Automatic tagging of medical reports based on International Classification of Functioning, Disability and Health

Žiga Zupanec and Luka Šajn*

University of Ljubljana, Faculty of Computer and Information Science,
Tržaška 25, SI-1001 Ljubljana, Slovenia,
ziga.zupanec@gmail.com, luka.sajn@fri.uni-lj.si

Abstract. Patients coming from different countries bring their medical reports with them and sometimes doctors do not understand the content in full. World health organization provided a framework for measuring health and disability at both individual and population levels named "International Classification of Functioning, Disability and Health" (ICF). ICF is focusing on unifying framework for classifying health components of functioning and disability and thus enabling data comparison between countries. The paper presents an automated procedure for tagging medical reports with the belonging ICF classes. Our final result will present a webpage service that will allow physicians to upload documents describing the patient's status. The service will provide a list of most probable tags listed in the ICF classification. Matching is supported by methods such as parsing, eliminating stop words, lemmatization and stemming of words.

Keywords: machine learning, natural language processing, medical report annotation, classification of functioning, disability, ICF, WHO

1 Introduction

The International Classification of Functioning, Disability and Health, more commonly known as ICF, is a classification of health and health-related domains [7]. The ICF is World Health Organization's (WHO) framework for measuring health and disability at both individual and population levels. Domains are classified from body, individual and societal perspectives by means of two lists: a list of body functions and structure, and a list of domains of activity and participation. Since individual's functioning and disability occurs in a context, the ICF also includes a list of environmental factors. ICF is also available online [8].

The ICF puts the notions of 'health' and 'disability' in a new light. It acknowledges that every human being can experience a decrement in health and thereby experience some degree of disability. Disability is not something that only happens to a minority of humanity. The ICF thus 'mainstreams' the experience of disability and recognizes it as a universal human experience. By shifting the focus from cause to impact it places all

* Corresponding author: Luka Šajn, University of Ljubljana, Faculty of Computer and Information Science, Tržaška 25, SI-1001 Ljubljana, Slovenia, luka.sajn@fri.uni-lj.si

health conditions on an equal footing allowing them to be compared using a common metric - the ruler of health and disability. Furthermore, ICF takes into account the social aspects of disability and does not see the disability only as a 'medical' or 'biological' dysfunction. By including Contextual Factors, in which environmental factors are listed, ICF allows to record the impact of the environment on the person's functioning.

ICF classification in practice is utilized especially when a patient seeks medical attention in different countries and different speaking regions. The medical reports that the patient brings are far more informative to a new doctor, since there are no problems with language and institution's internal classification tags. This enables the rehabilitation process to start sooner and avoid misunderstandings.

Due to the vast amount of different domains (1420) covered by ICF classification, tagging the medical reports manually is tedious and time consuming process. Usually the physician cannot afford to spend 2 hours in average to annotate a single medical report. Physicians in University Rehabilitation Institute (RHC) in Slovenia [2] have contacted us to automate the process of annotating medical reports written in different word processing applications.

The paper presents an automated procedure for tagging the medical reports with the belonging ICF classes. Our final result will present a webpage service that will allow physicians to upload documents describing the patient's status. The service will provide a list of most probable tags listed in the ICF classification. Matching is supported through a series of methods, including:

- detecting the document type (Portable Document Format - PDF, MS Word, plain text ...)
- parsing document (extracting words from different types of documents, while keeping semantic structure)
- lemmatization of words (finding the canonical forms)
- stemming of words (reducing words to their stem)
- eliminating words not related to the subject (i.e. conjunctions, numbers, punctuation marks ...)

These methods are used both for constructing ICF classification database and for medical report processing.

2 Materials & Methods

Tagging medical reports is a very time consuming task. Our project aims to make this task faster using natural language processing tools. We are working together with physicians from RHC Soča [2] who provided us with several annotated medical reports and gave us guidance on how to improve certain medically related aspects of this project. Classification index data structure was constructed from the Slovenian ICF book (PDF)[10]. The presented methods are applicable to any language supported by WHO. Online browser for ICF classification is available in Chinese, French, English, Russian and Spanish language.

2.1 Converting PDF to plain text

This is one of the most crucial steps as the text has to be without any errors. It is important to catch different types of mistakes. Lowercase 'L' can easily be mistaken for '1' (one) during the process of conversion, line between classifications is often unclear due to PDF formatting, there are some printing bugs that mix some classifications. We have written a parser (syntactic analyzer) that can detect and correct such mistakes. Collected data was gathered in a file using CSV (comma separated value) format. This file was imported to a database in order to restore ICF structure. Possible further anomalies (wrong structure, incorrect classification) can be detected using database triggers and constraints.

2.2 Lemmatization of words

Lemmatization is finding canonical forms of words. It is used to reduce time and space complexity needed to process different forms of words while preserving semantic value. Diversity of forms for each word is language specific. Lemmatization for Germanic language families is mostly straightforward as one can get canonical form by just removing endings. Slavic languages on the other hand have a variety of different endings and often the meaning of the word is not clear without making lexical analysis first.

2.3 Eliminating words not related to the subject

Eliminating "stop words" is a technique used to remove words that have no meaning to the related subject. Stop words can be divided in two groups - generic and specific stop words. Generic stop words are words without meaning regardless of the category of document (i.e. conjunctions). Collection of such words is available in many languages. Text Mining library for statistical program R [3, 4] offers these words for Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Norwegian, Portuguese, Russian, Spanish and Swedish language.

Specific stop words are document dependent. The word "function" is not in generic stop words list but is the most common word in ICF classification and as such it has little to no semantic value since it is present in almost every classification. To get specific words, a number of repetitions is set for each word of the document and the most frequent words are added to the generic stop words list. In last step, stop words are stripped from the document.

2.4 Stemming words

Stemming is a term used in linguistics for transforming words to their root form [6]. Sometimes the root of the word is not an actual word *per se* so the term stem is used instead. We use stemming to further reduce time and space complexity but more importantly to increase semantic value and provide more accurate results when matching classification with given description in medical report [5].

2.5 Matching Classification

Methods mentioned above were used to pipe raw materials (ICF book in PDF format) through a series of processes to get proper data structure suitable for machine learning (see Figure 1). Each method was implemented using the most appropriate programming language for it's field. Parsing is done in PHP because medical reports are uploaded using web service. Parsed data is stored in a file and sent to Slovenian lemmatization service provided by Institute of Josef Stefan [1]. Received data is sent to stop words removing unit. Handling parsed data and removing stop words is programmed in Python. In order to keep the same code for different languages, important part of this stage is to use Unicode (UTF-8) aware programming language.

Data is then processed by the stemming unit. It consists of python script which passes the input flow of words to Snowball's word processing unit [9]. Snowball is string processing programming language created by Dr.Martin Porter. Different language packs provided by Snowball community can be implemented in prepared libstemmer library that is available in either Java or C. Compiled C binary files were used in this project to keep size of the package small and to be able to use our solution in systems without Java platform. Collected data is organized in indexed data structure. Each classification is presented with a set of stemmed keywords and frequency of each word in current classification. Process for identifying specific stop words can be repeated in this step to remove most common words for the whole indexed structure.

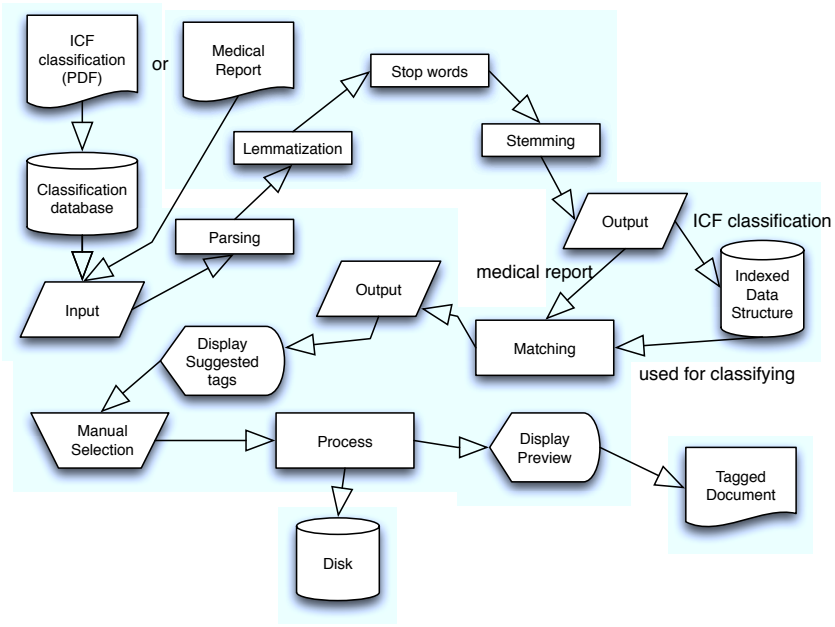

Fig. 1. Workflow of text processing for both ICF classification (PDF) and medical reports.

The above same steps apply for processing medical reports. After the last step, words collected from medical report are compared with keywords from classification index. Classifications with most relative hits are considered candidates. Each sentence in medical report is tagged with classifications from the list of the candidates where the ratio between candidates and words in that sentence exceeds certain threshold. Threshold can be adjusted in real time anytime during tagging process.

3 Results

Currently the correct classifications for each diagnosis in a medical report are within top 5 suggested tags for a domain (number of displayed tags is adjustable) which is, considered the limited amount of data available for this problem, supportive enough to further develop this project. Interface offers quick and easy navigation between classifications for highlighted part of the medical report (see Figure 2).

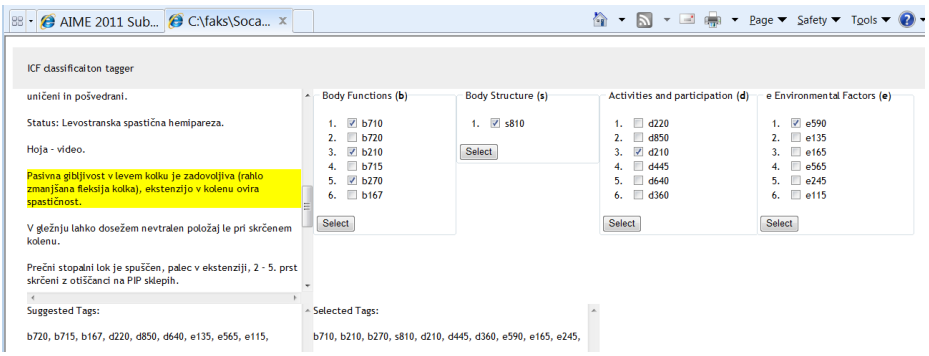

Fig. 2. User interface available through our web service. (Imported medical report on the left, suggested and selected tags for complete medical report below. Tags suggested for current sentence are displayed top right.)

4 Discussion & Further work

We have presented the preliminary results of automated tagging of medical reports based on International Classification of Functioning, Disability and Health. The results are encouraging and there is a good chance they will be used in everyday medical practice. The system will be presented to physicians dealing with patients that come from abroad in hope to reduce time needed to annotate medical reports with ICF classifications.

We plan to use machine learning algorithms on data collected from users of this service. This way incorrect results decrease in number over time. Tags chosen by users

are stored in a separate database. Once there is enough data a weighted voting algorithm constructed on decision tree, K-nearest neighbors and naïve bayes will be able to offer more accurate and smaller set of suggested tags next time, a similar case occurs.

Increased classification semantic value can also be achieved with thesaurus (group of words referencing same meaning) that can be linked to hierarchical structure of words to get better results. On more sophisticated data structure more advanced learning methods can be used - such as SVM (support vector machine) in contrast with current highest frequency method. We plan to implement this system on other *ICF* supported languages as well.

Acknowledgements

We thank prof.dr. Helena Burger, M. D., University Rehabilitation Institute, Ljubljana, for comments and cooperation. This work was supported by the Slovenian Ministry of Higher Education, Science, and Technology.

References

[1] Jos totale text analyser. http://nl.ijs.si/jos/analyse/.

[2] University rehabilitation institute, republic of Slovenia. http://www.ir-rs.si/en/.

[3] P. Dalgaard. *Introductory Statistics with R (Statistics and Computing)*. Springer, 2nd edition, 2008.

[4] I. Feinerer. Text mining package, 9 2010. http://cran.r-project.org/web/packages/tm/tm.pdf.

[5] I. Kononenko and M. Robnik. *Inteligentni sistemi*. Založba FE in FRI, 2010. In Slovene.

[6] J.B. Lovins. *Development of a stemming algorithm. Mechanical Translation and Computational Linguistics*. Number 11:22-31. 1968.

[7] World Health Organization. International classification of functioning, disability and health (ICF), 2007. http://www.who.int/classifications/icf/en/.

[8] World Health Organization. International statistical classification of diseases and related health problems (10th revision), 2007. http://apps.who.int/classifications/apps/icd/icd10online/.

[9] M. Porter. Snowball. http://snowball.tartarus.org/.

[10] J. Remškar, M. Seljak, and R. Cugelj. *Mednarodna klasifikacija funkcioniranja, zmanjšane zmožnosti in zdravja (ICF Translation)*. Ministry of health, Republic of Slovenia, 2008. In Slovene.