生化学分野のフルペーパーを対象としたリンキングと索引付け

辻村 有輝　井田 龍希　三輪 誠　佐々木 裕
豊田工業大学
{sd18602,sd18006,makoto-miwa,yutaka.sasaki}@toyota-ti.ac.jp

概要

2021 年 8 月に開催された BioCreative VII Track 2 に参加した．Track 2 は論文全体からの固有表現抽出，リンキング，Indexing の 3 つのタスクから構成され，各タスクごとにニューラルモデルと，TF-IDF ベースの Indexing モデルを構築して臨んだ．リンキングでは省略形の解決により性能向上が確認された．また，構築したニューラル Indexing モデルは TF-IDF ベースのモデルに比べエラー伝播に弱く，パイプライン設定では TF-IDF ベースのモデルより低い識別性能となることが分かった．

1　はじめに

医学・生化学分野の大規模文献検索システムである PubMed [1] において，最もよく検索に使用されるのが化学用語である．文献中の化学用語を自動抽出し，主要トピックを見つけて索引付けできるようになれば，検索システムの利用者にとって大きな助けとなり，また後続の言語処理システムにおける有用な情報となる．BioCreative VII Track 2 [2] はより高性能な化学用語の抽出・索引付けアルゴリズムの研究を目的としたコンペティション形式のワークショップトラックであり，参加者はアノテーション付与された 150 件の論文全体からのエンティティ抽出，リンキング，および Indexing タスクに取り組む．本稿では，BioCreative VII Track 2 における参加結果と，そこで使用したモデルについて述べる．なお，本稿の内容は BioCreative VII workshop にて報告した結果 [3] を日本語で再構成したものである．

2　関連研究

2.1　BioCreative VII Track 2

BioCreative VII workshop [4] は生化学分野を対象とした種々の情報抽出タスクを扱うコンペティション形式のワークショップである．対象テーマ別に 5 つのトラックに分かれており，特に Track 2 [2] では生化学分野の論文全体からの固有表現抽出，リンキングおよび Indexing を扱う．

2.1.1　固有表現抽出

固有表現抽出は，文献中に出現するエンティティを抽出する問題である．本トラックで扱うエンティティは生化学用語である．本トラックのユニークな点として，抽出対象が論文全体であることが挙げられ，図表のキャプションなどに登場するエンティティも抽出対象となる．論文はセクションやキャプションといった区切りに分割され，出現するエンティティのスパンがタグ付けされている．本トラックでは各エンティティは必ず連続した 1 つのスパンのみで構成される．公式の評価指標はエンティティスパンに対する F 値である．

2.1.2　リンキング

リンキングは文書中に出現するエンティティとデータベースに登録されたエントリーの対応をとるタスクである．本トラックでは抽出された生化学用語それぞれに対し MeSH シソーラス [5] 全体から対応する概念を割り当てる．MeSH シソーラスには 2021 年 8 月当時で 348,240 件の概念が登録されている．また，本トラックでは一つのエンティティ表層が複数の概念の組み合わせとなりうるマルチラベル分類の設定になっている．該当する概念が存在しない場合があり，その場合は一致なしを表すラベルを付与する．公式の評価指標は F 値であるが，正解数等はエンティティ表層単位ではなく，論文単位でまとめたユニーク概念集合に対して計測される．

2.1.3　Indexing

Indexing は文献に対する索引となる特に重要な話題を抽出するタスクである．本トラックでは各論文ごとに索引となる MeSH シソーラスの概念クラス集合を予測する形で行う．正解概念は，論文中で直

表 1 BioCreative VII Track 2 Indexing タスクの訓練コーパスにおける各条件での正解クラスの網羅数

候補範囲	概念数	正解網羅数
全正解概念数	364	364
論文中に直接出現する概念	5,039	287
直接出現する概念とその親	10,521	342
直接出現する概念とその全祖先	17,845	345

接出現したエンティティのほか，そのエンティティをより一般化した上位概念となることもある．訓練コーパスにおける，各区分に分けた時の正解クラスの網羅数を表 1 に示す．ここで親子関係は MeSH Tree 構造における上位下位関係である．表のとおり，多くの正解クラスは論文中に直接出現し，また直接出現した概念クラスの親クラスを含めることで，全正解クラスの約 94%を網羅できる．公式の評価指標は F 値である．

2.2 二段階学習を用いたニューラルリンキングモデル

我々は医療文書におけるリンキングを扱った N2C2 2019 workshop において，リンキングを分類問題として解くニューラルネットワークベースのリンキングモデルを構築した [6]．入力はリンキング対象のエンティティ表層であり，大規模事前学習済み言語モデルである SciBERT [7] を用いて入力エンティティをベクトル表現へエンコードする．このエンティティ表現と各概念の埋め込みベクトルのコサイン類似度をとることで類似度スコアを計算し，温度付きソフトマックス関数を適用することで，各概念の予測確率を得る．確率が最大となった概念クラスをモデルの最終的な予測として採用する．学習は，訓練コーパスに加えて，データベースの登録名を用いて作成した追加の訓練事例により行う．これにより訓練コーパスに出現しない概念クラスについても学習できるようにする．モデルの学習では，データベースからの訓練事例にのみ出現する概念クラスの訓練正解率が向上しないアンダーフィットの問題が発生する．このため，正解率が向上しなくなった段階で，出力層の重みをエンティティ表現から得た疑似的な収束値に上書きし，その後さらに追加の学習を行うことでこの問題に対処する．

3 提案手法

我々は BioCreative VII Track 2 の各タスクに対し，それぞれニューラルネットワークベースのモデルを構築し取り組んだ．また，Indexing に関しては TF-IDF ベースの手法も構築した．

3.1 固有表現抽出

固有表現抽出は，BILOU (Beginning, Inside, Last, Outside, Unit) スキーマによるエンティティ位置の系列ラベリング問題として解いた．使用したモデルは SciBERT ベースのモデルで，入力はサブワード分割されたセクション一つであり，出力は各サブワードの BILOU タグである．入力を SciBERT でエンコードし，SciBERT の最終層の表現ベクトルを全結合層による出力層への入力として，この出力にソフトマックス関数を適用することで，BILOU タグの確率分布を得る．出力層の入力ベクトルにはドロップ率 20%のドロップアウトを適用する．学習アルゴリズムには Adam を使用する．1 セクションに含まれるサブワード数が SciBERT の最大入力系列長を超える場合，ストライド幅 128 で最大系列長に収まるように分割する．デコード時には Viterbi アルゴリズムを使用し確率が最大となるタグ系列を選ぶ．

3.2 リンキング

リンキングでは先行研究 [6] のモデルをベースラインとして使用した．元々の訓練コーパスに加え，MeSH シソーラス上の全概念の各登録名について，対応する概念クラスを正解とする訓練事例を作成し訓練に用いた．また，BC5CDR データセット [8] の訓練，開発，テストセット全ても訓練事例として追加して利用した．省略形は Ab3P [9] によりあらかじめ正規化し，入力に用いた．

3.3 Indexing

Indexing タスクにおいては，TF-IDF ベースのモデルと，ニューラルネットワークベースのモデルを構築した．どちらのモデルも候補とする概念クラスに対して，それが索引対象かどうかを二値分類する形で予測を行う．候補クラスは論文中に出現した概念クラスと，その概念の MeSH Tree 構造における親概念の集合である．

3.3.1 TF-IDF モデル

このモデルは，各概念クラスの TF-IDF の値に対して，設定した閾値を超えるかどうかで索引対象となるかどうかを判断する．TF-IDF の値は各論文中の出現したエンティティ表層の数から計算する．親概念の出現回数は，子概念の出現回数の合計で

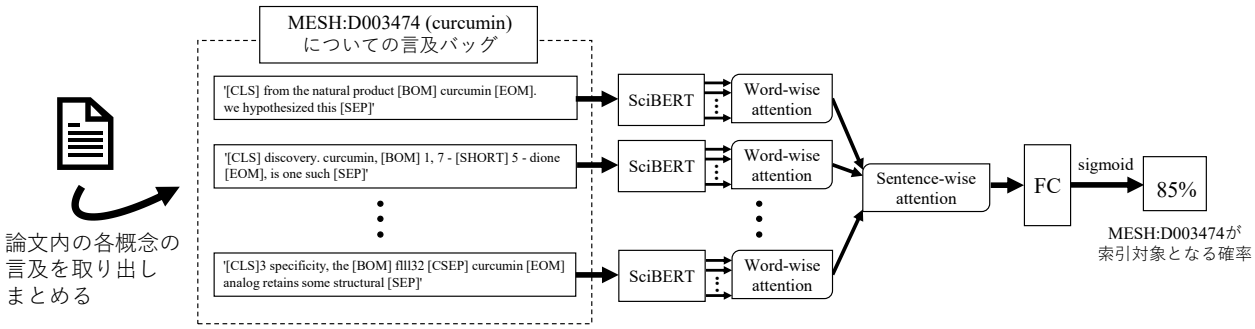

図 1　ニューラル Indexing モデルの概観.

あり，親概念自身も直接出現する場合はそれも合算する．親概念としての出現回数から計算された TF-IDF と，直接の出現数のみから計算した（子概念としての）TF-IDF では，それぞれ別の閾値を使って索引対象の識別を行う．親概念が直接出現する場合は，両種の閾値判定を行い，いずれかでも閾値を超えた場合は索引対象とする．閾値は F 値が最大となるようにチューニングする．

3.3.2　ニューラル Indexing モデル

図 1 に提案する Indexing モデルの概観を示す．初めに，論文中に直接出現する各概念とその親概念クラスごとに言及集合をまとめ，各概念のバッグを作成する．親概念クラスのバッグは，その概念の子概念の言及と，直接の出現がある場合はそれらを加えた集合となる．各言及はサブワードに分割し，前後 4 サブワードまでの文脈を付け加える．エンティティの表層と文脈の間には，特殊トークン [BOM], [EOM] を挿入しスパンの境界を表す．エンティティの表層が 9 サブワードを超える場合，これに収まるように中央のトークンを特殊トークン [SHORT] で置き換える (例：図 1 中段の言及)．また，言及がバッグにおける子概念のものである場合，それを表す特殊トークン [CSEP] と，バッグの概念に対応する辞書登録名を付け加える (例：図 1 最下段の言及)．以上の工程で作成した言及バッグをモデルの入力とし，そのバッグに対応する概念クラスが索引対象かどうかの二値分類によって Indexing を行う．

モデルは各言及を SciBERT によりエンコードし，その最終層の表現ベクトルに対してトークンレベルのアテンション機構を適用することで言及表現を得る．これらにさらに言及レベルのアテンション機構を適用してバッグ表現を計算し，全結合層とシグモイド活性化関数によって索引対象である確率値を得る．アテンション機構にはマルチヘッドアテンション [10] を用いた．損失関数には負の対数尤度を使い，最適化アルゴリズムには Adam を採用した．

表 2　固有表現抽出の結果.

セット	モデル	Precision	Recall	F 値
開発	Ours	0.8671	0.8305	0.8484
テスト	Ours	0.8476	0.8101	0.8284
	参加者中央値 [2]	0.8476	0.8136	0.8373
	参加者最高位 [12]	0.8759	0.8587	0.8672

4　実験

4.1　実験設定

BioCreative VII Track 2 で提供される訓練コーパス 150 論文のうち，120 論文を訓練セット，30 論文を開発セットとして分割し，開発セットにおける各評価指標が最大になるようにチューニングを行った．チューニングには Optuna [11] を使用した．テスト用のモデルの学習には 150 論文全てを訓練データとして使用し，固有表現抽出，リンキング，Indexing の順にパイプラインで予測を行った．テスト用のリンキングモデル，Indexing モデルの学習時の入力には，それぞれアノテーションに基づいた正しいスパン・概念ラベルを使用した．各ニューラルモデルは，同一の設定で 5 回モデルの学習を行い，各出力確率の平均をとる形でアンサンブルを行った．実装には PyTorch 1.8.1 を利用し，GPU での計算には GeForce RTX 3090 を用いた．

4.2　実験結果

4.2.1　固有表現抽出

表 2 に固有表現抽出の実験結果を示す．投稿したモデルはトラック参加者全体の中央値から 0.89%ポイント低い F 値であった．使用されたモデルの多く

表 3 開発セットにおけるリンキングの結果．“-OW” は重み上書きとその後の追加学習の不利用を表す．“+Ab3P” および “+BC5CDR” は Ab3P と BC5CDR の利用を表す．

モデル	Precision	Recall	F 値
ベースライン	0.7325	0.8812	0.8000
-OW	0.7433	0.8380	0.7878
+Ab3P	0.7600	0.8855	0.8180
+Ab3P, -OW	0.7704	0.8445	0.8058
+BC5CDR	0.7400	0.8823	0.8049
+BC5CDR, -OW	0.7635	0.8542	0.8063
+Ab3P, BC5CDR	0.7640	0.8844	**0.8198**
+Ab3P, BC5CDR, -OW	0.7830	0.8575	0.8186

表 4 テストセットにおけるリンキングの結果．

モデル	Precision	Recall	F 値
ベースライン	0.7078	0.8698	0.7805
+Ab3P	0.7338	0.8683	0.7954
+BC5CDR	0.7038	0.8670	0.7769
+BC5CDR, Ab3P	0.7306	0.8658	0.7925
参加者中央値 [2]	0.7120	0.7760	0.7749
参加者最高位 [14]	0.8621	0.7702	0.8136

が BERT [13] ベースのモデルであった．参加チームは合計 15 チームであり，11 位のスコアだった．

4.2.2 リンキング

表 3 にリンキングタスクにおける各モデルの開発スコアを示す．本データにおいては重み上書きの利用による性能向上は，先行研究 [6] で報告されたもの比べあまり大きくない．特に，BC5CDR データセットを使用した場合は，重み上書きを使わずともほとんど同じ F 値を得られた．これは，追加の学習データによって学習が容易になり，重み上書きの導入の目的であるアンダーフィットが起こりづらかったためだと考えられる．Ab3P による省略形の解決では常に性能が向上した．BC5CDR データセットの利用による性能向上は比較的小さく，Ab3P を利用する場合はほとんど差が生まれない結果となった．

表 4 にテストセットにおけるリンキングのスコアを示す．各参加者が採用したシステムには，ニューラルベースのものと，ルールベースおよびルールとニューラルのハイブリットのモデル各種が見受けられた．最高位のスコアを獲得したチームは，ルールベースでリンキングを行ったのち，リンク先を見つけられなかったエンティティをニューラルベースのモデルで補うシステムであった [14]．参加チームは合計 15 チームであり，うち 2 チームは固有表現抽出タスクには参加していない．提案システムは全体で 4 位であった．

表 5 開発セットにおける Indexing の結果．“スパン・概念” は，モデルの入力がアノテーションに基づくか (“Gold”)，パイプラインにより得られた予測結果であるか (“Pipeline”) を表す．

モデル	スパン・概念	Precision	Recall	F 値
ニューラル	Gold	0.6557	0.5714	0.6107
ニューラル	Pipeline	0.2292	0.3143	0.2651
TF-IDF	Gold	0.5714	0.5714	0.5714
TF-IDF	Pipeline	0.3651	0.3286	0.3459

表 6 テストセットにおける Indexing の結果．

モデル	Precision	Recall	F 値
ニューラル	0.2352	0.2661	0.2497
TF-IDF	0.2753	0.6163	0.3806
参加者中央値 [2]	0.4411	0.3883	0.3971
参加者最高位 [15]	0.4397	0.5344	0.4825

4.2.3 Indexing

表 5 に開発セットにおける Indexing のスコアを示す．表のとおり，ニューラルモデルは与えられるエンティティのスパンと概念ラベルが正しいのであれば TF-IDF よりも高い F 値を得ることができるが，パイプラインの設定の時は大きくスコアを落とし，エラー伝播の影響を受けやすい．これは，ニューラルモデルの入力にはスパンの境界情報を利用しているためにスパンのずれが直接入力に影響を与える一方，TF-IDF モデルの入力は出現数のみであり，エラー伝播の影響が小さいためだと考えられる．表 6 にテストセットにおける Indexing のスコアを示す．提案モデルのテストセットのスコアは，開発セットにおけるパイプラインの設定と同等のスコアになった．最高位を記録したチームのモデルは PubMedBERT と特徴量を組み合わせた二値分類モデルであった [15]．Indexing タスクは前 2 つのタスクを解いたうえで取り組む必要があり，時間的な制約もあって最終的な参加チームは 4 チームと前 2 つのタスクより少ない．また，我々を含めた 3 チームは公式の提出期限後の追加実験としての提出である．我々は 3 位のスコアであった．

5 おわりに

本論文では，BioCreative VII Track 2 を対象に構築したモデルを評価した．今後の課題として，ニューラル Indexing モデルのエラー伝播の問題を改善することが挙げられる．

謝辞

本研究は JSPS 科研費 JP20K11962 の助成を受けたものです.

参考文献

[1] National Library of Medicine. PubMed, 1996. https://pubmed.ncbi.nlm.nih.gov/.

[2] Robert Leaman, Rezarta Islamaj, and Zhiyong Lu. The overview of the NLM-Chem BioCreative VII track. In **BioCreative VII Challenge Evaluation Workshop**, pp. 108–113, 2021.

[3] Tomoki Tsujimura, Ryuki Ida, Isanori Oiwa, Makoto Miwa, and Yutaka Sasaki. TTI-COIN at BioCreative VII Track 2. In **BioCreative VII Challenge Evaluation Workshop**, pp. 156–161, 2021.

[4] BioCreative VII, 2021. https://biocreative.bioinformatics.udel.edu/events/biocreative-vii/biocreative-vii/.

[5] National Library of Medicine. MeSH. https://www.ncbi.nlm.nih.gov/mesh/.

[6] 茂里憲之, 辻村有輝, 三輪誠, 佐々木裕. 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. 言語処理学会第 26 回年次大会発表論文集, pp. 1487–1490, 2020.

[7] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A Pretrained Language Model for Scientific Text. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3615–3620, 2019.

[8] BioCreative V, 2015. https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-3-cdr/.

[9] Sunghwan Sohn, Donald C Comeau, Won Kim, and W John Wilbur. Abbreviation definition identification based on automatic precision estimates. **BMC Bioinformatics**, Vol. 9, No. 402, 2008.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In **Proceedings of the 31st International Conference on Neural Information Processing Systems**, p. 6000–6010, 2017.

[11] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD ’19), pp. 2623–2631, 2019.

[12] Hyunjae Kim, Mujeen Sung, Wonjin Yoon, Sungjoon Park, and Jaewoo Kang. Improving Tagging Consistency and Entity Coverage for Chemical Identification in Full-text Articles. In **BioCreative VII Challenge Evaluation Workshop**, pp. 140–143, 2021.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **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, 2019.

[14] Tiago Almeida, Rui Antunes, João Figueira Silva, João Rafael Almeida, and Sérgio Matos. Chemical detection and indexing in PubMed full text articles using deep learning and rule-based method. In **BioCreative VII Challenge Evaluation Workshop**, pp. 119–123, 2021.

[15] Arslan Erdengasileng, Keqiao Li, Qing Han, Shubo Tian, Jian Wang, Ting Hu, and Jinfeng Zhang. A BERT-Based Hybrid System for Chemical Identification and Indexing in Full-Text Articles. In **BioCreative VII Challenge Evaluation Workshop**, pp. 130–134, 2021.