Creation and Use of Ontology Related to Genes, Syndromes, Diseases and Symptoms for the Classification of Biomedical Texts

C. Pérez de Celis
Facultad de Ciencias de la Computación, Universidad Autónoma de Puebla.

Fátima Ronquillo
Facultad de Ciencias de la Computación, Universidad Autónoma de Puebla.

Emilio Salceda
Instituto de Fisiología, Universidad Autónoma de Puebla

ABSTRACT
This research focuses on analyzing and classifying biomedical articles in the field of neuroscience, with a particular emphasis on scientific articles related to hearing loss. To carry out this task in a more efficient manner, resources as the elimination of stopwords were used. As well, it was implemented the n-gram-based text categorization system along with the use of a domain ontology related with genes, diseases and syndromes, obtaining promising results.

General Terms
Biomedical text mining.

Keywords
Multi-cataloguing, n-grams of letters, ontologies, hearing loss, genes.

1. INTRODUCTION
The text categorization process usually consists of two stages: the first consisting of the delimitation of the classes that divide the subject of our interest and the second focused on the categorization of the texts of interest.

In most applications, the categorization is resolved basing the model in the collection of classes that are dispersed, allowing the existing categorization algorithms having excellent results, given that between them there is a wide line of separation of the classes. The problem is when the evaluation of classes contains a line of narrow spacing between them.

This work presents a different approach to the traditional through the integration of two algorithms of categorization, the use of n-grams of letters for the categorization of partially distant classes and subsequently refining the categorization of documents using the terms of a domain ontology. It was applied this methodology of multi-categorization to biomedical texts, in particular of neuroscience, focusing on the hearing loss. For this type of texts, the process started with a set of categories, suggested by an expert from the area, based in the taxonomy of the hearing loss arising from its etiologic classification.

In subsequent sections the results obtained using the algorithm of n-gram of letters will be introduced. A discussion on the reasons that lead us to use the ontology to improve and allow the multi-classification will be presented too. Then, we'll discuss in detail the strategy of the use of ontologies to improve the classification. Finally we will present our results and conclusions.

2. PREVIOUS RESULTS
The goal of this work, as mentioned above, is the development of a classifier by categories of biomedical texts related to the symptoms of hearing impairments, in particular, hearing loss. For the definition of the classes on the texts related to hearing loss, the levels of the taxonomy of this hearing deficit by its etiology (Figure 1) were took into account.

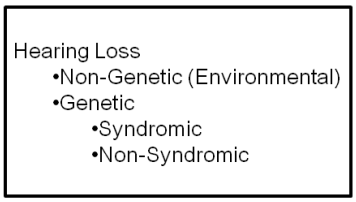

Figure 1. Taxonomy of the hearing loss arising from its etiologic classification.

In the first level of classification it was assumed that between the considered documents may be texts of general scope and texts about hearing loss, defining in this way the first two classes (general vs. hearing loss). At the same time, the texts that belong to the class hearing loss can be divided into texts on not genetics (environmental) or genetic hearing loss, and the latter be further subdivided in syndromic and non-syndromic hearing loss.

Once established the classes, it was developed a corpus of texts in English correlated with them. This corpus was composed of scientific articles selected by a panel of specialists in the topic and was divided into four subcorpus:

• Subcorpus 1 (General). Containing 300 articles in various fields: medicine, computer science, linguistics and computational algorithms applied to medicine.

• Subcorpus 2 (Not genetic hearing loss). Containing 85 articles that deal with cases of not genetic hearing loss, that is to say, hearing loss acquired during the course of a patient's life due to some disease or accident.

• Subcorpus 3 (Syndromic hearing loss). Containing 100 articles dealing with cases of genetic hearing loss, developed through a gene with a syndromic origin.

• Subcorpus 4 (Non-syndromic hearing loss). Containing 100 articles on cases in which the genetic hearing loss is related to a specific gene (or a set of genes) and the patient does not have any syndrome associated with it.

The texts were used in txt converted directly from the original pdf. For the test of the algorithm of n-grams of letters the texts were preprocessed. In the preprocessed text characters were converted to lowercase letters and only the alphanumeric characters were preserved. Established the corpus of work, it was selected the learning algorithm with which experiments were conducted; it was the algorithm proposed in [1].

Three classification experiments were carried out of in the following three levels: I) General vs. hearing loss, II) Not genetic hearing loss vs. genetic hearing loss, and III) Syndromic hearing loss vs. non-syndromic hearing loss. For these tests the corpus was distributed as follows: to the learning corpus was attached 90% of the texts, and to the test corpus 10 % of the texts, in each of the three levels. After the application of the algorithm, the results of the system were evaluated through the cross-validation technique [2] with 11 blocks of textual data (still divided in 90% of texts for the learning corpus and 10% of texts for the test corpus).

To undertake a rigorous evaluation of the system, three algorithms for classification of texts included in the environment Weka [3] were selected. Weka was used because it allows the execution of classification algorithms that use different approaches, such as SVM, decision trees, association rules, functions, and so on. In particular, were selected three algorithms: a classification algorithm based on rules (OneR), an algorithm based on decision trees (J48) and an algorithm based on functions (VFI).

In addition, it was designed and implemented a baseline algorithm to confirm that the system gets better results. The baseline algorithm is a classification algorithm that assigns a set of words to each one of the previously established classes, thus creating a bag of words. When this algorithm classifies a new document, it divides the text into words generating a new bag of words. Subsequently compares this bag of words with the bags of words already assigned to each class. The algorithm will be indexed the document to the class that match to a greater number of words.

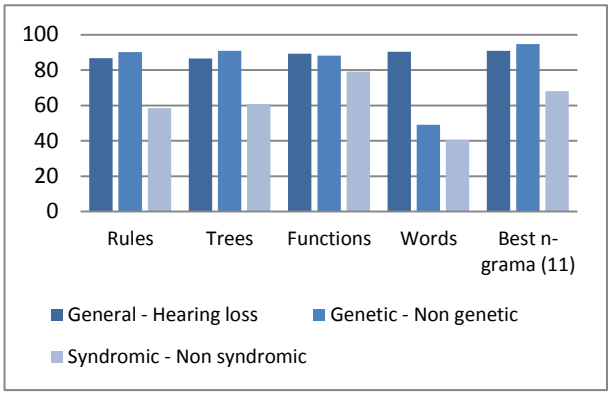

Figure 2. Graph showing the results of the systems for classification.

The graph in Figure 2 shows the average F-score for each of the three levels of classification of the taxonomy. The n-grams algorithm had a considerably good performance compared to the other algorithms. Comparing the n-grams algorithm with the functions algorithm, which had the best performance in the last level, it can be observed that the difference in the two other levels of classification belonging to the general class (hearing loss and environmental hearing loss) genetic hearing loss, where the algorithm of n-gram had a best performance, is constant both in F-score as in time. It should be noted that the system of n-grams takes much shorter time than systems with which it is compared.

The strategy used in the algorithm of n-grams of letters is acceptable; however, the problem with this classification is that there are genes that are associated with the two classes, as shown in Figure 3, in which it is observed that the MYO7A gene belongs to the class related with syndromic affectations caused by Usher's syndrome, but there is also the relationship with the not syndromic genes.

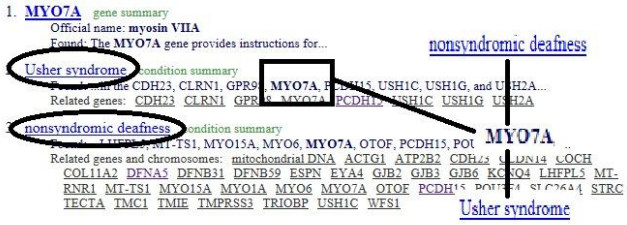

Figure 3. Sample of the relationships of the gen MYO7A.

There are then different genes that may be linked to both classes. This fact could be found verifying the documents poorly classified in which there are documents that make reference to the two bags of words, both syndromic and non-syndromic. This problem is also the reason why all classifiers increased their F-score by wanting to classify the last level. However, for the purposes stated above, it is noted that the system of n-grams maintains very high results and small execution times in all the experiments.

With regard to the last level of classification, it was performed the evaluation of the results to raise a strategy to improve its operation, proposing the use of a domain ontology that will help in the decision-making for the allocation of the class to which belongs the new document [4] [5] [6].

3. ALGORITHM OF ONTOLOGICAL SUPPORT FOR MULTI CATEGORIZATION OF TEXTS

This section presents the experimental results of the use of ontologies for the multi-classification of biomedical texts [7] [8]. For the creation of the ontology, different ontological tools were consulted in order to use existing resources, taking into account that 1) its implementation would be simple, and 2) the employed tool would return a document easy to see in the classification system since this phase would have to be attached to the algorithm of n-gram. The methods of classification are divided into two groups: one that includes supervised methods, as is the case of the algorithm of n-gram, which requires a set of training to create the model of the language that the algorithm will use for its classification, and that where is not generated the training set, as is the use of the ontology, in which is through the data of the ontology that the documents are classified [9].

The ontology developed in Protégé [10] for the field of study of syndromic and non-syndromic hearing loss consists of the following classes: *genes*, *syndromes*, *symptoms* and *diseases*. As it is shown in Figure 4, the relationship between the different classes focus on the class of genes, given that to perform the query of these classes is it departed from the existence of a gene, showing the relationships that it may have with the syndromes, symptoms or diseases. At the bottom of the figure are shown some examples of genes; on the right side it shows the genes and their association with the syndromes, in this particular example the relationship between the Usher syndrome and the Stickler syndrome, the left side shows the relationship between symptoms and diseases with the gene to which they were linked.

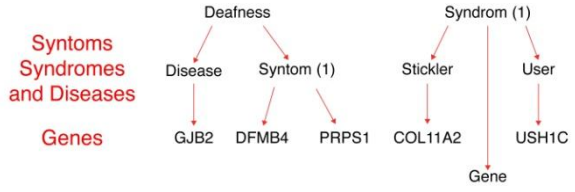

Figure 4. Ontological relations and data.

From the class *genes* is generated a relationship with the other three. The relationship that was named "May cause" allows us to relate the terms of the classes *gene-syndrome*, *gen-symptom* and *gene-disease*. The planning of these relationships was made for the system to have easy access to the data and subsequently does not hinder the query when the classification of the last level of the ontology has to be performed.

The class genes also has a unique property, which indicates whether the gene belongs to the class syndromic, to the class non-syndromic or both, so it is possible to generate the bag of words to be consulted when it is time to make the classification of the documents [11].

The ontology on hearing loss that we used is based on the data of the National Institute of Deafness and other Communication Disorders (NIDCD), and the genes that are found in it were consulted on two pages of genetic organizations, the same that provide information on what type of gene is, the class to which it belongs and, if necessary, the symptom or syndrome associated. The two references that were used for the filling of the ontology are: Genetics Home Reference [12] and HUGO [13].

The filling of the ontology was made in three steps: 1) Collection of the genes associated with hearing loss, as well as syndromes and symptoms, 2) Design of the ontology in the system Protégé, and 3) Filling of the ontology with the data obtained from the references mentioned above. The use of the ontology offers the advantage that after its creation it is possible to continue entering data, which ensures that if data were omitted or increases the existence of any of these 4 classes, the ontology can be updated.

Figure 5 shows a screenshot of the ontology class *genes*, with all its items, this figure serves to show the magnitude of the number of genes that are related to deafness and gives an idea of the information that can be found in the articles dealing with this type of disorders; this is the reason for which is important to have organized in a repository the knowledge base for the classification of the documents related to this type of disorders.

3.1 Operation of the algorithm of classification

The tool Protégé was used for two things, the first consists in generating an ontology with a correct logic, and the second is the representation of the knowledge embodied in the ontology for the implementation of a classification system of biomedical articles related to hearing loss combining the algorithm of n-gram of letters, which had acceptable results in the top two levels, and the use of the ontology for the last level of classification.

When generating the ontology, Protégé creates a document with termination OWL. This document has a description similar to XML, which enables you to view the ontological relationships, the classes and their elements in a hierarchical and relatively easy to read manner.

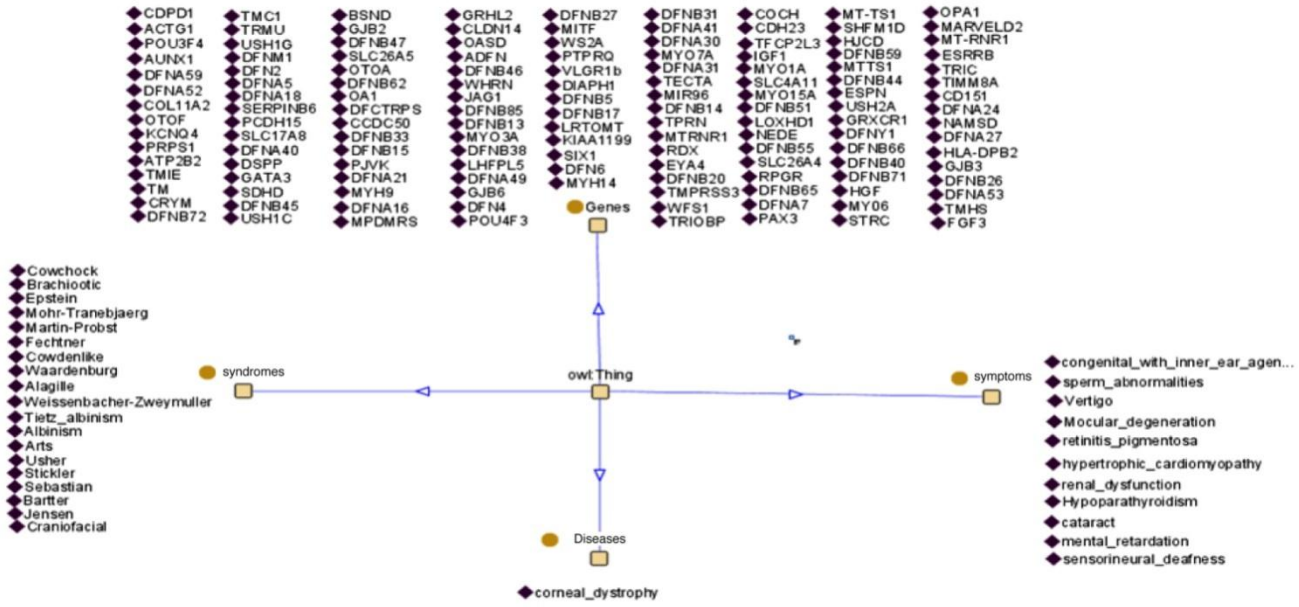

Figure 5. Screenshot of the ontology implemented.

Once having the bags of words, the next step is the classification of the documents. When the document has been cleaned, the words considered defining the content of the paper are stored in a repository. When it concludes the reading of the object document the words stored from the document are contrasted with those found in the 3 bags of words relating to the ontology, the terms in common are accounted for later to decide which class they belong to. The choice of the class is done based on the matches with the bag of words: if the document matches the bag belonging to the class syndromic, the document is assigned to this class, the same is true for the non-syndromic class. If the bag of words from the document to classify has information for both classes, the document is assigned to the third class. The same is true if the document contains genes that are related to both classes.

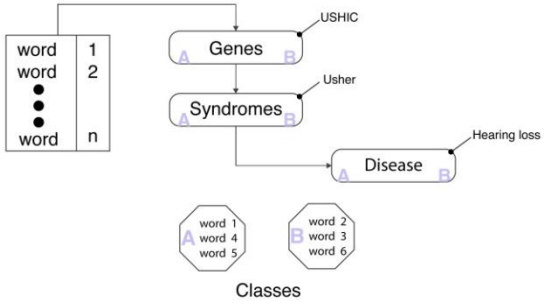

Figure 6. Representation of the classification algorithm.

Figure 6 shows the approach to the classification algorithm explained above and used for the classification level associated with syndromic and non-syndromic hearing loss. In this schema can be noted that the search is done by order of priority of the existing data in the ontology. First is the search in the repository of genes, subsequently syndromes, symptoms and finally diseases.

For the realization of these experiments the corpus before mentioned was used; in this case there was no need to perform the cross validation given that the proposed algorithm belongs to the non-supervised algorithms and does not need a set of training.

4. ANALYSIS OF RESULTS

The evaluation was conducted in the same way that the algorithm of n-grams, using the precision, the coverage and the F-score. The results of our experiments were splitted in order to be able to better show the operation of the classifier and to understand more precisely the reason why the algorithm has a good performance as well as to understand the need for creating a new class Both.

In the first place, the results were divided into two groups: those that belong to the class *syndromic* were assigned to a group, and those that belong to the class *non-syndromic* were assigned to a second group. Below are the results for both groups

Well classified: these documents are those who have genes that only belong to one class; for example, in the non-syndromic class, the GJB2 gene only belongs to this type of hearing loss if the article contains this gene in their recovered words; thus, the system retrieves it and concludes that the document belongs to the non-syndromic class, by what is counted as a document well classified.

Contains genes syndromic and non-syndromic: in this group are considered documents that include genes of the syndromic class, as well as genes belonging to the

nonsyndromic class. Does not consider the number of genes in each class, only the existence of the same, with that you have a gene from each class enters at this point.

Both classes: As mentioned before, there are genes that in the literature belong to the class *both*, syndromic and non-syndromic; this is the case with the MYO7A gene: when the system finds this gen, assigns the document to both classes.

Wrong classified: in this group are the documents that the system assigns to the class *a*, but the document belongs to the class *b*.

Impossible to classify: Includes documents for which the system could not find any gene in the ontology, thus could not infer the class to which the document belongs. The reasons why this occurs can be that the document could not be completely converted from pdf to txt and in the converted section there were not genes, or that the gene is misspelled in the paper and therefore it is not in ontology.

Impossible to access: these are documents to which the system could not access due to an error in the reading of the file. This can happen in pdf documents protected by their authors. In these cases the converter only encounters characters that do not belong to the alphanumeric alphabet and when the document is cleaned there is no usable information.

Once taken into account the above points, the results of the experiments are shown in Table 1.

Table 1. Results of the classification.

Non-Syndromic	%	% Total
Well classified	50.3875969	82.1705427
Contains genes syndromic and non-syndromic	20.1550388	
Both classes	11.6279070	
Wrong classified	5.4263566	17.8294574
Impossible to classify	9.3023256	
Impossible to access	3.1007752	
Syndromic	**%**	**% Total**
Well classified	27.6190476	78.0952381
Contains genes syndromic and non-syndromic	12.3809524	
Both classes	38.0952381	
Wrong classified	14.2857143	21.904762
Impossible to classify	6.6666667	
Impossible to access	0.9523810	

The overall performance (documents well classified) of the algorithm using an ontology for the creation of bags of words for the classes of syndromic and non-syndromic hearing loss was 80.12, presenting a considerable improvement in comparison with the algorithm of n-gram of letters, as was expected for this system.

Figure 7 shows the comparison with the different classification methods used in this research for the third categorization level corresponding to syndromic deafness and non-syndromic hearing loss.

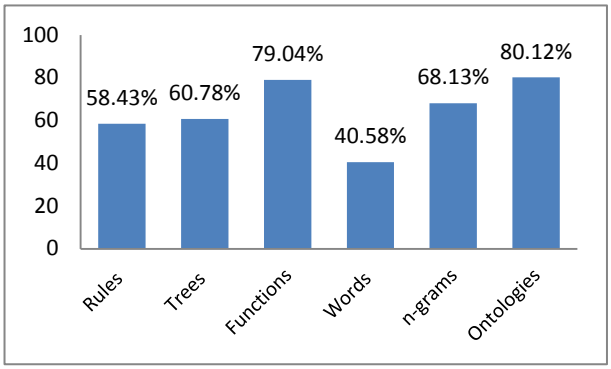

Figure 7. Results of the last level of classification.

5. CONCLUSIONS

Based on the results reported here, it is possible to assure that the algorithm of n-grams is a good candidate for the task proposed in this research. However, as we pointed out in the preceding paragraphs, the performance of the algorithm of n-grams in the last level of the taxonomy is not as good as in the two previous levels, so the coupling of a second algorithm was necessary to improve the classification at this level. In this way, the algorithm based on the ontology serves as a complement of the algorithm based on n-grams of letters, while maintaining a high F-score for the three levels of classification. In summary, with the proposed solution it is possible to have the efficiency and adaptability of the algorithm of n-grams of letters for documents where the context is wide or semi wide, in a short time compared with other systems, and it is still possible to improve the overall accuracy with the algorithm based on ontologies.

6. REFERENCES

[1] Ronquillo, Fátima-Itzel; Pérez de Celis, Concepción; Sierra , Gerardo; da Cunha , Iria; Torres-Moreno, Juan-Manuel (2011). «Automatic classification of biomedical texts: experiments with a hearing loss corpus». En Ding, Yongsheng; Peng, Yonghong; Shi, Riyi; Hao, Kuangrong; Wang, Lipo (eds.). 4th International Conference on Biomedical Engineering and Informatics,**BMEI** 2011. *1*674-1679. Shanghai, China: IEEE. ISBN 978-1-4244-9351-7

[2] Amari S., N. Murata, K. R. M□uller, M. Finke y H. H. Yang. 1997. Asymptotic statistical theory of overtraining and cross-validation. IEEE Transactions on Neural Networks, 8(5):985-996.

[3] Hall. M., Eibe F., Holmes G., Pfahringer B., Reutemann P. y Witten I. H. 2009. The WEKA Data Mining Software: An Up-date. SIGKDD Explorations, 11(1):10-18.

[4] Hmway Hmway Tar, Thi Thi Soe Nyunt. 2011. Enhancing Traditional Text Documents Clustering based on Ontology, International Journal of Computer Applications (0975 – 8887) Volume 33– No.10, November 2011, pp. 38-42.

[5] G.Bharathi, D.Venkatesan. 2012. Study of Ontology or Thesaurus based Document Clustering and Information Retrieval, Journal of Theoretical and Applied Information Technology, 15 June 2012. Vol. 40 No.1, Pags. 55-61.

[6] Stephan Bloehdorn, Philipp Cimiano, and Andreas Hotho. 2006. Learning Ontologies to Improve Text Clustering and Classification, From Data and Information Analysis to Knowledge Engineering Studies in Classification, Data Analysis, and Knowledge Organization 2006, pp 334-341.

[7] Irena Spasic, Sophia Ananiadou, John McNaught and Anand Kumar. 2005. Text mining and ontologies in biomedicine: Making sense of raw text, Henry Stewart Publications 1467-5463. Briefings in Bioinformatics. Vol 6. No 3. 239–251. September 2005, pp.. 239-251.

[8] Alexander Maedche and Ste_en Staab. 2000. Mining Ontologies from Text**,** R. Dieng and O. Corby (Eds.): EKAW 2000, LNAI 1937, pp. 189-202, Springer-Verlag Berlin Heidelberg 2000.

[9] N. Dragu, F. Elkhoury, T. Miyazaki, R.A. Morelli, and N.D. Tada, 2010. Ontology-Based Text Mining for Predicting Disease Outbreaks. ;In Proceedings of FLAIRS Conference. 2010.

[10] Protégé, http://protege.stanford.edu/

[11] S. Bloehdorn and P. Cimiano and A. Hotho and S.Staab. 2005. An Ontology-based Framework for Text Mining, LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, Vol. 20, Nr. 1 (May 2005) , pp. 87-112.

[12] Genetics Home Reference, http://ghr.nlm.nih.gov/

[13] HUGO Gene Nomenclature Committee (HGNC), http://www.genenames.org/