MeSHx-Notes: Web-System for Clinical Notes Information Extraction

Henrique D. P. dos Santos, Rafael O. Nunes, João E. Soares, and Renata Vieira

School of Technology at Pontifical Catholic University of Rio Grande do Sul
Email: {henrique.santos.003, rafael.oleques,
joao.etchichury}@acad.pucrs.br, renata.vieira@pucrs.br

Abstract. We present MeSHx-Notes, MeSH eXtended for clinical notes, a multi-language web system based on the Django framework to present information selected in clinical notes. MeSHx-Notes extends Medical Subject Headings (MeSH) terms with Word Embeddings with similar semantic/syntactic words. Since MeSH is available for 15 languages, MeSHx-Notes is easily extendable by replacing the MeSH thesaurus with the target language. In this demo, we show examples with Portuguese and English.

Keywords: Multi-language, Web System, Clinical Notes, Information Extraction, Word Embeddings, MeSH

1 Introduction

Electronic Health Records (EHR) play an important role in hospital environments, bringing many benefits in terms of patient safety, effectiveness and efficiency of care, and patient satisfaction [1]. Records of health care practices in hospitals generate a rich and large amount of patient information and an intrinsic relation between symptoms, diseases, drug interaction, and diagnoses that may be used for many purposes [4].

This study aims to help healthcare professionals concerning the understanding of what has been informed by a clinical note. This is possible through the use of Natural Language Processing (NLP), combined with the MeSH dictionary. The system consists of a web application that exhibits the meaning and the related words of the main terms used in clinical notes, thus enhancing the understanding of what is reported.

Other systems, such as cTAKES [3], rely on several UMLS sources for English to provide several information from clinical notes. We focus on developing a user-friendly and easy-handling UI through a web application, portable for languages other than English, using a language-specific MeSH thesaurus.

In this context, we present an easy-to-use system that provides users with extra knowledge of the information given in clinical notes, which can be used by anyone with access to the internet.

2 System Description

The system consists of a web application that receives clinical notes, identifies the main terms, and then, returns their definition, similar words and a link to the MeSH dictionary. Its development is based on Python, Django, Pandas, Bootstrap, JQuery, Word Embeddings, XPath, and the MeSH thesaurus.

2.1 Data Source

Three resources are used to develop MeSHx-Notes, as described below. In addition, we describe the process to generate the word embedding vectors.

Electronic Health Records The Portuguese dataset was obtained from Hospital Nossa Senhora da Conceição (HNSC). The English dataset was obtained from i2b2 Challenge [5] from 2008 to 2012. It is a set of nine datasets from several shared tasks promoted by Informatics for Integrating Biology and the Bedside (i2b2).

Medical Subject Headings (MeSH) MeSH is a "controlled vocabulary" Metathesaurus, developed by the National Library of Medicine (NLM). As of 2013, MeSH has 54,935 entries where each entry has a unique tree number and consists of 26,851 main headings and 213,000 entry terms that increase the power of classification of medical documents.

Word Embeddings Word vectors are a way of mapping words in a numerical space. A latent syntactic/semantic vector for each word is induced from a large unlabeled corpus. The Portuguese and English model for the word embeddings was trained with Word2Vec [2]. For the Portuguese version, we used 21 million sentences from HNSC's medical records, trained with 50 dimensions per word and 100 minimum word count. This training resulted in 63 thousand word vectors used as a semantic model in the neural network below. For the English version, we used 171 thousand sentences from the i2b2 challenge dataset, trained with 50 dimensions and 10 minimum word count, resulting in 17 thousand word vectors.

2.2 Back-end

First, the MeSH dictionary is generated, using previously saved data in an XML file, containing ID, name, scope, terms, and qualifier. The dictionary is enriched by identifying similar words using Word Embeddings, so that we provide a greater range of terms, which are stored in the terms field. We consider higher similarity degrees to identify those words. An initial evaluation made by the authors resulted in 67% accuracy.

Table 1. Enrichment of MeSH Terms

Heading	Original Terms	New Similar Terms
Abdomem	abdomem, belly	abd, abdome...
Celecoxib	celecoxib, celebrex	norvasc, losartan...
Abscess	abcesso, absceso	abscess, abscesses...

In Table 1 we show some concepts that are commonly used in clinical notes. Each concept has a heading in the MeSh dictionary, its terms, and the new identified terms. For example, the heading "Abscesso" had "abcesso" and "absceso" as the original terms, and "abscess" and "abscesses" were added as new terms.

After that step, we read the clinical notes, using Pandas, in the web application, using Django as the development framework. Each word found in the dictionary is captured and the lists of original and new similar words are stored.

2.3 Front-end

When a clinical note is shown to the user, the words from the (enriched) dictionary are highlighted. These words are shown in different colors, according to the classes: medication, diagnosis, procedure or anatomy. We provide users with not only that, but also a navigation bar to go through all the desired clinical notes. This page is developed using JQuery and Bootstrap.

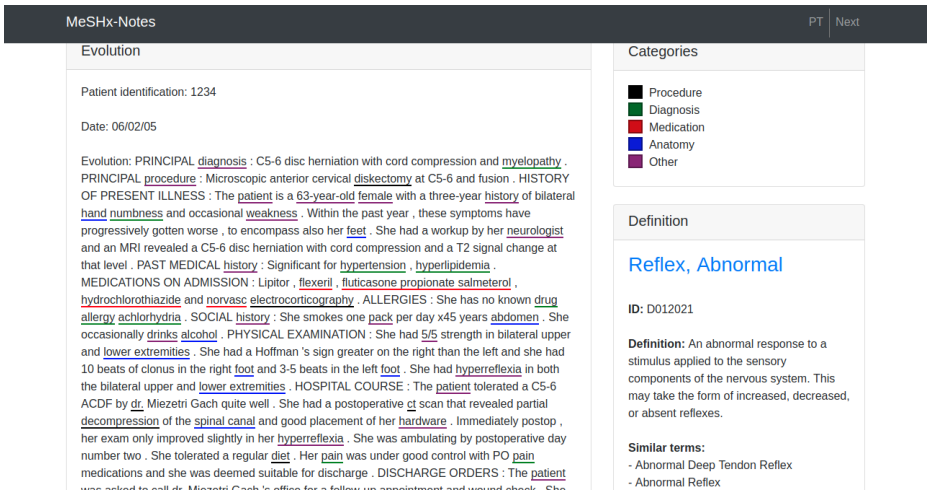

Fig. 1. Example of clinical notes with highlighted terms, color legend for each category, and MeSH descriptor.

3 Demo

MeSHx-Notes is presented for Portuguese and English samples. We use Word Embeddings for the dictionary expansion, in Portuguese and English.

In the web page, buttons are provided to navigate between clinical notes and to change the language. Besides, the clinical note description is given with data about the patient record and its modification date with a concomitant section of legends that are related to the classification of the terms. Nonetheless, identified words are underlined according to their classification, so that, when clicked, they show their technical name, ID, description, terms with similar meanings, and a link to the MeSH description website.

4 Conclusion and Further Work

MeSHx-Notes is able to provide, both for health professional and for non-specialists, a simple tool that enables a better understanding of the terms used in clinical notes in a clear, concise, accessible way. The source code is available on the project's Github page[1], and the demo is found on the group's website[2]. As further work, we plan to use bigram and trigram embeddings to find similar multi-word expressions.

Acknowledgments This work was partially supported by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) Foundation (Brazil), PUCRS (Pontifical Catholic University of Rio Grande do Sul) and UFRGS (Federal University of Rio Grande do Sul).

References

1. Buntin, M.B., Burke, M.F., Hoaglin, M.C., Blumenthal, D.: The benefits of health information technology: a review of the recent literature shows predominantly positive results. Health affairs **30**(3), 464–471 (2011)
2. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
3. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. Journal of the American Medical Informatics Association **17**(5), 507–513 (2010)
4. Silveira, M., Nogueira, V.B., Rodrigues, I.: Ferramentas e tecnologias para a integração e extração de informação hospitalar. INF - Artigos em Livros de Actas/Proceedings (2015)
5. Uzuner, Ö., South, B.R., Shen, S., DuVall, S.L.: 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association **18**(5), 552–556 (2011)

[1] https://github.com/nlp-pucrs/meshx-notes
[2] http://grupopln.inf.pucrs.br/meshx