Combining Compositionality and Pagerank for the Identification of Semantic Relations between Biomedical Words

Thierry Hamon
LIM&BIO UFR SMBH
Université Paris 13, France
thierry.hamon@univ-paris13.fr

Christopher Engström
Division of Applied Mathematics
Mälardalen University
Västerås, Sweden

Mounira Manser
LIM&BIO UFR SMBH
Université Paris 13, France

Zina Badji and Natalia Grabar
CNRS UMR 8163 STL
Université Lille 1&3
59653 Villeneuve d'Ascq, France
natalia.grabar@univ-lille3.fr

Sergei Silvestrov
Division of Applied Mathematics
Mälardalen University
Västerås, Sweden

Abstract

The acquisition of semantic resources and relations is an important task for several applications, such as query expansion, information retrieval and extraction, machine translation. However, their validity should also be computed and indicated, especially for automatic systems and applications. We exploit the compositionality based methods for the acquisition of synonymy relations and of indicators of these synonyms. We then apply pagerank-derived algorithm to the obtained semantic graph in order to filter out the acquired synonyms. Evaluation performed with two independent experts indicates that the quality of synonyms is systematically improved by 10 to 15% after their filtering.

1 Introduction

Natural languages have extremely rich means to express or to hide semantic relations: these can be more or less explicit. Nevertheless, the semantic relations are important to various NLP tasks within general or specialized languages (*i.e.*, query expansions, information retrieval and extraction, text mining or machine translation) and their deciphering must be tackled by automatic approaches. We focus in this work on synonymy relations. Thus, it is important to be able to decide whether two terms (*i.e.*, *anabolism* and *acetone anabolism*, *acetone anabolism* and *acetone biosynthesis*, *replication of mitochondrial DNA* and *mtDNA replication*) convey the same, close or different meanings. According to the ability of an automatic system to decipher such relations, the answers of the system will be more or less exhaustive. Several solutions may be exploited when deciphering the synonymy relations:

1. Exploitation of the existing resources in which the synonyms are already encoded. However, in the biomedical domain, such resources are not well described. If the morphological description is the most complete (NLM, 2007; Schulz et al., 1999; Zweigenbaum et al., 2003), little or no freely available synonym resources can be found, while the existing terminologies often lack the synonyms.
2. Exploitation and adaptation of the existing methods (Grefenstette, 1994; Hamon et al., 1998; Jacquemin et al., 1997; Shimizu et al., 2008; Wang and Hirst, 2011).
3. Proposition of new methods specifically adapted to the processed data.

Due to the lack of resources, we propose to exploit the solutions 2 and 3. In either of these situations, the question arises about the robustness and the validity of the acquired relations. For instance, (Hamon and Grabar, 2008) face two problems: (1) contextual character of synonymy relations (Cruse, 1986), *i.e.*, two 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 words taken out of their context can convey different relations than the one expected. Our objective is to assess the reliability of synonymy resources. We propose to weight and to filter the synonym relations with the pagerank-derived algorithm (Brin and Page, 1998). When

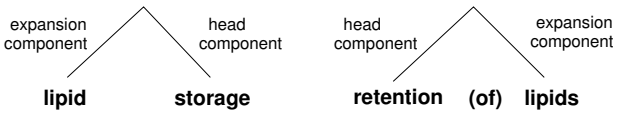


Figure 1: Parsing tree of the terms *lipid storage* and *retention of lipids*

processing textual data, this algorithm has been previously applied in different contexts such as semantic disambiguation (Mihalcea et al., 2004; Sinha and Mihalcea, 2007; Agirre and Soroa, 2009), summarization (Fernandez et al., 2009) and, more recently, for the identification of synonyms (Sinha and Mihalcea, 2011). This last work takes into account the usage of a given word in corpora and its known synonyms from lexical resources. Other related works propose also the exploitation of the random walk algorithm for the detection of semantic relatedness of words (Gaume, 2006; Hughes and Ramage, 2007) and of documents (Hassan et al., 2007). Our work is different from the previous work in several ways: (1) the acquisition of synonymy is done on resources provided by a specialized domain; (2) the pagerank algorithm is exploited for the filtering of semantic relations generated with linguistically-based approaches; (3) the pagerank algorithm is adapted to the small size of the processed data.

In the following of this paper, we present first the material (section 2), then the method we propose (section 3). We then describe the experiments performed and the results (section 4), as well as their evaluation and discussion (section 5). Finally, we conclude and indicate some perspectives (section 6).

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 vocabulary for describing the roles of genes and their products in any organism. *GO* terms are structured with four types of relations: subsumption `is-a`, meronymy `part-of`, `synonymy` and `regulates`. The version used in the current work is issued from the UMLS 2011AA. It provides 54,453 concepts and their 94,161 terms. The generated pairs of terms have 119,430 `is-a` and 101,254 `synonymy` relations.

3 Methods

Our method has several steps: preprocessing of *GO* terms (section 3.1), induction of elementary synonyms (section 3.2) and their characterization with lexical and linguistic indicators (section 3.3), analysis of the synonymy graph, its weighting thanks to the pagerank algorithm and its filtering (section 3.4). We also perform an evaluation of the generated and filtered synonymy relations (section 3.5).

In the following, we call *original synonyms* those synonyms which are provided by *GO*, and we call *elementary synonyms* those synonyms which are induced by the compositionality based approach.

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[1] and perform: segmentation into words and sentences; POS-tagging and lemmatization (Tsuruoka et al., 2005); and syntactic analysis[2]. Syntactic dependencies between term components are 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, *lipid* is the head component of the two terms analyzed on figure 1.

3.2 Compositionality based induction of synonyms

GO terms present compositional structure (Verspoor et al., 2003; Mungall, 2004; Ogren et al., 2005). In the example below (concept GO:0009073) the 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

We propose to exploit the compositionality for induction of synonym resources (*i.e.*, *biosynthesis*, *anabolism*, *formation*, *synthesis* in the given example).

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

While the cited works are based on the string matching, our approach exploits their syntactic analysis, which makes it independent on their surface graphical form (like examples on figure 1).

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). This assumption is very often true in specialized langages, which are known to be compositional. 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*}. It should noticed that *rel* depends on the original relations: if the original terms are synonyms then the elementary terms are also synonyms, if the original terms are hierarchically related then the elementary terms are also hierarchically related, etc.

3.3 Lexically-based profiling of the induced elementary synonyms

In order to test and improve the quality of the induced synonymy relations, we confront these synonyms with approaches which allow to acquire the hyperonymy relations. All these resources are endogeneously acquired from the same terminology *GO*:

- Each induced pair of synonyms is controlled for the lexical inclusion (Kleiber and Tamba, 1990; Bodenreider et al., 2001). If the test is positive, like in the pair {*DNA binding*, *binding*} this would suggest that this pair may convey a hierarchical relation. Indeed, it has been observed that lexical subsumption marks often a hierarchical subsumption. Thus, in the pair {*DNA binding*, *binding*}, *binding* is the hierarchical parent of *DNA binding*, while *DNA binding* has a more specific meaning than *binding*. One can assume that the cooccurrence of synonymy with the lexical subsumption makes the synonymy less reliable;
- The same compositional method, as described in the previous section, is applied to original *GO* term pairs related through `is-a` relations. In this way, we can also infer `is-a` elementary relations. Thus, if a pair of induced synonyms is also induced through `is-a` relations, *i.e.* {*binding*, *DNA binding*}, this also makes the synonymy relations less reliable.

In summary, an induced synonymy relation is considered to be less reliable when it cooccurs with a lexical inclusion or with `is-a` relation. For instance, several edges from figure 2 present the cooccurrence of synonymy relations with the `is-a` relations (such as, {*holding*, *retention*}, {*retention*, *storage*} or {*retention*, *sequestering*}).

3.4 Pagerank-derived filtering of the induced elementary synonyms

The induced semantic relations can be represented as graphs where the nodes correspond to words and the edges to one or more relations between given two words. An example of what it can look like can be seen on figure 2: the induced synonymy relations may indeed cooccur with non-synonymy relations, like the hierarchical relations `is-a`. We propose to use a pagerank approach (Brin and Page, 1998) in order to separate a given graph of synonym relations into subsets (or groups) within which all the words are considered as synonyms with each other but not with any other word outside their subset. In order not to influence the results by the varying size of the graphs, we exploit a non-normalized version of pagerank (Engström, 2011). Thus, given the usual

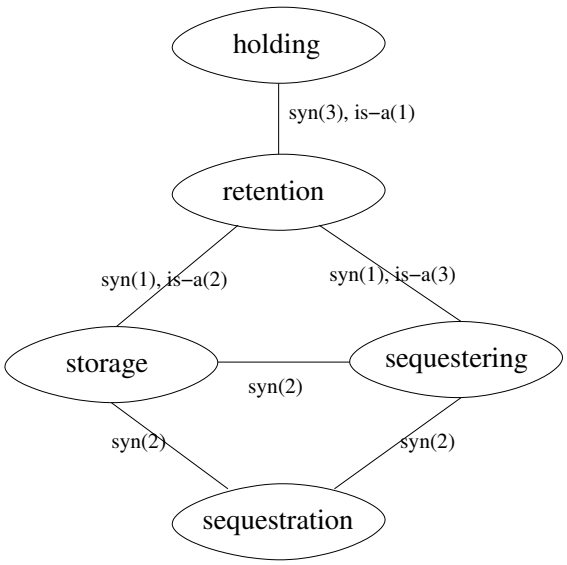


Figure 2: An example of graph generated thanks to the induced semantic relations: pairs related with synonymy relations `syn` may also be related with non-synonymy relations (like hierarchical relation `is-a`)

normalized version $P_{S_i}^{(1)}$ of pagerank:

Definition 1 *$P_S^{(1)}$ for system S is defined as the eigenvector with eigenvalue one to the matrix*

$$M = c(A + gv^T)^T + (1 - c)v1^T$$

where g is a $n \times 1$ vector with zeros for nodes with outgoing nodes and 1 for all dangling nodes, $0 < c < 1$, A is the linkmatrix with sum of every row equal to one, v is a non-negative weightvector with sum one.

As we mentioned, with the processed data we have to use the non-normalized version of pagerank:

Definition 2 *$P_S^{(2)}$ for system S is defined as:*

$$P_S^{(2)} = \frac{P_S^{(1)}||V||_1}{d}, \text{with } d = 1 - \sum cA^T P_S^{(1)}$$

where V is the part of a global weightvector corresponding to the system S. We let V be the one vector such that all words are weighted equally.

Looking at the example from figure 2, we start from any node and then randomly either stop by a probability c or choose (possibly weighted by edge-weights) a new node by the probability $1 - c$ from any of those linked to the chosen node. The pagerank of a node can then be seen as the sum of the probabilities of all paths to the node in question (starting in every node once including itself).

Usually A is a two-dimensional matrix in which the sum of every row is equal to one and all non-zero elements are equal between them. In order to use different types of relations and different weights on these relations we calculate cA. Given B, where B contains the weights of different edges and their type, we calculate A as:

$$A_{i,j} = (B_{i,j,SYN}/(B_{i,j,OTHER} + 1))/n_i$$

where n_i is the total number of edges connected to node i. We treat all relations as symmetric relations for the filtering algorithm when creating B. While some relations aren't symmetric it seems reasonable to assume they affect the likelihood of synonyms in both directions. We also do not distinguish non-synonym relations among them. However, we try a few variations on how to weight A such as assigning different weights to synonym and non-synonym relations or using a logarithmic scale to decrease the effect of very different weights in B.

Further to the weighting, the rows of A do not necessarily sum to one. We propose then not to choose a specific value for c, but to threshold the sum of every row in cA to 0.95. This means that for most of the rows we set $c_{row} = 1/\sum A_{row} \cdot 0.95$, but for rows with a low sum we don't increase the strength of the links but rather keep them as they are ($c_{row} = 1$). Choosing the threshold can be seen as choosing c in the ordinary pagerank formulation. A low threshold means that only the immediate surrounding of a node may impact its pagerank, while a high threshold means that distant nodes may also have an impact. Higher threshold is also useful to separate the pagerank of nodes and to make slower the convergence when calculating the pagerank. When the sum of all rows is less than one and all non-zero elements are positive we can guarantee that the pagerank algorithm converges (Bryan and Leise, 2006). We also use the *Power Method* modified for the non-normalized version of pagerank (Engström, 2011). On the basis of these elements, we apply the following algorithm for segmenting the graph into groups of nodes:

1. Calculate weighted linkmatrix;
2. Calculate pagerank from uniform weightvector v_i;

3. Select the node with the highest pagerank;
4. Calculate pagerank from non-uniform weightvector (zero vector with a single 1 for the selected node);
5. Nodes with $P^{(2)} >$ *cutoff* are selected as synonyms with selected node and each other;
6. Remove the found synonym nodes from the graph;
7. If the graph is non empty, restart from step 1;
8. Otherwise end: words belonging to the same group are considered as synonyms.

We present the application of the algorithm on the example from figure 2 using the *cutoff* $= 1.5$. We start by calculating the weights on the links (weighted linkmatrix). For instance, given the relation from *storage* to *retention* we have: $A_{i,j} = (B_{i,j,SYN}/(B_{i,j,OTHER} + 1))/n_i = (1/(2 + 1))/3 = 1/9$. After computing the weights for all the relations and thresholding the sum of rows to 0.95, when the sum of weights out of a node is larger than 0.95, we obtain figure 3. This gives the pagerank from uniform vector $[4.8590, 7.7182, 16.4029, 16.1573, 15.4152]$, in which we select the node *storage* with the highest pagerank. Pagerank from non-uniform weightvector is then $[0.5490, 1.0970, 4.7875, 4.0467, 3.9079]$, in which we select the nodes with rank larger than *cutoff* $= 1.5$ (*storage*, *sequestration*, *sequestering*) as synonyms. After removing these nodes, we recalculate the weight matrix and repeate the algorithm: the two remaining nodes are found to belong to the same group. We then terminate the algorithm.

3.5 Evaluation protocol

The evaluation is performed against the manually validated synonymy relations. This validation has been done by two independent experts with the background in biology. They were asked to validate the induced synonyms acquired as the step 3.2 of the method. The inter-expert Cohen's kappa is 0.75. On the basis of this evaluation, we compute the precision: percentage of relations which allow to correctly group terms within the connected components and the groups. We compute two kinds of precision (Sebastiani, 2002): micro-precision which is the classical conception of this measure obtained at the level of the relations, and macro-precision which corresponds to the mean of the precisions obtained at the level of connected components or groups. The evaluation is done with the induced synonyms and also after their filtering with the pagerank-derived algorithm. This last evaluation leads to a better observation of the efficiency of the pagerank algorithm.

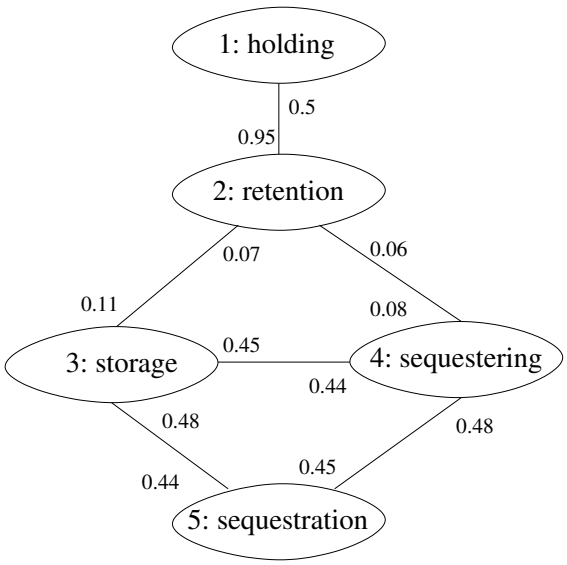


Figure 3: Example from figure 2 with weighted links

4 Experiments and Results

The *GO* terms have been fully processed with the NLP tools (POS-tagging and syntactic analysis) in order to prepare the next step, during which the elementary relations and the indicators are acquired.

4.1 Application of the lexical NLP methods

We applied the NLP method to the *GO* terms. The application of the compositionality approach to original synonymy and hierarchical relations generated 3,707 and 10,068 elementary relations, synonymous and hierarchical respectivelly. Depending on the syntactic structure of the original terms, the synonymy relations are induced between simple or complex terms, but also between their abbreviated and full forms, between the morpho-syntactic variants, etc. Very few of these synonyms exist within *GO* or within the WordNet resource (Fellbaum, 1998). We also detected 1,608 lexical inclusions. The lexical inclusions and the `is-a` relations are preserved only if they cooccur with in-

duced synonymy relations. All these relations are then grouped into connected components (figure 2): the synonymy relations correspond to edges, term components correspond to nodes, while the information on `is-a` relations and on lexical inclusions appears as reliability indicators of the synonymy edges. A total of 2,017 connected components are generated. The biggest connected component contains 140 nodes and 183 edges. At this step, the connected components are evaluated against the reference data: we compute the precision.

4.2 Filtering of the induced synonyms with the pagerank-derived algorithm

We the apply the pagerank-derived algorithm to the induced synonyms, but also to the combinations of these synonyms with `is-a` relations and/or with lexical inclusions. The objective is then to filter the induced synonyms and to improve their reliability. We perform seven experiments, in which the synonymy and the indicators may receive the same importance or may be weighted:

1. syn: only the elementary synonymy relations are considered;
2. syn-isa: combination of synonymy and hierarchical `is-a` relations;
3. syn-incl: combination of synonymy relations with lexical inclusions;
4. syn-isa-incl: combination of synonymy and hierarchical relations with lexical inclusions;
5. syn-isa(535): combination of synonymy relations with lexical inclusions, using different weights: $(A_{i,j} = 5B_{i,j,SYN}/(3B_{i,j,OTHER} + 5))/n_i$;
6. syn-isa(353): combination of synonymy relations with lexical inclusions, using different weights: $(A_{i,j} = 3B_{i,j,SYN}/(5B_{i,j,OTHER} + 3))/n_i$.
7. syn-isa(log): combination of synonymy relations with lexical inclusions, using logarithmic weights: $(A_{i,j} = ((1/ln(2))ln(B_{i,j,SYN} + 1)/((1/ln(2))ln(B_{i,j,OTHER} + 2)))/n_i$.

According to the method described in section 3.4, the connected components of the synonymy relations obtained in section 3.2 are segmented again into one or more smaller and more homogeneous groups. The number of groups varies between 745 and 1,798 across the experiments. Moreover, around 25% of the synonymy relations may be removed by pagerank. These connected components and groups can also be evaluated against the reference data and we can compute the precision.

5 Evaluation and Discussion

The evaluation has been done by two independent experts, with the Cohen's kappa inter-expert agreement 0.75. We exploit the reference data of the two experts separately (we distinguish expert_1 and expert_2) and in common. We also distinguish macro-precision and micro-precision. Finally, the precision is first evaluated after the induction step with the NLP methods, and then after the processing of the acquired synonymy relations through the pagerank-derived algorithm and their filtering.

For the weighting of the non-synonymy and synonymy relations, we tested and applied several coefficients: 5, 3 and 5 in experiment 5 (*syn-isa535*); 3, 5 and 3 in experiment 6 (*syn-isa353*), etc. Different weights have been tested ranging from 1 to 7, as well as the log variations. On the whole, these variations have no significant impact on the results. But then, it is very important to respect the dependence among these coefficients and not to set them randomly.

The filtering of the synonymy relations has to control two factors: (1) the first is related to the fact that the removed relations are to be true negatives and that among them there should be no or a small number of correct relations; while (2) the second is related to the fact that the remaining relations are to be true positives and that among them there should be no or a small number of wrong relations.

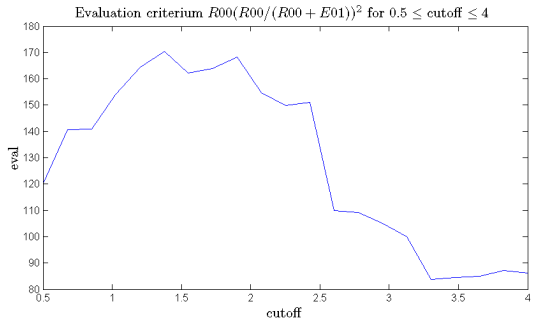
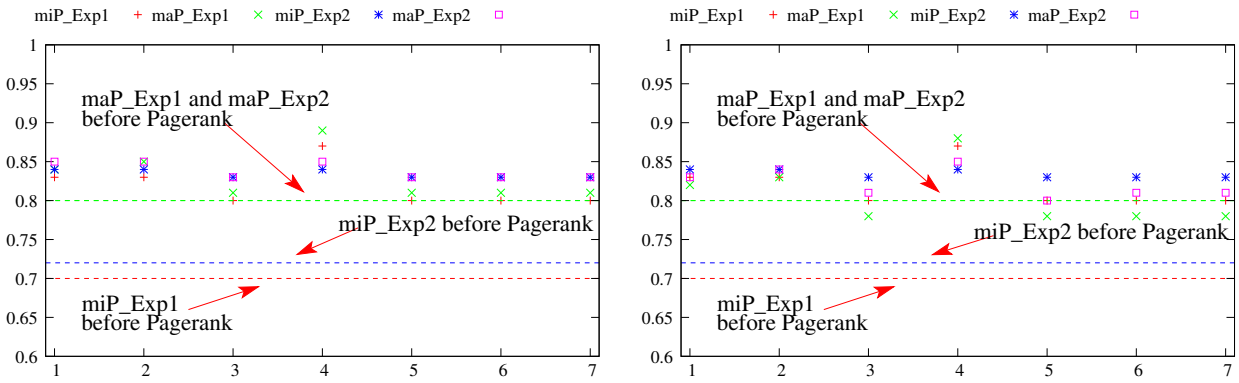


Figure 4: Impact of the cutoff values on the filtering of synonymy relations



(a) Connected components with terms

(b) Groups of terms

Figure 5: Evaluation of the results in terms of micro-precision miP and of macro-precision maP for connected components and for groups of terms (performed according to the reference data provided by two experts)

On figure 4, we present the impact of the cutoff values on the selection and filtering of the synonyms. Like with other parameters, we have tested several values between 0.5 and 4. This figure illustrates the distribution of the correctly removed relations. The cutoff values have an important impact on the results: we can observe that the optimal cutoff values are set between 1.5 and 2 because they allow to remove the highest number of the wrong relations. We have set the cutoff value to 1.5. The choice of cutoff is an important factor for the definition of the amount of the links that are to be removed: the higher the *cutoff* the higher the number of clusters. On the data processed in this work, the cutoff value has been defined experimentally thanks to the observation of the processed data. For the generalization of this method to new unknown but similar linguistic data (new terminology, new langage, new domain...), the *cutoff* will be either set in order to remove a certain predefined number of links or will be defined from a typical sample of the data.

Contrary to the cutoff values, the choice of threshold doesn't greatly impact the results, although using a lower threshold makes it harder to choose a good cutoff values since the ranking of different nodes will be closer to each other.

As for the analysis of the precision and of the relations which are correctly kept within the connected components, let's observe figure 5. On this figure, we present the evaluation results performed within the connected components with induced synonyms (figure 5(a)) and within the groups of filtered synonyms (figure 5(b)). On the y-axis we indicate the precision values, and on the x-axis, we indicate the different experiments performed as mentioned above: 1 in which only synonyms are exploited, 2 in which synonyms are combined with hierarchical `is-a` relations, 3 in which synonyms are combined with lexical inclusions, etc. Horizontal lines correspond to the precision obtained before the application of the pagerank: they remain the same whatever the experiment. These lines correspond to three reference data provided by the expert$_1$, the expert$_2$ and by their common data. As for the points, they indicate the precision obtained further to the pagerank: it varies according to experiments and experts. On the basis of figure 5, we can observe that:

- the difference between the expert evaluations is very low (0.02);
- the pagerank allows to increase the precision (between 0.10 and 0.15 for micro-precision, while macro-precision varies by 0.05);
- the consideration of synonymy alone provides performant results;
- the consideration of `is-a` relations improves the results but lexical inclusions decrease them;
- the increased weight of some of the quality indicators has no effect on the evaluation;
- macro-precision is superior to micro-precision because our data contain mainly small groups,

while the few large connected components have a very low precision;

- there is but a small difference between connected components (figure 5(a)) and groups (figure 5(b));
- the consideration of `is-a` relations and of lexical inclusions provides the best precision but the amount of the remaining synonyms is then the lowest. As we explained, it is important to keep the highest number of the correct relations, although when a lot of relations is removed, it is logical to obtain a higher precision. This means that the combination of `is-a` relations and of lexical inclusions is not suitable because it removes too much of synonyms.

In relation with this last observation, is should be noted that the balance between the removed and the remaining relations is a subtle parameter.

The obtained results indicate that the pagerank is indeed useful for the filtering of synonyms, although the parameters exploited by this algorithm must be defined accurately. Thus, it appears that synonymy alone may be sufficient for this filtering. When the quality indicators are considered, `is-a` relations are suitable for this filtering because very often they propose true hierarchical relations. However, the lexical inclusions have a negative effect of the filtering. We assume this is due to the fact that the lexical inclusions are ambiguous: they may convey hierarchical relations but also equivalence relations (Haralambous and Lavagnino, 2011). Indeed, contextually some terms may be shortened or may be subject to an elision while their meaning is not impacted.

Currently, the pagerank is limited by the fact that it is applied to a relatively small set of data while it is designed to process very large data. Then, it can be interesting to enrich the model and to be able to take into account other quality indicators, such as frequencies, productivity or other semantic relations proposed within *GO* (`part-of` and `regulates`). Moreover, we can also give a lesser weight to some indicators (such as lexical inclusions) with penalties and keep the strong weight for other indicators. In the current model of the pagerank, we threshold rows to < 0.95. However, we assume that the algorithm may have problems with very large and very connected graphs: the pagerank may spread out in the graph too much and possibly allow the first words with the highest pagerank to make groups with only one word. This can be corrected if an additional calculation is added and when the group contains only one word at step 5.

6 Conclusion and Perspectives

We propose an original approach for inducing synonyms from terminologies and for their filtering. The methods exploit the NLP methods, compositionality principle and pagerank-derived algorithm. This work is motivated by the fact that synonymy is a contextual relation and its validity and universality are not guaranteed. We assume the semantic cohesiveness of synonymy relations should be qualified and quantified. The compositionality and NLP methods allow to acquire endogeneously the synonymy relations and the quality indicators, while the pagerank-derived algorithm leads to the filtering of the acquired synonyms. Its functionning is based upon the synonymy relations and also upon the acquired indicators (*is-a* relations and lexical inclusions). It appears that the synonymy relations alone provide good clues for their filtering. The *is-a* relations are also fruitful, while the use of the lexical inclusions appears not to be suitable.

In the future, we plan to add and test other indicators. Other experiments will also be done with the pagerank approach. For instance, it will be interesting to propose a model which takes into account that, within a cluster, words may be synonym with some cluster words but not with all the words of the cluster. This method can be adapted for the processing of corpora and also applied to terms from other terminologies. The acquired and filtered synonymy relations will be exploited within the NLP applications in order to test the efficiency of these resources and also the usefulness and efficiency of their filtering. Moreover, the compositionality approach can be adapted and exploited for the paraphrasing of the biomedical terms and for the improvement of their understanding by non expert people.

References

E Agirre and A Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *EACL 2009*, pages 33–41, Athens, Greece, March.

O Bodenreider, A Burgun, and TC Rindflesch. 2001. Lexically-suggested hyponymic relations among medical terms and their representation in the UMLS. In URI INIST CNRS, editor, *Terminologie et Intelligence artificielle (TIA)*, pages 11–21, Nancy.

S Brin and L Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117.

K Bryan and T Leise. 2006. The $\$25,000,000,000$ eigenvector: the linear algebra behind google. *SIAM Rev.*, 48(3):569–581.

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

C Engström. 2011. Pagerank as a solution to a linear system, pagerank in changing systems and non-normalized versions of pagerank. Master's thesis, Mathematics, Centre for Mathematical sciences, Lund University. LUTFMA-3220-2011.

C 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.

S Fernandez, E SanJuan, and JM Torres-Moreno. 2009. Résumés de texte par extraction de phrases, algorithmes de graphe et énergie textuelle. In *Société Francophone de Classification*, pages 101–104.

B Gaume. 2006. Cartographier la forme du sens dans les petits mondes lexicaux. In *JADT*.

G Grefenstette. 1994. *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers.

T Hamon and N 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.

T Hamon, A Nazarenko, and C Gros. 1998. A step towards the detection of semantic variants of terms in technical documents. In *COLING-ACL'98*, pages 498–504.

Y Haralambous and E Lavagnino. 2011. La réduction de termes complexes dans les langues de spćialité. *TAL*, 52(1):37–68.

S Hassan, R Mihalcea, and C Banea. 2007. Random-walk term weighting for improved text classification. In *ICSC*, pages 242–249.

T Hughes and D Ramage. 2007. Lexical semantic relatedness with random graph walks. In *EMNLP-CoNLL*, pages 581–589. Association for Computational Linguistics.

C Jacquemin, JL Klavans, and E Tzoukerman. 1997. Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *ACL/EACL 97)*, pages 24–31, Barcelona, Spain.

G Kleiber and I Tamba. 1990. L'hyperonymie revisitée : inclusion et hiérarchie. *Langages*, 98:7–32, juin.

R Mihalcea, P Tarau, and E Figa. 2004. Pagerank on semantic networks, with application to word sense disambiguation. In *COLING*, pages 1126–1132.

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.

BH Partee, 1984. *Compositionality*. F Landman and F Veltman.

S Schulz, M Romacker, P Franz, A Zaiss, R Klar, and U Hahn. 1999. Towards a multilingual morpheme thesaurus for medical free-text retrieval. In *Medical Informatics in Europe (MIE)*, pages 891–4.

F Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

N Shimizu, M Hagiwara, Y Ogawa, K Toyama, and H Nakagawa. 2008. Metric learning for synonym acquisition. In *COLING*, pages 793–800.

R Sinha and R Mihalcea. 2007. Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In *IEEE International Conference on Semantic Computing (ICSC 2007)*, pages 363–369.

RS Sinha and RF Mihalcea. 2011. Using centrality algorithms on directed graphs for synonym expansion. In *FLAIRS*.

Y Tsuruoka, Y Tateishi, JD Kim, T Ohta, J McNaught, S Ananiadou, and J Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. *LNCS*, 3746:382–392.

CM Verspoor, C Joslyn, and GJ 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.

T Wang and G Hirst. 2011. Exploring patterns in dictionary definitions for synonym extraction. *Natural Language Engineering*, 17.

P Zweigenbaum, R Baud, A Burgun, F Namer, É Jarrousse, N Grabar, P Ruch, F Le Duff, B Thirion, and S Darmoni. 2003. Towards a Unified Medical Lexicon for French. In *Medical Informatics in Europe (MIE)*, pages 415–20.