言語理解と知識

－情報空間の構造化に向けて－

辻井潤一

東京大学大学院情報学環教授、マンチェスター大学計算機科学科教授

１．前提

言語と知識の問題は、チューリングテストに象徴されるように人工知能の究極の目的と密接に絡んでいる。思考・推論と言語、言語の意味と知識、対象世界のカテゴリ的な把握と言語など、人工知能研究の根幹に言語理解の研究がある。私自身、人間知能の解明に興味をもち、この分野の研究に携わってきた。

ただ、一方で人工知能の究極の目標を日常的な研究活動の直接的な目的に直結させることのむつかしさ、危険性も強く感じている。むしろ、役に立つ言語処理技術を開発していくことを直近の研究目的とし、その中で長期的には言語と意味、言語と知識の相互関係をより深く理解していくという研究戦略が良いのではないかと、研究室の学生たちには指導している。

このような観点から、過去１０年間ほど、生命科学文献を対象とする言語処理・テキストマイニング・情報統合の研究に従事してきた。この分野では、言語とは独立に存在する知識、生命科学の知識が想定でき、しかも、この「知識」というものが、生命科学者が集団で構築している事実のデータベースや分野オントロジーとして計算機に使用可能な形で存在している。テキストの意味をこの具体的な「知識」と有機的に関連付けるという、具体的な技術を開発する過程で、言語と知識との相互関係がより具体に詳細に捉えられるのではないか、と考えての研究である。

もちろん、言語と知識とはこのように明確に切り分けられるわけではない。言語的なものと我々の世界把握（知識）との間にある動的な相互関係を明らかにすることが、言語理解研究の本質である、とする立場もあろう。上で「知識」とカッコ付けで表現したのは、現時点での研究では、我々の知識（対象把握）への言語的なものの関与という問題は取り扱わずに（したがって、メタファーといった現象は埒外として)、言語に先行して「知識」がまず存在するとし、それとの対応で科学的言語（論文）を対象に言語理解の研究を行っている、ということを言いたかったため、である。

２．情報統合とテキストマイニング、セマンティック・ウェッブ

生命科学の分野では、一見多様に見える生命現象に DNA・遺伝子・タンパク質という共通の基盤があることが認識されるようになったことで、大規模な分野の統合が起こっている。この分野での研究の有効な遂行には、広い幅広い分野の論文から必要な情報を収集する必要性が強く認識されている。

また、生命という複雑な対象の理解には、個別的な知識断片（たとえば、特定のタンパク質）だけでなく、それらが全体としてどのようなシステムを構成しているかを理解する必要がある。これがシステム生物学(Systems Biology)の主張であるが、生命現象を引き起こす基本要素をタンパク質に限定しても、数十万種類という個々のタンパク質すべてに関

する知識、あるいは、これらのタンパク質の相互関係、その相互関係によって構成されるネットワークの詳細に関する知識は膨大であり、これらすべての理解を一人の生命科学者に期待することはできない。

システム生物学、医療科学においては、現在、多くの論文に分散的に報告されている断片的な知識を寄せ集めて、それらの相互関係に一貫性のある解釈を与えるネットワークの構築が人手で行われている。このようなネットワークには各種のものがあるが、一般に Pathway と呼ばれている。図１に、我々のグループの共同研究者が作った信号伝達の Pathway (Signaling Pathway)を示す。この Pathway を構築した生命科学者は、６００件以上の論文に目を通し、その中に散在する情報の断片を統合することでこの Pathway を構築している。このような Pathway は、生命科学における情報統合のプラットフォームになっている。

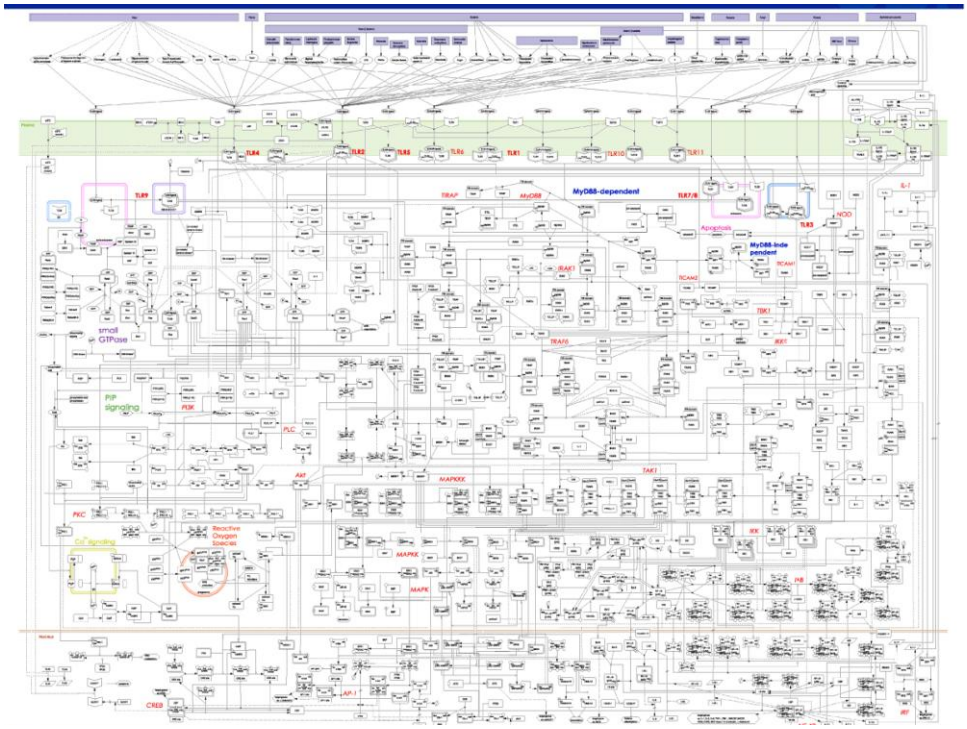


図 Signaling Pathway の例
（Oda K, Matsuoka Y, Funahashi A, Kitano H: A comprehensive pathway map of epidermal growth factor receptor signaling. *Mol Syst Biol* 2005, 1:2005）

また、このように Pathway はすでに出版された論文を読み、そこから必要な情報をキュレート(Curate)することで構築される。前述したように、生命・医療関係の論文の数は急激に増大している。アメリカ医学図書間(NLM：National Library　of Medicine)が構築している文献データベース MEDLINE には、年間５０万件以上、１日あたり千数百件のペースで新しい論文が付け加えられている。これらの論文に散在する情報を統合して Pathway 構築を行うこと、また、構築された Pathway を新しい論文の知見にあわせて修正、改編するメンテナンス作業は、とても人手だけでできるものではない。これら一連の作業を手助けするシステムの構築が、我々の PathText プロジェクトの目的である。

講演ではその一部のデモを紹介するが、我々の研究に関するより詳細な情報は

http://www.nactem.ac.uk/pathtext/

から得られる。

3　情報統合とオントロジー

生命科学での情報・知識は、さまざまな形で構築されている。たとえば、Pathway だけでなく、タンパク質や遺伝子に関する情報も、個々の論文や実験結果から人手により、キュレートされ定型的なデータベースの形式に整理されている。各タンパク質に関しては、たとえば、その構造、どのような生命事象に関係したか、どのような機能を果たすかなどが、コード化されてデータベースに蓄積されている。

関与する生命事象や機能をコード化することで、特定の生命事象に関与するタンパク質をすべて検索するといった操作が可能となる。このように、言語で記述されている内容のコード化・定型化により、情報アクセスは容易になるが、Pathway の場合と同様、論文内容をコード化するキュレーション・コストは高く、言語処理の手助けが必要となる。

また、生命現象、タンパク質機能のコード化の前提条件として、生命科学にとって有意味な現象や機能の分類が多くの生命科学研究者が合意する形で設定されていること、すなわち、分野オントロジーを構築されていることが不可欠となる。生命科学の分野は、この分野オントロジーの構築に関しては最も進んだ分野であり、本講演でも紹介する GO(Gene Ontology : http://www.geneontology.org/)など、様々なオントロジーが作られている。生命・医療科学は、大規模オントロジーの構築とその利用がもっとも進んでいる分野であり、その標準化と各種オントロジーのアーカイブが http://www.obofoundry.org/　（Open Biological and Biomedical Ontolgies）にある。このサイトに集積されたオントロジーの集合からわかるように、生命・医療化学分野でのオントロジーの多くは中途半端なおもちゃ的なオントロジーではなく、大規模で、かつ、実際の情報統合に使われているオントロジーとなっていることが理解できよう。

生命科学の情報統合が複雑になる要因の一つに、生命現象の文脈依存性がある。たとえば、あるタンパク質が特定の機能を持つと言っても、その機能が常に発現するわけではない。ある生命文脈に限って機能が発現するといった場合が多い。また、生命科学者は、他の生命科学者のグループがキュレートした Pathway やデータベースをそのまま使うわけではなく、情報の信頼性の確認に原論文を読む場合も非常に多い。このように、構造化データベースや Pathway の有効活用には、文脈を捨象したキュレート結果だけでなく、豊富な文脈情報を持った原論文との相互関係がきっちりととられている必要がある（図２）。

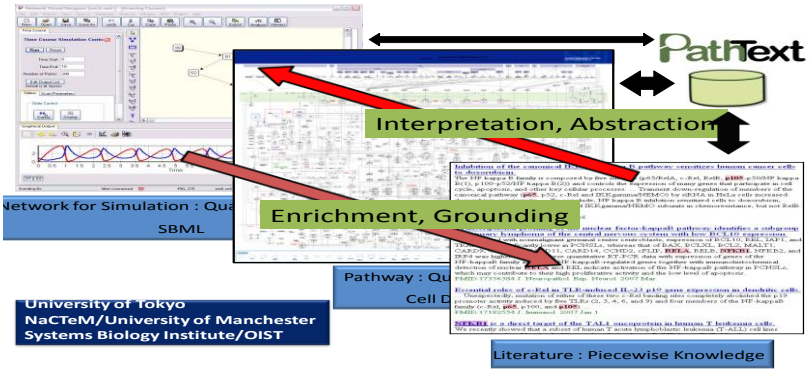


図２　定量モデル・定性モデルと文献情報

４．Semantic　Search

単純なキーワード検索やその Boolean 結合による検索ではなく、テキストや質問の意味を取り扱う Semantic Search が次世代の検索の形態であるというのは、多くの人が合意するところである。あるいは、Web を介した情報のリンケージも、現在の物理的な所在情報(URL)によるリンクではなく、意味を中心にした(URI)に移行すべきであるという Semantic Web の主張もある。

PathText システムは、意味に関する処理を

（１）ユーザの Query の意味（User Semantics）
（２）テキスト中にある情報の意味(Textual Semantics)

の２つに分けて考える。ユーザにとって使い易いシステムとするためには、ユーザに負担の掛からない方式、すなわち、できるだけすくないユーザの入力から、ユーザの置かれた文脈を考慮することでユーザの意図を汲み取ること、また、そのユーザ意図をテキストの意味を取り扱う文献検索・事実検索システムへの Query として実現することが重要だと考えている（図３）。文献検索の Query 言語としては、領域代数(Region Algebra)に基づく GCL 言語を使うが、ユーザが直接この Query 言語で Query を書くわけではない、というところが我々のシステムの特徴である。

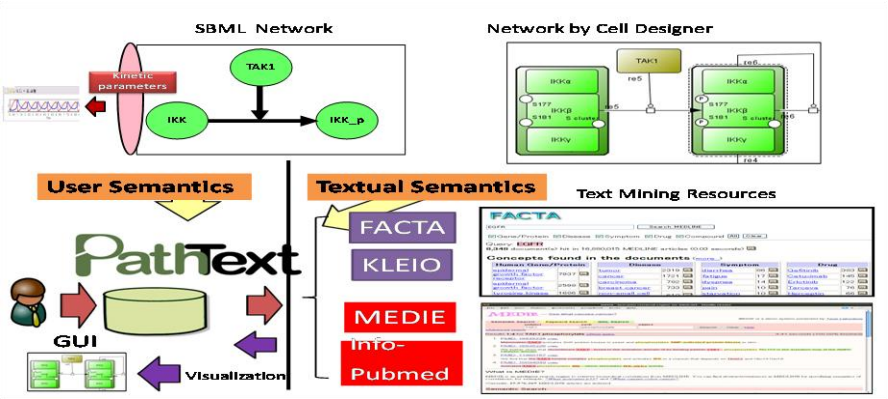


図３　PathText の概念図：ユーザセマンティクスとテキストセマンティクス

５．PathText のデモ

具体的にシステムの動作を見てみよう。

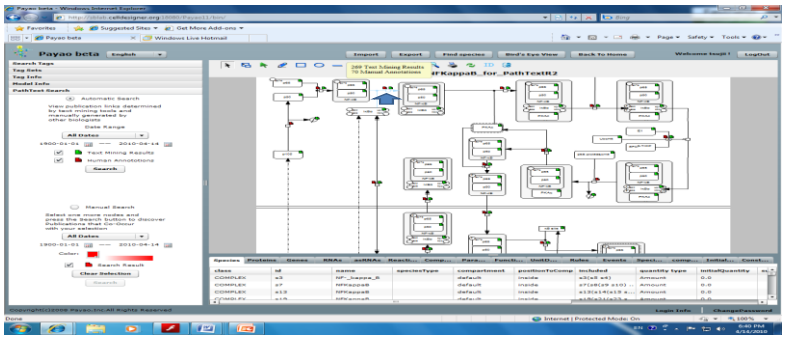


図４(a)
Pathway の図が表示される。節点間のリンクには、どれだけの論文がそのリンクに添付されているかが示される。

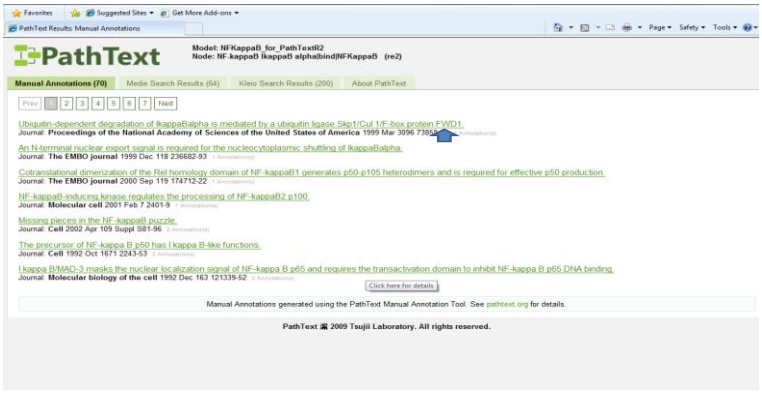


図４(b)
Pathway のリンク((a)の矢印)を一箇所クリクしたところ。生命科学者によって人手でキュレートされた論文が表示されている。

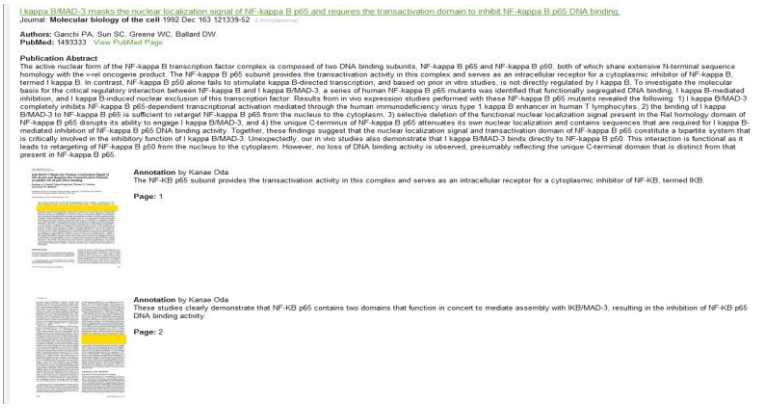


図４(c)
人手によりキュレートされた論文を選択（(b)の矢印)、クリックしたところ。キュレートされた論文の抄録と本論文中で関係すると判断された箇所が表示される。他の生命科学者は、その論文をよみ、コメントできる。

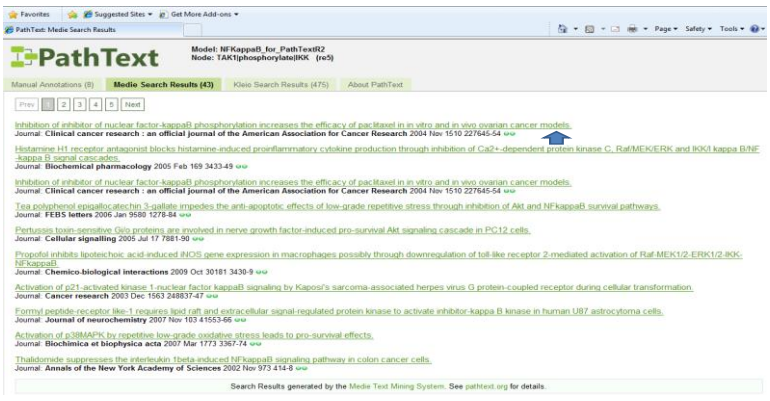


図４(d)
テキストマイニングによってPathwayの特定箇所に関連が深いと判断され検索された論文の一覧

図４(e)
(d)の矢印の論文をクリックしたところ。
システムがPathwayの特定箇所と関係が深い判断した文と抄録が示される。ユーザは文、抄録が本当に関連していたかどうかの判断をシステムにフィードバックできる。

６．テキストから知識への写像

文献情報を構造化データベース、Pathway、実験からの生データなどと統合していくためには、テキストに埋れた情報を顕在化し、言語から独立に設定されたオントロジーと関係づける必要がある。Semantic Web が主張するように、自然言語表現が本質的に持つ曖昧性(Ambiguity)や多様性(Diversity)を解消する必要がある。

言語表現が持つ曖昧さと多様性の典型的な例として、科学技術論文に頻出する Acronym の問題を考えてみよう。図５は、
（１）[曖昧性] 表面上全く同じ文字列(PCR)が、文脈によっていくつもの意味(Expanded Forms)に対応すること
（２）[多様性] 人間にとって同一のものと理解されるもの（PCR の定義・意味）、したがって、オントロジーでは単一の概念となるべきものに違った表記が対応すること

を示したものである。

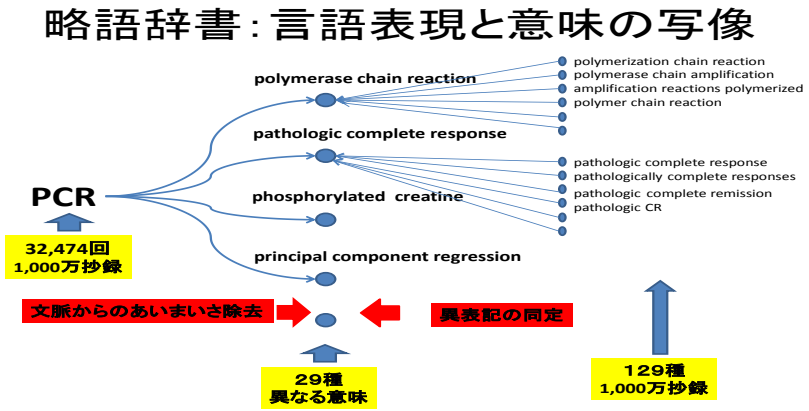


図５　言語表現の曖昧性と多様性(Acronym の場合)

前者が多義性解消(Disambiguation)、後者が標準化(Normalization)の処理と呼ばれ、Acronym 処理だけでなく、人名や組織名などの処理に典型的に見られる。また、後述するように、文が表現する事象や状態に対しても一意に意味が確定できるオントロジー中の ID に対応させることを考えられる。多義の解消と標準化処理は、テキスト情報をオントロジーを介してそれ以外の情報源中の情報にリンクしたい場合の基本的な処理となっている。

過去１０年間の言語処理の研究は、情報抽出という外部知識との対応に焦点を強くあてることで、この２つの処理技術を大きく進展させてきた。機械学習と構造に基づく言語処理の成果を統合により、この２つの処理が現実問題に適用できる技術となった。

７．深い言語処理と事実検索システム MEDIE

多義解消、標準化処理の研究成果を使った事実検索システム MEDIE を紹介する。このシステム MEDIE は、我々の研究グループの言語処理研究の成果を統合することで、現時点での技術の到達点を示すために開発したものである。また、MEDIE は、PathText のテキストマイニング・コンポーネントとして使われている。MEDIE は、

(1) 生命現象に関与する Named Eities(タンパク質、DNA、細胞中の場所など)を認識、標準化する
(2) 生命現象に関与する Event（事象）を認識し、標準化する。

(1)の技術は、基本的には前述の Acronym での２つの処理と基本的には同じである。これに対して、(2)は文の深い構造解析を行い、その結果を使うことで標準化的な処理を行っている（図６）。

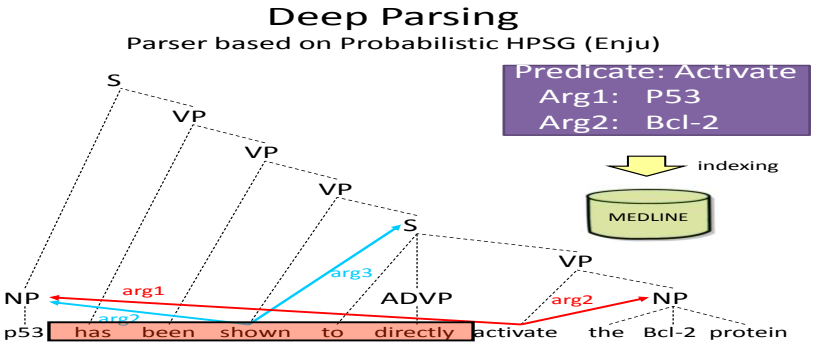


図６　深い文解析と事象の標準化処理

図６の深い解析が得られていることから、

P53 has been shown to directly activate the Bcl-2 protein

という文から、p53 が不定詞句”to activate …”の意味上の主語になっていることが認識さできることから、”p53　activate !X” といった質問に対して、この文を検索することができる。従来のキーワード検索では非常に雑音の多い（Recall は高いが、Precision が非常低い）検索しか出来ない。図７に”p53　activate !X”で検索される様々な文を示す。この MEDIE は、一般に公開されており、興味のある人は

http://www-tsujii.is.s.u-tokyo.ac.jp/medie/index.cgi

を参照してほしい。

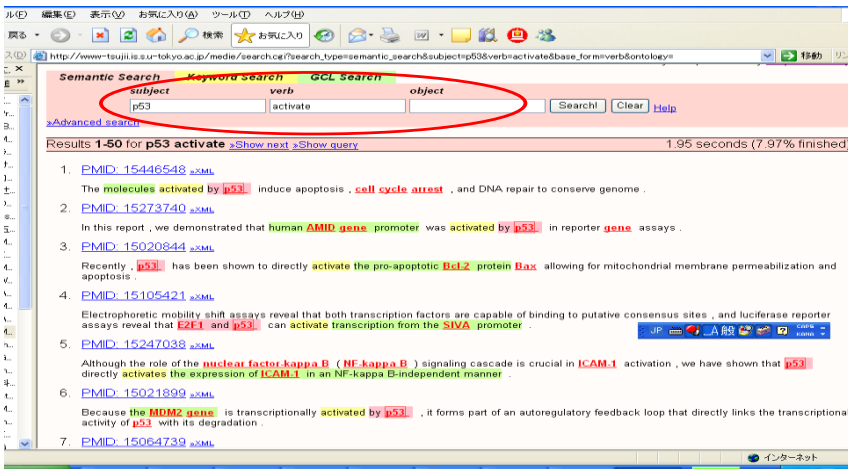


図７検索例
“activate”の意味的主語が p53 になっている文を検索している。受身文や不定詞句の主語などもうまく認識されている。

深い文解析だけでは、事象認識の標準化処理は十分ではない（８、９節参照)。

(1)に関しては、外部データベースの Unique　Key へと写像することを標準化処理としておこなっており、これを使うことで特定のタンパク質や DNA に関する外部データベースの情報が得られるようになっている。実例は、講演時にデモで紹介。

８．深い文の解析

言語処理技術の典型である文解析技術の現況に触れてこう。文解析は組み合わせ爆発の典型的な問題で、長い間、計算的に重い処理とされてきた。これは、組み合わせ爆発を回避するための評価関数がうまく作れなかったことに大きな原因がある。この困難は、構造を Annotate（付記）した大規模なテキスト集合が使えるようになったことから、確率モデルで評価関数を作ることが一般化し、解消されることになった。

また、文解析のもう一つの問題点である耐性の低さも、文法が持つ拘束条件を極力緩くし、従来の拘束条件を確率モデルによる優先度条件として使うことで大幅に解消されることになった。このことから、過去５年間に、英語や日本語に対して、浅い依存構造解析や句構造解析を非常に効率良く、かつ、耐性高く実行するプログラムが開発されてきている。

これに対して、英語の関係節での空所や、受身・使役・コントロール動詞などの現象を取り扱うことができる深い解析器の研究も、同時に進んでいる。一般に、言語学理論に基づく深い文解析器は、いわゆる記号的な規則の集合に依存する部分が大きく、拘束条件が厳しすぎるために耐性が低く、また、規則が複雑化することで処理速度も遅いという考え方が広がっているが、実際には、耐性・処理速度ともに、浅い解析器と同程度か、それ以上の文解析器が深い解析器で実現されてきている（図８)。

Results on PTB-WSJ

Parser	grammar	Accuracy	Speed
MST parser	dependency	90.02% (LAS)	4.5 snt/sec
Sagae's parser	dependency	89.01% (LAS)	21.6 snt/sec
Berkeley parser	CFG	89.27% (LF_1)	4.7 snt/sec
Charniak's parser	CFG	89.55% (LF_1)	2.2 snt/sec
Charniak's parser reranker	CFG	91.40 % (LF_1)	1.9 snt/sec
Enju parser	HPSG	88.87% (PAS-LF_1)	2.7 snt/sec
Mogura parser	HPSG	88.07% (PAS-LF_1)	22.8 snt/sec

図８　文解析器の効率
赤枠で示したものが、我々のグループが開発した文解析器、Enju/Mogura が深い文解析器。浅い文解析器と深い文解析器は正解構造が異なるために精度は比較できない。

従来の文法規則で規定されていた拘束条件を確率モデルによる優先度情報として使うという考え方は、文法理論による深い解析にも適用可能で、浅い解析器と深い解析器の差は、深いレベルでの構造要因をモデル中に明示的に取り込むかどうかだけである。実際には、深いレベルを取り込むことで探索空間がより限定され、深い解析器の効率性を高めていると考えられる。

我々の研究グループでは、Super Tagging という処理のかなり前半の段階、まだ構造が組み上がらない段階で、POS　Tagger などと同じ系列 Tagging のモデルを使うことで、かな

りの精度で構造解析を先取り的に実行出来ること、この結果、浅い解析器よりも実行速度が早い深い文解析が可能であることを示している。最も処理効率の高い Mogura が、この Super-tagging による Staged Architecture である(図９)。

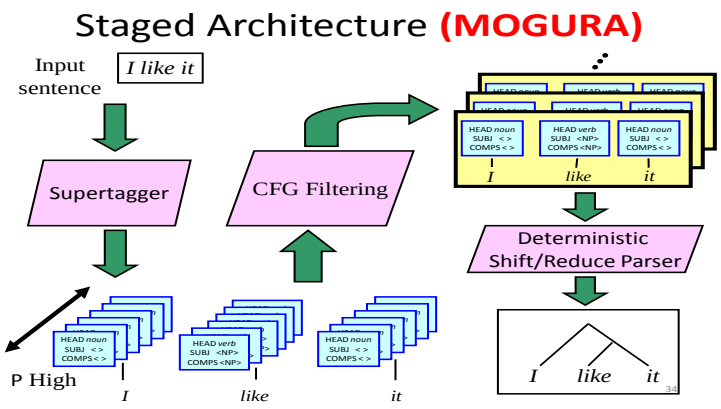


図９

文解析のアーキテクチャ

9. 事象の標準化、曖昧さ解消

深い文解析器では、受動態、使役、コントロール、関係節など、文の統語的な構造変換に対応した標準化処理を行う。しかしながら、外部知識の観点からは同一事象とみなしたい現象が自然言語では、表面上非常に多様な構造をとって現れる。

前述の生命科学のオントロジーGO(Gene Ontology)では、生命事象(biological process)や生命機能(function)という分類で「こと」的なもののオントロジーが作られている。たとえば、このオントロジーには、“STAT protein nuclear translocation”という事象分類があるが、このような「こと」的な概念は、「もの」的なオントロジー概念とはテキスト中での出現の仕方が大きくことなる。図１０は、生命科学者が８００の論文抄録中での表現で、このオントロジー概念であると認定したものを列挙したものである。ここには、“STAT protein nuclear translocation”という表現はなく、様々な変形（そのなかには、文表現もある）が見られる。

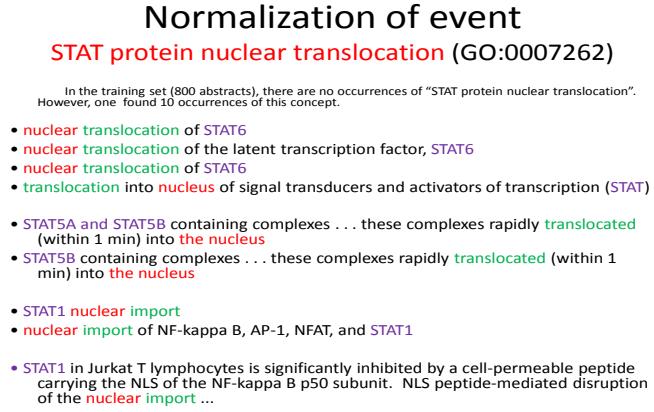


図１０「こと」的な概念のテキスト中での現れ

名詞句的なものから、文としての表現まで、様々なものがある。同義語だけでなく、構造的な側面からの同義性の定義が必要となる。

このようなことから、GO での事象、機能概念がテキスト中でどのように現れるかをみるために、我々のグループでは、２０００件の抄録に対して事象に関する意味アノテーションを行った。図１１に、この事象アノテーションに使ったオントロジーを示すが、これは GO ではかなり上位の中間ノードに相当する。前述の“STAT protein nuclear translocation”は、このオントロジーでは、localization の下位ノードとなっている。

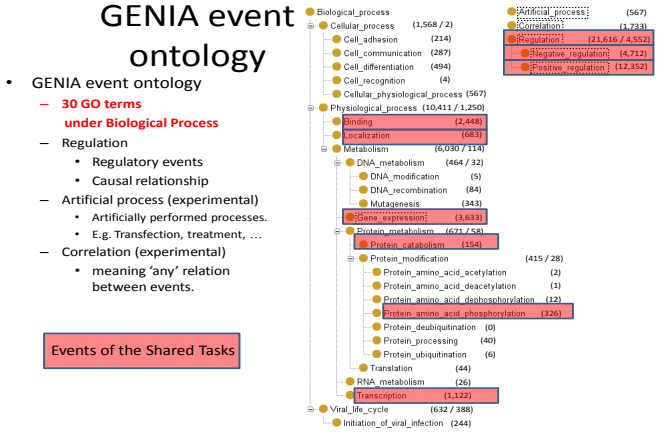


図１１　事象アノテーションのためのオントロジー

赤枠で示したものは、NA-ACL2009 の shared tasks で採用した事象クラス（頻度が多いもの）

事象の意味的な標準化は、この意味アノテーションを施したコーパスを訓練コーパスとして Classifier を機械学習させることによって行われる。この機械学習のための素性には、８節で述べた深い文解析器の結果が使われる（図１２）。現時点での Classifier の精度を図１３に示す。

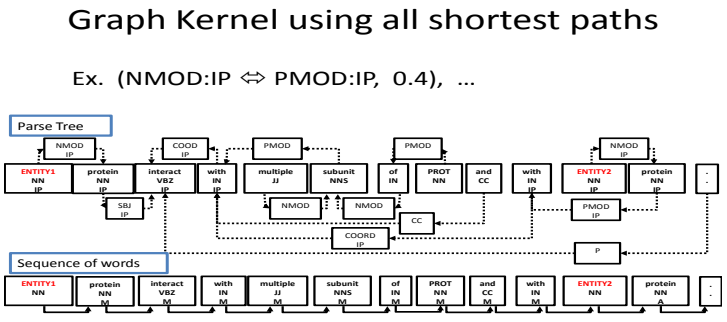
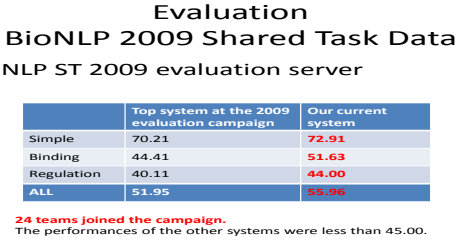


図１２　Classifier と文解析

事象認識は、NER と同様に Classifier の機械学習の問題と捉えられる。事象は複数個の NE（一個の場合もある）とそれらの文脈とにより分類される。この複数の NE を取り囲む文脈をみるために文解析器の結果が使われる。



	Top system at the 2009 evaluation campaign	Our current system
Simple	70.21	72.91
Binding	44.41	51.63
Regulation	40.11	44.00
ALL	51.95	55.96

図１３　事象認識の精度

図 13 に示した精度は、全体では５５％程度を低いように見える。ただ、複数の NE と特定の動詞の共起だけでは、精度は 10%を切るほどに低いために、文解析と機械学習による精度の向上はかなり大きなものである。55%の精度でも検索結果の質は著しく向上することになる。

１０．おわりに

言語処理は、過去１０年間に長足の進歩をとげ、文の構造やその意味を明示的に把握する技術として使うことができるようになった。特に、テキスト中に埋れた情報を言語とは独立に措定された知識とリンクすることで明示的に捉える技術(多義性解消と標準化の技術)は、Semantic web、オントロジー工学と連携することで、ウェエブ中の情報の構造化、組織化に大きな役割を果たすものとなってきている。

言語処理技術が現実の問題を解くための技術として、広範な分野に影響を与えるためには図１４に示したような課題を、今後研究していく必要があろう。

- Generic NLP tools
  - POS taggers, Dependency Parsers, Deep Parsers
  - Spelling variants normalizers, acronym recognizers
- Generic Semantic-Processing tools
  - Named entity recognizers, Event recognizers, Relation recognizers
  - Normalizers (Disambiguators)
- Domain and task adaptation
  - Semi-supervised learners, Annotation accelerators
  - Transfer learning
- Platform for interoperable NLP tools (U-Compare)
  - University of Manchester (NaCTeM), Colorado University, University of Tokyo
- Integration of NLP modules in application environments
  - Garuda Consortium (OIST, U-Manchester, U-Tokyo, U-Edinburg, SRI, etc.)
  - Taverna + U-Compare (NaCTeM)
- HPC for text mining
  - Parsing, NER, ER and intelligent indexing on the whole Medline in a day

図１４　今後の研究課題

（参考文献）
以下は、本講演での話題に関係する我々のグループの研究発表
[Pathtext]
(1) Brian Kemper, Takuya Matsuzaki, Yukiko Matsuoka, Yoshimasa Tsuruoka, Hiroaki Kitano, Sophia Ananiadou and Jun'ichi Tsujii: PathText: A Text Mining Integrator for Biological Pathway Visualizations, Proc. of ISMB 2010, (to appear).
(2) Oda, Kanae, Jin-Dong Kim, Tomoko Ohta, Daisuke Okanohara, Takuya Matsuzaki, Yuka Tateisi and Jun'ichi Tsujii. **New challenges for text mining: Mapping between text and manually curated pathways.** BMC Bioinformatics. 9(Suppl 3). pp. S5, BioMed Central, Apr 2008. ISSN 1471-2105.

[MEDIE]
(3) Miyao, Yusuke, Tomoko Ohta, Katsuya Masuda, Yoshimasa Tsuruoka, Kazuhiro

Yoshida, Takashi Ninomiya and Jun'ichi Tsujii. **Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases.** In the Proceedings of COLING-ACL 2006. Sydney, Australia, pp. 1017--1024, July 2006.

[Acronym]
(4) Okazaki, Naoaki, Sophia Ananiadou and Jun'ichi Tsujii. **Building a High Quality Sense Inventory for Improved Abbreviation Disambiguation.** Bioinformatics. Oxford University Press, 2010.

[Deep parser]
(5) Miyao, Yusuke and Jun'ichi Tsujii. **Feature Forest Models for Probabilistic HPSG Parsing.** Computational Linguistics. 34(1). pp. 35–80, MIT Press, March 2008.
(6) Ninomiya, Takashi, Yoshimasa Tsuruoka, Yusuke Miyao, Kenjiro Taura and Jun'ichi Tsujii. **Fast and Scalable HPSG Parsing.** Traitement automatique des langues (TAL). 46(2). Association pour le Traitement Automatique des Langues, 2006.
(7) Matsuzaki, Takuya, Yusuke Miyao, Jun'ichi Tsujii. **Probabilistic Context-Free Grammars with Latent Annotations.** In Srinivas Bangalore and Aravind K. Joshi (Eds.), Supertagging - Using Complex Lexical Descriptions in Natural Language Processing. pp. 337-354, MIT Press, March 2010.

[Event Recognizer]
(8) Miwa, Makoto, Rune Sætre, Jin-Dong Kim, and Jun'ichi Tsujii. **Event Extraction with Complex Event Classification Using Rich Features.** Journal of Bioinformatics and Computational Biology (JBCB) . 8(1). pp. 131--146, February 2010.
(9) Miwa, Makoto, Rune Sætre, Yusuke Miyao, and Jun'ichi Tsujii. **Protein-Protein Interaction Extraction by Leveraging Multiple Kernels and Parsers.** International Journal of Medical Informatics. 78(12). pp. e39--e46, April 2009. Mining of Clinical and Biomedical Text and Data Special Issue.

[Semantic Annotation]
(10) Kim, Jin-Dong, Tomoko Ohta and Jun'ichi Tsujii. **Corpus annotation for mining biomedical events from lterature.** BMC Bioinformatics. 9(1). pp. 10, BioMed Central, January 2008. ISSN 1471-2105.
(11) Pyysalo, Sampo, Tomoko Ohta, Jin-Dong Kim and Jun'ichi Tsujii. **Static Relations: a Piece in the Biomedical Information Extraction Puzzle.** In the Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop. pp. 1--9, 2009.