生命科学辞書とオントロジーの自動構築法の開発

●大久保 公策
国立遺伝学研究所 生命情報・DDBJ 研究センター 遺伝子発現解析研究室

＜研究の目的と進め方＞

ゲノムワイドな測定などの探索的なデータをあらゆる知識を動員して解釈したり、マイニングしたりするために大量のテキストを内容にしたがって整理分類し、またそれらの内容を研究者に簡潔に表現したりするための仕組みを開発してきた。

同じタスクを行うためにはMeSH用語やＧＯ用語などの宣言的に作成した構造化索引用語を人手で付与する方法がとられ一定の水準での目的を達成している。

このきわめて古典的な方法は索引用語集と構造の絶え間ない更新と新規文書への絶え間ない索引付けを必要とし、文書の爆発に対応できる機械化が必ず必要であると考えている。

本研究では専門家による古典的な当該分野知識の濃縮整理の結果である教科書が構造として持っている、相互独立なトピック配列と標準化された用語の分布を材料として用い、当該領域の索引付けのための構造化用語集を自動生成し(オントロジーの自動生成) それらを機械的に質問文書にマップしたのちに教科書中の各ページとの類似度の計算を通じて各文書を整理し、また独立トピックとの類似パタンとしてその内容を表現するものである。

＜2008年度の研究の当初計画＞

今年度までに大まかな仕組みを完成し試験公開サイトを通じてチューニングとデータの打ち込みミスの修正の段階にはいっているが 本年度は特にアイデアの提出とサービスの提供に終わらず「再利用可能な様々の言語資源」を残すべく手作業による用語辞書の手直しなど基礎固めとその公開にも注力する。

＜2008年度の成果＞

用語の重み付けについて遺伝子の機能の表現を例に研究を続けた。チューニングを行ってゆくにあたり使用する遺伝子列として前年まではマイクロアレイデータやゲノムデータにより序列化された遺伝子を使用していたが今回はより確実な検定のために「同じファミリーに属する遺伝子群がいかに機能表現を受けているか」に注目した。 すなわち個々の遺伝子の機能はファミリーに共通な部分(generic function) とメンバー間で違う(specific function)が知られているはずであるが、前者と後者に区別してそれぞれがどの程度強く正しく表現されているかに注目し目視検定を行った。

まず総てのヒト遺伝子機能を新たにEntrez Geneの参照文献の要旨の連結文およびサマリー記述に含まれるＢＯＢ辞書用語で用語ベクトル化し、２５冊の平均評価書空間に投射し、各ページとの余弦パタンによってそれぞれの遺伝子機能を表現した。一方で検定を容易にするためにInterProを用いヒト遺伝子の約半数を何十にも重複を許すファミリーの組にまとめ、それぞれから前年度までに作成していたＢＯＢ教科書マップ表示へ直接移行できるインターフェイスを作成した。

目視による検定は始まったばかりであるが 一般的に遺伝子機能はファミリー内では区別されて表現されておらず、ＧＯでも同じ傾向を大学院生が指摘している。結局 ファミリーメンバー間の機能の違いがEntrexに記載されているケースも必ずしも多くはない印象を得ている。

これは、遺伝子機能の定式表現において一般に注意すべき問題である。

そもそも遺伝子の機能記述については多くの教科書ではファミリーの存在を意識せずに書かれている場合も多く、総説でも明確な区別なくさらには別の生物由来のものも混在している。

将来、両方の機能を別々に明確に表現する方法を確立することを遺伝子機能の定式表現においては一つの目標にするべきであると確信した。

まずこの原因が Entrex Geneの文献の選び方にあるのか それともＢＯＢの用語ベクトルの重み付けにあるのかをシステマティックに検討するために、ヒト遺伝子を総てＰｆａｍを頼りにファミリー化してそれぞれのファミリーを１セットとして基本教科書総てに対する関連を計算しデータベース化した。

現在は ファミリーごとの機能パタンを一望しながら メンバーが区別できているケースと区別できていないケースの特徴を目視で検討している。

＜国内外での成果の位置づけ＞

開始時に比べるとマイクロアレイの凋落に加えてＧＯの流行がはなはだしく、ＧＯ以外で遺伝子機能の表現を行う研究はほとんど見られない。

しかしながら ＧＯの表現にもさまざまな課題があり、遺伝子機能の自動表現は完成すれば十分に役立つものと考える。

また遺伝子に限らず あらゆる文書をベクトル化して教科書ページとの関連性パタンで表現する方法は文書検索や、情報取得へと一般化できる。

検索後の数百のヒッツの中身の表現はＶＩＶＩＳＩＭＯに代表されるクラスタリングとグループの名づけの巧みさの問題とするのが一般的である。

ＢＯＢではクラスタリング後のグループの名づけは絶対知識座標である教科書の項目見出しで行うわけであるので対象文書世界によってはＶＩＶＩＳＩＭＯ的手法をしのぐと期待される。

＜達成できなかったこと、予想外の困難、その理由＞

手作業でしか誤りの検出やチューニングの評価が出来ない困難がおおきい。

ＢＯＢ辞書中での教科書間での表記のわずかなズレによるマージしそこないが多く見られるがシステマティックな検出とデータの修正が進まなかった。

ＢＯＢ辞書の二つの用語を一つの意味であると認めたとき、用語空間を変更せねばならないが、空間作成時のＳＶＤだけでなくあらゆる遺伝子やPubMedアブストラクトの折込計算をやり直し、出力パタンの画像を総て生成しＤＢに記録するという工程をすべてやり直すという作業になり、現在の資源ではやく一週間の時間を要す。

＜今後の課題＞

上記の問題をクリアすることが今後の課題と考えられる。

＜成果公表リスト＞

1）論文／プロシーディング（査読付きのものに限る）

なし

2）データベース

なし