Sentence-Embedding による対話破綻検出を用いたコーパススクリーニング

北海道科学大学 大学院工学研究科 情報工学専攻 1 年 澤野太一
北海道科学大学 大江亮介，川上敬

1. はじめに

近年の AI ブームによって，以前から存在する学習モデルのブラッシュアップや，新規の学習モデルの考案が盛んに行われるようになった．この AI ブームには，CNN や Encoder-Decoder モデルなどの画期的な学習アルゴリズムが発案されたことの他に，GPU や TPU などによって計算機の能力が飛躍的に向上したことが寄与している．

この計算機能力の向上によって，大規模なデータを学習用データとして利用出来るようになった結果，良質かつ大規模のデータセットを用いることで，学習の精度が上昇することが判明した．

しかし，企業や研究機関と提携していない一個人が，そのような大規模で良質なコーパスを入手することは難しい．中でも，良質な会話データの収集は特に難しいと考えられる．会話データの入手方法としては，Twitter 上での Tweet とそれに対する Reply や，LINE でのやり取り，またはネット掲示板での書き込みなどが挙げられる．だが，LINE などの個人間またはグループ間用の SNS では情報が非公開であり，Twitter などのデータ収集用の API を公開している SNS であっても，良いデータだけを収集できる訳ではない．また，ネット掲示板の会話データを対話システムの学習に用いた場合，相応しくない応答が生成されてしまう可能性がある．

また，収集したデータを人手でふるい分けするのは，データの規模を考慮すると非常に難しい．仮に人手でふるい分けをするとしても，大人数を雇わなければならないため，とても多くの人件費がかかってしまう．

このような背景から，大規模な会話データの中から対話システムのコーパスとして相応しいものだけを抽出する手法が必要になると考えられる．

そこで，本研究では，対話破綻検出を用いてのネット上から収集した会話コーパスのスクリーニングを提案する．対話破綻検出の詳細については 3 章で述べる。本稿では，対話破綻検出チャレンジ[1]の評価方法に則り，筆者が作成した対話破綻検出器を評価する．本研究では，Random Forest と Light GBM[2]の 2 つの学習モデルを破綻検出器に用いている．最後に，作成した対話破綻検出器を使い，筆者がインターネット上で収集した対話データに対してラベル付けを行う．その後，そのラベル付けが妥当なものかをラベル付けされた対話から判断する．収集した対話データは， Twitter での Tweet とそれに対する Reply である．

2. 関連研究

2015 年から，対話破綻検出チャレンジというコンペティションが現在までに 4 回開催されている．4 回行われたコンペティションの中で，最も性能が良かった破綻検出器は，第 3 回目の Sugiyama によって提案されたもの[3]である．この論文中では，Word-based Similarities や Dialogue act，Sentence embedding などの 10 種類の素性について，どの素性がどの程度破綻検出器の性能に寄与するかを検証している．結果として，wmd(Word mover's distance)-noun と uni-gram の Language Model が accuracy の改善に効果的であり，分布関連のメトリックの改善においては，uni-gram の Language Model，Sentence embedding などが重要な素性となっている．また，学習モデルとしては，複数の回帰手法をスタックするという形態を取っている．最上位のレイヤーは Extra Trees Regressor で，下位のレイヤーにはその他の回帰手法を配置している．

Takayama らの手法[4]では，単語またはフレーズレベルでの破綻を検出する CNN ベースの検出器と，文章または文脈レベルの破綻を検出する RNN ベースの検出器を用いて対話破綻検出を行なっている．この論文中では，2 つの検出器を別個に利用する他，2 つの検出器の平均を取るアンサンブル学習の手法も取り入れている．また，学習に使うデータとしては，対話破綻検出コーパスをアノテーターのラベル分布によって k-means++でクラスタリングし，各クラスターに分配する．そして，分配されたデータにアノテーションの平均ラベルが割り当てられたものを用いている．これらの文章は，CBoW によって単語毎にベクトル化される．

また，Inaba らは対話システムの性能向上を目的として対話破綻検出を用いている[5]．この研究では IRS(用例ベース対話システム)，NCM(Neural Conversational Model)，NUR(Neural Utterance Ranking モデル)の 3 つの対話システムを用いている．また，破綻検出の手法として，分類ベース，非破綻確率ベース，線形回帰ベースの 3 つ手法を実験に用いている．これらの対話システムと破綻検出器を組み合わせ，ユーザー発話とシステム応答から破綻検出を行い，システムの応答をリランキングしていくことで，対話システムの応答性能の向上を図っている．

3. 対話破綻検出について

本研究では，対話破綻検出の大部分を，対話破綻検出チャレンジ(以降，DBDC と略す)の問題定義やデータセット，評価

表 1:対話破綻検出のデータセットの詳細

	Chat Dialogue corpus		DBDC1	DCDB2		
	init100	rest1024		DCM	DIT	IRS
対話数	100	1046	100	100	100	100
アノテータ数	24	2or3	30	30	30	30
NB(Not a Breakdown)	59.2%	58.3%	37.1%	39.8%	33.0%	37.4%
PB(Possible Breakdown)	22.2%	25.3%	32.2%	30.2%	27.4%	24.3%
B(Breakdown)	18.6%	16.4%	30.6%	29.9%	39.5%	38.3%
Fleiss' κ(NB, PB, B)	0.28	0.28	0.20	0.31	0.24	0.36
Fleiss' κ(NB, PB+B)	0.40	0.40	0.27	0.44	0.38	0.48

方法に則る形式で行なっている.

DBDC では，ユーザーの発言に対するシステムの応答が，対話行為として成立しているかを評価する．そのために，DBDC では表 1 に挙げるような，ユーザーと 3 つの雑談対話システムとの対話に，破綻ラベルやコメントなどのアノテーションを付与したコーパスを提供している．今回利用するアノテーションは，NB(破綻していない)・PB(破綻の可能性がある)・B(破綻している)が付与されている破綻ラベルのみである.

DBDC における，作成した対話破綻検出器の評価方法は以下の通りである.

1. クラス分類
   - Accuracy : 全ラベルの一致率
   - Precision，Recall，F-measure (B) : 破綻ラベル B に対する精度・再現率・F 値
   - Precision，Recall，F-measure (PB+B) : 破綻ラベル PB と B を同一視した際の精度・再現率・F 値
2. 分布予測
   - JS Divergence (NB,PB,B) : Jensen-Shannon divergence (以降 JSD)による分布間の距離
   - JS Divergence (NB,PB+B) : 破綻ラベル PB と B を同一視した際の JSD による分布間の距離
   - JS Divergence (NB+PB,B) : 破綻ラベル NB と PB を同一視した際の JSD による分布間の距離
   - Mean Squared Error (NB,PB,B) : 分布間の平均二乗誤差 (MSE)
   - Mean Squared Error(NB,PB+B) : 破綻ラベル PB と B を同一視した際の MSE
   - Mean Squared Error(NB+PB,B) : 破綻ラベル NB と PB を同一視した際の MSE

本研究でも，この評価方法を用い，作成した対話破綻検出器の性能を評価する.

4. 対話破綻検出実験

4.1 提案手法

Random Forest と Light GBM の 2 つの機械学習の手法でそれぞれ破綻検出器を作成し，回帰と分類での性能の比較検討を行う．また，今回の実験では，回帰の結果において最も確率が高く出力されたものから破綻ラベルを一意に決定し，それを用いての分類も行う.

今回の実験で破綻検出器の学習に使う素性は，発話とそれに対する応答の Sentence-Embedding(文章の埋め込み表現)を連結したものである．Sentence-Embedding を獲得する手段としては，Google から提案された Universal Sentence Encoder[6] (以降 USE)を用いている．今回の実験では，TensorFlow Hub で提供されている「universal-sentence-encoder-multilingual-large /3」というモデルを使用した．USE で得られる Sentence-Embedding は 512 次元のため，素性となる連結された Sentence-Embedding は 1024 次元のデータとなる.

4.2 学習用データ

表 1 に示したデータセットを使い，対話破綻検出器を作成する．本論文では，表 1 のデータセットの 7 割を訓練データ，残りの 3 割をテストデータとした．クラス分類を行う際，対話に付与されている破綻ラベルの割合から，その対話が 3 つの破綻ラベルの内どれに属するかを一意に決定している．分布予測では，破綻ラベルの個数を，アノテータの総数で割ったものを目的変数としている.

4.3 実験結果

Random Forest と Light GBM での回帰，分類，回帰結果を用いての分類の 3 つの結果を比較すると，全ての結果において Light GBM が Random Forest を上回っていたため, 本稿には Light GBM の結果のみを記載する.

分布予測の結果を図 1，クラス分類の結果を表 2 に示す.

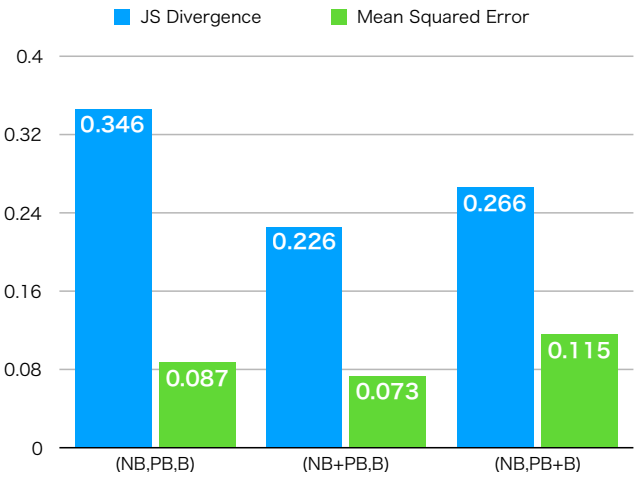

図 1 : Light GBM による分布予測結果

表 2 : Light GBM による分類結果

	ラベル	Accuracy	Precision	Recall	F-measure
回帰結果での 3 クラス分類	NB	0.61	0.65	0.90	0.76
	PB		0.41	0.13	0.19
	B		0.48	0.35	0.40
3 クラス分類	NB	0.60	0.64	0.91	0.75
	PB		0.41	0.13	0.19
	B		0.50	0.30	0.38
回帰結果での 2 クラス分類(NB+PB)	NB+PB	0.86	0.88	0.98	0.93
	B		0.54	0.14	0.22
2 クラス分類(NB+PB)	NB+PB	0.86	0.87	0.99	0.93
	B		0.58	0.09	0.16
回帰結果での 2 クラス分類(PB+B)	NB	0.68	0.76	0.65	0.70
	PB+B		0.60	0.71	0.65
2 クラス分類(PB+B)	NB	0.69	0.71	0.80	0.75
	PB+B		0.67	0.55	0.60

分布予測では，MSE に対して JSD が非常に高くなってしまった．また，過去に行われた DBDC において，最も性能が高い Sugiyama の手法における JSD，MSE には及ばない結果となった．

分類結果では，純粋なクラス分類を行うより，回帰結果を用いてのクラス分類を行った方が，結果が改善する傾向が見られた．特に，Recall において大きな改善が期待でき，PB+B に注目した際の 2 クラス分類では，Recall が 0.55 から 0.71 にまで上昇している．しかし，Precision や NB に注目した際の Recall に若干の減少が見られる．こうしたデメリットはあるものの，回帰結果によるクラス分類の方が F 値の向上が見込める．

また，PB+B でのクラス分類以外の結果では，NB を除くラベルに注目した際の Precision，Recall，F-measure が非常に低くなっている．特に，NB+PB での B に注目した際の Recall が極端に低くなってしまった．

4.4 考察

PB+B でのクラス分類以外の結果において，Precision, Recall, F-measure が，NB または NB+PB に注目したときは高く，B に注目したときには低くなってしまった原因として，表 3 に挙げるような，データセットの偏りが考えられる．

表 3 : データセット内のラベルの個数

3 クラスのラベルの個数	
NB	8626
PB	3345
B	3780
2 クラス(NB+PB)のラベルの個数	
NB+PB	13323
B	2128
2 クラス(PB+B)のラベルの個数	
NB	9017
PB+B	6434

3 クラスでのラベル間の個数では，NB に対して PB，B 共に半分以下のサンプルしか存在しない．また，2 クラス(NB+PB)では，ラベル間での個数の差が 6 倍近くなってしまっている．

このような，ラベル間での差が大きくなっている根本的な原因として，今回の実験で使用したデータセットに含まれている「init100」と「rest1024」のラベル間の割合の乖離が激しいことが考えられる．このことは，DBDC2 での Sugiyama の結果[7]でも，これらのデータセットを使用すると JSD，MSE 共に結果が悪化するとの記述があることからも推測される．このように，不均衡なデータを用いたため，PB，B，NB+PB に注目したときの Precision，Recall、F-measure が低いにも関わらず，Accuracy が高くなってしまった．

また, PB+B 以外での分布予測と回帰結果を用いてのクラス分類においても，同様の理由で検出結果が悪化してしまったと考えられる．

PB+B の場合のクラス分類では, ラベル間の個数の差がそこまで大きいわけではないのにも関わらず，期待した性能に到達していないため，Light GBM のハイパーパラメータの調整不足や，学習に用いる素性が適当ではない可能性がある．

分布予測で, MSE に対して JSD が高くなってしまった原因としては，目的変数と予測値が局所的に大きく乖離してしまった，もしくは JSD の算出方法に誤りがある可能性がある．

今回の実験では，Sentence-Embedding を素性とし，破綻検出器を作成した．性能としては満足のいくものは得られなかったが，PB+B における結果を見ると，最低でも F 値が 0.6 を超えているため，Light GBM が 1024 次元という多量の説明変数を与えられても性能を維持できる学習器であることと，Sentence-Embedding が対話破綻検出において有効な素性であることが判明した．

今後の展開として，破綻検出器の性能がラベルの不均衡によるものなのかを，Undersampling や Downsampling によって調査する予定である．また，「init100」と「rest1024」をデータセットから除外した場合の性能についても検証を行う．

表 4 : Twitter 上の対話データの分類結果(結果の一部を抜粋したもの)

	ラベル	対話文
NB	Ture NB	発話：ホットミルクを飲みました…。……温かくて、ほっとする…… 応答：寒い朝には欠かせない一杯ですね苦手ではなければ、黒糖やチューブの生姜も一緒に入れてみて下さい。香りや甘さでいつもと変わったホットミルクになりますよ…！
	False NB	発話：大丈夫ですよ ？ またここで話せますし連れてきてくれてありがとうございます 、嬉しいなあ 勿論です 応答：ツイ…
B	Ture B	発話：ありがとうございます。よろしくお願いします。さっそくですが何とお呼びすればいいですか？？ 応答：答えが読める！優しい問題ですね！！！！
	False B	発話：腰痛発生のため仕事、ポケ活休み。フレンドの皆さんギフト申し訳ないです。 応答：大丈夫ですか？。お大事に〜。

5. Twitter での対話データのスクリーニング

5.1 実験目的

作成した破綻検出器を用いて，Twitter 上での Tweet と，それに対する Reply を収集したコーパスに対し破綻検出を行う．この実験の目的として，DBDC コーパスによって学習された破綻検出器が人間同士の支離滅裂な会話などについても破綻を検知できるかを検証することが挙げられる．この試みが成功すると，対話システムを作成する際に用いるデータセットが事前に人手で選別されていないランダムなものだった場合に，自動で質が高いコーパスに変換することが可能になると予測される．

5.2 実験結果

実験結果の一部を表 4 に示す．ラベル列は，筆者が分類結果を確認し判断したもので，True NB は「正確に破綻していないことを検出」，False NB は「破綻しているが破綻していないと判断」，True B は「正確に破綻していることを検出」，False B は「破綻していないが破綻と判断」としている．破綻検出対象とした対話は 500 個で，その内 NB と判断された対話は 112 件，B と判断された対話は 388 件だった．

破綻検出器の性能が不足しているため，誤検出が多々見受けられる．また，NB と判断される対話が B に比べ少ないため，この破綻検出器を用いると，データセットが縮小してしまう恐れがある．そのため，破綻検出の精度を改善することが第一に求められる．

今回の結果から，同一の単語が双方に含まれている場合には比較的 NB と検出され易いと推測される．また，発話と応答がどちらも長い場合には B と検出されることが多く，どちらも短い場合には NB と検出されることが多かった．

6. おわりに

本研究では, Sentence-Embedding による対話破綻検出手法を提案すると共に，作成した破綻検出器によるコーパスのスクリーニングを試みた．

破綻検出器として，Random Forest と Light GBM の 2 つで検証を行った結果，1024 次元の Sentence-Embedding を素性とする場合，Light GBM の方が破綻検出の性能が高くなることが確認できた．また，純粋なクラス分類の結果と，回帰結果を用いてのクラス分類の結果を比較した場合，後者の方が，F 値が改善する傾向が見られた．

また，今回の実験では，Light GBM で 1024 次元の巨大な説明変数を用いて学習を行っても, PB+B の結果においては極端に性能が低下することはなかった．このことから，Light GBM は多量の説明変数について頑健であり, Sentence-Embedding が対話破綻検出に有効であることが分かった．

今回の実験では，満足な破綻検出器の性能が得られなかった．今後は破綻検出器の性能向上のために，学習に用いるデータセットの精査や，学習器のハイパーパラメータの調整を行う他，Sentence-Embedding 以外の素性を用いての学習や，Sentence-Embedding を次元削減したものを素性とする実験も行っていく．また, 今回の実験では Random Forest と Light GBM を用いたが， 今後は Neural Network を用いての対話破綻検出についても検証する予定である．

参 考 文 献

[1] 対話破綻検出チャレンジ 2
https://sites.google.com/site/dialoguebreakdowndetection2/

[2] Guolin Ke，Qi Meng，Thomas Finley，Taifeng Wang
Wei Chen，Weidong Ma…
LightGBM: A Highly Efficient Gradient Boosting Decision Tree，2017

[3] Hiroaki Sugiyama
Dialogue Breakdown Detection based on Estimating Appropriateness of Topic Transition，2017

[4] Junya Takayama，Eriko Nomotok，Yuki Arase
Dialogue Breakdown Detection Considering Annotation Biases，2017

[5] Michimasa Inaba，Kenichi Takahashi
対話破綻検出の対話システムへの適用，2019

[6] Daniel Cera，Yinfei Yanga，Sheng-yi Konga，
Nan Hua，Nicole Limtiacob，Rhomni St. John…
Universal Sentence Encoder，2018

[7] Hiroaki Sugiyama
https://sites.google.com/site/dialoguebreakdowndetection2/presentation