


基于实体语义关系的中文 问题-答案关系研究

上海交通大学
计算机科学与技术系
王睿 姚天昉
marswang@sjtu.edu.cn




提纲

- 背景介绍
- 巨人的肩膀
- 我的工作
 - 我为什么要做问答系统
 - 我的方法
 - 具体工作
- 未来的工作




背景介绍：为什么需要问答系统？

- 信息电子化、网络化
- 知识隐藏
- 关键字检索的缺点
 - 准确性
 - 人性化
- 自然语言回答



前人的工作：现行的系统

- MIT开发的Start系统
- Michigan大学的AnswerBus系统
- 中文问答式系统
 - 尤里卡搜索引擎
 - 百度(孙悟空)搜索引擎
 - 问一问搜索引擎



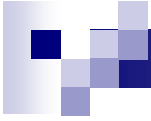
前人的工作：我站在哪里？

- 上海交通大学计算机科学与工程系姚天昉副教授开发的基于体育领域中文命名实体及其关系识别技术为本系统提供了实体语义关系库。
- 例子：郝海东<-[属于/拥有关系]->大连队
翻译为自然语言是：“郝海东是大连队的。”



我的工作：我为什么做问答系统

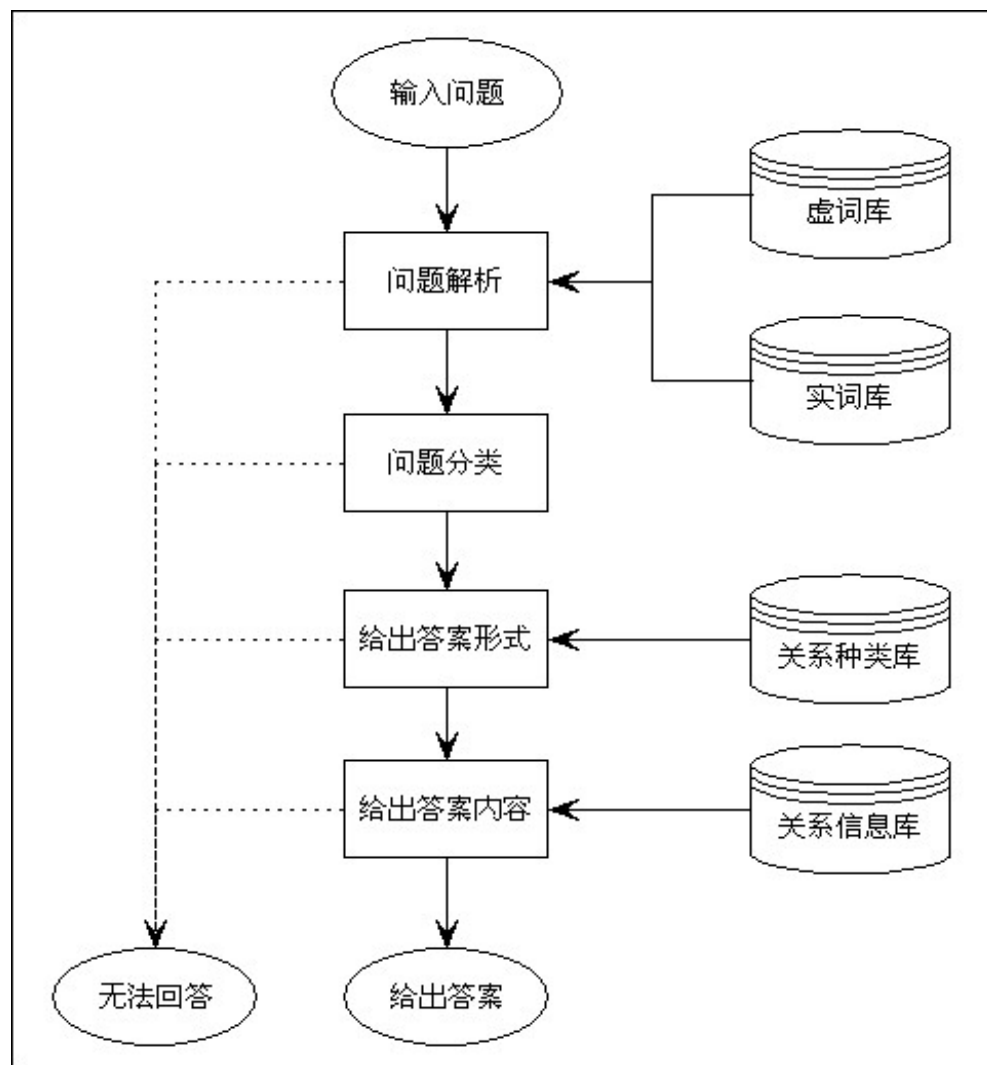
- 对现有系统的支持和补充
- 验证命名实体的抽取是否有效
- 探索类似关系库的应用前景
- 自然语言的深入理解
- 为开放领域的中文问答系统做准备



我的工作：方法定义

- 采用的方法：
 - 语言学的方法
 - 对话经验的总结
- 未来的方法：
 - 公式化
 - 概念化

我的工作：系统的总流程图





我的工作：问题的解析（一）

- **假设1** 用户提出的问句一般都是简单句[3]。
- 问句的特点：简短，从句情况少，疑问词、语气词能提供相当大的信息
- 关系库的特点：二元，因此问题最终要归结为两个（种）实体间的关系
- 领域信息的特点：名词需要细化，比如人名、队名、比赛名称等，不能仅限于标注出是名词

我的工作：问题的解析（二）

表1 词性标注表					
Tab.1 Labels of POS					
词的分类	标记	省份	Ns	“有”	Vy
否定词	B	城市	Nm	“在”	Vz
日期	Dd	地点	Nl	标点符号	W
时间	Dt	代词	R	“?” 或 “? ”	?
一般名词	N	人称代词	Rp	ABA型词	X
人名	Np	指示代词	Rd	A不A	Xb
队名	Nt	助词	U	A没A	Xm
角色名	Nr	“的”	Ud	疑问词	Y
比赛名	Nc	动词	V	“谁”	Yp
国家	Nn	“是”	Vs		

我的工作：问题的解析（三）

- 如果用户输入的问题为：

郝海东 是 在 大连队 踢球 的 吗 ?

- 系统解析后的结果为：

Np VsVzNt N V UdU ?

我的工作：问题的分类

- **规则1.1**（特殊疑问句判别规则）
一个（种）实体，有疑问词（并且提供了提问的方面），语气词一般为“呢”、“啊”等
- **规则2.1**（特征疑问句判别规则）
两个（种）或两个（种）以上实体，句子的主语、谓语和宾语都已经具有，有疑问词（并且提供了提问的方面）
- **规则3.1**（一般疑问句判别规则）
两个（种）或两个（种）以上实体，句子的主语、谓语和宾语都已经具有，没有疑问词，语气词一般为“吗”、“吧”等
- **注：**选择疑问句包括在一般疑问句中。

我的工作：给出答案形式（一）

■ 特殊疑问句

□ 问: **Y+V+N或N+V+Y**

答: **N+V+N**

□ 问: 郝海东是哪个队的?

答: 郝海东是大连队的。

■ 特征疑问句

□ 问: **Y+N+V+N或N+Y+V+N或N+V+N+Y**

答: **DD/DT/NL+N+V+N或N+DD/DT/NL+V+N或N+V+N+DD/DT/NL**

□ 问: 郝海东是什么时候加入大连队的?

答: 郝海东是在1995年7月加入大连队的。

■ 一般疑问句

□ 问: **N+V+N**

答: **V**

□ 问: 郝海东是大连队的吗?

答: 是的。



我的工作：给出答案形式（二）

- 需要用到的不止一种关系的情况，比如：
郝海东是谁？
- 无法定位到一种关系的情况，比如：
郝海东为什么要加入大连队？
- 解决方案：
 - 一是搜索全部可能的关系；
 - 二是随机选择一种关系搜索。

我的工作：给出答案形式（三）

- 不存在的关系提问：
郝海东参加了亚洲杯赛吗？
- 关系的推理：
人和比赛的关系=人和队伍的关系+队伍和比赛的关系
- 解决方案：划分为现有二元关系或者现有二元关系的组合。



我的工作：确定关系类型

- 对于特殊疑问句：
根据二元关系中已知的一个（种）实体查询另外一个（种）实体；
- 对于特征疑问句：
则根据现有的每个（种）实体查询时间、地点等信息再进行整合；
- 对于一般疑问句：
根据问题中给出的二元关系的双方与关系库进行比较，如果关系库中找到了这一对实体则返回正确，反之，则返回错误。



我的工作：检索答案信息

- 一般情况：

问：郝海东是哪个队的？ $(N+V+Y)$

答：郝海东是大连队的。 $(N+V+N)$

- 答案不止一个

问：谁在大连队？

答：郝海东、李明在大连队。




我的工作：给出最终答案

- 给出最终答案还得注意一些自然语言的特点
 - 比如一般疑问句中“是”和动词同时说的时候，答案中的动词用“是”等。
 - 助词、语气词以及标点的正确使用。
- 让系统的回答更加人性化
 - 建立答案形式库，在多种形式可以选择的情况下，随机挑选，每次对于同一问题的回答有可能不一样（形式上），这些作为本系统的扩充内容在将来会引入。

我的工作：代词的处理


- **假设2** 一般代词指的就是前面说到的最近的事物[2]，远指代词指由一般代词所指示的事物往前推最近的事物

表2 代词数组表				
Tab.2 Array of contents of Pronouns				
.....	他	这个队	那个人
.....	李明	大连队	郝海东



我的工作：问题集

- 问答集的建立与中间过程的修改都是便于机器学习和人为控制的
- 用户输入完问题后，可以根据系统给出的答案进行对系统的评价，打分等级为：正确、错误和部分正确三种。
- 系统将问题集输出为文件后，可以便于系统管理员收集整理，从而更好地改善系统。



我的工作：实验（一）

■ 问题集（部分）

- 郝海东是大连队的吗？
- 郝海东是在大连队的么？
- 郝海东属于大连队么？
- 郝海东在什么（球）队（呢）？
- 郝海东在哪个队？
- 大连队里有哪些球员？
- 郝海东什么时候加入大连队的？
- 大连队什么时候买入郝海东的？
-

我的工作：实验（二）

表5 问题集3						
Tab.5 Question Set 3						
测试日期	不包含代词			包含代词		
	正确	错误	部分正确	正确	错误	部分正确
2009-8-4	90%	5%	5%	87%	5%	8%



我的工作：小结

- 本系统是基于实体语义关系的中文问答式系统，具有以下特点：
 - 限于关系库信息，提问内容有限，形式可以随意
 - 内容提取优先，语法分析辅助，疑问词标注细化
 - 提供对中间结果的修改和问答集，为机器学习作好准备
 - 准确率比较高，引入代词处理
- 作为中文问答式系统的一个初步实现，本系统大致探询了一下问答式系统除文章信息抽取外其他部分的工作，给出了一种实现方案，为全开放的中文问答式系统提供了一些参考。



未来的工作

- 人工输入特殊句式、句型库，用以辅助机器学习
[3]
- 建立模型描述问题与答案之间的对应关系，构成形式化关系库
- 建立完整的评价系统进行更加科学、完全的评价
- 答案形式更加丰富，更加人性化，让用户使用起来更加舒适



参考文献

- [1] Tianfang Yao. Hybrid Approach Based Chinese Named Entity Extraction on a Specific Domain. International Workshop ILT&CIP on Innovative Language Technology and Chinese Information Processing, 2001: p74.
- [2] 贾娇燕. 实用汉语语法. 安徽教育出版社, 2003.
- [3] Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Michael Junk, Chin-Yew Lin. Question Answering in Webclopedia.
- [4] Scott Robertson. Where Can I Get Tickets to the Redwings Game?. 2001.
- [5] Deepak Ravichandran and Eduard Hovy. Learning Surface Text Patterns for a Question Answering System.
- [6] Eduard Hovy, Ulf Hermjakob, and Deepak Ravichandran. A Question/Answer Typology with Surface Text Patterns.

鸣谢

- 感谢姚天昉教授的指导和同学们的帮助！

