Customization of biomedical terminologies

Julien HOMO[a], Laëtitia DUPUCH[b], Allel BENBRAHIM[c],
Natalia GRABAR[d] and Marie DUPUCH [d]
[a]*Antidot, Paris, France;* [b]*Université Toulouse III Paul Sabatier, France;*
[c]*OBJECT'IVE, Paris, France;* [d]*CNRS UMR8163, Université Lille 1&3, France*

Abstract. Within the biomedical area over one hundred terminologies exist and are merged in the Unified Medical Language System Metathesaurus, which gives over 1 million concepts. When such huge terminological resources are available, the users must deal with them and specifically they must deal with irrelevant parts of these terminologies. We propose to exploit seed terms and semantic distance algorithms in order to customize the terminologies and to limit within them a semantically homogeneous space. An evaluation performed by a medical expert indicates that the proposed approach is relevant for the customization of terminologies and that the extracted terms are mostly relevant to the seeds. It also indicates that different algorithms provide with similar or identical results within a given terminology. The difference is due to the terminologies exploited. A special attention must be paid to the definition of optimal association between the semantic similarity algorithms and the thresholds specific to a given terminology.

Keywords. Automatic Data Processing; Medical Informatics/methods; Reproducibility of Results; Terminology as Topic

1. Introduction

Thanks to the recent research in Artificial Intelligence, Natural Language Processing and Knowledge Engineering, an increasing number of terminological resources exists. Besides, these resources become larger and more complex when they tend to cover whole domains. For instance, within the biomedical area, several terminologies are available, such as MeSH (Medical Subject Headings) [1] or SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) [2], while the UMLS (Unified Medical Language System) Metathesaurus [3], making freely available over million concepts from over 100 terminologies, emphasizes this situation. Therefore, the user has to deal also with the irrelevant parts of these terminologies [4] because in several situations and contexts it is suitable to exploit a more constrained terminology or its subset. Because of the growing complexity and of the availability of terminologies, the research concerns move towards their customization. Customization leads indeed to the reduction of the semantic space and to generation of smaller and more coherent subsets of terms. The customization of terminologies has been discussed previously [5], but a real evolution has been marked only recently. Thus, this process may lead to the coherence of the top concepts [6], to the revision of description logic-based ontologies in order to incorporate new concepts [7], to the logical definition of modularity and of ontological modules, and to their application and combination [8-9]. Let us also cite *Slim GO* [10] subsets of Gene Ontology (GO) [11] terms created thanks to the functional annotations of genes. The slims provide functionally homogeneous sets of terms which concentrate on a given species or on a set of genes and group only those terms which are exploited for their annotation. We propose to exploit the semantic distance algorithms often applied within tree-structured resources, such as terminologies, and allowing the computing of the relatedness between two terms. The algorithms rely on the number of edges between two terms and may also exploit other factors such as depth and density of concepts, type of relationships or link orientation [12-16]. These algorithms are exploited within applications in which it is necessary to compute the semantically close terms, often in order to increase the sensitivity of the automatic systems: information indexing, retrieval and filtering [14,17-18], word sens disambiguation [15,19], malapropisms [16], terminology and dictionary development [20-21], novelty and redundancy detection [22], detection of similarity between genes [23]. We propose to exploit the semantic distance for the terminology customization. In this paper, we present the material and the methods exploited for the terminology customization, we then discuss the results and conclude with some perspectives.

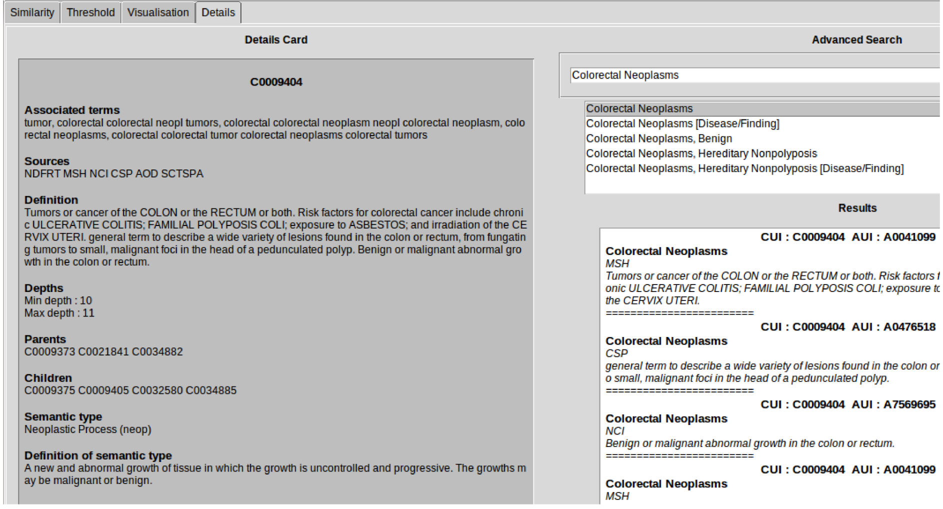
2. Material and Methods

Material. UMLS is our main material. It merges over one million concepts from over one hundred terminologies. Each concept has an unique identifier *CUI* and is assigned to a semantic type (ie, *Atrial fibrillation* belongs to the *Pathologic Function* type, *Colorectal neoplasms* to *Neoplastic Process*). The concepts are linked between them with 16 categories of UMLS-specific relationships (*PAR has parent, CHD has child...).*
Methods. Our method relies on the *s*emantic similarity approaches, such as those implemented within the UMLS-Similarity [24] package (Leacock and Chodorov *LCH* [19], *path* [24]...). This package has been created for the computing of semantic similarity between two UMLS CUIs explicitly indicated by users. The computing is done within a given terminology and with a given semantic measure. We propose to augment this module and to go beyond the processing of pairs of CUIs. The new function will exploit this module for the customization of UMLS or of a given terminology: on the basis of the seed CUIs provided by users, the module will be able to extract a set of the related CUIs and to semantically constrain this set. Additionally, we implement new measures. We perform three types of evaluation: (a) technical validity and reproducibility of the results; (b) evaluation of newly implemented algorithms; (c) judgment on the relevance of the results within the context of terminology customization performed with an expert who validated the sets of terms and also rated all the pairs of terms within these sets according to a previously proposed scale [25]: (1) non related concepts, (2) marginally related concepts, (3) closely related concepts, and (4) synonym or almost synonym concepts.

3. Results and Discussion

We present the following results: design and functionalities of the module for the customization of terminologies, implementation of new semantic similarity measures, and the evaluation of the results according to the proposed evaluation methods.

Customization of terminologies. The UMLS-Threshold package has been developed in order to allow the customization of terminologies (whole UMLS or its individual terminologies). It exploits the existing UMLS-Similarity module and provides new functionalities to it. The UMLS-Threshold package accepts as input several parameters: one or a set of the UMLS seed *CUIs*, and also the information on the names of the terminologies, on the relations to be exploited, on the semantic similarity measures and on the thresholds to be applied. Thanks to these parameters and data, the package then finds out all the neighboring *CUIs* within a given terminology or within the whole UMLS though the application of the Dikjstra algorithm [26]. It then applies the similarity measures and exploits them to weight the paths between the *CUIs*. These paths are then cut according to the threshold and the similarity measure indicated by the user. This last function leads to the customization of the terminology through the limitation of the paths and the reduction of the semantic space around each seed *CUI*. An additional customization function is applied when the type of relationships is taken into account: it can be only hierarchical UMLS relationships, like in the UMLS-Similarity module, or it can also cover other types of relationships. Indeed, we assume that it can be interesting to go beyond the hierarchical relationships and to exploit other relationships: terms linked by these may also be interesting for several applications.



Details Card

C0009404

Associated terms
tumor, colorectal colorectal neopl tumors, colorectal colorectal neoplasm neopl colorectal neoplasm, colo
rectal neoplasms, colorectal colorectal tumor colorectal neoplasms colorectal tumors

Sources
NDFRT MSH NCI CSP AOD SCTSPA

Definition
Tumors or cancer of the COLON or the RECTUM or both. Risk factors for colorectal cancer include chroni
c ULCERATIVE COLITIS; FAMILIAL POLYPOSIS COLI; exposure to ASBESTOS; and irradiation of the CE
RVIX UTERI. general term to describe a wide variety of lesions found in the colon or rectum, from fungatin
g tumors to small, malignant foci in the head of a pedunculated polyp. Benign or malignant abnormal gro
wth in the colon or rectum.

Depths
Min depth : 10
Max depth : 11

Parents
C0009373 C0021841 C0034882

Children
C0009375 C0009405 C0032580 C0034885

Semantic type
Neoplastic Process (neop)

Definition of semantic type
A new and abnormal growth of tissue in which the growth is uncontrolled and progressive. The growths m
ay be malignant or benign.

Advanced Search

Colorectal Neoplasms

Colorectal Neoplasms
Colorectal Neoplasms [Disease/Finding]
Colorectal Neoplasms, Benign
Colorectal Neoplasms, Hereditary Nonpolyposis
Colorectal Neoplasms, Hereditary Nonpolyposis [Disease/Finding]

Results

CUI : C0009404 AUI : A0041099
Colorectal Neoplasms
MSH
Tumors or cancer of the COLON or the RECTUM or both. Risk factors f
onic ULCERATIVE COLITIS; FAMILIAL POLYPOSIS COLI; exposure to
the CERVIX UTERI.
========================
CUI : C0009404 AUI : A0476518
Colorectal Neoplasms
CSP
general term to describe a wide variety of lesions found in the colon or
o small, malignant foci in the head of a pedunculated polyp.
========================
CUI : C0009404 AUI : A7569695
Colorectal Neoplasms
NCI
Benign or malignant abnormal growth in the colon or rectum.
========================
CUI : C0009404 AUI : A0041099
Colorectal Neoplasms
MSH

Figure 1: An excerpt from the graphical user interface for the seed CUI C0009404 Colorectal neoplasms.

The user graphical interface GUI (Figure 1) has been developed to access all these functionalities and their results. Figure 1 shows the details (parent and child *CUIs*, depth of the *CUIs*, synonyms, definition, source terminologies, semantic types...) for the seed *CUI* C0009404 *Colorectal neoplasms*. The Visualization tab of the interface generates the graph with Graphviz (www.graphviz.org) and presents the selected *CUIs*. The execution time is extremely short when queries rely on the indexed hierarchical UMLS relationships. It becomes longer (several seconds)when the whole UMLS is queried.

For the customization tests, we exploited several settings and parameters for seven seed *CUIs*. For instance, we exploited the same relations (PAR *has parent* and CHD *has child*) provided either by the SNOMED CT or by MeSH. We then compared the extracted graphs of terms. With the SNOMED CT, we observe that the selected terms belong to five hierarchical levels, while they belong to three hierarchical levels when the MeSH is exploited. We also observed that the extracted sets of terms are different: not surprisingly this set is richer with the SNOMED CT. Another difference in results appears depending on the similarity measures and on the thresholds exploited. As a matter of fact this is the core point for a successful exploitation of this kind of approaches: the user has to define experimentally or from a previous work the optimal association between the terminology, the semantic measures and the thresholds. In our experience with SNOMED CT and with MeSH, we have for instance observed that the measure *LCH* is optimal with the 2.4 threshold. Additionally, this measure provides very close results to those obtained with the *path* measure and the threshold 0.32.

New semantic similarity measures. Two new measures have been implemented. One is specific to the exploitation of the Gene Ontology [27], the other one respects the specificities of UMLS (different terminologies and relationships among the *CUIs)* and is adapted to the exploitation of the whole UMLS graph [8]. The implementation of these new measures is possible thanks to various data computed and gathered by the module, such as those presented on figure 1.

Evaluation. The integration of the UMLS-Similarity and UMLS-Threshold packages guarantees the reproducibility of the results generated though the UMLS-Similarity package. Thus, the technical reproducibility of results and the communication between the packages are performed correctly. The evaluation of two newly implemented algorithms indicates that the result are reproduced nearly exactly for the algorithm adapted to UMLS [8]. The few differences observed are due to the evolution between the two exploited versions of UMLS. As for the Gene Ontology specific algorithm [27], the comparison with previously reported results appears to be difficult. The main reason is again due to the evolution of the terminology. Indeed, since the publication of the measure in 2007, several hundreds of new terms and relations among them have been added. This causes a difference in the results, although they remain coherent. The aspect related to the evolution of the terminological resources is an interesting issue.

Generated sets of terms have been presented to the expert, who was asked to rate each pair of terms from 1 (non related terms) to 4 (synonyms), and to evaluate the relevance of these terms to the seeds. The evaluated sets of terms contain mainly the ratings 4 and 3, which is a positive result. As for the relevance of these sets to the seed terms, for the *CUI* C0004238 *Atrial fibrillation* (sets extracted from SNOMED CT and MeSH) the expert validation indicated that the precision is 73% and 62.5% respectively. For the SNOMED CT set, the expert selected 16 and rejected 6 terms, while for the MeSH sets, the expert selected 10 and rejected 6 terms. When we presented the expert with more neighboring *CUIs* not selected by the UMLS-Threshold, he selected 16 more terms out of 94. These results indicate that the semantic similarity approach may ease a lot the customization of terminologies and especially that the reduction of the potentially available *CUIs* with semantic similarity approaches is reliable.

4. Conclusion and Perspectives

We propose to exploit the semantic similarity measures for customization of terminologies. We rely on the existing module UMLS-Similarity and enrich it with the thresholding functions (UMLS-Threshold package). The results are available through the graphical user interface. The evaluation of the results by a medical expert provides with several indications, although additional evaluation is necessary to make these observations stronger. First of all, the semantic similarity measures seem to be suitable for the customization of the existing large terminologies. A possible difficulty for the use of such approach may be related to the fact that the thresholds vary according to the measures and to the terminologies exploited, but the previous experience of users may help the definition of the optimal thresholds. Another positive point is that, with the settings exploited, the majority of the relevant terms were included in the extracted sets of terms. The combination of measures, of the UMLS relations, and the study of their complementarity is another perspective of this work. We also implemented two new measures, but their full evaluation is also a perspective to this work.

References

[1] NLM (2001). Medical Subject Headings. National Library of Medicine, Bethesda, Maryland.

[2] Stearns MQ, Price S, Spackman KA & Wang AY. SNOMED clinical terms: overview of the development process and project status. In AMIA, pages 662--666, 2001

[3] NLM (2008). UMLS Knowledge Sources Manual. National Library of Medicine, Bethesda, Maryland.

[4] Dzbor M & Motta E. (2008). Engineering and customizing Ontologies. The Human-Computer Challenge in Ontology Engineering, 25–57.

[5] Guarino N. (1997). Understanding, building, and using ontologies. IJHCS, **46**(2-3), 293–310.

[6] Wang Z, Wang K, Topor R. & Pan JZ. (2008). Forgetting concepts in DL-Lite. In Springer Verlag, editor, WWTP, 245–57.

[7] Qi G, Haase P, Huang Z, Ji Q, Pan JZ & Volker J. (2008). A kernel revision operator for terminologies - Algorithms and evaluation. In International Conference on The Semantic Web, 419–34.

[8] D'Aquin M, Schlicht A, Stuckenschmidt H & Sabou M. (2007). Ontology Modularization for Knowledge Selection: Experiments and Evaluations. In Database and Expert Systems Applications, 874–83.

[9] Stuckenschmidt H, Parent C & Spaccapietra S. (2009). Modular Ontologies. Concepts, Theories and Techniques for Knowledge Modularization. Springer.

[10] www.geneontology.org/GO.slims.html

[11] Gene Ontology Consortium. (2000). Gene Ontology: tool for the unification of biology. Nature Genetics, **25**, 25–29.

[12] Rada R, Mili H, Bocknell E, Blettner M & De Freitas R. (1989). Development and application of a metric on semantic nets. IEEE Transactions on systems, man and cybernetics, **19**(1), 17–30.

[13] Zhong J, Zhu H, LI J & Yu Y.(2002). Conceptual graph mathching for semantic search. Drug Saf, **25**(6), 459–65.

[14] Sussna M. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. In Conference on Information and Knowledge Management.

[15] Resnik P. (1995). Disambiguating noun groupings with respect to wordnet senses.

[16] HIRST G & ST ONGE D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. The MIT Press, (1).

[17] Ide NC, Loane RF & Demner-Fushman D. (2007). Essie: a concept-based search engine for structured biomedical text. J Am Med Inform Assoc, **14**(3), 253–63.

[18] Dupuch M, Trinquart L, Colombet I, Jaulent MC & Grabar N. (2010). Exploitation of semantic similarity for adaptation of existing terminologies within biomedical area. Ekaw Workshop.

[19] Leacock C. & Chodorov M. (1998). Combining local context and wordnet similarity for sens identification. The MIT Press, **11**(1), 265–83.

[20] Church KW. & Hanks P. (1989). Word association norms, mutual information, and lexicography. In ACL, 76–83.

[21] Parekh V. (2004). Mining domain specific texts and glossaries to evaluate and enrich domain ontologies. In International Conference of Information and Knowledge Engineerig.

[22] Zhang Y, Callan J & Linka T. (2002). Novelty and redundancy detection in adaptative filtering. In ACM SIGIR.

[23] Lord PW, Stevens RD, Brass A & Goble CA. (2003). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics, (10), 1275–83.

[24] McInnes B., Pederson T. & Pakhomov S. (2009). UMLS-interface and UMLS-similarity: Open source software for measuring paths and semantic similarity. In AMIA, 431–5.

[25] Pedersen T, Pakhomov S, Patwardhan S, & Chute C. Measures of semantic similarity and relatedness in the biomedical domain. *J Biomed Inform*, 40(3):288--99, 2007

[26] Cormen TH, Leiserson CE, Rivest RL & Stein C. (2001). Introduction to Algorithms. MIT Press and McGraw-Hill, second edition.

[27] Wang J, Du Z, Payattakool R, Yu P & Chen CF. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23(10):1274--81, 2007