Mixture of Language Models Utilization in Score-Based Sentiment Classification on Clinical Narratives

Dang Tran Thai(1450207)

School of Knowledge Science
Japan Advanced Institute of Science and Technology

September 2016

Keywords: Electronic Medical Records, clinical text, sentiment classification, linear combination, language models, negation, short text.

In recent years, the emergence of Electronic Medical Records (EMRs) opens an opportunity to improve the quality of healthcare and reduce medical cost. EMRs are well known as digitized medical records mostly created by doctors and nurses in hospital. EMR data is a rich and valuable resource including various data types such as digitized images, laboratory test, clinical text, in which the clinical text that contains information of patient health status such as symptoms, observations, physician's assessments plays an important role in EMRs exploitation. Therefore, EMRs exploitation mostly refers to clinical text exploitation.

Clinical text exploitation is still in infancy stage and poses a lot of challenges in analyzing and mining. As the clinical text is almost spoken language that is also called informal text, so it makes processing and mining on such kind of text become more challenging. The challenges can come from several features of this text such as: ungrammatical text and disambiguation of abbreviations; the shortness of text that makes classification become ineffective; implicit and vague expressions make sentiment analysis on clinical text to evaluate patient health status become more challenging; the text is strongly related to time. Moreover, we also have to face with the lack of annotated data, lexicon resources for higher level analysis such as information extraction, sentiment analysis, adverse drug reactions detection.

One of perspectives in clinical text exploitation that our study concentrates on is to evaluate patient health status through symptoms, diseases, conditions observed

during the treatment, and doctor's assessments which are noted in the clinical text. Patient health status evaluation can support doctor diagnosis and treatment. Moreover, it can be exploited for researches related to drug usage such as adverse drug reactions detection, drug repositioning. Evaluating the health status means that we have to determine whether observed symptoms, conditions, and physician's assessments are positive or negative. That inspires us to pose a problem of doing sentiment analysis on clinical text to evaluate patient health status.

Sentiment analysis that is also called opinion mining is a study field aiming to build methods for automatically analyzing people's opinions, sentiments, attitudes, emotions towards entities. Sentiment analysis is specified through many concrete problems in which document/sentence-level sentiment classification and aspect-based sentiment analysis are backbone in this study field.

Relying on similar points and different points between patient health status evaluation and the original sentiment analysis, we extend sentiment analysis for medical domain. In our initial study, we focus on sentiment classification on clinical text at sentence level. Doing sentiment classification on clinical text is significantly different from on normal text due to some specific features of the clinical text. In order to solve this problem, we have to face with four main challenges which are lack of domain-specific sentiment lexicon resources, implicit sentiment, various forms of negation, shortness of text.

In this thesis, we present our study of using a mixture of language models for score-based sentiment classification on clinical narratives. Our proposed method is a score-based classification method that can deal with the lack of sentiment lexicon resources, the variety of negation forms in clinical text, the shortness of text. The key idea is to use a linear combination of terms extracted from different language models to estimate an overall sentiment score of a sentence. Additionally, through using the linear combination, we derive a new vector representation called language-model-based representation that can help classification method work more effective on short text.

In conclusion, our study aims to build a groundwork for sentiment analysis on clinical text. We start with a backbone problem in sentiment analysis – sentiment classification. In order to effectively do sentiment classification on clinical text, we propose a score-based classification method that can deal with several challenges in this problem. This study initially reaches our proposed objectives.