CCKS Task_2 Summary

En Ouyang

2017-08-03

Outline

- **Introduction**
  - Background
  - Datasets
- **Method**
  - RNN
  - Model
  - Results
- **Discussions**
- **Future works**

Background

CCKS 2017-全国知识图谱与语义计算大会

China Conference on Knowledge Graph and Semantic Computing – 成都, 2017年8月26- 29日

Task_2 电子病历命名实体识别, CNER（ Clinical Named Entity Recognition ）。即对于给定的一组电子病历文档，任务的目标是识别并抽取出与医学临床相关的实体名字（ entity mention ），并将它们归类到预先定义好的类别（ pre-defined categories ），比如疾病、症状、检查等。

Datasets

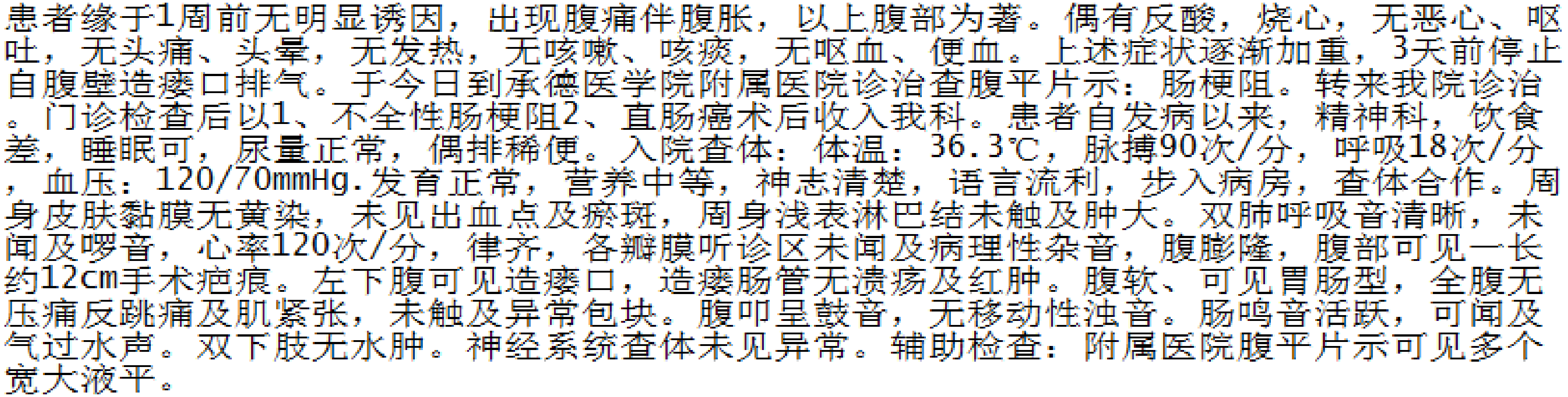
EMRs: 一般项目(general items), 病史特点(medical history), 诊疗经过(diagnosis & treatment), 出院情况(discharge summary)

Entities:身体部位(body), 症状和体征(symptom), 疾病和诊断(disease), 检查和检验(exam), 治疗(treatment)

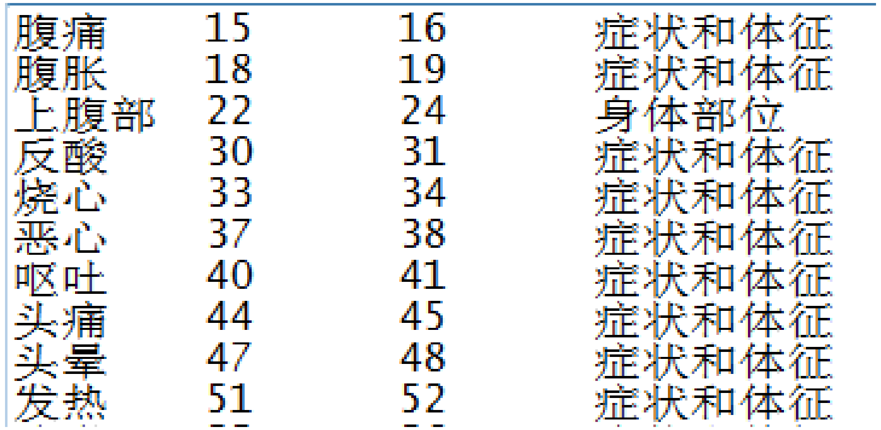
Labeled data(train : test = 300 : 100) & Unlabeled data(2205*4)

	body	symptom	treatment	disease	exam
General items(**300**/100)	**181**/67	**558**/200	**0**/2	**74**/10	**1**/1
Medical history(**300**/100)	**6373**/1771	**4608**/1364	**138**/115	**570**/368	**5902**/1912
Diagnosis & treatment(**299**/99)	**875**/310	**547**/95	**902**/347	**74**/175	**794**/358
Discharge summary(**299**/99)	**3290**/873	**2118**/652	**8**/1	**4**/0	**2849**/872

Datasets



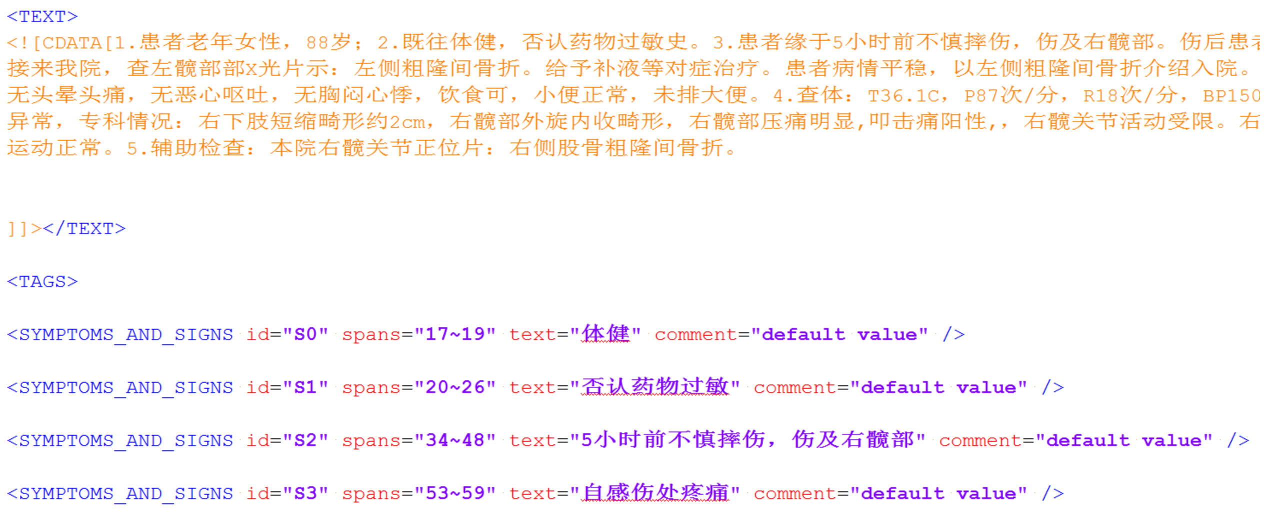
患者缘于1周前无明显诱因，出现腹痛伴腹胀，以上腹部为著。偶有反酸，烧心，无恶心、呕吐，无头痛、头晕，无发热，无咳嗽、咳痰，无呕血、便血。上述症状逐渐加重，3天前停止自腹壁造瘘口排气。于今日到承德医学院附属医院诊治查腹平片示：肠梗阻。转来我院诊治。门诊检查后以1、不全性肠梗阻2、直肠癌术后收入我科。患者自发病以来，精神科，饮食差，睡眠可，尿量正常，偶排稀便。入院查体：体温：36.3℃，脉搏90次/分，呼吸18次/分，血压：120/70mmHg.发育正常，营养中等，神志清楚，语言流利，步入病房，查体合作。周身皮肤黏膜无黄染，未见出血点及瘀斑，周身浅表淋巴结未触及肿大。双肺呼吸音清晰，未闻及啰音，心率120次/分，律齐，各瓣膜听诊区未闻及病理性杂音，腹膨隆，腹部可见一长约12cm手术疤痕。左下腹可见造瘘口，造瘘肠管无溃疡及红肿。腹软、可见胃肠型，全腹无压痛反跳痛及肌紧张，未触及异常包块。腹叩呈鼓音，无移动性浊音。肠鸣音活跃，可闻及气过水声。双下肢无水肿。神经系统查体未见异常。辅助检查：附属医院腹平片示可见多个宽大液平。



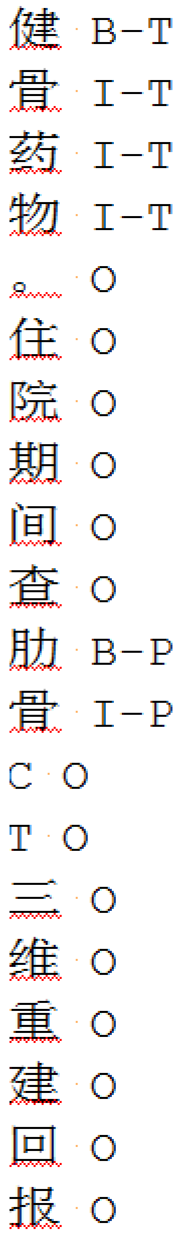
腹痛	15	16	症状和体征
腹胀	18	19	症状和体征
上腹部	22	24	身体部位
反酸	30	31	症状和体征
烧心	33	34	症状和体征
恶心	37	38	症状和体征
呕吐	40	41	症状和体征
头痛	44	45	症状和体征
头晕	47	48	症状和体征
发热	51	52	症状和体征

Data Conversion

- For modification and input
- Raw text + labeled text → MAE file(xml) → CoNLL file



```
<TEXT>
<![CDATA[1.患者老年女性，88岁；2.既往体健，否认药物过敏史。3.患者缘于5小时前不慎摔伤，伤及右髋部。伤后患
接来我院，查左髋部部x光片示：左侧粗隆间骨折。给予补液等对症治疗。患者病情平稳，以左侧粗隆间骨折介绍入院。
无头晕头痛，无恶心呕吐，无胸闷心悸，饮食可，小便正常，未排大便。4.查体：T36.1C，P87次/分，R18次/分，BP150
异常，专科情况：右下肢短缩畸形约2cm，右髋部外旋内收畸形，右髋部压痛明显,叩击痛阳性,，右髋关节活动受限。右
运动正常。5.辅助检查：本院右髋关节正位片：右侧股骨粗隆间骨折。


]]></TEXT>

<TAGS>

<SYMPTOMS_AND_SIGNS id="S0" spans="17~19" text="体健" comment="default value" />

<SYMPTOMS_AND_SIGNS id="S1" spans="20~26" text="否认药物过敏" comment="default value" />

<SYMPTOMS_AND_SIGNS id="S2" spans="34~48" text="5小时前不慎摔伤，伤及右髋部" comment="default value" />

<SYMPTOMS_AND_SIGNS id="S3" spans="53~59" text="自感伤处疼痛" comment="default value" />
```



```
健 B-T
骨 I-T
药 I-T
物 I-T
。 O
住 O
院 O
期 O
间 O
查 O
肋 B-P
骨 I-P
C O
T O
三 O
维 O
重 O
建 O
回 O
报 O
```

Method

Rule-based -> Machine Learning (CRF) -> Deep Learning (RNN, 序列标注)

Advantage:

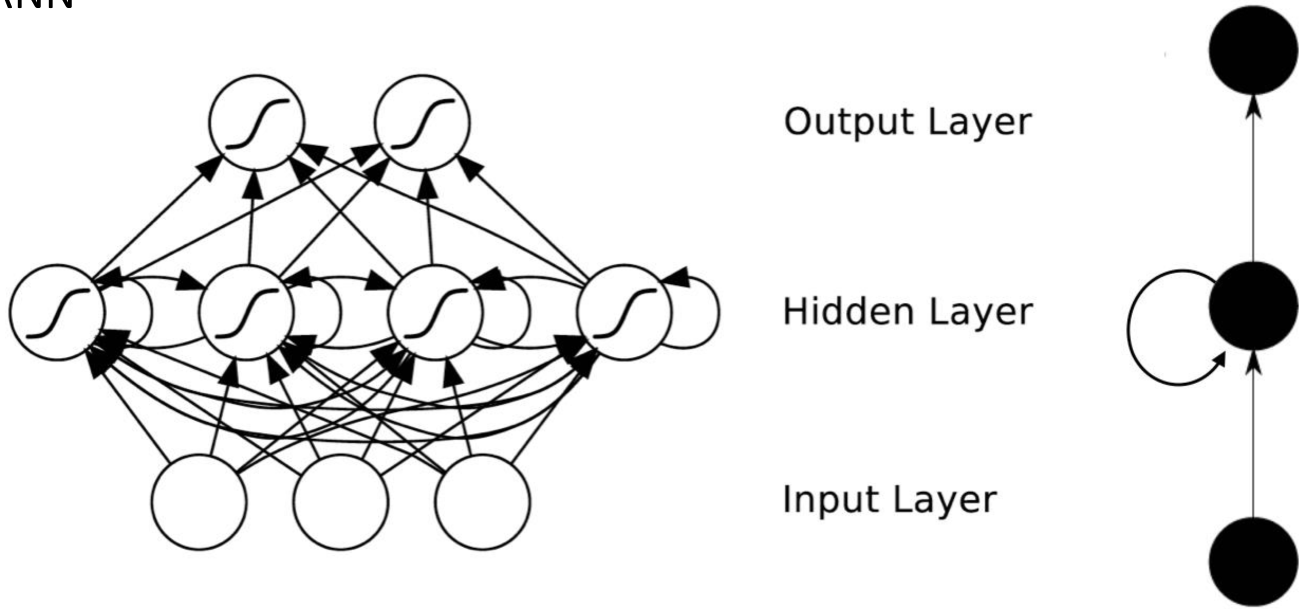
Word embedding (word2vec)

Automatic features learning

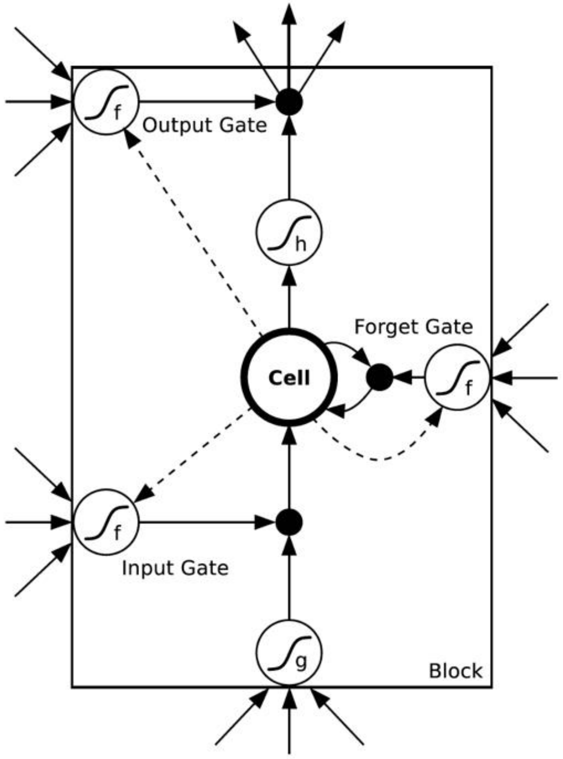
High performance (ML + hand-craft features ~ DL)

RNN (Recurrent Neural Network)

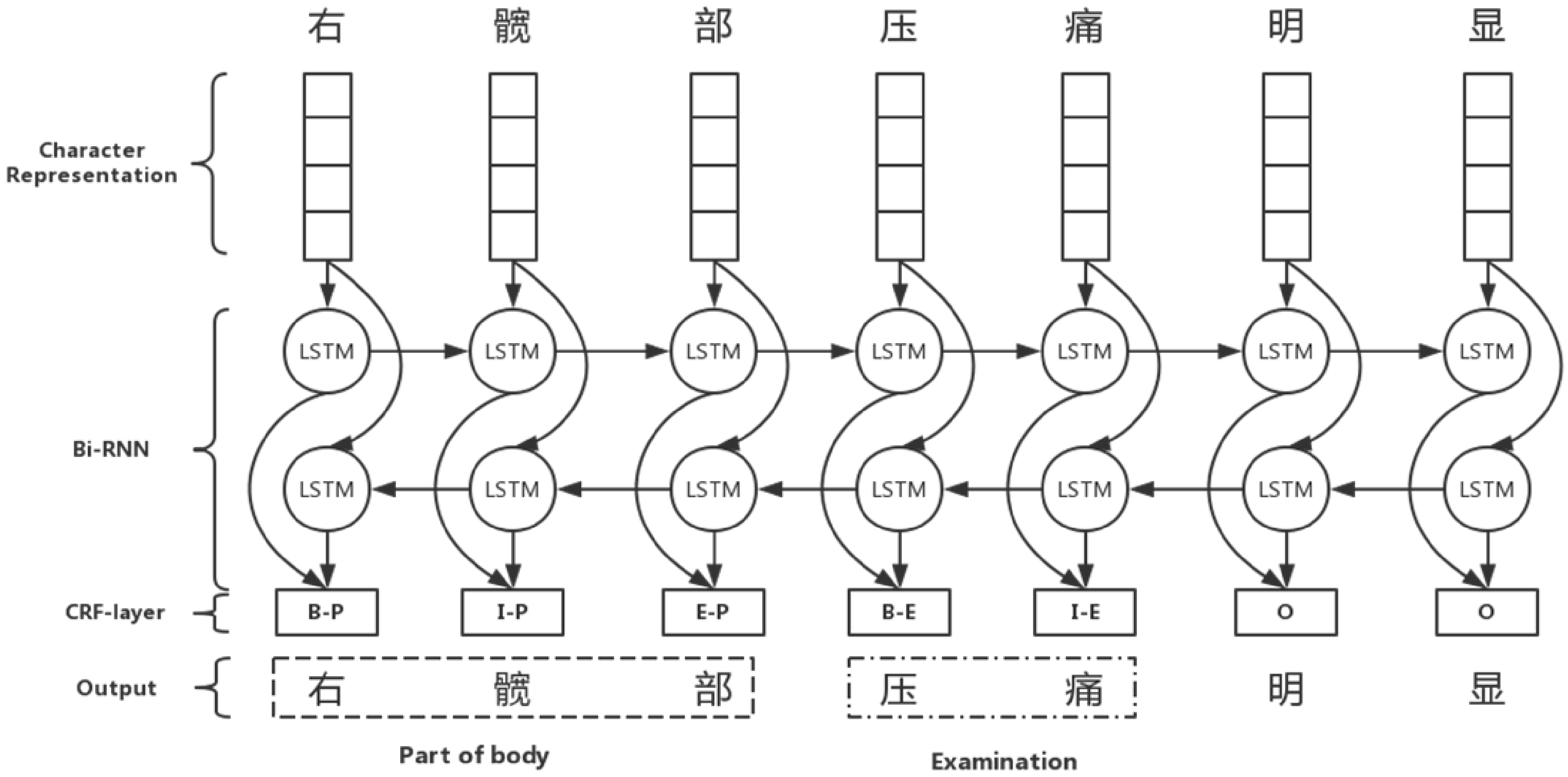
RNN



LSTM (Long-Short Term Memory)



Architecture



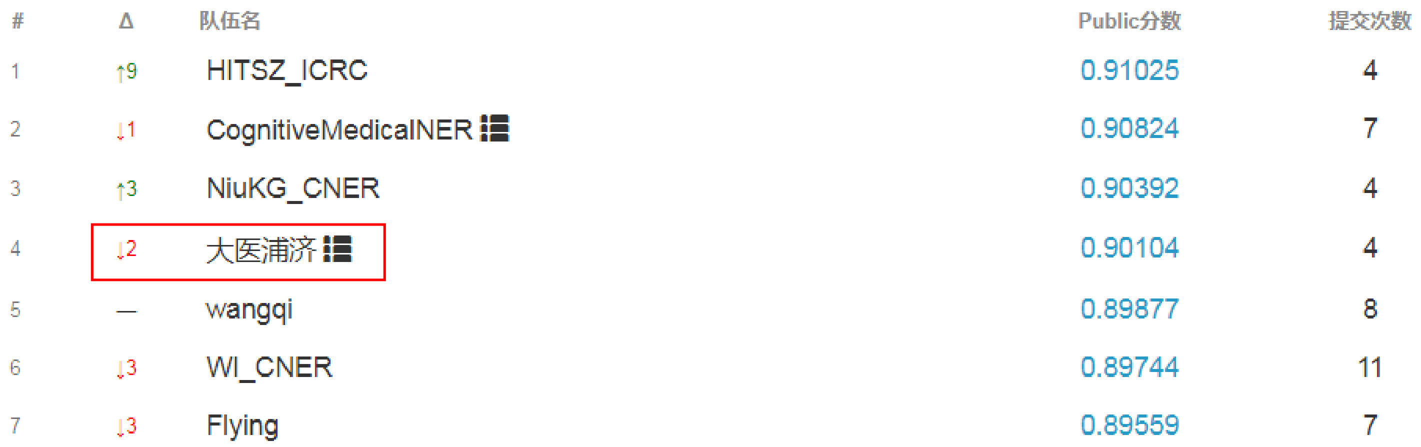
Main contribution

- Bi-RNN-CRF Architecture
- Concatenated n-gram character representation
- Introducing semantic and vocabulary features
- Combined results from models trained using each type of EMRs

Results

Model	body	symptom	disease	exam	treatment	overall
Bi-RNN-CRF	0.8332	0.9473	**0.7622**	0.9274	0.7337	0.8843
Bi-RNN-CRF_N	0.8352	0.9457	0.7470	**0.9328**	0.7497	0.8864
Bi-RNN-CRF_N_F	**0.8377**	0.9481	0.7610	0.9299	0.7551	0.8877
Our model	0.8361	**0.9507**	0.7610	0.9319	**0.7551**	**0.8885***

*the public score is 0.9010



#	Δ	队伍名	Public分数	提交次数
1	↑9	HITSZ_ICRC	0.91025	4
2	↓1	CognitiveMedicalNER	0.90824	7
3	↑3	NiuKG_CNER	0.90392	4
4	↓2	大医浦济	0.90104	4
5	—	wangqi	0.89877	8
6	↓3	WI_CNER	0.89744	11
7	↓3	Flying	0.89559	7

Discussion

- n-gram character representation, sematic features and results combination improve the performance.
- Imbalanced entities distribution influence the performance.
- Decrease of the number of entity categories improve the performance (Discharge summary).
- Inconsistency in annotation
  - Annotation missing
  - Complex entity

Future works

- More effective character representation and feature introduction method
- Extension of EMR categories information (word embedding)
- Annotation strategy with high consistency
- Application (existing works improvement, new tasks)

Tips

- Share out the works and cooperation
- Reading and writing for ‘Introduction’ and ‘Method’
- Discuss more
- Coding capability

Thanks!