Combination of endogenous clues for profiling inferred semantic relations: experiments with Gene Ontology

Natalia Grabar[1,2], PhD, Marie-Christine Jaulent[1], PhD, Thierry Hamon[3], PhD

[1]**Centre de Recherche des Cordeliers, Université Pierre et Marie Curie - Paris6, UMR_S 872, Paris, F-75006; Université Paris Descartes, UMR_S 872, Paris, F-75006; INSERM, U872, Paris, F-75006 France**

[2]**HEGP AP-HP, 20 rue Leblanc, Paris, F-75015 France**

[3]**LIPN – UMR 7030, Université Paris 13 – CNRS, 99 av. J-B Clément, F-93430 Villetaneuse, France**

Abstract

Acquisition and enrichment of lexical resources is acknowledged as an important research in the area of computational linguistics. While such resources are often missing, specialized domains, i.e. *biomedicine, propose several structured terminologies. In this paper, we propose a high-quality method for inferring elementary synonym lexicon through a structured terminology. The method is based on the analysis of syntactic structure of complex terms. The inferred synonym pairs are then profiled according to different clues endogenously computed within the same terminology. We apply and evaluate the approach on the Gene Ontology biomedical terminology.*

Introduction

The decision as to whether two terms (*i.e.*, *acetone anabolism* and *acetone biosynthesis*) convey the same or different meaning is important for various applications of the biomedical informatics, especially for deciphering and computing semantic similarity between words and terms within tasks like query expansions, information retrieval, knowledge extraction or terminology matching. Lexicon of synonyms and of morphological or orthographic variants can be used for this. Depending on languages and domains, such resources are not equally well described. The morphological description of languages is the most complete thanks to databases like Celex[1] for English and German, MorTal[2] for French, UMLS Specialized Lexicon[3] for medical English, and similar resources for German[4] and French[5]. At the semantic level, when one looks for the description of synonym relations, little available resources can be found: WordNet[6] proposes synonym relations for English, but the corresponding resources for other languages are not freely available; and the initiative for fitting this resource for the biomedical area[7] is still ongoing. Otherwise, various existing biomedical terminologies (*i.e.*, Gene Ontology,[8] Snomed,[9] MeSH[10] or UMLS[3]) provide complex terms, which use, compared to the lexical resources, is less suitable for the biomedical applications. In our work, we aim at filling this gap in specialized domains. We propose to use the existing terminologies in order to infer a lexicon of elementary synonyms specific to the biomedical area and/or one of its sub-domains. Such synonyms are indeed often "hidden" within complex terms. The proposed method exploits the compositionality of complex terms extracted from structured terminologies and is based on the identification of their syntactic invariants. We position our research in the domain of biology. In order to automatically evaluate the validity of the inferred synonyms we intend to profile them through exploiting endogenous information acquired within the same terminology.

Material

The goal of the Gene Ontology (*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 used version 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.

Methods

Often within *GO*, terms are coined on the same 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

As the compositionality has been recognized to be a characteristic feature of the GO, [11,12,13] we propose to exploit it in order to induce paradigms of elementary synonyms (*i.e.*, *biosynthesis*, *anabolism*, *formation*, *synthesis*). Like in the given examples, our method exploits compositional structure of terms and relies on existence of structured terminologies. The notion of

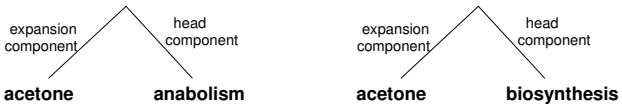

Figure 1: Parsing tree of the terms *acetone anabolism* and *acetone biosynthesis*.

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.[14] The syntactic analysis of terms is crucial: it normalizes the representation of terms through their head and expansion components and it prepares the syntactic dependencies computing.

Preprocessing of *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,[a] suitable for the processing of large amounts of data and tunable to specialized areas. Through the platform, we perform: recognition of named entities[15] (*i.e.*, gene names, chemical products); segmentation into words and sentences; POS-tagging and lemmatization;[16] syntactic analysis.[b] 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 (see fig. 1) composed of two elements: head component and expansion component. For instance, *anabolism* is the head component of *acetone anabolism* and *acetone* is its expansion component.

Acquisition of elementary synonymous relations
We adapt our previous work[17] to inferring elementary synonym relations between simple terms. Thus, if the meaning $\mathcal{M}$ of two complex terms $A\ rel\ B$ and $A'\ rel\ B$ are 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 synonym relations is based on the three rules:

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

R2 If both terms are synonymous and their head components are identical, then an elementary synonym relation is inferred: the pair {*endocytic*, *endocytotic*} is inferred from the synonym 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 synonyms, then an elementary synonym relation is inferred: the pair {*nicotinamide adenine dinucleotide*, *NAD*} is inferred from the synonym relation between *nicotinamide adenine dinucleotide catabolism* and *NAD breakdown* where the head components {*catabolism*, *breakdown*} are already known synonyms.

The method is recursive and each inferred elementary synonym relation can then be propagated in order to infer new elementary relations, which allows to generate a more exhaustive lexicon of synonyms.

Profiling of the inferred resource of synonyms
We have previously evaluated the inferred elementary synonyms of *GO* through their comparison with the general language synonyms from the WordNet synsets. This comparison showed mainly that the overlap is very low: any of the inferred synonym sets could be completely matched with any of the synsets, although we could find partial overlapping between them.[18] Then, an additional evaluation was performed manually: each inferred pair was examined as well as its source synonyms. Over 90% accuracy of the inferred pairs was thus found. Quality of the results is high which is due to the method and data used but also to the contextual nature of synonymy[19,6]: as far as we can observe at least one context within which terms are synonymous we record these terms as true synonyms. But sometimes this can lead to questionable results: {*cell*, *lymphocyte*}, {*T-cell*, *T-lymphocyte*}, {*binding*, *DNA binding*} are found to be synonyms and their original terms support this finding, although the intuition would suggest they are probably not. In the current work, we propose to combine several types of endogenously generated clues for profiling the inferred lexicon and for defining the reliability zones:

- Elementary synonym relations are generated, according to the method described, within original *GO* terms linked through synonymy relations;

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

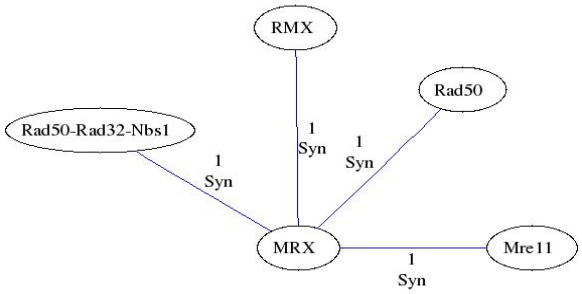

Figure 2: The *MRX* connected component.

- For each inferred pair its productivity within *GO* (or number of original pairs from which this elementary relation is inferred) is computed. For instance, {*binding*, *DNA 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: {*T-cell*, *T-lymphocyte*} pair appears to be more reliable;
- Terms within each inferred pair are controlled for the lexical inclusion.[20] 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;
- The same compositional method is applied to original *GO* term pairs related through `is-a` and `part-of` relations. Thus, we can infer `is-a` and `part-of` elementary relations. If a pair of inferred synonyms is also inferred through `is-a` or `part-of` relations, *i.e.* {*binding*, *DNA binding*}, this would weaken this synonymy relation.

In summary, the co-occurrence of a synonymy relation with the lexical inclusion or with `is-a` or `part-of` relations is supposed to weaken it, as well as its lower productivity.

Results and Discussion

23,899 *GO* terms have been fully parsed through the Ogmios platform. The three compositional rules have been applied and allowed to infer elementary synonyms (n=921), `is-a` (n=1,273) and `part-of` pairs (n=178). Productivity of the inferred synonyms within original complex *GO* terms and their lexical inclusions have been computed.

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, figure 2 contains five elementary synonyms (*MRX*, *Rad50-Rad32-Nbs1*, *RMX*, *Rad50* and *Mre11*) inferred from the *GO* concept GO:0030870 which preferred term is *MRX complex*. Synonym relations are labelled *Syn* on the figures. In this CC, the preferred elementary term *MRX* is linked to its synonyms – this CC is star-shaped. Observing synonyms through their CCs presents their semantics in a more global view. Thus, the contextual nature of synonyms, which can alter their universality, can be more easily detected within CCs. We will analyze several CCs and profile them through the proposed endogenous clues.

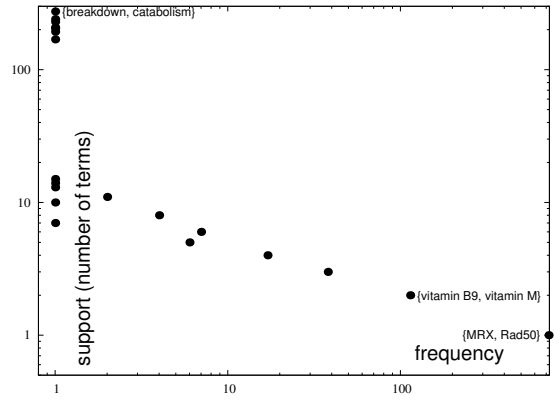

Figure 3: Productivity of the inferred elementary synonyms within *GO* (logarithmic scale).

Productivity of the elementary synonyms

Figure 3 (scaled logarithmically) represents the productivity of the inferred synonyms through their 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 acception 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 acception 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 inferred synonyms and their validity cannot rely only on this clue.

Lexical inclusion

Further to the lexical inclusion test, 84 elementary synonym pairs appear to present such relation. They are labelled *H(IL)* on figures. For instance, the left CC of fig. 4 contains four synonyms. The link {*death*, *cell death*} is inferred from terms *synergid death* and *synergid cell death* (GO:0010198), but it presents also

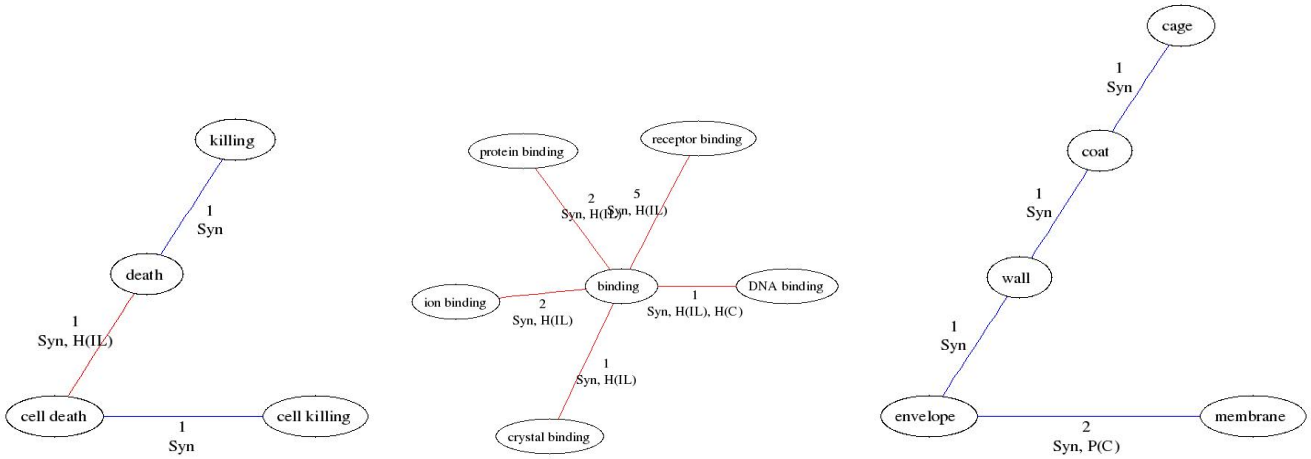

Figure 4: Connected components with lexical inclusions, hierarchical and part-of relations (possible weak points of connected components).

the lexical inclusion. We noticed that synonymy is a contextual relationship: in the example {*death*, *cell death*}, original terms as well as the elementary synonyms convey the same semantics. The elision of *cell* within *synergid death* seems to denote a kind of implicit knowledge: all the biological functions and processes are located at the cell level. As for the middle CC of fig. 4, *binding* contains only lexically subsumed relations with its synonyms. If they are correct in the original pairs of *GO* synonyms, their co-occurrence with the lexical inclusions weaken them, especially if a generalization of these relations is planned. Indeed, if these relations were to be used in a different context comparing to their source, lexical inclusions would certainly indicate the possible weak points, especially if productivity of the relations is low. For instance, the left CC can be split on the weaker relation {*death*, *cell death*} which would lead to two (sub)CCs: {*death*, *killing*} and {*cell death*, *cell killing*}.

Elementary `is-a` **and** `part-of` **relations**

Inferring elementary `is-a` and `part-of` relations allows to find out that 11 elementary synonyms are also inferred through `is-a` pairs, and 3 more from `part-of` pairs. They are labelled *H(C)* and *P(C)* on figures. Any of these relations have been validated: they are used as possibly relevant clues for the profiling the inferred synonyms. We can observe, on the middle CC of fig. 4, that {*binding*, *DNA binding*} pair is inferred within `is-a` relation in addition to have been inferred within *GO* synonyms and to show the lexical inclusion relation. On the right CC of fig. 4, the pair {*envelope*, *membrane*} has been inferred within two *GO* original synonym pairs; at the same time, it has been observed in four *GO* term pairs related with `part-of` relation. This decreases the validity of this synonymy relation.

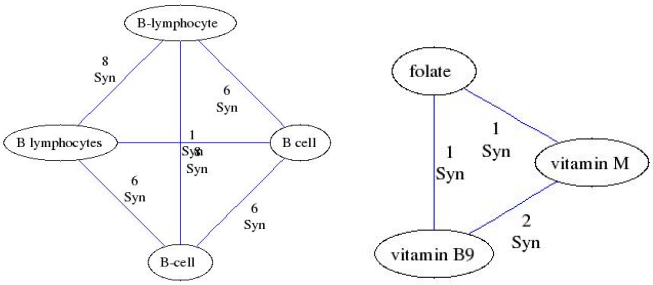

Figure 5: Cliques, or strongly connected components, convey strong synonymy relations.

These various clues can be useful for profiling the inferred synonym pairs of simple synonyms and, in this way, to associate them with their confidence rate. The use of this rate can be helpful within both manual and automatic approaches. In case of manual validation by human experts, this profiling can provide helpful indications and decrease the time necessary for the validation. In case of applying automatic heuristics, this profiling can be used for the definition of threshold rates. Use of these clues is indeed complementary to the analysis of the original synonym pairs, upon which the first validation relied. Another observation is concerned with shape and chaining of CCs. For instance, a CC, which nodes are strongly connected among them (fig. 5), may be more reliable. While graphs, that are star or path-shaped (fig. 4), may indicate that the substitution between their nodes is less suitable.

Conclusion and Perspectives

In this paper, we propose a novel method for inferring simple synonyms from structured terminologies in order to help the natural language processing-based applications. This method exploits the compositionality principle and three rules based on syntactic dependency analysis of terms. We noticed that synonymy is a contextual relation and, for this reason, validity and universality of the inferred pairs should be verified. Thus, we propose several clues for profiling the inferred synonymy relations and for detecting possible weak points. All clues are generated endogenously: they are acquired on the same source terminology from which the synonymy lexicon is inferred. These clues can be used for preparing the manual validation or automatic filtering steps. Moreover, use of endogenous clues for the pre-validation step is suitable especially as no reference resource exists. These first experiments have been performed on Gene Ontology but the method is easily applicable to other structured terminologies. For a more precise profiling, the four relationships of *GO* (synonymy, `is-a`, `part-of` and `regulates`) can be cross-validated, while currently, we aim at validating synonymy through `is-a` and `part-of` relations (and other clues). The topography (*i.e.*, shape and chaining of nodes) of connected components should be studied more thoroughly, as well as the density of links within each graph. We plan also to use the inferred relations and propagate them through corpora and discover some of the missing synonyms.[21] In this way, applying the same compositionality principle, we can enrich and extend the Gene Ontology: new synonym or other relations between GO terms can be detected. From a more ontological perspective, our method can be used for the consistency checking of a terminology.[12] Besides, it can provide data for transforming a pre-coordinated approach into a post-coordinated one.

REFERENCES

1. Burnage G. *CELEX - A Guide for Users*. Centre for Lexical Information, University of Nijmegen, 1990.
2. Hathout N, Namer F, and 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, et al. Towards a multilingual morpheme thesaurus for medical free-text retrieval. In: Medical Informatics in Europe (MIE), 1999.
5. Zweigenbaum P, Baud R, Burgun A, et al. 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* 1998;32(2-3):209–20.
7. Smith B and Fellbaum C. Medical wordnet: a new methodology for the construction and validation of information. In: Proc of 20th CoLing, Geneva, Switzerland. 2004:371–82.
8. Gene Ontology Consortium . Creating the Gene Ontology resource: design and implementation. *Genome Research* 2001;11:1425–33.
9. Cté RA, Brochu L, and Cabana L. SNOMED Internationale – Répertoire d'anatomie pathologique. Secrétariat francophone international de nomenclature médicale, Sherbrooke, Québec, 1997.
10. National Library of Medicine, Bethesda, Maryland. Medical Subject Headings, 2001. *http://www.nlm.nih.gov/mesh/meshhome.html*.
11. Verspoor CM, Joslyn C, and Papcun GJ. 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, 2003:51–6.
12. Mungall C. Obol: integrating language and meaning in bio-ontologies. *Comparative and Functional Genomics* 2004;5(6-7):509–20.
13. Ogren P, Cohen K, and Hunter L. Implications of compositionality in the Gene Ontology for its curation and usage. In: Pacific Symposium of Biocomputing, 2005:174–85.
14. Partee BH. Compositionality. F Landman and F Veltman, 1984.
15. Berroyer JF. Tagen, un analyseur d'entités nommées : conception, développement et évaluation. Mmoire de D.E.A. d'intelligence artificielle, Universit Paris-Nord, 2004.
16. Tsuruoka Y, Tateishi Y, Kim JD, et al. Developing a robust part-of-speech tagger for biomedical text. *LNCS* 2005;3746:382–92.
17. Hamon T and Nazarenko A. Detection of synonymy links between terms: experiment and results. In: *Recent Advances in Computational Terminology*. John Benjamins, 2001:185–208.
18. Hamon T and Grabar N. Acquisition of elementary synonym relations from biological structured terminology. In: Computational Linguistics and Intelligent Text Processing (5th International Conference on NLP, FinTAL 2006), number 4919 in LNCS. Springer, 2008:40–51.
19. Cruse DA. *Lexical Semantics*. Cambridge University Press, Cambridge, 1986.
20. Bodenreider O, Burgun A, and Rindflesch TC. Lexically-suggested hyponymic relations among medical terms and their representation in the UMLS. In: URI INIST CNRS , ed, Terminologie et Intelligence artificielle (TIA), Nancy. 2001:11–21.
21. Hole W and Srinivasan S. Discovering missed synonymy in a large concept-oriented metathesaurus. In: AMIA 2000, 2000:354–8.