症例報告を対象とした個人情報匿名化のための病名の識別と評価

細谷 健登　北橋 竜雄
株式会社インテック
{hosoya_kento, kitahashi_tatsuo}@intec.co.jp

概要

本稿では, データ利活用を見据えて, 営業日報などのテキストデータから個人情報に相当する用語を識別, 除去することを目的に症例報告コーパスを用いて, 固有表現認識による病名識別精度の改善および実用性を重視した固有表現認識の評価方法について提案する. 固有表現認識を個人情報の匿名化に活用する際には, 精度が 100%とはならない以上, 人の手による確認作業が必要になるため, そのことを前提としてどれだけ目的の固有表現の存在を示せるかの評価指標を提案する. 症例報告のような専門的な内容のコーパスにおける固有表現認識の改善方法として, そのドメインコーパスのみを用いた分散表現のモデルを作成し, 固有表現認識の学習に加えることで精度を向上させる手法を提案する.

1　はじめに

近年様々な分野において, 営業日報やコールセンターの応対記録, 臨床データなどのテキストデータの利活用が進んでいる, しかし, 個人情報の取り扱いや情報漏えい防止に関する各種法令およびガイドラインが強化されていく中で, データの活用には個人情報や機微情報の匿名化が必要となっている. 構造化データの場合は個人情報の位置の特定は比較的容易のため, k-匿名化 [1] 等の確立された手法で自動的に行うことは可能である. 一方, テキストデータのような非構造化データの場合, どこに個人情報が埋め込まれているかわからないため, 匿名化するには手作業で行わなければならないが, その作業負荷は大きい. これがテキストデータの利活用が進まない一つの要因となっている. これらの作業の負荷軽減のために, 個人情報の自動マスキングというタスクのコンペティションが行われる [2] など, 固有表現認識の活用が求められている.

本研究では, 機微情報として病名に注目した固有表現認識を行う. 病名は非常にセンシティブな情報であるため, 機微情報として扱う必要性があるためである. 機微情報である病名を自動的に識別することにより, 匿名化の作業負荷軽減を図ることができる. しかし, 人名や地名などは固有表現認識のためのデータセットが充実し始めているが [3], 病名についてはそのようなデータセットが少なく, 一般的に使用することのできる固有表現認識のためのコーパスは確立されていない. 本研究においては, 病名の識別のため, 奈良先端科学技術大学院大学 ソーシャル・コンピューティング研究室（SOCIOCOM）より提供される MedTxt-CR: 症例報告 (Case Reports) コーパス [4] を用いて病名の固有表現認識モデルを構築し, さらにモデルの精度向上のための手法を提案する.

固有表現認識を個人情報の識別に活用する際に, 課題となる点として精度の問題がある. 個人情報の識別では限りなく 100%に近い精度が求められるが, 固有表現認識の精度を 100%にすることは困難であり, 実際に固有表現認識を活用する場合には, 人の目で確認するという作業が必要となる [5]. 一般的に使用される固有表現認識の精度評価では, 固有表現のアノテーションを正確に予測できた場合のみが正解となり, その正解数に応じて精度の計算を行う. しかし, 個人情報を匿名化する場合は, 予測したアノテーション位置が完全である必要性は高くない. 予測したい部分の一部でもアノテーションされていれば, 手作業で修正することは容易であるためである. そのため本稿では, より実用性を重視し, どれだけ目的の固有表現の存在を示せるかについての固有表現認識の評価方法を提案する.

2　関連研究

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

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識別し, それらのラベルを予測する自然言語処理技術である. 固有表現とは人名, 地名, 団体名などの固有名詞や, 時刻や数量などのあらかじめ定義された表現の総称である. 文書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の応用において重要となる.

固有表現認識の多くの研究は系列ラベリングによる手法に基づいており, 近年では BERT[6] を用いた事前学習によるニューラル言語モデルを用いた BIO タグ付けによる手法において高い精度が確認されている.

2.2 MedTxt-CR: 症例報告 (Case Reports) コーパス

MedTxt-CR: 症例報告 (Case Reports) コーパスは, J-Stage でオープンアクセス公開されている症例報告論文 PDF から OCR 抽出したテキストのコーパスである [4]. OCR エラーによる非文を削除した OCR 抽出テキスト全文書 (3148 件), および単語レベルの OCR エラー修正・NER アノテーション [7] 済みの頻度バランスサブセット (224 件) が提供されているおり, NER アノテーションは病名/症状, 臓器/部位, 特徴/尺度, 変化, 時間表現など, 症例にかかわる様々なアノテーションがされている. 本研究では, 特に病名の識別に注目するため, NER アノテーション済みのテキストから病名/症状のアノテーションのみを残し, その他のアノテーションを除去したものを用いる.

3 実験

ニューラル固有表現認識の精度は文字や単語の埋め込みの手法に影響される. 本実験では単語埋め込みの手法を変化させ, 固有表現認識モデルを作成する. また, 本研究における, 固有表現認識モデルには, 固有表現認識タスクにおいて高い精度が報告されている, Bi-LSTM + CRF[8] を用いる.

3.1 ベースラインモデル

ベースラインモデルとして, 埋め込み表現に文字ベースの言語モデルを使って単語分散表現を構築する Akbik らの手法 (Flair embedding)[9] を用いる. また比較のため, 一般的に高い精度が出るとされる Flair embedding と BERT の分散表現を組み合わせたものでも実験を行う. ここで, BERT の事前学習モデルとしては情報通信研究機構 データ駆動知能システム研究センターより公開されている NICT BERT 日本語 Pre-trained モデル BPE あり [10] を使用する.

3.2 提案手法: ドメインコーパスのみを用いた分散表現の作成

BERT 等の事前学習モデルを用いる手法の場合, 事前学習モデルの学習には大量のデータが必要になるため, 基本的に Wikipedia など, 一般的に見られるテキストを使用して学習が行われている. しかし, 症例報告のようなあるドメインに特化したコーパスを用いる場合, そのドメイン内でのみ見られる単語が多く使用されており, 事前学習モデルでは補い切れないという問題がある. 本研究ではその問題の解決のために, 入力する単語埋め込みに事前学習モデルの分散表現に加え, ドメインコーパスのみを用いて学習した分散表現を加えることで精度の向上を図る. ドメインコーパスにおける分散表現の学習には Word2Vec[11] を用いる. 本実験では, 下記の通りに入力する単語埋め込みを組み合わせたもので実験を行う.

- Flair embedding + Word2Vec
- BERT + Word2Vec
- Flair embedding + BERT + Word2Vec

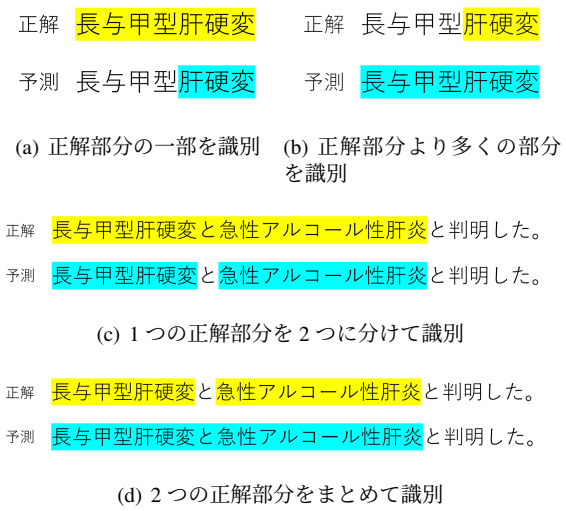
3.3 症例報告コーパスの追加アノテーション

学習データ量を増やすため, OCR 抽出テキスト全文書から NER アノテーション済みの頻度バランスサブセットに使用されていない文書を取り出し, 手動でアノテーションを行った. アノテーションの手がかりとして, SOCIOCOM で提供される万病辞書（MANBYO_202106）を用いた形態素解析結果 [12], および標準病名マスター病名検索結果 [13] を元に実施した. 今回追加でアノテーションしたのは 159 件である. ここでアノテーションしたコーパスを学習データに追加し, ベースモデルと同様の手法で学習を行った.

3.4 提案評価指標

本実験で作成したモデルの評価指標として, F1 値に加えて実用性を重視した評価指標 (正解指示率) を用いる. 正解指示率では, テストデータのアノテーションにおいて, 一部でもアノテーションがされている場合には正解としてカウントする. テス

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).



(a) 正解部分の一部を識別　(b) 正解部分より多くの部分を識別

(c) 1 つの正解部分を 2 つに分けて識別

(d) 2 つの正解部分をまとめて識別

図 1　正解指示率において正解とするアノテーション例

トデータ内の正解アノテーション数を N_c, 予測したアノテーション数を N_p とし, 正解アノテーションを $s_{ci}(b_{ci}, e_{ci})$ $(i = 1, 2, \ldots, N_c)$, 予測したアノテーションを $s_{pj}(b_{pj}, e_{pj})$ $(j = 1, 2, \ldots, N_p)$ と表すとする. ここで, b, e はそれぞれアノテーションの開始位置, 終了位置である. このとき, 正解指示率 Score は

$$\text{Score} = \frac{\sum_{i=1}^{N_c} f(s_{ci}, \boldsymbol{S}_p)}{N_c}, \tag{1}$$

$$\boldsymbol{S}_p = \{s_{p1}, s_{p2}, \ldots, s_{pN_p}\} \tag{2}$$

$$f(s_{ci}, \boldsymbol{S}_p) = \begin{cases} 1 & \exists s_{pj}(b_{pj}, e_{pj}) \in S_p \\ & : (b_{ci} \leq b_{pj} \leq e_{ci}) \cup (b_{ci} \leq e_{pj} \leq e_{ci}), \\ 0 & \text{otherwize} \end{cases} \tag{3}$$

と表される.

図 1 に正解指示率で正解とするアノテーションの例を示す.

4　結果

表 1 に固有表現認識の結果を示す. ※を付与したものが提案手法である.

ベースラインの手法と比較すると埋め込みの手法を複数組み合わせた場合は精度が高くなることがわかる. F1 値について比較すると, 提案手法である Flair+W2V, BERT+W2V の結果は, Flair+BERT には劣るものの, ベースラインを上回る精度となり, Flair+BERT+W2V と 3 つを組み合わせた場合が最も高いという結果になった. また, 正解指示率について比較すると, Flair+BERT, および BERT+W2V が同等の値で最も高いという結果となった.

表 1　実験結果

Embedding	F1-score	正解指示率
Flair(baseline)	0.7285	0.8378
Flair+BERT	0.7463	0.8867
Flair+W2V ※	0.7436	0.8587
BERT+W2V ※	0.7445	0.8867
Flair+BERT+W2V ※	0.7646	0.8853
Flair(データ追加)	0.6924	0.8210

新たにアノテーションをしたデータを追加した場合の結果については, F1 値および正解指示率ともに減少するという結果となった. これは, 提供されている NER アノテーション済みの頻度バランスサブセットと新規でアノテーションしたデータでアノテーションの規準が統一されていなかったためであると考えられる.

5　おわりに

本稿では症例報告コーパスにおける病名の識別モデルの構築と実用のための固有表現認識の評価指標を提案した. 症例報告コーパスでの病名の識別では, 単語埋め込みにドメインコーパスより学習した分散表現を追加することで, 固有表現認識の精度を向上させる手法を提案した. 固有表現認識の評価において正解指示率を用いることで, どの程度目的の固有表現の存在を示すことができるのかの指標となるため, 匿名化に必要なタスクについても固有表現認識モデルの有用度を示すことができる.

個人情報の匿名化について実用化を意識した場合の課題として, まず提案評価指標においての精度を 100%に近づける必要がある. そのための今後の展望として, モデルの学習方法工夫だけでなく, 辞書やルールベースの手法を組み合わせることでより精度を向上させることが挙げられる. また, データの不足を補うために, 同基準によるアノテーションデータを拡充するための作業フローの整備をしていく必要があると考えられる.

謝辞

本研究を行うにあたり, 症例報告コーパスのデータを提供いただいた奈良先端科学技術大学院大学ソーシャル・コンピューティング研究室の皆様に心より感謝いたします.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] Pierangela Samarati and Latanya Sweeney. Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression. Technical report, Computer Science Laboratory, SRI International, 1998.

[2] Nishika 株式会社. 判例の個人情報の自動マスキング　コンペ振り返り, 2021. https://note.com/nishika_inc/n/n78447a423abe#N58Rd.

[3] 近江崇宏. Wikipedia を用いた日本語の固有表現抽出のデータセットの構築. 言語処理学会 第 27 回年次大会, 2021.

[4] MedTxt-CR: 症例報告 (Case Reports) コーパス, 2020. https://sociocom.naist.jp/medtxt/cr/.

[5] 荒牧英治, 奥村学. 医療言語処理. コロナ社, 2017.

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

[7] Shuntaro Yada, Ayami Joh, Ribeka Tanaka, Fei Cheng, Eiji Aramaki, and Sadao Kurohashi. Towards a Versatile Medical-Annotation Guideline Feasible Without Heavy Medical Knowledge: Starting From Critical Lung Diseases. **Proceedings of The 12th Language Resources and Evaluation Conference**, pp. 4567–4574, 2020.

[8] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF Models for Sequence Tagging, 2015.

[9] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual String Embeddings for Sequence Labeling. **Proceedings of the 27th International Conference on Computational Lingistics**, pp. 1638–1649, 2018.

[10] NICT BERT 日本語 Pre-trained モデル, 2020. https://alaginrc.nict.go.jp/nict-bert/.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space, 2013.

[12] 万病辞書 Manbyo-Dictionary, 2021. https://sociocom.naist.jp/manbyou-dic/.

[13] 標準病名マスター病名検索, 2021. http://www.byomei.org/Scripts/Search/index_search.asp.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).