

# 信息抽取前沿动态综述

---

韩先培

xianpei@nfs.iscas.ac.cn



中国科学院软件研究所

# 信息抽取

- 从非结构化/半结构化文本（如网页、新闻、论文文献、微博等）中提取指定类型的信息知识（如实体、属性、关系、事件、商品记录等）
- 通过信息归并、冗余消除和冲突消解等手段将非结构化文本转换为结构化信息



# 前沿动态调研

- 调研了过去一年在ACL、EMNLP、NAACL、SIGIR、CIKM、IJCAI、AAAI等会议上发表的81篇相关信息抽取论文



# 研究趋势

---

- **更合适的模型**：深度学习开始统治
- **更精准的推理**：Joint/Global Inference
- **更多的资源**：海量文本、背景知识、知识库
- **更少的监督**：半监督、弱监督、远距离监督
- **更通用的模型**：领域无关、语言无关

# 更好的模型：深度学习占据统治地位

- 相关论文数目： **$28/81 = 34.6\%$ 论文**
- 标题里面出现：**LSTM : 10 , CNN: 7 , Embedding: 5, Attention: 3**
- 神经网络主要的提升机制
  - **基本单元的Embeddings**：实体、词、关系等基本单元更好的表示
  - **自然语言结构的表示学习**：句法树、依存路径、词序列等结构化对象的组合表示学习
  - **更好的任务建模**：使用LSTM来进行序列标注，Attention机制用来解决远距离监督的实例选择
  - **联合建模多个任务**：更少的中间错误，更强的约束

# 更好的推理：Joint/Global Inference

- 相关论文数目：**19/81 = 23.5%论文**
- 相关关键词：**joint : 13 , collective : 4 , global : 3 , cross-document: 2**
- **基本假设**：Facts are inter-connected, Tasks are Interdependent
- **基本做法**：
  - **协同建模多个任务**，解决Pipeline的错误传递问题，Multi-task Learning，更通用的表示，信息的相互利用（Coreference和Entity Linking）
  - **建模相互关联的抽取对象**，相互约束，相互增强（实体和关系）
  - **建模相互依赖的决策**（Collective Entity Linking）

# 更多的资源

- 相关论文数目：**18篇/81篇=22%**
- 没有显著关键词
- 充分利用外在的资源来提升现有模型的性能
  - **背景知识的约束和指导** ( FrameNet , MLN , Soft Logic )
  - **额外的信息补充** ( 同时使用知识库embedding、文本信息、路径 )
  - **海量数据下的冗余性**
  - **多模态信息的冗余性** ( 文本、media、图片 )

# 更少的监督

- 相关论文数目：**13篇/81篇=16%**
- 关键词：**Unsupervised : 4 , Distant Supervision : 3 , semi-supervised : 1 , zero-shot, one-shot: 1**
- 基于训练语料的构建成本过高，如何降低构建信息抽取系统的成本？
  - **无监督**：聚类、Topic、模式发现
  - **半监督**：Bootstrapping , Propagation
  - **远距离监督**：使用大规模知识库作为监督



# 更通用

- 相关论文数目：**8篇/81篇=10%**
- 解决现有模型跨语言、跨领域、跨媒体时模型无法使用或性能大幅下降的问题
- **核心方法**
  - Language Independent
  - Domain Independent
  - 领域自适应
  - 领域/媒体/语言无关特征

# 总结

---

- **提升信息抽取结果的质量**

- 深度学习模型
- 利用额外的信息资源
- Joint/Global Inference

更好、更便宜、更通用

- **降低信息抽取系统的成本**

- 降低监督要求
- 提升通用性