おーぷん2ちゃんねる対話コーパスを用いた用例ベース対話システム

A Example Based Dialogue System using the Open 2channel Dialogue Corpus

稲葉 通将 [1*]

[1] 電気通信大学
[1]The University of Electro-Cummunications

Abstract: This paper describes NoneJupiterBot that participated in Dialogue System Live Competition 2. NoneJupiterBot is an example-based dialogue system using "Open 2 Channel Dialogue Corpus". This corpus is newly constructed by the author using data from a textboard site, Open 2 Channel. This paper also describes the construction method of this corpus.

1 はじめに

本稿では，対話システムライブコンペティション2[1]に出場した対話システム「NoneJupiterBot」について述べる．NoneJupiterBot は「おーぷん 2 ちゃんねる対話コーパス」を用いて構築した用例ベース対話システムである．おーぷん 2 ちゃんねる対話コーパスは，電子掲示板サイトおーぷん 2 ちゃんねる[1]のデータを用いて筆者が新たに構築した対話コーパスであり，本コーパスの構築方法についても本稿で述べる．

2 おーぷん2ちゃんねる対話コーパス

近年，Ubuntu Dialogue Corpus [2] や Conversational Datasets [3] など，Web データを用いて構築された大規模対話コーパスが次々と公開されており，深層学習を用いた対話システム研究を中心に活発な利活用が行われている．しかしながら，そのほとんどが英語，もしくは中国語 [4] であり，日本語のコーパスは著者が知る限り存在しない．

そこで，パブリックドメインの電子掲示板サイトおーぷん 2 ちゃんねるのデータを用いて日本語の対話コーパスを構築した．以下では，本コーパスの構築方法について述べる．本コーパスは GitHub 上で公開している[2]．

2.1 クローリング

おーぷん 2 ちゃんねるは，板と呼ばれるカテゴリに分かれており．ユーザは板の中でスレッドを立て，スレッドに対して投稿を行う．本コーパスでは，おーぷん 2 ちゃんねるにおいて人気が高い板である「なんでも実況 (ジュピター)」「ニュー速 VIP」「ニュース速報+」の 3 つを対象に，スレッドのクローリングを行った．

おーぷん 2 ちゃんねるでは，現在書き込むことができない過去のスレッド (過去ログと呼ばれる) の一覧を取得する方法が存在しないため，過去ログの保存・検索サイトであるログ速[3]の検索機能を用いてスレッドの一覧を取得した．なお，ログ速ではほぼリアルタイムでおーぷん 2 ちゃんねるに対する新規投稿も反映されるため，現行スレッドの情報も取得可能である．スレッドのデータはログ速でもおーぷん 2 ちゃんねると同等のデータが取得可能であるが，個人で運営しているログ速よりもおーぷん 2 ちゃんねるのほうが負荷耐性が強いと考え，スレッドのクローリングはおーぷん 2 ちゃんねるに対して行った．

2.2 対話データの抽出

スレッドにおける投稿は投稿順に ID が割り振られる．利用者は投稿の冒頭でアンカー「>>」により ID を指定することで，どの投稿に対する返信かを明示することができる．おーぷん 2 ちゃんねる対話コーパスでは，このアンカーで関連付けされた投稿間の関係を，対話における応答関係と捉え，対話データを抽出した．抽出した対話をコーパスに収録する際は，アンカーと

*連絡先： 電気通信大学
〒 182-8585 東京都調布市調布ケ丘 1 丁目 5 − 1 Email: m-inaba@uec.ac.jp

[1]https://open2ch.net/

[2]https://github.com/1never/open2ch-dialogue-corpus

[3]https://www.logsoku.com/

表 1: コーパスの統計情報

取得元	対話数	平均対話長
なんでも実況 (ジュピター)	5948218	2.24
ニュー速 VIP	1983626	2.41
ニュース速報+	217296	2.09

表 2: おーぷん 2 ちゃんねる対話コーパス中の対話例

菅野が澤村に勝ちを消されていなかったらと考えたら切なくなる
4 回勝ち消したんやっけ？
眼精疲労ですね
それなんや？治るか？
ワイは目薬挿したら治ったで
さして見るわ
会ったことない人好きになれるなら別れてもすぐ忘れるでしょ
彼氏とは付き合う前にあってる
おにぎりバーガーとか開発して！
モスにライスバーガーってのがあるよ

ID は削除した．3 投稿以上からなる対話については，2 名の利用者が交互に投稿を行っているもののみを抽出した．

また，以下のいずれかに該当する投稿はコーパスから除外した．

- 文字数が 5 文字未満，および 150 文字より大きい投稿
- URL および画像を含む投稿
- 2 つ以上のアンカー，もしくは冒頭以外でアンカーを含む投稿
- ひらがな，カタカナ，漢字のいずれも含まない投稿
- 4 回以上の改行を含む投稿

最後の 4 回以上の改行を含む投稿を除外する理由は，日程や野球における打順の列挙など，対話としては不適切なものを含む場合があるためである．

2.3 コーパス構築

おーぷん 2 ちゃんねる開設時から 2019 年 7 月 20 日までのデータを使用し，コーパスを構築した．コーパスの統計情報を表 1 に示す．

また，コーパス中の対話例を表 2 に示す．

3 NoneJupiterBot

NoneJupiterBot はおーぷん 2 ちゃんねる対話コーパスを用いて作成した用例ベースの対話システムである．用例の選択には全文検索エンジンの Elasticsearch と BERT[5] を用いている．

まず，おーぷん 2 ちゃんねる対話コーパスから，キーワードリストを用いて放送禁止用語などを含む不適切な投稿を除外する[4]．次に，コーパス中の応答関係にある 2 投稿をペアにして Elasticsearch に投入する．なお，NoneJupiterBot というシステム名はコーパスの構築の際に用いた板の 1 つである「なんでも実況 (ジュピター)」からとっているが，Elasticsearch には他の 2 つの板のデータも含め，すべてのデータを投入した．また，Elasticsearch には，日本語を扱うためのプラグインである Japanese (kuromoji) Analysis プラグインをインストールした．

BERT については Google が公開している多言語の事前学習済みモデル (BERT-base, Multilingual Cased) を用いて，Elasticsearch に投入したものと同様の投稿ペアを正解データ，ランダムに抽出した 2 つの投稿ペアを不正解データとし，BERT に入力したペアが応答関係を有するペアかランダムのペアかを 2 値分類するファインチューニングを行った．

ユーザの入力発話に対して応答を選択する方法は以下の通りである．まず，ユーザの発話をクエリとし，Elasticsearch で投入したペアの 1 投稿目を検索し，最大 20 件のペアを獲得する．これら獲得したペアの 2 投稿目が応答候補とする．次に，ユーザ発話とそれぞれの応答候補に対し，ファインチューニング済みの BERT を用いて応答関係を有する確率の最も高い応答候補を応答として選択する．

以上の手法により構築したシステムを著者がテストしたところ，ユーザが「そうですね」などの短い応答を行うと，返ってくる応答も簡潔なものになる傾向が見られた．そのままだと，ユーザとシステムがともに簡潔な応答に終止してしまう可能性があったため，ライブコンペティションでは所定の回数以上応答を行った場合，Elasticsearch と BERT による応答選択を行わず，事前に設定した発話を出力することでユーザから最近の出来事について話すよう促すこととした．

4 評価結果と分析

NoneJupiterBot は，第 10 回対話システムシンポジウム内で開催された対話システムライブコンペティション 2 に登録し，本コンペティションの予選において評

[4] キーワードリストおよび除外のためのスクリプトについてもコーパスと同じリポジトリで公開している

表 3: NoneJupiterBot の対話例 (N:NoneJupiterBot, U:ユーザ)

N	2 ちゃんねるのなん J から来たンゴ！よろしくニキ！早速やけど、最近あった楽しかったことを聞かせてほしいんやで
U	立ち食い蕎麦屋の店員がマツコ・デラックスに似てたことかな。
N	直営の蕎麦屋ってほぼってか全くないんちゃう
U	チェーン店の蕎麦屋だったわ。
N	ワイはチェーン店すこ
U	気軽に利用できるからね。
N	軽くってのはそういう意味じゃないんだ。このスレで君が書き込んだ内容を流し読みしたって意味なんだ。おわかり？
U	え？なんかゴメン。
N	ええんやで
U	ワイさんのお仕事はなんですか？
N	ワイ公務員やから利害関係者と仲良くなるのタブーなんや…
U	そうなんだ。いろいろ大変ですね。
N	次は逆に最近しんどかったことを聞かせてクレメンス
U	残業して完成した書類がボツになったことかな。
N	求人票の時点でサービス残業確定とかいかんでしょ
U	どこの会社でもやってそうだけどね。
N	うちの会社の仕組みが悪いのか

価された．予選では,「どれくらいまた話したいと思うか」という評価軸にて 5 段階評価 (1:全くそう思わない〜5:とてもそう思う) が行われた．評価は 50 人のクラウドワーカーにより行われ，NoneJupiterBot は平均評価 3.15，全 9 システム中 4 位であった．予選通過は 3 位以上であったため，NoneJupiterBot は予選で敗退する結果となった．

表 3 にライブコンペティションの予選における対話例を示した．表からもわかるように，NoneJupiterBot は単語レベルでは関連性の高い応答ができているものの，対話としてはやや不自然な応答となる場合が多く見られた．これは，Elasticsearch により単語の一致を考慮した投稿の検索は成功している一方，BERT による応答選択の性能が十分では無いことが要因であると思われる．なお，表中の「次は逆に最近しんどかったことを聞かせてクレメンス」という発話が，前節で述べた最近の出来事について話すよう促すための発話である．

クラウドワーカーによるコメントを表 4 に示した．高い評価をしたワーカーのコメントからは特徴的な話し方が好印象であったという感想が多く見られた．一方，低評価をしたワーカーはシステムの応答が意味的に不自然であったこと指摘していた．また，2 ちゃんねるで使われる特徴的な口調に対し，好意的な意見もあったが，不快感を表すワーカーも存在した．

表 4: クラウドワーカーによる評価とコメント

評価値	コメント
5	非常にナチュラルな会話が進み楽しかったです。口調は少し個性的でしたがコメントもとても人間ぽいものが多く、心に寄り添ってくれるようなキャラクターが素敵だと思いました。
5	しゃべっていておもしろかったです。すごく参考になるアドバイスをくれました。
4	特殊なネットスラングを多用して実際に居そうな性格だと思った。
4	会話がかみ合いにくい箇所がしばしばありましたが、楽しい雰囲気のボットでした。こちらの会話を理解してくれたら話が弾みそうです。
3	ロボットが一方的に話してるだけのように感じました。
2	話がかみ合ってないように感じられ、やり難かったです。
2	変な日本語使われて、あまり愉快な気持ちがしなかった。
1	今一会話がかみ合いませんでした。

5 まとめ

本稿では，おーぷん 2 ちゃんねる対話コーパスの構築方法について述べるとともに，本コーパスを用いて作成した用例ベース対話システム NoneJupiterBot の構築手法について述べた．本システムは，Elasticsearch によりユーザの発話と関連した用例を検索し，BERT により用例をリランキングすることで応答を選択する用例ベースの対話システムである．対話システムライブコンペティション 2 の予選では全 9 システム中 4 位であり，応答の自然さに課題が残る結果となった．今後は，より適切な応答のリランキング手法について検討していきたい．

参考文献

[1] 東中竜一郎, 船越孝太郎, 稲葉通将, 角森唯子, 高橋哲朗, 赤間怜奈, 宇佐美まゆみ, 川端良子, 水上雅博.

対話システムライブコンペティション 2. 人工知能学会 言語・音声理解と対話処理研究会第 87 回 (第 10 回対話システムシンポジウム), 2019.

[2] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 285–294, 2015.

[3] Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulic, and Tsung-Hsien Wen. A repository of conversational datasets. In *Proceedings of the Workshop on NLP for Conversational AI*, jul 2019. Data available at github.com/PolyAI-LDN/conversational-datasets.

[4] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 496–505, 2017.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.