アノテーション分布を考慮した対話破綻検出

河東　宗祐†　　酒井　哲也†

† 早稲田大学基幹理工学研究科　〒 169–8555 東京都新宿区大久保 3–4–1
E-mail: †sow@suou.waseda.jp, ††tetsuyasakai@acm.org

あらまし　近年，対話システムの実用化に向けて様々な研究が行われている．その中に，対話システムの精度向上にむけて，対話の破綻を検出する試みがある．本論文では，学習用データ中の正解ラベルの分布に注目して対話の破綻検出を試みる．データセットとして対話破綻検出チャレンジ 2 の学習用データと評価用データを用いる．データは，人間とシステムの対話集合からなり，対話内の各発話に複数のアノテータによって破綻の正解ラベル（破綻ではない・破綻とは言い切れないが違和感を感じる・あきらかにおかしい）が付与されている．正解ラベル上の確率分布やアノテータ毎の正解ラベルを考慮することで破綻検出の精度が向上するのか検証する．

キーワード　対話システム，破綻検出

1. はじめに

人工知能の発展のために対話システムの研究は重要であり，実用化に向けて様々な研究が行われている．近年，対話破綻検出チャレンジ [10,11] という，対話システムとユーザ間の破綻を自動検出することを目的にした評価型ワークショップが行われた．対話システムの出力の破綻を事前に検出することができれば，より実用的な対話システムの実現が可能になると考えられる．対話破綻検出チャレンジで提供されるデータは，人間とシステムの対話集合からなり，対話内の各システム発話に，図 1 に示すように，破綻ラベル (O,　T,　X) が複数の “annotator-id” を持つアノテータにより付加されている．また，データは学習用データと評価用データに分けられ，評価用システム発話の破綻ラベルの確率分布を予測するタスクとなっている．

本稿では，正解ラベルの分布に注目して対話の破綻検出手法を検討する．各破綻ラベルに相関があることを考慮した手法，アノテーションの難易度を考慮した手法とアノテータ毎の正解ラベルを考慮した手法を提案する．データセットとして，対話破綻検出チャレンジ 2 [10](注1) の学習用データと評価用データを用いる．2. 節で関連研究，3. 節で提案手法を説明し，4. 節で実験と結果，5. 節で分析と考察，6. 節でまとめと今後の課題を述べる．

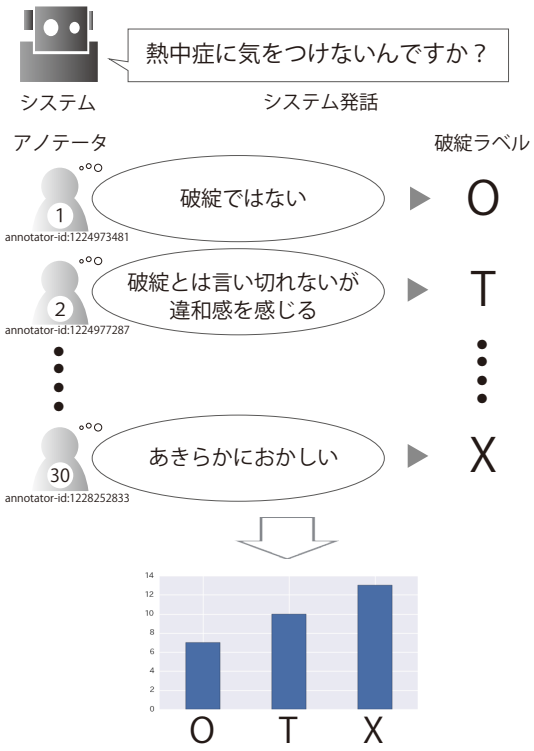


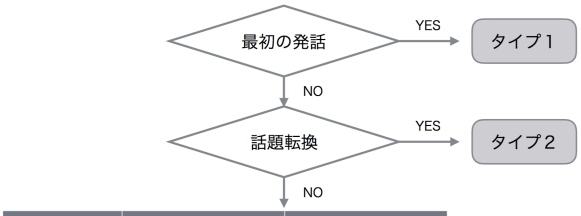
図 1　システム発話とアノテーション

2. 関連研究

対話破綻検出チャレンジ 2 では，対話破綻検出アルゴリズムの汎用性を考慮し，3 つのシステム (DCM，DIT，IRS) を用いた，学習用，評価用データセットが作られている．また，各システム発話に付与された破綻ラベルから破綻ラベルの確率分布が計算され，その確率分布との分布間距離，Jensen-Shannon Divergence (JS) や平均二乗誤差 (MSE) が主な指標とされている．河東ら [6] のラン 2 では，システム発話を図 2 に示す 6 種類の発話タイプに分類した上で破綻確率を予測している．発話タイプは，対話内で何番目の発話か，話題展開の発話であるか，1 つ前のユーザ発話が質問文であるかに基づき決定している．分類後，評価用発話と同じタイプの学習用データ内のシステム発話の集合の破綻確率の平均値を用いて O, T, X 各ラベル間の確率をそれぞれ計算し破綻確率予測を行っている．そのため，各ラベル間の相関が考慮されていない．対話破綻検出チャレンジ 2 において，最も安定して高い性能を示した杉山 [9] は，発話から特徴量を抽出し，Random Forest の派生形である Extra Trees Regressor を用いて対話破綻検出を行っている．また，データ量を増やした時に Deep Neural Network などの抽象度の高いアルゴリズムが利用できる可能性に言及してい

(注1)：https://sites.google.com/site/dialoguebreakdowndetection2/

る．Cho ら [2] は翻訳システムに Recurrent Neural Network (RNN) を用いているが，対話破綻検出チャレンジ 2 に参加した稲葉ら [5] や堀井ら [12]，久保ら [7] は対話破綻検出に RNN を用いている，更に，発話のエンコーディングとは別に，破綻確率の分布を出力する部分で，堀井ら [12] は 3 層のパーセプトロンを，稲葉ら [5] は 2 層の Recurrent Neural Network を用いており，破綻ラベル間の相関を考慮していると考えられる．また，確率分布を Neural Network のモデルとして学習するものに Kingma ら [3] の研究がある．しかし，複雑な機械学習モデルを用いた時の失敗分析では，モデルが適切でないからか，データ量が十分でないからか判断が難しい．



	出現位置が2	出現位置が4以上
質問文である	タイプ3	タイプ4
質問文でない	タイプ5	タイプ6

図 2　発話タイプの分類

アノテーションの難易度に関連して，Li ら [4] は文書間の関連の有無を判別するタスクにおいて，正解ラベルのばらつきとタスクのパフォーマンスに相関があることを示している．

3. 提 案 手 法

本稿では，ルールベースな発話タイプの分類に基づく河東らのラン 2 の手法をベースに，以下の手法を提案する．各発話のアノテーションの分布に注目した手法とアノテーションの難易度に注目した手法，各対話のアノテータ毎の正解ラベルに注目した手法の 3 つを示す．

3.1　発話毎のアノテーション分布

各破綻ラベル間で相関があることを確認し，発話毎のアノテーション分布をベータ分布と仮定することによって，ベータ分布のパラメータを用いて，各破綻ラベルの確率を予測する．

まず，各破綻ラベル間の相関を確認するため，学習データ内の破綻ラベル間の関係を分析する．学習用データ内の各システム発話に付与された破綻ラベルの数は合わせて 30 個であるが，O ラベルの数と T ラベルの数の関係を表した散布図を図 3 に，O ラベルの数と X ラベルの数の関係を表した散布図を図 4 に，T ラベルの数と X ラベルの数の関係を表した散布図を図 5 に示す．例えば，図 3 の左側，最も高い位置にプロットされている点は，学習用データ内に，O ラベルが 2 個，T ラベルが 25 個付与された発話を表している．データ内には対話を始めるためのシステム発話 (全てのラベルが O ラベルである発話) も含んでいる．また，各破綻ラベル間のピアソンの相関係数および 95% 信頼区間を表 1 に示す．ラベル O・T 間の上限値を見ても，強い相関があるとは言えない．また，ラベル T・X 間の上限値を見ても，ほとんど相関がないと考えられる．しかし，図 3，5 において，T ラベルの数が 5 以下のものに注目すると，O ラベルの数，X ラベルの数ともに 0 または 30 に偏っていることが見てとれる．各破綻ラベル間になんらかの相関があることがわかる．

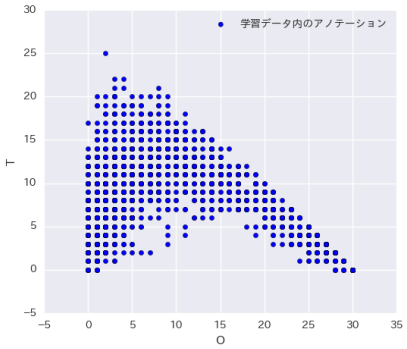


図 3　破綻ラベル数 O・T 間の散布図

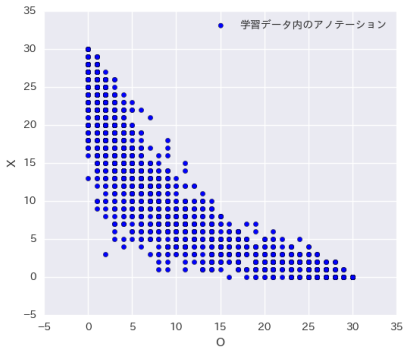


図 4　破綻ラベル数 O・X 間の散布図

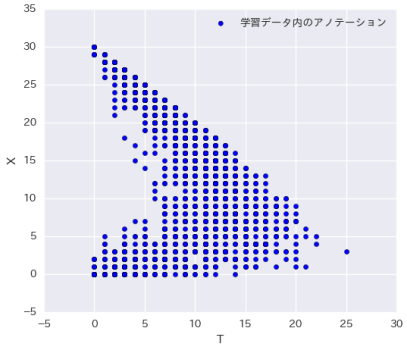


図 5　破綻ラベル数 T・X 間の散布図

次に，各破綻ラベルの確率の求め方を示す．1 つのシステム発話に注目した時のアノテーション分布，つまり破綻ラベルの数をアノテータの数で正規化した確率分布であるが，この確率分布がなんらかの確率分布に従うと仮定しパラメータを求める．

表 1　破綻ラベル間の相関係数

破綻ラベル	相関係数 [95% 信頼区間]
O・T	−0.48 [−0.51, −0.44]
O・X	−0.86 [−0.87, −0.84]
T・X	−0.04 [−0.09, 0.00]

本実験では，上限下限があり，且つ非対称な分布を表現できる簡単な分布としてベータ分布を用いる．まず，ベータ分布のパラメータ α，β を求める．各破綻ラベルを [0,1] の値に変換する必要があるため，表 2(a) に示す代表値に置換する．

表 2　破綻ラベルの区間と代表値

破綻ラベル	代表値 (a)　区間 (b)	離散値 (c)
O	$\frac{1}{6}$　$[0, \frac{1}{3}]$	0
T	$\frac{1}{2}$　$(\frac{1}{3}, \frac{2}{3}]$	1
X	$\frac{5}{6}$　$(\frac{2}{3}, 1]$	2
発話 u	X_u	L_u

あるシステム発話 u に付与された破綻ラベル 30 個を表 2(a) により代表値に置換したものを X_u とする．システム発話 u のパラメータ α_u，β_u は次式で計算する．

$$\alpha_u = \frac{E[X_u]^2(1 - E[X_u])}{V[X_u]} - E[X_u]$$
$$\beta_u = \frac{\alpha_u}{E[X_u]} - \alpha_u$$

ここで，$E[X_u]$ は X_u の平均，$V[X_u]$ は X_u の分散である．分散が 0 の時，つまり，アノテータによるラベルが全て一致した時は近似的に T ラベルを 1 つ追加した (破綻ラベルが全て T ラベルな発話は学習データ内には存在しなかった)．ニュートン法を用いて確率分布のパラメータを推定する方法 (注2) があるが，パラメータが 2 つなので用いなかった．

ある評価用発話の発話タイプと同タイプの学習データ内のあるシステム発話集合 U が与えられた時に，ベータ関数の平均パラメータ $\overline{\alpha} = \frac{1}{|U|}\sum_{u \in U} \alpha_u$，$\overline{\beta} = \frac{1}{|U|}\sum_{u \in U} \beta_u$ を求め，各破綻ラベルの確率 p_O，p_T，p_X を表 2(b) の区間を用いて次のように求める．

$$p_O = I(\frac{1}{3}|\overline{\alpha}, \overline{\beta})$$
$$p_T = I(\frac{2}{3}|\overline{\alpha}, \overline{\beta}) - I(\frac{1}{3}|\overline{\alpha}, \overline{\beta})$$
$$p_X = 1 - I(\frac{2}{3}|\overline{\alpha}, \overline{\beta})$$

ここで，$I(x|\alpha, \beta)$ はパラメータが α，β のベータ分布の累積分布関数である．

(注2)：Estimating a Dirichlet distribution: https://tminka.github.io/papers/dirichlet/

3.2　アノテーションの難易度

Li らは，文書間の関連の有無を判別するタスクにおいて，正解ラベルのばらつきとタスクのパフォーマンスに相関があることを示している．本稿では，対話破綻検出タスクにおける正解破綻ラベルのばらつきがタスクのパフォーマンスに影響を与えうるのか検証する．アノテーションの難易度は，あるシステム発話に破綻ラベルを付与する際の難しさであるが，難易度が高ければ破綻ラベルにばらつきがでると仮定し，難易度を表す指標を考える．まず，Li らと同様にアノテーションの最大値と最小値の差を考える．学習データ内のシステム発話 u に関して，破綻ラベルを表 2(c) の離散値を用いて数値に変換したものを L_u とする．x 軸に L_u の平均値，y 軸に $\max(L_u) - \min(L_u)$ をとった散布図を図 6 に示す．ラベルの最大値・最小値の差が小さいほど，平均値の値は 0 または 2 に偏っているように見える．しかし，最大値・最小値の差は 0，1，2 のいずれかであるため，閾値を設けた時の選択肢が少ない．そこで，T ラベルの数をアノテーションの難易度の指標とすることを考える．学習データ内のシステム発話 u に関して，x 軸に L_u の平均値，y 軸に付与された破綻ラベル T の数をとった散布図を図 7 に示す．T ラベルの数が少ない時に，平均値の値は 0 または 2 に偏っていることがわかる．T ラベルの数をアノテーションの難易度の指標とし，簡単のため，ある閾値 θ_T を用いて，学習データを選別する．

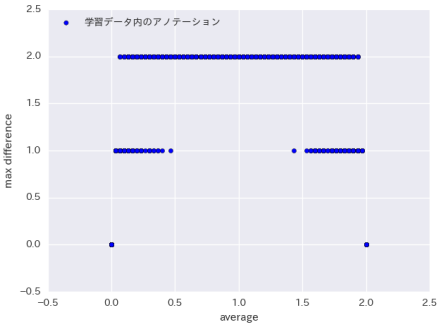


図 6　アノテーションの最大値・最小値の差と平均値の関係

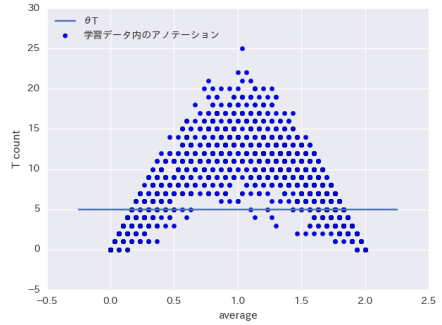


図 7　アノテーションの T ラベルの数と平均値の関係

ある評価用発話の発話タイプと同タイプの学習データ内のあるシステム発話集合 U が与えられた時に，次のように予測に

用いる集合 $U'(\subseteqq U)$ を選別する．

$$U' = \{u | n_T(u) < \theta_T, u \in U\}$$

ここで，$n_T(u)$ は発話 u に付与された T ラベルの数である．発話タイプ毎に選別された発話集合 U' を用いて，河東らのラン 2 と同様に各破綻ラベルの確率を求める．

3.3 アノテータ毎の学習データの選別

上記では，発話単位で学習に用いるデータを選別したが，アノテータ単位での学習データの選別を試みる．1. 節で述べたように，対話破綻検出チャレンジ 2 のデータセットでは，各システム発話の破綻ラベルに annotator-id が付与されている．簡単のため，同一の対話ログ内で同一の annotator-id を持つ破綻ラベルは同一のアノテータにより付与されたものだと考える．アノテータ i のスコア s_i を次式で求める．

$$s_i = \frac{1}{|D(i)|} \sum_{u \in D(i)} z(i,u)$$

ここで，$D(i)$ はアノテータ i が破綻ラベルを付与した対話ログ内のシステム発話集合である．上で用いた，$z(i,u)$ の説明を示す．$z(i,u)$ は L_u と L_u 内のアノテータ i により付与された破綻ラベルを数値に変換した値 l_i を用いて，次のように表せる．

$$z(i,u) = \frac{l_i - E[L_u]}{\sqrt{V[L_u]}}$$

学習データ内の破綻ラベルを付与している全てのアノテータに関して，アノテータのスコア s_i の分布を図 8 に示す．

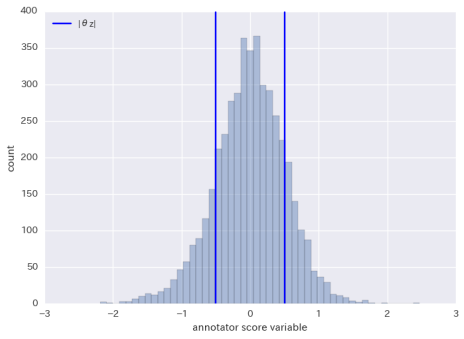


図 8　s_i の分布図

ある閾値 θ_z を用いて，$|s_i| > \theta_z$ となるようなアノテータ i の付与した破綻ラベルを全て無視し，河東らのラン 2 と同様に各破綻ラベルの確率を求める．

4. 実験と結果

表 3 に，3.1 小節に示した手法を手法 1，3.2 小節に示した手法を手法 2，3.3 小節に示した手法を手法 3，河東らのラン 2 の手法をベースラインとして，ラベル一致系統と分布間距離系統の結果を示す．Precision，Recall，F-measure ともに正解ラベルを X のみとしている．手法 2 で用いる閾値 θ_T は，図 3，5 をもとに $\theta_T = 5$ とした．手法 3 で用いる閾値 θ_z は，$\theta_z = \sqrt{V[S]}$ とした．ここで $V[S]$ は，学習用データ内の全てのアノテータのスコア s_i の集合 S の分散である．

次に，対話破綻検出チャレンジ 2 で主な指標とされた JS と MSE に関して，分散分析およびランダム化 Tukey HSD 検定 [1, 8] を行った．表 4，6 に 3 種の対話システム，4 種の手法を用いた表 3 の JS(O,T,X) および MSE(O,T,X) の値に対する，繰り返しのない 2 元配置の分散分析表を示す．有意水準を $\alpha = 0.05$ とするとき，JS(O,T,X) および MSE(O,T,X) の値に対する手法による効果，対話システムによる効果の p 値はそれぞれ $p = 0.007$，$p = 0.002$ および $p = 0.011$，$p = 0.004$ であり，JS(O,T,X)，MSE(O,T,X) のどちらの値に対しても，手法，対話システムによる効果共に統計的に有意となる．4 種の手法の各対の JS(O,T,X)，MSE(O,T,X) の平均値の差についてランダム化 Tukey HSD 検定を施行回数 $B = 10000$ で行ったところ，p 値は表 5，7 のようになった．また，各表の括弧内に標本効果量として $\overline{d}/\sqrt{(V_E)}$ の値を示した．ここで，$\overline{d}$ は各表毎に各システム対の JS(O,T,X)，MSE(O,T,X) の平均値の差，V_E は各表毎に表 4，6 の繰り返しのない 2 元配置の分散分析表から得た誤差分散である．統計的には，いずれの提案手法もベースラインより優れているとは言えなかった．

5. 分析と考察

手法 1，2，3 に関して行った分析を示す．

5.1 手　法　1

図 9，10 に，各対話システム (DCM，DIT，IRS) について，評価用データの正解ラベルとベースラインおよび手法 1 による予測ラベルのヒートマップを示す．ヒートマップの横軸が正解ラベル，縦軸が予測ラベルである．左上から右下の対角線に近いセルの色が濃いほど，正解ラベルと予測ラベルの分布間距離が近いことを示している．

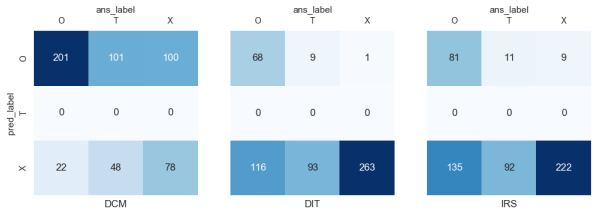


図 9　ベースラインの予測ラベルと正解ラベルの関係

図 10　手法 1 の予測ラベルと正解ラベルの関係

表 3　ラベル一致系統と分布間距離系統

対話システム	手法	Precision(X)	Recall(X)	F-measure(X)	JS(O,T,X)	MSE(O,T,X)
DCM	手法 1	0.000	0.000	0.000	0.119	0.070
	手法 2	0.000	0.000	0.000	0.180	0.095
	手法 3	**0.527**	**0.438**	**0.479**	**0.094**	**0.049**
	ベースライン	**0.527**	**0.438**	**0.479**	0.095	**0.049**
DIT	手法 1	**0.724**	0.496	0.589	0.064	0.037
	手法 2	0.557	0.947	0.701	0.101	0.056
	手法 3	0.706	0.545	0.615	0.056	0.032
	ベースライン	0.557	**0.996**	**0.715**	**0.053**	**0.030**
IRS	手法 1	0.450	0.329	0.380	0.150	0.084
	手法 2	0.491	**0.974**	**0.653**	0.154	0.079
	手法 3	**0.494**	0.961	**0.653**	**0.109**	**0.058**
	ベースライン	**0.494**	0.961	**0.653**	**0.109**	**0.058**

表 4　JS(O,T,X) 値の分散分析表

変動要因	平方和	自由度	平均平方	F 値
手法間変動	$S_A = 0.0070$	$\phi_A = 4-1$	$V_A = 0.0023$	$F_0 = 11.567$
対話システム間変動	$S_B = 0.0090$	$\phi_B = 3-1$	$V_B = 0.0045$	$F_0 = 22.255$
誤差間変動	$S_E = 0.0012$	$\phi_E = (4-1)(3-1)$	$V_E = 0.0002$	
全変動	$S_T = 0.0173$			

表 5　JS(O,T,X) 値のランダム化 Tukey HSD 検定 (施行回数 $B = 10000$ 回) による p 値，および標本効果量 $\overline{d}/\sqrt{V_E}$

	手法 2	手法 3	ベースライン
手法 1	$p = 0.406(-2.4094)$	$p = 0.648(1.7166)$	$p = 0.628(1.7608)$
手法 2	-	$p = 0.033(4.1260)$	$p = 0.029(4.1701)$
手法 3	-	-	$p = 1.000(0.0442)$

表 6　MSE(O,T,X) 値の分散分析表

変動要因	平方和	自由度	平均平方	F 値
手法間変動	$S_A = 0.0020$	$\phi_A = 4-1$	$V_A = 0.0007$	$F_0 = 9.512$
対話システム間変動	$S_B = 0.0023$	$\phi_B = 3-1$	$V_B = 0.0011$	$F_0 = 16.220$
誤差間変動	$S_E = 0.0004$	$\phi_E = (4-1)(3-1)$	$V_E = 0.0001$	
全変動	$S_T = 0.0047$			

表 7　MSE(O,T,X) 値のランダム化 Tukey HSD 検定 (施行回数 $B = 10000$ 回) による p 値，および標本効果量 $\overline{d}/\sqrt{V_E}$

	手法 2	手法 3	ベースライン
手法 1	$p = 0.612(-1.5740)$	$p = 0.430(2.0366)$	$p = 0.385(2.1402)$
手法 2	-	$p = 0.041(3.6106)$	$p = 0.021(3.7142)$
手法 3	-	-	$p = 1.000(0.1036)$

図 10 の左 (DCM) のヒートマップを見ると，予測ラベルが T ラベルに偏っており X ラベルを 1 つも予測していない．表 3 において，手法 1 の Precision，Recall，F-measure が 0 であることに対応している．各図を比較すると，ベースラインに比べ手法 1 は予測ラベルが T ラベルに寄る傾向があることがわかる．全ての対話システムに共通して，正解が X ラベルな発話を T ラベルと多く予測してしまっている．各ラベルを実数値に置換する際の $[0, 1]$ 内での最適な区間を検証する必要がある．

5.2 手　法　2

図 11 に，各対話システム (DCM，DIT，IRS) について，手法 2 の $\theta_T (1 \leqq \theta_T \leqq 31)$ を変化させた時の分布間距離の平均値を示す．図 11 における JS'(O,T,X)(青) と MSE'(O,T,X)(緑) を説明する．手法 2 において，T ラベルの数が閾値 θ_T を上回っている発話を学習データから除外したが，評価用発話も同様に閾値 θ_T を上回っている発話を除外し，分布間距離の平均値を求めたものが，JS'(O,T,X) および MSE'(O,T,X) である．

JS，MSE の値は $\theta_T = 31$ の時にベースラインによる値と一致するはずであるが，図 11 の JS(O,T,X) と MSE(O,T,X) を見ると，右肩下がりのように見え，手法 2 が有用とは考えにくい．しかし，JS(O,T,X) と JS'(O,T,X)，MSE(O,T,X) と MSE'(O,T,X) はある θ_T を境に値が乖離しており，JS'(O,T,X)，MSE'(O,T,X) の値は $\theta_T = 1$ 付近で，ベースライン ($\theta_T = 31$)

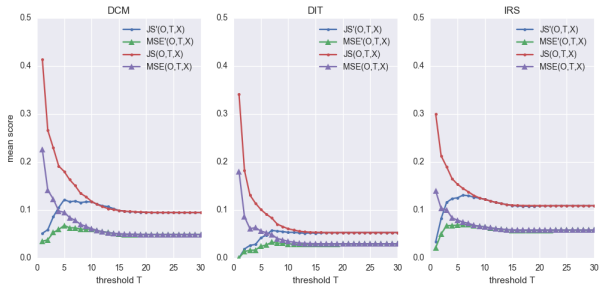


図 11　θ_T と各指標の関係

より分布間距離の平均値が下回るのがわかる．T ラベルの数が少ない発話に限定すると予測性能の向上が見込める．評価用発話の T ラベルの数を予測し 0 付近かそうでないかによって，予測アルゴリズムを変更することが考えられる．

5.3 手　法　3

図 12 に，手法 3 の $\theta_Z(0.01 \leqq \theta_Z \leqq 2.00)$ を変化させた時の JS(O,T,X)，MSE(O,T,X) の値を各対話システムについて示す．左が JS(O,T,X)，右が MSE(O,T,X) の値である．θ_Z はステップ 0.01 で変動させた．また，表 8 に変動させた θ_Z 内での最小値および括弧内に最小値を記録した θ_Z を示す．

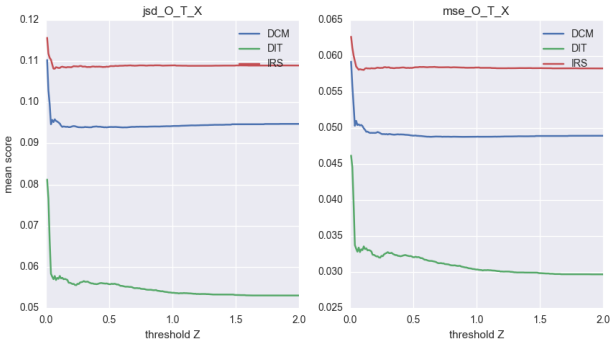


図 12　θ_Z と各指標の関係

表 8　JS，MSE の最小値と θ_Z

	JS(O,T,X)	MSE(O,T,X)
DCM	0.094($\theta_Z = 0.61$)	0.049($\theta_Z = 0.88$)
DIT	0.053($1.67 \leqq \theta_Z \leqq 1.71$)	0.030($1.90 \leqq \theta_Z \leqq 2.00$)
IRS	0.108($\theta_Z = 0.07$)	0.058($\theta_Z = 0.11$)

DCM と IRS に関する JS(O,T,X) の最小値はベースラインをわずかながら下回っていることがわかる．極小値があるのであれば，その時の θ_Z を予測できれば，分布間距離を小さくすることができることがわかる．どのようなラベル付けをしたアノテータが手法 3 により除外されたか細かく分析する必要がある．

6.　まとめと今後の課題

今回の実験では提案手法によりベースラインを統計的に有意に上回る結果が得られなかったが，改良の可能性は確認できた．

破綻ラベル間の相関を考慮した手法 1 では，アノテーション分布をベータ分布と仮定し，表 2(a)(b) の代表値と区間を用いて破綻確率を予測した．予測ラベルが T ラベルに偏ってしまう問題が確認できた．アノテーションの難易度を考慮した手法 2 では，アノテーションの難易度の指標として T ラベルの数を用いたが，アノテーションの難易度が破綻ラベル予測の難易度に影響を与えていることが確認できた．アノテータ毎の正解ラベルを考慮した手法 3 では，個のアノテータのラベル付けにおいて，周りのアノテータのラベルとどれほど逸脱しているかにスコア付けをして，閾値を用いてアノテータを除外し破綻確率を予測した．対話システムや指標により，ベースラインよりも破綻確率の性能を向上させる閾値が存在することを確認できた．

手法 1 に関しては，ラベルを置換する代表値と区間を調整することや，様々な確率分布を仮定し検証する必要がある．手法 3 に関しては，適切な閾値を求めることができるのか検証する必要がある．

文　　献

[1] Benjamin A. Carterette. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM TOIS*, Vol. 30, No. 1, 2012.

[2] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *arXiv preprint arXiv:1406.1078*, 2014.

[3] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *arXiv preprint arXiv:1312.6114*, 2013.

[4] Jiyi Li and Masatoshi Yoshikawa. Evaluation with confusable ground truth. In *Information Retrieval Technology*, pp. 363–369. 12th Asia Information Retrieval Societies Conference, Springer International Publishing, 2016.

[5] 稲葉通将, 高橋健一. RNN エンコーダによる文脈を考慮した対話破綻検出. 人工知能学会 言語・音声理解と対話処理研究会 SIG-SLUD-B505-27, pp. 98–101, 2016.

[6] 河東宗祐, 酒井哲也. word2vec による発話ベクトルの類似度を用いた対話破綻予測. 人工知能学会 言語・音声理解と対話処理研究会 SIG-SLUD-B505-20, pp. 70–71, 2016.

[7] 久保隆宏, 中山光樹. Neural conversational model を用いた対話と破綻の同時学習. 人工知能学会 言語・音声理解と対話処理研究会 SIG-SLUD-B505-26, pp. 94–97, 2016.

[8] 酒井哲也. 情報アクセス評価方法論: 検索エンジンの進のために. コロナ社, 2015.

[9] 杉山弘晃. 発話生成における誤りパターンの分析に基づく対話破綻検出. 人工知能学会 言語・音声理解と対話処理研究会 SIG-SLUD-B505-23, pp. 81–84, 2016.

[10] 東中竜一郎, 船越孝太郎, 稲葉通将, 荒瀬由紀, 角森唯子. 対話破綻検出チャレンジ 2. 人工知能学会 言語・音声理解と対話処理研究会 SIG-SLUD-B505-19, pp. 64–69, 2016.

[11] 東中竜一郎, 船越孝太郎, 小林優佳, 稲葉通将. 対話破綻検出チャレンジ. 人工知能学会 言語・音声理解と対話処理研究会 SIG-SLUD-075-07, pp. 27–32, 2015.

[12] 堀井朋, 森秀晃, 林卓矢, 荒木雅弘. 破綻類型情報に基づく雑談対話破綻検出. 人工知能学会 言語・音声理解と対話処理研究会 SIG-SLUD-B505-22, pp. 75–80, 2016.