Cache を用いた英語での類似症例検索の試み

田中達規[1]　鈴木隆弘[2]　高林克日己[2]
千葉大学工学部メディカルシステム工学科[1]
千葉大学医学部附属病院企画情報部[2]

背景と目的

近年、テキストマイニング技術を用いて、診療情報から診療に役立つ情報を抽出する試みが注目されている。これまで千葉大学医学部附属病院では、テキストマイニングにより様々な医療統計情報を取得し、データベース Cache を利用して日本語での類似症例検索に関する研究を行ってきた。本研究では、英語医学文献の情報を収集したデータベース MEDLINE の症例報告を対象とし、ベクトル空間モデルを用いて各文献の類似度を測定する。これによって、まだ例の少ない英語文献での類似症例検索を目的とする。

対象と方法

はじめに、対象となる MEDLINE に掲載された症例報告を取得し、1 文書 1 ファイルとして約 16 万件保存した。ベクトル空間モデルを作成するために、集めたテキストデータに対して形態素解析を行った。形態素解析には、TreeTagger を用いた。

次に、形態素解析したテキストから名詞単語を抽出し、TF-IDF 法を用いて単語重要度ベクトルを求めた。このベクトルより、ベクトル空間モデルを作成した。

最後に作成したモデル間で内積演算を行い、類似度を算出した。いくつかのテキストについて類似度の高い上位 5 件の症例報告を集め、医師に比較評価を依頼した。

結果

比較評価の結果、病名や患者の特徴などで類似しているとの評価を受けた。しかし、内容に関しては文献によって類似性にばらつきが見られた。特に類似度が 3 位以降の文献は類似していないものが多いと判断された。

考察

MEDLINE の症例報告を対象にしたテキストマイニングにより、類似症例を検索できることを示した。しかし、内容の類似性は類似度の高さに関係なくばらつきがあり、検索精度の向上が必要と考えられる。今後は、単語の重要度ベクトル作成の際、病名や体の部位を示す単語に対して重要度が大きくなるように重みづけの改善を図りたい。

参考文献

[1] 土井俊祐、複数病院間でのテキストマイニングによる DPC 判定の試み、第 28 回医療情報学連合大会論文集、2008