

分散表現による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

{tada,yamamoto}@jnlp.org, h-matsumoto@ysreading.co.jp

1 はじめに

近年, 自然言語処理分野の様々なタスクに分散表現の手法が用いられている. 本稿では, その分散表現手法を用いて文書ベクトルを生成し, 医療文書である読影レポートの類似文書検索に取り組む.

現在, 医療文書においては, 日本語の文書を対象とするいくつかの個別のタスクに対しては分散表現が用いられているが, 論文としては数としてはあまり多くの発表がされていない. しかし, 分散表現は自然言語処理分野の様々なタスクにおいてその有用性を示している. そこで, 読影レポートから分散表現手法による文書ベクトルを生成し, 類似文書検索の評価を行うことで医療ドメインでの有用性を改めて確認したい.

ここで取り組む医療ドメインでの類似文書検索タスクについては, 先行研究はいくつか存在するが, 活発に発表が行われてはいない現状がある. 本稿では, 医療文書である読影レポートを用いて類似文書検索を行った.

2 関連研究

電子カルテ等を始めとする医療ドメイン文書での自然言語処理の研究は英語圏では活発であり, 多くの利用可能な資源が存在し, 学会でも多く取り組みが発表がされている [1]. しかし, 日本語の医療ドメイン文書を用いた研究は英語圏のそれと比較し, 活発であるとは言えない. 日本語医療ドメイン文書の検索システムの研究は以前より行われており, 論文として発表されているものについては, 小野ら (2004) や岡本ら (2006), 土井ら (2011), 荒牧ら (2018) などが代表的であるが, その他あまり多くの件数は発表がされていない. これは, 医療文書データ自体が入手困難であることや, 二次利用の取り組みが活発になってからあまり長くないため [2] であると考えられる. また, データの匿名化の必要性や医療機関での文書の位置づけ, ドメイン特有の専門用語が現れることも医療従事者の協力なしに文書を取り扱うことを難しくしている. 例えば, 疾病名の特徴として部位や症状の名称に基づく組み合わせ的な複合名詞が多く現

| | | | |
|-----------|--|------|-----------------------------|
| 患者ID | 〇〇〇〇-〇〇〇〇 | | |
| 患者氏名 | 〇〇〇〇 様 (男) | 生年月日 | 19XX 年 XX 月 XX 日 (検査時 68 歳) |
| 臨床診断 | 胃癌。 | | |
| 検査目的 | 上記精査。 | | |
| 臨床情報 | 胃癌で化学療法を予定。胃体部～噴門、前壁～小弯に癌腫あり。 | | |
| 部位/モダリティ | 胸部～骨盤 / CT | 検査日時 | 20XX 年 XX 月 XX 日 XXXX |
| 検査実施 医療機関 | 〇〇病院 | 造影剤 | オイハロミン |
| 依頼医 依頼科 | △△ △△ / 外科 | 病種 | 外来 |
| 希望対応 | 通常 | | |
| 読影担当医 | 〇〇 〇〇 | | |
| 読影所見 | <p>前回 CT 検査(20XX/XX/XX)と比較しました。</p> <p>胃噴門部から胃体上部にかけてほぼ全周性に胃壁の肥厚を認めます。明らかな周囲組織への直接浸潤を疑う所見は認めません。T2-3 相当と思われず。胃噴門部から小弯部にかけて認める多発リンパ節腫大(#1~#3)は最大のもので径 42×52mm とサイズに変化はありません。7 個以上は認めず(N3a, UICC 7th)。</p> <p>肝外側区には癌腫を疑う低吸収域を認めます。明らかな肝転移を疑う腫瘍は認めません。その他、腹腔内臓器に転移を疑う所見はありません。</p> <p>胆嚢底部には萎縮、壁肥厚を認めます。局所的な慢性炎症を疑います。内部に結石を認めます。上行結腸には多発憩室を認めます。</p> <p>単純 CT 上、明らかな転移を疑う所見はありません。播種を疑う結節および腹水は認めません。</p> <p>肺野に明らかな結節、浸潤影は指摘できません。有意なサイズの縦隔リンパ節腫大、その他の縦隔臓器病変はみられません。胸水ありません。</p> <p>骨粗大病変は認めません。</p> <p>その他、有意な所見はありません。</p> | | |
| 読影診断 | 胃癌、多発リンパ節転移疑い(T2-3, N3a, M0; UICC 7th) | | |
| | 読影医: 〇〇 〇〇 | | |

図 1 読影レポートの例

れる点が挙げられる. このような理由から医療文書の扱いが複雑となり, 研究対象とすることへの障壁となっていると考えられる.

医療ドメインでの類似文書検索の先行研究としては, 岡本ら (2006)[3] および土井ら (2011)[4] は, TF-IDF を用いた類似文書の検索を行っている. 岡本ら (2006) の論文では, 診療文書内の文章に Observation, Diagnosis, Treatment の分類を付与し検索に応用している. 土井ら (2011) 論文では, TF-IDF による単語を基底とした文書ベクトル空間法を用いた類似症例検索を行っている. 小野ら (2004)[5] は国際疾病分類である ICD コードと TF-IDF による単語の重み付けから最大全域

木のデンドログラムを作成し、医療文書である退院サマリーのテキストの分類を行った。荒牧ら (2018)[6] は、連想検索エンジンである GETA をベースとした全文検索を行う症例検索システムを実際に 7 年間運用し報告をしている。本稿では、TF-IDF 法をベースラインとし、分散表現手法を用いた文書ベクトルを生成し、医療文書である読影レポートの類似文書検索を行った。

3 読影レポートについて

本研究で取り扱う医療文書は、読影レポートとその検査オーダーである。検査オーダーとは、患者を診察した医師がその患者の症状に関する所見と検査目的を記載した文書である。読影レポートは、検査オーダーに基づき病院の検査部署で検査を行い、撮影された画像を踏まえて検査部署の医師が所見と診断を記載したレポートである。図 1 に読影レポートの例を示す。一次読影レポートと二次読影レポートとがあり、二次読影レポートは検査画像に加えて一次読影レポートを参考に作成されるという関係性がある。二次レポートは一次レポートと合わせダブルチェックの意味を持っている。1 日の検査数でも非常に多いため、これらの読影レポート作成は担当する医師の大きな負担となっている [7]。

本稿での、読影レポート作成の際に参考となりうる類似文書を検索するタスクへの取り組みは、レポート作成者の負担軽減にも繋がるのが期待される。本稿では上記の通り、検査オーダーおよび一次、二次の読影レポートのテキストのみを用い類似文書検索を行った。

4 実験

4.1 データ・セット

実験に使用するデータの内訳は、1 つの検査に対する検査オーダー、一次読影レポートおよび二次読影レポートの 3 つを 1 セットとする 181,875 検査分のテキストデータである。これらの文書から 80 検査分を選定し評価用データとし、その二次レポートを用い実験を行った。残りの 181,795 検査分の文書を学習用データとする。

4.2 前処理について

医療文書への前処理として、MeCab による分かち書きと評価用データのクリーニングを行った。分かち書きに使用した辞書は、MeCab のデフォルトで利用可能となっている IPADic に加え、医療用語辞書である

ComeJisyo^{*1}および万病辞書^{*2}の 2 点を併用した。分かち書きを行う際に使用する医療用語辞書の順番による実験結果への影響も考慮し、使用する辞書の順番を入れ替えて 2 通りの分かち書きを行った。

また、読影レポートの特徴であるが、検査手法によって検査画像内に映る検査対象の部位周辺の臓器などに対して、本来の検査対象でなくとも読影レポート内で触れる文が記載される。この検査目的でない部位に触れる文は、ほとんどが陰性の所見であり、部位が連続した表 1 のような文が現れる。これは、ほぼ全てのレポート中に現れ、文書間の類似度を求める際にノイズとなってしまうことが想定される。そのため、評価用データから定型文の様に現れる部位名称や画像から分かる代表的な疾病（胸水、心拡大など）が 1 つの文に連続して含まれる陰性の文書の削除を行った。合わせて、“その他、有意な所見はありません。”という文も定型でほぼすべての文書に現れるため、削除を行った。この陰性を示す文言も“有意な所見を認めません（はありません）”、“異常を認めません（はありません）”と定型であるため、判断基準とした。また、扱ったテキストでは、所見内にレポート受取り手への挨拶文や他検査との比較依頼も定型で現れるため、その削除も行った。文書の頻度は扱うデータの確認の際、人手および自動で行った。今回は評価データの量が少ないため、削除作業は人手で行った。表 1 は削除した文の例である。

表 1 削除した高頻出文書の例

肝、胆、膵、脾に有意な所見を認めません。
胃、小腸、大腸などに有意な所見を認めません。
副腎、尿管、膀胱などに有意な所見を認めません。
骨折、足関節に異常はありません。
胸水、心拡大はありません。
その他、有意な所見はありません。
エコーなどとも比較されてください。
精査をお願いします。
お世話になっております。

4.3 実験方法

類似文書検索のため文書類似度を算出する。手法としては先行研究で用いられている TF-IDF により重み付けした文書ベクトル間でのコサイン類似度をベースラインとし、doc2vec および fastText を用いて生成した

^{*1} <https://ja.osdn.net/projects/comedic/>

^{*2} <http://sociocom.jp/~data/2018-manbyo/index.html>

文書ベクトル間でのコサイン類似度により各文書間の類似度を求めていく。doc2vec[8]についてはgensim実装^{*3}を用い、fastText[9, 10]はfacebookInc.が提供するオリジナル実装^{*4}を用いた。ハイパーパラメータについては、doc2vecのファインチューニングを行ったLauら(2016)[11]を参考にし実験を行った。

4.4 評価

評価用データの各文書から生成した文書ベクトルを用い、文書間の類似度を算出する。疾病が単一である読影レポートは少なく、複数の疾病が組み合わせ的にレポート内の所見や診断で出現する。異なる疾病が検査の主対象であっても、共通する疾病を併発している場合も多い。さらに各疾病間での類似度を定めるのも困難である。そのため、医師の様に医療知識が十分にある人間でないと文書類似度の優劣を判断することが困難である。そこで、先行研究の岡本ら(2006)に準じ、8つの疾病(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, 内側側副韌帯損傷, 肘頭骨折, アキレス腱損傷)をカテゴリとして設定し、疾病が複数のカテゴリに含まれない様に各疾病で10文書, 計80文書選定し評価用データとした。この評価用データから1文書を選び出し、評価データ内の残り79文書に対する類似度を求め、その類似度の上位9文書に対し同じ疾病カテゴリの文書がどれだけ含まれているかを確認し精度とした。

5 結果・考察

今回の読影レポートを用いた類似文書検索の実験では、分散表現手法により生成した文書ベクトルによる検索がTF-IDFで重み付けした文書ベクトルによる検索よりも高い精度となった。表2および図2はTF-IDFおよびdoc2vec, fastTextでスコアの高かったモデルでの実験結果、表3は実験で精度の高かったハイパーパラメータの一覧である。表2は各文書に対する類似度上位9文書の内、同じ疾病の文書が得られた精度(中央値および平均)である。図2は、横軸に各文書に対して同じ疾病カテゴリの文書が得られた精度(類似度上位9文書中の同疾病カテゴリの文書数)、縦軸は精度に対する文書数である。縦軸の文書数は手法毎に合計値が評価分散表現手法での異なるハイパーパラメータを変更しての実験は、その全てを網羅する規模では行えなかったが、TF-IDFと比較しても、doc2vecのデフォルトのハ

表2 より精度の高かったモデルでの実験結果

| | accuracy(median) | accuracy(average) |
|----------|------------------|-------------------|
| TFIDF | 0.111 | 0.108 |
| doc2vec | 0.778 | 0.743 |
| fastText | 0.667 | 0.613 |

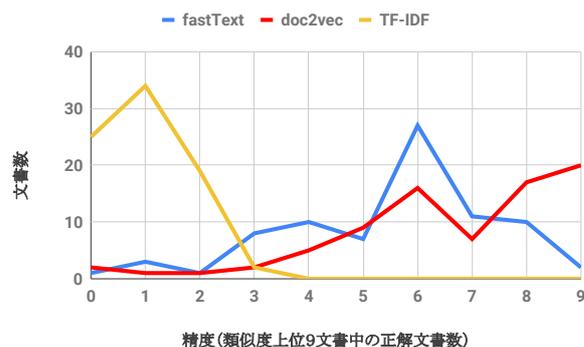


図2 より精度の高かったモデルで精度に対する文書数

表3 より精度の高かったモデルでのハイパーパラメータ

| Hyper parameters | |
|------------------|--|
| doc2vec | Method : dbow, Dim : 300, Window Size : 15, Min Count : 5, Sub-sampling : 10^{-5} , Negative sampling : 5, Epoch : 20 |
| fastText | Method : skipgram, the other parameters are default |

イパーパラメータ設定のものと比較しても高い精度での類似文書検索を行うことができた。

fastTextではハイパーパラメータの変更を含め16通りと実験のバリエーションが少ないが、実験した範囲での結果では、skipgramを用いハイパーパラメータはfastTextデフォルト値で学習したモデルで精度が高い結果となった。実験結果で精度が高くなったモデルのハイパーパラメータ設定の傾向としては、skipgramを用いたもの、損失関数をsoftmaxとしたものでより高い結果となり、次に階層的softmaxが続く結果となった。これらを用いると他の手法と比較し計算コストが高くなるが、今回の類似文書検索の用途では、学習時に医療機関が保持する大きな規模の文書量を使用することができれば再学習が必要となる頻度は低くなるため、障壁は少ないと考えられる。MeCabでの分かち書きの際に

^{*3} <https://radimrehurek.com/gensim/index.html>

^{*4} <https://fasttext.cc/>

用いる医療辞書の順番については、入れ替えた際に精度が僅かに異なるが、どちらの順番が優位といえる結果は今回の実験では得られなかった。

doc2vec では 20 通りの実験を行い、dbow を用いた際に精度が高い結果となり、損失関数を階層的 softmax としたものでは精度が低い結果となった。こちらでも MeCab での分かち書きに用いる医療辞書の順番による優位差は統計結果からも得られなかった。

今回の実験で、同じ疾病カテゴリの文書を上位類似文書として得られなかった文書については、評価データと同疾病カテゴリ内の他の文書に無い疾病を複数患っていたもの、同疾病カテゴリ内の他の文書と異なり既に治療が行われており経過観察での検査であったもの、“軟部組織”や“高信号”、“断裂”などの組織や検査、怪我に共通する表現があり、他の疾病カテゴリの文書の方が高類似度となってしまったものが見受けられた。これらから出現する単語からの影響を大きく受けていることが確認できる。しかし、人が見て印象が異なる文書であることが確認できる文書内容であったため、その点ではより文書の内容が反映された文書ベクトルとも言えるのではないか。怪我や疾病の所見を記す際に共通して現れる“軟部組織”や“高信号”、“断裂”などの単語に影響を受けてしまう点に関しては、これらが現れた文書同士は広い意味では近い文書と言えるかもしれないが、やはり課題が残る。

また、評価用データで頻出する文を削除する前のものでも同様の実験を行ったが、やはり評価データ内の全文書間での類似度が上がり、より多くの異なる疾病の文書が上位類似度文書に現れる結果となった。

6 おわりに

本稿では、分散表現手法を用いた文書ベクトルによる読影レポートの類似文書検索を行った。実験では、分散表現手法でベースラインとする TF-IDF よりも精度が高い結果となり、医療ドメインにおける分散表現手法の有用性を改めて確認できた。これは同時に、医療文書からある程度の質の文書ベクトルが生成できたとも言えるのではないか。また、実験を通して、読影レポートを扱う際の前処理や文分散表現手法での学習時の適したパラメータの傾向などの知見を得た。

参考文献

[1] Pranjul Yadav, Michael Steinbach, Vipin Kumar, Gyorgy Simon, Mining Electronic Health

Records (EHR): A Survey, ACM Computing Surveys, Vol. 1, No. 1, Article 1, Publication date: April 2016.

- [2] 武田 理宏, 真鍋 史朗, 松村 泰志, 電子カルテデータ二次利用の現状と課題, 生体医工学 55 巻 4 号 pp. 151-158 2017
- [3] 岡本 和也, 竹村 匡正, 黒田 知宏, 長瀬 啓介, 吉原 博幸, 文脈に基づく類似診療文書検索システム, 生体医工学 44 巻 1 号 pp. 199-206 2006
- [4] 土井 俊祐, 木村 隆, 関根 正樹, 鈴木 隆弘, 高林 克日己, 田村 俊世, 学会ホームページにおける類似症例検索システムの実運用と評価, 生体医工学 49 巻 6 号 pp. 870-876 2011
- [5] 小野 大樹, 高林 克日己, 鈴木 隆弘, 横井 英人, 井宮 淳, テキストマイニングによる退院サマリー自動分類の試み, 医療情報学 24 巻 1 号 pp. 35-44 2004
- [6] 荒牧 英治, 岩尾 友秀, 若宮 翔子, 伊藤 薫, 矢野 憲, 大江 和彦, 症例検索システムの思考運用に基づいた利用状況に関する基礎的検討, 医療情報学 38 巻 4 号 pp. 245-256 2018
- [7] 牧野 恭子, 早川 ルミ, 寺井 公一, 深津 博, 知識処理を活用した読影レポート作成支援システムの開発と評価, 人工知能学会論文誌, 23 巻 6 号 pp. 463-472 2008
- [8] Quoc Le, Tomas Mikolov, Distributed Representations of Sentences and Documents, Proceedings of the 31 st International Conference on Machine Learning, JMLR: W&CP volume 32, pp. 1188-1196 ,2014
- [9] Piotr Bojanowski, Edouard Grave, Armand Joulin and Tomas Mikolov, Enriching Word Vectors with Subword Information, Transactions of the Association for Computational Linguistics, vol.5 pp. 135-146 2017
- [10] Armand Joulin, Edouard Grave, Piotr Bojanowski, Tomas Mikolov, Bag of Tricks for Efficient Text Classification, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2, Short Papers, pp. 427-431 2017
- [11] Jey Han Lau and Timothy Baldwin, An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation, Proceedings of the 1st Workshop on Representation Learning for NLP, pp. 78-86 2016