雑談要約技術に向けた取り組み

東中竜一郎　光田航　増村亮　斉藤いつみ　青野裕司

NTT メディアインテリジェンス研究所

1　はじめに

人間の会話のうち，約 6 割は雑談に分類される [10]．対面で，電話で，チャットアプリなどで多くの雑談がなされている．雑談のデータが多くなってくると，過去のやり取りを把握することが困難になってくる．そんな時，雑談の要約が実現できれば，大幅な時間の節約になるだろう．また，自分のやり取りだけではなく，第三者同士のやり取りの要約ができれば，途中から会話に参加することも容易になると考えられる．

本稿では，雑談，特に，人間同士のテキストチャットの自動要約に向けた取り組みについて述べる．要約の研究は一般に抜粋型要約と生成型要約に分けられるが [11]，本稿では生成型要約に着目する．これは，対話型コンテンツは生成型要約の方が好まれることによる [8]．対話に対する生成型要約の研究は複数ある [9, 4]．しかし，いずれもミーティング等の対話に対してのものであり，雑談を扱ったものは見られない．

本稿では，幅広い話題で内容も多様な雑談を対象とした自動要約手法について検討する．具体的には，雑談要約のデータセットを構築し，ベースライン手法や人手要約の精度を評価することで，雑談の要約がどの程度可能であり，何が難しいのかを考察する．

2　雑談要約データセットの構築

雑談の要約データセットを構築するにあたり，どのような要約が望まれるかを検討する必要がある．我々は，対話において誰が何を話したかを第三者がすぐに理解できる要約をよい要約と考えた．その上で，自分が何を話していたか，対話相手が何を話していたかを把握できるようにすることが重要と考え，特定の話者に着目した要約を作成することにした．加えて，二人の間でどういうやり取りがあったかを把握することも重要と考え，特定の話者だけでなく，やり取りに着目した要約も作成することにした．まとめると，以下の二種類の要約を作成することにした．

- 話者に着目した要約
- やり取りに着目した要約

	ショート	ミドル	ロング
対話数	2400		
ターン数	5	10	15
対話に含まれる発話数	9.09	20.97	32.41
対話における話者 A の発話数	5.28	10.11	16.95
対話における話者 B の発話数	3.82	10.86	15.46
対話に含まれる文字数	125.22	320.94	498.26
対話における話者 A の文字数	73.75	150.43	259.68
対話における話者 B の文字数	51.47	167.52	238.58

表 1: ショート，ミドル，ロングデータセットの統計量

データセット構築にあたり，要約のもととなる雑談のデータとして，Higashinaka らが収集したもの [5] を用いた．これらは二者間のテキストチャットであり，全部で 3,680 対話からなる．要約は三種類の長さの対話について行うことにした．それぞれ，5 ターン，10 ターン，15 ターンである．ここで，ターンとは話者が切り替わるまでの一連の発話列のことを指す．

我々は 3,680 対話のうち，15 ターン以上の長さを持つものから，2,400 対話を選択した．そして，これらの対話の始めの 5 ターン，10 ターン，15 ターンの発話を抽出した．これらの長さの対話を含むデータをそれぞれショート，ミドル，ロングデータセットと呼ぶ．統計量を表 1に示す．対話において最初に発話した話者を A とし，もう一方の話者を B としている．話者 B の発話数や文字数が若干話者 A のそれよりも短いのは，奇数ターンからなる対話には 1 ターン分，話者 A の発話が多くなるからである．

ショート，ミドル，ロングデータセットについて，クラウドソーシングによって人手で要約を作成した．ワーカーはランダムに対話テキストを割り当てられ，25 文字以上，50 文字以内で 3 つの要約を作成した．3 つの要約とは，話者 A に着目した要約，話者 B に着目した要約，やり取りに着目した要約である．教示は以下の通りであった．

話者 A/B に着目した要約: 対話において，話者 A/B が何を話して，どう相手に反応したかを把握できる要約を作成する．

やり取りに着目した要約: 対話において，話者の二人

発話 ID	話者 ID	発話
U1	A	先程は私の趣味について話しましたが、
U2	A	あなたの趣味を教えて下さい。
U3	B	母の影響で、最近は海外旅行にはまってます。
U4	A	すごいですね。
U5	B	この間、スペインとポルトガルに行ってきました。
U6	A	何日間行かれたのですか?
U7	B	8 日間ですよ。
U8	A	すてきですね。
U9	B	ツアーで行ったんで、
U10	B	弾丸ツアーでした。
U11	A	忙しかったのですね。
U12	B	そうですね。
U13	B	でも、いろんな都市を回れたんでよかったです。

最初の 5 ターンに対する要約 (U1–U6)	
話者 A に着目した要約	自分の趣味を話した後に、B さんの趣味の海外旅行の期間を訪ねた。
話者 B に着目した要約	趣味が海外旅行で最近スペインとポルトガルに行ったと伝えた。
やり取りに着目した要約	A さんが B さんに趣味を聞き、海外旅行にはまっていて先日もスペイン、ポルトガルに行ってきたとのこと。

最初の 10 ターンに対する要約 (U1–U13)	
話者 A に着目した要約	スペインとポルトガルに行った B さんの日程と内容を聞いている
話者 B に着目した要約	趣味の海外旅行で、先日スペイン、ポルトガル 旅行に行った。8 日間の弾丸ツアーだったが楽しかった。
やり取りに着目した要約	お互いの趣味を話し、B さんは海外旅行が好きで、最近スペインとポルトガルに行った。

図 1: 雑談の要約例．話者 A と話者 B のテキストチャットについて，最初の 5 ターンおよび 10 ターンについて人手で要約したもの．

が何を話し合ったかを把握できる要約を作成する．

各ワーカーは与えられた対話の最初の 5 ターン，次の 5 ターン，さらに次の 5 ターンを順に見ながら，それぞれの長さの対話について要約を作成した．各ワーカーは 36 対話について 3 つずつ要約を作成したので，一人につき 108 の要約を作成した．対話は各ワーカーにランダムに割り振られ，同じ対話について複数のワーカーが要約を作成した．つまり，本データはマルチレファレンス（各対話に対する平均要約数は 5.27）のデータセットである．今回 1,054 名のワーカーに作業を依頼したため，作成された要約の総数は 113,832 である．

図 1にテキストチャットの例と人手で作成された要約の例（5 ターンに対するものと 10 ターンに対するもの）を示す．表 2に，ショート，ミドル，ロングデータセットについて作成された要約の平均文字数を示す．25 文字以上，50 文字以下という制約で作成したが，全体的に 30 文字から 40 文字の要約が作成された．

	ショート	ミドル	ロング
話者 A に着目した要約	34.73	37.38	38.27
話者 B に着目した要約	33.84	38.23	37.84
やり取りに着目した要約	37.43	39.07	39.47

表 2: ショート，ミドル，ロングデータセットについて作成された要約の平均文字数．

3 評価実験

雑談要約データセットをもとに，雑談を要約するというタスクの妥当性を検証する．具体的には，ベースライン手法を実装し，客観評価および主観評価によって，現状どの程度の要約が実現できるのか，何が難しいのかを把握する．

3.1 ベースライン手法

ベースライン手法として，Seq2Seq のモデルを用いたものと近年言語生成にも利用される BERT [3] を用いたものの二種類を用意した．話者に着目した要約については，入力形式の違いにより，以下の 5 つのベースライン手法 (a)–(e) を準備した．

(a) Seq2Seq-attn (tgt): アテンション付きの Seq2Seq モデル [1]．実装には OpenNMT[1]をデフォルト設定で用いた．入力は，対話テキストに含まれる，着目している話者（ターゲット話者）の発話のそれぞれを SentencePiece [6] で分割し，空白区切りで連結したものである．出力側についても，SentencePiece で分割した．

(b) Seq2Seq-attn (both): (a) と基本的には同じだが，入力は，対話テキストに含まれる両方の話者の発話のそれぞれを SentencePiece で分割し，空白区切りで連結したものである．話者 ID などは入力トークン列に含めていない．

(c) BERT-U(tgt): 入出力は (a) と同じだが，Encoder と Decoder に BERT のモデルを用いたもの [2]．実装には OpenNMT-APE[2] をデフォルト設定で用いた．事前学習された BERT のモデルは，`bert-base-multilingual-cased` のものを用いた．

(d) BERT-U(both): 入出力は (b) と同じだが，BERT を用いたもの．

(e) BERT-U(both)+U(tgt): 入力として，両方の話者の発話列のあとに，セパレータ（`[SEP]`）を

[1] `https://github.com/OpenNMT/OpenNMT-py`
[2] `https://github.com/deep-spin/OpenNMT-APE`

	ショート						ミドル					
	話者 A に着目			話者 B に着目			話者 A に着目			話者 B に着目		
ベースライン手法	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
(a) Seq2Seq-attn (tgt)	0.239	0.048	0.200	0.250	0.043	0.200	0.213	0.025	0.161	0.224	0.035	0.182
(b) Seq2Seq-attn (both)	0.258	0.040	0.204	0.233	0.046	0.185	0.243	0.039	0.195	0.223	0.029	0.175
(c) BERT-U(tgt)	0.416	0.186	0.360	0.423	0.197	0.370	**0.346**	**0.119**	**0.282**	**0.321**	**0.103**	**0.257**
(d) BERT-U(both)	0.416	0.186	0.354	0.425	0.194	0.360	0.322	0.096	0.264	0.319	0.096	0.255
(e) BERT-U(both)+U(tgt)	**0.429**	**0.200**	**0.365**	**0.457**	**0.232**	**0.397**	0.307	0.092	0.256	0.317	0.096	0.256
(*) 人手要約	0.420	0.170	0.350	0.440	0.190	0.370	0.370	0.120	0.290	0.360	0.120	0.280

表 3: ショートおよびミドルデータセットに対するベースライン手法による話者に着目した要約の評価結果.

	ショート			ミドル		
ベースライン手法	R-1	R-2	R-L	R-1	R-2	R-L
(f) Seq2Seq-attn (both)	0.300	0.064	0.228	0.226	0.045	0.180
(g) BERT-U(both)	0.390	0.148	0.313	0.314	0.091	0.251
(h) BERT-U(both)+U(A)+U(B)	**0.400**	**0.150**	**0.316**	**0.316**	**0.095**	**0.257**
(*) 人手要約	0.400	0.140	0.310	0.340	0.100	0.260

表 4: ショートおよびミドルデータセットに対するベースライン手法によるやり取りに着目した要約の評価結果.

入れ，そのあとに，ターゲット話者の発話列を追加したもの．対話全体の流れと，ターゲット話者の発話内容の両方を考慮できる可能性がある．

やり取りに着目した要約については，以下の 3 つのベースライン手法 (f)–(h) を準備した．

(f) Seq2Seq-attn (both): (b) と同じ．

(g) BERT-U(both): (d) と同じ．

(h) BERT-U(both)+U(A)+U(B): (d) と同じだが，入力として，両方の話者の発話列のあとに，セパレータを入れ，そのあとに，話者 A，そして，話者 B の発話列を追加したもの．

3.2 客観評価結果

表 3および表 4に，ベースライン手法を ROUGE（R-1, R-2, R-L) [7] で評価した結果について示す．学習は，雑談要約データセットを，対話が重複しないよう 18:1:1 の比率で分け，18 を学習，1 を開発，残りの 1 を評価に用いた．評価は，ショートとミドルのデータセットを対象に行った．ロングは対話が長く，入力が BERT の制限である 512 トークンを超えることが多かったため，今回は評価の対象としなかった．

話者，および，やり取りのどちらに着目した要約についても，Seq2Seq-attn のモデルよりも，BERT のモデルの方が精度が高いことが分かる．また，BERT を用いたものの中でも，(c) や (e)，(h) の ROUGE の値が高いことから，話者それぞれの発話系列を考慮することが重要であることが示唆される．

どちらの表にも人手要約の ROUGE の値を含めている．この値は，複数ある要約のうち，一つの要約をシステム出力，その他を正解要約と見なして ROUGE を計算するということを複数の要約のそれぞれについて行い，その平均を取った値である．これを見ると，ミドルについては，人手要約までには若干の伸びしろがあるものの，ショートについては (e) や (h) のスコアはすでに人手要約を超えている．人手要約よりもベースライン手法の ROUGE が高い理由については，次節で考察する．

3.3 主観評価結果

主観評価を大規模に実施することはコストが高いため，予備実験的に，ショートデータセットに対する，話者 A に着目した要約の結果について，人手による主観評価を行った．評価者数は 4 名である．評価者は要約結果の流暢性と妥当性（以下）を，7 段階のリッカート尺度（7 が一番よく，1 が一番悪い）で評価した．

流暢性 要約文を単体の文章として読んだ際に，日本語の文章として読みやすいかどうか．

妥当性 対話テキストの要約文として妥当かどうか．要約文として妥当とは，対話内容（誰が何を発言し，どう反応したか）を十分把握できることを言う．

表 5に主観評価の結果（4 名の評価者の平均）を示す．表から，表 3とは異なる傾向が見て取れる．つまり，人手要約の方が，流暢性と妥当性のどちらについても大幅に高い．このことから，ROUGE では雑談要約の評価を適切にできていないと考えられる．

ベースライン手法	流暢性	妥当性
(a) Seq2Seq-attn (tgt)	4.98	1.01
(b) Seq2Seq-attn (both)	6.07	1.03
(c) BERT-U(tgt)	4.58	4.14
(d) BERT-U(both)	4.89	3.77
(e) BERT-U(both)+U(tgt)	4.91	4.41
(*) 人手要約	5.57	5.55

表 5: ショートデータセットの話者 A に着目した要約の主観評価結果.

Maximum Likelihood に基づく生成モデルは平均的な要約を生成するため，それなりに正解とマッチし ROUGE が高くなっていると考えられる．一方，人間の要約は，よりバリエーションがあり，ROUGE は低くなるものの，流暢かつ適切に対話内容を捉えたものになっていると考えられる．雑談要約では，ROUGE 以外の尺度による評価が必要ということが示唆される．

3.4 出力結果の例

図 2に，ショートデータセットに対するベースライン手法の話者 B に着目した要約の出力例を示す．Seq2Seq-attn (tgt) では，学習データの影響か，お酒についての要約となってしまっている．また，一般的で具体性のない dull summary になっている．Seq2Seq-attn (both) は全く関係のないビーズアクセサリについて言及している．BERT に基づく手法は，対話の内容を比較的踏まえている．ただ，BERT-U(tgt) は，同じ内容を繰り返しており，BERT-U(both) は内容に誤りがある（イージーリスニング系が多いのは A であり B ではない）．ここでは，BERT-U(both)+U(tgt) のみが適切な要約を出力できている．

発話 ID	話者 ID	発話
U1	A	こんにちは。
U2	B	こんにちは。
U3	B	音楽は聞かれますか?
U4	A	はい。
U5	A	結構聞きます。
U6	B	どんなものを聞かれるのでしょう?
U7	B	私はパンクやテクノが多いです。
U8	A	最近勉強したりすることも多いので、イージーリスニング系が多いです。
U9	A	あと、クラシックとか。
U10	A	変わった所で、サウンドホライズンも聞きます。

正解（人手）: A さんに好きな音楽について尋ね、自身はパンクやテクノを好む。
Seq2Seq-attn (tgt): A さんにお酒を飲むか聞かれたので、あまり行っていないが、A さんはどうか尋ねた。
Seq2Seq-attn (both): A さんに趣味を尋ね、ビーズアクセサリー作りと聞き、自分も好きだと答え、A さんの答えた。
BERT-U(tgt): 音楽はパンクやテクノが多いが自分はパンクやテクノが多い
BERT-U(both): 音楽はパンクやテクノが多いので、イージーリスニング系が多い。
BERT-U(both)+U(tgt): A さんに音楽を聴くか尋ね、自分はパンクやテクノが多いと答えた。

図 2: ショートデータセットに対するベースラインモデルによる話者 B に着目した要約の出力例.

4 おわりに

本稿では，雑談要約技術に向けた取り組みとして，話者，および，やりとりに着目した要約を新たに設定した．そして，雑談要約データセットの構築とベースライン手法，および，人手要約の評価について述べた．ベースライン手法の BERT の ROUGE スコアは高かったが，人手による主観評価では人手要約との差は依然大きく，流暢性・妥当性において改善の余地があることが確認できた．今後は，主観評価値が高くなるような要約アルゴリズムの検討を行いたい．また，雑談要約に適切な評価尺度についても検討したい．

参考文献

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proc. ICLR*, 2015.

[2] Gonçalo M Correia and André FT Martins. A simple and effective approach to automatic post-editing with transfer learning. In *Proc. ACL*, 2019.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL*, 2019.

[4] Chih-Wen Goo and Yun-Nung Chen. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *Proc. SLT*, 2018.

[5] Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. Towards an open-domain conversational system fully based on natural language processing. In *Proc. COLING*, pp. 928–939, 2014.

[6] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proc. EMNLP*, 2018.

[7] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

[8] Gabriel Murray, Giuseppe Carenini, and Raymond Ng. Generating and validating abstracts of meeting conversations: A user study. In *Proc. INLG*, pp. 105–113, 2010.

[9] Tatsuro Oya, Yashar Mehdad, Giuseppe Carenini, and Raymond Ng. A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In *Proc. INLG*, pp. 45–53, 2014.

[10] 小磯花絵, 石本祐一, 菊池英明, 坊農真弓, 坂井田渡部 涼子, 田中弥生, 伝康晴. 大規模日常会話コーパスの構築に向けた取り組み—会話収録法を中心に—. *SIG-SLUD*, Vol. 5, No. 01, pp. 37–42, 2015.

[11] 奥村学, 難波英嗣. テキスト自動要約. オーム社.