安田女子大学紀要　42，313–316　2014.

教育用電子カルテのデータベース上に収録された模擬患者診療記録のレコード群に対する自然言語処理ツールを用いた単語の種類や出現頻度に関する考察

大　塚　敬　義

A Study on the Frequency and Type of Words that Use Natural Language Processing Tools for Simulated Medical Records in a Database of an Educational Electronic Medical Record System

Takayoshi Otsuka

1. 背　　景

団体「教育用電子カルテ共同利用協議会」では，加盟各校での授業で電子カルテおよび電子教材としての模擬患者診療記録（以下「模擬カルテ」と表記）を共同で利用している。

模擬カルテは，文部科学省の「大学教育充実のための戦略的大学支援プログラム」のひとつとして採択された「コメディカル養成のための教育用電子カルテシステムおよびデータベースの構築と実践」において開発された電子教材が原型となっている。

加盟各校が利用している模擬カルテのデータベースには，2013年 6 月末時点で368件のレコードが蓄積されている。レコード群に対してタグ付けを行った結果は言語資源協会から提供されているものの，各レコード上における医療用語の出現傾向を詳述した先行研究[1–11]は，NTCIR-10医療言語処理（MedNLP）にほぼ限定されており決して多くない。

2. 目　　的

そこで加盟各校の科目担当者にとってより良い授業を実施するための一助を提供すべく，我々は模擬カルテのテキスト内において出現する単語の種類や出現頻度の概要を把握する。

また模擬カルテに含まれる語彙の豊富さを測定すべく，出現頻度の少ない単語群にどのようなものがあり，かつ何回出現するか等を観測することで，模擬カルテに含まれる語彙の豊富さを測定でき，教材としての模擬カルテの質を示す指標を知ることができる。

3. 方　　法

なお模擬カルテからテキスト部分を抜粋したデータ（GSK2012-D）が言語資源協会（略称GSK）から提供されている。

当該データは注意書きである README0206.txt を除くと，実質的に次の 2 つのテキストデータから構成されている。ひとつは，タグが付与されていないプレーンテキスト EHR_0131.txt である。もうひとつは，<a>（age：年齢）や <c>（complaint：症状名・疾患名）などの文節タグが，文節を越えない範囲で付与されている EHR_0327.xml である。

著者はここでは前者を Windows 版の形態素解析器「茶筌」（chasen-2.3.3）にかけ，単語の品詞や出現回数を調査した。

4. 結 果

当該データを98,200個の形態素に分割できた。その中から品詞が「名詞 -」で始まる29種類の形態素を37,180個抽出でき，異なり語数は3,705種類であった。頻出順でみると，数字語句の「１」が588回，「２」が440回であった。これら 2 つの数字語句に次ぐ頻出語は「時」（322回），「日」（272回），「入院」（262回），「検査」（254回），「性」（244回）であった。これに対し出現回数が最も少ない出現回数 1 回の語句には「喀痰」「疝痛」「蕁麻疹」「齲歯」「癒合」「敗血症」「播種」「膿疱」等があった。

なお茶筌による解析作業とは別に，Excel でデータ EHR_0131.txt を読み込み，排液管を指す「ドレーン」で検索したところ85回の出現を確認できた。しかしながら今回利用した茶筌による解析では「ドレーン」の出現回数は 0 回であり，「ド」が93回，「レーン」が85回含まれていた。

5. 考 察

まず形態素解析の正確さについて考察する。今回利用した茶筌には，配布されたままの原型辞書データ内に医療の専門用語すべてが完全収録されていないゆえ，形態素解析の結果は完全正確ではない。ひとつの手段として「ドレーン」を含むいくつかの専門用語を辞書に手動で追加し，可能な限り形態素への分割が正しくなるようにして再度解析を試みる必要があるとも考えられる。

ただし，確かに専門用語を辞書登録すれば正しく形態素解析を行えるようにはなる。しかし専門用語自体が時代の進展と共に新規に続々と発生していくので，それに追随して形態素解析の辞書を更新し続けるのは現実的には厳しい。そのために専門用語の自動抽出という研究分野が存在するほどである。よって本格的な専門用語抽出とまではゆかなくとも未知語の名詞が連続して出現する場合は複合名詞として単一の用語とみなす等の処理を試行したり，MeCab 等の他の形態素解析ツールも用いる余地がある。

次に EHR_0131.txt に含まれる語彙の多様性について考察すると，当初に危惧されていたような偏りや単純性は決して無かった。「蕁麻疹」（じんましん）のような比較的なじみのある単語を含むと同時に，高校卒業後に医療系の学科で初めて目にする傾向のある「喀痰」「疝痛」「齲歯」「癒合」「播種」「膿疱」といった単語を含んでいた。よって一定の水準で語彙の多様性は確保されていることといって良さそうである。

付記：本稿の執筆に当たり言語資源協会から供与されたデータ「GSK2012-D」を用いた。また模擬患者診療記録のデータベースへアクセスするに当たり「教育用電子カルテ共同利用協議会」と本学関係者との間で事前に契約を交わした。

参 考 文 献

1) Mizuki MORITA, Yoshinobu KANO, Tomoko OHKUMA, Mai MIYABE, Eiji ARAMAKI: Overview of the NTCIR-10 MedNLP Task.
http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/MedNLP/01-NTCIR10-OV-MEDNLP-MoritaM.pdf.
2) Ryo Fujii, Masashi Tada: Improvement Recall of NTCIR-MedNLP Using Hierarchical Bayesian Language Models.
http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/MedNLP/02-NTCIR10-MEDNLP-FujiiR.pdf.
3) Shohei Higashiyama, Kazuhiro Seki, Kuniaki Uehara: Clinical Entity Recognition Using Cost-Sensitive StructuredPerceptron for NTCIR-10 MedNLP.
http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/MedNLP/03-NTCIR10-MEDNLP-HigashiyamaS.pdf.
4) Hiroto Imachi, Mizuki Morita, Eiji Aramaki: NTCIR-10 MedNLP Task Baseline System.
http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/MedNLP/04-NTCIR10-MEDNLP-ImachiH.pdf.
5) Kota Kanno, Kazuyoshi Osanai, Kyoji Umemura: An Trial Report to NTCIR10 MedNLP: Extracting Medical Diagnostic Term.
http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/MedNLP/07-NTCIR10-MEDNLP-KannoK.pdf.
6) Lun-Wei Ku, Edward T.-H. Chu, Cheng-Wei Sun, Wan-Lun Li: Identifying Symptoms and Diseases in MedNLP Japanese Materials Using Chinese Resources.
http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/MedNLP/08-NTCIR10-MEDNLP-KuL.pdf.
7) Pierre-François Laquerre, Christopher Malon: NECLA at the Medical Natural Language Processing Pilot Task (MedNLP).
http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/MedNLP/09-NTCIR10-MEDNLP-LaquerreP.pdf.
8) Yasuhide Miura, Tomoko Ohkuma, Hiroshi Masuichi, Emiko Yamada Shinohara, Eiji Aramaki, Kazuhiko Ohe: UT-FX at NTCIR-10 MedNLP: Incorporating Medical Knowledge to Enhance Medical Information Extraction.
http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/MedNLP/10-NTCIR10-MEDNLP-MiuraY.pdf.
9) Yuji Nomura, Tkashi Suenaga, Daisuke Satoh, Megumi Ohki, Toru Takaki: Medical Information Extracting System by Bootstrapping of NTTDRDH at NTCIR-10 MedNLP Task.
http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/MedNLP/11-NTCIR10-MEDNLP-NomuraY.pdf.
10) Koichi Takeuchi, Shozaburo Minamoto, Motoki Yamasaki: Complaint and Diagnosis Extraction System Utilizing Rulebased Term Extraction System.
http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/MedNLP/12-NTCIR10-MEDNLP-TakeuchiK.pdf.
11) Yuka Tateisi, Takashi Okumura: A Simple Approach to NTCIR-10 MedNLP task.
http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/MedNLP/13-NTCIR10-MEDNLPTateisiY_20130528.pdf.

Summary

In member schools of the organization JUCEE (The Joint Use Conference for Electronic Health

Care Education), simulated medical records are being used as teaching materials in class.

There are 368 records in the database of simulated medical records (in 2013 at the end of June).

Indeed the tagged text data for the part of 368 records has been provided by GSK, but few previous studies have mentioned the tendency of the medical terms that appears in the records.

The data GSK2012-D includes the text data EHR_0131.txt that is an excerpt from the text of simulated medical records.

We used morphological analyzer “ChaSen” in order to analyze the text data EHR_0131.txt, and we report the result.

Key words: Educational Medical Record System, Natural Language Processing, Simulated Medical Record

〔2013. 9. 26 受理〕