

文脈化埋め込み表現を用いた対照学習による病名正規化

氏家 翔吾 磯 颯 荒牧 英治

奈良先端科学技術大学院大学

{ujiie, iso.hayate.id3, aramaki}@is.naist.jp

1 はじめに

病名の表記揺れ解消（本稿では病名正規化と呼ぶ）は、医学論文や診療録などのテキスト（医療文書）の解析における重要な要素技術の一つである。例えば、医療文書からの情報抽出 [1] や関係抽出 [2] では、病名を有用な特徴量として扱っているが、表記揺れにより同じ概念を同一視できない問題がある。この場合、病名を正規化することはこれらのタスクに必須の処理となる。

病名正規化の主なアプローチとして、病名の埋め込み表現を学習し、辞書中の最近傍のエントリに対応づける手法が挙げられる。これは、大量の学習データを用いた教師あり学習 [3, 4, 5] や、大規模な類義語辞書を用いた自己教師あり学習 [6] により実現されるが、一般に前提となる大規模なリソースを構築することが困難である場合が多い。例えば日本語では、人手でメンテナンスされている大規模な類義語辞書は存在せず、病名正規化のデータセットも公開されていない。そのため、十分に大きな辞書に依存しない手法の開発が必要である。

そこで本研究は、低リソース上でも有用な病名正規化手法の開発を目的とする。具体的には、人手によりアノテーションされた学習データがなく、類義語辞書が小規模な状況を想定する。これは、医学概念の小規模な辞書が存在する一方で、専門知識を必要とするためにアノテーションコストの高い医療分野の現状と矛盾しない自然な設定である。

本研究では、大規模な類義語辞書や大量の学習データを用いる代わりに、医療文書中に現れる病名の文脈を手がかりに病名を正規化する手法を提案する。例えば、「TdP」と「TDP-43 proteinopathy」は似た語を含む一方で、「TdP」は不整脈の一種、「TDP-43 proteinopathy」はタンパク質異常による疾患と、両者は全く異なる疾患であり、医療文書中で現れる文脈は大きく異なる。図 1 に示す提案手法では、このような文脈の違いを正規化の手掛かりとして利用す

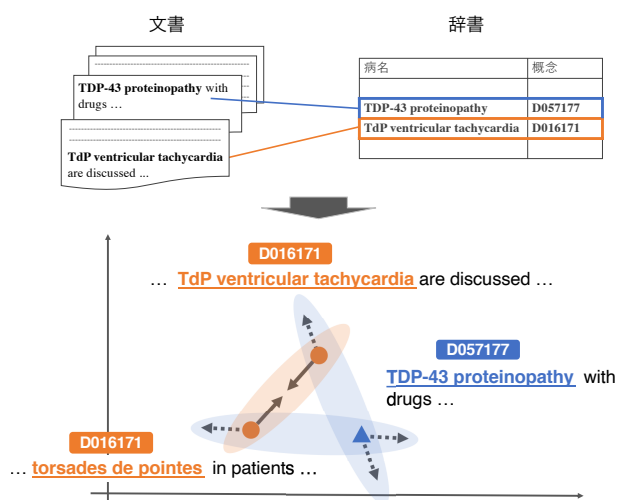


図 1 病名の文脈化埋め込み表現. 埋め込み表現は辞書マッチにより病名が自動認識された文から得られる。これらは、同じ概念を持つ病名を近く（実線）、異なる概念を持つ病名を遠く（破線）埋め込むよう、対照学習により最適化される。

るため、小規模辞書を用いて医療文書中の病名を自動認識し、その病名の文脈化埋め込み表現を用いて近傍探索を行う。この際、同じ概念を持つ病名が近く埋め込まれるように、辞書の情報をもとにモデルを対照学習によって最適化する。実験では、病名正規化における標準的なデータセットを用いて、提案手法の有効性を示す。本研究の貢献は以下の通りである。

- 生テキストと小規模な外部リソースを用いた、病名正規化モデルの学習の枠組みを導入した。
- 実験の結果、外部リソースの利用が制限されている状況でも、提案手法は高い精度を示すことを確認した。

2 関連研究

病名正規化を扱う研究では、医療分野に最適化された埋め込み表現を用いて病名間の類似度を学習する手法が多く提案されている [3, 7]。特に近年、双方向言語モデル BERT [8] を、大規模な医療文書を用い

て事前学習した BioBERT [9] や, PubMedBERT [10] が, 病名正規化タスクにおいても高い有用性を示している. 例えば Sung らは, BioBERT を用いて得られた病名の埋め込み表現に対し, 類義語間の類似度が高くなるように Fine-tuning することで, 既存手法よりも高い精度で病名正規化ができることを示した [5]. しかし, この手法はアノテーションされた病名が一定量必要であり, 学習データの作成コストが高い.

一方で, 本研究と同様, 明示的な学習データを用いることなく, 病名正規化モデルを構築する手法も提案されている. 例えば Liu らは, 大規模な類義語辞書を用いて自己教師あり学習を行うことで, 人手によりアノテーションされたデータを必要としない手法を開発した [6]. 同研究では, 医学概念における大規模なメタソーラスである UMLS [11] を用いて, 類義語を学習データとみなすことで病名表現を学習し, 高精度に病名正規化が行えることを示した. この手法では人手でアノテーションされた学習データを必要としない一方で, 大規模な類義語辞書の存在が前提となっている. 実際, 類義語辞書の利用を制限した場合に, 同病名正規化モデルの精度が大幅に劣化することを確認した. 本研究では, 病名のみを入力とする同研究と異なり, 学習・推論時に病名が自動認識された文から得られた文脈化埋め込み表現を用いることで, 類義語が少量の場合にも文脈を手掛かりに正規化できる.

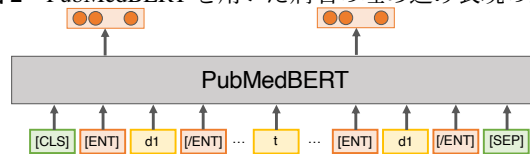
3 手法

病名正規化タスクは, 入力文 X における i 番目の病名 x_i について, それに対応する概念 $y_i \in \mathbb{C}$ を予測するタスクとして表される. ここで, \mathbb{C} は類義語辞書収載の概念集合である.

提案手法では, 病名が自動認識された文集合 (用例集合) から得た文脈化埋め込み表現を用いて, 近傍探索による病名正規化を行う. 文脈化埋め込み表現は, 同じ概念を持つ病名が近く埋め込まれるように, 対照学習により最適化する.

以降, 本研究では, 類義語の最長一致により病名を自動認識した用例集合 $\mathbb{S} = \{X'_0, X'_1, \dots, X'_M\}$, 及び \mathbb{S} 内の各文の病名, 概念系列 $x^{(m)}, y^{(m)}$ から構成されるデータベース $\mathbb{D} = \{(x_j^{(m)}, y_j^{(m)})\}_{m=1, j=1}^{M, |x^{(m)}|}$ にアクセス可能であると仮定する. $x_j^{(m)}, y_j^{(m)}$ はそれぞれ, $x^{(m)}, y^{(m)}$ における j 番目の病名と概念を表す.

図2 PubMedBERT を用いた病名の埋め込み表現の獲得



3.1 正規化モデル

病名 x_i を正規化するため, K 近傍法を用いる. 本研究では, 近傍探索時の距離として, 文脈化埋め込み表現のコサイン類似度を用いる.

x_i に対する各概念の予測確率を, x_i の近傍 K サンプル $\mathbb{N}_K(x_i, \mathbb{D})$ を用いて以下のように定義する.

$$p(y_i = c | x_i, \mathbb{D}, K) = \frac{1}{K} \sum_{j \in \mathbb{N}_K(x_i, \mathbb{D})} \mathbb{1}(y'_j = c) \quad (1)$$

ここで, $\mathbb{1}(e)$ は指示関数である. この予測確率を用いて, y_i は $\arg \max_{c \in \mathbb{C}} p(y_i = c | x_i, \mathbb{D}, K)$ で正規化される.

文脈化埋め込み表現を得るため, 双方向言語モデルの一つである BERT [8], 特に医療文書を用いて最適化された PubMedBERT [10] を用いる. 図2に示す通り, まず, 文中の対象とする病名を特殊トークンで “[ENT]”, “[/ENT]” で囲み, PubMedBERT の入力とする. その後, “[ENT]” トークンに対応する埋め込み表現を, 対象とする病名の埋め込み表現として用いる [12].

3.2 対照学習

対照学習とは, あるデータに対して, 同一ラベルが付与されたデータ (正例) を近く, 異なるラベルが付与されたデータ (負例) を遠く埋め込むように学習する距離学習手法である [13, 14]. これにより, 同じ概念の病名同士は近く, 異なる概念の病名は遠く埋め込まれる.

対照学習には, 正例, 負例の病名のペアが必要である. 本研究では, それらを用例のミニバッチ内で動的に構成する. つまり, ミニバッチ内の用例における全ての病名 $x_j^{(m)}$ について, 正例として $y_j^{(m)}$ と同じ概念の病名を, 負例として $y_j^{(m)}$ と異なる概念の病名をミニバッチ内から抽出し, それらを正例・負例ペアとする.

損失は Multi-Similarity loss [15] を用いる. この損失関数は正例, 負例ペアの相対的な類似度を考慮した距離学習手法である. これは, ミニバッチ内のすべての病名 \mathbb{B} において, 病名 $x_i, x_j \in \mathbb{B}$ 間のコサイン

類似度 $S_{i,j}$ を要素にもつ類似度行列 $\mathbf{S} = \in \mathbb{R}^{|\mathbb{B}| \times |\mathbb{B}|}$ と、 x_i の正例、負例集合 $\mathbb{P}_i, \mathbb{N}_i$ を用いて、以下の損失関数で表される。

$$\mathcal{L}_{MS} = \frac{1}{|\mathbb{B}|} \sum_{i=1}^{|\mathbb{B}|} \left\{ \frac{1}{\alpha} \log \left[1 + \sum_{k \in \mathbb{P}_i} e^{-\alpha(S_{ik}-\lambda)} \right] + \frac{1}{\beta} \log \left[1 + \sum_{k \in \mathbb{N}_i} e^{\beta(S_{ik}-\lambda)} \right] \right\}. \quad (2)$$

ここで、 α, β は温度パラメータ、 λ は類似度のオフセットを表す。学習には、正例のうち、最も類似度の小さい負例より類似していないもの、負例のうち、最も類似度の大きい正例より類似しているもののみを用いる [15]。

4 実験

4.1 データ

リソース 病名の類義語辞書として、MEDIC [16] を用いる。MEDIC は、13063 個の疾患に対して平均して 6 個の病名表現が類義語として収録されている。本研究では、より低リソースとなるよう、各概念に対する類義語を半数になるように無作為に抽出し類義語辞書とする。つまり、今回の実験では、1 つの概念に対して平均して 3 個の類義語が収録されている。

学習・推論に用いる用例は、医学論文の検索システムである PubMed に掲載された論文のタイトル及びアブストラクト (約 27GB) から作成する。なお、評価データに含まれる論文は除外した。

評価用データセット 病名正規化で一般的に用いられる 3 つのデータセット、NCBI disease corpus (NCBID) [17], BioCreative V Chemical DiseaseRelation (BC5CDR) [18], MedMentions [19] を用いて評価を行う。各データセットの詳細は付録 A.1 節に記載する。

4.2 実験設定

モデルの詳細 推論時の kNN の近傍数は、検証用データセットにおいて最も精度の高かった 15 とした。その他のハイパーパラメータは付録 A.2 節に述べる。

ミニバッチの作成 ミニバッチの作成にあたり、用例集合から無作為に用例を抽出した場合、多くの病名はミニバッチ内に正例を持たない恐れがある。そのため、ミニバッチ作成時に、一定数の概念を無

作為に選択し、各概念から 2 文を無作為に抽出した。これにより、各用例内の少なくとも 1 つの病名がミニバッチ内に正例を持つことが保証される。

評価方法 先行研究 [5, 6] に従い、Accuracy で評価する。各データセットには、MEDIC に含まれない医療概念もアノテーションされている。本研究は、類義語辞書への紐付けを目的としており、類義語辞書外の概念は対象外であるため、それらは評価から除いた。

比較手法 提案手法の比較のため、以下のモデルを比較手法として用いた。いずれも病名のみを入力としている。

- TF-IDF: 文字単位 1-, 2-gram の TF-IDF のコサイン類似度による正規化。Sung らの研究において、病名正規化に対する有用性が示されている [5]。
- PubMedBERT [10]: 同モデルを用いて得られた埋め込み表現を用いて、Fine-tuning なしに近傍探索を行う手法。
- SapBERT [6]: 類義語辞書を用いた自己教師あり学習手法。Liu らの実験とは異なり、類義語辞書として、UMLS 全体でなく提案手法と同様の辞書を用いた。

4.3 結果

表 1 各データセットにおける Accuracy。病名のリンク先が異なるため、先行研究 [5, 6] とは異なる。

手法	NCBID	BC5CDR	MedMentions
TF-IDF	0.4533	0.5993	0.5783
PubMedBERT	0.3455	0.4216	0.4522
SapBERT	0.5647	0.6523	0.6511
提案手法	0.6031	0.6896	0.7141

表 1 に実験結果を示す。提案手法はいずれのデータセットにおいても Accuracy が最も高くなっている。これは、病名が自動認識されたデータに対する対照学習により得られた病名の文脈化埋め込み表現が、病名正規化が有用であることを示している。

5 考察

5.1 辞書サイズと正規化精度

類義語辞書の規模と病名正規化の精度の関係を分析するため、以下の 4 つの辞書サイズにおいて実験を行った。カッコ内は類義語辞書に含まれる類義語

表 2 MedMentions におけるモデルの結果. 各例は, [予測概念] 最近傍データの形式となっている. 対象となる病名を太字, 手掛かりとなる語をイタリックで示す. いずれの例も類義語辞書に存在しない病名が入力となっている.

Label: D014178	Input: ... phenotype B-I for <i>B-ALL</i> , phenotype T-I-II-IV for T-ALL; LDH ... the presence of translocation t(4;11)
SapBERT	[D054868] partial 11q monosomy syndrome
Proposed	[D014178] ... involved in genetic translocations characteristic of b-cell acute lymphoblastic leukemia (<i>B-cell ALL</i>).
Label: D000077273	Input: ... histopathology consisted of papillary thyroid carcinoma (PTC) (n = 91, 86.7%).
SapBERT	[C536943] TCC
Proposed	[D000077273] ... for detecting cervical lymph node (LN) metastasis in papillary thyroid carcinoma (PTC) .
Label: D002292	Input: High occurrence of non-clear cell renal cell carcinoma
SapBERT	[D002292] renal cell carcinomas
Proposed	[D002289] ... used to treat renal cell carcinoma, non-small-cell lung cancer and colon cancer ...

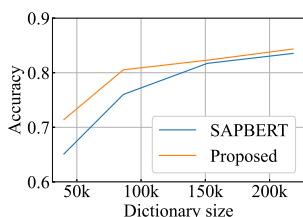


図 3 辞書サイズと Accuracy

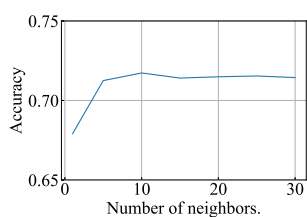


図 4 kNN の近傍数と Accuracy

数である.

- MEDIC の 50%を無作為に抽出 (39,959)
- MEDIC をそのまま使用 (86,205)
- UMLS の関連病名の 50%を MEDIC に追加 (151,559)
- UMLS の関 x 連病名を MEDIC に追加 (218,325)

図3は MedMentions における実験結果である. 提案手法は SapBERT と比較し, 辞書サイズが小さいほど有効であることがわかる. これは, 辞書サイズが小さく病名表現が限られた場合に, 文脈でその情報を補うことができるためと考えられる.

5.2 近傍数と正規化精度

kNN の近傍数と正規化精度を比較するため, k が 1, 5, 10, 15, 20, 25, 30 の場合で, MedMentions において実験を行った. 図 4にその結果を示す. k が 10 の時に Accuracy が最大となるが, その他の値の場合にも精度は飽和しており, 入力病名に対応する概念が正しく近傍として抽出されていると考えられる.

5.3 事例の分析

文脈を考慮する有効性を質的に議論するため, MedMentions におけるモデルの予測結果を分析した, 表 2に, 文脈を考慮することにより予測が変化した例を示す. なお, 例示した全ての入力病名は類義語辞書に収載されていない.

一つ目と二つ目の例は, 文脈が予測に有効に働い

た例である. 一つ目の例では, 急性リンパ性白血病を表す “translocation t(4;11)”が入力である. 提案手法では, 急性リンパ性白血病の一種である “B-ALL”を手がかりに, “B-cell ALL”の含まれる用例を最近傍として抽出できていることがわかる.

二つ目の例は同様に, 正式名称 “papillary thyroid carcinoma”と共に “PTC”を手掛かりにして提案手法が奏功した例である.

三つ目の例は, SapBERT が正解し, 提案手法が予測を誤った例である. 提案手法の予測文は, “carcinoma, non-small-cell lung”が誤って病名として認識されており, 正しくは “renal cell carcinoma”と認識されるべきである. この場合, 文脈を考慮することにより, “carcinoma, non-small-cell lung”が病名として認識されている場合にも “renal cell”に注目してしまい, “renal cell carcinoma”の近傍として誤って予測されている. このように, 辞書マッチによる認識誤りが致命的な予測誤りに繋がり得る. 今回の実験では, 最長一致により病名を認識したが, リソースとして固有表現抽出器を使用可能な場合, この問題は緩和可能である.

6 おわりに

本研究では, 文脈の一致による病名正規化手法を提案した. 標準的な病名正規化データセットにおける実験により, 文脈を考慮することで正規化精度が向上することが示された. 文脈の一致による正規化は, 低リソース下において精度の向上が顕著であり, 病名の多様性を文脈情報で補えることが示唆された. 本研究は, 医療ドメインで自然な問題設定をとっており, 今後, さまざまな応用システムにおいて本手法を活用できると考えている

参考文献

- [1]Sunwon Lee, Donghyeon Kim, Kyubum Lee, Jaehoon Choi, Seongssoon Kim, Minji Jeon, Sangrak Lim, Donghee Choi,

- Sunkyu Kim, Aik-Choon Tan, and Jaewoo Kang. BEST: Next-Generation biomedical entity search tool for knowledge discovery from biomedical literature. *PLoS One*, Vol. 11, No. 10, p. e0164680, October 2016.
- [2]Jun Xu, Yonghui Wu, Yaoyun Zhang, Jingqi Wang, Hee-Jin Lee, and Hua Xu. CD-REST: a system for extracting chemical-induced disease relation in literature. *Database*, Vol. 2016, , March 2016.
- [3]Ishani Mondal, Sukannya Purkayastha, Sudeshna Sarkar, Pawan Goyal, Jitesh Pillai, Amitava Bhattacharyya, and Mahanandeeswar Gattu. Medical entity linking using triplet network. In *Proc. of the 2nd Clinical Natural Language Processing Workshop*, pp. 95–100, June 2019.
- [4]Jennifer D’Souza and Vincent Ng. Sieve-Based entity linking for the biomedical domain. In *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 297–302, July 2015.
- [5]Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. Biomedical entity representations with synonym marginalization. In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3641–3650, 2020.
- [6]Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pre-training for biomedical entity representations. October 2020.
- [7]Haodi Li, Qingcai Chen, Buzhou Tang, Xiaolong Wang, Hua Xu, Baohua Wang, and Dong Huang. CNN-based ranking for biomedical entity normalization. *BMC Bioinformatics*, Vol. 18, No. Suppl 11, p. 385, October 2017.
- [8]Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. 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, June 2019.
- [9]Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, Vol. 36, No. 4, pp. 1234–1240, February 2020.
- [10]Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pre-training for biomedical natural language processing, 2020.
- [11]Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, Vol. 32, No. Database issue, pp. D267–70, January 2004.
- [12]Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, July 2019.
- [13]Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.
- [14]Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proc. of the 37th International Conference on Machine Learning*, Vol. 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, Virtual, 13–18 Jul 2020.
- [15]Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [16]Allan Peter Davis, Thomas C Wieggers, Michael C Rosenstein, and Carolyn J Mattingly. MEDIC: a practical disease vocabulary used at the comparative toxicogenomics database. *Database*, Vol. 2012, p. bar065, March 2012.
- [17]Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. NCBI disease corpus: a resource for disease name recognition and concept normalization. *J. Biomed. Inform.*, Vol. 47, pp. 1–10, February 2014.
- [18]Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wieggers, and Zhiyong Lu. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, Vol. 2016, , May 2016.
- [19]Sunil Mohan and Donghui Li. MedMentions: A large biomedical corpus annotated with } concepts. In *Proc of the AKBC*, 2019.

A 付録

A.1 データセットの詳細

本研究において使用した3つのデータセットでは、医学論文のタイトルとアブストラクトに対して医学概念のスパンと概念識別子がアノテーションされている。表3に、各データセットの文書数、病名数を示す。

dataset	文書数	病名数
NCBID	100	960
BC5CDR	500	4,424
MedMentions	879	3,795

表3 各データセットにおける文書数と病名数

A.2 ハイパーパラメータ

学習におけるハイパーパラメータとして、バッチ数は16 (32文)、温度パラメータ α, β は2, 50, オフセット λ は1に設定した。推論時は各概念について用例集合から無作為に抽出した25文の用例を用いた。

A.3 エンティティタグの付与方法

実際の運用では、用例は辞書内の病名が認識される一方で、入力文は固有表現抽出結果のすべてを病名として認識するため、病名の認識粒度の不一致が起こる。つまり、推論時において過剰に病名が認識され、学習時に認識されていない病名にもENTタグを付与してしまう問題がある。これは、対象となる病名のみならずENTタグを付与することで緩和可能である。この問題のAccuracyへの影響を分析するため、用例・テストデータ共に、対象とする病名のみならずENTタグを付与し、学習・評価を行った。

実験結果を表5に示す。対象とする病名のみならずENTタグを付与した場合は、すべてにENTタグを付与した場合と同等、もしくは精度が悪化した。この原因の一つとして、学習時のミニバッチにおける病名数がある。対象となる病名のみならずENTタグを付与した場合、すべての病名にENTタグを付与する場合に比べて、負例の数が少なくなる。これにより、学習が効率的に進まなかった可能性がある。

手法	NCBID	BC5CDR	MedMentions
提案手法 (all)	0.6031	0.6896	0.7141
提案手法 (one)	0.5927	0.6692	0.7067

表4 実験結果.

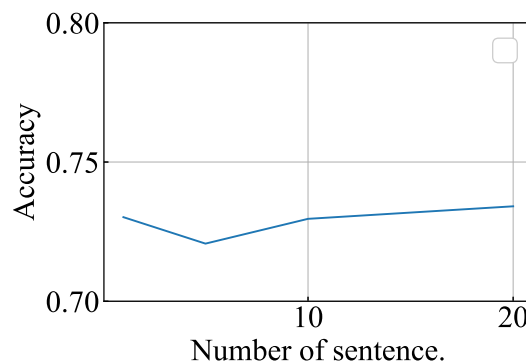


図5 用例数と正規化精度

A.4 用例数と正規化精度

用例集合の作成に用いる生テキストの数が膨大な場合、全探索に大きな計算コストが生じる。これは、病名ごとに用例を無作為に抽出することで緩和可能である。用例数のAccuracyへの影響を分析するため、用例数を1, 5, 10, 20としてMedMentionsにおいて実験を行った。図5に、各用例数に対するAccuracyを示す。Accuracyは用例数20の場合に最大になるが、用例数が少ない場合にも大きな減少はない。このことから、対照学習により汎用的な病名の埋め込み表現を獲得しており、無作為な抽出により用例数が減少しても精度が保たれることが示唆された。