対話破綻検出チャレンジにおける対話破綻データと破綻検出結果の分析

－主観性の高い言語データにおける言語処理に関して－

船越孝太郎（(株）ホンダ・リサーチ・インスティチュート・ジャパン）・東中竜一郎（NTT メディアインテリジェンス研究所）・稲葉通将（広島市立大学）・小林優佳（(株）東芝 ）・菅原朔（東京大学）・高梨克也（京都大学）・大塚裕子（公立はこだて未来大学）・小磯花絵（国立国語研究所）・坊農真弓（国立情報学研究所）

1 はじめに

我々は 2014 年度に行われた Project Next NLP エラー分析プロジェクトにおいて，対話タスクとして対話破綻に関するデータの収集と分析を行った [2]．そしてそれを受けた対話破綻検出技術に関する参加型ワークショップを 2015 年に開催した [3]．本稿では，参加型ワークショップ開催に際して新たに収集した 30 名による対話破綻アノテーション結果の分析と，ワークショップ参加者 6 チームによる破綻検出結果のメタ的な分析について報告する．対話破綻の厳密な定義が難しいこともあり，その認定は主観性が高く，個人間での揺らぎが大きい．本稿では対象データのそのような性質を意識しながら，分析と考察を行う．

2 データ

本稿で分析の対象となるデータは 2 種類ある．1 つは，一般から募集した [1]1 対話あたり 30 名によって対話破綻ラベルを付与したラベル付き対話データで，これを「対話破綻データ」とよぶ．もう 1 つは，対話破綻検出チャレンジにおいて評価用データとして使用した「対話破綻データ」に対して，6 つのチャレンジ参加チームが提出した総数 15 の「破綻検出結果」（それぞれを run とよぶ）である．どちらもより詳細な説明が [3] にあるが，以下に概要を示す．

2.1 対話破綻データ

クラウドソーシングで募集した不特定多数のユーザによる 100 の雑談対話を集めた．各ユーザは指定された web サイトにブラウザでアクセスして，20 ターンの雑談を行った．このうち 20 対話を開発用データ (**dev**) とし，80 対話を評価用データ (**test**) とした．本稿での分析はこの **test** を対象とする．

評価用データ中の各システム発話には，別のクラウドソーシングサービスによって募集したアノテータ 30 名により，○・△・×の 3 段階で，対話破綻のラベルが付与されている．アノテータは対話単位での募集なので，対話毎にアノテータの集団は異なるものの，1 つの対話中の 10 のシステム発話は同一の 30 名によりアノテーションされている．

○・△・×の基準は，以下の通りで，アノテータの主観を重視している．

○ **破綻ではない** 当該システム発話のあと対話を問題無く継続できる．

△ **破綻と言い切れないが，違和感を感じる発話** 当該システム発話のあと対話をスムーズに継続することが困難．

× **あきらかにおかしいと思う発話．破綻** 当該システム発話のあと対話を継続することが困難．

[1]学習用データとして提供した init100 および rest1046 データセット [2] は対話研究者および関係業務に携わる者によってアノテートされている．

評価用データに対する正解ラベルは 30 名の多数決を基本としているが，閾値 t によって△・×となる発話を調整できるようにしている．すなわち，最多数を占めるラベルを求め，その割合が閾値 t 以上の場合，その最多数ラベルを正解ラベルとする．閾値を超えるものがない場合は，正解ラベルを○としている．

次節以降の分析では $t = 0.5$ と設定した．ラベルが△あるいは×の場合，30 名のうち 15 名以上がそのラベルを選択していることになる．

2.2 破綻検出結果

破綻検出結果は，前述の評価用データに対して，対話破綻チャレンジに参加した 6 チームが 1 チーム最大 3 つの破綻検出処理の結果を提出したものである．合計 800 のシステム発話に対して，○・△・×のいずれかのラベルと，それぞれの確率分布が出力されている．各チームの手法はルールベースから深層学習に基づくものまで多様であり，各チームによる解説が [1] に収録されている．評価用データに対するラベルアノテーションおよび正解ラベルは，破綻検出結果提出時には公開されていない．

3 破綻アノテーションの分析

3.1 ラベルの妥当性の検証

今回の多数決に基づく破綻ラベルの決定が妥当な結果になっているかどうかを調べるため，まずは 1 名（著者外の協力者）による主観判断で，多数決により決定した正解ラベルが許容できるかどうかを確認した．

閾値 $t = 0.5$ の設定で定めた正解ラベルをもとに確認を行い，対象となる 800 箇所のうち，32 箇所（4%）のみが違和感を感じる，という結果であった．まずこの段階で，96%は（少なくとも確認をした 1 名にとっては）問題ない結果であったので，正解ラベルは概ね妥当という結果が示唆されたといえる．

残り 4%についても，$t = 0.5$ という閾値設定が問題となっていた．例えば，以下のシステム発話では，質問に対して無関係な質問を返してしまっているが，○のラベルがついている．30 名のアノテーションの内訳をみると，○・△・×が 5，14，11 となっていて，最も数の少ない○になってしまっている．

ID:1437630091
01 U: こんにちは、あなたの名前は？
02 S: ご家族とは一緒に暮らしていますか？

閾値 t はもともと，破綻（×および△）の検出という前提において，アノテータ間で一致の程度が低いものを評価データとしての検出対象から除外するために導入したものであるので，○を含めたデータの分析や，学習用データと

して利用する際には注意が必要である．このような目的では，単純に多数決で正解ラベルを決定したり，アノテーションの内訳（ラベルの頻度分布）を参照しながらデータを用いるのが適切である．実際，後で見るように，学習用データとして使用する場合は，1 つに確定した正解ラベルを用いるよりも，頻度（確率）分布を利用するほうがよい結果を得られる見込みがある．

一連の取り組みにおいて，我々は「破綻」の定義を 2.1 節で示したように主観的なものにとどめており，アノテータ間での揺れが大きいことは否めないが，今回の妥当性の検証からも，システムが行った不適切な応答例を収集するという目的に対して，上記の破綻の定義に基づくデータ構築と多数決による正解決定は，大筋としては機能したとみることができる．

一方で，上記の ID:1437630091 の事例からもわかるように，細部においては検討が必要な事項が存在することも確かである．上記の例は，質問に対して無関係な質問を返す，という一見明確な破綻（×）の事例のように見えるが，実際には△が最多数を占めている．実際，今回の「そのあと対話を続けられるか？」という破綻の定義だけから見れば，自分が質問していた流れを放棄しさえすれば対話を続ける（家族と一緒に暮らしているかどうか答える）ことは問題なくできるので,「○」であってもおかしくはないといえる．しかしながら，質問に対して全然関係ない質問を返すのは、常識的な会話の「規範」からははずれていると思われ，システムが行うべき適切な応答ではないであろう.「規範からのずれ」と「破綻」の関係をどう考え，定義に反映するかは，今回の分析から見出された課題の 1 つである．

3.2 破綻の種類と分布

閾値 $t = 0.5$ で定めた 80 対話中の破綻箇所（△および×）について，対話破綻類型 [2] に基づき予備的な分類を行った．2 名の作業者が [2] 中の類型の説明だけを元に独立に作業を行い，事前の意識合わせ等は行っていない．アノテーションの一致率を高めることよりも，どのような破綻が多いのかを知ることを優先したため，大分類の間での優先順位も定めていない．

表 1 に結果を示す．それぞれの結果を，作業者 A および B の列に示す．判断に迷う場合に複数の類型に分類することを許したので，それぞれの列の頻度の総和は，破綻の数と一致しない．

A と B の頻度の傾向の違いをみると，発話・文脈の大分類については，A と B の判断に大きな乖離はなさそうにみえる．違いが目に付くのは，応答の大分類で，情報過不足と不理解の頻度が逆転しているように見える部分である．また，環境についても A と B で逆転しているように見える．これらの部分については，類型の定義付けに関して，詳しい検討が必要であろう．

発話・文脈の大分類についても，細かくみると A と B で異なる発話に付与していることがまま観察されるので，頻度的に乖離が少ないからといってよく一致しているようではなさそうであり，注意が必要である．

破綻の類型化がそれなりの一致度と整合性をもって行えることは，破綻検出を部分問題に分解してアプローチし，技術的ボトルネックを特定していく上で重要であるので，類型の定義と分類に当たってのガイドラインのさらなる改善が必要であろう．そのためには，主観性／客観性という概念自体の整理 [4] も必要となるかもしれない．

4 破綻検出システムの結果の分析

各参加チームの手法および run 毎の検出性能は [3] にまとめられているが，表 2 に各チームが採用した手法の概要を示す．ここでは，対話破綻データと破綻検出結果の関係について，全体の傾向を分析する．以下，システムという語で各 run の結果を出力した検出器を指す．

表 1: 類型毎の頻度

大分類	類型名	作業者 A	作業者 B
発話	構文的誤り	11	4
	意味的誤り	33	18
	解釈不能	2	7
応答	情報過不足	205	9
	不理解	22	100
	無関係	118	74
	意図不明	78	78
	誤解	4	0
文脈	不要情報	41	35
	矛盾	13	8
	無関係話題	5	3
	関連性不明	12	18
	不追随	43	16
環境	共通基盤欠如	36	0
	一般常識欠如	11	0
	社会性欠如	2	10

4.1 破綻ラベルのエントロピーとシステムの正答率の相関関係

正解ラベルは 30 名のアノテータの多数決によって決まるが，その数（一致の度合い）にはばらつきがある．表 3 に示すように，×ラベルが正解として付与されていても，それは 30 名中 15 名（50%）の判断にすぎない場合もあれば，30 名中 27 名（90%）の判断による場合もある．

一致の度合いが高いものは人間にとって破綻であることの判断が易しいものと考えられ，それがシステムにとっても同じであるかどうかを調べる．具体的には，破綻ラベルのエントロピーとシステムの正答率（15 の run のうち，いくつが正解しているか）との相関関係を見る．負の相関が認められれば，エントロピーが低い（ラベルの偏りが大きい）ほど正答率が高いということであり，人間とシステムの間で判断の難易が似ていることが示唆される．

$t = 0.5$ での×ラベルのデータに関して，相関係数を求めると -0.127 となる．非常に弱い負の相関であり仮説と矛盾はしないが，散布図（図 1）を見ても人の目でははっきりとした傾向は見出せない程度である．結論としては，人間の意見が一致しやすい破綻がシステムにとっても容易とはなっていないようである.（手法・システムによって傾向が大きく違う可能性は残るが，ここでは結果全体の分析に止める.）

4.2 システム間の破綻検出の傾向

人間の判断傾向とシステムの判断傾向には強い類似性はない模様であったが，システム間での判断の類似性はどうであろうか．すなわち，各システムは同じ問題に正答し，同じ問題で間違えやすい傾向があるであろうか．

この点について調べるために，破綻ラベル×が正解としてついている発話について，いくつの run が正解したかを数えた（表 4）．これによると，どのシステムも正答できなかったものは 9 つしかないが，半数以上の run が正答した発話もたかだか 30 個（15%弱）しかない．つまり回答の傾向は全体としてはあまり似ておらず，システム間で大きく異なる模様である．破綻検出技術としてのボトルネックを

表 2: 各チームの手法の概要（[3] より引用．各 run の違いは省略）

team	学習手法	素性
team1	RNN, LSTM-RNN	単語頻度ベクトル（BoW），発話応答間の単語の共起頻度ベクトル，単語頻度・共起頻度ベクトルを Sent2Vec を用いて変換したベクトル
team2	LSTM-RNN	単語頻度ベクトルを Word2Vec を用いて変換したベクトル
team3	ルール	形態素解析器 kuromoji[2] によって抽出したキーワード
team4	Poly kernel SVM	当該システム発話と 1 つ前のユーザ発話の単語頻度ベクトル
team5	6 層 DNN	対象文と直前の文の対話行為，直前の文から予測されたシステムが出すべき対話行為，パープレキシティ，自動評価値，システムパーソナリティを尋ねる質問か否かのフラグ
team6	RNN	単語頻度ベクトルを NCM，LSTM エンコーダ，Bag-of-Words Embedding，NCM の拡張モデルの 4 種類の方法で変換したベクトル
baseline	CRF	単語頻度ベクトル

(RNN:Recurrent Neural Network, LSTM:Long Short-Term Memory, DNN:Deep Neural Network, NCM:Neural Conversational Model)

表 3: 評価用データにおける×の数に対する発話数の分布

×の数	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	合計
発話数	27	30	21	21	17	17	9	12	15	5	11	10	3	0	0	0	198

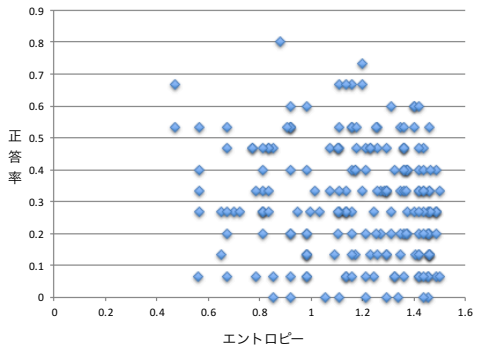

図 1: 発話毎のエントロピーと正答率の関係

議論するにはもう少し手法が成熟するのを待つ必要がありそうである．

4.3 システム統合の有効性の検討

システム間で傾向が大きく異なるという上記の結果から，単純に複数のシステムを統合することで，結果を改善できる可能性がある．そこで，システム統合による性能の変化を検討した．

表 5 に各実験の結果を示す．team5r2 は，[3] の表 3 に示されている結果の中で，破綻ラベル（×）についての性能が F 値で最も高かった team5 run2 を指す．以下では同表 3 中の team1 run1 から team6 run1 までを s1 から s15 までの ID で示す．

4.3.1 システムの OR 統合

×について単純にシステムの結果の論理和をとると，当然再現率は向上するが，精度は悪くなる．15 のシステムのうち 1 つでも×と判定した場合の結果が OR(s1,...,15) である．この場合 800 発話のうちの 710（89%）を破綻と判定していることになってしまう．

15 のシステムから 2 つ，3 つ，4 つを選んで OR で統合してみると，s2 (team1 run2), s3 (team2 run1), s6 (team3 run1), s15 (team5 run1) の 4 つのシステムを統合した場合に，F1 値で 0.491 まで改善できた．team5 run2 (s13) 単独にくらべると再現率は下がってしまうが，精度が高まる．

4.3.2 15 システムの多数決

15 のシステムのうち N 以上が×と判断した場合に，×とする方法も考えられる．表 3 で見たように，複数のシステムが×だと判定する場合の数はあまり多くないので，当然再現率は大きく下がると予想されるが，その分精度は改善することも期待できる．

N を 4 とした時に F 値で最もよい結果が得られた（VOTE(≥ 4)）が，大きな改善はなかった．

4.3.3 15 システムの出力を入力とする分類器

単純に OR や多数決を取るのではなく，システム毎に重みを変えて結果を統合することも考えられる．そこで提出されている結果を用いて，Weka[3] による 10 分割交差検証を行った．

単純に破綻検出結果と対応する正解ラベルから 3 値分類器を構築しても，○のラベルの数が多いために，それに引きずられて目標とする×の検出に焦点が合っていない結果しか得られない（○の出力傾向が高い学習結果になる）．そこで Weka のデータインスタンスへの重み付け機能を利用して，○・△・×の割合が均等になるように調整して学習を行った．また，各システムのラベル出力を単純に入力素性としても性能がでないため，各システムが出力する確率分布（システムによっては出力ラベルに対応した 1/0 を出力）を素性とした．これにより，SMO RBF-Kernel で F 値が 0.541 まで向上した（表 5 の SMO (RBF) [4]）．

4.3.4 同一手法による並列モデルの統合

ここまでは多人数によるアノテーションを多数決により縮退して，正解ラベルを 1 つに決定してから，それを学習データとして与え，集合知的な 1 人の仮想人格による判断を模倣するモデルを構築してきた．ここでは複数のアノテー

[3] http://www.cs.waikato.ac.nz/~ml/weka/

[4] 重み付けにより，×ラベルを持つインスタンスの数はおよそ 470 になるが，表 5 では比較しやすいように 198 にそろえて分数表示している．

表 4: ×が正解の発話に対して正答できたシステムの数の分布

正答システム数	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	合計
発話数	9	22	19	22	31	25	18	22	17	6	5	1	1	0	0	0	198

表 5: システム統合による破綻（×）の検出性能の比較

方式	Precision	Recall	F1
team5r2 (s13)	.33 (155/465)	.78 (155/198)	.48
OR(s1,. . . ,15)	.27 (189/707)	.96 (189/198)	.42
OR(s2,3,6,15)	.43 (112/258)	.57 (112/198)	.49
VOTE(≥ 4)	.37 (126/338)	.64 (126/198)	.47
SMO (RBF)	.47 (126/267)	.64 (126/198)	.54

表 6: 24 システムの統合による破綻（×）の検出性能

	unified			split		
機械学習手法	Pre.	Rec.	F1	Pre.	Rec.	F1
RandomForest	.390	.374	.382	.444	.399	.420
SMO (Poly-1)	.488	.338	.400	.455	.444	.450
SMO (RBF)	.480	.273	.348	.448	.444	.446

タ各人に対するモデルを学習し，それを統合して最終的な推定を行うことを試みる．

学習用データとして，init100 データセット [2] を用いる．init100 では，24 人のアノテータ全員が同じ 100 対話に対してラベルを付与している．まず baseline 手法（CRF）を用いて，init100 のデータから 24 人のアノテータそれぞれに対応する 24 個のシステムを構築する．そして，4.3.3 節と同じように，24 のシステムの結果を統合する分類器を構築する．このような学習のさせ方においては，多人数のアノテーション結果から正解を 1 つに定める必要がなくなる．

結果を表 6 に示す．unified は，24 人のアノテーションから正解ラベルを決め，それを元に 1 つのモデルを構築したものである．split は，24 人のアノテーションから 24 のモデルを構築したものである．条件を揃えるために，どちらも学習したモデルでラベルの確率分布を予測させて素性ベクトルとし（それぞれ 3 次元と 72 次元），それを元に，表の左列に示す機械学習手法で 10 分割交差検証を行った．

init100 の 100 対話のデータしか用いていないため，表 5 に示した結果（rest1046 の 1046 対話データを学習に用いている）に比べて全般に性能は低めであるが，3 つの機械学習手法で一貫して，split のほうが性能が高い．

この性能向上においてアノテータの個性・主観的な判断基準を保っていることが重要なのかどうかを調べるために，対話単位で 24 人の間のアノーテションをランダムに入れ替えて 24 個のモデルを構築し，表 6 の split と同じように，10 分割交差検証で性能を確認した．

結果を表 7 に示す．学習手法毎に性能が上下しているが，概ね split とほぼ同等のようである．このことから 24 個のモデルが 24 人の判断基準を保持している必要はないことが示唆される．これが正しいならば，対話毎に別々のアノテータを採用しても問題ないということになるので，クラウドソーシングなどを使った低コストのアノテーションを採用しやすくなる．

今回の実験では，最終的な評価について多数決で 1 つに縮退してしまった正解ラベルとの比較で行っているため，結局評価の段階で「正解を決める」という課題が発生してしまっている．[3] でも行ったように，分布と比較するようにすれば，最終結果についても「正解を決める」必要がなくなり，主観性の高いデータを扱う課題の枠組みとして，より都合がよいように思われる．

表 7: ランダマイズされたアノテーションに基づく 24 システムの統合による破綻（×）の検出性能

	shuffled		
機械学習手法	Pre.	Rec.	F1
RandomForest	.468	.460	.464
SMO (Poly-1)	.451	.419	.435
SMO (RBF)	.456	.475	.465

学習に使用できるデータ量が増えた場合や，並列数が少ない場合にも，今回と同様な結果が得られるかはまだ不明であるが，正解を 1 つにしてしまってモデルを作るよりも性能が向上するという現象が一般的に成り立ちうるのであれば，多人数による主観性の高いデータを扱う際の一般的なアプローチとして学習・評価時共に「正解を決めない」アプローチがスタンダードとなるかもしれない．

5 おわりに

本稿では，30 名による対話破綻アノテーションの結果と，それを評価に用いた 6 チーム 15 システムの振る舞いについて，主観性の高さに注意を向けながら，分析を行った．

システムの分析から，人間の感じる破綻の認定の容易さとシステムにとってのそれは現時点では一致していないようであること，まだシステム毎の振る舞いに大きな違いがあり技術的なボトルネックを議論するには早い段階にあること，が示唆された．それと並行して，多人数アノテーション結果の分析とその類型化の試行から破綻の定義についてさらなる検討の必要性が明らかになったが，これは破綻検出技術の向上のためにも重要であろう．

目視による確認により，多数決による正解ラベルの決定は概ね妥当に機能していることが確認された一方で，それぞれ 24 人と 30 人のアノテータにより作られた学習データと評価データを用いた予備的な実験から，主観性の高いデータにおける「正解を決めない」アプローチの優位性が示唆された．

参考文献

[1] 第 75 回言語・音声理解と対話処理研究会 (第 6 回対話システムシンポジウム), 人工知能学会研究会資料 SIG-SLUD-75-B502, 2015.

[2] 東中竜一郎, 船越孝太郎, 荒木雅弘, 塚原裕史, 小林優佳, 水上雅博. テキストチャットを用いた雑談対話コーパスの構築と対話破綻の分析. 自然言語処理, Vol. 23, No. 1, 2016.

[3] 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. 第 75 回言語・音声理解と対話処理研究会 (第 6 回対話システムシンポジウム), 人工知能学会研究会資料 SIG-SLUD-75-B502, pp. 27–32, 2015.

[4] 佐藤啓介. 客観性の梯子: 実践的客観性へ向けての一試論. 往還する考古学-近江貝塚研究会論集 1, pp. 149–158. 近江貝塚研究会, 2002. http://www.h7.dion.ne.jp/~pensiero/archives/ex_objectivity.html.