Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 100 (2016) 55 – 61

Procedia Computer Science

Conference on ENTERprise Information Systems / International Conference on Project MANagement / Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2016, October 5-7, 2016

Named entity recognition over electronic health records through a combined dictionary-based approach

Alexandra Pomares Quimbaya[a*], Alejandro Sierra Múnera[b], Rafael Andrés González Rivera[a], Julián Camilo Daza Rodríguez[b], Oscar Mauricio Muñoz Velandia[a,b], Angel Alberto Garcia Peña[a,b] , Cyril Labbé[c]

[a]*Pontificia Universidad Javeriana, Cra. 7 #40-62, Bogotá 110231, Colombia*
[b]*Hospital Universitario San Ignacio, Cra. 7 #40-62, Bogotá 110231, Colombia*
[c]*Laboratoire d'Informatique de Grenoble équipe SIGMA, 700 avenue centrale,Saint-Martin-d'Hères 38400, France*

Abstract

In health care information systems, electronic health records are an important part of the knowledge concerning individual health histories. Extracting valuable knowledge from these records represents a challenging task because they are composed of data of different kind: images, test results, narrative texts that include both highly codified and a variety of notes which are diverse in language and detail, as well as ad hoc terminology, including acronyms and jargon, far from being highly codified. This paper proposes a combined approach for the recognition of named entities in such narrative texts. This approach is a composition of three different methods. The possible combinations are evaluated and the resulting composition shows an improvement of the recall and a limited impact on precision for the named entity recognition process.

Peer-review under responsibility of the organizing committee of CENTERIS 2016

Keywords: Named Entity Recognition; Electronic Health Records; Text Mining

* Corresponding author. Tel.: + 57-1-3208320 ext 5291; fax: + 57-1-3208320 ext 5338
E-mail address: pomares@javeriana.edu.co

Peer-review under responsibility of the organizing committee of CENTERIS 2016
doi:10.1016/j.procs.2016.09.123

1. Introduction

Electronic health records (EHR) constitute an important resource not just for tracing single patient histories, but for population studies with clinical or administrative purposes. The nature of EHR, however, presents multiple challenges for doing so. Essentially, what we have is a typical knowledge extraction task where a combination of structured and unstructured data must be processed, including medically codified classifications, images, test results and narrative text. This paper will focus on the challenge of extracting information from narrative text contained within EHR through combined named entity recognition.

A large body of work has presented an array of methods to deal with biomedical text[1,2]. However, as pointed out in Leaman et.al[3], biomedical text is a highly codified result edited for clarity and intended at a large audience, while clinical narrative text contained in EHR is written by healthcare professionals about a single patient and is aimed at colleagues or themselves. This implies a variety of notes which are diverse in language and detail, as well as ad hoc terminology, including acronyms and jargon, far from being highly codified and standard. In addition, EHR are often filled under time pressure and with low motivation due to the fact that it takes time away from actual patient care. As a result, EHR narrative text usually suffers from low quality reflected in: variable semantics, structure without formal sentences, missing punctuation, missing expected words, misspelling or heterogeneous styles and jargon[3]. Moreover, independently of the motivation or resulting quality, the clinical language implies additional challenges, including term variability, ambiguity and complexity, lack of fine-grained classifications and data availability[4]. As such, many existing natural language processing approaches become ineffective or insufficient for it.

In this paper, we place attention particularly in named entity recognition (NER), for which specific challenges have also been identified, including which inference algorithm to use and how to use external knowledge resources (e.g. gazetteers)[5] or dealing with diverse medical fields, costly text annotations and different languages[6]. This paper, proposes an approach for dictionary-based, combined NER aimed at improving entity recall dealing with the aforementioned challenges associated to clinical narratives. To do so, we present related works in Section 2. We then go on to present the proposed strategy in Section 3. Section 4 then presents the evaluation results of applying the strategy to a standard data set. Finally, Section 5 presents some conclusions and suggests avenues for future work.

2. Related work

There have been proposed a wide variety of tools and methods to improve natural language processing of medical text. Table 1 summarizes the characteristics of a group of relevant works in NER in medical and biomedical fields.

	NER Technique	Entities	Features	Limitations
Chang 2002[11]	Pre-defined Dictionary	Acronyms and initials for health information resources, Human genome acronyms	Entity Recognition Search Abbreviations	Limited to abbreviations and acronyms
Jiang 2011[12]	Conditional Random Fields CRF	Medical Problems, Test, Treatments, Status	Entity Recognition	It needs a model training process
Jimeno 2008[13]	Exact Matching Flexile Matching	Pathologic Functions, Sign or Symptom, Cell or Molecular Dysfunctions, Findings	Entity Recognition	NER is limited to scientific texts
Aramaki 2009[15]	Conditional Random Fields CRF	Remedy, Medical Operations, Test, Examinations, Diseases, Symptoms, Medications	Entity Recognition Events Recognitions Date Times Negative Events	It needs a model training process
Zhang 2013[16]	Seed Term Collection Boundary Detection Entity Classification	Disorders, Therapeutic or Preventive Procedures, Laboratory Procedures, Clinical Drug, Test	Entity Recognition	It does not recognize typing and orthographic errors
Skeppstedt 2014[18]	Conditional Random Fields CRF Lemmatization Part-of-speech tagging	Disorders, Findings, Pharmaceutical, Body Structure	Entity Recognition NER in English and Swedish medical texts	It needs a model training process Previous annotation process developed by Physician or

	Terminology match		Errors recognition	Computational linguist
Xu 2012[19]	CRF-based NER Labeled sequential pattern LSP	Exams, Impressions, Conclusions	Entity Recognition Recognition of sentences with follow-up information	It needs a model training process

Table 1: Comparative analysis various methods of NER in medical text.

Based on these characteristics, we found there is not a general NER method considering EHR quality characteristics. Some of the analyzed works are focused on specific entities and others are unable to find entities without having a training set or a strong human involvement during the process. Our proposed approach aims to perform NER in medical text, especially EHR's text, despite their quality or variability, ambiguity or complexity.

3. Combined dictionary-based NER over EHR

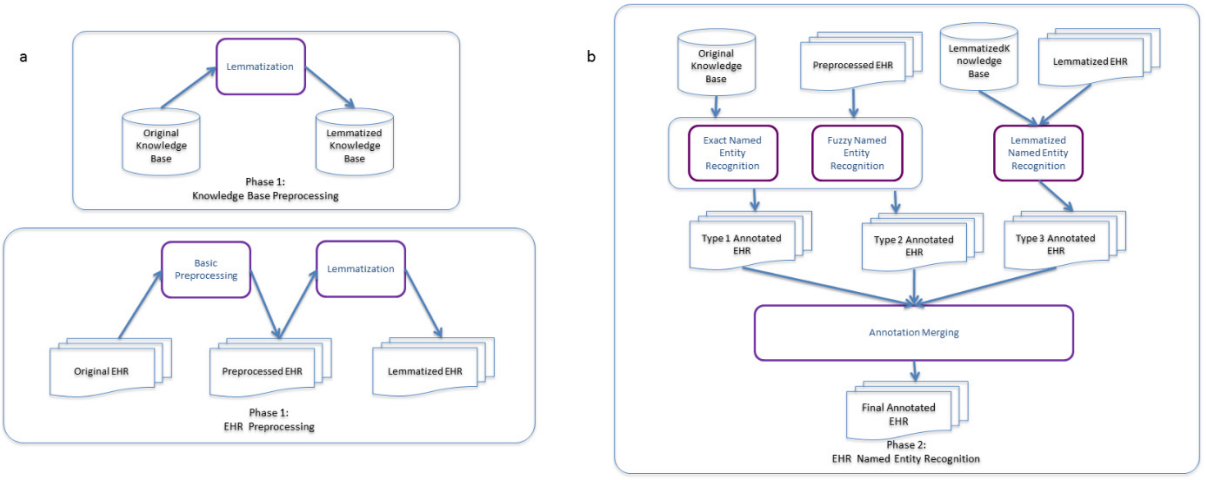

Fig. 1. Strategy proposed. (a) Preprocessing phase; (b) NER recognition phase.

Figures 1a and 1b present the strategy proposed to improve the recall of NER in EHR. It considers the quality characteristics of the texts contained in them and has two phases: the first one executes preprocessing tasks (Figure 1a), and the second one performs the NER including three types of recognition (Figure 1b).

The preprocessing phase extracts the lemmas from the domain knowledge base. This task generates a new knowledge base with the lemmas of the concepts that have to be recognized. Besides, it performs preprocessing tasks including tokenization and lemmatization over the EHR.

The objective of the NER phase is to recognize the relevant entities (e.g. diagnosis, treatment), even if they have misspellings or are written with differences with respect to the concepts included in the domain knowledge base. To achieve its purpose it has three types of NER processes that are going to be explained in section 3.1, 3.2 and 3.3. Finally, it includes a merging process that combines the overlapped entities recognized.

3.1. One to one recognition

The most basic approach when using a dictionary or gazetteer in NER is simply finding matches of the dictionary words within the text. That means having a one to one match between a dictionary term and a portion of the text.

For instance let *G* be a diagnosis dictionary, and *DocX* a document in the corpus

$$G = \{'diabetes', 'hypertension', 'COPD', 'influenza'\} \quad (1)$$

DocX='John Doe is a 67 year-old diabetic white male with a history of COPD, and hypertension. Mr. Doe was hospitalized 20 days ago at High Plains Hospital for pneumonia resulting from influezna ...'

Running NER with the dictionary would result in an annotated document DocXAnnot

DocXAnnot='John Doe is a 67 year-old diabetic white male with a history of [**COPD**], and [**hypertension**]. Mr. Doe was hospitalized 20 days ago at High Plains Hospital for pneumonia resulting from influezna ...'

This annotated document contains references for the terms 'COPD' and 'hypertension' and such annotations can be used to flag this document as relevant for such medical terms.

3.2. Fuzzy recognition

EHR are prone to contain typing and orthographic errors. This characteristic degrades the recall of dictionary-based NER, due to the mismatch between the dictionary and the actual words in the documents. For instance, a dictionary term 'pneumonia' will not match a mistyped 'npeumonia'.

We propose using a Fuzzy Gazetteer approach[20] in order to find more instances of the dictionary concepts in the texts, not only including direct matches, but also mistyped instances.

In order to compare EHR terms and dictionary terms we make use of the concept of edit distance[21] which is a metric representing the number of character changes among words. We define the threshold for accepting annotations as a proportion to the length of the dictionary term. For instance, we admit more changes in a long term like 'hypertension' than in a short one like 'COPD'. The threshold parameter $T \in [0,1]$ reflects such proportion and corresponds to the division of the edit distance over the number of characters in the dictionary term.

Using the document *DocX* and the dictionary *G* with the Fuzzy Gazetteer match approach would result in the following annotated document

DocXAnnFuzz= 'John Doe is a 67 year-old diabetic white male with a history of [**COPD:COPD(0)**], and [**hypertension:hypertension(0)**]. Mr. Doe was hospitalized 20 days ago at High Plains Hospital for pneumonia resulting from [**influezna:influenza(2)**] ...'

Each annotation in the document has the form [*text : dictionaryTerm(editDistance)*] where *text* represents the matched string in the document, *dictionaryTerm* the term in the gazetteer and *editDistance* the distance among them. In this case $T = 0.25$. Therefore, the edit distance 2 between *influenza* and *influenza* satisfies the defined threshold.

3.3. Stemmed recognition

Recognizing mentions based on the stem of a dictionary involves having the dictionary lemmas or stems, then finding the stem of each word in the documents and comparing the stem and not the actual words in the dictionary against the stemmed dictionary. In the example we first compute *G':* the stemmed version of the dictionary *G* and then, document *DocX* is processed with the same stemmer and as a result we have a parallel document *DocX'*

$$G' = \{diabet', hypertens', copd', influenza'\} \quad (2)$$

DocX'= 'john doe is a 67 year-old diabet white male with a histori of copd and hypertens mr doe was hospit 20 day ago at high plain hospit for pneumonia result from influezna ...'

Note that in *DocX'* words like 'diabetic' and 'hypertension' are respectively replaced with 'diabet' and 'hypertens'.

Once the documents and the dictionary are stemmed, they are matched and annotated. Here the annotation of 'diabet' corresponding to the stem 'diabet' shared by the words 'diabetes' and 'diabetic'.

DocXAnnStem= 'john doe is a yearold [**diabet**] white male with a histori of [**copd**] and [**hypertens**] mr doe was hospit day ago at high plain hospit for pneumonia result from influezna...'

4. Evaluation

4.1. Data set and evaluation configuration

For testing the impact of fuzzy annotations and stemmed annotations on NER we use the i2b2 NLP Data Set #7b: 'Heart Disease Risk Factors Challenge Data Set'[22,23] which contains annotations of risk factors related to Coronary Artery Disease (CAD). There are several annotations within the data set, which includes mentions to risk factors,

tests, events, symptoms among others. Since we are testing a dictionary NER approach we focus on direct mentions to entities. Specifically, we test mentions of diseases which include *obesity*, *diabetes*, *CAD*, *hyperlipidemia*, *hypercholesterolemia* and *hypertension* and mentions of medications. The data set contains several files from which we use the Complete Set for the Risk Factors Task combining the training and test sets to have 1304 texts in total.

To build the dictionary of terms we use the UMLT metathesaurus[14] to obtain names of the diseases and the medications. The synonyms of the aforementioned diseases were taken from the MEDLINE data source. For medications we combine the synonyms of the medication as well as the name of their children and their trade names. For instance *ACE Inhibitor* has synonym *angiotensin-converting-enzyme inhibitor*, has child *Trandolapril,* which has trade name *Mavik*. The medications dictionary was build using the NCI, MeSH and USPMG sources.

For evaluation we wrote a Java program which uses Gate Embedded to tokenize, stem and annotate the documents with the three kind of gazetteers. To study the impact of the fuzzy and stemmed versions of the NER, we compare four different combinations of the results against the gold standard: *e, ef, es, efs* where *e* means it contains the exact, *f* the fuzzy and *s* the stemmed annotations. For instance *ef* does not contain stemmed annotations.

For each configuration the threshold T is set to 0.15 meaning we accept an edit distance not bigger than 15% of the dictionary term's length. After running the annotation process we count the number of true positives (*tp*) and false positives (*fp*) and compute the false negatives (*fn*) according to the gold standard annotations. Then we compute Precision, Recall, F1 and F2. An annotation counts as a *tp* only if it starts and ends exactly where an annotation in the gold standard does. Additionally, we study the impact of the fuzziness threshold on the aforementioned measures by running the *ef* configuration with different values of *T* starting from 0.05 to 0.3 with steps of 0.05.

4.2. Results

Table 2 shows the different configurations and their values of precision, recall, F1 and F2. We note that adding the fuzzy and stemmed approaches increases recall. In terms of recall the stemmed annotations have more positive impact than the fuzzy annotations and the combination of all approaches (efs) achieves the best results.

Table 2. Comparison of the different approaches.

	tp	fp	# Ann	fn	P	R	F1	F2
e	9124	**4750**	13874	7442	**0,658**	0,551	0,599	0,569
ef	9267	5007	14274	7299	0,649	0,559	**0,601**	0,575
es	9371	5361	14732	7195	0,636	0,566	0,599	0,578
efs	**9491**	5585	**15076**	**7075**	0,630	**0,573**	0,600	**0,583**

Table 3. Impact of the fuzziness threshold T for the *ef* approach

T	tp	fp	# Ann	fn	P	R	F1	F2
0,05	9146	**4768**	13914	7420	**0,657**	0,552	0,600	0,570
0,10	9210	4849	14059	7356	0,655	0,556	**0,601**	0,573
0,15	9267	5007	14274	7299	0,649	0,559	**0,601**	**0,575**
0,20	9301	5908	15209	7265	0,612	0,561	0,585	0,571
0,25	**9356**	6379	15735	7210	0,595	**0,565**	0,579	0,570
0,30	**9356**	8088	**17444**	**7210**	0,536	**0,565**	0,550	0,559

We also note that, as expected, when trying to improve recall, the precision is affected negatively. By looking at the F1 measure, which combines precision and recall, we note that it never diminishes indicating that the decrease in precision is compensated by the increase in recall. Also the measure F2, which favors recall over precision, improves in all the models. In terms of the fuzziness threshold (table 3) it is clear that slightly increasing the value improves recall initially, but after a value of 0.15 precision is strongly degraded in comparison to the gain in recall.

The number of true positives and annotations suggests that by including the fuzzy and stemmed annotations we gain a big number of new annotations (9% more corresponding to 1202 annotations) where new information could be found. Although the precision decreases, all those new annotations can be weighted in terms of their distance to the original terms, meaning that fuzzy and stemmed matches have some degree of confidence. Such condition allows us to open the window of annotation candidates without ignoring the quality loss of those new candidates, which is better than reaching just the exact annotations.

5. Conclusions

This paper proposes an approach for named entity recognition in narrative texts of EHR. This task is difficult because of the particular and unique characteristics of texts in health records (codified, condensed, jerky language). This approach combines a direct match technique with fuzzy matching and stemmed matching. The proposed method was tested using the i2b2 NLP Data Set which includes a gold standard with annotations. Experimental results on this data set shows an improvement of the recall while having a limited impact on precision. Nevertheless, when analyzing the text, the proposed methods do not take into account the surrounding words (the context) appearing near a named entity candidate. In future works, taking this context into account may be of great interests for improving both precision and recall.

Acknowledgements

This work is part of the projects funded by Hospital Universitario San Ignacio and Pontificia Universidad Javeriana for improving the analysis of electronic health records. Deidentified clinical records used in this research were provided by the i2b2 National Center for biomedical Computing funded by U54LM008748 and were originally prepared for the Shared Tasks for Challenges in NLP for Clinical Data organized by Dr. Ozlem Uzuner, i2b2 and SUNY.

References

1. A. Cohen, W. Hersh, A survey of current work in biomedical text mining, Briefings in Bioinformatics 6 (1) (2005) 57–71.
2. M. Simpson, D. Demner-Fushman, Biomedical text mining: A survey of recent progress, in: Mining Text Data, 2012, pp. 465–517.
3. R. Leaman, R. Khare, Z. Lu, Challenges in clinical natural language processing for automated disorder normalization, Journal of Biomedical Informatics.
4. A. Dehghan, J. Keane, G. Nenadic, Challenges in clinical named entity recognition for decision support, 2013, pp. 947–951.
5. L. Ratinov, D. Roth, Design challenges and misconceptions in named entity recognition, 2009, pp. 147–155.
6. M. Skeppstedt, M. Kvist, G. Nilsson, H. Dalianis, Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study, Journal of Biomedical Informatics 49 (2014) 148–158.
7. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, Journal of Machine Learning Research 12 (2011) 2493–2537.
8. M. Patawar, M. Potey, Approaches to Named Entity Recognition: A Survey, International Journal of Innovative Research in Computer and Communication Engineering 3 (12) (2015) 12201–12208.
9. I. Spasi, J. Livsey, J. Keane, G. Nenadi, Text mining of cancer-related information: Review of current status and future directions, International Journal of Medical Informatics 83 (9) (2014) 605–623.
10. D. Demner-Fushman, W. W. Chapman, C. J. McDonald, What can natural language processing do for clinical decision support?, Journal of Biomedical Informatics 42 (5) (2009) 760–772.
11. J. T. Chang, H. Sch¨utze, R. B. Altman, Creating an online dictionary of abbreviations from medline, Journal of the American Medical Informatics Association 9 (6) (2002) 612–620
12. M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, H. Xu, A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries, Journal of the American Medical Informatics Association 18 (5) (2011) 601–606.
13. A. Jimeno, E. Jimenez-Ruiz, V. Lee, S. Gaudan, R. Berlanga, D. Rebholz-Schuhmann, Assessment of disease named entity recognition on a corpus of annotated sentences, BMC bioinformatics 9 (Suppl 3) (2008) S3.
14. D. A. Lindberg, B. L. Humphreys, A. T. McCray, The unified medical language system., Methods of information in medicine 32 (4) (1993) 281–291.
15. E. Aramaki, Y. Miura, M. Tonoike, T. Ohkuma, H. Mashuichi, K. Ohe, Text2table: Medical text summarization system based on named entity recognition and modality identification, in: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, Association for Computational Linguistics, 2009, pp. 185–192.
16. S. Zhang, N. Elhadad, Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts, Journal of Biomedical Informatics 46 (6) (2013) 1088 – 1098, special Section: Social Media Environments.
17. O¨ . Uzuner, B. R. South, S. Shen, S. L. DuVall, 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text, Journal of the American Medical Informatics Association 18 (5) (2011) 552–556.
18. M. Skeppstedt, M. Kvist, G. H. Nilsson, H. Dalianis, Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study, Journal of Biomedical Informatics 49 (2014) 148 – 158.
19. Y. Xu, J. Tsujii, E. I.-C. Chang, Named entity recognition of follow-up and time information in 20 000 radiology reports, Journal of the American Medical Informatics Association 19 (5) (2012) 792–799.
20. B. W. Paleo, An approximate gazetteer for gate based on levenshtein distance, ESSLLI 2007 (2007) 197.

21. V. Levenshtein, Binary Codes Capable of Correcting Deletions, Insertions and Reversals, Soviet Physics Doklady 10 (1966) 707.
22. A. Stubbs, zlem Uzuner, Annotating risk factors for heart disease in clinical narratives for diabetic patients, Journal of Biomedical Informatics 58, Supplement (2015) S78 – S91, proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.
23. A. Stubbs, C. Kotfila, H. Xu, zlem Uzuner, Identifying risk factors for heart disease over time: Overview of 2014 i2b2/uthealth shared task track 2, Journal of Biomedical Informatics 58, Supplement (2015) S67 – S77, proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data