生物学知識データの構築

○真栄城 哲也†，藤原 譲‡，下原 勝憲†
†ATR 人間情報科学研究所
‡独立行政法人 工業所有権総合情報館

Construction of Biological Knowledge System

○ Tetsuya Maeshiro†, Yuzuru Fujiwara‡, Katsunori Shimohara†
†ATR Human Information Science Laboratories
‡National Center for Industrial Property Information

Abstract 133,251 terms extracted from biological books and dictionaries were processed to automatically extract semantic relations among terms such as hierarchical and equivalent relations. A total of 2,574,099 relations were extracted. Extracted terms represent biological concepts, thus extracted relations define the biological conceptual structure. The network structure of constructed conceptual structure from a fixed viewpoint shows a structure of double scale-free network, with the border on concepts with 1,000 relations.

1 はじめに

ヒトを含む様々な生物種の設計図であるゲノム配列が決定されつつあり，それに伴い生物の機能を担う蛋白質のアミノ酸配列や構造，さらには生体内の様々な分子の発現パターン等，蓄積されるデータ量は指数的に増加しつつある．これらのデータを処理する手法として，ホモロジー検索等の類似性の検出方法が多数提案されつつあり，さらなるデータ量の増加を招いている．一方，これらのデータは様々な研究グループや用途の都合に応じて，固有の形式で記述，保持や運営されており，例えば同じ生物種を対象とするデータベースであっても，複数のデータベースにまたがった使用は困難な状況にある．

生物学に限らないが，生物学のデータベースにおいては少なくとも (i) データベースの統合，(ii) 分野毎に使われる用語の違い，の２つの問題がある．本論文では，これらの問題を解決するための１つの手段として，生物学の基礎知識を専門書や辞典から抽出し，データベースの統合や用語の違いを吸収するために構築した基礎概念の基盤について述べる．

情報科学の分野では，オントロジーやコーパスの研究が盛んに行われている [1]．一方，生物学では，以前から MEDLINE 文献データベース [2] の整備が行われており，NIH が Medline で使用する用語集と用語間の階層関係を定義した MeSH [3] を整備している．また，最近はオントロジーの構築が盛んであり，BioTermBank [4]，Interactions Ontology [5] 等が挙げられる．

[1]本研究は通信・放送機構の研究委託により実施したものである

2 概念と意味関係の抽出

2.1 概念の抽出

概念間の意味関係を抽出するには，概念が必要である．今回対象としている生物学のような自然科学の場合には，専門用語が対象分野の概念を表している．従って，専門用語を MeSH，そして分子生物学，脳神経学，生化学，免疫学，微生物学等の分野の 26 冊の専門書の索引と辞典から抽出した．なお，約 30 万語ある MeSH の化合物用語集は含めていない．英語と日本語の 2ヶ国語の専門用語を扱い，対訳関係も抽出した．このようにして抽出した専門用語は，重複部分を除いて 133,251 語あり，これらを基に意味関係を抽出した．

2.2 概念間の意味関係の抽出

専門用語 (概念) の収集より重要なのは，概念間を関係付ける意味関係の抽出である．概念間の関係を基に概念を構造化して，初めて収集した専門用語の活用が可能となる．

抽出した概念間の意味関係の種類は，同義関係，関連関係，類似関係，階層関係，属性，である．意味関係の抽出は専門用語の生成規則に基づいた C-TRAN と SS-KWIC [6, 7, 8]，そして専門用語の抽出に利用した 26 冊の専門書の本文を用いて様々な意味関係を抽出した．さらに，専門書の目次からは概念間の階層関係を抽出した．また，MeSH に記述されている概念の階層構造も利用した．

このようにして 2,574,099 個の意味関係を抽出した．1 つの概念当りの意味関係の平均数は 19.3 個である．これらの意味関係を表現する概念構造のモデルとして HBM [8] を用いている．

意味関係抽出の処理で，132,563 個の概念について何らかの意味関係を見付けた．これは，全体の 99.5%の概念が意味的に関連付けられた概念構造を構築したことになる．

専門書の本文から意味関係を抽出する単位として，節，段落，文等があり，ここでは文を意味関係抽出の処理単位とした．文からの抽出では，複数の専門用語が共出現する抽象的な関連関係のレベルから，特定の概念を介した関係まで，様々な抽象度の意味関係を抽出できる．これは，意味関係の階層構造に相当する．

3 考察

今回抽出した概念は 133,251 個であり，これは NIH の MeSH に含まれる専門用語数 (約 2 万語) より多く，専門用語 (概念) の抽出処理は有効であることを示している．また，概念当りの意味関係数は最大 7,435 個であり，1 つの意味関係を持つ概念が全体の 74.5%を占める．

意味関係の抽出処理で問題になるのは，誤りの検出である．概念間の意味関係を自動抽出する場合，誤った意味関係がほぼ間違いなく混入する．従って，自動生成された概念構造に含まれる誤った意味関係を自動的に検出する必要がある．しかし，同じ間

違いでも False positive よりも深刻なのは False negative である．前者の場合，意味関係の抽出時 (フェーズ 1) に洩れても後の誤り検出時 (フェーズ 2) で除くことができるため，検出の機会が 2 回ある．一方，前者とは異なり，後者は，概念間の意味関係の抽出時 (フェーズ 1) にしか処理されず，誤り検出の機会が 1 回しかない．従って，False positive が増加しても，可能な限り False negative が少なくなる手法をフェーズ 1 で採用する必要がある．そしてフェーズ 2 で False positive を排除する手法を用いる．システムに含まれる概念と正しい概念間の関係が増加するにつれ，意味関係の矛盾点を検出が可能となるため，意味関係の誤りを自動検出できる割合が増す．

概念構造は様々な観点から捉えることができるが，観点を固定した場合に概念毎の意味関係の数のヒストグラムを取ると，1,000 個前後の意味関係を持つ概念を境に 2 重のスケールフリー構造を持つことが解る (図 1)．図 1は，得られた概念構造が Small World Network (SWN) [9, 10] と類似の構造を持っていることを示している．このことは，概念間の距離を，固定された観点において概念を接続する意味関係の最短数と定義した場合，観点に依存するが，多くの概念と意味関係を持つハブとして機能する少数の概念の存在によって，ある概念から他の概念へ短距離で到達できることを意味する．なお，このような 2 重のスケールフリー構造が，生物学の概念構造特有か，概念構造に共通するかの解析が必要である．

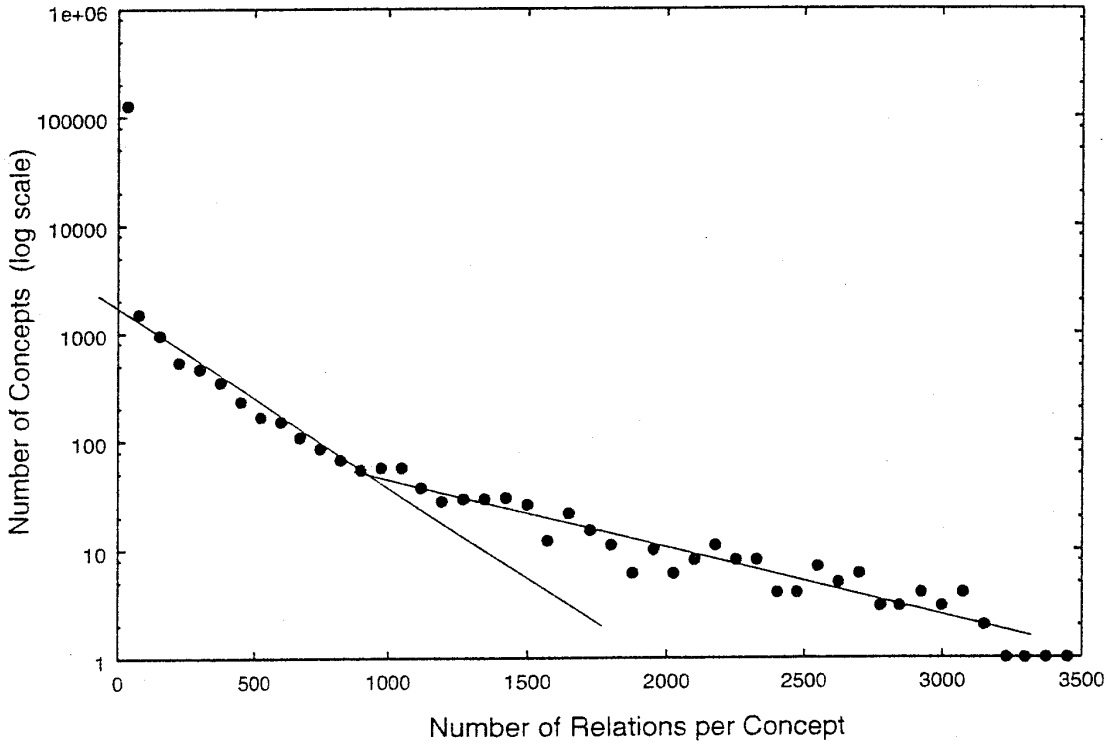

図 1: 概念毎の意味関係の数のヒストグラム．意味関係の数が 3,500 個までの部分を拡大してある．3,500 個以降は X 軸上に点が並ぶ．

参考文献

[1] 情報処理学会誌, Vol41, No.7, 2000

[2] http://www.nlm.nih.gov/pubs/factsheets/medline.html

[3] http://www.nlm.nih.gov/mesh

[4] http://ontology.ims.u-tokyo.ac.jp/BTB/index_Jhtml

[5] http://www.ai.sri.com/pkarp/interactions.html

[6] Y. Fujiwara and J. Lai, "An information-base system based on the self-organization of concepts represented by terms", *Terminology* **3**, 313-334, 1997.

[7] Y. Fujiwara, W. G. Lee, Y. Ishikawa, A. Nishioka, H. Hatada and S. Fujiwara, "Dynamic thesaurus for intelligent access to research databases", *Proc. 47 FID Conf. (Helsinki)* , 173-181, 1988.

[8] Y. Fujiwara and Y. Lie, "The homogenized bipartite model for self organization of knowledge and information", *IFID* **2**, 13-17, 1998.

[9] D.J. Watts and S.H. Strogatz, "Collective dynamics of 'smal-world' networks", *Nature* **393**, 440-42, 1998.

[10] R. Albert and A.-L. Barabasi, "Statistical mechanics of complex networks", *Reviews of Modern Physics* **74**, 47-97, 2002.