dHealth 2019 – From eHealth to dHealth
D. Hayn et al. (Eds.)

doi:10.3233/978-1-61499-971-3-41

Application of Named Entity Recognition Methods to Extract Information from Echocardiography Reports

Szabolcs SZEKÉR[a], György FOGARASSY[b], Károly MACHALIK[a] and Ágnes VATHY-FOGARASSY[a, 1]
[a]*Department of Computer Science and Systems Technology, University of Pannonia, Hungary*
[b]*State Hospital of Cardiology, Hungary*

Abstract. As there is no consensus about how to store the results of echocardiography examinations, information extraction from them is a non-trivial task. Successful named entity recognition (NER) is key to getting access to the stored information and the process of identification has been recognized as a bottleneck in text mining. Our goal was to develop and compare such NER methods that are capable of achieving this task. Our practical results show that the text mining-based NER method is able to perform at a similar level in finding and identifying terms as the regular expression-based NER method. The paper highlights the advantages and disadvantages of both methods.

Keywords. data mining, text mining, named entity recognition, echocardiography, electronic medical record

1. Introduction

Most medical institutes generally use Electric Medical Record (EMR) to record and store information about their patients, including diagnostics, performed treatments and their results. EMR is a valuable information source for medical analysis, however it is usually incomplete or redundant, making data mining a difficult and challenging task. It is especially true in case of echocardiography reports. Generally, echocardiography reports can be divided into two parts in terms of diagnostic content: in the first semi-structured part diagnostic results are stored in the form of term-value pairs (e.g.: *interventricular septum: 14 mm*), and in the second part results are recorded as free text written in natural language (e.g.: *mild left ventricular hypertrophy*). As there is no consensus about how to store the results of echocardiography examinations and it is varying across different medical institutes, processing of echocardiography reports is a nontrivial task. Present paper is focusing on how to process the first, semi-structured part of echocardiography reports. As processing of the free text part requires quite different methods, including Natural Language Processing (NLP) techniques, we do not deal with them in this paper.

[1] Corresponding Author: Ágnes Vathy-Fogarassy, Department of Computer Science and Systems Technology, University of Pannonia, 2. Egyetem Str., 8200 Veszprém, Hungary, E-Mail: vathy@dcs.uni-pannon.hu

Generally, information extraction from medical texts focuses on the following two tasks: named-entity recognition (NER, or term extraction), and relation extraction (RE). Named-entity recognition refers to the process of identifying particular types of names, terminologies or symbols in documents, while relation extraction identifies the relation between them [1]. Successful term identification is key to getting access to the stored information and the process of identification has been recognized as a bottleneck in text mining [2]. The process of term identification is usually done in three steps: the first step is term recognition; the second step is term classification; and the last step is term mapping [2].

There are two possible approaches to solve this task. The first approach is to directly search for specific terms (e.g. *aortic root*, *ejection fraction*) in documents. Direct term search always relies on a specialized dictionary to recognize and classify medical terminology, and the performance of this approach heavily depends on the coverage and quality of the dictionary. The acquisition of such knowledge is a time-consuming task. Direct search can also be extended by pattern search, which requires a priori knowledge about the structure of the processed text (e.g. use of colon between terms and values, order of terms, various expletives). With this extension, it becomes possible to recognize terms and their measured value (e.g. *aortic root: 27 mm*) together.

Other term extraction methods also exist which utilize classical text mining techniques. These text mining-based solutions do not need a predefined dictionary to extract terms from the text, but simply collect every occurrence of word sequences that are possibly valid terms. However, these methods require a text pre-processing phase (including text cleaning), and term candidates must be identified and mapped onto a dictionary after term extraction.

In the literature, several international studies have been published which are engaged in echocardiography report processing [3-10]. They are mostly based on the direct search approach, but some of them apply text-mining methods as well. In the published studies, typically only the extraction of one specific parameter is the aim, such as ejection fraction (EF). Garvin et al., Kim et al., and Xie et al. all successfully extracted this parameter from free text documents and described practical extraction techniques [3-5]. In [6] a natural language-based method was presented which uses a predefined dictionary, expert rules and predefined patterns to extract echocardiography measurements from documents. In this study, a pattern-matching algorithm was created and tested to extract term candidates from a large set of clinical notes. The presented method relies heavily on pattern matching, but it can also identify possible misspellings and synonyms by iterative extraction. Wells et al. also successfully extracted a set of predefined parameters, including wall thicknesses, chamber dimensions or flow velocities [7]. They applied NLP to parse the most frequently measured dimensions and used outlier analysis to filter out unrealistic values. Toepfer et al. developed and evaluated an information extraction component with fine-grained terminology that enabled them to recognize almost all relevant information stated in German transthoracic echocardiography reports at the University Hospital of Würzburg [8]. Jonnalagadda et al. described an information extraction-based approach that automatically converts unstructured text into structured data, which is cross-referenced against eligibility criteria using a rule-based system to determine which patients qualify for a heart failure with preserved ejection fraction (HFpEF) clinical trial [9]. In [10], Renganathan proposed text mining techniques that enable the extraction of unknown knowledge from unstructured documents.

As we can see, all the suggested methods report successful medical text processing but were implemented in different ways. However, until now, there was no analysis

published that would compare the two basic approaches. The purpose of our research was to examine how well a text mining-based solution fares against a direct term search-based method in processing medical, especially echocardiography documents, and whether it is able to outperform it or not. For this purpose, we implemented both approaches, processed the same corpus of echocardiography reports with them, and compared the results. Our results are primarily valid for the analysis of echocardiography reports, but we believe that they might be valid in case of disclosure information stored in term-value pairs from other medical documents as well.

The structure of this document is as follows. In Section 2, we give a brief overview of the challenges faced when processing echocardiography reports and present two fundamentally different methods to extract, identify, and map terms. In Section 3, the used dataset and the evaluation process are described and the result of the analysis is presented. Finally, in Section 4, general experiences and future developments are discussed.

2. Methods

The most vexatious problem with echocardiography reports is that there is no unified process on how to record data of patients. The form of recorded information varies from medical institute to medical institute. Furthermore, not only the location of data recording is an influencing factor, but medical assistants or doctors record the results according to their own habits and, arising from the lack of a unified recording interface, free text contains many typos as well.

In our study, two methods have been realized to extract terms from the first, semi-structured part of echocardiography reports. The first method is a general regular expression-based method which processes raw text, meaning that there is no pre-cleaning applied, and assumes that terms and their measurement results are separated by a colon. The second method is based on traditional text mining methods. In this case, the raw text is first cleaned and then the cleaned text is processed. This method searches for numerical values and assumes that there is a term before and a unit of measurement (if needed) after each numerical value.

The main difference between the two methods lies in the text preparation phase. The regular expression-based method processes raw text and assumes, based on a priori knowledge, that the term and their measurement result pairs follow a certain pattern, e.g. they are separated by a colon, while the text mining-based method cleans and manipulates the text in such a way that it becomes easier to process. Furthermore, the text mining-based method does not rely on any a priori knowledge about the medical text to process them. These two methods are introduced in detail in the following subsections.

2.1. Regular expression-based NER

The regular expression-based NER method uses regular expressions to extract terms from echocardiography reports. A regular expression is a sequence of characters that defines a search pattern. Usually this pattern is used by string searching algorithms for "find" or "find and replace" operations on strings. The regular expression-based method processes raw text, meaning that the data is processed as it is, no pre-cleaning methods are applied. Furthermore, the regular expression-based processing method, based on a priori knowledge, presumes that every term and the adherent value is separated by a

colon (in our case, but it can be separated by any other predefined separator character as well) and the applied regular expressions are built upon this assumption and knowledge.

In our study, firstly simpler regular expressions have been defined on which more complex expressions were based. These rudimentary regular expressions include expressions for *terms*, *values*, *units* and *extended units*. Sample expressions for terms and values are the following:

terms `r'(?P<term>(?!\d)\w\D+)'` (1)

values `r'(?P<digits>\d[\d,.+\-x/*]*)'` (2)

The first expression defines that terms cannot start with a number and one or more non-numeric characters follow a word character. The second expression defines the values in such a way that they can be integers (e.g. *27*), decimal numbers (e.g. *12.5*), ranges of values separated by a hyphen (e.g. *25-28*, *12.4-12.7*) or a multi-dimensional value specified by an "x" character (e.g. *27x13*).

Using these rudimentary expressions more complex expressions can also be constructed. For example, the *measurement result* is a complex expression, which is a concatenation of *values*, some separating *whitespace characters* and a measurement *unit* with affixation taken into account (*measurement_result = [values][whitespace characters][unit][affixation]*). An expression for a *term–measurement_result* pair is the concatenated form of the *term*, *whitespace characters*, a *colon*, *whitespace characters* and the *measurement result* expressions (*term–measurement_result = [term][whitespace characters]**:**[whitespace characters][measurement_result]*). The flexibility of the regular expression-based NER comes from its ability to find character sequences matching the defined patterns regardless of their position in a longer sequence.

Using the previously defined regular expression set, the raw text can be processed. Based on the indeterministic nature of regular expression matching, the echocardiography reports were processed from start to end. If a string matching an expression was found, the string was processed, stored, and removed from the document. These few steps were executed iteratively until no processable string was left.

2.2. Text mining-based NER

The second method is a more straightforward approach which utilizes traditional data mining and cleaning methods. It pre-processes the raw text without any a priori knowledge about the contents of the documents. As part of the cleaning process, this method unifies whitespaces, removes all colons, parentheses and unneeded characters. It is important to note that however, it does not modify any commas or dots as they can also be used as decimal comma or decimal point based on the localization of the recording software. It also unifies the units of measurement based on a predefined list. The unified and found measures are concatenated to the preceding numerical values during the pre-processing phase.

The trick of this method is that it assumes that all measured values are numerical and before every numerical value there is a term and after every numerical value there can be a unit of measurement present. To remove the numerical values which do not express measurements results, the algorithm in the pre-processing phase modifies some of the measures like *mm2* or *cm2* to, respectively, *sqrmm* and *sqrcm*.

After the text pre-processing phase, the text mining-based NER splits the cleaned documents into "words" (sequences of characters separated by whitespaces) and searches

for the first occurrence of a "word" starting with a numerical value. The preceding n words, if n words are present, are considered term candidates. The candidates are then checked, marked, and stored for later usage. In our case, $n = 4$ was chosen for the threshold of the number of words for candidate terms.

2.3. Identification of complex terms and measurement results

The previously introduced methods are capable of recognizing *term–measurement result* pairs, however more complex sequences are also present in the echocardiography reports e.g. *term1–term2–measurement_result1–measurement_result2* (e.g. *left ventricular diameter end-diastolic/end-systolic: 54/35 mm*).

To find these kinds of expressions as well, in case of the regular expression-based method, the execution was improved by adding more complex regular expressions. After repeated testing, we came to the conclusion that by processing the raw text it is not possible to find every term present in the documents and expanding the regular expressions for every special occurrence is a difficult, time-consuming task.

The text mining-based method was extended as follows. During the execution, the found sequence was checked whether it fits the predefined rules, e.g. simple *term–measurement_result* pair (e.g. *ejection fraction: 56%*), *term1–measurement_result1–subterm2–measurement_result2* sequence *(*e.g. *ejection fraction Teichholz: 56%, Simpson: 52%")*, or *term1–term2–measurement_result1–measurement_result2* sequence (e.g. *E/A: 0.4/0.8 m/s*). If one of the rules fits the word list, the terms and the numerical values were stored. These rules were defined as *IF-THEN* rules. In this form the rule definition is much easier than the creation of the previously described, more complex regular expressions.

2.4. Dictionary-based mapping

The recognized named entities from the processed documents were checked whether they are valid terms or not by using a dictionary of terms. This dictionary has been created with the help of a medical expert. The used dictionary contains more than 30 terms from the field of cardiology probably present in some form in echocardiography reports and over 100 synonyms have also been defined for the 30 terms. The previously defined $n = 4$ word length of terms stems from the dictionary, for the reason that the maximum length of terms recorded in the dictionary is 4.

The terms extracted in the previous phases were compared against the elements of the dictionary. The Jaro-Winkler distance [11] was calculated for each comparison and if the distance was lower than a specified distance threshold, the term was considered valid and identified. This threshold parameter was defined as the lowest, non-zero intra-distance of the terms stored in the dictionary. The Jaro-Winkler distance (d_w) can be calculated in the following way:

$$d_w(\mathrm{s}_1, \mathrm{s}_2) = 1 - sim_w(\mathrm{s}_1, \mathrm{s}_2) \tag{3}$$

$$sim_w(\mathrm{s}_1, \mathrm{s}_2) = sim_j(\mathrm{s}_1, \mathrm{s}_2) + lp(1 - sim_j(\mathrm{s}_1, \mathrm{s}_2)) \tag{4}$$

where sim_j is the Jaro similarity for s_1 and s_2 strings, l is the length of a common prefix up to 4 characters and p is a constant scaling factor with a standard value of 0.1. The Jaro similarity (sim_j) is calculated in the following way:

$$sim_j(s_1, s_2) = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases} \quad (5)$$

where $|s_i|$ is the length of s_i, m is the number of matching characters and t is half of the number of transpositions. The concept of matching and transpositions is detailed in [11].

3. Results

To compare the effectiveness of the previously presented regular expression-based NER (RE-NER) and text mining-based NER (TM-NER) a corpus containing 20 089 anonymized echocardiography reports has been processed. Each document had a unique identifier and a basic description about the diagnosis. The first, semi-structured part contained various terms mainly in the form of *[term]:[measurement_result]*. As any other free text stored medical records, the echocardiography reports under study also included typing errors or deficiencies.

Results of both algorithms were evaluated in the following way: for each method–term pair we counted the number of documents in which the method has found the specific term. Furthermore, an important part of the evaluation was to identify the documents in which only one method was able to identify the given term. The number of matched documents by any methods (*N*), by RE-NER (N_{RE}), by TM-NER (N_{TM}), and the number of documents exclusively matched by RE-NER (Ne_{RE}), and by TM-NER (Ne_{TM}) can be seen in Table 1. To evaluate the relative success of the methods, we also calculated the frequencies of the matched documents relative to the number of the documents matched by any method (q_{RE}, q_{TM}). Furthermore, the rate of the exclusively matched documents was also calculated (qe_{RE}, qe_{TM}). These results are presented in Table 2.

Table 1. The number of the most common terms identified by RE-NER and TM-NER methods.

		RE-NER		*TM-NER*	
Term	***N***	N_{RE}	Ne_{RE}	N_{TM}	Ne_{TM}
Left ventricular end-systolic diameter	19 598	19 464	42	19 549	116
Interventricular septum (end-diastolic)	19 562	19 498	109	19 491	43
Aortic root	19 537	19 492	66	19 476	21
Posterior wall (end-diastolic)	19 496	15 696	116	19 386	3 800
Left ventricular end-diastolic diameter	19 240	19 096	102	19 147	81
Left atrium (M-mode)	19 344	19 259	208	19 144	85
E	18 759	18 719	44	18 723	59
EF	18 768	18 640	636	18 135	131
A	18 458	18 421	977	17 483	41
Interventricular septum (end-systolic)	14 372	2	2	14 370	14 370
Posterior wall (end-systolic)	14 310	41	1	14 309	14 269
Right ventricle (M-mode)	10 656	10 448	239	10 432	17
2D right atrial dimensions	10 492	10 398	237	10 264	11

Most differences of N_{RE} and N_{TM} occur from typos, missing spaces or non-numerical values. During testing, we found that there are cases when spaces are missing between some named entities (terms). As TM-NER is based on the list of words, in this case, this method is unable to find the appropriate term. To handle this kind of failure it is suggested to insert separator space characters into the text (for example after the measurements) during text cleaning. Furthermore, there were occurrences of *term–measurement_result* pairs where the measurement result part was a non-numeric value. The TM-NER method is unable to identify these kind of results, but RE-NER may be able to find these occurrences based on the presumption that terms and values are separated by a colon regardless the type of value. The biggest difference occurred during the exploration of terms *Interventricular septum (end-systolic)* and *Posterior wall (end-systolic)*. These terms are composite terms. They follow the *term1–measurement_result1–subterm2–measurement_result2* pattern. RE-NER struggles to find and to process these kinds of terms in a humanly consumable way.

Table 2. The relative occurrence of most common terms identified by RE-NER and TM-NER methods.

		RE-NER		***TM-NER***	
term	***N***	q_{RE}	qe_{RE}	q_{TM}	qe_{TM}
Left ventricular end-systolic diameter	19 598	99,32%	0,21%	99,75%	0,59%
Interventricular septum (end-diastolic)	19 562	99,67%	0,56%	99,64%	0,22%
Aortic root	19 537	99,77%	0,34%	99,69%	0,11%
Posterior wall (end-diastolic)	19 496	80,51%	0,59%	99,44%	19,49%
Left ventricular end-diastolic diameter	19 240	99,25%	0,53%	99,52%	0,42%
Left atrium (M-mode)	19 344	99,56%	1,08%	98,97%	0,44%
E	18 759	99,79%	0,23%	99,81%	0,31%
EF	18 768	99,32%	3,39%	96,63%	0,70%
A	18 458	99,80%	5,29%	94,72%	0,22%
Interventricular septum (end-systolic)	14 372	0,01%	0,01%	99,99%	99,99%
Posterior wall (end-systolic)	14 310	0,29%	0,01%	99,99%	99,71%
Right ventricle (M-mode)	10 656	98,05%	2,24%	97,90%	0,16%
2D right atrial dimensions	10 492	99,10%	2,26%	97,83%	0,10%

4. Discussion

Information extraction from echocardiography reports stored as free text is a challenging task in medical analysis. The success of information extraction mainly depends on the quality of the source documents, and the algorithms have to overcome many difficulties. The main goal of this study was to compare two types of information extraction methods and to highlight their strengths and drawbacks. The algorithms in the study were (i) a classical regular expression-based information extractor algorithm, which is based on some prior knowledge about the text to be processed and (ii) a general text mining algorithm, which operates without any prior knowledge about the text. Our practical

results show that the text mining-based method is able to perform at a similar level in finding and identifying terms as the regular expression-based method. Both methods have advantages over the other. The text mining-based algorithm has difficulty in handling missing space characters. As the text mining-based method is based on the assumption that measured results are stored as numerical values, this method is unable to find non-numerical values. In case of the regular expression-based method, the formulation of the expression set is a difficult task and it is even harder to extend this regular expression set to recognize complex terms. Furthermore, not all occurrences can be expressed with a general expression. These special occurrences require more and more unique expressions to be added to the set, which increases the processing time.

Our primary finding is that the text mining-based NER method is able to perform at a similar level in finding and identifying terms as the regular expression-based method and in case of extracting complex terms and their measurement results it outperforms the regular expression-based NER method. Information extraction can be further improved by implementing a hybrid NER which merges the advantages and negates the disadvantages of both methods. This hybrid NER is part of our future research.

Acknowledgment

We acknowledge the financial support of Széchenyi 2020 under EFOP-3.6.1-16-2016-00015 and UNKP-18-3 New National Excellence Program of the Ministry of Human Capacities, and the professional support of GINOP-2.2.1-15-2016-00019 "Development of intelligent, process-based decision support system for cardiologists".

References

[1] Wencheng Sun, Zhiping Cai, Yangyang Li, et al. Data Processing and Text Mining Technologies on Electronic Medical Records: A Review. *Journal of Healthcare Engineering*, 2018; Article ID 4302425

[2] Krauthammer M, Nenadic G. Term identification in the biomedical literature. *J Biomed Inform*. 2004;37(6):512–526.

[3] Xie F, Zheng C, Yuh-Jer Shen A, Chen W. Extracting and analyzing ejection fraction values from electronic echocardiography reports in a large health maintenance organization. *Health Inform J.* 2017;23(4):319–328.

[4] Garvin JH, DuVall SL, SOuth BR, et al. Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architechture (UIMA) fro heart failure. *J Am Med Inform Assoc.* 2012;19(5):859–866.

[5] Kim Y, Garvin JH, Goldstein MK, et al. Extraction of left ventricular ejectionfraction information from various types of clinical reports. *J Biomed Inform.* 2017;67:42–48.

[6] Patterson OV, Freiberg MS, Skanderson M, et al.. Unlocking echocardiogram measurements for heart disease research through natural language processing. *BMC Cardiovasc Disord.* 2017;17(1):151.

[7] Wells QS, Farber-Eger E, Crawford DC. Extraction of echocardiographic data from the electronic medical record is a rapid and efficient method for study of cardiac structure and function. *J Clin Bioinforma.* 2014;4(1):12.

[8] Toepfer, M., Corovic, H., Fette, et al. Fine-grained information extraction from German transthoracic echocardiography reports. *BMC Medical Informatics and Decision Making*, 2015; 15(1):91

[9] Jonnalagadda, S.R., Adupa, A.K., Garg, R.P. et al. Text Mining of the Electronic Health Record: An Information Extraction Approach for Automated Identification and Subphenotyping of HFpEF Patients for Clinical Trials. *J. of Cardiovasc. Trans. Res.* 2017; 10(3), 313–321.

[10] Renganathan, V. Text Mining in Biomedical Domain with Emphasis on Document Clustering. *Healthcare Informatics Research*, 2017; 23(3), 141–146.

[11] Piskorski J., Sydow M. String Distance Metrics for Reference Matching and Search Query Correction. In: Abramowicz W. (eds) Business Information Systems. BIS 2007. LNCS, Vol 4439. Springer, Berlin