，均衡性，分布間距離，対話破綻検出，学習データ選別

An investigation into Training Data Selection for Dialogue Breakdown Detection Corpora

Kato Sosuke[1,a)]　Sakai Tetsuya[1,b)]

1. はじめに

対話システムなどの自然言語処理システムの評価においては人間の主観的な判断を扱う必要があり，このためのひとつのアプローチとして，正解を 1 つのラベルではなく複数の判定者によるラベルの分布として扱う評価方法がある．この代表的なものに，与えられたシステム・ユーザ間対話の各システム発話の破綻ラベル上の判定者分布を予測する対話破綻検出チャレンジ [2] がある．ここで，破綻ラベルは NB (対話を継続することができないあきらかにおかしい発話)，PB (違和感を感じる発話)，B (破綻ではない) の三種類からなり、複数の種類の対話システムを用いて収集された対話集合が与えられる。本論文では，各対話システムを用いて収集された対話集合それぞれをコーパスと呼ぶ．対話破綻検出チャレンジ 3 [2] では日本語と英語のデータセットが使用され，英語データセットは 4 つの対話システムから収集された 4 つのコーパス (CIC，IRIS，TickTock，YI) からなる．分布距離系統の評価尺度である Jensen-Shannon Divergence (JSD) のコーパス毎の結果を見ると，コーパス毎にトップのシステムのスコアに大きく違いがあることがわかる．ここでは，よりコーパス毎の結果に違いがある英語データセットに注目する．

実際に対話システムとユーザが対話している際に対話の破綻を検出する場合を考えた時，どの対話システムが使用されているのかがわかっていることは自然である．問題設定として，破綻ラベルの確率分布を予測する上で評価用対話が属するコーパスを既知とした．つまり，システム発話 u とその発話が属するコーパス s が与えられた時に，破綻ラベル i ($i \in \{\mathrm{NB}, \mathrm{PB}, \mathrm{B}\}$) の確率 $p(i|u,s)$ を予測する．

本研究では，対話集合に付与された破綻ラベルの確率分

[1] 早稲田大学
Waseda University, Shinjuku, Tokyo 169–8555, Japan
[a)] sow@suou.waseda.jp
[b)] tetsuyasakai@acm.org

布群の特性に注目し，コーパス単位での学習データの選別を試みた．確率分布群の特性として，不均衡性と不均一性を考えた．破綻ラベルの確率分布群の不均衡とは，破綻を示す確率分布と非破綻を示す確率分布の釣り合いのとれていないことで，不均一とは，確率分布群がより似た分布で占められていないことと定義する．

2. 関連研究

2.1 不均衡なデータセットに対する学習データの選別

Li [4] らは，感情分類タスクにおいて，不均衡なデータセットに対する学習データの選別を行なっている． 感情分類タスクにおけるデータセットでは，付与されているラベルは確率分布としてではなく単一のラベルである．Li らの学習データの選別の主な方針は，学習データセット内のラベルの割合が最も多い多数派なラベルと少ない少数派なラベルの数を同じに近づけることである．

本研究でとり組んでいる対話破綻検出タスクでは，単一のラベルではなく確率分布が付与されているため，確率分布群に対する不均衡性の指標を 3.2 節で示す．

2.2 キーワードと発話間類似度を用いた対話破綻検出

河東ら [3] は，対象のシステム発話の特徴量として，*offer weight* [7] を用いて選出したキーワードの有無や，1 つ前のユーザ発話および 1 つ前のシステム発話間の発話間類似度などを用いている．発話間類似度を計算するために，単語頻度をベースとしたベクトル [1,5] と事前学習した単語埋め込みベクトルを利用したベクトル [5] を用いている．また，推定アルゴリズムとして ExtraTreesRegressor [6][*1] (ETR) を用いているが，単に各破綻ラベルの確率を予測対象とするのではなく，各破綻ラベルを実数値にマッピングしたものの平均と分散の値を予測対象としている．

3. 提案手法

本研究で使用する手法はベースラインを含め全て，2.2 節で示した河東ら [3] の Run2 をベースとしている．ベースラインを含めた各手法間の違いは，モデルの選択方法の違いのみであり，特徴量および推定アルゴリズムとして ETR を用いる点は同じである．各コーパスのみを学習に用いたモデルと全てのコーパスを学習に用いたモデルを用意して，評価用発話の属するコーパスによりモデルを選択することで，学習データの選別を行なっている．

発話の特徴量は，河東らの Run2 で用いられた特徴量と一部を除き同じものを用いている．河東らは，学習データ内の対象のシステム発話，1 つ前のユーザ発話，1 つ前のシステム発話からそれぞれ抽出したキーワードに対して，対象の評価用発話内のキーワードの有無を用いて，キーワードフラグを生成している．本論文で用いる手法では全て，学習データ内の対象のシステム発話から抽出したキーワードは対象の評価用発話内のキーワードの有無を用いて，1 つ前のユーザ発話から抽出したキーワードは対象の評価用発話の 1 つ前のユーザ発話内のキーワードの有無を用いて，1 つ前のシステム発話から抽出したキーワードは対象の評価用発話の 1 つ前のシステム発話内のキーワードの有無を用いて，キーワードフラグを生成している．また，河東らは ETR の予測対象として，各破綻ラベルを実数値にマッピングしたものの平均と分散の値を用いているが，単純に各破綻ラベルの確率を ETR の予測対象とした場合の結果も 4 節で示している．

3.1 ベースライン

1 つ目のベースラインは，対象の評価用発話の属するコーパスによらず，全てのコーパスの学習用データを用いたモデルを選択する．評価用発話 u と属するコーパス s が与えられた時に，全てのコーパスの学習用データを使い学習したモデルを用いて予測した破綻ラベルの確率を $p_G(i|u)$ とすると，予測する破綻ラベルの確率 $\hat{p}(i|u,s)$ は次のように表すことができる．

$$\hat{p}(i|u,s) = p_G(i|u) \tag{1}$$

2 つ目のベースラインとして，コーパス毎に特性があることを考慮した単純な手法を考える．対象の評価用発話の属するコーパスと同じコーパスの学習用データのみを用いたモデルを選択する．評価用発話 u と属するコーパス s が与えられた時に，コーパス s の学習用データのみを使い学習したモデルを用いて予測した破綻ラベルの確率を $p_s(i|u)$ とすると，予測する破綻ラベルの確率 $\hat{p}(i|u,s)$ は次のように表すことができる．

$$\hat{p}(i|u,s) = p_s(i|u) \tag{2}$$

3.2 不均衡性

対話集合に付与された破綻ラベルの確率分布群の不均衡性に注目する．より均衡な学習データセット，つまり，より破綻を示す確率分布と非破綻を示す確率分布の釣り合いのとれた学習データセットの方が学習データセット内の確率分布群の偏りによらない予測ができるという仮定のもと，モデルの選択を行う．まず，確率分布の集合 P が与えられた時に，その不均衡さを示す指標として次のような式を考える．

$$Imb(P) = \frac{1}{|P|}\sum_{p\in P}\{D_{JSD}(p,p_+) - D_{JSD}(p,p_-)\} \tag{3}$$

ここで， $D_{JSD}(p,q)$ は 確率分布 p，q 間の JSD の距離であり，p_+，p_- は次のような両極端な確率分布である．

[*1] http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html

$$p_+(i) = \begin{cases} 1.0 & i = \mathrm{NB} \\ 0.0 & i \neq \mathrm{NB} \end{cases} \quad (4)$$

$$p_-(i) = \begin{cases} 1.0 & i = \mathrm{B} \\ 0.0 & i \neq \mathrm{B} \end{cases} \quad (5)$$

次に，対象の評価用発話と属するコーパスが与えられた時のモデル選択を考える．対象の評価用発話の属するコーパスに付与された確率分布群の不均衡性の絶対値が全てのコーパスに付与された確率分布群の不均衡性の絶対値よりも小さい時のみ，同じコーパスの学習用データのみを用いたモデルを選択する．つまり，評価用発話 u と属するコーパス s が与えられた時に，予測する破綻ラベルの確率 $\hat{p}(i|u,s)$ は次のように表すことができる．

$$\hat{p}(i|u,s) = \begin{cases} p_s(i|u) & |Imb(P_s)| < \left|Imb\left(\bigcup_{s' \in S} P_{s'}\right)\right| \\ p_G(i|u) & |Imb(P_s)| \geq \left|Imb\left(\bigcup_{s' \in S} P_{s'}\right)\right| \end{cases} \quad (6)$$

ここで，P_s はコーパス s 内の学習データに付与された確率分布の集合である．また，S はコーパス集合であり，$S = \{\mathrm{CIC}, \mathrm{IRIS}, \mathrm{TickTock}, \mathrm{YI}\}$ である．

3.3 不均一性

対話集合に付与された破綻ラベルの確率分布群の不均一性に注目する．より均一な評価データセット，つまり，データセット内の確率分布群がより似た分布同士で占められている評価データセットの場合，モデルの学習に用いるデータセット内の確率分布群も似た分布同士で占められている学習データセットの方が良く，且つ，同じコーパスの評価データ内の学習データの特性は似るという仮定のもと，モデルの選択を行う．まず，確率分布の集合 P が与えられた時に，その不均一さを示す指標として次のような式を考える．

$$Het(P) = \frac{1}{|P|} \sum_{p \in P} D_{JSD}(p, \overline{p}_P) \quad (7)$$

ここで，$\overline{p}_P$ は P 内の各破綻ラベル i の確率の平均をそれぞれ値とする確率分布であり，次の式で表すことができる．

$$\overline{p}_P(i) = \frac{1}{|P|} \sum_{p' \in P} p'(i) \quad (8)$$

次に，対象の評価用発話と属するコーパスが与えられた時のモデル選択を考える．対象の評価用発話の属するコーパスに付与された確率分布群の不均一性が全てのコーパスに付与された確率分布群の不均一性よりも小さい時のみ，同じコーパスの学習用データのみを用いたモデルを選択する．つまり，評価用発話 u と属するコーパス s が与えられた時に，予測する破綻ラベルの確率 $\hat{p}(i|u,s)$ は次のように表すことができる．

$$\hat{p}(i|u,s) = \begin{cases} p_s(i|u) & Het(P_s) < Het\left(\bigcup_{s' \in S} P_{s'}\right) \\ p_G(i|u) & Het(P_s) \geq Het\left(\bigcup_{s' \in S} P_{s'}\right) \end{cases} \quad (9)$$

4. 実験と結果

データセットは，対話破綻検出チャレンジ 3 [2] の英語サブタスクで用いられたものを使用した．データセット内には 4 つのコーパス，CIC，IRIS，TickTock および YI があり，発話はいずれかのコーパスに属している．問題設定として，評価用発話が与えられた時に，その発話が属するコーパスも与えられることとした．

表 1 に，各手法の全評価用発話に関する F 値と JSD の平均値を示す．表 1 において，3.1 節で述べた 1 つ目のベースラインを G-BL，2 つ目のベースラインを s-BL，3.2 節で述べた提案手法を IMB，3.3 節で述べた提案手法を HET とした．また，河東ら [3] の Run2 と同様に ETR の予測対象として，各破綻ラベルを実数値にマッピングしたものの平均と分散の値を用いた場合 (STAT)，モデル選択手法の後ろの山括弧内に STAT と表記し，単純に各破綻ラベルの確率を ETR の予測対象とした場合 (OTX)，山括弧内に OTX と表記した．また，STAT である全手法対についてと OTX である全手法対について，Tukey HSD 検定を行なった時の p 値と括弧内の標本効果量 [8] を表 2，3 に示す．表 1 において，最もスコアの良いものを太文字で示し，上位 2 つに下線を引いた．

表 1 より，F 値と JSD の値で最も良いスコアを示したモデルは違うものの s-BL<OTX> と HET<OTX> が良いスコアを示していることがわかる．JSD の値に注目すると，表 2，3 より，STAT と OTX どちらにおいても，モデル選択手法 G-BL と HET の間に，つまり，全てのコーパスを学習データとして用いる手法と不均一性によって学習データを選別する手法の間に $\alpha = 0.05$ で統計的有意差が得られた．

表 1 手法別の F 値と平均 JSD

手法	F1(B)	JSD(NB,PB,B)
G-BL<STAT>	0.3058	0.0414
G-BL<OTX>	0.3574	0.0375
s-BL<STAT>	0.3579	0.0403
s-BL<OTX>	**0.3992**	0.0371
IMB<STAT>	0.2958	0.0411
IMB<OTX>	0.3678	0.0381
HET<STAT>	0.3423	0.0394
HET<OTX>	0.3849	**0.0358**

5. 分析と考察

表 4 にモデル毎の評価用発話の属するコーパス別の JSD の平均値を示す．モデルは学習に用いたコーパスの種類 5 個と ETR の予測対象の種類 2 個 (STAT と OTX) で計 10 個である．また，コーパス別の不均衡性，不均一性の値を表 5，6 に示す．学習用データにおいて，全コーパスの不

表 2　Tukey HSD 検定による p 値および標本効果量 (STAT)

	s-BL(STAT)	IMB(STAT)	HET(STAT)
G-BL(STAT)	$p = 0.395(-0.050)$	$p = 0.973(-0.014)$	$p = 0.021(-0.091)$
s-BL(STAT)	-	$p = 0.664(0.036)$	$p = 0.554(-0.042)$
IMB(STAT)	-	-	$p = 0.067(-0.078)$

表 3　Tukey HSD 検定による p 値および標本効果量 (OTX)

	s-BL(OTX)	IMB(OTX)	HET(OTX)
G-BL(OTX)	$p = 0.867(-0.024)$	$p = 0.751(0.032)$	$p = 0.015(-0.094)$
s-BL(OTX)	-	$p = 0.288(0.056)$	$p = 0.119(-0.070)$
IMB(OTX)	-	-	$p = 0.000(-0.126)$

均衡性の絶対値，不均一性の値よりも値が小さいものを太文字で示した．

表 4　モデル毎のコーパス別平均 JSD

モデル	評価対象のコーパス			
	CIC	IRIS	TickTock	YI
CIC<STAT>	0.0241	0.0505	0.0862	0.0347
CIC<OTX>	0.0271	0.0464	0.0895	0.0374
IRIS<STAT>	0.0334	0.0567	0.0752	0.0425
IRIS<OTX>	0.0360	0.0466	0.0716	0.0438
TickTock<STAT>	0.0376	0.0634	0.0648	0.0393
TickTock<OTX>	0.0379	0.0533	0.0584	0.0415
YI<STAT>	0.0196	0.0549	0.0741	0.0156
YI<OTX>	0.0193	0.0520	0.0672	0.0162
全コーパス <STAT>	0.0266	0.0518	0.0660	0.0213
全コーパス <OTX>	0.0288	0.0439	0.0562	0.0213

表 5　コーパス別の不均衡性

P の属する	$Imb(P)$		$\lvert Imb(P)\rvert$	
コーパス	学習用	評価用	学習用	評価用
CIC	0.1276	0.0273	0.1276	0.0273
IRIS	0.0670	0.0169	0.0670	0.0169
TickTock	0.0273	−0.1808	**0.0273**	0.1808
YI	−0.0573	−0.1197	0.0573	0.1197
全コーパス	0.0442	−0.0641	0.0442	0.0641

表 6　コーパス別の不均一性

P の属する	$Het(P)$	
コーパス	学習用	評価用
CIC	**0.0323**	0.0140
IRIS	0.0525	0.0462
TickTock	0.0658	0.0621
YI	**0.0163**	0.0104
全コーパス	0.0434	0.0372

表 6 より，モデル選択手法 HET では評価用発話が IRIS または TickTock の時に全コーパスを学習データに用いたモデルを選択していることがわかる．更に，表 4 より，モデル選択手法 HET<OTX> では評価用発話のコーパスが IRIS または TickTock の場合，より良いモデルを選択していることがわかる．

評価用発話のコーパスが YI の場合，モデル YI<OTX> よりモデル YI<STAT> の方が低い平均値を示しており，破綻ラベルの確率を予測する際に，破綻ラベルを実数値にマッピングして求めた平均と分散の値を介して予測した方がより性能が向上していることがわかる．

評価用発話のコーパスが CIC の場合，学習データに CIC を用いたモデルよりも YI を用いたモデルの方が低い平均値を示している．ここで，モデル YI<OTX>，YI<STAT> とモデル CIC<STAT> を比較する．評価用発話の属するコーパスが CIC であるものだけに注目し，モデル CIC<STAT> の JSD の値からモデル YI<OTX>，YI<STAT> の JSD の値を引いた値を評価用発話の対話内での出現位置毎に平均値を求めたものを図 1 に示す．図 1 より，差の値の平均値が大きくなっている発話の出現位置は 1 である．差の値が大きくなるということは，モデル CIC<STAT> の予測した破綻ラベルの確率分布が正解の確率分布と大きく異なっているか，モデル YI<OTX>，YI<STAT> の予測した破綻ラベルの確率分布が正解の確率分布と非常に近い分布になっているかである．実際に各モデルの予想した確率分布を見たところ，CIC コーパスに属する対話内での出現位置が 1 な評価用発話の破綻ラベルの確率分布の予測において，モデル CIC<STAT> の予測した破綻ラベルの確率分布が大きく異なっていることが多々見られた．ユーザの最初の発話に対するシステムの返答は性質が大きく異なることが考えられる．

6. 結論と今後

本研究では，対話破綻検出チャレンジの英語データセットに対して，評価用発話の属するコーパスによってコーパス単位での学習データの選別を試みた．評価指標として JSD を用いた時に，全てのコーパスの学習データを学習に用いるよりも，3.3 節に示した提案手法により学習データを選別した時の方が性能が向上し，統計的に有意な差が得られた．コーパス単位での学習データ選別では性能の向上が確認できたので，今後，対話単位，発話単位での学習デー

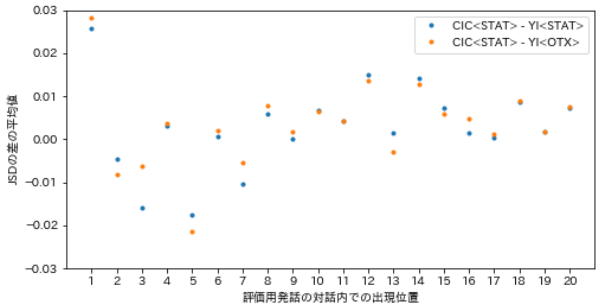

図 1 評価発話の対話内での出現位置と JSD の差の平均値の関係

タの選別も検討したい．

評価用発話のコーパス毎に各モデルのスコアを見たところ，CIC コーパス内の評価用データに関して，学習データとして CIC のみを用いたモデルよりも YI のみを用いたモデルの方が良いスコアを示した．対話内での出現位置が 1 である評価用発話に対して，予測した確率分布が大きく外れていることが多々見られた．本研究では，学習データの選別に評価用発話の属するコーパス情報のみを用いたが，評価用発話の対話内での出現位置も考慮した学習データの選別方法が必要であると考えられる．

参考文献

[1] Allan, J., Wade, C. and Bolivar, A.: Retrieval and Novelty Detection at the Sentence Level, Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 314–321 (2003).

[2] Higashinaka, R., Funakoshi, K., Inaba, M., Tsunomori, Y., Takahashi, T. and Kaji, N.: Overview of Dialogue Breakdown Detection Challenge 3, Proceedings of Dialog System Technology Challenge 6 (DSTC6) Workshop (2017).

[3] Kato, S. and Sakai, T.: RSL17BD at DBDC3: Computing Utterance Similarities based on Term Frequency and Word Embedding Vectors, Proceedings of Dialog System Technology Challenge 6 (DSTC6) Workshop (2017).

[4] Li, S., Zhou, G., Wang, Z., Lee, S. Y. M. and Wang, R.: Imbalanced sentiment classification, Proceedings of the 20th ACM international conference on Information and knowledge management (2011).

[5] Omari, A., Carmel, D., Rokhlenko, O. and Szpektor, I.: Novelty Based Ranking of Human Answers for Community Questions, Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 215–224 (2016).

[6] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, Vol. 12, pp. 2825–2830 (2011).

[7] Robertson, S. and Spärck Jones, K.: Simple, proven approaches to text retrieval, Technical Report UCAM-CL-TR-356, University of Cambridge, Computer Laboratory (1994).

[8] Sakai, T.: Statistical Reform in Information Retrieval?, SIGIR Forum, Vol. 48, No. 1, pp. 3–12 (2014).