

# Thesaurus or logical ontology, which do we need for mining text?

Junichi TSUJII

Department of Computer Science

Graduate School of Information Science and Technology

University of Tokyo

[tsujii@is.s.u-tokyo.ac.jp](mailto:tsujii@is.s.u-tokyo.ac.jp)

## Abstract

It has been claimed that domain ontology is necessary not only for effective and efficient information sharing but also for information extraction and text mining. In particular, the need of common ontology for information sharing among different research communities has been recognized in bio-medical fields, and several domain ontologies are now being built (GO 2003). This is because of the sheer complexity of the semantic space and a huge number of concepts or terms used in these specific domains.

We have been engaged in annotation of Medline abstracts in molecular biology (GENIA corpus: <http://www-tsuji.is.s.u-tokyo.ac.jp/GENIA/>) for 5 years and, in the process, we also developed our own ontology (GENIA ontology).

However, there seem to be confusions on what kinds of ontology we really need and how it can be used for effective information management systems for bio-medical research.

We argue in this paper that there are several different views of ontology and that, while logical ontology a la OWL would be useful, it may be neither practical nor possible to build a single large logical ontology of the domain. We would also like to claim that urgent issues we have to resolve are more concerned with construction of thesauri than logical ontology, i.e. lexical resources that treat convoluted nature of a mapping from linguistic forms to concepts.

Although we need ontological consideration to construct thesauri, consistency and coherency of the whole system that logical ontology usually requires should not be the main concern.

## 1. BACKGROUND

In recent years, data (DNA sequences, micro array data, etc.) that bio-medical sciences produce have increased drastically. Those raw data thus produced have to be interpreted by biologists in terms of the existing body of knowledge. Since knowledge of the fields is represented in the form of published papers, interpretation of data actually mean to relate observed data with linguistic expressions. Most of the results of data interpretation are in turn published again as text and provide a renewed body of knowledge, in terms of which new data are to be interpreted. 40,000 bibliographical units (papers) are added to the Medline data base in a month.

Considering the amount of published papers, it is inevitable for biologists to make serious efforts to make the existing body of knowledge (text bases of published papers) more transparent and accessible. Medline, for example, contains 12,000,000 bibliographical units in molecular biology, bio-chemistry and medicine, as of 2003.

Various types of information of specific genes, for example, have been manually extracted from published papers and stored in data bases, and thus, biologists need not to read original papers to order to find relevant information on specific genes. Biologists call such accumulation of relevant information (their functions expressed by language) attached to genes as *gene annotation*. Similar attempts have been made on proteins and many data bases of proteins have been constructed.

Now, molecular biologists are interested in identifying functions of proteins coded by genes. This means that they are interested in what roles specific proteins play in interaction networks like the one in Figure 1.

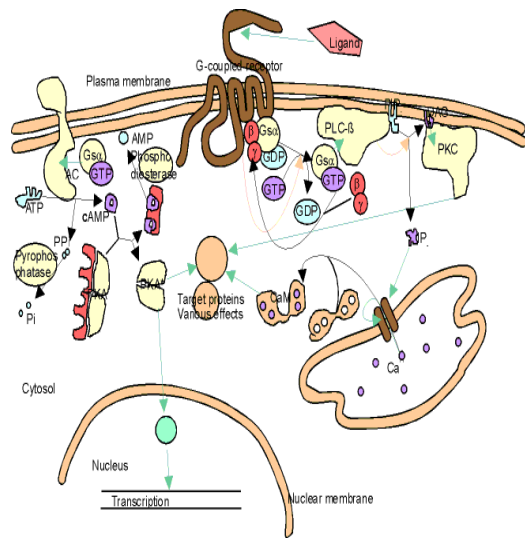


Figure 1 G-protein coupled receptor pathway model (from TRANSPATH)

Such networks of interactions of biological entities are called signal or metabolic pathways, depending on what is actually transferred through a network, either information or energy. It is one of the main claims of system biology that functions of given proteins or genes can be understood only through understanding of such pathways.

Furthermore, since genes, proteins and pathways are shared across different species through evolution, findings in different species are to be used in order to understand a network in a given species, e.g. human. That is,

*Scientists in areas such as molecular biology and biochemistry aim to discover new biological entities and their functions. Typical cases could be discoveries of the implications of new proteins and genes in an already known process, or implication of proteins with previously characterized functions in a separate process.*

*The use of available information (published papers, etc.) is a key step for the discovery process, since in many cases weak or indirect evidences about possible relations hidden in the literature are used to substantiate working hypothesis that are experimentally explored (Valencia 2001).*

Since data bases of single proteins have been constructed, serious efforts are now being made to curate information on protein-protein interactions

from published papers that constitute pathways. However, considering the amount of papers published in the field (40,000 bibliographical units per month), it is by no means easy to curate them manually, e.g. by reading papers.

Curating a large number of protein-protein interactions, storing them in databases and using them to hypothesize a whole pathway network is just an example of a huge step in the transition of scientific methodology that bio-medical fields are now experiencing.

Traditional units of knowledge, published papers or text, are deconstructed and used as data for creating new knowledge. Pieces of information embedded in text are to be extracted (information extraction), accumulated and accessed to create new knowledge. In text mining, these extracted information are treated as data to discover new pieces of knowledge.

These deconstruction, re-integration and interpretation processes of raw data and published papers are inevitable, since traditional boundaries of biological sciences were dismantled and massive integration of knowledge across different fields in biology has started and been accelerating.

Furthermore, in order to associate anomaly of genes (and thus proteins coded by them) with specific diseases, to understand the mechanisms and thereby, to design new drugs, one has to link knowledge of molecular biology with clinical or pathological observations, which in turn are accumulated in the form of published papers or case reports.

The whole integration process of bio-medical knowledge has been motivated by the current belief of biologists that all biological processes in different contexts and species share common rules, those determined by DNA sequences and processes involving proteins and other biological entities like protein-protein interactions.

However, we also believe that

- (1) Similar integration processes are hugely beneficial in other scientific endeavors as well, including economics, social sciences, etc.,
- (2) Such integration is now possible due to recent advances of data processing, computer networks, electronic publications and archiving,

- (3) Therefore the following discussion in this paper is relevant to all the attempts variously called as e-science, data-grid, semantic-grid or semantic web.

## 2. ONTOLOGY IN BIOMEDICAL DOMAINS

Whenever different communities share their knowledge, both terminological and ontological problems arise. Different communities may use different terms to denote the same concepts and the same terms to denote different concepts (terminological problems). It is also often the case that different communities view the same things from different perspectives and thus conceptualize them differently (ontological problems).

In application such as those in e-business, different communities may be able to reach explicit agreements on single standard reference ontology with a set of terms to refer to concepts (or events, processes) in the ontology. At least, one can explicitly grasp ontological differences among different communities and define explicit partial mappings among them when agreement on a single ontology cannot be reached.

Since bio-medical domains are now experiencing massive integration of knowledge of different areas, it is natural that terminological and ontological problems have become one of their major concerns and that they take the same methodology as e-business, i.e. to define a common reference ontology through which knowledge of different communities have are to be shared. They are now being engaged in building standard reference ontologies such as GO (Gene Ontology).

However, the nature of the bio-medical domains seems different from those in e-business or other application in some crucial aspects.

- (1) Size
- (2) Context Dependency
- (3) Evolving nature of ontology
- (4) Inconsistencies

### 2.1 Size of Ontology

While ontologies for e-business or for meta-data of bibliographical entities (like Dublin Core) only need a limited number of concepts or terms, to describe the content of bio-medical knowledge require a huge collection of them. For example,

the meta-thesaurus of UMLS, which is arguably the largest collection of terms in the field but nonetheless many researchers complain is not at all comprehensive enough to cover the domains, contain terms such as follows.

- (1) In total, as of July 2003,  
900,551 concepts  
1,852,501 English strings
- (2) For the tissues, organs, and body parts,  
81,435 concepts  
177,540 English strings
- (3) For the diseases and disorders,  
114,444 concepts  
350,495 English strings

Although it is possible to manage mutual relationships among several hundreds of concepts, it becomes intractable, if one has to manage consistency of complex relationships of more than one million terms/concepts. The size really matters.

### 2.2 Context Dependency

Logical ontology assumes that categories are explicitly defined by a set of their defining properties and that, once an entity is judged as an instance of a category, it inherits a set of other properties. The power of logical ontology comes from the inference capability that presupposes such static, context-independent relationships between categories and properties (or features).

However, such context-independent relationships between categories and features are exceptions but not norm in the bio-medical domains. Genes may have the same names across different species but they are not exactly the same. Protein names have similar problems. They may appear in pathways in different species and play similar functions, but since they have different amino-acid sequences, their properties are not identical.

Furthermore, whether a certain protein shows a certain set of properties depend not only on species but also on other factors like locations inside a cell, location of the cell, states of other biological entities surrounding them, etc.

In short, biological features and events are highly dependent on contexts, which is the reason why biologists would read original papers to check the contexts of observed phenomena once they identify relevant properties or events in

curated database. In a sense, curated data bases provide indexes to published papers.

The very basic assumption of static, context-independent relationships between concepts (categories) and properties (features), which bestows logical ontology the inference capability, does not hold in bio-medical domains.

### 2.3 Evolving Nature of Ontology

The term “gene” had existed well before genome science started. Names can exist without explicit understanding of the things denoted by them.

Ontologies in e-business are defined for the sake of business, i.e. facilitating effective communication in business. On the other hand, to define a proper ontology for biology is the ultimate goal of biology.

Although the former assumes that complete understanding of relevant aspects of given domains (relevant aspects of business transactions of specific kinds at hand) is possible, the latter assume that understanding of biological systems are always partial and reaching fuller understanding itself is the goal.

In the former, terms are introduced as labels to denote concepts that are fully understood and whose meanings are shared by communicators. In the latter, terms are often introduced to denote concepts that are not fully understood. To understand the meanings of terms thus introduced is their goal. Due to this dynamism between terms and ontological entities, it often happens that a single term that has been considered to denote a single category in ontology is split into several categories or several terms are merged into one.

Ontology at a given time simply reflects the state of understanding at the time and of a specific community of scientists.

We can see this hypothetical and dynamic nature of ontology clearly manifested in development of anatomical ontology.

In order to support generalization of scientific claims across different species, one has to establish categorization of organs or part of organs that can be applied across different species. Without ontology of anatomy that can be applied across different species, one cannot compare and transfer biological knowledge of one species to

another, since much of biological events like protein-protein interaction are dependent on anatomical locations.

However, since organs or parts of organs in different species are “different”, one encounters the essential issues of ontology, i.e. from which perspective one should establish categorization of organs and organ parts, what set of properties one should use for categorization, and what set of properties are to be shared by the entities that belong to the same categories or ontological entities.

One can use similarities/dissimilarities of physical properties to identify categories of anatomical organs and their parts, and by using those terms, build up theories of explaining their functions. Alternatively, one can use evolutionary roots to identify the same organs and their parts across different species and use them to generalize knowledge of other kinds like protein-protein interaction across different species.

Actual anatomical ontologies are constructed by using eclectic criteria, and constantly revised. To reach an ontology that explains observational facts systematically is one of the ultimate goals of biology.

Terms or names exist before complete understanding of what they denote.

### 2.4 Inconsistency

Biologists use many hierarchical classification schemes but they are really not schemes for logical inferences. Their classification schemes are like UDC codes for information access and tend to be eclectic.

A classification scheme of viruses, for example, is based on their shapes at one level, and on methods of detecting them at another level.

Since they are not designed for logical inferences, they have many problems in terms of property inheritance if we see them as logical ontology.

Due to the context-dependent nature of biological knowledge, even simple concept hierarchies with property inheritance may introduce a lot of inconsistencies.

### 3. ONTOLOGY vs. TERMINOLOGY

Terminological and ontological problems are intertwined. In order to judge whether two terms in different communities are used to denote the same “concepts” or not (terminological problem), one has to be able to judge, in the first place, whether two concepts referred in the different communities are the same or not (ontological problem).

Since the meanings of concepts are determined in terms of other concepts in systems, this way of thinking may lead to explicit representation of all concepts and their relationships. Such ontology-first approach, which sees terms as mere labels of concepts, emphasizes the importance of establishing ontology beforehand that is consistent as a logical system. While the approach may succeed in static and small domains such as those in e-business or the domain of meta-data of bibliographical units, it is our contention that

- (1) The ontology-first approach cannot capture the dynamic aspects of human communication and the evolutionary nature of scientific endeavors,
- (2) The term-first approach, which starts with surface forms and restricts ontological consideration to the minimum necessity, is more effective for information management systems for domains such as bio-medicine

It may sound strange to claim that language can work as communication media without explicit shared ontology or meanings. However, the claim sounds natural if one considers the use of words in everyday language. In everyday life, despite the fact that the meaning of a word is highly dependent on context and despite the fact that different parties may not use the same words with the same meanings, we can still communicate by language.

#### 3.1 Synonyms by Surface Forms

Since the terminological problems, synonyms and ambiguous terms, arise in the mapping between linguistic forms and ontological entities, one may think that the issues cannot be properly addressed without explicit reference to ontological entities. However, quite a large proportion of terminological problems can be resolved just by looking at the surface forms or

with minimum commitment to ontology.

According to (Nenadic 2002)(Nenadic 2004), synonyms that denote the same ontological entities are classified into

- (1) Spelling variants
- (2) Morphological variants
- (3) Acronyms
- (4) Structural variants
- (5) Lexical variants

They claim that (1)-(3) can be treated without ontological commitment.

[Ex 1]

**nuclear factor kappa B**  
**NF-kappa B**  
**NF kappa B**  
**NFKB factor**  
**NF-KB**  
**NF kB**  
**nuclear-factor kappa B**  
**nuclear factor B**  
**Nuclear Factor kappa B**

We also show that these variants can be recognized as such without referring to abstract ontological entities, but just by simple string matching with edit distance (Tsuruoka 2003) and algorithms of associating acronyms with their expanded forms.

While variants that are structurally correlated with sometimes denote different ontological entities, these are rare and algorithms similar to acronyms are devised for identifying structural variants (Jacquemin 2001)(Nenadic 2004).

#### 3.2 Synonyms through ontology

However, variants called “lexical variants” require ontological judgment concerning whether two expressions used in different communities denote the same protein or not.

[Ex 2]

**PKB, Akt**

Since general function/role names and names that imply certain properties are used to denote extensionally single proteins (Morgan 2003), some lexical variants can only be captured by ontological consideration.

In particular, whether a common noun phrase denotes a specific protein can only be made by

biologists.

[Ex 3]

**“cyclin-dependent kinase inhibitor” are the same as p27, p27kip1**

While the variants in [Ex 1] can be identified just by looking at linguistic expressions, those in [Ex 3] require ontological judgement of individual linguistic forms and therefore should be treated by lexicon.

### 3.3 Ambiguous terms

Because of the nature of the problem, to identify ambiguous expressions requires ontological consideration, except for cases of ambiguous acronyms. Ambiguous acronyms can be recognized and thus collected automatically by gathering pairs of acronyms and their expanded forms.

The other ambiguous terms, which require ontological consideration to recognize ambiguity and thus should be explicitly listed in lexical resources, include:

#### (1) Systematic ambiguities

- (1-1) Common nouns that describe function or properties turn to be the names of specific proteins and thus introduce ambiguities between the two readings
- (1-2) Names of proteins are often used to denote the gene names that encode them
- (1-3) Domain names that are names of proteins are used to denote the proteins that contain them

[Ex 4](Morgan 2003)

#### **suppressor of sable**

specific gene that suppresses expression of another gene called sable  
this is first used as the name of protein that suppresses the expression, then used as the gene name that codes for that protein

- (2) Application specific ambiguities: These are not ambiguities introduced only by the convoluted mapping between linguistic expressions and concepts, but by finer distinctions required by biological application.

- (2-1) Some family names of proteins are used in text as protein names, but highly ambiguous in the sense of denoting

many different proteins. As an example of extreme cases, MAP kinase or MAPK, which itself denotes a function like “suppressor of sable”, contain more than 40 different individual proteins, unlike suppressor of sable. Since other family names can be effectively treated as single proteins, the degree of disambiguation

- (2-2) A gene found in a species has their homologues in other species. But if the two homologues are close enough, they are denoted by the same names. While homologues are accompanied by the names of species in their official nominations, they are often dropped in actual text. Furthermore, since the proteins coded by them often show different properties, they may have different protein names. According to different applications, homologues or proteins coded by them have to be distinguished.

[Ex 5]

#### **NFKB2**

The gene name for one of the subunits (proteins) of NFKB is used to denote both of its homologues of chicken and human. The proteins they code are also referred by the same name. But SwissProt have different ids for the two proteins, since they have different amino-acid sequences and therefore have somewhat different properties. Whether we have to disambiguate these homologues depend on application.

p52shc, p52(Shc) are homologues that, in some application, have to be distinguished.

- (3) Non-systematic ambiguities: Since nomenclature in the fields has not been established, researchers use arbitrary names that often conflict with common words. (Morgan 2003) reported that THE, TO etc. are used as gene names. Although these ambiguities cause serious difficulties in NLP, they do not require any ontological consideration (Ananiadou 2003).

## 4. INFORMATION MANAGEMENT SYSTEMS THROUGH TERMS AND EVENTS

As we discussed in Section 1, a database that contains information curated from published papers can be taken as indexes for information retrieval. It enables one to access all kinds of information, some fragments of text and others

factual data (Ananiadou 2003)(Mima 2002)(See Figure 2).

Same as ordinary IR systems using uncontrolled indexing terms, it is crucial in such a system that terms in text are properly disambiguated and that synonyms of terms are to be expanded according to users' information demand.

While ontological ambiguities of certain types should be disambiguated in some cases, they should be taken non-ambiguous in other cases. Explicit ontology would be more useful for enabling such flexible adjustment of granularity of ontological ambiguities than for inference capabilities that logical ontology aims to attain.

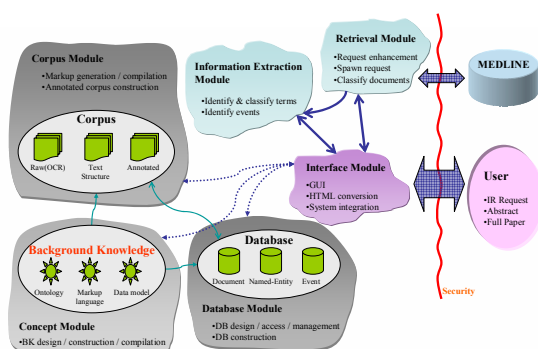


Figure 2: Term-based information management system

For the sake of simplicity, we focus on nominal terminology in this paper. We are now extending the same idea to cover verbs and use events like protein-protein interactions extracted from text as indexes. Since events expressed by verbs are more abstract than nominal concepts such as proteins, genes, etc, decision on synonyms and ambiguous terms need more careful ontological considerations than nominal terms. However, we keep our principle of minimum commitment to ontological consideration.

## 5 Concluding Remark

Language allows people with different backgrounds and with different levels of understanding to communicate. In everyday communication, it is norm than exception that words have multiple meanings depending on context.

While such dynamisms are less apparent in

language in science, the essential nature of language that words are used without complete understanding remains and this nature of language supports our thought/communicative process. Science like biology that relies on language exploits this nature of language extensively in expanding their understanding or science.

Due to the dynamic relationship between language and science, the ontology-first approach that sees terms as mere labels to ontological entities will never work for the bio-medical science.

I suspect that this is the case for all knowledge-sharing situations. While static and detailed common ontologies may be possible for well-circumscribed domains such as meta-data for bibliographical information, the terminology-first approach with minimum commitment to ontology would be inevitable for facilitating broader knowledge sharing.

## [References]

[Ananiadou 2003] Ananiadou, S. And Tsujii, J. (eds.): Proc. of workshop on "Natural Language Processing for Biomedical domains", ACL, Sapporo, 2003.

[Ananiadou 2001] Ananiadou, S., Mima, H., Nenadic, G.: A terminology management workbench for molecular biology, Information extraction in molecular biology (eds: van del Vet, P., et.al), University of Twente, the Netherlands, 2001

[Chang 2002] Chang, J., Schutze, D., Altman, R.: Creating on-line dictionary of abbreviations from Medline, Jour. of the American Medical Informatics Association, 2002.

[GO 2004] The Gene ontology (GO) database and information resource, Nucleic Acid Research, 32: D258-D261, 2004

[Jacquemin 2001] Jacquemin, C.: Spotting and discovering terms through NLP, MIT Press, 2001.

[Morgan 2003] Morgan, A., Yeh, A., and Hirshman, L.: Gene name extraction using FlyBase resources, in Proc. of workshop on NLP for Biomedical domains" (eds:Ananiadou and Tsujii), ACL, Sapporo, 2003.

[Nenadic 2002] Nenadic, G., Mima, H., et.al.:

Terminology-based literature mining and knowledge acquisition in Biomedicine, International Journal of Medical Informatics, 2002.

[Nenadic 2004] Nenadic, G., Spasic, I., and Ananiadou. S.: Mining biomedical abstracts: What's in a term?, in Proc. of IJCNLP, Hainan, 2004.

[Ohta 2002] Ohta, T., Tateishi, Y., Tsujii, J., et.al.: GENIA corpus: an annotated research abstract corpus in Molecular biology domain, in Proc. of HLT 2002, San Diego, 2002.

[Pustejovsky 2001] Pustejovsky, J., Castano, B., Cochran, B., et.al.: Extraction and disambiguation of acronym-meaning pairs in Medlone, in Proc. of Medinfo, 2001.

[Tateishi 2004] Tateishi, Y., Ohta, T., Tsujii, J.: Annotation of predicate-argument structure on molecular biology text, in Proc. of the workshop on "Beyond shallow analyses", IJCNLP-04, Hainan, 2004.

[Tsuruoka 2003] Tsuruoka, Y., Tsujii, J.: Probabilistic term variant generator for biomedical terms, ACM SIGIR, Toronto, 2003.

[Tuason 2004] Tuason, O., Chen, L., et.al.: Biological nomenclatures: A source of lexical knowledge and ambiguities, in Proc. of PSB, Hawaii, 2004.

[Valencia 2001] Valencia, A.: Text mining for biology, in Proc. of NLP and Ontology for biology (ed: J.Tsujii), University of Tokyo, 2001