様式 C-19

# 科学研究費補助金研究成果報告書

平成２３年６月５日現在

機関番号：12601
研究種目：若手研究（A）
研究期間：2008～2010
課題番号：20680006
研究課題名（和文）非文法的かつ断片化したテキストからの情報抽出に関する研究

研究課題名（英文）Information Extraction from ungrammatical and fragmented texts

研究代表者
荒牧　英治（ARAMAKI EIJI）
東京大学・知の構造化センター・特任講師
研究者番号：70401037

研究成果の概要（和文）：

電子カルテが急速に普及しつつある現在，電子カルテ文章からの情報抽出技術が待ち望まれている．しかし，カルテ文章は自然言語で記述されるため，現状では扱いが困難である．本プロジェクトでは，カルテに特化した情報抽出システムの開発を目指す．また，同時に研究利用可能な模擬カルテデータの構築も行い，研究材料として利用可能にする．

研究成果の概要（英文）：

With the rapidly growing use of electronic health records, the possibility of large-scale clinical information extraction has drawn much attention. It is not, however, easy to extract information because these reports are written in natural language. To address this problem, this project aimed to develop an information extraction system that handles medical records. We also develop a dummy text data, which could be a helpful material for the future studies.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| 2008 年度 | 3,000,000 | 900,000 | 3,900,000 |
| 2009 年度 | 4,000,000 | 1,200,000 | 5,200,000 |
| 2010 年度 | 600,000 | 180,000 | 780,000 |
| 年度 | | | |
| 年度 | | | |
| 総　計 | 7,600,000 | 2,280,000 | 9,880,000 |

研究分野：情報学
科研費の分科・細目：知能情報学・1005
キーワード：言語処理・医療情報・テキストマイニング・知識処理

## 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の普及が課題の一つとして掲げられた．以降，急速に電子カルテが普及し，その結果，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．このデータを利用できれば，過去類をみない大規模な統計的な臨床研究が実現可能であり，大きな期待がよせられている．しかし，カルテ中の一部の情報は自然言語で記述されており，データをフルに利用するためには，自然言語処理技術が必須となる．

## ２．研究の目的

本研究ではカルテの一種である退院サマリー（退院時に記述される患者の経過をまとめた文書）を対象とし，そこから患者情報を抽出することを目標とする

３．研究の方法
本提案の目的は，従来とは異なる性質（非文法的・断片化）をもつカルテ文章から「いつ何が起こったのか」という情報を抽出することであり，このために，まず，ダミーのカルテ文章を作成し（STEP1），次に情報抽出対象となる表現（「いつ（時間）」と「何が」（事象））のアノテーションを行い（STEP2），最後に，複数手法をハイブリッドした提案手法を研究／開発し，実験／考察を行う（STEP3）．

４．研究成果

ダミーのカルテ文章を作成を行った．可能な限り，現実的なデータが望ましいため，これは疾患の分布を実際のデータに基づき行った． カルテ文章には，既往歴，現病歴などいくつかのセクションが存在するのが一般的であるが，これらすべてを記述するのではなく，検索要求が高いと考えられる現病歴（患者の過去の病歴）のみを対象とした．
また，次に，作成したカルテ文章に対してアノテーションを行った．カルテデータ中に不要な情報は記述されないと考えられるため，あらゆる事象表現（サ変名詞，動詞）と時間表現のアノテーションを行い，事象表現が疾患を含む場合は，その正式な病名（標準病名），時間表現については正規化した値(”1996-10-31”など)を付与した． 現在，これらの公開準備を進めている最中である．
また，これらを検索するシステムを構築し，日本内科学会や日本循環器学会に提供を行っている．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計２件）
Emiko Yamada Eiji Aramaki, Takeshi Imai, Kazuhiko Ohe: The Internal Structure of a Disease Name and its Application for ICD Coding, Stud Health Technol Inform. 2010, Vol., No., pp.1010-1014, 2010.

Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Mashuichi, Kayo Waki, Kazuhiko Ohe: Extraction of Adverse Drug Effects from Clinical Records, Stud Health Technol Inform. 2010, Vol., No., pp.739-743, 2010.

〔学会発表〕（計１件）
Yasuhide Miura, Eiji Aramaki, Tomoko Ohkuma, Masatsugu Tonoike, Daigo Sugihara, Hiroshi Masuichi and Kazuhiko Ohe: Adverse-Effect Relations Extraction from Massive Clinical Records , COLING 2010 Workshop （In cooperation with Info-plosion） The Second International Workshop on NLP Challenges in the Information Explosion Era (NLPIX 2010), 2010.8.29 中国（北京）

〔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〕
ホームページ等
http://mednlp.jp

６．研究組織
(1)研究代表者
（荒牧英治）
研究者番号：70401073