Supplementary A: Some statistics on the collected Chinese discharge summaries

Table S1 the number of discharge summaries of different diseases as used

Diseases	No. of Records	Diseases	No. of Records
Cancers and tumors	50	Diabetes	20
Cardiovascular diseases	35	ENT diseases	15
Gynecological diseases	32	Stones	14
Fractures	24	Anemia	13
Pediatrics diseases	23	Cerebrovascular diseases	11
Leukemia	22	Hepatitis	12
Bronchitis	21	Other	44

Table S2 information for our corpus

No.	Value
The average number of characters for each discharge summary	438.19
The average number of sentences for each discharge	12.19
The average number of medical problems for each discharge	26.22 (8811)
The average number of treatments for each discharge	3.33 (1188)
The average number of medications for each discharge	2.33 (782)
The average number of tests for each discharge	3.87 (1299)
The average number of anatomy for each discharge	12.60 (4234)
The average percentage of characters of non- named entities for each discharge	32.51%

Supplementary B: Detailed annotation guideline for name entity

Annotation guidelines

In order to make the annotation guideline, three experts in computational linguistics randomly picked 10 discharge summaries with annotations from the doctors and collaboratively designed an annotation guideline that corrects the inconsistency in doctors' annotations. In general, our annotation guideline is based on I2B2 challenge 2010's task specification [7]. Namely, named entities (markable participants) in a discharge summary include

Medical Problems: patients' diseases, symptoms, test abnormality, etc. For instance, "感冒 (cold)", "咳嗽 (cough)", "白细胞升高 (increase of white blood cells count)" are marked as problems.

Treatments: treatments other than drug names that aim to cure diseases or alleviate symptoms. For instance, "支架 (stent)", "降血糖 (lower blood sugar)", "抗炎 (anti-inflammation)", "胃癌切

除术 (resection of gastric carcinoma)" are marked as treatments.

Medication: drugs used in treatments. For instance, "阿司匹林片 (Aspirin tablet)", "头孢曲松 (Ceftriaxone)" are marked as medications. Note that modes and dosages are not marked as medication entities, like "皮下注射 (hypodermic injection)" and "一次一片 (one tablet each time)".

Tests: medical tests performed for diagnose purposes. For instance, "体温 (anatomy temperature)", "白细胞 (white blood cells)", "乙肝核心抗体 (HBcAb)" are marked as tests. Note that test results (whether normal or abnormal) are not marked, like "阴性 (negative)", "升高 (increase)", etc.

Anatomies: patients' anatomies mentioned in the texts. For instance, "胃 (stomach)", "前降支 (anterior descending)", "支气管 (bronchus)" are marked as anatomy.

Apart from the fact that we are using different tags from I2B2 challenge, there are other notable differences between our annotation principle and the I2B2 task's as well. First, verb phrases are allowed to be included in a markable participant. For instance, "不能提重物" ([T] unable to lift heavy things) ([E] difficulty with lifting heavy things) is marked as a problem and "祛痰" (dispel phlegm) is marked as a treatment. This allows us to include important observations and detailed treatment methods into our marked entities. Nevertheless, we decide that verb phrases should only be included if necessary (i.e. if a markable participant can stand alone, we must not include any verb phrases around it). For instance, in the phrase "开始咳嗽 (start to cough)", only "咳嗽 (cough)" is marked as a problem, because "咳嗽 (cough)" itself already stands for a problem and the verb "开始 (start)" fails our "necessity test". (For comparison, in the phrase "不能提重物" ([T] unable to lift heavy things) ([E] difficulty with lifting heavy things) we cannot find a stand-alone noun phrase for a problem without including the verb "提 (lift)".)

Another remarkable difference is that we allow anatomies to be nested into other markable participants. For instance, in phrases "胸部 CT (chest CT)" and "腹部疼痛 (abdominal pain)", the marked anatomies "胸部 (chest)" and "腹部 (abdominal)" are nested into the larger markable

participants (a test and a problem). Marking anatomy inside of a markable participant which enables us to locate a problem or medical test on specific anatomy certainly benefits us since most phrases of problems and tests include anatomies.

We are also using a slightly complicated way to treat modifiers. In most cases, modifiers of a named entity should be included into the markable participants, regardless of the positions of the modifiers in phrases, such as "严重咳嗽 (severe cough)" or "白细胞升高明显 (significant increase in white blood cell count)". However, if a modifier modifies multiple named entities, we decide not to include the modifier into any markable participants it modifies. For instance, in the phrase "严重咳嗽咳痰 (severe cough and expectoration)", "咳嗽 (cough)" and "咳痰 (expectoration)" are marked as two separate problems and the modifier "严重 (severe)" is not marked at all, because it modifies both the following problems.

Compared to other markable named entities, problems have the most complicated surface structures. To help annotators capture these complicated structures better, we propose a context-free grammar (CFG) to mark some of the problems. Note that the CFG only covers a portion of the problems (mostly with complicated surface structures) and is not intended to be a grammar that covers all possible problem entity structures in the medical records.

1. Terminals: *Modifier* (except assertions like "无" (no), "可疑" (possible), etc.), atomic named entities *Problem*, *Test* and *Anatomy.*
2. Deduction rules.
    a. Problem → Modifiers *Problem* Modifiers
    b. Problem → Subject Problem
    c. Subject → Modifiers *Anatomy** | Modifiers *Test**
    d. Modifiers → *Modifier**

To better understand how the CFG works, we use some examples to show how we can expand the starting symbol (i.e. Problem) using deduction rules in order to match a specific surface form.

1. 淋巴结少许肿大 ([T] lymph nodes a little swelling) ([E] mild lymph node swelling)
    - Problem → Subject Problem → *Anatomy* Problem → 淋巴结 Problem → 淋巴结 *Modifier Problem* → 淋巴结 少许 肿大
2. 轻度二尖瓣三尖瓣肺动脉瓣反流 ([T] mild mitral tricuspid regurgitation pulmonary valve regurgitation) ([E] mild regurgitation in mitral valve and tricuspid and pulmonary artery flap)
    - Problem → *Modifier* Problem → 轻度 Problem → 轻度 Subject Problem → 轻度 *Anatomy1 Anatomy2 Anatomy3* Problem → 轻度 二尖瓣 三尖瓣 肺动脉

瓣 Problem → 轻度 二尖瓣 三尖瓣 肺动脉瓣 反流

3. 血糖波动大 ([T] blood glucose fluctuations sharply) ([E] labile blood sugar)

Problem → Subject Problem → *Test* Problem → 血糖 Problem → 血糖 *Problem Modifier* → 血糖 波动 大

Supplementary C: Detailed annotation guideline for segmentation

Annotation guidelines

1. Medication entities are labeled as single words. For instance, the phrase "硝苯地平控释片 (nifedipine controlled release tablets)" is usually segmented into two words (i.e. "硝苯地平 (nifedipine)" and "控释片 (controlled release tablets)"). However, since we have labeled the whole phrase as a medication entity, the whole phrase must be segmented as a single word.
2. Expressions of body parts are labeled as independent words. For instance, the phrase "腹部疼痛 (abdominal pain)" is labeled as two words (i.e. "腹部 (abdominal)" and "疼痛 (pain)") because "腹部 (abdominal)" is an expression of a body part while "疼痛 (pain)" is an expression of a symptom. However, when a body part expression consists of only one Chinese character, it is wired to label them as separated words. Therefore, in such cases the character together with the expression of symptom (if any) following it forms a single word. For instance, the phrase "腹痛" (which also means abdominal pain) is labeled as a single word since the expression of body part "腹 (abdominal)" contains only one character. Note that only expressions of symptoms can be attached to the single-character body part. In the phrase "肝胆脾肺 (liver, gallbladder, spleen and lung)" is labeled as four separated words.
3. Location words, such as "上 (upper)", "下 (lower)", "左 (left)", "右 (right)", "前 (front)" and "后 (back)", are labeled as single words. For example, the phrase "右胸 (right chest)" is labeled as two words and "左上肢 (left upper limb)") is labeled as three separated words.

In contrast to location words, characters like “针 (pin)”, “病 (disease)”, “术 (operation)” and “片 (tablets)” are not labeled as independent words. They are usually labeled with their preceding characters together as one single word. For instance, the phrases “切除术 (resection)”, “高血压病 (disease of hypertension)” and “多西他赛针 (docetaxel-pin)” are labeled as single words in our annotation.

Supplementary D: Inter-annotator Agreement for name entity and segmentation

Table S3 Pair-wise inter-annotator agreement results for each annotator and the gold standard in named entity annotation

		Problem	Treatment	Medication	Test	Anatomy	All
A1	True positive	5450	747	624	810	3568	11199
	False positive	2487	441	29	183	365	3505
	False negative	3361	558	158	489	666	5232
	Precision	68.67%	62.88%	95.56%	81.57%	90.72%	76.16%
	Recall	61.85%	57.24%	79.80%	62.36%	84.27%	68.16%
	F-measure	65.08%	59.93%	86.97%	70.68%	87.38%	71.94%
	k	65.08%	59.93%	86.97%	70.68%	87.38%	71.94%
A2	True positive	5346	838	677	851	3684	11396
	False positive	2780	513	100	221	388	4002
	False negative	3465	350	105	448	550	4918
	Precision	65.79%	62.03%	87.13%	79.38%	90.47%	74.01%
	Recall	60.67%	70.54%	86.57%	65.51%	87.01%	69.85%
	F-measure	63.13%	66.01%	86.85%	71.78%	88.71%	71.87%
	k	63.13%	66.01%	86.85%	71.78%	88.71%	71.87%
A3	True positive	6835	963	729	1132	4001	13660
	False positive	1004	217	50	122	123	1516
	False negative	1976	225	53	167	233	2654
	Precision	87.19%	81.61%	93.58%	90.27%	97.02%	90.01%
	Recall	77.57%	81.06%	93.22%	87.14%	94.50%	83.73%

	F-measure	82.10%	81.33%	93.40%	88.68%	95.74%	86.76%
	k	82.10%	81.33%	93.40%	88.68%	95.74%	86.76%
B1	True positive	7874	1059	765	1211	4102	15011
	False positive	1111	137	37	71	99	1455
	False negative	937	129	17	88	132	1303
	Precision	87.63%	88.55%	95.39%	94.46%	97.64%	91.16%
	Recall	89.37%	89.14%	97.83%	93.23%	96.88%	92.02%
	F-measure	88.49%	88.84%	96.59%	93.84%	97.26%	91.59%
	k	88.49%	88.84%	96.59%	93.84%	97.26%	91.59%
B2	True positive	7788	1024	760	1255	4143	14970
	False positive	1260	142	28	92	103	1625
	False negative	1023	164	22	44	91	1344
	Precision	86.07%	87.82%	96.45%	93.17%	97.57%	90.21%
	Recall	88.39%	86.20%	97.19%	96.61%	97.85%	91.76%
	F-measure	87.22%	87.00%	96.82%	94.86%	97.71%	90.98%
	k	87.22%	87.00%	96.82%	94.86%	97.71%	90.98%

Index “A” means a doctor annotator and index “B” means an annotator with a background in computer linguistics. A3 means the results from the agreement of the three doctors.

Table S4 Pair-wise inter-annotator agreement results for each annotator and the gold standard in segmentation

		segmentation
1	True positive	69301
	False positive	2013
	False negative	2034
	Precision	0.9718
	Recall	0.9715
	F-measure	0.9716
	k	0.9716
2	True positive	69018
	False positive	2156
	False negative	2317
	Precision	0.9697

Recall	0.9675
F-measure	0.9686
k	0.9686

Supplementary E: A detailed description of some features used in the CRF model

Segmentation results: Word segmentation is an important step for Chinese language processing. The segmentation tool we used is from Microsoft Research Asia. In our task, segmentation results can assist to determine the boundary of named entities. For example, the word "支气管炎 (bronchitis)" will be separated as a word by the segmentator, which helps the NER system to accurately capture the boundary of the entity.

Conjunctions: Conjunctions like "、", "及 (and)", "与 (and)", "和 (and)" and "伴 (with)". These words play an important role in recognizing medical problems with more than one anatomies, such as "胸部及腹部疼痛 (chest *and* abdomen pain)".

Positions: Position indicators like "上 (upper)", "下 (lower)", "左 (left)", "右 (right)", "前 (anterior)" and "后 (posterior)". These words are usually prefixes of an anatomy (e.g. "左肺 (left lung)") and hence parts of named entities.

Radicals: Special radicals convey strong information of a Chinese character. For instance, the radical "疒" denotes diseases, such as "病 (disease)", and "疼 (pain)", while the radical "月" denotes anatomies, such as "腰 (kidney)" and "肘 (elbow)". We used features to indicate whether a character has radicals "疒" or "月".

Thesauruses: thesauruses are essential resources, especially when the training data is limited. We collected three thesauruses from the web and used them as features to enhance the performance of our system. Please see [34] for the collection method in detail.