人工知能学会研究会資料
SIG-SLUD-B502-08

類型毎の検出手法の組合せによる雑談発話破綻検出の検討

A breakdown detection method based on taxonomy of errors in chat-oriented dialogue

堀井 朋　荒木 雅弘 *
Tomo Horii , Masahiro Araki

京都工芸繊維大学
Kyoto Institute of Technology

Abstract: It is difficult to detect a breakdown phenomena in the dialogue with a chat-oriented dialogue system because of the variety of causes of breakdown. To deal with this problem, we analyzed a chat dialogue corpus and made a taxonomy of the errors that yield the dialogue breakdown. In this paper, we propose a breakdown detection method that consists of combinations of the classifiers for the different cause of errors based on the taxonomy.

1　はじめに

雑談対話はユーザの継続的な対話システム使用を促し，信頼感の醸成などの面で有用である．長く使い続けると，ユーザが音声インタフェースに慣れ，高齢者などが様々なサービスを音声対話サービスで受けられるなどのメリットがある．対話を長く続けるためには，破綻しそうなシステム発話を出力前に検出することが有効である．

しかし，雑談対話の破綻の原因は様々であり，単一の識別器による検出は難しい．また，既存手法は，様々な特徴を用いているが，その抽出には言語依存のツールが必要になることが多い．我々は言語に依存しない単純な特徴を用いた複数の識別器の組み合わせによる破綻検出を試みる．

2　提案手法

一口に対話が破綻すると言っても，それが引き起こされる要因は一つではない．発話の文構造が誤っている場合や，文脈の流れに添えていない場合など，様々な原因が考えられる．そこで，雑談対話コーパスの分析の結果，作成された類型化案 [1],[2] に基づき，その類型毎に識別機を作成して組み合わせる手法を提案する．

類型化案の内容は表 1 の通りである．大分類は，対話のどの範囲に関連した破綻であるかという点を基準にして，以下のように分類が行われている．

- 発話 : 当該システム発話のみから破綻が認定できるケース
- 応答 : 直前のユーザ発話と当該システム発話から破綻が認定できるケース
- 文脈 : 対話開始時点から当該システム発話までの情報から破綻が認定できるケース
- 環境 : 破綻原因が，上記の 3 分類には当てはまらないケース

そして，各大分類に対して破綻要因を更に細かく突き詰めたものが小分類である．本稿ではまず，この 16 分類に則して，破綻検出を行うことを考える．

3　実験

3.1　実験 1 : 単語ベクトルを特徴とした検出

まず，ベースラインを定めるために類型を用いていないデータで破綻検出を行った．直前のユーザ発話とシステム発話からなる単語ベクトルを特徴とし，SVM（poly–kernel,1 次）を識別アルゴリズムに用いた．本実験で扱う学習データは Project Next NLP 対話タスク内の類型化ワーキンググループが作成した，ユーザ–システム間におけるシステム発話の破綻の有無と破綻類型がアノテーション付けされた 2100 発話である．テストデータは本チャレンジで配布された開発データ 420 発話とした．また，評価は配布された評価スクリプトによって行った．結果を表 2 に示す．

*連絡先：京都工芸繊維大学工芸科学部
〒 606–8585 京都府左京区松ヶ崎
E-mail: horii@ii.is.kit.ac.jp, araki@kit.ac.jp

表 1: Project Next NLP による破綻類型化

大分類	小分類	内容
発話	構文制約違反	構文的な誤り
	意味制約違反	意味的な誤り
	不適切発話	発話としての機能を持たない
応答	構文制約違反	構文的な誤り
	意味制約違反	意味的な誤り
	関係の公準違反	発話対を形成しない応答（意味的側面も含む）
	構文制約違反	構文的な誤り
	意味制約違反	意味的な誤り
応答	構文制約違反	構文的な誤り
	意味制約違反	意味的な誤り
	関係の公準違反	発話対を形成しない応答（意味的側面も含む）
	構文制約違反	構文的な誤り
	意味制約違反	意味的な誤り
発話	構文制約違反	構文的な誤り
	意味制約違反	意味的な誤り
	不適切発話	発話としての機能を持たない

表 2: 類型なしの場合の破綻検出スコア

Accuracy	Precision	Recall	F-measure
0.42	0.73	0.73	0.73

表 3: 16 類型に則した場合の破綻検出スコア

Accuracy	Precision	Recall	F-measure
0.56	0.76	0.29	0.42

表 4: 大分類に則した場合の破綻検出スコア

Accuracy	Precision	Recall	F-measure
0.49	0.76	0.70	0.73

続いて，破綻の類型化に従って検出を行うことの有用性を確かめるために，16 分類に則した 16 種の検出器を上記と同じ条件で作成した．複数の破綻検出器のうち，一つでも破綻であるという結果になれば，その発話は破綻であると識別している．この実験結果を表 3 に示す．これを見ると，破綻分類に則した識別器の方がベースラインよりも精度が下回った．これは破綻類型が細分化されているため，類型毎の破綻を学習するデータの数も非常に少なくなってしまったが故に，破綻を検出できていないのはないかと考えた．つまり，学習データの数が増えたならば，より多くの破綻を検出できる可能性がある．

そこで，16 種の小分類で分けるのでなく，4 種の大分類に沿った学習データを作成し，同様の実験を行った．結果は表 3 の通りである．16 分類に則した場合と比べて正解率は低下したものの，F 値は上昇している．つまり，破綻をより多く検出できていることが分かる．また，ベースラインと比べても F 値こそ僅かに小さいものの，正解率はこちらのほうが高い．破綻検出においては，この 2 つの値の釣り合いが重要となるため，ベースラインよりも類型毎に検出を行った方が，良い結果になっているのではないかと考えられる．また，大分類の 4 種に類型化された学習データの量も，未だ不充分である可能性もある．

以上のことから，破綻類型に従って学習データを充分に揃え検出器を作成すれば，単語ベクトルによる特徴抽出という簡単な方法で，ある程度の有効性を持った検出精度を得られることが示された．この方法は日本語のみでなく，他の言語にも適用させることが容易であると考えられる．

3.2 実験 2：類型毎に特徴を設定した検出

類型ごとの破綻検出スコアが低い原因は、それぞれの破綻検出に適した特徴が設定できていないことも考えられる．そこで，大分類に則した破綻検出に有効と思われる手法を以下のように設定し，破綻検出を行った．

- 発話
  - システム発話のみが特徴
  - 構文解析結果のスコアを特徴とする

表 5: 「応答」のみ 2 次カーネルを用いた検出スコア

Accuracy	Precision	Recall	F-measure
0.50	0.75	0.65	0.70

- 応答
  - 直前のユーザ発話とシステム発話からなる単語ベクトルが特徴
  - 2 次カーネルによって単語共起情報を用いて学習
- 文脈
  - システム発話とそれ以前の発話全てが特徴
  - 上記 2 つの単語ベクトルに対するコサイン類似度を特徴とする
- 環境
  - システム発話の単語ベクトルのみが特徴

実験の結果,「発話」・「応答」類型については，KNP[1] のスコア及びコサイン類似度に関して，正例・負例の分布に顕著な違いは見られなかった．

また,「応答」類型において，2 次カーネルを用いて学習を行い，実験 1 と同データに対して破綻検出を行った結果が表 5 である．この場合もあまり改善は見られない．

これらの結果より，4 分類で単純に単語ベクトルを用いて学習・検出を行った場合と変化がない，もしくは悪化するという結果になった．原因として,「発話」については，文が短く，2〜3 文節程度の係り受けスコアでは文の妥当性が判定できなかったためと考えられる．「文脈」については，コサイン類似度を求める際,「車」や「自動車」と言ったような，言葉は違えど意味は同じである言葉を同一と見なせていなかったことが挙げられる.「応答」については，特別な原因を挙げることはできないが学習データが不足している可能性があると考える．

3.3 総括

今後の展望としては，学習データを増やすと共に，16 種の類型毎に適切と思われる特徴や識別アルゴリズムを新たに設定・改善を行い検出器を作成し，検出精度の向上を目指す．

[1]http://nlp.ist.i.kyoto-u.ac.jp/?KNP

4 関連研究

雑談対話システムにおける破綻検出についての先行研究が複数ある．

Xiang ら [3] はシステムとの対話の中で，ユーザ発話に問題がある状況を検出するといった手法を提案している．彼らは，ユーザ意図や感情について，対話のシーケンス・モデルを用いて検出を行っている．彼らの提案手法と我々の提案手法の違いは，破綻直近のユーザ発話を用いるか否かというものである．我々の問題設定では，システムが発話を生成する前に，その発話が破綻となるかどうかを検出するものとなっている．つまり，我々は破綻の後，ユーザ発話を検出に使用することはしない．

また，東中ら [4] は雑談対話システムとの対話の中で，システム発話の一貫性を評価して破綻を検出する手法を提案した．彼らは，システム発話の一貫性を評価するために，発話ペア・対話内容・述語項構造などの各種特徴を使用している．しかし，このような多様な特徴抽出は，言語依存ツールを必要とする．これに対して，我々の提案手法では，言語非依存の単純な特徴抽出を行う識別器を組み合わせて判定を行う．

5 おわりに

本稿では，破綻を類型化し，その分類に則して破綻を検出する手法を検討した．

結果として，類型毎の破綻検出手法に一定の有効性があることが示されたが，破綻類型のそれぞれに適した特徴を発見するまでには至らなかった．

今後は，今回発見した問題点を取り除き，その結果を含めて，再度類型毎の検出器構築手法を検討する予定である．破綻検出に有用な検出器を 16 分類各々に対して，構築することが最終的な目標である．

参考文献

[1] Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi and Masahiro Mizukami: Towards Taxonomy of Errors in Chat-oriented DialogueSystems, *In Proc. SIGDIAL 2015, pp.87-95. (2015)*

[2] Ryuichiro Higashinaka, Masahiro Mizukami, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara and Yuka Kobayashi: Fatal or not? Finding errors that lead to dialogue breakdowns in chat-oriented dialogue systems, *In Proc. EMNLP 2015, pp.2243-2248. (2015)*

[3] Yang Xiang, Yaoyun Zhang, Xiaoqiang Zhou, Xiaolong Wang, Yang Qin: Problematic Situation-Analysis and Automatic Recognition for Chinese Online Conversational System, *In Proc. of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, pp.43-51. (2014)*

[4] Ryuichiro Higashinaka, Toyomi Meguro, Kenji Imamura, Hiroaki Sugiyama, Toshiro Makino, Yoshihiro Matsuo: Evaluating Coherence in Open Domain Conversational Systems, *In Proc. of Interspeech 2014, pp. 130-134.*