エンティティの階層的分類体系を用いた遠距離教師あり固有表現抽出

芝原 隆善 [1]　山田 育矢 [1,2]　西田 典起 [1]　寺西 裕紀 [1]

大内 啓樹 [1,3]　古崎 晃司 [1,4]　渡辺 太郎 [3]　松本 裕治 [1]

[1] 理化学研究所 [2]Studio Ousia [3] 奈良先端科学技術大学院大学 [4] 大阪電気通信大学

takayoshi.shibahara@a.riken.jp　{ikuya.yamada, noriki.nishida, hiroki.teranishi, hiroki.ouchi, kouji.kozaki, yuji.matsumoto}@riken.jp　{hiroki.ouchi, taro}@is.naist.jp　ikuya@ousia.jp　kozaki@osakac.ac.jp

概要

本論文では辞書を疑似データの作成に活用した固有表現抽出（遠距離教師あり固有表現抽出：Distant Supervision Named Entity Recognition）に取り組む．本論文では取得したいカテゴリの辞書のみを活用してきた従来研究とは異なり，分類体系全体を考慮する手法を提案する．具体的には分類体系に含まれる全てのカテゴリを含んだ疑似データを作成し，モデルの学習に利用する．実際に生物医学系のエンティティ・リンキング及び固有表現抽出のデータセット：MedMentions とリンキング対象の知識ベース：UMLS を活用して遠距離教師あり固有表現抽出を行ったところ，ベースラインに比べて F1 値において 4.19 pt の改善を達成し階層的な分類体系全体を考慮する重要性を明らかにすることができた．

1　導入

文章に記述されているエンティティの位置とクラスを特定する固有表現抽出（Named Entity Recognition：NER）は自然言語処理における基本的なタスクである．応用例として，情報抽出 [1] や情報検索 [2] などをあげることができる．一方で多様な情報抽出・検索のニーズに応じた教師ありデータを準備するには多大なアノテーションコストがかかってしまう．

このため本論文では教師データがない状況における固有表現抽出に取り組む．その中でも特に遠距離教師あり学習（Distant Supervision）に基づく固有表現抽出（Distant Supervision Named Entity Recognition：DS NER）[3, 4, 5] に取り組む．遠距離教師あり学習では取得したいカテゴリに応じた辞書を用いて疑似データを作成することで学習データの不在に対処する．遠距離教師あり学習は DBpedia [6], UMLS [7] などの大規模知識ベースの発達により，これらの知識ベースに含まれる様々なカテゴリに対して教師データの不在を補うことができる．

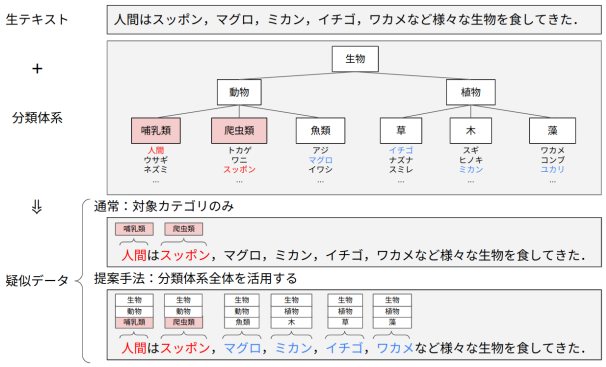

図 1　既存の Distant Supervision NER(DS NER) との比較．通常の DS NER では疑似データ作成時に対象となるカテゴリ（哺乳類・爬虫類）に含まれる語句のみを活用する．一方で提案手法では全てのエンティティと全てのカテゴリを利用する．例えば「マグロ」や「ミカン」などのエンティティや魚類・植物などのカテゴリも疑似データに活用する．

DS NER の先行研究 [3, 4, 5] ではエンティティ全体の属する分類体系を利用せず，関心のあるクラス（以下対象カテゴリと呼ぶ）の情報しか利用してこなかった．例えば図 1 のように，哺乳類・爬虫類のみを検出する固有表現抽出器を作成する場合を考えてみるとわかりやすい．疑似データ作成の際に哺乳類・爬虫類という分類体系の一部のみを利用し生物全体の分類体系に含まれる他のカテゴリの情報を利用してこなかったのである．より具体的には哺乳類・爬虫類に所属する「人間」や「スッポン」などの語句しか利用せず，知識ベース

内に存在しているが対象カテゴリでないクラス (e.g. 動物・魚類) や語句 (e.g. 「マグロ」・「ミカン」) を活用してこなかったのである．

そこで本論文では「分類体系全体の情報を使う」というアイデアを提案・実装する．具体的には分類体系の全クラスを含んだ疑似データを作成し，活用する手法を提案する．

実際に生物医学系のエンティティ・リンキング及び固有表現抽出のデータセット：MedMentions [8] とリンキング対象の知識ベース：UMLS [7] を活用して検証を行った．ベースラインに比べて F1 値において 4.19 pt の改善を達成し分類体系全体を考慮する重要性を明らかにすることができた．

2 手法

提案手法では「分類体系の全カテゴリを認識するモデル」によって「対象カテゴリの抽出」を行う．ただし教師データのない状況を考えているので「分類体系の全カテゴリを抽出するモデル」は「分類体系の全カテゴリを含んだ疑似データ」で訓練することによって実現する．そこで「分類体系の全カテゴリを含んだ疑似データ」をどのように作成したか，「分類体系の全カテゴリを含んだ疑似データ」によって「分類体系の全カテゴリを認識するモデル」をどのように訓練し実現したか，「分類体系の全カテゴリを認識するモデル」によってどのように「対象カテゴリの抽出」をしたかを述べていく．

2.1 分類体系の全カテゴリを含んだ疑似データ作成

「分類体系の各カテゴリに所属する語句」に対して辞書マッチを行うことで「分類体系の全カテゴリを含んだ疑似データ」を作成する．「分類体系の各カテゴリに所属する語句」は「各カテゴリに所属する語句の情報」と「各カテゴリ間の階層構造」を用いて取得する．

具体例として図 1 の「人間」という語句を考える．「人間」という語句は哺乳類に直接所属している．さらに哺乳類は動物・生物の下位概念でもある．このことから「人間」という語句は哺乳類・動物・生物の 3 つのカテゴリに所属していることがわかる．この情報から，「人間はスッポン，マグロ，ミカン，イチゴ，ワカメなど様々な生物を食してきた」という文の「人間」の部分に哺乳類・動物・生物という 3 つのラベルを文字列マッチによって付与する[1)]．同文に出現する「スッポン」・「マグロ」・「ミカン」・「イチゴ」・「ワカメ」などの他の語句についても同様に複数のラベルを付与する．

2.2 疑似データに基づく全カテゴリを認識するモデルの学習

2.1 で作成された疑似データ上でモデルを学習させ，「分類体系の全カテゴリを認識するモデル」の取得を目指す．本研究では先行研究 [9] と同じ BERT ベースのスパン分類モデルを学習させる．このスパン分類モデルは BERT により文を符号化し，スパンの始端・終端のベクトルを連結して線形分類を行うものである（図 2）．

まず前処理として 2.1 で作成された NER 疑似データをスパン分類のデータに変換した．具体的にはスパン最大長を決め文中のスパンを列挙する．その後ラベルのあるスパンはそのままのラベルを利用し，ラベルの付与されていないスパンは“O”ラベルを持つスパンとした．また，複数の正解に対処できるように負の対数尤度の正解クラスに対する平均をロス関数として学習を行った．

単純に学習を行うと，スパン分類モデルが既知の（辞書に含まれている）語句は抽出できる一方で未知の（辞書に含まれていない）語句を抽出しにくくなってしまう．例えば文：「新種の哺乳類オリンギートがアンデス山脈で人間に発見された．」の「人間」は抽出できる一方で，分類体系に含まれない新種の「オリンギート」は抽出しにくくなってしまう．そこで Li らの手法 [10] を使い，ラベル付与されていないスパンの影響を“O”スパンのアンダーサンプリングによって割り引く．

2.3 対象カテゴリの抽出

2.1 の疑似データ上で学習した「分類体系の全カテゴリを認識するモデル」を用いて「対象カテゴリの抽出」を行う．具体的には i) スパン最大長を決め文中の候補スパンを列挙する．次に ii) 予測対象のカテゴリを絞って分類する．最後に iii) これらスパン分類の出力を固有表現抽出の出力に変換する．ii) において予測対象のカテゴリは対象カテゴリとそれを補完し分類体系全体を被覆するような最小限のカテゴリ（以下補完カテゴリと呼ぶ）及び“O”ラベルに限定した（図 2）．iii) の際には，Yamada ら [11] と同様にスパンを予測確率の高いものから重複の

1) ただし修飾語も含んだ文字列も取得するために名詞句チャンカーを利用し，名詞句の末端に語句がある場合にラベルを付与した．

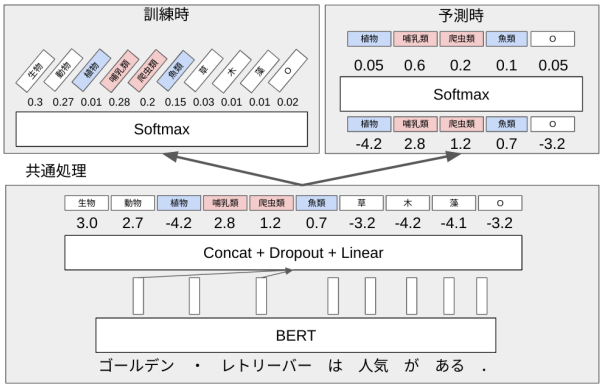

図 2 訓練時と予測時のスパン分類モデルの挙動の違い．訓練時には分類体系の全カテゴリ＋“O”ラベルを対象にスパン分類確率を計算するが，予測時には対象カテゴリ＋補完カテゴリ（対象カテゴリを補完し分類体系全体を被覆するような最小限のカテゴリ）＋“O”ラベルの予測スコアからスパン分類確率を計算する．

ないように選んでいくという方法で変換する．（ただしこの際に“O”ラベルが予測されたスパンは除いた．）

3 実験設定

本論文では UMLS(2021AA 版) [7] という知識ベースの分類体系の一部のカテゴリ[2] を対象カテゴリとして，UMLS [7] を対象としたエンティティ・リンキング及び固有表現抽出データセットである MedMentions データセット [8] を用いて提案手法の評価を行った．UMLS は生物医学分野の知識ベースで，127 のクラス (Semantic Types) と 16,132,274 の用語を持つ．MedMentions は，4,392 の生物医学論文抄録に 352,496 のスパンが UMLS の概念に対応付けられるようにアノテーションされているデータセットである．train/dev/test の文書数はそれぞれ 2,635/878/879 である．train 部分のデータセットは Distant Supervision の擬似アノテーションの対象として，dev/test の分割は正解アノテーションまま利用する．

本論文では提案手法をいくつかのベースラインと比較する．まず疑似データ作成手法と教師あり設定との比較を行うことで，遠距離教師あり学習としての妥当性を確認する．つまり，提案手法が辞書マッチよりも改善し，教師あり学習と同等の精度を達成するという理想にどこまで近づけているかを確認する．次に「分類体系の持つカテゴリの一部ではなく全体を活用する」という本論文のアイデアの是非を確認するために（通常の DS NER と同様に）対象カテゴリのみを疑似データ作成に活用した BERT ベーススパン分類モデルとの比較を行う．最後に「分類体系の持つカテゴリの一部ではなく全体を活用する」というアイデアが提案手法においてよりうまく実現できているかを確かめるために，類似した目的意識を持つ先行研究 [9] の手法と比較する．この先行研究は対象カテゴリに追加して補完カテゴリを疑似アノテーションに活用する研究である．

また補足実験として，「分類体系の全カテゴリを活用する」というアイデアが教師あり設定でも有用であるかどうかを確かめた．ただし，本論文の主題から外れるため付録 A に結果を示す．

2） 具体的には MedMentions[8] で固有表現抽出タスクの対象として指定されている 21 個のカテゴリ

4 結果

実験結果は表 1 のようになった．結果の評価尺度としては通常固有表現抽出に利用されるスパン・クラス完全一致によって予測・正解スパンの一致を測った Strict Precision/Recall/F1 スコアとスパン部分一致に条件を緩和した Lenient Precision/Recall/F1 スコアを利用した．ここで部分一致を許容した尺度を利用した理由は，辞書とマッチするエンティティに修飾語が掛かってエンティティ全体のスパンの曖昧性が生じてしまうためである．例えば，辞書に「水」が含まれる際に「純水」の「純」が正解スパンに含まれるかどうかはアノテーション基準に依存し統一的な判断をすることが難しい．

4.1 遠距離教師あり学習としての妥当性

辞書マッチ手法の結果（表 1 行目）と提案手法の結果（表 4 行目）を比較すると，Strict F./Lenient F. の両方において改善を達成していることがわかる．また，提案手法（表 4 行目）を教師あり学習（表 5 行目）と比較すると Strict F./Lenient F. スコアにおいて 31.8 pt/18.73 pt もの差があることがわかる．以上から，辞書マッチと比べれば妥当なスコアである一方で，まだまだ教師あり設定と同等の精度という理想には大きく及ばないことがわかる．

4.2 論文のアイデアの成否

分類体系全体を利用する提案手法（表 4 行目）は分類体系の一部にしか着目しないベースライン（表 2 行目）に比べて Strict F./Lenient F. スコアにおいて 4.19 pt/3.7 pt の改善を示している．この結果は分類体系の一部しか評価対象でない場合であっても分類

手法	学習データ種別	Strict			Lenient		
		P.	R.	F.	P.	R.	F.
データ作成手法	学習データなし	**26.88**	17.38	21.11	**57.50**	38.59	46.19
スパン分類（対象カテゴリのみ）	疑似データ	21.78	17.67	19.51	53.26	44.21	48.31
＋補完カテゴリ [9]	疑似データ	22.40	17.39	19.58	55.69	43.81	49.04
＋分類体系の残りのカテゴリ	疑似データ	26.09	**21.71**	**23.70**	55.82	**48.75**	**52.04**
スパン分類（対象カテゴリのみ）	教師データ	54.84	56.19	55.50	71.06	70.47	70.77

表 1　実験結果：疑似データ作成に利用した辞書マッチ，遠距離教師あり設定，教師あり設定における固有表現抽出のスコアを表示している．遠距離教師あり設定ではスパン分類モデルを利用し，通常の DS NER と同様に対象カテゴリのみを利用する方法・先行研究 [9] と同様に補完カテゴリを追加して利用する方法・分類体系の残りのカテゴリを追加して利用する提案手法の結果を示している．Strict P./R./F. はスパン・クラス完全一致によって予測・正解スパンの一致を測った Precision/Recall/F1 スコアである．Lenient P./R./F. はスパン部分一致を許容するようにしたスコアである．太字は教師データを必要としない手法の中で最も高いスコアである．

体系全体を利用することが有用であると示しているといえる．

この理由について次のような仮説で説明することができる．まず対象カテゴリに対して子クラスはクラス内の多様性の理解につながり，精度の増大につながる（例えば対象カテゴリが 動物 の際に，その下位カテゴリである 魚類 を利用することで，動物 には泳ぐものがいることを理解できる）．また，対象カテゴリの親クラスは子クラスに対する制限として働き Precision を増加させていると考えられる（例えば対象カテゴリが 哺乳類 の際に，動物 でないものは 哺乳類 と分類されにくくなる）．同様に補完カテゴリに対しても Precision/Recall が増大することで，さらに対象カテゴリの Recall/Precision の増大に寄与していると考えることができる．

4.3　補完カテゴリの追加と階層構造

対象カテゴリに追加して補完カテゴリを利用する先行研究 [9]（表 3 行目）では対象カテゴリのみを利用するベースライン（表 2 行目）に比べて Strict(Lenient) P. では改善が見られるものの Strict(Lenient) R. においては減少してしまっている．このことは補完カテゴリの追加により補完カテゴリの予測を増やすことで，対象カテゴリの Recall を犠牲に Precision を増加せているのだと解釈できる．例えば，対象カテゴリ 哺乳類 に対する補完カテゴリ 魚類 の追加によって，泳ぐなら 魚類 であって 哺乳類 ではないとしてしまっているようなことが起きてしまっていると考えられる．

一方で分類体系の全カテゴリを活用する提案手法（表 2 行目）ではこの Strict(Lenient) R. の減少を補っている．このことは補完カテゴリに近い特徴をもつ対象カテゴリの下位クラスによって，本来対象カテゴリであるエンティティを補完カテゴリとして予測する失敗が減っているためであると考えられる．例えば，対象カテゴリ 哺乳類 の下位カテゴリ クジラ目 の追加によって，「哺乳類 の中にも泳ぐものが存在することをモデルが理解し，本来 哺乳類 クラスである事例への 魚類 クラスの過剰な予測を防いでいるというようなことが起きていると考えることができる．他にもラベル不均衡により対象カテゴリがうまく学習できていないのが原因であると考えることもできる．なぜなら，補完カテゴリは対象カテゴリより階層の浅く学習事例の多いクラスになりやすい（例：哺乳類 に対する 植物 ）と考えられるからである．

5　結論

本研究では遠距離教師あり学習および教師あり学習（付録 A）においてたとえ分類体系の一部のカテゴリに興味がある場合であっても，分類体系全体に含まれる全カテゴリを活用することは有用であることを明らかにした．一方で今回の手法が一般ドメインでも有用かどうか，同じく教師データがない状態で利用可能な Few-Shot 設定と比べてどれほどの優位性があるかを検証しきれていない．

本研究の発展としてラベル定義などラベル階層構造以外の様々な情報の Distant Supervision NER への活用や Fine-Grained NER データセットの Distant Supervision NER 事前学習データとしての活用などが考えられる．

参考文献

[1] Tianyi Liu, Xinsong Zhang, Wanhao Zhou, and Weijia Jia. Neural relation extraction via Inner-Sentence noise reduction and transfer learning. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 2195–2204, Brussels, Belgium, 2018. Association for Computational Linguistics.

[2] Ahmed Aliwy, Ayad Abbas, and Ahmed Alkhayyat. NERWS: Towards improving information retrieval of digital library management system using named entity recognition and word sense. **Big Data and Cognitive Computing**, Vol. 5, No. 4, p. 59, October 2021.

[3] Zhanming Jie, Pengjun Xie, Wei Lu, Ruixue Ding, and Linlin Li. Better modeling of incomplete annotations for named entity recognition. 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. 729–734, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[4] Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. Distantly supervised named entity recognition using Positive-Unlabeled learning. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 2409–2419, Florence, Italy, July 2019. Association for Computational Linguistics.

[5] Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. BOND: BERT-Assisted Open-Domain named entity recognition with distant supervision. In **Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**, KDD '20, pp. 1054–1064, New York, NY, USA, August 2020. Association for Computing Machinery.

[6] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, and Others. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. **Semantic web**, Vol. 6, No. 2, pp. 167–195, 2015.

[7] Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. **Nucleic Acids Res.**, Vol. 32, No. Database issue, pp. D267–270, January 2004.

[8] Sunil Mohan and Donghui Li. MedMentions: A large biomedical corpus annotated with UMLS concepts. p. 13, 2018.

[9] 芝原 隆善, 大内 啓樹, 山田 育矢, 西田 典起, 寺西 裕紀, 古崎 晃司, 渡辺 太郎, 松本 裕治. ユーザの興味があるカテゴリに応じた NER システム構築フレームワーク. 言語処理学会年次大会, 2022.

[10] Yangming Li, Lemao Liu, and Shuming Shi. Empirical analysis of unlabeled entity problem in named entity recognition. September 2020.

[11] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6442–6454, Online, November 2020. Association for Computational Linguistics.

[12] Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. Few-Shot named entity recognition: An empirical baseline study. pp. 10408–10423, November 2021.

[13] Alexander Fritzler, Varvara Logacheva, and Maksim Kretov. Few-shot classification in named entity recognition task. In **Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing**, New York, NY, USA, April 2019. ACM.

[14] Xiao Ling and Daniel S Weld. Fine-Grained entity recognition. In **AAAI**, 2012.

[15] Satoshi Sekine and Chikashi Nobata. Definition, dictionaries and tagger for extended named entity hierarchy. In **LREC**, 2004.

A　教師あり設定における提案手法のアイデアの有効性

手法	P.	R.	F..
スパン分類（対象カテゴリのみ）	54.84	56.19	55.50
＋補完カテゴリ [9]	56.20	55.66	55.93
＋分類体系の残りのカテゴリ	59.37	60.85	60.10

表 2　教師あり設定において利用するカテゴリを変更した際の Strict P./R./F.

教師あり設定においても通常の固有表現抽出と同じように対象カテゴリのみを活用する方法，先行研究 [9] と同様に補完カテゴリを追加し活用する方法，提案手法のように分類体系に含まれる全カテゴリを活用する方法の三種類を比較した．表 2 では実験によって得られた Strict P./R./F. が表示されている．

実験結果からは 4.3 と同様に，補完カテゴリの追加によって Recall が減少するが Precision が上がること，また分類体系の残りのカテゴリを追加するとこの Recall の減少を補い，Precision/Recall の両方が対象カテゴリのみの活用より改善するという結果が得られた．以上のことから分類体系全体の階層構造を活用することはたとえ教師あり設定であっても有用であると言える．

B　関連研究

本研究ではアノテーションデータが存在しない状況において利用可能な固有表現抽出技術に取り組む．このような状況に対応しうる実験設定として 2 つのものが挙げられる．一つは辞書（語句をカテゴリごとにまとめたもの）に基づいた文字列マッチによって疑似データを作成し活用する方法（Distant Supervision NER [3, 4, 5]）であり，もう一つはごく少量のアノテーションデータを利用する方法（Few-Shot NER [12, 13]）である．本研究は Distant Supervision NER に分類されるものである．

通常の Distant Supervision NER が疑似データ作成時に対象カテゴリのみを活用するのに対して，本研究は分類体系の持つカテゴリ全てを活用する (図 1)．この点において今回の研究は分類体系に含まれるカテゴリ階層全体を考える Fine-Grained NER[14] や Extended Named Entity [15] の観点を Distant Supervision NER に導入した研究であると言える．また，この研究は類似した問題意識を持つ先行研究 [9] を発展させたものである．