Mining Term Similarities from Corpora*

Goran Nenadic	Irena Spasic	Sophia Ananiadou
Dept. of Computation UMIST, Manchester Manchester M60 1QD,UK	Computer Science University of Salford Salford M5 4WT, UK	Computer Science University of Salford Salford M5 4WT, UK
G.Nenadic@umist.ac.uk	I.Spasic@salford.ac.uk	S.Ananiadou@salford.ac.uk

Abstract

In this article we present an approach to the automatic discovery of term similarities, which may serve as a basis for a number of term-oriented knowledge mining tasks. The method for term comparison combines internal (lexical similarity) and two types of external criteria (syntactic and contextual similarities). Lexical similarity is based on sharing lexical constituents (i.e. term heads and modifiers). Syntactic similarity relies on a set of specific lexico-syntactic co-occurrence patterns indicating the parallel usage of terms (e.g. within an enumeration or within a term coordination/conjunction structure), while contextual similarity is based on the usage of terms in similar contexts. Such contexts are automatically identified by a pattern mining approach, and a procedure is proposed to assess their domain-specific and terminological relevance. Although automatically collected, these patterns are domain dependent and identify contexts in which terms are used. Different types of similarities are combined into a hybrid similarity measure, which can be tuned for a specific domain by learning optimal weights for individual similarities. The suggested similarity measure has been tested in the domain of biomedicine, and some experiments are presented.

Keywords: automatic terminology management, term similarity, contextual similarity, pattern mining, term clustering

* To appear in International Journal of Terminology, 2004

1. Introduction

The vast and constantly increasing amount of information in electronic format demands innovative techniques to gather and systematically structure knowledge, usually available only (or predominantly) from textual resources. Such resources still contain the most relevant and useful information, and are, so far, the principle knowledge sources for both researchers and knowledge mining systems. In order to extract knowledge represented in documents, main concepts, linguistically represented by terms in a given sublanguage (cf. (Sager 1990)), need to be identified and linked. The problem, however, is the huge volume of the textual resources, which are constantly expanding both in size and thematic coverage (Hirschman, Park, Tsujii, Wong and Wu 2002). In addition, dynamic development and new discoveries in many domains have resulted in the deluge of newly coined terms and relationships representing and linking newly identified or created concepts, events, associations, etc. These facts make the existing terminological resources rarely up-to-date. Since they need to adapt constantly to the advent of new knowledge, automatic terminology management tools are indispensable for dynamic management of terminological resources.

Traditional terminology management systems are maintained manually, even when the computational support is used to back the terminological work (Sager 1990). We define automatic terminology management (ATM) as a set of procedures used to support creation, storage, maintenance, update and curation of a terminology. An ATM system typically relies on three modules: an automatic term recognition (ATR) module, an automatic term structuring (ATS) module, and, an intelligent term manager (ITM). The ATR module is used to automatically recognise and extract lexical expressions used to label domain concepts in a free text, while the ATS module attempts to organise terms by discovering and establishing relationships among them. Finally, the ITM module stores terminological data in an appropriate format (e.g. a database) and links terminological entries to respective factual databases that provide additional information (e.g. definitions, references to the corresponding textual resources or factual databases).

ATR denotes a set of procedures that are used to systematically extract pertinent terms and their variants used in a collection of documents. The main aim of the ATR module is to highlight and extract lexical units related to relevant domain concepts, i.e. to extract sequences that have potential terminological relevance. However, the main problem in designing and implementing ATR procedures in the majority of domains is the lack of clear naming conventions in the form of stable term formation patterns and controlled lexical resources. Namely, there are no formal or reliable morpho-syntactic criteria that could be used to distinguish terms from non-terms based on their internal structure (e.g. terms typically have a structure of a noun phrase (NP), but not all NPs are necessarily terms). Additionally, there are no firm lexical criteria that can be used in this sense, as term components are not strictly confined to controlled vocabularies and “general” words frequently appear as term constituents.

ATR is not the ultimate aim from the terminology management point of view: terms should be also related to existing knowledge and/or to each other. Terminological structuring typically entails classification or clustering, which play a key role in organising knowledge in specialised scientific fields. Term constituents solely cannot be used as reliable criteria for terminology structuring, since they rarely systematically reflect functional properties or relatedness between entities. Naming conventions are typically defined by some guidelines, which may vary in sophistication depending on the subject field (Ananiadou 1994). For example, the Guidelines for Human Gene Nomenclature (Lander et al. 2001) include even principles such as starting a gene name with a lower case letter and avoiding molecular weight

designations. Still, these are only guidelines and as such do not impose restrictions to domain experts. In addition, they apply only to a subset of terms, while the rest of a terminology often remains highly non-standardised. For instance, ad-hoc or arbitrary names can be found frequently (such as a gene name "*Bride of sevenless*" or "*Boss*"). Such terms cannot be placed into an existing knowledge network without referring to features other than lexical.

Terms typically have a very large number of different features, but only a portion of these features are actually used when producing a structured model of a domain. Relevant features (or their subsets) are used to establish a notion of similarity among terms, which can be used as a basis for term structuring. The notion of term similarity has been defined and considered in different ways: terms may have functional, structural, causal, hoponymous, localisation or other similarities (Skuce and Meyer, 1991). The key problem is that genuine features (i.e. the ones that refer to the relevant properties of concepts denoted by terms) are typically unavailable, so other attributes (e.g. corpus-based features of term occurrences) need to be used. Therefore, the selection and discovery of relevant features and estimation of term similarities are the basic and most challenging problems to be solved (Blake and Pratt 2001).

In this article we suggest a domain-independent method for automatic mining of term similarities, which can serve as a basis for a number of term-oriented knowledge acquisition tasks. The method for term comparison combines internal evidence (lexical similarities) and two types of external criteria (syntactic and contextual similarities). While lexical and syntactic similarities rely on manually defined patterns, contextual similarity is based on the automatic discovery of significant term features through contextual pattern mining.

The article is organised as follows. In Section 2 we overview related terminology management approaches. Section 3 introduces the term similarity measure, and Section 4 presents experiments and discussion. Finally, Section 5 concludes the article.

2. Related work

Several text mining systems have been developed for the extraction of terminological knowledge from corpora. Numerous approaches to ATR have been proposed. Some methods (e.g. (Bourigault 1992; Ananiadou 1994)) rely purely on linguistic information, typically morpho-syntactic features of term candidates. They are frequently combined with statistical approaches (e.g. (Nakagawa and Mori 1998; Frantzi, Ananiadou and Mima 2000)). Further, machine-learning techniques have been used to acquire and disambiguate terms from specialised corpora (Hatzivassiloglou, Duboue and Rzetsky 2001; Kazama, Makino, Ohta and Tsujii 2002).

Apart from distinguishing terms from non-terms, an additional problem for ATR is dealing with terminological variation. In theory, terms should be mono-referential (one-to-one correspondence between terms and concepts), but in practice we deal with ambiguities (the same term corresponding to many concepts) and variants (many terms leading to the same concept). If aiming at systematic acquisition and structuring of domain-specific knowledge, then handling term variation needs to be treated as an essential part of terminology mining. Few methods for term variation handling have been developed. For example, in the FASTR system (Jacquemin 2001) morphological and syntactic variations are handled by means of lexicalised meta-rules used to describe term normalisation, while semantic variants are handled via a specialised WordNet. Similarly, the C/NC-value method (Frantzi, Ananiadou and Mima 2000) for term extraction has been extended to handle orthographic, morphological and structural term variants, as well as acronyms (Nenadic, Spasic and Ananiadou 2002a).

Several methods have been developed to automatically structure terminological knowledge. For example, Bourigault and Jacquemin (1999) used lexical similarities to cluster

terms. Their idea is based on adapting the term normalisation process proposed within the FASTR framework (Jacquemin 2001). A cluster is produced by linking terms that are associated by specific syntactic variation links (namely lexical characteristics and possible term-formation decompositions), which reflect the internal term structures. Although undoubtedly useful, this information is typically insufficient to discover similarities among many terms. Even the authors themselves suggested that only 10% of multi-word terms automatically extracted from documents benefited from variation-based links produced by FASTR. Still, the proposed approach is useful for relating terms within systematically structured terminologies, but it would be rather limited when dealing with ad-hoc names, as well as with single-word terms, as such terms cannot be compared to other terms without using additional features.

Additional features are typically extracted from a domain-specific corpus by exploring different terminological relationships. Such relations can be expressed via a variety of surface lexical and syntactic realisations. Approaches based on shallow-parsing range from lexical pattern matching (e.g. (Hearst 1992)), via template-based approaches (e.g. (Maynard and Ananiadou 1999)), to full parsing of documents using domain-specific grammars (e.g. (Yakushiji, Tateisi, Miyao and Tsujii 2001)), and they typically extract specific "named" relations. There has been much debate as what types of patterns are the most reliable for the extraction of term similarities (cf. (Maynard and Ananiadou 1999)). Lexical patterns are in particular effective for the extraction of basic conceptual relationships (such as hyponymy), while semantic-frame based approaches can, on the other hand, collect very precise and reliable task-specific information. Beside manually engineered patterns, several approaches for automatic learning of contextual patterns for general conceptual relations have been proposed (cf. (Hearst 1992; Riloff 1996; Finkelstein-Landau and Morin 1999; Thelen and Riloff 2002)). However, these approaches are basically oriented towards information extraction (IE) tasks, as they are based on predefined types of relationships. Similarly, machine-learning approaches have been used to learn lexical contexts expressing a given relationship. Many reports (e.g. (Craven and Kumlien 1999; Marcotte, Xenarios and Eisenberg 2001)) suggested that such approaches proved to be reliable for retrieving pre-defined, domain-specific relations.

On the other hand, various statistical methods (such as co-occurrence frequency counts) have also been used to link terms. For example, Maynard and Ananiadou (2000b) and Mima, Ananiadou and Nenadic (2001) analysed terms co-occurring in a close proximity to one another as a basis for estimating similarities. However, term co-occurrences and statistical distributions over larger text units (e.g. documents) may not reveal significant associations for some types of relationships (cf. (Hindle 1990; Ding, Berleant, Nettleton and Wurtele 2002)). Therefore, statistical and shallow-parsing methods have been combined. For example, Hindle (1990) suggested a similarity measure among nouns based on mutual information of subject-verb and verb-object co-occurrences. His main assumption is that a noun appears as subject or object of a restricted set of verbs, and that, consequently, each noun can be characterised by the verbs it co-occurs with. Grefenstette (1994) extended this approach by considering other grammatical roles.

In the following Section we suggest a hybrid approach that combines pattern-based and machine-leaning techniques with a statistical scoring mechanism to mine similarities among terms, which may indicate different types of links among them. We also combine similarities based on internal lexical correspondences and different types of corpus-based distributions.

3. Mining term similarities

Our approach to discovering term similarities incorporates three aspects of term similarity, namely lexical, syntactic and contextual similarity. These similarities are linearly combined in order to estimate similarity among terms. In the following subsections we describe each of the three similarity measures, and the process of the supervised optimisation of their combination.

3.1 Mining lexical similarities

The most straightforward approach to measuring term similarities is to measure lexical similarity among the words that constitute terms. This idea was exploited by Bourigault and Jacquemin (1999) by adapting the term variation conflation process, and by Dagan and Church (1994) via "grouping" the list of term candidates according to their heads. We, however, generalise these approaches by considering constituents (head and modifiers) shared by terms. The rationale behind lexical similarity involves the following hypotheses: (1) Terms sharing a head are assumed to be (in)direct hyponyms of the same term (e.g. *progesterone receptor* and *oestrogen receptor* are both *receptors*). (2) A term derived by modifying another term may indicate concept specialisation (e.g. *orphan nuclear receptor* is a kind of *nuclear receptor*). More generally, when a term is nested inside another term, we assume that the terms in question are related (e.g. *retinoic acid* and *retinoic acid receptor* are associated). In order to neutralise inflectional and simple structural variations, we compare only *normalised* terms (i.e. singular terms containing no prepositions; terms containing prepositions are transformed into the corresponding forms without them (Nenadic, Spasic and Ananiadou 2002a)).

Let us now describe the calculation of lexical similarity between terms, which is based on common subsequences shared by the terms. By comparing all non-empty sub-sequences and not only single modifiers, we want to give more credit to pairs of terms that share longer nested constituents, with an additional weight given to the similarity if the two terms have common heads. Given a sequence of words *s*, we will use P(*s*) to refer to a set of all non-empty sub-sequences in *s*. For example, P(*orphan nuclear receptor*) = {*orphan*, *nuclear*, *receptor*, *orphan nuclear*, *nuclear receptor*, *orphan nuclear receptor*}. Formally, lexical similarity between terms t_1 and t_2 (whose heads are denoted by h_1 and h_2 respectively) is calculated according to a Dice-like coefficient formula:

$$LS(t_1, t_2) = \frac{|\,\mathrm{P}(h_1) \cap \mathrm{P}(h_2)\,|}{|\,\mathrm{P}(h_1)\,| + |\,\mathrm{P}(h_2)\,|} + \frac{|\,\mathrm{P}(t_1) \cap \mathrm{P}(t_2)\,|}{|\,\mathrm{P}(t_1)\,| + |\,\mathrm{P}(t_2)\,|} \quad (1)$$

The numerators in the formula (1) denote the number of shared constituents, while the denominators refer to the sums of total numbers of constituents. Table 1 gives some examples.

Table 1. Examples of lexical similarities

i	t_i	$P(t_i)$
1	*nuclear receptor*	{*nuclear, receptor, nuclear receptor*}
2	*orphan receptor*	{*orphan, receptor, orphan receptor*}
3	*orphan nuclear receptor*	{*orphan, nuclear, receptor, orphan nuclear, nuclear receptor, orphan nuclear receptor*}
4	*nuclear orphan receptor*	{*nuclear, orphan, receptor, nuclear orphan, orphan receptor, nuclear orphan receptor*}

$LS(t_1, t_2) = 0.67$ $LS(t_1, t_3) = 0.83$ $LS(t_1, t_4) = 0.72$ $LS(t_2, t_3) = 0.72$ $LS(t_3, t_4) = 0.75$

Lexical similarity is useful for comparing multi-word terms, but it is rather limited when it comes to ad-hoc names (since they can have arbitrary constituents) or single-word terms. Note that the lexical similarity between two different terms can have a positive value only if at least one of them is a multi-word term. Still, in some cases, terms are frequently represented by single words (e.g. standardised protein/gene names), so alternative methods need to be used to compare them lexically (e.g. approximate string matching). More importantly, lexical similarity can capture only restricted types of similarities (typically hyponymy and meronymy). Therefore, external similarities are needed in addition to lexical term similarity.

3.2 Mining syntactic similarities

It has been previously reported (e.g. (Hearst 1992)) that some general (i.e. domain independent) lexico-syntactic patterns may indicate functional similarity among terms. For instance, the following sequence:

> ... *steroid receptors* such as *estrogen receptor*, *glucocorticoid receptor*, and *progesterone receptor* ...

suggests that all terms involved in this excerpt are highly correlated, since they appear in an enumeration (represented by the *such-as* pattern), which indicates their similarity. Similar patterns have been previously used to infer hyponym relations among words (Hearst 1992). We generalise this approach by taking into account patterns in which terms are used *concurrently* within the same context. In our approach, two types of "parallel" lexico-syntactic patterns are considered: term enumeration expressions and term coordination/conjunctions. We hypothesise that a specific type of co-occurrence of terms in the parallel patterns (i.e. within identical contexts) shows their functional similarity. More precisely, all terms within a parallel structure perform the same *syntactic* function within a sentence (e.g. an object or a subject) and are used in combination with the same verb or preposition. This fact indicates their semantic similarity. For example, the parallel usage of terms *estrogen receptor* and *progesterone receptor* in the following sentence indicates their similarity with respect to the transactivation process:

> *"Transactivation by* **either** estrogen receptor **or** progesterone receptor *involves a conserved AF-2 domain."*

Manually defined enumeration patterns (see Table 2) are applied as syntactic filters in order to retrieve sets of similar terms. In addition, we used term conjunction and term coordination patterns (Klavans, Tzoukermann and Jacquemin 1997) as another type of parallel syntactic structure. Two types of argument coordination and two types of head coordination patterns are considered (see Table 3). However, these patterns are ambiguous, as they may retrieve both coordinated terms and conjunctions of terms (see Table 4). In either case, the retrieved terms are associated. The retrieval of terms from a coordinated structure requires transformation of coordination constituents, while this is not needed in case of a simple conjunction. In an argument coordination (where term arguments are coordinated), the coordinated terms could be retrieved by "multiplying" the arguments with the shared head. On the other hand, in a head coordination (where term heads are coordinated), the coordinated terms could be retrieved by "multiplying" the heads with the shared arguments. We supported the two coordination types by the LR(1) grammar rules (Mima, Ando and Aoe 1995), which extracted terms from coordinated patterns. In order to differentiate between coordination structures and nominal conjunctions, we employed a simple heuristic approach, where the "multiplied" candidates were accepted as terms if they occurred independently elsewhere in the corpus. In such case, it was hypothesised that the coordination structure was a correct one. Otherwise, we considered that that the nominal conjunction was the case, and no transformation was performed on the involved terms.

Table 2. Example of term enumeration lexico-syntactic patterns[1]

<TERM>([**(**](**such as**)|**like** | (**e.g.**[,])) <TERM> (**,**<TERM>)* [[**,**] <**&**> <TERM>] [**)**]
<TERM> (**,**<TERM>)* [,] <**&**> **other** <TERM>
<TERM> [**,**] (**including** | **especially**) <TERM> (**,**<TERM>)* [[**,**] <**&**><TERM>]
both <TERM> **and** <TERM>
either <TERM> **or** <TERM>
neither <TERM> **nor** <TERM>

Table 3: Example of term coordination patterns[1]

<N>|<Adj>) (**,**(<N>|<Adj>))* [**,**] <**&**> (<N>|<Adj>) <TERM>
(<N>|<Adj>)***/***(<N>|<Adj>) <TERM>
(<N>|<Adj>) <TERM> (,<TERM>)* [,] <**&**> <TERM>
(<N>|<Adj>) <TERM>***/***<TERM>

Table 4. Example of ambiguities of coordinated structures

head coordination	[*adrenal* [*glands and gonads*]] → *adrenal glands, adrenal gonads*
term conjunction	[*adrenal glands*] *and* [*gonads*] → *adrenal glands, gonads*

When calculating the syntactic similarity, we do not discriminate among different syntactic relationships among terms (represented by different patterns), but instead, we consider terms appearing in the same syntactic roles as highly semantically correlated. Based on co-occurrence of terms in these parallel lexico-syntactic patterns, we define the syntactic similarity (*SS*) measure for a pair of terms as 1 if the two terms appear together in any of the patterns, and 0 otherwise.

The parallel lexico-syntactic patterns provide a term similarity measure with high precision, but low recall, as terms do not frequently appear in parallel patterns relative to the number of term occurrences (in particular for smaller corpora). For this reason, we need other important contextual patterns in which terms tend to appear in order to compare them.

3.3 Mining contextual similarities

Determining the similarity of terms based on their contexts is a standard approach based on the hypothesis that similar terms tend to appear in similar contexts (Maynard and Ananiadou 2000a). Contextual similarity, however, may be determined in a variety of ways depending on the definition of context.

Our approach to contextual similarity is mainly based on the Harris' notion of substitutability: if two terms can substitute each other in many *similar* contexts, then they can be deemed similar. In our approach, the main hypothesis is that if two terms appear in a number of similar, *domain important* contexts, then they can be deemed similar. Take, as an example, the term "*ligand-inducible transcription factor*" (or LITF, for short). By exploring its occurrences (see Table 5), we can see that this term typically appears in a context that can be described as "*belonging to a superfamily of*". This description follows a certain contextual pattern:

<TERM(s)> (*belong* to | *be member*(*s*) *of*) *superfamily of* LITF

Furthermore, *nuclear receptor* is frequently used to modify *superfamily*, while typically less significant and less content-bearing words (e.g. *novel member*, or *large superfamily*, or *structurally related* LITF) can be inserted in the pattern, without significantly affecting its structure or meaning. Similarly, terms *ligand-dependent transcription factors* and *ligand-activated transcription factor* appear in similar contexts, following the same pattern (see Table 6). Also, other terms (such as *nuclear receptors, nuclear hormone receptors, nuclear steroid hormone receptors, hormone-dependent transcription factors*) can be found in the similar patterns, and all these terms are mutually associated. Based on this, we hypothesise that such context patterns can be used to establish term similarities.

Table 5. Sample contexts of the term "*ligand-inducible transcription factor*"

... T3R belongs to the nuclear receptor superfamily of ligand-inducible transcription factors...
... The retinoid receptors belong to a large superfamily of ligand-inducible transcription factors ...
... VDR, which belongs to the nuclear receptor superfamily of ligand-inducible transcription factors...
... ER, a member of a large superfamily of nuclear receptors, is a ligand-inducible transcription factor...
... This receptor is a novel member of the superfamily of ligand-inducible transcription factors, ...
... RXR, a member of the superfamily of nuclear receptors, is a ligand-inducible transcription factor...

Table 6. Sample contexts of terms similar to the term "*ligand-inducible transcription factor*"

... an unique nuclear receptor belonging to the superfamily of ligand-dependent transcription factors ..
... a family of ligand-dependent transcription factors ...
... TRs and steroid hormone receptors belong to a large superfamily of nuclear hormone receptors ...
... a novel orphan receptor in the nuclear receptor superfamily of ligand-activated transcription factors ...
... PPARs are members of the steroid/thyroid nuclear receptor superfamily of ligand-activated transcription factors....

However, there are two problems. Firstly, there is a variety of similar patterns, whose lexical variability (including the variability in length) needs to be neutralised if we intend to use them a basis for term comparison. Secondly, the problem is to distinguish terminologically important contexts, as terms may also appear in contexts that are not relevant for establishing their similarities. Consider, for example, frequently used, but non-informative and non-discriminative pattern *<TERM> has been recently reported to ...*. In order to resolve these problems, several authors restricted contexts to either "bag-of-specific-entities" approach (e.g. Grefenstette (1994) considered only lexicalised subject/object and modifier relations), or a pre-defined set of patterns was used (e.g. Maynard and Ananiadou (1999) used a set of clustered, pre-defined semantic frames that were deemed domain relevant). Our idea is to observe substitution restrictions and substitution relations in a corpus, and to use automatically extracted, terminologically relevant contextual patterns as features for mining similarities among terms.

More precisely, our approach to contextual similarity is based on automatic *pattern mining*. The aim is to automatically identify, normalise and harvest the most important context patterns in which terms appear. *Context pattern* (CP) is defined as a generalised regular expression that corresponds to either left or right context of a term, which are treated separately. The following example shows two left context patterns of the term *ligand-inducible transcription factor*:

(2)

V:*belong*	PREP:*to*	TERM:*nuclear_receptor*	NP:*superfamily*	PREP:*of*
V:*belong*	PREP	TERM:*nuclear_receptor*	NP	PREP

Different types of context constituents and different levels of generalisation can be considered as important for characterising terms. The main challenge is to select information relevant for measuring similarity that will include the maximal possible generalisation with the minimal loss of "term identity". For example, in (2), the second pattern is more general, but may be discriminative enough to correctly link similar terms.

We consider two types of constituents: morpho-syntactic (such as noun and verb phrases, prepositions, etc.) and terminological (i.e. term occurrences). Morpho-syntactic constituents can be identified by applying a tagger and appropriate local grammars (which recognise chunks, such as NPs, VPs), while terminological entities can be recognised either by an ATR processor or by a controlled vocabulary. In the simplest case, contexts are mapped into the syntactic categories of their constituents. However, lemmatised forms for each of the syntactic categories can be used as well to instantiate the constituents in question. For example, the context "*belongs to the nuclear receptor superfamily of*" of the term "*ligand-inducible transcription factor"* can be mapped into any of the following CPs:

V PREP TERM NP PREP (non-instantiated pattern)
V PREP TERM:*nuclear_receptor* NP PREP:*of* (partially instantiated pattern)
V:*belong* PREP:to TERM:*nuclear_receptor* NP:*superfamily* PREP:*of* (instantiated pattern)

Some of the syntactic categories can be removed from the context patterns, as not all of them are equally significant in providing useful contextual information (Maynard and Ananiadou 2000a). For example, adjectives (that are not part of terms), adverbs and determiners can be removed from context patterns as they rarely bare some specific information. In addition, so-called linking words (e.g. *however, moreover,* etc.), or, more generally, linking devices (e.g. verb phrases such as *result in, lead to, entail*, etc.) are frequently used in special languages in order to achieve more effective communication (Sager, Dungworth and McDonald 1980). However, these constituents are typically non informative and can be eliminated. At the same time, the important constituents can further be instantiated, in order to specify their semantic content. CPs that have certain types of constituents instantiated and some constituent types discarded, will be called *canonical* CPs.

In the experiments reported in this article, we instantiated terms and either verbs or prepositions, as these categories were regarded as significant for term comparison. As indicated in many studies, terms are the most informative entities in documents, and characterise them semantically. Also, similar terms typically co-occur in near proximity with their "friends" (Maynard and Ananiadou 2000a), so it is justifiable to instantiate them as they seem to be good indicators of similarity among terms. Further, verbs proved to be useful for characterising NPs appearing as subjects and objects (cf. (Hindle 1990; Grefenstette, 1994)), as well as anchors for many IE tasks (Riloff 1996). Also verbs (and their complementation patterns) have been used to guide the term classification tasks (Spasic, Nenadic and Ananiadou 2003b). Finally, as prepositions may denote some relationships (e.g. the preposition *with* may denote the meronymy (“has part”) relation, while *in* may denote localisation), they are beneficial for indicating term similarities.

Finally, in order to address the problem of variable pattern lengths, we have decided to generate all possible “linearly nested” patterns for each given context. Precisely, when considering left contexts, contexts of the maximal length (without crossing the sentence boundary) are initially selected, and they are then iteratively trimmed on the left side until the minimal length is reached. Right contexts are treated analogously. Maximal and minimal lengths are chosen empirically: in the experiments reported in this article, we have set the minimal pattern length to 2, and the maximal length to 10. The following example illustrates the left linear pattern generation process:

V PREP TERM NP PREP (the maximal pattern)
PREP TERM NP PREP
TERM NP PREP
NP PREP (the minimal pattern)

Although nested CPs may seem to be redundant in a given context, they may be relevant for comparison with other (shorter) contexts. However, we will assume that longer contexts are more important for assessing term similarities.

Let us now describe the process of constructing CPs and determining their importance. First, we collect concordances for all terms for which term similarities are to be analysed. For each term occurrence, the maximal left and right canonical CPs are extracted, and nested CPs are generated. Once we have canonical CPs, we calculate the values of a measure called CP-value in order to estimate the importance of the CPs. CP-value is inspired by the C-value

measure for assigning termhoods to term candidates (Frantzi, Ananiadou and Mima 2000), and by the cost criteria introduced in (Kita, Kato, Omoto and Yano 1994).

The CP-value measure assigns importance weights as follows: the weights for CPs that do not appear as nested elsewhere (e.g. some of the maximal length CPs) are proportional to their frequency and length. As indicated above, we assume that longer CPs should be given more credit, as it is less probable that a longer CP (that is terminologically important for a domain) will appear with the same frequency as shorter CPs. Thus, if a longer CP appears as frequently as a shorter one, we assign a higher value to the longer CP, as we believe that the fact that the longer CP appears frequently is more significant (this effect is moderated by the application of the logarithm function, see later).

If a CP appears as nested, we take into account both the number of times it appears as maximal (positive impact – CPs appearing frequently as maximal CPs should be given more credit), and the number of times it appears as nested (small negative impact). Therefore, the absolute frequency of the nested CP is reduced by its frequency as nested, resulting in its independent frequency of occurrence. Further, since a CP is more relevant if it appears as nested in fewer CPs, we want to "normalise" the frequency of "nested occurrence" by dividing it by the number of other CPs that contain the CP in question. More precisely, CP-value of a pattern p is defined as

$$\text{(3)} \qquad \text{CP}-\text{value}(p) = \begin{cases} \log_2 |p| \cdot f(p), & p \text{ is a maximal, not - nested CP} \\ \log_2 |p| \cdot \left(f(p) - \frac{1}{|T_p|} \sum_{q \in T_p} f(q) \right), & p \text{ is a lineary nested CP} \end{cases}$$

where $f(p)$ is the absolute frequency of p, $|p|$ is its length (as the number of constituents), T_p is a set of all CPs that contain p, and consequently $|T_p|$ is the frequency of its occurrence within other CPs.

As indicated earlier, left and right CPs are treated separately. Table 7 shows examples of left context patterns extracted from a collection of Medline abstracts (Medline 2002). The CPs whose CP-values are within the threshold boundaries can be deemed important: CPs with very high CP-values are typically general patterns, while CPs with low CP-values may be irrelevant (they typically have low frequency). Note, however, that middle-ranked patterns are domain-specific and that they are automatically extracted from a corpus.

Table 7. Example of left CPs (terms and most frequent verbs are instantiated)

Contextual pattern	**CP-value**
PREP NP	272.65
PREP NP PREP	186.47
. . .	. . .
PREP NP V:*stimulate*	9.32
V:*indicate* NP	5.00
PREP NP PREP V:*involve* NP	4.64
PREP TERM:*transcriptional_activity*	4.47
V:*require* NP PREP	4.38
PREP TERM:*nuclear_receptor* PREP	4.00

At this point, each term is associated with a set of the most characteristic left and right patterns in which it occurs. As we treat CPs as term features, we have used a Dice-like coefficient to estimate contextual similarity between terms as a function of both common and

distinctive features. Let C_{L1}, C_{R1}, C_{L2} and C_{R2} be sets of left and right CPs associated with terms t_1 and t_2 respectively. The contextual similarity (*CS*) between t_1 and t_2 corresponds to the ratio between the number of common and all significant CPs they appear in:

$$CS(t_1,t_2) = 2 \cdot \frac{|C_{L1} \cap C_{L2}| + |C_{R1} \cap C_{R2}|}{|C_{L1}| + |C_{L2}| + |C_{R1}| + |C_{R2}|} \quad (4)$$

The multiplication parameter in the formula (4) is used to normalise the value of the *CS*, so that it has the value of 1 when two identical terms are compared.

3.4 Combining similarities

None of the similarities introduced so far is sufficient on its own to define term similarity measure between two arbitrary terms (see also Section 4). For example, our experiments have shown that syntactic similarity provides high precision, but extremely low recall (less than 1%) when used on its own, as not all terms appear in parallel lexico-syntactic expressions. Furthermore, if a term appears infrequently or within very specific CPs, the number of its significant CPs will influence its contextual similarity to other terms. On the other hand, there are concepts that have idiosyncratic names, which thus cannot be classified relying exclusively on lexical similarity.

In order to make use of all possible information, we introduce a hybrid term similarity measure (called the *CLS* similarity) as a linear combination of the three similarity measures:

$$CLS(t_1, t_2) = \alpha\, CS(t_1,t_2) + \beta\, LS(t_1,t_2) + \gamma\, SS(t_1,t_2)$$

where $\alpha + \beta + \gamma = 1$. Since *CS*, *LS* and *SS* are similarity measures (they are reflexive and symmetric), their linear combination also has these properties. Still, the choice of the weights α, β and γ in the previous formula is not a trivial problem. In our preliminary experiments we used manually chosen values, but, then, an automatic learning method was used to suggest the optimal weights. The learning method uses an existing ontology to provide a training set of terms. Ontology-based term similarities (used as a "gold" training standard) are calculated using both the vertical position of terms and their horizontal distance in the ontology. Namely, we use a *commonality measure* as the number of shared ancestors between two terms in the ontology, and a *positional measure* as a sum of their tree depths, i.e. distances from the root (Maynard and Ananiadou 2000b). Similarity between two terms then corresponds to the ratio between commonality (common) and positional (depth) measures:

$$O(t_1,t_2) = \frac{2 \cdot \text{common}(t_1,t_2)}{\text{depth}(t_1) + \text{depth}(t_2)} \quad (5)$$

In order to determine an optimal solution for the weights, we used a genetic algorithm approach. Genetic algorithms (GAs) are meta-heuristics incorporating the principles of natural evolution and the idea of "survival of the fittest" (Reeves 1996). A solution is encoded as a sequence of "genes", referred to as an individual. In our case, an individual is represented as a triple[2] (α, β, γ), where $\alpha + \beta + \gamma = 1$. In the initial phase of the GA we generate in a random manner a number of triples (α, β, γ) such that α, β, $\gamma > 0$ and $\alpha + \beta + \gamma = 1$.

Operators typical of GAs, namely selection, crossover, mutation, and replacement, are applied, in that order, in each iteration of the GA. We use the tournament selection, a

technique where a group of individuals is singled out randomly, and after that the fittest ones are selected for crossover. A uniform crossover is used, so that each gene position is chosen with 50% probability for the genes at that position to be swapped. Finally, the mutation operator introduces diversity into a population by modifying a small portion of newly formed solutions. Random triples (α, β, γ) are changed in the following manner: one of α, β, γ is randomly chosen and its value is changed randomly, and other values are adjusted so that $\alpha + \beta + \gamma = 1$.

Once a sufficient number of new solutions have been created by applying the three GA operators, they are evaluated according to a predefined quality criterion, called *fitness*. As we want to minimise the deviation of the *CLS* similarity values from the similarity values derived from the ontology, we estimate the fitness of a triple (α, β, γ) through the Euclidean distance as follows:

$$f(\alpha,\beta,\gamma) = \sum_{t_1,t_2 \in T} \left(CLS_{\alpha\beta\gamma}(t_1,t_2) - O(t_1,t_2) \right)^2 \quad (6)$$

In formula (6), T is a set of training terms appearing in the training corpus, $CLS_{\alpha\beta\gamma}(t_1, t_2)$ is the *CLS* similarity measure calculated for the given weights (α, β, γ), and $O(t_1, t_2)$ is the similarity measure derived from the ontology by formula (5). The goal is thus to find a triple that minimises the value of the fitness function.

Once all the new individuals have been evaluated, the fittest ones replace the appropriate number of the less fit old solutions, thus forming a new population. This process is repeated until a stopping condition is fulfilled. The stopping condition is satisfied if the current generation contains an individual for which the value of the evaluation function is smaller than a given threshold, or if a certain number of iterations have been performed.

4. Experiments and discussions

We have performed a series of experiments with mining term similarities from biomedical corpora. We have firstly experimented with manually extracted and marked terms from the Genia biomedical corpus (Ohta, Tateisi, Kim, Mima and Tsujii 2002), which contains 2,000 abstracts and around 30,000 terms. The experiments have shown that *LS* is highly accurate in predicting associations between terms that have high values for *LS*, but that it has typically revealed only the hyponymy relationships. Although the majority of terms have at least one lexically similar term, only 2-3% of term pairs can be compared lexically, with only 5% of the semantically closest terms having positive *LS*.

Syntactic, co-occurrence based similarity has been even more sparse: less than 1% of the semantically closest terms have appeared in parallel patterns. Also, term coordination expressions are infrequent (in the Genia corpus, around 2% of term occurrences). Furthermore, just one third of terms appearing in coordination expressions could be found elsewhere in corpora, while the precision of the proposed approach for differentiating between term coordination structures and conjunctions was approximately 70%. On the other hand, coordinations typically assume that coordinated terms share constituents (either arguments or/and heads), and, thus, terms involved in a coordination expression can be typically compared lexically with high values for the respective *LS*. For example, the average lexical similarity among terms that co-occurred in a coordination expression in the Genia corpus was 2.4 times the average lexical similarity for all terms.

While lexical and syntactic similarities have low coverage, the CS similarity provides a similarity measure that covers the majority of semantically linked term pairs. The experiments have shown that its recall is significantly higher than recall of other two measures at all precision points. The maximal recall for the Genia terms was above 80% at 60% precision.

In the second set of experiments, the hybrid *CLS* measure has been tested with multi-word terms automatically extracted by the C/NC-value method (Frantzi, Ananiadou and Mima 2000) from a corpus of 2,082 abstracts retrieved from the Medline database (Medline 2002). The corpus has been tagged by the EngCG shallow parser (Voutilainen and Heikkila 1993) coupled with a set of simple local grammars for the NP/VP chunking. The first experiments were performed with manually chosen values 0.3, 0.3 and 0.4 for α, β, and γ respectively. For the CP mining task, the minimal and maximal pattern lengths have been set to 2 and 10 respectively, while the interval for relevant CPs has been chosen as follows: 5% of the top ranked patterns were discarded as general, while the lower CP-value threshold has been set empirically at 2.0.

Random samples of results have been evaluated by a domain expert. Table 8 shows similarity of the term *retinoic acid receptor* to a number of terms. The examples point out the importance of combining different types of term similarities. For instance, the low value of contextual similarity for *retinoic X receptor* (caused by relatively low frequency of its occurrence in the corpus) is balanced out by the other two similarity values, thus correctly indicating it as a term similar to the term *retinoic acid receptor*. On the other hand, the high value of contextual similarity for *signal transduction pathway* is neutralised by the other two similarity values, hence preventing it as being labelled as highly similar to *retinoic acid receptor*.

Table 8. Example of similarity values between *retinoic acid receptor* and other terms

Term	***LS***	***SS***	***CS***	CLS_{manual}
nuclear receptor	0.61	1.00	0.58	0.76
retinoic X receptor	0.67	1.00	0.32	0.70
progesteron receptor	0.61	0.00	0.35	0.29
signal transduction pathway	0.00	0.00	0.75	0.23
retinoic acid	0.33	0.00	0.31	0.20
receptor complex	0.11	0.00	0.52	0.19

The *CLS* measure proved to be consistent in the sense that similar terms shared the same "friends". For example, the similarity values of two highly associated terms *glucocorticoid receptor* and *estrogen receptor* (the value of their similarity is 0.68) with respect to other terms are mainly approximate (see Table 9).

We have also used automatically tuned parameters for the calculation of *CLS* similarities. A simplified ontology (produced by a domain expert) was used as a source for establishing term similarity weights. The supervised learning of parameters (described in Section 3.4) resulted in the values 0.13, 0.81 and 0.06 for α, β, and γ respectively (cf. (Spasic, Nenadic, Manios and Ananiadou 2002)). Note that lexical similarity appeared to be the most important, and syntactic similarity to be insignificant. We believe that there are several reasons for that. First, the ontology used as a seed for learning term similarity weights contained well-structured and standardised terms, which resulted in the lexical similarity being promoted as the most significant. On the other hand, the syntactic similarity is corpus-

dependent: the size of the corpus and the frequency with which the concurrent lexico-syntactic patterns are realised in it, affect its relevance. In the training corpus such patterns occurred infrequently relative to the number of terms.

Table 9. Example of similarity values for *glucocorticoid receptor* and *estrogen receptor* and other terms

Term	*glucocorticoid receptor*	*estrogen receptor*
steroid receptor	0.66	0.64
progesterone receptor	0.55	0.59
human estrogen receptor	0.28	0.37
retinoid x receptor	0.27	0.36
nuclear receptor	0.30	0.33
receptor complex	0.31	0.33
retinoic acid receptor	0.29	0.29
retinoid nuclear receptor	0.26	0.26

In Table 10 we compare the similarities of the term *retinoic acid receptor* to a number of terms. The first column represents the similarity values calculated with manually chosen weights, the second shows the corresponding values obtained with automatically learned weights, while the third column stands for the similarity values derived from the ontology. The measure with automatically determined weights showed a higher degree of stability relative to ontology-based similarity measure. For example, the ratio between the values in the first and third column ranged from 1.05 to 2.31, whilst the same ratio for the second and third column ranged from 1.26 to 1.54.

Table 10. The comparison of similarity values for term *retinoic acid receptor*

Term	***CLS*** α= 0.3, β=0.30, γ= 0.40	***CLS*** α= 0.13, β=0.81, γ=0.06	***O***
nuclear receptor	0.76	0.63	0.80
progesterone receptor	0.29	0.45	0.67
estrogen receptor	0.29	0.49	0.67
glucocorticoid receptor	0.29	0.49	0.67
human estrogen receptor	0.28	0.37	0.57

We have further experimented with term clustering using the *CLS* similarity. Clustering has been applied to a set of 174 top-ranked terms automatically extracted from the corpus using the C/NC-value method (Frantzi, Ananiadou and Mima 2000). Each row in the similarity matrix represented a similarity vector corresponding to the *CLS* similarity values between a given term and other terms from the set. The Euclidian distances between such vectors were used to establish clusters. We used hierarchical clustering based on two different clustering methods: the nearest neighbour and the Ward's method (cf. (Theodoridis and Koutroumbas 1999). These two methods are opposed to each other in the sense that the

nearest neighbour tends to produce long chain-like clusters, since the clusters are linked via their nearest members, while the Ward's method favours spherical clusters by minimising the increase in the sum of the distances between the members of a cluster. In both cases, the resulting hierarchy (dendrogram) was subsequently decomposed into a set of clusters by cutting off the hierarchy at the certain depth (chosen empirically) and collecting the leaves corresponding to sub-trees being cut off (see Figure 1).

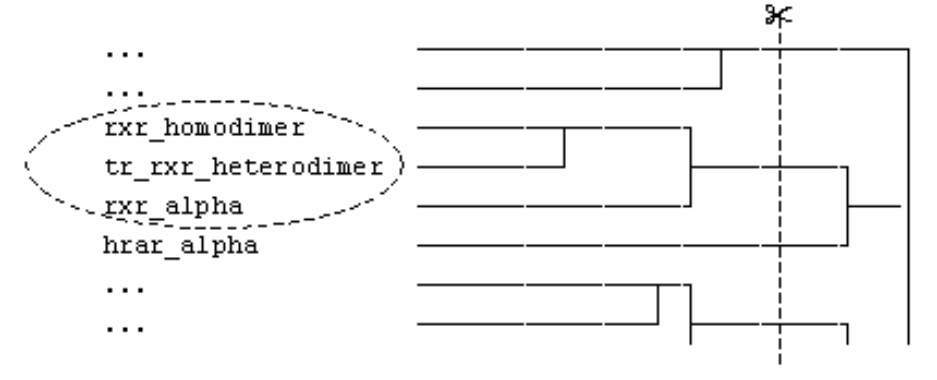

Figure 1. Producing clusters by cutting off the subtrees of the dendrogram

The resulting clusters have been evaluated by a domain expert, and the results, after discarding the singleton clusters, are given in Table 11 (see also (Nenadic, Spasic and Ananiadou 2002b)). Although the distribution of clusters differed significantly for the two clustering methods, the overall precision did not considerably vary: both methods achieved around 70% precision in clustering semantically similar terms. However, the higher number of small clusters produced by the Ward's method was preferred by the human evaluators, as the clusters were more coherent.

Table 11. Clustering results

Cardinality of a cluster	Nearest neighbour			Ward's method		
	# of clusters	# of correct		# of clusters	# of correct	
		clusters	terms		clusters	terms
2	16	7 (44%)	14	33	22 (67%)	44
3	7	6 (86%)	18	19	10 (53%)	30
4	4	2 (50%)	8	5	3 (60%)	12
≥ 5	10	7 (70%)	47	2	1 (50%)	8
Total:	37	22 (59%)	87 (63%)	59	36 (61%)	114 (71%)

Beside term clustering, the *CLS* similarity measure can be used for a number of term-oriented knowledge mining tasks. For example, a term-based corpus query engine has been presented in (Spasic, Nenadic, Manios and Ananiadou 2003a). Its main aim is to help domain specialists in locating and extracting related knowledge from scientific literature by using similarities among terms. Before querying, a corpus is automatically terminologically processed (the terminology recognition is based on the C/NC-value method, and similarities are mined by the method presented here). The results of terminology processing are stored in

an XML-native database (used as an ITM), which is subsequently used to query a corpus. Users can then formulate queries that generalise the classical IE task by retrieving, for example, entities that are “similar” or “associated” with given query terms.

5. Conclusions and further research

In this article we have presented a method for automatic mining of term similarities from documents. The method is based on combining lexical, syntactic and contextual similarities among terms and their occurrences. Lexical similarity exposes the resemblance among the words that constitute terms. Syntactic similarity is based on the co-occurrence in parallel lexico-syntactic patterns, while contextual similarity is based on the discovery of significant contexts through contextual pattern mining. Although the approach is domain independent and knowledge-poor, automatically collected patterns are domain dependent and they identify significant terminological contexts in which terms tend to appear. While lexical and syntactic similarities have low coverage, contextual similarity provides a similarity measure that covers the majority of semantically linked term pairs, and therefore it can be used for effective mining of associations among terms.

The presented measures are linearly combined in order to make use of all possible information that is mined for a pair of terms. While lexical similarity is typically limited to hyponymy relations, contextual and syntactic similarities reveal different types of domain-specific and functional associations among terms. In order to learn domain-appropriate term similarity parameters, we have used an ontology as a means of representing domain-specific knowledge needed for tuning the method for a specific domain.

The results in the domain of biomedicine have shown that the *CLS* measure proves to be a consistent indicator of semantic associations among terms, as similar terms tend to share the same “associates”. Our experiments with contextual similarity have demonstrated that terms belonging to semantically most related classes have a significantly higher degree of contextual similarity than terms belonging to weakly-related classes. This means that the contextual measure is coherent with semantic relatedness among terms. The clustering experiments have also shown that in 70% of cases terms were reliably grouped into clusters with their semantically related counterparts.

Still, further improvements can be made. Contextual similarity can be enhanced by incorporating weights and statistical properties for comparing term contexts. For example, if two terms appear exclusively in a certain context, then this fact is more important than an “incidental” sharing of a context by other terms. Similarly, different syntactic similarity relationships among terms (represented by different patterns) may be weighted: the values of syntactic similarity can be parameterised by the number and type of patterns in which two terms appear simultaneously. In order to increase the number of concurrent patterns, additional patterns (such as patterns describing appositions) can be considered. Lexical similarity can be generalised (in particular for single word terms) by combining alternative methods for lexical comparison (e.g. approximate string matching).

We believe that term similarity measures presented here also can be used for term sense disambiguation (e.g. by comparison of a contextual pattern corresponding to an ambiguous term occurrence with patterns relevant to each of the term senses), which is essential for resolving the terminological confusion occurring in many domains. Besides, our future work will also focus on term classification and consistent population and update of ontologies. However, in this case a specific term relationship identification rather than general term similarity is needed to place terms in a hierarchy.

Acknowledgements

This research was partially funded by BioPath, a Eureka project sponsored by the German Ministry of Research and coordinated by LION BioScience (Heidelberg, Germany). We would like to thank Dr. Dietrich Schuhmann, Dr. Sylvie Albert and Harald Kirsch (LION BioScience) for support and for the evaluation of results, and Dr. Hideki Mima (University of Tokyo) for fruitful discussions on terminology management issues.

Notes

[1] Non-terminal categories are given in angle brackets (<TERM>, <N>, <Adj> and <&>, the last denoting the following regular expression: (as well as) | and[/or] | or[/and]. Special characters (such as (,), [,], |, and *) have the usual interpretation in the regular expression notation.

[2] Note that we could learn only two parameters (and calculate the third at the end of the learning process). However, by learning three parameters we wanted to introduce more options during the mutation phase. Note also that any other optimisation or interpolation procedure could be used to learn the parameters.

References

Ananiadou, S. 1994. “A methodology for automatic term recognition.” In *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto, Japan, 1034-1038.

Blake, C. and W. Pratt. 2001. “Better rules, fewer features: a semantic approach to selecting features from text”. In *Proceedings of IEEE Data Mining Conference*, San Jose, CA.

Bourigault, D. 1992. “Surface grammatical analysis for the extraction of terminological noun phrases.” In *Proceedings of 14th International Conference on Computational Linguistics*, Nantes, France, 977-981.

Bourigault, D. and C. Jacquemin. 1999. “Term extraction + term clustering: an integrated platform for computer-aided terminology.” In *Proceedings of the 8th Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, Norway, 15-22.

Craven, M. and J. Kumlien. 1999: “Constructing biological knowledge bases by extracting information from text sources.” in *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*, Heidelberg, Germany, 77-86.

Dagan, I. and K. Church. 1994. “Termight: identifying and translating technical terminology”, In *Proceedings of Applied NLP (ANLP)*, 1994, 34-40.

Ding, J., D. Berleant, D. Nettleton and E. Wurtele. 2002. “Mining Medline: abstracts, sentences, or phrases?” in *Proceedings of Pacific Symposium on Bioinformatics 2002*, Hawaii, USA, 326-337.

Finkelstein-Landau, M. and E. Morin. 1999. “Extracting semantic relationships between terms: supervised vs. unsupervised methods.” In *Proceedings of International Workshop on Ontological Engineering on the Global Information Infrastructure*, Dagstuhl Castle, Germany, 71-80.

Frantzi, K.T., S. Ananiadou and H. Mima. 2000. “Automatic recognition of multi-word terms: the C-value/NC-value method.” *International Journal on Digital Libraries* 3(2), 115-130.

Grefenstette, G. 1994. *Exploration in Automatic Thesaurus Discovery*. Massachusetts: Kluwer Academic Publishers.

Hatzivassiloglou, V., P. Duboue and A. Rzetsky. 2001. “Disambiguating Proteins, Genes, and RNA in Text: A Machine Learning Approach.” *Bioinformatics* 17(1), S97-S106.

Hearst, M.A. 1992. “Automatic acquisition of hyponyms from large text corpora.” In *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France.

Hindle, D. 1990. “Noun classification from predicate-argument structures.” In *Proceedings of 28th Annual Meeting of the Association for Computational Linguistic* ,Pittsburgh, PA, USA, 268 275.

Hirschman, L., J. Park, J. Tsujii, L. Wong and C. Wu. 2002. “Accomplishments and challenges in literature data mining for biology.” *Bioinformatics* 18(12), 1553-1561.

Jacquemin, C. 2001. *Spotting and Discovering Terms through NLP*. Cambridge MA: MIT Press.

Kazama, J., T. Makino, Y. Ohta and J. Tsujii. 2002. “Tuning support vector machines for biomedical named entity recognition.” In *Proceedings of the Natural Language Processing in the Biomedical Domain (ACL 2002)*. Philadelphia, PA, USA, 1-8.

Kita, K., Y. Kato, T. Omoto and Y. Yano. 1994. “A comparative study of automatic extraction of collocations from corpora: Mutual information vs. cost criteria.” *Journal of Natural Language Processing* 1(1), 21-33.

Klavans, J. L., E. Tzoukermann and C. Jacquemin. 1997. “A natural language approach to multi-word term conflation.” In *Proceedings of DELOS Workshop on Cross-Language Information Retrieval*, Zurich, Switzerland, 33-40.

Lander, E. et al. (International Human Genome Sequencing Consortium). 2001. “Initial sequencing and analysis of the human genome.” *Nature* 409, 813-958.

Marcotte, E., I. Xenarios and D. Eisenberg. 2001. “Mining literature for protein-protein interactions.” *Bioinformatics* 17(4), 359-363.

Maynard, D. and S. Ananiadou. 1999. “A linguistic approach to terminological context clustering.” in *Proceedings of 5th Natural Language Processing Pacific Rim Symposium*, Beijing, China.

Maynard D. and S. Ananiadou, 2000a. “Identifying terms by their family and friends.” in *Proceedings of COLING 2000*, Luxembourg, 530-536.

Maynard, D. and S. Ananiadou. 2000b. “TRUCKS: a model for automatic multi-word term recognition.” *Journal of Natural Language Processing*, Vol. 8, No. 1, 101-125

Medline (2002): National Library of Medicine. http://www.ncbi.nlm.nih.gov/PubMed/

Mima, H., K. Ando and J. Aoe. 1995. “Incremental generation of LR(1) parse tables.” In *Proceedings of the third Natural Language Processing Pacific Rim Symposium*, Seoul, Korea, 600-605.

Mima, H., S. Ananiadou and G. Nenadic. 2001. “ATRACT workbench: an automatic term recognition and clustering of terms.” In Matousek, V. et al. (eds), *Text, Speech and Dialogue - TSD 2001*, LNAI 2166, Springer-Verlag, Berlin, 126-133.

Nakagawa, H. and T. Mori. 1998. “Nested collocation and compound noun for term recognition.” In *Proceedings of the First Workshop on Computational Terminology COMPUTERM 98*, 64-70.

Nenadic, G., I. Spasic and S. Ananiadou. 2002a. “Automatic acronym acquisition and term variation management within domain-specific texts.” In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Spain, 2155-2162.

Nenadic, G., I. Spasic and S. Ananiadou. 2002b. “Term clustering using a corpus-based similarity measure,” In Sojka, P. et al. (eds), *Text, Speech and Dialogue - TSD 2002*, LNAI 2448, Springer Verlag, 151-154.

Ohta, T., Y. Tateisi, J. Kim, H. Mima and J. Tsujii. 2002. “GENIA corpus: an annotated research abstract corpus in molecular biology domain”. In *Proceedings of HLT-2002*, 73-77.

Riloff, E. 1996. “Automatically generating extraction patterns from untagged text.” In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96)*, Portland, Oregon, USA, 1044-1049.

Reeves, C. 1996. “Modern heuristic techniques.” In Rayward-Smith, V. et. al (eds) *Modern Heuristic Search Methods*. New York: John Wiley & Sons Ltd. 1-25.

Sager, J.C. 1990. *A Practical Course in Terminology Processing*. Amsterdam/Philadelphia: Benjamins.

Sager, J.C., D. Dungworth and P.F. McDonald. 1980. *English Special Languages: Principles and Practice in Science and Technology*. Wiesbaden: Brandstetter.

Skuce, D. and I. Meyer. 1991. “Terminology and Knowledge Engineering: Exploring a Symbiotic Relationship,” in Proceedings of 6th International Workshop on Knowledge Acquisition for Knowledge-Based Systems, Banff, 29.1-29.21

Spasic, I., G. Nenadic, K. Manios and S. Ananiadou. 2002. “Supervised learning of term similarities." In Yin, Hujun et al. (eds), *Intelligent Data Engineering and Automated Learning – IDEAL 2002*, LNCS 2412, Springer Verlag, 429-434.

Spasic, I., G. Nenadic, K. Manios and S. Ananiadou. 2003a. “An integrated term-based corpus query system.” In *Proceedings of 10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, 243-250.

Spasic, I., G. Nenadic and S. Ananiadou. 2003b. "Using domain-specific verbs for term classification." In *Proceedings of ACL Workshop on Natural Language Processing in Biomedicine*, Sapporo, Japan.

Thelen, M. and E. Riloff. 2002. “A bootstrapping method for learning semantic lexicons using extraction pattern contexts.” In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, Pennsylvania, Philadelphia, PA, USA.

Theodoridis, S. and K. Koutroumbas. 1999. *Pattern Recognition*. Academic Press, Elsevier Science.

Voutilainen, A. and J. Heikkila. 1993. “An English Constraint Grammar (ENGCG) a surface-syntactic parser of English.” in Fries, U. et al. (eds), *Creating and Using English Language Corpora*, Rodopi, Amsterdam/Atlanta, 189-199.

Yakushiji, A., Y. Tateisi, Y. Miyao and J. Tsujii. 2001. “Event extraction from biomedical papers using a full parser.” In *Proceedings of Pacific Symposium on Biocomputing 2001*, Hawaii, USA, 408-419.