医療言語処理ツールとコーパスの互換化：
Kachako をベースとした大規模処理に向けて

狩野芳伸*
*科学技術振興機構 さきがけ
kano@nii.ac.jp

1　はじめに

医療言語処理、特に電子カルテの処理は、今後大きな需要が見込まれる。同時に、大量のデータが利用可能になることから、大規模処理の必要性は高い。一方で、医療言語処理の最終的なユーザは医療の現場に関わる方々であり、言語処理や大規模処理の専門的知識が必要となるとハードルが高い。また、計算機が専門の開発者であっても、既存のツールやコーパスを適切かつ効率的に再利用できなければ、多くの時間を本質的でない部分にとられてしまいかねない。そのため、これから新たに作成されていくであろうものも含めた、ツールやコーパスの再利用と共有が必要である。本稿では、これまで開発してきた Kachako システム[1][2]を用いてそういった問題を解決するための計画について記述する。

2　関連研究

医療言語処理に用いるツールは、言語処理ツールとしてそれほど特殊なものではなく、言語処理一般の再利用性や大規模処理についての議論がそのまま適用できる。以下では、主にこれまでの狩野の議論[1][2]を紹介する。

言語処理ツールの標準化、互換性と相互運用

言語処理ツールを用いたアプリケーションの構築においては、多かれ少なかれ既存ツールの再利用と組み合わせが必要となる。アプリケーションが目的でツールは手段に過ぎないのであれば、極力既存のものを再利用すること、そしてその再利用にかかる手間を最低限にするのが最良の方策であるといえる。

ツールの組み合わせを容易にできるかどうかは、相互運用を考慮した標準化・互換化が十分になされているか、またツールの粒度が再利用に適した細かさになっているかがポイントである。また、統一的な定義に基づいて入出力を明確にし、組み合わせ可否の判定をユーザサイドのみで可能にすることが重要である。

再利用性の向上のためには、ツールのインストールおよび実行をできるだけ自動化するとともに、ポータビリティに配慮する必要がある。

これまでのツール開発では、こういった点の配慮が不十分であり、結果として多くの有用なツールが利用されず放置されている。Kachako では、逆にこれらの点を徹底的にサポートすることで、ツールのインストール・組み合わせ・実行・比較評価・結果の視覚化までを全自動で行うシステムを実現している。

テキスト大規模データの処理と個人情報保護

ビッグデータという言葉に象徴されるように、大規模なテキストデータの処理の需要は日々高まっている。本稿の対象とする医療言語処理においても、大規模な電子カルテデータの処理に対する期待は大きい。多くの場合、データが大規模だからと言って、ユーザのやりたいことが大きく違うわけではない。乱暴にいってしまうと、対象となるデータを現実的な期間内に処理できれば、ユーザにとって処理機構の詳細はどうでもよいことである。すなわち、どれだけ容易に、かつ確実に処理が実行できるかがポイントであると考える。

医療情報処理において特に配慮が必要なのは、個人情報保護という点である。大規模処理ではいわゆるクラウドシステムなど、他者の提供する計算資源を用いることが多いが、その場合以下に個人情報保護を確保するかが問題となる。

Kachako では、容易で確実な大規模処理と個人情報保護の双方を同時に解決している。まず、機械学習や検索など一部のタスクを除けば、ほとんどの大規模テキスト処理はドキュメントごとに独立して実行できる。Kachako ではこの点を生かし、要求をみたす標準互換化されたツールであれば自動的に Hadoop/HDFS システム[3]上に展開し実行することを可能とした。Hadoop/HDFS のインストールは誰でも手軽にで

きるというようなものではないため、それ自体のインストールも自動化し、ユーザはサーバ名を指定するだけで全自動実行できるようにした。

もう一つの特徴は、ハイブリッド・クラウドと呼んでいる機構である。Kachako では大規模処理においてもポータビリティを徹底し、任意のサーバクラスタ上で自動インストール・実行を可能にしている。この機構のもともとの目的は、計算資源をユーザ自身の管理下におくことでプロバイダに起因するダウンタイムをなくすと同時に、プロバイダの管理コストを実質的にゼロにすることであった。この仕組みにより、処理対象のデータをもユーザの管理下におくことができるため、医療情報処理のような個人情報保護が必要なタスクにも適している。

NTCIR-10 MedNLP シェアドタスク

我々は NTCIR-10[4]のパイロットタスクとして、MedNLP パイロットタスク[5]を主催している。NTCIR（NII-Test Collection for IR）は国立情報学研究所が中心となって一年半ごとに開催しているシェアドタスクである。MedNLP は日本語の医療言語処理シェアドタスクで、アノテーション付の模擬電子カルテデータを配布する。アノテーションなしの状態からアノテーションの復元を試みる「匿名化タスク」「症状と診断タスク」と、参加者自身の発想に基づく「自由タスク」を設定している。本稿の計画については MedNLP シェアドタスクの一環・発展として位置づけている。

3　医療言語処理の大規模化に向けて

個人情報保護に配慮しつつ、医療言語処理の大規模データへの適用を容易にすることが、本計画の目的である。具体的には、主に MedNLP タスクで作成されたコーパス・ツールを互換・ポータブル・再利用可能にし、Kachako 上で自動実行できるようにする。

コーパス

MedNLP で配布するコーパスのアノテーションは、XML 的な埋め込みタグ形式になっている。Kachako が標準形式として採用している UIMA[6]に対応するには、アノテーションとテキストを分離して保持するスタンドオフ形式に変換するほか、データ型定義が必要となる。そこで、既存の Kachako 互換ツール群やそのデータ型定義と調和するように、必要な諸定義とコーパスの UIMA 準拠リーダ・ライタを作成する。

コーパスは配布できないが

ツール

MedNLP シェアドタスクの開催にあたり、参加者の作成する医療言語処理ツール群は貴重な成果である。シェアドタスクの開催後、ツール群より特に有用なものを選定し、ポータブルかつ Kachako 互換にした上で自動インストール・自動実行可能な形で配布することを計画している。可能であればオープンソースあるいはフリーウェアとして一般に無償で利用できるよう公開したいと考えている。それらのツールと Kachako システムとをあわせて利用すれば、専門的知識がなくとも容易に大規模テキストデータに対し医療言語処理が行えるようになると期待できる。

4　おわりに

自然言語処理を「役立つ」という本来の工学的な目的から考えると、ユーザの負担をどれだけ減らせるかということは非常に重要であるが、研究成果の公開においてあまり考慮されることはなかった。医療言語処理の場合は特に、計算機の専門家ではないユーザも想定する必要がある。本稿で記述した実装計画は、ユーザの負担を極限まで減らすためのものであり、同じ目的の下に実装の進んでいる他の Kachako 互換言語処理ツール群とともに言語処理ユーザの裾野を広げ、社会的な還元を進めることができるのではないだろうか。

謝辞:本研究は JST 戦略的創造研究推進事業さきがけ「情報環境と人」及び文部科学省科学研究費補助金（基盤研究 C）による。NTCIR-10 MedNLP シェアドタスクオーガナイザの森田瑞樹、大熊智子、宮部真衣、荒牧英治の各氏に感謝いたします。

参考文献

[1]狩野芳伸, “Kachako: 誰でも使える全自動自然言語処理プラットホーム,” in 2012 年度人工知能学会全国大会（第 26 回）, 2012.

[2] Y. Kano, “Kachako: a Hybrid-Cloud Unstructured Information Platform for Full Automation of Service Composition, Scalable Deployment and Evaluation,” in the 1st International Workshop on Analytics Services on the Cloud (ASC), the 10th International Conference on Services Oriented Computing (ICSOC 2012), 2012.

[3] “Apache Hadoop.” [Online]. Available: http://hadoop.apache.org/.

[4] “NTCIR-10.” [Online]. Available: http://research.nii.ac.jp/ntcir/ntcir-10/.

[5] “NTCIR-10 MedNLP Pilot Task.” [Online]. Available: http://mednlp.jp/medist/.

[6] “Apache UIMA.” [Online]. Available: http://uima.apache.org/.