

知识分析比较试验结果

丁如江 2019.10.30

I. 知识抽取范围比较

评价指标

设由Converter得到的实体集合（包含疾病和症状）为 \mathbb{G} ，由文本匹配得到的为 \mathbb{D} 。则测量以下指标：

1. 集合大小。简单比较提取知识的覆盖面。

- $|\mathbb{G}|$
- $|\mathbb{D}|$

2. 集合相似度。比较提取知识的相似程度。

- $J = \frac{|\mathbb{G} \cap \mathbb{D}|}{|\mathbb{G} \cup \mathbb{D}|}$

3. 集合覆盖率。结合集合相似度比较两个结果的大致包含关系。如果两个集合的 J 较小，而其中一个集合的 F 较大，则可以得出集合的包含关系

- $F_{\mathbb{G}} = \frac{|\mathbb{G} \cap \mathbb{D}|}{|\mathbb{G}|}$
- $F_{\mathbb{D}} = \frac{|\mathbb{G} \cap \mathbb{D}|}{|\mathbb{D}|}$

实验结果

- KB中随机症状集

n 表示症状集的大小,我一共使用了 $n=2,4,6,8,10$ 五种大小的症状集.在每一种大小的症状集下,分别从KB中随机抽取了5组的症状集.

- 文本匹配匹配度最少50%的关键词

n=2					
	1	2	3	4	5
G D	5 98	4225 153	4246 32	26 20	1 17
J	0	0.0106	0.0033	0.0222	0
Fg Fd	0 0	0.0108 0.0314	0.0033 0.4375	0.0385 0.500	0 0

n=4					
	1	2	3	4	5
G D	27 0	4131 0	4158 0	4073 0	1 17
J	0	0	0	0	0
Fg Fd	0 -1(-1表示除0错误)	0 -1	0 -1	0 -1	0 -1

n=6					
	1	2	3	4	5
G D	4083 0	4079 0	4530 0	4508 0	25 0
J	0	0	0	0	0
Fg Fd	0 -1	0 -1	0 -1	0 -1	0 -1

n=8					
	1	2	3	4	5
G D	4307 0	4118 0	4133 0	4112 0	4104 0
J	0	0	0	0	0
Fg Fd	0 -1	0 -1	0 -1	0 -1	0 -1

n=10					
	1	2	3	4	5
G D	4174 0	4555 0	4220 0	4138 0	4086 0
J	0	0	0	0	0
Fg Fd	0 -1	0 -1	0 -1	0 -1	0 -1

- 病例主述集

- 文本匹配匹配度最少50%的关键词

	乏力, 光过敏	前列腺癌术后, 发热, 寒战, 嗜睡	间断发热,尿检异常	纳差恶心,乏力,间断黑便	下腹痛,停经,阴道出血
G D	4520 164	5546 268	4867 83	4520 175	4142 626
J	0.0209	0.0236	0.0073	0.0235	0.0399
Fg Fd	0.0212 0.5853	0.0242 0.5000	0.0074 0.4337	0.0239 0.6171	0.0442 0.2923

- 病例病史集

- 文本匹配匹配度最少50%的关键词

	脱发, 大小便正常, 乏力, 光过敏, 睡眠可, 精神可, 眼睛不适感, 皮肤发痒, 食欲可	激素非依赖性前列腺癌, 嗜睡, 寒战, 体重明显下降, 双下肢疼痛浮肿, 前列腺结节合并左侧颈部淋巴结肿大, 双侧睾丸切除术, 冷汗, 症状稍缓解, 精神欠佳, 体温升高, 肌注柴胡, 体温39°C, PSA有所上升, 小便略费力, 颈部淋巴结转移癌, 多西他赛单药化疗, 发热, 口服头孢类抗生素, 恶心呕吐, 大便干结, 饮食可, 前列腺癌远处转移, 呼之有回应, 口服阿比特龙, 前列腺癌术后, PSA持续明显升高, 前列腺癌	间断发热, 尿检异常,尿潜血, 左肾静脉受压	气短, 小便正常, 乏力, 黑便, 输液治疗, 间断黑便, 睡眠可, 成形黑便, 食欲差, 纳差, 房性心动过速陈旧性下壁及后壁心肌梗死, 胸憋, 体重下降, 精神差, 抑酸输血, 纳差恶心, 胃癌, 全胃壁弥漫性病变性质待查胃潴留	肌壁回声尚均, 疼痛明显时不能站立, 异位妊娠, 食欲好, 心慌, 量中, 停经, 左附件区不均质包块, 盆腔积液, 缓解时活动自如, 精神好, 间断下腹出现牵扯样胀痛, 睡眠欠佳, 下腹痛较前加重, 下腹痛, 大小便正常, 阴道出血, 透声差, 量稍多于平素月经量, 月经规律
G D	4520 0	5555 0	4867 0	4923 0	4142 0
J	0	0	0	0	0
Fg Fd	0 -1	0 -1	0 -1	0 -1	0 -1

- 文本匹配匹配度最少匹配2个关键词

	脱发,大小便正常,乏力,光过敏,睡眠可,精神可,眼睛不适感,皮肤发痒,食欲可	激素非依赖性前列腺癌,嗜睡,寒战,体重明显下降,双下肢疼痛浮肿,前列腺结节合并左侧颈部淋巴结肿大,双侧睾丸切除术,冷汗,症状稍缓解,精神欠佳,体温升高,肌注柴胡,体温39℃,PSA有所上升,小便略费力,颈部淋巴结转移癌,多西他赛单药化疗,发热,口服头孢类抗生素,恶心呕吐,大便干结,饮食可,前列腺癌远处转移,呼之有回应,口服阿比特龙,前列腺癌术后,PSA持续明显升高,前列腺癌	间断发热,尿检异常,尿潜血,左肾静脉受压	气短,小便正常,乏力,黑便,输液治疗,间断黑便,睡眠可,成形黑便,食欲差,纳差,房性心动过速陈旧性下壁及后壁心肌梗死,胸憋,体重下降,精神差,抑酸输血,纳差恶心,胃癌,全胃壁弥漫性病变性质待查胃潴留	肌壁回声尚均,疼痛明显时不能站立,异位妊娠,食欲好,心慌,量中,停经,左附件区不均质包块,盆腔积液,缓解时活动自如,精神好,间断下腹出现牵扯样胀痛,睡眠欠佳,下腹痛较前加重,下腹痛,大小便正常,阴道出血,透声差,量稍多于平素月经量,月经规律
G D	4520 52	5555 294	4867 0	4923 281	4142 88
J	0.0072	0.0341	0	0.0360	0.0035
Fg Fd	0.0072 0.6154	0.0347 0.6667	0 -1	0.0367 0.6584	0.0036 0.1705

II. 疾病收敛能力比较

评价指标

设由Converter得到的实体集合（只有疾病）为 G ，由文本匹配得到的为 D 。则测量以下指标：

- 集合大小。简单比较提取知识的覆盖面。
 - $|G|$
 - $|D|$
- 集合相似度。比较提取知识的相似程度。
 - $J = \frac{|G \cap D|}{|G \cup D|}$
- 集合覆盖率。结合集合相似度比较两个结果的大致包含关系。如果两个集合的 J 较小，而其中一个集合的 F 较大，则可以得出集合的包含关系。
 - $F_G = \frac{|G \cap D|}{|G|}$
 - $F_D = \frac{|G \cap D|}{|D|}$

实验结果

- 文本匹配度为最少匹配50%关键词

	1	2	3	4	5
G D	1306 0	0(unsat) 0	920 0	1150 0	0(unsat) 0
J	0	-1	0	0	-1
Fg Fd	0 -1	-1 -1	0 -1	0 -1	-1 -1

- 文本匹配度为最少匹配2个关键词

	1	2	3	4	5
G D	1306 15	0(unsat) 76	920 0	1150 70	0(unsat) 33
J	0.0046	0	0	0.0226	0
Fg Fd	0.0046 0.4000	-1 0	0 -1	0.0235 0.3857	-1 0