表層が同じ文字列の同一性を表現した深層固有表現抽出

吉村 貴紀　牧野 晃平　三輪 誠　佐々木 裕
豊田工業大学
{sd23447,sd21505,makoto-miwa,yutaka.sasaki}@toyota-ti.ac.jp

概要

本稿では，BERT の内部で文書中の表層が同じ文字列について，その同一性を表現する機構を取り入れた，One Sense Per Discourse 仮説を考慮した深層固有表現抽出モデルを提案する．日本語医療文書データセット Real-MedNLP において提案手法における固有表現抽出の性能は向上しなかったが，解析によりモデルが予測する固有表現のタイプの一貫性が向上していることや，表層が同じ固有表現のスパンを正しく認識できた場合にそのタイプの正解率が向上することを確認した．

1　はじめに

固有表現抽出（Named Entity Recognition; NER）は文章中から人名や組織名などの固有表現を特定し，分類するタスクである [1]．これは自然言語で記述された情報を計算機で扱うための基礎技術で，取得した固有表現は関係抽出や質問応答などの下流タスクに利用される [2]．

NER においては，BERT (Bidirectional Encoder Representations from Transformers) [3] を用いた手法 [4, 5] が高い性能を示しており，未知の単語を含む固有表現が登場しても，文脈情報からある程度正しく予測できる．一方で，文脈のみでは固有表現を正しく認識出来ない場合がある．例えば，略語の場合では，一度だけ用語の種類も特定できるように定義し，以降の文脈では用語の種類が特定できない形式で記述されて，正しく予測できないことがある．

人はこのような場合においても，最初に定義された用語に関する文脈に基づいて，中盤以降の同じ用語を解釈することで，文書を正しく読むことができる．この経験則に関連する仮説として，**議論や文書中に現れる表層が同じ語句同士は，それぞれ同じ意味を示す傾向が強い**とする One Sense Per Discourse 仮説（OSPD 仮説）[6] が提唱されている．OSPD 仮説を考慮して，文書中の表層が同じ語句すべての意味が同じと解釈できれば，語句の解釈の曖昧性を減少できる可能性がある．

しかしながら，BERT を用いた NER においては，OSPD 仮説を表現できず，考慮もされていない．BERT を用いた NER においては，表層が同じトークンに対して，同じ単語表現を使うことはあっても，それらのトークンの表層が同一であるという情報は BERT 内では表現されていない．さらに，入力の単語や単語表現が同じでも BERT によって文脈化された単語表現に変更されるため，同じトークンに対しても，異なった単語表現が使われる．

そこで本研究では，BERT を用いた NER[5] を基盤に，OSPD 仮説を考慮した深層固有表現抽出モデルを実現し，その固有表現抽出に与える影響を明らかにすることを目指す．そのために，BERT の内部で文書中の表層が同じ文字列の同一性を表現する機構を取り入れ，それを利用した固有表現抽出モデルを構築する．OSPD 仮説をモデル内で明示的に表現できれば，略語の例のように，解釈の曖昧性を減らし，より正しく固有表現の認識を行うことができると期待される．

2　関連研究

OSPD 仮説は 1992 年に Gale らによって提唱された [6]．彼らは，Brown Corpus の同じ文書から抽出された表層が同じ多義語のペアの 96% は同じ意味を持っていることを確認した．この結果から，議論や文書において同じ表層で複数回登場する単語は一つの意味を示す傾向が強いと示した．

OSPD 仮説を仮定して利用する手法が，語義曖昧性解消や固有表現抽出，共参照解析などの自然言語処理タスクで性能向上を示している [7, 8]．Barrena らの研究 [7] では，固有表現に Wikipedia 記事へのリンクが付与された AIDA データセット [9] において，文書中で複数回同じ表層で登場した固有表現の 96.01% が同時に同じリンクに結び付いていることを挙げている．これを利用して，同じ表層の固有表

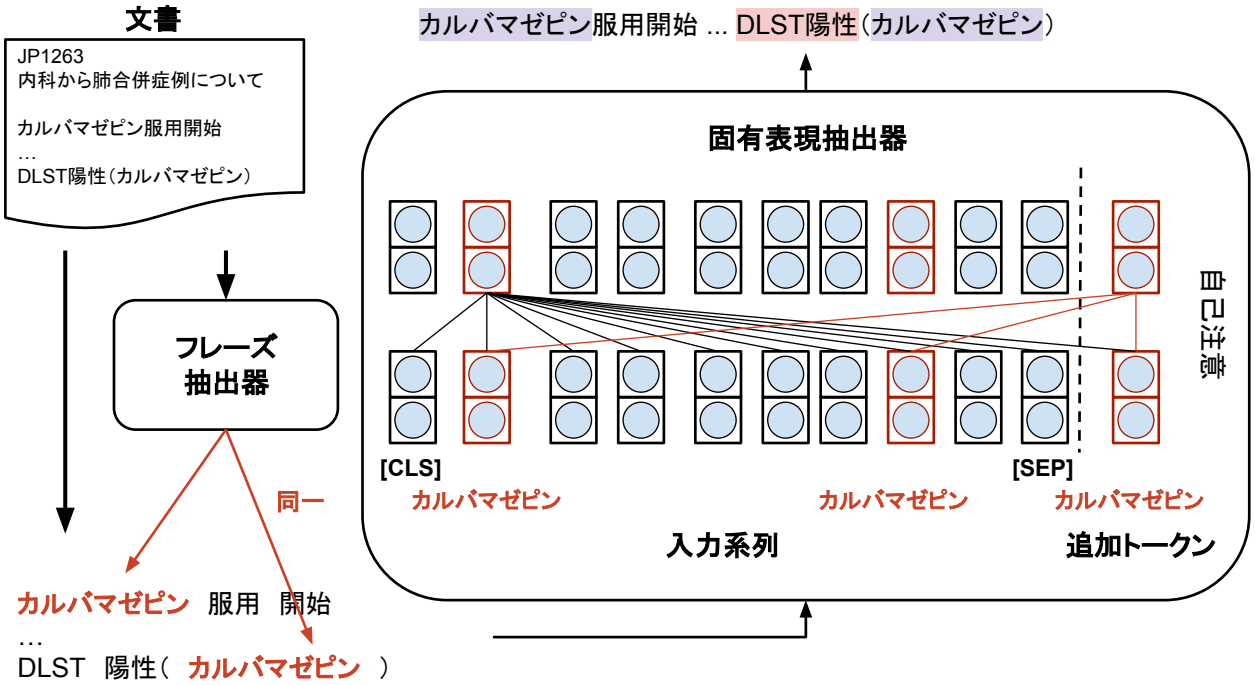

図 1　提案手法の概要

現に結び付くリンクを最も頻繁に予測されたリンクで揃える後処理により，OSPD 仮説を固有表現に適用し，語義曖昧性解消の性能の向上を報告している．Garcia らの研究 [8] では，固有表現抽出と共参照解析の両者の性能を向上させるために，既存の手法を用いた固有表現抽出の結果を，OSPD 仮説に基づいた後処理アルゴリズムで編集し，ラベルの一貫性を向上させることで，両タスクの性能が向上することを示した．

3　提案手法

OSPD 仮説を考慮した深層固有表現抽出モデルを実現し，その固有表現抽出における影響を明らかにするために，同一の表層で登場するフレーズを抽出するフレーズ抽出器と，フレーズの表層の同一性を考慮した BERT をベースにした固有表現抽出器の 2 つのモジュールを作成する．図 1 に概要を示す．

3.1　フレーズ抽出器

フレーズ抽出器では文書中に同じ表層で複数回登場する文字列としてフレーズを抽出する．抽出したフレーズには OSPD 仮説の影響があると考える．しかし，文中においてフレーズの候補は一意ではなく，文字列の範囲は明らかではない．本手法では，単体で意味を持つ内容語であり，固有表現を構成する品詞として最も多くを占めていることから，名詞が OSPD 仮説を表す文字列であると仮定する．形態素解析器を用いて名詞と判定された形態素のうち，文書中に複数回登場するものをフレーズとして抽出する．固有表現を構成している品詞に関する予備調査の結果を付録 A の表 6 に示す．

3.2　フレーズの表層の同一性を利用した深層固有表現抽出

BERT の自己注意機構を利用して，フレーズの表層の同一性を考慮した予測を行うモデルを実現する．各フレーズに対応する追加トークンを作成し，入力系列に追加することで，拡張入力系列として新たな系列を作成し，これをモデルに入力する．

追加トークンはフレーズ抽出器によって抽出されたすべてのフレーズをそれぞれトークナイズすることで得る．得られた追加トークンを入力系列中の “[SEP]” トークンの後ろに結合することで拡張入力系列を得る．追加トークンの埋め込みには，BERT の単語埋め込みと，フレーズ用に別に用意した位置埋め込みを用いて行う．具体的には，フレーズを構成するトークンの順序を表すための位置埋め込みを用意し，元の入力系列とは別の位置埋め込みとする．

モデル内では入力された拡張入力系列に対し選択的な自己注意を行う．入力系列中のトークンは，入

表 1　すべての固有表現に対する各モデルの F 値 [%]

比較手法	F 値
ベースラインモデル	75.3±0.2
ベースラインモデル+訂正	75.4±0.2
提案モデル	74.7±0.4

表 2　各モデルが予測した固有表現のタイプの一貫性 [%]

比較手法	一貫性
ベースラインモデル	91.5±1.9
ベースラインモデル+訂正	100.0±0.0
提案モデル	93.8±2.6

力系列に加えて，対応する追加トークンが存在する場合には，その追加トークンに対しても注意機構を適用する．追加トークンに対しては，入力系列中の同じ表層と判定されたフレーズのトークンと，追加トークン自身を対象とするようにする．

これらの操作により，フレーズの表層の同一性を利用した深層固有表現抽出モデルを実現する．文書中の表層が同じフレーズから取得した表現を各トークンで一つ保持しておき，それを入力系列中のトークンが参照することで，同じ表層で複数回登場するフレーズは一つの意味を持つと考える OSPD 仮説を考慮する．

4　実験

4.1　実験設定

提案手法による固有表現抽出への影響を評価するため，Real-MedNLP データセット [10] の症例報告コーパスを用いた．これはオープンアクセスの日本語症例報告から作成された文書単位のデータセットであり，「病名・症状」や「臓器・部位」といった 13 種類のタイプを含む BIO タグでアノテーションされている．入手可能な 148 件のデータを，訓練，開発，検証のため，70:30:48 に分割して用いた．データセットに含まれる固有表現数を付録 A の表 4 に示す．なお，開発データ中に同じ表層で複数回登場する固有表現のタイプがすべて同じであったことから，OSPD 仮説を仮定する妥当性を確認している．

ベースラインとして BIO ラベルを用いた固有表現抽出に用いられる BERT+CRF モデル [5] を使用し，これに提案手法を加えた提案モデルと比較する．また，ベースラインモデルの出力に後処理を行い，同じ表層で複数回登場する固有表現をそれぞれの表層において最も頻繁に登場するタイプに揃えて訂正する，ベースラインモデル+訂正についても比較する．

BERT には日本語文書で事前学習された NICT-BERT を，形態素解析器には MeCab+juman 辞書を使用した．学習設定の詳細は付録 B に示す．

4.2　結果と考察

検証データを用いて実験を行った．Real-MedNLP Subtask 1 [10] の設定に基づいて主要な 8 種類のタイプの固有表現を評価対象とし，評価指標にマイクロ F 値を用いて評価した．評価対象となる固有表現は付録 A の表 5 に示す．

表 1 の実験結果から，提案モデルは他の比較手法よりも低い F 値を示した．この結果から，本手法により OSPD 仮説を考慮することで固有表現抽出の性能が向上するとはいえないことがわかった．

5　解析

提案モデルが OSPD 仮説を考慮して固有表現抽出を行えているかどうかを確認し，固有表現抽出に与える詳細な影響を調べるため，開発データを用いて解析を行った．

5.1　モデルが予測する固有表現のタイプの一貫性の比較

モデルが予測する固有表現のうち，各文書中で同じ表層で複数回登場する固有表現を一貫して一つのタイプで予測しているかどうかを確認した．各表層で固有表現のタイプが揃っていれば 1，そうでなければ 0 とし，表層の数で平均を取ることで，一貫性を計算した．比較手法は 4.1 節のものと同じである．結果を表 2 に示す．

提案モデルが予測する固有表現のタイプの一貫性はベースラインモデルのものと比べて向上していることがわかった．後処理により訂正を行う手法に比べ一貫性は劣るが，モデル内で OSPD 仮説を考慮して固有表現のタイプを予測していることを確認した．

5.2　同じ表層で複数回登場する固有表現のみに対する性能の比較

正解データ中の，各文書中に含まれる同じ表層で複数回登場する固有表現を対象に以下の 3 つの基準で性能を評価した．結果を表 3 に示す．

- スパンとタイプを同時に正しく予測できたか．
- スパンのみを正しく予測できたか．

表 3 表層が同じで複数回登場する固有表現に対する 3 つ基準における各モデルの正解率 [%]

比較手法	スパンとタイプ同時	スパンのみ	スパンが正しいときのタイプ
ベースラインモデル	83.4±0.7	89.0±1.0	93.7±0.8
ベースラインモデル+訂正	83.8±1.2	89.0±1.0	94.1±1.3
提案モデル	83.0±0.7	88.0±0.7	94.3±0.9

- スパンを正しく予測できたとき，そのタイプを予測できたか．

比較手法と比べ，提案モデルはスパンとタイプを同時に予測した場合やスパンのみを予測した場合の正解率は低い値を示した．一方で，スパンを正しく予測できていたとき，そのタイプを予測する場合には比較手法よりも高い正解率を示した．ここから，OSPD 仮説の影響がある固有表現に対して，スパンとタイプを同時に予測した場合に提案モデルの正解率は低くなるが，その要因はスパンを正しく予測できていないことであるとわかる．

5.3 事例研究

ベースラインモデルと提案モデルの出力を比較することで予測性能の変化を事例レベルで確認した．「ACP 発現，nitroglycerin 奏効より狭心症を疑う．」「ACP 発現，これは nitroglycerin 内服で用意に消失す．」の 2 文を含む文書に含まれる「nitroglycerin」はどちらも薬品名の固有表現である．ベースラインモデルは後者中のタイプを正しく予測し，前者中のタイプを処置と予測した．一方，提案モデルは両者中で正しく予測した．前者の文に含まれる「奏功」は治療効果が現れることを意味するため，「nitroglycerin」のタイプに薬品名，処置の 2 つの曖昧性が生じていたと考えられる．一方，後者の文に含まれる内服は薬を飲むことを意味するため，「nitroglycerin」は薬品名である可能性が高い．提案モデルは「nitroglycerin」をフレーズとして扱っており，表層の同一性を利用して曖昧性を軽減したことで正しく予測できたと考えられる．提案モデルによって正しく予測されるようになった事例を 5 個確認した．

一方で，間違った予測がなされるようになった事例も存在した．「低栄養状態を伴う直腸癌患者に NST が介入し，栄養状態が著しく回復した一例」「10 病日に NST 介入となった．」の 2 文を含む文書に含まれる「NST(Nutrition Support Team)」において，固有表現であるのは後者中の「NST 介入」のみであり，前者中の「NST」は固有表現ではない．ベースラインモデルがこれらを正しく予測できたのに対して，提案モデルは前者中の「NST」と後者中の「NST 介入」を固有表現として予測した．提案モデルは「NST」をフレーズとして扱っており，固有表現である「NST 介入」中の「NST」を参照することで，これを前者中でも固有表現として認識したと考えられる．提案モデルによって間違った予測がなされるようになった事例を 3 個確認した．このような事例を抑制するためには，フレーズ抽出の対象となる文字列のレベルを変更し，「NST」と「NST 介入」をそれぞれ別のフレーズとして扱えるようにすることが考えられる．

6 おわりに

本稿では，BERT の内部で文書中の表層が同じ文字列について，その同一性を表現する機構を取り入れた，One Sense Per Discourse 仮説を考慮した深層固有表現抽出モデルを提案した．日本語医療文書データセット Real-MedNLP において提案手法における固有表現抽出の性能は向上しなかったが，解析によりモデルが予測する固有表現のタイプの一貫性が向上していることや，表層が同じ固有表現のスパンを正しく認識できた場合にそのタイプの正解率が向上したことを確認した．

現在の課題として，5.2 節での結果から，固有表現のスパンを正しく予測する性能が低いことが挙げられる．その要因として，5.3 節で確認されたように，本来は別々に扱うべきフレーズを同様に扱っていることが挙げられる．5.2 節の結果から，固有表現のスパンを正しく認識することができれば，OSPD 仮説を考慮してそのタイプをより正確に予測できることが確認できているため，今後はフレーズ抽出手法を改善することで，OSPD 仮説を考慮した深層固有表現抽出による固有表現抽出への影響を明確したいと考えている．

参考文献

[1] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In **Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003**, pp. 142–147, 2003.

[2] Zhaohui Yan, Zixia Jia, and Kewei Tu. An empirical study of pipeline vs. joint approaches to entity and relation extraction. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang, editors, **Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**, pp. 437–443, Online only, November 2022. Association for Computational Linguistics.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[4] Zihan Liu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. Ner-bert: A pre-trained model for low-resource entity tagging, 2021.

[5] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Portuguese named entity recognition using bert-crf, 2020.

[6] William A. Gale, Kenneth W. Church, and David Yarowsky. One sense per discourse. In **Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992**, 1992.

[7] Ander Barrena, Eneko Agirre, Bernardo Cabaleiro, Anselmo Peñas, and Aitor Soroa. "one entity per discourse" and "one entity per collocation" improve named-entity disambiguation. In Junichi Tsujii and Jan Hajic, editors, **Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers**, pp. 2260–2269, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

[8] Marcos Garcia. Incorporating lexico-semantic heuristics into coreference resolution sieves for named entity recognition at document-level. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**, pp. 3357–3361, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[9] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In Regina Barzilay and Mark Johnson, editors, **Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing**, pp. 782–792, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[10] Yada, et al. Real-mednlp: Overview of real document-based medical natural language processing task. NTCIR 16, 2022.

[11] Nict bert. https://alaginrc.nict.go.jp/nict-bert/index.html.

A　データセット

データセットの統計を表 4, 5, 6 に示す．

表 4　データセット中の固有表現数

データ	固有表現数	複数回登場する固有表現数
訓練	4,078	378
開発	1,633	163
評価	2,766	296
合計	8,558	837

表 5　評価対象となる固有表現

タイプ	評価対象
病名・症状	○
臓器・部位	○
特徴・尺度	×
変化	×
時間表現	○
検査（検査名）	○
検査（検査項目）	○
検査（検査値）	○
薬品（薬品名）	○
薬品（薬品値）	○
処置	×
クリニカルコンテキスト	×
保留	×

表 6　訓練データ中の固有表現を構成する品詞の数

品詞	数
名詞	8,454
接尾辞	1,301
特殊	953
助詞	350
形容詞	287
接頭辞	227
動詞	226
副詞	39
助動詞	10
指示詞	7
連体詞	6
接尾辞	2
合計	11862

B　学習設定

実装には Python 3.7.11，Pytorch 1.8.0，Transformers 2.2.2, pytorch-crf 0.7.2 を用いた．ハードウェアは CPU に Intel(R) Core(TM) i7-5960X，GPU に GeForce GTX TITAN X を用いた．事前学習モデルは NICT_BERT-base_JapaneseWikipedia_32K_BPE[11]，形態素解析器は Juman 辞書を用いた MeCab を使用した．最適化手法には AdamW を設定し，BERT の学習率を 3×10^{-5}，CRF の学習率を 5×10^{-3} とした．バッチサイズを 4 に設定し学習した．CRF の推移スコアを編集し不可能なタグの推移を抑制した．実験は 5 回行った．