一般口演 | 知識工学
一般口演9
自然言語処理・テキストマイニング
2021年11月20日(土) 09:10 〜 11:10 E会場 (2号館2階222+223)

[3-E-1-06] 医学文書を対象とした疾患症状関係抽出におけるゼロ照応の評価

*和田 聖哉[1,2]、武田 理宏[2]、岡田 佳築[1,2]、真鍋 史朗[2]、小西 正三[2]、松村 泰志[3,2]（1. 大阪大学大学院医学系研究科　変革的医療情報システム開発学寄附講座, 2. 大阪大学大学院医学系研究科　医療情報学, 3. 国立病院機構大阪医療センター）
*Shoya Wada[1,2], Toshihiro Takeda[2], Katsuki Okada[1,2], Shirou Manabe[2], Shozo Konishi[2], Yasushi Matsumura[3,2]（1. 大阪大学大学院医学系研究科　変革的医療情報システム開発学寄附講座, 2. 大阪大学大学院医学系研究科　医療情報学, 3. 国立病院機構大阪医療センター）

キーワード：Natural Language Processing, Relation Extraction, Zero Anaphora

【背景】深層学習がもたらした自然言語処理の精度向上により，医学文書から自動的に情報を抽出する技術の実現が期待されている．しかしながら，疾患とそれに起因する症状関係を抽出する課題を考えた際に，従来の関係抽出フレームワークでは解決し難い問題が日本語では頻繁に出現する．それは，項の省略，代名詞や指示詞での言い換えを行う「照応」という現象であり，これを解決するには，文単位を超えて関係抽出を行う必要がある．今回，医学文書の自動疾患症状関係抽出器を構築することを目的として，省略された項の推定が必須となるゼロ照応がどの程度出現するのかを調査した．
【方法】日本語 Wikipediaにおいて ICD-10コードが与えられている疾患記事と MSDマニュアルプロフェッショナル版（日本語）の記事から，それぞれ30項目を抽出して対象データとした．疾患症状関係については，対象疾患に起因する症状であることが記載表現のみで特定出来るものと定義した．また，そのパタンについて，1) 同一文中関係，2) 同一段落関係，3)見出し関係と細分類し集計した（このうち，ゼロ照応に該当するのは2及び3）．
【結果】 Wikipedia文書の対象文は全2,363文で，そのうち同一文中関係，同一段落関係，見出し関係はそれぞれ32，57，150文存在した． MSDマニュアルプロフェッショナル版の対象文は全2,757文で，疾患症状関係の細分類はそれぞれ40，58，190文であった．
【結語】日本語医学参考書として， Wikipediaと MSDマニュアルプロフェッショナル版を対象に疾患症状関係の記載パタンを調査した．見出しに記載されている疾患がゼロ照応の対象となっている表現が最も多く，書式も考慮して関係抽出モデルへの入力をデザインする必要がある．

医学文書を対象とした疾患症状関係抽出におけるゼロ照応の評価

和田 聖哉[*1、*2]、武田 理宏[*2]、岡田 佳築[*1、*2]、真鍋 史朗[*2]、小西 正三[*2]、松村 泰志[*3、*2]
*1 大阪大学大学院医学系研究科 変革的医療情報システム開発学寄附講座、
*2 大阪大学大学院医学系研究科 医療情報学、
*3 国立病院機構大阪医療センター

Evaluation of Zero Anaphora in Relation Extraction for Disease-Symptom from Medical Documents

Shoya Wada[*1, *2], Toshihiro Takeda[*2] , Katsuki Okada[*1, *2], Shirou Manabe[*2], Shozo Konishi[*2], Yasushi Matsumura[*3]
*1 Department of Transformative System for Medical Information, Osaka University Graduate School of Medicine,
*2 Department of Medical Informatics, Osaka University Graduate School of Medicine,
*3 National Hospital Organization Osaka National Hospital

Abstract in English comes here.

We are now expecting to realize technologies for automatically extracting information from medical documents. However, in relation-extraction between a disease and the symptoms caused by the disease, it is necessary to extract the relationship while taking "anaphora" into account. In this study, to construct an automatic disease/symptom relation extractor for medical documents, we investigated the occurrence of zero anaphora, which requires the guess of omitted terms. We extracted 30 pages from each disease article with ICD-10 codes in Japanese Wikipedia and the articles in the MSD Manual Professional Edition (Japanese) as the target data. The MSD Manual Professional Edition contains 2,757 sentences, of which we classified 32, 57, and 150 sentences into the following categories: 1) intra-sentence relations, 2) same paragraph relations, and 3) heading relations. The total number of target sentences in the Professional Edition of the MSD Manual was 2,757, of which we classified 40, 58, and 190 sentences into each category. This study investigated the patterns of disease symptom-related subdivisions in Wikipedia and the MSD Manual Professional Edition as Japanese medical documents. The most common expression is that the disease in the headline is the target of zero anaphora; therefore, it is necessary to design the input to the relation extraction model considering the format.

Keywords: Natural Language Processing, Relation Extraction, Zero Anaphora.

1. 緒論

自然言語処理は、人間が日常的に使用する自然言語を計算機に処理させる一連の技術である。深層学習がもたらした自然言語処理技術の発展により、従来のタスクを高精度に解くことが出来るようになった。現在では大規模コーパスで自己教師あり学習による事前学習を行ってニューラル言語モデルを構築し、その事前学習済み言語モデルの重みを初期値として少数のターゲットタスクデータで重みを調整する Few-shot Learning が主流となった。また、文書生成モデルにおいては、多量のコーパスで巨大なパラメータ数を持つニューラル言語モデルの事前学習を行った GPT-3 により、重みの更新を行わずに、タスクの説明のみを入力とした Zero-shot Learning による予測にも成功している [1)]。

このような自然言語処理技術の進歩を背景に、医学領域でも医学文書から高精度に情報を抽出する技術の開発が期待されており、それは医学領域のコーパスで事前学習を行ったニューラル言語モデルの公開により実現しつつある [2)3)]。

しかしながら疾患とそれに起因する症状関係を抽出する課題を考えた際に、従前の関係抽出フレームワークでは解決し難い問題が日本語では頻繁に出現する。それは、文章の中で、項の省略、代名詞や指示詞での言い換えを行う「照応」という現象であり、これを解決するには、文単位を超えて関係抽出を行う必要がある。

2. 目的

医学文書から自動的に疾患・症状関係を抽出する分類器の構築を目的として、省略された項の推定が必須となるゼロ照応がどの程度出現するのかについて調査する。

3. 方法

3.1 対象

疾患に関して記述したページを取得するために、日本語 Wikipedia において ICD-10 コードが与えられている疾患記事と MSD マニュアルプロフェッショナル版（日本語）の記事から、それぞれ 30 項目を抽出して対象データとした。疾患・症状関係の認定には、対応する症状が対象疾患に起因するものであることが記載表現のみで特定できることを判定基準とした。

3.2 照応解析と関係抽出

照応解析とは、代名詞や指示詞などの照応詞の指示対象の推定や、省略された名詞句（ゼロ代名詞）を補完する処理のことを指す。日本語では主語の省略が頻繁に生じるため、照応解析を前処理に含めることがタスク精度の改善に繋がることもある。

関係抽出とは、2 つの特定のエンティティ間の関係を取得する処理のことを指す。一般的な共用タスクでは、2 つのエンティティは与えられた一文の中に存在することが多い。

本研究では、関係抽出を目的タスクとするときに照応、特に項の省略によって発生するゼロ照応の有無による自然言語処理の難易度を評価するために、関係抽出のパタンを以下の 3 つに分類して集計した。

A) 同一文中関係:
対象疾患とそれに対応する症状が単一の文中に出現する。本分類には従前の関係抽出タスクにおける

典型的な表現パタンが含まれる。

B) 同一段落関係:
同一文中関係には該当しないが、段落単位で確認すれば症状に対応する対象疾患名が記載されているものを本分類とする。

C) 見出し関係:
見出しもしくはページタイトルに対象疾患名が出現する場合を、見出し関係と定義する。

上記分類について、一部の事例では排他的とならずに複数の分類が与えられることがある。AとBは排他関係のためこの 2 つが同時に同一事例に与えられることはない。しかし C の見出し関係は A と B のどちらにも重複して分類される可能性がある。例えばページタイトルで疾患名が提示され、単一文中に対象疾患名と対応する症状が記載されている場合には、同一文中関係かつ見出し関係ありと分類出来る。そのように分類されない事例としては、鑑別診断について、重要所見となる症状に関する記載が挙げられる。この場合では、文中に出現する疾患名とページタイトルや見出しに存在する疾患名は異なる場合がある。このときには見出し関係とは分類されずに、同一文中関係もしくは同一段落関係のどちらかが単独で与えられる。

なお、ゼロ照応は同一段落関係と、見出し関係にのみ出現しうる。同一段落関係には照応詞で対象疾患を指示する事例も存在するが、見出し関係では照応詞による指示は含まれない。すなわち、見出し関係に分類したものは全てゼロ照応となる。

4. 結果

Wikipedia 文書の対象文は全 2,363 文、MSD マニュアルプロフェッショナル版の対象文は全 2,757 文であった。疾患症状関係の分類を表 1 に示す。2 種類の文書のいずれにおいても見出し関係が最も多く出現した。

表 1　疾患症状関係の分類

	Wikipedia 文書	MSD マニュアル Professional
同一文中関係	32	40
同一段落関係	57	58
見出し関係	150	190

5. 考察

医学文書を対象に、対応する症状表現の存在する文章で疾患名の照応評価を行った。文書種の違いで照応の出現頻度に差があるかどうかを検証するために 2 種類の異なる由来を持つ医学文書で確認したところ、どちらの文書種でも見出し関係に分類されるゼロ照応が最も多く出現した。

一般的に、省略されている項(先行詞)を述語と同一文外にある項から探して当てる文間ゼロ照応解析は、既存の自然言語処理の中でも困難なタスクとして扱われている。その理由は、先行詞同定のために文書全体を探索する必要があるからである。また、文脈次第では先行詞の推定に常識知が必要になる場合がある。現行の自然言語処理技術では、大規模言語モデルを用いても常識知の獲得は難しい。しかしながら、医学文書においては、症状表現が出現する場合で疾患名に対応するガ格が省略されている場合でも、ある程度パタンを特定できる可能性がある。例えばページタイトルや見出しにその症状に対応する疾患名が存在し、それを先行詞として扱うことが出来る場合である。中には一般的な自然言語処理と同様に直近の文との間でゼロ照応解析が必要な場合もあると考えられるが、その場合でも常識知は不要と思われる。つまり、その文書はそもそも医学知識を伝えるために書かれた文書であり、常識知を前提とする省略は行われる可能性は低いという仮定が成立する可能性が高い。このことは関係抽出タスクの設計段階で入力する書式を工夫すれば、ゼロ照応の存在する関係抽出であっても、一般的な自然言語処理フレームワークで解くことの出来る問題にまで落とし込むことが可能であることを示唆する。

自然言語処理における関係抽出タスクは基本的に、2 つの対象エンティティが与えられた状態で解くのが通例である。深層学習を用いた関係抽出モデルでは、対象エンティティをダミー文字列に置換、もしくは特殊 token を対象エンティティの前後に挟んで強調して入力文として扱い、目的の関係かどうかを判定する分類器を構築するアプローチが取られることが多い。本研究では同一文中関係であれば入力文は比較的短くなる。またその中に 2 つの対象エンティティが含まれているため、本アプローチで解くことが出来る。しかしながら医学文書における疾患症状関係において、同一文中関係は最も少なかった。

同一段落関係でも、長い入力を許容する関係抽出モデルを設計すればこのまま対応可能である。しかし、最近のニューラル言語モデルでは Transformer 構造 [4]がベースにあり、入力長が長くなるほど指数的にパラメータ数が増加する特徴がある。そのため、2 つのエンティティ距離が離れすぎる際には計算資源の問題で分類モデルの構築が難しい可能性がある。

その一方で見出し関係に分類される場合には、入力文の長さは単一文中関係よりは長くなるものの、同一段落関係よりは短くて済む。ただし、この入力を設計する際には、本文で記載されていた文章と、見出しの文字列を何らかの方法で区別して扱う必要がある。見出しに出現する疾患名がそのままゼロ照応の先行詞になっており、その出現位置の情報を保持することがそのまま関係抽出の判断材料として有利に働くかもしれない。

そこで、文書の書式を保ちつつ、かつ互換性のある表現を検討する。最近の自然言語処理においては、分量や利用しやすさの観点から Wikipedia が選ばれることが多い。その中でも、Wiki-40b[5]は前処理済みのデータセットとして利便性が高い。Wiki-40b は 40 言語以上の Wikipedia を前処理して作られたデータセットであり、言語毎にダウンロードして利用することも可能である。前処理として、不要なページのフィルタリング(曖昧さ回避ページ、リダイレクトページ、削除済みページ)とページ内処理(マークアップやページ内非コンテンツ部分(参考文献、外部リンク、脚注、画像キャプション、リスト、テーブル等)の除去)が適用済みの、質の高いデータセットとなっている。文書の構造を表現するために、以下の 4 つのマークアップが定義されている。

- ✓ _START_ARTICLE_
  記事の始まりを表す。
  この後にページタイトルが続く。
- ✓ _START_SECTION_
  節の始まりを表す。
  この後に節のタイトルが続く。
- ✓ _START_PARAGRAPH_
  節のタイトルと節内のパラグラフの間に設置される。
- ✓ _NEWLINE_
  1 パラグラフの終わりを表す。

このマークアップを用いた表記例を図 1a に示す。また、このマークアップを用いて関係抽出を行う際の入力文を図 1b で提示する。このように、パラグラフに対応する記載を 1 文ずつ分割して入力文を構築することで、書式の構造を保ちつつ入力文を短くすることが出来る。

次に、このマークアップをニューラル言語モデルに適用する手法について考察する。ニューラル言語モデルではまず入力文に対し、その言語モデルが持つ vocabulary に従ってトークナイゼーションを行う。Byte Pair Encoding (BPE)[6]によるサブワード分割を適用するニューラル言語モデル（BERT[7]、ELECTRA[8]など）では、新規表現を vocabulary に追加すること自体は比較的容易に出来る。上記 4 つのマークアップを登録すれば特殊 token として使用できるようになり、簡便ではあるが書式情報も含めてニューラル言語モデル内で表現可能になる。Sentencepiece を用いてサブワード分割を行う言語モデル（RoBERTa[9]、T5[10]など）では、言語モデルベースの分割を行うため、新規表現を後から追加することは難しい。しかし事前学習に Wiki-40b から取得したコーパスを用いているのであれば、マークアップ記号自体も SentencePiece で上手く扱えるよう学習済みである可能性がある。以上から、マークアップを使用した入力文を用いる際には、事前学習に用いたコーパスと、サブワード分割の手法を確認した上で、適切なニューラル言語モデルを選択する必要があると言える。

6. 結論

日本語医学文書として，Wikipedia と MSD マニュアルプロフェッショナル版を対象に疾患症状関係の記載パタンを調査した．見出しに記載されている疾患がゼロ照応の対象となっている表現が最も多く出現することが明らかとなったものの、マークアップを用いて文書レイアウトを保持しつつ学習データの整備を行えば、疾患関係関係抽出モデルが構築できる可能性がある。

参考文献

1) Brown TB, Mann B, Ryder N, et al.. Language Models are Few-Shot Learners. arXiv preprint 2020; arXiv:2005.14165.
2) Lee J, Yoon W, Kim D, et al.. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020; 36(4): 1234-1240.
3) Gururangan S, Marasović A, Swayamdipta S, et al.. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics 2020; 8342-8360.
4) Vaswani A, Shazeer N, Parmar N, et al.. Attention is all you need. In Advances in neural information processing systems 2017: 5998-6008.
5) Guo M, Dai Z, Vrandečić D, and Al-Rfou R. Wiki-40b: Multilingual language model dataset. In Proceedings of The 12th Language Resources and Evaluation Conference 2020; 2440-2452.
6) Sennrich R, Haddow B, Birch A. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics 2016; Volume 1: Long Papers.
7) Devlin J, Chang MW, Lee K, and Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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.
8) Clark K, Luong MT, Le QV, and Manning CD. Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint 2020; arXiv:2003.10555.
9) Liu Y, Ott M, Goyal N, et al.. Roberta: A robustly optimized bert pretraining approach. arXiv preprint 2019; arXiv:1907.11692.
10) Raffel C, Shazeer N, Roberts A, et al.. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research 2020; 21: 1-67.

a. マークアップを用いた医学文書の表記例

```
_START_ARTICLE_
巨細胞性動脈炎
_START_SECTION_
症状と徴候
_START_PARAGRAPH_
症状は数週間にわたって徐々に発現することもあれば突然現れることもある。発熱(通常は微熱)，疲労，倦怠感，説明のつかない体重減少，発汗などの全身症状を呈することがある。…
_NEWLINE_
```

b. 関係抽出時の入力文

```
_START_ARTICLE_
巨細胞性動脈炎
_START_SECTION_
症状と徴候
_START_PARAGRAPH_
症状は数週間にわたって徐々に発現することもあれば突然現れることもある。
_NEWLINE_
        ------------------------------------------------------------
_START_ARTICLE_
巨細胞性動脈炎
_START_SECTION_
症状と徴候
_START_PARAGRAPH_
発熱(通常は微熱)，疲労，倦怠感，説明のつかない体重減少，発汗などの全身症状を呈することがある。
_NEWLINE_
```

図 1　マークアップによる医学文書の書式表現