

電子カルテから病態や病気の部位などの指定された情報を高精度に抽出する技術を開発 欧州研究学会主催の医療情報抽出技術コンペにおいて、日立チームが精度 86.8%で総合 1 位を獲得

Hitachi India Pvt. Ltd.(社長:飯野 一郎/以下、日立インド)と株式会社日立製作所(執行役社長兼 COO:東原 敏昭/以下、日立)は、このたび、インド情報技術大学ハイデラバード校(学長:Prof. P. J. Narayanan/以下、IIIT-H)と共同で、電子カルテから病態や病気の部位などの指定された情報を高精度で抽出する技術を開発しました。この技術を応用することで、医師などが自由形式で記入した文章や既定の書式に記入された検査数値など、さまざまな形式が混在する電子カルテの分析が可能となります。これにより、将来の病気の発症確率の予測をはじめ、さらなる医療サービスの高度化が期待されます。

なお、本技術は2014年1月から5月にかけて実施された情報検索技術の欧州研究学会 CLEF^{*1}が主催する医療情報抽出技術コンペ「CLEF e-Health 2014 Task2^{*2}」にて、情報抽出精度 86.8%で総合 1 位を獲得しました。また、本技術の内容は2014年9月15日～18日に英国で開催される CLEF 学会にて発表されます。

将来の病気の発症確率予測などを行うには、血液検査結果などの既定フォーマットに入力された情報に加え、医師が記入した電子カルテを分析して得られた情報の活用が必要です。しかし、電子カルテは、医師によってさまざまな表現で記入されるため、必要な医療情報を抽出する技術が必要であり、多くの機関で研究開発が行われています。

本コンペは、多くの医療専門用語を含む電子カルテを患者などが読めるようにすることを目的に2013年から開催されているもので、技術内容によって Task1 から Task3 に分かれて競われます。日立インド、日立と IIIT-H が参加した医療情報抽出精度を競う Task2 には、2014年は企業や大学などの10団体が参加しました。Task2 では、電子カルテから、病気の部位や病態などの医療情報を高精度で抽出する技術が競われ、日立インドと日立、IIIT-H は抽出精度 86.8%で総合 1 位を獲得しました。今回、開発した技術の概要は以下の通りです。

1. 判定規則を自動構築する機械学習技術

対象情報の抽出精度を向上させるため、あらかじめ用意した学習データに基づき、電子カルテに記入された情報から答えを判定する規則を自動構築する機械学習手法を導入しました。さらに文中の情報に加え、電子カルテの章や節タイトルなどの文書構造に関するテキスト情報も考慮することで精度を向上させました。

2. 文脈考慮による後処理技術

機械学習手法で抽出された情報の精度をさらに高めるため、2つの後処理技術を追加しました。最初の処理では抽出された語句の前後の文脈を考慮することで、より正確な情報に変換します。次の処理で

は、心電図検査所見や放射線検査所見など、電子カルテの種類を判別し、その種類に適した医療用辞書を用いることで曖昧な単語の意味を決定します。これらの後処理の導入により、機械学習手法だけでは60%だった病気部位の抽出精度を74%にまで向上させることができました。

また、本技術の開発では、オープンソースの医療テキスト解析プログラム Apache cTAKES^{*3}を活用することで、研究開発の効率を高めるとともに将来の機能拡張が容易になりました。

日立インドと日立は、今回開発した技術の抽出精度をさらに向上させるとともに、本技術を応用して医療サービスの高度化につながる技術開発を進めていきます。

■用語

- *1 CLEF (Conference and Labs of the Evaluation Forum) は情報抽出や可視化などの情報アクセス技術の研究促進を目的とした欧州の学会組織であり、CLEF e-Healthのような技術コンペの実施と年次カンファレンスを主催しています。今年の技術コンペでは CLEF e-Health の他に、ニュース記事推薦や画像認識など 8 件が実施されました。
<http://www.clef-initiative.eu/home>
- *2 CLEF e-Health 2014 は電子カルテからの情報アクセス技術の開発を目的とした技術コンペで、3 つのタスクに分かれます。Task1 は電子カルテを患者に見やすく表示する技術に関して、Task2 は電子カルテから病気に関する情報を抽出する技術に関して、Task3 は病気に関する情報を大量の文書の中から検索する技術に関して競われます。
<http://clefehealth2014.dcu.ie/>
- *3 Apache cTAKES (clinical Text Analysis and Knowledge Extraction System) はオープンソースで提供されている自然言語処理プログラムであり、医療関係の英語テキストを対象としてさまざまな情報を解析することができます。
<http://ctakes.apache.org/>

■照会先

株式会社日立製作所 中央研究所 情報企画部 [担当:木下]
〒185-8601 東京都国分寺市東恋ヶ窪一丁目 280 番地
電話 042-323-1111(代表)

以上

このニュースリリース記載の情報(製品価格、製品仕様、サービスの内容、発売日、お問い合わせ先、URL 等)は、発表日現在の情報です。予告なしに変更され、検索日と情報が異なる可能性もありますので、あらかじめご了承ください。
