Random Forestを用いた対話破綻検知器の作成

Making dialog breakdown detector with Random Forest

金井貴浩[1*]　松原仁[1]
Takahiro Kanai[1]　Hitoshi Matsubara[1]

[1] 公立はこだて未来大学
[1] Future University Hakodate

Abstract: This paper describes method of dialog breakdown detector. We developed dialog breakdown detector for the dialog breakdown detection challenge. The detector uses Random forest and Paragraph Vector. Input is user's utterance and system's utterance. Features are vector of user's utterance and system's utterance. We carried out experiment for learning corpus.

1　はじめに

近年，対話システムの需要は高まっており，多くのところで活用されている．これらは何らかの目的を持った対話システムであり，タスク志向と呼ばれている対話システムである．これらのタスク志向対話システムは，様々なところで成果を上げている．最近では，特に目的の持たない会話，雑談を目的とした雑談対話システムが多く登場し，注目を浴びている．Microsoft 社が，提供しているりんなが人気である [1]．しかし，雑談対話システムは自然な対話ができているとは言いづらく，不自然な対話が多い．この不自然な対話を対話破綻と呼び，この対話破綻を検出する対話破綻検出チャレンジが行われている [2]．本稿では，この対話破綻検出チャレンジに提案した手法などを紹介する．

2　対話破綻検出チャレンジ

対話破綻検出チャレンジとは，雑談対話システムとの対話ログを使用し，対話破綻を検出する評価型のワークショップである [2]．雑談対話システムとの対話ログには，複数のアノテータによる対話破綻のラベルがつけられている．ラベルは「O」,「T」,「X」の三つでつけられている.「O」は，破綻していない.「T」は，破綻とは言い切れないが違和感がある.「X」は，破綻している，の以上三つである．対話破綻チャレンジでは，対話ログ中のラベルを推定する．

対話破綻チャレンジでは，二つの方法によって，検出器の性能を評価する．一つ目はラベル一致系統である．これは対話ログ中のラベルを多数決によって正解ラベルを決める．Accuracy や，Precision(X)，Recall(X)，F-measure(X) など評価項目が設けられている． 二つ目は分布距離系統での評価である．これは，ラベル一致系統とは違い正解ラベルを決めない．分布距離系統は，対話ログ中の評価の割合を推定するものである．対話破綻検出チャレンジ 2 ではこちらを重視する．

*連絡先： 公立はこだて未来大学
北海道函館市亀田中野町 116 番地 2
E-mail: g2116010@fun.ac.jp

3　提案手法

本研究では，対話破綻はユーザ発話とシステム発話の意味の差から破綻できるもの，と考えた．そこで，ユーザ発話とシステム発話をベクトル化を行い，内積とコサイン尺度によって意味の差が表現できる，と仮定した．この内積と，コサイン尺度の二つの特徴量によって破綻している，していないの分類を行った．また,「破綻とは言い切れないが，違和感がある」のラベルについては破綻しているに組み込んだ．これは，破綻していないのほうに組み込んでしまった場合，違和感がある会話が残ってしまう可能性があるためである．本研究では，対話破綻検出チャレンジ 2 において重視する分布距離系統ではなく，ラベル一致系統を重視した．

3.1　アルゴリズム

今回，開発した対話破綻検出器の入力は，システム発話とその直前のユーザ発話の組み合わせとした．出力は対話破綻している，していない，を表す 1 と 0 のみを出力とした．特徴量は，当該のシステム発話のベクトル表現と直前のユーザ発話のベクトル化を行い，システム発話ベクトルと，ユーザ発話ベクトルの内積とコサイン尺度を使用した．ベクトル表現は MeCab に

表 1: 実験結果 (t=.0)

学習量	Accuracy	Precision(X)	Recall(X)	F(X)	Precision(T+X)	Recall(T+X)	F(T+X)
963 個 run1	.335	.187	.683	.294	.385	.633	.479
100 個 run1	.551	.141	.197	.164	.383	.214	.274
25 個 run1	.611	.216	.053	.086	.472	.053	.095

よるわかち書きの結果をもとに，Paragraph Vector を使用した [3]．分類器は Random Forest を使用した．

3.2 Paragraph Vector

文章の意味をとらえる素性として Paragraph Vector を使用した．これは，Word2Vec という単語分散表現を文章にまで拡張したものである [3]．

Paragraph Vector の特徴としては，文脈を保存したままベクトルにすることができる点である．これは，自然言語処理でよく使われる Bag of Words にはない特徴である．この特徴によって，より正確に文章の意味を正確に捉えることができると考えた．

実装には python ライブラリである，gensim を利用した．学習には昨年度の対話破綻チャレンジにて配布された rest1046 のコーパスを用いて行った．

3.3 Random Forest

分類器として Random Forest を使用した．Random Forest は木構造を複数使ったアンサンブル学習の一つである．特徴としては，学習が早いことが挙げられる．この特徴は，昨年度の対話破綻チャレンジで多く使われていた RNN などディープラーニングとは違い，計算量が少なくすむ．この特徴によってマシンパワーの弱いコンピュータでも対話破綻検出器を作成することができる．

実装には，python ライブラリである，scikit-learn を用いて実装を行った．

4 実験

評価実験として，実際に対話破綻検出器にコーパスを渡し，評価をした．どの学習量が一番性能が良いかを調べるため，Random Forest の学習量を変更して行った．それぞれの学習セットを学習させた．ラベル一致系統では，閾値を 0.0 にすることが基準となっている．これは正解ラベルを多数決によってのみ決めるということである．これは，閾値をどれくらいに設定するのかが難しいためである．

4.1 データセット

Random Forest の学習をするコーパスを 25 個，100 個，963 個，の 3 つに設定し，モデル作成を行った．この学習で使用したコーパスは昨年度配布された rest1046 と今年配布された DCM．IRS の中からランダムに抜き出したものを使用した．テストセットとして，使用したコーパスはそれぞれ，rest1046 から 100 個，DCM から 25 個，IRS から 25 個，計 150 個を使用した．

4.2 結果

実験の結果から，Accuracy は 25 個のコーパスを使用したモデルが一番高い結果であった．しかし，F(X) を見ると，.053 ととても低く，X をほとんど検出できていないことがわかる．次に，963 個のコーパスを学習したモデルだが，Accuracy が.335 ととても低いため，今後はこのモデルは使用しない．最後に 100 個のコーパスを学習したモデルは Accuracy は.551 と 25 個のコーパスを学習したモデルに比べ低くなっている．しかし，X の検出の性能を示す F(X) は.164 になっており，25 個のコーパスに比べ高くなっている．

これら 3 つのモデルを比べてみると，学習量を増やせば増やすほど，Accuracy が上がり，X の検出の性能が下がっていることがわかる．しかし，全体的に X の検出率が低い．今回のモデルでの T の扱いは，X に組み込んだモデルを設計した．T を X に組み込んだことによって，基本的には正解が増えるはずであり，正解数も増えるため，F(X) に比べ，F(T+X) が高くなると想定される．実際にすべてのモデルで，F(T+X) のほうが高くなっている．

次に，今回の対話破綻検出チャレンジ 2 において，重視されている分布間距離系統についての結果を表 2 にまとめた．本手法では，分布間距離系統は重視していないため，全体的に低い結果となった．953 個のコーパスを学習したモデルは Accuracy が.335 と高いのに対して，JS divergence が全体的に一番良い結果となっている．逆に一番 Accuracy が高い 25 個のコーパスを学習したモデルは，JS divergence が一番悪い結果になっている．

表 2: 分布距離系統

学習量	Accuracy	JS divergence (O,T,X)	JS divergence (O,T+X)	JS divergence (O+T,X
963 個 run1	.335	.493	.405	.353
100 個 run1	.551	.633	.445	.585
25 個 run1	.611	.695	.468	.685

4.3 考察

今回の評価実験によると，学習を行えば行うほど X の検出率が上がり，O の検出率が下がることが分かった．また，Accuracy についてみてみると．学習量を減らしたほうが上がることが分かった．しかし，Accuracy が上がったのはテストセットの影響が十分考えられる．これは今回使用したテストセットの約 6 割が O であることによって Accuracy が上がっていった可能性があり，違うデータセットで検証する必要がある．25 個のコーパスを学習したモデルでは，Accuracy が高く，Precision(X) や Recall(X)，F(X) の値が低くなっている．つまり，X をほとんど検出もせず，間違えて X を出力もほとんどしていないことがわかる． したがって，検知器のモデルとしてはあまり意味がない．100 個のコーパスを学習したモデルは Accuracy も平均的で，X の予想も 25 個のコーパスを学習したモデルに比べ，性能がいいため，このモデルを対話破綻チャレンジに提出をした．

5 まとめ

5.1 今後

今回の実験では，25 個，100 個，963 個のコーパスを使用して学習を行ったモデルの比較を行った．最終的に 100 個のコーパスを学習したモデルを提出した．しかし，25 個，100 個．963 個では学習量に差があると考えている．特に 100 個と 963 個の間にもっと良い学習量が存在する可能性がある．今後は学習量についてもう少し検討をしていく必要がある．

今回のモデルでは，rest1046 を用いて，Paragraph Vector の学習を行った．しかし，これでは単語の学習量があまりに少なすぎるため，正しいベクトル化ができていない．したがって，今後はより学習量を増やす必要がある．そのために今後コーパスを集める必要がある．

5.2 まとめ

本稿では，対話破綻チャレンジ提出するモデルの検討，実験を行った．本稿で作成したモデルは発話文，応答文を Paragraph Vector によりベクトル化し，内積とコサイン尺度を特徴量とした．二つの特徴量を元に Random Forest を用いて分類を行った．また，コーパスの学習量の違いによって性能に大きな差が出ることが分かったため，実験を行った．実験を行った結果，100 個のコーパスを学習したモデルを採用した．

参考文献

[1] 呉先超, 伊藤和重, 飯田勝也, 坪井一菜, クライアン桃, りんな : 女子高生人工知能, 言語処理学会第 22 回年次大会, 2016

[2] 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，第 78 回言語・音声理解と対話処理研究会（第 7 回対話システム　シンポジウム），2016.

[3] Quoc Le, Tomas mikolov, Distributed Representations of Sentences and Documents, Proceedings of the 31st International Conference on Machine Learning, 2014.