e-Health – For Continuity of Care
C. Lovis et al. (Eds.)

doi:10.3233/978-1-61499-432-9-570

Computer Based Extraction of Phenoptypic Features of Human Congenital Anomalies from the Digital Literature with Natural Language Processing Techniques

Gökhan KARAKÜLAH[a,1,2] Oğuz DİCLE[a,b,2], Özgün KOŞANER[c], Aslı SUNER[d], Çağdaş Can BİRANT[e], Tolga BERBER[f] and Sezin CANBEK[g]
[a] *Dokuz Eylül University, Health Sciences Institute, İzmir, Turkey*
[b] *Dokuz Eylül University, Faculty of Medicine, Department of Radiology, İzmir, Turkey*
[c] *Dokuz Dokuz Eylül University, Faculty of Letters, Department of Linguistics, İzmir, Turkey*
[d] *Ege University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, İzmir, Turkey*
[e] *Dokuz Eylül University, Faculty of Engineering, Department of Computer Engineering, İzmir, Turkey*
[f] *Karadeniz Technical University, Faculty of Science, Department of Statistics and Computer Science, Trabzon, Turkey.*
[g] *Adana Tranining and Research Hospital, Department of Medical Genetics, Adana, Turkey*

Abstract. The lack of laboratory tests for the diagnosis of most of the congenital anomalies renders the physical examination of the case crucial for the diagnosis of the anomaly; and the cases in the diagnostic phase are mostly being evaluated in the light of the literature knowledge. In this respect, for accurate diagnosis, ,it is of great importance to provide the decision maker with decision support by presenting the literature knowledge about a particular case. Here, we demonstrated a methodology for automated scanning and determining of the phenotypic features from the case reports related to congenital anomalies in the literature with text and natural language processing methods, and we created a framework of an information source for a potential diagnostic decision support system for congenital anomalies.

Keywords. Human congenital anomalies, information extraction, natural language processing, text processing, clinical decision support

Introduction

With reference to the World Health Organization 2012 data, it is estimated that 3.2 million fetuses and infants are being affected from human congenital anomalies (HCAs) per year and 270.000 newborns die in the first 28 days due to reasons related

[1] Corresponding author, e-mail: gokhan.karakulah@deu.edu.tr

[2] These authors contributed equally to this work.

with HCA [1]. Although there are variations among ethnical groups in frequency, HCAs, which affect almost 2-3% of the newborns, are accepted as common and major health issue for all ethnical groups [2-3]. HCAs are mostly challenging and require special expertise, considering the number of the phenotypic features and the rare symptoms in the diagnosis phase. In addition, most HCAs are not diseases that comprise of a particular combination of phenotypic features and that are exactly distinct from other HCAs in terms of the related features [4-6]. This situation renders the use of assistive tools such as clinical decision support systems (CDSSs) almost obligatory for the accurate and exact diagnosis of HCAs.

Natural language processing (NLP) techniques have gained a wide use in obtaining and evaluating the information especially in life sciences due to the excessive increase of information in recent years. Besides, the automatized extraction of clinical information, such as signs, symptoms, medications and/or observations, from scientific documents and free-text formatted medical records using NLP should be considered crucial for the development of the CDSSs [7]. Here we aimed at developing a computational strategy to extract the phenotypic features, which characterize HCAs from the case reports in the literature via text processing and NLP methods. In addition to this, by using the extracted information, we created an initial framework of an information base for a potential CDSS in the diagnosis of HCAs.

1. Methods

The term list, which comprises the C16.131-Congenital Abnormalities subcategory of a branch of the Medical Subject Headings (MeSH) tree structure, is downloaded from the MeSH database for determining HCAs [8]. During the creation of the corpus associated with the HCAs, the free-text literature data served via PubMed database was utilized. Using the query words in the structure of "MeSH term[MeSH Major Topic] AND Case Reports[ptyp]" for each HCA, the abstracts of the case reports pertaining each of the MeSH terms were extracted automatically from the database. Forming the query words and the download procedure were performed with the scripts written in Python programming language.

Natural Language Toolkit modules and the scripts written in Python were utilized in all text processing steps on the corpus [9]. At the first step, the string.lowercase() function was used for standardizing the characters in the corpus. In the following pre-processing step, the sentence segments for each abstract were determined with Punkt algorithm [10]. From these segmented sentences in each abstract, the white spaces were removed via tokenization in the third pre-processing steps. In the fourth step the stopwords with little lexical content such as "a", "an", "the", "at", etc. were removed from the sentence. Thus, preventing possible alignment issues to occur in the following steps, during the searching of pre-defined entities in the texts and mapping these with phrases, was aimed. In the last step, each token were parsed into their roots and affixes and token roots were obtained. The Lancaster stemmer in the nltk module was used for the stemming procedure [11]. By this means, possible issues in mapping a token or a token group with the dictionaries or ontologies due to plural suffixes, past tense suffixes, etc. were solved. The same text processing steps were conducted on the more than 10.000 phenotypic feature terms that comprise the Human Phenotype Ontology (HPO). Later, an inverted index was built up for enabling and facilitating the search of the HPO terms in the case reports. While searching and identification of phenotypic features, the following rules were applied; each token comprising the HPO term must (i) have the same document ID, (ii) be in the same sentence, and (iii) be ordered sequentially (the difference between the index positions must equal to 1).

For determining whether the terms defined in the case reports were accurately matched with the findings defined in the documents, the randomly selected case reports were manually examined. In addition, the percentage of the features, which could not be determined due to the use of the HPO, even though they were present in the document, was calculated. For this purpose, the features in a randomly selected document were marked and this was compared to the total numbers of features and the number of the HPO terms associated with that document. When the number of the features that could not be determined reached 100, this number was proportioned to the number of the computationally determined HPO terms.

The maximum entropy model for part-of-speech tagging algorithm was used to define the lexical categories of the tokens [12]. By this method, the tokens tagged as verb for each sentence bearing at least one phenotypic feature were identified and their frequencies were calculated. When the total occurrence frequencies of the verb tagged tokens were ordered descendingly, the first 75% of this total value was computationally determined. The tokens in this group were manually examined and it was established whether there were semantic relations between these tokens and the phenotypic features. With this approach, verb labeled tokens frequently used in the expressions for the presence of phenotypic features were identified.

In addition to this, N-gram analysis was performed to find whether the sentences containing HPO terms had common phrase patterns or not, and if there were, what their frequencies were. In the first 5.000 2, 3 and 4-tuple phrase patterns, the verb labeled tokens were searched, which facilitated the direct inference that whether the manually determined phenotypic features were related the HCA in question. Moreover, phrases such as “patient with”, which did not contain any verb token, but provided information about the presence of an abnormality in the patient, were identified. It was manually determined that the aforementioned phrase patterns could positionally precede or follow the findings; and they could be positive or negative semantically. If the phrase pattern had a positive meaning, the negative version of the phrase; or if it had a negative meaning, the positive versions were manually created. Then, a corpus search was conducted using the phrase patterns which facilitated the semantic inferences regarding the computationally determined phenotypic features in the sentences.

2. Results

As a result of the automatic search conducted with the programmatic access and advanced search features through PubMed database using a total of 486 MeSH terms, 115.542 distinct case reports were accessed and 60.306 abstracts belonging to these were downloaded in free-text format. After the text processing procedure steps, 79.544 distinct and unique character sequences comprising all the case reports in the corpus, in other words the token roots, were determined.

After searching the HPO terms, which were put through the text processing procedures, 428.797 phenotypic features were identified on the index table. In order to identify whether the HPO term mapped by the computer is the same with the phenotypic abnormality in the sentence, a total of 336 sentences were analyzed in the sampling process, and it was found that the HPO terms defined in these sentences and the terms in the ontology list matched with a 97.82% accuracy rate. Moreover, the investigation of the methodology for capturing the phenotypic abnormalities in the case reports showed that the approach used here could catch 89.20% of the features in the document.

With the computer-based approach, 203.907 different sentences including at least one HPO term were found in all the documents comprising the documents. 6.753 distinct tokens were determined with the verb tag. These tokens were used in the corpus for 1.369.478 times; and it was found that the most frequent 125 verbs constituted 75% of the total number tokens with verb tag. After the n-gram analyses conducted for determining the common phrase patterns in the sentences containing at least one HPO term, 5.102.623 bigrams, 4.898.686 trigrams and 4.694.751 four-grams were obtained and the frequencies of these double, triple, and quadruple token groups were determined. In addition to the n-gram analysis, the 125 most frequent verb tagged token were examined and 120 among these were found to have semantic relations with the phenotypic features. The tokens in this group, which did not have any relation with the features, were “performed”, “using”, “admitted”, “did” and “follow”.

As a result of the evaluation of the verb tagged tokens together with the double, triple and quadruple token groups obtained with n-gram analyses, the phrases which were present in the case reports and had semantic relations to the phenotypic features were manually tagged. A set of rules, which facilitated the extraction of the phenotypic features from the case reports, was devised using the 190 phrases and generating the antonyms of these phrases, as shown in Table 1. As a result of scanning the rules on the whole corpus 28.853 findings, symptoms and signs were determined from all case reports, which were associated with the 486 HCAs.

Table 1. Some examples of phrase patterns which have semantic relations to the features in the case reports.

Phrase	Polarity	Opposite meaning	Phenotypic feature is observed before/after the phrase
affected with	Positive	not affected with	after
biopsy showed	Positive	biopsy showed no	after
examination showed	Positive	examination showed no	after
found to have	Positive	not found to have	after
not been reported	Negative	previously not been reported, not been reported previously, not been reported before	after/before
was seen	Positive	was not seen	before
were observed in	Positive	were not observed in	after/before
who has	Positive	who has not, who has no	after

3. Discussion

Herein, we described an automated scanning and determining methodology for the phenotypic features related to the HCA, from the case reports in the scientific literature using text processing and NLP methods. In the search procedure performed by using the HPO terms in the sentences, the rate of accurate match between a phenotypic feature determined by the search algorithm and the HPO term is thought highly convincing for such a new designed computational process.

In addition, in the calculations for determining to what extent the features in the case reports could be determined using the HPO term list, we found that approximately nine out of ten of all features in a sentence could be determined with the devised algorithmic approach. We consider the success rate obtained is within acceptable limits and promising for future works. In the manual analysis for determining the reasons for failing to capture the features in the sentences, we found that the phenotypic features in a document could not be computationally captured due to three main reasons. The first one is that the feature mentioned in the document did not have correspondence in the terms list in the HPO. For instance, the "multiple nodular liver tumors" feature in the document with PubmedID: 2596519, did not have any corresponding entry in the HPO ontology. Second, as in the "inguinal herniae" example, although the term had correspondence in the HPO ontology, "Inguinal hernia" (HP:0000023), the differences in word morphology caused mismatches. The third one is the determination errors due to the abbreviated writing of terms in the documents, such as writing "Mental retardation" (HP:0001249) as retardation.

In this study, a novel approach was implemented, as the rules were constructed using both phrase patterns and the verbs, and these rules were used for semantic inference. Rule-based approach is one of the methods used in semantic inference from natural languages [13]. However, information extraction from case reports, as in this study, is conducted with regard to the rules, and information extraction could not be possible in the existence of structures which violate these rules. We believe that the information extracted with the rules, constructed using both phrase patterns and the verbs, might be a reliable information source for a potential diagnostic decision support system for HCAs. We are currently preparing to practically apply the methodology developed in this work and building a statistical model for defining major phenotypic features that characterize each HCA as wells as creating an easily updatable and completely automated living diagnostic decision support tool which facilitates the decision making processes of potential users.

References

[1] World Health Organization [Internet]. (cited: 2013 November 21) Available from: www.who.int/en/

[2] Rosano A, Botto LD, Botting B, Mastroiacovo P. Infant mortality and congenital anomalies from 1950 to 1994: an international perspective. J Epidemiol Community Health. 2000; 54(9)660–6.

[3] Corsello G, Giuffrè M. Congenital malformations. J Matern Fetal Neonatal Med. 2012; 25 Suppl 1:25-9.

[4] Kumar P, Burton B. Congenital Malformations: Evidence-Based Evaluation and Management. McGraw Hill Professional. 2007; 408.

[5] Rauch A, Hoyer J, Guth S, Zweier C, Kraus C, Becker C, et al. Diagnostic yield of various genetic approaches in patients with unexplained developmental delay or mental retardation. Am J Med Genet A. 2006; 140(19):2063–74.

[6] Srour M, Mazer B, Shevell MI. Analysis of clinical features predicting etiologic yield in the assessment of global developmental delay. Pediatrics. 2006; 118(1):139–45.

[7] Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support?. J Biomed Inform. 2009; 42(5):760–72.

[8] National Library of Medicine. Medical Subject Headings [Internet]. (cited: 2013 November 23) Available from: www.ncbi.nlm.nih.gov/mesh

[9] Natural Language Toolkit [Internet]. (cited: 2013 November 28) Available from: http://nltk.org/

[10] Kiss T, Strunk J. Unsupervised Multilingual Sentence Boundary Detection. Comput Linguist. 2006; 32:485–525.

[11] The Lancaster Stemming Algorithm. (cited: 2013 November 28) Available from: www.comp.lancs.ac.uk/computing/research/stemming/

[12] Ratnaparkhi A. Maximum entropy model for part-of-speech tagging. In: Proceedings of the Empirical Methods in Natural Language Processing Conference. 1996; Vol 1:133–42.

[13] Mykowiecka A, Marciniak M, Kupsc A. Rule-based information extraction from patients' clinical data. J Biomed Inform. 2009; 42:923–36.