Schools of Computer Science and Technology, Harbin Institute of Technology

Web Intelligence Lab, Research center of Language Technology

A Summary of Annotation Guidelines of Named Entities in Chinese Electronic Medical Records

Abstract: We summarize annotation guidelines of medical named entities in Chinese electronic medical records (CEMRs). EMRs are complete records of the patient's health status which contains a wealth of medical knowledge, and the construction of medical named entity (NE) annotated corpus on EMR is significant to the knowledge mining research in the clinical domain. In view of the fact that named entity annotated corpus on Chinese electronic medical records is still in blankness, under the guidance of professional doctors, we have developed the annotation guidelines on CEMRs based on I2B2 challenge with the analysis of a large number of CEMRs.

Key words: Chinese electronic medical records (CEMRs); named entity; annotated corpus; annotation guidelines; inter-annotator agreement (IAA)

1 Introduction

Electronic medical records (EMRs), which are captured by medical staffs using health information systems in clinical medical activities, contain words, symbols, charts, graphs, numbers, and images detailing the health conditions of patients. These EMRs can be stored, managed, and reproduced [1]. Medical knowledge in EMRs consists of medical named entities and their relations [2]. Therefore, recognizing these entities and relations from EMRs by information extraction technologies is benefit to build intelligence softwares for assisting medical staffs to accomplish clinical medical activities. In EMRs, phrases that describs diseases, symptoms, tests taken by patients, and treatments are defined as medical named entities.

2 Categories of Medical Named Entities

CEMRs include discharge summaries, progress notes, doctor-patient agreements, ultrasound reports, and so on. The discharge summaries and the progress notes are two most important types of free text. To illustrate categories of medical named entities, structural features of these two types of free texts are shown in Figures 1 and 2.

In the discharge summary, diagnosis made by doctors is recorded in the section of diagonose. Conditions of patients and test results are shown in patient conditions before and after discharge to support diagnosis and to confirm treatment effects. In the treatment procedure, treatment methods are shown in detailed. Follow-up treatments are indicated in the discharge summary dictated.

In the progress note, symptoms containing subject feelings, physical tests and results, and auxiliary tests and results are shown in clinical pathological characters. Clinical presumptive diagnosis is made by doctors on the basic of the clinical pathological characters. Treatments in the

treatment plan are taken based on the clinical presumptive diagnosis and the clinical pathological characters.

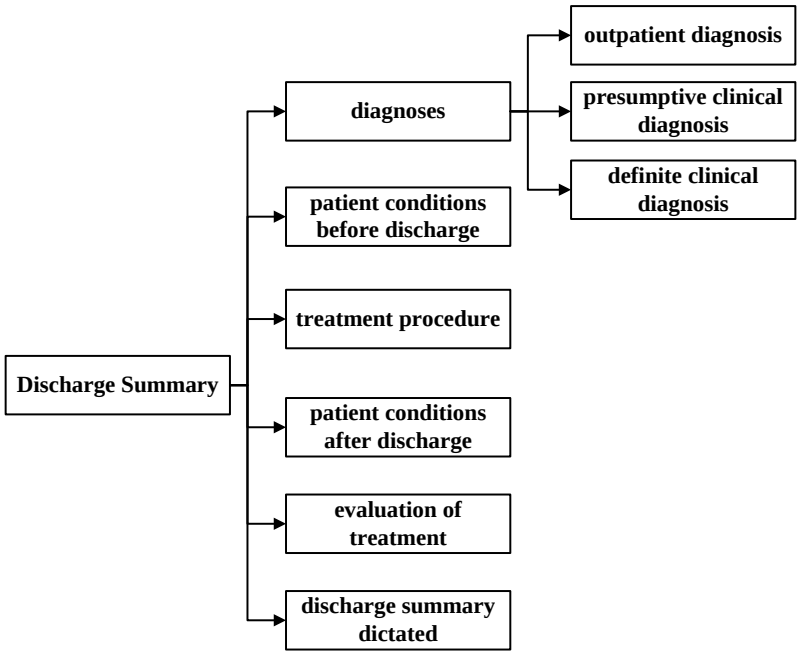

Figure 1. Structures of discharge summary

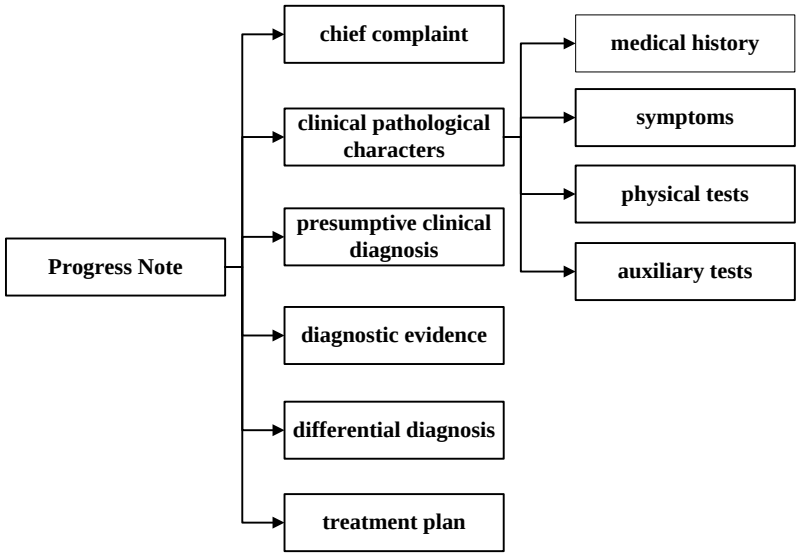

Figure 2. Structures of progress note

In general, clinical medical activities contain using (or not) tests to find symptoms of diseases, making diagonosis, and making treatments accoding to the diagonosis. Therefore, we classify medical named entities into four categories.

The first category of named entities are phrases describing diseases that refer to reasons to conditions that are wrong with the patient, and diagnosis which is made by doctors according to these conditions, where the diseases can be cured or improved.

The second category involves phrases showing symptoms of diseases. The symptoms used to discreb discomfort feelings and abnormal test results are divided into two sub-categories: private prosecution symptoms and abnormal test results.

The third category contains phrases showing various tests that consist of test equipments, test procedures, and test items which are used to obtain more abnormal symptoms caused by diseases for supporting diagnosis.

The last category refers to treatments such as drugs and surgeries which are used to cure diseases and to alleviate or improve symptoms.

Relations amongs these medical named entities are shown in Figure 3.

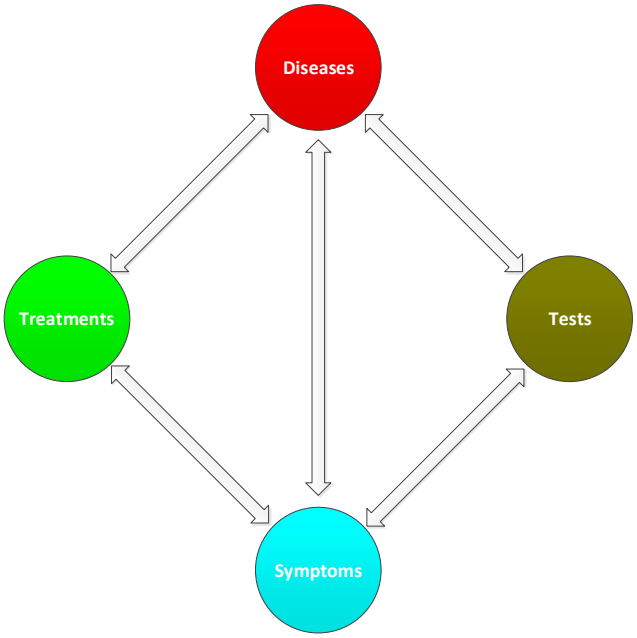

Figure 3. Relations amongs these named entities

The relations consist of those between diseases and treatments, those between treatements and symptoms, those between tests and diseases, those between diseases and sypmtoms, those between tests and symptoms.

3 Annotation Guidelines of Medical Named Entity

Four categories of medical named entities are defined in this annotation guideline, namely, diseases, symptoms, tests and treatments. Scopes of each category of medical named entities are defined according to semantic types in UMLS [3].

Annotators must follow three main principles when tagging medical named entities: (1) entities do not overlap; (2) entities do not nest; (3) entities contain no punctuation.

3.1 Disease

Phrases that describing causes that are wrong with the patient or diagnosis made by doctors are defined as disease named entities. Semantic types in UMLS contain disease or syndrome, injury or poisoning, congenital abnormality, virus/bacterium, pathologic function, cell or molecular dysfunction, acquired abnormality, anatomic abnormality, neoplastic process, and so on.

☐ Annotation examples of disease named entities

a) Disease names

✧ 冠心病 **(coronary heart disease)**

b) Viruses and bacteria

✧ 结核分枝杆菌（**mycobacterium tuberculosis)**

c) Trauma and allergy history

✧ 遗传病史（history of genetic disease)

d) Diseases found by tests

✧ 头 MRI 示: 腔隙性脑梗死（head MRI shows: lacunar infarction）

3.2 Symptom

Symptoms refer to uncomfortable feelings or abnormal test results. Semantic types in UMLS contain signs, mental or behavioral dysfunction, and abnormal test results.

3.2.1 Complaint symptoms

Complaint symptoms refer to uncomfortable or abnormal feelings which are told by patients or somebody who knows patient's conditions to doctors.

☐ Annotation examples of complaint symptoms

a) patient symptoms

✧ 耳鸣（**tinnitus)**

b) Status of behaviors or mental

✧ 反应迟钝（**slow response)**

3.2.2 Abnormal test results

Abnormal test results show explicity and abnormal changes in patients' bodies which are found by doctors or equipments.

☐ Annotation examples of abnormal test results

a) Sign

✧ 构音障碍（dysarthria)

b) Abnormal results

✧ 胸片示左肺炎症病变（Chest X-ray shows left lung inflammatory lesion)

3.3 Test

For discovering diseases or symptoms, related information of diseases or symptoms should be found by test processes, test items, and test euiptments. Semantic types in UMLS contain laboratory procedures and diagnostic procedures.

For avoiding ambiguities of annotations, test named entities are only three categories as follows.

(1) Auxiliary examinations, treatment plans, and examinations in the treatment procedure.

(2) Fluid examinations, physiological tests, physical signs that are followed by numbers.

(3) Phrases containing key words such as “显示(show)” and “测定(measure)”.

☐ Annotation examples of test

a) Auxiliary examinations, treatment plans, and tests in the treatment procedure.

✧ 心电图 (electrocardiography)

b) Fluid examinations, physiological tests, physical signs that are followed by numbers.

✧ 血常规 (blood routine examination)

c) Key words in contexts such as “显示(show)”, “试验 (test)” and “测定(measure)”.。

✧ 前庭功能试验 (vestibular function test)

3.4 Treatment

Medical named entities about treatments consist of treatment procedures, drugs, and intervening measures which are given to patients for allivating diseases or symptoms. Semantic types in UMLS contain pharmacologic substance, therapeutic or preventive procedure, drug delivery device, medical device, steroid, biomedical or dental material, antibiotic, clinical drug, and so on.

☐ Annotation examples of treatment

a) Drug names

✧ 20mg 布洛芬 (20mg **Ibuprofen**)

b) Treatment procudures

✧ 支具固定 (**braces fixation**)

c) Other clear treatment information

✧ 穿弹力袜 (wear **stretch sock**)

According to these guidelines above, professional annotators with medical backgrouds build a corpus including 992 EMRs annotated. IAA is used to evaluate the quality of the corpus [4,5]. And IAA of the corpus is larger than 92 percents, which means that the corpus tagged under the guidelines can be thought available [6].

4 Conclusions and future work

We proposed annotation guidelines for tagging medical named entities in Chinese EMRs. There are four categories of medical named entities: disease, symptoms, test, and treatment. Annotation results showed the reliability of the guidelines. However, there are also shortcomings in the guidelines. For example, entity category granularities defined are not enough to tag all entities clearly. Tagging granularity of treatment entities is too coarse-grained because drugs, surgeries, and treatment procedures are all annotated as treatment entities.

References

[1]中华人民共和国卫生部. 电子病历基本规范（试行）(People's Republic of China Ministry of Health. Basic Rules of Electronic Medical Records (Trail)), http://www.moh.gov.cn/publicfiles/business/htmlfiles/mohyzs/s3585/201003/46174.htm, 2010.

[2]R. C. Wasserman. Electronic medical records (EMRs), epidemiology, and epistemology: reflections on EMRs and future pediatric clinical research. Acad. Pediatr., 2011, 11(4): 280–287

[3]O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res., Jan. 2004, 32(Database issue): 267-270

[4]Ogren P V, Savova G, Buntrock JD, Chute CG. Building and Evaluating Annotated Corpora for Medical NLP Systems. In: AMIA AnnuSympProc.Vol 2006, American Medical Informatics Association; 2006:1050.

[5]Roberts A, Gaizauskas R, Hepple M, et al. Building a semantically annotated corpus of clinical texts. Journal of biomedical informatics, 2009,42(5):950-66.

[6]Artstein R, Poesio M. Inter-Coder Agreement for Computational Linguistics. Computational Linguistics,2008,34(4):555-596.