Symptom recognition issue

Laure Martin
MoDyCo
Paris Ouest University
laure.martin.1988
@gmail.com

Delphine Battistelli
MoDyCo
Paris Ouest University
del.battistelli
@gmail.com

Thierry Charnois
LIPN
Paris 13 University
thierry.charnois
@lipn.univ-paris13.fr

Abstract

This work focuses on signs and symptoms recognition in biomedical texts abstracts. First, this specific task is described from a linguistic point of view. Then a methodology combining pattern mining and language processing is proposed. In the absence of an authoritative annotated corpus, our approach has the advantage of being weakly-supervised. Preliminary experimental results are discussed and reveal promising avenues.

1 Introduction

Our work is part of the Hybride[1] Project, which aims to expand the Orphanet encyclopedia. Orphanet is the reference portal for information on rare diseases (RD) and orphan drugs, for all audiences. A disease is considered rare if it affects less than 1 person in 2,000. There are between 6,000 and 8,000 RD. 30 million people are concerned in Europe. Among its activities, Orphanet maintains an RD encyclopedia by manually monitoring scientific publications. Hybride Project attempts to automatically acquire new RD-related knowledge from large amounts of scientific publications. The elements of knowledge about a disease are varied: onset, prevalence, signs and symptoms, transmission mode, disease causes (etiology).

In this article, we investigate the automatic recognition of signs and symptoms in abstracts from scientific articles. Although named entity recognition in the biomedical domain has been extensively studied, signs and symptoms seem to have been left aside, for there is very little work on the subject. First, the linguistic issue of our study is presented in section 2, then the state of the art and the description of our lexical resources in section 3. Then our corpus and general method are presented in section 4. First experiments are introduced in section 5. Finally, the work to come is presented in section 6.

[1] http://hybride.loria.fr/

2 Signs and symptoms

Signs and symptoms both refer to the features of a disease, except that a symptom (or functional sign) is noticed and described by a patient, whilst a clinical sign is observed by a healthcare professional. In thesauri and medical ontologies, these two notions are generally put together in the same category. Moreover, in texts –particularly in our corpus of abstracts from scientific articles– there is no morphological or syntactic difference between sign and symptom. The difference is only semantic, so it is impossible for non-specialists in the medical field to tell the difference from the linguistic context alone. In example (1), clinical signs are in bold and symptoms are italicized.

> (1) Cluster headache (CH) is a primary *headache* disease characterized by recurrent short-lasting attacks of excruciating unilateral periorbital *pain* accompanied by **ipsilateral autonomic signs** (*lacrimation*, **nasal congestion**, **ptosis**, **miosis**, lid **edema**, and eye **redness**).

Furthermore, the diagnosis is established by the symptoms and the clinical signs together. We did not, therefore, try to distinguish them.

Signs and symptoms take on the most varied linguistic forms, as is noticeable in the corpus (which will be described in more detail below). In its simplest form, a sign or symptom is a noun, which may be extended by complements, such as adjectives or other nouns (example 2). They also appear in other, more complex, forms, ranging from a single phrase to a whole sentence (example 3).

> (2) With disease progression patients additionally develop **weakness** and

Proceedings of the 2014 Workshop on Biomedical Natural Language Processing (BioNLP 2014), pages 107–111,
Baltimore, Maryland USA, June 26-27 2014.

wasting of the limb and bulbar muscles.

(3) Diagnosis is based on clinical presentation, and **glycemia and lactacidemia levels, after a meal (hyperglycemia and hypolactacidemia), and after three to four hour fasting (hypoglycemia and hyperlactacidemia)**.

In addition to their variety, the linguistic units representing signs and symptoms present some syntactic ambiguities, particularly ambiguities concerning prepositional attachment and coordination scope. In example (2), the first occurrence of "and" is ambiguous, for we don't know if "weakness" and "wasting" should be grouped together as a single manifestation of the disease, or if "weakness" on the one hand and "wasting of the limbs and bulbar muscles" on the other hand are two separate entities, as annotated here.

In addition to these syntactic ambiguities, two annotation difficulties also arise. The first one consists in correctly delimiting the linguistic units of the signs and symptoms (example 4a). We agreed with experts in the field that, generally, pieces of information such as adjectives of intensity or anatomical localizations were not part of the units; nevertheless, this information is interesting in that it provides the linguistic context for the signs and symptoms. The second difficulty concerns elliptical constructions: where two signs can be distinguished, only one can be annotated because the two nouns have an adjective in common (example 4b).

(4) In the severe forms, **paralysis** (4a) concerns the neck, shoulder, and proximal muscles, followed by involvement of the muscles of the distal upper extremities, the diaphragm and respiratory muscles, which may result in **respiratory compromise or arrest** (4b).

Eventually, the last difficulty that was met during the corpus observation is the semantic ambiguity existing between sign or symptom and disease denominations. A disease can be the clinical sign of another disease. A clinical sign may be included in a disease name or conversely. In example (5), the clinical sign is in bold and the name of the disease is underlined.

(5) The adult form results in progressive limb-girdle **myopathy** beginning with the lower limbs, and affects the respiratory system.

3 State of the art

Signs and symptoms have seldom been studied for themselves in the field of biomedical information extraction. They are often included in more general categories such as "clinical concepts" (Wagholikar et al., 2013), "medical problems" (Uzuner et al., 2011) or "phenotypic information" (South et al., 2009). Moreover, most of the studies are based on clinical reports or narrative corpora –the Mayo Clinic corpus (Savova et al., 2010) or the 2010i2b2/VA Challenge corpus (Uzuner et al., 2011)–, except for the Swedish MEDLEX Corpus (Kokkinakis, 2006), which comprises teaching material, guidelines, official documents, scientific articles from medical journals, etc. Our work aims at scientific monitoring and is therefore based on a corpus of abstracts from scientific articles.

Most of the information extraction systems developed in the works previously cited use lexical resources, such as the Unified Medical Language System (UMLS) or Medical Subject Headings (MeSH) thesauri for the named entity extraction task. The UMLS comprises over 160 controlled vocabularies such as MeSH, which is a generic medical thesaurus containing over 25,000 descriptors. However, as Albright et al. (2013) pointed out, UMLS was not originally designed for annotation, so some of the semantic types overlap. They add that "the sheer size of the UMLS schema increases the complexity of the annotation task and slows annotation, while only a small proportion of the annotation types present are used." That is why they decided to work with UMLS semantic groups instead of types, except for signs and symptoms –originally a semantic type in the Disorders semantic group–, that they used independently.

In a genetic disease context, a sign or symptom may be phenotype-related. A phenotype is all the observable characteristics of a person, such as their morphology, biochemical or physiological properties. It results from the interactions between a genotype (expression of an organism's genes) and its environment. As many rare diseases are genetic, many signs and symptoms may be found in lists of phenotype anomalies. For that reason,

we chose to use the Human Phenotype Ontology – HPO (Khler et al., 2014) as a lexical resource. To our knowledge, HPO has not yet been used in any study on signs and symptoms extraction. Nevertheless, it should be recalled that phenotype anomalies are not always clinical signs, and signs or symptoms are not all phenotype-related. Even so, we decided to use HPO as a lexical resource because it lists 10,088 terms describing human phenotype anomalies and can be easily collected.

Just a very few studies take advantage of considering the linguistic contexts of sign and symptom entities. Kokkinakis (2006), after a first annotation step of his corpus with MeSH, states that 75% of the signs and symptoms co-occur with up to five other signs and symptoms in a sentence. This allowed him to develop new annotation rules. We can also mention the MedLEE system (Friedman, 1997), which provides, for each concept, its type (e.g. "problem"), value (e.g. "pain") and modifiers such as the degree (e.g. "severe") or the body location (e.g. "chest").

As far as we are concerned, our approach is based on the combination of NLP and pattern mining techniques. We will see that the linguistic contexts mentioned above are part of the patterns automatically discovered with our text mining tool.

4 Corpus and general method

As mentioned above, HPO was selected as the lexical resource for this project. With the list of phenotype anomalies as queries, we compiled a corpus of 306,606 abstracts from the MEDLINE database with the PubMed search engine. These abstracts are from articles published within the last 365 days. They consist of an ID, a title and a paragraph. Then, we applied HPO and kept only the sentences containing a unit annotated as a sign or symptom. As already pointed out, signs and symptoms are not all phenotype-related, so our pre-annotation is incomplete. Nonetheless, this first annotation is quick and cheap, and it initiates the process.

Figure 1 illustrates the successive steps in the approach. In step 1, HPO (f) is used to annotate a first corpus (a) by a single projection of HPO terms onto the texts. This annotated corpus provides a first learning corpus (b) to discover patterns (c) by a text mining method (step 2; this method is detailed below). These patterns are then validated by an expert (step 3), as linguistic patterns (d). Step

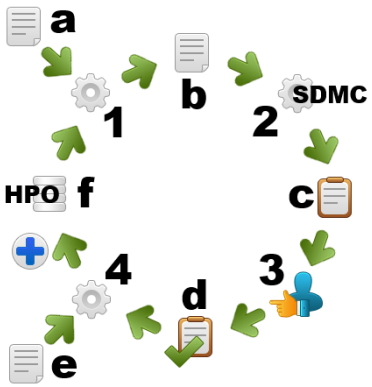

Figure 1: Iterative process of our sign and symptom extraction method

4 consists in using these patterns to annotate new corpora (e) and extract new terms (here with the semantic type of sign or symptom), which will be added to the resources (f). The process is finally repeated (back to step 1, with enriched lexical resources). This incremental process has the advantage of being weakly-supervised and non-dependent on the corpus type.

Sequential pattern mining was first introduced by Agrawal et al. (1995) in the data mining field. It was adapted to information extraction in texts by Béchet et al. (2012). It is a matter of locating, in a set of sequences, sequences of items having a frequency above a given threshold (called "support"). Pattern mining is done in an ordered sequence of items base, where each sequence corresponds to a text unit (the sentence here). An item represents a word in this sequence, generally the inflected form or the lemma or even the part of speech if the aim is to identify generic patterns. A number of parameters can be adapted along with the application.

Contrary to classical Machine Learning approaches which produce numerical models that are unintelligible for humans, data mining allows the discovery of symbolic patterns which can be interpreted by an expert. In the absence of authoritative annotated corpora for the recognition of signs and symptoms, manual validation of the patterns step is necessary, and often a large number of patterns still remains. To overcome this difficulty, Béchet et al. (2012) suggested adding constraints in order to reduce the results. In continuation of this work, we make use of the sequential patterns extraction tool SDMC[2], which makes it possible to

[2] https://sdmc.greyc.fr/

apply various constraints and condensed representations extraction (patterns without redundancy).

We adapted pattern mining to our field of application. Thus we first propose to use TreeTagger (Schmidt, 1994) as a pretreatment, in order to mark up different types of item (inflected form, lemma, part of speech). To narrow down the number of patterns returned by the tool, we introduce several constraints specific to our application: linguistic *membership* constraints (for example, we can choose to return only patterns containing at least one sign or symptom name), or the "gap" constraint (Dong and Pei, 2007), corresponding to possible gaps between items in the pattern. Thus a gap of maximal value *n* means that at most *n* items (words) are between each item of the pattern in the corresponding sequences (sentences).

5 First experiment

Annotating the first MEDLINE corpus of Abstracts with HPO provided us with a corpus of 10,000 annotated sentences. The 13,477 annotated units were replaced by a keyword –SYMPTOM– in order to facilitate the discovery of patterns. Then we used SDMC to mine the corpus for maximal patterns, with a minimal support of 10, a length between 3 and 50 words and a gap constraint of $g(0,0)$, i.e. the words are consecutive (no gap allowed). We were mining for lemma sequences only.

Results produced 988 patterns, among which 326 contained the keyword symptom. Based on these patterns, several remarks can already be made:

- Several annotated signs or symptoms are regularly associated with a third term, which can be another sign or symptom: {symptom}{symptom}{and}{stress};
- HPO annotation limitations (see section 3) are made visible by some contexts: {disease}{such}{as}{symptom};
- Some contexts are particularly recurrent, such as {be}{associate}{with}{symptom} or {characterize}{by}{symptom};
- Some temporal and chronological ordering contexts are present: {@card@}{%}{follow}{by}{symptom};
- The term "patient" is quite regular ({patient}{have}{severe}{symptom}), but after the evaluation, these occurrences turned out to be disease-related more than sign or symptom-related;
- The body location proved to be another regular context: {frontotemporal}{symptom}{ftd}.

Firstly, a linguistics expert selected the patterns that he considered the most relevant. These patterns were then classified in three categories: strong if they seem to strongly imply the presence of signs and symptoms (43 patterns), moderate (309 patterns) and weak (45 patterns). Secondly, these patterns were applied on a new corpus of MEDLINE abstracts in order to annotate the sign and symptom contexts. For the moment, only strong patterns have been applied.

25 abstracts were randomly selected among all the scientific articles published within the last month and dealing with Pompe disease. These 25 articles were manually annotated for signs and symptoms by an expert and thus constituted a gold standard. Then, we compared the manual annotation to our automatically annotated contexts. If the annotated sentence includes signs or symptoms, we consider that the annotation is relevant. Among the 25 abstracts (225 sentences), 27 contexts were extracted with our method. 23 were correct, 4 were irrelevant; 70 sentences were not annotated by the system. Thus the results were 23.7 in recall, reaching 82.2 in precision (36.8 in F-score).

6 Conclusions

Sign/disease ambiguity is the cause of 3 of the 4 irrelevant annotations, i.e. diseases were in the same linguistic context than signs. Thus the sentences were annotated but they contained diseases, not signs. The forth irrelevant annotation indicates a diagnosis test; it highlights that causes and consequences of a disease can be easily confused by non-specialists. Most of the left out sentences contain signs or symptoms expressed by complex units, such as Levels of creatinkinase in serum were high. (36%). 27% of these sentences are about gene mutations, which can be considered as causes of diseases or as clinical signs. Others contain patterns which have not been selected by the expert but can be easily added to improve the recall.

The context annotation is only a first step towards sign and symptom extraction. So far, we have not solved the problem of unit delimitation. In order to achieve this, we have two working hypotheses. We intend to compare chunking and syntactic analysis results in defining the scope of sign and symptom lexical units. Chunking will be conducted with an NLP tool such as TreeTagger, and syntactic analysis will use a dependency parser such as the Stanford Parser (ref.). The latter should allow us to delimit some recurring syntactic structures (e.g. agents, enumerations, etc.).

We also intend to compare our results with results provided by CRFs. First the features will be classical (bag of words, among others), and second, we will add the contexts obtained with the text mining to the features. This should enable us to compare our method to others. Finally, we are going to develop an evaluation interface to facilitate the work of the expert. In the absence of comparable corpora, the evaluation can only be manual. Our current sample of 50 abstracts is just a start, and needs to be expanded in order to strengthen the evaluation.

Acknowledgments

This research was supported by the Hybride Project ANR-11-BS02-002.

References

Rakesh Agrawal and Ramakrishnan Srikant. 1995. Mining Sequential Patterns. *Proceedings of ICDE'95*.

Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F. Styler IV, Colin Warner, Jena D. Hwang, Jinho D. Choi, Dmitriy Dligach, Rodney D. Nielsen, James Martin, Wayne Ward, Martha Palmer ans Guergana K. Savova. 2013. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*, 20:922–930.

Nicolas Béchet, Peggy Cellier, Thierry Charnois and Bruno Crémilleux. 2012. Discovering linguistic patterns using sequence mining. *Proceedings of Springer LNCS, 13th International Conference on Intelligent Text Processing and Computational Linguistics - CICLing'2012*, 1:154–165.

Guozhu Dong and Jian Pei. 2007. *Sequence Data Mining*. Springer.

Carol Friedman. 1997. Towards a Comprehensive Medical Language Processing System: Methods and Issues. *Proceedings of the AMIA Annual Fall Symposium*, 1997:595–599.

Sebastian Köhler, Sandra C. Doelken, Christopher J. Mungall, Sebastian Bauer, Helen V. Firth, Isabelle Bailleul-Forestier, Graeme C. M. Black, Danielle L. Brown, Michael Brudno, Jennifer Campbell, David R. FitzPatrick, Janan T. Eppig, Andrew P. Jackson, Kathleen Freson, Marta Girdea, Ingo Helbig, Jane A. Hurst, Johanna Jähn, Laird G. Jackson, Anne M. Kelly, David H. Ledbetter, Sahar Mansour, Christa L. Martin, Celia Moss, Andrew Mumford, Willem H. Ouwehand, Soo-Mi Park, Erin Rooney Riggs, Richard H. Scott, Sanjay Sisodiya, Steven Van Vooren, Ronald J. Wapner, Andrew O. M. Wilkie, Caroline F. Wright, Anneke T. Vulto-van Silfhout, Nicole de Leeuw, Bert B. A. de Vries, Nicole L. Washingthon, Cynthia L. Smith, Monte Westerfield, Paul Schofield, Barbara J. Ruef, Georgios V. Gkoutos, Melissa Haendel, Damian Smedley, Suzanna E. Lewis and Peter N. Robinson. 2014. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, 42:966–974.

Dimitrios Kokkinakis. 2006. Developing Resources for Swedish Bio-Medical Text-Mining. *Proceedings of the 2nd International Symposium on Semantic Mining in Biomedicine (SMBM)*

Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, Christopher G. Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17:507–513.

Helmut Schmidt. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*.

Brett R. South, Shuying Shen, Makoto Jones, Jennifer Garvin, Matthew H. Samore, Wendy W. Chapman and Adi V. Gundlapalli. 2009. Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. *Summit on Translational Bioinformatics 2009*

Özlem Uzuner, Brett R. South, Shuying Shen, Scott L. DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18:552–556.

Kavishwar B. Wagholikar, Manabu Torii, Siddartha R. Jonnalagadda and Hongfang Liu. 2013. Pooling annotated corpora for clinical concept extraction. *Journal of Biomedical Semantics*, 4:3.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.