様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 25 年 6 月 7 日現在

機関番号：12601
研究種目：挑戦的萌芽研究
研究期間：2010 ～ 2012
課題番号： 22650025
研究課題名（和文）　ダミー診療録の構築および自動構造化に関する研究

研究課題名（英文）　Automatic Structuring for Dummy Health Records

研究代表者
荒牧　英治（ARAMAKI EIJI）
東京大学・知の構造化センター・講師
研究者番号：70401073

研究成果の概要（和文）：
平成 13 年度に政府が発表した「保健医療分野の情報化にむけてのグランドデザイン」にて，電子カルテシステムを始め医療 IT 技術の普及が課題の一つとして掲げられました．以降，急速に医療の IT 化が進み，その結果，かつてない大量の臨床データが電子化された状態でストックされつつあります．このデータを有効に利用することができれば患者の生活習慣と疾患の相関（例えば，喫煙と癌）や，薬品と副作用の相関（タミフルと精神障害）について過去類を見ない大規模な調査を迅速に行うことが可能となり，臨床研究が加速的に進展するとして高い期待が寄せられています．しかし，単にデータを電子化しただけで，上記のような革新が実現できるわけではありません．電子カルテにおいても自然言語で入力される箇所が相当な割合で存在します．よって，電子カルテデータを臨床知識としてフルに活用にするためには，自然言語処理を活用することが必須となります．以上の背景のもと，本プロジェクトでは模擬診療録を用いた応用研究を進めます．

研究成果の概要（英文）：
The use of Electronic Health Records (EHR) in hospitals is increasing rapidly everywhere. They contain much clinical information about a patient's health, including the frequency of drug usage, related side-effects, and so on, which facilitates unprecedented large-scale research. Nevertheless, extracting clinical information from the reports is not easy because they are written in natural language. This study specifically examines clinical information using dummy records that are also build by this project.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,400,000	0	1,400,000
2011 年度	1,000,000	300,000	1,300,000
2012 年度	500,000	150,000	650,000
年度			
年度			
総　計	2,900,000	450,000	3,350,000

研究分野： 知識処理
科研費の分科・細目：知識処理・自然言語処理
キーワード： 自然言語処理，医療情報学

１．研究開始当初の背景

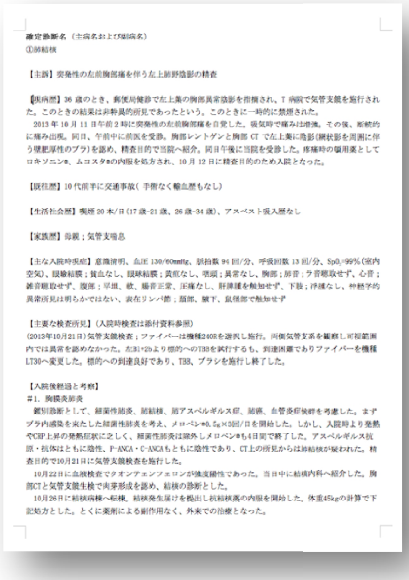
平成 13 年度に政府が発表した「保健医療分野の情報化にむけてのグランドデザイン」にて，電子カルテシステムを始め医療 IT 技術の普及が課題の一つとして掲げられた．以降，急速に医療の IT 化が進み，その結果，かつてない大量の臨床データが電子化された状態でストックされつつある．このデータを有効に利用することができれば患者の生活習慣と疾患の相関（例えば，喫煙と癌）や，薬品とその副作用の相関（タミフルと精神障害）について過去類を見ない大規模な調査が可能となり，臨床研究が加速的に進展するとして高い期待が寄せられている．
しかし，単にデータを電子化しただけで，上記のような革新が実現できるわけではない．この理由の一つは，電子カルテにおいても自然言語で入力される箇所が相当な割合で存在するためである．よって，電子カルテデータを臨床知識としてフルに活用にするためには，自然言語情報を活用することが必須となる．

２．研究の目的

本研究では言語処理技術を用い，カルテのテキスト情報を構造化された情報へ変換する技術を研究／開発し，その技術で臨床現場をサポートするシステムを構築／実証実験を行う．

３．研究の方法

カルテ文章は個人情報の塊ともいえる文章であり，研究利用が困難である．したがって，まず，ダミーのカルテデータを構築する(phase-1)．（図１参照）



確定診断名（主病名および随伴病名）
①肺結核

【主訴】喫煙性の左前胸部痛を伴う左上肺野陰影の精査

【病歴】36 歳のとき、胸部局健診で左上葉の胸部異常陰影を指摘され、T 病院で気管支鏡を施行された。このときの結果は非特異的所見であったという。このときに一時的に禁煙された。
2013 年 10 月 11 日午前 2 時に喫煙性の左前胸部痛を自覚した。夜気時で痛みは増強。その後、断続的に痛み出現。同日、午前中に病医を受診。胸部レントゲンと胸部 CT で左上葉に陰影(網状影を周囲に伴う境界厚性のプラ)を認め、精査目的で当院へ紹介。同日午後に当院を受診した。疼痛時の服用薬としてロキソニン、ムコスタの内服を処方され、10 月 12 日に精査目的のため入院となった。

【現住歴】10 代前半に交通事故（手術なく輸血歴もなし）

【生活社会歴】喫煙 20 本/日 (17 歳-21 歳、26 歳-34 歳)、アスベスト吸入歴なし

【家族歴】母親；気管支喘息

【主な入院時所見】意識清明。血圧 120/80mmHg、脈拍数 94 回/分、呼吸回数 13 回/分、SpO2 99%（室内空気）、眼瞼結膜；貧血なし、眼球結膜；黄疸なし、咽頭；発赤、ラ音軽度せず、心音；雑音聴取せず、腹部；平坦、軟、腸音正常、圧痛なし、肝脾腫を触知せず、下肢；浮腫なし、神経学的異常所見は明らかではない、表在リンパ節；頚部、腋下、鼠径部で触知せず

【主要な検査所見】(入院時検査は添付資料参照)
(2013年10月21日)気管支鏡検査；ファイバーは機種2408を選択し施行。両側気管支系を観察し可視範囲内では異常を認めなかった。左B1+2bより標的へのTBBを試行するも、到達困難でありファイバーを機種LT30へ変更した。標的への到達良好であり、TBB、ブラシを施行し終了した。

【入院後経過と考察】
#1．胸膜炎肺炎
鑑別診断として、腫瘍性肺炎、結核性、膠アスペルギルス症、結核、血管感症候群を考慮した。まずプラ内感染を疑した細菌性肺炎を考え、メロペネム 6g/日3回/日を開始した。しかし、入院時より発熱やCRP上昇の発熱症状に乏しく、細菌性肺炎は除外しメロペネム4日間で終了した。アスペルギルス抗原・抗体はともに陰性、P-ANCA・C-ANCAもともに陰性であり、CT上の所見からは結核が疑われた。精査目的で10月21日に気管支鏡検査を施行した。
10月22日に血液検査でクオンテフェロンが強度陽性であった。当日中に結核内科へ紹介した。胸部CTと気管支鏡生検で病巣形成を認め、結核の診断とした。
10月26日に結核菌核～石検、結核発生届けを提出し抗結核薬の内服を開始した。体重45kgの計算で下記処方とした。とくに薬剤による副作用なく、外来での治療となった。

図１：ダミーカルテの例
（平均的な１枚のカルテ）

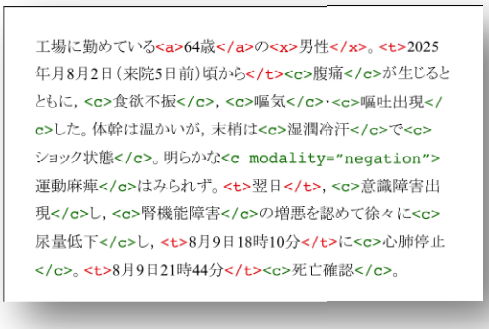


図２：ダミーカルテのアノテーション例

次にデータのアノテーションを行い (phase-2)，これを自動化する言語処理技術を研究開発する．研究対象となる技術は，固有表現特定／表記ゆれ吸収／モダリティ解析といった基礎研究に分解可能で，学術的意義と社会的重要性の両方に貢献可能なよう配慮する．
最後に，開発した技術を用いたアプリケーションを実装し，実験を行う．
　実験は，固有表現の自動特定，症状の特定などを行う．

実験の結果，表１のような精度が得られた．

tag	Precision	Recall	$F_{\beta=1}$
<a>: age	88.89	75.00	81.36
<t>: time	83.08	76.60	79.70
<h>: hospital	96.15	65.79	78.12
<l>: location	0.00	0.00	0.00
<x>: sex	100.00	50.00	66.67
<c>	87.37	71.86	78.86
<c> positive	62.58	62.08	62.33
<c> family	100	59.09	74.29
<c>negation	78.47	66.4	71.93
<c> suspicion	60	20	30

表 1：実験結果

age は年齢，time は日付，hospital は医院名，location は地域名，sex は性別を示す〈c〉は症状名を示す．

図３：匿名化コピー

また，このシステムを実装したハード（匿名化コピー）を開発し，発表を行った．この匿名化コピーは多くのメディアに取り上げられ商品化を複数社と検討している．

４．研究成果

本研究の結果，50 枚のダミーカルテ文章が完成した．このダミーカルテを研究目的にて配布した．また，ダミーカルテをもとに国際ワークショーク NTCIR10 MedNLP を主催した．これは本邦ではじめてのカルテデータを用いたシェアードタスクであり，国外を含め 16 カ国の研究チームが参加した．さらに，2015 年度の開催も決定し，今後の発展が期待される．

５．主な発表論文等

〔雑誌論文〕（計０件）

〔学会発表〕（計７件）

（１）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, In Proceedings of NTCIR-10, 2013.（2013 年 6 月 18 日，東京）

（２）Mizuki Morita, Kano Yoshinobu, Ohkuma Tomoko, Miyabe Mai, Eiji Aramaki: Overview of the NTCIR-10 MedNLP task, In Proceedings of NTCIR-10, 2013.（2013 年 6 月 18 日，東京）

（３）Hiroto Imachi, Mizuki Morita, Eiji Aramaki: NTCIR10 MedNLP Baseline Systemm, In Proceedings of NTCIR-10, 2013. (2013 年 6 月 18 日, 東京)

（４）森田瑞樹, 狩野芳伸, 大熊智子, 宮部真衣, 荒牧英治: NTCIR-10 “MedNLP” Pilot Task：医療分野の言語処理研究の環境整備に向けて, 言語処理学会 第 19 回年次大会, 2013. (2013 年 3 月 13-15 日)

（５）森田瑞樹, 荒牧英治: 医療分野における言語処理研究の環境整備に向けての提案, テキストアノテーションワークショップ 2012. (2012 年 8 月 6 日, 東京). (奨励賞; 16%=3 件/18 件)

（６）荒牧英治, 増川佐知子, 森田瑞樹, 保田祥: 日本人のオンライン・コミュニケーション上での平均使用語彙数は 8,000 語である, 情報処理学会 第 208 回 自然言語処理研究会 (SIG-NL), 2012. (2012 年 9 月 3 日, 仙台)

（７）森田瑞樹, 篠原(山田)恵美子, 宮部真衣, 荒牧英治: 受診理由のコーディング支援ツールの開発, 第 32 回医療情報学連合大会, 2012. (2012 年 11 月 16 日, 新潟)

〔図書〕(計０件)

〔産業財産権〕
○出願状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
http://mednlp.jp

６．研究組織
(1)研究代表者
荒牧 英治 (ARAMAKI EIJI)
東京大学・知の構造化センター・講師
研究者番号：70401073