二段階学習と概念クラスを用いた医療固有表現の正規化

茂里 憲之* 辻村 有輝 * 三輪 誠 佐々木 裕

豊田工業大学
{sd16087, sd18602, makoto-miwa, yutaka.sasaki}@toyota-ti.ac.jp

1 はじめに

医薬学の分野では，論文やカルテなどの文書における薬物や疾患のような概念とデータベースの情報を結びつけることが，研究者や医師が必要なデータを容易に参照できるようにし，効率的な作業を行うために重要である．しかしながら，文書は大量に存在し，かつ，急速に増加しているため，必要な文献を探すことが難しく，情報を処理しきれていないのが現状である．例として，生命科学の文献情報を収集したオンラインデータベース MEDLINE にある論文は約 1 [件/分] のペースで増え続けている．またカルテのような文書は，自動で処理することが難しい．そのため人手によって作成された文書を自動的に処理し，データベースの情報との関連付けを行うシステムの実現が重要である．

このような情報の関連付けに向けて，医療文献における自然言語処理を目的とした国際的な評価型ワークショップである National NLP Clinical Challenges（n2c2）2019 Shared Task Track 3 において，カルテを対象とした固有表現正規化タスクが開催された．このタスクの目的は，カルテのデータセット MCN[1] について，文献内の疾患や薬物等の概念を示す文字列の位置とそれが示す概念を医学系データベースシステム Unified Medical Language System (UMLS) [2] に登録された医学系の概念の識別子 Concept Unique Identifier (CUI)，あるいは該当する CUI が存在しないもの (CUI-less) が注釈として与えられた上で，新たに与えられた文書中の文字列を概念の識別子もしくは CUI-less に写像する高性能なシステムを構築することである．タスクの概観を図 1に示す．MCN はそれぞれ 50 件のカルテの学習データとテストデータからなる．データセットの内訳を表に示す．Track 3 では対象データベースに 434,056 種類の CUI が存在するため 43 万クラスへの分類問題を解く必要があるが，訓練データに現れる医学概念は 6,684 件，ユニークな CUI 数は 2,331 種類と活用できるデータが極めて少ない．

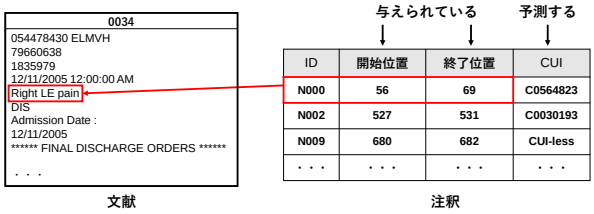

図1 n2c2 2019 Track3 の概観

本論文では，与えられた固有表現について，SciBERT[3] によりベクトル表現を獲得し，cosine 類似度により正規化を行う深層固有表現正規化モデルとその二段階学習を用いた学習を提案する．辻村が主導し考案した提案手法を用いて参加した TTI-COIN は Track 3 において，表 2に示す通り，最高の正解率を達成した．さらに，文脈を考慮できていないという現状の TTI-COIN のシステムの問題点の改善のため，文脈を考慮した固有表現分類による後処理による精度向上を試みた．

表1 MCN の内訳

データ	文献数	固有表現数	ユニークな CUI 数
訓練	50	6,684	2,331
テスト	50	6,925	2,579

表2 n2c2 2019 Track 3 の上位スコア

チーム	正解率
TTI-COIN	0.8526
Kaiser Permanente (KP)	0.8194
University of Arizona (UAZ)	0.8166
Alibaba (Ali)	0.8105
Med Data Quest, Inc (MDQ)	0.8101

* これらの著者は等しく貢献した．

2 関連研究

2.1 固有表現認識

固有表現認識は文中の固有表現の位置を予測するタスクである．固有表現とは地名や人名・団体名などのいわゆる固有名詞的表現であるが，時刻や数量などを認識の対象とすることもある．文献中の固有表現を正しく認識することは情報抽出や情報検索などの自然言語処理システムの実現において重要である．近年では BERT を用いた事前学習 [4] によるニューラル BIO タグ付けによる高い抽出性能が確認され，BERT を用いた固有表現認識のための文脈情報の埋め込み手法が Akbik ら [5] により提案されている．

2.2 固有表現正規化

固有表現正規化は，文献中に現れる固有表現が示す対象を同定し，知識ベースとの関連付けを行うタスクである．近年では単語のベクトル表現による深層学習の適用により精度向上が試みられている．医学薬学分野では，Li ら [6] により，固有表現と候補の組を畳み込みニューラルネットワークの入力とし，各組のスコアを計算し，ランク付けして，予測を行う方法が報告されている．

3 提案手法

3.1 二段階学習による医療概念正規化

3.1.1 モデルの構造

与えられた文書中の固有表現に対し，SciBERT[3] による埋め込みを用いた cos 類似度 [7] によるニューラル固有表現正規化モデルを提案する．モデルの概観を図 2に示す．

モデルは固有表現を示す文字列をサブワードに分割したものを入力とする．BERT エンコーダを用い，入力をベクトル表現に埋め込み，先頭と末尾の特殊トークン ([CLS]，[SEP] トークン) を除くベクトルの平均プーリング $\boldsymbol{x}$ をとる．これに重み行列をかけ，tanh により非線形変換した $\boldsymbol{h}$（式 1）を固有表現の表現ベクトルとする．

$$\boldsymbol{h} = \tanh(\boldsymbol{W}\boldsymbol{x}) \tag{1}$$

コーパス内の全 CUI の表現ベクトルと上記の固有表現の表現ベクトルの cos 類似度（式 2）を求め，類似度が最大の CUI を予測結果として決定する．

$$Sim(\boldsymbol{h}, \boldsymbol{w}_i) = cos\theta_i = \frac{\boldsymbol{h} \cdot \boldsymbol{w}_i}{|\boldsymbol{h}||\boldsymbol{w}_i|} \tag{2}$$

ここで $\boldsymbol{w}_i$ は i 番目の CUI の表現ベクトルである．

3.1.2 モデルの学習

学習の際は学習データの不足を補うためデータベース内の登録名を用い訓練データの増強を行い，学習データとデータベースによるデータの比が 1:1 になるようアップサンプリングを行う．学習時の損失関数には cos 類似度を用いたマージンベースの損失 ArcFace[7] を用いる．

$$L_y = -\log\left(\frac{e^{s\cos(\theta_y+m)}}{e^{s\cos(\theta_y+m)} + \sum_{i\neq y} e^{s\cos(\theta_i)}}\right) \tag{3}$$

ここで，s は正規化因子，m はマージンを示すハイパーパラメタ，y は正解の CUI である．

43 万クラス分類に起因する訓練データのスパース性に対応するために，(Step1)CUI の表現ベクトルの学習には，ランダムな初期値から一定回数の学習を行った後，(Step2) データベースの登録名から得た表現ベクトル $\boldsymbol{h}$ の平均値を各 CUI ごとに計算し，この値で対応する CUI の表現ベクトルを置き換え，再度少数の反復回数で学習を行うという二段階学習を考案した．

$$\boldsymbol{w}_i \leftarrow \frac{1}{|C_i|}\sum_{\boldsymbol{h}\in C_i} \boldsymbol{h} \tag{4}$$

ここで C_i は i 番目の CUI に対応する概念のデータベース上の登録名から得られた表現ベクトルの集合である．BERT エンコーダは事前学習済みの値で初期化し，ファインチューニングを行う．最適化手法には Adam[8] を用いる．

3.2 固有表現分類による後処理

現状のシステムの問題点として，参照しているのが固有表現を示す文字列のみであり，文脈を考慮できていないことがある．しかしながら，学習データ内の例は 6,684 件のみであるため，43 万件の候補から文脈を考慮し，CUI を直接予測することは現実的でない．そこで，文脈を用い，対象となる概念の属するクラスを予測し，予測先の絞り込みを行う．提案手法の概観を図 3に示す．

分類先のクラスとしては CUI とは別に，UMLS 内で定義された Semantic Type identifiers（TUI）を元に，

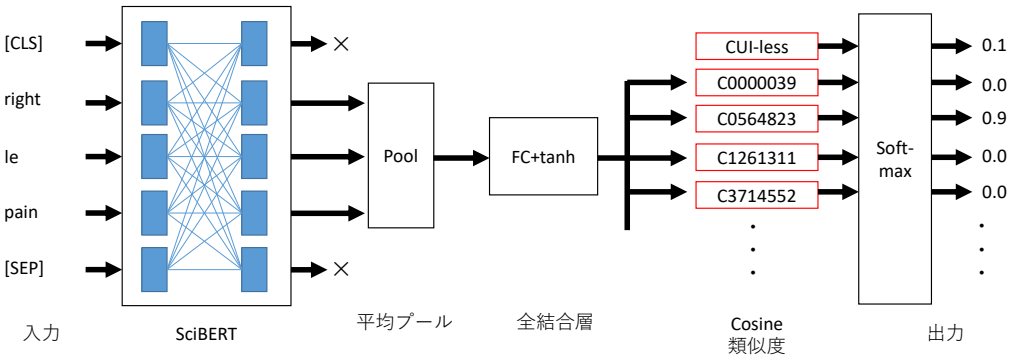

図2　医療概念正規化モデルの概観

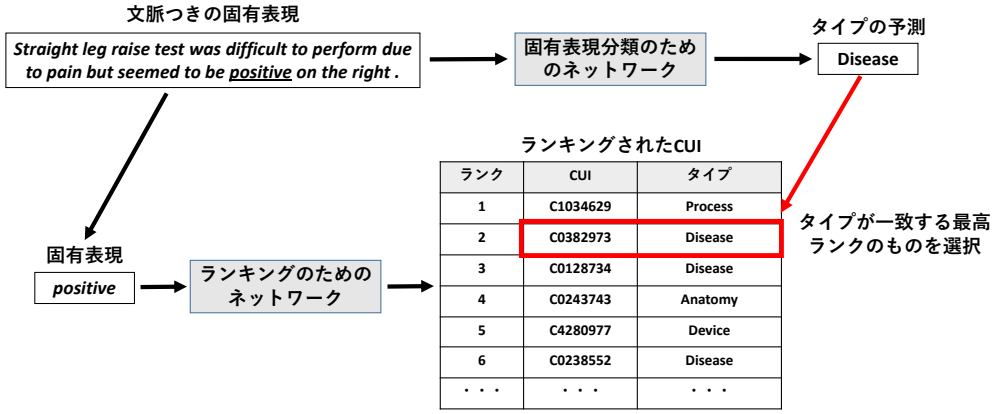

図3　固有表現分類による後処理の概観

ezDI が定義した概念クラス (以降，ezDI クラスと呼ぶ) を用いる．ezDI クラスの一覧を表 3に，提案手法による CUI の予測手順を以下に示す．

1. 対象概念を含む文を Transformer を用い系列ラベリングすることにより，ezDI によるクラス及び “other” に分類する．
2. 各概念について深層学習モデルにより予測されたクラス内の CUI のスコアを計算する．
3. 予測されたクラスの中で最大のスコアの CUI を予測先として決定する．

固有表現分類には．BERT 系列タガーを用いる．サブワードに分割された文脈つきの固有表現を入力とし，各サブワードに対し，固有表現でないものは “other”，固有表現に対してはそのクラスを出力するよう訓練する．予測時は固有表現のサブワードのスコアの平均から，予測クラスを決定する．

4　固有表現正規化モデルの評価

3.1に示したモデルの評価を行った．モデルの学習において，OPUTUNA を用い，学習率・正規化因子・マージン・隠れ層の次元数のハイパーパラメータのチュ

表3　ezDI による概念クラスの一覧

概念クラス
Anatomy
Disease
Lab
Procedure
Device
Medicine
Measurement
Activity
Modifier

ーニングを行った．二段階学習では，モデルの重みの学習にカルテの固有表現のそれぞれが平均で 1 回ずつ利用されるまでを 1 イテレーションとし，750 イテレーション訓練した後，出力層の重みを更新し，25 イテレーションの訓練を行った．

50 件を全ての訓練に用いた異なる乱数による 5 つのモデルのアンサンブルによるテストデータに対する正解率と各モデルの平均精度を表 4に示す．アンサンブルモデルは n2c2 2019 Track 3 での最高スコアである 0.8526 を獲得した．

5 後処理による精度向上の試み

5.1 ezDI クラスの有用性の検証

固有表現の ezDI クラスへの分類が，後処理として有効であるか調べるための実験を行った．正解 CUI と同じ TUI を持つもの，同じクラスに属するものの中で，TTI-COIN のシステムにより最大スコアのものを予測とした場合の正解率を調べた．TUI は全 75 種であり，ezDI による CUI クラスは全 10 種である．

実験結果を表 5に示す．75 種類の TUI がわかった場合 0.052 程度の正解率の向上が見込めるのに対し ezDI の分類を用いた場合 10 クラスの分類に落とし込んだ上で 0.041 程度の向上が見込める．ezDI によるクラスへの固有表現分類の後処理としての有効性が示された．

5.2 固有表現分類の評価

BERT 系列タガーを用い，固有表現分類の評価を行った．学習データを訓練データ 40 件と評価データ 10 件に分割し，Adam により 100 エポック訓練を行った．評価データにおける正解率は 0.4913 であり，現段階では後処理に用いるために十分な精度が得られなかった．

表4 MCN テストデータに対する正解率

モデル	正解率
平均	0.8440
アンサンブル	0.8526

表5 TUI と ezDI によるクラスの情報がある場合の正解率

予測コーパス	正解率
全コーパスからの予測	0.8526
TUI の同じ集合からの予測	0.9047
同クラスの集合からの予測	0.8940

6 おわりに

本研究では，n2c2 2019 Shared Task Track 3 の医療文献の固有表現正規化において，43 万クラス分類問題をニューラルネットにより解くために，SciBERT と cos 類似度を用いたモデルと 2 段階学習による学習手法を提案した．提案モデルは n2c2 2019 Task 3 で最高スコアである 0.8526 の正解率を獲得した．さらに，計算効率，及び，予測精度向上のため，文脈を考慮した固有表現分類による後処理の試みを考察した．後処理の評価では，BERT 系列タガーを用いても固有表現クラスの分類正解率は 0.4913 と低く，現段階では後処理として用いるには不十分な精度となったが，CUI の属するクラスの予測に成功した場合，0.041 程度の正解率の向上の余地があることがわかった．今後は，系列タガーの改良と利用法の検討を行い，計算効率・予測精度の向上を目指す．

参考文献

[1] Luo et al. Mcn: A comprehensive corpus for medical concept normalization. *Journal of biomedical informatics*, 2019.

[2] Bodenreider et al. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 2004.

[3] Beltagy et al. Scibert: A pretrained language model for scientific text. In *EMNLP-IJCNLP*, 2019.

[4] Devlin et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *ACL*, 2019.

[5] Akbik et al. Pooled contextualized embeddings for named entity recognition. In *ACL*, 2019.

[6] Li et al. Cnn-based ranking for biomedical entity normalization. *BMC bioinformatics*, 2017.

[7] Deng et al. Arcface: Additive angular margin loss for deep face recognition. In *IEEE*, 2019.

[8] Kingma et al. Adam: A method for stochastic optimization. *International Conference for Learning Representations*, 2014.