IBM WATSON analysis report

# 概述

IBM人工智能计算机系统"沃森"（Watson）在2011年2月美国热门的电视智力问答节目"危险边缘"（Jeopardy！）中战胜了两位人类冠军选手，随后 IBM 对Watson进行了开源。Watson研发力量主要集中在自动问答系统里面。这部分由该研发中心的DeepQA项目组负责。

DeepQA项目组有大概40多个研发人员，主要分为下面几个小组：

l 算法小组（负责各NLP模块和其他模块的算法）

l 策略小组（处理流程和处理策略）

l 系统小组（硬件使用和接入，系统效率提升）

l 语音小组（语音接入）

l 注释小组（标注语料库，扩展机器人理解范围）

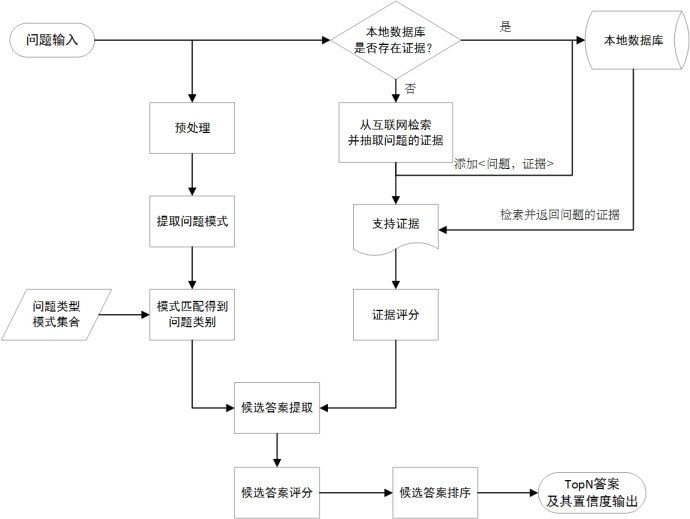
l 中国研究实验室（本体，元数据，语义网络和知识表示方面的研究）

l 东京的研究实验室（文本挖掘，文本抓取，半结构化信息使用）

l 海法的研究实验室（各种信息检索算法）

# 系统架构

该系统的架构如下图所示：



Watson技术实现简要分析

其中，最核心的可以分为证据获取、证据评分、问题分类和候选答案评分四个模块。这几个模块的目标及内容如下：

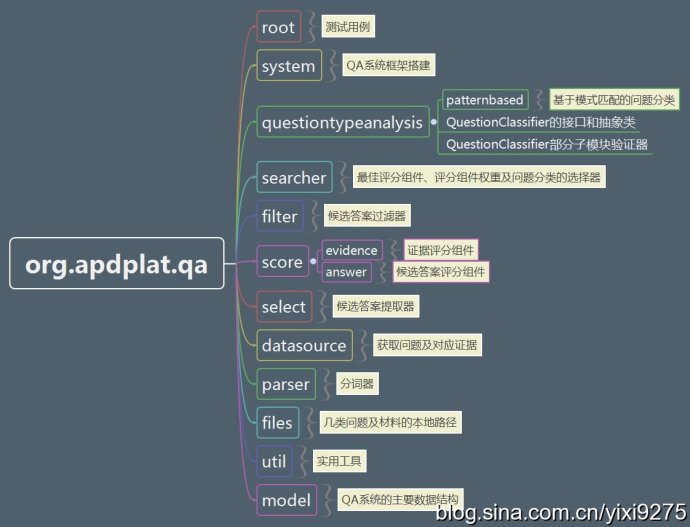
证据获取：从本地数据库或互联网上获取支撑问题的证据。若本地数据库存储有该问题的证据，则直接返回支撑问题的证据。否则，需要利用搜索引擎（如百度、谷歌）从互联网上抓取与该问题相关的片段，并抽取出其中的正文作为该问题的支撑证据。

证据评分：为评价不同证据对问题的支撑度，需建立一套证据评分机制。证据评分模型，采用了基于词频的、基于bigram的和基于skip-bigram的三种评价方法及基于上述三种方法加权的组合方法。评分过程中，可以由用户设定以上四种评价方法的任意一种。

问题分类：对问题所属的类别进行判定。该系统将可识别的问题类别划分为人名、地名、机构名、数字、时间、定义和对象七类（暂时仅支持前五类），并预先定义这几类问题的匹配模式。分类过程为：1）提取问题的模式，2）和预定义的问题类型模式进行正则匹配，3）根据匹配的结果确定问题的类别。针对同一类问题，系统又定义了五大类数小类的匹配模式，用户可以自由设定匹配模式为大类别或基于五大类的加权组合。

候选答案评分：为评估候选答案的质量，需建立一套候选答案评分机制。在候选答案评分模型中，基础的评价方法有基于词频的、基于词距的、基于最短词距的、基于文本对齐的、基于宽松文本对齐的、基于回带文本对齐的和基于热词的七类方法；综合的评价方法有基于基础评价方法加权的组合方法。评价过程为：1）根据问题类型确定答案类型，然后从证据词集中筛选出命名实体标记与答案类型一致的词，作为候选答案，2）针对每个候选答案，利用评分模型进行打分，用户可以自由设定八类评价方法中的一种作为评分模型。在打分过程中，每类评价方法均有一个权值，候选答案的得分是评价方法的打分与该评价方法权值的乘积。

# 代码研读



本节重点描述Watson先进的技术，对于和chatrobot类似的功能(如分词器)不做描述。

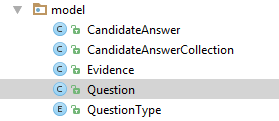
根据数据流向从上到下逐一分析。

1. System模块

提供基础框架和调度功能，是整个系统的入口和中心调度枢纽。

1. Model模块

模块包含以下数据结构



核心组件是Question，Watson的接收到问题请求会生成一个Question对象，关键参数为

**private** String **question**;  
**private final** List<Evidence> **evidences** = **new** ArrayList<>();

question为问题输入的字符串：

* “海的女儿”是哪个城市的城徽？
* 世界上流经国家最多的河流是哪一条？
* 世界上最长的河流是什么？
* 如何选购一个冰箱？

List<Evidence> 是一个问题的证据list：

* Title: 世界上最长的河流？（前5位）谢谢\_爱问知识人

Snippet: 世界上最长的河流是什么河？有多长？ 世界上最长的两大河流及其长度？ 21. 世界 上最长的河流是：(). 世界上最长的是什么河流. 世界上最长的河是?...

* Title: 北美洲最长的河|世界上最长的河流是什么河数米长的

Snippet: 尼罗河纵贯非洲大陆东北部，流经布隆迪、卢旺达、坦桑尼亚、乌干达、埃塞俄比亚、 苏丹、埃及，跨越世界上面积最大的撒哈拉沙漠，最后注入地中海。流域面积约335?...

* Title: 世界之最3

Snippet: 2011年2月28日 ... 世界上最长的河是：尼罗河. 世界上最长的河流在哪个国家？：埃及 ... 世界上毒性最 强的动物是一种什么动物？：水母. 世界上水流量最大的河：亚马孙?...

* Title: 世界地理之最\_互动百科

Snippet: 世界地理之最-含沙量最大的河流黄河是世界上含沙量最大的一条河流。根据位 ... 起 自鲍威尔湖口的大峡谷是其中最长、最宽、最深的一个，也是世界上最有名的峡谷。

* Title: 世界第一大河是什么?流域面积最大的河是什么?最深的湖是什么?最大 ...

Snippet: 亚马逊森蚺是当今世界上最大的蛇，最长可达十公尺，重达二百二十五公斤以上，粗 如 ... 最终注入地中海，是世界上流经国家最多的国际性河流之一。

* Title: 世界上最长的河流前十名\_荆楚养生网

Snippet: 2012年11月19日 ... 我们知道世界上最长的河流是尼罗河，那么世界上最长河流排名前十的都有哪些您 了解吗？让小编来告诉您.

* Title: 尼罗河- 维基百科，自由的百科全书

Snippet: 尼罗河長6,650公里，是世界上最长的河流。2007年雖有來自巴西的學者宣稱亞馬遜 河長度更勝一籌，但尚未獲全球地理學界的普遍認同。尼罗河有两条主要的支流，白 ?...

* Title: 世界上最大的胸部,世界上最贵的车,世界上最大的黑帮,最大的蛇等各类 ...

Snippet: 本栏为广大朋友们提供了各类世界之最的相关介绍，世界上最大的胸部,世界上最贵 的车,世界上最大的黑帮都是什么，有什么特点等各种不同类型的世界之最的相关?...

* Title: 欧洲最长的河流是什么合理?\_九千网

Snippet: 2012年11月9日 ... 伏尔加河(Volga River)是欧洲最长的河流，全长3500多千米，最后注入里海，流域 面积达136万平方千米，是世界上最长的内流河。俄罗斯内河航运?...

* Title: 天文地理- 世界之最百科

Snippet: 行家：性感的坏狗时间：2011年11月20日 贝加尔湖贝加尔湖是世界上最深和蓄水量 最大的淡水湖。位于布里亚特共和国(Buryatiya)和伊尔库次克州(Irkutsk)境内。

1. Datasource模块

获取问题的来源和证据，主要分为web查询(baidu,google,etc.)，本地文本库。可以选择单一数据源和组合数据源作为输入。

* Web查询

结合上面的question结构体，web爬取查询的问题字符串，作为question对象的question字段参数，爬取到的答案存放的List<Evidence>内。

* 本地查询

通过查询互联网很容易构造满足上面的question机构体，对于本地的文件，Watson同样要求提供满足上述的结构体，对于chatrobot构造问题会很困难。

需要预先知道问题的描述，组织对应的证据，如果两个问题只是描述方式不一样，不能得到答案。

**Question在构造过程中是特定的问题对应特定的答案组，要求唯一对应的，“谁写了著名的诗篇《将进酒》”和“著名的诗篇《将进酒》是谁写的”在Watson看来是两个完全不相关联的两个问题。**

本地数据源的标准格式如下：

?21. 谁写了著名的诗篇《将进酒》:李白

Title: 《将进酒》\_互动百科

Snippet: 写《将进酒》的时候，李白自觉怀才不遇，于是乎，要他的朋友和他一起喝酒，“与尔 ... 李白咏酒的诗篇极能表现他的个性，这类诗固然数长安放还以后所作思想内容更为 ...

Title: 将进酒\_百度百科

Snippet: 百科名片. 《将进酒》. 《将进酒》原是汉乐府短箫铙歌的曲调，题目意译即“劝酒歌”。 ..... 既是生逢知己，又是酒逢对手，不但“忘形到尔汝”，诗人甚而忘却是在写诗，笔下之 ...

Title: 《将进酒》---任志宏诗歌朗诵—在线播放—优酷网，视频高清在线观看

Snippet: 《将进酒》---任志宏诗歌朗诵李白《将进酒》---任志宏诗歌朗诵.

Title: 李白•嵩山•《将进酒》 中国嵩山|SongshanGeopark

Snippet: 2012年7月28日 ... 元丹丘姓林，名宗，乃大唐开元、天宝年间的著名道人，家居北嵩山脚下的颍阳 ... 怀才不遇的李白借酒消愁，挥笔写下了千古名篇乐府诗《将进酒》。

Title: 李白-诗词《将进酒》 古诗全诗赏析、作者及出处、年代！—诗词名句网

Snippet: 但全诗洋溢着豪情逸兴，取得出色的艺术成就。 李白咏酒的诗篇极能表现他的个性， 这类诗固然数长安放还以后所作思想内容更为深沉，艺术表现更为成熟。《将进酒》 ...

Title: 中国古代诗歌散文欣赏《将进酒》教案\_一松柏\_新浪博客

Snippet: 2013年3月18日 ... B.李白因酒而豪放、狂傲，因酒而成就了他的传世诗篇，就连他的死，也有 ... 成纪人， 唐代著名的浪漫主义诗人，后世称他为“诗仙”，又自称“酒中仙”。

Title: 李贺《将进酒》诗歌鉴赏资料论述文章解读- {TeacherCN.com}

Snippet: 李贺《将进酒》诗歌鉴赏 ... 一、多用精美名物，辞采瑰丽，且有丰富的形象暗示性， 诗歌形式富于绘画美。 此诗用大量篇幅烘托及时行乐情景，作者似乎不遗余力地搬出  ...

Title: 名篇欣赏李白的《将进酒》

Snippet: 注：李白《将进酒》原诗“朝如青丝暮成雪”大多数人都把“朝”字译为“早晨”，将“暮”译 .... 既是生逢知己，又是酒逢对手，不但“忘形到尔汝”，诗人甚而忘却是在写诗，笔下之 ...

Title: 同写《将进酒》:"诗仙"和"诗鬼"的区别——华夏文明——中国经济网

Snippet: 2007年11月28日 ... 他曾痛感生命无常，岁月不居，也挥毫写下了充满忧愁苦恨的诗篇《将进酒》，诗仙、诗 鬼，时代有别，人生有恨，借酒浇愁，一展歌喉，唱出了各有千秋 ...

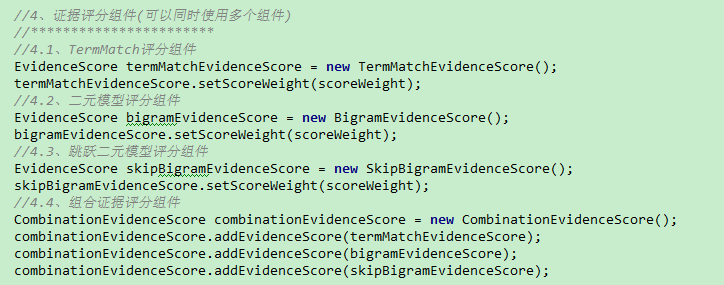
Title: 《将进酒》教案5-免费语文教学资源

Snippet: 2007年7月21日 ... 他一生写下 了许多咏酒的诗篇。他的《将进酒》就是酒与愁经过心理反应后留下的 结晶。 《将进酒》是一支劝酒歌，是汉乐府曲名，大约作于李白以梁园( ...

1. 证据评分模块

该模型的目标是评价支撑问题的证据的可信度，包含三个子模型（基于词频、基于bigram、基于skip-bigram）和一个组合模型（前三个子模型的线性加权），各评价子模型的初始权重均设置为1。

**关于三个评价模型的原理不做详细描述，可以google获取details，这里对chatrobot有借鉴意义的是采用多个模型线性加权的形式对证据进行评分，输出一个高可信的权重评分。**



1. 候选答案评分模块

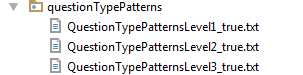
该模型的目标是评价候选答案的可信度，包含七个子模型（基于词频、词距、最短词距、文本对齐、宽松文本对齐、回带文本对齐和热词）和一个组合模型（前七个子模型的线性加权）。候选答案，是支撑证据中命名属性与问题类型一致的词。每个评价子模型的初始权重均设置为1。

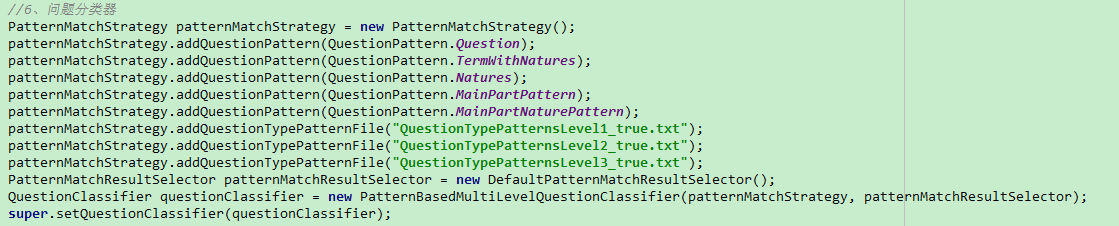
**在(4)证据评分模块的基础上，候选答案评分基于不同的证据评分值，采用多种组合模型再进一步的评分，关键的技术包括两个，一个是层级迭代的评分，一个是多个模型的组合加权评分，增加可靠性，这两点都是目前chatrobot里面缺少的，这部分思想可以借鉴，增加我们程序的准确度和稳定性。**



1. 问题分类模型模块

Watson处理的是结构化的问题，提供三级的问题模板，从Level3开始匹配，Level1级别最高，不同的级别对应不同的策略，匹配的级别越高对应问题答案的准确度越高。





1. 候选答案过滤和候选答案提取模块

主要对候选答案在上述评分后的基础上，进一步的筛选，提取，保证高可靠性的输出，本部分我没有理解掌握，不在这里展开误导大家了，有兴趣的同学可以研究一下。

综述：

Watson并没有使用太高大上的算法，采用细化又细化的处理逻辑，逻辑和处理流程的优化，对每一步的都采用多种评分和过滤方法，以保证本阶段向下传递的信息的准确度，最终输出的结果准确度是非常高的。

Reference：

1. “Building Watson: An Overview of the DeepQA Project”
2. “Text Comparison Using Machine-Generated Nuggets”
3. <https://developer.ibm.com/open/watson-developer-cloud-unity-sdk/>

样例问题：

* Question 水电验收需要注意什么？

Question 候选答案：

安全性 1.0

有经验 0.26635294094834805

平竖直 0.2663450585285952

有什么 0.13095680096143505

* Question “海的女儿”是哪个城市的城徽

Question 候选答案：

哥本哈根 1.0

丹麦 0.7111290328940997

首都 0.6996459146294252

博物馆 0.17386603527209313

公园 0.17385178331720308

罗马 0.15177306031531432

丹麦王国 0.15168690667472243

海滨公园 0.1406219503488705

Question 世界上最长的河流是什么

* Question 候选答案：

尼罗河 1.0

非洲 0.6125939404631553

长江 0.42658251180969714

密西西比河 0.4065074715975602

长河 0.34829873697179997

坦桑尼亚 0.3029948374356353

埃塞俄比亚 0.30294144365716363

苏丹 0.3029399920391388

埃及 0.30291561945339174

亚马孙河 0.2863676747172569

中国 0.21295377813048036

我国 0.20680078481427563

地中海 0.199909625153333

北美洲 0.18702228517751196

南美洲 0.18697183281852744

亚马逊河 0.18696713845076404

高原 0.1611478073112974

肯尼亚 0.16109411135073107

国家 0.16104121569108296

非洲大陆 0.14195092085068017

撒哈拉沙漠 0.14180619869183642

衣索匹亚 0.0581983021529526

三角洲 0.058155910014599965

* Question 大庆油田是哪一年发现的

Question 候选答案：

1959年 1.0

2013年 0.7824505078281574

1月 0.5977936020248433

2014年 0.5456505466665855

3日 0.4280351550050584

28日 0.42775710529839994

14日 0.4261021276429774

11月 0.3916648147429705

12月 0.39083319145875683

9月 0.38227175048370365

25日 0.3266700155318966

26日 0.26360663269166407

2010年 0.24447759506826014

今天 0.23428099751128453

21日 0.20568343689294077

1997年 0.20494017925011718

1960年 0.17304303211458083

50年 0.17180222730884231

是日 0.1568311147026074

国庆 0.15401377662755708

30年 0.1205459928376794

19日 0.12049933079900355

7月 0.12049085984429009

2015年 0.12048264558517399

* Question 如何选购一个冰箱？

Question 候选答案：

西门子 1.0

爱装网 0.8813835770896348

那句 0.6072131588400462

都有 0.5594895019207567

有必要 0.5190353398646035

齐全 0.444649232847631

* Question 如何选购家用轿车？

Question 候选答案：

成正比 1.0

有些人 0.9921891417101085

拉风 0.9898270809999397

宋晓辉 0.8696234267559182

越大 0.7929891813150866

有的 0.7920781764019422

车前 0.5328304912517725

能接受 0.3539286172193592

都来 0.20024042133506897