Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200180

On the Construction of Multilingual Corpora for Clinical Text Mining

Fabián VILLENA [a,b], Urs EISENMANN [a], Petra KNAUP [a], Jocelyn DUNSTAN [b,c], and Matthias GANZINGER [a,1]

[a] *Institute of Medical Biometry and Informatics, Heidelberg University, Germany*
[b] *Center of Medical Informatics and Telemedicine, University of Chile, Chile*
[c] *Center for Mathematical Modeling, University of Chile, Chile*

Abstract. The amount of digital data derived from healthcare processes have increased tremendously in the last years. This applies especially to unstructured data, which are often hard to analyze due to the lack of available tools to process and extract information. Natural language processing is often used in medicine, but the majority of tools used by researchers are developed primarily for the English language. For developing and testing natural language processing methods, it is important to have a suitable corpus, specific to the medical domain that covers the intended target language. To improve the potential of natural language processing research, we developed tools to derive language specific medical corpora from pub-licly available text sources. In order to extract medicine-specific unstructured text data, openly available pub-lications from biomedical journals were used in a four-step process: (1) medical journal databases were scraped to download the articles, (2) the articles were parsed and consolidated into a single repository, (3) the content of the repository was de-scribed, and (4) the text data and the codes were released. In total, 93 969 articles were retrieved, with a word count of 83 868 501 in three different languages (German, English, and Spanish) from two medical journal databases. Our results show that unstructured text data extraction from openly available medical journal databases for the construction of unified corpora of medical text data can be achieved through web scraping techniques.

Keywords. Natural Language Processing, Data Mining, Information Storage and Retrieval, Medical Informatics, Linguistics

1. Introduction

The amount of digital data derived from healthcare processes have increased tremendously in the last years [11]. Data available from healthcare sources can be roughly divided into two types: (1) structured data, namely information already organized into a known data model, for instance, results from psychological surveys or results of laboratory tests; and (2) unstructured data in the form of free-text data, medical images, or physiological signals. Examples for such free-texts are anamnesis records or discharge

[1] Corresponding Author: Institute of Medical Biometry and Informatics, Heidelberg University, Im Neuenheimer Feld 130.3, 69120 Heidelberg, Germany; E-mail: matthias.ganzinger@med.uni-heidelberg.de.

notes and they are commonly written by healthcare professionals [7]. Dalianis et al. estimated that over 40% of healthcare data are unstructured [3]. It is often hard to use such unstructured texts for research and other analyses due to the lack of available tools to process and extract information. These tools must be tailored specifically for the medical field, and moreover, tailored for a specific language [3].

Nowadays, the way to analyze texts is using large corpora from which computers can learn the meaning of a word from its context, replacing former rule-based and feature-selection-based approaches [5]. However, availability of medical records for such research purposes is usually restricted due to privacy issues. Consequently, collecting biomedical texts extracted from journals have been proposed as a possible solution for the lack of resources [1, 12].

Natural language processing (NLP) is a sub-field of computer science, concerned with interactions between computers and human languages [4]. Some of the most important NLP applications in medicine are (1) detection of patterns in electronic health records (EHR) (e.g. detection of healthcare associated infections, adverse drug events or cancer symptoms); (2) text summarizing and translation of EHR; and (3) automatic codification [2]. While there are successful applications of NLP in the medical field [6, 13, 9], the majority of tools used by researchers are developed for the English language [8]. Therefore, creating tools and training corpora for languages different to English are valuable contributions to empower NLP of clinical sources in non-English speaking countries. To improve the potential of NLP research, especially for non-English languages, we developed tools to derive language specific medical corpora from publicly available text sources. Specifically, we focus on Spanish and German texts.

2. Methods

In order to extract medicine-specific unstructured text data, openly available papers from biomedical journals were used. The general process was divided into (1) web scraping of the database to download each of the articles contained in the website, (2) parsing of the raw downloaded data to consolidate it into a local standardized repository, (3) descriptive analysis of the content, and (4) the release of the data to the scientific community. A general overview of the workflow is summarized in Figure 1. The entire pipeline was developed in the Python[2] programming language.

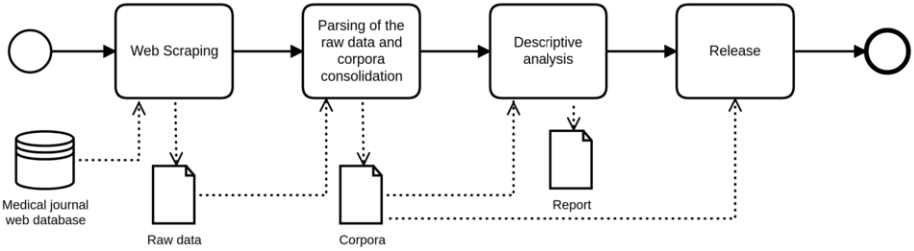
Web Scraping
Parsing of the raw data and corpora consolidation
Descriptive analysis
Release
Medical journal web database
Raw data
Corpora
Report

Figure 1. Overview of our proposed workflow to generate corpora from medical journal databases.

[2] Python - https://www.python.org/

2.1. Web scraping

Web scraping is a technique to extract automatically data from websites. The package Scrapy[3] was used to develop a script, which navigates trough the database and downloads every article in XML format.

The recursive algorithm first searched for a medical journal in the index of the database, then inside the journal looked up for the volumes available, and finally the script searched for all the articles inside that volume and downloaded the XML file associated with the article.

2.2. Parsing of the raw data and corpora consolidation

For parsing raw XML files, the package BeautifulSoup[4] was used. Additionally, we extracted the following attributes: title, language, abstract, day of release and journal / meeting name.

The parsed articles were then consolidated into separate corpora based on the language in which the article was written. Each corpus was word tokenized using the NLP framework Spacy[5].

2.3. Descriptive analysis

For the exploration of the corpora, each corpus was summarized by (1) the number of articles that is composed of, (2) number of word tokens, and (3) vocabulary size counted as the number of different words. To compare each of the corpora, we first deleted stopwords from the text using the list provided in Spacy and then frequency of appearance of each word in each corpus was calculated.

2.4. Release

The corpora are made available to the research community as a package, which contains each article labeled with the language of the content. The code used to retrieve the corpora is available on a git repository [6] and the corpora in a hosting platform [7].

3. Results

In total, 95 737 articles were retrieved from the German Medical Science Database [8] (GMS) and the Chilean Scientific Electronic Library Online [9] (SciELO). After parsing the raw data, 1 762 empty articles were discarded. A division by language was made, finding that 63% of the articles were written in German, 24% in English, 13% in Spanish and <1% in French. The metrics of each corpus are described in Table 1.

A word frequency analysis was made for each corpus and ten of the most frequent words are summarized in Table 2.

[3] Scrapy: A Fast and Powerful Scraping and Web Crawling Framework - https://scrapy.org

[4] Beautiful Soup - https://www.crummy.com/software/BeautifulSoup

[5] Spacy: Industrial-Strength Natural Language Processing - https://spacy.io

[6] https://github.com/fvillena/multilingual-medical-nlp

[7] https://zenodo.org/record/3463379.XY4RsUEzaV4

[8] https://www.egms.de

[9] https://scielo.conicyt.cl

Table 1. Summary statistics of the corpora

	corpus		
Metric	German	English	Spanish
Articles count	59 539	22 372	12 058
Number of word tokens	20 437 502	12 093 145	51 337 854
Vocabulary size	497 256	144 550	374 877

Table 2. Most frequent words in the corpora

	corpus		
Rank	German	English	Spanish
1	patienten (*patients*)	patients	pacientes (*patients*)
2	ergebnisse (*results*)	results	estudio (*study*)
2	methoden (*methods*)	study	años (*years*)
3	schlussfolgerung (*conclusion*)	treatment	casos (*cases*)
4	einleitung (*introduction*)	methods	tratamiento (*treatment*)
5	fragestellung (*research question*)	clinical	diagnóstico (*diagnostic*)
6	studie (*study*)	medical	riesgo (*risk*)
7	material (*material*)	group	figura (*figure*)
8	hintergrund (*background*)	used	enfermedad (*disease*)
9	therapie (*therapy*)	years	resultados (*results*)
10	durchgeführt (*carried out*)	time	forma (*form*)

4. Discussion

Current research on NLP in medicine mainly focuses in the English language; nevertheless there is an increasing effort in closing the gap of the development of NLP in languages beyond English [8]. Our work is a building block of a framework to retrieve medical unstructured text by the construction of biomedical corpora to support clinical text mining. Similar work was published by Soares et al. [10], but only SciELO database was scraped.

The toolset we describe in this manuscript has shown to be suitable for establishing generic medical corpora for multiple languages. To validate our approach we have queried a public Chilean and German database for medicine related publications. While we were focusing on German and Spanish texts, we retrieved a considerable amount of English texts because many the German site publishes in English as well.

Analysis has shown that the most frequent words conform across the corpora only partially (cf. Table 2). This result is to some extent unexpected, since scientific texts, as they are provided by the the underlying collections, typically share a very similar structure (IMRaD – Introduction, Methods, Results, and Discussion). While the reason for this finding needs further investigation, it might be due to differing focus areas in the collections. This result is supported by the figures shown in Table 1. Obviously, the GMS database contains shorter texts, possibly due to a large amount of conference abstracts published by them.

With a large amount of medical text data, the simplest model we can develop is a probability distribution, where we can predict upcoming words given a sentence. These models are essential in tasks where we have noisy and ambiguous inputs [4], as they occur, for instance, in spelling and syntax correction of clinical texts or speech recognition tasks for electronic health records (EHR).

5. Conclusion

The unstructured text data extraction from openly available medical journal databases for the construction of unified corpora of medical text data can be achieved through web scraping techniques.

References

[1] Zhiwei Chen et al. “Evaluating semantic relations in neural word embeddings with biomedical and general domain knowledge bases”. en. In: *BMC Medical Informat-ics and Decision Making* 18.S2 (July 2018). (Visited on 07/19/2019).

[2] Hercules Dalianis. *Clinical Text Mining*. en. Cham: Springer International Publish-ing, 2018. (Visited on 09/23/2019).

[3] Hercules Dalianis, Martin Hassel, and Sumithra Velupillai. “The Stockholm EPR Corpus – Characteristics and Some Initial Findings”. en. In: *Management Research* (2009), p. 7.

[4] Dan Jurafsky and James H. Martin. *Speech and Language Processing*. 3rd ed. draft. 2018.

[5] Yoav Goldberg. “A Primer on Neural Network Models for Natural Language Pro-cessing”. In: *CoRR* abs/1510.00726 (2015).

[6] Kory Kreimeyer et al. “Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review”. eng. In: *Journal of Biomedical Informatics* 73 (2017), pp. 14–29.

[7] Clement J. McDonald. “The Barriers to Electronic Medical Record Systems and How to Overcome Them”. In: *Journal of the American Medical Informatics Asso-ciation* 4.3 (May 1997), pp. 213–221.

[8] Aurélie Névéol et al. “Clinical Natural Language Processing in languages other than English: opportunities and challenges”. eng. In: *Journal of Biomedical Se-mantics* 9.1 (2018), p. 12.

[9] Ewoud Pons et al. “Natural Language Processing in Radiology: A Systematic Re-view”. eng. In: *Radiology* 279.2 (May 2016), pp. 329–343.

[10] Felipe Soares et al. “Medical Word Embeddings for Spanish: Development and Evaluation”. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Lin-guistics, June 2019, pp. 124–133.

[11] S. S. -L. Tan, G. Gao, and S. Koch. “Big Data and Analytics in Healthcare”. en. In: *Methods of Information in Medicine* 54.06 (2015), pp. 546–547. (Visited on 09/23/2019).

[12] Yanshan Wang et al. “A comparison of word embeddings for the biomedical natu-ral language processing”. In: *Journal of Biomedical Informatics* 87 (2018), pp. 12–20.

[13] Wen-Wai Yim et al. “Natural Language Processing in Oncology: A Review”. eng. In: *JAMA oncology* 2.6 (June 2016), pp. 797–804.