How can the term compositionality be useful for acquiring elementary semantic relations?

Thierry Hamon[1] and Natalia Grabar[2]

[1] LIPN – UMR 7030, Université Paris 13 – CNRS, 99 av. J-B Clément, F-93430 Villetaneuse, France
thierry.hamon@lipn.univ-paris13.fr
[2] Université Paris Descartes, UMR_S 872, Paris, F-75006 France; INSERM, U872, Paris, F-75006, France
natalia.grabar@spim.jussieu.fr

Abstract. Acquiring and enriching lexical resources is cruciel for various areas of the computational linguistics applications, espcially in specialised domains. In this paper, we propose a high-quality method exploiting the compositionality of complex terms issued from a structured terminology. to infer synonymy and hierarchical relations between words or terms. The approach has been applied and evaluated on the Gene Ontology biomedical terminology. 1,273 is-a, 178 part-of and 921 synonymy relations have been inferred from the 24,357 is-a, 2,726 part-of and 13,850 synonymy relations provided by GO. We analyse the results and their combination through a graph representation to identify clues helping their validation.

1 Introduction

Detection of semantic similarity between terms is an important but heavy step within various natural language processing (NLP) applications. For instance, tasks like query expansions, information retrieval, knowledge extraction or terminology matching highly rely on such information and would generate different results according to whether the semantic proximity between two terms (*i.e.*, *aromatic amino acid family anabolism* and *aromatic amino acid family biosynthesis*) is established or not. In order to help the NLP applications, specific lexica, offering various semantic relations (hyperonymy, meronymy and synonymy) as well as morphological and orthographic variants, can be used. But, depending on languages and on specialized areas, such resources are not equally well described.

We can mention the availability of morphological resources for common language [1, 2] and for medical area [3–5]. We can also mention the common language resource of synonyms WordNet [6] for English, although the corresponding resources for other languages are not freely available. Notice that the initiative for fitting this resource for the medical area [7] is still ongoing, and that there is no initiative for the creation of a similar resource for the NLP processing of biological documents. Besides, lexica with hierarchical or meronymy relations, especially in specialized areas, are not available. The purpose of our work is to

fill in this gap in the biological domain. Within this area, several terminologies are created and continuously updated. We propose to reuse them in order to infer lexica of semantically related words or terms specific to biology. The relationships aimed include synonymy, hyperonymy and meronymy. All these relationships can be used for computing the semantic similarity between words and terms [8–10]. Moreover, they are basic resources for structuring terminologies as well as a way to improving the sensitivity of information retrieval and extraction applications.

The proposed novel method provides high-quality results. This method is language-independent. It exploits the compositionality of complex terms extracted from structured terminologies and is based on the identification of their syntactic invariants. The main originality of our work is that the same method is applied for inferring various semantic relationships as far as input material is correctly constrained.

2 Material

Our main material is Gene Ontology [11] (*GO*), which goal 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. Within *GO*, terms are structured through three types of relations: (1) hierarchical subsumption or hyperonymy, also called `is-a` relation; (2) meronymy or `part-of` relation; (3) and synonymy. Synonyms are grouped within the same concept, which are related between them through hierarchical and `part-of` relations.

For instance, within the concept GO:0009073, the preferred term *aromatic amino acid family biosynthetic process* has several synonyms (*i.e.*, *aromatic amino acid family anabolism*, *aromatic amino acid family biosynthesis*). This concept is related to the concept GO:0008652 *amino acid biosynthetic process* through a hierarchical relation.

The used version of *GO* provides 24,537 `is-a` and 2,726 `part-of` relations, while synonymy relations are established among 18,315 terms and their 13,850 synonyms.

3 Method

Often within biomedical terminologies, terms are coined on the same syntactic and compositional scheme which can be exploited in order to induce the elementary relations between simple terms. For instance, the *GO* concept GO:0009073 contains the following synonyms, which show the compositionality through the substitution of one of their components (underlined):

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

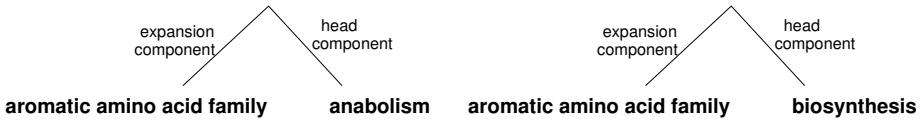


Fig. 1. Parsing syntactic trees of the terms *aromatic amino acid family anabolism* and *aromatic amino acid family biosynthesis* for the acquisition of synonymous relations.

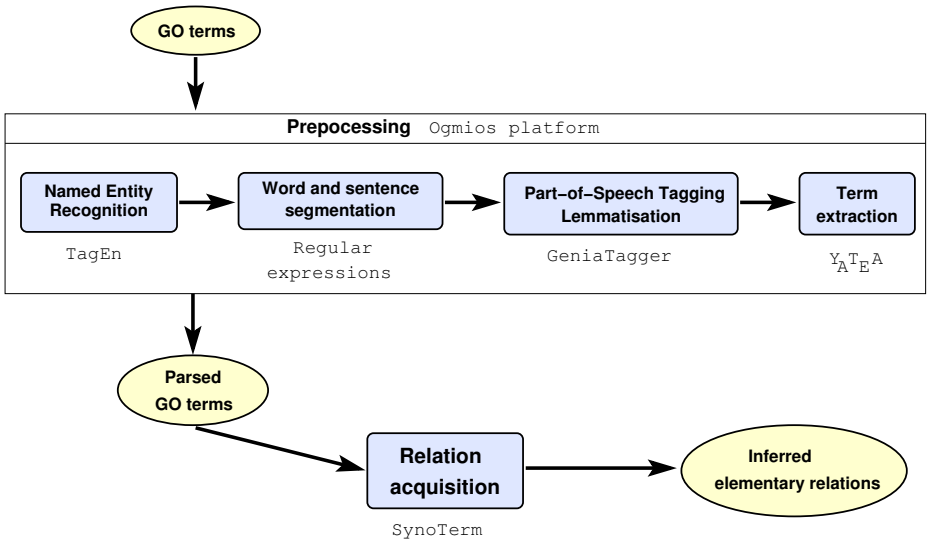


Fig. 2. General flowchart of the method.

It is possible to exploit this scheme and to induce the following paradigm of elementary synonyms: *biosynthesis*, *anabolism*, *formation*, *synthesis*. We propose a method for generalizing this observation for the acquisition of various elementary semantic relationships. Like in the given examples, the method exploits compositional structure of terms and relies on existence of structured terminologies. 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 [12]. In our work, the syntactic analysis of terms is crucial: it normalizes the representation of terms through their head and expansion components and it prepares thus the syntactic dependencies computing.

In the following of this section, we first present the approach for achieving the syntactic analysis of terms (section 3.1) and then the compositionality-based method for acquisition of lexical resources (section 3.2).

3.1 Preprocessing and syntactic analysis: Ogmios

Figure 2 presents general workflow scheme implemented for computing elementary relations existing within *GO*. For this, *GO* terms are preprocessed through

Fig. 3. Parsing syntactic trees of synonym terms *replication of mitochondrial DNA* and *mtDNA replication* with surface syntactic variation for the acquisition of synonymous relations.

the Ogmios linguistic annotation platform [13] in order to automatically analyze these terms and generate their syntactic analysis. As result, all terms are parsed into their head and expansion components. The used tools are developed in Perl5 language.

The `Ogmios` platform is adapted to the processing of large amount of data and, moreover, can be easily tuned to a specialized domain. Through this platform, several types of linguistic processing are performed. First, the `TagEN` [14] tool is applied for the recognition on named entities (*i.e.*, gene names, chemical products). Its application at the beginning of linguistic pipeline helps the forthcoming segmentation into words and sentences. Indeed, the recognition of named entities allows disambiguating special characters, such as punctuation marks, dashes, slashes, etc, widely used within named entities in biology and often altering the segmentation into words and sentences.

After the segmentation, the `GeniaTagger` [15] tool is applied in order to perform POS-tagging and lemmatization.

The final step within the Ogmios platform is the shallow syntactic analysis of terms in order to syntactically parse them. This task is carried out thanks to the rule-based term extractor YaTeA [16]. Syntactic dependencies between term components are computed according to assigned POS tags and shallow parsing rules. Thus, each term is considered as a syntactic binary tree (figure 1) composed of two elements: head component and expansion component. For instance, in terms of syntactic dependencies, *anabolism* is the head component of *aromatic amino acid family anabolism* term, while *aromatic amino acid family* is its expansion component. It goes without saying that each complex component (*i.e.*, *aromatic amino acid family*) is also syntactically parsed, which can give place to inferring even more elementary relations. Moreover, because we used the syntactic structure of the terms, their surface form is not an obstacle for their alignment. For instance, once two synonym terms like *replication of mitochondrial DNA* and *mtDNA replication* are lemmatized and syntactically analyzed, *replication* is recognized to be their head component and *mitochondrial DNA* and *mtDNA* their expansion components (figure 3).

Such analyzed *GO* terms are then aligned through specific compositional rules and ready for detection of elementary semantic relations within *GO* terms.

3.2 Acquisition of elementary relations

In this work, the compositionality-based method designed for the terminology structuring through corpora [17], is adapted to inferring elementary relations between simple terms. The method is applied for the acquisition of synonymy, hierarchical and `part-of` relations, as described below. It should be noticed that the method is recursive and each inferred elementary relation can then be propagated in order to infer new elementary relations, and thus to generate a more exhaustive lexicon with a given relationship.

Acquisition of synonymy relations. For the acquisition of synonymy relations, we consider that if the meaning $\mathcal{M}$ of two complex terms $A\ rel\ B$ and $A'\ rel\ B$ 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 synonymous terms and if B is identical, then the synonymy relation between simpler terms A and A' can be inferred. The method takes into account the syntactic structure of complex terms. The fully parsed terms are represented as a terminological network, within which the deduction of the elementary synonymy relations is based on the three rules:

R1 If two terms are synonymous and their expansion components are identical, then an elementary synonymy relation is inferred: the pair {*anabolism*, *biosynthesis*} is inferred from the original synonymy relation between *acetone anabolism* and *acetone biosynthesis* where the expansion component *acetone* is identical in both terms (figure 1).

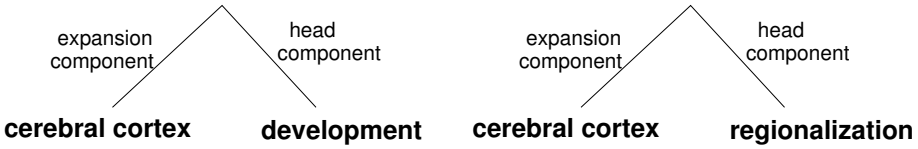
R2 If both terms are synonymous and their head components are identical, then an elementary synonymy relation is inferred: the pair {*endocytic*, *endocytotic*} is inferred from the synonymy relation between *endocytic vesicle* and *endocytotic vesicle* where the head component *vesicle* is identical.

R3 If both terms are synonymous and either their head or expansion components are synonymous, then an elementary synonymy relation is inferred: the pair {*nicotinamide adenine dinucleotide*, *NAD*} is inferred from the synonymy relation between *nicotinamide adenine dinucleotide catabolism* and *NAD breakdown* where the head components {*catabolism*, *breakdown*} are already known synonymous.

Acquisition of hierarchical and `part-of` relations. The same method is applied for the acquisition of hierarchical (or `part-of`) relations. For this, original pairs are composed of the *GO* hierarchical (or `part-of`) pairs. Thus, if the meaning $\mathcal{M}$ of two complex terms $A\ rel\ B$ and $C\ rel\ B$ are given as following:

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

(a) Parsing tree of the terms *cell activation* and *astrocyte activation* related through a hierarchical relation



(b) Parsing tree of the terms *cerebral cortex development* and *cerebral cortex regionalization* related through a `part-of` relation

Fig. 4. Parsing syntactic trees for the acquisition of hierarchical and `part-of` relations.

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

for a given composition function f, if $A\ rel\ B$ and $C\ rel\ B$ are complex terms related through hierarchy (or `part-of` relation) and if B is identical, then the hierarchical (or `part-of`) relation between simpler terms A and C can be inferred.

For the acquisition of these relations we exploit the same three rules. For instance, figure 4 exemplifies rules R2 and R1, where one of components of the original terms is identical. On figure 4(a), original terms are two biological processes: *cell activation* GO:0001775 and *astrocyte activation* GO:0048143. They have between them hierarchical relation: *cell activation* is the hierarchical parent to *astrocyte activation*. Further to their syntactic analysis and application of the compositional rule R2, the hierarchical relation between their expansion components *cell* $\Rightarrow$ *astrocyte* can be inferred.

Similarly, on figure 4(b), original terms are two biological processes: *cerebral cortex development* GO:0021987 and *cerebral cortex regionalization* GO:0021796. They have between them `part-of` relation: *cerebral cortex regionalization* is a part of a more large biological process *cerebral cortex development*. Further to their syntactic analysis and application of the compositional rule R1, the `part-of` relation between their head components *development* $\Rightarrow$ *regionalization* can be inferred.

Notice that another work [18] aimed at the acquisition of elementary hierarchical relations from structured terminologies. Their approach relies on string substitution within identical lexical contexts, while in the work we propose we perform a more rich NLP approach by syntactically analysing terms and applying compositionality-based transformation syntactic rules.

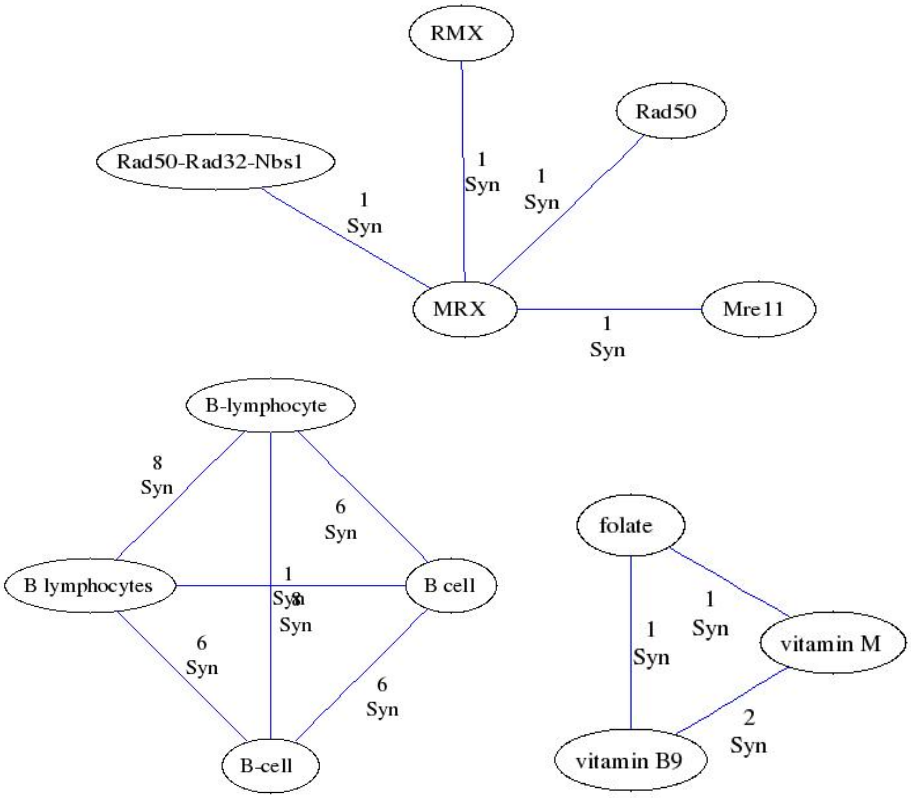


Fig. 5. Connected components (CCs) presenting elementary synonymy relations. Components can have various shapes: star-shaped connection between nodes for *MRX* CC, or cliques (strongly connected components) for the two other CCs.

4 Results and Discussion

23,899 *GO* terms have been fully parsed through the Ogmios platform. The three compositional rules have been then applied and allowed to infer elementary synonyms (n=921), `is-a` (n=1,273) and `part-of` pairs (n=178). We present and discuss these results. For the inferred synonymy relations, we present also their productivity (number of original *GO* pairs which allowed to infer them).

4.1 Elementary synonyms

The 921 inferred elementary synonyms have been grouped into 627 connected components (CCs) – groups of synonyms which are linked between them. For instance, the CC *MRX* of figure 5 contains five elementary synonyms (*MRX*, *Rad50-Rad32-Nbs1*, *RMX*, *Rad50* and *Mre11*) inferred from the *GO* concept

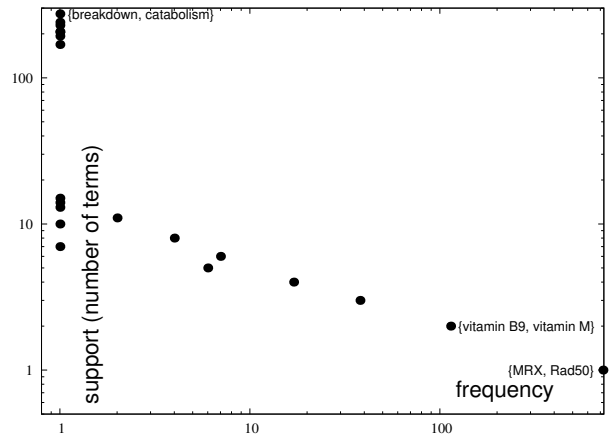
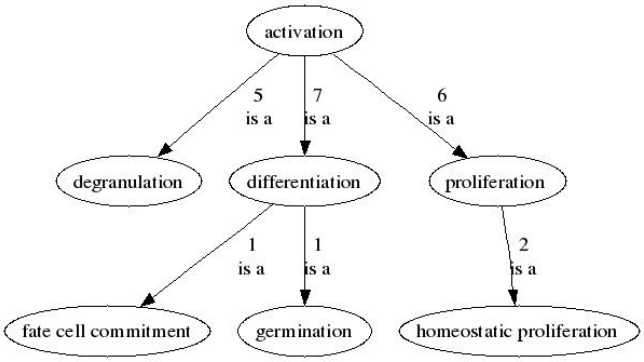


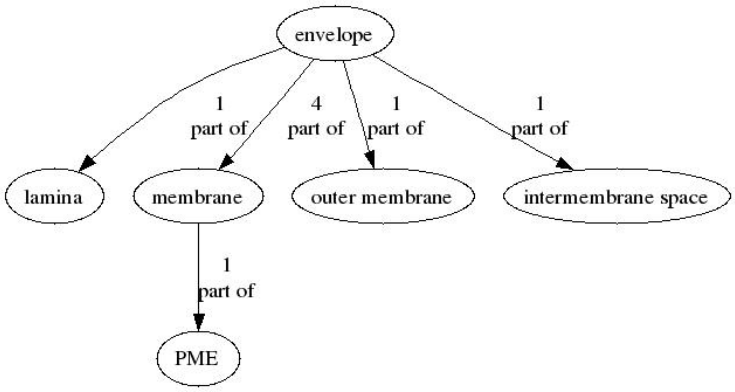
Fig. 6. Productivity of the inferred elementary synonyms within *GO* (logarithmic scale).

GO:0030870 which preferred term is *MRX complex*. Elementary synonymy relations are labelled *Syn* on figures, and numbers indicate their productivity (number of original *GO* pairs from which an elementary relation has been inferred). In this CC, the preferred elementary term *MRX* is linked to all its synonyms – this is a star-shaped CC. The two other CCs are strongly connected, or cliques: all their nodes are related between them. Observing synonyms through their CCs, rather than pairs of synonyms (*i.e.*, {*MRX*, *Rad50*}, {*B-cell*, *B-lymphocyte*}, {*B-cell*, *B cell*}, {*vitamin B9*, *vitamin M*}) gives a more global view of their semantics. In this way, the contextual nature of synonyms, which can influence their acceptance, is more easily detected within CCs though terms or their relations.

Productivity of the elementary synonyms within *GO* is presented on figure 6 (scaled logarithmically). Its axes represent the support (number of original *GO* synonyms that allow to infer a given pair of elementary synonyms) and the frequency of each support value. Pairs, which productivity values are concentrated near the top left corner, are the more reliable: their meaning and use are the most common. For instance, {*breakdown*, *catabolism*} is the most productive (and reliable) synonym pair: it is inferred within 274 *GO* synonyms and appears to be a fundamental notion in biology. At the other end, we have pairs like {*MRX*, *Rad50*} or {*vitamin B9*, *vitamin M*} inferred from one or two original synonyms. As their meaning seems to be more specific, they may convey more specific semantics. Besides, such rare pairs represent nearly 80% (n=722) of the whole set of the inferred synonyms.



(a) CC of elementary hierarchical relations of *activation*.



(b) CC of elementary `part-of` relations of *envelope*.

Fig. 7. Connected components of elementary hierarchical and `part-of` relations. Numbers indicate the productivity of the inferred pairs within original *GO* terms.

4.2 Elementary hierarchical and part-of relations

We inferred 1,273 hierarchical and 178 `part-of` elementary relations. Figure 7 presents two of the generated connected components.

Most of the acquired hierarchical pairs (85%, n=1089) are inferred from only one pair of original *GO* terms. This is the case for *differentiation* ⇒ *fate cell commitment* and *differentiation* ⇒ *germination* pairs on figure 7(a) and with the relations on figure 7(b). The most frequent elementary hierarchical pair is *membrane* ⇒ *part*. It is found within nine *GO* term pairs, for instance, within the following cellular components terms:

vacuolar part (GO:0044437) ⇒ *vacuolar membrane* (GO:0005774)
peroxisomal part (GO:0044439) ⇒ *peroxisome membrane* (GO:0005778)

endoplasmic reticulum part (GO:0044432) ⇒ *endoplasmic reticulum membrane* (GO:0005789)

As for elementary `part-of` relations, the most frequent one is *development* ⇒ *morphogenesis*. It is acquired within 46 *GO* term pairs, denoting mostly biological processes, for instance:

compound eye development (GO:0048749) ⇒ *compound eye morphogenesis* (GO:0001745)
Bolwig's organ development (GO:0055034) ⇒ *Bolwig's organ morphogenesis* (GO:0001746)
neural plate development (GO:0001840) ⇒ *neural plate morphogenesis* (GO:0001839)
endothelial cell development (GO:0001885) ⇒ *endothelial cell morphogenesis* (GO:0001886)

First evaluation of the acquired resources showed that the quality of synonyms is very good (over 90% precision). Evaluation of hierarchical and `part-of` relations is still ongoing. Recall of these resources could not be evaluated because there is no reference data for this kind of lexica. Nevertheless we assume that a better POS-tagging and then shallow parsing would improve detection of the semantic relations within terms. Our results showed also that the redundancy of use of elementary synonyms is very high: productivity of several pairs is greater than 100 original *GO* term pairs. Redundancy within hierarchical and `part-of` relationships is rather small.

5 Conclusion and Perspectives

In this paper, we propose a compositionality-based method for inferring different types of elementary semantic relations from structured terminologies in order to help the natural language processing applications. The method relies on syntactic analysis of terms and exploits three compositional rules. The main originality of this work is that the same approach is applied for inferring different types of semantic relations: synonymy, hierarchical and `part-of` relations. The semantic nature of the source term pairs has to be constrained, while the NLP part of the method remains the same.

The inferred resources are useful for different NLP applications, particularly for those initiated by Philip Resnik [8] for computing the semantic distance between terms.

The presented work has several perspectives. For instance, the inferred elementary relations can be used for enriching the existing terminologies through the detection of additional synonymous or hierarchically related terms. For this, the reverse method should be applied on corpora [17].

This method is language-independent, and it is possible to apply it to other languages as far as (1) the required linguistic processing can be realized and (2) semantic relations between complex terms are available. In the biomedical area,

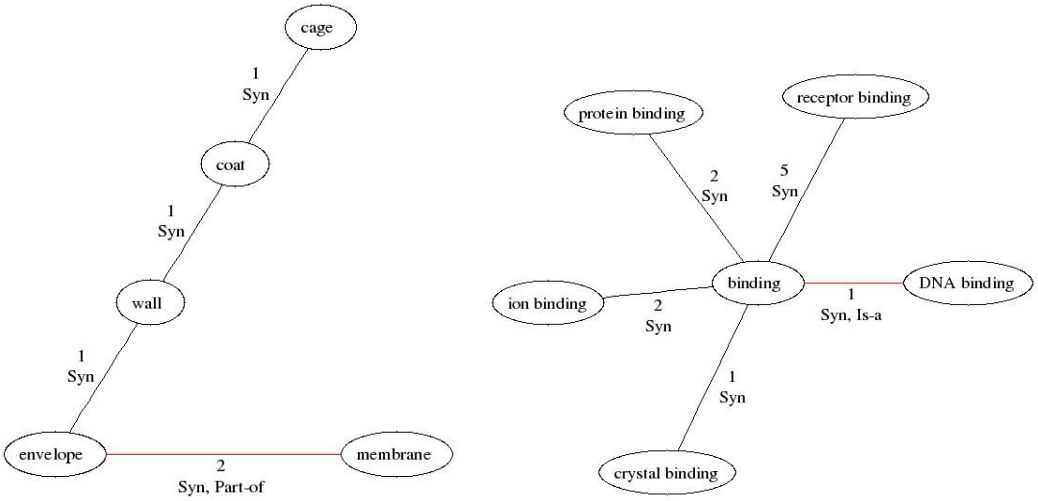


Fig. 8. Connected components with lexical inclusions, hierarchical and part-of relations (possible weak points of connected components).

we plan to apply our method on the `UMLS` resource [3], or more specifically the `MeSH` [19] or `Snomed` [20] terminologies which are available in several languages.

Additionally, the inferred resources can be used for their cross-validation. For instance, if the same elementary relation is inferred as being synonymy and hierarchical, it can help detecting possibly weak points within network of the generated relations and ambiguities or inconsistencies within original terminologies. As shows figure 8, some of the inferred relations are indeed ambiguous:

- *envelope* and *membrane* are found to be both synonymously and `part-of` related;
- *binding* and *DNA binding* are found to be both synonymously and hierarchically related.

In this perspective, the three relationships between elementary terms (synonymy, hierarchical and `part-of`) can be cross-validated between them. This would prepare the human validation of the inferred resources which must be thoroughly evaluated still.

References

1. Burnage, G.: CELEX - A Guide for Users. Centre for Lexical Information, University of Nijmegen (1990)
2. Hathout, N., Namer, F., Dal, G.: An experimental constructional database: the MorTAL project. In Boucher, P., ed.: Morphology book. Cascadilla Press, Cambridge, MA (2001)

3. NLM: UMLS Knowledge Sources Manual. National Library of Medicine, Bethesda, Maryland. (2007) www.nlm.nih.gov/research/umls/.
4. Schulz, S., Romacker, M., Franz, P., Zaiss, A., Klar, R., Hahn, U.: Towards a multilingual morpheme thesaurus for medical free-text retrieval. In: Medical Informatics in Europe (MIE). (1999)
5. Zweigenbaum, P., Baud, R., Burgun, A., Namer, F., Jarrousse, E., Grabar, N., Ruch, P., Duff, F.L., Thirion, B., Darmoni, S.: Towards a Unified Medical Lexicon for French. In: Medical Informatics in Europe (MIE). (2003)
6. Fellbaum, C.: A semantic network of english: the mother of all WordNets. Computers and Humanities. EuroWordNet: a multilingual database with lexical semantic network **32**(2-3) (1998) 209–220
7. Smith, B., Fellbaum, C.: Medical wordnet: a new methodology for the construction and validation of information. In: Proc of 20th CoLing, Geneva, Switzerland (2004) 371–382
8. Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research (JAIR) **11** (1999) 95–130
9. Bousquet, C., Jaulent, M.C., Chatellier, G., Degoulet, P.: Using semantic distance for the efficient coding of medical concepts. In: Annual Symposium of the American Medical Informatics Association (AMIA), Los Angeles, CA (2000) 96–100
10. Lord, P., Stevens, R., Brass, A., Goble, C.: Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics **19**(10) (2003) 1275–1283
11. Gene Ontology Consortium: Creating the Gene Ontology resource: design and implementation. Genome Research **11** (2001) 1425–1433
12. Partee, B.H. In: Compositionality. F Landman and F Veltman (1984)
13. Hamon, T., Nazarenko, A., Poibeau, T., Aubin, S., Derivière, J.: A robust linguistic platform for efficient and domain specific web content analysis. In: RIAO 2007, Pittsburgh, USA (2007)
14. Berroyer, J.F.: Tagen, un analyseur d"entits nommes : conception, dveloppement et valuation. Mmoire de D.E.A. d'intelligence artificielle, Universit Paris-Nord (2004)
15. Tsuruoka, Y., Tateishi, Y., Kim, J.D., Ohta, T., McNaught, J., Ananiadou, S., , Tsujii, J.: Developing a robust part-of-speech tagger for biomedical text. LNCS **3746** (2005) 382–392
16. Aubin, S., Hamon, T.: Improving term extraction with terminological resources. In Salakoski, T., Ginter, F., Pyysalo, S., Pahikkala, T., eds.: Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006). Number 4139 in LNAI, Springer (August 2006) 380–387
17. Hamon, T., Nazarenko, A., Gros, C.: A step towards the detection of semantic variants of terms in technical documents. In: International Conference on Computational Linguistics (COLING-ACL'98), Université de Montréal, Montréal, Quebec, Canada (1998) 498–504
18. Verspoor, C.M., Joslyn, C., Papcun, G.J.: 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. (2003) 51–56
19. National Library of Medicine Bethesda, Maryland: Medical Subject Headings. (2001) `http://www.nlm.nih.gov/mesh/meshhome.html`.
20. Côté, R.A.: Répertoire d'anatomopathologie de la SNOMED internationale, v3.4. Université de Sherbrooke, Sherbrooke, Québec. (1996)