Text mining and ontologies in biomedicine: Making sense of raw text

Irena Spasic, Sophia Ananiadou, John McNaught and Anand Kumar

Date received (in revised form): 7th June 2005

Irena Spasic
is a postdoctoral research associate in the School of Chemistry and the Manchester Interdisciplinary Biocentre at the University of Manchester. Her research interests include biomedical text mining, machine learning and bioinformatics.

Sophia Ananiadou
is co-director of the UK National Centre for Text Mining and a Reader in Computer Science at the University of Salford. Her research interests are in the areas of computational terminology and biomedical text mining.

John McNaught
is a Lecturer in the School of Informatics at the University of Manchester and an Associate Director of the UK National Centre for Text Mining. His research interests include information extraction and computational lexicography.

Anand Kumar
is Alexander von Humboldt research fellow in the Faculty of Medicine at the University of Leipzig and a member of the Institute for Formal Ontology and Medical Information Science at Saarland University in Saarbrücken. His research interests include medical and biomedical knowledge representation, data models and ontologies.

Keywords: *text mining, ontology, terminology, information extraction, information retrieval*

Irena Spasic,
School of Chemistry,
The University of Manchester,
Sackville Street,
PO Box 88,
Manchester M60 1QD,UK

Tel: +44 (0)161 306 4414
Fax: +44 (0)161 306 4556
E-mail: i.spasic@manchester.ac.uk

Abstract

The volume of biomedical literature is increasing at such a rate that it is becoming difficult to locate, retrieve and manage the reported information without text mining, which aims to automatically distill information, extract facts, discover implicit links and generate hypotheses relevant to user needs. Ontologies, as conceptual models, provide the necessary framework for semantic representation of textual information. The principal link between text and an ontology is terminology, which maps terms to domain-specific concepts. This paper summarises different approaches in which ontologies have been used for text-mining applications in biomedicine.

INTRODUCTION

Text is the predominant medium for information exchange among experts.[1] The volume of biomedical literature is increasing at such a rate that it is difficult to efficiently locate, retrieve and manage relevant information without the use of text-mining (TM) applications. In order to share the vast amounts of biomedical knowledge effectively, textual evidence needs to be linked to ontologies as the main repositories of formally represented knowledge. Ontologies are conceptual models that aim to support consistent and unambiguous knowledge sharing and that provide a framework for knowledge integration.[2] An ontology links concept labels to their interpretations, ie specifications of their meanings including concept definitions and relations to other concepts.[3] Apart from relations such as is-a and part-of, generally present in almost any domain, ontologies also model domain-specific relations, eg has-location, clinically-associated-with and has-manifestation are relations specific for the biomedical domain. Therefore, ontologies reflect the structure of the domain and constrain the potential interpretations of terms. As such, ontologies can be used to support automatic semantic interpretation of textual information (Figure 1), and thus provide a basis for sophisticated TM.

Table 1 lists some popular biomedical ontologies. Many such ontologies exhibit differing degrees of overlap, exhaustivity and specificity and indeed differing views over conceptual space. Therefore, TM applications that rely on multiple ontologies also need to include methods for mapping between such ontologies.[4] These methods, together with other biomedical applications (including TM) that rely on the use of ontologies, would benefit from a standard ontology language (eg using standard initiatives such as RDF[5] and OWL[6]). Still, even when a single standardised ontology is used, it is not always straightforward to link textual information with ontology owing to the inherent properties of language. Two major obstacles are: (1) inconsistent and imprecise practice in the naming of biomedical concepts (terminology),[7] and (2) incomplete ontologies as a result of rapid knowledge expansion.

Nonetheless, a comprehensive body of knowledge is currently stored in biomedical ontologies, which can be utilised in numerous ways by TM applications. Moreover, the results of TM can be curated and used to facilitate

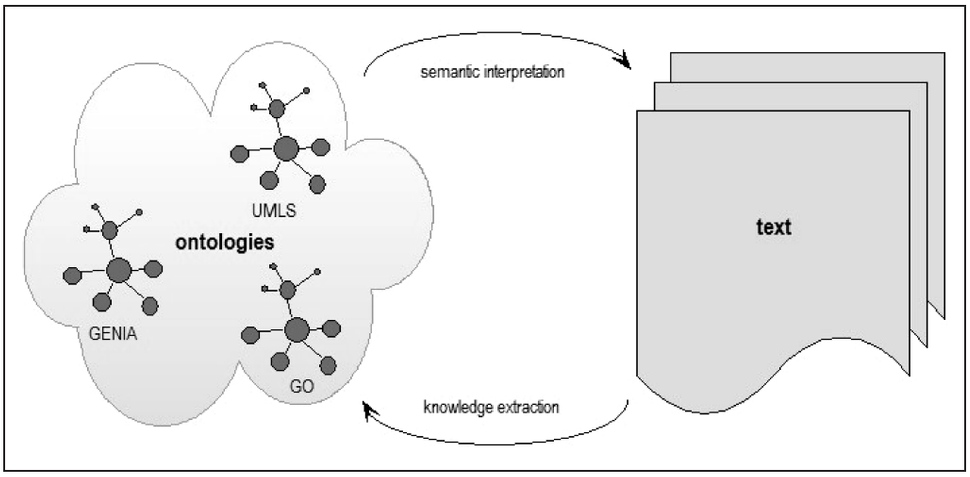

Figure 1: Ontologies provide machine-readable descriptions of biomedical concepts and their relations. Linking domain-specific terms, ie textual representation of these concepts, to their descriptions in the ontologies provides a platform for semantic interpretation of textual information. An explicit semantic layer supported by the use of ontologies allows text to be mined for interpretable information about biomedical concepts as opposed to simple correlations discovered by mining textual data using statistical information about co-occurrences between targeted classes of biomedical terms. The knowledge extracted from text using advanced TM can then be curated and used to update the content of biomedical ontologies, which currently lag behind in their attempts to keep abreast of new knowledge owing to its rapid expansion

update of biomedical ontologies (Figure 1). In this paper the focus is on only the former aspect of the relation between text mining and ontologies, ie problems, existing practice and prospects of using ontologies for different TM applications are reviewed. The section 'Terminology' focuses on the problem of linking text to ontologies. The section 'Text mining' provides an introduction to TM and discusses two of its principal tasks: information retrieval and information extraction. The ways in which ontologies can be used to support these applications are discussed separately in the following sections: 'Information retrieval' and 'Information extraction'. The latter section is divided into three subsections. The first subsection deals with named entity recognition as a key step in information extraction. The following two subsections discuss information extraction systems depending on the degree to which they rely on the use of ontologies. Since many TM applications resort to the use of machine learning methods as a way of tackling the complexity of both natural language and biomedical knowledge, it is explained how ontologies can be used for this purpose in the section 'Machine learning'. The conclusion completes the paper.

Table 1: Selected generic biomedical ontologies

Name	URL
UMLS	http://www.hlm.nih.gov/research/umls/
SNOMED	http://www.snomed.org/snomedct/
GENIA	http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/
GALEN	http://www.opengalen.org/about.html
TaO	http://imgproj.cs.man.ac.uk/tambis/
GO	http://www.geneontology.org/

OBO (Open Biomedical Ontologies) provides a more comprehensive list of ontologies and is available at http://obo.sourceforge.net/

TERMINOLOGY

The principal link between text and an ontology is a *terminology*, which aims to map concepts to terms (Figure 2). A *term* is defined as a textual realisation of a specialised concept, eg gene, protein,

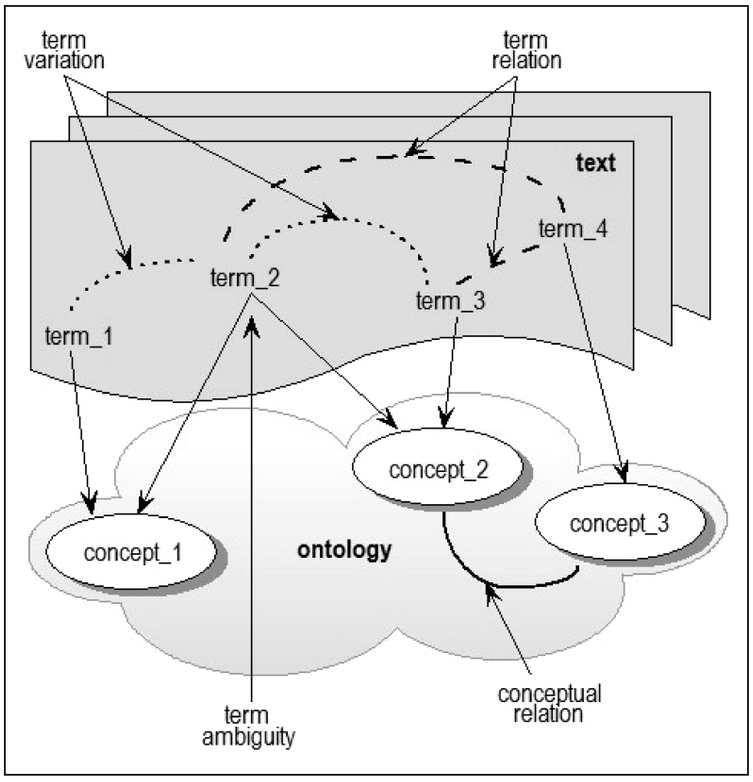

Figure 2: Conceptual relations reflect the connections between the concepts denoted by the given terms. These relations may be general relations commonly found in every domain (e.g. is-a, part-of or similarity relation) or they can be confined to a specific domain (e.g. *activation* of receptors by hormones). Conceptual relations are encoded in ontologies. Term ambiguity and term variation represent specialisation of general lexical relations, namely synonymy and homonymy. These relations exist on the lexical level and do not describe the relations between the underlying concepts

Biomedical terminology

disease. The introduction of a new term presupposes the establishment of a new concept which points to a specific area of the domain knowledge space.[8,9] This process assumes the mapping of a term to a concept in an ontology. This mapping is crucial for semantic interpretation in TM applications and is far from trivial. The main problems arise from the fact that there is often no one-to-one correspondence between concepts and terms. In practice, TM applications are faced with the problems of term variation and term ambiguity, which make the integration of information available in text and ontologies difficult.

Text mining challenges in biomedicine

Term variation originates from the ability of a natural language to express a single concept in a number of ways. For example, in biomedicine there are many synonyms for proteins, enzymes, genes, etc. Having six or seven synonyms for a single concept is not unusual in this domain.[10] The probability of two experts using the same term to refer to the same concept is less than 20 per cent.[11] In addition, biomedicine includes pharmacology, where numerous trademark names refer to the same compound (eg *Advil*, *Brufen*, *Motrin*, *Nuprin* and *Nurofen* all refer to *ibuprofen*).

Term ambiguity occurs when the same term is used to refer to multiple concepts. Ambiguity is an inherent feature of natural language. Words typically have multiple dictionary entries and the meaning of a word can be altered by its context. Sublanguages, as the languages confined to specialised domains,[12] provide a context which generally reduces the level of ambiguity. However, biomedicine encompasses a plethora of subdomains, which is an additional cause for the high level of ambiguity in biomedical terminology. For example, the term *promoter* refers to a 'binding site in a DNA chain at which RNA polymerase binds to initiate transcription of messenger RNA by one or more nearby structural genes' in biology, while in chemistry it denotes a 'substance that in very small amounts is able to increase the activity of a catalyst'. In addition, acronyms are extensively used in biomedicine (a new acronym is introduced in every five to ten abstracts)[13] and they are known to be highly ambiguous (>80 per cent of acronyms are ambiguous, the average number of possible interpretations being >15).[14] For example, *AR* could be expanded to any of the following terms: *Androgen Receptor*, *AmphiRegulin*, *Acyclic Retinoid*, *Agonist–Receptor*, *Adrenergic Receptor*, etc.

Furthermore, text is not the only origin of ambiguity in biomedicine. Ambiguity is inherent to the field, because the evolution of species gave rise to many homologues and analogues. For instance,

NFKB2 denotes a family of two individual proteins with separate identifiers in Swiss-Prot. These proteins are homologues belonging to different species, human and chicken.[15]

TEXT MINING

Originally, TM was defined as the automatic discovery of previously unknown information by extracting information from text.[1] However, in the biomedical community, the term TM is often reduced to the process of highlighting (ie retrieving or extracting) small nuggets of relevant information from large collections of textual data. Generally, TM is used to collectively denote computer applications that aim to aid experts in making sense of large amounts of text by distilling information, extracting facts, discovering implicit links and generating hypotheses relevant to user needs. TM typically consists of:

Text mining tasks

- information retrieval (IR), which gathers and filters relevant documents;[16]
- information extraction (IE), which selects specific facts about prespecified types of entities and relationships of interest;[17]
- data mining (DM), which is used to discover unsuspected associations between known facts.[18] For example, mining of textual data succeeded in linking magnesium deficiency to migraine, a correlation which was later experimentally confirmed.[19]

Text processing and representation

The techniques for IR, IE and textual DM can be applied to either raw or structured text (Figure 3) with different success rates. Raw text is digitally represented as a sequence of characters. Such plain text representation is usually processed to add structure explicitly in a machine-readable form. The initial step in automatic text processing is *tokenisation*,[20] which identifies the basic textual units which need not be further decomposed. Even this basic problem cannot be resolved straightforwardly by relying on white spaces and punctuation marks as explicit delimiters (eg *[3H]R1881* is a single token).

Tokenisation is typically followed by some form of *lexical processing*, which may include *part-of-speech tagging* (mapping of individual words to their lexical classes, eg noun, verb, adjective),[21] word *stemming* (reducing a word to its stem or root form, eg both *inhibitor* and *inhibited* are reduced to *inhibit*)[22] or *lemmatisation* (mapping a word to its lemma or the base form, eg *bind* is the lemma for *binds*, *bound*, *binding*).

Syntactic processing usually involves parsing as the process of determining the syntactic structure of a whole sentence (full or deep parsing) or some of its parts (partial or shallow parsing).[23] Syntactic structure often implies the semantic relations between the concepts described. The interpretation of the semantic content expressed by natural language requires linguistic knowledge and some degree of general knowledge. In specialised domains such as biomedicine, it also requires domain-specific knowledge. Scientific publications do not explicitly encode all the necessary information needed to understand the reported conclusions. The targeted reader is assumed to possess some expertise in the area. For example, a biomedical expert should be able to infer that *5 alpha-dihydrotestosterone* is a *hormone*, *[3H]R1881* is used as a *ligand* and *androgen receptor* is a *receptor*. For TM applications to take a step closer to natural language understanding,[24] such specialised knowledge needs to be encoded in a machine-readable form to a great extent. Biomedical ontologies currently provide (partial) coverage of the domain, and thus can be used in TM applications together with other forms of knowledge (eg linguistic) to aid semantic interpretation of biomedical publications.

With each layer of annotation (lexical, syntactic and semantic), better opportunities for more sophisticated analysis arise. For example, a simple search

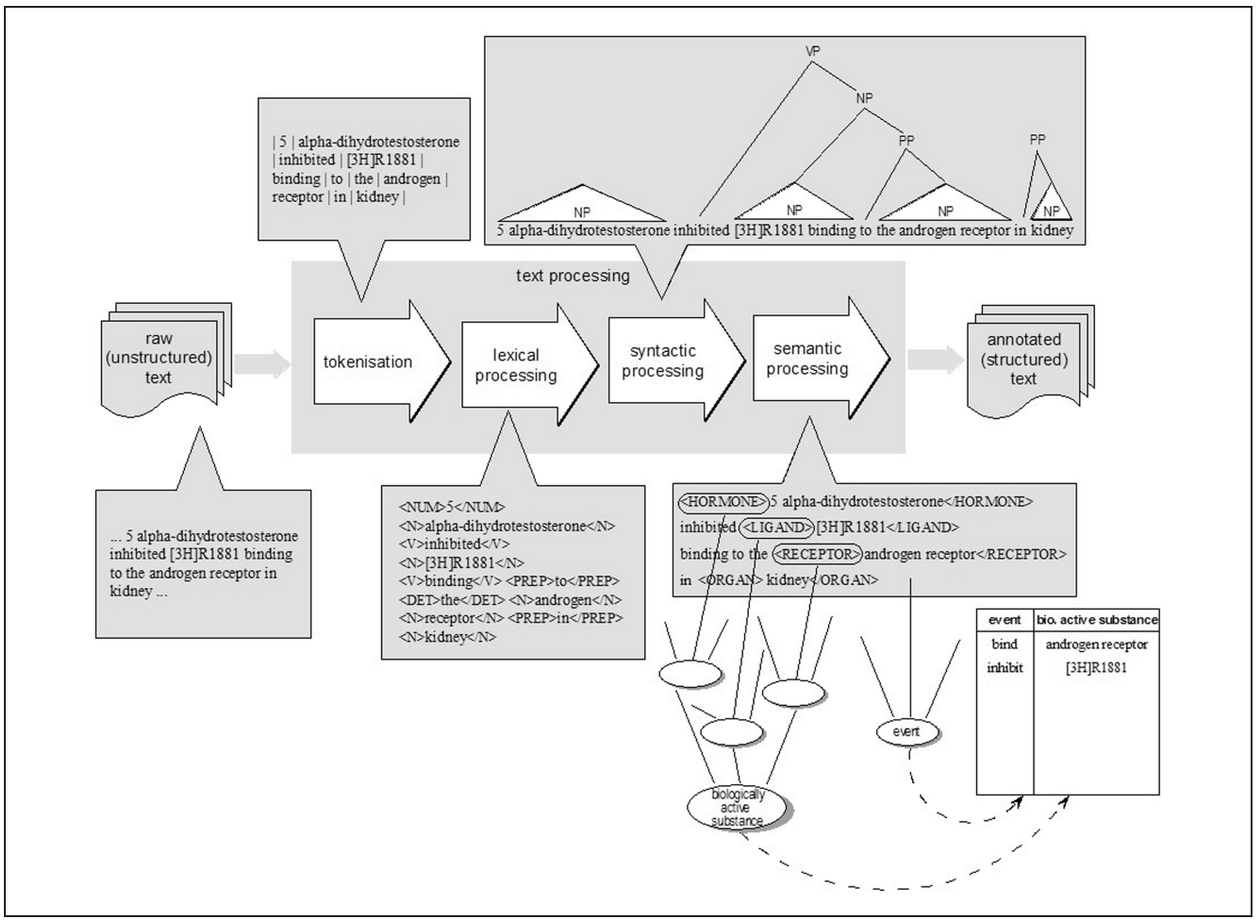

Figure 3: Generic illustration of a possible pipeline of text processing modules, not necessarily reflecting any particular analysis technique. For example, a sublanguage analysis would probably not separate syntactic and semantic processing, as it would use a syntactico-semantic approach. Each module enhances text representation with a layer of annotation, which represents explicit linguistic and/or semantic information attached to text in machine-usable form. Such information is inferred by a human reader through the use of (1) linguistic and general knowledge, and (2) domain-specific expertise, but for the text to be analysed automatically at a higher semantic level, a large part of such knowledge has to be explicitly represented in a machine-readable form. Here the focus is on domain-specific knowledge and ontologies as the means of their machine-readable descriptions. The figure illustrates a possible role of ontologies in semantic interpretation as part of overall text processing

with a query term *testosterone* over raw text is not able to differentiate between a single token *testosterone* and *testosterone* as part of other tokens (eg *5 alpha-dihydrotestosterone*). Similarly, searching for tokens of the *inhibit* relation by using a single search term *inhibit* over tokenised text would not retrieve other forms of the same word (eg *inhibits* or *inhibited*), while this is simply achieved in a lemmatised text by looking for the lemma *inhibit*. Further, syntactic information can be used to differentiate between the genuine occurrences of query terms and their nested occurrences within other terms (eg *androgen v. androgen receptor*).

While most of these problems can be tackled effectively to a certain degree using various heuristics, a real window of opportunity for sensible TM opens only by adding structured semantic information to text representation. An explicit semantic layer supported by the use of ontologies offers a higher expressive power for formulating semantic queries as opposed to simple

Semantic annotation

Boolean queries and keyword matching. Furthermore, semantically annotated text coupled with ontologies can be mined for higher-order relations between biomedical entities including temporal, causal, conditional and other types of relations (eg the conditions that produce a sequence of events that results in the expression of a disease with genetic predisposition) as a contrast to simple correlations between them (eg gene–disease associations).

Until recently, most TM systems have used neither a sophisticated terminological lexicon nor an ontology of entities or of events. They have used gazetteers, which map between a look-up string and a tactically useful semantic category, from a small set. However, the gazetteer-based approach is not suited for biomedical TM, because terminology plays a crucial part in characterising knowledge in the domain. This is one of the main reasons why biomedical TM systems generally provide poorer results compared with other domains (eg newswire-type data). The following sections describe how ontologies can be used to support various TM-related tasks.

Information extraction

INFORMATION RETRIEVAL

Information retrieval

IR is extensively used by biomedical experts to locate relevant information (most often in the form of relevant publications) on the Internet. Apart from general-purpose search engines such as Google[tm], many IR tools have been designed specifically to query the databases of biomedical publications such as PubMed.[25–29]

It is particularly important in biomedicine not to restrict IR to exact matching of query terms, because term ambiguity and variation phenomena may cause irrelevant information to be retrieved (low precision) and relevant information to be overlooked (low recall). Some biomedical ontologies (eg UMLS) explicitly store such terminological information (though not always complete). In addition, the hierarchical organisation of ontologies and relations[30] between the described concepts (and through them the corresponding terms) can be used to constrain or relax a search query and to navigate the user through huge volumes of published information.

For example, Suarez *et al.*[31] utilised UMLS for this purpose. Similarly, TIMS[32] uses an ontology to perform a sophisticated search, which enables users to access implicitly stated relevant information through hierarchical query expansion. More recently, Müller *et al.*[33] developed Textpresso, an IR system operating at the sentence level. It uses a specifically designed ontology to query a corpus for information on specific classes of biological concepts (gene, allele, cell, etc) and their relations (association, regulation, etc).

INFORMATION EXTRACTION

Early efforts in biomedical IE were devoted to named entity recognition (NER) – the recognition of terms denoting specific classes of biomedical entities (eg gene and protein names),[34] followed by the extraction of specific relations between such entities (eg protein–protein interactions),[35] progressing slowly towards extracting more complex types of information (eg metabolic pathways).[36] In this section, an overview is given of the existing approaches to these problems that rely on the use of ontologies. First, the focus is on NER as a crucial step in extracting more complex types of information (ie facts and events). The following subsections look at how ontologies are used in IE systems to extract facts and events, focusing on rule-based systems, bearing in mind that there have been few attempts to apply machine learning (ML) techniques to fact or event extraction.[35] Here an important distinction is made between ontology-based and ontology-driven systems.

Named entity recognition

IE depends on NER (ie term recognition, classification and mapping to designated concepts) as the main step in accessing

Named entity recognition

textually described domain-specific information.[38] As already mentioned, the mapping between terms (in text) and concepts (in an ontology) is not trivial. One of the main reasons is that terms exhibit a high degree of variation, which is not always explicitly reflected in biomedical ontologies.[39] For this reason, the UMLS ontology is distributed together with computational support for neutralisation of variation in the biomedical domain.[40]

Typically, one-third of term occurrences are variants,[41] which means that many new terms can be recognised as variants of known terms. Therefore, a list of classified terms that can often be derived from a biomedical ontology can be used as a training set to automatically detect new terms. (We differentiate between the use of the word term in this paper and the same word used in some of the biomedical ontologies, where it is used as a concept label (eg GO terms). Unfortunately, such concept labels have little to do with terms as they occur in text or as they are found in term banks. Many ontological 'terms' are not attested linguistic units. Instead, they have more in common with documentation thesaurus descriptors, facet labels or index terms from a controlled vocabulary than with terminological terms.)

Passive ongology use

Chiang and Yu[42] used a rule-based approach and the Gene Ontology to support robust dictionary-based term recognition. They consider variants arising from permutation (same words, but in different order, eg *inner mitochondrial membrane v. mitochondrial inner membrane*) and insertion/deletion (eg *focal adhesion associated kinase v. focal adhesion kinase*). In addition, edit distance is calculated to measure the reliability of the term variant recognition through the above rules.

Tsuruoka and Tsujii[43] implemented an approach to the recognition of orthographic variants (eg *EGR-1 v. EGR 1 v. EGR1*), which are a common type of variation in protein names. Such variants were automatically recognised by applying approximate string matching techniques for the known protein names against a domain-specific corpus. The UMLS ontology was used to provide training data.

Tsuruoka and coworkers[44,45] also developed a probabilistic term variant generator. In rule-based variant generators, arbitrary variants may be produced, resulting in a large number of non-existing variants, whose matching against a corpus consumes time and resources unnecessarily. In order to reduce this problem, each generated variant is assigned a probability factor corresponding to its plausibility. Rules are defined as applications of allowed operations (substitution, deletion and insertion) in a given context. They are learnt together with their probabilities from raw text.

Mukherjea *et al.*[46] used UMLS to extract biomedical term formation patterns and learn classification rules, which are then used to semantically annotate different classes of terms in text.

Ontology-based IE

Ontology-*based* IE systems attempt to map a term occurring in text to a concept in an ontology, typically in the absence of any explicit link between term and concept. This is passive ontology use. When such mapping is attempted depends on the type of approach adopted. For example, where syntactic chunking (identification of major syntactic constituents such as noun and verb phrases) is followed by syntactic parsing (linking syntactic constituents to build the representation(s) of an entire sentence), ontology look-up will occur after a syntactic parse has been obtained. Where a hybrid, syntactico-semantic approach is adopted, there can be early look-up of an ontology. Where term recognition is applied, ontological categories can be assigned early, instead of ad hoc semantic ones. Leroy and Chen[47] provide an example of late-stage ontology (GO, HUGO and UMLS) look-up. In another approach,[48] late-stage attempts to map

tokens of relations to concept labels in ontologies were a major source of failure: the technique called for at least one word from each argument of a relation to exist in GO or HUGO, and for at least one word forming the predicate to exist in a list of domain verb stems.

Kim and Park[49] applied full syntactic parsing, but only on sentences containing instances of predefined patterns involving keywords. Extracted general biological interaction information is annotated with GO concepts. There is an attempt to exploit similarities of sentential syntactic dependencies and ontology label syntactic structure to achieve mapping to concepts.

Ontology-driven IE

Active ontology use

Ontology-*driven* IE systems, unlike ontology-based ones, make active use of an ontology in processing, to strongly guide and constrain analysis. For example, Daraselia *et al.*[50] employ a full sentence parser[51] and a domain-specific filter to extract information on protein–protein interactions. Each of the potentially many thousands of semantic analyses per sentence is filtered against a custom-built frame-based ontology to yield a frame tree, a representation in which ontological frames are instantiated and linked according to the constraints expressed in the ontology. Frame trees are converted to conceptual graphs, which can then be subjected to querying or used as a basis for advanced mining.

PASTA[52] extracts information on the roles of specific amino acid residues in protein molecules. An ontology-based domain model is incrementally populated with the contents of predicate–argument structures, with inference and co-reference also contributing to enrich the domain model.

GenIE[53] extracts information on biochemical pathways, and on sequences, structures and functions of genomes and proteins. It makes use of an ontology linked to a semantic lexicon, in which fillers of verbal semantic subcategorisation slots are particular concepts, or specialisations thereof. It applies syntactic, semantic and ontological constraints to filter out implausible analyses, and integrates extracted information in discourse-level semantic representations.

GENIES[54,55] adopts a strong sublanguage approach, which leverages the specific informational structure of specialised texts to reduce ambiguity. This approach is applied to extraction of biomolecular interactions relevant to signal transduction and biochemical pathways, using hybrid syntactico-semantic rules. A small number of semantic categories relevant to the biomolecular domain is used. In addition, an ontology was developed,[56] covering both entities and events. Friedman *et al.*[55] describe how the semantic categories that verbs look for in their environment are mapped to the more general categories found in ontologies.

As evidenced by the results reported on the described systems, an ontology-driven IE approach is to be preferred to an ontology-based approach for extraction of relations, facts and events. Hybrid syntactico-semantic approaches offer promising results, particularly where these are based on a strong sublanguage approach and are linked with an ontology-driven approach.

MACHINE LEARNING

Machine learning in text mining

Previously, the potential of using an ontology as a training set for NER as a specific task of TM has been illustrated. For this purpose, an ontology is reduced to a list of classified terms. However, ontologies provide much richer information, which may be utilised by ML approaches to other TM tasks, such as term classification, term clustering and term relation extraction.

Numerous ML approaches have used the GENIA corpus, semantically annotated with its own custom ontology, as the training or testing set for different TM tasks:[57] eg NER[7] using methods such as hidden Markov models,[58,59] naive Bayes classification,[43,46] maximum entropy,[60,61] conditional random field,[62,63] support vector machines,[64,65]

decision trees[66] and a combination of different heuristics.[67]

Training corpora

The current version of the GENIA corpus consists of 2,000 manually annotated PubMed abstracts. While without doubt extremely useful for many ML approaches to TM tasks,[68] the manual building of semantically annotated resources is an expensive task.[46]

However, an ontology can be practically used to sense-tag raw text, ie to map a term occurrence to its sense (the concept designated by the given term). The relational information stored in the ontology can be used to automatically disambiguate terms that can be mapped to multiple concepts. For example, Liu *et al.*[14] used co-occurrence with related terms to resolve the meaning of an ambiguous term.

Ontology structure

Ontologies are typically organised in a hierarchy using the is-a relation between concepts. This property can be used to quantify the similarity between the concepts and, implicitly, the semantic similarity between the terms used to designate these concepts.[69] Such numerical information that can be inferred from an ontology, on top of the symbolic information it explicitly stores, is of particular value for TM applications. For example, semantic similarity measure can be used as a vehicle of ML approaches (instance-based approaches such as *k*-nearest neighbour and case-based reasoning)[70] to a variety of TM tasks (eg clustering[71] and classification[72] of both individual terms and the documents containing them).

Semantic similarity measures

A number of different approaches to inferring semantic similarity from an is-a hierarchy have been suggested. The tree similarity (ts) between two concepts, C_1 and C_2, is calculated according to the following formula:

$$\mathrm{ts}(C_1,\ C_2) = \frac{2 \cdot \mathrm{common}(C_1,\ C_2)}{\mathrm{depth}(C_1) + \mathrm{depth}(C_2)}$$

where common(C_1, C_2) denotes the number of common nodes in the paths between the root and the given concepts, and depth (C) is the number of nodes in the path connecting the root and the given concept C. This formula is a derivative of Dice's coefficient where ancestor concepts are treated as their features and the similarity corresponds to the ratio between the common and all features. It has been previously used to measure conceptual similarity in a hierarchically structured lexicon.[73] A 'probabilistic' variation of this model:[74]

$$\mathrm{ts}(C_1,\ C_2) = \frac{2 \cdot \log P[S(C_1,\ C_2)]}{\log P(C_1) + \log P(C_2)}$$

is obtained by 'normalising' Resnik's[75] variant of semantic similarity measure:

$$\mathrm{ts}(C_1,\ C_2) = -\log P[S(C_1,\ C_2)]$$

where $S(C_1, C_2)$ is the deepest common node that subsumes both of the given concepts, and $P(C)$ is an estimation of the probability of a textual realisation of the given concept C.

Term similarity measures need to be consistent in reflecting semantic similarity between the designated concepts, and an ontology can be used to assess such consistency. For example, Spasic *et al.*[76] used an ontology hand-crafted by a domain expert to automatically tune the parameters of a weighted corpus-based term similarity measure. The core similarity method is based on the lexical and contextual term similarities. In this approach, an ontology was used to provide the training values for the conceptual term similarity (calculated as Dice's tree similarity – see above), which should be approximated by the textual term similarity values. A consistent approximation of ontology-based similarity measure is important in biomedicine, because new concepts described in literature using new terms are not efficiently incorporated in an ontology.

In another approach, Spasic and Ananiadou[77] utilised UMLS to compare individual term occurrences in an edit distance (ED) approach to assessing their contextual similarity. Partial parsing was used to chunk the contextual information into major syntactic constituents, with

special consideration given to terms. The importance of terms as principal conveyors of domain-specific information was reflected in the high cost of deleting and inserting terms when aligning two contexts through ED. The cost of replacing (or matching) two terms in such an alignment depends on their semantic similarity, which is estimated via their tree similarity using their positions in the ontology (see above). Lexical similarity was used as an alternative for ontological tree similarity for terms not found in the ontology. In addition, the ontology was used to navigate through the conceptual space and efficiently select credibly similar contexts, ie the ones sharing semantically similar terms.[72]

CONCLUSIONS

Different layers of text annotation (lexical, syntactic and semantic) are required for sophisticated TM in biomedicine. High terminological variability, typical of the domain, emphasises the need for lexico-syntactic procedures and annotations that can be used to neutralise the effects of such variation. Such phenomena can be tackled effectively through the use of rule-based or machine learning techniques. However, traditional heuristic and ad hoc TM methods simply do not deliver in a complex sublanguage such as that of biomedicine. Encoding of the explicit semantic layer in biomedical text representation needs to be supported by ontologies as the formal means of representing domain-specific knowledge. Up until recently, most TM systems have not relied on ontologies or terminologies, which is the main reason why biomedical TM systems generally provide poorer results compared with other domains (eg newswire).

Therefore, ontologies together with terminological lexicons are prerequisites for advanced TM. It is not enough to rely on one or the other: both are needed if we wish to produce highly accurate results of the kind needed by biomedical experts and also to obtain broad coverage of biomedical text. TM applications should aim at deriving complex information from text, eg temporal, causal, conditional and other types of semantic relations between biomedical entities as opposed to simple associations. In order to achieve such objectives, biomedical text needs to be semantically annotated and actively linked to ontologies.

This leads us to the question of the types of ontologies needed for TM. As demonstrated by GENIES[54] and GenIE,[53] it is essential to focus on describing the syntactic and semantic behaviour of biomedical sublanguage and on the formal description of domain event concepts. These systems had to develop their own ontologies of events and their own terminological lexicons. Therefore, the challenge for the field is to develop appropriate ontology resources and link them to adequate terminological lexicons in order to support the kind of processing required – and also to support interoperability between such ontologies.

Terminology and ontology resources

This can be greatly facilitated by recent advances in reducing the cost of configuring and tuning systems based on biomedical sublanguage: lexical standards enabling reusability; ML techniques to discover patterns of sublanguage behaviour in large annotated text corpora to help grammar writers; development of ontologies that can act as domain models and major developments in extracting and characterising terminology, including compound terms and acronyms.

Acknowledgments

The National Centre for Text Mining (NaCTeM), UK is sponsored by the JISC. AK expresses his gratitude to Alexander von Humboldt Foundation.

References

1. Hearst, M. (1999), 'Untangling text data mining', in 'Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics' (ACL 1999), 20th–26 June, University of Maryland, pp. 3–10.
2. Stevens, R., Bechhofer, S. and Goble, C. (2000), 'Ontology-based knowledge representation for bioinformatics', *Brief. Bioinformatics*, Vol. 1(4), pp. 398–414.

3. Uschold, M., King, M., Moralee, S. and Zorgios, Y. (1998), 'The enterprise ontology', *Knowledge Eng. Rev.*, Vol. 13(1), pp. 31–89.
4. Koehler, J., Rawlings, C., Verrier, P. *et al.* (2005), 'Linking experimental results, biological networks and sequence analysis methods using ontologies and generalised data structures', *In Silico Biol.*, Vol. 5 (1), pp. 33–44.
5. URL: http://www.w3.org/RDF/
6. URL: http://www.w3.org/TR/owl-guide/
7. Ananiadou, S., Friedman, C. and Tsujii, J. (2004), *J. Biomed. Inform.*, Special issue: Named entity recognition in biomedicine, Vol. 37(6).
8. Sager, J. C. (1990), 'A Practical Course in Terminology Processing', John Benjamins, Amsterdam/Philadelphia.
9. Meyer, I., Eck, K. and Skuce, D. (1997), 'Systematic concept analysis within a knowledge-based approach to terminology', in 'Handbook of Terminology Management, Vol. 1: Basic Aspects of Terminology Management', John Benjamins, Amsterdam, pp. 98–118.
10. Bernardi, L., Ratsch, E., Kania, R. *et al.* (2002), 'Interdisciplinary work: The key to functional genomics', *IEEE Intell. Syst. Trends Controv.*, Vol. 17(3), pp. 66–68.
11. Grefenstette, G. (1994), 'Exploration in Automatic Thesaurus Discovery', Kluwer Academic Publishers, Boston, MA.
12. Harris, Z. (2002), 'The structure of science information', *J Biomed. Inform.*, Vol. 35(4), pp. 215–221.
13. Chang, J., Schutze, H. and Altman, R. (2002), 'Creating an online dictionary of abbreviations from Medline', *J. Amer. Med. Inform. Assoc.*, Vol. 9(6), pp. 612–620.
14. Liu, H., Johnson, S. and Friedman, C. (2002), 'Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS', *J. Amer. Med. Inform. Assc.*, Vol. 9(6), pp. 621–636.
15. Tsujii, J. and Ananiadou, S. (2005), 'Thesaurus or logical ontology, which one do we need for text mining?', *Language Res. Eval.*, in press.
16. Baeza-Yates, R. and Ribeiro-Neto, B. (1999), 'Modern Information Retrieval', ACM Press/ Addison-Wesley, New York.
17. Hobbs, J. (1993), 'The generic information extraction system', in 'Fifth Message Understanding Conference (MUC5)', Sundheim, B., Ed., Morgan Kaufmann Publishers, Inc., San Francisco, CA.
18. Fayyad, U. and Uthurusamy, R. (1996), 'Data mining and knowledge discovery in databases', *Commun. ACM*, Vol. 39 (11), pp. 24–26.
19. Swanson, D. and Smalheiser, N. (1994), 'Assessing a gap in the biomedical literature: Magnesium deficiency and neurologic disease', *Neuro-sci. Res. Commun.*, Vol. 15, pp. 1–9.
20. Webster, J. J. and Kit, C. (1992), 'Tokenization as the initial phase in NLP', in 'Proceedings of the 15th International Conference on Computational Linguistics', Nantes, France, pp. 1106–1110.
21. Brill, E. (1992), 'A simple rule-based part of speech tagger', in 'Proceedings of the 3rd Conference on Applied Natural Language Processing', Trento, Italy, pp. 152–155.
22. Hull, D. A. (1996), 'Stemming algorithms: A case study for detailed evaluation', *J. Amer. Soc. Information Sci.*, Vol. 47(1), pp. 70–84.
23. Jurafsky, D. and Martin, J. H. (2000), 'Speech and Language Processing: An introduction to natural language processing, computational linguistics and speech recognition', Prentice Hall, Upper Saddle River, NJ.
24. Allen, J. (1994), 'Natural Language Understanding', Addison Wesley, Reading, MA.
25. Srinivasan, P. (2001), 'MeshMap: A text mining tool for Medline', in 'Proceedings of the AMIA Symposium', AMIA, Bethesda, MD, pp. 642–646.
26. Perez-Iratxeta, C., Pérez, A., Bork, P. and Andrade, M. (2003), 'Update on XplorMed: A web server for exploring scientific literature', *Nucleic Acids Res.*, Vol. 31(13), pp. 3866–3868.
27. Fisk, J., Mutalik, P., Levin, F. *et al.* (2003), 'Integrating query of relational and textual data in clinical databases: A case study', *J. Amer. Med. Inform. Assoc.*, Vol. 10(1), pp. 21–38.
28. Becker, K., Hosack, D., Dennis Jr, G. *et al.* (2003), 'PubMatrix: A tool for multiplex literature mining', *BMC Bioinformatics*, Vol. 4, p. 61.
29. Ding, J., Viswanathan, K., Berleant, D. *et al.* (2005), 'Using the biological taxonomy to access biological literature with PathBinderH', *Bioinformatics*, Vol. 21(10), pp. 2560–2562.
30. Smith, B., Ceusters, W., Klagges, B. *et al.* (2005), 'Relations in biomedical ontologies', *Genome Biol.*, Vol. 6, p. R46.
31. Suarez, H., Hao, X. and Chang, I. (1997), 'Searching for information on the internet using the UMLS and medical world search', in 'Proceedings of the 1997 Annual AMIA Fall Symposium', Orlando, FL, Masys, D., Ed., pp. 824–828.
32. Mima, H., Ananiadou, S., Nenadic, G. and Tsujii, J. (2002), 'A methodology for terminology-based knowledge acquisition and integration', in 'Proceedings of COLING 2002', Tapai, Taiwan, pp. 667–673.
33. Müller, H., Kenny, E. and Sternberg, P. (2004), 'Textpresso: An ontology-based information retrieval and extraction system for biological literature', *PLoS Biol.*, Vol. 2(11), p. e309.
34. Fukuda, K., Tsunoda, T., Tamura, A. and

Takagi, T. (1998), 'Toward information extraction: Identifying protein names from biological papers', in 'Proceedings of the 3rd Pacific Symposium on Biocomputing', 4th–9th January, Hawaii, Altman, R. *et al.*, Eds, World Scientific Publishing Company, Singapore, pp. 705–716.

35. Thomas, J., Milward, D., Ouzounis, C. *et al.* (2000), 'Automatic extraction of protein interactions from scientific abstracts', in 'Proceedings of the 5th Pacific Symposium on Biocomputing', 4th–9th January, Hawaii, Altman, R. *et al.*, Eds, World Scientific Publishing Company, Singapore, pp. 538–549.
36. Humphreys, K., Demetriou, G. and Gaizauskas, R. (2000), 'Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures', in 'Proceedings of the 5th Pacific Symposium on Biocomputing', 5th–9th January , Hawaii, Altman, R. *et al.*, Eds, World Scientific Publishing Company, Singapore, pp. 505–516.
37. Nédellec, C. (2004), 'Machine learning for information extraction in genomics – state of the art and perspectives', in 'Text Mining and its Applications', Sirmakessis, S., Ed., Springer, Berlin, pp. 99–118.
38. Krauthammer, M. and Nenadic, G. (2004), 'Term identification in the biomedical literature', *J. Biomed. Inform.*, Vol. 37(6), pp. 512–526.
39. Nenadic, G., Spasic, I. and Ananiadou, S. (2004), 'Mining biomedical abstracts: What is in a term?', in 'Natural Language Processing – IJCNLP 2004', Su, K.-Y. *et al.*, Eds, LNAI 3248, Springer, Berlin, pp. 797–806.
40. McCray, A., Srinivasan, S. and Browne, A. (1994), 'Lexical methods for managing variation in biomedical terminologies', in '18th Annual Symposium on Computer Applications in Medical Care', Washington, DC, Ozbolt, J., Ed., pp. 235–239.
41. Jacquemin, C. (2001), 'Spotting and Discovering Terms Through Natural Language Processing', MIT Press, Cambridge, MA.
42. Chiang, J.-H. and Yu, H.-C. (2003), 'Meke: Discovering the functions of gene products from biomedical literature via sentence alignment', *Bioinformatics*, Vol. 19(11), pp. 1417–1422.
43. Tsuruoka, Y. and Tsujii, J. (2004), 'Improving the performance of dictionary-based approaches in protein name recognition', *J. Biomed. Inform.*, Vol. 37(6), pp. 461–470.
44. Tsuruoka, Y. and Tsujii, J. (2003), 'Probabilistic term variant generator for biomedical terms', in 'Proceedings of the 26th Annual International ACM SIGIR Conference', Toronto, Canada, pp. 167–173.
45. Tsuruoka, Y., Ananiadou, S. and Tsujii, J. (2005), 'A machine learning approach to acronym generation', in 'BioLINK SIG: Linking Literature, Information and Knowledge for Biology', ISMB, Detroit, MI.
46. Mukherjea, S., Subramaniam, L., Chanda, G. *et al.* (2004), 'Enhancing a biomedical information extraction system with dictionary mining and context disambiguation', *IBM J. Res. Develop.*, Vol. 48(5/6), pp. 693–701.
47. Leroy, G. and Chen, H. (2005), 'Genescene: An ontology-enhanced integration of linguistic and co-occurrence based relations in biomedical texts: Research articles', *J. Amer. Soc. Inf. Sci. Technol.*, Vol. 56(5), pp. 457–468.
48. McDonald, D., Chen, H., Su, H. and Marshall, B. (2004) 'Extracting gene pathway relations using a hybrid grammar: The Arizona relation parser', *Bioinformatics*, Vol. 20(18), pp. 3370–3378.
49. Kim, J. and Park, J. (2004), 'BioIE: Retargetable information extraction and ontological annotation of biological interactions from the literature', *J. Bioinform. Comput. Biol.*, Vol. 2(3), pp. 551–568.
50. Daraselia, N., Yuryev, A., Egorov, S. *et al.*(2004), 'Extracting human protein interactions from Medline using a full-sentence parser', *Bioinformatics*, Vol. 20(5), pp. 604–611.
51. Novichkova, S., Egorov, S. and Daraselia, N. (2003), 'Medscan, a natural language processing engine for medline abstracts', *Bioinformatics*, Vol. 19(13), pp. 1699–1706.
52. Gaizauskas, R., Demetriou, G., Artymiuk, P. and Willett, P. (2003), 'Protein structures and information extraction from biological texts: The pasta system', *Bioinformatics*, Vol. 19(1), pp. 135–143.
53. Cimiano, P., Reyle, U. and Saric, J. (2005), 'Ontology-driven discourse analysis for information extraction', *Data Knowledge Eng.*, in press.
54. Friedman, C., Kra, P., Yu, H. *et al.* (2001), 'GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles', *Bioinformatics*, Vol. 17, pp. S74–82.
55. Friedman, C., Kra, P. and Rzhetsky, A. (2002), 'Two biomedical sublanguages: A description based on the theories of Zellig Harris', *J. Biomed. Inform.*, Vol. 35(4), pp. 222–235.
56. Rzhetsky, A., Koike, T., Kalachikov, S. *et al.* (2000), 'A knowledge model for analysis and simulation of regulatory networks', *Bioinformatics*, Vol. 16(12), pp. 1120–1128.
57. Kim, J., Ohta, T., Tateisi, Y. and Tsujii, J. (2003), 'Genia corpus – semantically annotated corpus for bio-textmining', *Bioinformatics*, Vol. 19, pp. i180–182.

58. Collier, N., Nobata, C. and Tsujii, J. I. (2001), 'Automatic acquisition and classification of terminology using a tagged corpus in the molecular biology domain', *Terminology*, Vol. 7(2), pp. 239–257.

59. Zhang, J., Shen, D., Zhou, G. *et al.* (2004), 'Enhancing HMM-based biomedical named entity recognition by studying special phenomena', *J. Biomed. Inform.*, Vol. 37(6), pp. 411–422.

60. Dingare, S., Nissim, M., Finkel, J. *et al.* (2005), 'A system for identifying named entities in biomedical text: How results from two evaluations reflect on both the systems and the evaluations', *Comp. Functional Genomics*, Vol. 6(1–2), pp. 77–85.

61. Lin, Y.-F., Tsai, T.-H., Chou, W.-C. *et al.* (2004), 'A maximum entropy approach to biomedical named entity recognition', in 'Proceedings of the 4th ACM SIGKDD Workshop on Data Mining in Bioinformatics', Seattle, WA, pp. 56–61.

62. Settles, B. (2005), 'ABNER: An open source tool for automatically tagging genes, proteins, and other entity names in text', *Bioinformatics*, in press.

63. Song, Y., Kim, E., Lee, G.G. and Yi, B.-K. (2005), 'POSBIOTM-NER: A trainable biomedical named-entity recognition system', *Bioinformatics*, in press.

64. Kazama, J., Makino, T., Ohta, Y. and Tsujii, J. (2002), 'Tuning support vector machines for biomedical named entity recognition', in 'Proceedings of the ACL Workshop on Natural Language Processing in the Biomedical Domain', Philadelphia, PA, pp. 1–8.

65. Lee, K., Hwang, Y., Kim, S. and Rim, H. (2004), 'Biomedical named entity recognition using two-phase model based on SVMs', *J. Biomed. Inform.*, Vol. 37(6), pp. 436–447.

66. Nobata, C., Collier, N. and Tsujii, J. (2000), 'Automatic term identification and classification in biology texts', in 'Proceedings of the Natural Language Pacific Rim Symposium (NLPRS 2000)', Beijing, China, pp. 369–375.

67. Torii, M., Kamboj, S. and Vijay-Shanker, K. (2004), 'Using name-internal and contextual features to classify biological terms', *J. Biomed. Inform.*, Vol. 37(6), pp. 498–511.

68. Nobata, C., Collier, N. and Tsujii, J. (2000), 'Comparison between tagged corpora for the named entity task', in 'Proceedings of the Association for Computational Linguistics (ACL 2000) Workshop on Comparing Corpora', Hong Kong, Kilgariff, A. and Berber Sardinha, T., Eds., pp. 20–26.

69. Lord, P., Stevens, R., Brass, A. and Goble, C. (2003), 'Semantic similarity measures as tools for exploring the gene ontology,' in 'Proceedings of the 8th Pacific Symposium on Biocomputing', 3rd–7th January, Hawaii, Altman, R. *et al.*, Eds, World Scientific Publishing Company, Singapore, pp. 601–612.

70. Mitchell, T. (1997), 'Machine Learning', McGraw Hill, New York.

71. Nenadic, G., Spasic, I. and Ananiadou, S. (2004), 'Mining term similarities from corpora', *Terminology*, Vol. 10(1), pp. 55–80.

72. Spasic, I., Ananiadou, S. and Tsujii, J. (2005), 'MaSTerClass: A case-based reasoning system for the classification of biomedical terms', *Bioinformatics*, Vol. 21(11), pp. 2748–2758.

73. Wu, Z. and Palmer, M. S. (1994), 'Verb semantics and lexical selection', in 'Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL '94)', Maryland, USA, pp. 133–138.

74. Lin, D. (1998), 'An information-theoretic definition of similarity', in 'Proceedings of the 15th International Conference on Machine Learning', Morgan Kaufmann, San Francisco, CA, pp. 296–304.

75. Resnik, P. (1995), 'Using information content to evaluate semantic similarity in a taxonomy', in 'Fourteenth International Joint Conference on Artificial Intelligence', San Antonio, TX, pp. 448–453.

76. Spasic, I., Nenadic, G., Manios, K. and Ananiadou, S. (2002), 'Supervised learning of term similarities', in 'Intelligent Data Engineering and Automated Learning – IDEAL 2002', Yin, H. *et al.*, Eds, LNCS 2412, Springer, Berlin, pp. 429–434.

77. Spasic, I. and Ananiadou, S. (2005), 'A flexible measure of contextual similarity for biomedical terms', in 'Proceedings of the 10th Pacific Symposium on Biocomputing', 4th–8th January, Hawaii, Altman, R. *et al.*, Eds, World Scientific Publishing Company, Singapore, pp. 197–208.