Automated Extraction of Symptoms related to Rare Diseases from Scientific Publications

Charles Cousyn
LIARA Lab
Université du Québec à Chicoutimi
Saguenay, Canada
charles.cousyn1@uqac.ca

Kevin Bouchard
LIARA Lab
Université du Québec à Chicoutimi
Saguenay, Canada
Kevin.Bouchard@uqac.ca

Bruno Bouchard
IEEE Senior, LIARA Lab
Université du Québec à Chicoutimi
Saguenay, Canada
Bruno.Bouchard@uqac.ca

Sébastien Gaboury
LIARA Lab
Université du Québec à Chicoutimi
Saguenay, Canada
Sebastien.Gaboury@uqac.ca

ABSTRACT

Rare diseases constitute a poorly known subject among the population. Nevertheless, despite their name, a large number of persons are afflicted by one or many of them. Research on the nearly seven thousand rare diseases is insufficient, and even if some works have been done to exploit scientific publications and extract relevant information, knowledge is very difficult to obtain for the general population. This paper presents a new system that try to address the dissemination of knowledge on rare diseases. Particularly, we focus on the task of extracting automatically symptoms of rare diseases from publications with a new approach using a Named Entity Recognition (NER) algorithm based on the numerical statistic Term Frequency - Inverse Document Frequency (TF-IDF).

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics • **Networks** → Network reliability

KEYWORDS

Text mining, Rare Disease, Named Entity Recognition, Knowledge Aggregation

Goodtechs '18, November 28–30, 2018, Bologna, Italy

ACM ISBN 978-1-4503-6581-9/18/11...$15.00
https://doi.org/10.1145/3284869.3284892

1 INTRODUCTION

The Canadian government spending on healthcare is continuously increasing on a yearly basis. In 2017, a total of 242 $ billion was invested according to Canadian Institute for Health Information (2017) [1]. These investments are a positive point and a sign of how important healthcare to the population is. Nevertheless, the government choses to invest, with an understandable logic, in public medical research programs that are likely to benefit the large majority of the population [2]. This is where research for rare diseases is at disadvantage; while the private pharmaceutical industry is reluctant to invest in research for orphan drugs that carry a high risk of failure, it would also not result in a highly lucrative business, since it would target only a small number of persons for the majority of the rare diseases.

The consequence for persons affected with rare disease is that they are left without treatment options and often a lack of even a basic scientific understanding of their disease. This population is left to themselves, often getting several wrong diagnoses and taking years to learn what they are afflicted with [3-5]. Indeed, in the case of rare diseases, it is often difficult to establish a reliable diagnosis for several reasons: the number of specialists for a specific rare disease is limited (or simply non-existent), access to specialists is difficult and there is a lack of scientific knowledge and information about rare diseases. In addition, even when a diagnosis is established, physicians are often unable to transfer accurate and relevant information about the disease to the person and her family (due to lack of time or knowledge) [6].

Several portals exist where we can learn about rare diseases. These portals are built by devoted persons who often do it for free in their spare time or by non-profit organization working to the awareness of some of these diseases (Orphanet [7], EpiRare [8], NORD [9], etc.). There are three major drawbacks to this information strategy. First, only the common rare diseases are discussed on these portals or websites. Second, the knowledge

transfer process starts from highly qualified specialist publishing paper, to physician, and then, to persons working on the portal. The whole process may take years, and there are chances for the information to lose accuracy over time. Third, for the most accurate and complete of these portals, the information is not presented in a way to be understandable by the general population; the platform target highly skilled specialists.

The topic of rare disease is obviously extremely challenging on all fronts of research. This project aims to address the first two challenges by trying to deal with a maximum of diseases but also and especially by focusing on accelerating the process of knowledge transfer by automating the search for symptoms in scientific publications. To respond to these problems, this paper presents an automatic extraction system for rare disease symptoms. Our tool uses reliable sources made available free of charge. The tool is functional, and the source code is available online[1].

2 Related Work

Text mining and the understanding of unstructured text data by machines has been the object of research for several years now [10-12]. Nevertheless, this area of research is still very active, and it proposes several unsolved challenges. The automatic extraction of a specific information through its context and understanding is among the remaining questions. In this section, we try to summarize the most relevant approaches regarding our work.

2.1 Text mining of scientific papers

There are a certain number of tools for excavating scientific publications automatically. One can quote PolySearch 2 [13], which is a web server for text mining and semi-automated discovery of text associations between various types of biomedical entities (e.g.: human diseases, genes, proteins, drugs, metabolites, toxins, metabolic pathways, organs, etc.). Technology used by Polysearch to extract associations is sentence scoring with a pattern recognition system that used 3 types of patterns depending on the number of words in the sentence: compact patterns, general patterns and relaxed patterns. Other works, such as Mahmood et al. 2016 [14], extract very specific information such as, in that case, mutation-disease associations using pattern recognition, syntactic and semantic analysis.

2.2 Extraction of symptoms

The automatic extraction of the symptoms of diseases is a task that has already been subject of few projects. The closest work to our problematic is certainly that of Holat et al. [15]. As in their paper, the objective is the extraction and recognition of symptoms in publications related to rare diseases. Using 100 publications per disease, out of 100 diseases selected by an expert, the authors achieved an F-score of 29.38% using a combination of pattern mining and Conditional Random Field (CRF). As a reminder, F-score represent the harmonic mean of precision and recall, precision and recall being measures of relevance of system. Another study that considered using a combination of pattern mining and Natural language processing tools (Tree tagger[16] and Stanford Parser[17]) achieved an F-Score of 36.8% by evaluating 25 disease summaries.

There are two observations that can be drawn from these works. In both cases, few diseases are tested. Indeed, they reportedly selected a few well documented diseases which would represent less than 2% of the 7000 rare diseases. Secondly, both these works only treat the abstract of the publications. Our approach aims to specifically perform the task to as many diseases as possible as long as there is enough scientific material on them.

3 Architecture of the new platform

The new tool is built around three modules to cover all aspects of automated extraction. First, an information aggregator module is capable of routing rare disease datasets and scientific publications from the PubMed Central (PMC) [18] database and storing them in a database. The second module deals with the extraction of entities from scientific papers. The tool uses the named entity recognition (NER) algorithms of the LingPipe library [19]. Finally, an extraction quality evaluation module is available to generate performance files in JSON format and thus allow an easy exploration of the evaluation results.

3.1 Database

The tool, in its current version, stores information on a MongoDB database (version 3.4.10), which have the advantage of scaling well when the volume of data starts to grow. The main role of the database is to store the list of rare diseases and all their related information to accelerate the data retrieving process. Everything that was automatically retrieved and that can be saved without requiring too much space is stored. The information include basic disease information taken from Orphanet [7], some the most reputable publications linked to each disease, the symptoms related to their respective weight (which symbolizes the importance of the symptoms, the strength of the relationship between the symptoms and the diseases) computed by text mining.

Scientific papers are stored in the database to enable offline text mining operations. Nevertheless, to avoid an explosive load of data, the number of papers is currently limited to a maximum of 1000 papers per disease. Moreover, it would be difficult, if not impossible, to update the diseases' information in a reasonable amount of time without limiting the number of publications to explore.

3.2 Data gathering process

Good data sources are essential to complete the database with consistent and relevant information. First, the list of rare diseases is based on the list provided by Orphanet in one of the datasets available on Orphadata. Orphanet was created by INSERN (1997) and is known to be the one of the most accurate and up to date information source for rare diseases [7].

For scientific publications, the platform uses the Entrez Programming Utilities API [20] by the National Center for Biotechnology Information (NCBI) to retrieve and analyze them. This API provides free access to all databases/search engines of Entrez (PubMed, PMC, Gene, Nuccore and Protein). In our case, we only use the PMC (Pubmed Central) database which allows to search in the MEDLINE database the publications having the full

[1] https://github.com/CharlesCousyn/RDSearch4

text available. Medline is a bibliographic database produced by the National Library of Medicine [11] which covers all biomedical fields: biochemistry, biology, clinical medicine, economics, ethics, odontology, pharmacology, psychiatry, public health, toxicology, veterinary medicine.

To maintain the quality of the information in the database, a whole system is dedicated to its update. The process update is simple, every time the program is launched, if there is an update on data sources online, a local update takes place and the database is updated. When it occurs, it first updates the list of diseases to look for potential additions that could have been made. Then, it goes through these diseases and deals with all their symptoms by launching all update processes including text mining in publications retrieved from the API. The long-term goal is, however, to be able to perform this task continuously in the background on a dedicated server without limiting the number of scientific papers to explore.

4 Text mining on symptoms

The method used to extract this information has to remain fast to allow the update to be done in a few hours since there are almost 7000 rare diseases with a significant variation in the number of associated papers (from zero to more than a hundred thousand). Hence, in this project, different solutions were explored to tackle this task.

The solution we decided to use is to rely on an external library called LingPipe [19]. LingPipe is a recognized option working in Java and C#. This library makes it possible to do many text mining operations such as the named entities recognition, text classification, clustering [21], sentiment analysis [22], part-of-speech tagging, etc. Another alternative is the Standford NLP library by Manning et al. [17] which has similar capabilities.

4.1 Dictionary based NER

The task of the automatic extraction of symptoms can be considered as a task of *Named Entity Recognition* (NER). NER is the search field that consists of identifying words or groups of words belonging to the same category in a text. For example, this may involve recognizing terms such as "Dog" or "Cat" in the text when searching for words in the "Animal" category.

One of the ways to do symptom recognition with NER techniques is to do NER based on a dictionary. The principle is as follows, we assume that we have the complete list of possible entities and we will simply try to extract the words or groups of words from the text that are in this list. This approach has the advantage of being very simple because once the dictionary is found or established, it is enough to check the word groupings of the text and find matches. And in our case, the dictionary exploited is the Human Phenotype Ontology (HPO) [23]. HPO provides a standardized vocabulary of phenotypic abnormalities encountered in human disease. A phenotypic abnormality happens to be a notion relatively close to what is commonly called a symptom. In the case of rare diseases, we consider these notions as similar since they are, for the great majority, of genetic origin even if we know that HPO is not an exhaustive list of symptoms.

4.2 Observation and TF-IDF

The dictionary-based NER is a very simple approach that is not sufficient when exploited naively. On our problematic, this basic approach gives a large number of false positives (average recall of 59.66%, average accuracy of 1.83% and average F-score of 3.55%). In view of this observation, our team worked to develop an approach allowing to eliminate a maximum of these false positives and thus to make go up the precision and the F-score without making the recall fall.

4.2.1 A basic TF-IDF approach. The idea of the approach is as follows: For each disease, we assign a significance value to each extracted entity, rank them by decreasing value, and set a general threshold to eliminate entities that are not significant enough. From this idea, two things remain to be determined. First, the calculation of the significance should be done in a way that gives a high significance value to the true symptoms and a low significance value to the surplus symptoms (false positives). Then, once each extracted symptom has an assigned value, it is necessary to determine an optimal threshold to eliminate enough false positives but not too many true positives, i.e. which maximizes the F-score.

To answer the first question, we tried to find a statistical formula that could be as discriminating as possible between the real symptoms and the surplus symptoms. Our choice of this formula was "Term Frequency - Inverse Document Frequency (TF-IDF). TF-IDF, is a statistical measure that attempts to reflect how important a word/term is in a text corpus [24]. The inverse frequency in documents (IDF) can be defined as follows: "Factor that decreases the weight of terms that appear very frequently in the document set and increases the weight of terms that appear rarely". In other words, a word that will tend to be used in a lot of texts will see its value of importance diminished while a word used in few texts will be preferred. The mathematical formulation of TF-IDF looks like this:

$$TF - IDF(t, d, D) = TF(t, d) * IDF(t, D) \quad (1)$$

$TF(t, d)$ refers to the term frequency of term t in document d and $IDF(t, D)$ refers to the inverse document frequency of term t in the corpus of documents D.

4.2.2 TF-IDF modified. TF-IDF being capable of discriminating significant words from others with a mathematical formula, we wanted to apply same principles to discriminate real symptoms from surplus symptoms.

Regarding the first factor, the "Term Frequency" (TF) represents the frequency of a word in a document. In our case, the principle is similar but applied to all publications of a disease. To simplify, all publications become a single document. Our new Term Frequency (*TF(t, m)*) designates the frequency of a term t in all publications of a disease m. The raw frequency, denoted by $f_{t,m}$, represents the number of occurrences of the term t in all the publications $\{p\}_m$ of the disease m. In our work, different version of *TF(t, m)* were implemented. This was done in order to optimize our statistical measurement. Those implementations are grouped in Table 1.

Version	Name used in graphs	Formulation
Binary	Binary	{0, 1}
Raw count	RawCount	$f_{t,m}$

Logarithmic Normalisation	LogNorm	$log(1+f_{t,m})$
Double Normalisation 0	MinMaxNorm	$\frac{f_{t,m}}{max_{\{t' \in M\}} f_{t',m}}$

Table 1: Table of different versions of our Term Frequency

Concerning the term IDF, like TF, it has been similarly upgraded to support the notion of disease. Thus, the Inverse Document Frequency becomes the Inverse Disease Frequency. The idea is to create a factor that has the following behavior: the more a term is present in all disease publications, the less significant it is for any specific disease. To do this, it is important to clearly define the meaning of being *present*, that is, to define its inter-disease frequency. Two different definitions to define its inter-disease frequency (its frequency in the disease set) were adopted. First, the inter-disease frequency 1 $(f_{inter1,t})$ refers to the number of diseases for which the term t appears in all publications $\{p\}_m$, i.e. mathematically:

$$f_{inter1,t} = Card(\{m \in M : t \in \{p\}_m\}) \quad (2)$$

Second, the inter-disease frequency 2 ($f_{inter2,t}$) refers to the sum, among the disease set $\boldsymbol{M}$, of the TF linked to the term t, i.e. mathematically:

$$f_{inter2,t} = \sum_{m \in M} TF(t,m) \quad (3)$$

This second frequency attempts to add an additional notion to the first. Where the first adds 1.0 each time a term is found in disease publications, the second attempts to add a weighted representation of the presence of the term in disease publications, TF. Note that if you choose to use the "Binary" type TF in the second frequency formula, you get a measurement identical to the first frequency. The different versions of IDF that are used are in Table 2:

Version	Formulation
Unary	**1**
Inter-disease frequency 1	$f_{inter1,t}$
Inverse inter-disease frequency 1	$\frac{Card(M)}{f_{inter1,t}}$
IDF_1	$log\left(\frac{Card(M)}{f_{inter1,t}}\right)$
IDF_1 Smooth	$log\left(1+\frac{Card(M)}{f_{inter1,t}}\right)$
Probabilistic IDF_1	$log\left(\frac{Card(M)-f_{inter1,t}}{f_{inter1,t}}\right)$
Inter-disease frequency 2	$f_{inter2,t}$
Inverse inter-disease frequency 2	$\frac{Sum_{tot}}{f_{inter2,t}}$
IDF_2	$log\left(\frac{Sum_{tot}}{f_{inter2,t}}\right)$
IDF_2 Smooth	$log\left(1+\frac{Sum_{tot}}{f_{inter2,t}}\right)$
Probabilistic IDF_2	$log\left(\frac{Sum_{tot}-f_{inter2,t}}{f_{inter2,t}}\right)$

Table 2: Table of different versions of our "Inverse Disease Frequency"

Sum_{tot} represents the sum, among the set of terms corresponding to symptoms $\boldsymbol{T}$, of $f_{inter2,t}$ whose formula is the following:

$$Sum_{tot} = \sum_{t \in \boldsymbol{T}} f_{inter2,t} \quad (4)$$

4.2.2 Choice of versions used. The new measure of TF-IDF value of importance will be used to rank the extracted items in a decreasing order, thus classifying, for each disease, the symptoms from most to least important. Given the different versions that have been presented, each pair combination has been tested in this project (TF, IDF) to find the most efficient. The most efficient is defined by the one that tends to assign high values to the extracted elements that are really important.

5 Validation of the platform

5.1 How to Test the Platform?

Testing the quality of symptoms extracted by our system can be done in different ways. The first way to approach this task is to show the elements extracted by the system to one or more rare disease experts. This approach is obviously very time-consuming, and the challenge is to find and recruit knowledgeable experts to deal with it. The second way, that the team chose to implement, is to do it automatically with a program and a verification set. Luckily, a reputable verification dataset is available through the Orphadata[25] database. Orphadata's mission is to provide the scientific community with a comprehensive, high quality and freely accessible data set on rare diseases and orphan drugs in a reusable format. The Rare Disease Phenotypes dataset[2] is an XML file updated monthly which, as its name suggests, contains the phenotypes associated with each rare disease. It represents the ground truth of the real symptoms of our rare diseases. Each of these symptoms is given using the terms of the HPO. Consequently, the vocabulary in this dataset is the same as in the dictionary used for our NER.

5.2 Selection of Performance Measures

In order to measure the quality of our extraction, it is necessary to define the performance measures that will be used. First, the accuracy, recall and F-score measures will be calculated from the number of true positives, false positives and false negatives. A final statistical measure, which created specifically for this project, was implemented to give an idea of the quality of the formula used to calculate the importance value of our symptoms. Used to determine the best combination of couples (TF, IDF), this measurement depends on various other formulas that are detailed in this section.

In our approach, each extracted symptom has a significance value; a weight. Once this value is calculated, it is used by sorting the extracted symptoms of each disease in descending order (the higher the value is, the more important the element is and the closer its rank is to 1). This weight thus makes it possible to obtain a rank RS_{mi} for a symptom i of a disease m in this classification. Once all the ranks RS_{mi} are found for each symptom i, true symptoms are identified. Once all true symptoms are identified,

[2] http://www.orphadata.org/cgi-bin/inc/product4.inc.php

the average of all RS_{mi} where i nominates a true symptom of disease m is computed. This average, which is called $RMVS_m$, can be calculated as follows:

$$RMVS_m = \frac{\sum_{i \in TS_m} RS_{mi}}{n_m}, \quad (5)$$

where n_m denoting the number of true symptoms extracted (true positives) for disease m and TS_m denoting the set of all true symptoms of disease m. Since this measure concerns only one disease, a measure that achieves the mean for all diseases need to be used. This measure, which is called $RMVS$, has the formula (6):

$$RMVS = \frac{\sum_{m \in M} RMVS_m}{Card(M)} \quad (6)$$

Figure 1 illustrates the calculation of $RMVS_m$ from a sample of extracted symptoms.

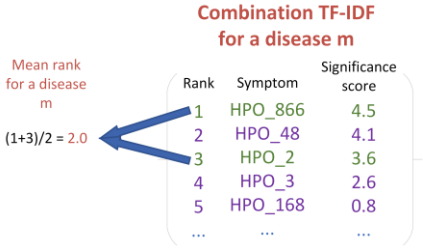

Figure 1: Illustration of computing $RMVS_m$ (Symptoms in green are true symptoms and symptoms in purple are false symptoms, i.e. false positives)

5.2 Dictionary based NER only

Before the comprehensive approach we detailed earlier, we wanted to evaluate what the NER dictionary could extract by default. Thus, by comparing our set of verification data with our extracted symptoms, we obtain performances of 1.83% in accuracy, 59.66% in recall, 3.55% in F-Score. Also, by displaying different performance measures according to the number of publications, we obtain the graph of figure 2.

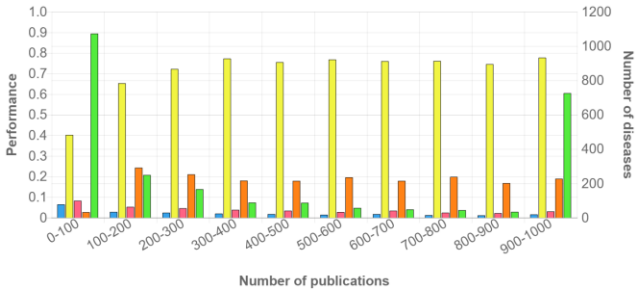

Figure 2: Performance by the number of publications (Precision in blue, Recall in yellow, F-score in pink, RMVS in orange, Number of diseases in green)

5.2.2 Find the best combinaison. The best possible combination (TF, IDF) is determined by achieving a minimum $RMVS$. With 11 distinct IDF types and 4 distinct TF types, 44 possibilities are tested. With a minimum $RMVS$ of 151.72, the combination with TF type "Normalisation double 0" and IDF type "IDF_1 Smooth" is the best according to this criterion.

5.2.3 Find the threshold. Once the best combination is determined, we determine the optimal threshold that maximizes the F-score. The approach envisaged for the threshold search is a simple approach which consists in testing a set of possible values in a given interval. For example, taking the interval [0.0, 5.0] and a step of 0.01, the set of threshold values tested will be of the form {0.0, 0.01, 0.02, 0.03, ..., 4.98, 4.99, 5.00}. We opted for this simple approach because during our tests, we saw that it could be achieved in a reasonable time. About the choice of interval and step, we proceeded by successive tests, trying each time to target the zone presenting the best performances, that is to say by increasing the step and by decreasing the interval towards this zone. Finally, the threshold that maximizes the F-score to a value of 17.17% has the value 0.0415. The Figure 3 shows the evolution of the performance measurements as a function of the threshold values.

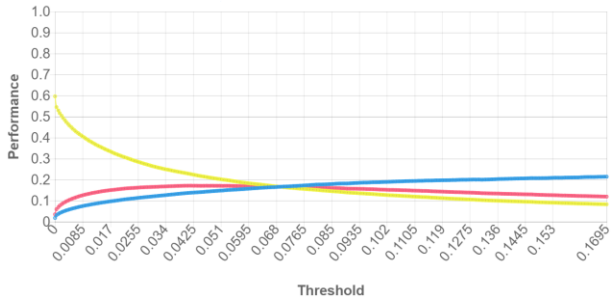

Figure 3: Evolution of performances by threshold value (Precision in blue, Recall in yellow, F-score in pink)

5.3 Interpretations and discussion

The experiments on automatic symptoms extraction was very beneficial for the future of our research. While there is still a lot of research to do, a deeper analysis made us learn valuable knowledge that other researchers may find useful if undertake a similar endeavor. The reader should keep in mind that this task is very novel in research and that only a few similar works can be found in the literature. Four major headwinds that our team faced are discussed in this section.

5.3.1 Presence of symptoms in the scientific literature. Concerning the results obtained using only the dictionary approach, as expected, a low accuracy and a rather high recall are obtained. Indeed, this approach makes it possible to find a majority of real symptoms as long as they are present in the scientific literature, but on the other hand, it also finds many symptoms that are not related to the disease. If the recall does not exceed 60% on all diseases, it means that, on average, almost 40% of symptoms are missing in the disease publications. Our assumption is that there must be a lack of publications to analyze in our database and that PMC does not contain all the necessary rare disease content. More concretely, it also means that our extraction can never have a recall higher than 60% and therefore the F-score is limited to 75%.

5.3.2 Rarity profiles. For those unfamiliar with rare diseases, it may sound surprising but rare diseases actually have a very different profiles of rarity. Four rarity profiles emerge from the lot. First, there are the common rare diseases. These diseases are well-known by experts and have been studied a lot in the

literature. For instance, Tuberculosis or Alzheimer's disease are in this category. Second, there are localized diseases. These diseases can be common but limited to a small geographical area. The Spastic ataxia Charlevoix-Saguenay and the dystrophia myotonica type 1 are to examples of localized diseases in the province of Quebec, Canada. Third, there are the average rare diseases, which are little known and affect a small percentage of the world population. Finally, there are the very rare rare diseases. In some case, only few persons on the planet were diagnosed. They often are the subject of only one or two scientific papers.

Of all rare diseases listed on Orphanet, 5894 give, at least, one publication when doing research on PMC, or 93.09%. We therefore have a small proportion (6.91%) of diseases from which it was impossible for us to extract anything. Also, it is important to note that the verification set from Orphadata used for evaluation, gives the real symptoms of, only, 3056 diseases. As a result, overall, we were only able to evaluate the symptoms of 2572 diseases that are present in this verification set and which possess, at least, one publication. And out of these 2572 diseases, 1073 diseases have less than 100 publications, or 41.72%. This heterogeneity in the rarity profiles probably affect significantly our results. Another problem is that many rare diseases are a combination of other rare diseases and therefore the symptoms of one may affect the symptoms of the other. The contamination factor is, however, difficult to evaluate.

5.3.3 Quality of the Data. The third challenge regards the data that we gather from the automated mining of papers. There is inherent bias in the extraction of symptoms using scientific publications. One of them is the point of view of the authors. Let's suppose that the algorithm mine the symptoms of a hypothetical syndrome X which affect the heart and change the skin properties. The symptoms regarding the skins might not significantly impact the patients, while the heart ones are mortal. In that case, it is likely that most papers on the subject will adopt a cardiovascular point of view which may impair the ability of the algorithm to detect the skin symptoms.

5.3.4 TF-IDF strategy. It is also possible that TF-IDF's own strategy may not be the best approach. Indeed, with our version of TF-IDF, the symptoms that are present globally, i.e. among a large number of diseases, see their value of importance diminished. This single criterion of "global" presence among diseases seems to work in part because we were able to raise the F-score to 17.17%. We believe that adding other criteria or statistical measures, for example concerning syntactic analysis of sentences containing terms referring to symptoms, could allow a significant improvement in extraction performance.

6 Conclusion

In this paper, we introduced a new approach to the problem of automatic symptom extraction based on a statistical measure inspired by TF-IDF. This work aims to contribute to the improvement of the quantity and the quality of information in the field of health, to promote the dissemination of knowledge and to reduce costs of information maintaining in health information platforms. The results of this work show that symptom extraction is an area that still needs the attention of researchers in the future to build a tool capable of quality extraction. In future work, the team plans to focus on improving text mining and extracting other types of elements such as remedies, drugs.

ACKNOWLEDGMENTS

This project success was conducted with the financial support received from the Université du Québec à Chicoutimi and the National Sciences and Engineering Research Council of Canada.

REFERENCES

[1] Canadian Institute for Health Information (CIHI), *National Health Expenditure Trends, 1975 to 2017 (report)*, 45 pages, 2017.
[2] Heinig, S. J., Dev, A. and Bonham, A. C. The US Public' s Investment in Medical Research: An Evolving Social Contract. *The American journal of the medical sciences*, 351, 1 (2016), 69-76.
[3] Slade, A., Isa, F., Kyte, D., Pankhurst, T., Kerecuk, L., Ferguson, J., Lipkin, G. and Calvert, M. Patient reported outcome measures in rare diseases: a narrative review. *Orphanet journal of rare diseases*, 13, 1 (2018), 61.
[4] Molster, C., Urwin, D., Di Pietro, L., Fookes, M., Petrie, D., van der Laan, S. and Dawkins, H. Survey of healthcare experiences of Australian adults living with rare diseases. *Orphanet Journal of Rare Diseases*, 11, 1 (March 24 2016), 30.
[5] Mebert, I. *Psychological Consequences of Rare Diseases*, Kantonsschule Zug, 105 pages, 2012.
[6] Zurynski, Y., Frith, K., Leonard, H. and Elliott, E. Rare childhood diseases: how should we respond? *Archives of Disease in Childhood*, 93, 12 (2008), 1071-1074.
[7] Inserm. 1999. Orphanet: À propos des maladies rares. Retrieved from http://www.orpha.net/consor/cgi-bin/Education_AboutRareDiseases.php?lng=FR.
[8] Epirare. 2015. Epirare - European Platform for Rare Diseases Registries. Retrieved from http://www.epirare.eu.
[9] NORD. 1969. Home - NORD (National Organization for Rare Disorders). Retrieved from https://rarediseases.org.
[10] Gupta, V. and Lehal, G. S. A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, 1, 1 (2009), 60-76.
[11] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B. and Kochut, K. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919* (2017).
[12] Sailaja, V. N., Padmasree, L. and Mangathayaru, N. Survey of Text Mining Techniques Challenges and their Applications. *International Journal of Computer Applications*, 146, 11 (2016), 30-35.
[13] Liu, Y., Liang, Y. and Wishart, D. PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. *Nucleic Acids Research*, 43, W1 (2015), W535-W542.
[14] Mahmood, A. S. M. A., Wu, T.-J., Mazumder, R. and Vijay-Shanker, K. DiMeX: A Text Mining System for Mutation-Disease Association Extraction. *PLoS ONE*, 11, 4 (2016), 1-26.
[15] Holat, P., Tomeh, N., Charnois, T., Battistelli, D., Jaulent, M.-C. and Métivier, J.-P. 2016. Weakly-Supervised Symptom Recognition for Rare Diseases in Biomedical Text. In *Proceedings of the International Symposium on Intelligent Data Analysis*. Springer, Stockholm, Number of 192-203.
[16] Schmid, H. Treetagger| a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43 (1995), 28.
[17] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. and McClosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, Number of 55-60.
[18] Europe_P.M.C._Consortium Europe PMC: a full-text literature database for the life sciences and platform for innovation. *Nucleic acids research*, 43, Database issue (2015/01// 2015), D1042-1048.
[19] Alias-i. 2008. LingPipe. Retrieved from http://alias-i.com/lingpipe/.
[20] Sayers, E. A General Introduction to the E-utilities. *Entrez Programming Utilities Help [Internet]. Bethesda: National Center for Biotechnology Information* (2010).
[21] Liu, Y., Liao, W.-k., Choudhary, A. and Li, J. *Parallel Data Mining Algorithms for Association Rules and Clustering*, 32-31 pages, 2007.
[22] Vukotic, V., Claveau, V. and Raymond, C. 2015. IRISA at DeFT 2015: Supervised and Unsupervised Methods in Sentiment Analysis. In *Proceedings of the DeFT, Défi Fouille de Texte, joint à la conférence TALN 2015*, Caen, France, Number of.
[23] Köhler, S., et al. The Human Phenotype Ontology in 2017. *Nucleic Acids Research*, 45, D1 (2017), D865-D876.
[24] Stephen, R. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60, 5 (2004), 503-520.
[25] Inserm. 1999. Orphadata: Free access data from Orphanet. Retrieved from http://www.orphadata.org.