

診断書作成支援のための後続文候補の提示手法の検討

松本 宏[†] 山本 和英[†] 相馬 章人[‡] 矢野 高宏[‡]

[†]長岡技術科学大学

[‡]株式会社ワイズ・リーディング

[†]{matsumoto, yamamoto}@jnlp.org [‡]{a-souma, t-yano}@ysreading.co.jp

1 はじめに

電子カルテ化の普及が広がっている [10]。これより、蓄積される電子医療文書から様々な応用が期待される。しかし、現状は電子カルテ化の有無に関わらず、医療文書作成がもっとも高負荷と報告されており [9]、負担軽減に貢献していない。そのため医療文書作成支援が求められる。

本稿では医療文書の中でも診断書を対象として論じる。現状、診断書作成支援にはテンプレート型、文リコメンド型と2種類存在する [7, 8]。テンプレート型は穴埋め形式のテンプレートを埋めて、診断書を完成させる。文リコメンド型は入力状況に応じて後続文を予測し、ユーザが選択する毎に候補が更新され、診断書を完成させる。ユーザ入力を削減することにより診断書作成の負担軽減に大きく貢献する。しかし、この文リコメンド・システムが抱える問題として候補数の多さや候補の順位付けのランキング問題がある。本稿ではこの2つの問題を解くためいくつかのモデルを提案し、評価を行う。

2 関連研究

医療言語処理の研究は世界的に言い換えタスクなどの研究が幅広く行われている [5]。日本においても NTCIR で MedNLP が設定され医療言語処理が着目され始めている。[4, 1, 2]。しかしながら、診断書作成支援の研究は数多くは研究されておらず、代わりに診断書自動生成の研究が行われている [6]。診断書作成支援は入力文より後続文を予測を行う。この様に予測システムという観点では類似タスクとして単語予測が挙げられる [3]。しかし、単語予測と後続文予測における違いは大きくある。単語予測では構文、意味などの情報が利用可能である。そして、文予測において

は文と文の繋がりにおける事象の因果関係などが利用可能である。

3 全体概要

本章では、本研究で扱うデータ仕様、診断書を対象としたときの後続文の予測の難しさ、そして既存ツールがもつ問題点について述べ、提案手法の概要について述べる。

3.1 データ仕様

本研究で扱うデータは放射線画像から読影された診断書となる。以下に1例を示す。構成の特徴としてスキャン画像から判別される診察結果からの知見の列挙となる。また、知見以外にも頭語や、他病院への紹介文などがある。

... 軽度の心拡大を認めます。心不全の可能性が
あります。心機能を評価して下さい。肺門に左
右差ありません。縦隔、肺門リンパ節腫大を疑
う像はありません。...

図 1: 診断書例一部抜粋

3.2 診断書における文予測の難しさ

予測の難しさとして以下の3点が考えられる。

1つ目に話題継続性である。ここでいう話題とは図1の場合において最初の3文に共通する「心臓」のように、文がもつテーマとする。そして話題継続性における難しさとは、同一の共通の話題が継続するかの予測問題である。例えば、図1では心臓の話題が3文続

いている。このように文同士が因果関係にある場合は容易に予測がつく。しかし、話題は同一であるが、文同士において因果関係もなく、ただ様々な症状の知見が列挙されるケースもある。

もう1つの難しさは話題遷移である。図1で、心機能の評価を促し、話題が「胸水」に遷移しているのが分かる。このように1つの話題から別の話題に遷移するときにおける問題として、次にあげる話題選択があげられる。

このように、2つの問題を同時に意識した予測システムの構築を行う際にどのような後続文候補を構築し、さらにその候補中の文に対してどのような順位付けを行うかが問題となる。

3.3 既存システムの問題点

既存システムにおける問題点について述べる。例として、既存ツールで「心拡大、背景に心不全を疑います。」の入力に対して得られる後続文候補は、数として60候補以上提示され、さらに候補リストにおいて、この文に続く文としてより適切である文「心機能の評価して下さい。」が、別の話題となる「胸水見られません。」より下位に位置づけられししまう。

3.4 システム概要

本研究では先述の問題に対して予測システムを構築して問題に挑む。以下の図が予測システムの概略図となる。入力文に対して医療用語抽出を行い、抽出された医療用語に基づき候補文リストの生成を行う。

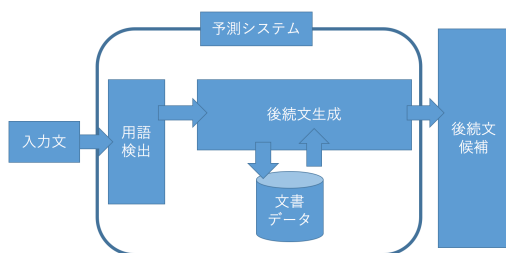


図 2: システム概要図

4 用語検出器の構築

本システムが取り扱う対象データは医療診断書である。図1を見て分かるように医療用語が多く含まれて

いる。本システムは医療用語より後続文の予測を行うため、医療用語を抽出する必要がある。しかし、既存の形態素解析では医療用語の抽出が行えるものはない。そのため、用語検出器の構築を固有表現抽出で多く利用されている手法である Conditional Random Field を用いて行う [1]。

4.1 用語タグ付け

診断書データから無作為に選んだ2,000文にある医療関連用語に対してタグ付けを行った。この医療関連用語の基準は特に医学知識を習得していない著者の主観的評価で行った。タグは以下の6種類を設定した。2,000文中におけるそれぞれのタグの例と種類数を以下に示す。

- 部位
 - 種類数: 789
 - 対象用語例: 頸椎, 肩鎖関節
- 症状/状態変化
 - 種類数: 620
 - 対象用語例: 陳旧性部分断裂, 肺炎合併
- 画像診断関連用語
 - 種類数: 180
 - 対象用語例: 単純 CT, 低吸収結節影, STIR 像
- 手術/療法/検査/分類法
 - 種類数: 33
 - 対象用語例: Hohl 分類, 腰椎固定術
- 器具名
 - 種類数: 5
 - 対象用語例: 腫瘍マーカー, スtent
- 上記以外
 - 種類数: 32
 - 対象用語例: 喫煙歴, 脳神経外科

4.2 素性

素性は対象形態素とその前後の形態素の表層形、文字列の種類、品詞を利用する。文字列の種類は、ひらがな、カタカナ、英字列、数字列の種類を指す。複数種類から構成されている場合は、それぞれを組み合わせ（ひらがな-カタカナ）表現する。

4.3 検証実験

タグ付けを行った 2,000 文に対して用語検出の性能を測るため、10 分割交差検定を行った。以下に平均値を示す。

表 1: 検証結果

適合率 [%]	再現率 [%]	F 値 [%]
92.5	92.6	92.5

5 予測モデル

予測モデルには用語 bigram モデル, trigram モデルの 2 つのモデル構築を行う。診断書例の図 1 中の 3 文を例にこのモデルへの経緯を説明する。

文例集 1

1. 「軽度の心拡大を認めます。」
2. 「心不全の可能性がります。」
3. 「心機能を評価して下さい。」

文例 1 で「心拡大」が認知され、文例 2 で「心不全」が疑われている。この様に、前文、後文で因果関係を持っていることが分かる。そのため、用語系列として「心拡大」が含まれる文に対して次に疑われる候補として「心不全」が挙げられ、これらに関する文を出力することが望ましいと考えられる。

この考えに基づき用語 bigram モデルを構築する。

5.1 用語 bigram モデル

診断書の全文に用語検出を行い、各文を用語列に変換する。得られた用語列に対して、直積集合を構築する。これは、文中に複数の用語がある場合においてそれぞれケースを考慮するためである。直積集合から用

語系列の bigram を作成し、それぞれの bigram の頻度を得る。

予測は入力文に含まれる用語を bigram の第 1 要素に含まれるエンタリーに対して、第 2 要素の用語を含む文を候補とする。それぞれの候補の順位は bigram 頻度に基づく。入力文に複数の用語が含まれる場合はそれぞれの用語で得られた候補同士を積結合を行う。

5.2 用語 trigram モデル

bigram モデルにおける問題点として、異なる文脈においても同一の用語が含まれている場合、生成される候補リストも同一候補となる一意性の問題がある。

例として、先の例文集 1 に対して以下のような文例があったとする。

文例集 1'

1. 「心拡大みられません。」
2. 「胸水みられません。」

文例集 1 の文例パターンが高頻度で存在した場合、文例集 1' の文系列を求めている場合でも「心拡大みられません。」という入力文に対して、内在する用語は「心拡大」だけなため、「心不全の可能性がります。」が候補上位に配置される。

この問題の解決方法として、対象文の前文に含まれる用語を用いて用語 trigram の構築を行う。

6 実験&評価

6.1 実験設定

医療診断書 13,000 文書 (平均文数 10 文/文書) に対して 10 分割交差検定で実験を行う。

6.2 評価方法

評価にあたって 3 種類の評価手法を設ける。入力文からシステムが生成する予測文候補中、正解文のランクを平均逆順位で評価する。正解文が予測候補に出現しない場合は対象外とする。全入力文中正解文が予測文候補に出現する割合を被覆率で評価する。そして、正解文が画面移動無しに得られる評価として、正解文が上位 25 件内に得られる割合を画面内被覆率で評価する。

- 平均逆順位 (Mean Reciprocal Rank)

$$MRR = \frac{1}{|S|} \sum_{i=1}^{|S|} \frac{1}{rank(s_i)}$$

- 被覆率 (Cover Rate)

$$C.R. = \frac{\text{候補リストに正解が含まれる文数}}{\text{全文数}}$$

- 画面内被覆率 (In-Window Cover Rate)

$$I.W.C.R. = \frac{\text{候補上位 } N \text{ 中に正解が含まれる文数}}{\text{全文数}}$$

6.3 評価結果

bigram, trigram モデルの評価結果を以下の表に示す。

表 2: 実験結果

	M.R.R	C. R. [%]	I. W. C. R. [%]
bigram	2.40 e -3	1.65	0.83
trigram	6.17 e -3	5.39	2.94

7 考察

実験結果の表 2 より bigram モデルと trigram モデルを比較すると、trigram モデルで全体的に向上している。平均逆順位、画面内被覆率両方共に向上していることから、trigram モデルを使うことにより、より適切な選択肢を上位に引き上げることができたことが分かる。さらに、被覆率と画面内被覆率を比較すると bigram モデルで得られる正解文全文の中で (0.83/1.65 より)50.3[%] の割合で上位 25 件で獲得できていたが、trigram モデルを用いることにより (2.94/5.39 より)54.5[%] の割合で向上できることがわかる。

しかし、bigram モデル、trigram モデル双方ともに被覆率が小さい。これは今回のモデルは医療用語の用語系列による候補取得のため、症状の発見/認知の有無に関わらず、用語系列が一致する場合、常に同一候補が提示されてしまう。このため、モデル化において用語のみだけでなく用語に係る動詞、またそれを否定する助動詞などの利用が必要だとわかる。

また、被覆率の低さの原因として考えられる別の要因として、入力文から複数個の候補集合が得られたとき、結合手法が積集合により複数の用語を持つ共通文のみを候補文対象とすることにあると思われる。

8 おわりに

本稿では、入力文からその入力文に後続する文を予測する文予測というタスクにおいて用語系列を利用した 2 つのモデルの比較を行った。正解文の順位の引き上げを用語 ngram モデルを利用し、bigram から trigram 変更することにより実現できた。

しかし、このモデルだけでは実利用において採用は難しいため、さらなる別の手法の検討が今後の課題となる。

参考文献

- [1] Eiji Aramaki, Mizuki Morita, Yoshinobu Kano, and Tomoko Ohkuma. Overview of the ntcir-11 mednlp-2 task. In *NTCIR*, pp. 147–154, 2014.
- [2] Eiji Aramaki, Mizuki Morita, Yoshinobu Kano, and Tomoko Ohkuma. Overview of the ntcir-12 mednlp-doc task. *Proceedings of NTCIR-12*, pp. 71–75, 2016.
- [3] Masood Ghayoomi and Saeedeh Momtazi. An overview on the existing language models for prediction systems as writing assistant tools. In *IEEE international conference on Systems, Man and Cybernetics*, pp. 5083–5087. IEEE, 2009.
- [4] Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, Mai Miyabe, and Eiji Aramaki. Overview of the ntcir-10 mednlp task. In *NTCIR*, pp. 696–701. Cite-seer, 2013.
- [5] Pracheta Sahoo, Asif Ekbal, Sriparna Saha, Diego Molla, and Kaushik Nandan. *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, chapter Semi-supervised Clustering of Medical Text, pp. 23–31. COLING, 2016.
- [6] Sebastian Varges, Heike Bieler, Manfred Stede, Lukas C Faulstich, Kristin Irsig, and Malik Atalla. Semscribe: Natural language generation for medical reports. In *LREC*, pp. 2674–2681, 2012.
- [7] インフォコム株式会社. Medi-support plus. <http://www.infocom.co.jp/healthcare/medi/medisupportplus.html>.
- [8] 株式会社ワイズ・リーディング. Y's chain. <http://labo.ysreading.co.jp/ys-chain>.
- [9] 瀬戸僚馬, 津村宏. 医師が電子カルテ操作に費やす業務時間に関する調査. *医療情報学*, Vol. 32, No. 2, pp. 59–63, 2012.
- [10] 保健医療福祉情報システム工業会. オーダリング・電子カルテシステム 病院導入状況調査報告書. [https://www.jahis.jp/files/user/\(印刷版\)2015導入調査報告書.pdf](https://www.jahis.jp/files/user/(印刷版)2015導入調査報告書.pdf).