科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１７　日現在

機関番号：　１４３０１
研究種目：　若手研究(A)
研究期間：　2011～2013
課題番号：　２３６８００１３

研究課題名（和文）表記ゆれ及びそれに類する現象の包括的言語処理に関する研究

研究課題名（英文）Orthographic Disambiguation Incorporating Transliteration, Typo and Other Similar Phenomena

研究代表者
荒牧　英治（ARAMAKI, Eiji）

京都大学・デザイン学ユニット・准教授

研究者番号：７０４０１０７３

交付決定額（研究期間全体）：（直接経費）　　13,800,000円、（間接経費）　　4,140,000円

研究成果の概要（和文）：近年，電子カルテやインターネットに接続可能な健康器具により大量の医療データが利用可能になりつつあり，これらを活用することで，過去例を見ない大規模な統計的研究や，大規模データに基づいた医療支援システムを実現可能であるとして大きな期待がよせられています．　しかし，現状では，電子化された言語データを処理する枠組みがないため，データは活用されるどころか，情報過多を起こし現場の医療者の負担をさらに増しているケースさえあります．以上の背景のもと，本プロジェクトでは，　カルテ文章に記述される疾患表現の表記ゆれを吸収する技術を開発します．

研究成果の概要（英文）：Medical records are increasingly written on electronic media instead of on paper, which has radically increased the importance of information processing techniques in medical fields. Nevertheless, the state of usage of information and communication technologies (ICT) in medical fields is said to 10 years behind that in other fields. By processing large amounts of medical records and obtaining knowledge from them, great potential exist in assisting more precise and timely treatments. Such assistance can save lives and provide better quality of life. The goal of this project is to disambiguate medical terms, especially disease names, in medical records. We also handle transliteration, typo and other similar phenomena.

研究分野：　人間情報学

科研費の分科・細目：　知能情報学

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．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．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る．特に，名詞句の羅列として表現されることが多いカルテ文章においては，名詞句の表記のゆれ（同一概念を指す複数の表記群）を吸収することが重要な課題となる．
本来，漢字／ひらがな／カタカナという 3 つの混合表記を用い，かつ，外来語を多く輸入するする本邦においては表記ゆれは他の言語に比べて大きな障害となっている．これに加え，多忙な業務の間に記述されるカルテでは，略語表記が頻出し，かつ，書き間違え，打ち間違え，英語表記，ドイツ語表記，記号表記が多様され，より一層の複雑さを呈している．現在まで，これら表記ゆれと表記ゆれと類似した現象は，同一概念を指す複数の表記群という点では同じ問題であるものの体系的に整理されていない．

２．研究の目的
本研究の課題は次の 3 点からなる．【1:表記ゆれ解消の研究】【2:表記ゆれと類似した現象の研究】【3:表記ゆれ及びそれと類似した現象の統合】．申請者はすでに 79-85%の精度で自由記載文章からの医療表現抽出及び表記ゆれに成功している[Aramaki2008]．本研究では，まず，この手法を拡張／整理を行う．
次に，表記ゆれと同じく同一概念を指す複数の用語群を扱う．現在は対象として(Ⅰ)同義語, (Ⅱ)翻字ペア, (Ⅲ)略語とその展開型, (Ⅳ)記号／絵文字化，(Ⅴ)書き間違い，を扱うものとし，すでに着手を開始しており[山田 2010][荒牧 2010]，これらの改善と整理（ライブラリ化）に務める．
最後に表記ゆれ及び表記ゆれと類似した現象の統合を行う．これは概念が想起され，記述される過程を，一つの通信路とみなし，それらの中に語彙レベル，音素レベル，文字レベル，打鍵レベルでノイズが混入すると考えることで行う（ノイジィ・チャンネルモデル）．ノイジィ・チャンネルモデルは機械翻訳において近年盛んに用いられており，統計的緻密さを増している．本研究においては，その知見をそのまま用いることができる．
さらに，語彙レベル，音素レベル，文字レベル，打鍵レベルという階層は認知言語学的にも研究する意義のあるモデルだと考える．

３．研究の方法
本申請はデータ構築(phase-0)と研究／開発(phase-1&2)と統合／実証実験 (phase-3)からなる．

平成 23 年度
次の 2 つのフェーズを行う．
【phase0: カルテデータ作成】実際に東大病院にて使用されているカルテを扱うためには，匿名化，倫理委員会への申請など複数の処理を経る必要がありオーバーヘッドが大きい，また，上記処理を経て自院のデータを閲覧できたとしても他研究機関への公開／配布などは事実上不可能である．これでは各研究グループ毎に高いコストをかけデータを構築ことになってしまい，我が国の今後の医療情報研究において非常な負担となる．そこで，本研究では，ダミーのカルテを記述することで，この問題をクリアを計る． 申請者はすでに，挑戦的萌芽研究にて同様の作業を行っており，構築されたリソースを速やかに利用することができる．また，一部の言語現象（略語，書き間違い）においては，ダミーカルテでは出現頻度／性質が実際のカルテと異なることが予想される．このため，倫理申請 を行い該当箇所を抜粋したデータを構築し，統計量のみ現実のカルテを利用する．

【phase1:表記ゆれ解消の研究】
構築したデータを用いて表記ゆれ現象の研究を行う．表記ゆれ研究には次の 2 つのアプローチが存在する．
≪アプローチ 1: 表記ゆれ生成≫ 医療テキスト中の医療表現(疾患名, 症状, 薬品名 etc)１語を入力とし，可能な表記ゆれを生成する技術．
≪アプローチ 2: 表記ゆれ同一判定≫医療用語 2 語を入力とし，それらが同一概念を指すかどうか判定する技術．

本研究では，まず，表記ゆれ生成を行い（アプローチ 1)，その確率モデルを用いて表記ゆれ同一判定を行う（アプローチ 2)．

平成 24 年度 【phase2:表記ゆれと類似した現象の研究】
表記ゆれと類似した現象としては次の 5 つを想定している：(Ⅰ)同義語，(Ⅱ)翻字ペア，(Ⅲ)略語とその展開型，(Ⅳ)記号／絵文字化，(Ⅴ)書き間違い．
これらの現象のうち，一部の現象，（Ⅰ）同義語と（Ⅱ）翻字ペア，は文脈／コンテキス

図 1: 研究成果の配布ページ．医療テキストから用語抽出を行うエクセル・プラグイン．Web ではコンソール版，Windows アプリケーション版なども同時に配布している．

トの影響を受けないため，カルテテキストを用いる必要がない．そこでまずはこちらから研究を開始し，他の現象へ拡張していくものとする．

平成 25 年度 【phase3:表記ゆれ及びそれと類似した現象の統合】
最後に表記ゆれ及び表記ゆれと類似した現象の統合を行う．また，この統合モデルを用いた実証実験を東大病院，及び他施設にて行う．医療アプリケーションの実際に利用に関しては，慎重を期す必要性があるため，初年度から実用に耐えうる汎用的な Java クラスライブラリ(Jstring)の構築を開始するものとする．このクラスライブラリは初年度（平成 23 年度）にて，表記ゆれ吸収メソッドから実装を開始し，最終年度まで逐次的に開発を行う．

４．研究成果

研究成果物を広く医療現場，臨床研究の場で利用できるようにするため，ライブラリをウェブページにて公開した（http://mednlp.jp/~miyabe/MAinNLP/）．
現在，複数の大学病院，医療施設から打診を受けており，普及につとめている．
また，2014，第 33 回医療情報学連合大会（第 14 回日本医療情報学会学術大会），研究奨励賞（3%=数件/約 200 件)など，学術学会からも一定の評価を得た．
また，現在，この技術移転と普及にあたって重要な要素となる技術文書の作成，解説情報の整備にも力を入れている．

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　０件）

〔学会発表〕（計　３件）

① Eiji Aramaki, Sachiko Maskawa, Mai Miyabe, Mizuki Morita and Sachi Yasuda: A Word in a Dictionary is used by Numerous Users, International Joint Conference on Natural Language Processing (IJCNLP2013), 2013 (2013/10/18, Nagoya, Japan).
② Yasuhide Miura, Tomoko Ohkuma, Hiroshi Masuichi, Emiko YamadaShinohara, Eiji Aramaki,Kazuhiko Ohe: UT-FX at NTCIR-10 MedNLP: Incorporating Medical Knowledge to Enhance Medical Information Extraction, In Proceedings of NTCIR-10, 2013. (2013/06/18, Tokyo, Japan)
③ Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, Mai Miyabe, Eiji Aramaki: Overview of the NTCIR-10 MedNLP task, In Proceedings of NTCIR-10, 2013. (2013/06/18, Tokyo, Japan)

〔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
MedNLP http://mednlp.jp

６．研究組織

(1)研究代表者

荒牧英治（Eiji ARAMAKI）
京都大学 学際融合教育研究推進センター

特定准教授

研究者番号：70401073

(2)研究分担者

(　　)

研究者番号：

(3)連携研究者

(　　)

研究者番号：