診断のサマリー作成支援に向けた会話記録のトピック分類

橋本和磨 [1]　北出祐 [2]　辻川剛範 [2]　久保雅洋 [2]
[1] 早稲田大学理工学術院
[2] 日本電気株式会社 バイオメトリクス研究所
84moto@fuji.waseda.jp
{t-kitade,tujikawa,masahirokubo}@nec.com

概要

2024 年度から施行された医師の働き方改革によって，医療現場では医師の業務の効率化が求められている．特に，書類の作成は大きな負担となっており，書類作成業務支援が医師の労働時間短縮に貢献することが期待されている．本研究では，医師による診察のサマリー作成を支援するため，外来診察における医師と患者の模擬会話記録を用いて，発話ごとにトピックのマルチラベル分類を行う手法を提案した．実験の結果，提案手法はベースライン手法を上回る性能を示し，診察会話における発話ごとのトピック分類において有効性が確認された．

1　はじめに

医師が行う書類作成業務の１つとして外来患者の診断時のサマリー作成業務がある．このサマリーは患者の診断情報を簡潔にまとめた記録であり，診療の継続性や医療連携を支える上で重要な役割を担っている．例えば，多くの疾患を抱える患者の医療情報やこれまでの治療の経時的変化を，電子カルテから迅速かつ正確に把握することは容易ではない．必要な情報が簡潔に整理されたサマリーは患者の全体像を把握し，適切な診断や治療方針の検討，さらには他の医療機関や診療科への引き継ぎ時にも有用である [1, 2].

しかしながら，このような書類作成業務は医師の所定外労働を引き起こす要因として挙げられている [3]．特に，日本の医療は医師の長時間労働によって支えられてきた背景があり，医師の労働環境の改善が課題とされていた．2024 年度から施行された医師の働き方改革によって時間外労働の上限規制が行われ，医療現場では業務効率化のさらなる推進と医師の負担軽減が求められるようになった [4]．医師の働き方改革に関する施行半年後の調査 [5] では，約 7 割の医師が AI 技術の活用に期待を寄せていることが示されており，特に AI による書類作成業務支援が労働時間短縮に貢献することが期待されている．

表 1　トピックの種類

トピック	説明
病名・病態	病気名と病気の状態
治療目的・代替治療	治療を行う目的と他の治療方法
手術内容	提案する手術の詳細
術後〜退院	手術後の治療や退院時期
合併症	疾患が及ぼす合併症リスク
同意撤回・SO	医療行為の同意取消しやセカンドオピニオン
質問	診察内で発生する質問
回答	質問に対する回答

そこで，医師のサマリー作成支援を目的とした，診断を行う医師と患者の会話記録から各発話のトピックを分類する機械学習手法を提案する．現状，医師はサマリーを作成する際に必要な医療情報を，電子カルテを参照することで抽出している [6]．診察時の会話からサマリーで重要とされるトピックについて自動的に情報を抽出し提示することで，医師のサマリー作成負担の軽減が見込まれる．具体的には，外来診察の文字起こしデータに対して，表 1 に示した８つのトピックをアノテーションした．そして，学習したモデルを用いて発話テキストのトピックを予測し，モデルの性能を評価した (図 1).

2　関連研究

病院やクリニックにおける会話に，言語モデルを適用させた研究はこれまでにいくつか提案されている [7, 8].

Liu らは，クリニックにおける糖尿病管理のフォ

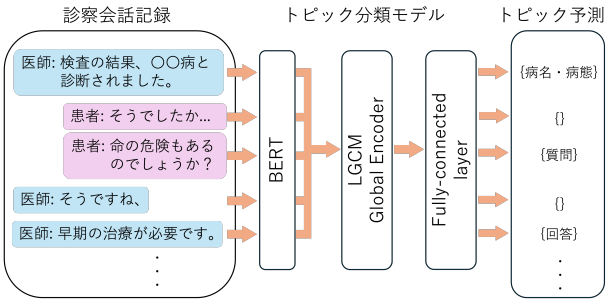

図 1　本研究の手法概要

ローアップ通話から得られる会話データを用い，看護師と患者の会話に対するトピック分類を実施した [7]．具体的には，会話中のトピックの始まりから終わりまでのセグメントを予測し，さらにそれぞれのセグメントが該当するトピックの予測を行う，セグメンテーションとクラス分類を組み合わせた結合モデルを提案している．この結合モデルは，発話レベルの埋め込み表現を得る BERT [9] モジュールと，会話レベルで処理する BiLSTM [10] モジュール，会話のトピックの切り替わりを予測するセグメンテーションモジュールおよびトピック分類モジュールによって構成される．

会話レベルの表現を活用する背景には，診察会話特有の性質が関係している．一般的に診察では医師と患者がそれぞれ特定の目的を持って会話を行うため，自由な会話とは異なり話し手同士が目標や議題に沿って会話の流れを調整する傾向がある [11]．一方で，文書ほど意味密度が高いわけではなく，会話中にトピックが頻繁に切り替わることやトピックが前後することもあるため，発話のみならず会話の文脈を理解してトピックを予測することが重要である．

このような会話における文脈理解に注目した研究の一つに LGCM [12] がある．LGCM では，会話における局所的な文脈と大域的な文脈を効果的に統合することで，会話全体の流れや発話間の関係を理解するモデルを提案している．具体的には，各発話を処理する Local Encoder と会話全体の文脈を捉える Global Encoder から構成されるエンコーダーを用いることで，統合的な文脈表現を学習し，モデルの性能向上を図っている．

3　データセット

本研究では，データセットとして外来診察における医師と患者の模擬会話記録を用いた．この模擬会話記録は，診察室での実際の会話を想定して作成された 290 件の診察会話から構成される．

表 2　診察会話の例

話者	発話文	トピック
患者	よろしくお願いします。	
医師	よろしくお願いします。	
医師	本日は、	
医師	B さんの診断結果と、	
医師	今後の治療方針についてご説明させていただきます。	
患者	はい、	
患者	お願いします。	
医師	まず、	
医師	B さんの病気は舌癌と診断されました。	病名・病態
医師	これは舌の細胞が異常に増殖する病気で、	病名・病態
医師	B さんの場合は扁平上皮癌という種類です。	病名・病態
患者	え、	
患者	癌？	質問
患者	うそでしょ。	
医師	いきなり癌と聞くとご不安になると思いますが、	
医師	幸いにも B さんの病状は初期のステージ 1 で、	病名・病態, 回答
医師	全身への転移は見られませんのでその点はご安心ください。	病名・病態, 回答
患者	そうなんですね。	

3.1　データ準備

すべてのデータは診察会話中に発生するポーズごとに区切られた発話単位によって構成されており，各発話には話者，発話テキストおよび該当するトピックの情報が用意されている．データセットには医師や患者に加え，看護師や患者の家族が話者である発話も含まれるが，本研究では話者を医師または患者に限定して扱った．具体的には，看護師の発話は医師として，患者の家族の発話は患者としてそれぞれ再分類した．また，トピックについては医師のサマリー作成支援を想定し，8 つのトピックについて各発話に手動でアノテーションを行った．一般的なトピック分類と異なり，サマリーに記載されるべき内容が含まれた発話に対してのみラベルが付与されている．そのため，サマリーに必要ない発話はトピックに関わらずラベルが付与されない．その結果，該当するトピックが複数ある発話や，いずれのトピックにも当てはまらない発話が存在するデータとなった．会話例を表 2 に示す．

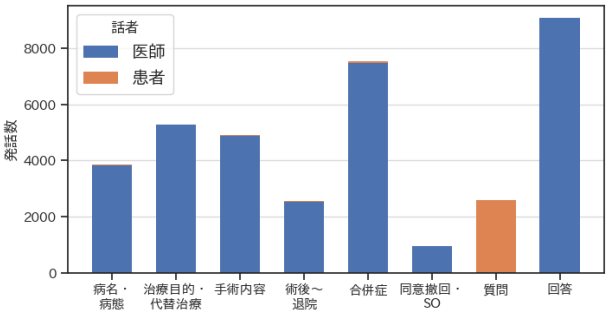

図 2　トピック別発話数

3.2　データ統計

本研究で用いた診察会話記録では，1 診察あたりの平均発話数はおよそ 258 であり，最大発話数は 421，最小発話数は 182 であった．全体の発話のうち，医師が話者であるものは 72.2%を占め，残りの 27.8%が患者による発話であった．

全ての発話のうち，58.4%はいずれのトピックにも該当せず，１つのトピックが与えられた発話は 35.8%あまりだった．トピックが２つ与えられた発話と３つ与えられた発話はそれぞれ 5.5%，0.2%であり，４つ以上のトピックが与えられた発話は存在しなかった．

トピックが与えられた発話の内訳を図 2 に示す．8 つのトピックの分布を見ると，患者の発話が該当するトピックの多くが「質問」であり，医師の発話が該当するトピックの多くは「質問」以外である．トピックごとの出現頻度には顕著なばらつきがあり，このデータセットはトピック間に偏りが存在する不均衡データであることがわかる．

4　提案手法

本研究では，東北大学が公開している事前学習済み日本語 BERT モデル（cl-tohoku/bert-base-japanese-v3[1)]）をファインチューニングすることで，会話文から発話ごとにトピックを予測する．

具体的には，BERT と LGCM の Global Encoder を組み合わせたモデルを作成する．この提案手法の枠組みは，発話の分散表現を得るための BERT，発話間の関係を得る LGCM の Global Encoder および判別器によるマルチラベル分類からなる．

モデルの入力として発話テキストのみを入れるモデルと発話テキストと話者情報を入れるモデルの２種類を作成した．

1）　https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

5　実験

5.1　ベースライン

ベースラインとして LGCM と，トピックセグメンテーションとクラス分類の結合モデルを用意した．LGCM は Decoder を判別器に置き換えて学習を行い，結合モデルの BERT は提案モデルと同じ事前学習済み BERT を用いてファインチューニングを行った．

5.2　実験設定

実験には我々の模擬診察会話データを学習データ，開発データ，テストデータでそれぞれ 8:1:1 の割合で使用する．1 回の診察で行われる会話量は非常に多いため，設定した最大系列長を超える診察についてはデータの分割を行った．

損失関数は，結合モデルではトピックセグメンテーションの損失とバイナリ交差エントロピーの和，その他のモデルではバイナリ交差エントロピーのみとした．

5.3　評価指標

モデルの性能を評価するために，マルチラベル分類性能の指標として標準的な Accuracy，Macro-F1 及び Macro-AUPRC に加え，実用上の懸念を考慮して Accuracy(overlap)(式 1) を導入した．Accuracy では１つのトピックだけ予測を外した場合も全く異なる予測をした場合であってもどちらも等しく不正解として扱われる．しかし，実用上の観点から考えると，ある発話が実際に該当するトピックを１つでも検出できていれば医師のサマリー作成支援に十分役立つと考えられる．そこで，N 件の発話の中で予測ラベル Yi と正解ラベル Ti が部分的に重複する場合にも正解とみなした時の精度である Accuracy(overlap) を含めた，４つの指標によってモデルを評価した．また，全てのモデルで評価指標は発話ごとに計算された．

$$Accuracy(overlap) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(\begin{cases} Y_i = \mathbf{0} & \text{if } T_i = \mathbf{0} \\ Y_i \cap T_i \neq \emptyset & \text{if } T_i \neq \mathbf{0} \end{cases}\right) \tag{1}$$

5.4　結果

提案手法とベースラインの実験結果を表 3 に示す．結果から提案手法はベースラインに比べて全て

表 3 トピック分類の実験結果

モデルタイプ	Accuracy	Accuracy(overlap)	F1	Macro-AUPRC
結合モデル	0.684	0.738	0.620	0.667
LGCM	0.721	0.780	0.617	0.640
BERT + Global Encoder	0.779	0.826	0.679	0.737
BERT + Global Encoder + 話者	**0.781**	**0.827**	**0.701**	**0.757**

表 4 トピックごとの Micro-AUPRC

	BERT + Global Encoder	BERT + Global Encoder + 話者
病名・病態	0.801	0.810
治療目的・代替治療	0.640	0.618
手術内容	0.726	0.753
術後〜退院	0.740	0.746
合併症	0.920	0.918
同意撤回・SO	0.894	0.931
質問	0.650	0.701
回答	0.524	0.577

の評価指標において上回っており，モデルの性能が高いことがわかる．また，話者情報を含めた提案手法は含めないモデルよりも全ての指標でわずかな向上が見られた．

トピックごとの性能を評価するために，提案手法における Micro-AUPRC を表 4 に示す．結果より，どちらの提案手法もトピックによって Micro-AUPRC に差があることがわかる．また，多くのトピックでは話者情報の有無がトピックの AUPRC にほとんど影響しない一方で，「質問」と「回答」では話者情報を含めることによる性能の向上が顕著に見られた．

6 考察

提案手法がベースラインよりも高い性能を示したことから，BERT による局所的な表現と Global Encoder による発話の前後関係を組み合わせることで，会話全体の文脈を考慮してトピックを予測できることが示唆された．また，事前学習済みモデルを Local Encoder の代わりに用いたことで，提案モデルが効果的な局所的表現を捉えたと考えられる．

一方で，結合モデルが高い性能を示さなかった要因として，セグメンテーションが本研究で用いたデータセットに適さなかった可能性が示唆される．先行研究のデータセットと本研究のデータセットの大きな違いとしてトピックの粒度がある．本研究のデータセットは先行研究と異なり，発話単位でトピックが細かくアノテーションされていたため，同じトピックが連続することが少なくセグメンテーションによる効果が薄かった可能性がある．

また，表 4 に示すように，トピック間で AUPRC に大きな違いが見られた．トピックが該当する発話総数が最も少ない「同意撤回・SO」においては予測性能が高く，最も多い「回答」では予測性能が低い．このことから，モデルの性能はトピックの出現頻度ではなく，トピックごとの判断の難しさが影響したと考えられる．例えば，「同意撤回・SO」では使用される表現がほぼ一定で話される内容に大きな違いがないため，予測が容易であると推測される．

本研究で使用した学習データにおいて，「質問」に該当する発話の話者は主に患者であった．このデータの特性により，話者情報が与えられることで，「質問」のトピックを判断する際に有効に機能したと考えられる．また，「回答」は一般的に会話の中で「質問」に続いて出現すると考えられる．したがって，「質問」の性能が向上したことでそれに関連する「回答」の性能も向上した可能性が高い．結果として，質問と回答の性能の向上が全体的な評価指標のわずかな向上に寄与したと考えられる．

7 おわりに

本研究では，診察における模擬会話データを用いて発話ごとのトピック分類を行った．実験の結果，事前学習済み BERT モデルをベースに LGCM の Global Encoder を組み合わせた提案手法は，ベースラインと比較して高い性能を示した．トピックごとの評価では，同じモデルでもトピックによって性能に大きな差があることが確認された．さらに，モデルに話者情報を加えて学習することによる全体的な性能向上はわずかであったが，一部のトピックでは性能が顕著に向上した．

本研究で提案した手法は，医師がサマリーの作成時の情報抽出を支援するだけでなく，LLM など生成モデルと組み合わせることによりサマリー作成の精度向上に貢献することもできると考えられる．

今後の課題としては，他環境のデータセットや診察現場での本手法の有効性を確認することである．

謝辞

本研究は日本電気株式会社の 2024 年度研究インターンシップの一部として実施されたものである．

参考文献

[1] Shuyi Zhou, Kazue Takayanagi, and Tetsuhiko Kimura. Research on the influences that affect the physician’s perception of discharge summaries. a survey of outpatient department physicians. **Journal of Nippon Medical School**, Vol. 66, No. 4, p. 270–278, 1999.

[2] 荒川迪生, 川出靖彦, 吉田麗己, 山北宜由, 遠渡豊寛, 宮治眞, 加藤憲, 吉田達彦. かかりつけ医における外来診療情報の年刊サマリー　—診療の質を高め, 情報を共有する—. 医療情報学, Vol. 32, No. 1, pp. 11–18, 2012.

[3] 厚生労働省. 平成 29 年度過労死等に関する実態把握のための労働・社会面の調査研究事業, 2018. [https://www.mhlw.go.jp/content/11200000/000511979.pdf](cited by 12/4/2024].

[4] 厚生労働省. 医師の働き方改革に関する検討会報告書, 2019. [https://www.mhlw.go.jp/content/10800000/000496522.pdf(cited by 12/4/2024)].

[5] Ubie 株式会社. 医師の約 7 割が「働き方改革による労働時間短縮を実感せず」、ai は 7 割・生成 ai は 5 割以上と活用への期待が高まる, 2024. [https://prtimes.jp/main/html/rd/p/000000096.000048083.html(cited by 12/4/2024)].

[6] 宇野裕, 石井亮, 柴田大作, 辻川剛範, 中川敦寛, 久保雅洋, 香取幸夫. 治療経過サマリ作成支援のための診察記事からの医療情報抽出. 医療情報学連合大会論文集, 第 43 巻, pp. 597–600. 日本医療情報学会, 11 2023. J-GLOBAL ID: 202402213685606649.

[7] Zhengyuan Liu, Siti Umairah Md Salleh, Hong Choon Oh, Pavitra Krishnaswamy, and Nancy Chen. Joint dialogue topic segmentation and categorization: A case study on clinical spoken conversations. In Mingxuan Wang and Imed Zitouni, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track**, pp. 185–193, Singapore, December 2023. Association for Computational Linguistics.

[8] Aleksei Artemiev, Daniil Parinov, Alexey Grishanov, Ivan Borisov, Alexey Vasilev, Daniil Muravetskii, Aleksey Rezvykh, Aleksei Goncharov, and Andrey Savchenko. Leveraging summarization for unsupervised dialogue topic segmentation. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Findings of the Association for Computational Linguistics: NAACL 2024**, pp. 4697–4704, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[10] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging, 2015.

[11] HARVEY SACKS, EMANUEL A. SCHEGLOFF, and GAIL JEFFERSON. chapter 1 - a simplest systematics for the organization of turn taking for conversation**this chapter is a variant version of “a simplest systematics for the organization of turn-taking for conversation,” which was printed in language, 50, 4 (1974), pp. 696–735. an earlier version of this paper was presented at the conference on “sociology of language and theory of speech acts,” held at the centre for interdisciplinary research of the university of bielefeld, germany. we thank dr. anita pomerantz and mr. richard faumann for pointing out to us a number of errors in the text. In JIM SCHENKEIN, editor, **Studies in the Organization of Conversational Interaction**, pp. 7–55. Academic Press, 1978.

[12] Zuoquan Lin and Xinyi Shen. Local and global contexts for conversation, 2024.