ここで，異なり語彙率はツイート内における同じ語句の繰り返しが多いものを除外するために設けた．また，Jaccard 係数は同じ語句がリプライ先のツイートに多く含まれるものを除外するために設けた．フィルタリングによって最終的に約 14 億個のチェーンを得た．

トークナイズ　Twitter のツイートには SNS 特有の表現や未知語が多く含まれることから，これに適したトークナイザを用意することが望ましい．本システムでは，トークナイザに sentencepiece[10] を使用し，フィルタリングされた 1,000 万のツイートから語彙数が 32,000 となるように学習した．学習したトークナイザによって，文脈付き入力発話とその応答をそれぞれ入力系列と出力系列へと変換し，それぞれの系列長が 128 以下となるもののみを擬似対話データに用いた[7]．最終的に集まった擬似対話データの事例数は約 14 億であり，事前学習の学習データには 3 億の事例を用いた．学習に用いた擬似対話データセットの事例数を表 4 に示す．

6.2　対話破綻検出チャレンジデータ

6.1 節のデータセットはリプライチェーンを擬似的な対話データとして作成したものであり，実際のユーザ・システム間で起こる対話とはギャップが存在することが想定される．そこで，対話破綻検出チャレンジ [11] のデータセットから実際におこなわれたユーザ・システム間対話を抽出し，系列変換器の fine-tuning に用いた．対話破綻検出チャレンジは，システムと人間の対話ログが与えられたとき，対話中から対話に破綻をもたらすシステム発話を検出するタスクである．タスクに用いられるラベル付き対話データセットから破綻をもたらす発話[8]を除いたうえで，連続して発話が残る部分のみを高品質なユーザ・システム間対話として取り出した．

第 1~3 回の対話破綻検出チャレンジの学習データ及び検証データ，第 4 回の対話破綻検出チャレンジの検証データを用いて，それぞれ系列変換器の学習データ，検証データを作成した．学習に用いた対話データセットの事例数を表 4 に示す．

6.3　話題転換対話データ

システムがユーザを飽きさせない対話を実現するうえで有効な方法として，定期的な話題の転換が考えられる．そこで，指定したタイミングで自然に話題の転換を促す応答を生成可能なシステムの実現に向けて，対話の話題転換前後のやりとりを集めたデータを作成し，系列変換器の fine-tuning に用いた．具体的には，人間の対話において話題が転換する際に「そういえば」などの合図となる語句がしばしば用いられることを利用し，6.1 節のリプライチェーンからこれらの語句を含むツイートを応答として，さらにその応答のリプライ元であるツイートを入力発話とすることで話題転換前後のやりとりを収集した．

応答生成時には適宜指定した話題への転換を促す応答を生成させるため，入力発話に追加で話題語を与えるとその話題語を含んだ応答を系列変換器が生成することを期待し，応答に含まれる話題語を入力発話に連結したデータを作成した．話題語は，日本語 Wikipedia のタイトルに含まれるものとした．学習に用いた対話データセットの事例数を表 4 に示す．

応答生成への利用　話題転換応答を生成させるタイミングで，転換先の話題語をユーザ発話へ連結し応答を生成させることを試みた．しかしながら，与えた話題語を不自然に組み込んだ応答が散見されたことや，話題転換をおこなうタイミングや話題語自体の決定の難しさから，系列変換器の学習データとしては用いたものの，本システムではこの形式の応答生成をおこなわないこととした．今後，対話システムに組み込んでいくためには，自然な話題転換を促す応答の生成方法の改善や話題転換のタイミング，話題語の決定方法のさらなる工夫が必要である．

7　評価・分析

本コンペティションでは，実際にシステムと対話をおこなった 51 人のクラウドワーカにより「自然性 : 対話が自然かどうか」「話題追随 : システムはユーザが選択した話題に関して適切に応答できたかどうか」「話題提供 : システムはユーザが選択した話題に関して新たな情報を提供できたかどうか」という三つの指標それぞれで 5 点を最高点として 5 段階評価がおこなわれた．全ワーカの評価の平均点がシステムのスコアとなり，本システムのスコアは 3.11（自然性 : 3.392，話題追随 : 3.216，話題提供 : 2.725）で予選 2 位の成績となった．実際に対話をおこなったワーカによる評価点の分布を表 5 に，ワーカからのコメントを表 6 に示す．

応答の自然性および話題追随　表 5 より，両指標で半数以上のワーカから 4 点以上の評価を得ており，話題に即した自然な応答ができていたことがわかる．

応答の一貫性　表 6 のコメントから，応答が矛盾を含む場合があることがわかった．原因として，5.1 節で述べた双方向言語モデルが評価する継続度スコアによるフィルタリングの精度が挙げられる．継続度スコアは応答の文脈に続く発話としての自然さの指標であり，矛盾した内容の応答には低いスコアが付与されることが

[7] 入力系列に含まれる複数のツイートの分割箇所を示すため，本システムでは分割トークンを用意し各ツイートの間に挿入した．

[8] 「破綻ではない」とラベルを付与したアノテータが 8 割に満たない発話を，破綻をもたらす発話とした．

表 5: 評価者による評価点の分布

評価点	自然性	話題追随	話題提供
5	3 (5.9%)	4 (7.8%)	3 (5.9%)
4	25 (49.0%)	22 (43.1%)	14 (27.5%)
3	13 (25.5%)	9 (17.6%)	10 (19.6%)
2	9 (17.6%)	13 (25.5%)	14 (27.5%)
1	1 (2.0%)	3 (5.9%)	10 (19.6%)

表 6: 評価者による本システムの評価

コメント [評価点:自然性, 話題追随, 話題提供]
自然に対話ができていると感じた。[評価点:4,4,1]
想像以上に自然だったのでびっくりしました。話の切り替えのついてき方がよかったです。[評価点:5,5,2]
矛盾した内容もありましたが、とても人との会話のような返し方が結構あって面白かったです。[評価点:4,3,4]
もっと、奥の深い専門的な話ができるところまで進めばもっと楽しいなと思います。[評価点:4,2,1]

期待される．しかし，継続度スコアの傾向として，応答が文脈と類似したトピックに関する語句を含む場合には，内容が矛盾していても高いスコアが付与されることがある．そのため，文脈と矛盾した内容の応答をフィルタリングで除去できず，システム応答として選択されてしまうことがある．文脈と矛盾した内容の応答に対して低いスコアを付与できるよう，双方向言語モデルの訓練のさらなる工夫が必要である．

応答による新情報の提供　表 6 のコメントから，会話の話題に関する新情報の提供が十分にできていないという課題が確認された．原因として，知識ベース応答生成モジュールによるルールベース応答を生成する条件が挙げられる．4.3 節で述べたように，知識ベース応答生成モジュールによる応答は，ユーザが知識を問う質問をした場合にのみ生成される．そのため，ユーザが質問をしない限り，知識ベース応答生成モジュールによる新情報の提供がおこなわれない．質問以外の発話に対しても適切なタイミングで，適切な知識を用いた応答を生成することが可能となれば，ユーザへの新情報の提供がさらにおこなえると考える．

8　おわりに

本稿では，対話システムライブコンペティションに参加した対話システム『ILYS aoba bot』について述べた．大規模な疑似対話データを用いて事前学習し，さらに少量の高品質なデータで fine-tuning した大規模系列変換器の応答と，知識ベースを用いたルールベース応答の中から，応答フィルタリングを経て自然かつ話題に即した応答を返すシステムを構築した．人手評価では，ワーカの評価点・コメントからユーザの選択した話題に即した自然な応答が可能であることが確認された．

今後の課題として，応答内容が矛盾している箇所が見受けられ，また，会話の話題に関する新情報の提供が十分にできていないため，人手評価結果を踏まえ，さらなる性能改善に取り組む予定である．

謝辞　本システムの試運転に協力いただいた東北大学乾研究室の皆様に感謝いたします．また，本論文の執筆にあたり，東北大学乾研究室の赤間怜奈氏，阿部香央莉氏から有意義なコメントをいただいたことに感謝いたします．

参考文献

[1] 東中竜一郎, 船越孝太郎, 高橋哲朗, 稲葉通将, 角森唯子, 赤間怜奈, 宇佐美まゆみ, 川端良子, 水上雅博, 小室允人, Dolça Tellols. 対話システムライブコンペティション 3. 第 90 回人工知能学会言語・音声理解と対話処理研究会 (第 11 回対話システムシンポジウム).

[2] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot. In *arXiv:2004.13637*, 2020.

[3] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL*, pp. 48–53, 2019.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pp. 5998–6008, 2017.

[5] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. In *arXiv:1610.02424*, 2018.

[6] 松田耕史, 鈴木正敏, 乾健太郎. Wikidata からの遠距離教師あり学習に基づく大規模関係知識獲得. 言語処理学会第 25 回年次大会 (NLP2019), pp. 659–662, 2019.

[7] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *EMNLP*, pp. 230–237, 2004.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pp. 4171–4186, 2019.

[9] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.

[10] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP*, pp. 66–71, 2018.

[11] 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.