情報知識学会関西部会2008年度第2回（通算第8回）究会報告

日　時：1月31日（土）14時半～17時
会　場：大阪市立浪速人権文化センター
論　題：医学生物学分野におけるテキストマイニング技術の展望
発表者：小池麻子氏（日立製作所中央研究所）
共　催：日本図書館研究会情報組織化研究グループ
後　援：科学技術振興機構、情報科学技術協会
出　席：22名
内　容：

１．テキストマイニングの必要性

・文献数の急速な増大に伴って情報検索や知見の抽出が困難になってきており、テキストマイニング技術を用いた自動解釈の必要性が増している。特に医学生物学分野においては、数式や化学式をほとんど用いず自然言語の表現に依存する割合が高いこと、ゲノム解読の進展に伴って多くの遺伝子を同時に対象とする大規模実験研究が行われるようになったこと、等から他分野に比しても重要性が大きい。
・医学生物学系は、米国医学図書館のPubMedによって網羅性を持った抄録レベルの情報が得られる。近年は論文本文のオープンアクセス化も進んでおり、マイニングを行う基礎条件が整っている。また、MeSH（Medical Subject Headings）やUMLS（Unified Medical Language System）などのシソーラスも利用できる。
・文献中の断片知識を利用して、仮説生成や知識発見を行う技術を追求している。これには、概念／用語の認識、概念間の関係性の抽出とその利用が必要である。

２．連想検索

・連想検索とは、ユーザが選択した文書群から特徴語群を抽出し（文書－単語連想）、データベース全体から特徴語群に関連する文書群を抽出する（単語－文書連想）もので、潜在的知識発見をマニュアルで行うものといえる。特徴語の関係性を視覚化するなどして利用者を支援するシステムを開発している。

３．辞書構築と概念認識

・文献から情報抽出を行うには、文章中の概念・用語を認識し、構文解析によって概念関係を明らかにする必要がある。語彙の問題と構文解析の問題の両面で、処理を円滑に行う手法の開発を行っている。
・特に、遺伝子及び蛋白質の名称は、同義語の多さ、異なるものが同一の名称を持つ曖昧性、スペリングのバリエーション、紛らわしい名称など多くの問題をはらんでいる。一定のコア情報から自動収集を行って成長していく半自動収集型の遺伝子辞書（GENA: Gene Name Dictionary）を開発している。
・機能用語の収集も行っている。遺伝子配列のアノテーションのために開発されたGene ontologyをベースに、高い共起性を持つ「関連語」、類似の局所文脈（collocation）を持つ「類似用語」の抽出等によって充実化をはかった機能用語辞書を開発している。

４．情報抽出

・構文解析によって、遺伝子／蛋白質／化合物の相互作用、遺伝子／蛋白質と機能との関係、遺伝子／蛋白質と疾患との関係の情報を抽出している。具体的には、遺伝子等の名称や機能用語を認識した後、品詞と係り受け関係をパーザーで付与し、構文解析によって特定のActor－Objectの関係を抽出する。
・相互作用に比べ、機能の表現には曖昧さや多様性が大きく、抽出ははるかに難しい。疾患との関係も曖昧な場合があるが、共起情報で何とかなる側面もある。
・現在、約300万の相互作用と約34万の機能情報を抽出し、データベースとして保持している。

５．仮説生成と知識発見

・膨大な文献中に埋もれている潜在知識を抽出することをめざしている。具体的には、概念A－概念B及び概念B－概念Cの関連が文献中に明示されていることを探知して、いまだ研究されていない概念A－概念Cの関連を発見するものである。あらかじめ概念AとCを指定して予測するClosed discoveryと、概念Aのみを指定するOpen discoveryがある。なお、実際には既知の研究で示された関連にはそれぞれ固有の条件があるため、新たな関連の発見は仮説生成と考えたほうがよい。
・この種の研究の初期のものとして、Swansonが予測したレイノー病と魚油の関係の発見（1986）があり、これは後に実験で証明が得られた。Swansonは人手で発見を行ったが、近年は自動処理による

研究も出てきている。

・単純に関連をつなぐだけではノイズが多く、解釈も難しい。発表者らが構築している潜在的知識発見支援システム BioTermNet では、UMLS 等のシソーラスを用いた意味クラスの指定、半自動収集によって構築した辞書を用いた用語の統一、概念間関係を抽出したデータベース、等によって精度の向上をはかっている。

・概念間の関係性を構文解析による情報抽出のみに依存することは難しく、統計解析も併用している。2 項関係の抽出には様々な手法が考えられるが、経験的な数式のほうがよい結果が得られるようである。

・Open discovery を目的とした場合の手法の評価は、現在知られている関係性を、発見前の段階での文献情報を用いて予測できるかどうかで行う。

・BioTermNet の利用実験では、ユーザの知識レベルによって問題が生じることも明らかになり、ユーザの知識を反映したシステム改良を行っているところである。

発表後、日本語文献への応用や辞書開発の機械処理とマニュアル処理のバランスなどについて質疑応答があった。

参考：小池麻子「テキストマイニングによる潜在的知識の発見支援」『情報処理』48(8), 2007. p.824-829