面向知识库的中文自然语言问句的语义理解

许坤，冯岩松，赵东岩，陈立伟，邹磊

北京大学

知识库

- **Yago**
- **Yago2**
- **DBpedia**
  - **340万个实体**
  - **300亿条三元组**
- **FreeBase**

Semantic Web

- **RDF**
  - <刘德华，父亲，刘礼>
- **SPARQL**
  - 刘德华的父亲出生于哪里？

```
Select  ?y
{
    刘德华   妻子   ?x
     ?x      出生于  ?y
}
```

“刘德华的父亲出生于哪里”

Bai du 百度 新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

刘德华的父亲出生于哪里 百度一下

刘德华出生于哪一年 - 已解决 - 搜搜问问
2个回答 - 最新回答: 2008年6月15日
最佳答案: 姓名:刘德华 族名:刘福荣 花名:华仔 华弟 华英雄 荣仔 英文名:Andy Lau 血型:ab型 身高:1.74cm 体重:64kg 出生年月日:1961.9.27 出生地:...
wenwen.soso.com/z/q677333...htm 2013-8-12 - 百度快照

刘德华:童年时因为赌钱挨过父亲暴打_网易娱乐
[图文] 2013年2月1日 - 刘德华:童年时因为赌钱挨过父亲暴打刘德华族名刘福荣,英文名Andy Lau,出生于1961年9月27日,入演艺界近20年的刘德华,1999年获得"香港十大杰出青年...
ent.163.com/13/0201/14/8MKQ9CIF00032... 2013-02-01 - 百度快照

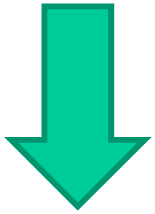
刘德华_百度百科
刘德华1961年9月27日生于香港,著名演员、歌手、制片人,影视歌多栖发展代表,是位有使命感的电影人。1981年以全优成绩毕业于TVB艺训班签约出道,1982年凭《猎鹰》...
baike.baidu.com/link?url=olK1ble5PKkhDYHj... - 百度快照

北京大学

The Problem

Map Sentences to SPARQL

刘德华的妻子毕业于哪里



```
Select  ?y
{
      刘德华   妻子   ?x
       ?x       毕业于  ?y
}
```

Several potential applications

- QA
- Dialogue System
- Natural Language Interfaces to Database

北京大学

Related Work

- **Yahya** 利用整数线性规划建模
- **Unger** 利用模板和句法分析树建模
- **Watson** 系统利用“线索”和“子线索”获取答案
- **FREyA** 系统依赖用户对问句的反馈来进行命名实体识别，需要有监督训练

Our Approach

- 构造查询语义图
  - 生成句法分析树
  - 分析名词性节点和动词性节点
  - 实体消歧和谓词消歧
- 利用**g-store**查询引擎搜索
  - 子图匹配
- 与之前工作的区别
  - 英文与中文（承担语法功能的结构与词性有无关联）
  - 资源（**WordNet, PATTY)**

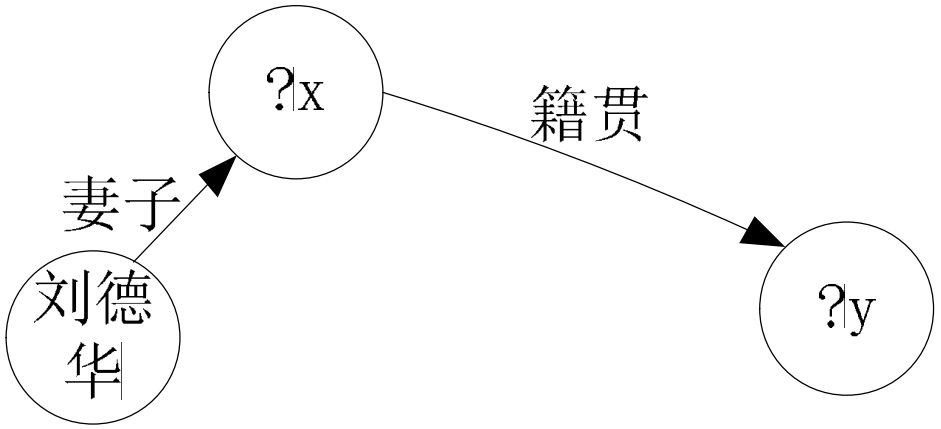
北京大学

Our Contribution

- **QALD-1, QALD-2, QALD-3**
  - **50**个不同语义复杂度的训练问题与**SPARQL**查询
  - **50**个测试问题
  - 英语问题
- **BaiduKnows42**
  - 人物、地点、组织
  - **42**个中文问题

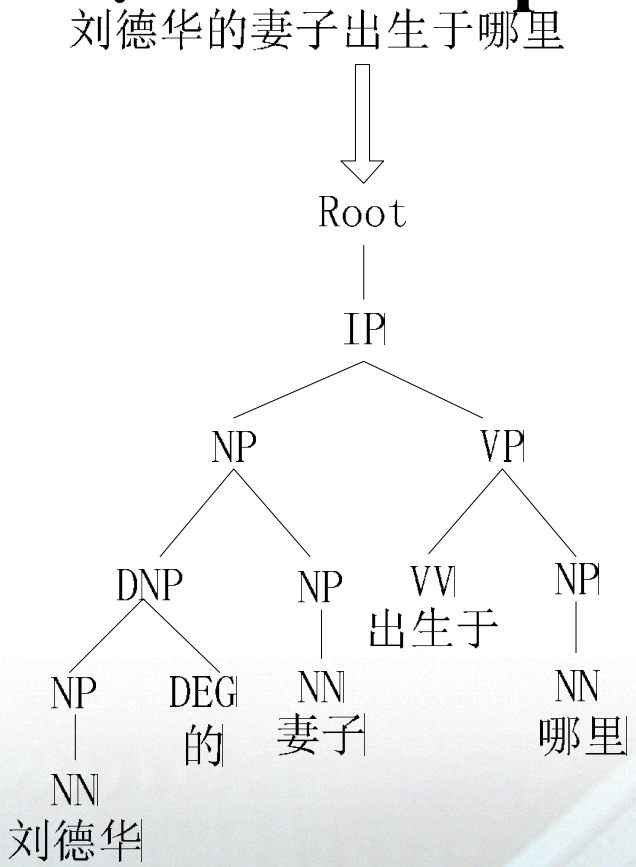
查询语义图

- 前提
  - 用户问题基于实体的属性或者实体间的关系
  - 事实性问题
- 实体
  - 点
- 关系
  - 边
- 单关系查询与多关系查询



构造查询语义图

- 生成句法分析树
  - **ICTCLAS、Stanford parser**



构造查询语义图

- 实体一般以名词的形式出现
  - 分析名词性节点
- 关系一般以名词或者动词词组的形式出现
  - 分析动词性节点
- 验证
  - 在百度知道上随机抽取20个问题
    - 90%的问句中的实体以名词形式出现
    - 80%的问句中的关系以名词或者动词词组的形式出现

输入：问句的句法分析树*T*

输出：问句的查询语义图*G*

1: 令***TreeNodes***表示句法分析树中所有节点的集合

2: for 每个节点***node*** ϵ ***TreeNodes*** **do**

3: **if** ***node***是名词性节点 **then**

4: 调用函数 分析名词性节点(***node***)

5: **Else If** ***node*** 是动词性节点 **then**

6: 调用函数 分析动词性节点(***node***)

7: End For

分析名词性节点

- 名词性节点
  - 命名实体
  - 名词性变量

1: 令*npNode*的子节点集合是*Children*
2: **if** *Children*中不包含名词性节点 **then**
2: 构造一个节点代表*npNode*加入*G*
3: **if** *npNode* 不是一个命名实体 **then**
4: *npNode*是一个变量
5: **End If**
3: **else**
4: **for** 每个节点*Child* ϵ *Children* **do**
5: **if** *Child* 是名词性节点 **then**
6: 调用函数 分析名词性节点(*Child*)
7: **End If**
8: **End For**
9: 在修饰性名词节点与被修饰名词节点之间引入边*l*， *l*从修饰性名词节点指向被修饰名词节点。如果被修饰名词节点不是命名实体而是一个名词性变量，则该名词是用来描述这两个节点之间的关系
10: **End Else**
11: **End If**

分析动词性节点

1: 令*vpNode*的子节点集合是*Children*

2: **for** 每个节点*Child* ϵ *Children* **do**

3: **if** *Child* 是名词性节点 **then**

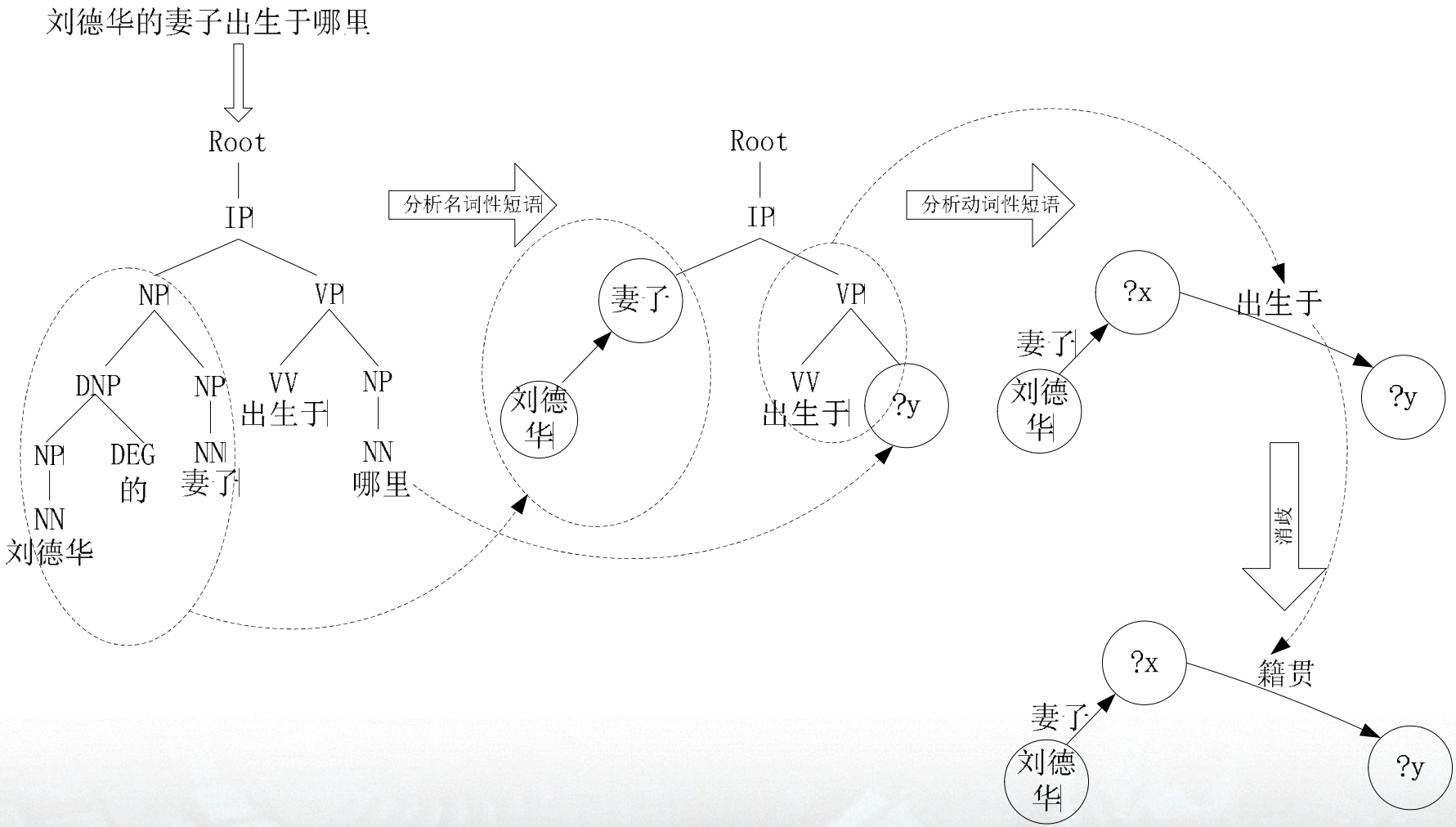
4: 调用函数 分析名词性节点(*Child*)

5: **End If**

6: **End For**

7: 在代表动词主体的名词性节点与代表客体的名词性节点之间构造一条边

Running Example



北京大学

消歧

- 实体消歧
  - 编辑距离(edit distance)
  - 构造包含中文名称、别名的词典
  - 穷举实体s的所有可能性
- 谓词消歧
  - “Y的作者是X”与“Y是X的一部作品”
  - 收集相关度较高的关键词

收集与关系相关度较高的词

- 互动百科
  - **Infobox**
  - **197种关系**
  - “刘德华在黄大仙天主教小学毕业后升读可立中学”
- 收集名词与动词

英文名:	AndyLau
艺名:	华仔、老大、华哥、华弟、刘天王
性别:	男
血型:	AB
出生年月:	1961年9月27日
毕业院校:	可立中学、第十期无线艺员训练班
身高:	175厘米
爱好:	保龄球、羽毛球、台球、驾驶

谓词映射

- 待映射的谓词是G，知识库的关系集合是L，l是任意一种关系，G与l的词语相关度是：

$$Rel(G,l) = \sum_{i=1\, w_i \epsilon W}^{m} Rel(G,w_i) * tf\_idf(w_i)$$

$$Rel(G,w_i) = \begin{cases} 1 ; G = w_i \\ 0 ; \mathrm{G} \neq w_i \end{cases}$$

实验

- 百度知道
  - 规模（230万个问题）
- 42个问题
  - 22个可以利用知识库回答
  - 3人测评生成的SPARQL查询
- 问题列表

问题列表

序号	类别	问题描述	序号	类别	问题描述	序号	类别	问题描述
1	人物	张杰是哪里人?	15	地点	北京有哪些地标性建筑?	29	组织	国民党是谁建立的?
2	人物	张杰的第一张专辑是什	16	地点	北京的人口有多少?	30	组织	中国共产党现任总书记是谁?
3	人物	张杰的生日是啥时候?	17	地点	北京市的现任市长是谁?	31	组织	九三学社的加入条件是什么?
4	人物	张杰的老婆是谁?	18	地点	北京市有哪些知名企业?	32	组织	国民党的政治理念是什么?
5	人物	刘德华的原名叫什么?	19	地点	北京市有哪些机场?	33	组织	共产党创建时间是哪一年?
6	人物	刘德华的爸爸是谁?	20	地点	北京市有哪些行政区?	34	组织	中国铁道部部长是谁?
7	人物	刘德华的女儿演过什么电	21	地点	朝阳区的邮政编码是多	35	组织	中国工会的会长是谁?
8	人物	梁朝伟演过什么电影?	22	地点	北京有哪些特产?	36	组织	国际红十字会的英文缩写是什
9	人物	刘德华出生在哪里?	23	地点	北京的著名景点有哪些?	37	组织	国民党什么时候成立的?
10	人物	梁朝伟的女朋友有哪些?	24	地点	北京的面积有多大?	38	组织	发改委的办公驻地在哪里?
11	人物	梁朝伟的身高有多少?	25	地点	北京有哪些地标?	39	组织	发改委的领导人是谁?
12	人物	梁朝伟写过那些书?	26	地点	上海的主要街道有哪些?	40	组织	发改委的全称是什么?
13	人物	梁朝伟高中毕业于哪里?	27	地点	上海的电话区号是多少?	41	组织	中国农业部的网站是什么?
14	人物	梁朝伟出生在哪里?	28	地点	上海的名人有哪些?	42	组织	中国农业部的职能是什么?

实验

- 回答准确率
  - **22个可以回答，答案只有一个，其中10个回答正确，45%**
- 生成SPARQL查询的准确率
  - 自动评价，**36%**
  - 人工评价，**48%**
- 谓词消歧的准确率
  - **HowNet，10%**
  - **Our approach，36%**

北京大学

实验分析

- **SPARQL**查询分析
  - 人工语义比较确定的准确率更高

    互动百科的关系体系并不完美，存在一些关系可以映射到多个关系
  - **80%**的错误由于谓词消歧造成

    部分属性所能收集到的语料比较少

    弱监督假设引入噪声

总结

- 提出查询语义图的概念
- 设计了从句法分析树构造查询语义图的方法
- 提出了启发式识别实体与关系的方法

Thank you!

北京大学