第７５回道南医学会大会医学研究奨励賞推薦演題

機械翻訳を利用した自然言語処理の試み−放射線画像診断レポートを素材として−

函館五稜郭病院　がんゲノム医療センター　○池田　健
同　メディカルインフォメーションセンター　坂本　勝・横山峰要
佐々木　愼・橋浦大希
村越翔太

【要旨】

機械翻訳の精度が著しく向上している状況を踏まえ、機械翻訳を利用した自然言語処理の手法を検討した。放射線画像読影レポート見逃し防止作業を対象とした。読影レポートを英訳し、自然言語処理に必要な前処理後に、機械学習モデルを作成した。このモデルにより感度 96%が得られ、スクリーニング検査として満足のいく結果と思われた。英訳することで定型的な自然言語処理技術を用いることができ、効率的かつ精度の高い予測モデルの構築が可能であった。医学用語の多い電子カルテは英訳になじみやすいので、本手法を他のテキストデータへと拡張することも容易と思われる。

【キーワード】：医療情報、機械翻訳、自然言語処理、機械学習

【はじめに】

プログラミング言語のように人工的に作られた言語を人工言語といい、我々が日常用いている自然発生的に生まれた言語を自然言語という[1)]。医療情報が持つ膨大なテキストデータの有効な利活用には、自然言語処理技術が必須である[2)]。一方、自然言語処理はおもに英語圏で発達してきた技術であり、英語と日本語の自然言語処理には異なる点が多く存在する[3)]。近年、機械翻訳の精度が著しく向上している状況を踏まえ、機械翻訳を介した自然言語処理の可能性を検討した。

本検討の素材として日本語で記載された放射線画像読影レポートを用いた。放射線画像診断を依頼した医師は、診断結果を確実に把握し患者の治療に活かす必要がある。読影レポート、特に悪性腫瘍に関連するレポートの見逃し防止のため、当院ではメディカルインフォメーションセンターのスタッフが要チェックレポートを抽出し、それをもとに依頼医へ注意を促している。この作業に割かれる人的・時間的資源は決して小さくない。今回の成果を利用することで要チェックレポートの自動抽出が可能になれば、大きな業務改善に繋がるものと期待される。

【方法】

テキストデータの処理；2021 年 2〜5 月の日本語で記載された放射線画像読影レポート 4,714 件を対象とした。これらのレポートにはスタッフによる要チェックか否かの「答え」が付加されている。それらの読影レポートを機械翻訳システム DeepL[4)]によって英語に翻訳した（図 1）。英訳データにステミング、小文字化、「特殊文字，ピリオド，数字」の除去、ストップワードの除去よりなる前処理を施した（図２）。ステミングとは、語幹化、複数の語形をまとめることで、例えば、read、reading などの語形を read に統一する操作である。また、ストップワードとは、冠詞の the や it など、ドキュメント内容に直接関係しない語をいう[5)]。これらの前処理によって得られた 4,942 単語より、出現頻度をもとに 577 単語を抽出した。これらを説明変数とし、スタッフによってラベリングされた「要チェック」の有無を目的変数とした。以上の自然言語処理には R パッケージ tm と同 SnowballC を用いた。

教師あり機械学習；対象レポート 4,714 件を教師データ 4,164 件とテストデータ 550 件に分割した。テストデータでの正解率（要チェックありの率）はリアルワールドデータと同じ 18%に設定し、訓練のための教師データの正解率は 40%に設定した。これは、訓練データの正解率を高めることで精度の高いモデル構築が可能になるためである。教師あり機械学習の手法として、ランダムフォレスト法を用いた（R パッケージ randomForest）。

【結果】

テストデータを用いたときの混同行列を示した（図

3)。この表より、感度96%、特異度89%、正確度91%が得られた。変数重要度は、figur (figure をステミングしたもの)、mass、increas (increase をステミングしたもの)、cancer の順であった (図4)。なお、figur、mass、increas、cancer は、それぞれ「図」、「腫瘤」、「増大 (あるいは増加)」、「癌」の英語訳に相当する。

【考察】

英訳データを使用する利点は、実績のある「英語用」の自然言語処理技術を適用でき、解析過程を単純化できることである。自然言語処理の基本的解析技術には、形態素解析、構文解析、文脈の解析、意味の解析などがある [1)]。今回用いた説明変数はレポート中の単語の出現パターンであり、技法としては文を単語に分割する形態素解析で十分であること、読影レポートは使用される単語が限られ定型文が多いことにより、翻訳によるデメリットは小さいと考えられた。

構築した機械学習モデルの精度は、感度96%、特異度89%、正確度91%であった。対象としたタスクは高い感度が求められるスクリーニング作業であったが、感度 96%は満足できる水準と考える。変数重要度は、日本語に直すと、図、腫瘤、増大、癌の順であった。機械学習モデルは、説明変数とした577 単語のうち、これら4 語など重要度上位の語を含むレポートを「要チェック」とラベルした。その結果として、感度96%という高い精度が得られた。目的が癌に関連する読影レポートの見逃防止であることを考えると、理解しやすい結果と思われる。

今後に残された課題がいくつかある。まず、本研究は機械翻訳により日本語の自然言語処理が容易になることを期待しているが、標準的な日本語自然言語処理との比較は行っていない。日本語を対象とした自然言語処理手法には、MeCab が知られている [3)]。構築モデルの精度などを、日本語を対象とした手法と比較することが望まれる。

また、機械翻訳にはコストが掛かることも知っておくべきである。DeepL の料金体系は、5 ファイル/月 (最大ファイルサイズ 10MB) で年間 12,000 円、20 ファイル/月 (最大ファイルサイズ 20MB) で年間 30,000 円、100 ファイル/月 (最大ファイルサイズ 20MB) で年間 75,000 円である [6)]。今回は年間 30,000 円のプランで実行した。DeepL は高性能で知られるが、機械翻訳にどの程度の性能を求めるのか検討の余地がある。低コストの機械翻訳ソフトも候補となるだろう。

偽陰性や偽陽性症例の詳細なレビュー、ランダムフォレスト法以外の手法、特に深層学習の利用など、検討すべき点が残っている。ただし、スタッフによるラベリングが必ずしも「正解」とは限らない。そもそも、セーフティネットとして行っている作業であり、人間あるいは機械、どちらがマーキングするにしても本作業に過度の精度を期待するべきではないと感じる。

【文献】

1) 黒橋禎夫, 改訂版 自然言語処理. 放送大学教育振興会, 東京, 2019
2) 荒牧英治, 医療言語処理, 自然言語処理シリーズ 12, コロナ社, 東京, 2017
3) Bird S, Klein E, Loper E. 入門自然言語処理. オライリー・ジャパン, 東京, 2010
4) DeepL. 2023, https://static.deepl.com/files/press/companyProfile_JA.pdf.
5) 石田基広, R によるテキストマイニング入門. 第 2 版, 森北出版, 東京, 2017
6) DeepL. 2023, https://www.deepl.com/ja/pro/change-plan#single.

本論文に関する著者の利益相反；なし

CT(2019/06/26)と比較しました。

肺気腫を認めます。
右肺上葉に18×16mm大の不整形結節を認めます。辺縁にspiculaを認め、胸膜陥入像(図2)を伴います。前回と比較して増大しており、肺癌が疑われます。
その腹側にも小結節(図3)を認めますが、前回と大きな変化ありません。

⇩

Compared with CT (06/26/2019).

Emphysema is present.
An irregular nodule of 18×16 mm in size is seen in the upper lobe of the right lung. There is a spicula at the margin with a pleural depression image (Figure 2). The nodule is larger than the previous one and is suspected to be lung cancer.
A small nodule (Fig. 3) is also seen on the ventral side of the spicula, but there is no significant change from the previous examination.

図1　DeepLによる英語への翻訳例

There is a spicula at the margin with a pleural depression (Figure 2).

⇩　ステミング（語幹抽出）

There is a spicula at the margin with a pleural depress (Figur 2).

⇩　小文字化

there is a spicula at the margin with a pleural depress (figur 2).

⇩　括弧等の特殊文字，ピリオド，数字などを除去

there is a spicula at the margin with a pleural depress figur

⇩　冠詞、前置詞、is などのストップワードを除去

spicula　margin　pleural　depress　figur

図2　形態素解析のための前処理

テストデータ

予測		要チェック	不要	
	要チェック	96	48	144
	不要	4	402	406
		100	450	550

図3　作成した機械学習モデルによる混同行列

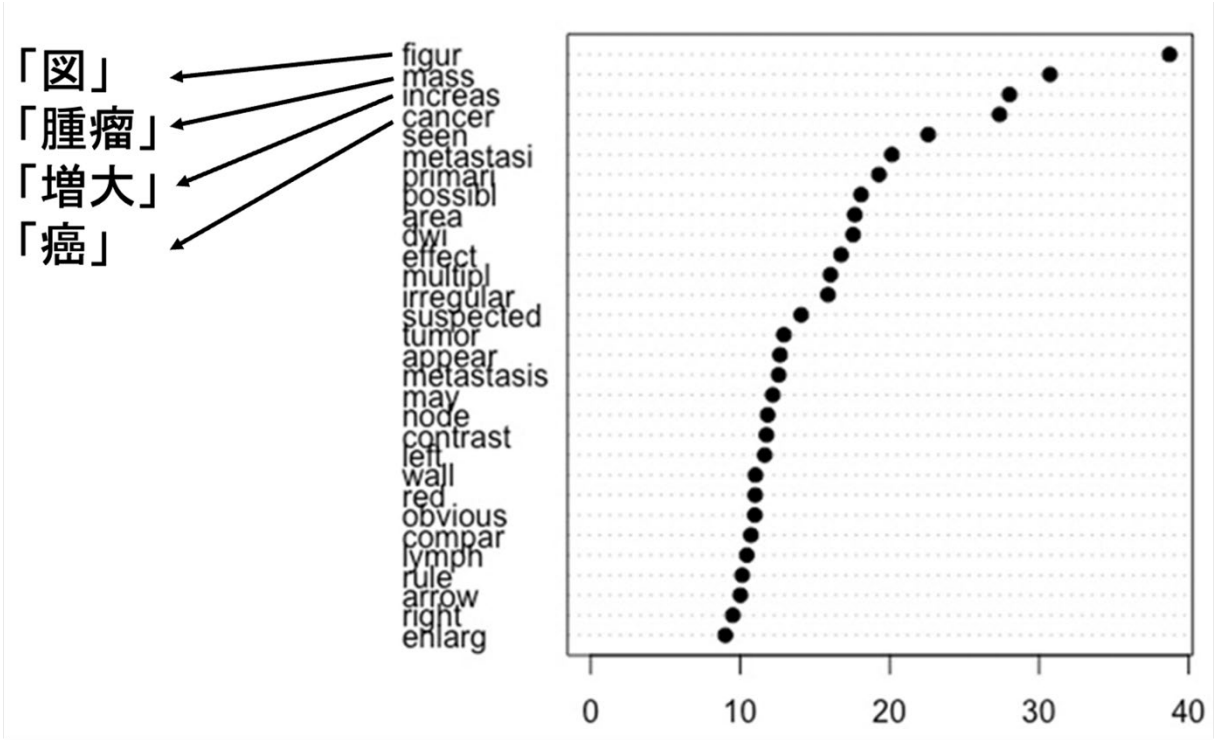

図4　変数重要度。横軸の値が大きいほど重要度が大きい。