基于层次化语义框架的知识库属性映射方法

李豫
华中师范大学/ 计算机学院
liyu@mails.ccnu.edu.cn

周光有
华中师范大学/ 计算机学院
gyzhou@mail.ccnu.edu.cn

摘要

面向知识库的自动问答是自然语言处理的一项重要任务，它旨在对用户提出的自然语言形式问题给出精炼、准确的回复。目前由于缺少数据集、特征不一致等因素，导致难以使用通用的数据和方法实现领域知识库问答。因此，本文将“问题意图”视作不同领域问答可能存在的共同特征，将“问题”与三元组知识库中“关系谓词”的映射过程作为问答核心工作。为了考虑多种层次的语义避免重要信息的损失，本文分别将“基于门控卷积的深层语义”和“基于交互注意力机制的浅层语义”两个方面通过门控感知机制相融合。我们在NLPCC-ICCPOL 2016 KBQA数据集上的实验表明，本文提出的方法与现有的基于CDSSM和BDSSM相比，效能有明显的提升。此外，本文通过构造天文常识知识库，将问题与关系谓词映射模型移植到特定领域，结合Bi-LSTM-CRF模型构建了天文常识自动问答系统。

关键词： 知识库 ； 属性映射 ； 深层语义

Property Mapping in Knowledge Base Under the Hierarchical Semantic Framework

Yu Li
School of Computer /
Central China Normal University
liyu@mails.ccnu.edu.cn

Guangyou Zhou
School of Computer /
Central China Normal University
gyzhou@mail.ccnu.edu.cn

Abstract

KBQA is an important task in natural language processing. It aims to give refined and accurate responses to natural language questions raised by users. At present, due to the lack of data sets and inconsistent features, it is difficult to use common data and methods to implement domain-specific. Therefore, this paper focuses on the core work of KBQA by identifying the intent of users' questions and attempting to map the questions and the predicates in the knowledge base. To avoid the loss of the important information, we combine the "deep semantics based gate convolution" and "shallow semantics based on interactive attention mechanism" into a unified framework using the gated perception mechanism. We conduct experiments on NLPCC-ICCPOL 2016 KBQA dataset. The results show that our proposed method significantly outperforms the existing CDSSM and BDSSM. Besides, we also construct a commonsense knowledge base under the domain of astronomy. Furthermore, we build a commonsense automatic question answering system by applying the proposed model and Bi-LSTM-CRF into the astronomy domain.

Keywords: Knowledge base , Property mapping , Deep semantics

第十九届中国计算语言学大会论文集，第246页-第255页，海南，中国，2020年10月30日至11月1日。

1 引言

问答任务（Question Answer，QA）是人工智能的核心研究之一。与传统搜索引擎相比，自动问答的便捷性和高效性增强了用户信息获取的体验，也使更多的学者开始对问答系统进行深入的研究。大规模知识库的迅速发展为实现自动问答目标提供了丰富有效的资源支撑，这使得面向知识库的自动问答（Knowledge Base Question Answer，KBQA）在工业界和学术界均受到了广泛的关注。知识库问答的目的就是根据用户提出的自然语言问题找到知识库中与之相关的知识，最后返回一个简洁、准确的答案。KBQA任务的核心工作是建立起问题到知识库的关系映射，而如何让机器理解自然语言问题与知识库三元组之间的语义等价关系是一个具有挑战性的难点。因此，本文探索的知识库属性映射方法可以作为KBQA系统中由问题关联到知识库的一种有效途径。

目前，随着如DBPedia (Auer et al., 2007)，Freebase (Bollacker et al., 2008)，Yago2 (Hoffart et al., 2011)，WikiData (Vrandečić and Krötzsch, 2014)等比较成熟的大型知识库相继涌现，自动问答的学术研究热度也在这些典型知识库基础上日益升温。知识库问答的实现有两大类主流方法，一种是语义解析（semantic parsingbased，SP-based），另一种是信息检索（information retrieve-based，IR-based） (Dong et al., 2015)。基于语义解析的方法是将问句分析成特有的表达形式或查询语句，如SPARQL、SQL语句等从知识库中搜索出答案；基于信息检索的方法是对候选答案通过特定方式进行排序得到最佳回答。早期针对小规模的知识库，多以语义解析方法为主，但这类方法往往会耗费大量精力去标注逻辑规则，也难以扩展到大规模知识库。Liang等 (2013)利用问答对语料，使用弱监督学习方法对问题进行过语义解析研究。Berant等 (2013)开发过一种语义解析器，可以训练无注释的逻辑形式，也可以扩展到大型知识库。Berant等 (2014)提出了一种基于释义的学习语义解析器的新方法以利用知识库中未涵盖的大量文本，但是其中一些工作仍依赖于手工标注和预定义规则，人工和时间成本较高。信息检索的方法则侧重于特征抽取以及对候选项的匹配和排序模型研究，其基本步骤是：“主题实体抽取”和“问题与关系谓词映射”。Yao等 (2014)使用句法分析技术，获得问句中的关键实体以及查询图。其他一些研究 (Zettlemoyer and Collins, 2012; Kwiatkowski et al., 2010)使用基于嵌入的模型来学习问题词和知识库构成的低维向量，并使用这些向量的总和来表示问题和候选答案，但是忽略了词序信息。Lai等 (2016)人使用主语谓语抽取算法，通过基于词向量的相似度结合分词技术实现属性映射，并利用人工定义的模板和规则取得了很好的效果。Wang等 (2016)使用分类器判断三元组中谓词与问题的映射。Yang等 (2016)使用了基于短语-实体字典的主题短语检测模型来检测问题与主题短语，之后使用排序模型对候选者进行排名。周博通等 (2018)在知识库问答属性映射问题上采用双向LSTM结合两种不同的注意力机制计算问谓相关度，在问题与谓词的映射测试上取得了91.77%的准确率。Xie等人 (2016)将CDSSM（Convolutional Deep Structured Semantic Models）与BDSSM（Bi-LSTM Deep Structured Semantic Models）相结合并利用余弦相似度计算问题与知识库关系谓词的匹配分数，但是其中采用的余弦距离是一个无参的匹配公式，并且仅使用深层神经网络可能会丢失一些重要的浅层词向量语义信息。赵小虎等 (zhao et al., 2020)通过将问题和知识库中三元组整体进行语义和字符的多特征匹配，并使用有参的全连接层计算相似分数，但是尚未考虑到浅层词向量的直接影响。

综合以上工作来看，问题与知识库的属性映射在KBQA任务中十分重要，同时也存在进一步改进的空间。本文着重关注问题与知识库谓词之间的映射方法，从表征和匹配两个角度改进前人所提到的CDSSM“问谓属性映射模型” (Xie et al., 2016)。在表征层中，首先针对问题的表述通过增设卷积门来过滤问题中与谓词无关的词级噪声，再使用两种共享的语义获取模型得到待匹配项的深层语义与浅层语义，最后利用门控机制平衡两种不同层次的语义得到层次化待匹配向量。在匹配层中，本文获取问题与关系谓词之间的多种联系，再由多层感知机融合，经池化操作获取最终的语义匹配得分。本文在NLPCC-ICCPOL 2016发布的中文问答数据集上进行属性映射实验，实验结果表明了该方法的有效性。另外，由于问题与谓词的映射是一种较为通用的问题意图识别过程，例如时间、地点、概念、因果、人物等通用询问意图在其他领域问答中也多有涉及，因此适合迁移到其他领域的问答。依照这种思路，本文构建了中文天文常识知识库，将天文命名实体识别作为基础任务，将面向知识库的“问谓属性映射”作为重点研究内容，构建了天文常识自动问答系统。综上，本文的贡献如下：

（1）针对问题表达，通过增设卷积门，适当过滤问题中与谓词无关的词级噪声。

（2）采用交互注意力机制获取浅层词向量全局语义。通过门控感知机制在表征层面有效地融合了层次化语义信息，既考虑了深层语义又防止浅层语义信息的丢失。

（3）最后，本文通过构建天文常识知识库以及将上述“问谓属性映射”方法迁移到特定领域知识问答中，与Bi-LSTM-CRF模型相结合构建天文常识自动问答系统。

2 面向知识库的属性映射

2.1 数据来源

由于中文知识库和相关问答语料较为欠缺，所以在中文知识库问答方面一直鲜有研究。在2016年和2017年NLPCC-ICCPOL发布了中文知识库以及问答对语料后，许多学者都开始围绕此项语料数据展开研究工作。为了扩展问答意图的范围，本文也选用了NLPCC-ICCPOL 2016年评测中公开的基于知识库的问答数据。这类数据来源于百科信息栏三元组，将其运用到领域知识库的问答中会具有较为全面的覆盖度。数据集的原始数据格式是：<问题，三元组，答案>三元组，例如：“机械设计基础这本书的作者是谁？”；“机械设计基础|||作者|||杨可桢，程光蕴，李仲生”；“杨可桢，程光蕴，李仲生”。在实际深度学习之前，需要针对所要用到的方法对原数据集进行修正和加工。

2.2 数据处理

数据重新处理的过程中，保留问题以及对应三元组中的关系谓词，如上例中的问题和谓词“作者”，对数据集中的全体谓词构造谓词词典，在词典中随机抽取9个谓词负例与正确谓词合并作为谓词候选集。随机初始化标签顺序，生成对应的谓词标签，其中正确谓词对应的标签为1，错误谓词对应的标签是0。在数据本身的标准性上，由于数据存在大量谓词中间空格现象，如“作者”这一谓词的原数据格式为“作 者”，需要去除空格保持数据一致性。另外，在数据预处理的过程中，本文严格控制谓词候选集中不存在重复项，从而防止训练和测试产生误差。同时，通过人工核查尽可能避免候选谓词中存在与正确谓词是同义词的情况，如“俗名”和“俗称”，“位置”和“地域所属”等，以免在准确率上出现偏差。为了探究主题实体在不同方法中产生的影响，我们将数据集分为掩盖主实体和不掩盖主实体两类。未掩盖主实体的数据集中，问句不做任何处理。在掩盖主实体的数据集中，首先根据每个问句对应的知识库三元组找到具体的主题实体，再将这些实体在问句中用特殊符号掩盖。例如问句“波色-爱因斯坦凝聚态有哪些比喻？”，经过掩盖后为：“E有哪些比喻？”；问句“大熊座47c多长时间一个周期？”掩盖后为：“E多长时间一个周期？”。转换完毕的数据集新格式为：<问题，谓词候选集，标签>。

2.3 属性映射框架

首先介绍本文提出的问题与关系谓词的属性映射网络架构GHSMM（Gate Hierarchical Semantic Match Model），结构图如图1所示。

第十九届中国计算语言学大会论文集，第246页-第255页，海南，中国，2020年10月30日至11月1日。

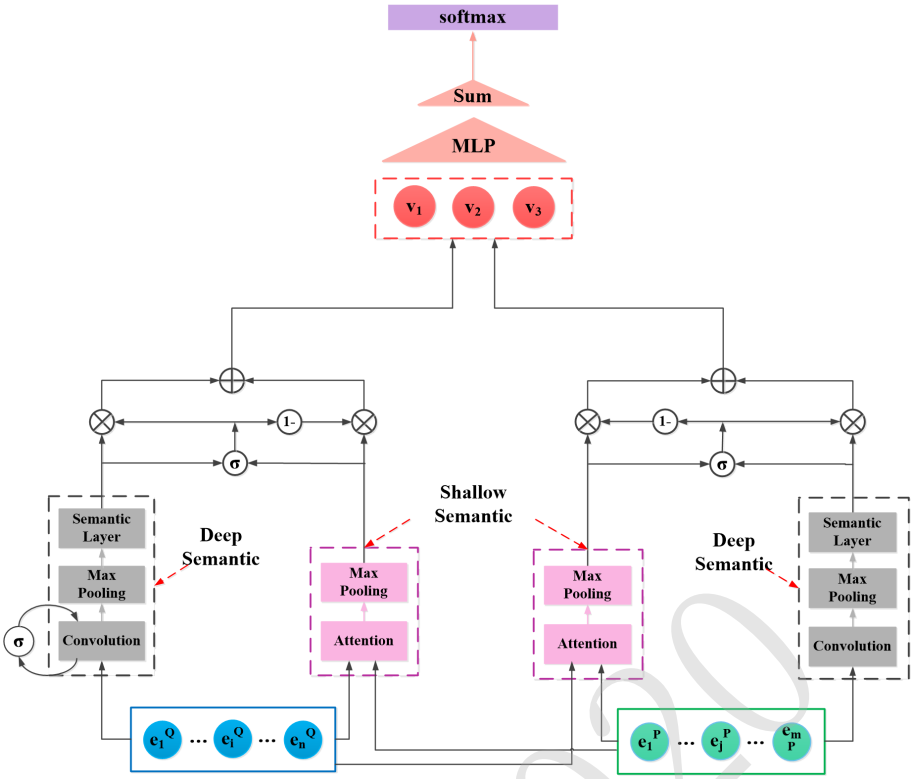

图 1: 融合门控感知的层次化语义匹配框架GHSMM

问句与候选谓词集合在初始状态下均以Word2vec词向量矩阵表示。在使用模型时，首先将问题与关系谓词的向量矩阵同时输送至“融合卷积门的深层语义模块”与“基于交互注意力机制的浅层语义模块”中得到两种层次语义向量，之后由“基于门控感知的层次化语义融合模块”将深层和浅层语义有效融合，经过“匹配层”和“决策层”得到候选谓词集与问题的匹配概率分布。

相较于传统的CDSSM模型，我们在获取深层语义的同时还考虑了上下文的浅层语义，同时，我们也改进了匹配层中简单的余弦夹角计算，利用多层感知机以及池化操作对问题和谓词的交互信息进行打分，在匹配效果上得到了一定提升。

2.3.1 融合卷积门的深度语义模型

问题语句较长且存在大量与谓词无关的噪声将会影响匹配，因此本文在CDSSM基础上增设卷积门，对问句进行门控过滤。

卷积神经网络（Convolutional Neural Network，简称CNN）可以有效地提取出矩阵的局部特征并在此之上进行全局的预测。给定句子序列的词向量$E = \{e_1, e_2, \cdots, e_n\}$，其中表示第$i$个词的词向量表示。通过设置卷积核的大小，使用卷积操作矩阵w对向量矩阵E进行卷积操作，得到结果$c = \{c_1, c_2, \cdots, c_{|E|-w+1}\}$。其中$\otimes$代表卷积运算，$f_x$为非线性激活函数，$b_c$是偏置项。

$$c_i = f_x(w \otimes E_{(i-w+1):i} + b_c) \tag{1}$$

为了进一步过滤问句中与谓词无关的词级噪声，本文采用卷积门控制问句的卷积输出，而对谓词则不使用门控机制。谓词的卷积结果c_p经过Relu激活函数直接输出。问句的卷积结果c_q由问句向量E_q通过两个卷积网络得到，其中一个是原始的CNN，另一个在sgmoid函数激活下生成门控向量。

$$c_p = \text{ReLu}(w \otimes E_p + b_1) \tag{2}$$

$$c_q = \text{ReLu}(w \otimes E_q + b_1) \odot sigmoid(v \otimes E_q + b_2) \tag{3}$$

我们采用最大池化操作提取特征图中的重要信息，将通过不同大小卷积核得到的卷积输出分别进行池化和拼接，最后再通过一个全连接层与tanh非线性函数将问句或谓词投影到一定维度大小的语义空间。

$$h_d = \tanh([\max(c^{(1)}); \max(c^{(2)}); ...; \max(c^{(win)})] \cdot W_d + b_d) \tag{4}$$

2.3.2 基于交互注意力的浅层语义表示

浅层语义可以反映全局的文本信息，我们使用交互注意力机制为词向量添加注意力信息，如图2所示。经过交互后，问句的每一个序列都对应一个待匹配谓词的全局语义向量；同样，谓词的每一个序列也对应着一个问句的全局语义向量。经过最大池化操作获取浅层语义信息。

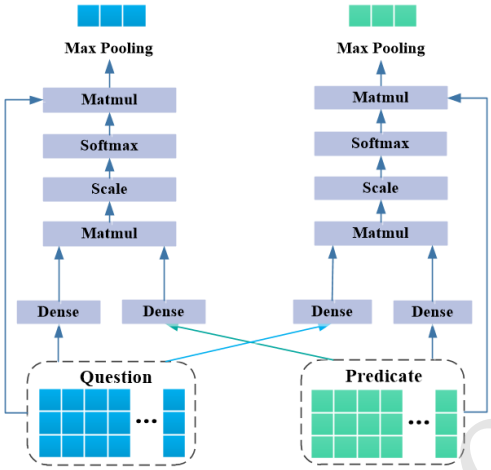

图 2: 基于交互注意力机制的浅层语义表示

注意力机制 (Vaswani et al., 2017)目前已广泛应用于自然语言处理领域，用来增大重要信息的权重系数，使模型关注重要的部分。公式(1)是注意力机制的计算过程。Q，K，V分别是输入向量与三个权重矩阵W^Q,W^K,W^V相乘的结果，各自代表了查询（query）、键（key）、真实值（value），其中Q与K的输出维度相同。Q与K的交叉乘积除以d_k的平方根是为了防止内积过大而影响梯度，经过softmax函数归一化后得到注意力权重。最后将V在Q的每一个位置上的向量进行一次加权求和，得到$Attention(K, Q, V)$，表达了对各个词的注意力程度。

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{5}$$

在使用注意力机制时，问句的目标语句是待匹配的谓词，谓词的目标语句是待匹配的问句，value值均为句子本身的表示（即Word2vec词向量）。E_q 与E_p分别表示问句q与谓词p的Word2vec词向量矩阵，最后得到问句和谓词的交互注意力语义向量S_q、S_p。

对于问句q，注意力运算如下：

$$S_q = Attention(E_p \cdot W^Q, E_q \cdot W^K, E_q) \tag{6}$$

对于谓词p，注意力运算如下：

$$S_p = Attention(E_q \cdot W^Q, E_p \cdot W^K, E_p) \tag{7}$$

为了获取全局语义信息，我们采用最大池化的操作对上述得到的注意力信息进行池化过滤，得到添加了注意力机制的词向量全局语义信息h_s。

$$h_s = \max(S) \tag{8}$$

2.3.3 基于门控感知的层次化语义融合

交互注意力机制可以融合问句与谓词之间的长期依赖信息，卷积神经网络可以较好地提取词与词之间的短距离特征，我们增设门控感知机制，将深层语义特征h_d与浅层词级特征h_s进行融合。

g表示它们生成的门控向量。⊙表示逐点乘积操作。

$$g = sigmoid(h_d \cdot W_g^{(1)} + h_s \cdot W_g^{(2)} + b_g) \tag{9}$$

门控机制将层次化语义信息进行平衡，其中g中每一个元素代表将多少浅层语义特征替换为深层语义特征，从而得到最终的层次化语义向量y。

$$y = g \odot h_d + (1 - g) \odot h_s \tag{10}$$

2.3.4 交互匹配层

CDSSM中采用了余弦距离计算问句与谓词之间的相似性，但余弦计算是一种简单无参数匹配方法。为了更充分地度量问谓相似程度，我们一方面将问谓间的多种交互特征进行融合，另一方面通过多层感知机获取问题与谓词之间多个方面上的相似得分，之后采用加和池化得到整体相似度。

这里我们选用三种方式对问题与谓词进行语义交互。v_1代表全局语义向量之和，v_2表示向量间的绝对差，v_3是向量之间的逐个点积运算。

$$v_1 = y_q + y_p \tag{11}$$

$$v_2 = \mid y_q - y_p \mid \tag{12}$$

$$v_3 = y_q \odot y_p \tag{13}$$

将v_1，v_2，v_3三者拼接形成一个扁平的向量，并输入到一个两层的全连接层中，将输出结果投射到m维作为评价相似度的m个方面，即$s = \{s_1, s_2, \cdots, s_m\}$，$w_s$、$w_h$是多层感知机中学习的权重，$b_s$、$b_h$是学习的偏置项。

$$s = w_s \cdot \tanh(w_h \cdot [v_1; v_2; v_3] + b_h) + b_s \tag{14}$$

在进行计算匹配得分前，需要将上一步得到的结果进行池化操作，这里我们选择加和池化计算匹配分数，并用sigmoid激活函数把匹配值压缩到0到1之间。将问题q与谓词p_i（其中$i = 1, 2, 3 \cdots k$，k为谓词集合的元素数目）的语义向量按照本文提出的方式进行一对一匹配，每一个问句对应得到k个语义相关性$SMS(p_i|q)$。

$$SMS(p_i \mid q) = sigmoid(\sum_{i=1}^{m} s_i) \tag{15}$$

2.3.5 决策层

将语义匹配得分送入softmax分类器中来预测最终的正确匹配项，并计算添加了正则项后的交叉熵目标函数。

$p(p_i|q)$是问题q与第i个谓词p_i相匹配的概率。其中，P是问题q的一组候选谓词，包括几个否定谓词样本和一个肯定谓词样本。p'代表候选谓词集P中的任意谓词元素。

$$p(p_i \mid q) = \frac{\exp(SMS(p_i \mid q))}{\sum_{p' \in P} \exp(SMS(p' \mid q))} \tag{16}$$

训练语义模型以最大化肯定谓词的可能性为训练目标，L是目标损失函数。其中，q_r代表R个问题中的第r个问题，p^+代表正确的谓词，$p(p^+|q_r)$是第r个问题中给定正谓词的条件概率，λ是L_2正则化参数，θ是模型的参数。最后使用优化算法对目标函数进行优化，此过程采用误差反向传播的方式更新各层权重和偏置值。

$$L(\theta) = -\log \prod_{r}^{R} p(p^+ \mid q_r) + \lambda \parallel \theta \parallel^2 \tag{17}$$

3 实验

3.1 实验设置与评测指标

实验采用NLPCC-ICCPOL公开数据集，其中训练集包含14609个问句，测试集包含9870个问句。本文从训练集中选取3000句作为开发集，剩下的11609个问句作为实际的训练集。实验环境：本实验的环境为tensorflow框架，编程语言为Python3.5，重要超参数设置情况：选取300维的Word2Vec词向量模型，batch size大小为50，学习率为0.005，卷积核窗口大小为1，2，3，卷积核数目为100，卷积步幅为1。为防止梯度爆炸使用了梯度裁剪。实验选择Momentum优化器。评测指标采用通用的准确率。

3.2 属性映射实验

	模型	测试集Acc(%) 无主实体	有主实体
对比模型	BiLSTM_AC1 (周博通等人, 2018)	86.74	-
	BiLSTM_AC12 (周博通等人, 2018)	87.64	-
	BiLSTM_AC12_Overlap (周博通等人, 2018)	91.77	-
	BDSSM	91.81	91.42
	CDSSM	92.15	91.49
本文模型	Attention+MLP（浅层语义匹配）	93.49	93.57
	CNN+MLP（无卷积门深层语义匹配）	93.68	94.01
	GCNN+MLP（融合卷积门深层语义匹配）	93.78	94.07
	HSMM（未引入平衡门的层次化语义匹配）	93.51	94.22
	GHSMM (综合模型：融合门控机制的层次化语义匹配)	**93.99**	**94.68**

表 1: 各种不同方法的属性映射实验结果

本文选取前人提出的神经网络与简单文本特征的结合模型、BDSSM模型以及CDSSM模型作为对比模型，进行属性映射实验。通过表1可以看出，本文的浅层语义匹配方法和深层语义匹配方法与对比模型相比效果均有提高，而将浅层语义和深层语义融合后得到GHSMM综合模型，比单独使用这两种语义时取得了更优的效果，证明了本文层次化语义匹配方法的有效性。

另外，实验显示，保留问句中的主题实体提高了所有本文改进模型的准确率，但没有对CDSSM、BDSSM起到提升作用。说明加入主实体更有利于本文模型的匹配。

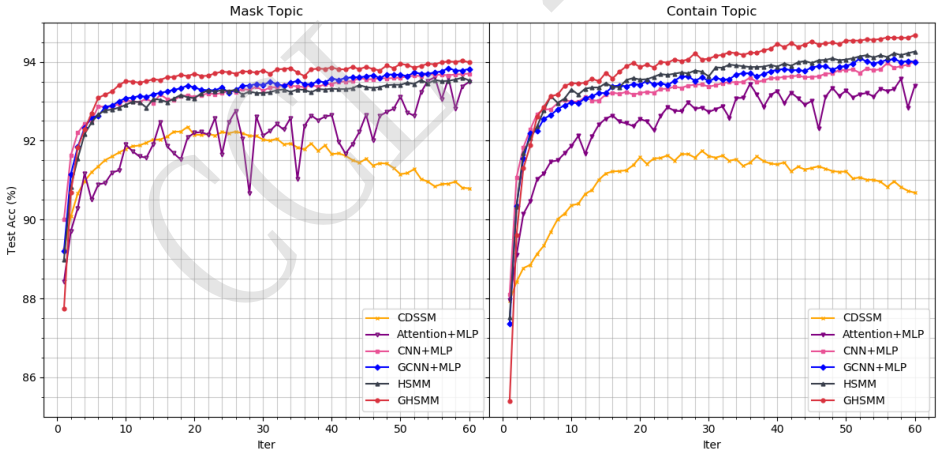

图 3: 本文各项改进方法在不同迭代次数下的效能比较

为了更好地说明本文在原CDSSM模型基础上所做的各项改进模块的有效性，我们对模型做了消融实验，即比较了不同迭代次数下各项改进模块的准确率。图3分别展示了基线模型（CDSSM）、基于交互注意力机制的浅层语义匹配（Attention+MLP）、无卷积门的深度语义匹配（CNN+MLP）、有卷积门的深度语义匹配（GCNN+MLP）、通过简单加和实现的层次化语义匹配（HSMM）以及组合起来的融合语义平衡门的层次化语义匹配（GHSMM）在一次训练中各迭代轮数下的性能。实验分别在包含主题实体的数据集（图3右部分）与掩盖主题实体的数据集（图3左部分）上进行。从图中可以看出，在实体被掩盖时有卷积门和无卷积门两者在每次迭代中的实际效能差距很小，说明此时卷积门对问句的过滤作用还比较微弱，而在保留主题实体时有卷积门的过滤效果相对更明显一些；单纯使用注意力机制得到的浅层语义匹配在

整体训练效果上具有较大的起伏，总体上弱于其他改进模型；HSMM将深层语义与浅层语义通过简单加和的方式相结合，效果要弱于GHSMM，这进一步证明了加入语义平衡门能使两种语义自适应地达到更好的结合效果。整体上看，层次化语义模型训练趋势平稳，且在两组实验数据集中呈现的总体效果为最优，这表明本文将浅层语义与深层语义成功地融合在一起，减少了语义信息的损失。

3.3 天文问答系统示范性应用

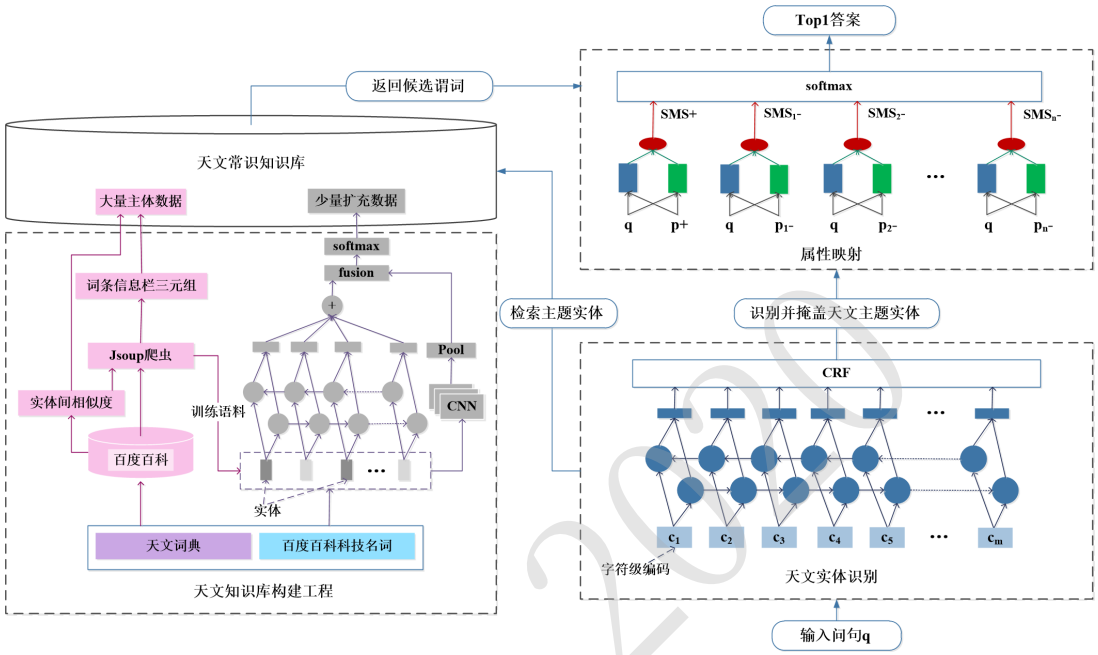

图 4: 天文问答系统构建

本文将提出的基于层次化语义框架的知识库映射方法应用到实际的领域知识库问答中，实现了天文问答系统的示范性应用。图4是整个天文问答系统的构建，包含“天文知识库构建”和“问答实现”两个大模块。

3.3.1 天文常识知识库的构建

目前，已经开放的天文知识库和相关成型的语料十分罕见，而百科知识语料属于互联网上开放的知识文本数据，具有规模庞大、持续更新扩展的特点，如著名的DBPedia知识库就是从维基百科中抽取构建的。因此，本文选择从百度百科网络资源中获取天文知识语料并进一步展开结构化信息抽取工作。百度百科具有较强的专业性和全面性，而且在词条页面中附带结构化的数据信息，这使得后期的信息抽取工作变得方便有效。本文首先下载了天文词典以及爬取百度百科“科技名词栏”中的天文术语，构建出了一份天文实体集，该实体集一共包含24376个天文学名词。本文依照实体集爬取了对应的百度百科信息框，将这些信息以三元组的格式存储，例如：{“木卫四”，“外文名”，“Dione”}。对于某些特定关系的尾实体（如“类别”、“别称”、“成分”等）要进行字符串的切割操作，将含多个并列成分的句子分开，例如：将{“日冕”，“结构”，“内冕、中冕和外冕3层”}这一个三元组拆分成：{“日冕”，“结构”，“内冕”}、{“日冕”，“结构”，“中冕”}、{“日冕”，“结构”，“外冕”}三个三元组。由于基于上述爬虫方法获取到的大多数是“实体-属性-属性值”类别的三元组，缺少“实体-关系-实体”这类三元组，因此为了进一步扩展知识库，本文事先回召了已有三元组中特定关系的关系句子，并作为语料通过基于注意力机制的Bi-LSTM+CNN网络训练关系抽取模型，再使用基于链入词条的TF-IDF的相似度计算获取关联度高的实体对，选择包含这些实体对的关系句子送入模型进行关系标签预测，经过人工评估后作为少量扩充数据入库。最终构建的知识库包含53975个三元组，以及10258个关系度较高的实体对和它们的关系程度数值。

3.3.2 天文问答的实现

上述方法构建的天文常识知识库与NLPCC-ICCPOL发布的知识库存在类似的问题，例如许多三元组宾语是以字符串形式表示而非知识库中的实体，因此难以形成像知识图谱这种网络拓扑结构；其次，对于同一个主实体可能存在多个几乎同义的关系谓词。很多在Freebase等外文知识库的研究并不适合直接应用在此类中文的数据集中，也无法处理需要多个三元组进行回答的复杂问题，而比较适合用来用来回答单实体单关系类型的问题。综合这些因素，本文将天文问答分为Bi-LSTM-CRF天文命名实体识别步骤与GHSMM属性映射步骤。考虑到大多数天文主实体较为生僻，因此使用主实体掩盖方法，使问句更接近自然语法结构。首先将问句经过实体识别层找出问句中对应的天文主实体，再将实体替换为特殊符号与检索出的相关候选谓词一并输送到属性映射层进行匹配。由于难以获取符合条件的数据集和标签，我们训练Bi-LSTM-CRF模型的语料是由爬取的天文百度百科文本经过分句、分字以及采用字符串匹配算法对照天文实体集添加字标签获取的，标签采用的序列格式为BIOES。在属性映射层中，本文模型可以处理不同数目的关系谓词，系统将匹配概率最高的谓词作为该问题的核心意图，联合主题实体和关系谓词，返回对应的尾实体作为最终答案。

3.4 样例分析

为了显示证明模型理解问题语义与选取关系谓词的有效性，本文抽取了一个天文问句例子，从已构建的天文知识库中检索相关实体和全体关系谓词，经过上述问答系统的识别后展示候选谓词分数经过归一化后得到的概率分布。

贝利珠名字的由来是什么?					
	外文名	命名原因	出现时间	人物	学科
CDSSM	**0.25513**	0.25404	0.16021	0.18954	0.14108
GHSMM	0.19977	**0.20090**	0.19977	0.19979	0.19977

表 2: 样例分析

如表2所示，该样例包含一个有关主题实体“贝利珠”的问题，并且询问的意图与“名字”有关。针对“贝利珠”主题实体，从知识库中检索出来的候选谓词有表中所示的5个。输入问句“贝利珠名字的由来是什么？”，在CDSSM中，带有“名字”意义的两个关系谓词“外文名”与“命名原因”在谓词候选集中的匹配概率都很高，模型最终错误地选择了与“名字”语义接近的“外文名”。同时，我们发现CDSSM中其他候选谓词也随着本身语义的相关度差异显示不同的概率，例如“人物”这种类型的词从某种程度上也经常和“名字”同时出现，语义相关度较高，故概率值也偏高一些。

而经过本文提出的GHSMM处理后，正确谓词匹配概率相对较高，而其他错误候选项的概率值接近一样，降低了其他词语在语义远近上的影响，例如在上述问句中与名字有关的“外文名”与不相关谓词“出现时间”、“学科”等概率值几乎相同。

上述分析再次表明本文提出的GHSMM比CDSSM具有更好的性能，在一定程度上避免了词语之间语义相近导致的错误匹配情况。

4 结语

基于知识库的自动问答大多数做法是通过主题实体识别和问句与关系谓语的属性映射，由于领域知识库缺少问答数据集，本文以NLPCC-ICCPOL数据进行问谓属性映射训练并迁移到天文领域知识库中实现自动问答。在属性映射研究过程中，本文将前人的CDSSM模型进行了改进，在特征抽取步骤中利用门控感知机制融合层次化语义信息，使最后的向量表示同时具备浅层的和深层的语义；在匹配层中我们将CDSSM中原有的无参余弦匹配改进为融合多种交互信息的多层感知机，通过池化等操作得到最终语义匹配分数，实验证明最终改进的属性映射方法比CDSSM与BDSSM等模型的匹配效果有明显提高。但是本文所实现的领域知识库自动问答系统仍然存在许多限制和不足，在下一步工作中，笔者希望将模型进一步改进，并关注问句中主题实体的提取环节，以实现更高性能的特定领域自动问答应用。

致谢

本文的工作作为毕业论文的一部分，受到国家自然科学基金（No. 61972173）支持。感谢匿名评审专家对我们工作提出的建设性修改意见。

参考文献

Auer S, Bizer C, Kobilarov G, et al. *Dbpedia: A nucleus for a web of open data*[M]. The semantic web. Springer, Berlin, Heidelberg, 2007: 722 735.

Bollacker K, Evans C, Paritosh P, et al. *Freebase: a collaboratively created graph database for structuring human knowledge*[C]. Proceedings of the 2008 ACM SIGMOD international conference on Management of data. ACM, 2008: 1247 1250.

Berant J, Chou A, Frostig R, et al. *Semantic parsing on freebase from question answer pairs*[C]. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 1533-1544.

Berant J, Liang P. *Semantic parsing via paraphrasing*[C]. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014, 1: 1415 1425.

Dong L, Wei F, Zhou M, et al. *Question answering over freebase with multi-column convolutional neural networks.* // The Association for Computational Linguistics. Beijing, 2015: 260–269.

Hoffart J, Suchanek F M, Berberich K , et al. *YAGO2: exploring and querying world knowledge in time, space, context, and many languages*[C]. Proceedings of the 20th international conference companion on World wide web. ACM, 2011: 229 232.

Kwiatkowski T, Zettlemoyer L, Gold water S, et al. *Inducing probabilistic CCG grammars from logical form with higher order unification*[C]. Proceedings of the 2010 conference on empirical methods in natural language processing. Association for Computational Linguistics, 2010: 1223 1233.

Lian g P, Jordan M I, Klein D. *Learning dependency based compositional semantics*[J]. Computational Linguistics, 2013, 39(2): 389 446.

Lai Y, Lin Y, Chen J, et al. *Open domain question answering system based on knowledge base*[M]. Natural Language Understanding and Intelligent Applications. Springer, Cham, 2016: 722 733.

Vrandecic D, Krotzsch M. Wikidata. *a free collaborative knowledgebase*[J]. Communications of the ACM, 2014, 57(10): 78 85.

Vaswani A, Shazeer N, Parmar N, et al. *Attention is all you need*[C]. NIPS 2017: Advances in Neural Information Processing Systems, 2017: 5998 6008.

Wang L, Zhang Y, Liu T. *A deep learning approach for question answering over knowledge base*[M]. Natural Language Understanding and Intelligent Applications. Springer, Cham, 2016: 8 85 892.

Xie Z, Zeng Z, Zhou G, et al. *Topic enhanced deep structured semantic models for knowledge base question answering*[J]. SCIENCE CHINA: Information Sciences,2017,60(11):28-42.

Yao X, Van Durme B. *Information extraction over structured data: Question answering with freebase*[C]. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014, 1: 956 966.

Zettlemoyer L S, Collins M. *Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars*[J]. arXiv preprint arXiv:1207.1420, 2012.

Yang F, Gan L, Li A, et al. *Combining deep learning with information retrieval for question answering*[M]. Natural Language Understanding and Intelligent Applications. Springer, Cham, 2016: 917 925.

周博通，孙承杰，林磊等. 基于*LSTM* 的大规模知识库自动问答[J]. 北京大学学报（自然科学版）, 2018, 54(2): 286 292.

赵小虎，赵成龙. 基于多特征语义匹配的知识库问答系统[J/OL]. 计算机应用:1-6[2020-06-09].http://kns.cnki.net/kcms/detail/51.1307.TP.20200519.1403.004.html.