MedTruth 论文介绍

1 背景

近年来，知识图谱在医疗领域得到了不少应用。医学知识图谱由包含头部实体、尾部实体以及它们之间的关系的医学知识三元组组成。但现有的知识图谱还缺乏一个重要组成部分：知识三元组的条件。

由于医学领域知识的不确定性和复杂性，知识三元组与某些特定的条件密切相关，如性别、年龄等。例如，三元组（胸痛、症状-疾病、乳腺增生）在性别（女性）的条件下应该比性别（男性）更容易被检索到，而三元组（感冒、疾病-药物、小儿用药）与年龄（10岁）的条件关系更大。因此，补充医学知识图谱中的条件信息至关重要。

电子病历（EMR）是患者健康信息和医学知识的集合，其中包含有价值的相关信息。因此，它可以成为发现医学知识条件的优质资源。但是，由于规范化、隐私问题等诸多原因，EMR的可用数据量是有限的。有限的数据量会导致重要的病情信息的丢失和一些不准确的挖掘知识。

同时，在线医疗健康问答社区的迅速兴起，促进了医疗知识和信息的互通，产生了大量的医疗问答数据，对知识条件信息发现有很大的帮助。但与EMR数据不同的是，在线医疗问答数据的质量很难保证，因为网站用户提供的建议或诊断都是基于患者有限的描述，而且回答者的专业水平也参差不齐。如果不分青红皂白地采用所有的QA数据，可能会引入大量的噪声，降低发现的病情质量。

作者提出了一种方法，MedTruth。试图将电子病历（EMR）信息作为可靠数据源，对不可靠数据源（医疗问答）做半监督学习，抽取可靠的信息。同时，对相关条件信息的可靠性做出判断。

2 动机

1. 能够有效利用可靠信息对不可靠信息的质量作出判断
2. 能够考虑到出现在同一病例(人)中的病症间的关系(病症共现关系)
3. 能够利用症状与诊断构成的知识图谱的信息
4. 能够有效判断条件信息对知识三元组的关联(置信)程度

3 关键概念介绍

3.1 关键符号介绍

符号	定义
f_m	第m个三元组
c_n	第n个条件
$P(m,n)$	第n个条件之于第m个三元组的置信度
v_m^*	第m个三元组对应的真相向量
u_n	第n个条件对应的条件向量
ω_k	第k个源的置信度
F_k	第k个源提供的三元组和条件
F_{ref}	可靠源提供的三元组和条件
C_m^n	三元组m和条件n出现的源中的所有三元组

其中，

- 源是：单个病历或者单次医疗问答的数据
- 可靠源是：来自病历的源
- 不可靠源是：来自医疗问答中的源

3.2 例子

如下：

Table 2: An Example of QA Data

Item	Content
GENDER	Male
AGE	66
QUESTION	I got hemoptysis, cough, chest pain, general malaise, trembling. What caused these symptoms?
ANSWER	According to your situation, consider the bronchiolitis or bronchitis that causes capillary bleeding.
...	...

- 三元组如：(chest pain, symptom-disease, bron- chiolitis), (chest pain, symptom-disease, bronchitis)
- 条件如：gender(male) and age(66)
- 共现关系：在同一个病历/问答中出现的三元组可以称之为共现

3.3 输入

- 从可靠源(病历)和不可靠源(问答网站)中提取的三元组和条件
- 将来自可靠源(病历)的数据标记为可靠，将来自不可靠源(问答网站)的数据标记为不可靠
- 知识图谱表示学习产生的embedding(见6.1.2节)
- 利用共现信息通过CBOW模型产生的embedding(见6.1.2节)

3.4 输出

- 算法估计的第k个源的置信度(可靠度)$\{\omega_k\}$
- $P(m,n)$，第n个条件之于第m个三元组的置信度

4 目标函数

$$\min_{\{\omega_k\},\{v_m^*\},\{u_n\}} \sum_{k=1}^{K} \frac{\omega_k}{|F_k|+\lambda|\Delta_k|} \sum_{(m,n)\in F_k} (1+\lambda^*) \left\{ (v_m^* - u_n)^2 + \mu \sum_{i\in C_m^n} \alpha_i (v_m^* - x_i)^2 \right\}$$

算法试图将信息源的可靠性ω_k、病症的共现信息和病症关联信息x_i(见6.1.2节)、病症发生的条件信息同时纳入考虑。

公式比较复杂，我们拆分来研究：

5 简化公式

$$\min \quad \sum_{k=1}^{K} \omega_k \sum_{(m,n)\in F_k} \left\{ (v_m^* - u_n)^2 + \sum_{i\in C_m^n} (v_m^* - x_i)^2 \right\}$$

即：

$$\min \sum_{k=1}^{K\text{个源}} \text{第}k\text{个源的置信度} \cdot \sum_{(\text{三元组}_m,\text{条件}_n)\in\text{第}k\text{个源提供的三元组与条件}} \left\{ \left(\text{真相向量}_m^* - \text{条件向量}_n\right)^2 + \sum_{i\in\text{三元组}_m\text{和条件}_n\text{出现的源中的所有三元组}} \left(\text{真相向量}_m^* - \text{预训练向量}_i\right)^2 \right\}$$

5.1 公式解释

5.1.1 第一部分

$$(v_m^* - u_n)^2$$

即：

$$\left(\text{真相向量}_m^* - \text{条件向量}_n\right)^2$$

目的是使真相向量与条件向量更加相近。

5.1.2 第二部分

$$\sum_{i\in C_m^n} (v_m^* - x_i)^2$$

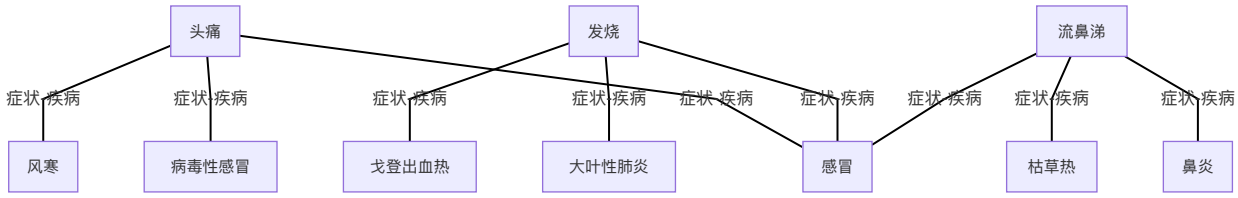
即：

$$\sum_{i\in\text{三元组}_m\text{和条件}_n\text{出现的源中的所有三元组}} \left(\text{真相向量}_m^* - \text{预训练向量}_i\right)^2$$

其中，预训练向量x_i的定义为：

$$x_i = [t_i : \frac{1}{2}(e_h + e_t)]$$

e_h与e_t为，**知识三元组构造的知识图谱**(如下图)，通过表示学习得到的嵌入向量embedding，e_h为头节点，e_t为尾节点。



t_i为，将三元组之于数据源看做文字之于句子(同一病例中的多个三元组如6.2.2节配图)，利用CBOW模型，学习到包含上下文环境信息的嵌入向量embedding。

> CBOW模型参考网页：
>
> https://www.zhihu.com/question/44832436/answer/266068967
>
> https://blog.csdn.net/u010665216/article/details/78724856

通过这种方式，作者试图将病症知识图谱的语义信息和来自每个病例身上，不同病症之间共现的关联信息，进行糅合。

5.1.3 第三部分

为了判断不同信息来源的可靠程度，作者引入参数ω_k，值越大，意味着来源的可靠性越高，运算中相关向量会以更快速度靠近。

5.2 进一步补充公式

5.2.1 考虑不同信息来源(质量)

为了将可靠来源(电子病历)和不可靠来源(在线医疗问答)的数据进行区分，或者说更有效的利用可靠信息对不可靠信息的质量进行判断，可以对公式做如下改变(增加$(1+\lambda^*)$)：

$$\min \quad \sum_{k=1}^{K} \omega_k \sum_{(m,n)\in F_k} (1+\lambda^*) \left\{ (v_m^* - u_n)^2 + \sum_{i\in C_m^n} (v_m^* - x_i)^2 \right\}$$

即：

$$\min \quad \sum_{k=1}^{K\text{个源}} \text{第}k\text{个源的置信度} \cdot \sum_{(\text{三元组}_m,\text{条件}_n)\in\text{第}k\text{个源提供的三元组与条件}} (1+\lambda^*) \left\{ \left(\text{真相向量}_m^* - \text{条件向量}_n \right)^2 + \sum_{i\in\text{三元组}_m\text{和条件}_n\text{出现的源中的所有三元组}} \left(\text{真相向量}_m^* - \text{预训练向量}_i \right)^2 \right\}$$

其中，

$$\lambda^* = \begin{cases} \lambda, & (\text{三元组}_m, \text{条件}_n) \in \text{可靠来源} \\ 0, & (\text{三元组}_m, \text{条件}_n) \notin \text{不可靠来源} \end{cases}$$

λ为作者设置的超参数，用于调节半监督的程度。

可见，当数据来源于可靠源时，公式有更大权重。相比于不可靠源数据，真相向量_m将更快的接近于条件向量_n，同时也将更快的接近预训练向量_i。

5.2.2 自注意力机制

在与病症相关的大量病症中，并不是所有病症的影响都一样，作者引入注意力机制改善这种问题(α_i)：

$$\min \quad \sum_{k=1}^{K} \omega_k \sum_{(m,n)\in F_k} (1+\lambda^*) \left\{ (v_m^* - u_n)^2 + \sum_{i\in C_m^n} \alpha_i (v_m^* - x_i)^2 \right\}$$

即：

$$\min \quad \text{第}k\text{个源的置信度} \cdot \sum_{(\text{三元组}_m,\text{条件}_n)\in\text{第}k\text{个源提供的三元组与条件}} (1+\lambda^*) \left\{ \left(\text{真相向量}_m^* - \text{条件向量}_n \right)^2 + \sum_{i\in\text{三元组}_m\text{和条件}_n\text{出现的源中的所有三元组}} \alpha_i \left(\text{真相向量}_m^* - \text{预训练向量}_i \right)^2 \right\}$$

其中：

$$\alpha_i = \frac{exp(x_i^T x_m)}{\sum_{j\in c_m^n} exp(x_j^T x_m)}$$

为接近当前运算的三元组赋予更高的权重，同时也起到归一化的作用。

效果如下：

(coronary heart disease, chest pain)	(coronary heart disease, chest pain) (coronary heart disease, chest distress) (heart disease, chest pain) (heart disease, chest distress) (disc herniation, pain) (disc herniation, tenderness) (bone hyperplasia, pain) (bone hyperplasia, tenderness)

6 完整公式

$$\min_{\{\omega_k\},\{v_m^*\},\{u_n\}} \sum_{k=1}^{K} \frac{\omega_k}{|F_k|+\lambda|\Delta_k|} \sum_{(m,n)\in F_k} (1+\lambda^*) \left\{ (v_m^* - u_n)^2 + \mu \sum_{i\in C_m^n} \alpha_i (v_m^* - x_i)^2 \right\}$$

$$s.t. \quad \sum_{k=1}^{K} \exp(-\omega_k) = 1$$

$$where \quad \Delta_k = F_k \cap F_{ref} \ , \ \lambda^* = \begin{cases} \lambda, & (m,n) \in \Delta_k \\ 0, & (m,n) \notin \Delta_k \end{cases}$$

其中，μ用于在条件向量和预训练的三元组向量之间权衡。

7 计算

7.1 真相向量$\{v_m^*\}$计算

将$\{u_n\}$、$\{\omega_k\}$固定，当做常量，得到：

$$\min_{\{v_m^*\}} \sum_{m=1}^{M} \sum_{k\in K_m} \sum_{n\in N_m^k} \frac{\omega_k \left(1+\lambda^*\right)}{|F_k| + \lambda \left|\Delta_k\right|} \left\{ \left(v_m^* - u_n\right)^2 + \mu \sum_{i\in C_m^n} \alpha_i \left(v_m^* - x_i\right)^2 \right\}$$

求导并赋值为0后可推导出：

$$v_m^* = \frac{\sum_{k\in K_m} \sum_{n\in N_m^k} \frac{\omega_k(1+\lambda^*)}{|F_k|+\lambda|\Delta_k|} \left(u_n + \mu \sum_{i\in C_m} \alpha_i x_i\right)}{\sum_{k\in K_m} \sum_{n\in N_m^k |} \frac{\omega_k(1+\lambda^*)}{|F_k|+\lambda|\Delta_k|} (1+\mu)}$$

7.2 条件向量$\{u_n\}$计算

将$\{v_m^*\}$、$\{\omega_k\}$固定，当做常量，得到：

$$\min_{\{u_n\}} \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{m\in M_n^k} \frac{\omega_k \left(1+\lambda^*\right)}{|F_k| + \lambda \left|\Delta_k\right|} \left\{ \left(v_m^* - u_n\right)^2 + b_m \right\}$$

同理可得：

$$u_n = \frac{\sum_{k=1}^{K} \sum_{m\in M_n^k} \frac{\omega_k}{|F_k|+\lambda|\Delta_k|} (1+\lambda^*) \, v_m^*}{\sum_{k=1}^{K} \sum_{m\in M_n^k} \frac{\omega_k}{|F_k|+\lambda|\Delta_k|} (1+\lambda^*)}$$

7.3 源置信度值$\{\omega_k\}$计算

将$\{v_m^*\}$、$\{u_n\}$固定，当做常量，得到：

$$\min_{\{\omega_k\}} \sum_{k=1}^{K} \omega_k \cdot \theta_k, \quad \text{s.t.} \quad \sigma\left(\omega_k\right) = \sum_{k=1}^{K} \exp(-\omega_k) = 1$$

$$\theta_k = \frac{1}{|F_k| + \lambda \left|\Delta_k\right|} \sum_{(m,n)\in F_k} (1+\lambda^*) \left\{ \left(v_m^* - u_n\right)^2 + \mu \sum_{i\in C_m^n} \alpha_i \left(v_m^* - x_i\right)^2 \right\}$$

同理可得：

$$\omega_k = -log(\frac{\theta_k}{\sum_{k=1}^{K}\theta_k})$$

8 初始化

8.1 源置信度值$\{\omega_k\}$初始化

$$\omega_k^{(init)} = \frac{|F_k \cap F_{ref}|}{|F_k| + |F_{ref}|}$$

8.2 真相向量$\{v_m^*\}$初始化

$$v_m^* = \frac{\sum_{k\in K_m}\sum_{n\in N_m^k}\frac{\omega_k(1+\lambda^*)}{|F_k|+\lambda|\Delta_k|}\sum_{i\in C_m}\alpha_i x_i}{\sum_{k\in K_m}\sum_{n\in N_m^k|}\frac{\omega_k(1+\lambda^*)}{|F_k|+\lambda|\Delta_k|}}$$

9 后期处理

$$P(m,n) = \frac{min_j(v_m^* - u_j)^2}{(v_m^* - u_n)^2}$$

该公式计算$_{\text{条件}\,n}$之于$_{\text{三元组}\,m}$的置信度。

如果真相向量v_m^*距离对应条件向量u_n最近，易知分子部分与分母将相同，最终$P=1$

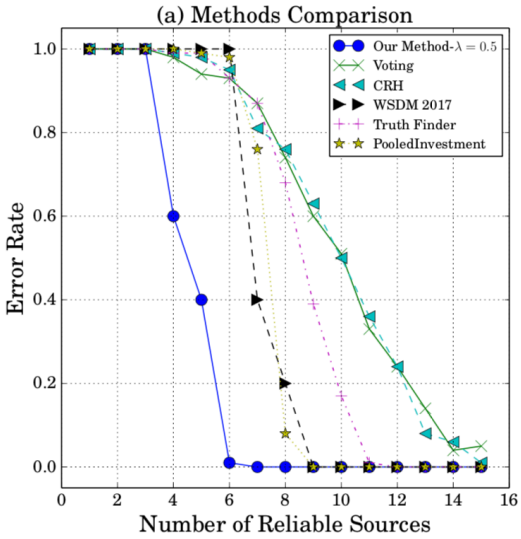
如果有与v_m^*距离更近的条件，则分子部分小于分母，最终$P<1$。偏离越远，说明$_{\text{条件}\,n}$之于$_{\text{三元组}\,m}$的置信度越低。

10 算法效果

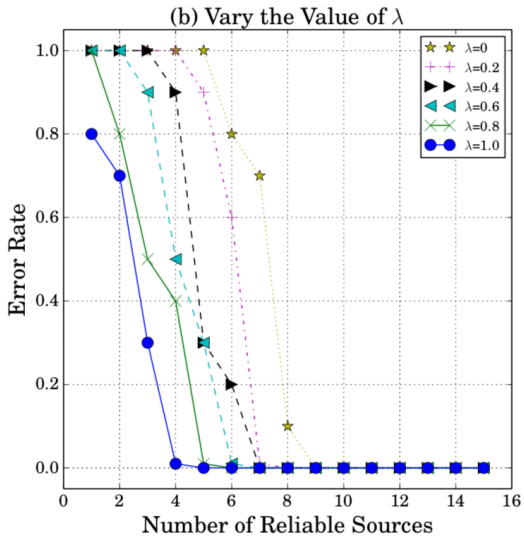
10.1 人造实验数据

作者设定了一个包含100个三元组和各自对应的10个条件作为标准先验知识，并构造测试数据集，其中，可靠源中的条件有5%的噪音，而不可靠源终的条件有95%的噪音。噪音即将对应正确的条件随机替换为其他条件。作者将源的数量固定为100。

10.1.1 不同方法比较

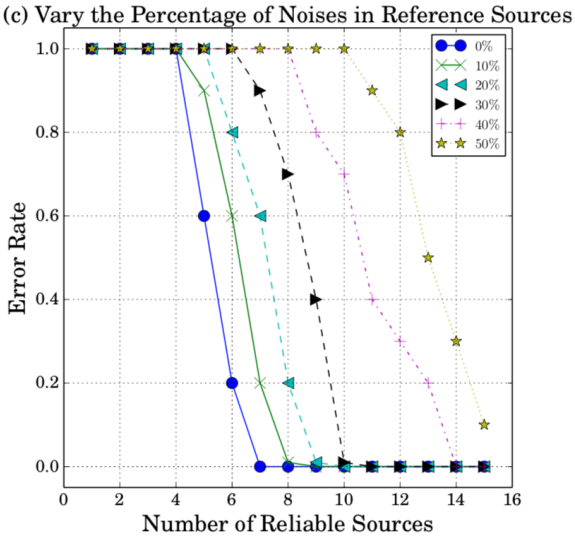


10.1.2 调整半监督权重λ



λ在公式中起到调整半监督(相信不可靠源)比重的作用，值越高，可靠源数据比重越大，所需要的可靠源数量越少。

10.1.3 调整可靠源中噪音占比



10.1.4 源置信度

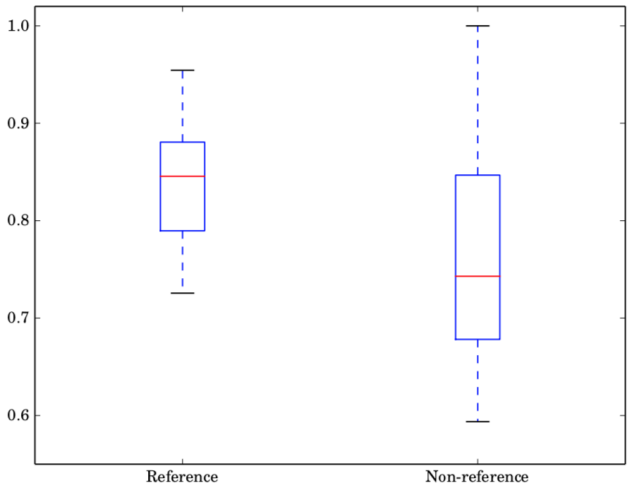
Method	Pearson's Correlation (Reliability & Error Rate)
Investment	-0.9316
PooledInvestment	-0.9216
TruthFinder	-0.8858
CRH	-0.9816
WSDM 2017	-0.9944
The Proposed Method	**-0.9950**

将源的可靠性按照噪音比例进行排序，与算法结果的置信度进行比较(皮尔森相关系数)，值越接近-1越好。

> 皮尔森相关系数：
>
> https://www.jianshu.com/p/b43c8731a309
>
> 系数衡量的是两组变量间的线性相关关系，取值为[-1 , 1]，-1表示完全的负相关，+1表示完全的正相关，0表示没有线性相关。

10.1.5 可靠、不可靠源的置信度比较

构造100个源，其中20个可靠源、80个不可靠源。可靠源的噪音居于0%–50%，不可靠源噪音0%–100%。



可见可靠源置信度整体较高，但部分不可靠源的置信度也不低于可靠源。

10.2 真实数据

10.2.1 条件置信度

10.2.1.1 (mycoplasma pneumoniae pneumonia, chest pain)和(pneumonia, chest pain)的案例

即(支原体肺炎，胸痛)和(肺炎，胸痛)

(mycoplasma pneumoniae pneumonia, chest pain)				(pneumonia, chest pain)	
Condition (Gender&Age)	EMR Count	QA Count	Confidence Score	All Count	Confidence Score
male	0	4	1.00	1698	1.00
female	1	1	0.92	991	0.91
20	1	3	1.00	815	0.99
10	0	0	0.96	255	1.00
30	0	1	0.87	425	0.86
0	0	0	0.81	267	0.86
40	0	1	0.67	329	0.69
90	0	0	0.54	19	0.55
50	0	0	0.54	262	0.58
60	0	0	0.51	200	0.54
80	0	0	0.49	25	0.50
70	0	0	0.48	79	0.50

可以看出，两种疾病对性别并不敏感，但年轻人较为易感。

10.2.1.2 消融实验-不区分可靠与不可靠源

(coronary heart disease, chest pain)					(bronchitis, chest pain)				
Condition (Age)	All Count	Confidence Score w/ ref	Confidence Score w/o ref	Statistic Probability	Condition (Age)	All Count	Confidence Score w/ ref	Confidence Score w/o ref	Statistic Probability
60	1146	1.00	1.00	18.42%	0	116	1.00	1.00	6.66%
50	1416	0.98	0.99	22.75%	10	196	0.99	0.99	11.24%
70	556	0.94	0.95	8.93%	20	591	0.95	0.96	33.91%
80	128	0.88	0.90	2.06%	30	367	0.87	0.90	21.06%
40	1253	0.86	0.87	20.14%	40	233	0.71	0.79	13.37%
90	8	0.75	0.80	0.13%	50	134	0.59	0.66	7.69%
30	770	0.69	0.76	12.37%	90	3	0.58	0.66	1.72%
10	77	0.66	0.73	1.24%	60	59	0.55	0.65	3.38%
20	659	0.64	0.68	10.59%	70	39	0.51	0.59	2.24%
0	210	0.45	0.52	3.37%	80	5	0.51	0.62	2.87%

左侧疾病对老年人更加易感，而右侧疾病则易感年轻人。

从表中可以看出，利用可靠数据进行半监督学习可以提高这种区分度。

10.2.2 消融实验-分别考虑图谱表示学习嵌入与共现嵌入(CBOW)的效果

以(coronary heart disease, chest pain)，即(冠心病, 胸痛)，为例

Only Entity Embeddings		Only Co-occurrence Embeddings		Combine Two Kinds of Embeddings	
Similar Triple	Distance	Similar Triple	Distance	Similar Triple	Distance
(heart disease, chest pain)	0.9872	(heart disease, chest pain)	0.6546	(heart disease, chest pain)	0.9444
(myocardial infarction, chest pain)	1.0204	(cardio-cerebrovascular disease, chest pain)	0.7095	(cardio-cerebrovascular disease, chest pain)	1.0933
(coronary heart disease, left chest pain)	1.2282	(cardiovascular disease, coronary insufficiency)	0.7275	(coronary heart disease, dorsal distending pain)	1.2294
(coronary heart disease, chest distress)	1.2822	(myocardial infarction, filling defect)	0.7504	(heart disease, limited activity)	1.2520
(heart failure, chest pain)	1.4299	(coronary heart disease, shoulder pain)	0.8885	(myocardial infarction, filling defect)	1.2703

在第一列(仅考虑图谱表示学习嵌入)，算法找到的最相似病症均与心脏病或胸痛有关。

在第二列(仅考虑贡献嵌入)，算法找到的最相似病症，其实是经常共现的病症，与心脏病等关系较远。

第三列综合考虑两者，算法可以同时考虑经常共现于同一病例中的疾病和本身就比较相似的病症。