人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第5回)
SIG-AM-05-12

NTCIR MedNLP-2: 医療分野の言語処理

NTCIR MedNLP-2: Natural Language Processing for Medixal Field

荒牧英治 [1, 2]　大熊智子 [3]　狩野芳伸 [2]　森田瑞樹 [4]

Eiji ARAMAKI [1, 2]　Tomoko OHKUMA [3]　Yoshinobu KANO [2]　Mizuki MORITA [4]

[1] 京都大学デザイン学ユニット
[1]Design Unit Kyoto University
[2] 科学技術振興機構 さきがけ
[2] JST PRESTO
[3] 富士ゼロックス（株）
[3] Fuji Xerox Co., Ltd.
[4] 東京大学
[4] The University of Tokyo

Abstract: Recently, medical records are increasingly written on electronic media instead of on paper, thereby increasing the importance of information processing in medical fields. We have organized an NTCIR-11 MedNLP task for medical records. Our pilot task, MedNLP, comprises three tasks: (1) term extraction for complaint and diagnosis, and (2) term normalization. These tasks represent elemental technologies used to develop computational systems supporting widely diverse medical services. This report presents the objective and data of this shared task.

はじめに

近年，急速に紙カルテから電子カルテへの移行がはじまっている．新規に開業する診療所で電子カルテの導入率は 7〜8 割に上るとされ[1]，今後さらに電子カルテの普及が進んでいくことは確実である．カルテの電子化に伴い，紙カルテの時代には事実上不可能であった大規模な医療情報の利活用が進むと期待されている．この活用先として，例えば次のような例が想定されている[2]:

- 新たな医学的知見の抽出
- 診療行為の結果評価
- 類似症例の検索
- まれな副作用や疾患の頻度の正確な把握

これらの実現のためには，病名や医薬品名などの専門用語やそれに対応するコードの標準化，データ記録形式の標準化などが必須となり，それを効率よく行うために言語処理技術の利用が注目されている．このため, 欧米では 1960 年代から医療分野の言語処理研究が盛んになっている．しかし，同時に他の分野と比べて進歩が遅いことも指摘されている[3]．その理由として，主なものは次の 3 つが指摘されている：

- 入手できるコーパスが不足している
- アノテーション済みのコーパスはさらに不足している
- アノテーションされる場合もその方針が統一されていない

我が国においても, こうした問題意識は同様である．したがって，我が国の医療分野の言語処理を高度化するためには，日本語で書かれたアノテーション済みの文書を研究者が共有できる仕組みが望まれる．

以上のような問題意識から，私たちは研究利用が可能な日本語のアノテーション済み医療文書を構築し，これを用いた解析タスクと共に研究コミュニティに提供する活動を 2011 年より続けている. この活動により，解析技術の客観的な評価を行うことが可能となる．また，医療文書の解析技術の開発に興味はあるがコーパスを持っていない研究者や解析技術を適用する場を模索している企業を集めて産学連携のコミュニティを形成し，課題共有と技術の発展・向上も図ることもできる．

関連するシェアドタスク

さまざまな分野において，実験材料を共有して解析手法の評価を行う，ということが行われている．こうした催しの呼び方はいろいろであるが（sharedtask, contest，competitio など），ここでは「**シェアドタスク**」に統一する．シェアドタスクに参加するグループには実験材料が配られるため，研究者がまず直面する実験材料の入手という壁が取り払われる．また，複数のグループで同じ実験材料を共有することで，手法ごとの解析精度の特徴を評価や議論が可能となる．このような利点から，言語処理分野では，TREC[1]をはじめとして CoNLL[2]や CLEFInitiative[3]，国立情報学研究所による NTCIR[4]，生命科学分野の BioNLP-ST[5]，BioCreative[6]など多くのシェアドタスクが開かれている．

医療分野でも，医療分野の言語処理シェアドタスクとして 2006 年から米国国立衛生研究所(NIH;National Institutesof Health)の主導による i2b2[7]が開催されている．また，TREC は 2011 年から Medical Records Track を開始した．これらのシェアドタスクは英語圏の文書解析技術の向上に貢献している．日本語での医療文書のシェアドタスクはこれまでのところ MedNLP が唯一のものとなる．

MedNLP-2 の概要

タスクの概要

先に挙げた目的を達成するためには，我々は日本語の医療文書を用いた言語処理シェアドタスクを NTCIR-10[8]のパイロット・タスクとして開催した．本年度は以下の２つのタスクを課題とする．

- **病名・症状抽出タスク**：文章から症状や診断病名を抽出する．
- **病名・症状正規化タスク**：抽出された病名と症状に ICD-10 コードを付与する．

[1] http://trec.nist.gov/
[2] http://conll.cemantix.org/
[3] http://www.clef-initiative.eu/
[4] http://research.nii.ac.jp/ntcir/
[5] http://www.bionlp-st.org/
[6] http://www.biocreative.org/
[7] http://www.i2b2.org/
[8] http://mednlp.jp/medistj-ja/

コーパスの概要

本パイロット・タスクのためのコーパスとして，医師によって書かれた患者の病歴要約(medical history)を用意している．ただし，個人情報保護の観点から現実の要約をそのまま研究に利用することは困難である．そこで，疑似的な病歴要約を書き起こしたものを用いる．この際，作成にあたっては，医師免許を持った医師に依頼した．

アノテーションの概要

コーパスに対し，次の２つのタグを付与している

- <t>：日時(time)
- <c>：症状と診断表現(complaint&diagnosis)

さらに，症状に診断表現に関しては，次の２つの属性を付与している．

- **モダリティ**：その症状が実際にあったのか（positive)，または，認められなかったのか(例；胸水なし（negation)），または，家族歴としての症状なのか（family)，単なる疑いであったのか（suspicion）を区別する．
- **ICD コード（疾病分類コード）**： WHO による病名分類コード．アルファベット１桁と数字３桁からなる．病名正規化タスクは，この属性を出力するタスクとなる．

図１にアノテーションされたコーパス例を示す．

本シェアドタスクを達成するために実装されるツールは，様々な医療システムの構築に必要となる基礎的な要素技術である．たとえば，患者にいつ，どのような症状が実際に出現したのかを自動集計できるシステムが構築される．これは医療システムの検索機能を実現するための基盤となるシステムである．

こうした応用アプリケーションの構築，また要素技術のさらなる性能向上のためには，ツールが互換性や再利用性を考慮した形で利用可能であることが理想である．そこで，コーパスデータとその評価器に加え，再配布可能なツールを国際標準 UIMA[9]準拠のコンポーネントとして実装し，Kachako プラットフォーム[10]に統合して大規模処理も含め自動実行できるように配布することを予定している．

[9] http://uima.apache.org/
[10] http://kachako.org/

図 1: コーパス例.

□ **入力データ例**：原文

工場に勤めている64歳の男性。
2025年月8月2日（来院5日前）頃から腹痛が生じるとともに，食欲不振，嘔気・嘔吐出現した。
体幹は温かいが，末梢は湿潤冷汗でショック状態。
明らかな運動麻痺はみられず。
翌日，意識障害出現し，腎機能障害の増悪を認めて徐々に尿量低下し，8月9日18時10分に心肺停止。
8月9日21時44分死亡確認。

□ **出力データ例**：タグ付き

工場に勤めている64歳の男性。
2025年月8月2日（来院5日前）頃から<c icd="R104">腹痛</c>が生じるとともに，<c icd="R630">食欲不振</c>，<c icd="R11">嘔気</c>・<c icd="R11">嘔吐出現</c>した。
体幹は温かいが，末梢は<c icd="unk">湿潤冷汗</c>で<c icd="R579">ショック状態</c>。
明らかな<c icd="G839" modality="negation">運動麻痺</c>はみられず。
翌日，<c icd="R402">意識障害出現</c>し，<c icd="N289">腎機能障害</c>の増悪を認めて徐々に<c icd="unk">尿量低下</c>し，8月9日18時10分に<c icd="I469">心肺停止</c>。
8月9日21時44分<c icd="R99">死亡確認</c>。

おわりに

本稿では，医療分野の言語処理シェアドタスク MedNLP2 について述べた．他の分野と異なり，材料であるコーパスが入手困難である医療分野において，シェアドタスクは，産学連携をすすめるよい装置となる可能性がある．これをきっかけとして，我が国において医療分野の言語処理を担う研究者が増加することを信じている．
また，このような試みは継続的に開催をすることでコミュニティが形成され，さらに開発が促進される．今後の継続開催に向け，様々な方の協力を得ながら努力を続ける予定である．

謝辞

本研究の一部は JST 戦略的創造研究推進事業(さきがけタイプ)「情報環境と人」および科研費補助金(若手研究 A)による．本シェアドテスクの開催にご協力して頂いた NTCIR 事務局および医師，アノテーター，参加者の皆様に感謝いたします．

参考文献

1. 株式会社シード・プランニング, *2011-2012 年版電子カルテの市場動向調査*. 2012.
2. 大江和彦 and 今井健, *臨床医学知識処理を目指した医療オントロジー開発*, in *オントロジーの普及と応用* 2012. p. 131-148.
3. Chapman, W.W., et al., *Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions.* J Am Med Inform Assoc, 2011. **18**: p. 540-543.