Automatic Symptom Extraction from Texts to Enhance Knowledge Discovery on Rare Diseases

Jean-Philippe Métivier[1], Laurie Serrano[2(✉)], Thierry Charnois[3], Bertrand Cuissart[1], and Antoine Widlöcher[1]

[1] Laboratoire GREYC, Univ. de Caen B.-N., CNRS, UMR6072, Caen, France
{jean-philippe.metivier,bertrand.cuissart,antoine.widlocher}@unicaen.fr
[2] Laboratoire MoDyCo, Univ. Paris X, CNRS, UMR7114, Nanterre, France
laurie.serrano@u-paris10.fr
[3] Laboratoire LIPN, Univ. Paris-Nord, CNRS, UMR7030, Villetaneuse, France
thierry.charnois@lipn.univ-paris13.fr

Abstract. This paper reports ongoing researches on automatic symptom recognition towards diagnosis of rare diseases and knowledge acquisition on this subject. We describe a hybrid approach combining sequential pattern mining and natural language processing techniques in order to automate the discovery of symptoms from textual content. More precisely, our weakly supervised approach uses linguistic knowledge to enhance an incremental pattern mining process, in order to filter and make a relevant use of the discovered patterns.

Keywords: Biomedical knowledge acquisition · Symptoms and rare diseases · Text mining · Incremental sequential data mining · Biomedical natural language processing

1 Introduction

A disease which affects less than 1 over 2,000 people is called a rare disease (RD): RDs are often disabling and life-threatening and, for most of these, there is no available cure. The Orphanet initiative maintains a reference portal providing services for knowledge sharing among the RD community. The related database includes expert-authored and peer-reviewed syntheses describing current knowledge about each RD. These syntheses result from a manual and time-consuming monitoring of literature made by specialists of RDs.

The work introduced in this paper aims at automating the update of the Orphanet knowledge by automatically identifying new symptoms associated to RDs. Symptoms have been rarely studied for themselves within biomedical information extraction (IE) literature but are often included in more general categories such as "clinical concepts", "medical problems" or "phenotypic information". In this work, we use the term "symptoms" to refer indifferently to functional and clinical signs of a disease. A very few studies consider linguistic contexts of symptoms towards their

J.H. Holmes et al. (Eds.): AIME 2015, LNAI 9105, pp. 249–254, 2015.
DOI: 10.1007/978-3-319-19551-3_33

automatic recognition. Some tackled symptom recognition with manually developed annotation rules [4] whereas others like [8] used a statistical approach (CRFs) to process clinical records. More recently, [7] automatically identified phenotypic information using the HPO ontology in order to recognize already known symptoms and enhance online search facilities. Most of existing studies process clinical reports or narrative corpora [6], whereas our work aims at scientific monitoring and analyses abstracts from research articles.

The overview of literature put forward two main difficulties: first, few works have tackled the problem of mining symptoms from texts and therefore existing resources are limited and incomplete; second, a given symptom can be expressed by multiple and diverse textual expressions which makes its automatic extraction very complex [5]. The contribution of this paper is to address these problems, designing a hybrid approach that combines data mining and natural language processing (NLP). On the one hand, we use an incremental process of frequent sequential data mining in order to automatically discover regularities (patterns) over textual expression of symptoms. The extracted patterns are then used for symptom recognition, each step of the incremental process allowing to discover more numerous symptoms. On the other hand, NLP techniques are involved to enhance the mining process and select relevant patterns, dealing with the well-known limitation of pattern mining techniques which produce very large, and hard to use, sets of patterns. It is worth noting that our method is fully automated and weakly supervised: to boot the process, the corpora are automatically annotated using public resources and the patterns are not validated manually.

Section 2 presents our hybrid approach. In Section 3, we give some preliminary results with their analysis and some possible technical improvements. Finally, Section 4 summarizes our contributions and provides perspectives.

2 Discovery of New Symptoms through Data Mining and Natural Language Processing

The proposed method relies on an iterative process where each loop runs three steps (see Figure 1). A corpus of medical texts is required as input, and the output is an annotated version of these texts where symptoms are tagged. As a first step, the texts of the initial corpus are annotated with the current list of already known diseases and symptoms. Then the annotated texts are mined to extract frequent sequential patterns which contain at least one symptom. During the third step, the most relevant patterns are selected thanks to a quality measure. Selected patterns are applied to the corpus providing new potential symptoms which enhance the resources for the first step of the next iteration. Then, this allows new patterns to be discovered, and so on.

Tagging of the Corpus. Two existing resources are used to annotate abstracts during the preprocessing step:

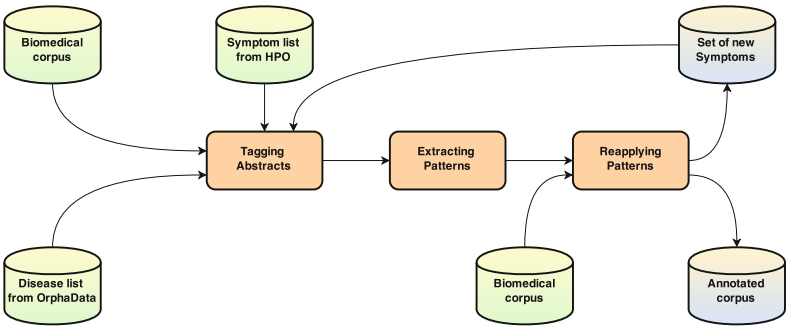

Fig. 1. Overall approach

1. OrphaData[1] provides a comprehensive, high-quality dataset related to rare diseases and orphan drugs. This resource allows to retrieve all the names of rare diseases and their aliases.
2. HPO[2] encapsulates a simple hierarchy of phenotypic anomalies. It does not provide a complete list of symptoms but this is a solid base to bootstrap the pattern mining phase.

These resources have been chosen for their specificity, rather than mostly used thesaurus (like UMLS *Unified Medical Language System* or MeSH *Medical Subject Headings*) that prove to be too large and generic for our objectives.

The corpus and the lists of terms coming from OrphaData and HPO are preprocessed using TreeTagger[3]: texts are tokenized, and each token is lemmatised and POS tagged. Each term (possibly composed of several tokens) coming from the external resources is matched against the corpus by comparing its tokens with those from the corpus. Terms coming from HPO are often generic (e.g. "weakness") and may be supplemented in medical texts with adjectives or object complements (e.g. "severe weakness of the tongue"). Thus, once a term matches, it can be expanded using the POS tags associated to surrounding terms.

Linguistic Pattern Mining. Sequential pattern mining was first introduced by [1] in the data mining field and was adapted to textual Information Extraction for instance by [2]. It is a matter of locating, in a set of sequences, sub-sequences (not necessarily contiguous) having a frequency above a given threshold. This mining process is applied to a base containing ordered sequences of itemsets where each sequence corresponds to a text unit (here sentences) and each itemset is a collection of features describing one word of a sequence. The goal is then to discover frequent sub-sequences of itemsets (called patterns).

[1] http://www.orphadata.org/
[2] http://www.human-phenotype-ontology.org/
[3] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

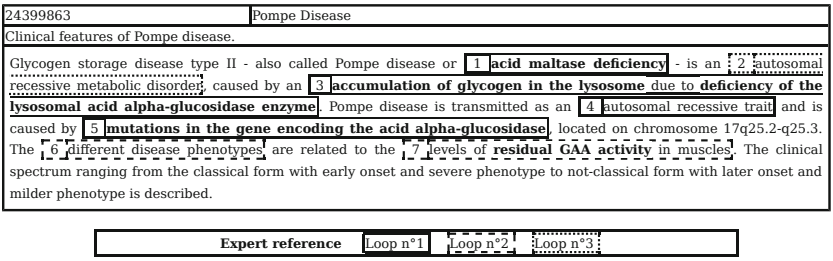

Fig. 2. Three iterations applied on one abstract

In this work, we use the SDMC[4] extractor [2]: this tool provides several constraints (e.g. patterns length and gap between itemsets) to guide the search for useful patterns [3] and allows to get a condensed representation of the patterns without loss of information ("closed patterns"). Depending on the frequency threshold fixed, a very large number of sequential patterns may be extracted and it is necessary to filter them. To avoid a manual filtering, we use a quality measure relying on the following principle : if the number of occurrences of a pattern is by far larger than the number of occurrences from which it was discovered, this pattern probably involves too much noise.

Finding New Symptoms. Once the sequential patterns are extracted and filtered, these ones can be used to find new symptoms. Since they are recurrent structures announcing symptoms, they are matched against the corpus in order to detect symptoms (already known symptoms as well as new ones). New symptoms can improve the previous annotations and may allow to extract new sequential patterns, and so on.

3 Preliminary Results

Experimental Corpus and Parameters. The initial corpus is composed of 150 raw abstracts collected from PubMed[5]. Sentences are used as sequences for the pattern mining process, which is set as follows: a frequency threshold of $0,25\%$, at least one symptom annotation included in each sequential pattern, a minimal length of 3 itemsets, and a gap fixed to 0 (patterns of contiguous elements). To filter irrelevant patterns the quality threshold is set to 2 during the first loop. This threshold is decreased of 10% at each loop to avoid a snowball effect.

Abstract-Example and Qualitative Analysis by an Expert. Figure 2 depicts an example of an abstract processed using our method. Bold annotations correspond to symptoms manually asserted by an expert, and squared annotations to symptoms that have been automatically recognized (a different style

[4] http://sdmc.greyc.fr
[5] www.ncbi.nlm.nih.gov/pubmed

of square is used for each loop). Here, the expert marked five symptoms which were all recognized by our extractor, meaning that our method may have a good recall. Nevertheless, some annotations remain imperfect. Different kind of errors arise: firstly, the symptoms *accumulation of glycogen in the lysosome* and *deficiency of the lysosomal acid alpha-glucosidase enzyme* were tagged as a single annotation (3); secondly, two other annotations (2) (4) do not refer to symptoms but to inheritance expressions; thirdly, another false positive appears during the second loop (6).

On-Going Improvements. Most of the erroneous annotations could be resolved by inserting deeper linguistic knowledge at different stages of the method. The first problem (3) could be quite easily avoided by adding specific linguistic rules to detect causality expressions (e.g. *due to* or *leading to*). We also plan to develop similar rules to better annotate complex expressions of symptoms like enumerations. Furthermore, the integration of a syntactic analysis (particularly verbal and nominal groups segmentation) could fix wrong delimitations of entities (e.g. (7) compared to the expert choice). Finally, we noticed that many of the false positives annotated within the corpus were due to patterns having similar characteristics that can be classified and processed accordingly. For example, some errors come from too generic patterns like the ones containing only POS tags items and being quite short (a length of 3 itemsets).

4 Conclusions and Future Work

The outcomes of this research concern both the rare diseases and the IE domains: on one hand, we tackle automatic recognition of symptoms associated to RDs which is a problem poorly studied until now; on the other hand, our work explores a hybrid approach combining data mining and NLP techniques, following recent trends in the IE research community. It is worth noting that our system is weakly supervised, completely automatic and does not imply any manual operations from RD experts. The first experiments emphasize good results regarding the development cost and the (still basic) linguistic knowledge involved. The remaining extraction problems could be avoided by adding more sophisticated NLP techniques: a syntactic analysis of the corpus in order to better determine the symptoms' frontiers; the addition of linguistic constraints within the pattern mining process; the definition of linguistic rules to define and filter out categories of patterns that involve erroneous detections.

Acknowledgments. This research was supported by the Hybride project (ANR-11-BS02-002).

References

1. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Proc. of the 11th International Conference on Data Engineering (1995)
2. Béchet, N., Cellier, P., Charnois, T., Crémilleux, B.: Discovering linguistic patterns using sequence mining. In: Proc. of the 13th International Conference on Computational Linguistics and Intelligent Text Processing (2012)
3. Dong, G., Pei, J.: Sequence Data Mining. Kluwer (2007)
4. Kokkinakis, D.: Developing resources for swedish bio-medical text mining. In: Proc. of the 2nd International Symposium on Semantic Mining in Biomedicine (2006)
5. Martin, L., Battistelli, D., Charnois, T.: Symptom extraction issue. In: Proc. of BioNLP 2014. Association for Computational Linguistics (2014)
6. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. Journal of the American Medical Informatics Association 17 (2010)
7. Taboada, M., Rodriguez, H., Martinez, D., Pardo, M., Sobrido, M.J.: Automated semantic annotation of rare disease cases: a case study. In: Database 2014 (2014)
8. Wang, Y., Liu, Y., Yu, Z., Chen, L., Jiang, Y.: A preliminary work on symptom name recognition from free-text clinical records of traditional chinese medicine using conditional random fields and reasonable features. In: Proc. of BioNLP (2012)