A symbolic approach to automatic multiword term structuring.

Eric Sanjuan, James Dowdall, Fidelia Ibekwe-Sanjuan, Fabio Rinaldi

A Symbolic Approach to Automatic MultiWord Term Structuring

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

[a] *LITA EA3097. University of Metz, France*
[b] *NLP Group, Dept. of Informatics, University of Sussex. UK*
[c] *ERSICOM, University of Lyon3, France*
[d] *Institute of Computational Linguistics, University of Zurich, Switzerland*

Abstract

This paper presents a three-level structuring of multiword terms (MWTs) 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 structures, 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.

Key words: Term variation, Automatic terminology structuring, Clustering technique, Topic maps, Knowledge discovery, Question Answering.

1 Introduction

Technical domains represent specialist knowledge gained through training or experience. Such specialization uses the foundation of general knowledge to build a level of expertise within a given domain necessitating an expansion in vocabulary to include domain specific objects and concepts. This results in the abundance of technical terms, realized linguistically as nominal compounds. Technical writing is an attractive domain in which to explore compounds for two reasons. First, it presents many examples [1] and secondly, it restricts semantic interpretation by excluding compounds with an idiomatic interpretation. This results in multiword terms which are both compositional, their formation is a function of their constituent elements (Kageura, 2002) and endocentric, the compound is a hyponym of its head (Barker

[1] 78.5% occurrences of simplex NPs in the GENIA corpus used in this study are terms (Kim et al., 2003).

and Szpakowicz, 1998). An extracted list of multi-word terms (MWTs) brings dramatic improvements for syntactic analysis but little else in solving the problems MWTs create for many natural language processing applications. To this end, a wealth of research has been directed toward identifying and organizing semantically related terms for different applications. The two families of approaches used for this task are distributional (statistical) and symbolic (Corpus-based linguistics) methods.

Distributional Similarity is taken as an indication of semantic similarity. The focus of many studies has been in relating single words statistically: (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. For instance, in Lin (1998), 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.

Corpus-based Linguistic analysis is used to identify linguistic markers which point to certain morphological, syntactic or semantic relations between MWTs (Morin and Jacquemin, 2003; Nenadic et al., 2004; Grabar and Zweigenbaum, 2004). The dominant methodology is shallow, bottom-up parsing around contextual clues like word insertion for syntactic variations or the use of phrases like "such as" or "also known as" for hypernym/hyponyms and synonyms identification respectively. For instance, in the sentence

"*In contrast to the purely enhancer-dependent effect of cytokines* **such as** *TNF on the activity of the HIV regulatory region (LTR), we observed that okadaic acid (OKA) activates HIV transcription through both the enhancer, responding to the factor NF-kappa B, and the promoter domain of the LTR*",

a hypernym relation will be acquired between "*purely enhancer-dependent effect of cytokines*" and "*TNF*". This approach has the advantage of computational tractability but is inherently limited to uncovering only relations explicitly identified through the targeted lexico-syntactic patterns. For example, (Morin and Jacquemin, 2003) 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%).

MWT Variation has been explored for a variety of applications such as building lexical resources from corpora (Daille, 2003; Hamon and Nazarenko, 2001; Grabar and Zweigenbaum, 2004), automatic thesaurus enrichment (Morin and Jacquemin, 2003), domain knowledge mapping and textmining (Ibekwe-SanJuan, 1998; Ibekwe-SanJuan and SanJuan, 2004), terminology knowledge base construction (Condamines and Reyberolle, 1998) and ontology building (Aussenac-Gilles and Séguéla, 2000). Syntactic variations involve basically three types of linguistic operations in a term: the addition of modifier or head words to an existing term, the substitution of a word in a term or the structural transformation of a term. The first type of operation is diversely called lexical inclusion (Nenadic et al., 2004; Grabar and Zweigenbaum, 2004), expansions (Jacquemin, 2001; Ibekwe-SanJuan, 1998) or modification (Daille, 2003). It denotes the simple fact that a term A is a subpart of term B ("*gene expres-*

sion → human beta globin gene expression"). The resulting term is a variant of the more generic term. Substitutions denote a change of a modifier or head word in a term ("*immature bone marrow cell ↔ murine bone marrow cell*"), while structural transformation (also called permutation) involves the passage from a syntagmatic structure with a PP attachment ("*system for database file transfer*") to a compound one ("*database file transfer system*").

Current research on computational terminology has reached the consensus that simple lists of terms extracted from corpora are not very useful for many applications. Indeed, it is very fastidious and quite inefficient to labor through lists of thousands of terms in a database or even to try to grasp the conceptual organization of terms in the domain if no synthesis of the information contained therein is offered. There are different ways in which this synthesis can be approached. The most classical and well known tools for organizing the conceptual structures of a field are thesauri, taxonomies and ontologies. 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 (Aussenac-Gilles and Séguéla, 2000; Biébow and Szulman, 1999). Also, these resources cannot synthesize the information contained in a huge corpus because every term is listed, albeit in a hierarchy. 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 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.

The need to structure domain concepts is even more acute for applications like scientific and technological watch or textmining where experts are required to grasp the topic emergence, shifts and obsolescence in limited time. Research on methods to this end, known as domain knowledge mapping (DKM) rely on powerful visualization tools for result presentation. While a lot of research has been carried out separately on computational terminology (see Jacquemin and Bourigault (2003) for a review) and on DKM (see Schiffrin and Börner (2004) for a review), very few attempts have, to our knowledge, been made to bring the two domains together. Research on DKM has always relied heavily on statistical models (co-occurrence models) to build clusters of frequently co-occurring item sets (Mane and Börner, 2004; Hearst, 1999; Small, 1999; Feldman et al., 1998). The challenge in our approach lies in combining symbolic representations (variations) and a data clustering algorithm. Parts of this methodology have been published in (Ibekwe-SanJuan, 1998; Ibekwe-SanJuan and SanJuan, 2004). Here we aim to test it on a technical corpus, the GENIA corpus and to compare the clusters against a gold standard, the hand built GENIA ontology. The idea is to evaluate how close the clusters come

to reflecting a human semantic organization of domain concepts. The outcome of such an evaluation will determine if the methodology has uses for other knowledge organization tasks such as terminology knowledge base or ontology population. We also explore the possibilities of using such a structuring in a Question Answering (Q-A) system focused on technical domains.

The rest of the paper is organized as follows: after an overview of the methodology (Section 1.1), we briefly describe the corpus used in this experiment and the normalization of the MWTs contained therein (Section 1.2). Section 2 describes the three-level structuring of the MWTs. Section 3 evaluates the similarity of the automatic structuring against the hand-built GENIA ontology. Section 4 is devoted to discussions on the potentials of the term variant clustering for Q-A.

1.1 Overview of the methodology

First, MWTs are extracted from a corpus (see Section 1.2) before subjecting them to a three-level structuring. Normally, MWT extraction is performed in our system using the LTPOS and LT_CHUNK package developed by the University of Edinburgh. LTPOS is a probabilistic part-of-speech tagger based on Hidden Markov Models. It uses the Penn Treebank tag set which ensures the portability of the output with many other systems. Since LT_CHUNK only identifies simplex NPs without prepositional attachments, we wrote contextual rules [2] to identify complex terminological NPs. In the current experiment however, the result of our extraction module was not used in further processing since the biological terms were already manually annotated in the corpus (see Section 1.2). Our main objective in this experiment is to evaluate our MWTs structuring against the GENIA ontology. Hence, it was necessary to adopt the same term list. Extracting the terms ourselves would have distorted the comparison of the two structurings and weakened its conclusions. On a more lexical note, using the MWTs annotated by domain experts creates the ideal situation for further processing as the MWTs are not influenced by a particular extraction technique. Furthermore, BioMedical Entity Recognition for which the GENIA corpus provides an interesting annotated resource, is a challenging task receiving ongoing attention.

The first level structuring consists in establishing binary "term-term" relations using the variation relations. Basing on these 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. Before this grouping is performed, noisy relations are filtered out using WordNet. Components thus obtained are sets of terms formed around a particular domain paradigm or a monothematic family (see examples below). This constitutes the second-level of structuring. The components are grouped into clusters 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 [3] graphic display package. This 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 variation relations. A

[2] Based on the POS information surrounding an NP structure

[3] www.aisee.com

detailed description is given in Section 2.3.1. Below is an example of a cluster 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 cluster is suggesting is that research word 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.*)

1.2 Normalizing the MWTs from the GENIA corpus

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, according to its authors, as 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 MeSH headings "*human*", "*blood cell*" and "*transcription factor*", the titles and abstracts of 2 000 articles[4] were collected 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 an embryon of a humanly constructed ontology, the GENIA ontology (see Figure 3). This 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>
```

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

Notice that the underlying XML markup of the terms 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-case form of 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. For instance, "*Ca(2+)-dependent_pathway*" becomes "*Ca(2+) dependent pathway*". This is an ad-hoc processing which will have to be adapted to each corpus, especially in technical domains where variation phenomena are frequent.

2 Structuring MultiWord Terms

We describe in detail the types of variations used to relate the MWTs (Section 2.1), the filtering process performed for certain variation types (Section 2.2). These variations then serve as basis for the three-level structuring effected on MWTs in order to map them into domain topics (Section 2.3).

2.1 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 [5], syntactic variations yield conceptual relations between terms. This hypothesis is current in computational terminology studies. 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 (lexical inclusion) 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

[5] The position, the morphological category and the grammatical role of inserted words in a term variant

addition of adjacent nominal elements (nouns, adjectives). This lessens the possibility of relating as variants, terms which are semantically distant.

Substitutions also subdivide into two types, modifier substitution (M-Sub) and head substitution (H-Sub) and 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 :

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

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 this creates across the document collection.

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

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

86% (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.

2.2 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 "*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 (Ins-1) modifier and left-expansions are used to create `IS_A` hierarchies around concept families. This permits a MWT to have more than one parent (see Figure 1).

These observations suggest that among the variations that do not change the head word, left-expansions (L-Exp) and (Ins-1) should be given priority for building components (2nd level structuring) if we want to obtain homogeneous clusters vis-à-vis the Genia ontology.

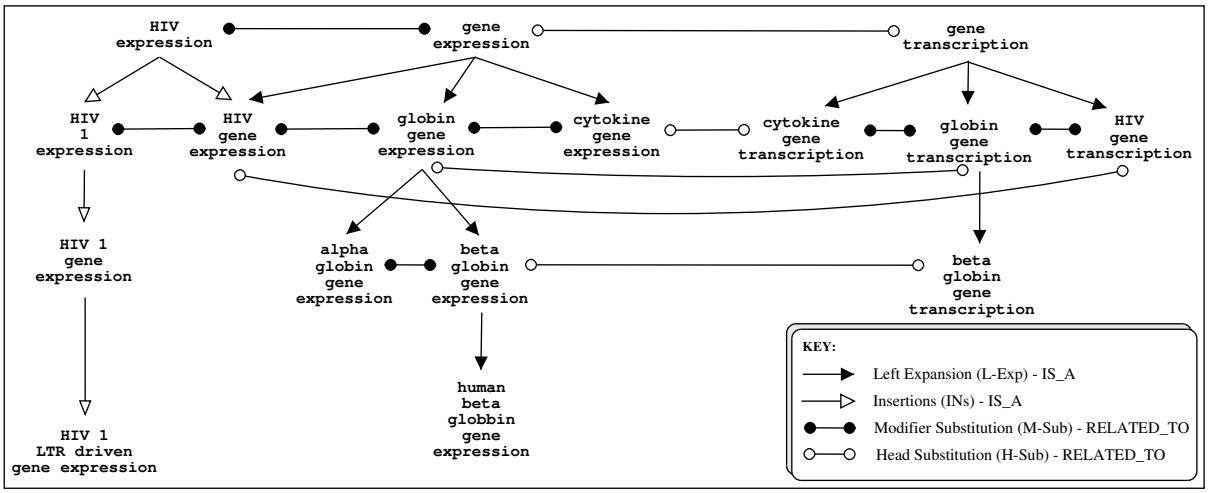


Fig. 1. Fragment of the conceptual hierarchy

Substitutions engender transversal 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, 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 Figure 1, "*gene expression*" and "*gene transcription*" are head substitutions and the conceptual link is close: the "*expression*" of a "*gene*" is the result of its "*transcription*". However, the same variation also links "*gene*" as it 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).

WordNet Substitutions (WN-Sub) are those in which the substituted words belong to the same synset. These variations follow the head/modifier distinction but, unlike syntactic substitutions, both the head and modifier can be substituted between two MWTs as in *hormone effect* and *endocrine event*.

WordNet	Example MWT 1	Example MWT 2
M-Sub	***strong*** *transcriptional repressor*	***potent*** *transcriptional repressor*
H-Sub	*inflammatory* ***reaction***	*inflammatory* ***response***
HM-Sub	***hormone effect***	***endocrine event***

Table 2. Semantic variations identified through WordNet

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 they produce in a specialized field can differ from the generic one suggested by a general language re-

source. For instance, within the genomic domain, "***strong***" refers to the degree to which a "*repressor*" 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 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. Of course, such high precision score implies very low recall. Only 304 links were present in WordNet among the 548 359 possible substitutions found in the corpus. This low recall score does not seem to be a major drawback in our approach. On the contrary, we will see later in Section 2.3.2 that we need to severely restrict the set of substitutions in other to avoid the chain effect, well known in clustering approaches that compute connected components.

2.3 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 2.3.1) and its application to the GENIA MWTs (Section 2.3.2).

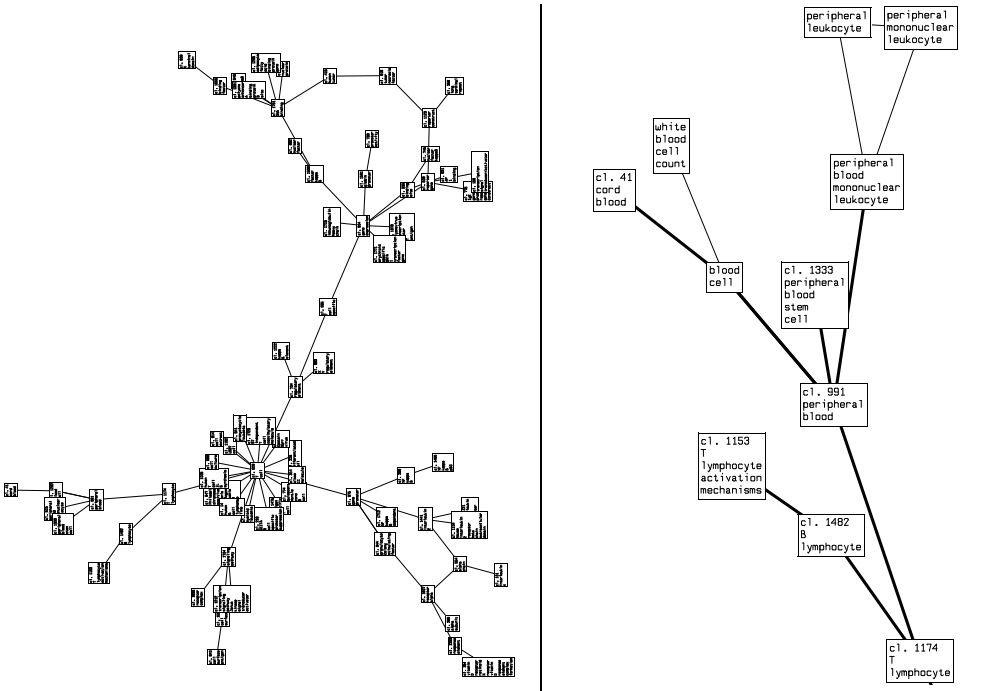
2.3.1 Term variant clustering

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) but observations made in Section 2.2 induced the choice of only constrained expansions and WordNet Substitutions (see Section 2.3.2 for more details) in the present experiment. We call these COMP relations.

The transitive closure COMP* of 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. More details can be found in (Ibekwe-SanJuan and SanJuan, 2004).



a) Central Network of GENIA clusters | b) Clusters related to "blood cell"

Fig. 2. Graphs of Term Variant clusters displayed with AiSee

2.3.2 Clustering the GENIA term variants

Preliminary clustering tests and the observations made in Section 2.2 led us to modify the roles usually assigned to the syntactic variations in the TermWatch system for the Genia corpus. Following observations in this section, we further split L-Exp into two sub-relations: strong-L-Exp and weak-L-Exp according to 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 ($\geq$ 2 000 terms) on this corpus.

Empirical tests showed the clusters produced at the 2nd iteration of the algorithm to be the most legible in terms of size and content. This produced 1 664 clusters, 6 151 components and a total of 10 285 MWTs in the clusters. 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 clusters, the user can temporarily hide very weak links between them. This gives Figure 2(a). Each cluster is labelled automatically by the term that shares the most number of variation links outside the cluster. The global image obtained exhibits a star shape with a central core, related to a cyclic subgraph (see Figure 2(a) that shows the structure of this graph). By order of importance, the central position is occupied by a big cluster labeled *"T-Cell"* with 374 terms. A second smaller sub network is formed around the cluster labelled *"gene expression"* with

235 terms.

Each cluster can be unfolded to show its internal structure: the connected components, the most active term variants. The user can thus immediately perceive the most salient features of a cluster. In Figure 2 (b), we unfolded the cluster "*blood cell*" to show an internal link between the main component "*blood cell*" (that gave its name to the cluster) and other related topics like "*white blood cell count*", "*cord blood*". This kind of interactive manipulations using the AiSee interface allows the user to access simultaneously the three levels of the clustering results: clusters, components and terms. The length of an arc has a straightforward meaning here. The higher is the number of variants between two clusters, the shorter is the arc between them. The sub-network labelled "*T lymphocyte*", "*B lymphocyte*", "*T lymphocyte activation mechanism*" forms a linear graph, that is 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 naturally aligns the elements of these linear graphs, thus highlighting them.

The two clusters "*T lymphocyte*" and "*B lymphocyte*" contain respectively terms like "*activated T lymphocyte*", "*human peripheral lymphocyte*", "*activated peripheral blood lymphocyte*", "*B lymphocyte specific mb 1 gene*", "*normal B lymphocyte*" and "*B lymphocyte growth transformation*". Their link with the clusters 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*" [6]. TermWatch thus seems to have effected coherent thematic associations in the domain via syntactic variations and few WN-Subs found in the corpus. We will now examine to what extent the clusters are consistent with the hand-built GENIA ontology.

3 Evaluation of the clusters 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 produce a sort of hierarchy (the inclusion of one cluster 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 interpretations hold a strategic dimension, [7] 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 are 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

[6] Concise Oxford Dictionary, Allen R.E. (eds.). 8th Edition, 708-709

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

concept is not too far from the nodes considered.

Assumption 2: TermWatch's clusters supposes a weaker "semantic proximity" between terms in the same component and by extension, in the same cluster.

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 cluster in the TermWatch output ?

For that purpose, let us call *atomic category* the categories at the leaves of the GENIA ontology that are different from "other_name". Then we map the set of clusters into the GENIA ontology by associating each cluster with its dominant atomic category, i.e. the atomic category that has the highest number of terms in the cluster.

By way of example, component 363 has five terms. Four of them: *"NF kappaB", "lung NF kappaB", "mammalian NF kappaB", "nuclear NF kappaB"* are in the "protein_complex" category, and only the fifth one: "*cytoplasmic NF kappaB*" comes from a different category: "protein_molecule". Since four terms out of five in this component belong to category "protein_complex" in the GENIA ontology, this is the category 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 cluster 646 that has the same label "NF kappaB" but is not associated to the same dominant GENIA group as shown in Table 3.

Indeed, this table shows the categories associated with the nine clusters having more than 50 terms in categories different from "other_name". The numbers and labels of the clusters are given in the fourth and the fifth columns. The associated category (the dominant one) is given in the last column. The first column "Nb_G" shows the number of terms in the cluster that share the dominant category, the second column "Nb_C" shows the total number of terms in the cluster and the third column gives the ratio between the previous two numbers.

NB_G	NB_C	rate	cluster	label	category
32	81	0.39	646	NF kappaB	protein_molecule
30	67	0.44	1700	mouse gene	DNA_domain_or_region
43	96	0.44	1791	DNA binding	protein_family_or_group
31	54	0.57	1260	response element	DNA_domain_or_region
218	364	0.59	628	cell_line	T-cell
47	72	0.65	1561	E-Box	DNA_domain_or_region
42	63	0.66	618	human enhancer	DNA_domain_or_region
78	111	0.70	336	binding site	DNA_domain_or_region
45	61	0.73	808	N-terminal domain	protein_domain_or_region

Table 3. GENIA categories associated with the biggest clusters

Hence, table 3 shows that the biggest clusters produced by TermWatch have more than 40% of their terms in the same GENIA category, except for cluster 646. 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 cluster is an error with regard to the GENIA ontology. Analyzing cluster 646 whose dominant GENIA category ("protein_molecule") represents only 39% of its terms, we find out that all the GENIA categories of terms in this cluster 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 for to the majority of the components and clusters.

First, we compute the number of components and clusters 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 occurrence in the GENIA corpus.
- the distributions d_{comp} and d_{class} of dominant categories in components and in clusters 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 occurrence in the GENIA corpus having the category "c"= "protein_molecule" which is 15 348 in this case, divided by the total number of occurrences.

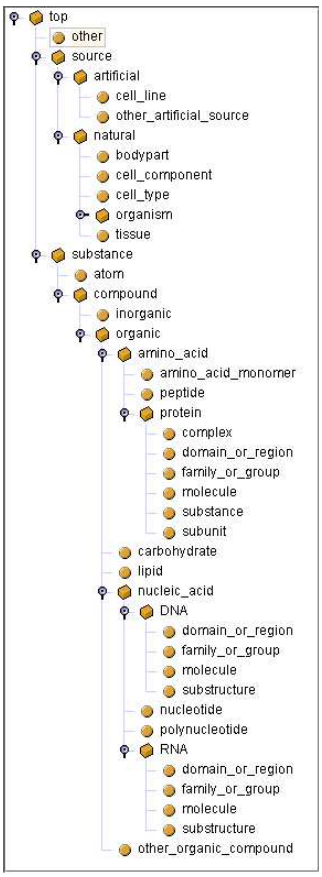
$d_{comp}(c)$ is the number of components in TermWatch output, of which the majority of the terms are in category c (4 493 in this case), divided by the total number of components.

$d_{class}(c)$ is the same as $d_{comp}(c)$ except that we consider clusters instead of components. 778 clusters are associated with "protein_molecule".

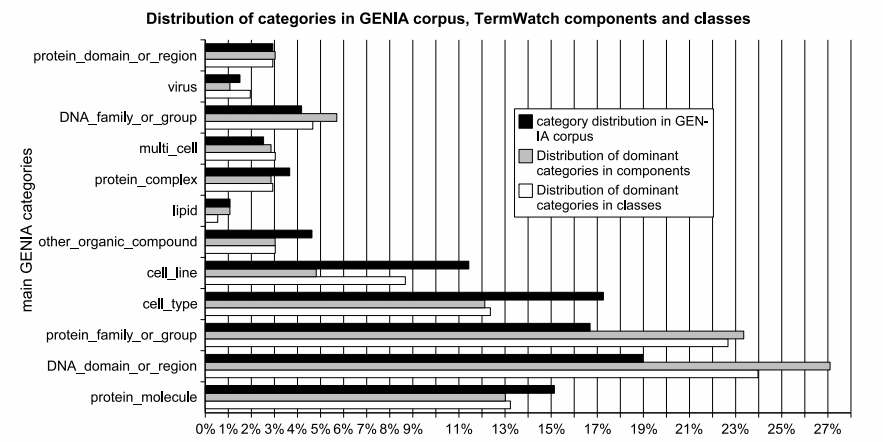
The right topmost graphic in Figure 3 ("Distribution of categories in GENIA corpus, TermWatch components and classes") allows us to compare the 12 topmost values of d_G (represented by the upper black bars) with the corresponding values of d_{comp} and d_{class} respectively represented by the middle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

This figure shows that clusters, more than components, lessens the deviation from the distribution of GENIA categories in the corpus (except for the small category "lipid". In fact, $d_{comp}(c)$ 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. This is more suited to evaluating the clustering performance of a system when a reference classification is known. We define formally the precision and recall functions we computed.



GENIA ontology



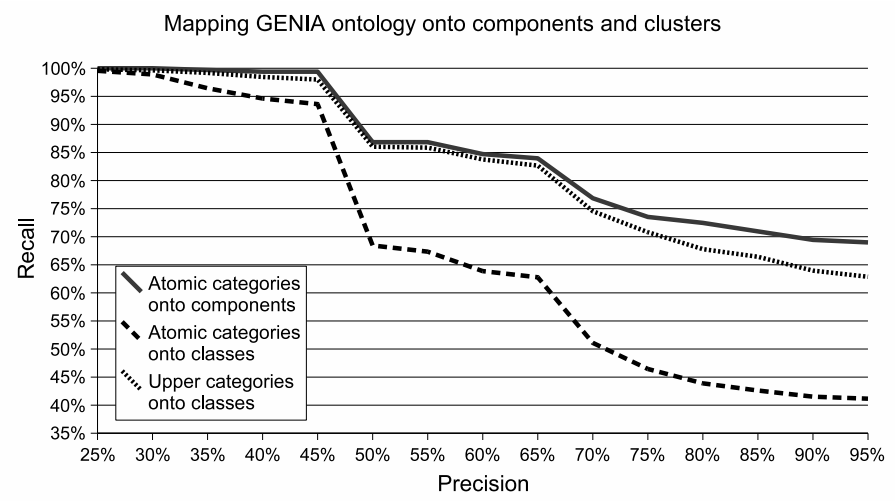


Fig. 3. Mapping GENIA categories onto TermWatch clusters.

We identify each GENIA category G with the set of included terms. Let $\mathcal{G}$ be a family of GENIA categories and let $\mathcal{X}$ be a family of clusters (components or classes). Using these notations we clearly have the equality: $|G_X \cap X| = \max\{|G \cap X| : G \in \mathcal{G}\}$.

Precision p can be defined for any a cluster 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:

$$r(v) = \frac{|\{X \in \mathcal{X} : p(X) \geq v\}|}{|\mathcal{X}|}$$

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 following 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 Figure 3 shows three precision/recall functions computed using different families of clusters $\mathcal{X}$ and different families $\mathcal{G}$ of categories. The uppermost bold line curve shows the function obtained by setting $\mathcal{X}$ to the whole set of components, and $\mathcal{G}$ 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, 48% of the components have terms from the same GENIA category, thus a 100% inclusion in one semantic type, while 95% of the components still attain 70% inclusion in one category. 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.

Clusters on the other hand group several components, thus variants with different heads. The lowest dashed curve shows the precision/recall function by setting $\mathcal{X}$ to the clusters and naturally, the semantic inclusion in one category is much lower than for the components. Still a comparable proportion of classes (46%) have a 100% semantic inclusion in one GENIA category but as we consider more classes, this figure drops. When we consider practically all the clusters (95%), only 41% (683 clusters) of them show 100% semantic inclusion in one GENIA category.

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 $\mathcal{G}$ 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 clusters. We observed that the semantic inclusion of the clusters increased and moved closer to the distribution of the ontology categories in the components. This is represented by the middle dotted curve on the Figure 3.

Hence the evaluation of the clusters against the GENIA ontology showed the contents to be thematically coherent. 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 clusters, while not being monolithic in terms of semantic class still group together coherent domain topics which are logically associated,
- that finally, whilst not targeting specifically the construction of a taxonomy or an ontology, the map of domain topics generated reflects this structure to a certain extent and thus offers a graphic and synthetic way of exploring a domain's knowledge structures.

4 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: [8]

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

(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 cluster (see Figure 2). By presenting the graph of clusters, 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 clusters 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 & Jacquemin (2003) 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.

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 [9]. It takes as input the number of co-occurrence of terms in GENIA corpus. We also computed the resulting precision/recall functions as in Figure 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

[9] FASTCLUST and CLUSTER procedures in SAS system for Windows 'V8 SAS Institute Inc., Cary, NC, USA.)

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 TermWatch'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.

Barker, K., Szpakowicz, S., August 10-14 1998. Semi-Automatic Recognition of Noun Modifier Relationships. In: Proc. of COLING-ACL98. Montreal, Quebec, Canada.

Biébow, B., Szulman, S., 1999. Terminae : A linguistics-based tool for building of a domain ontology. In: Fensel, D., Studer, R. (Eds.), proc. 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., August 1998. Ctkb : A corpus-based approach to a terminological knowledge base. In: Proceedings 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., 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: Proc. COLING'92. Nantes, pp. 539–545.

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

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'Information assistée par ordinateur (RIAO). Avignon, pp. 26–28.

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.

Kageura, K., 2002. The dynamics of Terminology: A descriptive theory of term formation and terminological growth. John Benjamins, Amsterdam.

Kim, J.-D., Ohta, T., Tateisi, Y., Tsujii, J., 2003. Genia corpus - a semanticall annotated corpus for bio-textmining. Bioinformatics 19 (1), i180–1182.

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., 2003. 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 ACL 2004 workshop on Question Answering in Restricted Domains. Accepted for publication.

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.