分散表現を用いた読影レポートの類似文書検索

多田 太郎[1]　松本 宏[2]　山本 和英[1]

[1]長岡技術科学大学　[2]株式会社ワイズ・リーディング

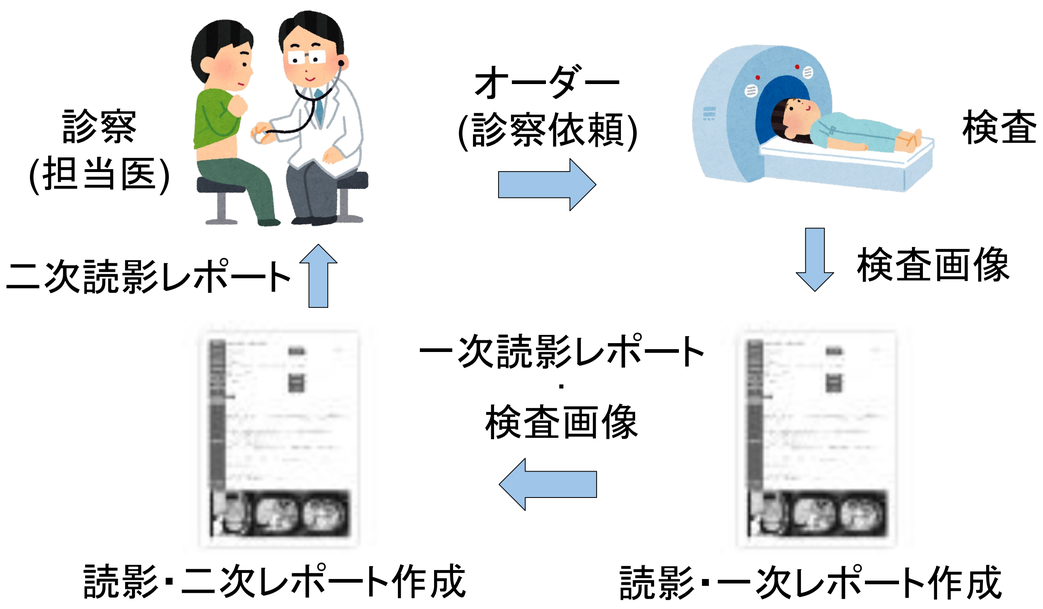
①目的

・放射線科医師の読影レポート作成は負担大
・日本語の医療文書検索で分散表現手法が少ない

→読影レポートを作成する医師の負担軽減
→分散表現の日本語医療文書での有用性確認

分散表現手法を用いて医療類似文書検索

②読影レポート



・オーダー
患者を診察した医師による検査依頼

・一次読影レポート
検査画像を基に作成

・二次読影レポート
検査画像と一次読影レポートを基に作成

③データ

オーダー文,一次読影レポート,二次読影レポート
学習データ: 181,875 検査分
評価データ: 二次読影レポート 80文書
（8疾病カテゴリ×10文書）

複数の疾病が出現する文書が多数
→文書間の類似度の優劣を定めるのが困難
→最初に取り組む課題として任意の疾病が複数カテゴリで出現しない様に評価データの文書を選定

8 つの疾病カテゴリ
(アルツハイマー型認知症, 肺がん, 心筋梗塞, 脂肪肝,椎間板ヘルニア, 内側側副靱帯損傷, 肘頭骨折, アキレス腱損傷)

④実験・結果

辞書: 万病辞書,ComeJisyo,UniDic, IPADic,IPADic-NEologd
　　上記に加え一文字区切りでも分割を行った
手法: doc2vec
学習手法, ハイパーパラメータ:

method	DBow	sub-sampling	10^{-5}
dimension	300	negative-sampling	10
min-count	5		

評価データ各文書に対し残り79文書との類似度を算出
・類似度上位9文書が同疾病カテゴリか
・全80文書に対して確認

dictionary	window-size	epoch	accuracy
万病辞書, ComeJisyo	15	20	0.822
1文字分割	15	500	0.803
UniDic	30	1000	0.894
IPADic	50	1000	**0.910**
IPADic-NEologd	15	20	0.907
IPADic-NEologd	30	1000	0.896

結果から分かち書きについて確認

例:右肺上葉切除術後	
万病辞書,ComeJisyo	右　肺上葉切除　術　後
UniDic	右肺　上葉　切除　術後
IPADic,IPADic-NEologd	右　肺　上　葉　切除　術　後

例:陳旧性心筋梗塞 (心尖部前壁中隔)	
万病辞書,ComeJisyo	陳旧性心筋梗塞　(　心尖部　前壁　中隔　)
UniDic	陳旧　性　心筋　梗塞　(　心尖　部　前壁　中隔　)
IPADic,IPADic-NEologd	陳　旧　性　心筋梗塞　(　心　尖　部　前　壁　中　隔　)

複合名詞は短い名詞で学習したほうが良い？

⑤まとめ・課題

・類似文書検索に分散表現を用い、疾病がカテゴリ間で独立した文書ではあるが90%を超える精度を達成

・一般的な辞書を使っての分かち書きで精度が高かった
→ドメイン特有の複合名詞の分散表現より、複合名詞の各要素個々で学習した分散表現の表現能力が優れていると考えられる
→データ量に依存する可能性がある

・肺がんのカテゴリで他カテゴリよりも精度がよくない
→上記で最も精度の良かった実験でも75.6%

・より実践的なタスクとするため、複数疾病が含まれる文書間での類似度の優劣を定める必要がある
→医療従事者の協力により実現できないか