中醫醫案術語抽取

實驗結果分享

2018/10/02

林依蓁

資料集

資料提供者：中醫在線 (北京果壳宇宙教育科技有限公司)

簡述：傳統中醫醫案的人工摘要，以利使用者快速找到症狀相符之醫案

檔名：case_summary_180910.xls

檔案格式：Mocrosoft Excel File

內容格式：共6個欄位(id, title, final_content, source_content, tags, summary)

資料筆數：共2107筆

資料集 - 範例

id	3
title	桂枝汤治寒热不适
final_content	李某，女，67岁。每于夏日洗澡后即寒热不适，自感发热而又需披衣裹护。并于不知不觉中汗透衣服，急又换衣，如此三四次方渐趋平复。年年如此，以至夏日视洗澡为畏途。曾住院治疗不效。中医鉴于盛夏炎热，每以涤暑益气，敛汗固表为治，总不见效。三年前来我处就诊。症如前述，脉虚大而迟，苔薄黄。乃处桂枝汤加黄芪、炮附片，1剂效，2剂愈。后每年夏日发作，均来服药2剂即愈。
source_content	{诊次:李某，{性别:女}，{年龄:67岁}。{刻下:每于夏日洗澡后即寒热不适，自感发热而又需披衣裹护。并于不知不觉中汗透衣服，急又换衣，如此三四次方渐趋平复。年年如此，以至夏日视洗澡为畏途。}曾住院治疗不效。中医鉴于盛夏炎热，每以涤暑益气，敛汗固表为治，总不见效。三年前来我处就诊。症如前述，{脉象:脉虚大而迟}，{舌象:苔薄黄}。乃处{处方名:桂枝汤加黄芪、炮附片}，1剂效，2剂愈。后每年夏日发作，均来服药2剂即愈。}
tags	[{"name": "诊次", "child": [{"name": "性别", "content": "女"}, {"name": "年龄", "content": "67岁"}, {"name": "刻下", "content": "每于夏日洗澡后即寒热不适，自感发热而又需披衣裹护。并于不知不觉中汗透衣服，急又换衣，如此三四次方渐趋平复。年年如此，以至夏日视洗澡为畏途。"}, {"name": "脉象", "content": "脉虚大而迟"}, {"name": "舌象", "content": "苔薄黄"}, {"name": "处方名", "content": "桂枝汤加黄芪、炮附片"}], "content": "李某，女，67岁。每于夏日洗澡后即寒热不适，自感发热而又需披衣裹护。并于不知不觉中汗透衣服，急又换衣，如此三四次方渐趋平复。年年如此，以至夏日视洗澡为畏途。曾住院治疗不效。中医鉴于盛夏炎热，每以涤暑益气，敛汗固表为治，总不见效。三年前来我处就诊。症如前述，脉虚大而迟，苔薄黄。乃处桂枝汤加黄芪、炮附片，1剂效，2剂愈。后每年夏日发作，均来服药2剂即愈。"}]
summary	{"main": "寒热不适，发热又需披衣，汗透衣服", "pulse": "脉虚大而迟", "recipe": {"raw": "", "memo": "", "title": "桂枝汤"}, "second": "", "tongue": "苔薄黄", "chinese": "发热", "therapy": "", "western": "", "dialectical": ""}

資料集 - 範例

原文 (final_content)	李某，女，67岁。每于夏日洗澡后即**寒热不适**，自感**发热而又需披衣**裹护。并于不知不觉中**汗透衣服**，急又换衣，如此三四次方渐趋平复。年年如此，以至夏日 视洗澡为畏途。曾住院治疗不效。中医鉴于盛夏炎热，每以涤暑益气，敛汗固表为治，总不见效。三年前来我处就诊。症如前述，**脉虚大而迟**，**苔薄黄**。乃处桂枝汤加黄芪、炮附片，1剂效，2剂愈。后每年夏日发作，均来服药2剂即愈。
摘要 (summary)	{**"main": "寒热不适，发热又需披衣，汗透衣服"**, **"pulse": "脉虚大而迟",** "recipe": {"raw": "", "memo": "", "title": "桂枝汤"}, **"second": "",** **"tongue": "苔薄黄",** "chinese": "发热", "therapy": "", "western": "", "dialectical": ""}

```
public $summaryKey = array(
    'main'        => '主症',
    'second'      => '兼症',
    'tongue'      => '舌象',
    'pulse'       => '脉象',
    'dialectical' => '辨证',
    'therapy'     => '治法',
    'chinese'     => '中病名',
    'western'     => '西病名',
    'title'       => '主方',
      'type'        => '主方类型',
      'raw'         => '主方详情',
      'memo'        => '主方备注'
);
```

預處理

- 移除標點
- 拆分為句
  - 以全形句點(。)將醫案拆分為句(sentence)，以句子作為資料點。
  - 拆分後共得**48901筆**，平均句長32.5個字。
- 詞位標注
  - 標注集：OS
  - 實體類別：SYM(合併主證與兼證)，PUL(脈象)，TON(舌象)
- 逐字標註
  - 共有3235個症狀詞(TON)，其中43個詞沒有完全對應，皆標為 O。
  - 範例
    - content：脈细，苔薄黄
    - tongue：“苔薄黄”
    - 標注結果：脉O细O，O苔TON薄TON黄TON

脉	O
细	O
，	O
苔	TON
薄	TON
黄	TON

預處理

- 移除標點
- 拆分為句
  - 以全形句點(。)將醫案拆分為句(sentence)，以句子作為資料點。
  - 拆分後共得**48901筆**，平均句長32.5個字。
- 詞位標注
  - 標注集：OSBIE
  - 實體類別：SYM(合併主證與兼證)，PUL(脈象)，TON(舌象)
- 逐字標註
  - 共有13452個症狀詞(SYM)，其中671個詞沒有完全對應，皆標為 O。
  - 範例：
    - content：每于夏日洗澡后即寒热不适，自感发热而又需披衣裹护
    - main": "寒热不适，发热又需披衣，汗透衣服"
    - 標注結果：每O于O夏O日O洗O澡O后O即O寒B-SYM热I-SYM不I-SYM适E-SYM，O自O感O发O热O而O又O需O披O衣O裹O护O

```
每 O
于 O
夏 O
日 O
洗 O
澡 O
后 O
即 O
寒 B-SYM
热 I-SYM
不 I-SYM
适 E-SYM
， O
自 O
感 O
发 O
热 O
而 O
又 O
需 O
披 O
衣 O
裹 O
护 O
```

模型訓練

- 模型：條件隨機場(CRF)
- 工具包：python-crfsuite
- 詞性標注系統：NLTK, ICTCLAS
- 資料集：
  - 將症狀,脈象,舌象分別標註為三個資料集
    - 症狀詞資料集：SYM
    - 脈象詞資料集：PUL
    - 舌象詞資料集：TON
  - 將資料集切分為訓練集和測試集
    - 80%作為訓練集(39120筆)
    - 20%作為測試集(9781筆)
- 特徵：字本身, 前一個字, 後一個字, 詞性

```
'word',
'word.isdigit',
'postag',
'-1:word',
'-1:word.isdigit',
'-1:postag',
'+1:word',
'+1:word.isdigit',
'+1:postag'
```

條件隨機場(Conditional Radom Fiels, CRF)

- 機率模型
- 常用於標注或分析序列資料
- 通過可觀測狀態(observeable/labels)判別隱含變量(hidden variables)，其概率亦通過標註集統計得來，是一個判別模型。
- 條件隨機場是邏輯迴歸的串行化版本。邏輯迴歸是用於分類的對數線性模型，條件隨機場是用於串行化標註的對數線性模型。

測試結果

- **資料集: 舌象詞資料集**
- 資料筆數: 2107筆
- 平均句長: 606字

- 演算法: CRF
- 類別標註: OS
- 詞性標註: ICTCLAS

- 詞位標注: 無
- 引用辭典: 無
- 去標點: 無
- 拆句: 無
- 斷詞: 無

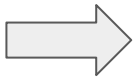
F1 score for training set:

	precision	recall	f1-score	support
O	1.00	1.00	1.00	1305864
S	0.61	0.31	0.41	8932
avg / total	0.99	0.99	0.99	1314796

F1 score for test set:

	precision	recall	f1-score	support
O	1.00	1.00	1.00	320988
S	0.53	0.26	0.35	2136
avg / total	0.99	0.99	0.99	323124

- f1-score 很差
- training set 與 test set 的分數差異不高



Underfitting

測試結果

實驗 #1 : 比較是否拆句 (recall 效果顯著)

實驗變因	不拆句 (2107筆 / 平均606字)	拆句 (48901筆 / 平均32.5字)
控制變因	- 斷詞: 無 - 詞位標注: 無	- 斷詞: 無 - 詞位標注: 無
數值	F1 score for training set: precision recall f1-score support O 1.00 1.00 1.00 1305864 S 0.61 0.31 0.41 8932 avg / total 0.99 0.99 0.99 1314796 F1 score for test set: precision recall f1-score support O 1.00 1.00 1.00 320988 S 0.53 0.26 0.35 2136 avg / total 0.99 0.99 0.99 323124	F1 score for training set: precision recall f1-score support O 1.00 1.00 1.00 1259654 S 0.67 0.86 0.75 12942 avg / total 1.00 0.99 0.99 1272596 F1 score for test set: precision recall f1-score support O 1.00 1.00 1.00 315324 S 0.65 0.84 0.74 3206 avg / total 0.99 0.99 0.99 318530

測試結果

實驗 #2：比較不同句長 (短句的recall效果顯著，overfitting?)

實驗變因	不拆句 (平均句長606.4字)	拆句 (平均句長32.5字)
控制變因	- 資料筆數：2107筆 - 斷詞：無 - 詞位標注: 無	- 資料筆數：2107筆 - 斷詞：無 - 詞位標注: 無
數值	F1 score for training set: precision recall f1-score support O 1.00 1.00 1.00 1305864 S 0.61 0.31 0.41 8932 avg / total 0.99 0.99 0.99 1314796 F1 score for test set: precision recall f1-score support O 1.00 1.00 1.00 320988 S 0.53 0.26 0.35 2136 avg / total 0.99 0.99 0.99 323124	F1 score for training set: precision recall f1-score support O 1.00 1.00 1.00 55059 S 0.78 0.90 0.84 482 avg / total 1.00 1.00 1.00 55541 F1 score for test set: precision recall f1-score support O 1.00 1.00 1.00 13036 S 0.72 0.64 0.68 126 avg / total 0.99 0.99 0.99 13162

測試結果

實驗 #3：比較是否斷詞 (precision 效果顯著)

實驗變因	斷詞：無	斷詞：有
控制變因	- 拆句 (48901筆 / 平均32.5字) - 詞位標注：無	- 拆句 (48901筆 / 平均32.5字) - 詞位標注：無
數值	F1 score for training set: precision recall f1-score support O 1.00 1.00 1.00 1259654 S 0.67 0.86 0.75 12942 avg / total 1.00 0.99 0.99 1272596 F1 score for test set: precision recall f1-score support O 1.00 1.00 1.00 315324 S 0.65 0.84 0.74 3206 avg / total 0.99 0.99 0.99 318530	F1 score for training set: precision recall f1-score support O 1.00 1.00 1.00 909469 S 0.83 0.84 0.84 16602 avg / total 0.99 0.99 0.99 926071 F1 score for test set: precision recall f1-score support O 1.00 1.00 1.00 224580 S 0.80 0.83 0.81 3917 avg / total 0.99 0.99 0.99 228497

測試結果

實驗 #4：增加詞位標註 (影響不大)

實驗變因	詞位標注: 無	詞位標注: 有
控制變因	- 拆句 (48901筆 / 平均32.5字) - 斷詞：無	- 拆句 (48901筆 / 平均32.5字) - 斷詞：無
數值	F1 score for training set: precision recall f1-score support O 1.00 1.00 1.00 1259654 S 0.67 0.86 0.75 12942 avg / total 1.00 0.99 0.99 1272596 F1 score for test set: precision recall f1-score support O 1.00 1.00 1.00 315324 S 0.65 0.84 0.74 3206 avg / total 0.99 0.99 0.99 318530	F1 score for training set: precision recall f1-score support O 1.00 1.00 1.00 1262835 B-SYM 0.68 0.85 0.75 3805 I-SYM 0.67 0.89 0.77 5239 E-SYM 0.67 0.85 0.75 3795 avg / total 1.00 0.99 0.99 1275674 F1 score for test set: precision recall f1-score support O 1.00 1.00 1.00 312143 B-SYM 0.65 0.80 0.72 977 I-SYM 0.64 0.81 0.72 1357 E-SYM 0.64 0.80 0.71 975 avg / total 0.99 0.99 0.99 315452

預測結果案例 - 1

[摘要]:舌淡苔白腻
[原文]:二诊(2009年12月26日):药后纳食增加，再未呕吐，面色较前红润，舌淡苔白略腻，脉细。

[摘要]:舌黯红、苔薄黄
[原文]:察其舌黯红，苔薄黄，左脉弦细，右脉沉滑

[摘要]:舌红，苔少
[原文]:舌红，苔少，脉沉细

True Positiive

預測結果案例 - 2

[摘要]:舌质黯淡、舌体偏大、舌苔白滑腻
[原文]:舌质淡红，舌体偏大，舌苔薄白，脉象沉细缓和

[摘要]:舌质紫黯，舌苔薄白
[原文]:四诊(2009年4月1日):胃脘胀满，大腿沉重，睡眠不佳，双膝疼痛微僵，气喘、胸闷，咳嗽减轻，痰少，双下肢浮肿，舌质黯淡，苔黄厚，脉象脉弦数

[摘要]:舌质黯红、苔厚腻
[原文]:三诊(2002年5月29日):患者诉胸闷显著减轻，心悸偶作，食眠皆正常，大便通畅，小便调，仍觉气短乏力，舌质稍黯、苔白，脉沉

True Positiive False Positive

預測結果案例 - 3

[摘要]:苔微黄腻、舌质黯红
[原文]:按:以失眠为主症并有脑血管病者，临床上除失眠、早醒或间断多醒外，常见头晕、头胀痛或脑响、耳鸣、心烦'易怒，记忆力减退或头颈板滞不适，时手抖手麻、口干苦、大便偏干，面赤或黯红、呆滞缺润彩，苔黄、舌红或偏绛红，脉弦紧

[摘要]:舌黯红而干、苔薄白
[原文]:《类证治裁》中说:“胸痹胸中阳微不运，久则阴乘阳位而为痹结也

[摘要]:苔薄白，舌质黯紫有瘀斑
[原文]:故仍有胸闷，神倦，夜寐欠安，喜太息，口唇青紫，舌黯紫，苔薄白，脉细涩

[摘要]:舌质淡红，无苔
[原文]:四诊(2008年11月13日):药后患者咳嗽明显好转，咳痰色白质稀量少，无咽痒，口舌略干燥，近1周眼睛干涩明显好转，舌质淡红，无苔，脉沉细数

True Positiive False Positive False Negative

測試結果

- **資料集: 脈象詞資料集**
- 資料筆數: 48901筆
- 平均句長: 32.5字

- 演算法: CRF
- 類別標註: O,PUL
- 詞性標註: ICTCLAS

- 詞位標注: 無
- 引用辭典: 無
- 去標點: 無
- **拆句: 有**
- **斷詞: 有**

F1 score for training set:

	precision	recall	f1-score	support
O	1.00	1.00	1.00	913259
PUL	0.86	0.89	0.88	12591
avg / total	1.00	1.00	1.00	925850

F1 score for test set:

	precision	recall	f1-score	support
O	1.00	1.00	1.00	225849
PUL	0.86	0.87	0.86	3132
avg / total	1.00	1.00	1.00	228981

測試結果

- **資料集: 症狀詞資料集**
- 資料筆數: 48901筆
- 平均句長: 32.5字

- 演算法: CRF
- 類別標註: OS
- 詞性標註: ICTCLAS

- 詞位標注: 無
- 引用辭典: 無
- 去標點: 無
- 拆句: 無
- 斷詞: 無

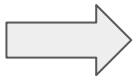
F1 score for training set:

	precision	recall	f1-score	support
O	0.97	0.98	0.97	851330
S	0.72	0.57	0.64	69776
avg / total	0.95	0.95	0.95	921106

F1 score for test set:

	precision	recall	f1-score	support
O	0.96	0.98	0.97	216777
S	0.65	0.53	0.58	16952
avg / total	0.94	0.95	0.94	233729

- f1-score 不佳
- training set 與 test set 的分數差異7%



Underfitting

測試結果

更多實驗:(效果不顯著)

資料集	症狀詞	症狀詞	症狀詞	症狀詞	症狀詞	症狀詞	症狀詞
資料筆數(拆句)	48901筆	48901筆	33752筆	2107筆	2107筆	2107筆	1436筆
詞性標註系統	NLTK	ICTCLAS	NLTK	NLTK	NLTK	NLTK	NLTK
去標點	無	無	無	無	有	無	無
引用辭典	無	無	無	無	無	有	無
詞位標注	BIEO	BIEO	BIEO	BIEO	BIEO	BIEO	BIEO
B-SYM (precision / recall)	0.59 / 0.46	0.60 / 0.46	0.61 / 0.46	0.45 / 0.18	0.45 / 0.12	0.44 / 0.18	0.48 / 0.20
I-SYM (precision / recall)	0.52 / 0.33	0.50 / 0.31	0.52 / 0.36	0.39 / 0.19	0.40 / 0.10	0.40 / 0.20	0.40 / 0.22
E-SYM (precision / recall)	0.60 / 0.47	0.61 / 0.46	0.63 / 0.48	0.45 / 0.18	0.47 / 0.13	0.46 / 0.19	0.49 / 0.21

問題探討(改進方向)

- 資料預處理
  - 逐字標註時，若包含了沒有完全對應的症狀詞，應該整筆移除。目前做法是標注為 O。
  - 以句點拆句之後的句長差異頗大，若超過一定長度則再切短。
- 引入辭典
  - 以辭典資訊作為特徵，是否有更好的方式？目前作法是有出現在辭典檔案中的字就標為 True,反之標為False
- 分詞效能
  - 從分詞工具著手，在分詞階段就取得較好的結果。目前作法是使用中科院 ICTCLAS系統
- 過濾停用字
  - 分詞後過濾停用字 (stop words)

簡報完畢 敬請指教