Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computer Speech and Language 19 (2005) 524–542
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

A symbolic approach to automatic multiword term structuring

Eric SanJuan [a,*], James Dowdall [b], Fidelia Ibekwe-SanJuan [c], Fabio Rinaldi [d]

[a] *LITA, Université Paul Verlaine & URI-INIST/CNRS, F-54514 France*
[b] *NLP Group, Dept. of Informatics, University of Sussex, BN1 9RH, UK*
[c] *University of LyonIII, 69007, France*
[d] *Institute of Computational Linguistics, University of Zurich, CH-8050, Switzerland*

Received 9 June 2004; received in revised form 6 January 2005; accepted 15 February 2005
Available online 19 March 2005

Abstract

This paper presents a three-level structuring of multiword terms basing on lexical inclusion, WordNet similarity and a clustering approach. Term clustering by automatic data analysis methods offers an interesting way of organizing a domain's knowledge structure, useful for several information-oriented tasks like science and technology watch, textmining, computer-assisted ontology population, Question Answering (Q–A). This paper explores how this three-level term structuring brings to light the knowledge structures from a corpus of genomics and compares the mapping of the domain topics against a hand-built ontology (the GENIA ontology). Ways of integrating the results into a Q–A system are discussed.

1. Introduction

It is a well-known fact that the majority of terminological units are multiword terms (henceforth MWTs). Current research on computational terminology has emphasized the need to dispose of structured terminology for several applications. To this end, a wealth of research has

* Corresponding address: IUT de Metz, dép. STID, Ile du Saulcy, 57045 Metz Cedex 1, France. Tel.: +33 3 87 31 51 60.
E-mail address: eric.sanjuan@iut.univ-metz.fr (E. SanJuan).

0885-2308/$ - see front matter
doi:10.1016/j.csl.2005.02.002

been directed toward identifying and organizing semantically related MWTs. The two families of approaches used for this task are distributional (statistical) and symbolic (linguistics) methods. Distributional similarity is taken as an indication of semantic similarity. The focus of many studies has been in creating classes of "similar" words: (Church and Hanks, 1990; Ushioda, 1996; Nenadić et al., 2002; Lin, 1998). All these methods result in a quantified similarity measure with the exact nature of the relations left undefined, and so heterogeneous or even antonymous concepts may end up in the same cluster. Lin (1998) creates similarity clusters by grouping words that occur in the same dependency relations in the SUSANNE corpus. By way of example, the most frequent words associated with the noun "*brief*" were "*affidavit, petition, memorandum, motion, lawsuit, deposition, slight, prospectus, document, paper*" which all hold different relations with the initial word, including collocational ones.

Alternatively, linguistic patterns are used to identify contexts which embody morphological, syntactic or semantic relations between MWTs. The linguistic approach can be subdivided into two main approaches: exogeneous methods relying on external semantic resource and endogeneous ones relying solely on evidence from corpora. Resources for exogeneous approaches are dictionaries (Hamon and Nazarenko, 2001), thesauri or ontologies. They allow to bootstrap semantic relations acquisition from corpora but clearly are dependent on the vocabulary coverage and availability of pre-existing resources.

Thesauri, taxonomies and ontologies are well-known tools for organizing the conceptual structures of a field. Yet these resources require considerable human effort and resources as well as time. As such, they are hardly readily available for every field and are rapidly overtaken by the constant appearance of new concepts. Although a huge effort is being dedicated towards semi- or fully-automated ontology building, the balk of the structuring still falls on the domain expert (Biébow and Szulman, 1999). Ontology expansion by populating an existing ontology with novel concepts provides a partial solution to the domain vocabulary coverage and structuring problem. Ontology populating tasks naturally utilize the existing conceptual structure. For the Unified Medical Language System (UMLS) (Humphreys et al., 1998), where the majority of related terms are identified manually, the thesaurus simply defines the set of possible relations. This process can be automated through compositional analysis of the MWTs by projecting relations between tokens onto relations between MWTs (Navigli and Velardi, 2004). However, for this technique to be successful, the ontology must already contain all of the tokens of a novel MWT. This is an unrealistic assumption in the case of GENIA corpus used in this study, where only 35.7% terminological tokens are in WordNet and 28.9% are in the UMLS.

In contrast with exogeneous methods, endogeneous approaches rely on shallow, bottom-up parsing and have the advantage of computational tractability. They are further subdivided into methods based on external or internal evidence.

Methods based on external evidence (Hearst, 1992; Morin and Jacquemin, 2004; Nenadic et al., 2004; Grabar and Zweigenbaum, 2004) search for lexico-syntactic cues like "such as" or "also known as" surrounding term structures which signal hypernym/hyponyms and synonyms relations, respectively. External evidence has been applied to terminology knowledge base construction (Condamines and Reyberolle, 1998) and ontology building (Aussenac-Gilles and Séguéla, 2000). However, this approach is inherently limited by the fact that it can only capture relations realized through the listed lexico-syntactic patterns. For instance, (Morin and Jacquemin, 2004)

report discovering 884 hypernyms relations in a corpus of almost 430,000 words (Jacquemin et al., 2002), with an average precision of 79% and an average recall of 46% (average *F*-score 58%).

Internal evidence refers to the case where evidence of the relation comes from within the term structure itself. This is generally called "variation" and covers operations of expansion (addition), structural transformation and substitution of lexical elements in a given term. For instance, the relation between ("*gene expression*" → "*human beta globin gene expression*") is one of hypernym/hyponym due to the addition of more specifiers to the generic term. The morpho-syntactic operations used to relate MWTs have been explored for a variety of applications such as building lexical resources from corpora (Daille, 2003; Jacquemin, 2001; Grabar and Zweigenbaum, 2004), automatic thesaurus enrichment (Morin and Jacquemin, 2004), domain knowledge mapping and textmining (Ibekwe-SanJuan, 1998; Ibekwe-SanJuan and SanJuan, 2004).

Although endogeneous methods offer sets of related terms, the structure proposed remains difficult to manage by human exploration. Indeed, it is very fastidious and quite inefficient to labor through thousands of terms in a database let alone try to grasp the conceptual organization of terms in the domain if no synthesis of the information is offered.

A way in which this synthesis can be approached is through data analysis methods and more specifically through clustering. Let us emphasize that this kind of structuring also differs from ontology building in which every term is listed, albeit in a hierarchy.

The need to achieve a meaningful synthesis of domain concepts is even more acute for applications like scientific and technological watch or textmining where experts are required to grasp topic emergence, shifts and obsolescence issues in limited time. Research on methods to this end, known as domain knowledge mapping (DKM) rely on powerful and suggestive visualization tools for result exploration. While a lot of research has been carried out separately on the two fields concerned here: computational terminology (see Jacquemin and Bourigault (2003) for a review) and DKM (see Schiffrin and Börner (2004) for a review), very few attempts have been made to bring the two together. Research on DKM traditionally relies on statistical models (co-occurrence models) to build clusters of frequently co-occurring items (Mane and Börner, 2004; Hearst, 1999; Small, 1999; Feldman et al., 1998). The challenge raised by our approach lies in extending further the integration of symbolic representations into a clustering algorithm for DKM. Earlier stages of this methodology have been published elsewhere (Ibekwe-SanJuan, 1998; Ibekwe-SanJuan and SanJuan, 2004). The focus of this paper is to evaluate the extent to which the automatic clustering of term variants based on symbolic relations reflects a hand-built knowledge structure. For this, after mining symbolic relations between terms gathered from the GENIA corpus, we cluster the terms based on these relations. The clusters so produced will be evaluated against a gold standard, the hand built GENIA ontology. The outcome of such an evaluation will determine if the methodology has uses for other knowledge organization tasks such as ontology population as it has up till now been solely applied to science and technology watch. We also discuss its potentialities in a Question Answering (Q–A) system focused on technical domains.

The rest of the paper is organized as follows: Section 2 describes the corpus used in this experiment and gives an overview of the methodology; Section 3 describes the three-level structuring of the MWTs; Section 4 evaluates the similarity of the automatic structuring against the hand-built GENIA ontology; Section 5 is devoted to discussions on the potentials of the term variant clustering for Q–A.

2. Corpus and methodology overview

The GENIA project (Kim et al., 2003) is an annotated corpus built to facilitate textmining in the field of genomics and thus promote bioinformatics using NLP techniques. It is also aimed to be a *"gold standard for the evaluation of textmining systems"* (Kim et al., 2003). This corpus deals with biological reactions concerning transcription factors in human blood cells. Utilizing the MEDLINE database and the MEdical Subject Headings (MeSH) thesaurus, records containing the keywords "*human", "blood cell*" and *"transcription factor*" were used to extract the titles and abstracts of 2000 articles[1] comprising more than 400,000 tokens. The corpus was manually enriched in XML by two domain experts. This led to almost 100,000 semantic annotations of which 26,789 unique terms were explicitly identified. Each biological term is assigned a semantic category from a small humanly constructed ontology, referred to as the GENIA ontology (see Fig. 3).

We exploited the GENIA annotation to extract MWTs, so that the terms we work on are the same the GENIA ontology is built on. This will ensure a plausible comparison between our MTWs structuring and the GENIA ontology itself.

2.1. Normalizing the MWTs from the GENIA corpus

Below is an example of a sentence from the GENIA corpus:

```
<cons lex = "IL-2_gene_expression" sem = "G#other_name">
<cons lex = "IL-2_gene" sem = "G#DNA_domain_or_region">
<w c = "NN">IL-2</w><w c = "NN">gene</w></cons>
<w c = "NN">expression</w></cons>
<w c = "CC">and</w>
<cons lex = "NF-kappa_B_activation" sem = "G#other_name">
<cons lex = "NF-kappa_B" sem = "G#protein_molecule">
<w c = "NN">NF-kappa</w> <w c = "NN">B</w></cons>
<w c = "NN">activation</w></cons>
```

Notice that the underlying XML markup of the terms (tag "cons lex") facilitates the identification of constituent MWTs, so "*IL-2 gene*" is a term in its own right which modifies the head *"expression*" to produce the full term in this instance, "*IL-2 gene expression*". Similarly, the GENIA annotation scheme disambiguates ellipsis in coordinated clauses by making explicit the terms involved. However, the GENIA corpus was not devoid of problems from an NLP perspective. There were many morphological variants amongst the terms which, unless corrected, would lead to spurious analyses in later stages. It was necessary to handle these variations in order to identify synonymous MWTs. We thus performed some normalizations on the terms which consisted in lower-casing every word whenever it exists in the corpus, harmonizing arbitrary punctuation use (for instance, "*gamma C chain"* & "*gamma (c) chain*"), harmonizing the irregular use of special characters (hyphens, slash, parenthesis, etc.) and retaining the singular form of each word.

[1] Version 3.0x, http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/.

For instance, "*Ca(2+)-dependent_pathway*" becomes "*Ca(2+) dependent pathway*". This is an ad-hoc pre-processing which will have to be adapted to each corpus, especially in technical domains where orthographic variations are frequent.

2.2. Methodology

Given that the terms were already annotated, the next step corresponding to the first level structuring consists in establishing binary "term–term" relations using the variation relations (see 3.1 for more details). Noisy relations are filtered out using WordNet (see 3.2). Basing on the remaining relations, connected components are formed by grouping together terms that share some modifier relations, i.e., terms that have the same head and a subset of common modifier words.

Components thus obtained are sets of terms formed around a particular domain paradigm or a mono-thematic family (see examples below). This constitutes the second-level of structuring. The components are grouped into classes iteratively according to the number of shared head variation links. This produces clusters of related domain topics that are mapped onto a 2D space using the AiSee[2] graphic display package and constitutes the third-level structuring. The whole methodology is embodied in the TermWatch system (Ibekwe-SanJuan and SanJuan, 2004) and relies on a hierarchical clustering algorithm specifically adapted to the linguistic nature of the relations. A detailed description of the clustering algorithm is given in Section 3.3.1. Below is an example of a class formed by four components. Terms within a component share modifier relations "*CD11b+ bone marrow cell*" is a modifier substitution of "*immature bone marrow cell*". Components are linked by head variation relations, i.e., "*bone marrow transplantation*" is a head expansion of "*bone marrow*".

- Comp1: CD11b+ *bone marrow cell; immature bone marrow cell; mouse bone marrow cell; normal bone marrow cell; normal bone marrow myeloid cell; normal* CD34+ *bone marrow cell; transgenic bonne marrow cell; murine bone marrow cell; primary murine bone marrow cell.*
- Comp2: *bone marrow transplantation; autologous bone marrow transplantation.*
- Comp3: *bone marrow; adult bone marrow; normal bone marrow.*
- Comp4: *bone marrow derived macrophage; murine bone marrow derived macrophage.*

What this class is suggesting is that research carried around *bone marrow* deals with the following topics (the added or substituted head words): *transplantation, cell, macrophage* whereas the modifier relations suggest the different "types" of *bone marrow* which are being studied (*CD11b+, immature, mouse, transgenic, murine, autologous, normal, adult, etc.*)

3. Structuring multiword terms

We describe in detail the types of variations used to relate the MWTs (Section 3.1) and the filtering process performed to remove some noisy variants (Section 3.2). These variations then serve as basis for the three-level structuring effected on MWTs in order to build classes (Section 3.3).

[2] www.aisee.com.

3.1. Lexical structuring of MWTs

The structuring capability of variation relations for a domain terminology has been attested in several studies. Under certain lexico-grammatical constraints[3] , syntactic variations yield conceptual relations between terms. Nenadic et al. (2004), Grabar and Zweigenbaum (2004) measured the "*lexical similarity*" between terms, i.e., "*the number of commonly shared words between a pair of terms*". In our study, we considered two types of syntactic variations: the addition (expansion) or substitution of nominal elements within a MWT. The two operations take place in the two syntactic structures: compound or syntagmatic (with a PP attachment) and can be viewed along the grammatical axis depending on whether they affect the head or modifier words. These variations have been described in (Ibekwe-SanJuan, 1998), we will recall them briefly here.

Expansions (or lexical inclusion) are subdivided into three types according to the position of the added words: left-expansion (L-Exp) is the addition of new modifier words and right-expansion (R-Exp) the addition of a new head. The combination of these two types results in left–right-expansions (LR-Exp). The addition of modifier words within a term results in an Insertion (Ins). Expansions engender asymmetrical relations in that they relate MWTs of different lengths, one being a subpart of the other. They are further constrained because we consider the addition of adjacent nominal elements (nouns, adjectives). This lessens the possibility of relating as variants, terms which portray arbitrary word order changes.

Substitutions are also subdivided into two types: modifier substitution (M-Sub) and head substitution (H-Sub). They identify variants of the same length (symmetrical links). This relation holds only between MWTs where one and only one word is different. An example of the rule identifying M-Sub is :

$$(t_2 \text{ is a } \mathrm{M-Sub} \text{ of } t_1) \Longleftrightarrow ((t_1 = M_1 m M_2 h) \text{ and } (t_2 = M_1 m' M_2 h) \text{ with } m' \neq m),$$

where t_1, t_2 are multiword terms, M_1, M_2 are strings of optional modifier words, m,m' are non-empty modifier words and h is the head noun.

Table 1 gives some examples of the syntactic variants found for "*blood cell*". The last two columns indicate the number of MWTs exhibiting each relation and the number of links created between the terms.

Eighty six percentage (23,314) of the MWTs found in the Genia corpus are involved in one or more types of syntactic variations. These represent general linguistic operations which can relate a high proportion of terms within the corpus, thus their coverage is very satisfactory.

3.2. Analyzing and filtering syntactic relations

The rationale in distinguishing modifier and head variations is that they do not convey the same linguistic information. Modifier variations affect the qualifiers whereas head variations fundamentally change the concept family. For this reason, left-expansion (L-Exp) naturally reflects the fact that more specific MWTs have more modifiers. However, the resulting conceptual relations are not straightforward for insertions (Ins) as changing the head–modifier relations of a MWT creates a structural (and therefore conceptual) ambiguity. For example, "*HIV 1 expression*" **IS_A** kind of

[3] The morphological category and the grammatical role of inserted words.

Table 1
Types and proportion of syntactic variations found in the GENIA corpus

Types	Example: *blood cell*	Terms	Links
Expansions			
L-Exp	*mononuclear blood cell*	5352	10,153
R-Exp	*blood cell receptor*	6641	7337
LR-Exp	*white blood cell count*	3698	3767
Ins	*blood mononuclear cell*	4821	6133
Substitutions			
M-Sub	*stromal cell*	14,865	437,291
H-Sub	*blood pressure*	11,702	111,068

"*HIV expression*" but this certainty diminishes as the number of inserted modifiers increases, "*HIV 2 gene expression*" and "*HIV LTR driven luciferase expression*". With this in mind, insertions that involve only a single additional modifier and left-expansions can be used to create **IS_A** hierarchies around concept families. This permits a MWT to have more than one parent (see "*HIV gene expression*" in Fig. 1).

These observations suggest that among the variations that do not change the head word, left-expansions (L-Exp) should be given priority for building components (second-level structuring) if we want to obtain homogeneous clusters vis-à-vis the GENIA ontology.

Substitutions engender horizontal relations between terms. Therefore, the resulting conceptual relation is a more general relatedness. Modifier substitutions (M-Sub) can denote members of the same concept family with alternative qualifications, otherwise known as "co-hyponyms" or "siblings" in an **IS_A** hierarchy. The conceptual shift engendered by head substitutions (H-Sub), on the other hand, links different **IS_A** hierarchies at the same level of specificity.

For example, in Fig. 1, although "*gene expression*" and "*gene transcription*" are head substitutions (thus normally implying a topical shift), there is still a conceptual link: the "*expression*" of a "*gene*" is the result of its "*transcription*". However, the same variation also links "*gene*" as it

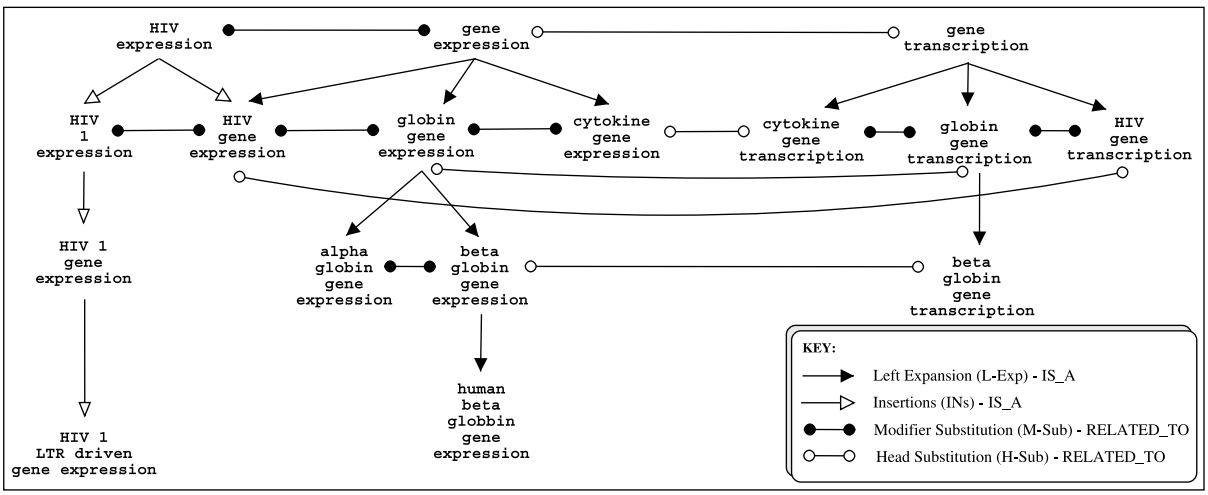


Fig. 1. Fragment of the conceptual hierarchy induced by morpho-syntactic variations.

modifies the two word MWTs headed by "*regulation*", "*knockout*", "*activation*" and "*product*", to name only a few. As Table 1 shows, substitutions are by far the most frequent type of variations with the vast majority of the links. For this reason, they are further filtered using WordNet's lexical taxonomy (Fellbaum, 1998) to obtain a category of semantically motivated relations, henceforth called "WordNet substitutions".

WordNet Substitutions (WN-Sub) are those pairs of variants found in the corpus, in which the substituted words belong to the same synset (see examples in Table 2). With the conceptual classes formed by WordNet synsets, we allow the substitution of both the head and modifier words as in "*hormone effect*" and "*endocrine event*".

Using a general lexical resource like WordNet to relate the MWTs identifies those words that are both related in a general vocabulary. Evaluating the overlap in "general knowledge" and "specialized knowledge" brings two observations. First, the coverage of WordNet over the GENIA corpus is limited with the result that WN-Subs are relatively rare. Second, the actual conceptual relation produced by MWTs in a specialized field can differ from the generic one suggested by a general language resource. For instance, within the genomic domain, "***strong***" refers to the degree to which a "*represser*" binds to the DNA whereas "***potent***" refers to the degree of its effect. Similarly, an "*inflammatory* ***response***"causes an "*inflammatory* ***reaction***"(the process of becoming inflamed). These are clearly more related than the syntactic substitutions but they are not synonyms in the genomics domain as WordNet synsets seem to suggest. However, the fact that WordNet relates them is good enough for the clustering task because they will end up in the same component, and thus be strongly related in the resulting domain knowledge structure. Despite the fact that general resources cannot capture the explicit conceptual relation between specialized domain terms, we still highly improved the precision of the substitutions variants using WordNet, in the sense that 97% of the WN-Subs linked semantically related terms. Only 304 links were present in WordNet among the 548 359 possible substitutions found in the corpus. Note however that this is not a measure of recall/precision since WordNet is not a specialized resource. A more adequate recall/precision measure would be obtained in comparing the corpus substitutions against the Mesh thesaurus. Let us recall that the corpus was extracted from Medline which relies on the Mesh thesaurus. In the present experiment, this low number of WN-Subs does not seem to be a drawback in our approach. On the contrary, we will see later in Section 3.3.2 that we need to severely restrict the set of substitutions in other to avoid the chain effect, well known in some clustering approaches.

3.3. Mapping a domain terminology

The aim is to produce knowledge maps of important clusters reflecting domain topics and their associations. We first describe the clustering algorithm (Section 3.3.1) and its application to the GENIA MWTs (Section 3.3.2).

Table 2
Semantic substitutions identified through WordNet

WordNet	WN-Sub1	WN-Sub2
M-Sub	***strong*** *transcriptional represser*	***potent*** *transcriptional represser*
H-Sub	*inflammatory* ***reaction***	*inflammatory* ***response***
HM-Sub	***hormone effect***	***endocrine event***

3.3.1. Term variant clustering

The variation relations used as basis for the clustering are represented as a graph. We recall briefly the functioning of the algorithm. Clustering is a two-stage process. First the algorithm builds connected components using a subset of the variation relations, usually the modifier relations (L-Exp, Ins, M-Sub). We call these COMP relations.

The transitive closure of COMP relations (COMP*) partitions the whole set of MWTs into components. These connected components are sub-graphs of MWT variants that share the same head word or a synonym attested by WordNet synsets. At the second stage, the connected components are clustered into classes using the head relations (R-Exp, LR-Exp, H-sub), this subset of relations is called CLAS. At this stage, components whose terms are in one of the CLAS relations are grouped basing on a similarity coefficient s computed thus:

$$s(i,j) = \sum_{R \in CLAS} \frac{N_R(i,j)}{|R|},$$

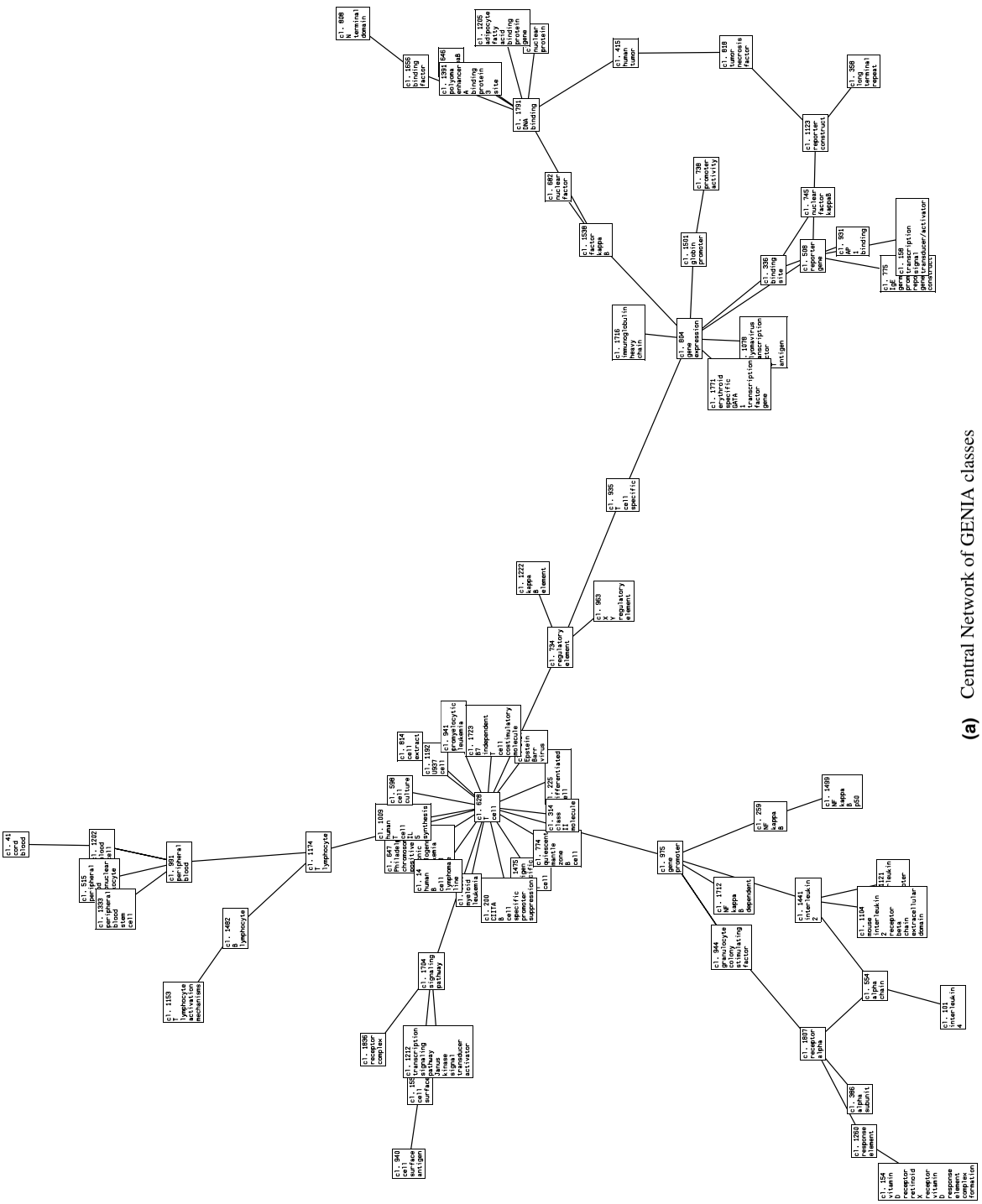
where R is a variation relation in CLAS, $|R|$ is the number of pairs of terms related by R and $N_R(i,j)$ is the number of these pairs between components i and j.

3.3.2. Clustering the GENIA term variants

Preliminary clustering tests and the observations made in Section 3.2 led us to modify here the roles usually assigned to the syntactic variations during clustering. Following observations in this section, we further split L-Exp into two sub-relations: strong-L-Exp and weak-L-Exp depending on if there was a unique or more appended modifiers. We selected WN-Sub and strong-L-Exp as COMP relations whereas Ins, weak-L-Exp, R-Exp, LR-Exp served as CLAS variations. Consequently, terms sharing the same connected component can have different heads, semantically related through WordNet synsets. Conversely, weak-expansions and insertions were excluded from the COMP set of relations because they led to too big components ($\geqslant$2000 terms) on this corpus.

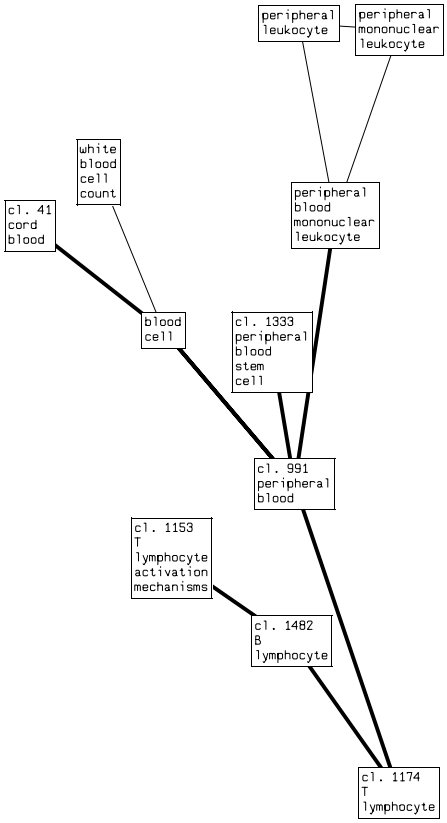
Empirical tests showed the classes produced at the 2nd iteration of the algorithm to be the most legible in terms of size and content. This produced 1664 classes, 6151 components and a total of 10285 MWTs in the classes. The output of the clustering module is automatically formatted in the Graph Description Language (GDL) used by AiSee for visualization. To visualize the underlying structure of the network of classes, the user can temporarily hide very weak links between them. This gives the image in Fig. 2(a) that shows the structure of the graph. Each class is labelled automatically by the term that shares the highest number of variation links outside the cluster. The global image obtained exhibits a star shape with a central core, related to a cyclic subgraph. By order of importance, the central position is occupied by a big class labeled "*T-Cell*" with 374 terms. A second smaller sub network is formed around the class labelled "*gene expression*" with 235 terms.

Each class can be unfolded to show its internal structure: the connected components and then its most active variants. The user can thus immediately perceive the most salient features of a class. Fig. 2(b) is a zoom on the classes surrounding "*blood cell*". This shows its internal links ("*white blood cell count*") but also its external links. A sub-network emerges from the structure of the three classes "*T lymphocyte*", "*B lymphocyte*", "*T lymphocyte activation mechanism*". These form a linear graph, i.e., chains of relatively long vertices starting from a central class to the border of the graph which have rarely more than one outgoing link. The visualization interface



(a) Central Network of GENIA classes

Fig. 2. Graphs of term variant classes displayed with AiSee.



(b) Classes related to "blood cell"

Fig. 2 (*continued*)

naturally aligns the elements of these linear graphs, thus highlighting them. The length of an arc has a straightforward meaning here. Strong variation links are symbolized by shorter arcs while weak links are symbolized by longer arcs. This kind of interactive manipulations using the AiSee interface allows the user to access simultaneously the three levels of the clustering results: classes, components and terms.

The two classes "*T lymphocyte*" and "*B lymphocyte*" contain, respectively, terms like "*activated T lymphocyte*", "*human peripheral lymphocyte*", "*activated peripheral blood lymphocyte*" for the former and "*B lymphocyte specific mb 1 gene*", "*normal B lymphocyte*", "*B lymphocyte growth transformation*" for the latter. Their link with the classes dealing with the "*blood cell*" and "*white blood cell*" or "*leucocytes*" is coherent because a "*lymphocyte is a form of leucocyte occurring in the blood*", "*in the lymph*" and a "*lymph is a colourless fluid containing white blood cells*".[4] Term Watch thus seems to have effected coherent thematic associations in the domain via syntactic variations and the few WN-Subs found in the corpus. We will now examine to what extent these classes also reflect the hand-built GENIA ontology.

[4] Concise Oxford Dictionary, Allen R.E. (Eds.), eighth ed., pp. 708–709.

4. Evaluation of the classes against the GENIA ontology

A clustering process is supposed to group together similar objects basing on some criteria. For domain knowledge mapping (DKM) and text mining systems, the criteria are usually statistical (co-occurrence of text units). Here we relied on symbolic criteria: the number and type of variation relations between terms which result in iteratively grouping sets of related MWTs. Although, we produced a sort of hierarchy (the inclusion of one class into another), it is a formal hierarchy stemming from a clustering algorithm, fundamentally different from the semantic hierarchy in an ontology. Mapped onto a 2D space, results from a clustering algorithm are meant to highlight spatial structures whose interpretation holds a strategic dimension,[5] for science and technology watch. This is quite different from the interpretations made on the hierarchy resulting from an ontology or any other semantic organization of domain concepts. However, any ontology induces an idea of similarity. The comparison of the two structures is based on the following assumptions:

Assumption 1: two terms from the GENIA ontology can be considered close if they were assigned the same semantic category, or if the level of the common subsuming concept is not too far from the nodes considered,

Assumption 2: TermWatch's classes supposes a "semantic proximity" between terms in the same component and "a weaker semantic proximity" between terms in the same class,

Assumption 3: for the evaluation task, we hypothesize that the distance between the two structurings may not be as big as the underlying organizing principles in both structures may suggest.

To test these assumptions, we try to answer the following question: *i*f two terms are close in the GENIA ontology (according to "assumption 1"), do they tend to appear in the same class in TermWatch's output?

For that purpose, let us call *atomic category* the categories at the leaves of the GENIA ontology. Then we map the set of TermWatch classes onto the GENIA ontology by associating each component and class with their dominant atomic category, i.e., the atomic category that has the highest number of terms in the class.

By way of example, the component labelled "NF kappaB" has five terms. Four of them: "*NF kappaB*", "*lung NF kappaB*", "*mammalian NF kappaB*", "*nuclear NF kappaB*" are assigned the "protein_complex" category, and only the fifth one: "*cytoplasmic NF kappaB*" comes from a different category: "protein_molecule". Thus "protein_complex" category will be associated with this component which clearly has a high degree of homogeneity (80%) vis-à-vis the GENIA ontology. This component is an element of a class that has the same label "NF kappaB" but not the same dominant GENIA group which is "protein_molecule".

Table 3 shows the categories (excluding "other_name") associated with the nine classes that have more than 50 terms. "other_name" was designed as a miscellaneous category to receive all the terms that could not be assigned a more specific semantic type by the GENIA ontology builders.

The labels of the classes are given in the fourth column. The associated category (the dominant one) is given in the last column. The first column "Nb_G" shows the number of terms in the class

[5] The notions of "central" vs. "border" topics, topic "growth" vs. "obsolescence" are crucial here.

Table 3
GENIA categories associated with the biggest classes

Nb_G	Nb_c	Rate	Class label	Associated GENIA category
32	86	0.37	NF kappaB	Protein_molecule
30	72	0.42	Mouse gene	DNA_domain_o r_region
43	99	0.43	DNA binding	Protein_family_or_group
31	60	0.52	Response element	DNA_domain_or_region
218	374	0.58	T-cell	Cell_line
47	73	0.64	E-Box	DNA_domain_or_region
41	64	0.64	Human enhancer	DNA_domain_or_region
78	112	0.70	Binding site	DNA_domain_or_region
45	63	0.71	N-terminal domain	Protein_domain_or_region

that share the dominant category, the second column "Nb_C" shows the total number of terms in the class and the following called "rate" gives the ratio between the previous two numbers.

Hence, Table 3 shows that the biggest classes produced by TermWatch have more than 40% of their terms in the same GENIA category, except for class "NF kappaB". These categories are also the most frequent in the GENIA corpus. However, we show in the sequel that other categories also appear in the clustering output, notwithstanding their low frequency. A low score does not however signify that a class is an error with regard to the GENIA ontology. Analyzing class "NF kappaB" whose dominant GENIA category ("protein_molecule") represents only 37% of its terms, we find out that all the GENIA categories of terms in this class are subsumed under the same common father concept in the ontology, namely "protein". We present now some statistics to verify if these local observations apply to the majority of the components and classes.

Before computing these statistics, we have to consider separately the miscellaneous category "other_name" which subsumes 21% of the GENIA ontology terms. A first observation is that our comparison showed the homogeneity of TermWatch's components and classes associated with "other_name" to be very high. The average proportion of terms belonging to this category is 98% for components and 85% for classes. This shows that the syntatic relations used in clustering were able to isolate terms in this miscellaneous category from the rest.

In the following, the rest of the comparison is performed on the remaining components and classes associated with the rest of the GENIA categories. These involve 1063 components, 659 classes and a total of 5674 terms.

We start by computing the number of components and classes associated to each atomic category of the GENIA ontology. For that purpose we consider:

- the distribution d_G of the most frequent GENIA categories in the original corpus over the total number of term occurrences in the GENIA corpus.
- the distributions d_{comp} and d_{class} of dominant categories in components and classes, respectively.

Thus, for a given category c like "protein_molecule" which is the most frequent category in the GENIA ontology,

$d_G(c)$ is the number of term occurrences in the GENIA corpus having the category "c" = "protein_molecule" which is 15348 in this case, divided by the total number of occurrences (95,138).

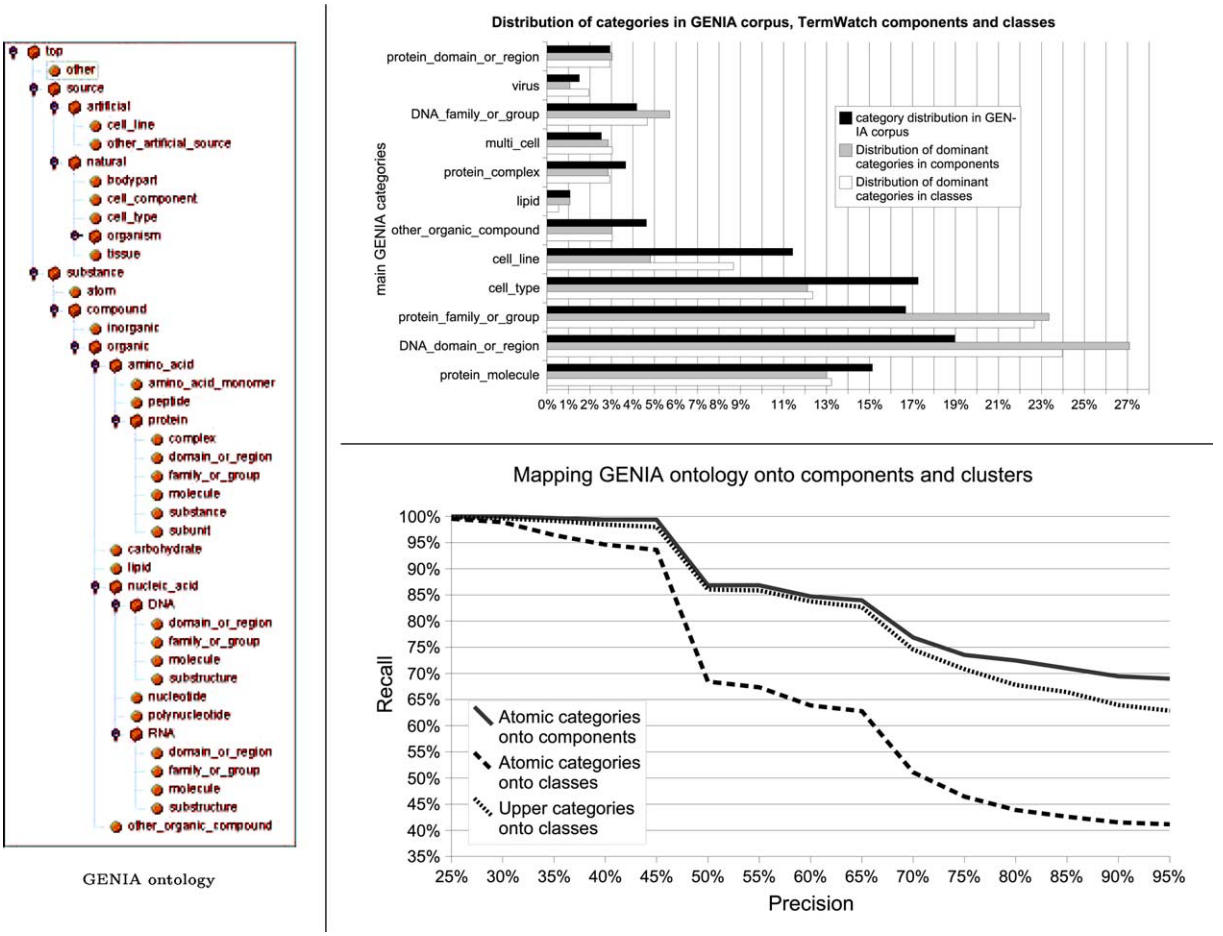


Fig. 3. Mapping GENIA categories onto TermWatch classes.

$d_{\text{comp}}(c)$ is the number of components in TermWatch output, of which the majority of the terms are in category c (122 in this case), divided by the total number of components (1063).

$d_{\text{class}}(c)$ is the same as $d_{\text{comp}}(c)$ except that we consider the 659 classes instead of components. 73 classes are associated with "protein_molecule".

The right topmost graphic in Fig. 3 ("Distribution of categories in GENIA corpus, TermWatch components and classes") allows us to compare the 12 topmost values of d_G (represented by the upper black bars) with the corresponding values of d_{comp} and d_{class}, respectively, represented by the middle grey bars and the lowest white bars, respectively.

This figure shows that classes, more than components, lessens the deviation from the distribution of GENIA categories in the corpus (except for the small category "lipid"). In fact, $d_{\text{comp}}(c)$ is much lower than $d_G(c)$ whenever category c contains terms like "*T-cell*" that generate huge components which only account for one occurrence of the category.

Now we use the concepts of precision and recall to analyze the quality of these mappings. Since we are not evaluating here a Q–A performance but the ability of a clustering algorithm to discern

terms from different semantic categories, we defined recall and precision slightly differently from the way in which they are used in Information Retrieval.

We identify each GENIA category G with the set of associated terms. Let $\mathscr{G}$ be a family of GENIA categories and let $\mathscr{X}$ be one of the families of components or classes. Using these notations, we clearly have for any $X \in \mathscr{X}$ the equality:

$$|G_X \cap X| = \max\{|G \cap X| : G \in \mathscr{G}\}.$$

Precision p can be defined for any cluster (component or class) X as the proportion of terms in X that are in G_X:

$$p(X) = \frac{|G_X \cap X|}{|X|}.$$

Hence, knowing that a term t is in a cluster X, the value $v = p(X)$ is the conditional probability $G_X|X$ of finding effectively t in the category G_X.

The recall r is defined for any precision value $v = p(X)$ as the proportion of clusters whose precision is higher than v:

$$r(v) = \frac{|\{X \in \mathscr{X} : p(X) \geqslant v\}|}{|\mathscr{X}|}.$$

Precision/recall functions associate with each value $v \in [0,1]$ the corresponding recall value. They are decreasing one-to-one functions. In fact, the precision/recall functions defined here roughly correspond to those induced by the theoretical IR system where documents are assumed to be the terms in the clusters, and the set of categories is viewed as a set of queries. Then for each category, the system would retrieve the list of terms in clusters where this category is dominant. The analogy would be perfect if all the clusters had the same size. Let us now apply these concepts to the clusters. The right bottom graphic in Fig. 3 shows three precision/recall functions computed on components and classes using different families $\mathscr{G}$ of categories. The uppermost bold line curve shows the function obtained by setting $\mathscr{X}$ to the whole set of components, and $_{\mathscr{G}}$ to the whole set of GENIA atomic categories.It shows that the syntactic variations used to cluster terms into components link essentially terms in the same GENIA category. For instance, all the components $X \in \mathscr{X}$ have at least 48% of their terms in the dominant GENIA category G_X associated with X, while 68% of the components attain a 100% inclusion in G_X, thus in one semantic type. This is not entirely surprising as components are formed by variations affecting the modifier elements in a term, thus components have the same head word or a synonym attested by WordNet synsets.

Classes on the other hand group several components, thus variants with different heads. The lowest dashed curve shows the precision/recall function by setting $\mathscr{X}$ to the clusters and $\mathscr{G}$ to the whole set of GENIA atomic categories. Naturally, the semantic inclusion in one category is much lower than for the components. Still a comparable proportion of clusters (95%), reach a precision of 41% and 40% of them have a 100% semantic inclusion in the category G_X to which they are associated.

We then considered the upper categories in the GENIA taxonomy by merging together terms belonging to the same common parent category, thus by changing the previous $\mathscr{G}$ family of considered categories. For instance, we merged on the one hand, terms on the super categories "DNA" and "RNA", and on the other hand, terms from categories containing "*cell*"

(“cell_type”, “cell_components”, “cell_line”) into their super category: “source”. We then mapped these upper-level categories onto the classes. We observed that the semantic inclusion of the classes increased and moved closer to the distribution of the ontology categories in the components. This is represented by the middle dotted curve on the Fig. 3.

These findings suggest:

- that forming clusters by syntactic variations is a sound linguistic approach which links together conceptually related terms;
- that naturally, components tend to be monolithic in terms of semantic class, i.e., they link together one family of concepts sharing different attributes;
- that TermWatch's classes, while not being monolithic in terms of semantic class still group together coherent domain topics which are logically associated;
- that as we move up a taxonomy to consider more generic GENIA categories, the semantic monolithy of the classes formed by TermWatch tends to increase.

5. Discussion

Structuring multiword terms using symbolic criteria is a promising research concern as it enables us to discover automatically meaningful associations between domain concepts which are useful for several tasks. We are currently seeking ways to integrate this multi-level structuring in a Question Answering (Q–A) application. We briefly describe the Q–A system and discuss ways of integrating the two approaches as well as other points of improvement.

ExtrAns is a Question Answering system aimed at restricted domains, in particular terminology-rich domains (Rinaldi et al., 2004b). While open domain Question Answering systems are targeted at large text collections and use relatively little linguistic information, ExtrAns answers questions over such domains by exploiting linguistic knowledge from the documents and terminological knowledge about a specific domain. Various applications of the ExtrAns system have been developed, from the original prototype aimed at the Unix documentation files to a version targeting the Aircraft Maintenance Manuals (AMM) of the Airbus A320 (Mollá et al., 2003). Recently the system has been applied to document collections based on scientific literature in the “Life Sciences” area (Rinaldi et al., 2004a). ExtrAns's approach to Question Answering is particularly computationally intensive: this allows a deeper linguistic analysis to be performed, at the cost of higher processing time. The documents are analyzed in an off-line stage and transformed in a semantic representation, based on logical forms which is stored in a Knowledge Base (KB). Documents (and queries) are subjected to the same processing stages: first they are tokenized, then they go through a terminology-processing module. If a term belonging to a synset in the terminological knowledge base is detected, then the term is replaced by a synset identifier in the logical form. This results in a canonical form, where the synset identifier denotes the concept that each of the terms in the synset names. In this way any term contained in a user query is automatically mapped to all its variants. This approach amounts to an implicit “terminological normalization” for the domain, where the synset identifier can be taken as a reference to the “concept” that each of the terms in the synset describes.

Unlike sentences in documents, user queries are processed on-line and the resulting semantic representations are proved by deduction over the contents of the KB. When no direct answer for a user query can be found, the system is able to relax the proof criteria in a stepwise manner. First, hyponyms are added to the query terms. This makes the query more general but maintains its logical correctness. If no answers can be found or the user determines that they are not good answers, the system will attempt approximate matching, in which the sentence that has the highest overlap of predicates with the query is retrieved. The matching sentences are scored and the best matches are returned.

The multi-level terminology structuring scheme presented here can be effectively exploited in locating answers. The answer strategy that we are considering can be summarized as:[6]

(1) First, extract potential answers that involve strictly synonymous MWTs.
(2) Second, look for potential answers with WordNet related MWTs.
(3) Third, try hypernyms/hyponyms acquired through lexico-syntactic patterns.
(4) Finally, allow the user to browse the clusters of MWTs to comprehend the conceptual organization of the research topics and identify which terms are of interest to his query.

This set then becomes the basis of a second round of answering specific questions. In this way the system can provide useful access to users by facilitating navigation through a domain of unfamiliar MWTs. For example, when looking for general information on "*blood cell*" a user may well be interested in its "*count*", the second different head word in this class (see Fig. 2). By presenting the graph of classes, the user can also browse related topics *(T lymphocyte, Peripheral blood, Peripheral blood mononuclear leucocyte, cord blood, T lymphocyte, B lymphocyte)* and thus grasp the different topics addressed in the corpus in connection with "*blood cell*" before deciding on more precise terms for the query. The classes can thus assist the query refinement process. However, experiments involving real users are still to be carried out in order to test these hypotheses.

Other areas of improvement on the current work are the acquisition of semantically related terms through the use of lexico-syntactic patterns found in the corpus. We have seen that some of the syntactic variations needed to be filtered through semantic constraints, and that using an external resource is often limited in terms of corpus vocabulary coverage. This resulted in a drastic drop in the number of semantically related terms recovered. To overcome this handicap, we identified semantically related terms using the lexico-syntactic cues basing on works done by Hearst (1992) and Morin and Jacquemin (2004) for hypernym/hyponym relations. In this case, the evidence for a semantic relation between MWTs comes from the corpus itself. The underlying hypothesis is that semantic relations can be expressed via a variety of surface lexical and syntactic patterns. These relations will augment the ones already used for clustering and will constitute a higher order level of structuring which selects semantically related terms from amongst the other lexical associations. They are yet to be integrated into the clustering algorithm. This will involve a re-ordering of the whole set of relations according to a scale of "semantic proximity" they engender between two terms. Following the outcome, each relation type will be assigned a role (COMP or CLAS) during the classification.

[6] While steps (1–3) are actually implemented, step (4) is currently under experimentation.

Lastly, there is need to compare the output of the clustering algorithm used in TermWatch with other existing algorithms based on statistical criterion (co-occurrence). To this end, we tried clustering the list of GENIA terms using a standard clustering method.[7] It takes as input the number of co-occurrence of terms in GENIA corpus. We also computed the resulting precision/recall functions as in Fig. 3, but none of them reached 35% of recall for 50% of precision. This poor performance is due to very low co-occurrence values (more than 33% of terms have less than two occurrences in the abstracts). To increase these values, it is necessary to take into account the variation phenomena. This can be done only by taking into account symbolic relations between the clustered units. Further and more profound experiments need to be carried out to compare Term Watch's output to other statistical clustering methods. Meanwhile, from this experiment, it appears that the co-occurrence paradigm is not suited to uncovering, from the corpus, the semantic links annotated in the GENIA ontology.

References

Aussenac-Gilles, N., Séguéla, P., 2000. Les relations sémantiques: du linguistique au formel. Cahiers de Grammaire 25, 175–198.

Biébow, B., Szulman, S., 1999. Terminae: a linguistics-based tool for building of a domain ontology. In: Fensel, D., Studer, R. (Eds.), Proceedings of the 11th European Workshop EKAW'99. Springer-Verlag, pp. 49–66.

Church, K.W., Hanks, P., 1990. Word association norms, mutual information and lexicography. Computational Linguistics 16 (1), 22–29.

Condamines, A., Reyberolle, J., 1998. Ctkb: a corpus-based approach to a terminological knowledge base. In: Proceedings of the 1st International Workshop on Computerm. In COLING-ACL'98. Berlin-Springer, Montreal, pp. 29–35.

Daille, B., July 2003. Conceptual structuring through term variations. In: Proceedings of the ACL-2003 Workshop on MultiWord Expressions: Analysis, Acquisition and Treatment, Saporro, Japan. pp. 9–16.

Feldman, R., Fresko, M., Kinar, Y., et al., 1998. Text mining at the term level. In: Zytkow, J.M., Quafafou, M. (Eds.), Principles of Datamining and Knowledge Discovery. Proceedings of the 2nd European Symposium PKDD'98. Berlin-Springer, Nantes-France, pp. 65–73.

Fellbaum, C. (Ed.), 1998. WordNet, an Electronic Lexical Database. MIT Press.

Grabar, N., Zweigenbaum, P., 2004. Lexically based terminology structuring: some inherent limitations. Recent trends in computational terminology: special issue of terminology 10 (1), 23–53.

Hamon, T., Nazarenko, A., 2001. Detection of synonymy links between terms. In: Bourigault, D., Jacquemin, C., L'Homme, M.-C. (Eds.), Recent Advances in Computational Terminology. John Benjamins, pp. 185–208.

Hearst, M., 1992. Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the COLING'92, Nantes, pp. 539–545.

Hearst, M., June 1999. Untangling text data mining. In: Proceedings of the 37th Annual meeting of the Association for Computational Linguistics, MD, USA.

Humphreys, B.L., Lindberg, D.A.B., Schoolman, H.M., Barnett, G.O., 1998. The unified medical language system: an informatics research collaboration. JAMIA 5, 1–11.

Ibekwe-SanJuan, F., August 1998. A linguistic and mathematical method for mapping thematic trends from texts. In: Proceedings of the 13th European Conference on Artificial Intelligence (ECAI). Brighton, UK, pp. 170–174.

Ibekwe-SanJuan, F., SanJuan, E., April 2004. Mining textual data through term variant clustering: the termwatch system. In: Proceedings of Recherche d'lnformation assistée par ordinateur (RIAO). Avignon, pp. 26–28.

[7] FASTCLUST (*k*-means) and CLUSTER (complete linkage) procedures in SAS system for Windows V8 (SAS Institute Inc., Cary, NC, USA).

Jacquemin, C., 2001. Spotting and Discovering Terms through Natural Language Processing. MIT Press.

Jacquemin, C., Bourigault, D., 2003. Term extraction and automatic indexing. In: Mitkov, R. (Ed.), The Oxford Handbook of Computational Linguistics. Oxford University Press, pp. 599–615.

Jacquemin, C., Daille, B., Royauté, J., Polanco, X., 2002. In vitro evaluation of a program for machine-aided indexing. Source Information Processing and Management 38 (6), 765–792.

Kim, J.-D., Ohta, T., Tateisi, Y., Tsujii, J., 2003. Genia corpus – a semanticall annotated corpus for bio-textmining. Bioinformatics 19 (1), 1180–1182.

Lin, D., August 1998. Automatic retrieval and clustering of similar words. In: Proceedings of the Joint international conference ACL-COLING. Montreal, pp. 768–773.

Mane, K., Börner, K., 2004. Mapping topics and topic bursts in pnas. Publication of the National Academy of Science (PNAS) 101 (1), 5287–5290.

Mollá, D., Rinaldi, F., Schwitter, R., Dowdall, J., Hess, M., 2003. Answer extraction from technical texts. IEEE Intelligent Systems.

Morin, E., Jacquemin, C., 2004. Automatic acquisition and expansion of hypernym links. Computer and the Humanities, 36.

Navigli, R., Velardi, P., 2004. Learning domain ontologies from document warehouses and dedicated web sites. Computational Linguistics 30 (2), 151–179.

Nenadić, G., Spasić, I., Ananiadou, S., 2002. Automatic discovery of term similarities using patter mining. In: Proceedings of the Second International Workshop on Computational Terminology (CompuTerm). Taipei, Taiwan.

Nenadic, G., Spassic, I., Ananiadou, S., 2004. Mining term simililarities from corpora. Recent Trends in Computational Terminology: Special Issue of Terminology 10 (1), 34.

Rinaldi, F., Dowdall, J., Schneider, G., Persidis, A., 2004a. Answering questions in the genomics domain. In: The ACL2004 Workshop on Question Answering in Restricted Domains, Barcelona, July 2004.

Rinaldi, F., Hess, M., Dowdall, J., Mollá, D., Schwitter, R., 2004b. Question answering in terminology-rich technical domains. In: Maybury, M. (Ed.), New Directions in Question Answering. MIT/AAAI Press.

Schiffrin, R., Börner, K., 2004. Mapping knowledge domains. Publication of the National Academy of Science (PNAS) 101 (Suppl. 1), 5183–5185.

Small, H., 1999. Visualizing science by citation mapping. JASIS 50 (9), 799–813.

Ushioda, A., 1996. Hierarchical clustering of words. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING). pp. 1159–1162.