Extracting Clinical entities and their assertions from Chinese Electronic Medical Records Based on Machine Learning

Jianhong Wang[1, a], Yousong Peng[*, 1, b], Bin Liu[1, c], Zhiqiang Wu[1, d], Lizong Deng[2, 3, e], and Taijiao Jiang[*, 1, 2, 3, f]

[1]College of Computer Science and Electronic Engineering,Hunan University,Changsha,China

[2]Center of System Medicine, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

[3]Center of Systems Medicine,Chinese Academy of Medical Sciences,SuZhou,China

[a]jianhong@hnu.edu.cn, [b]pys2013@hnu.edu.cn, [c]yjygalb@hnu.edu.cn, [d]cxiaodiao@hnu.edu.cn, [e]denglizong@hotmail.com, [f]taijiao@moon.ibp.ac.cn, [*]Corresponding author

Keywords: Chinese Electronic Medical Records, Information extraction, Named Entity Recognition, assertion classification, Machine Learning

Abstract. With the rapid growth of electronic medical records (EMRs) in China, large amounts of clinical data have been accumulated. However, limited work for extracting information from EMRs in Chinese has been conducted. In this work, using manually annotated dataset of EMRs in Chinese, we investigated the clinical Named Entities Recognition (NER) based on Conditional Random Field (CRF) and further built a Support Vector Machine (SVM) classifier to determine their assertion status and evaluate the contributions of different features for assertion classification. For Chinese clinical NER, our CRF-based classifier achieved the best F-measure of 89.07%, while the SVM-based assertion classifier achieved a maximum F-measure of 94.10%. Our work suggests that machine learning methods are helpful in NER and assertion determination for Chinese medical clinical records.

Introduction

Electronic Medical Records (EMRs), generated in the process of clinical treatments, refer to systematized collections of patients' clinical information stored in electronic medical records systems [1]. EMRs contain a range of data, including demographics, medical history, medication and allergies, immunization status, laboratory test results, radiology images, vital signs, personal statistics like age and weight, and billing information. A large amount of medical knowledge, closely related to patients, can be discovered through analyzing these medical records [2].

Generally, EMRs are always described in the form of natural language, from which mining patient-related health-care medical knowledge needs applications of information extraction related technologies. NER and assertion classification are the fundamental issues in information extraction of EMRs [2], which not only serves for Clinical Decision Support (CDS) research in medical informatics, but also supports for the user's healthy state modeling and personalized healthcare information services in Consumer Health Informatics. A number of information extraction studies on EMRs in English have been conducted. In 2010, the Center of Informatics for Integrating Biology and the Bedside (i2b2) at Partners Health Care System and Veterans Affairs (VA) Salt Lake City Health Care System organized a challenge in natural-language processing (NLP) for clinical data. Two of the main tasks for the challenge are extracting clinical concepts from natural language text and classifying modification (assertion) made about medical problems [3]. Most clinical NER systems achieved good performance using CRF algorithm for modeling[4-8], while SVM algorithm gets more attention for assertion classification [4-6,9,10].

With the rapid growth of EMRs in China, large amounts of clinical data have been accumulated. However, limited work for extracting information from EMRs in Chinese has been conducted. Lei et al. [11] compared different machine learning algorithms and various types of features for NER in

Chinese admission notes. Recently, Wu et al. [14] designed a deep learning based NLP systems that could automatically learn useful feature representations from unlabeled corpora through unsupervised learning in Chinese clinical text. Most of work focused on clinical NER [11-14]. Besides, there is a lack of publicly available EMRs in Chinese for medical information extraction research. In this study, we manually annotated a corpus of Chinese typical EMRs and described a machine-learning based system to extract the clinical entities and their assertions. Furthermore, we investigated the contributions of different types of features for assertion classifiers. To the best of our knowledge, this is one of the first extensive studies for Chinese medical entity assertion classification.

Datasets and Annotation

A corpus of typical EMRs, provided by one Chinese electronic medical records system online (http://www.easyemr.com.cn/activitys/activity-list.aspx?activity_id=32), were collected and collated. This corpus contains 200 internal medicine records. The information which each record includes is shown in Fig. 1.


主诉（chief complaint）：上腹胀痛3天，加重1天。
现病史(history of present illness)：3天前，由于在外地打工，饮酒与进油腻食物后，自觉上腹胀，无恶心呕吐，未在意，仍坚持日常劳作。但腹胀不减轻，有加重之趋势，而且疼痛，喜按，到当地医院就诊为“胰腺炎”，未注，而返回当地医院住院治疗。病来无发热、腹泻；饮食差、二便、睡眠正常。
既往史(past history)：既往体健，无“肝炎、结核、伤寒、疟疾”等传染病史；无手术、外伤及输血史；无药物过敏史。
个人史（Personal history）：生于原籍，久居当地，未到过血吸虫病流行区和牧区。无不良嗜好，无性病冶游史。
婚育史（obstetric history）：
月经史(menstrual history)：
家族史(family history)：否认家族中遗传病及传染病史。
体 格 检 查(physical examination)
体温:T37.0℃；脉搏：80次/分；呼吸：18次/分；血压：120/70mmHg
一般情况：发育正常，营养中等；步入病房，神志清楚，语言流利，痛苦面容；平卧位，查体合作。皮肤粘膜：全身皮肤无黄染，未见出血点、瘀斑及皮疹。未见蜘蛛痣。浅表淋巴结：表浅淋巴结未触及肿大。头部及其器官：头颅无畸形，发布均匀；额纹对称，眼裂对称，双侧眼睑无浮肿，结膜无充血，巩膜无黄染，双侧瞳孔等大正圆，对光反射灵敏，未见眼震。耳鼻无异常分泌物，鼻唇沟对称；咽部未见充血，双侧乳突区及副鼻窦区无压痛，口唇无发绀；伸舌居中，悬雍垂居中，咽部无充血，扁桃体不大。颈部：颈软，双侧对称，无颈静脉怒张及异常搏动，气管居中，甲状腺无肿大。胸部：对称无畸形，双侧呼吸动度均匀，节律整，触觉语颤正常，叩诊呈清音，双肺呼吸音清，未闻及干湿罗音。心前区无隆起，心尖搏动无弥散，未触及震颤，叩诊心界不大，心率80次/分，律齐，心音有力，各瓣膜听诊区未闻及病理性杂音。腹部：平坦，无腹壁静脉曲张，未见胃肠型及蠕动波；腹软，剑下及左上腹部压之不适，未扪及包块，无压痛、反跳痛及肌紧张，肝脾肋下未触及；腹部叩诊正常，无移动性浊音。肛门和外生殖器：未查。脊柱和四肢：无畸形，各关节无红肿、压痛，活动自如。双下肢无水肿。
专 科 检 查(specialist examination)
辅 助 检 查(laboratory examination)
2011-05-14 血常规 WBC10.6×109/L RBC2.87×1012/L HGB98g/L PLT418×109/L GRA 69.6%。
尿常规 OB X线一胸腹透：未见异常。
血生化 K＋3.5 mmol/L；Na+145mmol/L；Cl－101mmol/L；Ca＋＋2.34mmol/L；BUN5.46mmol/L；Cr161umol/L；BS5.7mmol/L； 血 AMY368 尿 AMY1496 u/L
超声 胆囊增大 胰管扩张 腹部CT 肝肾隐窝可见水样密度影；胆囊增大，壁不厚；胰腺胰头增大。
鉴 别 诊 断(differential diagnosis)
初 步 诊 断(initial diagnosis)
1，急性胰腺炎
2，胆囊炎
诊断依据：①年轻男性；②饮酒与食用油腻食物后出现上腹部胀、腹痛；③上腹部饱满，剑下及左上腹部压之不适，无肌紧张及反跳痛；④WBC10.6×109/L；腹部CT所示，尿淀粉酶1496。
诊 疗 计 划(treatment plan)
1， 首先完善相关检查，如离子、肾功、血糖、血脂、血尿淀粉酶；再行胰腺超声或CT，以明确诊断；
2， 予内科Ⅰ级护理。抑制胰腺分泌——应用奥曲肽；抗炎、补液，维持水、电解质平衡。对症支持。

Fig. 1 A sample of a typical Chinese EMRs

Recently, under the guidance of several professional doctors, Yang et al. developed an annotation guideline (http://wi.hit.edu.cn/dev/YuLiao/NER.pdf) for Chinese EMRs according to that of i2b2 [3]. The annotation guideline defines six types of entities, including disease, disease type, complaint-symptom, test-result, test and treatment. For concepts of disease, complaint-symptom, test-result and treatment, eight assertions were further defined: present, absent, possible, conditional, hypothetical, family, history and occasional. Following the above annotation guidelines, two post-graduates of computer science independently marked all the entities and their assertions in our collated EMRs. To calculate the inter-rater agreement for annotation, 40 records were annotated by both annotators. The remaining 160 records were annotated by a single annotator only. The initial agreement between two annotators on 40 records as measured by kappa [15] was 0.926, which indicates the annotation is reliable.

Methods

CRF-based approaches for clinical NER. Chinese medical NER task requires determination of the boundaries of clinical entities and assignment of their entity types mentioned above. We converted this NER task to a classification task, for which the CRF algorithm was used (http://crfpp.googlecode.com/svn/trunk/doc/index.html). Firstly, we transformed the annotated data into the 'BIESO' format, in which each word was assigned into a label as follows: B=beginning of an entity, I=inside an entity, E=end of an entity S=single word entity and O=outside of an entity. Because the entity type 'disease type' occurred rarely in our collated EMRs, it would be not used in our analyses. Twenty-one tags in total were generated: B-disease, B-complaintsymptom, B-testresult, B-test, B-treatment, I-disease, I-complaintsymptom, I-testresult, I-test, I-treatment, E-disease, E-complaintsymptom, E-testresult, E-test, E-treatment, S-disease, S-complaintsymptom, S-testresult, S-test, S-treatment, and O. Fig. 2 shows a sentence of annotated entities labeled with BIESO tags.

```
#Sentence 1              上腹部疼痛伴恶心4天。

BIESO representation     上/B-complaintsymptom 腹/I-complaintsymptom 部/I-complaintsymptom
                         疼/I-complaintsymptom 痛/E-complaintsymptom
                         伴/O 恶B-complaintsymptom 心/E-complaintsymptom 4/O 天/O。
```

Fig. 2 A sample of Chinese medical NER represented in BIESO format.

Then, we defined the following four types of features to build the CRF classifier for NER task:

1) Single word features: individual Chinese character, punctuation, English alphabet and number.

2) Part of Speech (POS): We use NLRIR (http://ictclas.nlpir.org/), developed by Dr. Zhang Huaping, to get POS tags.

3) Dictionary features: We integrated ICD-10 and Chinese EMRs common terms which contain 3000 phrases into NLPIR. After analyzing concepts of diseases and symptoms appeared in Chinese EMRs, we discovered that most concepts consist of a body part or a subject, such as "体重(weight)", and a basic disease or a symptom such as "减轻 (loss)". Then, a dictionary containing body parts, subjects and basic diseases and symptoms was further built for our CRF-based NER.

4) Section headings: An EMR is a semi-structured narrative text divided into many sections such as chief complaint, present illness and so on. Different sections record different clinical data. For example, the physical examination has concepts of test and abnormal test results, while the section of preliminary diagnosis contains disease concepts. Therefore, we manually reviewed some records and defined 12 different section headings as section headings feature.

SVM-based approaches for assertion classifier. Assertion classification is to assign one of the eight possible assertion labels mentioned above to the entity type of disease, complaint-symptom, test-result, test and treatment. The SVM algorithm was used to conduct the classification, which was implemented using libsvm (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) with a linear kernel. Four types of features were used as follows:

1) Context feature: a$\pm$4 word window, i.e., words that appear within t$\pm$4 word window of the target concept. Given the concept at the nth position in the sentence, the $\pm$4 word window captures the words found in the (t-1)th, (t-2)th, (t-3)th, (t-4)th,(t+1)th, (t+2)th, (t+3)th, and (t+4)th positions in the sentence;

2) Section headings: 12 different section headings were used as mentioned above.

3) Cues information feature: before or after a target concept, some word cues can indicate the target concept is present or absent. For example, in the sentence of "双侧眼睑无浮肿(no bilateral eyelid edema)", "浮肿(edema)" is a target concept, and "无(no)" before "浮肿(edema)" indicates "浮肿(edema)" is absent for patient. Then, we developed a system to automatically extract these cues.

4) POS features.

The above features were transformed into binary features. For each feature, the binary feature vector lists all possible values of that feature in the corpus as its columns, and for each target medical problem to be classified (row), it sets the columns observed to 1, leaving the rest at zero. If the target has no value for a feature, then all columns representing this feature will be set to zero.

Experiments and evaluation. The internal medicine typical records mentioned above were divided into two subsets: two-thirds (120 records) for training and one-third (80 records) for test. The parameters of CRF classifier and SVM classifier were optimized using the training set via the five-fold cross-validations. Then we tested the classifiers using the independent test set. The precision, recall and F-measure were used to measure the performance of the classifiers. To evaluate the contributions of different types of features for assertion classification, we started with the context feature firstly, and then added additional types of features and reported corresponding results.

Results and Discussion

Table 1 describes the statistics of entities in the annotated EMRs used in our study. There are 13119 entities in the training dataset and 8026 entities in the test dataset. Distribution of different types of entities' assertions in the training set and test set were shown in Table 2. The proportion of present, absent, possible and history assertions is far larger than other assertion types. In our annotated dataset, the number of hypothetical, conditional and family assertion is lower than 20, so we ignore these assertion types.

Table 1 Distribution of different types of entities in training set and test set

Data set	disease	complaintsymptom	testresult	test	treatment	All
training	1403	1623	5362	2086	845	11319
test	1031	976	3786	1620	613	8026
total	2434	2599	9148	3706	1458	19345

Table 2 Distribution of different types of assertions in training set and test set

Data set	Present	absent	possible	occasional	history	All
training	1921	6146	527	26	175	8795
test	1395	4143	371	20	110	6039
total	3316	10289	898	46	285	14834

The detailed results of CRF classifier for each entity type were shown in Table 3. F-measures ranged from 78.47% to 94.58% among five types of entity. Performance was best for the type “test” and worst for the type “treatment”. For each type of entity, precision was higher than recall.

Table 3 Result of CRF-based named entity recognition for each entity

Entity type	Precision	recall	F-measure
disease	90.22%	82.66%	86.27%
complaint-symptom	87.24%	78.40%	82.58%
test-result	94.65%	94.51%	94.58%
test	84.85%	82.77%	83.80%
treatment	92.90%	67.92%	78.47%
overall	91.31%	86.94%	89.07%

Our best performance results are similar to the results in [13] for discharge summary and the NER task in i2b2 2010 challenge [3, 8]. In this study, we found that our CRF model could correctly identify the entity that did not appear in the training set. The reason may mainly lie in the dictionary features used in our study. In Chinese EMRs, most entities have their syntactic structures. For example, the concept of "左侧肢体(left limb)无力(weakness)" is made up of a body part "左侧肢体(left limb)" and a basic symptom "无力(weakness)". In our dictionary, we defined these common body parts and basic symptoms and diseases. Table 4 shows the performance of SVM classifier in assertion determination when using different feature sets. Context feature set was observed to contribute most to assertion classification with the F-measure as high as 92.07%. Both section headings feature and cues information feature improved assertion classifier performance. For example, section headings

feature improved F-measure from 92.07% to 93.14%, while cues information improved F-measure from 92.07% to 93.56%. POS did not further improve assertion classification performance.

Table 4 Results of SVM-based assertion classifier when different sets of features were used

Feature set	F-measure
context feature	92.07%
context feature+section headings	93.14%
context feature+POS	92.18%
context feature+cues information	93.56%
context feature+section headings+POS+cues information	94.10%

Table 5 Detailed results of the best SVM-based Assertion classifier for each entity type

Assertion category	Precision	recall	F-measure
present	95.68%	92.95%	94.30%
absent	96.93%	93.62%	95.25%
occasional	92.65%	89.51%	91.05%
possible	78.86%	63.47%	70.33%
history	81.31%	74.37%	77.69%
overall	95.06%	93.15%	94.10%

The detailed results of the best SVM assertion classifier for each assertion types are shown in Table 5.The overall F-measure is 94.10%.The classifier achieved the best performances for the present and absent assertion, with F-measure equaling to 94.30% and 95.25%, while it performed worst for the possible assertion (F-measure=70.33%). This may be due to the fact that human determination of 'possible' assertions was not always straightforward, as reported by Jiang's work [3].

Conclusion

In this study, we investigated clinical NER from EMRs in Chinese based on CRF algorithm and further built a SVM-based classifier to determine assertion status and evaluate the effects of different features for assertion classification. For the former, our CRF classifier achieved the best F-measure of 89.07%, while for the latter our SVM assertion classifier get a maximum F-measure of 94.10%. Although our experiment achieved a good performance, there are still some limitations. The data used in our study were limited to internal medicine records. In the future, we will cooperate with hospitals to get more medical records to improve our models. Overall, our work suggests machine-learning methods could be helpful in NER and assertion determination for the Chinese EMRs.

Acknowledgements

This study was supported by National Natural Science Foundation (31500126 and 31371338), and the Young Teachers Development Plan of Hunan University to YS (531107040720) and the International Scientific and Technological Cooperation project (2014DFB30010).We would like to thank the members of the Jiang lab for their help and deliberations. We would also like to thank the anonymous reviewers for their valuable comments and suggestions. No competing interests exist.

References

[1] Information on http://www.nhfpc.gov.cn.

[2] Yang Jinfeng,Yu Qiubin, Guan Yi, Jiang ZhiPeng: An Overview of Research on Electronic Medical Record Oriented Named Entity Recognition and Entity Relation Extraction. Acta Automatica Sinica. Vol. 40 (2014), p. 1537-1562.

[3] Özlem Uzuner, South B R, Shen S, Duvall S L: 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc. Vol. 18 (2011), p. 552-556.

[4] Jiang M, Chen Y, Liu M, Rosenbloom ST, Mani S, Denny JC and Xu H: A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. J Am Med Inform Assoc. Vol. 18 (2011), p. 601-606.

[5] Bruijn B D, Cherry C, Kiritchenko S, Martin J and Zhu X: Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. J Am Med Inform Assoc. Vol. 18 (2011), p. 557-62.

[6] Grouin C, Abacha A B, Bernhard D, CARAMBA: concept, assertion, and relation annotation using machine-learning based approaches. In: Proceedings of the 2010 I2B2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA (2010).

[7] Uzuner O, Solti I, Cadag E: Extracting medication information from clinical text. J Am Med Inform Assoc. Vol. 17 (2010), p. 514-518.

[8] Doan S, Collier N, Hua X, Duy P H and Tu M P: Recognition of medication information from discharge summaries using ensembles of classifiers. Bmc Medical Informatics & Decision Making. Vol. 12 (2012), p. 1-10.

[9] Uzuner O, Zhang X, Sibanda T: Machine learning and rule based approaches to assertion classification. J Am Med Inform Assoc. Vol. 16 (2009), p. 109-115.

[10] Clark C, Aberdeen J, Coarr M, Tresner-Kirsch D, Wellner B, Yeh A, Hirschman L: MITRE system for clinical assertion status classification. J Am Med Inform Assoc. Vol. 18 (2011), p. 563-7.

[11] Lei J, Tang B, Lu X, Gao K, Min J, Hua X: A comprehensive study of named entity recognition in Chinese clinical text. J Am Med Inform Assoc. Vol. 21 (2014), p. 808-14.

[12] Wang, Shi Kun, L. I. Shao zi, and T. S. Chen: Recognition of Chinese Medicine Named Entity Based on Condition Random Field. Journal of Xiamen University. Vol. 48 (2009), p. 359-364.

[13] YAN yang, WEN Dun-wei, WANG Yun-ji, WANG Ke: Named entity recognition in Chinese medical records based on cascaded conditional random field. Journal of Jilin University (Engineering and Technology Edition). Vol. 44 (2014), p. 1843-1848.

[14] Wu Y, Jiang M, Lei J, Xu H: Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network. Studies in Health Technology & Informatics. Vol. 216 (2015), p. 624-628.

[15] Hripcsak, George, and A. S. Rothschild: Agreement, the F-Measure, and Reliability in Information Retrieval. J Am Med Inform Assoc. Vol. 12 (2005), p. 296-298.