MeSHx-Notes: Web-System for Clinical Notes

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

School of Technology at Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, Brazil
{rafael.oleques,joao.etchichury,henrique.santos.003}@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 selected terms in clinical notes. MeSHx-Notes extends Medical Subject Headings (MeSH) terms with Word Embeddings with similar words. Since MeSH is available in 15 languages, MeSHx-Notes is easily extendable by replacing the MeSH thesaurus with the target language (plus the generation of the corresponding WE for the new language). Our version deals 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, satisfaction, and effectiveness/efficiency of care [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 interactions, and diagnoses that may be used for many purposes [2,7,8]. Clinical notes, such as discharge summaries, have a semi- or unstructured format. These documents contain information about diseases, treatments, drugs, etc. Extracting meaningful information from them becomes challenging due to their narrative format [5].

This work aims to help healthcare professionals concerning the understanding of what is informed in clinical notes. This is possible through the use of Natural Language Processing (NLP), combined with the MeSH dictionary[1]. We developed a web application that exhibits the meaning and the related words for terms of a set of categories used in clinical notes, thus enhancing the understanding of what is reported.

[1] https://www.ncbi.nlm.nih.gov/mesh.

F. Koch et al. (Eds.): AIH 2018, LNAI 11326, pp. 5–12, 2019.
https://doi.org/10.1007/978-3-030-12738-1_1

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.

The rest of this paper is organized as follows: Sect. 2 presents previous works on information extraction through clinical notes. In Sect. 3 we explain the concepts related to the term expansion. Section 4 describes the concepts used in the MeSHx-Notes system, followed by the results in Sect. 5. Finally, in Sect. 6 we summarize our contributions and present further research directions.

2 Related Work

One problem in clinical notes is that registers are not always is accordance with the standard language, therefore the identification of the right dictionary entry is challenging [12]. Clinical notes usually contain abbreviations, misspelled words, and word concatenations. To overcome such problems, we propose the use of Word Embedding models (generated on the basis of clinical notes) to spot terms that are similar to the dictionary entries.

The use of pre-established ontologies for the classification of medical documents has also become a trend, since such structures already bring a semantic knowledge of the data and help in the organization of texts [10]. The US National Library of Medicine has developed an ontology for medical systems to communicate, called the Unified Medical Language System (UMLS). The same project includes the medical subject ontology, known as MeSH, which relates the medical vocabularies of diseases, symptoms, organs, etc.

While other systems, such as cTAKES [9], rely on several UMLS sources for English to provide several information from clinical notes, we focus on developing a user-friendly and easy-to-handle web interface, portable for languages other than English, using a language-specific MeSH thesaurus.

Several efforts have been reported in the area of clinical text mining to bridge the gap between unstructured clinical notes and structured data representation, including tools such as MetaMap and KnowledgeMap, which have been developed to automatically annotate medical concepts in free text, along with systems to identify the patient's disease status, medication information, etc. [3].

3 MeSH Dictionary Expansion

3.1 Medical Subject Headings (MeSH)

The MeSH dictionary (Medical Subject Headings) (see footnote 1) is the National Library of Medicine controlled vocabulary thesaurus used to index articles for PubMed.

Started in 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. MeSH is available in 15

languages: English, Croatian, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Portuguese, Russian, Spanish, and Swedish.

In MeSH, each heading has information about a term - e.g., Unique ID, Scope Note, qualifiers, and Entry Terms. The unique ID refers to the working term, therefore, homonyms "perna" (leg) and "perna" (organism) have distinct IDs. The Scope Note refers to the term's meaning. Qualifiers divide terms into categories. Entry Terms are synonymous or alternative ways to write a term, for instance "ache" is an Entry Term of "pain".

MeSH has 81 qualifiers. In our study we selected five categories: pharmacology, anatomy, methods, diagnosis, and others. These are the most frequent qualifiers, according to our analysis. The category 'others' includes the least frequent 77 categories and terms that do not have a qualifier. We use the qualifiers as a way to classify not only terms previously found in MeSH but also the new terms with which our dictionary was enriched, as explained below.

3.2 Electronic Health Records

The Portuguese dataset was obtained from Hospital Nossa Senhora da Conceição (HNSC). We used a large cohort extracted from the administrative hospitalization database from this Hospital. HNSC is part of the Brazilian public healthcare system and provides tertiary care. The data comprises 1.5 million clinical notes from 48.9 thousand hospitalization records annotated with the Charlson comorbidity index between January 2012 and December 2017.

Ethical approval to use the hospital dataset in this research was granted by the Research Ethics Committee of Conceição Hospital Group under the number 71571717.7.0000.5530.

The English dataset was obtained from i2b2 Challenge [11] 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). In the 2012 i2b2 Challenge, 310 discharge summaries were annotated for temporal information. The challenge focused specifically on the identification of clinically relevant events in patient records and on the relative ordering of the events with respect to each other and with respect to time expressions included in the records.

3.3 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 [4]. For the Portuguese version, we used 21 million sentences from HNSC's medical records, trained with 50 dimensions per word and a minimum word count of 100 [6]. This training resulted in 73 thousand word vectors. For the English version, we used 171 thousand sentences from the i2b2 challenge dataset, trained with 50 dimensions and a minimum word count of 10, resulting in 17 thousand word vectors.

The original dictionary was expanded using Word Embeddings. The expansion process was made by analyzing the similarity of the MeSH's Entry Terms of each input with those from the Word Embeddings. The terms which were considered similar were linked to the specific Unique ID of the enriched dictionary and added to a reverse index.

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...

Table 1 shows some examples of heading terms in the MeSh dictionary, their alternative terms, and the corresponding new identified terms. For example, the heading "Abscesso" had "abcesso" and "absceso" as alternative MeSH terms, and "abscess" and "abscesses" were added as new terms found through the WE model. Originally the dictionary had 80,973 terms; with the expansion there was an enrichment of 40,588 new terms. The enrichment brings new terms due to abbreviations, orthography errors, and word concatenations. Table 2 shows examples of such cases.

Table 2. Enriched dictionary terms

MeSH terms	Expanded terms
Tomography	tomo, tc, tomographyexanms
General surgery	srg, surrgery, sugery
Enoxaparin	enoxa, enoxeparin
Fever	chills, hyperthermia
Behavior	behav, bhv

4 MeSHx-Notes: 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.

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

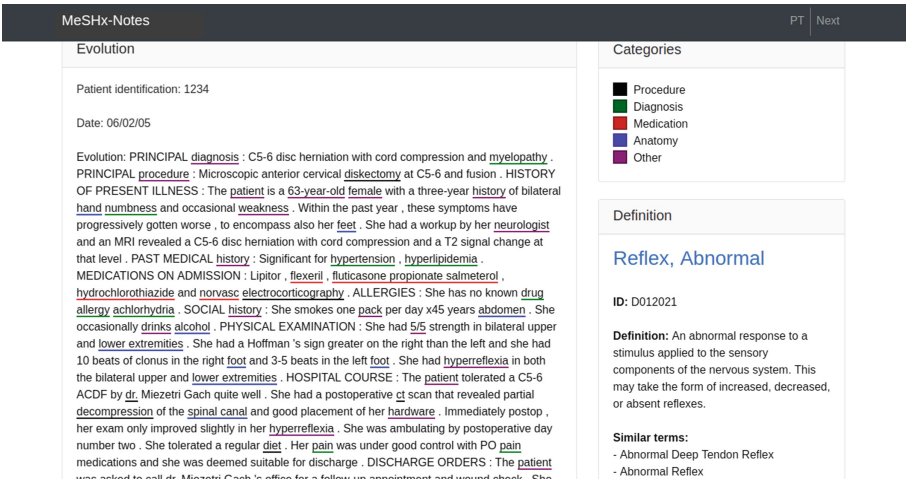

Fig. 1. Example of clinical notes with highlighted terms, color legend for each category, and MeSH descriptor.

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. Some special features were implemented, such as highlighting of terms of a specific category.

Having in mind the processing time for information to be presented to the user, we search for terms using a reverse index, previously generated with the terms and their IDs. MeSHx-Notes was built for Portuguese and English.

This system shows the definition of medical terms, helping in their understanding. It can be applied in various situations: some applications improve learning of nursing, medical or other health-related students, in addition to aiding multidisciplinary research groups in which not all members have technical medical knowledge. Besides, the system works not only with clinical notes, but also with any texts related to health, for example, journalistic and academic texts.

4.1 Back-End

First, the extended MeSH dictionary is generated, using previously saved data in an XML file, containing ID, name, scope, terms, and qualifier. The dictionary is enriched to provide a greater range of terms, which are stored in the terms field. We consider higher similarity degrees to identify those words.

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.

4.2 Front-End

When a clinical note is shown to the user, terms from the (enriched) dictionary are highlighted. These words are shown in different colors, according to the following classes: medication, diagnosis, procedure or anatomy. It is possible to select specific classes, providing better information visualization (e.g., only medication to look for what the patient is using in a treatment). This development used JQuery and Bootstrap.

5 Term Expansion Evaluation

For the expansion of alternative terms, there was a manual evaluation of the enriched terms from 42 clinical notes, whose expanded terms were annotated as "correct" or not. Based on this evaluation, a gold standard was generated with all the appended terms and their Unique IDs, to which each term should be properly related. As a result, we had 651 examples. Based on that, we tested several thresholds, to estimate the best threshold for each category.

Through the gold standard and different thresholds, we obtained values with the lowest failure rates using an algorithm. This algorithm analyzed the accuracy of each threshold, as shown in Table 3. The analysis started with a threshold of 0.80 to 0.99–1.00 returns the term itself. Thus, we accomplished a 58% accuracy rate assessing 691 terms contained in our gold standard. The obtained results are based on tests with real clinical notes.

This accuracy value of 58% is yet to be improved. At this stage, the user still has to judge for themselves the alternatives presented by the system. However, given the difficulty of the task, we consider that this initial result is promising and there are ways in which it could be further improved.

Table 3. Best thresholds per qualifier

Qualifier	Thresholds	Correct	Total	Accuracy
Methods	0.89	80	116	71%
Diagnosis	0.93	81	164	49%
Pharmacology	0.96	137	263	65%
Anatomy	0.95	44	83	54%
Others	0.94	97	188	51%

True positives (correctly enriched terms) are presented in Table 4. There are terms with lexical similarity (e.g., rehab and rehabilitation), but also terms that are semantically similar but lexically distant—e.g., amlodipine and norvasc.

Table 4. True positive terms

MeSH term	True positive
Arteries	Vessels
Angioplasty	Stenting
Rehabilitation	Rehab
Amlodipine	Norvasc

Table 5 presents examples of false positives, terms that were incorrectly identified as similar through Word Embeddings. These terms, on the established threshold, had a degree of similarity. New NLP techniques, based on word-sense disambiguation, are being studied to try to solve these problems.

Table 5. False positive terms

MeSH term	False positive
Thoracotomy	Parietal
Bicuspid	Ulcerative colitis
Ocular vision	Weakness

6 Conclusion and Further Work

MeSHx-Notes aims to provide, both for health professionals 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[2]. A web demo is also available[3]. As further work, we plan to use bigram and trigram embeddings to find similar multi-word expressions.

Aiming to improve the system and the accuracy rate (58%), we will use new disambiguation techniques and similarity analysis, besides the evaluation of enriched terms made by nurses. MeSH ambiguity is a problem to be studied in the continuation of this work. User pilot studies are an important phase to be pursued to test whether the system enhances the readability of medical notes, after we achieve better accuracy rates.

Another goal is to perform classification tasks in clinical notes written in Portuguese using MeSH codes. Then, we will validate the learned model in clinical notes in English using MeSH terms for those codes. These experiments intend to evaluate the cross-language ability of MeSH for classification tasks in different languages. Furthermore, our purpose is to use a new database with the MeSH definitions in Portuguese [13]. This way, we will be able to better identify ambiguous terms through their definition elements in clinical notes.

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

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 Aff. **30**(3), 464–471 (2011)
2. Jensen, P.B., Jensen, L.J., Brunak, S.: Mining electronic health records: towards better research applications and clinical care. Nat. Rev. Genet. **13**(6), 395 (2012)
3. Kovačević, A., Dehghan, A., Filannino, M., Keane, J.A., Nenadic, G.: Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. J. Am. Med. Inform. Assoc. **20**(5), 859–866 (2013)
4. 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)
5. Reátegui, R., Ratté, S.: Comparison of metamap and ctakes for entity extraction in clinical notes. BMC Med. Inform. Decis. Mak. **18**(3), 74 (2018)
6. dos Santos, H.D.P., Nunes, R.O., Soares, J.E., Vieira, R.: Meshx-notes: web system for clinical notes information extraction. In: AIH Joint Workshop on Artificial Intelligence for Health, p. 1. Stockholm, Sweden, July 2018
7. dos Santos, H.D.P., Ulbrich, A.H.D.P.S., Woloszyn, V., Vieira, R.: DDC-outlier: preventing medication errors using unsupervised learning. IEEE J. Biomed. Health Inform., 1 (2018). https://doi.org/10.1109/JBHI.2018.2828028
8. dos Santos, H.D.P., Ulbrich, A.H.D.P.S., Woloszyn, V., Vieira, R.: An initial investigation of Charlson comorbidity index regression based on clinical notes. In: 31st IEEE CBMS International Symposium on Computer-Based Medical Systems (CBMS), pp. 6–11. IEEE, Karlstad, June 2018. https://doi.org/10.1109/CBMS.2018.00009
9. Savova, G.K., et al.: Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J. Am. Med. Inform. Assoc. **17**(5), 507–513 (2010)
10. Trieschnigg, D., Pezik, P., Lee, V., De Jong, F., Kraaij, W., Rebholz-Schuhmann, D.: Mesh up: effective mesh text classification for improved document retrieval. Bioinformatics **25**(11), 1412–1418 (2009)
11. Uzuner, Ö., South, B.R., Shen, S., DuVall, S.L.: 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J. Am. Med. Inform. Assoc. **18**(5), 552–556 (2011)
12. Wang, Y., et al.: Clinical information extraction applications: a literature review. J. Biomed. Inform. **77**, 34 – 49 (2018). https://doi.org/10.1016/j.jbi.2017.11.011. http://www.sciencedirect.com/science/article/pii/S1532046417302563
13. Who, B.P.: Health sciences descriptors: DECS (2017). http://decs.bvsalud.org/I/homepagei.htm. Accessed 30 Sept 2018