医療テキストに対する網羅的な所見アノテーションのためのアノテーション基準の構築

篠原 恵美子[*1], 河添 悦昌[*1], 柴田 大作[*2], 嶋本 公徳[*1], 関 倫久[*1]
[*1] 東京大学, [*2] 奈良先端科学技術大学院大学

Development of finding-comprehensive Annotation Guideline for practical clinical text processing

Emiko Shinohara[*1], Yoshimasa Kawazoe[*1], Daisaku Shibata[*2]
Kiminori Shimamoto[*1], Tomohisa Seki[*1]
[*1] The University of Tokyo, [*2] Nara Institute of Science and Technology

抄録：診療録においては所見など自由記載テキストにのみ記録される重要な情報が存在し、これを自動抽出する技術が求められている。実用的な技術開発の促進のためには、記載されている情報を網羅するようなアノテーションが付いたコーパスの公開が必要である。現在、我々はそのようなアノテーション付きの症例報告コーパスを構築し公開の準備を進めている。本稿ではそのアノテーション基準の制定について報告する。症例報告へのアノテーションとアノテーション基準の修正を繰り返した結果、49 種のタグと 35 種の関係から成るアノテーション基準が得られ、先行研究よりも詳細な情報が表現できるだけでなく、これまで表現できなかった事実性の時間的変化を捉えられるようになった。

キーワード　自然言語処理、アノテーション、コーパス

1. はじめに

ICT や AI 技術を用いた電子カルテの利活用が期待されるなか、特に自由記載のテキスト中にのみ記録される情報を抽出する自然言語処理（NLP）技術が必要とされている。その研究開発には、テキストと、そこに含まれる個々の情報とその記載箇所のアノテーションから構成されるコーパスが必要であり、さらにコーパスが公開されることで研究開発が促進される。コーパス構築は高コストであるため、妊娠の有無判定など特定の応用に特化しない、汎用性のあるコーパスが有用である。また、実用的な技術の実現のためにはテキストに含まれる情報を網羅するアノテーションが必要である。このようなコーパスは特に日本語ではほとんど存在しない。汎用性を志向した事例 [1]は存在するものの、網羅性については触れられていない。我々は汎用的かつ網羅的なアノテーション付与を目指し、これまでに 300 以上の症例報告テキストに対して主に所見に焦点を当てたアノテーションを行っており、公開予定である。本稿ではこのコーパスのアノテーション基準の構築について述べる。

2. 方法

1) 材料

対象文書としては、退院時要約に近く、公開のハードルが比較的低い臨床医学系雑誌の症例報告を用いることにした。

症例報告はタイトルに厚生労働省の指定難病名と「例」を両方含み、2000 年以降に出版されたものを JSTAGE で検索し、本文が公開されているものから 1 疾患あたり最大 4 件について症例記載部分をコピー・ペーストしてテキストデータに変換した。指定難病を使ったのは診療科や疾患領域に限定が無く、幅広い症状・所見が記載されていると考えたためである。

2) アノテーション基準の構築

アノテーション基準は、最初に仮の基準を作成し、その後アノテーション実施と基準の修正を繰り返すことで行った。

アノテーション内容は文字列範囲に対するタグとその属性、およびタグ付けされた文字列範囲間の関係から成るものとする。アノテーション

した情報への UMLS 概念コード付与も行っているが、本稿では扱わない。

基準の方針としては、テキスト中の情報をできるだけ漏らさず表現可能とすることとした。また、外部用語集へのマッピングにおいて可能な限り細かい粒度でのマッピングを行えるようにするため、タグを付与する範囲は細かくすることとした。なお、アノテーションの対象とする内容は書き手が認識したことであり、真実であるとは限らない。

3. 結果

指定難病 333 疾患のうち 151 疾患について 362 症例報告を収集した。

構築したアノテーション基準は、タグ 49 種、関係 35 種から構成されている。図 1 にアノテーションの例を示す。タグは症状・所見を直接表すものだけでなく、人体部位や時間などさまざまであり、症状・所見はコアとなるタグ（以下、所見系タグ）から一定の規則で関係を辿ることで抽出される。所見に関わる他のタグや関係としては、臨床検査タグと観測手段関係、姿勢タグや行為タグ等と観測条件関係、時点タグや時区間タグと観測時関係・開始時関係等がある。また、所見が別の所見をより詳細に述べている場合には所見系タグ対は対象関係を持つ。さらに因果関係や判断根拠の関係がある。所見系のタグは、肯否や判断を表すタグと関係を持つことで、肯定・否定や疑いの情報を付与できる。

先行研究と比較して最も重要な差は、肯否タグ・判断タグを導入したことである。これと所見間の因果関係・根拠関係を併用することで、従来は表現できなかった、診療の過程で判断が変化するケースを表現できるようになった。ただし意味を注意深く定義する必要があった。

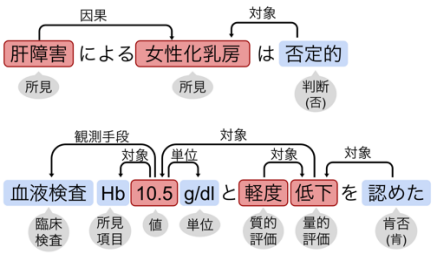

Fig 1. 構築した基準によるアノテーションの例
赤枠は所見系タグを示す。

4. 考察

構築したアノテーション基準は網羅性を志向したものであり、多くのタグ・関係が含まれる。NLP のタスクとしてはこのアノテーションを直接再現するほかにも、一部のみを対象とする・複数のタグから新しいタグを生成することで粒度を下げたものを再現する、といったものが考えられる。また、他の基準[1]等による粗く量の多いコーパスと、本研究の緻密で量の少ないコーパスを併用するような NLP の手法が有用と考えられる。

本研究では主に所見をアノテーション対象として扱ったが、実際のテキストでは治療との境界例も見られた。例えば「血糖コントロールを行った」は治療であるが、「血糖コントロール良好」はその結果としての患者状態である。このようなケースを汎用性を犠牲にすることなく、すなわち治療の意味を捨てることなく表現するためには、所見だけでなく他の種類の情報もアノテーション対象として包括的に扱うべきである。

本研究では症例報告を対象としたが、診療記録との相違点として、退院後の経過や剖検、同一報告中の他症例への言及があった。このような部分をあとから除外して利用できるようにアノテーションしておくことも有用と考えられる。

研究促進のためにはコーパスを公開することが重要である。現在、各発行団体に許諾を得るための手続きを取っている。

5. 結語

症例報告へのアノテーションを通じて、所見関連情報のアノテーション基準を構築した。今後は、基準の有用性の評価とコーパスの公開、診療記録への適用可能性の検討、治療等へのアノテーション基準の拡大を予定している。

参 考 文 献

[1] Shuntaro Yada, Ayami Joh, Ribeka Tanaka, et al: Towards a Versatile Medical-Annotation Guideline Feasible Without Heavy Medical Knowledge: Starting From Critical Lung Diseases. Proceedings of the 12th LREC, pp. 4565–4572, 2020.