A Tagged Corpus for Automatic Labeling of Disabilities in Medical Scientific Papers

Carlos Valmaseda, Juan Martinez-Romo, Lourdes Araujo
NLP&IR Group, Universidad Nacional de Educación a Distancia (UNED)
28040 Madrid, Spain
carlosvalmaseda@gmail.com, juaner@lsi.uned.es, lurdes@lsi.uned.es

Abstract

This paper presents the creation of a corpus of labeled disabilities in scientific papers. The identification of medical concepts in documents and, especially, the identification of disabilities, is a complex task mainly due to the variety of expressions that can make reference to the same problem. Currently there is not a set of documents manually annotated with disabilities with which to evaluate an automatic detection system of such concepts. This is the reason why this corpus arises, aiming to facilitate the evaluation of systems that implement an automatic annotation tool for extracting biomedical concepts such as disabilities. The result is a set of scientific papers manually annotated. For the selection of these scientific papers has been conducted a search using a list of rare diseases, since they generally have associated several disabilities of different kinds.

Keywords: corpus, annotation, medical concepts, disabilities

1. Introduction

The study of the relationships between different elements of the biomedical domain is essential for further progress in the area. Great efforts are being devoted to identify some of these relationships, such as interactions between proteins, genes-diseases associations or adverse drug effects. The way to deal with these problems usually involves the identification of some of these relationships by experts. For both addressing this problem, and finding specialized terminology related to a specific aspect of the domain it is essential the annotation of corresponding concepts such as diseases, genes, proteins, etc.

In this paper we address the annotation of a type of concept that is not collected in previous works, at least not specifically. This is the identification of expressions related to disabilities. While some disabilities are included among the symptoms of some biomedical domain ontologies, they are only a few cases and their identification needs to be tackled in a specific way. In this paper this problem is addressed, that although shares some aspects with the annotation of concepts in the biomedical domain, also presents particular aspects, since references to disabilities can be more freely expressed that references to diseases, genes, proteins, etc.

References to disabilities supports all kinds of syntactic, morphological, and semantic variations. For instance, for the same disabilities, the following variants could be found:

- I can not move my left leg
- Mobility limitations in lower limbs
- The leg does not respond to the patient

Therefore in this context, to apply natural language processing (NLP) techniques becomes more relevant.

Taking into account that there are currently no resources with which to evaluate a possible system that tries to detect these disabilities, we proposed to develop a corpus of documents which includes several disabilities.

As experimental framework, we focused on the disabilities associated with rare diseases (RD) for several reasons. On the one hand it is a problem of great importance given the limited available information, and therefore resources to promote their detection and treatment.

On the other hand Orphanet[1], the international organization of the RDs and orphan drugs, has created a specialized collection of texts dedicated to professionals and social service providers; the Orphanet Encyclopedia for professionals. It focuses on the disabilities associated with a specific RD. These profiles for every disability provide a brief overview of the medical aspects of the disease validated by medical experts, and include a description of disabilities experienced by patients.

This information will allow us to develop more accurate annotation criteria, with which the annotators do their work.

Finally, in Orphanet are indexing the functional consequences of each RD with the Orphanet Functioning Thesaurus (de Chalendar et al., 2014), an adaptation of the International Classification of Functioning, Disability and Health for Children and Youth (ICF-CY (Organization., 2007)), which includes additional terms to describe cognitive, sleep, temperament and behavior disorders.

This paper is structured as follows: Section 2 describes the way in which the documents that form the corpus were collected, and the source of the annotated disabilities. Section 3 shows the methodology used for the annotation process and some data about the resulting corpus. Finally, we draw conclusions and point out possible directions for future steps in Section 4.

2. Evaluation Corpus

Orphanet provides a group of diseases for which several experts have associated their disabilities. For this reason, a subset of these diseases was considered to build the evaluation corpus. The randomly selected diseases were:

[1] http://www.orpha.net

- Angelman syndrome (AS) is a neurogenetic disorder characterized by severe intellectual deficit and distinct facial dysmorphic features.
- Cockayne syndrome (CS) is a multi-system condition characterized by short stature, a characteristic facial appearance, premature aging, photosensitivity, progressive neurological dysfunction, and intellectual deficit.
- Dystrophic epidermolysis bullosa (DEB) is a form of inherited epidermolysis bullosa (EB) characterized by cutaneous and mucosal fragility resulting in blisters and superficial ulcerations that develop below the lamina densa of the cutaneous basement membrane and that heal with significant scarring and milia formation.
- Fragile X syndrome (FXS) is a rare genetic disease associated with mild to severe intellectual deficit that may be associated with behavioral disorders and characteristic physical features.
- Norrie disease (ND) is a rare X-linked genetic vitreoretinal condition characterized by abnormal retinal development with congenital blindness. Common associated manifestations include sensorineural hearing loss and developmental delay, intellectual disability and/or behavioral disorders.
- Pendred syndrome (PDS) is a clinically variable genetic disorder characterized by bilateral sensorineural hearing loss and euthyroid goiter.

We have chosen these diseases because the World Organization for Rare Disorders, Orphanet, provides documents to the associated disabilities. At the moment they only have a small number of diseases for which they have identified their disabilities, but they are working on expanding this set.

For harvesting scientific papers we have used Google Scholar (GS). For every disease, it has been conducted a search in GS by downloading only those documents in which the name of the disease appeared in the article title. This restriction has been used to ensure that the document does not make a simple mention of the disease, but the article treats the disease as its main theme in the text.

From the results, we have downloaded that papers for which there was a free PDF version available. Finally, the items have been transformed into text with the pdftotext tool[2].

2.1. Associated Disabilities to Rare Diseases

For the annotation of the corpus, we used the list of disabilities associated to diseases that the Orphanet experts had developed. In this article, the corpus has been designed for scientific articles in English and therefore the associated disabilities have been used in the same language. The disabilities associated with each of the rare diseases is as follows:

Angelman syndrome (AS):

- very low learning ability
- difficulty to mimic
- difficulty to memorize the gestures
- almost non-existent language
- slow execution of the instructions
- high fatigue
- attention disorders
- concentration disorders
- can not be completely autonomous

Cockayne syndrome(CS):

- neurological disorders
- intellectual deficit
- gradual loss of hearing
- gradual loss of sight
- difficulty performing certain activities of daily life
- difficulty to move
- difficulty to communicate with other
- late visual impairment (disability)
- late hearing impairment(disability)

Dystrophic epidermolysis bullosa (DEB):

- delayed walking age
- interfere walking
- aesthetic consequences
- psychological disability
- difficulty to accept
- difficulty to be accepted
- depression
- behavioral problems
- Sleep may also be affected
- psychological problems
- impact on the autonomy
- impact on the locomotion
- difficulty walking
- difficulty writing
- difficulty catching objects

[2] http://www.glyphandcog.com/XpdfText.html

- difficulty manipulating objects
- personal hygiene problems

Fragile X syndrome (FXS):

- mental retardation
- behavioral disorders
- low learning ability
- difficulty to reason
- difficulty to understand
- difficulty to memorize things
- difficulty speaking properly
- behavioral problems
- communication difficulties
- difficulty to reading
- difficulty to writing
- intellectual deficit
- communication problems
- socialization problems
- autonomy problems

Norrie disease (ND):

- visual deficit
- gradual loss of hearing
- mental retardation
- behavioral disorders
- difficulty performing activities of daily living
- difficulty to move
- difficulty to communicate with others
- disrupt communication
- intellectual deficit
- impaired concentration
- attention disorders
- memory disorders
- cognitive impairment
- difficulty speaking properly
- behavioral disorders
- autonomy problems

Pendred syndrome (PDS):

- congenital deafness
- hearing loss
- problems in language acquisition
- severe or profound hearing loss
- difficulty for learning
- difficulty for communication
- balance disorders
- social integration
- professional integration integration

3. Corpus Annotation

The annotation process was conducted by a group of 3 volunteers. Each person tagged the disabilities found in several scientific articles. The annotation criteria were made available to the annotators taking into account issues such as the specificity of the concept of disability, the scope of the problem or the discontinuity of the disease in the text. One of the main difficulties of the annotation process is the identification of a disability. According to Wikipedia: ”Disability is the consequence of an impairment that may be physical, cognitive, mental, sensory, emotional, developmental, or some combination of these. A disability may be present from birth, or occur during a person’s lifetime.”

One of the criteria used in the annotation was considering a disability as a permanent problem without solution. For example, the main causes of visual impairment according to WHO[3] are distributed as follows:

- Uncorrected refractive errors (myopia, hyperopia or astigmatism): 43%.
- Unoperated cataract: 33%.
- Glaucoma: 2%.

Therefore, in this paper myopia, hyperopia, astigmatism, cataracts or glaucoma are not considered a disability since they are not permanent. Another criteria used was the annotation of the most concrete form of a disability. For example, in the case of finding the string ”severe neuropsychiatric disorders”, the annotated disability should be ”neuropsychiatric disorders”.

For the annotation of disabilities, we have used the ”disc” xml label in order to mark both the start and the end of the disability in the text. The ”disc” label comes from an abbreviation of the word ”discapacidad”, which is the translation of disability in Spanish. A real example of annotation extracted from a document related to Fragile X syndrome (FXS) disease is shown below.

- *Fragile X syndrome, the most common form of $< disc >$ inherited intellectual disability $< /disc >$, is caused by a lack of FMRP, which is the product of the Fmr1 gene.*

[3] http://www.who.int/mediacentre/factsheets/fs282/en/

After the labeling process, only the disabilities that had been annotated by at least two people independently and with the same annotation were considered.

The evaluation of the agreement among annotators was measured by Fleiss kappa value (Fleiss, 1971) obtaining 0.68. In simple terms, the kappa coefficient corresponds to the ratio of observed concordances over the total of observations, having excluded all random concordances. The kappa coefficient takes values between -1 and +1. The value obtained for the annotation of this corpus corresponds to a level "substantial agreement" of agreement.

In Table 1 can be seen the percentage of scientific papers annotated by each rare disease that compose the corpus generated in this work.

Rare Disease	% Annotated Docs
Cockayne Syndrome	13%
Fragile X Syndrome	36%
Angelman Syndrome	3%
Norrie Disease	22%
SPendred Syndrome	26%

Table 1: Annotated documents per rare disease in the corpus.

Table 2 illustrates some statistics about the corpus such as the total number of documents or disabilities in the corpus, mean/min/max of words for every document or disability, and the number of unique names of disabilities in the corpus. We can see that even though the number of papers is still small, the corpus includes more than 1000 disabilities and therefore can be a very useful tool to evaluate a system.

Measure	Value
# Documents/corpus	31
Mean words/doc	5216
Min words/doc	1826
Max words/doc	13206
# Disabilities/corpus	1135
Unique disabilities/corpus	394
Mean words/disability	1.98
Min words/disability	1
Max words/disability	8

Table 2: Statistics about the corpus.

4. Conclusions and future work

This article describes an annotated corpus for evaluating detection of disabilities in medical texts. To create this corpus, we marked first a set of guidelines for that the experts could annotate disabilities in a more accurate manner. The annotators of the corpus were three people who followed these guidelines and labeled a set of documents containing disabilities in english texts about rare diseases.

The documents that do not reach a minimum agreement among the annotators were discarded, resulting a corpus with a value of agreement measured in the Fleiss kappa obtaining a value of 0.68.

In this paper the complexity of annotating disabilities is reflected and therefore value of the resulting corpus for the evaluation of automatic detection concepts such as disabilities.

Future work will include further extension of the gold standard corpus by manually annotating more documents. Furthermore, we believe that multilingualism can aid in the detection of such concepts and therefore we will work on creating a corpus in several languages in order to evaluate future systems that may arise following this line of work.

Acknowledgment

This work has been partially supported by the Spanish Ministry of Science and Innovation within the project EXTRECM (TIN2013-46616-C2-2-R), and the Universidad Nacional de Educación a Distancia (UNED) within the project TwiSE (2013-025-UNED-PROY).

5. Bibliographical References

de Chalendar, M., Daniel, M., Olry, A., and Rath, A. (2014). Rare diseases and disabilities: improving the information available with three orphanet projects. *Orphanet Journal of Rare Diseases*, 9(Suppl 1):O31.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Organization., W. H. (2007). *nternational classification of functioning, disability and health : children and youth*. World Health Organization Geneva.