Exploitation of linguistic indicators for automatic weighting of synonyms induced within three biomedical terminologies

Natalia Grabar[a,b], Thierry Hamon[c]

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

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

[c] *LIMBIO (EA3969), UFR SMBH Léonard de Vinci, Université Paris 13, 74, rue Marcel Cachin, 93017 Bobigny Cedex France*

Abstract

Acquisition and enrichment of lexical resources is an important research area for the computational linguistics. We propose a method for inducing a lexicon of synonyms and for its weighting in order to establish its reliability. The method is based on the analysis of syntactic structure of complex terms. We apply and evaluate the approach on three biomedical terminologies (MeSH, Snomed Int, Snomed CT). Between 7.7 and 33.6% of the induced synonyms are ambiguous and cooccur with other semantic relations. A virtual reference allows to validate 9 to 14% of the induced synonyms.

Keywords:

Natural language processing, Semantics, Vocabulary, Terminology, Medical Informatics, UMLS.

Introduction

Within the biomedical area, practitioners and institutions may use different terms, which can convey the same or a close meaning. For example, the terms heart attack, myocardial infarction, and MI present the same meaning to a medical expert, while these expressions remain different to a computer, unless suitable resources and tools are available and used. The purpose of these resources and tools is to compute the semantic similarity between terms and to guarantee semantic interoperability between automatic systems. Such need appears whenever applications like information exchange and retrieval, knowledge extraction, terminology matching are addressed. Lexica of synonyms and of morphological or orthographic variants are typically used for the computing of semantic similarity. Depending on languages and domains, these lexica 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 level of synonyms, little available resources can be found: WordNet [6] proposes synonym relations for English, but the corresponding resources for other languages are not freely available. Otherwise, various existing biomedical terminologies provide complex terms, but their use is less suitable for the biomedical applications [7].

In a previous work, we proposed a method for filling the gap and for acquisition of synonymy resources within biological area: we used an existing structured terminology Gene Ontology [8] in order to induce a lexicon of elementary synonyms. The induced synonyms were then profiled through endogenous information acquired within the same terminology [9]. In the current work, we propose to generalize this method and to apply it to three other biomedical terminological resources: MeSH [10], Snomed Int [11] and Snomed CT [12]. Since synonyms are a contextual phenomena and they may convey more or less close or ambiguous meaning, we propose also a method for transformation of linguistic profiling indicators into numeric values, which are to be used to automatically weight the acquired synonyms. The objective of this part of work addresses the degree of semantic similarity and reliability of synonyms.

Material: semantic relations between terms

Our material is provided by three biomedical terminological resources: MeSH [10], Snomed Int [11] and Snomed CT [12]. These three terminologies are generic to the biomedical area: they propose its general descriptions, although they aim at satisfying different needs. The goal of the MeSH thesaurus is to provide a terminological resource for information retrieval. The goal of the Snomed Int nomenclature is to help the computerization of clinical data. Finally, the goal of the Snomed CT nomenclature is to provide terminological resource for organizing and, more particularly, for exchanging clinical data.

These three terminologies are structured: their terms are related among them with various semantic relationships. We access this information through the UMLS [3], version 2008AB. We extract the semantic relations according to their broad categories as they are defined by the UMLS. These categories are the following: AQ (allowed qualifier), CHD (has

child), DEL (deleted concept), PAR (has parent), QB (can be qualifier by), RB (has a broader relationship), RL (has similar or like relationship), RN (has narrower relationship), RO (has relationship other than synonymous, narrower or broader), RQ (related and possibly synonymous), SIB (has sibling), SY (source-asserted synonymy). These UMLS categories of relationships are assigned on the basis of the source documentation or on the basis of the NLM understanding of the sources. For extraction of our material, we focus on four categories of relationships:

- synonymy relations provide identical or similar meanings. They are extracted within UMLS concepts and correspond to the category SY which links preferred term labels to their synonyms;
- is a relations provide the hierarchical structure for terms. We consider that they are indicated by four broad categories: PAR, RB, CHD or RN;
- sibling relations link terms that have the same hierarchical father. They are indicated by SIB category;
- associative relations may convey various kinds of semantic relations. We consider they are indicated by RO category.

The extracted terms related by these relationships are always restricted to the corresponding terminology. is a, sibling and associative relations take into account preferred and synonymous labels of terms.

Methods

Inducing and profiling synonymy relations

In order to induce and to profile resources of synonymy relations, we applied the method described in previous work [9]. Here, for the sake of clarity, we will mention the general principles of the proposed approach.

Biomedical terms are often coined on the same syntactic scheme and show the compositionality through the substitution of one of their components (underlined):

infection of navel cord; infection of umbilical stump

benign tumour of scrotal skin; benign neoplasm of scrotal skin

We proposed to exploit the compositionality and to induce paradigms of elementary semantic relations (i.e., {*navel cord, umbilical stump*}, {*tumour, neoplasm*} in the examples above). Compositionality of biomedical terms has been exploited previously, especially through Gene Ontology, for consistency checking [13], for adding missing synonym terms [14] or for deriving simple graphs from relations between complex terms [15]. While the cited works are based on the string matching within terms, our approach aims at exploiting the syntactic analysis of terms, according to the compositionality definition [16]: the meaning of a complex expression is fully determined by its syntactic structure, the meaning of its parts and the composition function. We assume that relationship between elementary terms is inherited from the relationship between complex terms having the same syntactic schema and components at the word or semantic level. In this work, we apply the method to several relationships: synonymy, is a, sibling and associative.

Terms are processed through the Ogmios NLP platform[1], and are syntactically analyzed by a dedicated term parser: syntactic dependencies between term components are computed according to assigned POS tags [17] and shallow parsing rules[2]. Thus, each term is considered as a syntactic binary tree (see fig. 1) composed of two elements: head component and expansion component. For instance, *infection* is the head component of *infection of navel cord* and *navel cord* is its expansion component. According to the compositionality principle, the synonymy terms from figure 1 enrich synonym lexicon with {*navel cord, umbilical stump*}. In these two terms, the variation occurs within the expansion components. Besides, the variation can also occur within head components, or even within both components (head and expansion). Each of these cases will be exploited for inducing semantic relations.

Figure ! - Parsing tree of the synonym complex terms infection of navel cord and infection of umbilical stump

However, semantic relationships as synonymy, are contextual [18]: for a given relation, its profile can vary according to contexts of its instances. In order to help the NLP to exploit such resources, we profile the induced synonymy relations through several types of linguistic indicators generated within the same terminologies:

- Cooccurrence of several elementary semantic relations induced by our approach;
- Lexical inclusion controlled within each induced synonymy pair, because lexical inclusions may convey a hierarchical relation: in the pair {*arterial embolism, embolism*}, *arterial embolism* is a kind of *embolism*;
- Productivity (or number of original pairs from which an elementary relation is inferred) for each induced semantic relation, including lexical inclusion.

Weighting and evaluating induced synonyms

The linguistic indicators (productivity, lexical inclusion, cooccurrence of semantic relation) will be used for automatic computing of weights for each induced synonymy relation. Currently, these indicators are descriptive and symbolic: they are meaningful to human users, but they have no exploitable meaning to a computer. In that respect, we have to: (1) transform the symbolic indicators into numeric values, and (2) pro-

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

pose an approach for combination of these values into a weight associated to each synonymy relation.

According to our general observation, reliability of the induced synonymy relations is closely related to its profile: productivity and cooccurrence with other semantic relations. For computing the numeric weight and reliability of each synonymy relation rel_i, we propose to sum weights of all the cooccurring indicators. The weight of each indicator corresponds to the product of its productivity $prod_j$ and coefficient α_j. The general formula is the following:

$$weight(rel_i) = \sum_{j \in \{syno, is\_a, asso, sib, incl\}} \alpha_j \times prod_j(rel_i)$$

Values of coefficients α_j were determined empirically, they are amplified by their productivity values.

- α_{syno} was set to 1: it is the highest value established, which gives more reliability to a given relation.
- Since is a relation weakens the synonymy reliability, its value α_{is_a} was set to 1.
- Lexical inclusion may convey both hierarchical relation, like is a, and synonymy through the elision phenomena. Its value α_{incl} was thus set to 0.5.
- associative and sibling relations also weaken reliability of synonymy but to a lesser extend than is a: there values α_{asso} and α_{sib} were set to 0.75.

With such set of α values, positive weights signify more reliable synonymy relations. The reliability increases as the positive values increase.

There is no gold standard for the evaluation of a lexicon of synonyms within the biomedical area: the only available WordNet resource appears to be unsatisfying [19, 20]. Here again, we propose to take advantage of the exploited terminologies in order to evaluate our results. We will generate a *virtual truth*: set of synonyms induced by our method, which are already present in the exploited terminologies.

Results and Discussion

Building the material

In table 1, we give indications on volume of material available in UMLS for the three processed terminological resources: numbers of terms (labels) and of the corresponding CUIs, and numbers of the extracted semantic relations (synonymy, is a, sibling and associative). We can observe that Snomed Int provides low number of semantic relations, but it has also the lesser number of terms. While MeSH and Snomed CT propose a richer network of relations and of the involved terms. Otherwise, sibling relationship is proposed only by MeSH.

Table 1 - Number of terms (labels and CUIs) and number of semantic relations (synonymy, is a, sibling, associative) provided by three exploited terminologies

	MeSH	Snomed Int	Snomed CT
number of terms	684,211	164,180	1,143,186
number of CUIs	291,746	112,709	313,612
Synonymy	469,847	57,111	399,712
is a	1,627,703	237,702	2,496,097
Sibling	7,870,078	—	—
Associative	265,178	213,108	6,166,776

Inducing and profiling the synonymy relations

All the semantic relations among complex terms from the three processed terminologies have been fully analyzed through the Ogmios platform. Compositional rules have been applied and allowed to induce elementary synonymy, is a, sibling and associative relations. Numbers for each type of the induced relations within each terminology are indicated in table 2. Lexical inclusions have been controlled for each synonymy relation: they are indicated in table 2, line *l.inclusion*. The last two lines of the table indicate the number of synonymy relations which cooccur with other semantic relations, and their percentage. Productivity of the induced relations within original complex terms have been also computed.

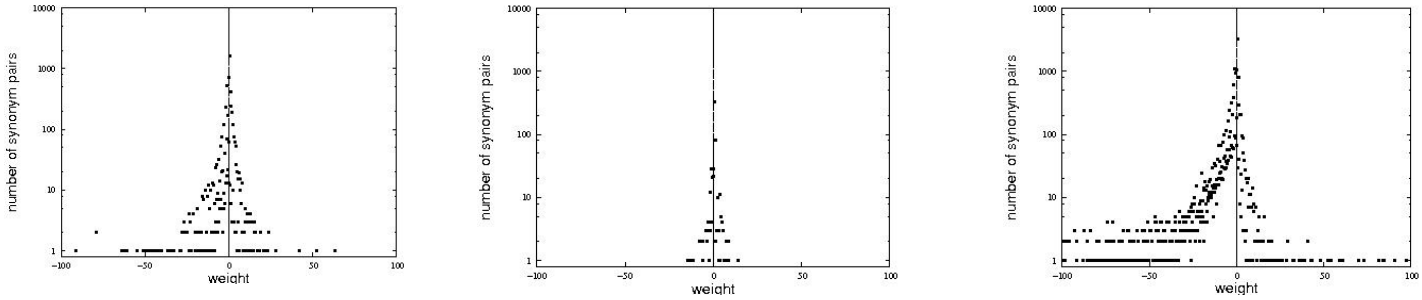
Table 2 - Number of induced semantic relations (synonymy, is a, sibling, associative and lexical inclusion) in three exploited terminologies, and number of ambiguous synonymy relations

	MeSH	Snomed Int	Snomed CT
Synonymy	29,741	7,950	39,921
is a	53,015	3,906	127,197
Sibling	(142,360)	—	—
Associative	4,623	2,248	96,862
l.inclusion	7,777	999	28,633
common (number)	3,847	611	13,409
common (%)	12.9%	7.7%	33.6%

7.7% of synonymy relations induced within the Snomed Int are cooccurring and ambiguous with other induced semantic relations, while within the MeSH ambiguous synonymy relations are more frequent (12.9%). As MeSH is the only terminology that proposes sibling relationship, these are not taken into account. If they are, number of ambiguous synonymy relations is 6,809 (22.9%). The highest ambiguity is observed within Snomed CT: up to 33.6%.

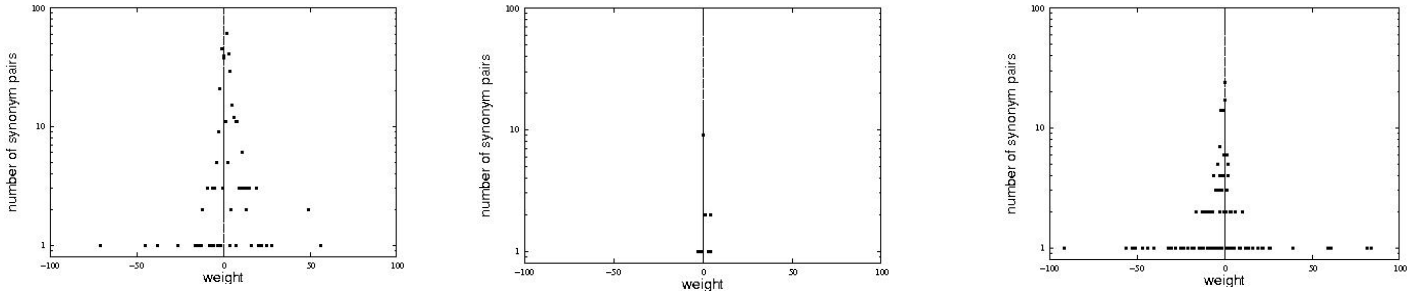
Weighting and evaluating induced synonyms

Weights of the synonyms induced within the three processed terminological resources have been computed according to the proposed formula. Figure 2 indicates distribution of these weights (x-axis) for synonyms that cooccur with other semantic relations. The central vertical line materializes the frontier between positive and negative weights. The y-axis of the figure is algorithmically scaled and indicates number of synonym pairs that show a given weight. For instance, within Snomed



MeSH Snomed Int Snomed CT

Figure 2 - Weights of induced synonymy relations cooccurring with other semantic relations



MeSH Snomed Int Snomed CT

Figure 3 -Weights of induced synonymy *relations within the set of the* virtual truth *(subset of relations from fig.* ***Erreur ! Source du renvoi introuvable.****)*

Int, 2 pairs of synonyms -{*bleeding, haemorrhage*} and {*bleeding, hemorrhage*}- have been assigned the weight of 8.75: they occur 11 times as synonyms and cooccur 3 times with associative relations. The extremities of the weight values can exceed the figure. Thus, the interval of values for MeSH is [-393, 388], [-14.25, 14] for Snomed Int, and [-510.5, 404] for Snomed CT. We can observe that the negative and positive frontiers of these intervals are parallel, except for Snomed CT; and that the amplitude is the highest within Snomed CT and lowest within Snomed Int. But the latter provides also the lowest number of terms and relations. There is a tendency of the point's cloud to be attracted to negative values, except for Snomed Int-induced synonyms.

Table 3 - of induced synonyms which are present in the three terminologies: validation through a virtual truth

	Existing synonyms				
	MeSH	***SNInt***	***SNCT***	***VT***	**%**
MeSH	2,438	198	560	2,692	9%
SN Int	290	438	979	1,102	13.9%
SNCT	1,043	1,322	5,211	5,575	14%

Table 3 indicates number of induced synonyms that are already known in the three processed terminologies. For instance, 2,438 MeSH-induced synonym pairs are already registered in this terminology, and 198 MeSH-induced synonyms are already registered in Snomed Int. We can observe a large number (5,211) of Snomed CT-induced synonyms that are already known in there: this resource provides many elementary, or non defined, terms of the biomedical area, although it doesn't allow to build an extensive set of the synonyms. The total number of the induced synonyms that exist within at least one of the exploited terminologies is 8,023. This set of synonyms is used to build up the virtual truth, on which basis we perform a further evaluation of the results. The last two columns of table 3 indicate number and percentage of the induced synonyms that are also in the virtual truth (VT) set: 9% of MeSH synonyms, 13,9% of Snomed Int and up to 14% of Snomed CT-induced synonyms are thus validated. Other induced synonyms are new. Figure 3 indicates the distribution of weights for the induced synonyms that are also part of the virtual truth set. We can observe that number of ambiguous synonymy relations is very small among Snomed Int-induced synonyms, and that the point's cloud of MeSH is now attracted to positive values. Within Snomed CT, the ambiguity of synonyms is still the most important.

Quality of results provided by this method depends (1) on precision of POS-tagging and we tried to apply the best currently known tagger [17]; (2) on quality of the source material; and (3) on the verification of a compositional structure of terms: up to now we have found only one pair of French terms where the compositional structure was not respected {*coup de soleil, sensibilité au soleil}* meaning (*Solar sensitiveness*), where *coup de soleil* is not compositional.

Conclusion and Perspectives

We proposed a novel method for inducing a lexicon of synonyms from structured terminologies and for its weighting 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 for inducing the synonyms. We exploit also a set of endogenously generated linguistic indicators (is a, sibling, associative, l.inclusion and their productivity) for profiling the induced synonymy relations and for computing their weight. If a synonymy relation is free of other semantic relations, its reliability is not hindered. Otherwise it suffers from these cooccurring relations. Thus, up to 33.6% of synonymy relations induced within Snomed CT are ambiguous with other semantic relations. The ambiguity is lower within MeSH (12.9%) and Snomed Int (7.7%). Weights of these ambiguous synonyms are attracted to negative values, which indicate less reliable synonyms. A virtual truth set of synonyms is built up with those induced synonyms that are also provided by the exploited terminologies. It allows to validate 9 to 14% of the induced synonyms. It also allows to observe that within this set, the ambiguity of the induced synonyms is lesser, particularly within MeSH and Snomed Int. Weights provided by the current work are helpful for the filtering step of synonyms and for preparing their validation. We noticed that the used material can be improved. For instance, it seems that there is an inconsistency in creating the broad categories of relations within UMLS: mapped to relations from source terminologies are currently assigned to RL, RQ, RN and RO relationships, which means that they may appear in both is a and associative categories. If a specific filter is applied, the material may provide less ambiguous set of induced synonyms. Besides, other NLP methods suitable for analysis of corpora may be used in order to enrich or cross-validate lexicon of synonyms acquired in this experience. Once thoroughly validated, this lexicon will be made available to the community. This lexicon can be exploited within various NLP tasks and applications.

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. Morphology book. Cascadilla Press, Cambridge, MA, 2001.

[3] 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] Poprat M, Beisswanger E, and Hahn U. 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", 2008:31–9.

[8] Gene Ontology Consortium . Gene Ontology: tool for the unification of biology. Nature genetics 2000;25:25–9.

[9] Grabar N, Jaulent MC, and Hamon T. Combination of endogenous clues for profiling inferred semantic relations: experiments with gene ontology. In: JAMIA (AMIA 2008), Washington, USA. 2008:252–6.

[10] National Library of Medicine, Bethesda, Maryland. Medical Subject Headings, 2001. www.nlm.nih.gov/mesh/meshhome.html.

[11] Côté RA, Brochu L, and Cabana L. SNOMED Internationale – Répertoire d'anatomie pathologique. Secrétariat francophone international de nomenclature médicale, Sherbrooke, Québec, 1997.

[12] Stearns M, Price C, Spackman K, and Wang A. Snomed clinical terms: overview of the development process and project status. In: AMIA, 2001:662–6.

[13] Mungall C. Obol: integrating language and meaning in bio-ontologies. Comparative and Functional Genomics 2004;5(6-7):509–20.

[14] 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.

[15] 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.

[16] Partee BH. Compositionality. F Landman and F Veltman, 1984.

[17] Tsuruoka Y, Tateishi Y, Kim JD, et al. Developing a robust part-of-speech tagger for biomedical text. LNCS 2005;3746:382–92.

[18] Cruse DA. Lexical Semantics. Cambridge University Press, Cambridge, 1986.

[19] Bodenreider O, Burgun A, and Mitchell JA. Evaluation of WordNet as a source of lay knowledge for molecular biology and genetic diseases: a feasibility study. In: Medical Informatics in Europe (MIE), 2003:379–84.

[20] Hamon T and Grabar N. Acquisition of elementary synonym relations from biological structured terminology. In: CICLING (5th International Conference on NLP, 2006), number 4919 in LNCS. Springer, 2008:40–51.

Address for correspondence

Natalia Grabar, Inserm U872 eq.20
15 rue de l'Ecole de Médecine
75006 Paris France
natalia.grabar@crc.jussieu.fr