检察院卷宗智能化处理技术交流会

石恩名

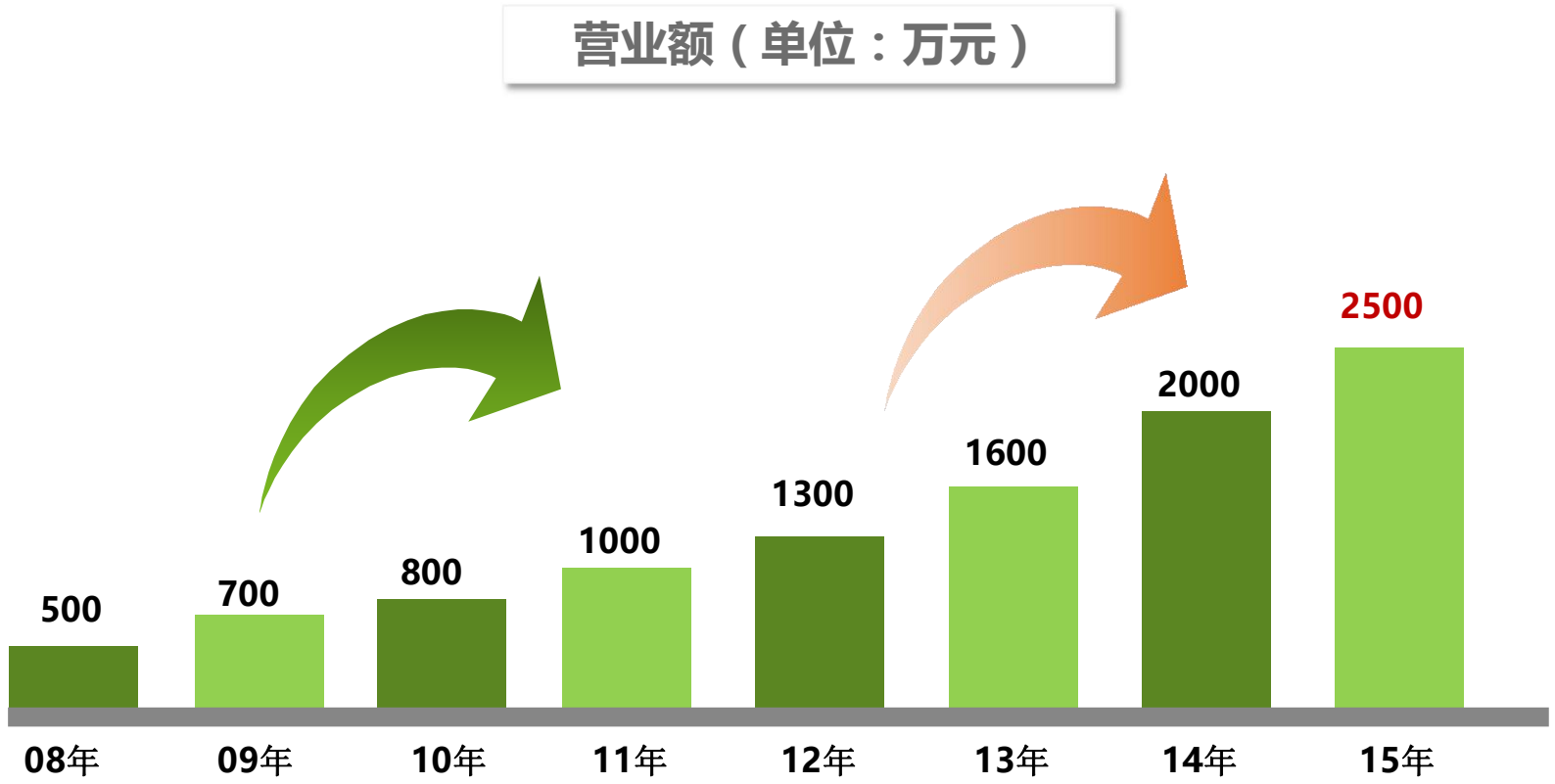
广州优亿信息科技有限公司

01 公司简介

我们的愿景：成为卓越的移动互联网信息化综合解决方案服务提供商
我们的使命：聚焦客户关注的压力和挑战，提供有竞争力的移动互联网产品及服务，帮助客户实现其战略愿景

公司简介

- 优亿科技位于广州市天河软件园金山园区，是一家专业提供移动互联网信息化综合解决方案的高科技企业
- “优亿”与“优异”谐音，“优”是优秀，“异”代表创新，代表了公司“追求卓越，敢于创新”的理念
- 此外，优亿也是英文Use Ease的两个字母缩写，意思是轻松使用，优亿科技从创立之日起即把客户放在首位，希望能够通过公司的专业化服务带给客户轻松、愉悦的体验。



优亿科技 USEEASE

公司简介

资质与荣誉

1、通过ISO9001：2008认证
2、通信系统集成丙级资质
3、双软认证企业
4、增值电信业务经营许可证
5、广州市高新技术企业认证
6、广通服优秀合作伙伴
7、广州市工程管理协会理事单位

优亿科技 USEEASE

公司简介

1 产品

- 大数据分析与挖掘系列
- 移动互联网产品
- 政企行业解决方案
- ……

2 服务

- 电信运营商各专业技术支撑
- 渠道运营服务
- 一体化设计
- ……

优亿科技
USEEASE

02 技术介绍

优亿公司在大数据方面有长期的积累、沉淀。主要产品包含：智能通用爬虫、DPI数据实时清洗与分类、电信客户画像、基于Spark的数据挖掘系统产品、大数据开发平台、移动互联网客户感知评估系统系统等。

技术背景介绍

Chapter 01 自动文摘系统

Chapter 02 自动归档

Chapter 03 问答系统（阅读理解）

优亿科技 USEEASE

自动文摘

自动文摘(Automatic Summarization):指通过机器学习的方法抽取文本的关键信息，对于源文本在长度上进行压缩和提炼。

主要的方法：

（1）基于无监督分类。

（2）基于监督分类。

（3）半监督分类。（迁移学习）

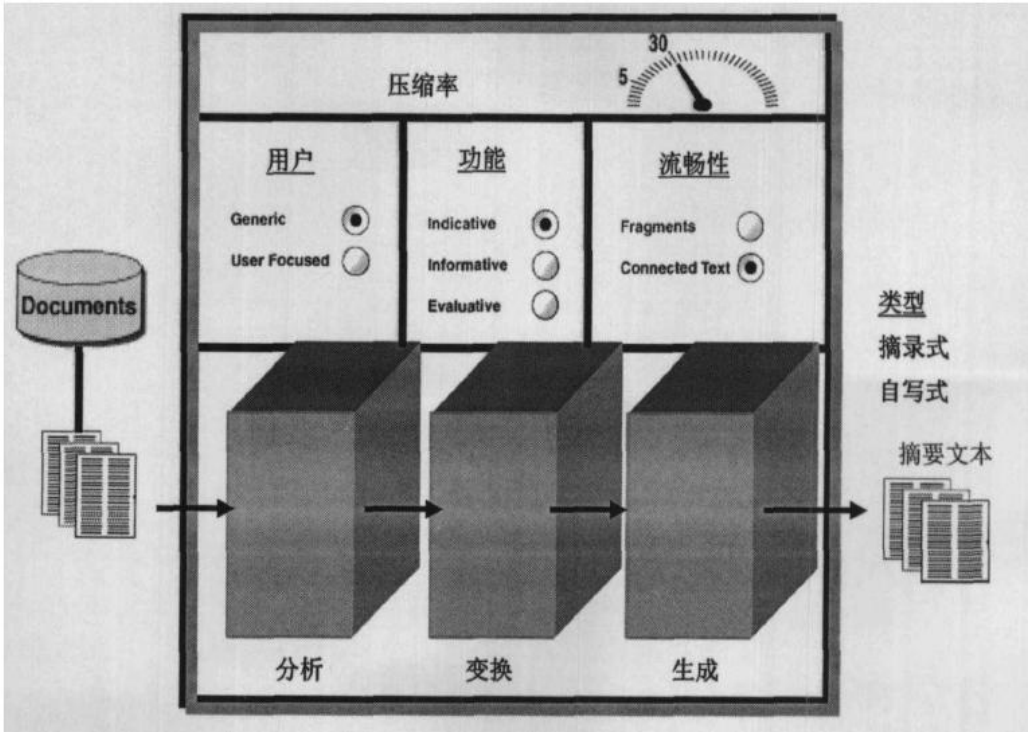
自动文摘

自动文摘系统主要可以分为三部分：

（1）分析。对输入的文本进行分析,得到文本内容的内部表示。

（2）变换。将本文内部表示转换成摘要表示。

（3）生成。将摘要表示翻译转换成人类可以理解的自然语言。



基于无监督分类自动文摘

基于无监督分类的自动文摘以摘句式文摘技术为主，即从源文本中抽取出合适的句子组成摘要,所以系统实现的核心就在于如何衡量一个句子是否能够作为摘要句。而摘要所要满足的两个基本要求就是信息量大和冗余性小。

方法步骤：

（1）抽取文本特征。如位置特征、指示词或指示短语、词频特征、指示词或指示短语、词汇链、句子相似度、文本结构。

（2）构建句子整体得分。如隐马尔可夫模型、条件最大嫡模型、条件最大嫡模型。

（3）通过对文中的句子进行聚类,将内容相近的句子聚为一类,然后从每类中抽出中心句组成摘要。或者，通过与文档的向量最相似的句向量，得到文章主题。

基于无监督分类自动文摘

AttSum系统一共分为三层：

1、CNN Layer，通过神经网络将句子映射到embedding上。

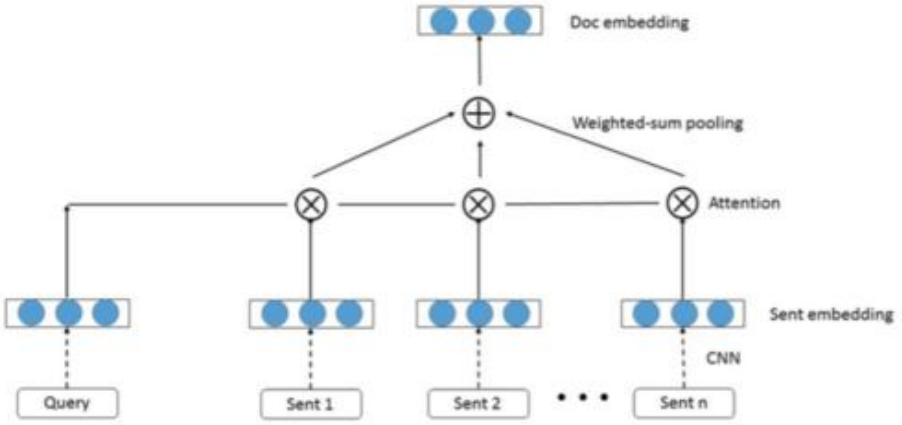
2、Pooling Layer，用注意力机制配合sentence embeddings构造document cluster embeddings。

3、Ranking Layer，计算sentence和document cluster之间的相似度，然后排序。

特点：

（1）应用了注意力机制。

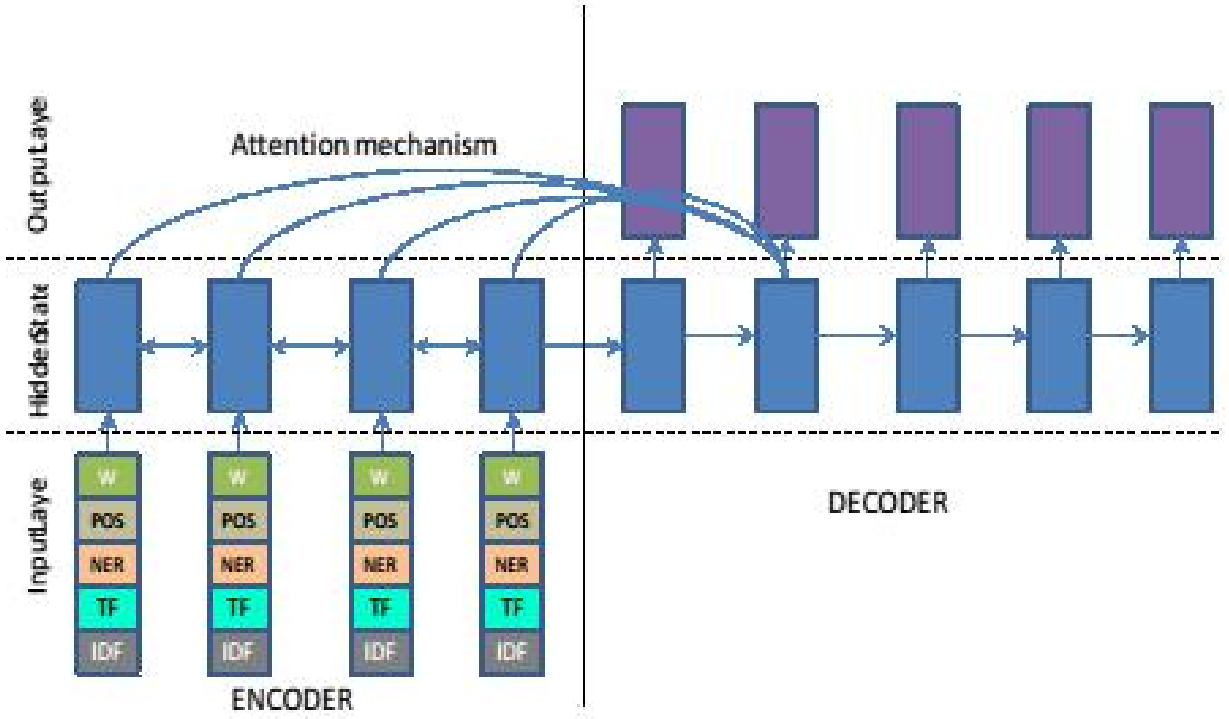
（2）一种联合查询相关性和句子显著性的神经网络模型。



基于监督分类自动文摘

传统的自动文摘主要基于Sentence compression，基本单元为句子j，是对核心句的抽取，无法形成新的句子。采用字词级别的摘要生成主要可以采用循环神经网络，如Google的textsum模型。

模型使用一种conditional RNN来生成摘要，条件是注意力模型，在给定输入句子的情况下，生成摘要的每个词，生成摘要依赖的条件是encoder的输出，encoder会计算输入中每个词的分数，这个分数可以理解为对输入作软对齐，模型可以进行end2end式地训练。



新闻自动文摘demo

采用基于Sentence级别进行构建，实现步骤如下：

（1）通过Word2vec构建通用词向量；

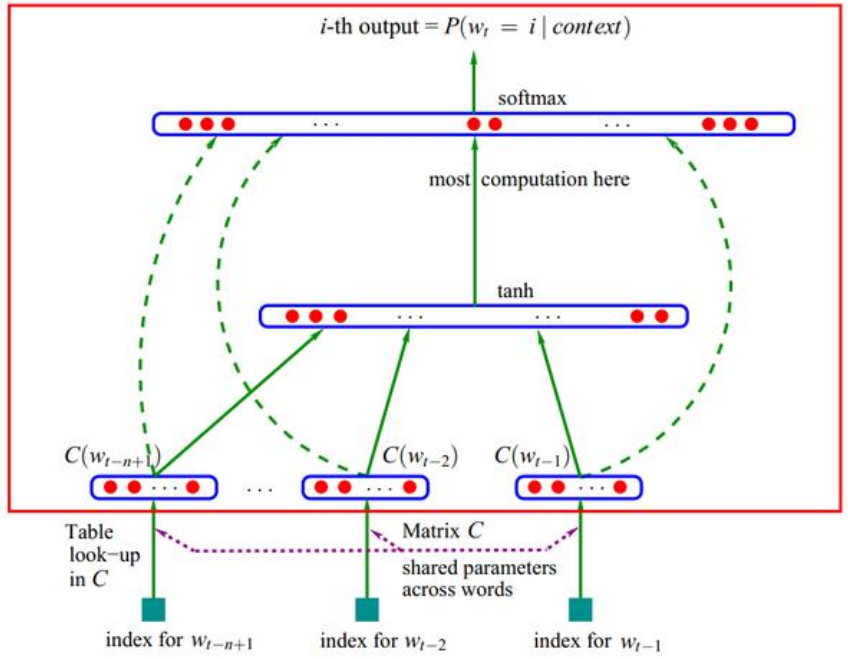
（2）将词向量相加得到句向量；

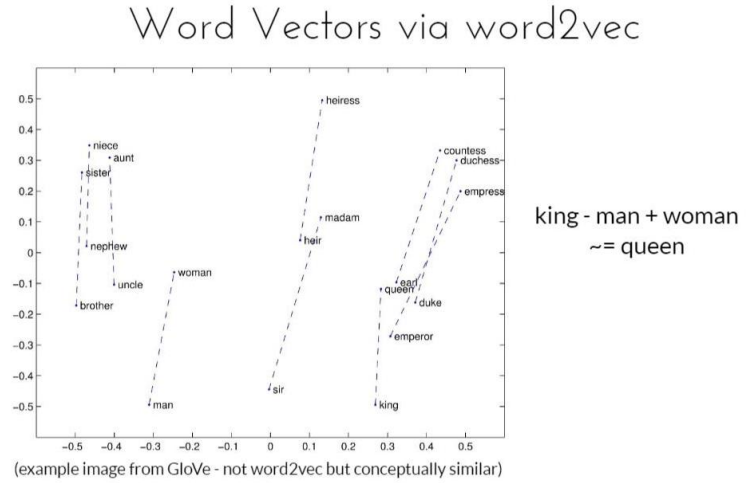
（3）将所有句向量相加得到文章向量；

（4）计算文章向量与句向量的相似度，然后排序；

（5）增加首句权重值；

（6）选取得分最高的句向量。





优亿科技 USEEASE

新闻自动文摘demo演示

演示

优亿科技
USEEASE

自动归档系统

自动归档，主要是通过对文本进行分析，从而实现按照某种特征对本文进行的聚类或分类任务。

主要步骤：

（1）抽取文本特征和关键词。传统的方法包括使用TF-IDF、词频、主题词。

（2）计算各文档之间的相似性。

（3）根据相似性对各文档进行聚类。

基于知识图谱的自动归档系统

基于知识图谱的自动归档方法：

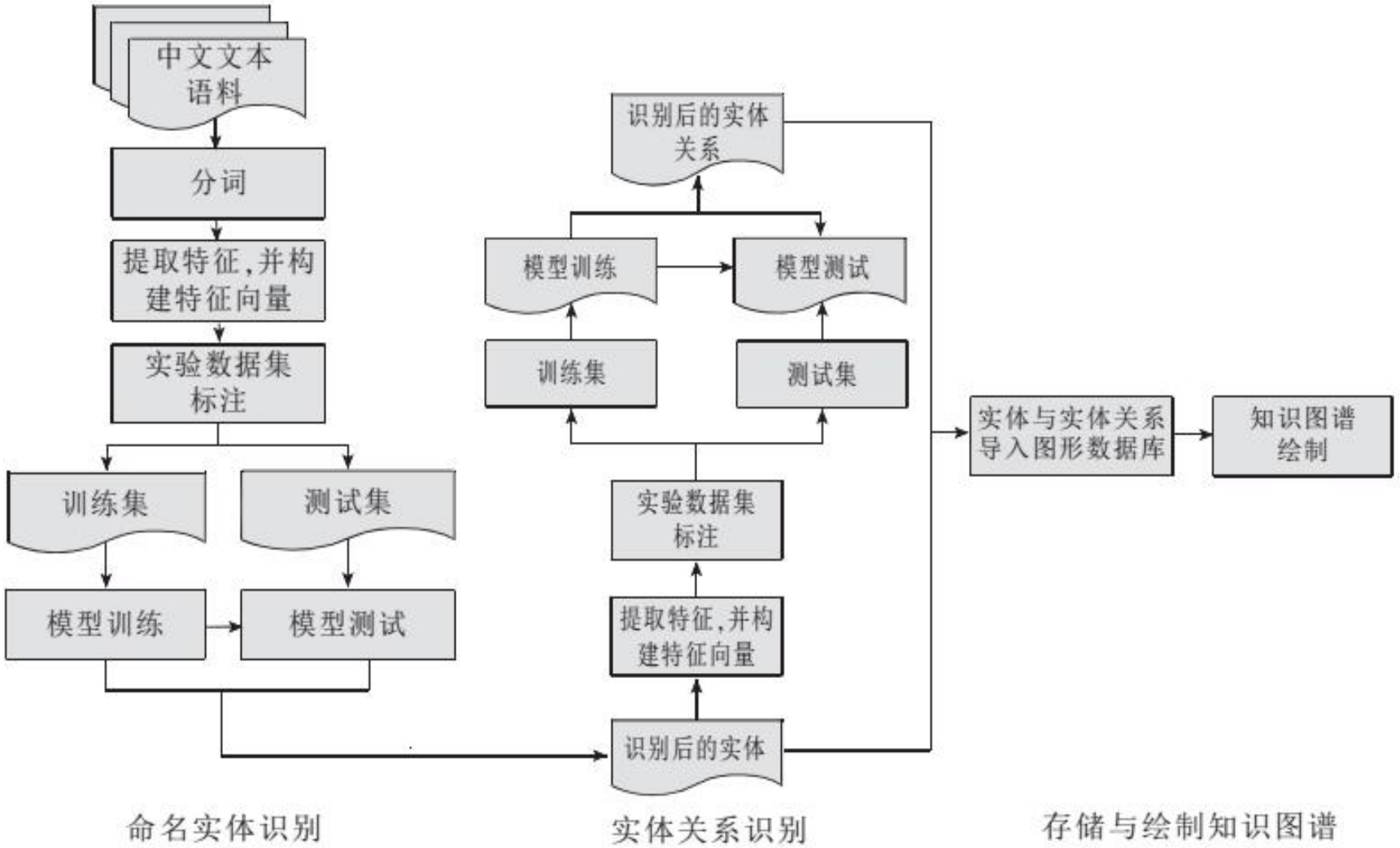
（1）构建基础知识库；

（2）抽取命名实体；

（3）将文档和知识库实体进行关联；

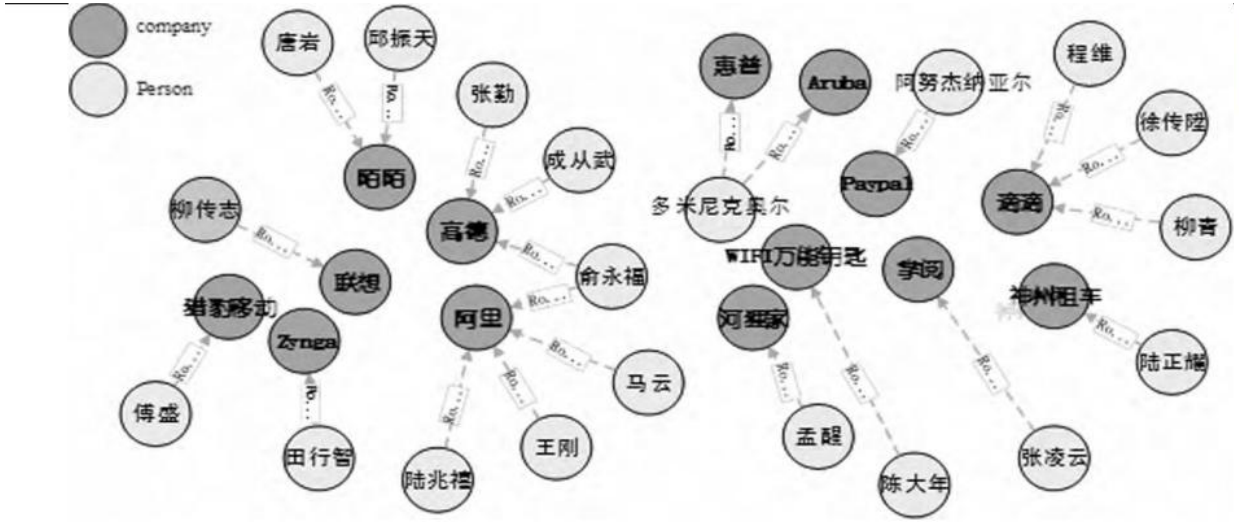
（5）通过实体对文章进行归档；

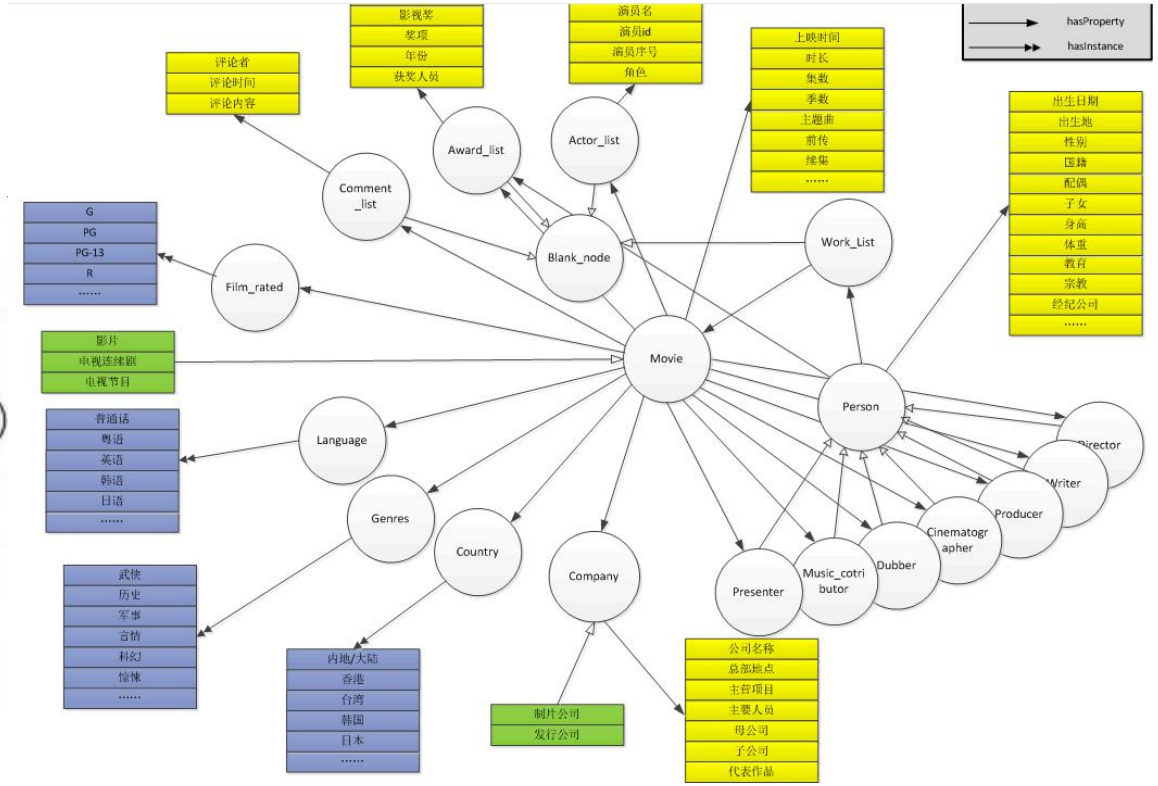
知识图谱构建



公司技术积累及相关成果

公司通过下载freebase数据，百度百科数据，通过实体识别和实体关系抽取技术，通过NEO4j作为三元组存储数据库，构建了通用知识图谱查询系统。

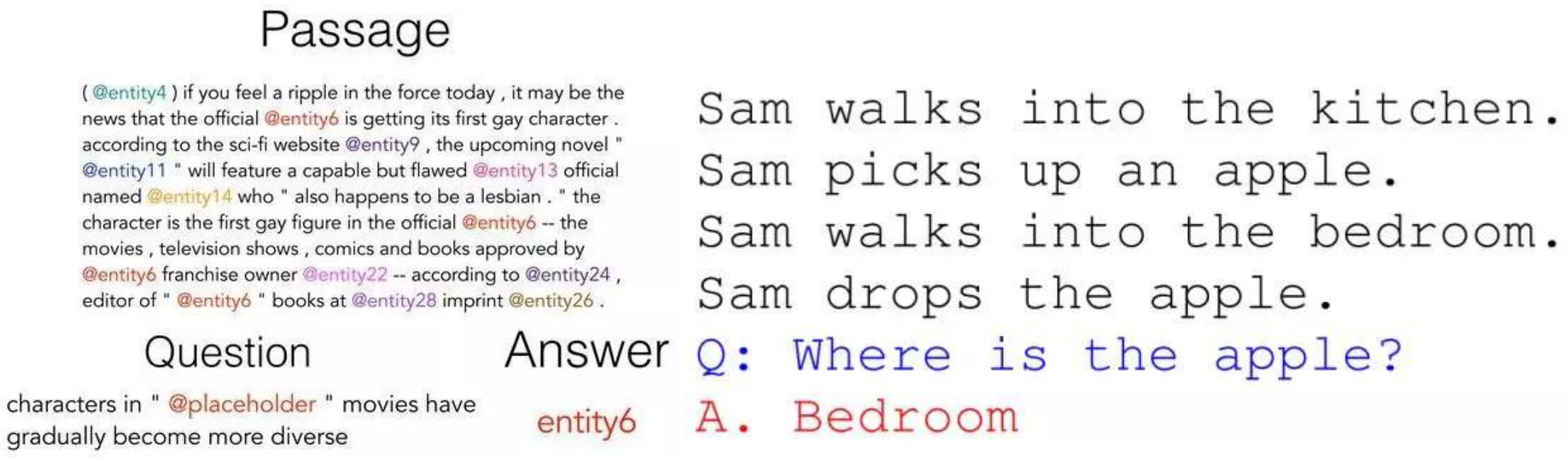




问答系统(阅读理解)

什么是机器阅读理解?

Cloze-style queries是类似于“完形填空”的任务，就是让计算机阅读并理解一篇文章内容后，对机器发出问题，问题往往是抽掉某个单词或者实体词的一个句子，而机器回答问题的过程就是将问题句子中被抽掉的单词或者实体词预测补全出来，一般要求这个被抽掉的单词或者实体词是在文章中出现过的。



一维匹配模型

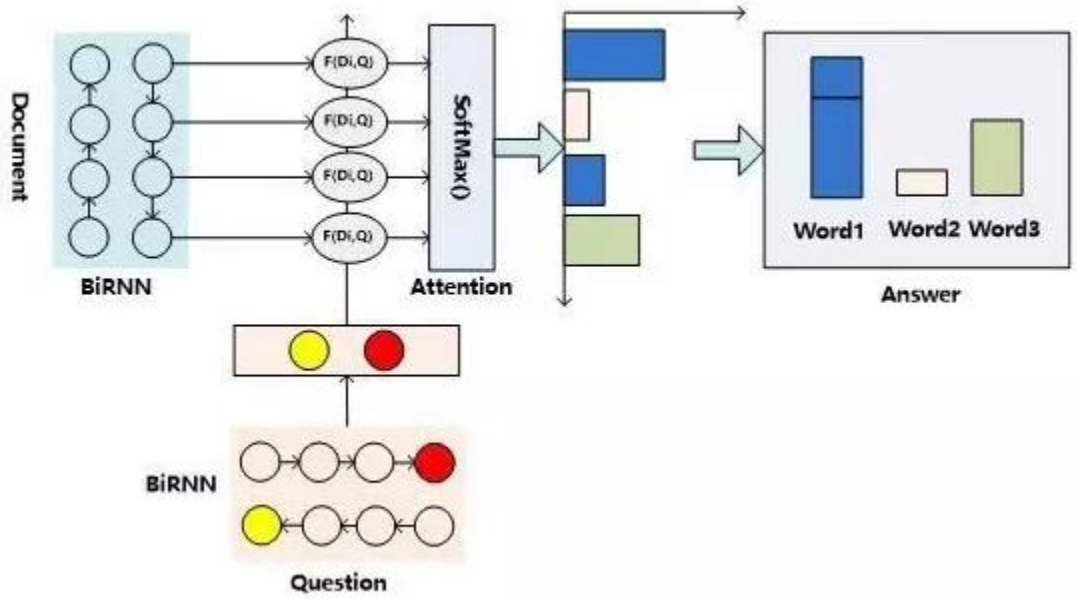
机器阅读理解任务常用的解决方案是使用“一维匹配模型”，主要架构：

（1）对文章内容使用BiRNN的方式对文章内容和问题进行语义编码。

（2）通过匹配函数来计算文章中每个单词Di语义和问题Q整体语义的匹配程度。如双线性（Bilinear）函数。

（3）对每个单词的匹配函数值通过SoftMax函数进行归一化，计算各类单词作为问题答案的可能性。

（4）将相同的单词概率进行累加。

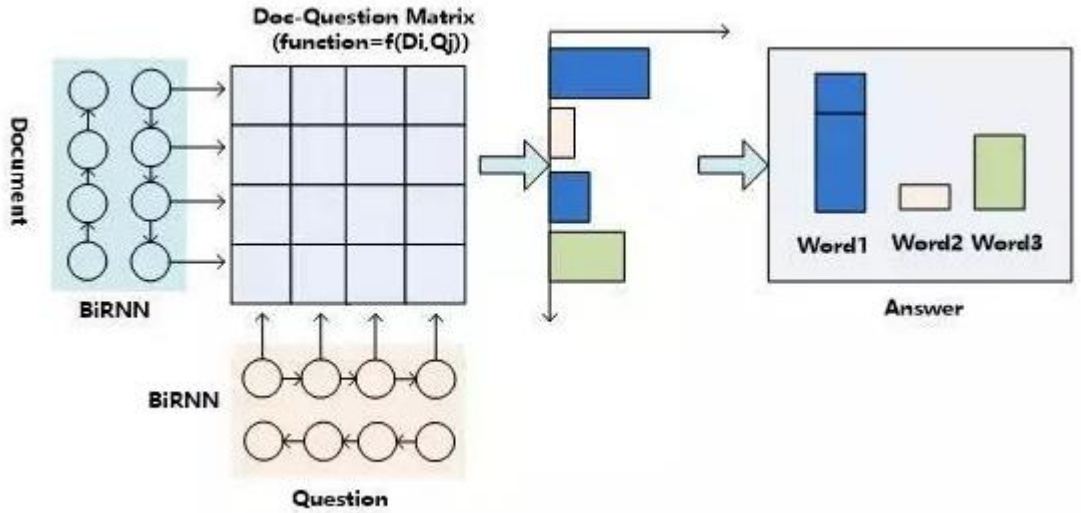


二维匹配模型

二维匹配模型结构上和一维匹配模型相似，主要区别在于匹配函数。具体：

（1）与一维匹配模型不同在于不是将问题的语义表达为一个整体，而是问题中的每个单词都单独用Word Embedding向量来表示。

（2）由于二维匹配模型将问题由整体表达语义的一维结构转换成为按照问题中每个单词及其上下文的语义的二维结构，明确引入了更多细节信息，所以整体而言模型效果要稍优于一维匹配模型。




Thanks for
your attention!