基于知识图谱的智能问答产品

武子越 马晓萱 邹曼琦 胡晓玥

Context

1. 背景和商业痛点
2. 具体解决方案
   a. 基于搜索引擎的知识问答
   b. NL2CQL进行本地图数据库查询
   c. 基于知识图谱的推理问答
   d. 基于图谱的自动文本生成
3. 产品创新与改进效果

01

背景与商业痛点

- ✓ 发明主题
- ✓ 现有问题
- ✓ 解决方案

1.1 发明主题

基于知识图谱的智能对话系统

具体功能：

1. 基于爬虫技术的实时搜索问答

2. 基于BERT等深度学习模型的自然语言转结构化查询语言

a. NL2SQL

b. NL2CQL

3. 基于知识图谱和图数据库技术进行逻辑推断、事理推理的对话问答

4. 基于知识图谱技术的文本自动生成

a. NL2PPT

b. NL2html

c. …

1.2 商业痛点和现有问题

1. 现有问答系统对于具体知识查询效果较差

- 还是以日常对话为主，知识类问答较少

  如：“太阳系中最大的行星”，“到2012年新中国成立多少年”
- 时效性不强，如“杭州当前的房价均价”

2. 人工查询关系型数据库操作困难

- 撰写复杂的SQL或其他查询语句
- 跨域查询需要繁杂的join操作

  如演员A和演员B共同出演过哪些电影，需要涉及演员、电影数据集

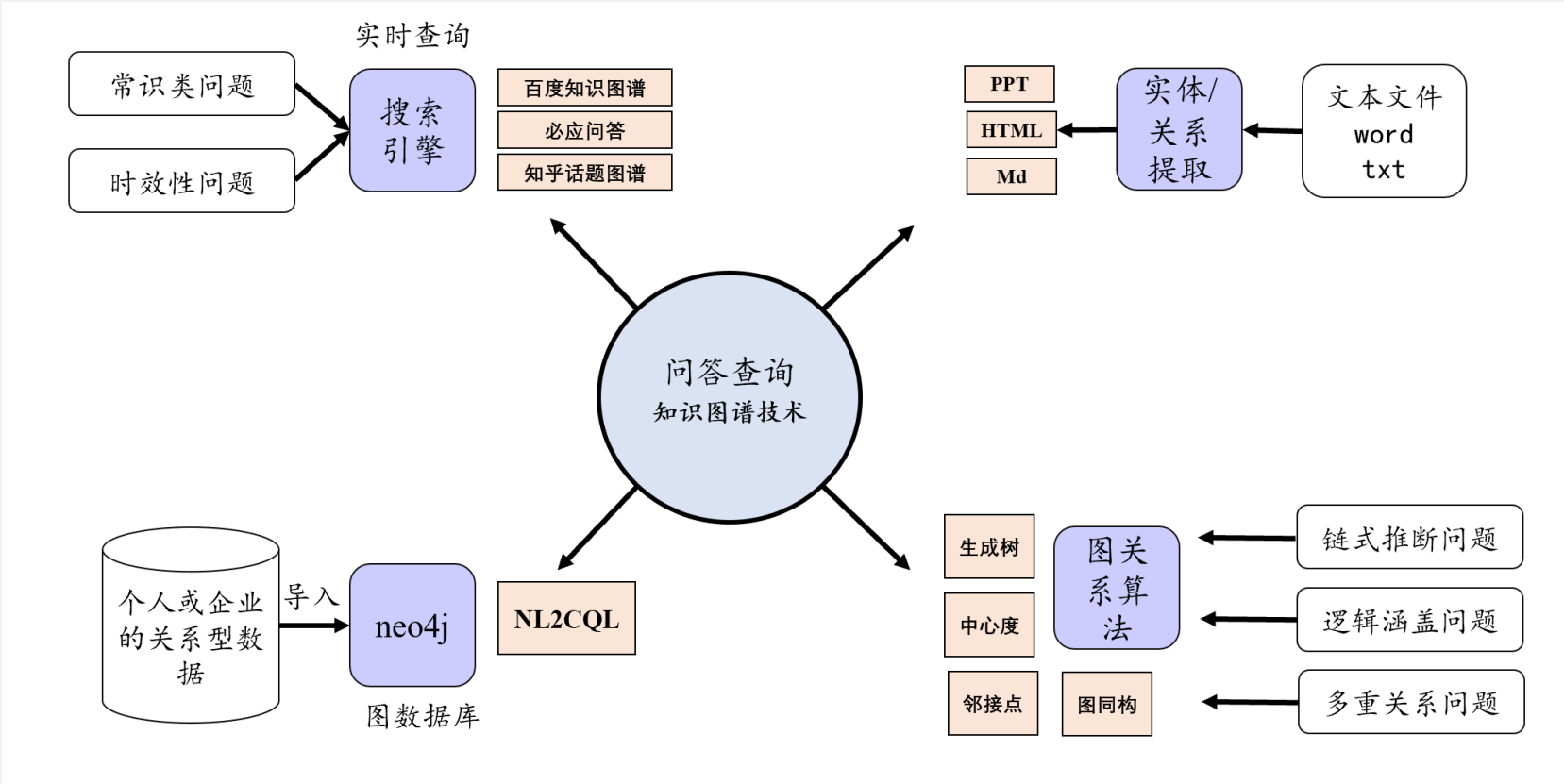
3. 对于未知的知识或关系，不能实现事理推理的效果

- 逻辑链推断，逻辑涵盖，多重关系问题

4. 直接将输入文档转化为PPT等流程繁杂

- 科研人员、公司经理需要花费大量时间将自己的报告制作为ppt

1.3 解决的整体思路


实时查询
常识类问题
时效性问题
搜索引擎
百度知识图谱
必应问答
知乎话题图谱
PPT
HTML
Md
实体/关系提取
文本文件
word
txt
问答查询
知识图谱技术
个人或企业的关系型数据
导入
neo4j
图数据库
NL2CQL
生成树
中心度
邻接点
图同构
图关系算法
链式推断问题
逻辑涵盖问题
多重关系问题

1.4 解决方案

1. 现有问答系统对于具体知识查询效果较差 **——结合搜索引擎的时效性问答**
   - 结合搜索引擎
   - 实体识别
   - 属性提取技术

2.人工查询关系型数据库操作困难 **——基于知识图谱及图数据库的智能问答**
   - NL2CQL算法
   - 本地图数据库

3. 对于未知的知识或关系，不能实现事理推理的效果 **——基于知识图谱的事理逻辑推断**
   - 图数据库保证了关系的连续性
   - 定位实体节点
   - 遍历相连节点寻找重合节点

4. 直接将输入文档转化为PPT等流程繁杂 **——基于知识图谱的文本自动生成**
   - 实体提取
   - 逻辑结构转化为图结构

02

具体技术方案

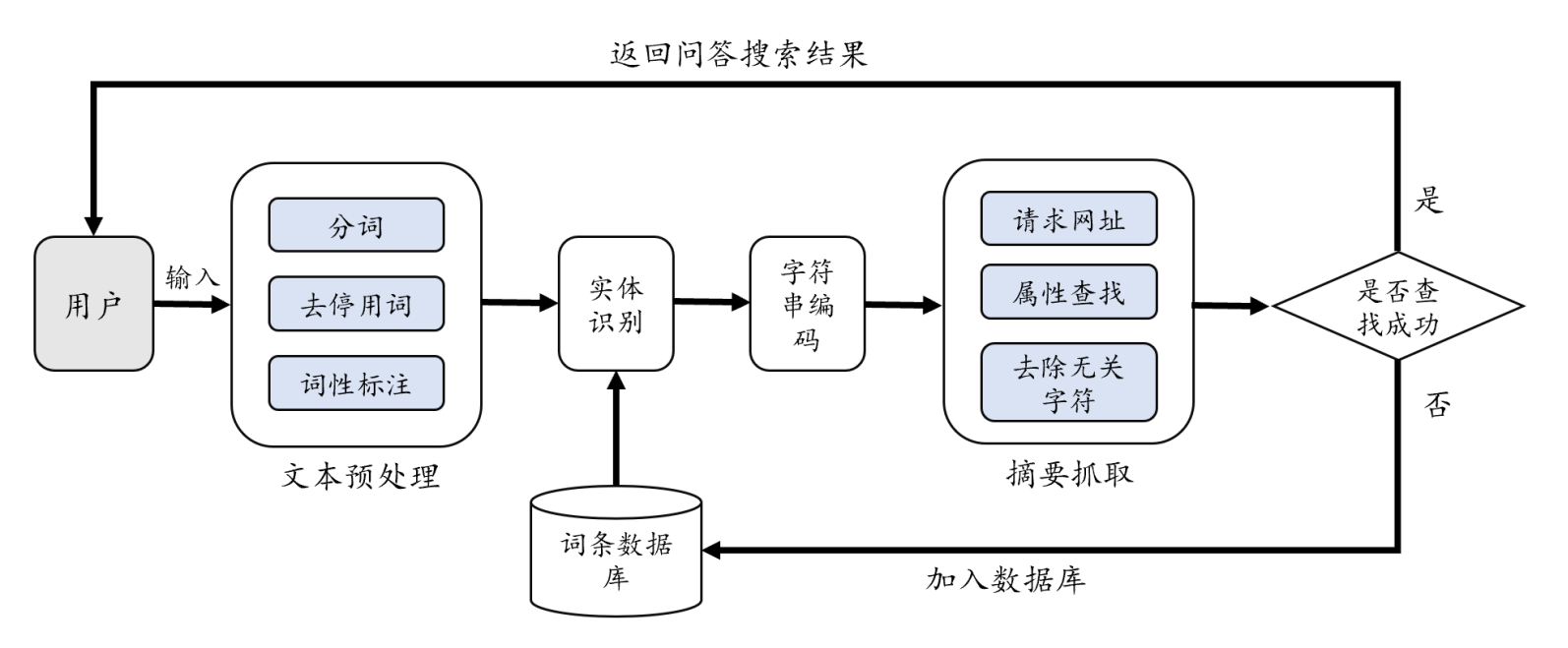
- ✓ 基于搜索引擎的知识问答
- ✓ NL2CQL的本地数据查询
- ✓ 基于知识图谱的推理
- ✓ 基于图谱技术的文本自动生成

2.1 基于搜索引擎的知识问答

概述

本智能问答系统根据问题类型的不同采取不同的方式获取答案。对于答案相对固定的知识形问题，如“中国的首都在哪？”，“2020年春节是什么时候？”，先将自然语言转化为结构化查询语言，通过网络爬虫的方式在网上搜索出对应的信息做出回复。

2.1 基于搜索引擎的知识问答——工作流程



基于搜索引擎知识问答流程

2.1 基于搜索引擎的知识问答——工作流程

1. 文本预处理
jieba库来进行分词
人工定义停用词表 (连词, 介词, 语气词等)

2. 实体识别
利用分词结果进行POS词性标注
向量化编码并使用深度学习模型进行处理

3. 字符编码转化
urllib.parse.quote对字符串进行编码
转化为可以进行网页搜索的字符

4. 请求站点网址
对指定网页利用requests.get()请求站点网址
一般抓取前十条摘要。

5. 属性查找
利用BeautifulSoup查找html文件中的指定元素
标签(tag)中的name、attr属性等， 以及id、href属性等

6. 去除冗余字符
使用extract()去除无关标签

7. 返回查询结果

2.1 基于搜索引擎的知识问答——例子

1. 基本知识问答

该产品能够回答包括但不限于如下基本知识型问题，如“浙江大学在哪个省？”，返回“浙江省”；“浙江大学在哪个市？”，返回“杭州市”；世界之最等常识类问题等。

2.1 基于搜索引擎的知识问答——例子

2. 日期等时效性回答

该产品能够回答具有时效性的问题，如“今年杭州房价均价是多少？”“到2012年新中国成立多少年了？”，该产品能够利用搜索引擎进行搜索返回最佳答案。

2.1 基于搜索引擎的知识问答——创新点

1. 返回简洁关键词答案

相比于一般搜索引擎，该搜索方式直接整合到了问答系统中，返回关键词，防止网页广告、冗余信息对于用户造成的干扰。

2. 借助搜索引擎具有时效性地进行知识更新

通过爬取网页的方式进行知识搜索可以保证信息得以及时更新，不用担心本地知识过时的问题。

3. 减少问答系统维护的人力成本

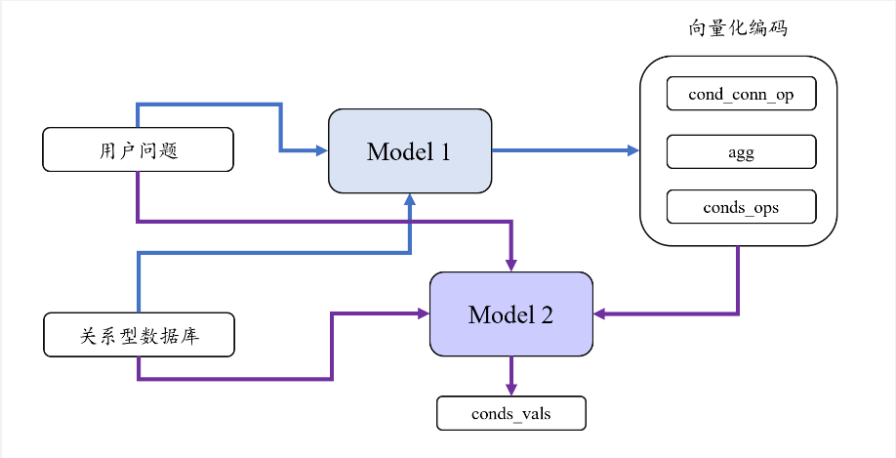
基于本地数据库的问答系统往往涉及到繁重的知识库更新的过程，而基于搜索引擎的问答能够减少维护成本。

2.2 NL2CQL进行本地图数据库查询

概述

对于具体到某个领域、频繁使用的知识类问题，可以建立相应的本地数据库，将自然语言转化成SQL语言对本地数据库进行查询。对于图数据库，可以使用CQL语言，其原理与NL2SQL非常相近，可以直接讲NL2SQL的算法用于NL2CQL中。

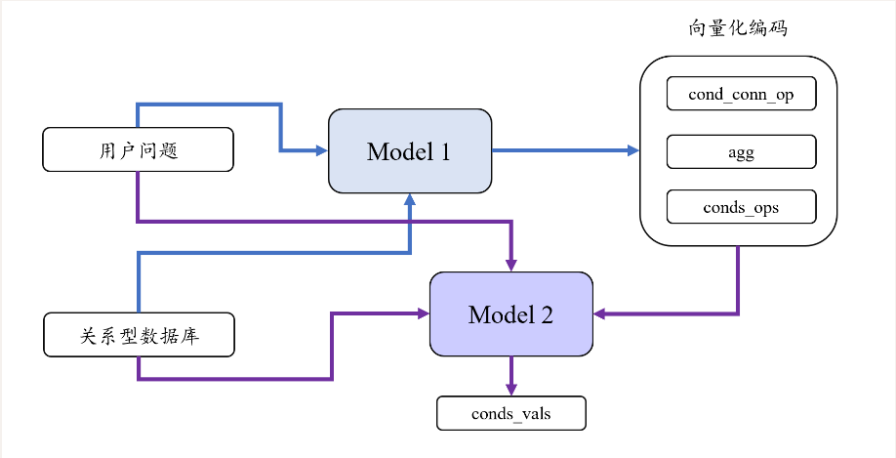
2.2 NL2CQL——工作流程



1. 向量化编码

- 将查询的一些语句及关键词进行向量化编码
- 如模型中agg表示对数据的运算操作，比如 sum, max, min 等；conds_ops代表 WHERE 语句中的一系列条件，每个条件是一个由 (列, 运算符, 条件值) 构成的三元组；cond_conn_代表 conds 中各条件之间的并列关系(and, or等)

2.2 NL2CQL——工作流程



2. 两阶段模型融合

- Model1预测SQL或CQL语句中conds_ops等向量化编码部分
- Model2将Model1传来最高概率的向量，将其重新转化为SQL或CQL语言并按顺序拼接。

3. 在图数据库中进行查询

2.2 NL2CQL——工作流程

Model1

将自然语言问题和所有表格表头顺序链接，每个部分之间用token区分，不同类型的column标记不同。

接着使用BERT模型对自然语言问句和表头进行编码。具体采取哈工大讯飞联合实验室发布的BERT-wwm, Chinese。BERT可以学习文本中单词的上下文关系，其中使用的Masked LM，用序列中未被标记的单词的上下文来预测被掩盖的单词，使结果的情景意识增加。

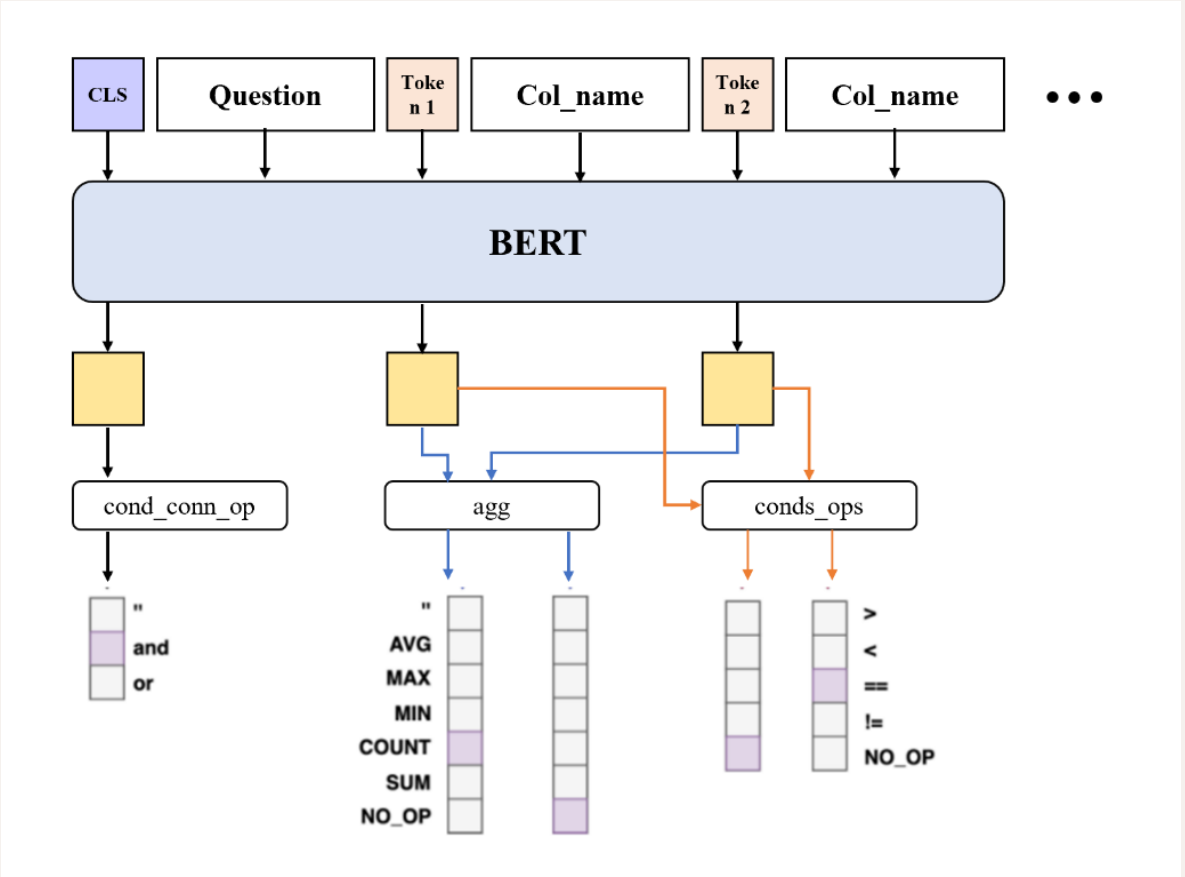
在输出的最后一层通过softmax来预测每种运算符的概率大小，返回概率最大的关系组合

Model2

根据Model1选择的cond_col, 枚举cond_op与cond_val,生成一系列候选组合，和自然语言问题顺序排列，通过BERT进行编码，编码后当成多个二分类问题进行分类，最后将预测结果进行合并

将最后生成的CQL语句输入本地neo4j数据库进行查询，并将结果返回用户。

2.2 NL2CQL——工作流程



2.2 NL2CQL进行本地数据库查询——创新点

1. 从NL2SQL推广到NL2CQL及其他结构化查询语言

该产品利用NL2SQL的技术，推广到生成CQL等结构化查询语言，该模型在这类语言中同样适用。

2. 将NL2CQL算法在问答产品上落地

目前虽有NL2SQL相关技术，但主要的实现还是基于一个算中文文本转SQL法比赛，而无现成的商业产品。而本系统可在导入本地数据库后通过自然语言直接对所需内容进行提取，在各行各业都可以得到广泛使用，也可使外行人更方便地获取行业内的信息。

3. 结合知识图谱的相关技术进行产品的完善

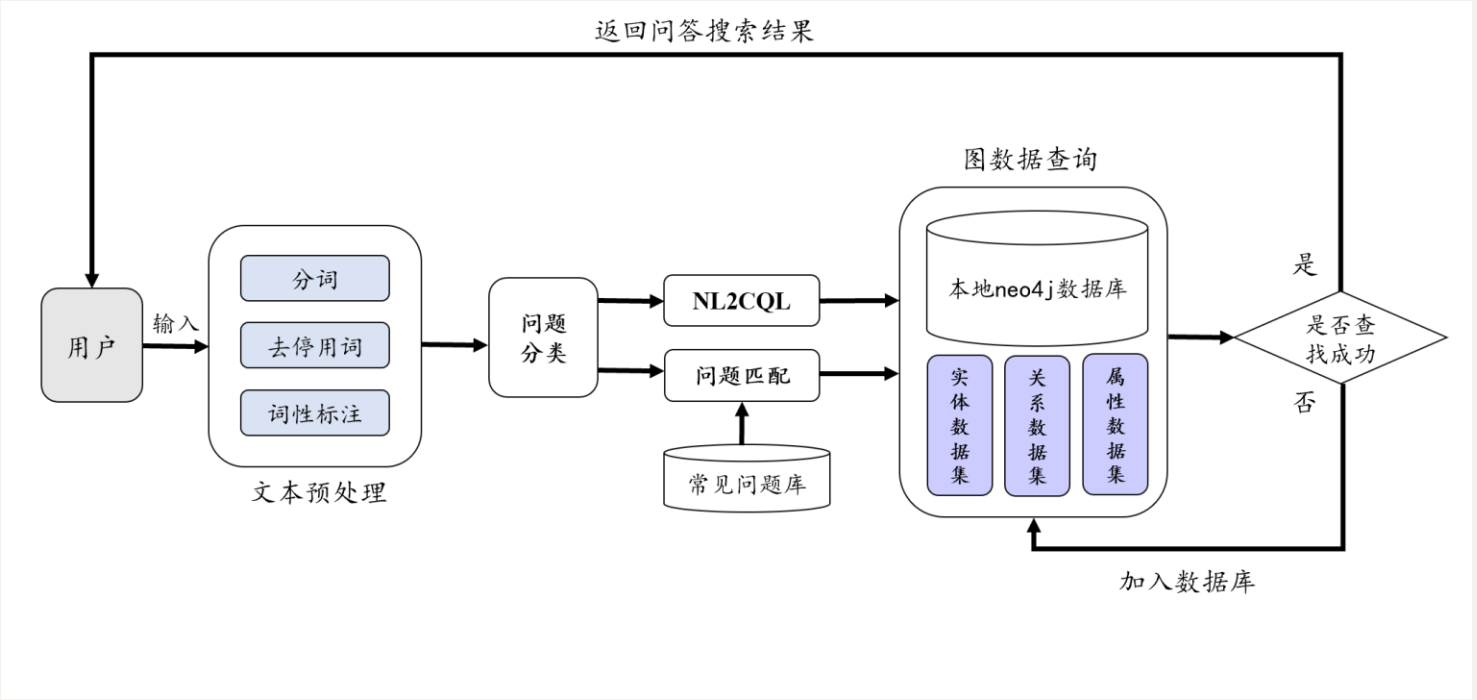
图数据库在各行各业应用前景较广，如对于会计事务所，可以将各公司的财务报表以图谱关系的形式录入公司本地数据库，通过自然语言即可调出所需要的财务数据，如“2016年XXX公司应收账款为多少？”等问题。对于医药行业可以将各药品成分、疗效、使用说明等信息录入本地数据库，在需要时可调取所需的药品信息，如“风寒感冒应该吃什么药？”，“XXX药品的疗效是什么？”等问题。

2.3 基于知识图谱的推理问答

概述

基于知识图谱的推理可以克服传统数据库中对于关系型问题难以查询的困难，同时结合到问答系统，能够完成一些语义推理、事理推理的基本功能，从而改进传统问答系统在此方面不够智能，与人类问答方式相差较远的局限性。

2.3 基于知识图谱的推理问答——工作流程



基于知识图谱的推理问答流程

2.3 基于知识图谱的推理问答——工作流程

1. 驱动本地neo4j数据库
使用Neo4j的Python驱动包py2neo进行实现

2. 文本预处理

3. 问题分类
利用朴素贝叶斯算法对于用户输入的问题进行分类，如对于电影相关的问题，分类为“查询电影信息，查询演员信息，查询电影与演员关系…”等

4. NL2CQL
将用户输入的自然语言转化为CQL语言，直接将CQL语句输入图数据库得到结果

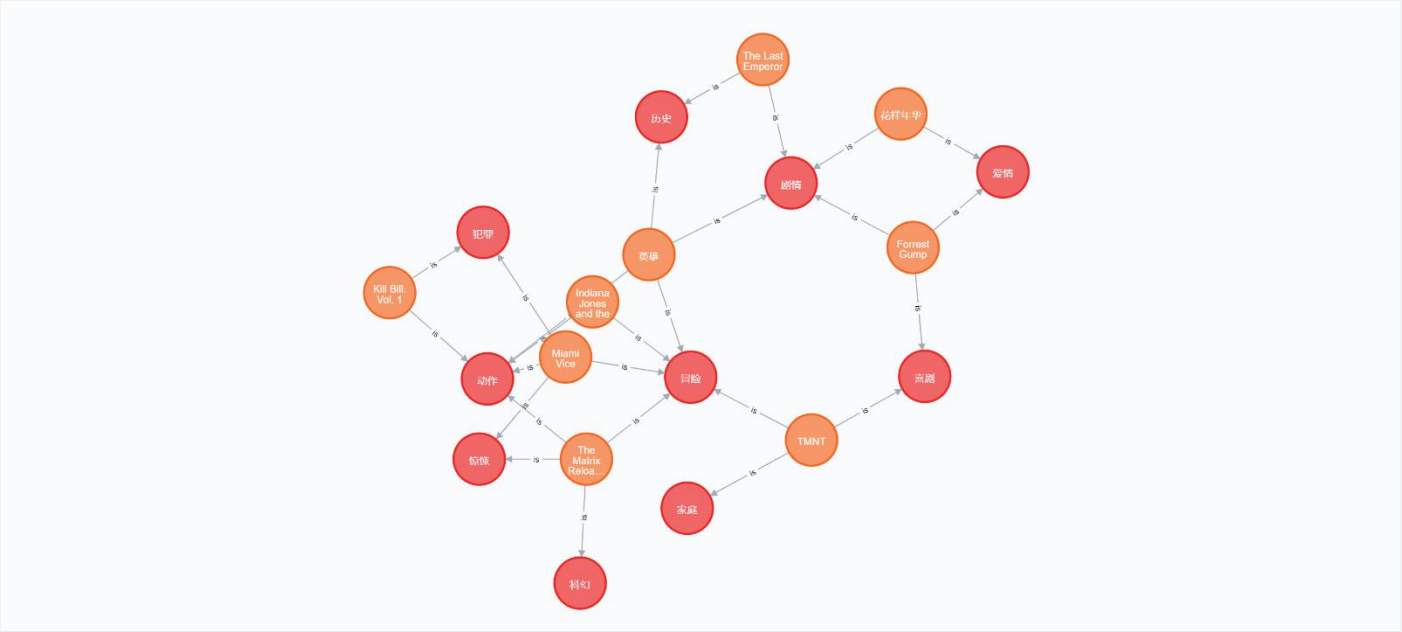
5. 问题匹配
直接进行问题的匹配，对于特定领域的问题，利用常见问题库进行匹配在精准度和效率上都会更好

6. 利用neo4j数据库构建图谱
企业使用本系统时，需要将自己的业务数据（包括实体数据、属性数据以及关系型数据）放入本地neo4j数据库中

7. 返回查询结果

8. 可视化
Neo4j数据库提供可视化的功能，当用户或企业对于问答结果有所疑问的时候，可以将搜索的结果可视化

2.3 基于知识图谱的推理问答——工作流程



Neo4j自带的关系图谱可视化

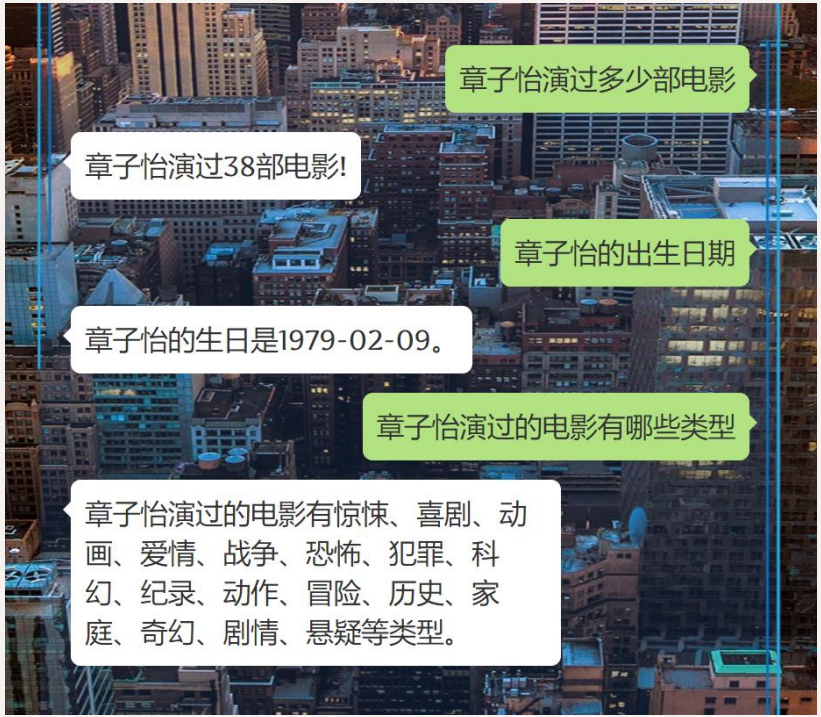
2.3 基于知识图谱的推理问答——例子

1. 关系型问题的准确回答

该产品能够回答包含但不局限于如下关系型问题，如“章子怡演过什么类型的电影”，只需要搜索到“章子怡”节点，并将与其连接的电影节点进行遍历统计即可。

2. 实体与实体的关系统计问答

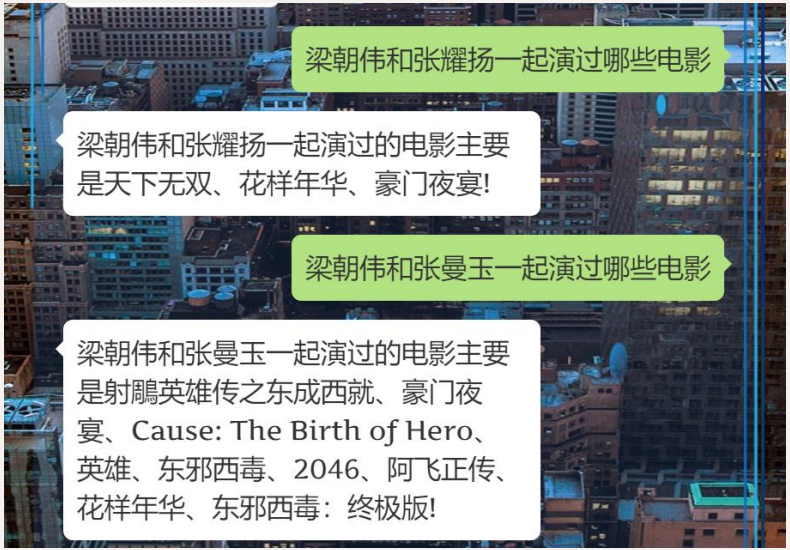
该产品能够回答包含但不局限于如下关系型问题，如“章子怡演过多少部电影”，只需要搜索到“章子怡”节点，并将与其连接的电影节点进行计数，或是对于特定节点的入度、出度进行计算即可。



2.3 基于知识图谱的推理问答——例子

3. 实体与实体关系的总结

该产品能够回答包含但不局限于如下关系型问题，如“梁朝伟和张曼玉共同演过哪些电影”，只需要搜索到“梁朝伟”“张曼玉”的节点，并搜索其共同连接的节点即可。

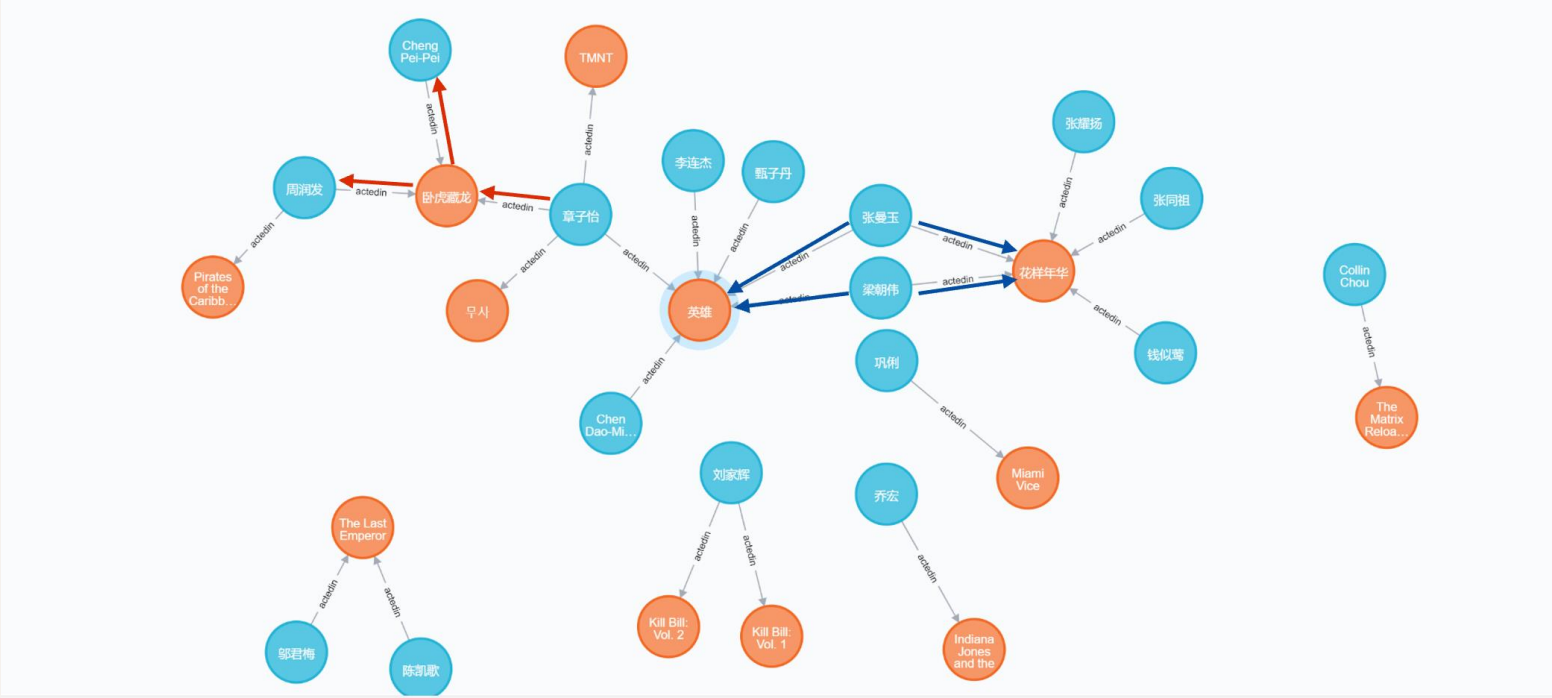


4. 实体关系的推断

该产品能够回答包含但不局限于如下关系型问题，如“章子怡出演电影的合作者是谁”，只需要定位到“章子怡”，搜索电影节点，并对子节点再进行深度搜索，两层搜索后便能够回答这种带有双重逻辑的问题。

2.3 基于知识图谱的推理问答——例子

可解释性



2.3 基于知识图谱的推理问答——创新点

1. 减少传统数据库繁琐的查询，提高查询的效率

无需繁杂的join操作，图的结构使得关系的查询转化为节点与边的查询问题

2. 具有关系汇总统计的功能

通过计算出入度，中心度等方式将关系进行汇总，完成基于传统数据库问答系统无法完成的工作。

3. 能够进行事理的推断

只要节点之间是连通的，其中的关系就可以用所进过路径的节点和边进行表示与推断。

4. 能够进行多层逻辑的推断

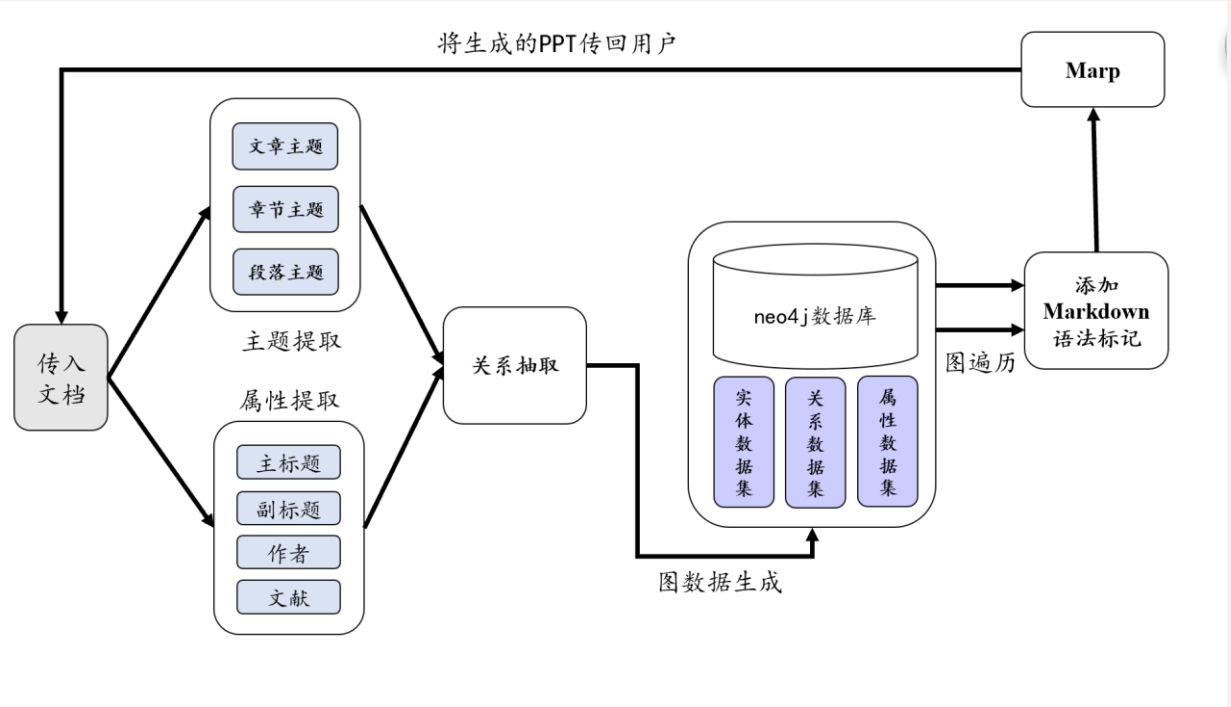
通过图的相关算法能够发现这些复杂的结构，同时拥有逻辑上的可解释性。

2.4 自然语言转PPT

概述

在ppt制作的过程中，使用者可以不必将文稿重新编辑成ppt所需要的大纲版式，通过实体识别等知识图谱技术判定文稿的逻辑关系，将文稿内容转化为图谱结构，基于此结构生成ppt

2.4 自然语言转PPT——工作流程



自然语言转PPT工作流程

2.4 自然语言转PPT——Marp简述

```
---
marp: true
theme: theme1
paginate: true
footer: '浙江大学管理学院'
---



![bg left:33% blur:1px](1.jpg)

# 用Marp制作PPT

1.这是第一章
2.这是第二章
3.这是第三章

---


## 这是章节标题

这是公式：

$$A=U\Sigma V^T, A_{m\times n}\approx U^1_{m\times r}\Sigma_{r\times r}^1 (V_{n\times r}^{1})^T $$

$$
\left[
 \begin{matrix}
  0 &  &  & \\
  1 & 0 &  &  \\
  1 & 1 & 0 &  \\
  1 & 1 & 1 & 0
  \end{matrix}
\right]
$$

$$T(X,Y)=\frac{XY^T}{\sqrt{|X|^2\times|Y|^2}}$$

---
```

用Marp制作PPT

1.这是第一章
2.这是第二章
3.这是第三章

浙江大学管理学院 1

这是章节标题

这是公式：

$$A = U\Sigma V^T, A_{m\times n} \approx U^1_{m\times r}\Sigma^1_{r\times r}(V^1_{n\times r})^T$$

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

$$T(X,Y) = \frac{XY^T}{\sqrt{|X|^2 \times |Y|^2}}$$

浙江大学管理学院 2

2.4 自然语言转PPT——工作流程

1. 属性提取
确定文本的已有的自然分段，确定文档中的文本、图片、公式等内容的大致位置归属。

2. 主题提取
TextRank

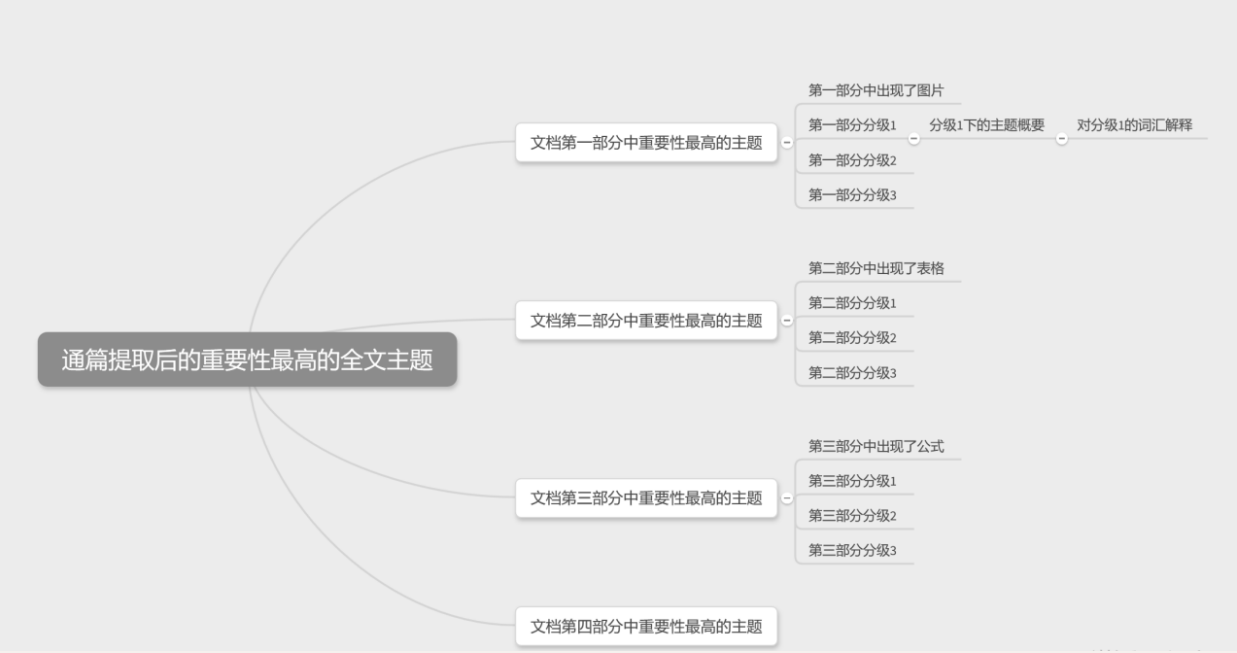
3. 关系抽取
将文章中不同逻辑结构的关系抽取出，如每一个段落之间的包含、顺承、并列等逻辑关系。

4. 图谱生成
根据已生成的不同层级的句段，产生知识图谱。

5. 对图谱进行遍历并添加Markdown标记
如利用广度优先算法，第一层遍历到的节点为大标题，下一层为大标题下的主题等，并根据markdown语言格式直接生成对应的markdown语言。对遍历到的一级、二级标题等，添加“#”，“##”等标题定位符；对遍历到的公式、表格、图片转化成对应的markdown格式；将图的一个分支遍历完后，添加“---”等Markdown分页符等。

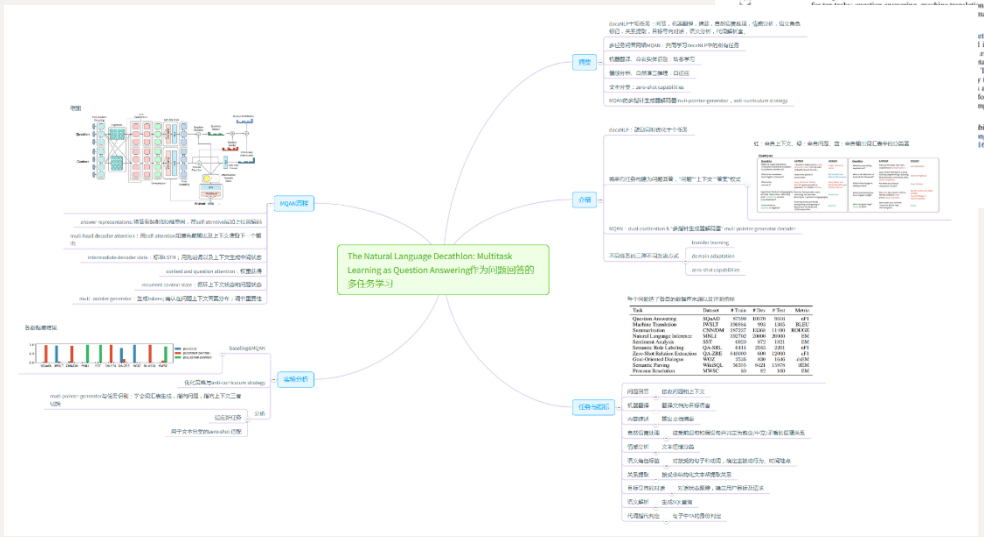
6. Markdown转PPT
利用Marp等第三方语言将生成的markdown文档转化为PPT等其他文档。

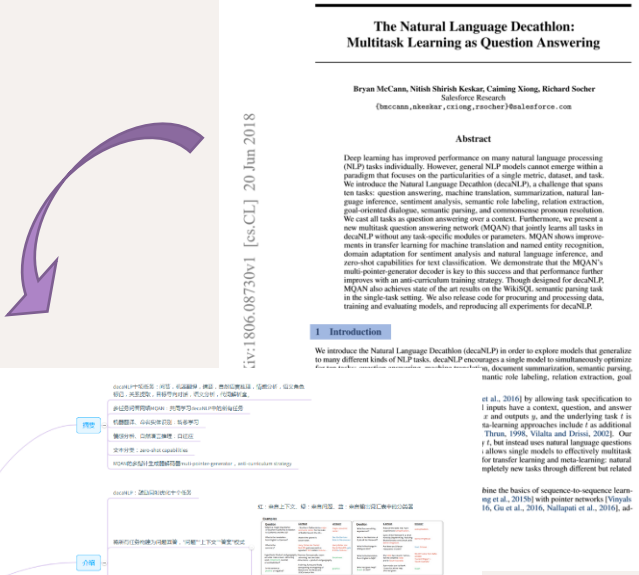
2.4 自然语言转PPT——工作流程



按照文章框架结构生成图谱

2.4 自然语言转PPT——例子





The Natural Language Decathlon:
Multitask Learning as Question Answering

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, Richard Socher
Salesforce Research
{bmccann,nkeskar,cxiong,rsocher}@salesforce.com

Abstract

Deep learning has improved performance on many natural language processing (NLP) tasks individually. However, general NLP models cannot emerge within a paradigm that focuses on the particularities of a single metric, dataset, and task. We introduce the Natural Language Decathlon (decaNLP), a challenge that spans ten tasks: question answering, machine translation, summarization, natural language inference, sentiment analysis, semantic role labeling, relation extraction, goal-oriented dialogue, semantic parsing, and commonsense pronoun resolution. We cast all tasks as question answering over a context. Furthermore, we present a new multitask question answering network (MQAN) that jointly learns all tasks in decaNLP without any task-specific modules or parameters. MQAN shows improvements in transfer learning for machine translation and named entity recognition, domain adaptation for sentiment analysis and natural language inference, and zero-shot capabilities for text classification. We demonstrate that the MQAN's multi-pointer-generator decoder is key to this success and that performance further improves with an anti-curriculum training strategy. Though designed for decaNLP, MQAN also achieves state of the art results on the WikiSQL semantic parsing task in the single-task setting. We also release code for procuring and processing data, training and evaluating models, and reproducing all experiments for decaNLP.

1 Introduction

We introduce the Natural Language Decathlon (decaNLP) in order to explore models that generalize to many different kinds of NLP tasks. decaNLP encourages a single model to simultaneously optimize

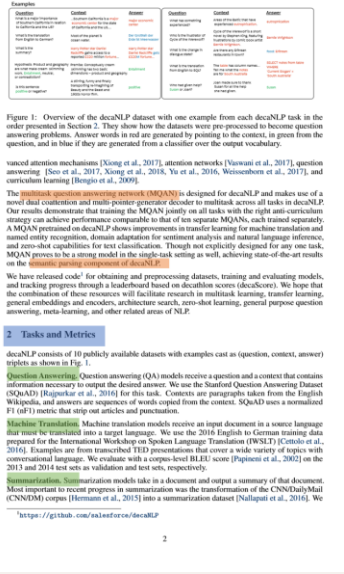


Figure 1: Overview of the decaNLP dataset with one example from each decaNLP task in the order presented in Section 2. They show how the datasets were pre-processed to become question answering problems. Answer words in red are generated by pointing to the context, in green from the question, and in blue if they are generated from a classifier over the output vocabulary.

vanced attention mechanisms [Xiong et al., 2017], attention networks [Vaswani et al., 2017], question answering [Seo et al., 2017, Xiong et al., 2018, Yu et al., 2016, Weissenborn et al., 2017], and curriculum learning [Bengio et al., 2009].

The multitask question answering network (MQAN) is designed for decaNLP and makes use of a novel dual coattention and multi-pointer-generator decoder to multitask across all tasks in decaNLP. Our results demonstrate that training the MQAN jointly on all tasks with the right anti-curriculum strategy can achieve performance comparable to that of ten separate MQANs, each trained separately. A MQAN pretrained on decaNLP shows improvements in transfer learning for machine translation and named entity recognition, domain adaptation for sentiment analysis and natural language inference, and zero-shot capabilities for text classification. Though not explicitly designed for any one task, MQAN proves to be a strong model in the single-task setting as well, achieving state-of-the-art results on the semantic parsing component of decaNLP.

We have released code[1] for obtaining and preprocessing datasets, training and evaluating models, and tracking progress through a leaderboard based on decathlon scores (decaScore). We hope that the combination of these resources will facilitate research in multitask learning, transfer learning, general embeddings and encoders, architecture search, zero-shot learning, general purpose question answering, meta-learning, and other related areas of NLP.

2 Tasks and Metrics

decaNLP consists of 10 publicly available datasets with examples cast as (question, context, answer) triplets as shown in Fig. 1.

Question Answering. Question answering (QA) models receive a question and a context that contains information necessary to output the desired answer. We use the Stanford Question Answering Dataset (SQuAD) [Rajpurkar et al., 2016] for this task. Contexts are paragraphs taken from the English Wikipedia, and answers are sequences of words copied from the context. SQuAD uses a normalized F1 (nF1) metric that strip out articles and punctuation.

Machine Translation. Machine translation models receive an input document in a source language that must be translated into a target language. We use the 2016 English to German training data prepared for the International Workshop on Spoken Language Translation (IWSLT) [Cettolo et al., 2016]. Examples are from transcribed TED presentations that cover a wide variety of topics with conversational language. We evaluate with a corpus-level BLEU score [Papineni et al., 2002] on the 2013 and 2014 test sets as validation and test sets, respectively.

Summarization. Summarization models take in a document and output a summary of that document. Most important to recent progress in summarization was the transformation of the CNN/DailyMail (CNN/DM) corpus [Hermann et al., 2015] into a summarization dataset [Nallapati et al., 2016]. We

[1] https://github.com/salesforce/decaNLP

2

2.4 自然语言转PPT——创新点

1. 将知识图谱技术拓展到自动文本生成领域

自动文本生成通常使用经典NLP中的一些方法，然而考虑到文本中存在着内在逻辑关系，因此从图谱的角度入手，从逻辑关系发现的方法出发，能够更高效精准地实现文本的转换。

2. 解决Markdown等语言学习曲线陡峭问题

对于还不熟悉markdown编辑特点的用户来说，将现有的自然语言文稿转化成ppt软件可识别的markdown产品只能通过人工手写输入，耗时耗力。将NLP的有关技术与markdown转ppt技术相结合可弥补自然语言转markdown这一空缺。

03

产品创新与改进效果

- ✓ 知识问答准确、简洁且具有时效性
- ✓ 能够识别多层逻辑关系
- ✓ 能够对未知关系发现与推断
- ✓ 文档自动生成PPT、HTML等

3 创新与改进效果

知识问答准确、简洁且具有时效性

解决了问答系统的时效性问题。利用搜索引擎的问答能够进行实时的回应，比如“现在的杭州房价是多少”，“到2020年新中国成立了多少年”，将会返回搜索中第一条结果，通常只要网页上有的词条和新闻都可以以关键词的形式返回用户。

3 创新与改进效果

能够识别多层逻辑关系，避免繁杂的查询

基于知识图谱和NL2CQL的算法技术，可以将用户查询的自然语言直接转化为结构化查询语言，在关系型图数据库中进行查询并返回结果。企业可以将自己的图数据库放在数据层，问答层可以直接将输入的问题进行分类、分词、实体抽取，并自动在本地图数据库中进行查询，利用图相关的算法识别其中的逻辑关系，并根据相应节点的性质（如出入度、深度、中心度等）有逻辑性返回结果。

3 创新与改进效果

能够对未知关系发现与推断

一般来说该问答系统能够对于未知的关系进行发现。比如基于一般数据库的问答只能返回数据库中已有的字段，如果需要进行更深层的推理需要大量不同库之间的join操作，而基于知识图谱技术的问答能够跟随图中的边和节点不断发现逻辑关系，因此即达到了推理的效果，同时节省了储存海量知识的空间（只需要在图数据库中存贮词条和关系即可）。

3 创新与改进效果

基于图的自动文本生成

将知识图谱的技术推广到自动文本生成上，用户可以将一般的word文档上传到系统中，系统自动进行文档内容主题的提取，存储到图数据库并通过PPT等方式输出。方便相关从业人员，具有较大的应用价值及广泛的使用场景。

Thank you