基于动态知识选择的预训练语言模型

Zhancheng Guo, Ting Song, Shizhu He ,Kang Liu, Jun Zhao and Shengping Liu

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences

Outline

- 知识图谱与预训练语言模型
- 动态知识选择预训练改进
- 具体的方法与结果
- 意外发现？

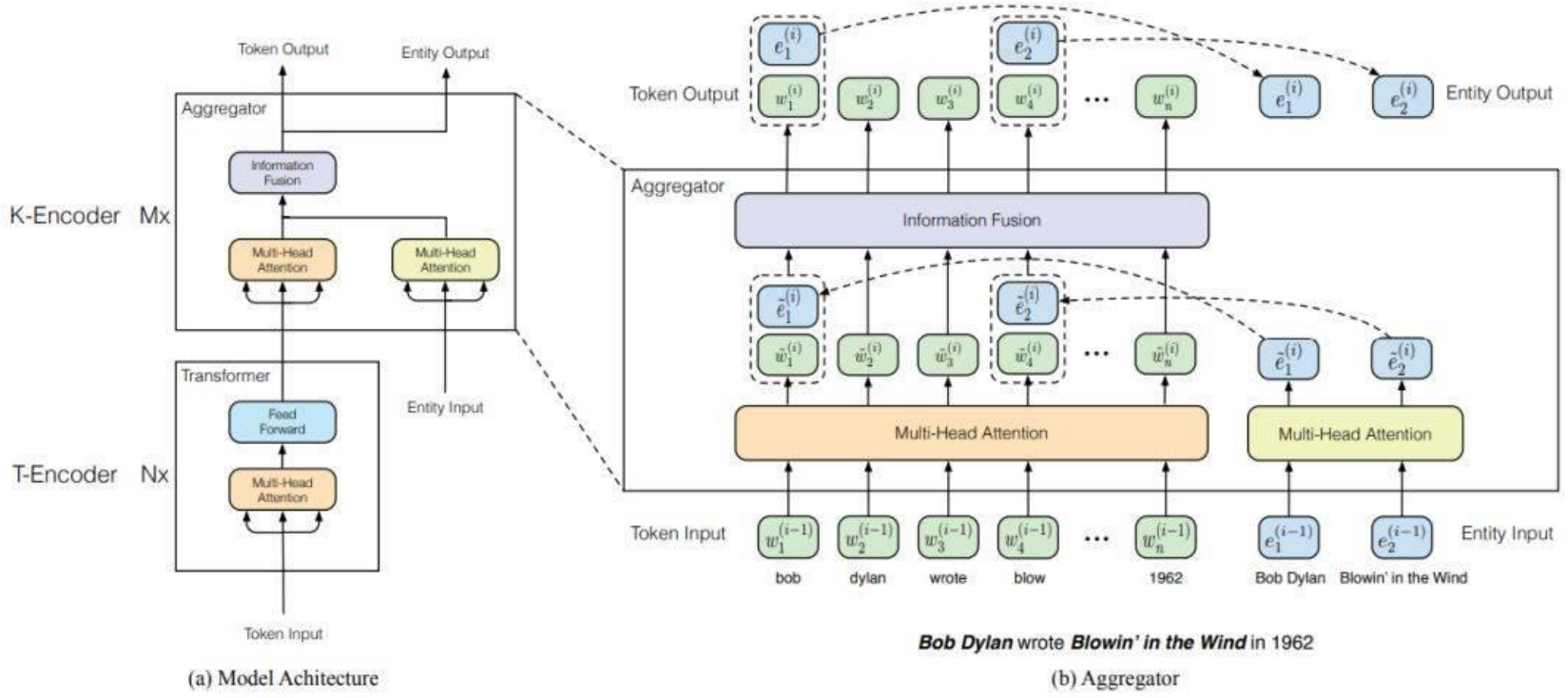
知识图谱与预训练语言模型

- 预训练模型增强知识图谱表示
  - KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation
  - CoLAKE: Contextualized Language and Knowledge Embedding

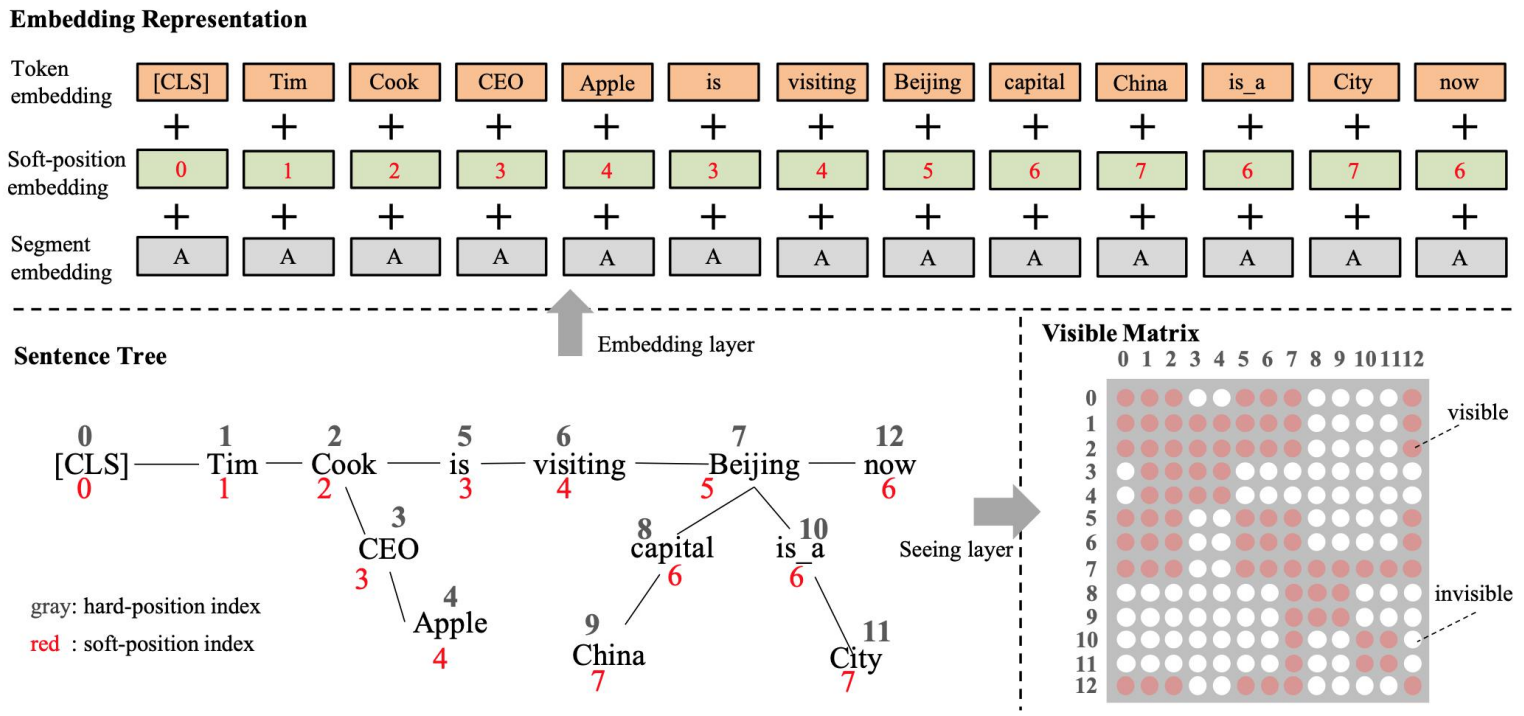
  ....
- 知识图谱增强预训练语言模型表示
  - ERNIE: Enhanced Language Representation with Informative Entities
  - K-BERT: Enabling Language Representation with Knowledge Graph

  ......

ERNIE: Enhanced Language Representation with Informative Entities



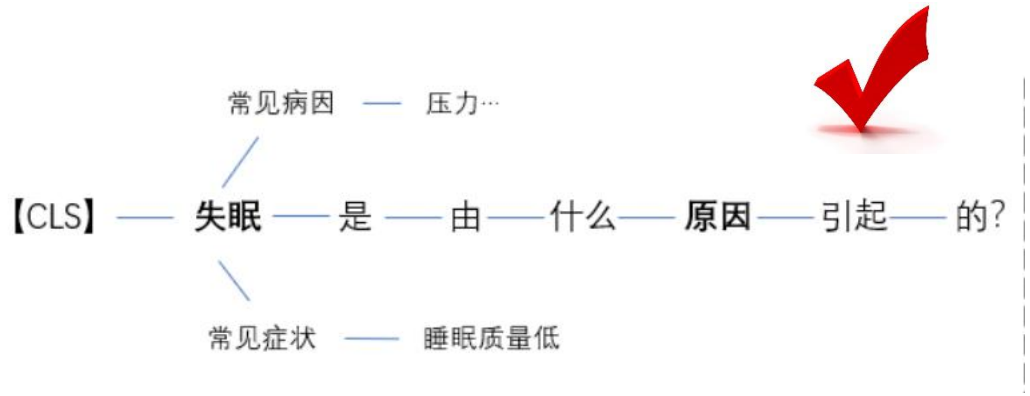
K-BERT: Enabling Language Representation with Knowledge Graph



动态知识选择

知识三元组
失眠，常见症状，睡眠质量低…
失眠，就诊科室，神经内科…
失眠，英文名称，insomnia
失眠，常见病因，压力…
原因，词性，名词
原因，外文名，Reason
……

动态知识选择



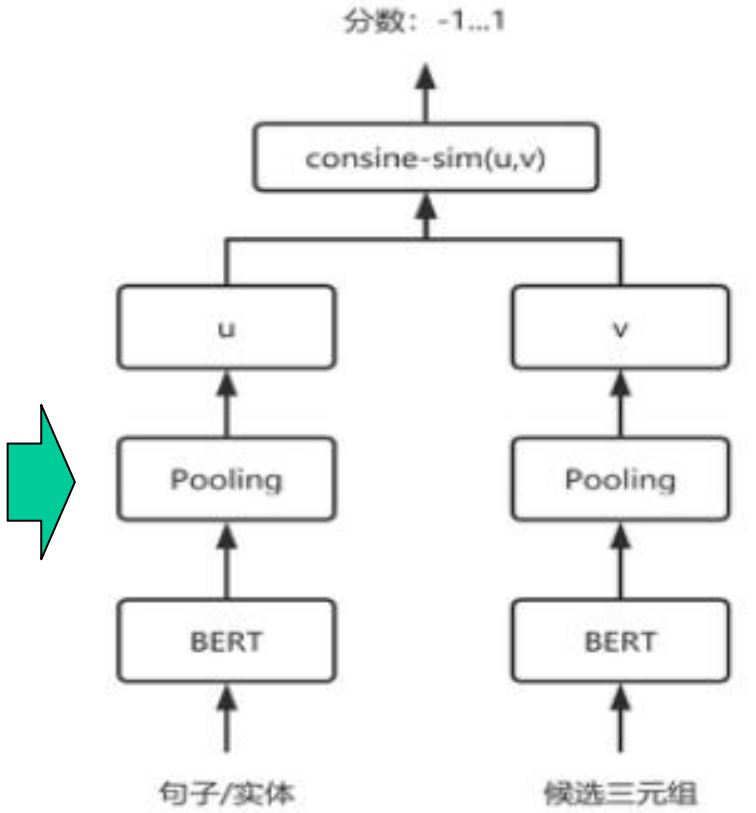
- 更符合语境
- 对语义理解有帮助

动态知识选择

句子或实体词

失眠是由什么原因引起的?

*失眠*是由什么原因引起的?



知识三元组文本

知识三元组
失眠，常见症状，睡眠质量低...
失眠，就诊科室，神经内科...
失眠，英文名称，insomnia
失眠，常见病因，压力...
原因，词性，名词
原因，外文名，Reason
.....

动态知识选择

- 动态句向量：

$$\boldsymbol{u} = BertMeanPooling(s)$$
$$E = K\_Query(s_w, K)$$
$$E_v = BertMeanPooling(E)$$
$$D = cosine(u, E_v)$$
$$t = K\_Inject(s, Sort_D(E))$$

- 动态词向量：

$$U = BertMeanPooling(e_{token})$$
$$E = K\_Query(s_w, K)$$
$$E_v = BertMeanPooling(E)$$
$$D = cosine(U, E_v)$$
$$t = K\_Inject(s, Sort_D(E))$$

动态知识选择

- 动态句向量选择：根据与句子向量相似度选择知识三元组

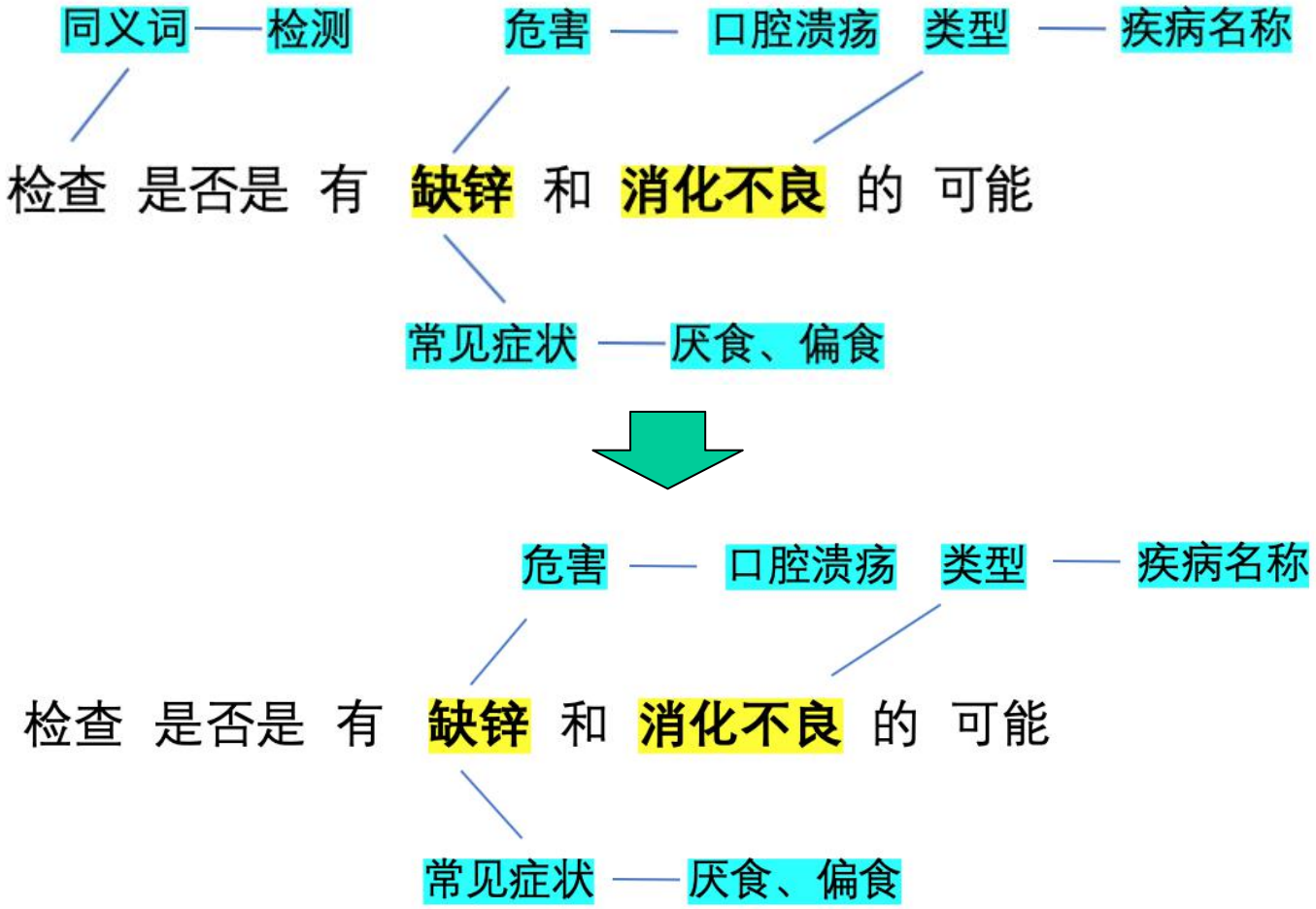
	cMedQNLI		cMedQQ		cMedTC		cMedIC	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Random2	88.6	88.6	85.4	84.8	73.7	75.2	87.0	88.1
Top2(sentence)	85.9	86.0	84.0	83.7	71.7	72.9	87.8	86.1
Last2(sentence)	90.8	**91.0**	86.0	**85.6**	77.0	**77.6**	89.3	**92.9**

- 动态词向量选择：根据与实体词向量相似度选择知识三元组

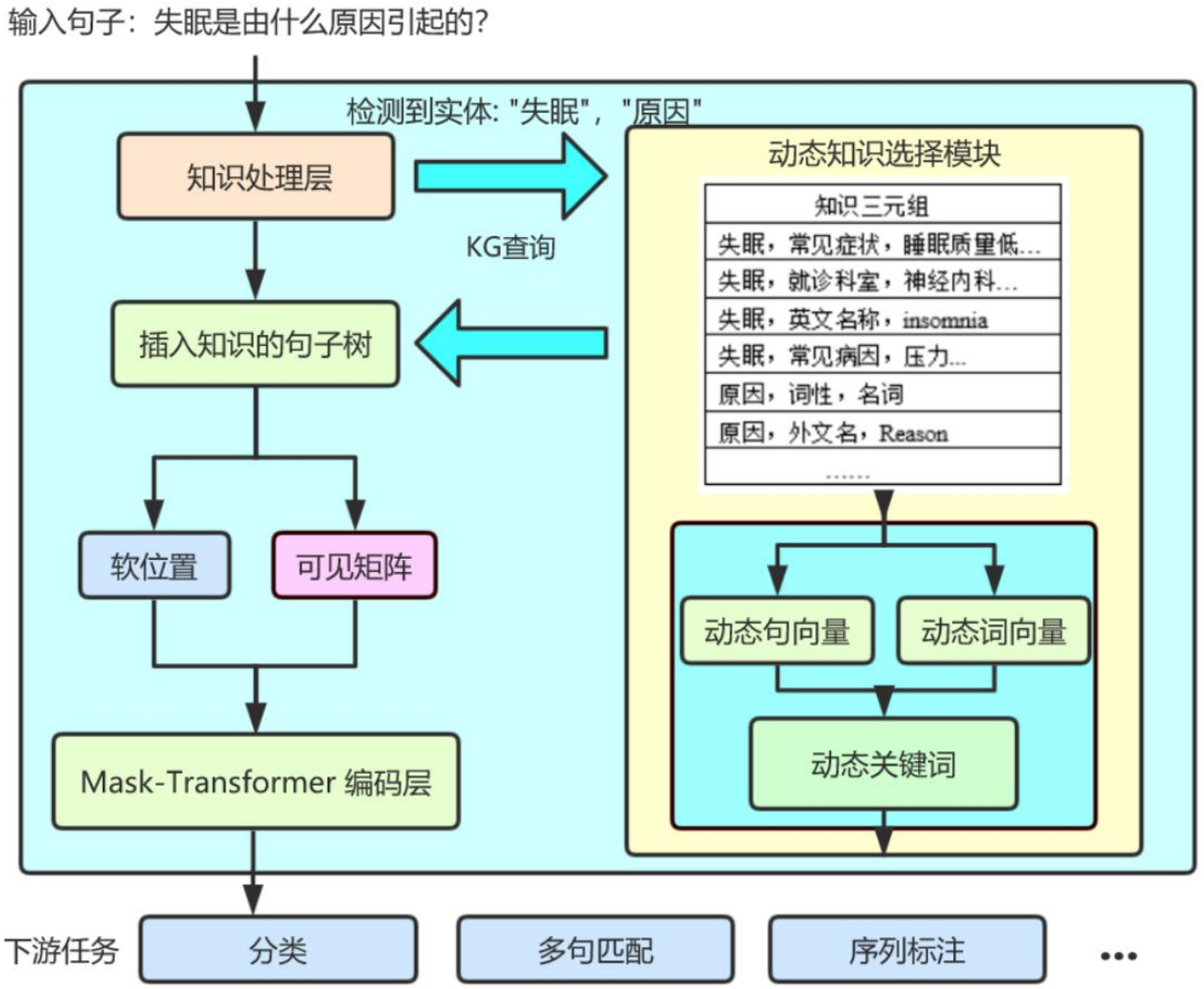
	cMedQNLI		cMedQQ		cMedTC		cMedIC	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Random2	88.6	88.6	85.4	84.8	73.7	75.2	87.0	88.1
Top2(token)	90.9	**91.0**	87.9	**85.6**	77.0	**78.4**	88.6	**90.5**
Last2(token)	87.7	88.0	86.0	84.4	73.7	74.4	87.0	87.5

动态知识选择

- 动态关键词：根据TF-IDF值筛选关键词



动态知识选择


输入句子：失眠是由什么原因引起的?
检测到实体: "失眠"， "原因"
知识处理层
KG查询
插入知识的句子树
软位置
可见矩阵
Mask-Transformer 编码层
动态知识选择模块
知识三元组
失眠，常见症状，睡眠质量低...
失眠，就诊科室，神经内科...
失眠，英文名称，insomnia
失眠，常见病因，压力...
原因，词性，名词
原因，外文名，Reason
.....
动态句向量
动态词向量
动态关键词
下游任务
分类
多句匹配
序列标注
...

动态知识选择

- 融入关键词动态选择

	cMedQNLI		cMedQQ		cMedTC		cMedIC	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Random2+key3	95.6	95.7	87.9	87.0	82.3	81.4	87.8	88.1
Sentence-Last2	90.8	91	86	85.6	77	77.6	89.3	92.9
Sentence-Last2+key3	96.0	**95.9**	87.7	87.5	81.7	82.0	89.3	**94.1**
Token-Top2	90.9	91	87.9	85.6	77	78.4	88.6	90.5
Token-Top2+key3	95.6	95.7	86	**87.9**	82.0	**82.3**	87.0	92.9
K-BERT(original)	88.6	88.6	85.4	84.8	73.7	75.2	87.0	88.1
BERT	-	93.3	-	86.5	-	79.0	-	86.0
MC-BERT	-	95.5	-	87.5	-	82.1	-	87.5

小结

- 我们的贡献
  - 设计了一套动态知识选择预训练的方法
  - 一定程度上解决了K-BERT模型容易在实体链接阶段产生同名实体链接错误及在知识图谱质量不佳的情况下模型鲁棒性差的问题
  - 在医疗领域数据集中国生物医学语言理解评估基准数据集ChineseBLUE 上效果取得显著提升

动态知识选择

	cMedQNLI		cMedQQ		cMedTC		cMedIC	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Random2	88.6	88.6	85.4	84.8	73.7	75.2	87.0	88.1
Top2(sentence)	85.9	86.0	84.0	83.7	71.7	72.9	87.8	86.1
Last2(sentence)	90.8	**91.0**	86.0	**85.6**	77.0	**77.6**	89.3	**92.9**

数据示例

CN-DBpedia 插入知识示例

螺内酯片*脱发*

- 螺内酯片--药品名称--螺内酯片
- 螺内酯片--不良反应--详见下文
- 螺内酯片--主要适用症--详见下文
- 螺内酯片--用途分类--低效利尿药
- 螺内酯片--用法用量--详见下文

六味地黄丸*有降糖作用!*

- 六味地黄丸--通用名--六味地黄丸
- 六味地黄丸--药品名称--六味地黄丸
- 六味地黄丸--商品名--六味地黄丸
- 六味地黄丸--功效--滋阴补肾

知识图谱改造

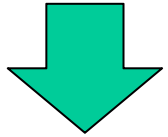
(a,b,c) → (a,b,a)

螺内酯片--不良反应--详见下文

螺内酯片--不良反应--螺内酯片

(a,b,c) → (a,a,a)

螺内酯片--不良反应--详见下文



螺内酯片--螺内酯片--螺内酯片

改造知识图谱结果

- *CN-DBpedia* “错误”知识图谱结果对比：

F1	cMedQA	cMedQNLI	cMedQQ	cMedTC
CnDbpedia	0.9846	0.9587	0.8713	0.8172
CnDbpedia (a,b,a)	0.9842	0.9587	0.8734	0.8172
CnDbpedia (a,a,a)	0.9855	0.9572	0.8636	0.8094

后续工作

- 对我们后续工作启示
  - 是否太过关注知识图谱与预训练语言模型形式上的结合？
  - 我们真正需要知识图谱中的什么信息，才能有效的帮助预训练语言模型获取知识信息？
  - 如何判定是否是真正利用了知识？

中国科学院
CHINESE ACADEMY OF SCIENCES

谢谢

Email:
zhancheng.guo@nlpr.ia.ac.cn

中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES