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

Junichi Tsujii
*Department of Computer Science*
*Graduate School of Information Science and Technology*
*University of Tokyo*
*Tokyo, Japan*
*tsujii@is.s.u-tokyo.ac.jp*

Sophia Ananiadou
*National Centre for Text Mining*
*School of Computing, Science and Engineering*
*Salford University*
*Manchester, UK*
*s.ananiadou@salford.ac.uk*

**Abstract.** Ontologies are recognised as important tools not only for effective and efficient information sharing but also for information extraction and text mining. In the biomedical domain, the need of a common ontology for information sharing has long been recognised and several ontologies are now widely used.

However, there is confusion among researchers on the type of ontology it is needed for text mining and how it can be used for effective knowledge management, sharing and integration in biomedicine.

We argue in this paper that there are several different views of the definition of ontology and that, while the logical view is popular for some applications, it may be neither possible nor necessary for text mining.

We propose as an alternative to formal ontologies a text-centred approach for knowledge sharing. We argue that a thesaurus (ie. an organised collection of terms enriched with relations) is more useful for text mining applications than formal ontologies.

**Keywords:** thesaurus, ontology, terminology, text mining

## 1. BACKGROUND

The currently dominant approach to knowledge sharing and integration is the ontology-centred approach. Ontologies are conceptual models which support consistent and unambiguous knowledge sharing and provide a framework for knowledge integration which ideally should be flexible, rigorous and consistent.

By thesaurus in this paper we mean a terminological thesaurus as distinct from a documentation thesaurus which is highly constrained as it typically has a narrow set of relationships (broader term, narrower term and related term) and a controlled vocabulary. A terminological thesaurus consists of a wider set of relationships pertinent to a subject domain, linking the defined concepts of that domain with the terms that realise them (including their variant forms).

While the ontology-centred approach has been successful in some applications, in others it has encountered difficulties. While small ontologies can be built manually in a top-down manner, constructing comprehensive ontologies for real applications is not a trivial task. Furthermore, in many fields of application, knowledge to be shared and integrated is presented mostly in text. Due to the inherent properties of human language, it is not straightforward to link knowledge in text with ontologies, even if comprehensive ontologies will ever be constructed.

More seriously, we suspect that in certain applications, such unambiguous and consistent conceptual models play a far less significant role in sharing knowledge than the ontology-centred approach assumes. In some cases, conceptual models across and within communities which intend to share their knowledge are inherently more fragmented and dynamic and less consistent than the ontology approach assumes.

In this paper, we propose a complementary approach, the *text-centred* approach, in which ontological commitment is kept to a minimum and, instead of using logical inferences for deriving *implicit* information, the emphasis is put on techniques of text mining and automatic knowledge acquisition for constructing ontologies from text.

We concentrate on Biomedicine, since Biomedicine has the hallmarks of a domain for which the ontology-centred approach fails to deliver effective knowledge sharing systems.

Knowledge sharing has become crucial in Biomedicine, because the recent developments in molecular biology has revealed that all creatures share, through history of evolution, common biological systems, i.e. gene/protein networks which are encoded in DNA sequences, and that all bio-medical phenomena (e.g. diseases, immunologic reaction, etc.) have their roots in these common biological systems. This implies that there is a high degree of interrelation between the areas of biology, medical and pharmaceutical sciences through such gene/protein networks, and thus knowledge in Biomedicine is highly inter-connected.

However, knowledge sharing in Biomedicine is not so straightforward. Firstly, knowledge to be shared is mostly presented in text, i.e. domain literature, and the amount of text to be shared is enormous. Although a great deal of crucial biomedical information is stored in factual databases, the most relevant and useful information is still represented in domain literature. Medline contains over 14 million records, extending its coverage by a large amount each month. Open access publishers such as BioMed Central have growing collections of full text scientific articles. There is increasing activity and interest in linking factual biodatabases to the literature, in using the literature to check, complete or complement the contents of such databases, however currently such curation is laborious, being done largely manually with few sophisticated aids, thus the risks of introducing errors or leaving unsuspected gaps are non-negligible.

Secondly, since communities which intend to share their knowledge have evolved independently of each other, they have their own vocabularies and language uses. The same proteins, for example, often have different names in different communities. More seriously, while different fields are interested in common biological systems, they are not exactly the same. Although similar proteins appear and may have similar functions in different species, their functions and properties are highly dependent on the surrounding context and not exactly the same.

Researchers who try to identify the function of a specific protein in a specific biological context, may gather all relevant facts reported in papers, including those on similar proteins. However, they do not assume that all reported facts in literature are valid for the protein in the context at hand. Rather, they will examine biological contexts in literature to choose a set of contexts similar to the one at hand and infer the function of the protein by considering and weighing all potential implications and consequences of reported facts.

Most of the widely used ontologies have been built on a top-down manner.They are limited in their conceptual coverage and they are mainly oriented for human (expert) use. The difficulties and limitations lie with the definition of concepts (classes, sets of instances) since one is expected to identify *all* instances of a concept. This task demands evidence from text.

Attempting to use ontologies to support knowledge management tasks such as classification, clustering, summarisation, indexing, information extraction, text mining etc reported disappointing results. One of the main reasons for this is the failure to match instances (terms) from text to concept labels of ontologies. This is due to the inherent ambiguous and diverse nature of language.

Inferences and knowledge-sharing in Biomedicine as such are very different from those envisaged by the ontology-centred approach in general and by formal ontologists in particular. They are more like abduction based on similarities than logical deduction. Reflecting on the nature of fragmented communities and the modes of inferences in Biomedicine, we argue in this paper that (i) terminological thesauri which maintain relationships among language uses in different communities are more important than logically consistent ontologies and (ii) bio-ontologies such as the GO (Gene Ontology) which biologists have found useful, though not completely satisfactory, are very different in nature from ontologies which the ontology-centred approach envisages.

## 2. Difficulties in the Ontology-Centred Approach

Whenever different communities want to share knowledge, both terminological and ontological problems arise. Different communities may use different terms to denote the same concept and the same terms to denote different concepts (terminological problems). It is also the case that different communities view the same entities from different facets and thus conceptualise them differently (ontological problems).

In some applications such as e-business, different communities can reach an explicit agreement on a standard ontology and a set of standard terms to denote concepts or entities in the ontology. However, in a constantly evolving domain such as biomedicine we encounter the following crucial differences: (1) Size of ontology (2) Context dependency (3) Evolving nature of science (4) Hypothetical nature of ontology (5) Inconsistency

### 2.1. Size of Ontology

The number of concepts covering ontologies in areas such as e-business is more limited than in biomedicine. For example, the UMLS metathesaurus contains

(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 may be possible to manage relationships for a small number of concepts, the task becomes intractable for a large amount of concepts such as in the above. Despite a huge number of concepts in UMLS, many of the recognised concepts do not appear simply because available resources do not represent these types of entities e.g. terms that refer to families or group of proteins (Blaschke and Valencia, 2002). Equally seriously, termforms which actually appear in text are often not registered in UMLS, since UMLS mainly focuses on conceptual information. This causes practical difficulties in sharing knowledge in text. In order to maintain such a large collection of concepts and termforms, one needs NLP tools to keep the collection up-to-date in relation to actual running text.

### 2.2. Context Dependency

The assumption in logical ontologies is that categories are explicitly defined by their defining properties and that, once an entity is judged as a member of a category, it inherits a set of other properties (derived properties). The attraction of logical ontology comes from such inference capability that presupposes static, context-independent relationships between categories and properties.

However, such context-independent relationships are not the norm in bio-medicine. Whether a protein contains certain properties or not depends on factors such as its location inside a cell, the anatomical position of a cell, the states of other bio-chemical entities around it, etc., as well as the protein category to which it belongs.

Context dependency is one of the hardest problems in logical modeling of everyday inferences in AI such as *qualification*, *non-monotonicity*, which severely restrict the utility of logically-based frameworks. Since biological entities and events portray a high-degree of context-dependency as everyday inferences, deduction would hardly be effective in Biomedicine either. It is also worthwhile to note that, because of context-dependency, bioscientists, even when they identify relevant events in curated data bases, also consult original papers from scientific literature.

### 2.3. The Evolving Nature of Science

If we compare diverse domains which ontologies are to model, we ascertain the following differences: while domains such as those in e-business are well circumscribed and understood, domains such as Biomedicine are open-ended and only partial understandings exist. In the former, ontologies are introduced in order to make the shared understanding explicit and thereby facilitate effective communication in business. On the other hand, ontologies in biology go beyond the level of effective communication: they are motivated by the need to fully understand and model the domain. One way of modeling or understanding a domain is through lexical means. That is, a new term is introduced to delineate knowledge about a concept which is considered to be useful or relevant, and to specify the properties or attributes characterising it (Sager 1990). In due course, new discoveries may change our understanding of the concept which the term denotes and subsequently change its meaning.

It is common in Biomedicine that a term introduced is subsequently found to denote several distinct concepts, thus raising a need to introduce new distinctive names. On the other hand, it is also very common that two distinct terms used in different communities are later found to denote the same concept and merged into a single term.

Due to the evolving nature of science, concepts often are not fully delineated, since they are themselves evolving. This is reflected in the degree of term variation observed in dynamic fields. Dynamically evolving fields, such as biomedicine, exhibit a high degree of term variation (Nenadic et al., 2005).

### 2.4. The Hypothetical Nature of Ontology

In scientific fields, not only the individual terms but also whole ontological frameworks are hypothetical in nature. Let us take as an example from anatomical ontologies.

In the NCI thesaurus, anatomic structure, system, or substance is classified into body cavity, body fluid or substance, body part, body region, organ, organ system, micro-anatomy etc. Within Organ, breast is classified as bronchial tree and diaphragm and differentiated between male and female breast.

Such an anatomical classification is not a transcendental object, but has been hypothesized, revised and established through the long history of medical science. There were many other classification schemas, some were based on functions of organs and others on their physical properties. For now, the NCI classification of human anatomy is, more or less, agreed upon by researchers, simply because the scheme is useful, more effective than other schemes, for explaining and understanding biomedical phenomena in humans.

In logical ontologies, classification schemas exist prior to a set of their logical consequences. On the contrary to this, in scientific ontologies a set of consequences (phenomena to be explained) pre-exists and researchers try to find an ontology by which they can derive or explain them in the most consistent manner. In other words, to build proper ontologies is a crucial step of science which looks for consistent and elegant ways of explaining reality. As we will see in Section 4-2, bio-ontologies such as the GO show characteristics of this type of ontology.

### 2.5. Inconsistency

As we discussed, deductive inferences based on formal consistent ontologies would be of limited use in Biomedicine. Closer examinations of the Gene ontology, Ancal ontologies, etc. show that logical inconsistency is abundant and that they are closer to UDC, a multilingual classification scheme, rather than a logical ontology. While researchers with formal orientation describe inconsistencies in biomedical ontologies as shortcomings, their criticisms are misplaced due to their misunderstanding of the nature of bio-ontologies, as pointed out by Ceusters et al.(2003).

## 3. Towards a Text-Centred Approach

As we have already mentioned, a complete and context-independent ontology is an unattainable goal in biomedicine. In the *text-centred* approach we take the position that most relationships among concepts as well as the concepts themselves remain implicit in text, waiting to be discovered. Thus, text mining and NLP techniques play a more important role in uncovering hidden and implicit information than logical deduction. This approach does not exclude the complementary use of explicit partial ontologies. Instead of explicit definitions, we assume that all term occurrences in text implicitly define the semantics of concepts. In addition by mining term associations, relationships among concepts are discovered.

### 3.1. The non trivial mapping between terms and concepts

As we have already reported , even within the same text, a term can take different forms. A term may be expressed via various mechanisms including orthographic variation (usage of hyphens and slashes (amino acid and amino-acid), lower and upper cases (NF-KB and NF-kb), spelling variations (tumour and tumor), various Latin/Greek transcriptions (oestrogen and estrogen) and abbreviations (RAR and retinoic acid receptor). Further complexity is introduced as authors vary the forms they use in different ways (e.g. different reductions: thyroid hormone receptor and thyroid receptor, or the SB2 gene and SB2) or use embedded variant forms within larger forms (CREB-binding protein, where CREB is in turn cAMP-response element-binding protein). This rich variety of termforms for each term is a stumbling block especially for language processing, as these forms have to be recognised, linked and mapped to terminological and ontological resources. It also causes problems to the human in cases where there is room for ambiguity or where some termform has never been seen before and its provenance (relationship to its term) is unclear.

Several approaches have been suggested to automatically integrate and map between resources (e.g. between GO and UMLS using exact string matching (Cantor et al, 2003), (Sarkar et al., 2003). Results revealed the difficulties inherent in the integration of biological terminologies, mainly in terms of extensive variability of lexical term representations, and the problem of term ambiguity with respect to mapping into a data source. For example, attempts to integrate gene names in UMLS were not successful since they increased ambiguity, and disambiguation information (particularly important for systematic polysemy) was not available in lexical resources examined.

In order to map successfully termforms in text to ontological concepts we have to deal with language variability. Several techniques dealing with term variation have been suggested.

Jacquemin and Tzoukermann conflate multiword terms by combining stemming and terminological look-up. Stemming was used to reduce words so that conceptually and linguistically related words were normalised to the same stem (thus resolving some orthographic and morphological variations), while a terminological thesaurus might be used for spotting synonyms and linking lexical variants.

Nenadic et al. 2005 incorporate different types of term variation into a base line method of automatic term recognition, the C/NC value (Frantzi et al., 2000). The incorporation of treatment of term variation enhanced the performance of the ATR system (where linking related occurrences is vital for successful terminology management).

Another approach to the recognition of term variants uses approximate string matching techniques to link or generate different term variants (Tsuruoka and Tsujii, 2003).

### 3.2. Thesauri

For biologists it is common to use two different names, e.g. *PKB* and *Akt* to denote the same protein. Taking into account the amount of new terms added daily in the field compounded by the high degree of term variability, it is not surprising that term synonyms are not recognised. Lexical variability is an important aspect of scientific communication and language use among different groups. Lexical variants and synonyms coexist with standardised terms. Synonymy relationships are often mentionned as comments in data base entries e.g. *"This protein is similar to Protein-B"*. Typically, these relationships remain hidden in the databases but are nevertheless significant for inferencing and biotext mining and as such they should be made explicit in any knowledge sharing system.

An example of a *text-centred* approach is the GENIA thesaurus which keeps track of such relationships. We assume that since the meanings of terms are only implicitly defined by all their occurrences in text, many of the relationships such as synonymy, hyponymy, meronymy etc are not further delineated. In order to make use of this hidden information existing in various heterogeneous resources we use an integrated terminological management system, TIMS, (Mima et al. 2002). TIMS (Tagged Information Management System) links term entries of the thesaurus with their occurrences in actual text, other surface terms such as synonyms, related terms such as homologues, orthologues and their ID record from various biodatabases.

### 3.3. Thesauri and Knowledge

Ideally, terms are monoreferential, ie. a term coincides to a concept. In reality, this is more of an exception than the norm. Let us observe the following examples from biomedicine: *Cycline-dependent kinase inhibitor* was first introduced to represent a protein family with only one extention, *p27*. However, *cycline-dependent kinase inhibitor* is used interchangeably with *p27* or *p27kip1*, as the name of the individual protein and not as the name of the protein family (Morgan 2003). In the case of *NFKB2*, the term is used to denote the name of a family of two individual proteins with separate **id**'s in SwissProt. These proteins are homologues belonging to different species, human and chicken.

The above examples demonstrate that it is rather difficult to establish equivalences between term forms and concepts. In effect, many proteins have dual names to also denote the protein family they belong to. *MAP kinase* is a family name including more than 40 individual proteins and because of the number of individual proteins in the family, it is never used as the name of individual proteins.

Since surface textual cues cannot distinguish between a genuine family name from individual protein names derived from family names, the thesaurus should include relationships of term forms with their actual denotations, i.e. **id**'s in various data bases.

A thesaurus links surface terms with data base **id**'s and other types of information in diverse data bases of proteins (SwissProt), genes (LocusLink), pathways(KEGG, TRANSFAC), etc. However, it is worth noting that a thesaurus does not presuppose a single, logically consistent ontology.

### 3.4. Minimum Ontology and Ambiguous Terms

In order for a thesaurus to be useful, it should maintain not only relationships among surface forms but should be able to deal with term ambiguity.

Gene names are often used to denote gene products (proteins) that they encode. Although there are many definitions of the term *gene*, it is nevertheless obvious that there are two distinct classes of entities, *genes* and *proteins*. A term like *suppressor of sable* is used ambiguously to refer to either one of these two classes genes and proteins which are ontologically very different. While domains are part of proteins, names of domains are sometimes used as the names of proteins that contain them as part.

It is important to note that, without commitment to the ontological distinction between *gene* and *protein* or *domain* and *protein*, we could not capture even such an obvious ambiguity. We need therefore an ontology which can represent and include term ambiguity; we call such an ontology a *minimum ontology.* The minimum ontology is *linguistically* motivated and acts as an interface to more detailed bio-ontologies. An example of a minimum ontology is GENIA (Ohta et al., 2002) which consists of 36 ontological classes. These classes are equivalent to the classes of named entity recognisers based on linguistic cues. Referential distinctions such as homologues, orthologues etc are not part of the minimum ontology.

## 4. The Nature of Inferences and Bio-Ontologies

In formal ontologies, there is emphasis on the soundness and completeness of the underlying deductive inference mechanism. In biology the nature of inferencing mechanism is different as more emphasis is given to the ability to make new plausible hypotheses.

### 4.1. An example of inferencing from biology

In order to illustrate our point about the nature of inferencing in biology, let us consider the following example.

**(1)** Results from a biological experiment (micro-array data) showed that three proteins, FLJ23251, BM002 and CGI-126, interacted with each other, and that this interaction was peculiar to patients with a specific disease. Based on these results, further information was needed to understand the mechanisms of the interaction.

**(2)** A comment from a bio-database recorded that *"this protein - ZK652.3 - is similar to human bone marrow protein BM002"* in the entry of ZK652.3. Further literature search, retrieved a paper on ZK652.3 with the statement that ZK652.3 has ubiquitin-like fold. From these two pieces of information, the biologist hypothesized that BM002 is actually ubiquitin and that the whole process is of ubiquitination (a type of protein degradation process).

**(3)** In another scientific paper we found that FLJ23251 has *ubiquitin-activating enzyme E1-domain.* This strengthened the hypothesis in step (2).

**(4)** Since the process of ubiquitination often involves another two enzymes, E2 or E3, we can hypothesize that CGI-126 would be either one of these two enzymes. From this hypothesis we can then look for further information of CGI-126.

The key to the whole process is Step (2), where two uncertain and vague statements are combined to form a hypothesis. This step is abductive in nature, and the subsequent steps help us to improve the plausibility of the hypothesis by gathering further evidence. Unlike in the process of deduction, as long as further evidence may improve the plausibility of an hypothesis, the hypothesis is not logically implied. Either the hypothesis would become plausible enough to be believed or it should to be validated by biological experimentation.

An additional point is that in step (2) we use a vague relationship of *being-similar-to* and that this similarity does not logically imply that *BM002 has also ubiquitin-like fold*. It only suggests that it is plausible to assume so.

Other relationships in biology such as homologues and orthologues are used in the same way as *being-similar-to*. They imply that part of the DNA sequences in different spieces are so similar that they are considered to be preserved across species through the history of evolution. It practical terms the implication is that two genes and their products (proteins) are likely to share common functional roles in similar networks in different spieces. Orthologues are most likely to share the same properties, while just similar proteins share the least properties. Such quantitative nature of inferences is a hallmark of abduction, and is being modeled, not by logical frameworks, but by models such Bayesian networks, etc.

### 4.2. Bio-ontology - the GO

The crucial step in abduction is making plausible hypotheses based on evidences. This step should involve biologists who have to search through a huge space of possibilities. In order to help biologists to gather evidence from large scale knowledge bases to form plausible hypotheses, classifications (ie classifying functions and processes and relating them to proteins and genes) and/or ontologies are needed. This is where the power of text mining can help: it can play a major role in the abductive process.

The Gene Ontology (GO), one of the most widely used bio-ontologies, aims to attain the same target as the text mining in the above. That is, by establishing an explicit classification schema, it intends to help biologists to gather facts on proteins which appear in similar biological contexts. As for such classification of biological contexts, the GO has three schemes, (1) cellular components (the location inside cells where proteins appear), (2) molecular functions and (3) biological processes. Under these three headings, the GO has a set of controlled vocabulary containing around 17,000 terms. Whether the GO is useful or not is judged by how effective the classification schemes are to retrieve relevant proteins in similar biological contexts, relevant for identification of unknown functions of a protein in a given biological context. As with anatomical ontologies, the whole classification scheme is based on hypotheses that factors chosen for classification are relevant to the task at hand.

It has also been suggested that the GO classes can be used as evidence in abductive reasoning. Thus, we can rank the plausibility of interactions of proteins by assuming that proteins reported to be in similar processes with similar roles and exist in similar locations are more likely to interact with each other.

## 5. Concluding Remarks

We have described a text centred approach to knowledge mining from large repositories of biomedical literature. One of the most important advantages of this approach is that it is data-driven, as the terminological information is collected dynamically from corpora. This is particularly important for domains such as biomedicine, as there is typically a gap between terms used in corpora and controlled vocabularies. If we take into account the pace of creating new terms, standardisation issues will still be a problem in the near future. Thus, the aim of a text centred approach to knowledge management is to provide tools to bridge that gap and facilitate effective mining and integration of scientific literature, experimental data, ontologies and databases.

Our system TIMS explores similar ideas such as that a major source of knowledge comes from text from which we derive information and that terms (instances in text) play a crucial role in the integration of knowledge sources, instead of a common ontology.

In TIMS, a set of operations on segments of text similar to those of *Regional Algebra* was the core for retrieving and deriving information from text. While such operations still play a central role, we plan to integrate them with more quantitative methods and with other text mining techniques.

We also plan to extend the linguistic units for integrating knowledge sources from simple terms to complex expressions of events. Events which are identified and extracted by information extraction techniques are to be annotated in text and used as units for accessing various knowledge sources. This method will make the links between records in curated data bases and relevant portions of text much clearer and will satisfy the users' demands to access and read original papers once relevant curated facts are located.

It is also a crucial step to integrate our work with the ontology-centred approach. One possible extention is to use our system to populate incomplete, existing ontologies. Classification of terms is essential for mapping to referent databases and knowledge integration. Some steps in this direction have been already made (Spasic and Ananiadou, 2004) for a term classification method that is guided by verb complementation patterns; also (Spasic and Ananiadou, 2005) presents a flexible variant of the edit distance to compare various contextual features for measuring term similarities that is used for classification).

Although to illustrate our point we used biomedicine as an example, our techniques are domain independent and applicable to other domains complementing the ontology-based approach in many knowledge management and sharing applications.

## References

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

Ananiadou, S., Friedman, C., Tsujii, J. (Eds) (2004) *Named Entity Recognition in Biomedicine*, Special Issue, *Journal of Biomedical Informatics*, vol. 37 (6),2004.

Blaschke, C., Hirschman, L., Valencia, A. (2002) Information Extraction in Molecular Biology. *Briefings in Bioinformatics*, 3(2): 154-165

Blaschke, C., Valencia, A. (2002) Molecular biology nomenclature thwarts information-extraction progress. *IEEE Intelligent Systems* 17(3). 73-76.

Ceusters, W., Smith, B., Kumar, A., Dhaen, C.(2003) Mistakes in Medical Ontologies: where do they come from and how can they be detected? In Pisanelli, D. (Ed.) *Ontologies in Medicine*. Proceedings of the workshop on medical ontologies, Rome.

Chang, J., Schutze, D., Altman, R.(2002) Creating on-line dictionary of abbreviations from Medline, *Journal of the American Medical Informatics Association.*

Frantzi, K., Ananiadou, S., Mima, H. (2000) Automatic Recognition of Multi-Word Terms: the C/NC value method. *International Journal of Digital Libraries*, vol. 3:2, pp. 115-130.

Hirschman, L., Park, J., Tsujii, J., Wong, L. Wu, C. (2002) Accomplishments and Challenges in Literature Data Mining for Biology, in *Bioinformatics*, vol. 18, no 12, pp. 1553-1561

Jacquemin, C., Tzoukermann, E. (1999) NLP for Term Variant Extraction: A Synergy of Morphology, Lexicon and Syntax. In T. Strzalkowski, Ed. *Natural Language Information Retrieval*, Kluwer: Boston. p. 25-74.

Jacquemin, C. (2001) *Spotting and discovering terms through NLP*, MIT Press.

Mima, H., Ananiadou, S., Nenadic, G., Tsujii, J., (2002) A methodology for terminology-based knowledge acquisition and integration, in *Proceedings of 19th Int. Conference on Computational Linguistics*, Taipei, Taiwan, 667-673

Mima, H., Ananiadou, S., Matsushima, K., (2004) Design and Implementation of a Terminology-based literature mining and knowledge structuring system, in *Proceedings of CompuTerm*, Coling, Geneva, Switzerland.

Morgan, A., Yeh, A., Hirshman, L. (2004) Gene name extraction using FlyBase resources. In Ananiadou, Friedman and Tsuji (eds) *Named Entity Recognition in Biomedicine*, Special Issue, *Journal of Biomedical Informatics*, vol. 37 (6).

Nenadic, G., Mima, H.,Ananiadou, S. Tsujii, J. (2002) Terminology-based literature mining and knowledge acquisition in Biomedicine, in *International Journal of Medical Informatics.*

Nenadic, G., Spasic, I., Ananiadou, S. (2005) Mining Biomedical Abstracts: What o?s in a Term? In Keh-Yih Su, Jun o?ichi Tsujii, Jong-Hyeok Lee, et al (Eds.) *Natural Language Processing IJCNLP 2004* First International Joint Conference , Lecture Notes in Computer Science vol. 3248, 2005

Ohta, T., Tateishi, Y., Tsujii, J., et.al. (2002) GENIA corpus: an annotated research abstract corpus in Molecular biology domain, in *Proceedings of HLT*, San Diego.

Pustejovsky, J., Castano, B., Cochran, B., et.al. (2001) Extraction and disambiguation of acronym-meaning pairs in Medline, in *Proceedings of Medinfo.*

Sager, J.C. (1990) *A Practical Course in Terminology Processing*, John Benjamins Publ. Company.

Spasic, I., Ananiadou, S. (2004) Using Automatically Learnt Verb Selectional Preferences for Classification of Biomedical Terms. In Ananiadou, S., Friedman, C., Tsujii, J. (Eds) *Named Entity Recognition in Biomedicine*, Special Issue, *Journal of Biomedical Informatics*,vol.37, (6), 483-497.

Spasic, I., Ananiadou, S., Tsujii, J. (forthcoming) MaSTerClass: a case-based reasoning system for the classification of biomedical terms, in Journal of Bioinformatics (accepted for publication), Oxford University Press.

Tateishi, Y., Ohta, T., Tsujii, J.(2004) Annotation of predicate-argument structure on molecular biology text, in *Proceedings of the workshop on Beyond shallow analyses* IJCNLP-04, Hainan, China.

Tauson, O., Chen, L., et.al. (2004) Biological nomenclatures: A source of lexical knowledge and ambiguities, in *Proceedings of PSB*, Hawaii.

Tsuruoka, Y., Tsujii,J (2003) Probabilistic term variant generator for biomedical terms, in *Proceedings of ACM SIGIR*, Toronto.

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

MEDLINE. 2004. National Library of Medicine. Available at: http://www.ncbi.nlm.nih.gov/PubMed

National Cancer Institute Thesaurus available at http://ncicb.nci.nih.gov/

UMLS http://www.nlm.nih.gov/research/umls/

Universal Decimal Classification (UDC) consortium available at http://www.udcc.org/