Computer Methods and Programs in Biomedicine 140 (2017) 53–59

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Detecting negation and scope in Chinese clinical notes using character and word embedding

Tian Kang[a], Shaodian Zhang[a], Nanfang Xu[b], Dong Wen[c], Xingting Zhang[c], Jianbo Lei[c,d,*]

[a] Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA
[b] Department of orthopedic surgery, Peking University Third Hospital, Beijing, China
[c] Center for Medical Informatics, Peking University, Beijing, China
[d] School of Medical Informatics and Engineering, Southwest Medical University, Luzhou city, Sichuan Province, PR. China

ARTICLE INFO

Article history:
Received 10 August 2016
Revised 4 November 2016
Accepted 22 November 2016

Keywords:
Chinese natural language processing
Negation detection
Clinical notes
Clinical natural language processing
Word embedding

ABSTRACT

Background and objectives: Researchers have developed effective methods to index free-text clinical notes into structured database, in which negation detection is a critical but challenging step. In Chinese clinical records, negation detection is particularly challenging because it may depend on upstream Chinese information processing components such as word segmentation [1]. Traditionally, negation detection was carried out mostly using rule-based methods, whose comprehensiveness and portability were usually limited. Our objectives in this paper are to: 1) Construct a large Chinese clinical notes corpus with negation annotated; 2) develop a negation detection tool for Chinese clinical notes; 3) evaluate the performance of character and word embedding features in Chinese clinical natural language processing.

Methods: In this paper, we construct a Chinese clinical corpus consisting of admission and discharge summaries, and propose sequence labeling based systems for negation and scope detection. Our systems rely on features from bag of characters, bag of words, character embedding and word embedding. For scopes, we introduce an additional feature to handle nested scopes with multiple negations.

Results: The two annotators reached an agreement of 0.79 measured by Kappa in manual annotation. In cue detection, our systems are able to achieve a performance as high as 99.0% measured by F score, which significantly outperform its rule-based counterpart (79% F). The best system uses word embedding as features, which yields precision of 99.0% and recall of 99.1%. In scope detection, our system is able to achieve a performance of 94.6% measured by F score.

Conclusions: Our study provides a state-of-the-art negation-detecting tool for Chinese clinical free-text notes; Experimental results demonstrate that word embedding is effective in identifying negations, and that nested scopes can be identified effectively by our method.

1. Introdcution

1.1. Background

An increasing amount of computerized clinical data is becoming available with the adoption of electronic medical records (EMR). Natural language processing (NLP) and information extraction (IE) techniques have been critical parts of the pipeline to automate the EMR data mining and knowledge discovery, and have become an active research field in biomedical informatics. For example, researchers have made significant progress on clinical named entity recognition (NER), which is to detect the boundaries and identify the categories of clinical entities, and map them to concepts in standardized terminologies [2–4].

One of the issues associated with clinical information extraction is that systems, without manual interference, cannot generally discriminate information that is present and negated. In clinical notes, presence of a term does not necessarily indicate presence of a medical condition. On the contrary, negation can be used to describe the absence of a finding or disease to rule out a suspected possibility, which is referred to as “pertinent negatives” in clinical reports [5]. For example, in “the patient has no prior history of head trauma”, an NER system may identify the concept “history of head trauma” or simply “head trauma”, but the concept alone will lead to a completely opposite clinical conclusion without detecting the negation cue “no”. Thus, accurately identifying whether a particular expression is negated is of great significance to decision making of diagnosis and prescription. This requires detecting not only cue of the negation, “no”, but also the scope of the negation,

* Corresponding author. Fax: +86 (10) 8280 5900.
E-mail address: jblei@hsc.pku.edu.cn (J. Lei).

http://dx.doi.org/10.1016/j.cmpb.2016.11.009
0169-2607/

i.e. statement or fact that is negated ("history of head trauma", in this example). From an information extraction point of view it will also affect the performance of downstream applications such as knowledge discovery and decision support.

Like English NLP, the pipeline of Chinese NLP includes part-of-speech tagging, named entity recognition (NER), syntactic parsing etc.. But unlike English, in which words are naturally delimited by spaces, sentences are represented as strings of Chinese characters without similar natural delimiters in Chinese [1,10]. Therefore, typically the first step in a Chinese language processing is to identify the sequence of words in a sentence, namely, "word segmentation" [11]. The need for word segmentation, which also exists in other languages like Arabic, Hebrew, and Japanese, creates additional challenges for automated NLP pipelines including named-entity recognition and other components (e.g., parsing, information extraction, etc.). What makes Chinese clinical NLP even more difficult is that existing word segmentation tools are usually trained on news text [11], which is dramatically different from clinical narratives. As such, a new pipeline for information extraction for Chinese clinical notes would require first building up a robust, domain specific word segmenter.

Negation is an ubiquitous phenomena in natural language. The study on negation identification in English has witnessed a significant boost. Most existing methods on negation detection are developed for English text. However, in contrast to a wide range of systems for English clinical notes, research in Chinese clinical NLP is relatively limited. In recent years, hospitals in China have been rapidly implementing and deploying EMR systems, which generate a great amount of unstructured clinical data. There is a strong need of clinical NLP methods for Chinese clinical notes, which enable reusing the large amount of collected EMR data to facilitate patient care and medical research. Efforts have been made in the research community to construct NLP components for Chinese clinical notes [6–9], but to our best knowledge there is no established system that identifies clinical negations.

In this paper, we construct a Chinese clinical corpus consisting of admission and discharge summaries with negations annotated by medical experts. Based on the corpus, we propose sequence labeling based systems for negation detection and scope identification, primarily relying on four groups of features: bag of character, bag of words, character embedding and word embedding. A word segmenter trained on clinical notes is applied before generating word-level features: bag of words and word embedding. For scope identification, we add an additional feature to handle nested negation, which cannot be directly solved by sequence labeling models. We compared our method with traditional rule-based approaches and demonstrated the effectiveness of machine learning-based method and word embedding features.

1.2. Related work

Previous studies on negation identification primarily focus on two sub-tasks: 1) cue detection, which is to identify the specific terms indicating negations; and 2) scope resolution, which aims at determining the linguistic coverage of the impact of a cue in the sentence [1].

Rule-based approaches. Whereas negation identification has been much explored in natural language processing, early systems usually rely on handcrafted rules based on grammatical patterns and keyword matching. Previous works for English text suggest that a small set of words cover a large portion of negation cues [5,12]. Morante et al. presented a list of negation cues in English texts, e.g. "exclude", "cannot", "neither…nor", and the description of their scopes in biomedical texts based on BioScope corpus [12]. Content negators, which typically are verbs such as "hampered", "lacked", "denied", etc. were also taken into consideration in applications like sentiment analysis [13,14]. Sometimes, soft negators such as adverbs "hardly", and "rarely" are treated as negation cues as well [15]. Without a general purpose corpus annotating the precise scope of negation, many studies incorporate negation terms through heuristics or soft-constraints in statistical models [16]. A study by Nakagawa et al. developed a semi-supervised model for sub-sentential sentiment analysis that predicts polarity based on the interactions between nodes in dependency graphs, which potentially can induce the scope of negation [17].

Detecting negation in clinical notes is of much significance for a wide range of applications, especially pertinent negatives analysis. Similar to negation detection in the general domain, rule-based approaches are widely applied in identifying clinical negations before the emergence of shared linguistic corpora. Terms such as "no", "not", "without", "denied" are the most frequent terms to indicate the absence of clinical observations [18]. Mutalik et al. developed a general-purpose negation detecting algorithm named Negfinder [19], using a left-to-right, rightmost-derivation parser to detect negations in surgical notes and discharge summaries. It achieved a sensitivity of 95.7% and a specificity of 91.8%. However it is unable to detect negation cues and corresponding scopes that are far away from each other. The current version of NegEx, developed by Mitchell et al. [20], utilizes 272 rules, a regular expressions, and has proved its effectiveness in detecting negations in discharge summaries with a recall of 95.93% and precision of 93.27%. Like other rule-based algorithms, NegEx has an advantage of being simple to implement but show a lower performance when applied to clinical texts in other domains. The limited portability largely results from inadequate coverage of negation phrases.

Machine learning based approaches. Since 2000, negation detection using machine-learning methods from biomedical text, especially scientific literatures, flourishes thanks to the emergence of shared linguistic corpora and pioneer works [21,22]. BioScope corpus, in particular, builds up a benchmark for speculation detection and negation detection, and has been used by following works [23,24]. This corpus consists of text from both biomedical literature and clinical notes, and part of the corpus is adopted by CoNLL-2010 shared task. Utilizing manually labeled corpora, machine learning-based approaches such as Support Vector Machine (SVM) [23], Conditional Random Field(CRF) [25], have show a success in detecting negations and scopes in biomedical field. For example, Patrick et al. developed a CRF model by adopting rich features and achieved more than 92% F-score on the "absent" category in the task [26].

Negation detection in Chinese text. In contrast to advances in English NLP, much fewer works have been carried out for Chinese negation detection, especially in the clinical domain. Chen et al. have applied a supervised machine learning method with CRF to detect negative and speculative information in scientific literature, using character-based and word-based framework, as well as the combination of features [27]. Their detecting system for negations achieves 94.70% of accuracy and the best performance attained by the system uses combination of features. For clinical notes, Zhang et al. developed an algorithm to detect negative expression in Chinese EMR combining rules with word co-occurrences, and achieved a predictive value of 99.85% [28]. Specificity of the system is not reported, though. Since there is no shared corpus and annotation protocol, it's hard to compare performance of existing Chinese clinical NLP systems.

2. Materials and methods

2.1. Data set and annotation

One month (March 2011) of admission notes and discharge summaries (36,828 notes in total) were collected from the EMR