DATA PAPER

Overview of CCKS 2020 Task 3: Named Entity Recognition and Event Extraction in Chinese Electronic Medical Records

Xia Li[1], Qinghua Wen[2], Hu Lin[1], Zengtao Jiao[3] & Jiangtao Zhang[1,2†]

[1]The 305th Hospital of the Chinese People's Liberation Army, Wenjin Street, Xicheng District, Beijing 100017, China

[2]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

[3]Yiducloud Beijing Technology Co., Ltd., Huayuan North Road, Haidian District, Beijing 100089, China

Keywords: Chinese electronic medical records; Event extraction; Named entity recognition; Clinical text; CCKS

Citation: Li, X., et al.: Overview of CCKS 2020 Task 3: Named entity recognition and event extraction in Chinese electronic medical records. Data Intelligence 3(3), 376-388 (2021). doi: 10.1162/dint_a_00093
Received: February 11, 2021; Revised: March 11, 2021; Accepted: March 15, 2021

ABSTRACT

The China Conference on Knowledge Graph and Semantic Computing (CCKS) 2020 Evaluation Task 3 presented clinical named entity recognition and event extraction for the Chinese electronic medical records. Two annotated data sets and some other additional resources for these two subtasks were provided for participators. This evaluation competition attracted 354 teams and 46 of them successfully submitted the valid results. The pre-trained language models are widely applied in this evaluation task. Data argumentation and external resources are also helpful.

1. INTRODUCTION

China Conference on Knowledge Graph and Semantic Computing (CCKS), which was founded in 2016, is organized by the Chinese Information Processing Society of China. To promote the development of technologies in knowledge graph and semantic computing, CCKS provides 8 evaluation tasks in 2020. Of these tasks, Task 3 focuses on named entity recognition (NER) and event extraction (EE) in the Chinese electronic medical records (EMRs).

† Corresponding author: Jiangtao Zhang (Email: zhang-jt13@tsinghua.org.cn; ORCID: 0000-0001-8462-3915).

EMRs are the core assets of a hospital. They are usually semi-structured data which contain lots of free text. Although a large amount of EMR data have been accumulated, most of them are not fully utilized. The difficulty of using free text blocks their usage.

NER and EE are commonly used techniques to acquire useful information from free text. NER in EMRs is also known as clinical named entity recognition (CNER). We can recognize diseases, drugs or other medical entity names from EMRs with the help of the NER model. The most popular NER method is sequence labeling, which can be based on long short-term memory (LSTM) [1,2,3] or bidirectional encoder representation from transformers (BERT) [4]. Clinical event extraction helps us identify medical events in EMRs, such as the tumor site, the tumor size and where the tumor transfers to. LSTM, BERT and other methods are applied in EE.

Traditional NER and EE are based on supervised models. However, the annotation of clinical information is much harder than the general domain information. Although there are some public medical data sets for the NER task, such as i2b2 [5], ShARe CLEF eHealth [6] and SemEval [7], there are barely public Chinese medical data sets. To promote the development of semantic analysis of the Chinese EMRs, the Knowledge Engineering Group of Tsinghua University and Yiducloud Beijing Technology Co., Ltd. organized this evaluation challenge at CCKS 2020. The data sets of this task provided by Yiducloud are restricted to CCKS evaluation only.

2. RELATED WORK

CCKS 2020 Task 3 focuses on NER and EE in the Chinese EMRs. NER and EE have been the core problems in natural language processing.

2.1 Chinese NER

NER is a task to locate and classify certain occurrences of words or expressions in unstructured text. In English NER, LSTM-CRF (Conditional Random Field) models [1,2,3] are a classic method to leverage both character-level and word-level representations, which can achieve the state-of-art results. Compared with NER in English, Chinese NER is more difficult since sentences in Chinese are not naturally segmented. A common practice for Chinese NER is to first perform word segmentation using an existing Chinese word segmentation (CWS) system and then apply a word-based NER model to infer the NER tags. However, the pipeline method suffers from error propagation, since the error of CWS may inevitably affect the performance of NER. Therefore, some approaches directly use a character-based NER model [8,9]. A drawback of the purely character-based NER model is that the word information is not fully exploited. To incorporate word information in Chinese NER, some recent methods, such as [10,11,12,13,14], resort to an automatically constructed lexicon.

2.2 Event Extraction

Event is a common but non-negligible knowledge type. Therefore, identifying events from texts and extracting their arguments are important for many applications. DMCNN [15] is a classic EE model, which uses the convolutional neural network (CNN) method to learn semantic features from raw texts, including lexical-level and sentence-level features. JRNN [16] is a recurrent neural networks (RNNs) based method for EE, aiming to integrate the discrete features with the automatically learned features. JMEE [17] is a method based on graph convolution networks (GCNs), which jointly extracts multiple event triggers and arguments by introducing syntactic shortcut arcs to enhance information flow and using attention-based GCNs to model graph information. Recently, event extraction is explicitly casted as a machine reading comprehension (MRC) problem [18] and the MRC model is used to solve event extraction.

2.3 NER and EE in Clinical Text

The information extraction of clinical text is getting more and more important in recent years. The TREC is the first shared tasks in clinical natural language processing (NLP), which focus on identifying relevant and irrelevant documents. Other evaluation tasks inculde ImageCLEFmed [19] and i2B2 [5]. For solving clinical NER, LSTM units and a conditional random field classifier [20] are used in the NER component. An unsupervised method [21] is used to build clinical NER systems which do not require any manual annotations and the models are trained on automatically annotated corpus followed by self-training iterations. For EE in clinical text, the bi-directional long short-term memory network assisted by the attention mechanism [22] is utilized to uncover the important aspects of the patient's medical conditions.

3. TASK DESCRIPTION

3.1 Clinical Named Entity Recognition

Given the free text from EMRs, this task aims to identify the clinical entity mentions and classify them into pre-defined categories. A novel method is presented for training clinical NER systems that do not require any manual annotations. It only requires a raw text corpus and a resource like Unified Medical Language System (UMLS) that can give a list of named entities along with their semantic types. Using these resources, annotations are automatically obtained to train machine learning methods. The methods were evaluated on the NER shared-task data sets of i2b2 2010 and SemEval 2014.

3.1.1 Formalized Definition

We define this task formally.

INPUT:

1). A document collection from EMR: $D = \{d_1,\ldots d_i,\ldots,d_N\}$, where $d_i = (w_{i1},\ldots,w_{in})$

2). A set of pre-defined categories: $C = \{c_1,\ldots,c_m\}$

OUTPUT:

Collections of entity mention-category pairs: $\{(m_1, c_{m1}),\ldots,(m_i, c_{mi}),\ldots,(m_p, c_{mp})\}$.

The $m_i = (d_i, b_i, e_i)$ represent the entity mention in document d_i, where b_i and e_i is the start and end position of m_i, respectively. $c_{mi} \in C$ represents the category of m_i. The overlap between mentions is not allowed, which is $e_i < b_{i+1}$.

3.1.2 Pre-defined Categories

There are 6 categories that are defined as follows.

1). Disease and diagnose (Dis)
2). Imaging examination (ImgExam)
3). Laboratory examination (LabExam)
4). Operation
5). Drug
6). Anatomy

3.2 Clinical Event Extraction

3.2.1 Formalized Definition

This task is formally defined as follows.

INPUT:

1). Event entity.
2). A document collection from EMR: $D = \{d_1,\ldots,d_N\}$, where $d_i = (w_{i1},\ldots,w_{in})$
3). A set of pre-defined attributes: $P = \{p_1, p_2,\ldots,p_m\}$

OUTPUT:

Collections of attribute entities: $\{[d_i,(p_j,(s_1,s_2,\ldots,s_k))]\}$, and $1 \le i \le N$, $1 \le j \le m$.

The s_k is the entity of attribute p_j from document d_i. There could be 0 or more than one entity for each attribute.

3.2.2 Pre-defined Attributes

The 3 pre-defined attributes are:

1). Tumor Primary Site
2). Tumor Size
3). Tumor Metastatic Site

4. DATA SETS

The data sets were provided by Yiducloud Beijing Technology Co., Ltd. Yiducloud organized a professional medical team to annotate these data. The data set is for CCKS evaluation only[①].

Compared with the CNER task in CCKS 2019, the annotated data set is about 4 times larger. Besides, Yiducloud provided an entity vocabulary and lots of unannotated data as additional resources that participators can use during the evaluation. The statistics of CNER data set are shown in Table 1.

The clinical event extraction data set includes a labeled training set, an unlabeled set and a vocabulary, which makes this challenge closer to the real-word scene. The statistics of clinical event extraction data sets are shown in Table 2.

Table 1. The statistics of clinical named entity recognition data set.

	Docs	Dis	ImgExam	LabExam	Operation	Drug	Anatomy	Total
Train	1500	6211	1490	1885	1327	2841	12660	26414
Test	300	1361	270	251	221	942	2661	5706
Unlabeled	1000	-	-	-	-	-	-	-

Table 2. The statistics of clinical event extraction data set.

	Docs	TumorPrimarySite	TumorSize	TumorMetastaticSite	Total
Train	1500	6211	1490	1885	1327
Test	300	1361	270	251	221
Unlabeled	1300	-	-	-	-

5. EVALUATION METRICS

5.1 Clinical Named Entity Recognition

5.1.1 Strict Metric

There are two evaluation metrics, the strict metric and relaxed metric. The extracted entities set is denoted as S and the gold entities set is denoted as G.

For the strict metric, $s_i \in S$ is equal to $g_j \in G$, which means they are exactly the same:

1). The start position of s_i equals to g_j
2). The end position of s_i equals to g_j
3). The category of s_i equals to g_j.

① To access the data sets, please contact the corresponding author after signing Data Usage Agreement.

The strict Precision, Recall and *F*1 can be calculated as follows:

$$P_s = \frac{|S \bigcap_s G|}{|S|} \tag{1}$$

$$R_s = \frac{|S \bigcap_s G|}{|G|} \tag{2}$$

$$F1_s = \frac{2P_sR_s}{P_s + R_s} \tag{3}$$

5.1.2 Relaxed Metrics

The relaxed metric does not require that $s_i \in S$ and $g_j \in G$ are exactly the same, and they only need to meet the following requirements:

1). The maximum value of the start position of s_i and g_j is less or equal to the minimum value of the end position of s_i and g_j;
2). The category of s_i is equal to g_j.

The relaxed Precision, Recall and *F*1 can be calculated as follows:

$$P_r = \frac{|S \bigcap_r G|}{|S|} \tag{4}$$

$$R_r = \frac{|S \bigcap_r G|}{|G|} \tag{5}$$

$$F1_r = \frac{2P_rR_r}{P_r + R_r} \tag{6}$$

5.2 Clinical Event Extraction

There could be more than one attribute entity for an event attribute. The Precision, Recall and *F*1 are calculated based on the attribute entity rather then attribute.

6. RESULTS AND DISCUSSION

This evaluation attracted 354 teams, and 46 of them successfully submitted their results. There are 32 teams which submitted results and 5 evaluation papers on the clinical named entity recognition task. Fourteen teams submitted their results and 3 papers on the clinical event extraction task. We list the top teams in Table 3 and Table 4, respectively.

Table 3. The results of clinical named entity recognition.

Rank	Team	Affiliation	Score
1	CASIA_Unisound	CASIA&Unisound AI Technology Co., Ltd.	0.91564
2	TMAIL	Medical AI Lab, Tencent Holdings Ltd.	0.91541
3	ywm	Lantone	0.91242
4	ChongChongChong	HFUT&SCUT	0.90801
5	SZU_IC	Shenzhen University	0.90511
6	mAI@pumc	Peking Union Medical College	0.90461

Table 4. The results of clinical event extraction.

Rank	Team	Affiliation	Score
1	dst	Knowledge Graph Group, Baidu, Inc.	0.76234
2	TMAIL	Medical AI Lab, Tencent Holdings Ltd.	0.74579
3	LHJB	National University of Defense Technology	0.73521
4	araloak	National University of Defense Technology	0.72730
5	zhjohnchan	Individual	0.71247
6	cecbrain	CEC Cloud Brain	0.67958

6.1 Clinical Named Entity Recognition

For the clinical named entity recognition task, Top 1 team and Top 2 team achieved very close scores. Both of them focus on the label inconsistency problem in CNER.

Top 1 team comes from the Institute of Automation, Chinese Academy of Sciences (CA-SIA) and Unisound AI Technology Co., Ltd. They proposed a hybrid system composed of a semi-supervised noisy label learning model based on adversarial training and a rule based post-processing module. They adopted a five-fold cross-voting mechanism to handle the annotation inconsistency problem in the data set. They used model ensemble and semi-supervised training to alleviate the insufficient training data problem. They also applied adversarial training to decrease aleatoric uncertainty and epistemic uncertainty simultaneously.

Based on the submitted papers, we have come to the following conclusions.

1). The pre-trained language models (PLMs) like BERT or ELMO [23] are widely applied. Using PLMs in their models have been a common sense among participants. Most teams did not simply apply the general BERT model, but the model pre-trained on the Chinese documents, such as RoBERTa-wwm [24]. Furthermore, some of them collected Chinese medical documents and pre-trained PLMs on these in-domain documents. The usage of PLMs in this year's evaluation challenge is more diverse than the previous competitions held at CCKS.
2). Model ensemble. Most teams applied this technique in their submission. The ensembled models usually achieve better results than a single model. The two-stage and *k*-folder ensemble methods are effective.

3). Feature engineering and rules are still valuable. In the clinical domain, there are lots of regular patterns and less annotated data. Therefore, participants can benefit from feature engineering. Some of the teams added the features of Chinese words into their model and gained stable improvements. They also introduced some rules to alleviate the data noise.
4). Semi-supervised methods. This evaluation provides 1,000 unlabeled data as additional resources. Some participants generated pseudo labels with a supervised model for the unlabeled data and trained the final model with both supervised and pseudo data.
5). Adversarial training. There are some unavoided label noises in the training data. To train a robust model not sensitive to the noises, some teams added turbulence to the word embeddings during training.
6). Domain vocabulary. Vocabulary is usually an important resource for the CNER task. In the past CNER evaluation, participants usually collected and extended the clinical vocabularies in various ways. The most popular vocabularies include ICD-10, the DrugBank database and some health websites such as "haodf.com" and "xywy.com". However, the top 3 teams in this year did not apply any vocabularies in their models. The main reason is their sufficient usage of PLMs. It may be a trend to replace vocabularies by PLMs.

6.2 Clinical Event Extraction

For the clinical event extraction task, Top 1 team achieved 0.76234 *F*1 score and Top 2 team achieved 0.74597 *F*1 score. The competition is fierce.

Top 1 team comes from Knowledge Graph Group, Baidu, Inc. They proposed a system mainly based on pre-trained language model. They applied domain adaption and task adaption during the pre-training, in order to improve the modeling ability of the pre-trained language model. To handle the insufficient training data challenge, they applied back translation to expand the training data. They also used entity vocabulary as the model input.

Based on the submitted papers, we have come to the following conclusions.

1). Pre-trained language models are widely used. Like the CNER task, Top 2 team applied PLMs in their models. Both of them chose RoBERTa [25] as the backbone. The usefulness of PLMs has been proved.
2). Data argumentation. The annotation of clinical documents is very difficult. Therefore, there are no sufficient labeled data for training. The top teams tried various data argumentation methods to enhance the model's robustness. Some of them doubled the training set by randomly re-arranging the sentence order in each training instance. Another team applied the back translation strategy to double the training set. They translated the training instance into English version and then back translated them into Chinese. Other teams tried random replacement of key information in the whole field.

3). Feature engineering and rules. Although the deep learning models are the key parts of the top teams' models, all of them utilized feature engineering or rules in pre-processing and post-processing. The pre-processing mainly focuses on gaining cleaner data. Some teams also cut the documents into certain length to meet the requirements of their model. The post-processing rules are applied on the model outputs to filter the meaningless results.

7. CONCLUSION

This paper presents a detailed introduction of CCKS 2020 Task3 for clinical named entity recognition and clinical event extraction for Chinese EMRs. From the evaluation results, the participants achieved exciting performances, especially in the CNER task. The models are more varified in this year's evaluation than in the previous year's evaluation. Participants modeled the evaluation problems in different aspects. Through this evaluation, we hope there could be more researchers who focus on semantic analysis of the Chinese EMRs and more companies pay attention to the industrialization of Chinese EMRs.

AUTHOR CONTRIBUTIONS

All of the authors contributed equally to the work. X. Li (lixia2012@139.com) summarized the evaluation task and drafted the paper. Q.H. Wen (wtsinghua1@163.com) reviewed the method documents submitted by the participating teams and undertook the code running test of the participating teams to ensure that the results were correct and fair. H. Lin (laohu20021210@163.com) summarized the result and discussion part of this paper. Z.T. Jiao (zengtao.jiao@yiducloud.cn) was responsible for producing data sets and labeling results by medical experts. J.T. Zhang (zhang-jt13@tsinghua.org.cn) was the organizer of this evaluation task who designed and released the shared task. All the authors have made meaningful and valuable contributions in revising and proofreading the resulting manuscript.

DATA AVAILABILITY STATEMENT

The data sets generated and/or analyzed during the current study are not publicly available due to the fact that the data sets are produced by medical expert consultants of Yidu Cloud based on their own experience. The publicly released version of the data sets needs the consent of all expert consultants, and they are available from the corresponding author on reasonable request.

REFERENCES

[1] Lample, G., et al.: Neural architectures for named entity recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 260–270 (2016)

[2] Chiu, J.P.C., Nichols, E.: Named entity recognition with bidirectional LSTM-CNNs. In: Transactions of the Association for Computational Linguistics, pp. 357–370 (2016)

[3] Ma, X.Z., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1064–1074 (2016)
[4] Devlin, J., et al.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
[5] Uzuner, Ö., et al.: 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association 18(5), 552–556 (2011)
[6] Suominen, H., et al.: Overview of the ShARe/CLEF eHealth evaluation lab 2013. In: The Fourth CLEF Conference, pp. 212–231 (2013)
[7] Pradhan, S., et al.: SemEval-2014 task 7: Analysis of clinical text. In: Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, pp. 54–62 (2014)
[8] He, J.Z., Wang, H.F.: Chinese named entity recognition and word segmentation based on character. In: Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing, pp. 128–132 (2008)
[9] Liu, Z.X., Zhu, C.H., Zhao, T.J.: Chinese named entity recognition with a sequence labeling approach: Based on characters, or based on words? In: International Conference on Intelligent Computing, pp. 634–640 (2010)
[10] Zhang, Y., Yang, J.: Chinese NER using lattice LSTM. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 1554–1564 (2018)
[11] Ding, R.X., et al.: A neural multi-digraph model for Chinese NER with gazetteers. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1462–1467 (2019)
[12] Liu, W., et al.: An encoding strategy-based word-character LSTM for Chinese NER. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2379–2389 (2019)
[13] Sui, D.B., et al.: Leverage lexical knowledge for Chinese named entity recognition via collaborative graph network. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3830–3840 (2019)
[14] Xue, M.G., et al.: Porous lattice transformer encoder for Chinese NER. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 3831–3841 (2020)
[15] Chen, Y.B., et al.: Event extraction via dynamic multi-pooling convolutional neural networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 167–176 (2015)
[16] Nguyen, T.H., Cho, K., Grishman, R.: Joint event extraction via recurrent neural networks. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 300–309 (2016)
[17] Liu, X., Luo, Z.C., Huang, H.Y.: Jointly multiple event extraction via attention-based graph information aggregation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 1247–1256 (2018)
[18] Liu, J., et al.: Event extraction as machine reading comprehension. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1641–1651 (2020)
[19] Kalpathy-Cramer, J., et al.: Evaluating performance of biomedical image retrieval systems– an overview of the medical image retrieval task at ImageCLEF 2004–2014. Computerized Medical Imaging and Graphics 39, 55–61 (2015)
[20] Magge, A., Scotch, M., Gonzalez-Hernandez, G.: Clinical NER and relation extraction using Bi-Char-LSTMs and random forest classifiers. In: The International Workshop on Medication and Adverse Drug Event Detection, pp. 25–30 (2018)

[21] Ghiasvand, O., Kate, R.J.: Learning for clinical named entity recognition without manual annotation. Informatics in Medicine Unlocked 13, 122–127 (2018)
[22] Yadav, S., et al.: Exploring disorder-aware attention for clinical event extraction. ACM Transactions on Multimedia Computing, Communications, and Applications 16(1s), Article 31 (2020)
[23] Peters, M., et al.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 2227–2237 (2018)
[24] Cui, Y.M., et al.: Revisiting pre-trained models for Chinese natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 657–668 (2020)
[25] Liu, Y.H., et al.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

AUTHOR BIOGRAPHY

Xia Li is currently the head of the Clinical Pharmacy Department of the 305th Hospital of the Chinese People's Liberation Army. She received her Bachelor's degree from the Second Military Medical University in 2002. Recently her research interests center around clinical knowledge graph and intelligent decision support for medication.
ORCID: 0000-0002-8420-1226

Qinghua Wen is currently a graduate student in the Department of Computer Science and Technology, Tsinghua University. His research interests include knowledge engineering, relation extraction and data mining.
ORCID: 0000-0002-4116-2140

Hu Lin is currently the head of the Department of Medical Administration of the 305th Hospital of the Chinese People's Liberation Army. His research interests focus on health management, Intelligent follow-up systems and medical knowledge graph.
ORCID: 0000-0003-1525-5922

Zengtao Jiao is currently the director of AI lab in Yidu Cloud Technology Co., Ltd. His research interests focus on the key and difficult problems in the field of medical artificial intelligence, such as medical text information extraction, disease prediction model, and medical knowledge mining.
ORCID: 0000-0002-3534-479X

Jiangtao Zhang received his PhD degree in Computer Science from Tsinghua University in 2018. He is now working as the director of the Information Center of the 305th Hospital of the Chinese People's Liberation Army. His research interests include knowledge graph, data mining and natural language processing in the medical domain. He organized and released a series of shared evaluation tasks for clinical knowledge discovery in CCKS 2017, CCKS 2018, CCKS 2019 and CCKS 2020.
ORCID: 0000-0001-8462-3915