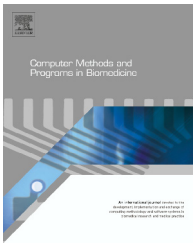
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 119 (2015) 88–100

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



Enhancing medical named entity recognition with an extended segment representation technique

Sara Keretna[a,*], *Chee Peng Lim*[a], *Doug Creighton*[a], *Khaled Bashir Shaban*[b]

[a] *Centre for Intelligent Systems Research, Deakin University, Australia*
[b] *Computer Science and Engineering Department, College of Engineering, Qatar University, Qatar*

ARTICLE INFO

Article history:
Received 16 September 2014
Received in revised form
18 February 2015
Accepted 24 February 2015

Keywords:
Biomedical text mining
Information extraction
Unstructured electronic medical records
Natural language processing
Biomedical text annotation

ABSTRACT

Objective: The objective of this paper is to formulate an extended segment representation (SR) technique to enhance named entity recognition (NER) in medical applications.
Methods: An extension to the IOBES (Inside/Outside/Begin/End/Single) SR technique is formulated. In the proposed extension, a new class is assigned to words that do not belong to a named entity (NE) in one context but appear as an NE in other contexts. Ambiguity in such cases can negatively affect the results of classification-based NER techniques. Assigning a separate class to words that can potentially cause ambiguity in NER allows a classifier to detect NEs more accurately; therefore increasing classification accuracy.
Results: The proposed SR technique is evaluated using the i2b2 2010 medical challenge data set with eight different classifiers. Each classifier is trained separately to extract three different medical NEs, namely treatment, problem, and test. From the three experimental results, the extended SR technique is able to improve the average F1-measure results pertaining to seven out of eight classifiers. The kNN classifier shows an average reduction of 0.18% across three experiments, while the C4.5 classifier records an average improvement of 9.33%.

1. Introduction

Text-based information has become increasingly important in the medical field [1,2]. It forms an important aspect of the scientific literature in medicine, electronic health records, medical databases, and medical discussion forums. Previous studies [3] have indicated that unstructured medical text contains valuable unexplored knowledge. In this aspect, text mining is concerned with the challenges in processing and extracting useful information from unstructured text.

In text mining, named entity recognition (NER) [4,5], also known as concept extraction, is a fundamental step. NER involves the process of detecting specific concepts from textual data [4,6]. The concepts can be general, e.g. human names, or domain specific, e.g. drug names which are known as medical named entities (MNEs) [2]. The process of extracting MNEs is called medical named entity recognition (MNER). MNER serves as the basis in many medical applications [6,7], e.g. drug–drug interaction [8,9], adverse drug effect detection [10,11], diagnosis classification [12], protein–protein and gene–gene interactions [13], and biomedical question–answer systems [1].

* *Corresponding author at.*
E-mail addresses: skeretna@deakin.edu.au (S. Keretna), chee.lim@deakin.edu.au (C.P. Lim), dougc@deakin.edu.au (D. Creighton), khaled.shaban@qu.edu.qa (K.B. Shaban).

http://dx.doi.org/10.1016/j.cmpb.2015.02.007
0169-2607/

MNER techniques have evolved over times since the inception of lexicon-based techniques [14]. Later, rule-based [15], machine learning (ML) [16], and hybrid techniques [17,18] have been introduced. In this aspect, ML and hybrid techniques are able to produce better results as compared with other techniques [19]. As such, there are a lot of studies focusing on ML models for undertaking MNER tasks. Most of the investigations are concerned with designing a set of features and an appropriate classifier that can contribute to improved results. From a literature review, relatively few studies examine the effects of segment representation (SR) of training data in supervised ML models for NER. This gap motivates us to focus on SR techniques for MNER in this study.

SR [20] involves the process of tagging NEs to training data. The tagged or annotated data samples are then used as the inputs to train an ML algorithm. SR has been applied to a variety of natural language processing tasks, which include part-of-speech (POS) tagging [21]. In NER, selecting a classification method and designing the corresponding feature set that can be identified and extracted from the data set are the main concerns. Indeed, the use of an appropriate SR technique can affect the detection results significantly [22]. Nevertheless, this issue is normally overlooked in NER, since only few investigations are reported in the literature thus far. As a result, we focus on the use of SR for MNER in this study. Specifically, an extended IOBES (Inside/Outside/Begin/End/Single) SR technique is formulated in this study. Although IOBES is not the most frequently used SR technique in MNER, it is able to produce better results than other SR techniques such as IOB2 (Inside/Outside/Begin version2) and IO (Inside/Outside) [20,22]. Here, a case study to evaluate the effects of using the proposed SR technique to undertake MNER tasks is presented. Based on the i2b2 2010 medical challenge data set [45], the task of detecting three MNEs, namely, treatment, problem, and test, using eight classifiers is conducted. From the three experimental results, the proposed SR technique is able to improve the average F1-measure results pertaining to seven out of eight classifiers.

The rest of this paper is organised as follows. In Section 2, a review of related work in the literature is presented. In Section, the proposed SR technique is explained. The experimental setup is presented in Section 4, while the detailed results and discussion are presented in Section 5. Conclusions and suggestions for future work are included in Section 6.

2. Background and related work

Unstructured textual data in the medical field contain valuable knowledge that has not been fully exploited. Studies indicate that the majority of this knowledge base, although being potentially important in disease diagnosis/prognosis, does not exist in any form of structured medical data [23]. Nevertheless, extracting useful information from unstructured medical text is a challenging task owing to the complex and ambiguous nature of the language used in the medical field. During the last decade, the interest in medical text mining has increased significantly, and MNER has become an important medical text-mining topic. Indeed, the statistics in Appendix A (Figs. A1 and A2) clearly show the increase in the number of publications related to MNER between 2000 and 2013 in two different scientific databases, i.e., Science Direct [24] and PubMed [25], respectively.

In general, NER techniques can be categorised into four types [11,17,26]: lexicon-based, rule-based, ML, and hybrid techniques. ML is a key technique in NER, while hybrid techniques usually comprise a combination of ML and other methods. Recently, there has been an increased interest in applying supervised ML models to undertaking NER tasks. This is because supervised ML models are able to produce satisfactory results, and can be extended to other domains [17]. In this aspect, classifiers are supervised ML models useful for tackling NER problems. A classifier can be used independently or as part of a hybrid technique. Examples of useful classifiers include conditional random field (CRF) [16,18,19,27], maximum entropy (ME) [28], hidden Markov model (HMM) [17] and support vector machine (SVM) [29–31]. All these classifiers require data sets with annotated NEs. In this aspect, SR offers a straightforward and useful technique for annotating data samples.

Introduced initially as a classification pre-processing step, SR serves as an alternative to the bracket structure representation method [32]. It is designed to aid the extraction process of non-overlapping, non-recursive chunks of text for shallow parsing and noun phrase chunking applications. The fundamental idea of SR is to provide every word a tag and a label. The label shows the class that the word belongs to, while the tag indicate the position of the word in the class, as classes can be formed using more than one word (i.e., text chunk). Examples of classes are POS and NEs, while examples of labels are 'Begin', 'Middle', and 'End'. The process of representing segments of texts is also known as chunk tagging, chunk encoding, and chunk representation [33].

A number of popular SR techniques include IO, IOB1 (Inside/Outside/Begin version-1) [34], IOB2 (Inside/Outside/Begin version-2) [35], IOE2 (Inside/Outside/End version-2) [36,37], and IOBES [21,22,38]. The most frequently used SR technique in NER is IOB2 [18,27,39,40]. Table 1 shows an example of how a medical text is tagged using three SR techniques [20], namely I/O, IOB2, and IOBES. The text is '*ventral hernia is treated with ventral hernia repair which is a surgery*'. Note that '*ventral hernia repair*' and '*surgery*' are treatment NEs, while others are regular words. In I/O, all words forming NEs are tagged with '**I**nside', while other words are tagged with '**O**utside'. In IOB2, the first word of every NE is tagged with '**B**egin', the subsequent words of NEs are tagged with '**I**nside', and the rest of the words are tagged with '**O**utside'. In IOBES, the first word of every multi-word NE is tagged with '**B**egin', a middle word in a multi-word NE is tagged with '**I**nside', the last word in a multi-word NE is tagged with '**E**nd'. In addition, a single-word NE is tagged with '**S**ingle, while other words are tagged with '**O**utside' in IOBES.

Being the first SR technique, IO has a simple representation composed of only two labels '**I**nside' and '**O**utside'. It is not able to recognise the boundary of two or more consecutive classes. As such, IOB1 is introduced to solve the boundary limitation problem of IO. IOB1 tags all the words in a class as '**I**nside', except the first word of a class that directly follows words having the same class, which is tagged as '**B**egin'. Words that do not belong to a class in IOB1 are tagged as '**O**utside'.