所望の患者データを作る：Variational Auto-Encoder による症例報告生成

清水聖司，矢田竣太郎，荒牧英治 [1]
[1] 奈良先端科学技術大学院大学
{shimizu.seiji.so8,s-yada,aramaki}@is.naist.jp

概要

医療言語処理の分野では，患者のプライバシーの問題から共有可能な医療ドメインコーパスが少ない．そのため，コーパスを自動生成する研究がなされている．現状ではコーパス生成の手法として事前学習済みモデルを用いた，プロンプトを条件として生成する手法 (prompt-based の手法) が広く用いられているが，コントロール性の高い生成が困難である．Variational Auto-Encoder (VAE) を使った連続的条件からの生成では，prompt-based にはできないコントロール性の高い生成が実現できる可能性がある．本論文では VAE ベースの症例報告生成モデルを学習し，(1) 医学概念が潜在空間から再構成できるか，(2) 条件同士の距離関係と，生成される症例報告に含まれる医学概念の類似度の対応関係，の 2 つの観点から評価する．

1 はじめに

自然言語処理研究において，異なる手法の統一的な比較は，特定のタスクにおける手法の有効性の検証や，それに基づく手法の改善のために必須である．よって，そのような比較を可能にする共有可能なコーパスは，分野発展において重要である．しかし，医療言語処理の分野では，患者のプライバシーの問題から共有可能な医療ドメインコーパスが少ない [1, 2]．そのため，個人情報を削除する匿名化，またはコーパスの言語生成モデルを使った生成，といったアプローチで共有可能なコーパスを構築する試みがなされてきた．

生成を使ったアプローチにおいて，現状では事前学習済みモデルを用いたプロンプトを条件としてコーパスを生成する手法 (prompt-based の手法) が広く用いられている．しかし，この手法では，生成の条件が離散的であり，プロンプトのデザインが恣意的であるため，コントロールが困難である．

一方で，Variational Auto-Encoder (VAE) を使った連続的条件からの生成では，テキストを潜在空間に次元削減し，そこからテキストを再構成することによって，コーパス全体を一つの連続的なベクトル空間に埋め込むことができる．このような埋め込みが出来ると，Prompt-based の手法にはできないコントロール性の高い生成が実現できる可能性がある [3]．

そこで，本論文では VAE ベースの症例報告生成モデルを評価した．具体的には，(1) 入力された症例報告を潜在変数に変換し，変換された潜在変数から元の症例報告に含まれる医学概念再構成できるか，(2) 条件同士の距離関係と，生成される症例報告に含まれる医学概念の類似度が対応するかを評価した．評価の結果，医学概念の再構成に関しては疾病、傷害及び死因の統計を国際比較するために用いられる ICD コードの, 頭文字の粒度で可能であることが示された．また，潜在変数の距離関係が，生成された症例報告に含まれる医学概念の類似度と対応することが示された．

2 関連研究

共有可能な医療コーパス構築において, コーパスを公開することで，患者のプライバシーが危険にさらされる可能性がある．この問題に対し，主に匿名化または生成のアプローチが用いられる．匿名化はコーパスから個人情報を消す手法であり [4]，匿名化された有名なコーパスに MIMIC [5] がある．しかし，匿名化には，コーパスのサイズが大きくなるほど完全には個人情報を消しきれないという問題がある．

生成は近年発展を遂げたテキスト生成技術を用いて，人工的なコーパスを生成する手法である．この手法では，医療コーパスを用いて言語生成モデルを学習し，学習したモデルから自動的に医療テキスト

表 1 再構成された症例報告の例．上段が，Optimus に入力された症例報告，下段が潜在変数から生成された症例報告で，下線は入力された症例報告から再構成されている要素を示す．

入力	症例は０回経妊の２９歳で，下腹部痛および下腹部膨満感を主訴に受診し，精査加療目的で入院となった．内診所見では子宮は正常大で，子宮前面に有痛性で児頭大の軟らかい腫りゅうを触知した．骨盤ＭＲＩでは子宮前面に巨大嚢胞性腫りゅうを認めた．巨大卵巣嚢腫茎捻転の疑いにて緊急開腹手術を施行した．術中所見では，左卵管が嚢胞状に腫大し骨盤を占拠していた．捻転はなく，左卵管切除術を施行した．病理組織所見は，卵管留水腫であった．ＭＲＩを再読影すると，卵巣嚢腫とは明らかに異なる管状構造が認められた．
出力	症例は 2 回経妊0 産婦で，下腹部痛を主訴に近医を受診し，経ちつ超音波検査にて子宮留を指摘され，精査目的に入院となった．内診所見では，右卵巣嚢腫大，骨盤腔内に不整な腫りゅうを認めた．mri 検査では子宮内膜症を疑われ，両側付属器切除術を施行した．病理組織学的には成熟嚢胞性奇形腫と判明した．術後経過は順調であったが，妊娠 37 週 0 日で正常分娩となり，産婦人科に紹介された．画像では両側卵管間結合組織母斑を疑い，腫よう茎捻転と診断し緊急開腹手術を行った．摘出標本の病理診断は乳頭状腺腫であった．

を生成することでコーパスを構築する．

2.1 医療テキスト生成

現状の医療テキスト生成には主に，Transformer ベースの学習済みモデルにプロンプトを条件として生成する手法が用いられている．

GPT-3 [6] を用いて病名を条件として，その定義を生成した手法 [7] や，医療ドメインにおけるファクトイド質問応答のために T5 [8] を用いて疑問文を生成した手法 [9] などが存在する．Amin-Nejad らは，Prompt-based の手法を用いて電子カルテを自動生成することでコーパスを構築し，構築されたコーパスを下流の分類タスクにおいて評価した [10]．この研究において，下流タスクにおいて，生成されたコーパスのみを用いた時の精度は，実際のコーパスを用いた時の精度に及ばなかった．

2.2 連続的条件からののテキスト生成

望ましい医学概念の分布を持ったコーパスを構築するには，コントロール性の高いテキスト生成が必要である．例えば，肺がんのあらゆるケースを想定して，肺がん患者という医学概念を固定して，あらゆる年齢・性別などの個人特性を表す言語表現を持ったコーパスを構築できることが望ましい．

Prompt-based の手法を含む，学習済みモデルをそのまま Fine-tuning する手法は，汎用的な使用を目的とした事前学習済みモデルをそのまま用いるため，望ましい医学概念の分布を持ったコーパス構築するという使用目的に適さない可能性がある．特に，文書同士のグローバルな関係性を捉えることができず，条件に対する操作もプロンプトの入力変更に限られているという問題がある．

一方で，VAE [11] は，連続的条件からの生成が可能であり，テキスト同士のグローバルな関係もモデルの内的表現として得ることができる．Liu らの研究では，潜在空間においてエネジー関数を学習し，学習したエナジー関数をもとに，生成の条件となる潜在変数数をサンプルすることで，ある特定の特性を持ったテキストを生成可能なことが示された [12]．この手法では，様々な初期条件から連続的に潜在変数を変化させ特定の特性を持ったテキストを生成可能であり，ある医学概念を固定して，他の特性を連続的に変化させるといった操作が実現可能である．

3 材料

本研究では，医療テキストデータとして 1975 年から J-Stage に投稿された論文データに含まれる症例報告論文を使用した．ここから以下の条件で，対象となる症例報告論文を抽出した．

1. タイトルに “症例” を含む
2. アブストラクトが “症例は” から始まる
3. アブストラクトの文字数が 350 文字以下である

以下で使う症例報告は, 上記全ての条件を満たすデータ 11,181 件のアブストラクトとし，**JST-CR** と呼ぶ．

4 生成手法

4.1 Optimus

VAE ベースの症例報告生成モデルとして Optimus [13] を用いた．Optimus は VAE をベースとして，Encoder に BERT+Linear 層を用い，Decoder に Linear 層+GPT2 を用いたモデルである．

Optimus の学習では，Decode の時に，潜在変数の情報が無視されて生成される KL collapse を防ぐために KL annealing が用いられる．KL annealing では，学習において，ハイパーパラメータ β を周期的に変化させる．Optimus の学習の目的関数は以下の通り

である．

$$
\begin{aligned}
L_\beta &= L_E + \beta L_R \quad \text{with} \\
L_E &= \mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right] \\
L_R &= \mathrm{KL}\left(q_\theta(\mathbf{x}) || p(\mathbf{z})\right)
\end{aligned}
$$

L_E は入力と出力が同じになるようにする Reconstruction Loss であり，L_R は潜在変数の分布を事前分布と近くする KL Divergence Loss である．症例報告生成のために，Encoder の BERT には UTH-BERT [14]，Decoder の GPT-2 には日本語コーパスで事前学習した GPT-2 [15] を用いた．UTH-BERT は東大病院に蓄積された約 1 億 2 千万文の電子カルテ文書を使って学習された BERT モデルである．

4.2 生成の流れと例

JST-CR のアブストラクトを入力として，Encoder により入力を潜在変数に変換，そして潜在変数を再構成し，アブストラクトを生成した．学習した Optimus によって出力された文章の例を表 1 に示す．下線は入力された症例報告の要素がモデルによって再構成された部分を示す．この例では，元の症例報告の大部分を再構成できていることがわかる．

5 評価実験

潜在変数からの生成には，潜在空間において医学概念の関係性が表現されており，そこから症例報告の中に特定の医学概念を再構成できることが重要である．また，潜在空間での操作を生成に反映させるためには，ある潜在変数を変化させたときに，それに対応して，生成される医学概念が変化することが重要である．そこで，(1) 医学概念が潜在空間から再構成できるか (2) 条件同士の距離関係と，生成される症例報告に含まれる医学概念の類似度が対応するかを評価した．

5.1 評価方法

医学概念の再構成は Optimus に入力された症例報告と Optimus から出力された症例報告を比較することにより評価した．潜在空間における条件同士の距離と，生成される症例報告に含まれる医療概念の類似度が対応するかは，ある潜在変数に異なるノルムの摂動を加え，そこから生成される症例報告と元の潜在変数から生成される症例報告とを比較することにより評価した．

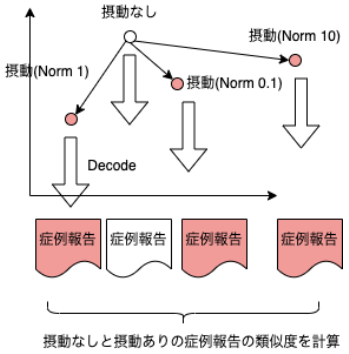

図 1 **EvalCorr** での潜在空間における操作．摂動なしは，摂動を加える前の潜在変数を表し，赤い点は様々な大きさのノルムの摂動が加えられた潜在変数を表す．それぞれの潜在変数から症例報告を生成し，評価した．

5.1.1 EvalRecon : 医学概念再構成の評価

医学概念が潜在空間から再構成できていれば，入力された症例報告と出力された症例報告は，その中に含まれる医学概念が類似するはずである．そこで，Optimus に入力された症例報告と Optimus から出力された症例報告の類似度を測ることで，医学概念の再構成の評価とした．以下ではこの評価を **EvalRecon** とする．

5.1.2 EvalCorr : 条件の距離と類似性の対応評価

潜在空間での操作が生成される症例報告に反映されるかを，条件同士の距離関係と，症例報告に含まれる医学概念の類似度が対応するかを基に評価した．具体的には，まず，図 1 のように，評価データ中の症例報告を潜在変数に Encode して，その潜在変数に対して，0.01 から 10 のノルムを持つランダムな摂動を加え, 元の潜在変数から生成された症例報告と摂動を加えられた潜在変数から生成された症例報告の類似度を測った．ノルムが大きくなるにつれ，類似度がそれに対応して下がるかどうかを評価した．以下ではこの評価を **EvalCorr** とする．

5.2 評価指標

2 文書の症例報告の類似度を定量的に評価する．

そのために，参照文と生成文の n-gram ベースの一致度を測る指標である BLEU と，文中に含まれる医学概念の一致度を測る指標である NER 一致度を用いた．NER 一致度の計算方法は以下の通りである．

1. 2 文書の症例報告に対して，NE 抽出器の RealMedNLP_CR_JA [16] を使い NER を施す．
2. NER 結果から病名と臓器・部位を抽出する．

表 2　**EvalRecon** の評価結果（NER 一致度）．Optimus に入力された症例報告と，Optimus から出力された症例報告に NER を施し，一致の Precision と Recall を計算した．

	Precision	Recall
NER 臓器・部位	0.027	0.029
NER 病名	0.132	0.137
NER　**ICD**	0.543	0.520

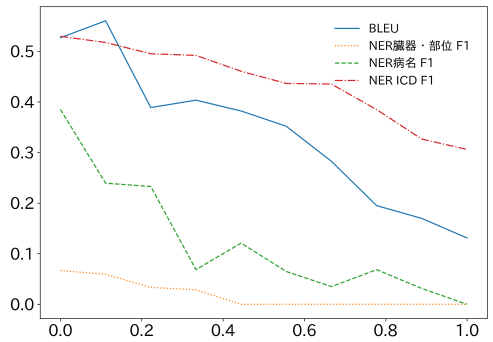

図 2　**EvalCorr** の結果．縦軸は評価指標のスコア (F1)，横軸は摂動のノルムに対し，常用対数を取ったものを表す．

3. 2 つの症例報告から抽出された病名と臓器・部位を比較し，Recall と Precision を計算する．

病名の NER 一致度に関しては，NER の結果得られた病名を DNorm-J により正規化したものと，病名を ICD10 コードに変換し，その頭文字をとったもの (**ICD**) を使って NER 一致度を計算した．

6　結果と考察

実験では，**JST-CR** を 8:1:1 に分割し，それぞれを学習・検証・評価データとして用いた．学習したモデルを 5 章で述べた方法と指標で，**JST-CR** の評価データを使って評価した．**EvalRecon** の結果から，**ICD** の粒度では，医学概念の再構成ができていることが示唆される．**EvalCorr** の結果，ノルムの大きさに対応して，生成された症例報告に含まれる医学概念の類似度が下がっていることから，条件同士の距離関係と，生成される症例報告に含まれる医学概念の類似度が対応していることが示唆される．

6.1　EvalRecon

入力された症例報告と出力された症例報告に対して BLEU スコアと NER 一致度を計算した．評価データ全体の平均の BLEU スコアは 0.315 であった．NER 一致度は表 2 の通りである．**ICD** の NER 一致度は，Precision が 0.543，Recall が 0.520 であった．

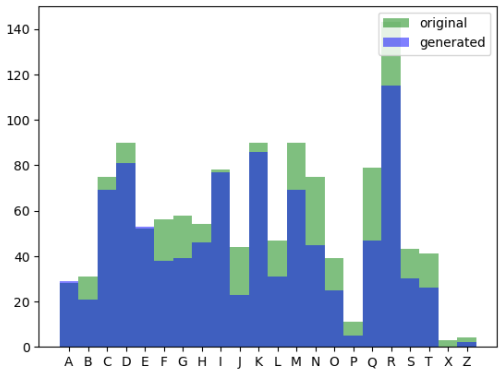

図 3　**JST-CR** の評価データにおける ICD コードの分布 (original) と，Optimus により生成されたコーパス (generated) における ICD コードの分布．横軸は ICD コードの頭文字，縦軸は頻度を表す．

6.2　EvalCorr

各症例報告に対して 10 回ランダムな方向の摂動を施し，スコアを平均したものを一つの症例報告に対するスコアとした．ノルムと各スコアの相関係数は，BLEU が -0.970，NER 臓器・部位が -0.875，NER 病名が -0.885，**ICD** が -0.969 であった．評価データ全体のスコアを摂動のノルムごとに平均した結果を図 2 に示す．図 2 から臓器・部位以外はスコアがおおよそノルムと対応し，下がっていることがわかる．

6.3　ICD コード分布の比較

生成されたコーパスが，**JST-CR** 評価データの ICD コード分布をどの程度再現できるかを評価した．**JST-CR** の評価データに含まれる ICD コードの頭文字の分布と，生成された症例報告における ICD コードの頭文字の分布を図 3 に示す．2 つの分布の KL Divergence は 0.020，分布内の順位相関係数は 0.941（p 値 ≤0.01) であった．この結果から，生成されたコーパスは，ICD コードの頭文字の分布を再現できていることが示唆される．

7　おわりに

本研究では，VAE ベースのモデルである Optimus を使い，症例報告生成を再構成して生成した．実験の結果，**ICD** の粒度で医学概念の再構成が可能であることと，症例報告に含まれる医学概念を，潜在空間における条件同士の距離に対応させて変化させることが可能であることを示した．

謝辞

本研究は，JST AIP 日独仏 AI 研究 JPMJCR20G9，JST CREST JPMJCR22N1，国立情報学研究所 (NII) CRIS の支援を受けたものである．

参考文献

[1] Eiji Aramaki, Shoko Wakamiya, Shuntaro Yada, and Yuta Nakamura. Natural Language Processing: from Bedside to Everywhere. **Yearbook of Medical Informatics**, June 2022.

[2] Udo Hahn and Michel Oleynik. Medical Information Extraction in the Age of Deep Learning. **Yearbook of Medical Informatics**, Vol. 29, No. 1, pp. 208–220, August 2020.

[3] Dayiheng Liu, Jie Fu, Yidan Zhang, Chris Pal, and Jiancheng Lv. Revision in continuous space: Unsupervised text style transfer without adversarial learning. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, pp. 8376–8383, 2020.

[4] Tawanda Sibanda and Ozlem Uzuner. Role of local context in automatic deidentification of ungrammatical, fragmented text. In **Proceedings of the Human Language Technology Conference of the NAACL, Main Conference**, pp. 65–73, June 2006.

[5] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. **Scientific data**, Vol. 3, No. 160035, pp. 1–9, 2016.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.

[7] Bosung Kim and Ndapa Nakashole. Data augmentation for rare symptoms in vaccine side-effect detection. In **Proceedings of the 21st Workshop on Biomedical Language Processing**, pp. 310–315, May 2022.

[8] Hermann Bujard, Reiner Gentz, Michael Lanzer, Dietrich Stueber, Michael Mueller, Ibrahim Ibrahimi, Marie-Therese Haeuptle, and Bernhard Dobberstein. [26] a t5 promoter-based transcription-translation system for the analysis of proteins in vitro and in vivo. In **Methods in enzymology**, Vol. 155, pp. 416–433. Elsevier, 1987.

[9] Dimitris Pappas, Prodromos Malakasiotis, and Ion Androutsopoulos. Data Augmentation for Biomedical Factoid Question Answering. **arXiv preprint**, Vol. arXiv:2204.04711, , 2022.

[10] Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. Exploring transformer text generation for medical dataset augmentation. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 4699–4708, 2020.

[11] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. **arXiv preprint arXiv:1312.6114**, 2013.

[12] Guangyi Liu, Zeyu Feng, Yuan Gao, Zichao Yang, Xiaodan Liang, Junwei Bao, Xiaodong He, Shuguang Cui, Zhen Li, and Zhiting Hu. Composable text control operations in latent space with ordinary differential equations. **arXiv preprint arXiv:2208.00638**, 2022.

[13] Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. Optimus: Organizing sentences via pre-trained modeling of a latent space. **arXiv preprint arXiv:2004.04092**, 2020.

[14] Yoshimasa Kawazoe, Daisaku Shibata, Emiko Shinohara, Eiji Aramaki, and Kazuhiko Ohe. A clinical specific bert developed using a huge japanese clinical text corpus. **Plos one**, Vol. 16, No. 11, p. e0259763, 2021.

[15] 趙天雨, 沢田慶. 日本語自然言語処理における事前学習モデルの公開. 人工知能学会研究会資料 言語・音声理解と対話処理研究会, Vol. 93, pp. 169–170, 2021.

[16] Tomohiro Nishiyama, Mihiro Nishidani, Aki Ando, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. Naist-soc at the ntcir-16 real-mednlp task.

A　参考情報

A.1　潜在空間の可視化

Optimus によって構築された潜在空間を可視化したものを図 4 に示す．タイトルに含まれる病名の ICD コードの頭文字（ICD の軸) ごとに色分けした．

同じ ICD コードをもつ症例報告同士で固まっているのは C，K，H で，それぞれ腫瘍，消化器系，眼科・耳鼻科系の病気を表すコードである．このことから，これら 3 つに関する症例報告は，ある程度一貫した言語的特徴を持っていると推測される．

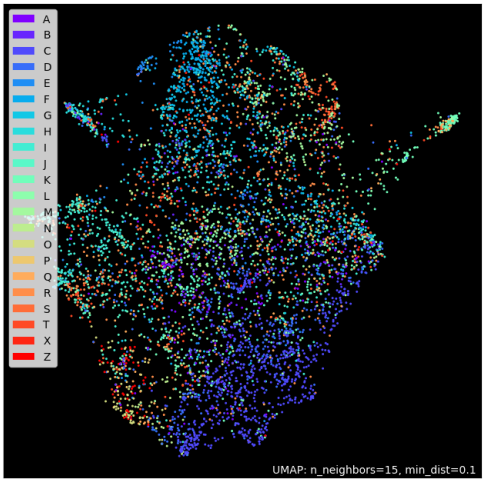

図 4　潜在空間の可視化．UMAP を用いて潜在変数を 2 次元に表現した．アルファベットは ICD コードの頭文字を表す．