Exploring graph structure for detection of reliability zones within synonym resources: Experiment with the *Gene Ontology*

Thierry Hamon
LIPN – UMR 7030
Université Paris 13 – CNRS
99 av. J-B Clément
F-93430 Villetaneuse, France
thierry.hamon@lipn.univ-paris13.fr

Natalia Grabar
Centre de Recherche des Cordeliers
Université Paris Descartes, UMR_S 872
INSERM, U872
HEGP AP-HP, 20 rue Leblanc
Paris, France
natalia.grabar@spim.jussieu.fr

Abstract

Computing the semantic similarity between terms relies on existence and usage of semantic resources. However, these resources, often composed of equivalent units, or synonyms, must be first analyzed and weighted in order to define within them the reliability zones where the semantic cohesiveness is stronger. We propose an original method for acquisition of elementary synonyms based on exploitation of structured terminologies, analysis of syntactic structure of complex (multi-unit) terms and their compositionality. The acquired synonyms are then profiled thanks to endogenous lexical and linguistic indicators (other types of relations, lexical inclusions, productivity), which are automatically inferred within the same terminologies. Additionally, synonymy relations are observed within graph, and its structure is analyzed. Particularly, we explore the usefulness of the graph theory notions such as connected component, clique, density, bridge, articulation vertex, and centrality of vertices.

1 Introduction

In various tasks and applications of natural language processing and of biomedical informatics (*i.e.*, query expansions, information retrieval, text mining, information extraction or terminology matching), it is important to be able to decide whether two terms (*i.e.*, *acetone anabolism* and *acetone biosynthesis*, *replication of mitochondrial DNA* and *mtDNA replication*) convey the same or different meaning. This is particularly important for deciphering and computing semantic similarity between words and terms. Lexicon of specific resources (synonym, morphological or orthographic variants) can be used for detection of semantic similarity. However, depending on languages and domains, such resources are not equally well described. Morphological description is the most complete for both general (Burnage, 1990; Hathout et al., 2001) and biomedical (NLM, 2007; Schulz et al., 1999; Zweigenbaum et al., 2003) languages. But the situation is not as successful at the semantic level: little synonym resources can be found. If WordNet (Fellbaum, 1998) proposes general language synonym relations for English, the corresponding resources for other languages are not freely available. Moreover, the initiative for fitting WordNet to the biomedical area (Smith and Fellbaum, 2004) seems to have been abandoned, although there is a huge need for this kind of resources.

In our previous work, we proposed to use the existing biomedical terminologies (*i.e.*, *Gene Ontology* (Gene Ontology Consortium, 2001), Snomed (Côté et al., 1997), UMLS (NLM, 2007)), wich provide complex terms, and to acquire from them lexical resources of synonyms. Indeed, the use of complex biomedical terms seems to be less suitable and generalizable as compared to lexical resources (Poprat et al., 2008). Within the biological area, we proposed to exploit the *Gene Ontology* (*GO*), and more specifically to exploit compositional structure of its terms (Hamon and Grabar, 2008). However, with the acquisition of synonymy we faced two problems: (1) contextual character of these relations (Cruse, 1986), *i.e.*, two terms or words are considered as synonyms if they can occur within the

same context, which makes this relation more or less broad depending on the usage; (2) ability of automatic tools to detect and characterize these relations, *i.e.*, two terms or words taken out of their context can convey different relations than the one expected. Because we aim at acquiring synonymy resources which could be used by various applications and on various corpora, we need to profile them and possibly to detect the reliability zones. We proposed to do this profiling through lexical and linguistic indicators generated within the same terminology (Grabar et al., 2008), such as productivity, cooccurence with other types of relations (`is-a`, `part-of`) and with lexical inclusion. These indicators on reliability zones will be used for defining the synonymity degree of terms and for preparing the validation of the acquired synonym resources. In the current work, we continue profiling the acquired synonyms, but rely on the form of the graph built from pairs of synonyms. We exploit for this some notions of the graph theory (Diestel, 2005). In the following of this paper, we first present our material (sec. 2) and methods (sec. 3), we then present and discuss results (sec. 4) and conclude with some perspectives (sec. 5).

2 Material

We use the *Gene Ontology* (*GO*) as the original resource from which synonym lexicon (or elementary synonym relations) are induced. The goal of the *GO* is to produce a structured, common, controlled vocabulary for describing the roles of genes and their products in any organism. *GO* terms convey three types of biological meanings: biological processes, molecular functions and cellular components. Terms are structured through four types of relationships: subsumption `is-a`, meronymy `part-of`, synonymy and `regulates`. The version, we used in the current work, was downloaded in February 2008[1]. It provides 26,057 concepts and their 79,994 terms. When we create pairs of terms, which we exploit with our methods, we obtain 260,399 `is-a`, 29,573 `part-of` and 459,834 synonymy relations. There are very few `regulates` relations, therefore we don't exploit them in our work.

[1]Our previous work has been performed with an anterior version of the *GO*.

3 Methods

GO terms present compositional structure, like within the concept GO:0009073, where compositionality can be observed through the substitution of one of the components (underlined):

aromatic amino acid family biosynthesis
aromatic amino acid family anabolism
aromatic amino acid family formation
aromatic amino acid family synthesis

Compositionality of the *GO* terms has been exploited previously, for instance (Verspoor et al., 2003) propose to derive simple graphs from relations between complex *GO* terms, (Mungall, 2004) exploits the compositionality as a mean for consistency checking of the *GO*, (Ogren et al., 2005) use it for enriching the *GO* with missing synonym terms. We propose to exploit the compositionality for induction of synonym lexical resources (*i.e.*, *biosynthesis*, *anabolism*, *formation*, *synthesis* in the given example). While the cited works are based on the string matching within *GO* terms, our approach aims at exploiting the syntactic analysis of terms, which makes it independent from the graphical form of the analyzed terms (like examples on fig. 1). Our method has several steps: linguistic preprocessing of the *GO* terms (sec. 3.1), induction of elementary semantic lexicon (sec. 3.2), and then the profiling the synonymy lexicon through the lexical and linguistic indicators (sec. 3.3), and through the analysis of connected components built from the induced synonym pairs (sec. 3.4). Steps 3.1 to 3.3 have been already described in our previous work: we mention here the main notions for the sake of clarity.

3.1 Preprocessing the *GO* terms: Ogmios NLP platform

The aim of terminology preprocessing step is to provide syntactic analysis of terms for computing their syntactic dependency relations. We use the `Ogmios` platform[2] and perform: segmentation into words and sentences; POS-tagging and lemmatization (Schmid, 1994); and syntactic analysis[3]. Syntactic dependencies between term components are

[2]http://search.cpan.org/∼thhamon/Alvis-NLPPlatform/
[3]http://search.cpan.org/∼thhamon/Lingua-YaTeA/

Figure 1: Parsing tree of the terms *replication of mitochondrial DNA* and *mtDNA replication*.

computed according to assigned POS tags and shallow parsing rules. Each term is considered as a syntactic binary tree composed of two elements: head component and expansion component. For instance, *replication* is the head component of the two terms analyzed on figure 1.

3.2 Acquiring the elementary semantic relations

The notion of compositionality assumes that the meaning of a complex expression is fully determined by its syntactic structure, the meaning of its parts and the composition function (Partee, 1984). On the basis of syntactically analysed terms, we apply a set of compositional rules: if the meaning $\mathcal{M}$ of two complex terms $A\ rel\ B$ and $A'\ rel\ B$, where A is its head and B its expansion components, is given as following:

$$\mathcal{M}(A\ rel\ B) = f(\mathcal{M}(A), \mathcal{M}(B), \mathcal{M}(rel))$$

$$\mathcal{M}(A'\ rel\ B) = f(\mathcal{M}(A'), \mathcal{M}(B), \mathcal{M}(rel))$$

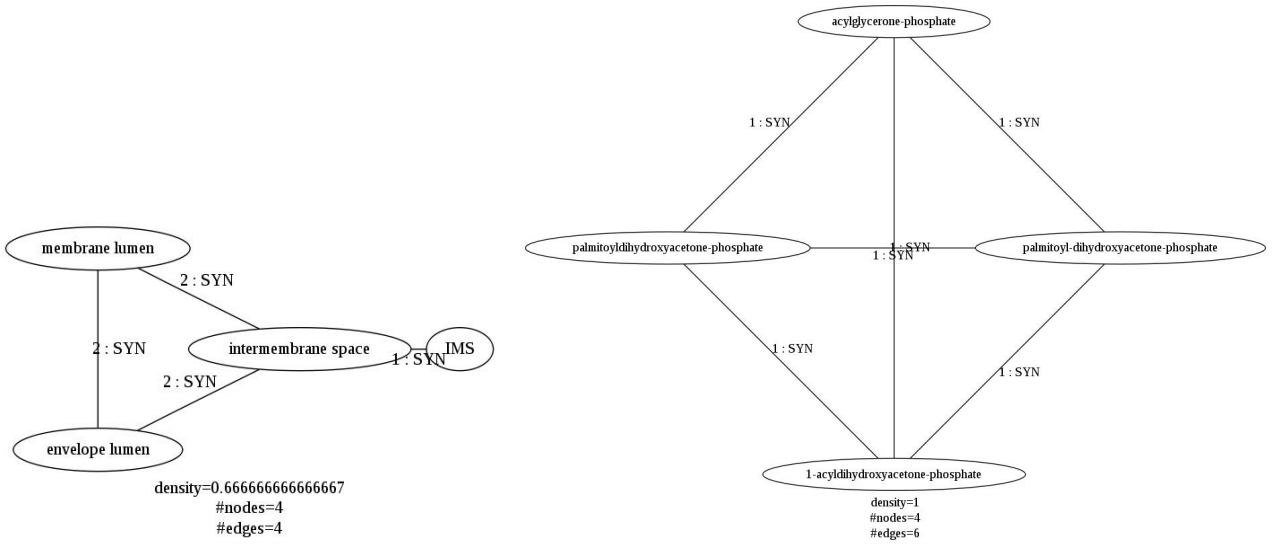
for a given composition function f, if $A\ rel\ B$ and $A'\ rel\ B$ are complex synonym terms and if B components are identical (such as *acetone* within *acetone catabolism* and *acetone breakdown*), then the synonymy relation between components A and A' {*catabolism*, *breakdown*} can be induced. The modification is also accepted on expansion component B: from terms *replication of mitochondrial DNA* and *mtDNA replication* (fig. 1), we can induce synonymy between *mitochondrial DNA* and *mtDNA*. Finally, the modification is also accepted for both components $A\ rel\ B$ and $A'\ rel\ B'$, such as in *nicotinamide adenine dinucleotide catabolism* and *NAD breakdown*, where one pair, *i.e.* {*catabolism*, *breakdown*}, can be known from previously processed synonyms and allow to induce the new pair {*nicotinamide adenine dinucleotide*, *NAD*}. The method is recursive and each induced elementary synonym relation can then be propagated in order to induce new elementary relations, which allows to generate a more exhaustive lexicon of synonyms.

This method is not specific to the synonymy. As it works at the syntactic level of terms, it therefore can be applied to other relationships: relationship between elementary terms is inherited from the relationship between complex terms. If we exploit complex terms related with `part-of` relations and if the compositionality rules can be applied, then we can induce elementary `part-of` relations. For instance, complex terms *cerebral cortex development* GO:0021987 and *cerebral cortex regionalization* GO:0021796 have a `part-of` relation between them, and we can induce the elementary `part-of` relation between their components *development* and *regionalization*. Similarly, on the basis of two *GO* terms that have `is-a` relation between them, *cell activation* GO:0001775 and *astrocyte activation* GO:0048143, we can induce the elementary `is-a` relation between *cell* and *astrocyte*.

3.3 Exploiting lexical and linguistic indicators

Several endogenously generated indicators are used for profiling the induced lexicon of synonyms:

- Elementary `is-a` relations;
- Elementary `part-of` relations;
- Lexical inclusion: terms within each induced synonymy pair are controlled for the lexical inclusion. If the test is positive, like in {*DNA binding*, *binding*}, this would suggest that the analyzed terms may convey a hierarchical relation: indeed, lexical subsumption marks often a hierarchical subsumption (Kleiber and Tamba, 1990), which can be either `is-a` or `part-of` relations;
- Productivity: number of original *GO* pairs from which this elementary relation is inferred. For instance, synonymy relations {*binding*, *DNA*



(a) Connected component of synonyms

(b) Clique of synonyms

Figure 2: Connected components formed with pairs of elementary synonym relations.

binding$\}$ and $\{$*cell*, *lymphocyte*$\}$ are inferred from only one original pair of *GO* synonyms, while the pair $\{$*T-cell*, *T-lymphocyte*$\}$ is supported by eight original *GO* synonym pairs.

Factors that would weaken synonymy relations and make them less reliable are their co-occurrence with lexical inclusions, `is-a` or `part-of` relations, and their low productivity.

3.4 Exploiting the graph theory notions

Pairs of induced synonyms are observed through the connected components they form: lexical entries are *nodes or vertices* and relations between them are *edges or paths*. For instance, connected component 2(a) contains four pairs of synonyms: $\{$*membrane lumen*, *envelope lumen*$\}$, $\{$*membrane lumen*, *intermembrane space*$\}$, $\{$*envelope lumen*, *intermembrane space*$\}$ and $\{$*intermembrane space*, *IMS*$\}$. On each edge, we projected information associated with the relation corresponding to this edge. For instance, $\{$*membrane lumen*, *intermembrane space*$\}$ relation is labelled as synonymy SYN and shows 2 as productivity value (it has been acquired from two original pairs of synonyms within *GO*). If other relationships ($INCL$, PAR, $HIER$) are associated to a given synonymy relation, they are also indicated together with their productivity.

As a matter of fact, figure 2 presents two typical examples of connected components we can obtain (in these examples, both of them have four nodes):

- *Connected component* (fig. 2(a)) is a graph in which any two vertices are connected to each other by edges. Connected components have not orphan vertices, which would remain not connected to any other vertex.

- *Clique*, also called *block* (fig. 2(b)) is a particular case of connected components: clique is a maximally connected component. In such graphs, all the vertices are interconnected between them.

We propose to exploit four more notions of the graph theory, which we assume can be useful for further profiling of the acquired synonymy relations:

- *Density* of a connected component is the ratio between the number of its edges and the number of edges of the corresponding clique. For instance, the connected component on figure 2(a) has 4 edges while the corresponding clique would have 6 edges. In that respect, this connected component has the dentisty of 0.67. Besides, the clique on figure 2(b) shows the maximum density (*i.e.*, 1). (For all the fig-

ures, we indicate their density, together with the number of vertices and edges).

- *Bridge* is defined as an edge which removal would increase the number of connected components. For instance, within connected component 2(a), removing the edge {*intermembrane space*, *IMS*} would lead to the creation of two new connected components: (1) single-vertex component *IMS*, and (2) connected component with three vertices *intermembrane space*, *membrane lumen* and *envelope lumen*. Consequently *articulation vertices* are defined as vertices which removal would increase the number of connected components. At figure 2(a), the articulation vertex is *intermembrane space*.

- The *centrality of a vertex* is defined as the number of shortest paths passing through it. For instance, on figure 2(a), *intermembrane space*'s centrality is 4, while the centrality of other vertices is null.

4 Results and Discussion

4.1 Acquiring the elementary synonymy relations and their lexical and linguistic profiling

79 994 *GO* terms have been fully analyzed through the Ogmios platform. Compositional rules (sec. 3.2) have been applied and allowed to induce 9,085 semantic relations among which: 3,019 synonyms, 3,243 `is-a` and 1,205 `part-of`. 876 lexical inclusions have discovered within all these elementary pairs. 2,533 synonymy pairs are free of the lexical profiling indicators. However, 486 synonymy relations (16%) cooccur with other relations, and the details of this cooccurrence is showed in table 1. We can observe for instance that 142 synonym pairs are also labelled as `is-a` relations, and 34 as `part-of` relations. Productivity of the induced synonyms is between 1 and 422 original complex *GO* terms.

Connected component on figure 3 illustrates coocurrence of synonymy relations with other types of relations: the pair {*import*, *ion import*} shows synonym and inclusion relations; the pair {*import*, *uptake*} shows synonym and hierarchical relations, both acquired on seven original pairs of *GO* terms.

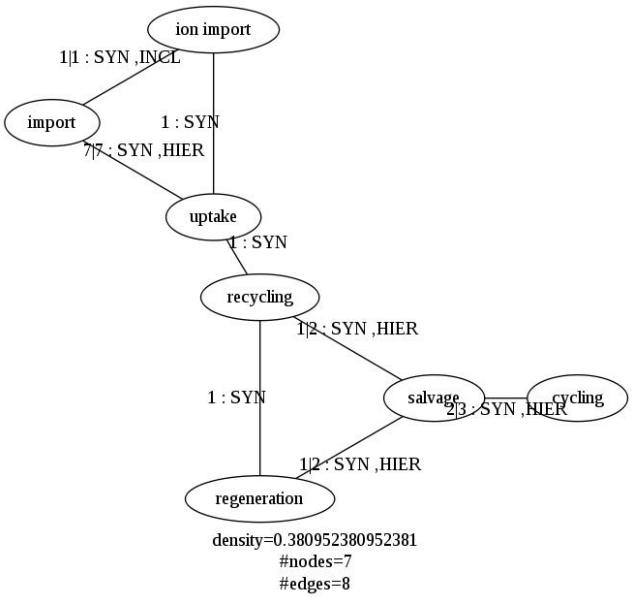


Figure 3: Connected component where synonymy relations cooccur with other relations.

Synonymy and other relations	Number
syno ∩ is-a	142
syno ∩ par	34
syno ∩ incl	309
syno ∩ par ∩ is-a	14
syno ∩ incl ∩ is-a \ par	40
syno ∩ incl ∩ par \ is-a	2
syno ∩ incl ∩ is-a ∩ par	1

Table 1: Number of synonymy relations which cooccur with other relations (`is-a`, `part-of` and lexical inclusions `incl`).

4.2 Analysing the induced synonym pairs through the graph theory

3,019 induced synonym pairs have been grouped into 1,018 connected components. These components contain 2 to 69 nodes, related among them by 1 to 132 edges. Analyses of the connected components have been performed with Perl package `Graph` and additionnal Perl scripts. Among the studied connected components, we have 914 cliques composed of 2 (n=708), 3 (n=66), 4 (n=88), 5 (n=44) or 6 (n=8) nodes. The remaining 104 connected components are less dense with edges. The density of the connected components is between

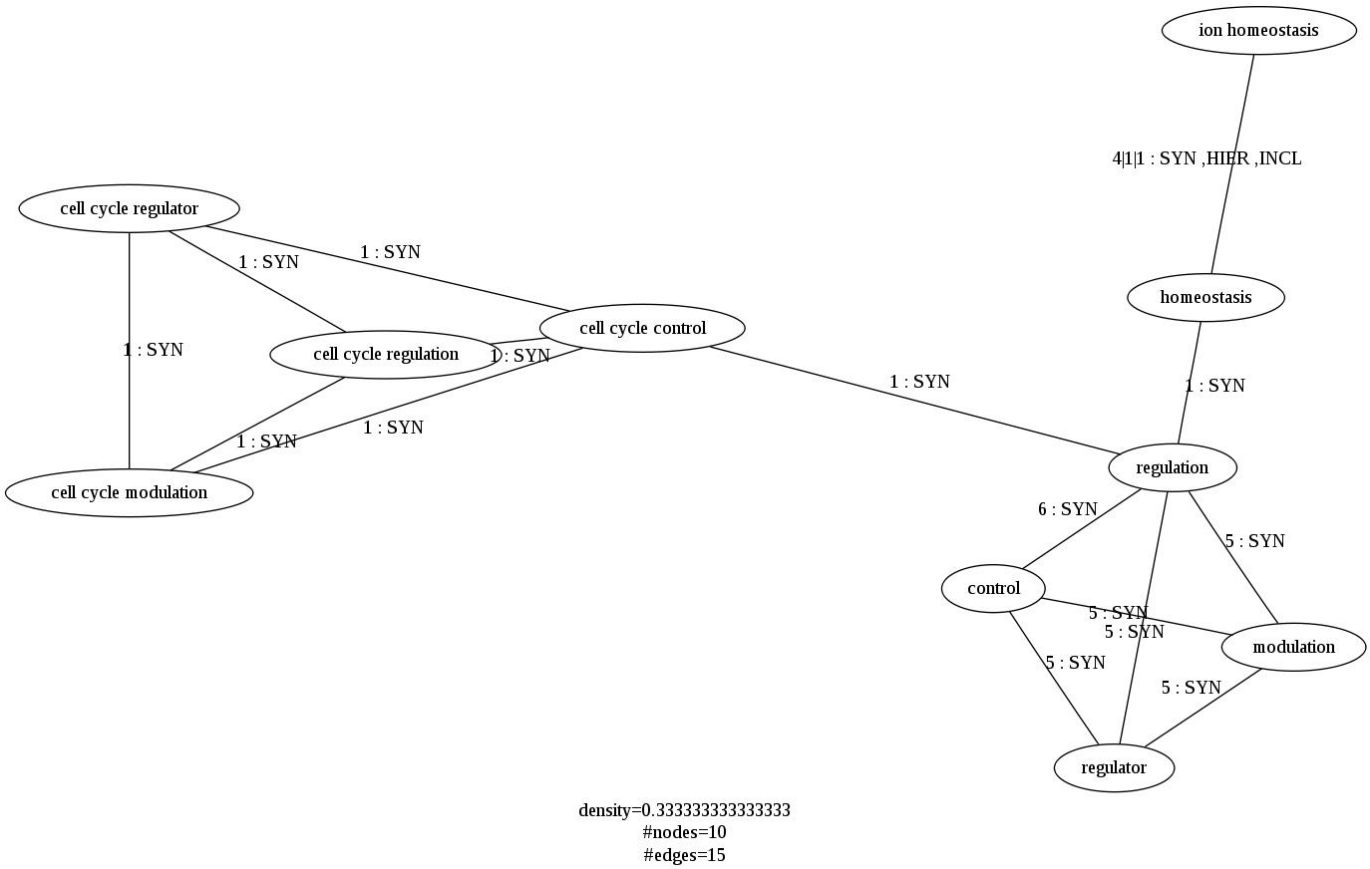


Figure 4: Connected component with three bridges: {*ion homeostasis*, *homeostasis*}, {*homeostasis*, *regulation*} and {*cell cycle control*, *regulation*}.

0.0467 and 1 (in case of cliques). Among the 104 connected components, which are not cliques, we detected 249 bridges: 0 to 35 depending on connected components. In order to propose a general approach exploiting graph theory notions for synonym profiling we analyse the structure of three representative connected components.

Density of the connected component 2(a) is 0.67. It contains one bridge: {*intermembrane space*, *IMS*}. This edge corresponds to the acronym and its expanded form, which can cause its contextual character. Moreover, *intermembrane space* is the central node of this connected component.

Connected component 3 (density=0.38) contains two bridges {*uptake*, *recycling*} and {*salvage*, *cycling*}, and three articulation vertices *uptake*, *recicling* and *salvage* with the measures of centrality 16, 18 and 10 respectively. Indeed, the majority of shortest paths pass by *uptake* and *recicling* nodes. Otherwise, edges around the *salvage* vertex are weakened because of the cooccurrence of synonymy and hierarchical relations. As we have already noticed, the edge {*import*, *uptake*} shows the cooccurrence of synonymy and hierarchical relations, but its productivity is rather high (seven for each relation), which stregthens this edge.

Finally, connected component 4 (density=0.33) contains three bridges {*ion homeostasis*, *homeostasis*}, {*homeostasis*, *regulation*} and {*cell cycle control*, *regulation*} and three articulation vertices: *regulation*, *cell cycle control* and *homeostasis* with the measures of centrality 52, 37 and 16 respectively. The bridge {*ion homeostasis*, *homeostasis*} is weakened by the cooccurrence of synonymy, hierarchical and lexical inclusion relations. Otherwise, other edges seem to convey non ambiguous synonymy.

From the analyzed examples, we can see that the graph theory may have several implications on profiling of synonyms. However, these implications must still be formalized and, possibly, expressed as a single reliability indicator, alone or combined with the lexical and linguistic clues.

First, within a connected component, with a given number of nodes, higher the number of edges, higher will be its density and closer it will be to a clique (fig. 2(b)). Consequently, within a clique, the semantic cohesion is more strong. Indeed, in these cases, terms are far more strongly related between them. But when the density value decreases the semantic cohesiveness of connected components decreases as well. In other words, density is an indication on the semantic cohesiveness between terms within connected components. As for bridges, we assume that they indicate breaking points within connected components, such as {*cell cycle control*, *regulation*} within figure 4. The weak character of these points can increased when the synonymy relation co-occurs with other relationships (`is-a`, `part-of`, lexical inclusion). Consequently, removal of bridges can create connected components with higher density and therefore with stronger synonymy relations. Finally, the centrality of vertices measure may be useful for identification of polysemic words or terms.

The connected components analysis can also indicate the missing relations. For instance, if a connected component, which is not a clique, has no bridges but its density is not maximal, this would indicate that it misses some correct synonymy relations which can be easily induced.

5 Conclusion and Perspectives

In this paper, we propose an original method for inducing synonym lexicon from structured terminologies. This method exploits the compositionality principle and three rules based on syntactic dependency analysis of terms. More specifically, we explore various indicators for profiling the acquired synonym relations, which is motivated by the fact that synonymy is a contextual relation and its validity and universality is not guaranteed. We assume the semantic cohesiveness of synonymy relations should be qualified and quantified. Thus, we propose several indicators for profiling the inferred synonymy relations and for detecting possible weak and strong points. First, lexical and linguistic clues are generated endogenously within the same terminology: other types of elementary semantic relations (`is-a` and `part-of`), lexical inclusions and productivity of the acquired semantic relations. Then, more specifically, this work is dedicated to exploring of the usefulness of notions of the graph theory. We propose to study the form and specificities of connected components formed by synonymy relations. We exploited the following notions from the graph theory: distinction between connected components and cliques, their density, bridges and articulation vertices within connected components, and the centrality of their vertices. We observed that the lexical indicators as well as connected components characteristics are helpful for profiling the acquired synonymy relations. These clues are intended to be used for preparing the validation of this lexicon by experts and also for its weighting in order to control and guarantee the specificity of lexicon during its use by automatic tools.

Currently, we study separately the endogeneous lexical indicators, and the characteristics of the connected components. However, in the future, these two types of clues should be combined. For this, these indicators should be modelized in order to provide a weight of each edge. This weight can be used for profiling of connected component through the detection of strong and weak points. Notice that the current version of the `Graph` package cannot take into account this additional information on edges and should be modified. Another perspective is the better exploitation of the *Gene Ontology* and taking into account the nature of synonymy relations as they are labelled by thier creators: *exact*, *broad*, *narrow* or *related*. Additionnally, for a more precise profiling, the four relationships of *GO* (synonymy, `is-a`, `part-of` and `regulates`) can be cross-validated, while currently, we perform the validation of synonymy relations through `is-a` and `part-of` (and other indicators). We plan also to use the induced relations and propagate them through corpora and discover some of the missing synonyms (Hole and Srinivasan, 2000). In this way, applying the same compositionality principle, we can enrich and extend the *Gene Ontology*: new synonyms of *GO*

terms and even other relations between *GO* terms and terms from corpora can be detected. As noticed, this method can be applied to other terminologies and languages as far as structured terminological resources and NLP tools exist. For instance, within the context of search of clinical documents, we successfully tested this method on the French part of the UMLS (Grabar et al., 2009). From a more ontological perspective, our method can be used for consistency checking of a terminologies, like in (Mungall, 2004). Moreover, as this method performs syntactic analysis of terms and their decomposition into semantically independent components, it can be used for the transformation of a pre-coordinated terminology into a post-coordinated one.

References

G. Burnage. 1990. *CELEX - A Guide for Users*. Centre for Lexical Information, University of Nijmegen.

Roger A. Côté, Louise Brochu, and Lyne Cabana, 1997. *SNOMED Internationale – Répertoire d'anatomie pathologique*. Secrétariat francophone international de nomenclature médicale, Sherbrooke, Québec.

David A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge.

Reinhard Diestel. 2005. *Graph Theory*. Springer-Verlag Heidelberg, New-York.

Christian Fellbaum. 1998. A semantic network of english: the mother of all WordNets. *Computers and Humanities. EuroWordNet: a multilingual database with lexical semantic network*, 32(2-3):209–220.

Gene Ontology Consortium. 2001. Creating the Gene Ontology resource: design and implementation. *Genome Research*, 11:1425–1433.

Natalia Grabar, Marie-Christine Jaulent, and Thierry Hamon. 2008. Combination of endogenous clues for profiling inferred semantic relations: experiments with gene ontology. In *JAMIA (AMIA 2008)*, pages 252–6, Washington, USA.

Natalia Grabar, Paul-Christophe Varoutas, Philippe Rizand, Alain Livartowski, and Thierry Hamon. 2009. Automatic acquisition of synonym ressources and assessment of their impact on the enhanced search in ehrs. *Methods of Information in Medicine*, 48(2):149–154. PMID 19283312.

Thierry Hamon and Natalia Grabar. 2008. Acquisition of elementary synonym relations from biological structured terminology. In *Computational Linguistics and Intelligent Text Processing (5th International Conference on NLP, 2006)*, number 4919 in LNCS, pages 40–51. Springer.

Nabil Hathout, Fiammetta Namer, and Georgette Dal. 2001. An experimental constructional database: the MorTAL project. In P. Boucher, editor, *Morphology book*. Cascadilla Press, Cambridge, MA.

WT Hole and S Srinivasan. 2000. Discovering missed synonymy in a large concept-oriented metathesaurus. In *AMIA 2000*, pages 354–8.

Georges Kleiber and Irène Tamba. 1990. L'hyperonymie revisitée : inclusion et hiérarchie. *Langages*, 98:7–32, juin. L'hyponymie et l'hyperonymie (dir. Marie-Françoise Mortureux).

CJ Mungall. 2004. Obol: integrating language and meaning in bio-ontologies. *Comparative and Functional Genomics*, 5(6-7):509–520.

NLM, 2007. *UMLS Knowledge Sources Manual*. National Library of Medicine, Bethesda, Maryland. www.nlm.nih.gov/research/umls/.

PV Ogren, KB Cohen, and L Hunter. 2005. Implications of compositionality in the Gene Ontology for its curation and usage. In *Pacific Symposium of Biocomputing*, pages 174–185.

Barbara H Partee, 1984. *Compositionality*. F Landman and F Veltman.

Michael Poprat, Elena Beisswanger, and Udo Hahn. 2008. Building a biowordnet using wordnet data structures and wordnet's software infrastructure - a failure story. In *ACL 2008 workshop "Software Engineering, Testing, and Quality Assurance for Natural Language Processing"*, pages 31–9.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Stefan Schulz, Martin Romacker, Pius Franz, Albrecht Zaiss, Rdiger Klar, and Udo Hahn. 1999. Towards a multilingual morpheme thesaurus for medical free-text retrieval. In *Medical Informatics in Europe (MIE)*.

Barry Smith and Christian Fellbaum. 2004. Medical wordnet: a new methodology for the construction and validation of information. In *Proc of 20th CoLing*, pages 371–382, Geneva, Switzerland.

Cornelia M Verspoor, Cliff Joslyn, and George J Papcun. 2003. The gene ontology as a source of lexical semantic knowledge for a biological natural language processing application. In *SIGIR workshop on Text Analysis and Search for Bioinformatics*, pages 51–56.

Pierre Zweigenbaum, Robert Baud, Anita Burgun, Fiammetta Namer, Éric Jarrousse, Natalia Grabar, Patrick Ruch, Franck Le Duff, Benot Thirion, and Stéfan Darmoni. 2003. Towards a Unified Medical Lexicon for French. In *Medical Informatics in Europe (MIE)*.