2A3-01

ことばの処理と生命科学

情報抽出、オントロジー、テキストマイニング

Natural Language Processing for Biology – Information Extraction, Ontology, Text Mining

辻井 潤一
TSUJII, Jun-ichi

東京大学大学院情報学環、情報理工学系研究科・コンピュータ科学専攻
Graduate School of IIS and Graduate School of IST, University of Tokyo

Unlike physics, there are scientific fields in which language plays important roles, much more than mathematical formula. Biology, medicine, economics, etc. are such fields. Researchers in these fields use language as primary means to introduce scientifically meaningful classifications in their domains to organize knowledge. In particular, after deciphering genome sequences, researchers in molecular biology are interested in identification of function of proteins encoded by genome sequences. They have to re-organize their knowledge about proteins already known to the community, because functions of proteins can only be identified by understanding a whole network of protein-protein interaction and this is not a simple task. Protein-protein interactions known to the community are actually embedded in a huge collection of papers already published or to be published. In this talk, I will discuss how NLP techniques can help biologists to challenge this enormous task. I also argue that we can use the same NLP techniques used for information exploitation in other domains.

1. 言語処理、情報管理、知識発見

言語処理の応用は、これまで仮名漢字変換とワープロ、機械翻訳システムといった、言語処理単体のものが多かった。しかし、Web中でのテキスト情報の爆発的な増加、ウェアラブル・ユビキタスコンピューティングなど、周辺技術の発展に伴い、言語処理を一部として取り込んだ統合的なシステムが開発されようになってきている。今回は、言語処理が、情報検索、テキストマイニング、知識発見などの分野でどのような貢献をするかを、生命科学を例にとって議論する。

2. 言語科学の革命と生命科学の革命

言語科学、生命科学は、近年、ともに大きな方法論的な変革を遂げてきた分野である。まずは、この2つの分野の現況と、なぜ、この2つの分野がいまつながりつつあるかを考える。

2.1 言語科学の革命

言語処理は、観察できる言語という対象を、知識とか情報という抽象的なものと結びつける技術である（図1）。知識・情報が物理的な実体から乖離しているために、言語処理技術は、たとえば、音声技術に比べても、思弁的な傾向が強かった。

また、言語は表面的に観察できるとはいえ、印刷されたテキストは、それ自体操作の対象とはならず、実証的方法論がとりにくい分野となっていた。言語科学の主体は、理論言語学や狭義の計算言語学のような合理主義的な潮流であった。

この状況は、計算機利用の急激な変化から、ここ十数年で大きく変化している。いまでは、数百万語、数兆語のテキストが計算機に蓄積され、プログラムで操作できる。このことが言語に関する実証主義的な研究、技術の発展を可能にし、いまでは、コーパス言語学や機械学習の手法を使った処理手法が主流となってきている。言語科学、言語処理の方法論的な革命である。

しかも、この変革に続いて、大規模辞書、シソーラスなど、知識資源も計算機に大量に蓄積されるようになり、現在では、非常に大規模な分野知識の蓄積も行われている。すなわち、抽象的な存在であった知識も計算機の対象となることで、大量の知識を使った言語処理、あるいは、テキストからの知識構築、知識獲得といった研究が可能になってきている。

1980年代の AI 研究者が夢想した、知識にもとづく言語理解の研究が、本格的にできる時代、合理主義と経験主義の方法論を融合した言語科学を構想できるようになった。方法論の変革の第二段階である。

とくに、後述するように、生命科学では、言語処理への要求と同時に、大規模知識の構築への要求も大きく、変革の第二段階にある言語処理研究に格好の研究の対象を与えてくれる。

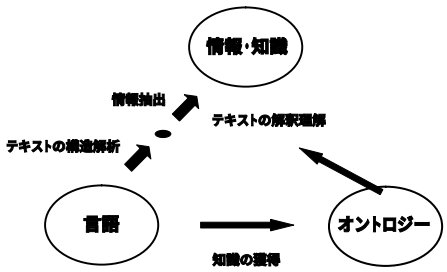

図1　言語・知識・情報

2.2 生命科学の革命

DNA の発見とそれに続くゲノム系列の解読は、生命科学の革命であった。あらゆる生命体は、DNA 配列を通して関係し、独立した研究と考えられていた個別生物の研究が相互に関係していることを明らかになった。また、生命現象を物質の間の、あるいは、たんぱく質間の相互関係で説明しようとする分子生物学が遺伝学と結びつき、同時に、現象面をみる病理学、医学とも結びつくことで、縦割り的な研究分野の壁が破れつつある。

連絡先：辻井潤一、tsujii@is.s.u-tokyo.ac.jp

生命に関する学問分野の知見の総合化によって、生命現象の総体を理解する時代を迎えている。

生命現象は、物質（たんぱく質、生体分子）間のインターラクションの連鎖（ネットワーク）によって生じている。しかし、生命現象の総体は、個々の生命体の、個々の部位、その部位中での細胞、また、細胞中の個別的な場所で生じる物質間の相互関係ネットワークを、個別に実験的に解明していくという方法論だけでは、捉えられない。この素朴な方法論では、禁止的な量の実験が必要となると同時に、科学としての一般化された理解に到達できないからである。

現実的には、膨大な数のたんぱく質（10万種）の性質、あるいは、その対の反応可能性に関する既存研究の成果を活用することで、個別ネットワークの機能をなるべく少ない実験で解明できなくてはならない。あるいは、既存のネットワークの知見から、予測性、説明性をもつ一般的な知識の系を構成することが、科学として不可欠である。

ポストゲノム時代の生命科学は、人工知能研究の観点からは、既存知識を再活用する知識共有、知識管理、あるいは、既存の知見の統合からあらたな知識を作り出す知識発見の研究を行う格好の場となる。

同質の問題は、生命科学の別の分野にも現れる。既存患者の病歴、治療経過、あるいは、症例報告を蓄積し、これを参照することで、個々の患者に対する診断や有効な治療方法を見つけていこうというEBM(Evidence-Based Medicine)においても、テキスト中に埋もれた情報をいかに有効に検索、活用するが鍵となる。さらに、SNP の研究に見られるように、遺伝子->たんぱく質->たんぱく質ネットワークという方向からの生命現象へのアプローチと、EBM とが結びつき、生命-病理現象の全体像を明らかにする研究も構想されている。

これら生命と病理に関する科学分野では、対象の分類(Classification)、構造把握(全体と部分など)という、「言語」的なものがその方法論の中核にあることから、この知識の発見と統合の試みのなかで、言語科学と言語処理の関与がつよく期待されている。

3. シグナル･パスウエイ

3.1 ネットワークとその部品

システム生物学の典型的な問題を、図2に示す。この図は、細胞外からの情報が細胞内にどのように伝達されて、細胞が外部からの情報に適切に反応するかを示したものである。

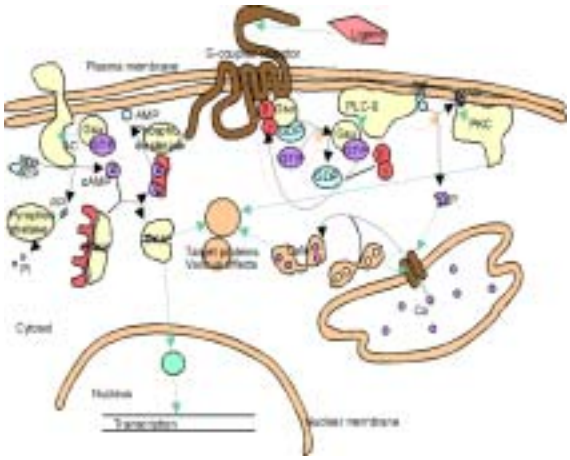

図2　典型的なシグナル･パスウェイ

すでに述べたように、図2のような伝達系は、一つ一つ生物実験で解明されていくわけではない。むしろ、既存のネットワークに関する知見を参照すること、ネットワークに関する仮説が立てられ、その仮説を検証するために生物実験が行われる。

仮説の構築、検証が可能となるためには、既存ネットワークの知見が分節化され、それを構成する部品に関する知識として整理されている必要がある。ここでネットワークの部品とは、

(1) 個々のたんぱく質や生体分子
(2) たんぱく質間の関係、たんぱく質と生体分子の関係
(3) 複数の対がつながったパス（進化の過程で保存されたモジュール）

などが考えられる。(1)、(2)は一見自明な単位である。しかし、これらの単位も、実際には、分子生物学という特定の科学分野が自らの知識を構造化するために持ち込んだ単位、オントロジカルな単位であることは忘れてはならない。

オントロジカルな見方は、(3)により顕著に現れる。異種生物間に保存されるモジュールという見方は、ゲノム科学が分子生物学のオントロジーに与えた影響とみることができる。このような対象の分節化が、実りある知識の構築に結びつくかどうかを、既存の文献情報や事実情報を駆使して検証することが、この分野でのテキストマイニングの課題となる。

3.2 知識部品の再利用と情報抽出

ネットワークを構成する単位が同定できると、(1)それらの単位に関するに個別知識を集積していくこと、(2)個別知識を一般化して再利用することが、次の課題となる。

実際、生命科学分野では、たとえば、個々のたんぱく質に関する情報を集積したSwissProtのようなデータベースが数多く作られている。これらのデータベースは、研究者が論文を読み、そこでのデータを人手で整理する形で構築されており、言語処理における情報抽出（IE）の格好の研究課題となっている。

また、生命科学における情報は、きわめて文脈依存性が高く、抽出された情報だけを自律的に取り扱うことはできない。多くのデータベースでは、抽出された情報にはその原著論文へのポインターが付与されており、ここにも、事実データと文献データとの有機統合という、情報･知識管理システムの研究課題がある。

4. おわりに：仮説としての生命オントロジー

現在、生命科学の分野では、分野オントロジーの構築が盛んに行われている。これは、3･2で議論したような知識の統合と再利用とを目指したものである。代表的な DB 間の相互関係をつける参照オントロジー（Reference Ontology）を目指す GO(Gene Ontology)の試みは、その代表的なものである。

しかし生命科学におけるオントロジーの役割は、知識共用という、受身的なものだけではない。対象の分類化と構造化は、生命科学における仮説であり、その仮説が適切化どうかは、個別知見がそのオントロジーによってどの程度うまく解釈できるかで判断される。この場合には、言語処理の技術は科学的仮説の検証に使われることになる。

あるいは、テキストとデータとをマイニングすることで、新たなオントロジークラスが同定できる可能性もある。知識の発見である。このように、方法論的な変革を経た2つの分野は、いま、次の野心的な科学方法論の構築に向かっている。

参考文献

[辻井　2001] 生命の理解とオントロジー、数理科学、No.458、2001.