知识库构建中的关键技术研究

Research on Key Technologies in Knowledge Base Construction

答辩人：胡伟龙
指导老师：彭敏 教授

huweilong@whu.edu.cn

武汉大学语言与信息研究中心
武汉大学计算机学院

2019 年 11 月 26 日

提纲

1 研究背景及意义

2 相关工作

3 知识库构建方法
- 命名实体识别
- 实体关系抽取
- 知识表示学习

4 研究计划

提纲

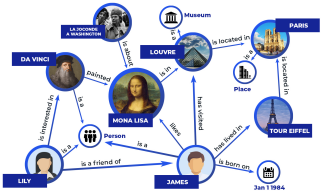
1 研究背景及意义

2 相关工作

3 知识库构建方法
- 命名实体识别
- 实体关系抽取
- 知识表示学习

4 研究计划

1 研究背景及意义



- 提高搜索质量与用户体验
- 百度知心、搜狗知立方
- 搜索、问答、个性化推荐

- 完全由专家人工构造
- 数据驱动的自动信息抽取
- 标注资源匮乏、噪声较大

- 基于知识库的智能问答
- 计算问句与知识库相似度
- 智能问答产品相对普及

基于标注资源匮乏的文本构建面向智能问答的知识库，进而推动问答类产品的落地，不仅具有研究意义，也具有应用价值.

提纲

1 研究背景及意义

2 相关工作

3 知识库构建方法
- 命名实体识别
- 实体关系抽取
- 知识表示学习

4 研究计划

2.1 相关工作 知识库的发展

表: 国内外主要知识库

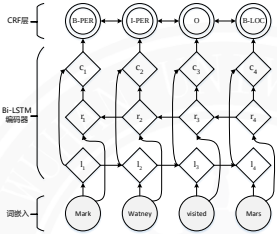
领域专家编辑	WordNet, Cyc
自动信息抽取	YAGO, DBPedia, Freebase
多语言知识库	Zhishi.me, CN-DBPedia, PKU-PIE, XLORE, Belief-Engine, Knowledge Vault, 搜狗知立方, 百度知心

上述知识库都是基于结构规范的百科类网站构建，但如何基于标注资源匮乏 的特定领域非结构化文本构建特定知识库仍是难题

2.2 相关工作 命名实体识别

现有方法：

- 规则式方法：结合启发式算法和人工规则从文本中抽取公司名称
- 统计学习方法：K-近邻算法、条件随机场
- 有监督深度学习：LSTM-CNN, Bi-LSTM-CRF, Attention-based Bi-LSTM-CRF, ID-CNN-CRF
- 其他方式：如半监督学习、主动学习、迁移学习



难题—如何减少人工标注量并以弱监督的方式进行命名实体识别

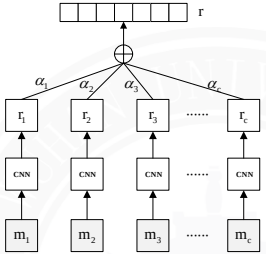
现有弱监督实体识别方法：

1. 系统冷启动：主动学习在初始时仍然需要部分标注数据
2. 初始标注集：半监督方法（如自训练）受限于初始标注集的质量

2.3 相关工作 实体关系抽取

现有方法：

- 有监督：基于特征的方法，基于各类核的方法
- 无监督：对实体上下文聚类推断关系类型
- 远程监督：Mintz，MultiR，MIML，CNN/PCNN+ATT，Att-BLSTM
- 其他方式：联合学习，自训练，主动学习，标签传播



难题—如何降低远程监督数据集中噪声对实体关系抽取的影响

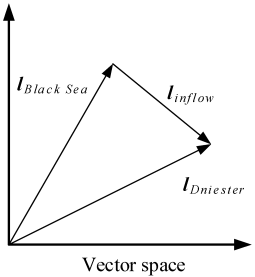
现有远程监督监督关系抽取方法：

1. 文本编码器：未考虑到文本编码器的表征能力与运算效率之间的权衡
2. 辅助性信息：未考虑到引入辅助性信息（外部知识）所带来的额外噪声

2.4 相关工作 知识表示学习

现有方法：

- 翻译模型：TransE, TransH, TransR, TransD, TransSparse, TransA, TransG, KG2E
- 图模型：R-GCN, L-GCN
- 其他方式：距离模型，单层神经网络模型，能量模型，双线性模型，张量网络模型，矩阵分解模型



难题—如何利用知识库的语义网络结构学习实体和关系的语义表示

现有翻译模型或者图模型：

1. 翻译模型孤立学习三元组：以三元组为单位，难以捕获邻近实体或关系之间的依赖
2. 图模型学习目标不太契合：以节点嵌入为学习目标，不能直接获得关系向量

提纲

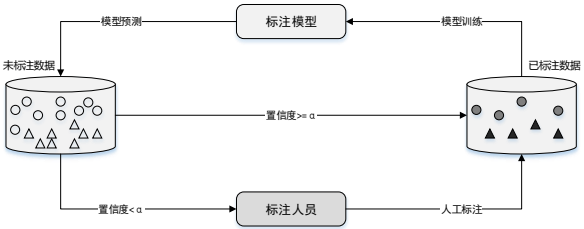
1 研究背景及意义

2 相关工作

3 知识库构建方法
- 命名实体识别
- 实体关系抽取
- 知识表示学习

4 研究计划

3.1 知识库构建方法 基于主动学习和自训练的弱监督实体识别



整体流程：

- 多标准主动学习：根据不确定性标准和信息密度标准采样待标注样本
- 自训练学习：对于置信度高于阈值的样本进行机器标注，置信度低于阈值的样本进行人工标注

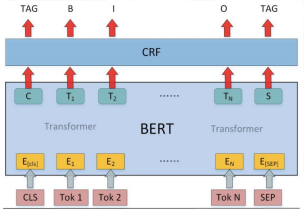
预训练语言模型：

- 作为特征编码器

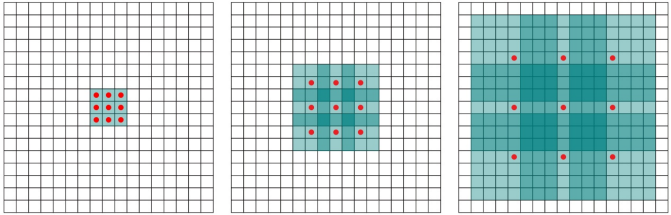
$$P(y|x;\theta) = \mathrm{CRF}(lm(x))$$

- 计算样本相似度

$$\mathrm{sim}(x, x^{(u)}) = \begin{cases} \dfrac{lm(x) \cdot lm(x^{(u)})}{\|lm(x)\| \times \|lm(x^{(u)})\|} \\ \sigma(lm(x, x^{(u)})) \end{cases}$$



3.2 知识库构建方法 结合空洞卷积和软实体类型约束的关系抽取



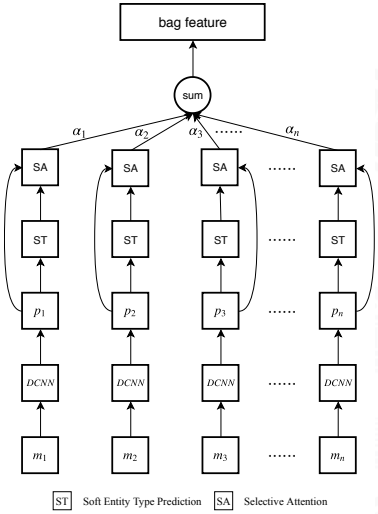
整体流程：

- 空洞卷积文本编码器：捕获长距离依赖关系的同时保持运算的高效

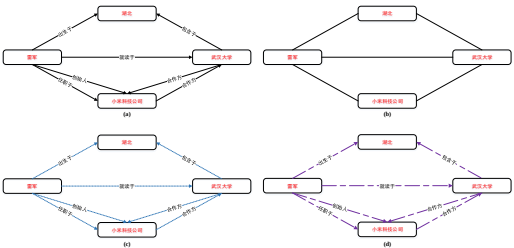
$$d_i = \mathbf{W}_c \bigoplus_{k=0}^{r} \mathbf{s}_{i \pm k\delta} + \mathbf{b}$$

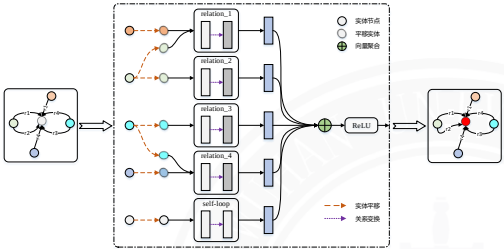
- 软实体类型约束：预测关系的同时预测对应的实体类型（软实体类型）

$$\mathbf{t}_s = \arg\max(\mathbf{o}_t + \beta \cdot (\mathbf{o}_t) \odot \mathbf{t})$$



3.3 知识库构建方法 基于多重图卷积和翻译框架的知识表示学习





整体流程

- 全局关系嵌入与局部关系嵌入：全局关系嵌入作为最终的关系向量表示

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\frac{1}{c_i}\sum_{r\in\mathcal{R}}\left([\mathbf{w}_s^r\mathbf{h}_i^{(l)} + \sum_{j\in\mathcal{N}_i} f(\mathbf{w}_{ij}^r, \mathbf{r}_r)\cdot\mathbf{h}_j^{(l)}]\mathbf{W}_r^{(l)}\right)\right)$$

- 消息传递框架结合翻译框架：头实体向量在进行消息传递之前，首先经过全局关系嵌入进行平移

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\frac{1}{c_i}\sum_{r\in\mathcal{R}}\left([\mathbf{w}_s^r\mathbf{h}_i^{(l)} + \sum_{j\in\mathcal{N}_i} (f(\mathbf{w}_{ij}^r, \mathbf{r}_r)\cdot\mathbf{h}_j^{(l)} + \mathbf{r}_r)]\mathbf{W}_r^{(l)}\right)\right)$$

提纲

1 研究背景及意义

2 相关工作

3 知识库构建方法
- 命名实体识别
- 实体关系抽取
- 知识表示学习

4 研究计划

4.1 研究计划 创新点总结

本文的主要创新点主要如下：

- 针对命名实体识别任务中存在的过于依赖标注数据问题，提出了基于主动学习和自训练的弱监督实体识别方法．针对系统冷启动问题设计了基于多标准的主动学习采样策略，并且将主动学习与自训练学习结合，充分利用预训练语言模型的表征能力进一步降低对标注数据的依赖．
- 针对实体关系抽取任务中远程监督数据集噪声过大问题，提出了结合空洞卷积和软实体类型约束的关系抽取方法．为了克服卷积神经网络和循环神经网络的不足，引入空洞卷积网络作为文本编码器，在捕获长距离依赖关系的同时保持运算的高效性．同时，将实体类型约束加入到注意力机制中，通过显式考虑外部知识中的噪声学习更加精确的注意力权重．
- 针对知识表示学习任务中存在的翻译模型孤立学习三元组、图模型学习目标不契合的问题，提出了基于多重图卷积网络和翻译框架的知识表示学习方法．在图模型的基础上引入全局关系嵌入和局部关系嵌入，将全局关系嵌入作为最终的关系向量．并结合消息传递框架与翻译框架，同时学习实体和关系的丰富语义表示．

4.2 研究计划 实验方案

1. 命名实体识别子任务：序列标注任务，同时识别实体词的边界和实体词类型
   - 实验数据：CONLL-2003
   - 基准模型：Bi-LSTM-CRF, ID-CNN-CRF
   - 评价标准：Accuracy，F1
2. 实体关系抽取子任务：多实例多标签学习任务，针对每对实体进行关系多分类
   - 实验数据：New York Times
   - 基准模型：Mintz, MultiR, MIML, CNN/PCNN+ATT
   - 评价标准：Precision@N, Precision recall curve
3. 知识表示学习子任务：利用知识图谱网络的结构信息学习实体和关系的向量表示
   - 实验数据：WN18, FB15K
   - 基准模型：DistMult, R-GCN, TransE
   - 评价标准：Mean reciptocal rank(MRR), Hits at n(H@n)

4.3 研究计划 进度安排

表: 进度安排表

2019.10-2019.11	查阅相关文献，确定课题内容及方案.
2019.12-2020.01	对现有的知识库构建方法进行综述与分析，并准备实验环境.
2020.01-2020.02	完成基于主动学习和自训练的弱监督实体识别.
2020.02-2020.03	完成结合空洞卷积和软实体类型约束的关系抽取.
2020.03-2020.04	完成基于多重图卷积网络和翻译框架的知识表示学习.
2020.05-2020.06	撰写毕业论文，准备答辩.

结语

恳请老师和同学们批评指正！

答辩人：胡伟龙
武汉大学计算机学院