A Way to Automatically Enrich Biomedical Ontologies

Juan Antonio Lossio-Ventura[1], Clement Jonquet[1], Mathieu Roche[1,2], Maguelonne Teisseire[1,3]
[1] LIRMM - University of Montpellier, France
[2] Cirad, TETIS, France
[3] Irstea, TETIS, France
juan.lossio@lirmm.fr, jonquet@lirmm.fr, mathieu.roche@cirad.fr, maguelonne.teisseire@teledetection.fr

ABSTRACT

Biomedical ontologies play an important role for information extraction in the biomedical domain. We present a workflow for updating automatically biomedical ontologies, composed of four steps. We detail two contributions concerning the concept extraction and semantic linkage of extracted terminology.

1. INTRODUCTION

Biomedical big data raises a major issue: the analysis of large volumes of heterogeneous data. Ontologies, i.e. conceptual models of the reality, can play a crucial role in biomedical fields for automating data processing, querying, and integration of heterogeneous data. Few semi-automatic methodologies to build ontologies have been proposed in recent years. Semi-automatic construction/enrichment of ontologies are mostly achieved using natural language processing (NLP) [1] techniques to assess text corpus. However, besides the existence of various English tools, there are considerably fewer ontologies and tools available in French and Spanish. This shortcoming is out of line with the huge amount of biomedical data produced for several languages, especially in the clinical world. This paper proposes a workflow to enrich biomedical ontologies or terminologies from texts, addressing the lexical/syntactic and semantic complexity of this process. The lexical/syntactic complexity involves the extraction of biomedical complex terms from a specialized text corpus. The semantic complexity is related to concept induction and semantic linkage of new terms. Our methodology has been applied for English, French, and Spanish.

2. PROPOSED APPROACH

Our approach consists of four steps: (I) Term Extraction, (II) Polysemy Detection, (III) Sense Induction, and (IV) Semantic Linkage. The lexical/complexity complexity is tackled by (I), and the semantic complexity is addressed by (II), (III), and (IV).

(I) Term Extraction: We use BIOTEX[1], our application to extract biomedical terms from documents from text databases (e.g. PubMed). This application implements some measures presented in [4] allowing to extract terms that might be added to a biomedical ontology, we called them "candidate terms".

(II) **Polysemy Detection:** This step seeks to predict if candidate terms are polysemic. We proposed new features based on statistical measures to characterize our text corpus. They are extracted directly from texts and from a graph itself induced from the text corpus. We used several machine learning algorithms to determine if a term is polysemic or not. Totally, 23 features were proposed, 11 direct and 12 from the induced graph. Their effectiveness showed an F-measure of 98%.

(III) **Term Sense Induction:** The objective of this step, is to induce the multiple or unique sense(s) (concept) of polysemic and not polysemic candidate terms. The senses are extracted according to the context of terms. For this, we execute two tasks. First, (a) *Number of senses prediction:* This task is performed only for the candidate terms predicted as polysemic in the previous step. Then, (b) *Clustering for concept induction:* This task executes a clustering algorithm taking as input the predicted k, then for each cluster it selects the most important features, which represent the induced concept. Note that $k = 1$ when the candidate term is not polysemic.

The prediction of the sense number of a term falls directly in clustering-based issues. In clustering tasks, one of the most difficult problems is to determine the number of clusters k, which is a basic input parameter for most clustering algorithms. In the biomedical domain, according to the statistics on UMLS (see Table 1), polysemic terms trend to be linked to only to 2 and 5 senses (i.e. 2 and 5 clusters). Therefore, as we aim at identifying the possible senses for a new biomedical candidate term, we will limit the number of senses between 2 and 5. Table 1 shows the details of polysemic terms statistics in UMLS and MeSH for English, French, and Spanish. The English version of UMLS contains about 9 919 000 distinct terms of which about 54 257 are polysemic. It means that approximately for 200 biomedical terms there exists just 1 polysemic term.

# of Senses	UMLS			MeSH		
k	***EN***	***FR***	***ES***	***EN***	***FR***	***ES***
2	54 257	1 292	10 906	178	11	0
3	7 770	36	414	1	0	0
4	1 842	1	56	0	0	0
5+	1 677	1	18	0	0	0

Table 1: Details of Polysemic Terms in UMLS and MeSH.

To evaluate the clustering solutions, there exist two kinds of quality indexes [2]: external and internal. External indexes use pre-labelled data sets with "known" cluster configurations. Internal indexes are used to evaluate the"goodness" of a cluster configuration without any priory knowledge of the clusters, in our case, we propose to focus on internal indexes. We use the following measures: (i) the intra-cluser similarity (*ISIM*), and (ii) the inter-cluster

[1] http://tubo.lirmm.fr/biotex/

similarity (*ESIM*), in order to create new indexes. They focus on choosing the minimum or maximum value. That allows to have an idea if the reached clusters are homogeneous. New internal indexes are described in Table 2. **Notation:** $\mid S_i \mid$ is the number of objects assigned to the i_{th} cluster.

1) Average of ISIM: represented as a_k, is the average of the *ISIM* value of each cluster of a solution clustering with number of clusters $= k$. $max(a_k) = max\left(\frac{\sum_{i=1}^{k} ISIM_i}{k}\right)$
2) Average of ESIM: represented as b_k, is the average of the *ESIM* value of each cluster of a solution clustering with number of clusters $= k$. $min(b_k) = min\left(\frac{\sum_{i=1}^{k} ESIM_i}{k}\right)$
3) Average of the difference between ISIM and ESIM: represented as c_k, is the average of the difference between *ISIM* and *ESIM* multiplied by the number of objects in such cluster $\mid S_i \mid$. $max(c_k) = max\left(\frac{1}{k}\sum_{i=1}^{k} \mid S_i \mid \times (ISIM_i - ESIM_k)\right)$
4) Division between the ISIM sum and ESIM sum: represented as e_k, is the division between the sum of *ISIM* multiplied by the number of objects in such cluster $\mid S_i \mid$, and the sum of *ESIM* multiplied by the number of objects in such cluster. $max(e_k) = max\left(\frac{\sum_{i=1}^{k} \mid S_i \mid \times ISIM_k}{\sum_{i=1}^{k} \mid S_i \mid \times ESIM_i}\right)$
5) Global objective function divided by the logarithm: represented as f_k, is the division between the value of the average of ISIM and the logarithm of k to base 10. $max(f_k) = max\left(\frac{\frac{\sum_{i=1}^{k} ISIM_i}{k}}{\log_{10}(k)}\right)$

Table 2: New Internal Indexes.

For this purpose, we represented our text corpus of two different manners: (i) bag-of-words representation, and (ii) graph representation. We used clustering algorithms and computed the new internal indexes.

(IV) **Semantic Linkage:** This step aims to add a candidate term in an existing biomedical ontology, i.e., how to find the correct position in the ontology. (1) Creation of term co-occurrence graph with terms extracted in (I), selecting only the MeSH neighborhood of a candidate term, then (2) we evaluate the semantic similarity of the candidate term with: (i) its MeSH neighbors, and, (ii) the fathers/sons of those neighbors in MeSH ontology. The semantic linkage is based essentially on a context similarity using the cosine measure between the new biomedical candidate term and those appearing in an ontology. At the end, a list of terms is proposed where the new biomedical candidate term could be positioned.

3. DATA AND RESULTS

In this section, we report experiments done to evaluate the performance of our proposal for (ii) prediction of sense number, and (ii) semantic linkage.

(i) Prediction of Sense Number: We will describe the text database used and the experiments in the following paragraphs.

Text corpus: MSH WSD[2] [3], which is composed of 203 polysemic entities in English, linked to a number of concepts (2,3,4,5). This data set is well-known in Word Sense Disambiguation literature applied to the biomedical domain.

Results: We use five well-known clustering algorithms implemented in the CLUTO[3] software, such as: *rb, rbr, direct, agglo, graph.* In general, bag-of-words and graph representations obtain similar accuracy values. For these two cases, the maximum value is 93.1% obtained by $max(f_k)$ index (See Table 2). Which means that for 100 terms, our approach can determine correctly the number of concepts of 93 terms.

(ii) Semantic Linkage: *Text corpus:* We collect 60 MeSH terms that have been added between 2009 and 2015, for instance the term *"corneal injuries"*. Each MeSH term will represent a *"biomedical candidate term"*. Then, we retrieve the context of these terms using PubMed, this context is composed of 333 073 311 tokens. Then, we create a co-occurrence graph per term from the retrieved context.

[2] http://wsd.nlm.nih.gov/

[3] http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

Results: We use cosine similarity between contexts and we propose 10 positions to add candidate terms in the MeSH ontology. For instance, we take the term *"corneal injuries"* added in MeSH between 2009 and 2015. Its synonyms in MeSH are *corneal injury, corneal damage,* and *corneal trauma*. Its fathers are *corneal diseases* and *eye injuries*. Then, we apply our methodology to locate *"corneal injuries"* in MeSH. Table 3 shows the first 10 best propositions done by our methodology. From our 10 propositions, 5 are correct, i.e. we found the correct synonyms and fathers of *"corneal injuries"* in MeSH version 2015 (yellow rows).

№	*Where*	*Cosine*
1	***corneal injury***	0.4251
2	***corneal damage***	0.4181
3	*chemical burns*	0.4081
4	***corneal diseases***	0.3696
5	*corneal ulcer*	0.3689

№	*Where*	*Cosine*
6	***eye injuries***	0.3681
7	*amniotic membrane*	0.3639
8	*re-epithelialization*	0.3588
9	***corneal trauma***	0.3582
10	*wound*	0.3472

Table 3: Propositions about where to add the term *corneal injuries*.

Table 4 shows the precision of the number of terms which have at least 1 correct proposition with our methodology for the *Top 1, Top 2, Top 5* and *Top 10* propositions; taking into account the paradigmatic relations, i.e. synonyms, hyperonyms (fathers), and hyponyms (sons). For instance, the yellow cell shows that there exist at least 1 correct proposition (i.e. existent in MeSH ontology) for the 36 of the 60 terms (i.e. 40%).

Top 1	*Top 2*	*Top 5*	*Top 10*
0.333	0.400	0.500	0.583

Table 4: Precision of the number of terms which have at least 1 correct proposition with our methodology.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we present an entire workflow to enrich biomedical ontologies. We focus on the last two steps of the global process. We presented new internal indexes to predict the number of clusters (number of senses) for a new biomedical candidate term. They are based on the clustering task by using bag-of-words and graph approaches. Another contribution is to find the right position in an already established ontology for new biomedical terms associated with their senses. We extracted the possible relations for a term. Those were based only on the similarity context, using the cosine measure between contexts.

A perspective of this work is to extract the type of relations. This could be performed with the linguistic patterns (e.g. the verbs used between two terms) and the associated contexts.

Acknowledgments

This work was supported in part by the French National Research Agency under JCJC program, grant ANR-12-JS02-01001, as well as by University of Montpellier, CNRS, IBC of Montpellier project and the FINCyT program, Peru.

5. REFERENCES

[1] A. Gangemi. A comparison of knowledge extraction tools for the semantic web. In *The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC*, pages 351–366. Springer, 2013.

[2] A. D. Gordon. Classification, (chapman & hall/crc monographs on statistics & applied probability). 1999.

[3] A. J. Jimeno-Yepes, B. T. McInnes, and A. R. Aronson. Exploiting mesh indexing in medline to generate a data set for word sense disambiguation. *BMC Bioinf*, 12(1):223, 2011.

[4] J. A. Lossio-Ventura, C. Jonquet, M. Roche, and M. Teisseire. Biomedical term extraction: overview and a new methodology. *Information Retrieval Journal*, to appear 2016.

A Way to Automatically Enrich Biomedical Ontologies

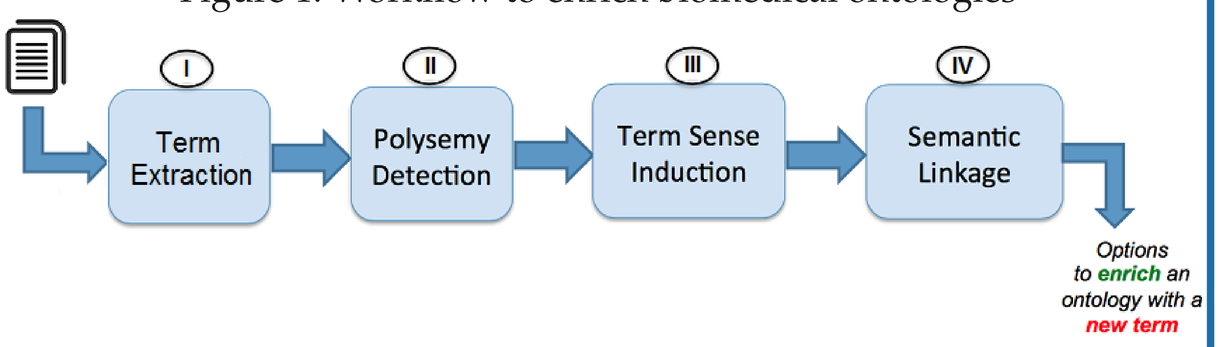
Juan Antonio Lossio-Ventura[1], Clement Jonquet[1], Mathieu Roche[1,2], Maguelonne Teisseire[1,2]
(1) LIRMM, University of Montpellier, CNRS, Montpellier - France
(2) Irstea, Cirad, TETIS, Montpellier - France
juan.lossio@lirmm.fr, jonquet@lirmm.fr, mathieu.roche@cirad.fr, teisseire@teledetection.fr

Introduction and Motivation

- Biomedical big data raises a major issue: the analysis of large volumes of heterogeneous data.
- Ontologies play a crucial role for data processing, querying, and integration.
- Few semi-automatic methodologies have been proposed to enrich/build ontologies. In addition, even if multiple tools do exist for extracting or using data in English, there is a strong lack in other languages such as French or Spanish.

Updating Biomedical Ontologies in 4 steps

Figure 1: Workflow to enrich biomedical ontologies



I) Term Extraction: We use BIOTEX[a], to extract/rank biomedical terms from text. BIOTEX implements several measures presented in [1]. We are interested in terms that do not appear in any ontology: **candidate terms**.

II) Polysemy Detection: We predict if a candidate term is polysemic or not. We presented in [2] new features to characterize our corpus in order to classify terms as polysemic or not.

III) Term Sense Induction: To induce the sense or senses (concepts) for polysemic and not polysemic terms. For this, we execute two tasks:

- **Number of Sense Prediction:** This is based on clustering issues, which seeks to determine the number of clusters. In the biomedical domain, according to statistics on UMLS, polysemic terms tend to be linked between 2 and 5 senses (clusters). We propose new internal indexes (see Figure 2) to predict the correct number of clusters. For this, we use the following measures: (i) the intra-cluser similarity (ISIM), and (ii) the inter-cluster similarity (ESIM).

Figure 2: New Internal Indexes

1) Average of ISIM: represented as a_k, is the average of the *ISIM* value of each cluster of a solution clustering with number of clusters $= k$.

$$max(a_k) = max\left(\frac{\sum_{i=1}^{k} ISIM_i}{k}\right)$$

2) Average of ESIM: represented as b_k, is the average of the *ESIM* value of each cluster of a solution clustering with number of clusters $= k$.

$$min(b_k) = min\left(\frac{\sum_{i=1}^{k} ESIM_i}{k}\right)$$

3) Average of the difference between ISIM and ESIM: represented as c_k, is the average of the difference between *ISIM* and *ESIM* multiplied by the number of objects in such cluster $\mid S_i \mid$.

$$max(c_k) = max\left(\frac{1}{k}\sum_{i=1}^{k} \mid S_i \mid \times (ISIM_i - ESIM_k)\right)$$

4) Division between the ISIM sum and ESIM sum: represented as e_k, is the division between the sum of *ISIM* multiplied by the number of objects in such cluster $\mid S_i \mid$, and the sum of *ESIM* multiplied by the number of objects in such cluster.

$$max(e_k) = max\left(\frac{\sum_{i=1}^{k} \mid S_i \mid \times ISIM_k}{\sum_{i=1}^{k} \mid S_i \mid \times ESIM_i}\right)$$

5) Global objective function divided by the logarithm: represented as f_k, is the division between the value of the average of ISIM and the logarithm of k to base 10.

$$max(f_k) = max\left(\frac{\frac{\sum_{i=1}^{k} ISIM_i}{k}}{\log_{10}(k)}\right)$$

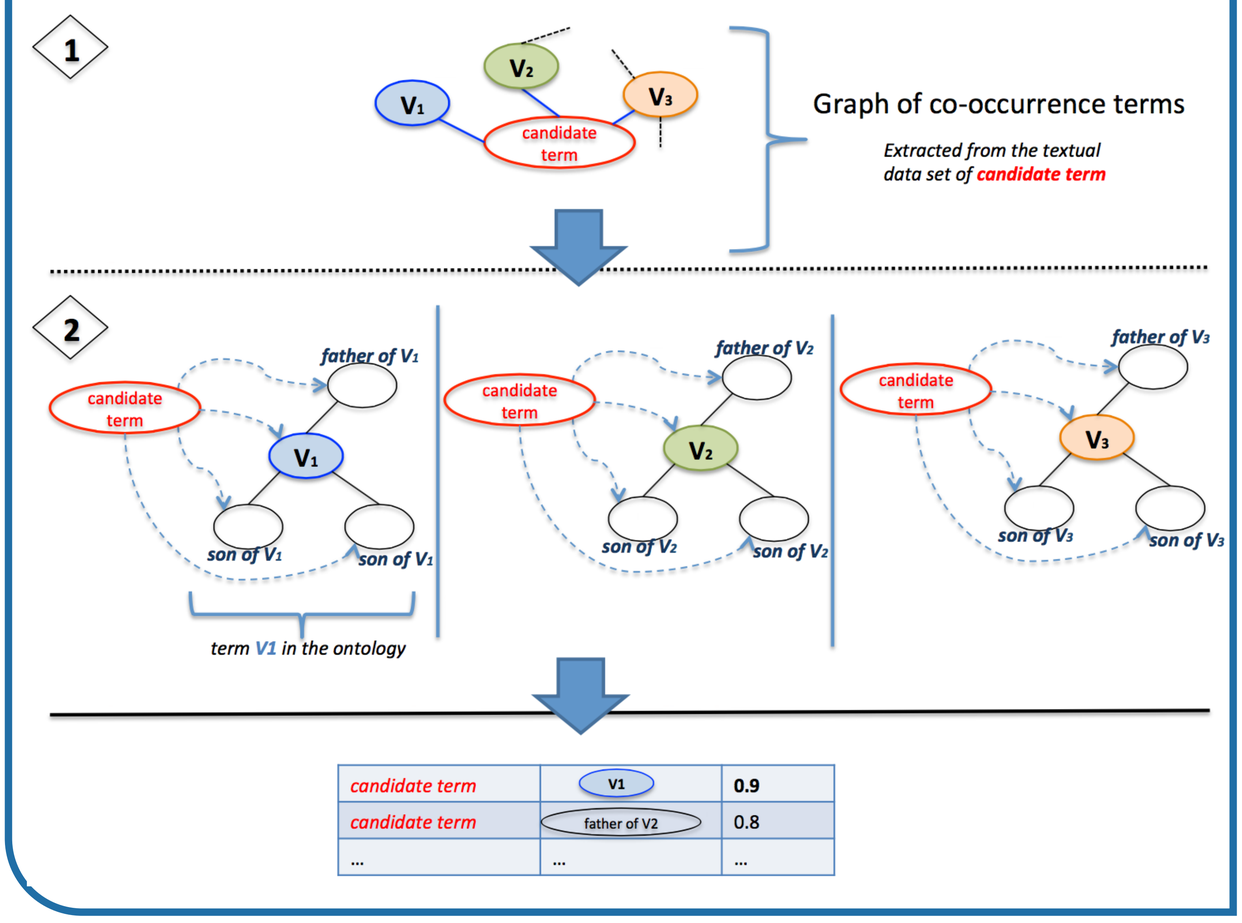
- **b) Clustering for concept induction:** A clustering algorithm is executed with the predicted k, selecting the most important features, which represent the induced concept.

IV) Semantic Linkage: We add a candidate term in an existing ontology. For this, we execute two tasks (see Figure 3): 1) Creation of term co-occurrence graph with terms extracted in (I), selecting only the MeSH neighborhood; 2) Semantic similarity of the candidate term with: (i) its MeSH neighbors, and, (ii) the parents/children of those neighbors in MeSH ontology. A list of terms is proposed where the candidate term could be positioned.

[a] http://tubo.lirmm.fr/biotex/

Updating Biomedical Ontologies in 4 steps

Figure 3: Semantic Linkage



Data and Results

(i) Prediction of Sense Number:

1. Text corpus: MSH WSD[a], which represents our gold standard corpus.
2. Results: In general, the index f_k gets the best *precision* = 93.1%. It means that for 100 terms, our approach can determine correctly the number of concepts of 93 terms.

(ii) Semantic Linkage:

1. Text corpus: 60 MeSH terms that have been added between 2009 and 2015, for instance the term "corneal injuries".
2. e apply our methodology to locate "corneal injuries" in MeSH. Figure 4 shows the first 10 best propositions done by our methodology. From our 10 propositions, 5 are correct, i.e. we found the correct synonyms and parents of "corneal injuries" in MeSH version 2015 (yellow rows).

Figure 4: Propositions about where to add the term corneal injuries

№	Where	Cosine
1	***corneal injury***	0.4251
2	***corneal damage***	0.4181
3	*chemical burns*	0.4081
4	***corneal diseases***	0.3696
5	*corneal ulcer*	0.3689
6	***eye injuries***	0.3681
7	*amniotic membrane*	0.3639
8	*re-epithelialization*	0.3588
9	***corneal trauma***	0.3582
10	*wound*	0.3472

[a] https://wsd.nlm.nih.gov/

Conclusions

We present a workflow to enrich biomedical ontologies. We focus on the last two steps of the global process. We presented new internal indexes to predict the number of clusters (number of senses) for a candidate term. Another contribution is to find the right position in an ontology for new biomedical terms associated with their senses.

References

[1] J. A. Lossio-Ventura, C. Jonquet, M. Roche, and M. Teisseire. Biomedical term extraction: overview and a new methodology. *Information Retrieval Journal*, 19(1):59–99, 2016.

[2] J. A. Lossio-Ventura, C. Jonquet, M. Roche, and M. Teisseire. Automatic biomedical term polysemy detection. In *Proceedings of the 10th Language Resources and Evaluation Conference*, LREC'16, 2016 *(to appear)*.