Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5654–5664
Marseille, 20-25 June 2022

ChiMST: A Chinese Medical Corpus for Word Segmentation and Medical Term Recognition

Yuanhe Tian[♥], **Han Qin**[♠], **Fei Xia**[♥], **Yan Song**[♠†]
[♥]University of Washington [♠]The Chinese University of Hong Kong (Shenzhen)
[♥]{yhtian, fxia}@uw.edu [♠]hanqin@link.cuhk.edu.cn [♠]songyan@cuhk.edu.cn

Abstract

Chinese word segmentation (CWS) and named entity recognition (NER) are two important tasks in Chinese natural language processing. To achieve good model performance on these tasks, existing neural approaches normally require a large amount of labeled training data, which is often unavailable for specific domains such as the Chinese medical domain due to privacy and legal issues. To address this problem, we have developed a Chinese medical corpus named *ChiMST* which consists of question-answer pairs collected from an online medical healthcare platform and is annotated with word boundary and medical term information. For word boundary, we mainly follow the word segmentation guidelines for the Penn Chinese Treebank (Xia, 2000); for medical terms, we define 9 categories and 18 sub-categories after consulting medical experts. To provide baselines on this corpus, we train existing state-of-the-art models on it and achieve good performance. We believe that the corpus and the baseline systems will be a valuable resource for CWS and NER research on the medical domain.

Keywords: Chinese medical corpus, Chinese word segmentation, named entity recognition

1. Introduction

In the past decade, there have been tremendous progress in applying natural language processing (NLP) techniques to medical text, with the hope that the progress will lead to real-world applications that improve healthcare quality. When analyzing Chinese medical text, Chinese word segmentation (CWS) and named entity recognition (NER) are often two of the early steps whose outputs are needed by downstream modules, and they have thus attracted attentions from both academia and industry (e.g., (Xing et al., 2018; Wang et al., 2018; Zhang et al., 2019; Wang et al., 2019; Luo et al., 2019; Chang et al., 2021; Li et al., 2021)).

Building high-quality systems for the two tasks often requires large labeled datasets, which are usually unavailable for special domains such as the Chinese medical domain. While there are studies (Xu et al., 2014; Jinfeng et al., 2016; He et al., 2017; Gao et al., 2019; Xiong et al., 2019; Su et al., 2019; Zan et al., 2020; Zhang et al., 2020a) that annotated Chinese medical text with CWS and NER labels, most of them use electronic medical records (EMR) collected from hospitals and thus are not publicly available due to privacy and legal issues. Meanwhile, little attention has been paid to medical text from online healthcare platforms especially the ones with questions from patients and answers from doctors. In addition, the named entities in most studies use common categories such as disease and symptom, without including fine-grained medical term types such as abnormality types (e.g., 二型 (*Type II*)) and body functions (e.g., 消化 (*digestion*)).

In this paper, we introduce *ChiMST*, a <u>Chi</u>nese <u>M</u>edical Corpus with Word <u>S</u>egmentation and Medical <u>T</u>erm Annotation. Specifically, the raw text in the corpus comes from *ChiMed* (Tian et al., 2019), a Chinese medical question answering (QA) corpus collected from a Chinese online healthcare platform[1], where the registered doctors verified by the platform answer the questions raised by the patients (i.e., the platform users). For word segmentation, our annotation guidelines follow the segmentation guidelines for the Penn Chinese Treebank (CTB) (Xia, 2000), with more detailed specifications and examples added to accommodate the language use in the medical domain. For medical terms, we start with the medical taxonomy proposed by unified medical language system (UMLS) semantic groups (Lindberg et al., 1993). After consulting physicians, we choose a label set with 9 categories and 18 sub-categories for medical terms (see Section 3.2 for more details). Because our corpus consists of QA pairs from online QA healthcare platform rather than EMRs from hospitals, it is available to the public under a data license agreement.[2]

To test the usefulness of the *ChiMST* corpus as well as providing baseline results on CWS and medical term recognition (MTR) tasks,[3] we train several state-of-the-art models for CWS and MTR on the corpus; some of the models are enhanced by word n-grams and syntactic information, and the experimental results show that leveraging such extra information is able to improve model performance on both tasks.

[†] Corresponding author.

[1] http://ask.39.net/

[2] See https://github.com/synlp/ChiMST for details.

[3] Because some medical term categories such as *Medical Order* are not normally regarded as named entity types, for the rest of the paper we will call the task of identifying medical terms *Medical Term Recognition (MTR)*, instead of *Named Entity Recognition (NER)*.

2. Related Work

Applying NLP techniques to processing Chinese medical text has attract much attention in recent years (e.g., Xue et al. (2012), Xu et al. (2015), Li et al. (2019), Wang et al. (2020), Song et al. (2020), Chang et al. (2021), etc.). Because the performance of the models trained on general domain data tends to drop significantly when they are tested on medical text, previous studies (Xu et al., 2015; Li et al., 2015; Zhang et al., 2016; He et al., 2017; Chowdhury et al., 2018) often rely on annotated medical corpus to achieve better performance. For CWS and MTR, many datasets have been created for them (Xu et al., 2014; Jinfeng et al., 2016; He et al., 2017; Gao et al., 2019; Xiong et al., 2019; Su et al., 2019; Zan et al., 2020; Zhang et al., 2020a; Zhang et al., 2020b).

For example, Jinfeng et al. (2016) constructed a Chinese corpus containing 992 Chinese EMRs with annotations of named entities and their relations. They mark five types of named entities in Chinese EMRs: *diseases*, *type of disease diagnosis*, *symptoms and signs*, *test*, and *treatments*. Their innovation in the definition of named entities was to split medical problems into diseases and symptoms, and further split the symptoms into two subcategories: self-reported symptoms and abnormal examination results. Zhang et al. (2020a) divided medical entities into nine categories (i.e., *disease*, *clinical manifestation*, *medical procedure*, *medical equipment*, *drug*, *medical test*, *body*, *department*, and *microbe*) for primary classification, where some of them are further divided into sub-categories.

Most existing datasets for CWS and MTR use EMRs and thus are not publicly available due to privacy and legal issues. In our study, we annotate medical text collected from online healthcare platforms with a more fine-grained set of categories.

3. The *ChiMST* Corpus

The raw data of the current version of the *ChiMST* corpus come from the QA records in the *ChiMed* corpus (Tian et al., 2019).[4] After a brief introduction to *ChiMed*, we describe *ChiMST*'s annotation guidelines and report annotation results.

3.1. Data from the *ChiMed* Corpus

The *ChiMed* corpus contains over 200 thousand QA records collected from a Chinese online healthcare platform called 39ask. A QA record contains five main fields: department, title, keyphrases, the question, and the answers. Among the five fields, the department and keyphrases are provided by the platform managers; the question and the title fields, which describe the problem the patient has, are written by the patient; the answers are written by physicians who have registered and verified by the platform. All the QA records come from 15 different departments and there are exactly two answers in each QA record.

To create the *ChiMST* corpus, we randomly sampled 1,000 QA records from *ChiMed*. The annotation is done only on the question and answer fields of the QA records. We segment text in those fields into sentences by three punctuation marks (i.e., delimiters period, question marks, and exclamation marks), resulting in a corpus of 6,646 sentences and 222,465 Chinese characters. Table 1 shows an example QA record[5], with word boundary and medical terms annotated as explained below.

3.2. Annotation Guidelines

Two graduate students in the field of NLP developed the annotation guidelines for CWS and MTR. They consulted two physicians when defining the category set for the medical terms.

For word segmentation, we follow the segmentation guidelines of the Penn Chinese Treebank (CTB) (Xia, 2000), which is one of the most widely used guidelines for CWS. Since the CTB guidelines focus on the general domain and do not include many examples from the medical domain, we add more specifications and examples from the medical domain to help our annotators. For example, the CTB guidelines do not specify how to segment medical terms. Following CTB's approach for segmenting names of organization/country/school, we add examples for medical terms, such as "神经/内科"[6] (*nerve internal medicine*), "营养科" (*nutrition*), "淋巴/细胞" (*lymphocytes*), and "膝跳/反射" (*knee-jerk reflex*).

For the medical terms, we start with the medical taxonomy proposed by unified medical language system (UMLS) semantic groups (Lindberg et al., 1993) and choose categories that are widely used in our corpus. After consulting two physicians, we define a set consisting of 9 medical term categories and 18 subcategories, shown in Table 2.

Compared with previous Chinese datasets with medical term annotations, our medical term categories are more fine-grained. For instance, the category "*Abnormality*" has two subtypes: "*Symptom*" is the patient's self experience and feeling to the physiological function of the body abnormal, such as "瘙痒" (*pruritus*), "疼痛" (*pain*), "紧张" (*distension*), "胀闷" (*stuffy*), and "头晕" (*dizziness*), whereas "*Sign*" represents the perceptible changes in the body's internal structure, such as 心脏杂音 (*heart murmur*) and 肝脾大 (*enlarged liver and spleen*). The category "*Illness*" includes "*Disease*" and "*Injured or Poisoned*“, which correspond to different physical condition for patients. "*Medical Test*" has three subcategories: "*Imaging Test*", "*Laboratory Test*", and "*General Test*", where the first two labels are for tests involving images and laboratory examinations, respectively. "*Treatment*" has four subcate-

[4] In the future, we plan to extend *ChiMST* with data from other sources.

[5] The English translation is not part of the corpus.

[6] "/" marks the word boundary.

Department	眼科 > 近视 *Ophthalmology > Nearsightedness*
Title	近视不想戴眼镜改怎么办? *What should I do if I am nearsighted but do not want to wear glasses?*
Keypharses	眼睛，视力，眼科检查，屈光不正 *eyes, vision, eye examination, refractive error*
Question	我 [近视@DI] ， 五六 百 度 了 。 *I am [nearsighted@DI], five or six hundred degrees.* 看 电脑 都 费劲 ， 晚上 玩 电脑 的 时候 戴戴 ， 平常 不 戴 。 *It takes a lot of effort to work on the computer. I wear glasses when playing (games) on computer at night, otherwise I do not normally wear glasses.* 请 问 [近视@DI] 会 越来越 严重 吗 ， 不 想 戴 眼镜 该 怎么 办 ？ *Will my [nearsightedness@DI] get worse? What should I do if I do not want to wear glasses?*
Answer 1	从 你 的 描述 来 看 ， 考虑 和 平时 [睡眠@BF] 缺乏 用眼 不 合理 ， 导致 [疲劳@SP] 和 [屈光 不 正@DI] 有 关系 。 *From your description, your condition could have something to do with lack of [sleep@BF] and unreasonable use of eyes in daily life, which cause [fatigue@SP] and [refractive errors@DI].* 建议 去 正规 医院 的 [眼科@DEP] 检查 ， 然后 佩戴 合适 的 眼镜 ， 平时 [注意 休息@MO] ， [合理 用眼@MO] ， 控制 手机 ， 电脑 的 使用 ， 做 [眼部@BP] 的 [热敷@TT] ， 可以 使用 滋润性 的 眼药水 ， 比如 [贝复舒 滴眼液@DR] 。 *It is recommended to go to the [ophthalmology department@DEP] of a hospital for an eye examination, and then wear suitable glasses, [rest more@MO] at ordinary times, [use your eyes rationally@MO], control the use of mobile phones and computers, do [hot compress@TT] for [eyes@BP], and use moisturizing eye drops, such as [Beifu Shu eye drops@DR].*
Answer 2	你 好 ， [近视@DI] 了 ， 建议 还是 配戴 眼镜 。 *Hello. If you are [nearsighted@DI], it is still recommended to wear glasses.* 配戴 眼镜 的 作用 ： *The purpose of wearing glasses:* 一 是 为了 矫正 屈光 ， 达到 清晰 视物 的 效果 ； *One is to correct the refractive and achieve the effect of clear vision;* 二 是 为了 防止 造成 [眼睛@BP] 的 [集合 和 调节 功能@BF] 失去 平衡 。 *The second is to prevent the [assembly and adjustment function@BF] of [eyes@BP] from losing balance.* 如果 确实 不 想 戴 眼镜 ， 可 考虑 通过 [近视 手术@SR] 来 提高 裸眼 视力 。 *If you really do not want to wear glasses, you can consider [myopia surgery@SR] to improve your vision.*

Table 1: An example of annotated QA record in *ChiMST*, where the question and two answers are annotated with word boundaries and medical terms. Word boundaries are illustrated by white spaces. Medical terms are highlighted in blue and marked by the brackets. The label of each medical term follows the schema in Table 2 and it is attached to the corresponding medical term by "@". For example, [贝复舒 滴眼液@DR] (*[Beifu Shu eye drops@DR]*) is a medical term of type *Drug (DR)* consisting of two words (i.e., 贝复舒 and 滴眼液). The English translation is included only for reference; it is not part of the corpus.

gories depending on the types of treatment, namely, ”*Drug*“, "*Surgery*", "*Targeted Treatment*", and "*Medical Order*". "*Body*" and "*Abnormality Type*" are categories that are rarely used in previous studies, but they are useful when we want to accurately identify the detail of a disease and obtain more accurate information for the follow-up diagnosis and treatment. The remaining three categories (namely, "*Department*", "*Medical Equipment*", and "*Pathogen*") are widely used categories, which also play an important role in the clinical process of diagnosis and treatment.

To help annotators to distinguish similar medical term (sub-)categories, our annotation guidelines include detailed specifications and examples such as the following:

- **Symptom and Sign:** If the abnormality can be detected without requiring professional medical equipment (excluding common home medical equipment such as thermometer, sphygmomanometer, blood glucose meter, etc.), annotate it as "*Symptom*" (e.g., "疼痛" (*pain*) and "发烧" (*fever*)). Otherwise, annotate it as "*Sign*".

	Category	Sub-category	Labels	Explanations
1	Abnormality	Symptom	SP	Symptom that is directly observed or felt by the patient, such as “疼痛” (*pain*) and “恶心” (*nausea*).
		Sign	SG	Sign that is diagnosed or detected by the doctor or medical equipment, such as “心脏杂音” (*heart murmur*) and “肝脾大” (*enlarged liver and spleen*)
2	Abnormality Type	Abnormality Type	DT	The type and degree of an abnormality, such as “极高危险组” (*very high risk group*) and “轻度” (*mild*).
3	Body	Body Part	BP	Body part such as “头” (*head*) and “腰” (*waist*).
		Body Substance	BS	Substance produced or excreted from the body, such as “血” (*blood*) and “尿” (*urine*).
		Body Function	BF	The general function of human body, such as “消化” (*digestion*) and “怀孕” (*pregnant*).
4	Department	Department	DEP	A hospital department such as “外科” (*surgical department*) and “妇科” (*department of gynecology*).
5	Illness	Injured or Poisoned	IOP	The name of injury and poison, such as “皮肤受伤” (*skin injury*) and “酒精中毒” (*alcoholism*).
		Disease	DI	The name of disease excluding injury and poison, such as “肺癌” (*lung cancer*) and “皮疹” (*rush*).
6	Medical Equipment	Medical Equipment	ME	Medical equipment, such as “鼻腔镜” (*rhinoscope*) and “呼吸机” (*ventilator*).
7	Pathogen	Pathogen	PG	Pathogen, such as “细菌” (*germ*) and “病毒” (*virus*).
8	Medical Test	Imaging Test	IMT	Image medical examination, such as “腹部CT” (*abdominal CT*) and “心脏彩超” (*echocardiography*).
		Laboratory Test	LT	Laboratory medical examination, such as “血常规” (*complete blood count*) and “凝血四项” (*blood clotting tetrachoric*).
		General Test	GT	General medical examination excluding image tests and laboratory tests, such as “体温” (*body temperature*) and “呼吸” (*breathe*).
9	Treatment	Drug	DR	The name of drugs, such as “青霉素” (*penicillin*) and “泰勒” (*Taylor*).
		Surgery	SR	The name of surgeries, such as “吸脂手术” (*liposuction*) and “心脏搭桥术” (*heart bypass surgery*).
		Targeted Treatment	TT	Targeted treatment, such as “抗病毒治疗” (*antiviral therapy*) and “中频疗法” (*intermediate frequency therapy*).
		Medical Order	MO	The suggestions from doctors for daily medical care, such as “严禁饮酒” (*drinking alcohol is strictly prohibited*) and “多喝水” (*drink more water*).

Table 2: Types of medical terms (there are 9 categories and 18 sub-categories) used in *ChiMST*. The last two columns show the label and explanation for each sub-category.

- **Body function, Imaging test, and Laboratory test:** If the word “做” (*do*) can be added before the medical term, then annotate the term as *Imaging Test* or *Laboratory Test* according to its meaning (e.g., “血常规” (*complete blood count*)). Otherwise, annotate the term as *Body Function* (e.g., “视力” (*vision*) and “肝功能” (*liver function*)).

Moreover, following the convention in previous studies, we do not annotate medical terms embedded in another larger medical term. For example, in the *Body Function* term “肝功能” (*liver function*), we do not mark the body part “肝” (*liver*) as a *Body Part*.

3.3. Annotation Process

Since questions and answers are the most important fields of QA records, we ask two annotators who have background in the medical domain to annotate them: word segmentation first, followed by medical term annotation.

We train the two annotators and ask them to first annotate a small set of sample data to ensure they fully un-

# of QA records annotated by both annotators	300
# of QA records annotated by one annotator	700
Agreement (Cohen's kappa) for CWS	0.934
Agreement (Cohen's kappa) for MTR	0.907
Avg. F1 of annotators for CWS	0.974
Avg. F1 of annotators for MTR	0.931

Table 3: The inter-annotator agreement (Cohen's kappa) of two annotators on the 300 double-annotated QA records. The average F1 scores of the annotators are computed with respect to the agreed annotation.

# of QA records	1,000
# of questions	1,000
# of answers	2,000
# of characters	222,465
# of words	142,455
# of medical terms	13,695
# of sentences	6,646
# of character types	2,455
# of word types	8,982
# of medical term (sub-)categories	18
Avg. # of characters per sentence	33.5
Avg. # of words per sentence	21.4
Avg. # of medical terms per sentence	2.1

Table 4: The statistics of the *ChiMST* corpus.

derstand the annotation guidelines. Then we split the 1,000 QA records in *ChiMST* into two subsets: one has 300 records which are annotated by both annotators, and the other 700 records are annotated by only one annotator (i.e., 350 records per annotator); that is, each annotator gets 650 records without knowing which of them will be double annotated.

For double-annotated records, if the two annotations disagree, the two annotators would discuss and resolve the disagreements. If they cannot reach an agreement, one of the annotation guideline designers would step in to make the final decisions. Table 3 shows the inter-annotator agreement in terms of Cohen's kappa and each annotation's quality in terms of F-score on the 300 double annotated QA records. Both kappa and F-scores fall in the range of $(0.8, 1)$ (Landis and Koch, 1977), indicating the annotation is of high quality.

3.4. *ChiMST* Statistics

To summarize, *ChiMST* contains 1,000 QA records randomly selected from *ChiMed* (Tian et al., 2019). The question and answer fields in the records are annotated with word boundary and medical term information. The statistics of *ChiMST* is in Table 4, where the number of characters, words, and sentences only considers the question and answer fields. Table 5 reports the occurrences of the 18 medical term subcategories.

	Sub-categories	**Labels**	**Count**
1	Symptom	SP	1,946
2	Sign	SG	164
3	Abnormality Type	DT	374
4	Body Parts	BP	2,408
5	Body Substance	BS	407
6	Body Function	BF	913
7	Department	DEP	231
8	Injured or Poisoned	IOP	72
9	Disease	DI	3,520
10	Medical Equipment	ME	78
11	Pathogen	PG	116
12	Imaging Test	IMT	242
13	Laboratory Test	LT	171
14	General Test	GT	124
15	Drug	DR	1,338
16	Surgery	SR	142
17	Targeted Treatment	TT	517
18	Medical Order	MO	932
Total			13,695

Table 5: The number of medical terms in each subcategory in the annotated *ChiMST* corpus.

4. Experiments

To test the usefulness of *ChiMST*, we train and evaluate CWS and MTR systems on the corpus.

4.1. Tasks

Following the convention in most previous studies (Tseng et al., 2005; Song et al., 2009; Zhang et al., 2010; Song and Xia, 2012; Pei et al., 2014; Lample et al., 2016; Nie et al., 2020b; Tian et al., 2020a; Tian et al., 2021), we regard CWS and MTR as sequence labeling tasks using the *BIO* scheme (or its variants such as the *BIES* scheme). That is, each character is assigned a CWS and an MTR label according to its position in a word or in a medical term.[7] For example, if a character is the first character of a word, its CWS label would be "B"; if it is part of a medical term of type department (DEP) but is not in the beginning position, its MTR label would be "I-DEP". The object of the models for CWS and MTR is to predict the corresponding CWS and MTR label sequence $\widehat{\mathcal{Y}} = \widehat{y}_1, \cdots, \widehat{y}_n$ for an input character sequence $\mathcal{X} = x_1, \cdots, x_n$ (n denotes the number of characters in the sequence).

4.2. Models

Since pre-trained word embeddings and language models have demonstrated their effectiveness in modeling

[7] In our experiments, the input to the MTR models is a character sequence, not a word sequence. Therefore, the BIO label is on each character.

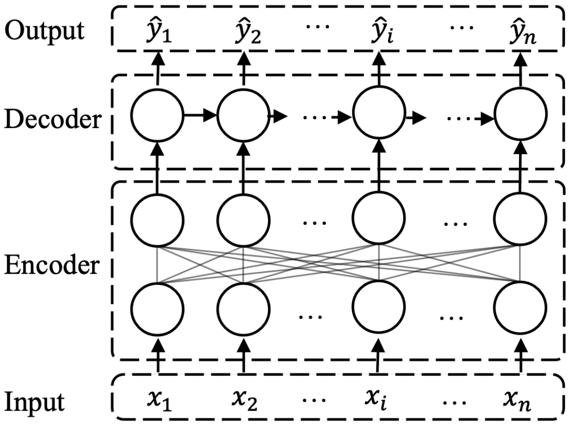
Output
$\hat{y}_1$ $\hat{y}_2$ ··· $\hat{y}_i$ ··· $\hat{y}_n$
Decoder
Encoder
Input
x_1 x_2 ··· x_i ··· x_n

Figure 1: The typical architecture of character-based models following the encoder-decoder paradigm.

	Train	Dev	Test
# of QA records	700	100	200
# of questions	700	100	200
# of answers	1,400	200	400
# of characters	154,342	23,649	44,474
# of words	98,874	15,068	28,513
# of medical terms	9,459	1,494	2,742
# of sentences	4,584	702	1,360
Double annotated	No	Yes	Yes

Table 6: The statistics of the training, development, and test sets of *ChiMST* in the experiments. "Double annotated" indicates whether the data is annotated by both annotators.

Hyper-parameters	Values
Learning Rate	$5e-6, 1e-5, \mathbf{2e-5}$
Warmup Rate	$0.06, \mathbf{0.1}$
Dropout Rate	$\mathbf{0.1}$
Batch Size	$8, \mathbf{16}, 32$

Table 7: The hyper-parameter values tested when tuning our models, and the ones used in our final experiments are in boldface.

the context information for different tasks (Pennington et al., 2014; Song and Shi, 2018; Bae et al., 2019; Chen et al., 2020; Raffel et al., 2020; Tian et al., 2020b; Mandya et al., 2020; Ribeiro et al., 2021; Qin et al., 2021a; Herzig and Berant, 2021; Qin et al., 2021b; Paolini et al., 2021; Rothe et al., 2021; Pasupat et al., 2021; Tian et al., 2022), in the experiments, we train four state-of-the-art character-based models for CWS and MTR with widely used pre-trained language models for Chinese. Specifically, the four models are BERT (Devlin et al., 2019), ZEN (Diao et al., 2020; Song et al., 2021), WMSeg (Tian et al., 2020c), and AESINER (Nie et al., 2020a), where all of them follow the standard encoding-decoding paradigm in Figure 1. The details of the models are illustrated as follows:

BERT BERT[8] is a pre-trained language model in the general domain that has achieved state-of-the-art performance on many NLP tasks. We used the base version of Chinese BERT as the text encoder and use softmax or conditional random field (CRF) as the decoder to predict the CWS and MTR labels for each character. We use the default setting for the BERT encoder, which uses 12 layers of multi-head attentions and 768 dimensional vectors for the hidden states.

ZEN ZEN[9] is another pre-trained language model for Chinese in the general domain. It enhances BERT by modeling n-gram information through the encoding process of the running text and thus achieves better performance on many Chinese NLP tasks. We use the large version of ZEN with the default setting (i.e., 24 layers of multi-head attentions and 1024 dimensional hidden states) and use softmax or CRF as the decoder.

WMSeg WMSeg[10] is a model for CWS, which uses either BERT or ZEN as the text encoder and CRF as the decoder. WMSeg improves CWS by leveraging the wordhood information in n-grams to achieve better performance. We follow the default settings specified in the original paper and apply it to CWS only.

AESINER AESINER[11] proposes an attentive ensemble mechanism to leverage different types of syntactic information (namely, part-of-speech (POS) labels, syntactic constituents, and dependency relations) and achieves state-of-the-art performance on NER in the general domain. Similar to WMSeg, the text encoder of AESINER is either BERT or ZEN and the decoder is CRF. We use the default settings of AESINER in the original paper and apply it only to MTR.

4.3. Settings

For the data, we split *ChiMST* into training, development, and test sets, where both CWS and MTR tasks use the same train/dev/test split. The training set contains the 700 QA records, each being annotated by one of the annotators; the development and test set contains 100 and 200 QA records, respectively, and each being double annotated. The statistics of the train/dev/test sets are reported in Table 6. In addition, since our CWS annotation follows the CTB guideline, for CWS we use the training set of CTB5[12] (Xue et al., 2005) in the gen-

[8] We obtain BERT from `https://github.com/google-research/bert`.

[9] We obtain ZEN from `https://github.com/sinovation/ZEN2`.

[10] We use the official code from `https://github.com/SVAIGBA/WMSeg`.

[11] We use the offcial code from `https://github.com/cuhksz-nlp/AESINER`.

[12] We obtain the official data of CTB5 from `https://catalog.ldc.upenn.edu/LDC2005T01` and use the training set specified in Zhang et al. (2014).

Methods	Prec.	Recall	F1
CTB Only			
BERT	94.61	94.81	94.71
BERT + CRF	94.50	95.11	94.80
WMSeg (BERT)	94.81	**95.02**	94.92
ZEN	94.58	94.93	94.75
ZEN + CRF	94.70	94.89	94.79
WMSeg (ZEN)	**95.14**	94.90	**95.02**
CTB+ChiMST			
BERT	97.93	97.65	97.79
BERT + CRF	97.89	97.76	97.82
WMSeg (BERT)	98.04	97.91	97.97
ZEN	98.21	97.99	98.10
ZEN + CRF	98.23	98.01	98.12
WMSeg (ZEN)	**98.39**	**98.04**	**98.21**
ChiMST Only			
BERT	97.82	98.15	97.98
BERT + CRF	97.90	98.16	98.03
WMSeg (BERT)	98.00	98.23	98.11
ZEN	98.22	98.18	98.20
ZEN + CRF	98.29	98.21	98.25
WMSeg (ZEN)	**98.05**	**98.41**	**98.38**

Table 8: CWS performance for different composition of training data, namely, the CTP data only (*CTB only*), the CTB data and our data (*CTB+ChiMST*), and our data only (*ChiMST Only*). BERT and ZEN refer to the models with the softmax decoder. WMSeg (BERT) and WMSeg (ZEN) denote the WMSeg models using BERT and ZEN encoder, respectively.

eral domain as extra data and conduct cross-domain experiments where the models are trained on CTB5 and evaluated on the test set of *ChiMST*.

To obtain the syntactic information required by the AESINER model, we use Stanford CoreNLP Toolkits[13] (Manning et al., 2014) to process the text and obtain the POS labels, syntactic constituents, and dependency relations for each sentence. In training, we update all parameters in the models, including the ones in the pretrained language models. Table 7 reports the hyperparameters tested when tuning the models. We test all combinations of them for each model and use the one that achieves the highest performance on the development set as the final model for evaluation. Following previous studies, for both CWS and MTR, we evaluate all models based on the precision, recall, and F1 scores.

Methods	Prec.	Recall	F1
BERT	79.19	83.81	81.43
BERT + CRF	78.60	84.79	81.57
AESINER (BERT)	79.15	84.87	81.91
ZEN	82.17	85.52	83.81
ZEN + CRF	82.44	85.59	83.98
AESINER (ZEN)	**82.67**	**86.25**	**84.41**

Table 9: The performance of different models on the MTR task in terms of precision, recall, and F1 scores. Herein, AESINER models use all three types of syntactic information (i.e., POS labels, syntactic constituents, and dependency relations).

Syntactic Info.	Prec.	Recall	F1
POS Only	82.31	85.83	84.03
SC Only	**82.55**	**86.04**	**84.25**
DR Only	82.30	85.98	84.10
All	82.67	86.25	84.41

Table 10: The performance of AESINER model (using ZEN encoder) using different types of syntactic information. "POS", "SC", and "DR" denote POS labels, syntactic constituents, and dependency relations, respectively. The performance of the model where all types of syntactic information are used (i.e., "All") is also reported for reference.

5. Results and Discussion

5.1. Performance on CWS

For CWS, we run the three models (i.e., BERT, ZEN, and WMSeg) with different settings and report system performance (precision, recall, and F1 scores) on the test set of *ChiMST* in Table 8. Herein, "*CTB only*" refers to the cross-domain experiment setting where the models are trained on the training set of CTB5; "*CTB+ChiMST*" denotes the setting where the training data is the union of CTB5 and *ChiMST* training sets; "*ChiMST Only*" is the in-domain setting where the models are trained and evaluated on *ChiMST*.

There are some observations. First, all models achieve outstanding performance when the training data is *ChiMST Only*. Second, *CTB only* results are worse than both *CTB+ChiMST* and *ChiMST only*, confirming that the gaps between the domains remarkably affect the performance of WSD models and demonstrating the benefits of constructing annotated corpus for special domains such as medical domain. Third, for all three settings, WMSeg with ZEN encoder achieves the highest F1 scores, because this model can leverage n-gram information to improve CWS performance.

5.2. Performance on MTR

For MTR, we run BERT, ZEN, and AESINER models with different configurations on *ChiMST*. We report

[13] `https://stanfordnlp.github.io/CoreNLP/`.

the experimental results (i.e., precision, recall, and F1 score) in Table 9, where the AESINER models with BERT or ZEN encoder use all three types of syntactic information (i.e., POS labels, syntactic constituents, and dependency relations). Overall, the model performance on MTR is much lower than that on CWS, which is expected because MTR is generally considered harder than CWS and thus has more room for future improvement. Among all models, AESINER with ZEN encoder achieves the highest performance with respect to the F1 score, indicating that n-gram and syntactic information provides useful cues for MTR.

To understand the effect of different types of syntactic information on model performance, we conduct an ablation study on the syntactic information used in AESINER with ZEN encoder. The experimental results are in Table 10: the first three rows use only one type of syntactic information, and the last row uses all three types. Among the three types, syntactic information (SC) works the best. One possible explanation is that the syntactic constituents suggest the boundary information of a medical term (medical terms are more likely to be noun phrases) and thus contribute more to identifying the medical terms than the other two types of syntactic information.

6. Conclusion

In this paper, we describe a Chinese medical corpus named *ChiMST* with annotations for CWS and MTR. Specifically, the corpus includes 1,000 QA records from the *ChiMed* corpus, where the question and answer fields are annotated with word segmentation and medical term information. The annotation guidelines for CWS follows CTB's segmentation guidelines; medical term annotation uses a label set with 9 categories and 18 sub-categories.

Compared with existing datasets for CWS and MTR, *ChiMST* uses QA records from a public healthcare platform rather than EMRs in the hospitals; thus we are able to release the corpus to the public. In addition, the 18-subcategory label set is more fine-grained than the label sets in previous studies.

We further conduct experiments on *ChiMST* with state-of-the-art sequence labeling models. The experimental results on CWS show the performance in the cross-domain setting is much worse than the in-domain setting, demonstrating the benefit of having labeled data for special domains. In addition, WMSeg and AESINER, which leverage n-gram and syntactic information, achieve the best performance for CWS and MTR, respectively, demonstrating that extra information (such as n-gram and syntactic information) is helpful to improve model performance.

7. Acknowledgements

This work is supported by Shenzhen Science and Technology Program under the project "Fundamental Algorithms of Natural Language Understanding for Chinese Medical Text Processing" (JCYJ20210324130208022) and the Natural Science Foundation of Guangdong Province, China, under the project "Deep Learning based Chinese Combination Category Grammar Parsing and its Application in Relation Extraction". We would also like to thank three anonymous reviewers for helpful feedback.

8. Bibliographical References

Bae, S., Kim, T., Kim, J., and Lee, S.-g. (2019). Summary Level Training of Sentence Rewriting for Abstractive Summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 10–20, Hong Kong, China, November.

Chang, Y., Kong, L., Jia, K., and Meng, Q. (2021). Chinese Named Entity Recognition Method based on BERT. In *2021 IEEE International Conference on Data Science and Computer Application (ICDSCA)*, pages 294–299. IEEE.

Chen, G., Tian, Y., and Song, Y. (2020). Joint Aspect Extraction and Sentiment Analysis with Directional Graph Convolutional Networks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 272–279, December.

Chowdhury, S., Dong, X., Qian, L., Li, X., Guan, Y., Yang, J., and Yu, Q. (2018). A Multitask Bidirectional RNN Model for Named Entity Recognition on Chinese Electronic Medical Records. *BMC bioinformatics*, 19(17):75–84.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Diao, S., Bai, J., Song, Y., Zhang, T., and Wang, Y. (2020). ZEN: Pre-training Chinese Text Encoder Enhanced by N-gram Representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4729–4740, November.

Gao, Y., Gu, L., Wang, Y., Wang, Y., and Yang, F. (2019). Constructing a chinese electronic medical record corpus for named entity recognition on resident admit notes. *BMC medical informatics and decision making*, 19(Suppl 2):56.

He, B., Dong, B., Guan, Y., Yang, J., Jiang, Z., Yu, Q., Cheng, J., and Qu, C. (2017). Building a Comprehensive Syntactic and Semantic Corpus of Chinese Clinical Texts. *Journal of biomedical informatics*, 69:203–217.

Herzig, J. and Berant, J. (2021). Span-based Semantic Parsing for Compositional Generalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 908–921, Online, August.

Jinfeng, Y., Yi, G., Bin, H., Chunyan, Q., Qiubin, Y., Yaxin, L., and Yongjie, Z. (2016). Corpus Construction for Named Entities and Entity Relations on Chinese Electronic Medical Records. *Journal of Software*, 27(11):2725–2746.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). 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*, pages 260–270, San Diego, California, June.

Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *biometrics*, pages 159–174.

Li, Y.-B., Zhou, X.-Z., Zhang, R.-S., Wang, Y.-H., Peng, Y., Hu, J.-Q., Xie, Q., Xue, Y.-X., Xu, L.-L., Liu, X.-F., et al. (2015). Detection of Herb-Symptom Associations from Traditional Chinese Medicine Clinical Data. *Evidence-Based Complementary and Alternative Medicine*, 2015.

Li, X., Wang, H., He, H., Du, J., Chen, J., and Wu, J. (2019). Intelligent Diagnosis with Chinese Electronic Medical Records based on Convolutional Neural Networks. *BMC bioinformatics*, 20(1):1–12.

Li, X., Wen, Q., Lin, H., Jiao, Z., and Zhang, J. (2021). Overview of CCKS 2020 Task 3: Named Entity Recognition and Event Extraction in Chinese Electronic Medical Records. *Data Intelligence*, 3(3):376–388, 09.

Lindberg, D., Humphreys, B., and McCray, A. (1993). The Unified Medical Language System. *Methods of information in medicine*, 32(4):281.

Luo, R., Xu, J., Zhang, Y., Ren, X., and Sun, X. (2019). PKUSEG: A Toolkit for Multi-domain Chinese Word Segmentation. *arXiv preprint arXiv:1906.11455*.

Mandya, A., Bollegala, D., and Coenen, F. (2020). Graph Convolution over Multiple Dependency Subgraphs for Relation Extraction. In *COLING*, pages 6424–6435.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Nie, Y., Tian, Y., Song, Y., Ao, X., and Wan, X. (2020a). Improving Named Entity Recognition with Attentive Ensemble of Syntactic Information. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4231—4245, Online, November.

Nie, Y., Tian, Y., Wan, X., Song, Y., and Dai, B. (2020b). Named Entity Recognition for Social Media Texts with Semantic Augmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1383–1391, Online, November.

Paolini, G., Athiwaratkun, B., Krone, J., Ma, J., Achille, A., Anubhai, R., Santos, C. N. d., Xiang, B., and Soatto, S. (2021). Structured Prediction as Translation between Augmented Natural Languages. *arXiv preprint arXiv:2101.05779*.

Pasupat, P., Zhang, Y., and Guu, K. (2021). Controllable Semantic Parsing via Retrieval Augmentation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7683–7698, Online and Punta Cana, Dominican Republic, November.

Pei, W., Ge, T., and Chang, B. (2014). Max-margin Tensor Neural Network for Chinese Word Segmentation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 293–303.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October.

Qin, H., Chen, G., Tian, Y., and Song, Y. (2021a). Improving Arabic Diacritization with Regularized Decoding and Adversarial Training. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Qin, H., Tian, Y., and Song, Y. (2021b). Relation Extraction with Word Graphs from N-grams. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2860–2868, Online and Punta Cana, Dominican Republic, November.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21:1–67.

Ribeiro, L. F. R., Zhang, Y., and Gurevych, I. (2021). Structural Adapters in Pretrained Language Models for AMR-to-Text Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4269–4282, Online and Punta Cana, Dominican Republic, November.

Rothe, S., Maynez, J., and Narayan, S. (2021). A Thorough Evaluation of Task-Specific Pretraining for Summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 140–145, Online and Punta Cana, Dominican Republic, November.

Song, Y. and Shi, S. (2018). Complementary Learning of Word Embeddings. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4368–4374, 7.

Song, Y. and Xia, F. (2012). Using a Goodness Measurement for Domain Adaptation: A Case Study on Chinese Word Segmentation. In *LREC*, pages 3853–3860.

Song, Y., Cai, D., Zhang, G., and Zhao, H. (2009). Approach to Chinese Word Segmentation based on Character-word Joint Decoding. *Journal of Software*, 20(9):2236–2376.

Song, Y., Tian, Y., Wang, N., and Xia, F. (2020). Summarizing Medical Conversations via Identifying Important Utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 717–729, December.

Song, Y., Zhang, T., Wang, Y., and Lee, K.-F. (2021). ZEN 2.0: Continue Training and Adaption for N-gram Enhanced Text Encoders. *arXiv preprint arXiv:2105.01279*.

Su, J., He, B., Wu, H., Yang, J., Guan, Y., Jing, J., Wang, H., and Yu, Q. (2019). Annotation Scheme and Corpus Construction for Cardiovascular Diseases Risk Factors from Chinese Electronic Medical Records. *ACTA AUTOMATICA SINICA*, 45(zdhxb-45-2-420):420.

Tian, Y., Ma, W., Xia, F., and Song, Y. (2019). ChiMed: A Chinese Medical Corpus for Question Answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 250–260.

Tian, Y., Song, Y., and Xia, F. (2020a). Joint Chinese Word Segmentation and Part-of-speech Tagging via Multi-channel Attention of Character N-grams. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2073–2084, December.

Tian, Y., Song, Y., Xia, F., and Zhang, T. (2020b). Improving Constituency Parsing with Span Attention. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1691–1703, November.

Tian, Y., Song, Y., Xia, F., Zhang, T., and Wang, Y. (2020c). Improving Chinese Word Segmentation with Wordhood Memory Networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8274–8285, Online, July.

Tian, Y., Chen, G., Qin, H., and Song, Y. (2021). Federated Chinese Word Segmentation with Global Character Associations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Tian, Y., Song, Y., and Xia, F. (2022). Improving Relation Extraction through Syntax-induced Pre-training with Dependency Masking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, May.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D., and Manning, C. (2005). A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*, pages 168–171.

Wang, R., Zhao, J., Peng, L., Yang, B., Wang, L., and Li, B. (2018). Medical Entity Recognition of Esophageal Carcinoma Based on Word Clustering. In *2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, pages 348–353.

Wang, Q., Zhou, Y., Ruan, T., Gao, D., Xia, Y., and He, P. (2019). Incorporating Dictionaries into Deep Neural Networks for the Chinese Clinical Named Entity Recognition. *Journal of biomedical informatics*, 92:103133.

Wang, N., Song, Y., and Xia, F. (2020). Studying Challenges in Medical Conversation with Structured Annotation. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 12–21, Online, July.

Xia, F. (2000). The Segmentation Guidelines for the Penn Chinese Treebank (3.0).

Xing, J., Zhu, K., and Zhang, S. (2018). Adaptive Multi-task Transfer Learning for Chinese Word Segmentation in Medical Text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3619–3630.

Xiong, Y., Wang, Z., Jiang, D., Wang, X., Chen, Q., Xu, H., Yan, J., and Tang, B. (2019). A Fine-grained Chinese Word Segmentation and Part-of-speech Tagging Corpus for Clinical Text. *BMC medical informatics and decision making*, 19(2):179–184.

Xu, Y., Wang, Y., Liu, T., Liu, J., Fan, Y., Qian, Y., Tsujii, J., and Chang, E. I. (2014). Joint Segmentation and Named Entity Recognition using Dual Decomposition in Chinese discharge summaries. *Journal of the American Medical Informatics Association*, 21(e1):e84–e92.

Xu, D., Zhang, M., Zhao, T., Ge, C., Gao, W., Wei, J., and Zhu, K. Q. (2015). Data-driven Information Extraction from Chinese Electronic Medical Records. *PloS one*, 10(8):e0136270.

Xue, N., Xia, F., Chiou, F.-D., and Palmer, M. (2005). The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. *Natural language engineering*, 11(2):207–238.

Xue, Y., Liang, H., Wu, X., Gong, H., Li, B., and Zhang, Y. (2012). Effects of Electronic Medical Record in a Chinese hospital: A Time Series Study. *International journal of medical informatics*, 81(10):683–689.

Zan, H., Liu, T., Niu, C., Yueshu, Z., Kunli, Z., and Sui, Z. (2020). Construction and Application of Named Entity and Entity Relations Corpus for Pediatric Diseases. *Journal of Chinese Information Processing*, 34(5):19.

Zhang, X., Song, Y., and Fang, A. C. (2010). Term Recognition Using Conditional Random Fields. In *Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010)*, pages 1–6.

Zhang, M., Zhang, Y., Che, W., and Liu, T. (2014). Type-Supervised Domain Adaptation for Joint Segmentation and POS-Tagging. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 588–597, Gothenburg, Sweden, April.

Zhang, S., Kang, T., Zhang, X., Wen, D., Elhadad, N., and Lei, J. (2016). Speculation Detection for Chinese Clinical Notes: Impacts of Word Segmentation and Embedding Models. *Journal of biomedical informatics*, 60:334–341.

Zhang, J., Li, J., Jiao, Z., and Yan, J. (2019). Overview of CCKS 2018 Task 1: Named Entity Recognition in Chinese Electronic Medical Records. In *China Conference on Knowledge Graph and Semantic Computing*, pages 158–164. Springer.

Zhang, H., Zong, Y., Chang, B., Sui, Z., Zan, H., and Zhang, K. (2020a). Medical Entity Annotation Standard for Medical Text Processing). In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 561–571, Haikou, China, October.

Zhang, T., Wang, Y., Wang, X., Yang, Y., and Ye, Y. (2020b). Constructing Fine-grained Entity Recognition Corpora based on Clinical Records of Traditional Chinese Medicine. *BMC medical informatics and decision making*, 20(1):1–17.