E-042

対話要約における話者情報を持つ Embedding の効果
The Effects of Embedding with Speaker Identity Information in Dialogue Summarization

楢木悠士
早稲田大学理工学術院
yuji.1277@akane.waseda.jp

酒井哲也
早稲田大学理工学術院
tetsuyasakai@acm.org

林 良彦
早稲田大学理工学術院
yshk.hayashi@aoni.waseda.jp

1 はじめに

膨大なテキスト情報を的確に把握するための手段として自動要約の重要性が増している．また，インターネット上での会議などが増えるにつれ，そこで発生する対話内容を簡潔に要約する対話要約の必要性も増大している．

自動対話要約は，一般的な文書要約と異なり，客観的に発言を捉えなければならない．本研究で扱う対話の形式 (図 1) では，話者名にその話者の発言が続く．この場合，それぞれの発言は発言の前に示された話者のものと理解しなければ，正しく要約することはできない．現在主流の要約手法である PEGASUS [1] や ERNIE-GEN [2] などの Transformer ベースの Encoder-Decoder モデルは通常のテキストコーパスに対する事前学習を前提とした手法であるが，そのままでは発話とその話者の関係性を維持した対話要約の生成には適さない．

対話において話者情報を陽に扱う研究として，Gu ら [3] は話者交代を意味する Speaker Embedding (SE) を導入し，返答発言選択タスクにおいて性能を向上させた．楢木ら [4] は，SE の次元数を減らし，Position Embedding の情報性の乏しい箇所にのみ SE を加算することで，対話要約学習における収束性を改善し，さらに ROUGE スコアの向上を達成した．

しかし，楢木らの報告 [4] では，Speaker Embedding の加算箇所の違いによる効果の違いについての系統的な検証は行われておらず，また，SE がどのような観点で生成要約を改善するかについての考察も行われていない．そこで本論文では，新たな検証実験・人手評価を通して楢木らの手法の有効性を検証する．

Dialogue
Mary: Hi Mike!
Mike: Hello :)
Mary: do u have any plans for tonight?
Mike: I'm going to visit my grandma. You can go with me. She likes u very much.
Mary: Good idea, i'll buy some chocolate for her.

Summary
Mike and Mary are going to visit Mike's grandma tonight. Mary will buy her some chocolate.

図 1 SAMSum の対話文と要約の例

2 検証実験

本研究では対話要約モデルとして最先端の文書要約モデルである PEGASUS [1] を用いる．PEGASUS は Transformer を基にしたモデルであり，入力には対話文から得られる Token Embedding と Position Embedding を用いる．本研究では，対話文を扱うため，これら 2 種類の Embedding に加え，Speaker Embedding も入力の一部として用いる．

2.1 Speaker Embedding の加算箇所

Gu ら [3] の提案する SE は話者ごとに異なる埋め込み表現をトークン毎に与える手法で，対話の返答発話選択タスクにおいて性能の向上を示した．しかし，Gu らの SE は対話要約タスクに対して悪影響を及ぼすことが報告されている [4]．

楢木ら [4] は，実験に用いた PEGASUS モデルの Position Embedding である Sinusoidal Positional Embedding の持つパラメータの傾向に着目し，SE

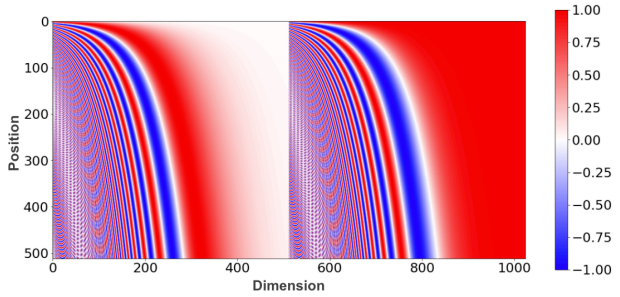

図 2　Sinusoidal Positional Embedding のパラメータのヒートマップ

表 1　Speaker Embedding を加算する次元とそのパターンの名称

	SE の次元数	前半次元	後半次元
whole SE	1024	0〜511	512〜1023
former half SE	512	0〜255	512〜767
latter half SE	512	256〜511	768〜1023
first quarter SE	256	0〜127	512〜639
second quarter SE	256	128〜255	640〜767
third quarter SE	256	256〜383	768〜895
fourth quarter SE	256	384〜511	896〜1023

の加算箇所の提案を行った．Sinusoidal Positional Embedding は式 1，2 によって定義された固定のパラメータを持つ Embedding であり，そのパラメータの分布は図 2 のヒートマップで表される．ここで，*pos* はシーケンス上の位置，*i* は次元を表し，*dim* はモデルに入力する Embedding の次元数を表す．

$$PE_{(pos,i)} = \sin\left(pos/10000^{2i/dim}\right) \quad (1)$$

$$PE_{(pos,i+dim/2)} = \cos\left(pos/10000^{2i/dim}\right) \quad (2)$$

楢木らは図 2 における色の変化が小さい，つまり，情報性の乏しい次元に限定して SE を加算することで，生成要約の ROUGE スコアの向上を達成した．この手法は Partial SE と呼ばれている．しかし，彼らの分析では加算箇所による効果の違いを系統的に検証していない．そこで本研究では，Partial SE の加算箇所について，いくつかのパターンで追加実験を実施し，有効な SE の加算箇所を分析した．実験で用いた加算箇所のパターン一覧を表 1 に示す．本研究に用いた PEGASUS モデルへの入力の全体の次元数は 1024 である．

2.2　人手評価

評価指標としては，楢木ら [4] と同様の ROUGE に加え，要約のプリファレンスおよび要約の誤りに関する人手評価を Amazon Mechanical Turk を用いて実施した．どちらの評価においても 1 つのデータに対して 5 人が評価を行なった．

表 2　SAMSum とフィルタリングした MediaSum のデータ数

	train	validation	test
SAMSum	14732	818	819
Filtered MediaSum	18888	472	456

Before Preprocessing
Amanda: I baked cookies. Do you want some?
Jerry: Sure!
Amanda: I'll bring you tomorrow :-)

After Preprocessing
Amanda *[SAYS]* I baked cookies. Do you want some? *[SEP]* Jerry *[SAYS]* Sure! *[SEP]* Amanda *[SAYS]* I'll bring you tomorrow :-)

図 3　前処理前と前処理後の対話例

1 つ目の人手評価は SE を用いない場合と fourth quarter SE を用いる場合の生成要約のプリファレンスの評価である．評価者には 1 つの対話文とランダムな順序に並んだ 2 つの生成要約が表示される．評価に用いるデータは，対話文が 512 トークン以下で，2 つの場合による生成要約の Jaccard 係数が 0.8 以下のデータとした．このフィルタリングにより，モデルが対話文を全て入力でき，かつ，2 つの要約に十分に差のある場合を抽出した上で，ランダムに選択した 100 個のデータを評価に用いた．また，評価者は優劣の判断理由を 4 つの選択肢から複数選択可で選んだ．これらは，1) より自然な文になっている，2) 事実の誤りがより少ない，3) より重要な箇所を含んでいる，4) その他の 4 つである．

2 つ目の人手評価は生成された要約に主述関係の誤りがあるかどうかの評価である．プリファレンスの評価と同様に，SE を用いない場合と fourth quarter SE を用いる場合の生成要約の誤りを評価した．評価者には 1 つの対話文と 1 つの生成要約が与えられ，要約に主述関係の誤りがあるかどうかを評価する．モデルが対話文を全て入力できる場合を考えるため，対話文が 512 トークン以下のデータから，それぞれの手法についてランダムに 50 個ずつ評価に用いるデータを選んだ．

2.3　データセット

SAMSum [5] と MediaSum [6] の 2 つの対話要約データセットを用いてファインチューニングを行った．SAMSum は言語学者によって人手で構築され

た高品質なデータセットであり，世間話や会議の手配などあらゆる日常会話と抽象要約で構成されている．SAMSum において，要約は話者名を含めるべきとされており，話者名と発言の繋がりを考慮しなければならない．MediaSum は実際のインタビューを収集した大規模なデータセットであり，対話記録と抽象要約で構成されている．MediaSum に収録されている対話は比較的長く，本研究で用いるモデルの最大入力長を超えるデータが約 95%を占める．そこで本研究では 512 トークン以下の対話を持つデータのみを扱った．SAMSum のデータ数とフィルタリングした MediaSum のデータ数を表 2 に示す．

図 1 に対話文と要約の例を示した SAMSum データの前処理においては，対話文中の話者名と発言の境界を明示するために 2 つの特殊トークンを挿入した．[SEP] トークンは話者の交代を表し，[SAYS] トークンは話者名と発言の区切りを表す．2 種類の特殊トークンを挿入した入力例を図 3 に示す．モデルの初期パラメータは，C4 [7] と HugeNews [1] で事前学習し，XSum [8] で学習した重みとした．

3 検証実験の結果と考察

3.1 Speaker Embedding の加算箇所と効果

表 1 にあげた 7 手法と SE を用いない場合の ROUGE スコアを表 3 に示す．また，いくつかの既存手法によるスコアを論文から引用した．表 3 から latter half SE や third quarter SE，fourth quarter SE のような PE の情報性の乏しい箇所に加算する場合は SE を用いない場合よりも ROUGE スコアが向上し，PE の情報性が高い箇所に加算する場合は ROUGE スコアが低下することが示された．この傾向は SAMSum においてもフィルタリングした MediaSum においても確認できるため，データセットに依存しない効果であると示唆される．どちらのデータセットにおいても，fourth quarter SE が最も良いスコアを出した．

さらに，加算箇所と収束性の関係を分析するため，一部手法による validation 損失の推移を図 4 に示す．PE の情報性の高い箇所に SE を与えた場合には，SE を用いない場合と比べて，損失が十分に下がっていないことがわかる．一方，latter half SE や fourth quarter SE では，十分に損失が下がり，かつ，収束速度が速いことがわかる．ROUGE スコアと同様に，この傾向はフィルタリングした MediaSum においても確認できた．

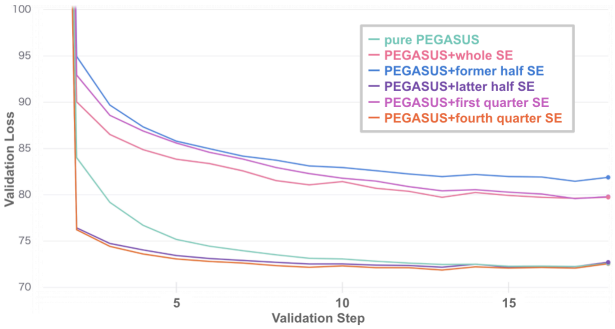

図 4 Validation 損失の変遷

以上の結果から，PE の情報性の高い箇所に SE を加算すると，損失が十分に収束せず，生成要約の ROUGE スコアが低下する．一方，PE の情報性の乏しい箇所に SE を加算すると，ROUGE スコアの向上と，収束性の向上をもたらした．PE はトークンの絶対的な位置・相対的な位置をモデルに示す重要な埋め込み表現である．ファインチューニングから新しく SE を追加する際，PE の情報性の高い箇所に加算すると，入力表現の内のトークンの位置関係に関する情報に悪影響を及ぼし，事前学習による収束解から大きく離れてしまう．よって，モデルはトークンの位置を上手く考慮できず，十分に収束できないと考えられる．一方，PE の情報性の乏しい箇所のみに SE を加算すると，トークンの位置関係に関する情報に大きな影響を与えないため，事前学習による収束解から大幅に外れないと考えられる．結果として，モデルはトークンの位置だけでなく，話者を特定する情報を考慮することができ，対話要約タスクにおけるより良い解に収束できたと考えられる．

3.2 人手評価

SE を用いない場合と fourth quarter SE を用いる場合の生成要約のプリファレンスの評価の結果を表 4 に示す．fourth quarter SE を用いた場合の方が，多数決による勝利数が多いことから，生成要約の質を向上していると考えられる．さらに，優劣の判断理由として得票数に十分な差がついた理由は「より重要な箇所を含んでいる」であり，fourth quarter SE を用いることで，対話文から重要な箇所をまとめることができるようになったと示唆される．他の 2 つの理由には大きな差が見られないため，fourth quarter SE は自然な文を生成するものの，事実の誤りを減らす効果は明確ではないと考えられる．

次に，生成された要約に主述関係の誤りがあるかどうかを評価した結果について考察する．多数決に

表 3 テストデータにおける ROUGE スコア

Method	SAMSum (n = 819)			filtered MediaSum (n = 456)		
	R-1	R-2	R-L	R-1	R-2	R-L
Longest-3	32.46	10.27	29.92	-	-	-
Transformer [9]	36.62	11.18	33.06	-	-	-
TGDGA [10]	43.11	19.15	40.49	-	-	-
PEGASUS	50.82	26.62	42.65	38.65	20.54	34.79
PEGASUS + *whole SE*	46.40	21.92	38.52	34.84	15.77	30.53
PEGASUS + *former half SE*	44.4	19.48	36.20	33.62	15.24	29.80
PEGASUS + *latter half SE*	51.39	27.03	43.39	37.80	20.76	34.39
PEGASUS + *first quarter SE*	46.19	20.99	38.04	34.36	15.15	30.11
PEGASUS + *second quarter SE*	48.56	24.81	40.80	38.29	20.65	34.21
PEGASUS + *third quarter SE*	51.17	27.37	43.51	38.85	21.01	35.11
PEGASUS + *fourth quarter SE*	**51.39**	**27.58**	**43.61**	**40.05**	**21.90**	**36.20**

表 4 プリファレンス評価の結果: 理由 1) より自然な文になっている，理由 2) 事実の誤りがより少ない，理由 3) より重要な箇所を含んでいる

手法	プリファレンス勝利数	理由 1	理由 2	理由 3
PEGASUS	37	28	20	47
PEGASUS + *fourth quarter SE*	**63**	24	12	28

より主述関係の誤りがあると認められたデータ数は，SE を用いない場合は 10 個，fourth quarter SE を用いた場合は 12 個であり，2 つの手法に大きな差は見られなかった．よって，fourth quarter SE が主述関係の誤りの低減に効果があるとは言えない．主述関係の誤りも事実の誤りの一種であり，この結果は fourth quarter SE が事実の誤りの低減に効果が乏しいというプリファレンス評価の結果と矛盾しない．

4 おわりに

本研究では対話要約において，楢木らの既存研究において未検証であった話者識別情報を含んだ Speaker Embedding を加算する箇所を変えた場合の効果を実験的に分析した．Position Embedding の情報性の乏しい箇所に SE を与えることで，生成要約の ROUGE スコアの向上と，収束性の向上を確認した．これらの効果は入力表現に含まれるトークンの位置情報を阻害することなく，話者アイデンティティ情報を与えたからだと考えられる．さらに，人手評価においても fourth quarter SE は生成要約の質の向上が認められ，より重要な箇所をまとめられていることが示唆された．

今後は，学習可能な Position Embedding に対しての Speaker Embedding を導入した実験を行い，より多くのモデルに対しての有効性を検証する．また，対話要約以外の対話を扱うタスクに対して Speaker Embedding の加算箇所を変えることが与える影響についても実験的に分析する．

参考文献

[1] Jingqing Zhang et al. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[2] Dongling Xiao et al. Ernie-gen: An enhanced multi-flow pre-training and fine-tuning framework for natural language generation. *arXiv preprint arXiv:2001.11314*, 2020.

[3] Jia-Chen Gu et al. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, 2020.

[4] 楢木悠士, 酒井哲也. 話者情報を考慮した対話要約. 言語処理学会第 27 回年次大会, B7-4, 2021.

[5] Bogdan Gliwa et al. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 70–79, Hong Kong, China, November 2019. Association for Computational Linguistics.

[6] Chenguang Zhu et al. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5927–5934, Online, June 2021. Association for Computational Linguistics.

[7] Colin Raffel et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, Vol. 21, No. 140, pp. 1–67, 2020.

[8] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[9] Ashish Vaswani et al. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[10] Lulu Zhao, Weiran Xu, and Jun Guo. Improving abstractive dialogue summarization with graph structures and topic words. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 437–449, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.