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

森田 瑞樹 [1,2], 狩野 芳伸 [3,4], 大熊 智子 [5], 宮部 真衣 [1], 荒牧 英治 [1,3]
[1] 東京大学 知の構造化センター, [2] 独立行政法人 医薬基盤研究所,
[3] 科学技術振興機構 さきがけ, [4] 国立情報学研究所, [5] 富士ゼロックス株式会社
morita@cks.u-tokyo.ac.jp

1　はじめに

近年，大規模データの利活用が様々な分野で注目されている。しかし，我が国の医療分野における ICT（Information and Communication Technology）の利活用は，他の分野と比べて 10 年遅れていると言われている。医療の現場で発生する情報の多くは自然言語によって記述されるため，医療分野において言語処理の技術はきわめて重要である。そこで私たちは，医療分野における大規模データ（特に言語データ）の利活用に向けた環境整備を目指している。

医療分野における情報の利活用

1999 年にカルテの電子的な保存が法的に認められるようになり [1]，診療情報の記録は紙カルテから電子カルテへの移行がはじまった。我が国における 2010 年時点での電子カルテの普及率はまだ 2 割程度に過ぎないが，新規に開業する診療所での導入率は 7〜8 割に上るとされ [2]，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。医療現場で蓄積されるデータの活用先として，大江らは次のような例を挙げている [3]：

- 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，臨床医学知識の体系としてのオントロジーの整備などが必要であり，現在，国や学会などの主導によって進められている [4-7]。

データ利用に伴うクレンジング作業などの負担を考慮すると，データ入力後に標準化などの処理を施すよりも，データ入力と同時に標準化されることが望ましいという発想があり得る。そこで，多くの電子カルテ・ソフトウェアにはカルテの入力を補助するための「テンプレート」が用意されている。しかし，このようなカルテの入力方法は医師の感覚に馴染まず，そのためテンプレートは限定的な使用にとどまっている [3]。よって，カルテには患者の状態や医師の考察などが，それぞれの医師の言葉で記載されており，そこから医学的な知見を抽出して利活用するためには言語処理が必要となる。

医療分野における言語処理

英語圏では 1960 年代から医療分野の言語処理研究が盛んになったが，Chapman らは他の分野と比べて進歩が遅いことを指摘している [8]。その理由として Chapman らはいくつかの要因を挙げているが，主なものは次の 3 つである：

1. 入手できるコーパスが不足している
2. アノテーション済みのコーパスが不足している
3. アノテーション方針が統一されていない

医療文書は患者のセンシティブな個人情報を含んでおり，そのため流通は制限される。また，アノテーション済みのコーパスが無ければ自動アノテーション手法の客観的な評価はできず，教師付き学習もできない。さらに，アノテーション方針が共通していなければ研究グループ間での研究成果の比較が難しくなる。

英語圏よりも医療分野の言語処理研究が進んでいないと考えられる我が国においても，こうした問題意識は同様である。従って，我が国の医療分野の言語処理研究の隆盛のためには，日本語で書かれたアノテーション済みの医療文書を研究者が共有できる仕組みが望まれる。

本研究の目的

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用意し，本コーパス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ことを計画した。こうすることで，解析技術の客観的な評価を行うことを目的としている。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，および解析技術を適用する場を持っている企業を集めて産学連携のコミュニティを形成することで，課題の共有と解析技術の発展・向上を図ることを目的としている。

2 関連研究

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（shared task，contest，competition，challenge evaluation，critical assessment 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

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，解析手法を開発する研究者が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，アノテーション方法や手法ごとの解析精度などの特徴を評価したり議論したりできる。さらに，その分野で現在解くべきタスクが整理・共有される，研究者がその分野に流入することを促す，などの効果もある。

言語処理に関連した現行のシェアドタスクとしては，TREC [9]をはじめとして CoNLL [10]や CLEF Initiative [11]，国立情報学研究所による NTCIR [12]，生命科学分野の BioNLP-ST [13]，BioCreative [14]，CALBC [15]などがある。また言語処理以外のタスクでは，生体分子の立体構造予測を題材とした CASP [16]や CAPRI [17]といったシェアドタスクがその分野においてはよく知られている [18]。

医療分野におけるシェアドタスク

海外では，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 Institutes of Health）の主導による i2b2（Informatics for Integrating Biology and the Bedside）[19]が開催されている [20]。また，TREC [9]は 2011 年開催の TREC 2011 から Medical Records Track を開始した。これらのシェアドタスクは英語で書かれた医療文書の解析技術の向上に貢献している。しかし，現在我が国ではカルテは日本語で書かれることが多く，電子カルテでは特にこの傾向が強いと言われている。そのため i2b2 や TREC とは別に，日本語の医療文書の言語処理技術を培っていく場が必要である。

3 タスクの概要

先に挙げた目的を達成するために，私たち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 [21]のパイロット・タスクとして開催することとした。

ワークショップ型共同研究 NTCIR

NTCIR（NII-Test Collection for IR）[12]とは，国立情報学研究所が 1998 年より開催しているシェアドタスクである。主催者によって用意された共通のデータ（テストコレクション）を参加者に配布することで，参加者のシステム間の相互比較を可能にし，また，研究者フォーラムを開催することで参加者間でのアイデアや技術の交換と移転を促進している。

NTCIR は約 1 年半に 1 度開催されており，第 10 回目となる NTCIR-10 は 2012 年から 2013 年にかけて開催されている。NTCIR の枠組みで個別のタスクを開催するには，タスクの提案をして審査を受けることになる。NTCIR-10 では 6 つのタスク（CrossLink，INTENT，1Click，PatentMT，RITE，SpokenDoc）と 2 つのパイロット・タスク（Math，MedNLP）が開催されることになった。

MedNLP パイロット・タスクの概要

私たちが主催する MedNLP パイロット・タスクでは，配布した医療文書コーパスを利用した次の 3 種類のタスクを設定した：

- **匿名化タスク**
- **症状と診断タスク**
- **自由タスク**

医療文書を研究利用するためには，そこに患者や関係者の個人情報が含まれていないことが望ましい（個人情報が含まれている場合でも，厚生労働省のガイドライン [22]に則っていれば研究利用は可能であるが，その場合には様々な制限がかかる）。そこで**匿名化タスク**は，医療文章に含まれる個人情報を抽出する（個人情報にタグを付与する）タスクとする。

文章から症状や診断病名を抽出することは基礎的ではあるが，様々な応用場面で必須の処理となっており，高い精度が求められる。そこで**症状と診断タスク**は，医療文書に含まれる症状や診断病名などを抽出するタスクとする。

上記の 2 タスク以外に，与えられたデータを用いて何が出来るか，実用的で創造的なアイデアを募集するタスクとして**自由タスク**を用意した。

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）を用意した。

医療文書として真っ先に思い浮かぶのはカルテ（medical record）であるが，カルテは問診の際に記載されるものであるため，最小限の情報が整理されずに時間順に書かれ，また整った文章として書かれないことが多い（体言止めや単語の列挙など）。そのためカルテからの情報抽出は難易度が高い。そこで，私たちは病歴要約に注目した。病歴要約は，たとえば入院していた患者が退院する際や患者を他の医師に紹介する際などに，第三者がその症例を理解できることを目的として書かれるもので，自然言語で記述される（図 1 にその例を示す）。

【現病歴】1994年8月11日頃より全身倦怠感，労作時の息切れを自覚。14日午前3時頃に黒色吐物を嘔吐したため救急車を要請し，当院救急外来を受診。診察中に吐血したため上部消化管出血を疑い，精査加療目的に緊急入院。

【既往歴】20代前半:交通事故（手術なし，輸血なし）。30歳代:胃潰瘍（保存療法）。

図 1. 病歴要約の例

医師は，他の医師によって書かれた図 1 のような病歴要約を読み，図 2 のように情報を抽出・整理して臨床推論を経て診断仮説を立て，それを確かめるための検査の決定や治療方針の組み立てを行う。

【現病歴】1994年8月11日頃より**全身倦怠感**，**労作時の息切れ**を自覚。14日午前3時頃に**黒色吐物を嘔吐**したため救急車を要請し，当院救急外来を受診。診察中に**吐血**したため**上部消化管出血を疑い**，精査加療目的に緊急入院。

【既往歴】20代前半:**交通事故**（**手術なし**，**輸血なし**）。30歳代:**胃潰瘍**（**保存療法**）。

図 2. 病歴要約からの重要情報抽出の例

このように，病歴要約には診察をした医師によって整理された情報が凝縮されており，またその内容には臨床推論に必要となる健康情報と病名などが含まれる。よって，病歴要約は診断・診療支援システムのためのデータ取得先の 1 つとしても用いることができる。

コーパスの作成

生コーパスとして，疾患に罹患している（もしくは罹患が疑われる）患者の病歴要約を複数の医師から収集した。

病歴要約にはその患者本人の年齢や健康情報などをはじめとして，社会生活像，家庭の事情や家族の病歴などのセンシティブな個人情報ないし準個人情報が含まれる。また，医療従事者の個人情報が含まれることもある。よって，個人情報保護の観点からこれをそのまま研究に利用することはできない。そこで，実際の患者の病歴要約を収集するのではなく，疑似的な病歴要約を書き起こしたものを収集した。ただし，医学の知識のない者が書いた病歴要約が実際の患者像を反映することは大変難しいと考えられる。そのため，書き起こしは医師免許を持った臨床医に依頼した。この際，医師に研究の主旨を説明し，研究利用への同意を得た。

アノテーション

架空の患者の確定診断名，症状，現病歴，既往歴，検査所見などが記述されているコーパスに対し，次のような 2 タイプのタグを付与した（括弧の中の数字は開発用コーパス 2,244 文中の各タグ数）：

- 個人情報タグ：
  - `<a>` 年齢（age, 56)
  - `<t>` 日時（time, 355)
  - `<h>` 病院名（hospital, 75）
  - `<l>` 場所（location, 2）
  - `<p>` 個人名（person, 0）
  - `<x>` 性別（sex, 4）
- 医療情報タグ：
  - `<c>` 症状と診断名（complaint & diagnosis, 1,922)

個人情報タグとは，匿名化の対象となる情報（日時，年齢，性別，地名・施設名，個人名など）である。医療情報タグとは，たとえば医師が患者の病歴を理解するのに重要な情報（病名，症状など）である。

図 3 にアノテーション済みのコーパスの例を示す。

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

図 3. アノテーション済みコーパスの例

開催概要

はじめに，本パイロット・タスクの参加者にアノテーションが付与された開発用コーパス（2,244 文）およびアノテーション・ガイドラインを公開した。2 ヶ月間のシステム開発期間を経てテスト用コーパス（1,121 文）を配布し，1 週間以内に各チームからアノテーション結果を回収した。

参加資格は特に設けず，大学，研究所，企業などから広く参加者を募集し，また個人でもグループでも参加できるものとした。参加登録〆切までに国内外から 16 チームの参加申し込みがあった。

各チームによるアノテーション結果の評価は 2013 年 6 月開催のワークショップにて公表する。

4 おわりに

本プロジェクトで提案するシェアドタスクは，日本語の医療文書の言語処理技術が向上することを狙いとしている。これ以外にも，シェアドタスクを通じて産学連携コミュニティが形成され，解くべきタスクがコミュニティで共有される，企業が求めている医療分野の言語処理技術が明らかにされる，医療文書のアノテーション方針が洗練される，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，といった効果も期待できる。

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，コミュニティ駆動型の開発が促進される。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] 診療録等の電子媒体による保存について（厚生省）．http://www1.mhlw.go.jp/houdou/1104/h0423-1_10.html (2013 年 1 月 7 日に取得).

[2] 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.

[3] 大江 和彦, 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*, pp. 131–148, 2012.

[4] 山本 隆一. 医療情報システムの相互運用性 (1) 医療情報システムの相互運用性の意義. *医学のあゆみ*, **221**, 939–943, 2007.

[5] 波多野 賢二, 大江 和彦. 医療情報システムの相互運用性 (2) 医療情報の電子化と用語・コードの標準化. *医学のあゆみ*, **221**, 1013–1017, 2007.

[6] 木村 通男. 医療情報システムの相互運用性 (3) データ形式-HL7, HL7 CDA, DICOM で医療情報システムの標準化. *医学のあゆみ*, **222**, 147–154, 2007.

[7] 大江 和彦. 病名用語の標準化と臨床医学オントロジーの開発. *情報管理*, **52**, 701–709, 2010.

[8] Wendy Chapman, Prakash Nadkarni, Lynette Hirschman, Leonard D'Avolio, Guergana Savova, Ozlem Uzuner. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc*, **18**, 540–543, 2011.

[9] TREC. http://trec.nist.gov/.

[10] CoNLL. http://conll.cemantix.org/.

[11] CLEF Initiative. http://www.clef-initiative.eu/.

[12] NTCIR. http://research.nii.ac.jp/ntcir/.

[13] BioNLP-ST. http://www.bionlp-st.org/.

[14] BioCreative. http://www.biocreative.org/.

[15] CALBC. http://www.calbc.eu/.

[16] CASP. http://predictioncenter.org/.

[17] CAPRI. http://www.ebi.ac.uk/msd-srv/capri/.

[18] 中村 周吾, 森田 瑞樹, 本野 千恵. CASP8 会議参加報告. *生物物理*, **49**, 151-152, 2009.

[19] i2b2. http://www.i2b2.org/.

[20] 荒牧 英治. i2b2-NLP シェアードタスク・ワークショップに参加して. *医療情報学*, **26**, 395–399, 2006.

[21] NTCIR-10. http://research.nii.ac.jp/ntcir/ntcir-10/.

[22] 厚生労働分野における個人情報の適切な取扱いのための ガイドライン等．http://www.mhlw.go.jp/topics/bukyoku/seisaku/kojin/ (2013 年 1 月 7 日に取得).