Towards Automated Knowledge Discovery of Hepatocellular Carcinoma: Extract Patient Information from Chinese Clinical Reports

Hongmei Yang
Sun Yat-Sen University,
Department of Biomedical Engineering,
No.74 Zhongshan Rd.2, Guangzhou, P.R. China
homayyoung@163.com

Lin Li
XinJiang Medical University,
Department of Public Health,
No.339 Xin Yi Road, Urumqi, P.R. China
lilin10001@126.com

Ridong Yang
Sun Yat-Sen University,
Department of Biomedical Engineering,
No.74 Zhongshan Rd.2, Guangzhou, P.R. China
yangridong9386@foxmail.com

Yi Zhou*
Sun Yat-Sen University,
Department of Biomedical Engineering,
No.74 Zhongshan Rd.2, Guangzhou, P.R. China
zhouyi@mail.sysu.edu.cn

ABSTRACT

Objectives: To accurately determine significant prognostic risk factors, patient information must be quantified accurately according to their extent of disease. An essential step for prediction of prognostic risk factors requires the determination of patient features which are typically hidden in electronic medical record(EMR). The goal of this study is to extract clinical entities of Chinese clinical reports, enabling automated hepatocellular carcinoma knowledge extraction.

Materials and Methods: In this paper, we annotated hepatocellular carcinoma corpora with patient records from EMR database. We present an information extraction solution based on assembled method. Our evaluation dataset contains 3996 training sentences and 1570 test sentences. The evaluation metrics are precision, recall, F1 of extract matching.

Results and Conclusions: NER of admission reports, radiology reports and discharge summaries with F1 of 0.8449, 0.5935 and 0.7320 respectively. RE of overall F1 is 0.9129. This study prepares a foundation for larger population studies to identify clinical features of hepatocellular carcinoma.

CCS Concepts

• **Information systems→Information retrieval.**

Keywords

EMR; information extraction; machine learning; NER.

1. INTRODUCTION

Hepatocellular carcinoma (HCC) is one of the most common malignancies worldwide, especially in east Asia and China. The clinical features and risk factors of poor prognosis of HCC have not been fully evaluated. To accurately determine significant prognostic risk factors given a specific set of HCC patient characteristics, patients must be quantified accurately according to their extent of disease. An important part of this requires determining patient characteristics such as the disorder, medical history, image features of tumors, pathology features of tumors, number of tumors, size of tumors, degree of tumor spread, etc., all of which typically hidden in electronic medical record(EMR) as free text [1-3], which are easily processed and perceived by humans, but hinder automatic machine understanding and consumer using. Information extraction(IE) methods offer an automated means of extracting free text, with the advantage of being scalable to volumes of historical data [4, 5]. In clinical domain, IE systems were widely used to identify clinical syndromes and common biomedical concepts from radiology reports, discharge summaries, problem lists, nursing documentation, and medical education documents [6, 7].

In this paper, we describe our annotated corpora and IE system which extracts entities and relations of patient information from free text EMR, including admission reports, radiology reports, discharge summaries. This work prepares a foundation for granular classification of hepatocellular carcinoma patients, facilitating automated methods of quantifying and cohort identification.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
ICMHI 2018, June 8–10, 2018, Tsukuba, Japan
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6389-1/18/06...$15.00
https://doi.org/10.1145/3239438.3239445

2. RELATED WORKS

Previous work considered the use of information extraction to facilitate the secondary use of EMR data, which automatically extracts and encodes clinical information from text. In recent years, machine learning approaches are popular-investigated in biomedical language processing community.

In the past years, several models and approaches had been proposed for the recognition of semantics types related to biomedical named entity recognition [10]. Generally, all the previous systems fell into three categories: rule-based approaches [11], dictionary-based approaches. And more recently, machine learning approaches were more investigated in biomedical NER community[1-7]. Rule-based method required domain-specific knowledge and large amounts of rules. Machine learning based algorithms considered biomedical NER as a sequence labeling problem which aimed at finding the best label sequence [12-18].

Most state-of-the-art NLP systems used machine learning approaches, which required large amounts of hand-crafted features and domain-specific knowledge to achieve high performance [4-6, 9, 10]. The top-ranked systems in 2017 CCKS CNER Challenge were primarily based on recurrent neural networks [11-13]. Hu, et al.[14] used a hybrid approach based on rule, CRF (conditional random fields) and RNN (recurrent neural network) methods for the CNER task.

We are interested in extracting meaningful terms regardless of whether they are existing terminologies or not, and relations between these entities. Many character-based Chinese NERs studies demonstrated their effectiveness by deep learning methods, with advantage of avoiding error propagation and hand-crafted features. Based on previous research, our system applies deep neural networks assembled rules on NER tasks. For the RE task, we train a classifier based on classification and regression tree (CART) method.

3. METHOD

3.1 Overall Flowchart

The annotated corpora are used to train and test our extraction system. Fig.1 presents the overall flowchart. A sentences identification module identified sentences using regular, e.g. "(。|\\n)". Based on the analysis of our corpora, we found that 99.9% of relations were from entities on the same line, Relations were identified using a direct classification of enumerated pairwise entity to entity candidate relations.

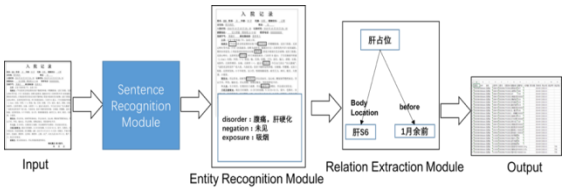

Figure 1. Overall flowchart: A report is first processed to identify sentences, before entities and relations are identified and assembled into templates.

3.2 Entity Recognition

In this paper, we use BiLSTM-CRF (bidirectional LSTM with CRF) assembles with rules to build NER models. BiLSTM have advantages of making full use of previous context and taking future information into consideration. More specifically, our models have a forward LSTM layer and a backward LSTM layer. Figure 2 presents character-level BiLSTM-CRF with the clinical sentence "既往有高血压" (There was high blood pressure in the past). Given a sentence $X = (X_1, X_2, X_3, \dots, X_T)$, for each word X_t there is a word vector e_t to represent it. The previous k words were set as a window, and features of X_t are represented as $x_t = < e_{t-k}, e_{t-k+1}, e_{i-k+2}, \dots, e_{t-1}, e_t >$. Where $[x]_1^T = (x_1, x_2, x_3, \dots, x_T)$ is an input sequence of BiLSTM. The output of BiLSTM is defined as logits:

$$\text{logits} = \sigma([h]_1^T) \quad (1)$$

where $[h]_1^T$ is concatenation of forward LSTM output $[h_f]_1^T$ and backward LSTM output $[h_b]_1^T$. And σ is a softmax function. In CRF algorithm, we define a state transition matrix $T_{i,j}$ to predict the current tags, where $T_{i,j}$ is score of jumping from tag i to tag j. The score of true tag sequence is defined as the sum of network and transition scores:

$$S([x]_1^T, [O]_1^T, \theta, T,) = \sum_{t=1}^{T}(\sigma(h_t) + T_{t-1,t}) \quad (2)$$

where $[O]_1^T$ is true tag sequence. The training procedure focuses on adjusting the network parameters θ and T of CRF model λ in order to minimize loss function, and the loss function can be written as:

$$\mathcal{L}(\theta, T) = S([x]_1^T, [O]_1^T, \theta, T) - \log \sum_{\forall [j]_1^T} e^{S([x]_1^T, [j]_1^T, \theta, T,)} \quad (3)$$

where log-likelihood is obtained by normalizing the above score over all possible tag-sequences $[j]_1^T$. Finally, we applied Adam optimizer to minimize loss value. The Adam is based on adaptive estimates of lower-order moments, which is computationally efficient, has little memory requirements[15]. Most [16-18]of the training parameters follow previous study[16-19], which show effectiveness. Both LSTM layers have a basic LSTM layer and a dropout layer, hidden dimensions are 100, window size is 7, we run 105 epochs and save the model which achieved best on validation dataset.

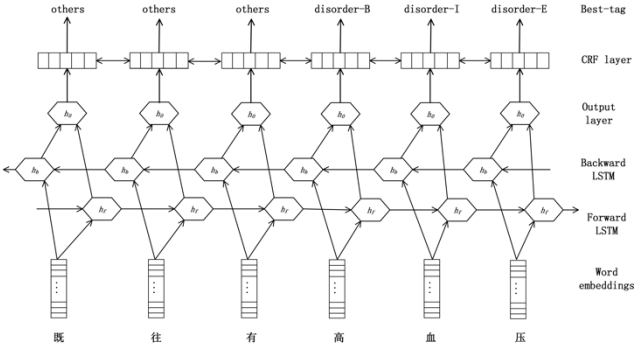

Figure 2. Graphical illustration of character-level bidirectional LSTMs with the clinical sentence "既往有高血压" (There was high blood pressure in the past).

Table 1. Describe patterns of time and size, "/" represent "or". And we use regular expressions to extract entity "time" and entity "size".

Patterns of time	1. yyyy/yy 年(Year)MM/M 月 (Month)dd/d 日(day); 2. yyyy/yy 年(Year); 3. yyyy/yy 年(Year)MM/M 月(Month); 4. yyyy/yy-MM/MM-dd/d; 5. yyyy/yy-MM/MM; 6. yyyy/yy.MM/M.dd; 7. yyyy/yy.MM; 8. n 天/日(n days); 9. n 年(n years); 10. n 个月(n months); 11. n 周(n week); 12. n 余天(more than n days); 13. n 余年(more than n years); 14. n 月又余(more than n months); 15. 半年(one half year); 16. 半月(one half month).

Patterns of size	1. 大小约(Size of about)l cm/mm ×w cm/mm ×h cm/mm; 2. 大小约(Size of about)l cm/mm ×w cm/mm; 3. 直径(Diameter)/内径(ID)/长轴(Long axis)n cm; 4. 范围(Range) l cm/mm ×w cm/mm.

We pre-trained character vectors with our closed dataset based on the skip-gram by word2vec tools, and most of the hyper parameters were set as Mikolov et al.(2013) [20], excepts vector dimension using 128 and minimize count using 5. And then we take character vectors as LTSM network input.

At Inference time, we use Viterbi algorithm to find the best tag sequence that maximizes path scores[20].

Based on analysis of our corpora, we found that recognition of some kind of entities achieves better F1 by rules, for instance, entity "time" and "size". Thus, we summarized patterns of entity "time" and "size" (They are described in detail in Table 3) and use regular expressions to extract them. Except for entity "time" and "size", recognition of other entities is by BiLSTM-CRF method.

3.3 Relation Extraction

Once sentences were identified with entities, they were run through relation extractor, all possible pairwise relations between entities were enumerated and classified based on CART method implemented by scikit-learn toolkit. Our features are related to the entities involved, the distance between two entities, the words around them. They are described in detail in Table 2.

Table 2. Description of relation extraction features.

Feature	Feature Description
Entity-1 type	The type head entity.
Entity-2 type	The type of tail entity.
Entity-1 word vector	Word vector of Entity-1.
Entity-1 word vector	Word vector of Entity-2.
Entity-1 start position	The number of characters from the beginning of the document to the first character of Entity-1.
Entity-2 start position	The number of characters from the beginning of the document to the first character of Entity-2.
Entity-1 end position	The number of characters from the beginning of the document to the first character of Entity-1.
Entity-2 end position	The number of characters from the beginning of the document to the last character of Entity-2.
Distance	The number of characters from the last character of Entity-1 to the first character of Entity-1.

3.4 Annotated Corpora

To build our system, we annotated clinical corpora (including admission reports, radiology reports and discharge summaries) originally from a cohort of 75 HCC patients from the from EMR database of the First Affiliated Hospital, Sun Yet-Sen University in China. An annotation guideline was developed by 2 clinicians and the author of this study. A particular annotation tool—Brat was adopted to record entities in the documents [21]. Figure 3 present Examples of the medical record annotated with entities and relations. To calculate the inter-annotation agreements(IAA) for annotation, 20 reports (377 sentences) were annotated by both annotators, a clinical researcher was in charge of dealing with inconsistency between the two annotations.

Three Chinese clinicians were recruited to annotate 15 types of clinical entities. entities descriptions are presented in Table 3. And relations were defined as directed links between the two entities types, often with either a tumor reference or a measurement as the source, or the head, of the directed relation. Relations are described as Table 4.

Table 3. Summary of annotation entities in corpora.

Entity	Entity Description	Train	Test
disorder	Related diseases, symptom.	5555	1894
negation	A negative word.	3077	1060
uncertainty	Expressed uncertainly words.	229	103
body_location	Body organs or body parts.	4848	1961
negative_symptom	Indicates negative characteristics.	1939	629
test	Including examinations items.	2110	732
test_result	Indicates a numerical test result.	1235	421
image_feature	Describe image features.	1315	820
pathology_feature	Describe pathological characteristics.	117	54
size	Tumor size, the size of the biopsy tissue or the size of the imaging lesion area.	232	115
number	Indicates the number of tumors or nodules.	189	64
drug	Drug treatment.	259	87
operation	Surgical name.	258	121
time	The exact point of time.	723	273
person	That people have a direct relationship with the patient.	64	18

Table 4. Description of annotation relations that we predefined.

Relation	Relation Description
result_of	Relation between test and test result.
size_of	Relation between a tumor reference to a tumor size.
num_of	Relation between a tumor reference to a tumor count.
negation_of	A negation cue, starting from the negation entity to other entity.
body_location_of	Marks in which disorder, test, pathology feature, image feature or operation happened.
time_related	Marks time when disorder, test, pathology feature, image feature or operation happened.

uncertainty_of	An uncertainty cue, starting from the uncertainty entity to other entity.

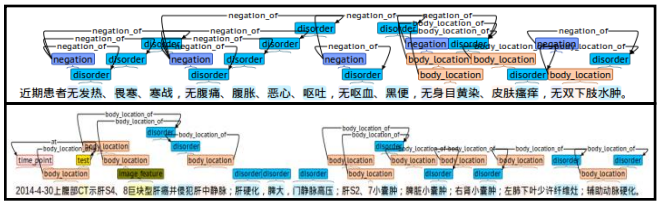

Figure 3. Examples of the medical record annotated with entities and relations.

4. RESULT

Based on inter annotations of 377 sentences, the IAA is 0.8098, which indicates that the annotation is reliable. The standard micro-averaged precision, recall, and F-measure are used to evaluate and gauge the performance of IE system[22]. We developed an evaluation tool to calculate their value, which can be written as:

$$Precision = \frac{True\ positive}{True\ positives + False\ positives} \quad (4)$$

$$Recall = \frac{True\ positive}{True\ positives + False\ negatives} \quad (5)$$

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

For NER task, we train three NER models with datasets of admission reports, discharge summaries and radiology reports respectively. For RE task, one model trained with all datasets. Table 5 presents the evaluation of NER system and RE system. Figure 4a reports the precision, recall and F1 of RE system. Figure 4b, Figure 4c and Figure 4d report F1 of each entity in admission reports, discharge summaries and radiology reports respectively.

Table 5. The result evaluation of our word-level model and character-level model.

Datasets	NER			RE		
	Precision	Recall	F1	Precision	Recall	F1
Admission Reports	0.8616	0.8288	0.8449	—	—	—
Discharge Summary	0.8275	0.6562	0.7320	—	—	—
Radiology Reports	0.6908	0.5201	0.5935	—	—	—
Overall	—	—	—	0.8984	0.9279	0.9129

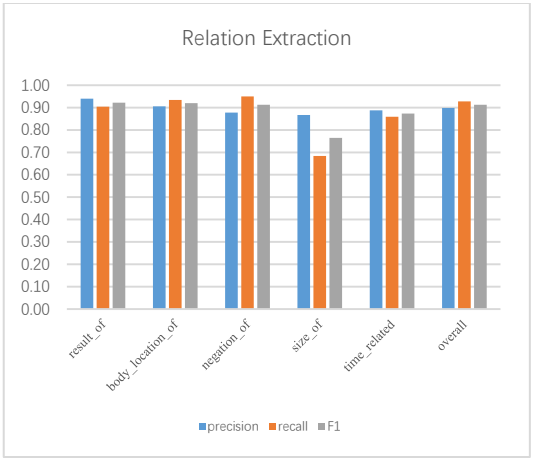

Figure 4a. The overall precision, recall and F1 of each relation.

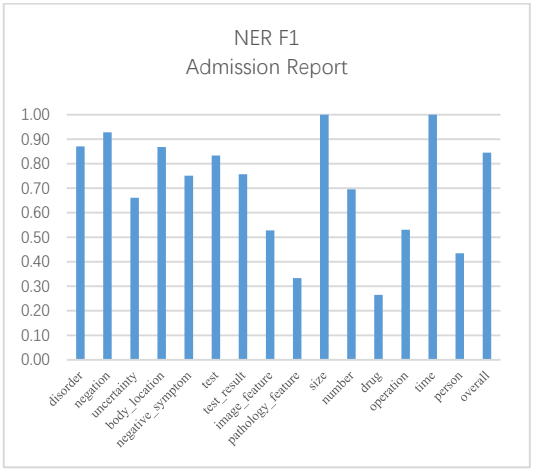

Figure 4b. The F1 score of recognition of each entities in discharge summaries.

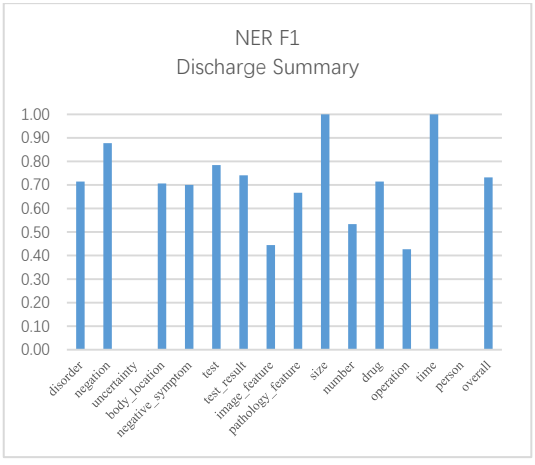

Figure 4c. The F1 score of recognition of each entities in admission reports.

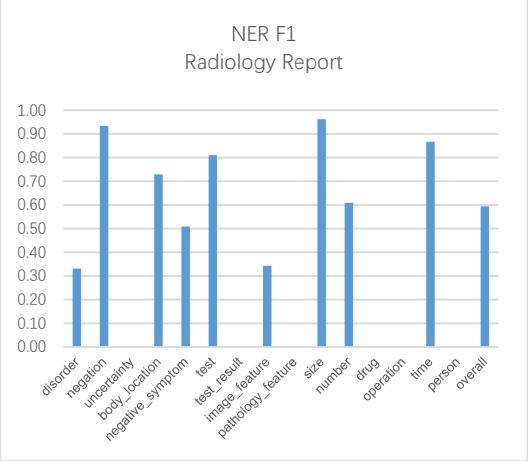

Figure 4d. The F1 score of recognition of each entities in radiology reports.

5. DISCUSSION

In this work, we annotated corpora and built an assembled approach based on deep learning and rules system for NER and RE. A significant hurdle for this information extraction task was complexity of annotation. Each annotator differs in understanding of the granularity and boundaries of the entity words, which result in poor IAA. NLP researchers should get involved as early as possible, such as writing detailed guidelines, calculating IAA and validating annotated data. Besides, they can discuss with physicians what kind of additional information would be beneficial to add.

For NER task, the most errors occurred in long entities with combined structures. For example, in a long disorder entity"肝占位性病变(Liver space occupying lesions)", two parts of it —"肝(Liver)" and "占位性病变(Occupying lesions)" were predicted to be body location and disorder respectively. The proportion of part-predicted in all error cases are presented in Figure 5. Maintaining different granular corpora and information about syntactic structures of Chinese sentences could potentially help in this scenario.

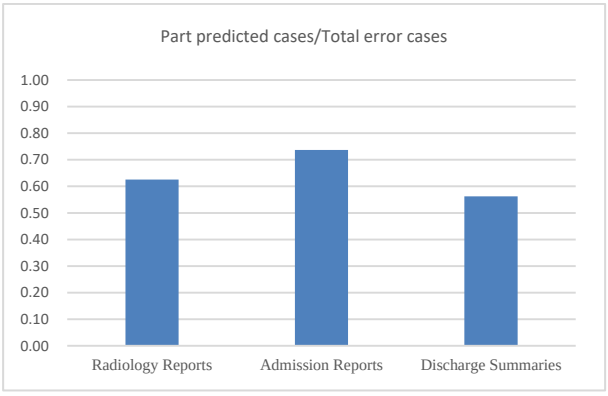

Figure 5. The proportion of part-predicted in all error cases.

Other issues included unbalance and sparsity of some entities and medical abbreviations which occur once in our corpus, especially in person and drug entities. The entity "person" and "uncertainty" hardly occurred in corpus, particular in Discharge Summary and Radiology Reports, the entity "person" and "uncertainty" were hardly predicted.

It is an early stage work of hepatocellular carcinoma knowledge discovery that extracting information from Chinese clinical reports, many practical problems remain to be solved. In our future work, we will improve our system and move towards granular classification of hepatocellular carcinoma patients, facilitating automated methods of quantifying and cohort identification.

6. CONCLUSION

In this work, we annotated hepatocellular carcinoma corpora on Chinese clinical reports, including admission report corpus, radiology reports corpus and discharge reports corpus. The corpora have rich entity types and relation types to represent patient features. Furthermore, we presented assemble methods based system for patient information extraction. Specifically, we used deep learning assembles rules for NER, which achieved F1 of 0.8449, 0.5935 and 0.7320 on admission report, radiology reports and discharge reports respectively. And we used CART for RE with our corpora, which achieved overall F1 of 0.9129. Considering the complexity of our annotation, our performance was very promising. This system provided a solution to extract patients information with hepatocellular carcinoma from Chinese EMR and prepares a foundation for larger population studies to identify clinical features of hepatocellular carcinoma.

7. ACKNOWLEDGEMENTS

I would like to express my gratitude to all those who have helped me during the writing of this paper. The funding was from the Ministry of Science and Technology of the People's Republic of China-National Key R & D Program and Major Project on Precision Medicine Research (No. 2016YFC0901602), Guangzhou City Major Projects of Collaborative Innovation in Health Care (No. 201604020016), Frontier and Key Technology Innovation Project of Guangdong Province in 2015 (Science and Technology Major Project) (No. 2015B010106008), Major Project of Collaborative Innovation in Industry-University-Research Cooperation of Guangzhou City in 2017 (No.201604016136), the NSFC and Guangdong Provincial Center for Big Data Science Joint Fund Project (No. U1611261) and the Major Project of Frontier and Key Technical Innovation of Guangdong Province in 2014 (Science and Technology Major Project) (No. 2014B010118003).

8. REFERENCES

[1] Wang, Y., et al. 2014. *Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: an empirical study*. J Biomed Inform. **47**: p. 91-104. DOI:https://doi.org/ 10.1016/j.jbi.2013.09.008.

[2] Zhang, S. and N. Elhadad. 2013. *Unsupervised biomedical named entity recognition: experiments with clinical and biological texts*. J Biomed Inform. **46**(6): p. 1088-98. DOI:https://doi.org/ 10.1016/j.jbi.2013.08.004.

[3] Han, L.F., D.F. Wong, and L.S. Chao, *Chinese Named Entity Recognition with Conditional Random Fields in the Light of Chinese Characteristics*. 2013: Springer Berlin Heidelberg. 74-85.

[4] Sun, C., et al. 2006. *Biomedical Named Entities Recognition Using Conditional Random Fields Model*. in *Fuzzy Systems and Knowledge Discovery, Third International Conference, FSKD 2006, Xi'an, China, September 24-28, 2006, Proceedings*.

[5] Zhao, S. 2004. *Named entity recognition in biomedical texts using an HMM model*. in *International Joint Workshop on*

Natural Language Processing in Biomedicine and ITS Applications.

[6] Su, J. and J. Su. 2002. *Named entity recognition using an HMM-based chunk tagger*. in *Meeting on Association for Computational Linguistics*.

[7] Lafferty, et al., *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. 2001.

[8] Varone, M., et al. 2017. *Conditional random fields with semantic enhancement for named-entity recognition*. in *International Conference on Web Intelligence, Mining and Semantics*.

[9] YukunChen, et al. 2011. *A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries.* Journal of the American Medical Informatics Association Jamia. **18**(5): p. 601.

[10] Settles, B. 2004. *Biomedical named intity recognition using conditional random fields and rich feature sets.* In Proceedings of COLING 2004, International Joint Workshop On Natural Language Processing in Biomedicine and its Applications (NLPBA. p. 104--107.

[11] Xia, Y. and Q. Wang. *Clinical Named Entity Recognition: ECUST in the CCKS-2017 Shared Task 2.*

[12] Wu, J., et al. *Clinical Named Entity Recognition via Bi-directional LSTM-CRF Model.*

[13] Author, et al. *Chinese Named Entity Recognition.*

[14] Jianglu Hu, X.S., Zengjian Liu. *HITSZ_CNER: A hybrid system for entity recognition from Chinese clinical text.*

[15] Kingma, D.P. and J. Ba. 2014. *Adam: A Method for Stochastic Optimization.* Computer Science.

[16] Shao, Y., et al. 2017. *Character-based Joint Segmentation and POS Tagging for Chinese using Bidirectional RNN-CRF*.

[17] Cogswell, M., et al. 2015. *Reducing Overfitting in Deep Networks by Decorrelating Representations.* Computer Science.

[18] Srivastava, N., et al. 2014. *Dropout: a simple way to prevent neural networks from overfitting.* Journal of Machine Learning Research. **15**(1): p. 1929-1958.

[19] Huang, Z., W. Xu, and K. Yu. 2015. *Bidirectional LSTM-CRF Models for Sequence Tagging.* Computer Science.

[20] Mikolov, T., et al. 2013. *Efficient Estimation of Word Representations in Vector Space.* Computer Science.

[21] . http://brat.nlplab.org/.

[22]Uzuner, Ö., et al. 2011. *2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text.* Journal of the American Medical Informatics Association Jamia. **18**(5): p. 552.