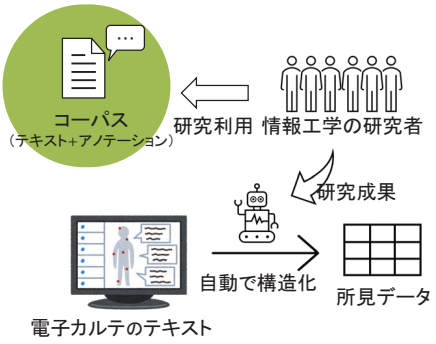
22世紀医療センター
22nd Century Medical and Research Center
医療AI開発学講座
Department of Artificial Intelligence in Healthcare

演題名：診療記録における症状・所見の網羅的なアノテーション基準の構築

演者名：篠原恵美子、河添悦昌

背景

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、特定の応用に依存しない、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。我々は主に所見に焦点を当てた、汎用的かつ網羅的なアノテーション付きの300以上の症例からなるコーパスを構築しており、公開予定である。



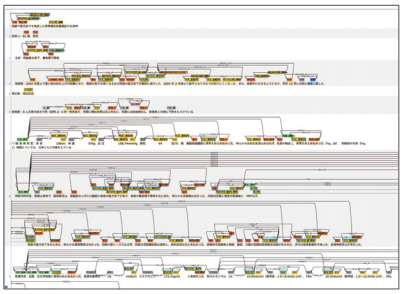
方法

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。

アノテーションは、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した。
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。

結果

指定難病333疾患のうち156疾患について358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34種、関係32種から構成されている。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。



アノテーションの例

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。自然言語処理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自然言語処理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。
今後、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、また機械学習等を用いた自動構造化手法の開発・精度評価を予定している。