

機械学習を用いた読影レポート解読手法の構築

Identification of clinical knowledge in free-text radiology reports using artificial intelligence

田中佑岳*1 橋本正弘*2 伊藤詩乃*1 奥田茂男*2 陣崎雅弘*2 榎原康文*1
 Yugaku Tanaka Masahiro Hashimoto Shino Ito Shigeo Okuda Masahiro Jinzaki Yasubumi Sakakibara

*1慶應義塾大学理工学部生命情報学科

Keio University Science and Technology Department of Biosciences and Informatics

*2慶應義塾大学医学部放射線科学教室

Keio University School of Medicine Department of Diagnostic Radiology

In the process of medical diagnosis, a radiologist examines clinical images of a patient and writes down a diagnostic report which is called a radiology report. To reduce the operations of radiologists, an automatic and precise extraction of information from free-text reports is demanded. In this study, we employed Named Entity Recognition analysis to extract essential elements from radiology reports. Six named entity tags, such as body-parts, features, conditions, filming-conditions, diagnoses, and conclusions, were defined. From the reports about MRI on pelvic cavity from Keio University Hospital, one hundred reports were randomly chosen and were manually annotated with named entity tags by a radiologist. Parameters for Support Vector Machine and features for machine learning were optimized. The supervised learning resulted in an F-value of 0.814. The F-value increased to 0.817 by the semi-supervised learning using 4900 unannotated reports. We utilized large amounts of data from medical domain and showed the effectiveness of semi-supervised learning in Named Entity Recognition analysis.

1. 背景

読影レポートは、放射線科の医師が作成する画像診断報告書である。膨大なレポートを作成する読影医の負担軽減を目的としたレポートの入力支援システムや要約の自動生成システムの開発が期待されている [Futami 10]。医師が自由に文章を記述するため特定の構造を持たないレポートから、正しく情報を抽出することがシステム開発に向けた第一の課題である。

本研究では、文章中に存在する特定の表現を抽出するため、自然言語処理技術である固有表現解析と機械学習を用いて読影レポートの解析を行った。固有表現タグとして、部位、所見、鑑別、特徴、撮影条件、結言の6種類用意し、文章中の単語の抽出と分類を行った。

2. 手法

慶應義塾大学病院より骨盤腔のMR画像についてのレポートを18,695件取得し、そのうち100件に正解タグを作成した。

まずレポートの文章を形態素解析し、各形態素の特徴選択を行ったのち、固有表現解析を行った。形態素解析とは文章を形態素と呼ばれる意味を持つ最小の単語に分割する解析であり、本研究ではMeCab[Kudo 02]を用いた。固有表現解析では*i*番目の固有表現タグを予測するために、赤枠で示したウィンドウ内にある情報を特徴量として使用した(図1)。特徴量となる要素は形態素、品詞、形態素の文字種、固有表現タグであり、それぞれについて予測精度を最大にするウィンドウを求めた。

機械学習手法の一つであるサポートベクターマシンによって学習と予測を行った。教師あり学習ではタグをつけたレポートについて機械学習パラメータと特徴量の最適化を行い、予測精度を求めた。半教師あり学習では、タグありデータに加えて正解タグのついていないレポートを使用した。まずタグありデータ

連絡先: 榎原康文, 慶應義塾大学理工学部, 横浜市港北区日吉3-14-1, 045-566-1791, yasu@bio.keio.ac.jp

	i-2	i-1	i	i+1	i+2		
形態素	...	DWI	で	は	強い	異常	...
品詞	...	一般名詞	格助詞	係助詞	形容詞	一般名詞	...
文字種	...	alphabet	かな	かな	漢字、かな	漢字	...
固有表現タグ	...				B-特徴	B-所見	...

← 予測方向

図1: ウィンドウと固有表現タグ予測

タから学習したモデルを使ってタグなしデータの固有表現を予測した。得られた予測結果をそのレポートの正解固有表現タグとして扱い、元の学習データとともに再び機械学習を行うことで新たなモデルを作成した。最後にテストデータを予測して精度を評価した。本研究では、画像から見て取れる状態を表す所見と、病名を表す鑑別についてはできるだけ大きな文節で区切るようにした。

精度がもっとも高かった予測結果について、正しく予測できなかった固有表現を分類した。領域不整合、タグ不一致、予測なし、その他の4つに分類した。領域不整合は、正解のうち一部は正しく予測できている固有表現を指す。タグ不一致は、固有表現は正しく抽出されたが、予測されたタグが正解と異なったものであり、予測なしとは、正解の固有表現にをどのタグとも予測しなかったものである。

3. 結果

3.1 教師あり学習

教師あり学習において機械学習のパラメータを最適化し、特徴量を最適化した。特徴量として用いる形態素は前後二つの形態素を使用する場合、すなわち*i*番目の形態素について*i-2*から*i+2*番目をウィンドウとして使用する場合に最も精度が高かった。同様に、品詞は*i-1*から*i*、文字種は*i-4*から*i+1*、固有表現タグは*i+1*を特徴量としたとき予測精度が最も高くなり、F値は0.814であった。

3.2 半教師あり学習

教師あり学習で最適化したウィンドウを特徴量として使用した場合の、タグなしデータの数に応じた予測精度を図2に示す。タグなしレポートを使用した場合のみに注目すると徐々に精度が上がり、レポートを4900件使用した半教師あり学習ではF値が0.817となり教師あり学習の精度を上回った。

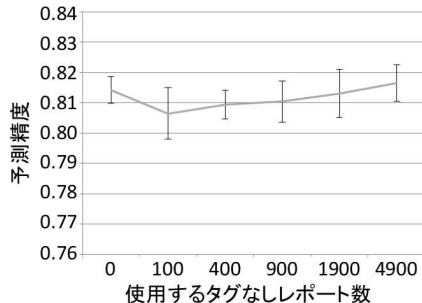


図2: 教師あり学習で最適化したウィンドウ

3.3 不正解の分類

4900件のタグなしレポートを使用した半教師あり学習についてタグごとの精度を図3に示す。

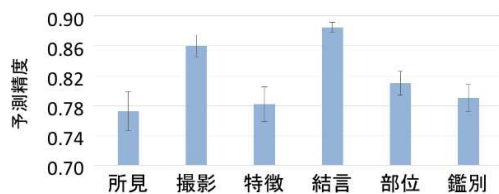


図3: 半教師あり学習でのタグごとの予測精度

正しく予測できなかった固有表現の総数は502であり、これらを分類した。

領域不整合は不正解全体の42%を占めた。例えば正解の固有表現が「変性した筋腫」という鑑別であるのに対して予測した固有表現は「変性した」という特徴の「筋腫」という鑑別となった。また、正解では「腫瘤」を所見、「腸骨部分外側」を部位とした箇所を「この腫瘤の腸骨部分外側」と一つの部位として予測した例もあった。

タグ不一致に分類されたのは22%であった。特に正解タグが「所見」であるものを「鑑別」と予測する場合、またその逆の場合が多かった。

予測なしに分類されたのは全体の13%であり、その他が23%であった。

4. 考察

本研究の予測精度を既存研究と比較した結果を図4に示す。それぞれ半教師あり学習の精度、教師あり学習の精度に加えて、既存ツールの特徴量を使用した結果、既存研究の結果である。既存ツールとは日本語固有表現解析ツール ChaSen[Yamada 02]で使用されているウィンドウを使用したものである。既存研究は2006年に安永らが読影レポートの固有表現抽出を行った論文で報告されている精度である[Yasunaga 06]。データセットが異なるため一概には比較できないが、本研究での精度が上回っている。

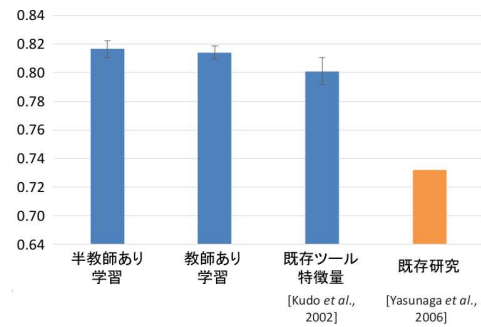


図4: 先行研究との比較

領域不整合とタグ不一致は手作業で作成した正解と異なったため不正解となったが、専門医によるタグ付けの場合でも文章の解釈によって正解が揺らぐことが指摘されている[Liu 05]。仮に所見と鑑別との間でタグ不一致に分類される固有表現を正解に含めると予測精度は0.885となった。それ以外の不正解の固有表現は、学習データに出現しないものが多かったため、正しく予測するために学習データを増やすことが必要と考えられる。

5. 結論と今後の展望

本研究では読影レポートの大規模な実データを使用して固有表現解析を行った。半教師あり学習を行うことで予測精度を向上させた。特に特徴量として用いる各要素のウィンドウを増やした場合の精度向上を確かめた。予測結果を詳細に分類し、タグの不一致や領域の不一致の場合の検証を行った。

今後は、機械学習による予測を手作業で修正することで正解データを増やし、学習の精度向上を図る。また、今回対象とした骨盤腔MRI以外のレポートの固有表現も抽出できるようにドメインの拡大を図る。

参考文献

- [Futami 10] 二見光, 山岸宏匡, 川口修, 塚本信宏, 藤井博史, 笠松智孝, 安藤裕, 長田雅和, 久保敦司. 構造化技術を用いた読影レポートの類似記載を特定する手法の開発. 日本放射線技術学会雑誌, Vol. 66, No. 9, pp. 1229-1236 (2010)
- [Kudo 02] 工藤拓, 山本薫, 松本裕治. Conditional Random Fields を用いた日本語形態素解析. 情報処理学会 研究報告, pp. 89-96 (2004)
- [Yamada 02] 山田寛康, 工藤拓, 松本裕治. Support Vector Machine を用いた日本語固有表現抽出. 情報処理学会論文誌 Vol. 43, No. 1, pp. 44-53 (2002)
- [Yasunaga 06] 安永晋, 川上洋一, 笹井浩介. Support Vector Machine (SVM) を用いた自然文読影レポートからの医学的知識の抽出. 医療情報学 (Japan journal of medical informatics), Vol. 25, No. 6, pp. 405-412 (2006)
- [Liu 05] Kaihong Liu, Kevin J Mitchell, Wendy W Chapman, and Rebecca S Crowley. Automating tissue bank annotation from pathology reports - comparison to a gold standard expert annotation set. AMIA Annual Symposium Proceedings, Vol. 2005, p. 460-464 (2005)