A proposal of lexical resources' development for ontological learning in the domain of speech disorders

Stephanie Vázquez G., María Somodevilla G., Ivo Pineda T., Concepción Pérez de Celis.
Facultad de Ciencias de la Computación, BUAP.
{stephanie.vazquez, mariajsomodevilla, ivopinedatorres, mpcelish} @gmail.com

Language & Knowledge Engineering

Abstract

Speech disorders in children are a condition that could reduce the opportunity to access education, health care and in the future could mean a worse socioeconomic outcome. Therefore, early diagnosis and timely therapy is really important to reduce their impact in later stages of life. This paper presents a method for the gathering of data for a corpus related to Speech Disorders in children; such corpus will serve as the base to generate a semi-automatic ontology intended as a tool for therapists to help in the diagnosis and shape up of a therapy strategy.

Introduction

A ***speech disorder*** is the difficulty to produce or to create the specific speech sounds to communicate. According to Global Disability Rights 7.5% of the population in Mexico has some disability (about 9.17 million people) and 4.87% of people with disability has some type of speech disorder (0.45 million people). In kids and young people the speech disabilities are in some cases twice or four times higher than in adults. The importance of the early detection and diagnosis of a speech disorder abides in the social, economic and educative impact that such disorders have in the life of infants. Technology is used in order to assist in the process of diagnosis and treatment of some speech disorders in children. Ontologies give an unambiguous and well defined structure for a clear and accurate representation of the data concerning a particular domain, in this case speech disorders, and thus, becoming a tool for diagnosis. One of the earlier steps in the development of an ontology is the conformation of a ***Corpus***, in this case of documents relatives to the domain of speech disorders. Corpus analysis provide lexical information, morphosyntactic information, semantic information and pragmatic information.

Figure 1. Percent of young people with disability by age.

State of the Art

Within the field of speech and language several works that use Information and Communication Technologies (ICT) have been conducted, focusing on some ailments such as dysphagia, on the automatic classification of the quality of pronunciation when treating disorders such as dyslalia or dysarthria, or an expert system for the initial evaluation of children with possible speech disorders.
Relevant to the building of corpus the main techniques have not varied a lot, and texts in a corpus need to be in electronic form.
In the present work, a method to gather information for the corpus building is proposed. This method also has the flexibility to feedback itself; once the initial dictionary is defined this can be updated with the extended dictionary obtained after completing the several steps into the method.

IR model for the definition of lexical resources

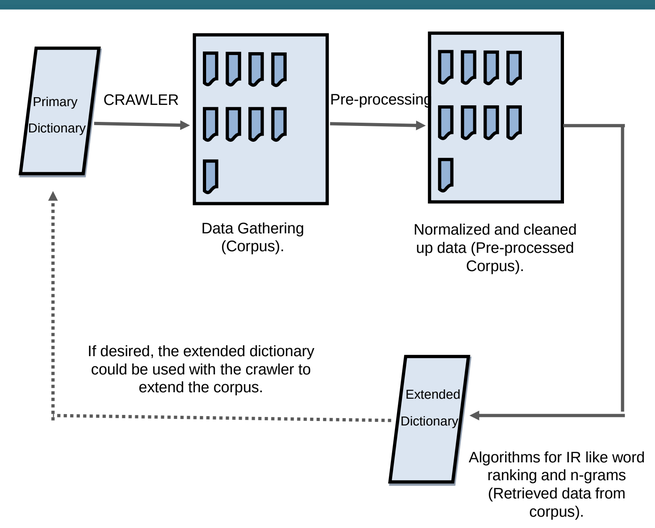

Figure 3. Diagram of the steps to build and process a corpus.

Corpus creation. The building of a corpus is divided into two stages: design and implementation. The main tool to gather the information to build a corpus is a Web crawler. This crawler is fed with some initial *seed* pages to start its task. To find documents relevant to the domain, and not just a list of links and random data contained into the seed page, it is necessary to establish a primary dictionary at the beginning of the crawling.

Dictionary creation. This dictionary is made of some of the more significative words into the domain. A simple way to identify these words is to take the domain taxonomy as a base to gather such list of words. Then the building of the primary dictionary to focus the results of the crawler can be started. The table 1 shows the very first version of the primary dictionary.

Table 1. List of terms from the primary dictionary.

No.	Term(s)	No.	Term(s)
1	Speech	9	Communication disorder
2	Disorder	10	Articulation disorder
3	Dyslalia	11	Rhythm disorder
4	Dysglosia	12	Therapy
5	Dysarthria	13	Speech therapy
6	Dysphemia	14	Logopedic therapy
7	Speech sound disorder	15	Speech development
8	Childhood-onset fluency disorder		

After retrieving relevant data for all the primary dictionary terms the first version of our corpus is finished, but the processing of the corpus is not done.
Data preprocessing. This is done through several algorithms that normalize the texts contained in the corpus. Once all the data gathered into the corpus is normalized the next step in the process can be done. In this step, information retrieval algorithms are implemented. Algorithms like word frequency and stemming are used. After this last step a new list of terms for the extended dictionary is obtained. The more frequent terms found into the corpus are taken and is made a comparison with the primary dictionary terms.

Testing

After applying the pre-preprocessing described in the previous section and the information retrieval algorithms, the terms shown in Table 2 were found to be the most frequent.

Table 2. 15 most frequent terms in corpus.

No.	Term(s)	No.	Term(s)
1	Speech	9	Sound
2	Disorder	10	Communication
3	Child	11	Research
4	Language	12	Services
5	Health	13	Words
6	Information	14	Development
7	Help	15	Medical
8	Therapy		

Observing this data from word frequency, not all of the proposed terms in the primary dictionary are equally relevant to the domain of knowledge. Therefore, the web crawler can be fed with the most frequent terms obtained from the corpus and thus, gather more relevant documents.
Another way to complement the corpus is to include synonyms to the original proposed terms.
Applying again the steps of crawling, pre-processing and IR algorithms more documents were added to the corpus and a new list of the most frequent terms is obtained.

Table 3. Comparison of most frequent terms in corpus.

Primary Dictionary Terms			Extended with Synonyms Dictionary Terms	
No.	Term	Freq	Term	Freq
1	Speech	4,036	Speech	9125
2	Disorder	2,798	Child	5877
3	Child	2,369	Language	5165
4	Language	1,695	Disorder	4792
5	Health	1,332	Sound	2968
6	Information	1,180	Word	2790
7	Help	963	Health	2786
8	Therapy	949	Information	2697
9	Sound	809	Therapy	2695
10	Communication	772	Help	2081
11	Research	742	Service	1939
12	Service	695	Communication	1687
13	Word	694	Development	1485
14	Development	651	Research	1476
15	Medical	640	Medical	1261

The 15 most frequent terms obtained after this expansion in the dictionary resulted to be the same as the ones obtained in the previous step non-using synonyms, just varying the order of appearance in the list. Terms as child and language resulted to be more frequent when synonyms were used as seeds

Acknowledgements

This work has been supported by the VIEP from the BUAP through the project Model of Teaching-Learning Process applying Ontological Engineering.

Conclusions

The corpus building process starts with a list of proposed terms followed by a crawling script execution. Afterwards, normalizing and IR algorithms were applied to include the resulting list of terms into the dictionary; the crawler can be fed again with the new dictionary. Ongoing work consists on the application of word ranking and n-grams algorithms to improve the terms into the dictionary. Besides, work has been doing in expanding with hyponyms and hyperonyms in the list of terms; this task allows adding an additional semantic level to the process and to gather more relevant documents for the corpus.

References

1. Disability in Mexico | Global Disability RightsNow! , http://www.globaldisabilityrightsnow.org/infographics/disability-mexico
2. 4. Loudon, K.: Developing Large Web Applications. O'Reilly Media, California (2010)
3. Robin: What is Corpus?, http://language.worldofcomputing.net/linguistics/introduction/what-is-corpus.html doi:10.3109/17549507.2012.689333
4. Haav, H.M.: A Semi-automatic Method to Ontology Design by Using FCA. CLA. 13-24 (2004)
5. Braga, F., Ebecken, N.: A semi-automatic method for extracting a taxonomy for nuclear knowledge using hierarchical document clustering based on concept sets Fabiane Braga. Int. J. Nucl. Knowl. Manag. 6, 155-169 (2013). doi:10.1504/IJNKM.2013.054496
6. Maedche, A., Staab, S.: Semi-automatic engineering of ontologies from text. Proc. 12th Int. Conf. Softw. Eng. Knowl. Eng. 231-239 (2000)
7. Wynne, M.: Developing Linguistic a Guide to Good Practice Corpora : (2005)
8. Mitchell, R.: Web scraping with Python: collecting data from the modern web. O'Reilly Media, Inc. (2015)
9. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders. (2013)
10. Python Software Foundation, https://www.python.org/
11. Caio Miyashiro: Text Mining and Natural Language Processing - Preprocessing, http://rstudio-pubs-static.s3.amazonaws.com/67435_ca0769f0dbbb4fc4bda5e4535e21fb54.html
12. Zhu, X.: Common Preprocessing Steps. CS769 Spring 2010 Adv. Nat. Lang. Process. 1-3 (2010)