

非構造化データを構造化することで、医療情報活用の幅と可能性が広がる

インターシステムズが「iKnow」と呼ばれる自然言語解析技術の日本語対応版を発表した。これは、例えば電子カルテであれば医師の自由入力欄等に記されたテキストデータから二次利用に必要な情報を抽出することができる最新テクノロジーである。日本では2017年1月、同技術を用いたDWHが熊本大学医学部附属病院で稼働予定。「iKnow」の臨床における期待について、同院の宇宿功市郎氏に聞いた。

非構造化データの構造化は情報活用に何をもたらすのか

——医療での非構造化データの活用には、どのような課題があるのでしょうか。

非構造化データは、電子カルテでいえば備考欄や自由入力欄といった医師が自由に入力した文章のようなデータのことであり、従来の技術でもそうしたデータから特定の単語を見つけ出すことは、ある程度できるようになっていきます。しかし、その単語が医療者が期待した文脈の中で使われているかどうかを調べるには、現状では少々手間がかかりすぎます。実際には必要のない無意味な検索結果が多すぎて、それらを整理するだけでも大変



宇宿 功市郎(うすく こういちろう)氏

1981年鹿児島大医学部卒。同大医学部第3内科、同大大学院医歯学総合研究科准教授(人間環境学講座医療システム情報学)などを経て、2006年より熊本大大学院医学教育部医療情報医学分野教授、同大医学部附属病院医療情報経営企画部長。

なのです。

そもそも現状では、ある単語が良い意味で使われているのか、悪い意味で使われているのかも分かりません。例えば「頭痛」という単語が検索で1000個ヒットしたとして、改善した症例に含まれるものが50個、逆に悪化した場合が50個であったと分かれば、その時点で自分に必要な情報は半分に絞り込めます。このように、ある単語がどのような文脈で、さらにいえば、医療者自身に必要な情報として使用されているかどうか分かれば、医療情報の活用は相当進むと思います。

——そうした課題を解決するツールとして、自然言語解析技術「iKnow」はどのような役割を果たすのでしょうか。

「iKnow」は簡単にいうと非構造化データを意味のある構造化データに変換する技術ですが、従来の非構造化データの分析は単語しかヒットしなかった印象が強く、ある1つの事象に対して、ある1つの単語が対応していれば、それだけで分類できたと解釈されていました。しかし、例えば「〇〇先生の発表会」という単語はヒットしても、その発表会が開かれた経緯とか、その後の反響などを調べようとすると、結局はその単語が含まれる文章を全部読み返さなければならなかったりするのが実状といえます。あるいは、発表会の中で良かった点、悪かった点を検索しようと思ったら、いくつかの検索ワードを組み合わせて命令を作り、それを入力しなければならなかったのも確かですね。つまり、辞書とかオントロジーといったものを、自分でメンテナンスなり設定なりする必要があったわけです。「iKnow」を活用すると、そうした手間をかけなくても求める文脈の中で文章を分析し、自分に必要な情報を抽出できるようにになります。それゆえ今後、医療における非構造化データの分析が大きく進むきっかけになるのではないかと期待しているのです。

非構造化データ活用の現状と課題、その解決法

——熊本大学病院における非構造化データ活用の現状についてお聞きします。

正直なところ、あまり活用が進んでい



あるニュースサイトでの記事で「オリンピック」を検索した例。関連するエンティティ、近似エンティティ、元の文章などを表示する

ないのが実状です。私自身も、以前、病名ごとに単語の出現数を調べてみたことはありますが、それを調べたからといって、自分が書きたい文章が出てくるわけではないので諦めました。カルテの中のある表現から、自動的に新しい文章が生成できれば、役に立つでしょうね。

また、診断書の作成は現状、医師の負担の1つになっていて、それを軽減するために医療クレークが補助していますが、これなどもある程度自動的に生成できるようにになれば、医師の業務もクレークの負担も軽減できます。カルテに入力した内容に関する記述が自動的にリストされて、他に流用もできて、さらに細かい単語のチェックなども不要になれば、大幅な効率化につながるでしょう。

さらに、例えば「薬を処方した」、「検査した」、「結果が良かった」という段階があったとして、各段階の情報は構造化されたデータを見れば容易に把握できます。しかし、それぞれの段階を結ぶ「中間」の情報、そこそが非構造化データなわけですが、その部分の情報が現状では乏しいのです。つまり、医師の判断であるとか、治療に対する患者さんの反応といった内容ですね。

最近では厚生労働省からの指導もあり記載はされているのですが、どんな内容をどのような形で記載しているかまでは、まだ十分に分析できていません。そうした分析が可能になれば、症例によって医師の判断と患者さんの反応にどのような

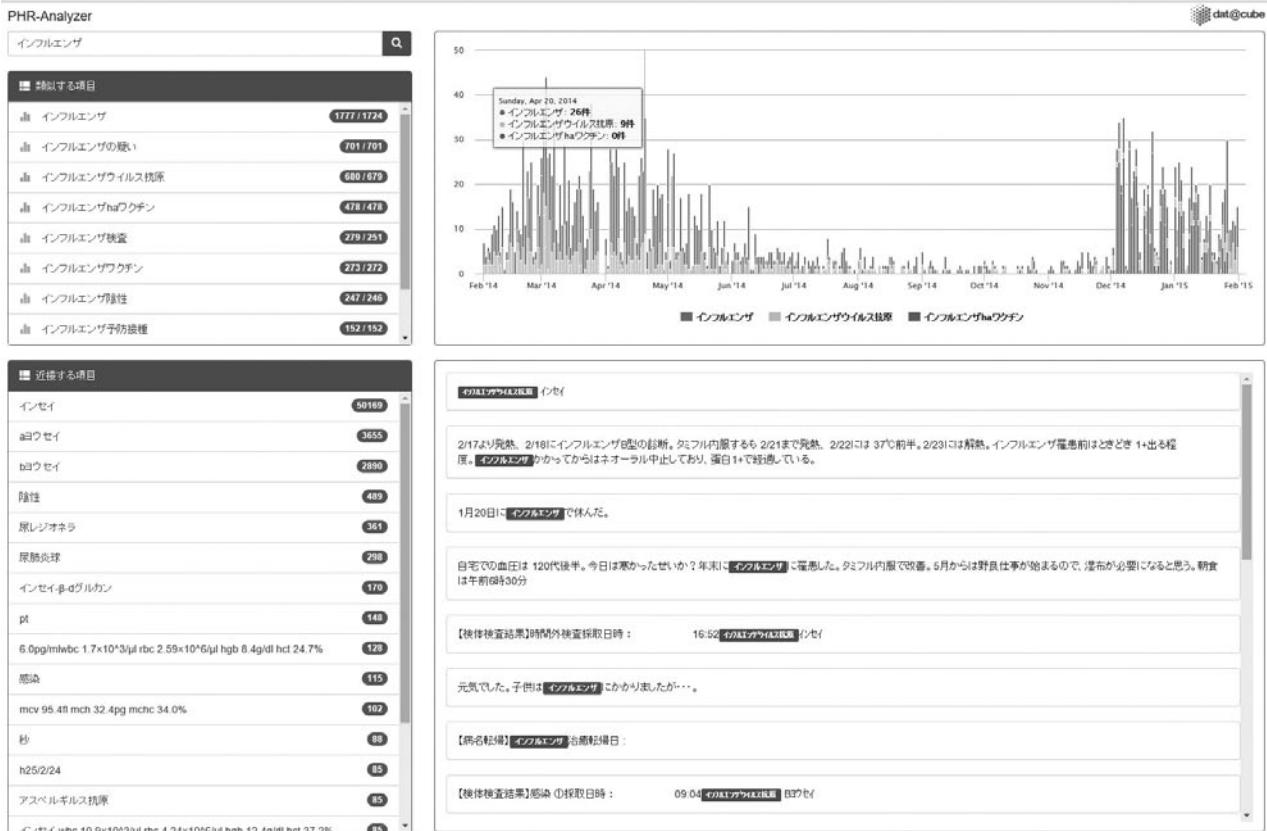
傾向があるのかが、ある程度見えてくるのではないのでしょうか。

——現状の課題に、「Know」はどのような解決策を与えてくれるのでしょうか。

例えば免疫性疾患などの場合、薬の効果で症状が改善しているのは明らかでも、患者さんは服薬を辛く感じることも多いようです。それを例えば朝と晩で飲む量を調整するといった医療現場の工夫で、患者さんの生活の質を改善したり、不満度を下げたりしているのが現状です。しかし、そうした工夫は「言い伝え」のレベルにとどまっています、はっきりとしたエビデンスにはなっていません。なぜなら、これまでのITの仕組みの中では、対象が煩雑すぎて検証にまで至らなかったからです。

このように現場のコミュニケーションの中から出たアイデアのようなものも、「Know」でカルテの記述をさかのぼって検証していくことで、データとしてうまく浮かび上がらせることができるかもしれません。経験的には知っていて現場でも活用していた手法なども、数値的に傾向が出れば、応用の幅も広がるのではないかと期待しています。

もちろん、現在も必要項目はカルテにきちんと入力されていますが、それ以外の「ちょっとしたこと」もテキストの形で書き加えておくと、「Know」により後で検索をかけたときに、リストとして出てきて、一目で確認できるわけです。そこから、何か新しい「気づき」が得ら



電子カルテの自由記述で「インフルエンザ」を検索した例。類似する項目、近接する項目と件数を表示。さらにグラフ表示も可能 (データキューブ社提供)

れる可能性は大きいと思います。
「Iknow」はそのような形のテキスト有効活用の支援ツールと成り得ると考えます。

「Iknow」活用の成果として 医学の進歩や治療の質向上を期待

——「Iknow」活用の成果をクリティカルパスへフィードバックすることなども可能になるのでしょうか。

例えば一定の範囲の患者さんがいるとして、そこには想定内の「外れ」(※パス通りに治療を進めた結果、当初想定した治療効果がみられないケース)の領域も含まれます。そうした場合に「Iknow」を活用して過去の非構造化データを検証することで、ある程度の「外れ」まで含めて、診療の自動化までは望めないとしても、現行のパスをもっとうまく運用するといった改善は可能になると考えています。また、同じ病名、同じ薬でも、副作用の有無をはじめ、カルテに記載されている内容が本当に患者さんの変化に有効だったかといったことも、もっと明確化できると思います。

実際問題として、パスを単純に適用でき、かつ何の問題もなく治療に至るケースはあまり多くはないと思われます。パスを適用した患者さんが退院して本当によくなっているのかどうかなどのデータを蓄積していけば、今後のパス自体の改善にもつながると思います。それは当然、病院長の面においてもよい影響を及ぼすでしょう。

——今後、どのような分野で「Iknow」の技術を活用したいとお考えですか。

「Iknow」は既存の電子カルテシステムの中で利用できるのを使い勝手がよく、現場の医師らが特別なトレーニングを受けなくても使えるレベルに達していると思います。従来は、何かを調べたいと思ったらデータ抽出の段階から専門家に依頼していましたから、この点は大きな進歩です。現場で扱えるということは、どの診療科でも、医師や看護師がいつでも振り返りができるといことです。それによって課題や改善すべき点が見つれば、より深く調べていけばよいのです。

将来的には、「Iknow」の技術応用により、もっと長い文章も分析できるようになるでしょう。特に似た意味の単語を集約できるようにすると便利です。

現状では、同じ病名でも診療科内と保険とDPCとでは記述が違う場合もあります。そうした単語を同じものとして集約して扱えるようになれば、現場は大変助かります。あるいはWHOによるICD10/11にしても、国内の医療現場とは乖離している場合もあり、標準とは違ってもカルテに書いておきたいことも出てきます。そうしたデータを後で振り返って整理できれば、必ず役に立つはずだと思います。「Iknow」というツール活用を前提に考えるなら、今は無駄と思っただけでもありあらず入力しておき、後で分析できるように備えておけばよいのです。データウェアハウスに蓄積したデータを、「言葉」

文書などの非構造化データを構造化データに変換する最新技術

「iKnow」は、人間が書いた一般的な「文章」を、整理・蓄積や分析に適した「データ項目」に変換する自然言語解析ツールである。ここでいう一般的な文章を「非構造化データ」、データ項目を「構造化データ」と捉えると、「iKnow」は「非構造化データを構造化データに変換するツール」ともいえる。

電子カルテであれば、病名や症状、薬剤の名称など、選択肢から選ぶようなものが構造化データ、それ以外の、例えば所見などを文章で入力したものが非構造化データに該当する。一般的に、こうした非構造化データは構造化データに比べて検索性が低く、従来は検索できても短い単語の羅列にすぎないなど、内容の分析や二次利用がしにくい状況にあった。「iKnow」は、従来のツールのように文章を単語に分割するのではなく、言語構造に基づいて、文章中から「意味のある一連のまとまり（これを「エンティティ」と呼ぶ）」を見つけてることによって解析を行う。この「ボトムアップ手法」によって、解析に先立ちあらかじめ辞書やオントロジーを設定しておく必要がなくなった。

なお「オントロジー」とは、文章を解析しやすくするために、文書に対して情報処理上の「意味づけ」を行う構造のことを指す。簡単にいえば、コンピューターのための用語辞典や文例集をまとめたものと考えればよく、従来の解析ツールでは辞書だけでなくオントロジーも常に設

定やメンテナンスが必要であった。これに対し「iKnow」は、文章中から最大単位のエンティティを自動的に認識し、エンティティ同士の近似性や優位性を算出することによって、文章の意味を踏まえた、よりユーザーの目的に沿った検索を可能にする。例えば、同じ症状に関する記載でも、それがどのような意味で書かれたか、ポジティブな意味かネガティブな意味か、といった点まで踏み込んだ検索ができるのである。

例えば下の図のような文章があるとする。従来のツールでは、検索時に単語単位に分割されてしまうため、その単語がどのような文脈で使用されているのかが判明しない。これに対し、「iKnow」はエンティティを単位として解析するため、文章が話題としている内容まで分かり、よりレベルの高い検索や分析が可能となる。

一般的な形態素解析ツールとiKnowによるテキスト解析の例

【例文1】 敗血症は腎盂腎炎から至ったケースという。

ツールAによる分かち書き：

敗血症／は腎盂／腎／炎／から／至っ／た／ケース／と／いう／。

ツールBによる分かち書き：

敗血／症／は／腎盂／腎炎／から／至っ／た／ケース／と／い／う／。

iKnowによる解析結果：

敗血症 は 腎盂腎炎 から 至った ケース という。

※iKnowは、細かい単語単位ではなく、意味あるフレーズ「敗血症」「腎盂腎炎」「ケース」を「エンティティ」として抽出し、敗血症と腎盂腎炎は何らかの関連があることを事前の定義や知識なく推測する。

を中心にも多様なデータと結びつけることで、「いま私たちが正しいか？ 患者さんにとっての正解は何なのか？」が分

かってくれば、それは医療者にとって大きな成果です。「iKnow」は、そうした成果が得られる期待を抱かせてくれるツールと感じています。