

# 基于 LSTM 的大规模知识库自动问答

周博通 孙承杰<sup>†</sup> 林磊 刘秉权

哈尔滨工业大学计算机科学与技术学院, 哈尔滨 150001; <sup>†</sup>通信作者, E-mail: cjsun@insun.hit.edu.cn

**摘要** 针对大规模知识库问答的特点, 构建一个包含 3 个主要步骤的问答系统: 问句中的命名实体识别、问句与属性的映射和答案选择。采用别名词典结合 LSTM 语言模型进行命名实体识别, 使用双向 LSTM 模型结合两种不同的注意力机制进行属性映射, 最后综合前两步的结果进行实体消歧和答案选择。该系统在 NLPCC-ICCPOL 2016 KBQA 任务提供的数据集上的平均 F1 值为 0.8106, 接近评测的最好水平。

**关键词** 知识库; 自动问答; 命名实体识别; 注意力机制

**中图分类号** TP914

## LSTM Based Question Answering for Large Scale Knowledge Base

ZHOU Botong, SUN Chengjie<sup>†</sup>, LIN Lei, LIU Bingquan

School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001;

<sup>†</sup> Corresponding author, E-mail: cjsun@insun.hit.edu.cn

**Abstract** To solve the specific problem in KBQA, a question answering system is built based on large scale Chinese knowledge base. This system consists of three main steps: recognition of named entity in question, mapping from question to property in KB, and answering selection. Alias dictionary and LSTM language model are used to recognize named entity contained in question, and two different attention mechanisms are combined with bidirectional LSTM for question-property mapping. Finally, exploit results of first two steps are exploited for entity disambiguation and answering selection. The average F1 value of proposed system in NLPCC-ICCPOL 2016 KBQA task is 0.8106, which is competitive with the best result.

**Key words** knowledge base; question answering; named entity recognition; attention mechanism

基于知识库的自动问答系统的核心在于对问句的语义理解。人们输入的问题句是自然语言形式, 而知识库中的信息却是结构化存储的, 同时, 问句的表述与知识库中存储的信息也存在一定的差异。例如用户输入问句“请问刘德华的妻子是谁?”, 且在知识库中存在三元组(“刘德华(香港演员、歌手)”, “配偶”, “朱丽倩”)时, 要想得到问题的正确答案, 首先需要找到问题中涉及的实体, 即“刘德华(香港演员、歌手)”, 然后找到问题询问的是关于这个实体的哪个属性, 如“配偶”, 最后才能找到正确的答案“朱丽倩”。然而, 这些联系仅仅通过表层的字面信息是很难得到的, 需要对问句进行足够深入的语

义分析。

目前主流的基于知识库的自动问答中采用的方法可以分为两类: 基于语义分析(semantic parsing-based, SP-based)的方法和基于信息检索(information retrieve-based, IR-based)的方法<sup>[1]</sup>。基于语义分析的方法首先将自然语言形式的问题句转换为某种特定类型的逻辑表达形式, 如带类型的 lambda 表达式(typed lambda calculus)<sup>[2]</sup>、简单 lambda 依存组合语义(simple  $\lambda$ -DCS)<sup>[3]</sup>等。这类逻辑表达形式通常也适用于在知识库中进行查询, 进而找出问题的答案。基于信息检索的方法首先通过比较粗糙的方法, 从知识库中获取一系列的候选答案, 然后对问

句和候选答案进行特征抽取, 利用它们对候选答案进行排序, 选择得分最高的作为最终答案。

本文将基于知识库的自动问答分为两个主要步骤和一个次要步骤: 命名实体识别步骤和属性映射步骤, 这两步是本文的重点; 最后一步为答案选择, 由于采取的方法较为粗略, 因此只在 2.4 节做简要介绍。在命名实体识别步骤中, 本文以长短时记忆网络(long short time memory, LSTM)作为语言模型来寻找问句中包含的命名实体; 在属性映射步骤中, 本文采用双向 LSTM 结合两种不同的注意力机制来计算属性与问句的语义相关程度。

## 1 国内外研究现状

基于知识库的自动问答系统在人工智能领域具有很长的发展历史。早期的研究主要针对小规模的知识库, 使用的方法以语义解析为主。这类方法往往需要人工标注的“自然语言语句-逻辑表达式”对, 需要花费大量精力, 也很难扩展到大规模知识库的情况。后来研究人员利用其他形式的语料(如问答对), 使用弱监督学习方法进行语义解析<sup>[4]</sup>。如前所述, 近期的研究方法主要分为基于语义分析的方法和基于信息检索的方法两大类。基于语义分析的方法首先将自然语言形式的问句转换为某种逻辑表达式, 常见的有  $\lambda$  表达式和依存组合语义树等。

基于信息检索的方法侧重于抽取有效特征, 对候选进行排序。Yao 等<sup>[5]</sup>使用依存分析技术, 获得问题的依存分析树, 进而找到问句中涉及的主要实体, 从知识库中找到该实体的 Topic Graph, 之后从问题的依存树和 Topic Graph 中抽取多种特征, 将其送入逻辑回归模型中进行分类。随着神经网络技术的不断进步, 有些研究人员将神经网络应用于自动问答领域。Bordes 等<sup>[6-7]</sup>使用简单的前馈神经网络, 对候选答案的各方面信息进行语义编码, 将问句和候选答案分别转换为相同维度的向量, 最后以两个向量的点积作为候选的得分。Dong 等<sup>[1]</sup>使用 MCCNN (multi-column CNN) 网络, 分别对问句中隐含的答案类型、关系和上下文信息进行语义编码, 对候选答案的这 3 类信息同样进行语义编码, 最后对这 3 类信息的语义向量的点积进行加权, 得到最后的得分。Zhang 等<sup>[8]</sup>在 Dong 等<sup>[1]</sup>的基础上, 使用注意力机制, 对候选答案的不同内容, 训练不同的问句表示。

在中文领域, 由于缺少大规模的通用知识库, 因此鲜有这方面的研究。2016 年 NLPCC-ICCPOL KBQA 任务提供了一个规模可观的知识库, 相关人员在此基础上进行了一定程度的研究。Wang 等<sup>[9]</sup>使用卷积神经网络和 GRU 模型, 进行问句的语义表示。Xie 等<sup>[10]</sup>使用 CNN 网络进行命名实体识别, 并使用 BiLSTM 和 CNN 进行属性映射。Yang 等<sup>[11]</sup>通过提取多种特征, 使用 GBDT 模型进行问句中的命名实体识别, 使用 NBSVM 和 CNN 进行属性映射。Lai 等<sup>[12]</sup>通过简单的词向量相似度运算, 结合细粒度的分词进行属性映射, 同时结合多种人工构造的规则和特征, 在该任务上取得最好的效果。

## 2 基于 LSTM 模型和注意力机制的问答系统

本文将基于知识库的自动问答拆分为两个主要步骤: 命名实体识别步骤和属性映射步骤。实体识别步骤的目的是找到问句中询问的实体名称, 属性映射步骤的目的在于找到问句中询问的是关于实体的什么内容。该系统的大体框架如图 1 所示。

图 1 中, 对于输入的问句“石头记是谁写的?”, 首先通过实体识别步骤找出问句中提及实体的部分“石头记”, 然后找到知识库中正式名或别名为“石头记”的所有实体, 如本例中的“红楼梦(小说)”和“红楼梦(电视剧)”; 属性映射将上一步得到的实体的所有相关属性作为候选, 在本例中即为“作者”、“创作年代”和“导演”, 然后计算每个候选属性与问句剩余部分, 即“\_\_是谁写的?”的语义相关度, 选择得分最高的属性“作者”; 最后, 通过查询知识库, 得到三元组(“红楼梦(小说)”, “作者”, “曹雪芹”), 最

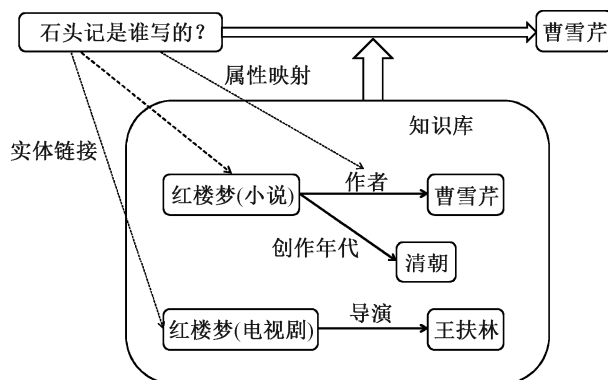


图 1 整体结构框架

Fig. 1 The overview of the KBQA system framework

终得到问题的答案“曹雪芹”。

## 2.1 知识库简介及预处理

本研究采用的知识库由 NLPCC-ICCPOL 2016 KBQA 任务提供。该知识库是中文领域第一个大规模的通用知识库, 包含 6502738 个实体和 43063796 个三元组。该知识库以每行一个三元组的形式存储, 第一列为实体, 第二列为属性, 第三列为属性值。因此, 该知识库本质上是多个三元组(实体, 属性, 属性值)构成的集合。该知识库中的内容如表 1 所示。

该知识库中的数据从百度百科的 Infobox 自动抽取, 因此存在不少的噪声, 特别是属性部分, 经常出现无用的字符。因此, 实验之前首先对该知识库进行一定程度的去噪处理。表 2 列举去噪的规则。

同时, 该知识库中存在同一实体的三元组分布在知识库文件中不同位置的情况。为了提高知识库

表 2 知识库去噪规则

Table 2 Rules for removing noise in knowledge base

规则	去噪前	去噪后
去除属性中的空白字符	书 名	书名
去除属性中所有非中文、数字和英文字幕的字符	- 行政村	行政村
将实体和属性中的所有大写外文字符转为小写	F-104 战斗机	f-104 战斗机

查询的效率, 本文将关于同一实体的所有三元组集中起来, 并为知识库建立索引文件。索引文件每行的格式为实体名、起始位置、内容长度, 表示每个实体的所有内容在知识库文件中的起始位置和内容的总长度(均以字节为单位)。在知识库中查找某一实体时, 首先读取索引文件, 获得该实体的起始位置和内容长度, 就可以直接从知识库文件中找到该实体的所有信息, 而不用从头至尾遍历知识库, 极大地提高了知识库的查询效率。

## 2.2 命名实体识别

实体识别步骤的目的在于找出问题中询问的命名实体。该部分的大体框架如图 2 所示。

首先根据别名词典, 查找问句中存在的命名实体作为候选。图 2 的候选集合中, 每个候选的浅色部分表示候选实体名及其在问句中的位置。最后通过模型为每个候选命名实体计算一个得分  $NER\_SCORE$ 。

表 1 知识库内容示例

Table 1 Examples in NLPCC knowledge base

实体	属性	属性值
西游记(中国古典长篇小说)	书名	西游记
西游记(中国古典长篇小说)	类别	浪漫主义长篇神魔小说
西游记(中国古典长篇小说)	又名	西游释厄传
西游记(中国古典长篇小说)	创作年代	明朝
西游记(中国古典长篇小说)	作者	吴承恩

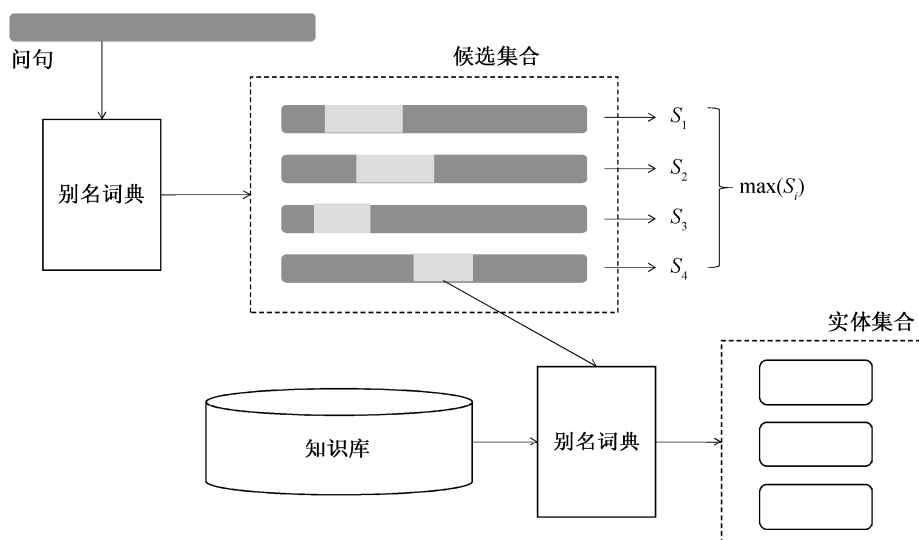


图 2 命名实体识别步骤

Fig. 2 Overview of named entity recognition

### 2.2.1 别名词典构建

首先根据知识库构建一个别名词典,用于命名实体识别步骤的候选实体获取,同时也可查找某个命名实体可以指代的所有实体。如“红楼梦”即可以指代实体“红楼梦(中国古典长篇小说)”,也可以指代实体“红楼梦(1987 年大陆央视版电视剧)”等。本文使用模版来匹配具有别名意义的属性,从知识库中抽取别名信息。如果一个属性能够匹配下列模版之一,则将其对应的属性值作为对应的实体的别名。1) 以“名”结尾:别名、中文名、英文名、原名等(第 X 名、排名等除外); 2) 以“称”结尾:别称、全称、简称、旧称等(XX 职称等除外); 3) 以“名称”结尾:中文名称、其他名称等(专辑名称、粉丝名称等除外)。

除此之外,如果实体名中存在括号,如“红楼梦(中国古典长篇小说)”,则将括号之外的部分作为该实体的别名,即“红楼梦”作为实体“红楼梦(中国古典长篇小说)”的别名。如果实体名中包含书名号,如“《计算机基础》”,则将书名号内的部分作为该实体的别名,即“计算机基础”作为实体“《计算机基础》”的别名。根据上述方法,最终得到一个包含 7304663 个别名的别名词典。

### 2.2.2 基于 LSTM 语言模型的特征

首先根据别名词典,查找问句中出现的命名实体,将其替换为特殊符号“\_NE\_”,用以表示命名实体。如输入问句“装备战伤理论与技术是哪个出版社出版的?”,经过该过程,得到下列的部分候选:

1) \_NE\_ 是哪个出版社出版的? ||| 装备战伤理论与技术

2) \_NE\_ 战伤理论与技术是哪个出版社出版的? ||| 装备

其中“|||”右边的是替换为“\_NE\_”之前的命名实体。

本文使用 LSTM 作为语言模型来计算每个候选的得分。语言模型是自然语言处理中虽简单却有效的模型。该模型计算的是语句  $S=w_1w_2...w_n$  作为一个自然语言的概率,符合常用语法结构的语句获得较高的概率,不符合常用语法结构的语句获得较低的概率。如果将表示命名实体的特殊符号“\_NE\_”视为一个普通词语,将替换后的语句送入语言模型,则正确的候选(如上述例 1)更符合自然语言语法结构,因此能够获得更高的概率。

使用 LSTM 作为语言模型的结构如图 3 所示,其中“<S>”和“</S>”分别表示语句的开始和结束。

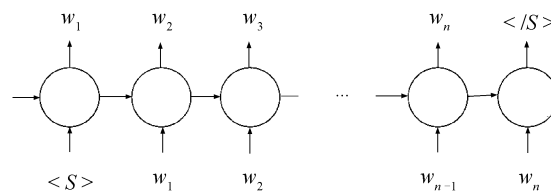


图 3 LSTM 语言模型

Fig. 3 LSTM language model

LSTM 语言模型使用之前的词序列来预测当前单词的概率。将输入的每个单词映射到固定维度的预训练的词向量后(“\_NE\_”的词向量初始为零向量,但训练过程中对词向量也进行更新),输入 LSTM 模型中。将每个 LSTM 单元的输出送入 softmax 分类器中,最后输出当前位置词表中每个单词的预测概率分布。整个句子的对数概率通过下式计算得到:

$$LP(S) = -\sum_{i=1}^n \log P(w_i), \quad (1)$$

其中,  $P(w_i)$  表示问句中第  $i$  个单词的预测概率。 $LP(S)$  越小,表示语句  $S$  符合语法的概率越高。

### 2.2.3 词表面特征

LSTM 语言模型得到的对数概率只利用了候选命名实体的位置和上下文信息,没有用到候选命名实体自身的任何信息。为了充分利用所有信息,提高命名实体识别的准确率,本文额外提取两个简单的特征:一个是候选命名实体的长度  $L(NE)$ (以字为单位),另一个是命名实体的 IDF 值  $IDF(NE)$ 。本文通过将训练集中每个问题视为一个文档,计算每个候选命名实体短语的 IDF 值。

最后将上述 3 个特征送入逻辑斯蒂回归模型中,得到每个候选的命名实体得分  $NER\_SCORE$ 。

### 2.3 属性映射

经过命名实体识别步骤,对于输入的问句“装备战伤理论与技术是哪个出版社出版的”,可以得到其询问的命名实体“装备战伤理论与技术”。依据别名词典,可以找到知识库中以该命名实体为正式名或别名的所有候选实体。本例中只有一个实体:一本名为“装备战伤理论与技术”的书籍。知识库中关于该实体的三元组共 5 个,其属性分别为“别名”、“中文名”、“出版社”、“平装”和“开本”。属性映射步骤的目的是找出这 5 个候选属性中与问句语义最相近的属性,该属性即为问句询问的内容。

NLPCC-ICCPOL 2016 KBQA 任务提供了一个

Baseline 系统, 该系统使用暹罗结构的卷积深度网络(Convolutional Deep Structure Semantic Model, C-DSSM)得到问句与属性的语义向量, 将这两个语义向量的余弦相似度作为最终得分。

在对数据进行观察和简单实验后, 我们发现属性映射阶段的重点在于找到属性中的单词与问句中的单词的对应关系。如问句“请问西游记是什么时候写的?”与知识库中的属性“创作年代”就存在这种对应关系: “创作”与“写”、“年代”与“时候”。这种对应关系可以通过这种注意力机制更好地被发现。因此, 本文提出结合两种不同注意力机制的双向 LSTM 模型, 对“问句-属性”对进行编码。一种是 Rocktäschel 等<sup>[13]</sup>提出的静态注意力机制, 另一种是 Yin 等<sup>[14]</sup>提出的起先应用于 CNN 模型的注意力机制。

### 2.3.1 结合静态注意力机制的双向 LSTM

结合静态注意力模型的双向 LSTM 模型的结构如图 4 所示。其中, 有两个双向 LSTM (BiLSTM) 模型, 左边 BiLSTM 的输入为去除命名实体后的问句, 右边 BiLSTM 的输入为候选属性。右边 BiLSTM 的初始状态为左边 BiLSTM 的最终状态, 在 BiLSTM 层之上有一个静态注意力层。

静态注意力机制的计算公式如下:

$$Y = [h_1 h_2 \dots h_n], \quad (2)$$

$$M = \tanh(W^y Y + W^h h_N \otimes e_L), \quad (3)$$

$$\alpha = \text{softmax}(w^T M), \quad (4)$$

$$r = Y\alpha, \quad (5)$$

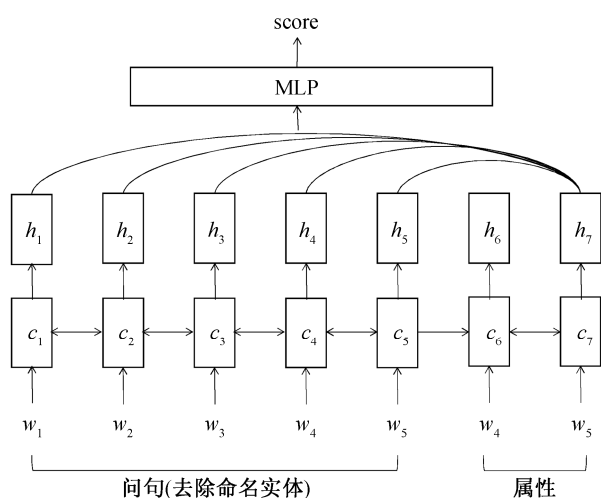


图 4 LSTM 和静态注意力结构

Fig. 4 Framework of LSTM combined with static attention

$$h^* = \tanh(W^p r + W^x h_N), \quad (6)$$

其中, 不同的  $W$  (用上角标区分) 表示网络中相应的权重矩阵;  $h_N \otimes e_L$  表示将列向量  $h_N$  重复  $L$  次, 形成一个  $L$  列的矩阵;  $\alpha$  为 Attention 的权重向量。最后得到的  $h^*$  作为输入, 通过全连接层 MLP, 得到最终得分。模型采用的损失函数如下:

$$\text{Loss} = \max\{0, 0.2 + S(Q, P^-) - S(Q, P^+)\}, \quad (7)$$

其中  $S(Q, P^+)$  表示正例的输出,  $S(Q, P^-)$  表示负例的输出。这种损失函数适用于排序类问题, 因为我们在最后使用该模型时, 将其视为排序模型而非分类模型。

### 2.3.2 基于单词语义相似度的注意力机制

该注意力机制首先计算一个注意力矩阵, 如图 5 所示, 其中  $S_p$  和  $S_q$  分别表示问句的词向量矩阵和属性的词向量矩阵; 注意力矩阵  $A$  中第  $i$  行第  $j$  个元素表示问句中的第  $i$  个单词和第  $j$  个单词的语义相似度;  $\|x - y\|^2$  表示两个向量  $x$  和  $y$  的欧式距离。得到  $A$  之后, 再根据两个参数矩阵  $W_p$  和  $W_q$  计算两个矩阵  $R_p$  和  $R_q$ , 这两个矩阵的大小分别与  $S_p$  和  $S_q$  相同。将  $R_p$  和  $R_q$  分别按行拼接到  $S_p$  和  $S_q$  之后, 作为双向 LSTM 的输入送入图 4 所示模型中。

最后, 将 LSTM 模型的输出结合人工选择的几个简单特征送入逻辑斯蒂回归模型中, 得到属性映射步骤的最后得分 PROP\_SCORE。人工选择的特征为属性与问句的 overlap (以字为单位) (即属性与问句中共同的字的个数) 除以属性的长度得到特征 OVP, 除以问句的长度得到特征 OVQ。

## 2.4 答案选择

在进行答案选择时, 将上述两个步骤的得分 NER\_SCORE 和 PROP\_SCORE 进行加权, 选择出得分最高的“命名实体-属性”对“NE\_PROP”。计算公式如下:

$$\text{Score} = \alpha \text{NER\_SCORE} + (1 - \alpha) \text{PROP\_SCORE}. \quad (8)$$

为减少计算量, 仅选取 NER\_SCORE 排在前 3 名的候选命名实体和 PROP\_SCORE 排在前 3 的候选相关属性。

得到“命名实体-属性”对后, 根据别名词典找到查找命名实体 NE 对应的实体集合  $E$ , 然后查看  $E$  中每个实体是否存在属性 PROP, 若存在, 则将其对应的属性值加入答案集 ANS。该步骤同时也实现了一定程度的实体消歧, 比如问句“红楼梦是

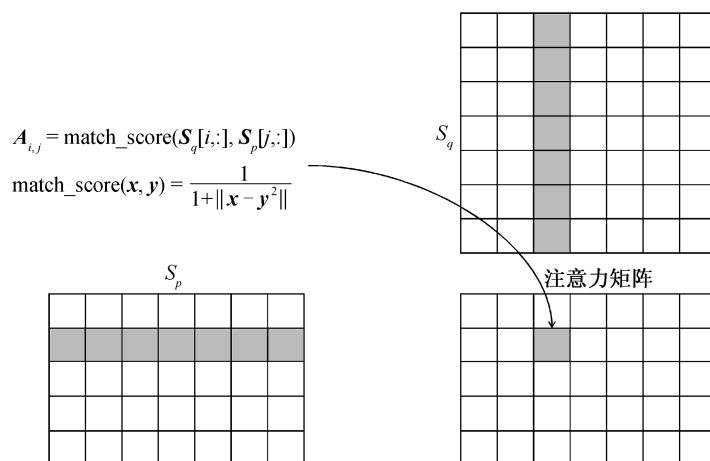


图5 注意力矩阵计算方法

Fig. 5 Calculation method of attention matrix

谁写的?”,询问对象为某一文学作品,与该问句语义相近的属性为“作者”,而其他名为“红楼梦”但不属于文学作品的实体不会包含“作者”属性,因此可以将其排除。

在实践中发现,知识库中存在含义相同但表述不同的属性,如同样表示“作者”,知识库中却有“作者”与“原作者”等属性。为了解决这类问题,我们将属性映射步骤的得分不小于  $\text{PROP\_SCORE}^* - \text{thresh}$  的属性同样考虑进来( $\text{PROP\_SCORE}^*$ 表示经过加权选择后的属性映射得分)。例如,如果上述加权选择后得到的属性为“作者”,得分为 0.9874,而“原作者”得分为“0.9779”,两者之差小于  $\text{thresh}$ ,则将对应属性为“原作者”的三元组中的属性值也加入答案集  $\text{ANS}$  中。

### 3 实验

NLPCC ICCPOL 2016 KBQA 任务提供了一个包含 14609 个问答对的训练集和包含 9870 个问答对的测试集。本文从总的训练集划分出 3000 个问答对作为开发集,剩余 11609 个问答对作为实际训练集。由于本文中实体链接步骤与属性映射步骤都需要对应的数据集,因此需要自行构建这两部分数据集。具体的方法为使用问答对中的标准答案,反向查找问句对应的“实体-属性”对,在构建命名实体识别数据集时,利用别名词典和人工规则进行命名实体的标注;在构建属性映射数据集时,也加入人工规则进行筛选,找出正确的“问句-属性”对。在剔除一些噪音数据后,最终的命名实体识别数据

集和属性映射数据集的规模如表 3 所示。

本文使用中文维基百科作为语料,使用 google word2vec 工具训练词向量,词向量的维度为 300。LSTM 语言模型的隐含层为 100 维,在命名实体识别数据集上进行训练和测试,结果如表 4 所示。可以看出,在测试集上该模型的准确率可以达到 97% 以上,而正确候选得分排在前 3 位(Top3)的样例已经达到 99.41%,证明模型可以很好地识别问句中的命名实体,为后续工作打下良好的基础。

属性映射步骤的 BiLSTM 的隐含层为 200 维,最后的 MLP 仅有一层,类似逻辑斯蒂回归模型。在进行模型训练时,由于数据集中正例远少于负例,因此本文将正例进行重复,达到与负例相同的程度,最终正负例比例为 1:1。表 5 为几种模型的训练和测试结果。

表 3 子任务数据集规模  
Table 3 Size of two subtask's data set

任务	训练集	开发集	测试集
命名实体识别	11369	2950	9483
属性映射	11369	2950	9483

表 4 命名实体识别准确率(%)  
Fig. 4 Accuracy of named entity recognition (%)

TopN	训练集	开发集	测试集
1	98.50	98.14	97.24
2	99.61	98.52	99.12
3	99.79	99.73	99.41

表 5 属性映射结果(%)

Table 5 Accuracy of property mapping (%)

模型	训练集	开发集	测试集
BiLSTM_AC1	92.37	91.76	86.74
BiLSTM_AC12	93.59	92.47	87.64
BiLSTM_AC12_Overlap	95.54	95.39	91.77

表 6 自动问答最终结果(%)

Table 6 Final results on question-answering(%)

指标	训练集	开发集	测试集
AvgPrecision	78.44	88.59	79.03
AvgRecall	93.23	92.97	87.42
AvgF1	81.78	89.387	81.06

官方评测最终结果采用的评价指标为平均 F1 值。由于每个样例的标准答案和候选答案均为集合的形式, 因此每个样例都可以得到一个 F1 值, 最后取所有样例 F1 值的平均。

选用各步骤取得最好结果的模型在数据集上进行实验, 当答案选择步骤的权重  $\alpha=0.7$ ,  $\text{thresh}=0.02$  时, 可以取得最好的结果。这两个值都通过枚举方法得到, 先选择取得最好实验结果的  $\alpha$ , 确定  $\alpha$  后, 再以枚举的方法得到  $\text{thresh}$ 。在 NLPCC 提供的数据集上进行实验, 得到表 6 的结果。

与参与 NLPCC-ICCPOL 2016 KBQA 评测任务的所有结果相比, 前两名的结果分别为 0.8244 与 0.8159, 但这两者都使用了其他较复杂的特征和人工规则(比如在命名实体识别阶段使用词性等特征)。本文在仅使用神经网络模型和少量简单文本特征的情况下, 取得接近的效果, 证明了本文模型的有效性。

## 4 结语

实体识别和属性映射是构建基于大规模知识库的问答系统的两个难点, 本文提出相应的方法解决这两个问题。在进行实体识别时, 首先采用别名词典获取候选实体, 然后使用 LSTM 语言模型结合简单的文本特征进行打分。在进行属性映射时, 考虑到属性与问句中的词对应关系, 结合两种不同的注意力机制, 使用双向 LSTM 结合简单文本特征获取正确的属性。最后, 结合前两步的得分, 进行实体消歧和答案选择。本文方法在 NLPCC-ICCPOL 2016 KBQA 任务的数据集上取得接近目前已发表的最好

成绩的效果。

## 参考文献

- [1] Dong L, Wei F, Zhou M, et al. Question answering over freebase with multi-column convolutional neural networks // ACL. Beijing, 2015: 260–269
- [2] Zettlemoyer L S, Collins M. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars // Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence. Edinburgh, 2005: 658–666
- [3] Berant J, Chou A, Frostig R, et al. Semantic parsing on freebase from question-answer Pairs // EMNLP. Seattle, 2013: 1533–1544
- [4] Liang P, Jordan M I, Dan K. Learning dependency-based compositional semantics. Computational Linguistics, 2011, 39(2): 389–446
- [5] Yao X, Van Durme B. Information extraction over structured data: question answering with freebase // ACL. Baltimore, 2014: 956–966
- [6] Bordes A, Chopra S, Weston J, et al. Question answering with subgraph embeddings // EMNLP. Doha, 2014: 615–620
- [7] Bordes A, Weston J, Usunier N. Open question answering with weakly supervised embedding models // ECML. Nancy, 2014: 165–180
- [8] Zhang Y, Liu K, He S, et al. Question answering over knowledge base with neural attention combining global knowledge information. arXiv preprint arXiv:1606.00979, 2016
- [9] Wang L, Zhang Y, Liu T. A deep learning approach for question answering over knowledge base // NLPCC 2016: Natural Language Understanding and Intelligent Applications. Kunming: Springer International Publishing, 2016: 885–892
- [10] Xie Z, Zeng Z, Zhou G, et al. Knowledge base question answering based on deep learning models // NLPCC. Kunming, 2016: 300–311
- [11] Yang F, Gan L, Li A, et al. Combining deep learning with information retrieval for question answering // NLPCC. Kunming, 2016: 917–925
- [12] Lai Y, Lin Y, Chen J, et al. Open domain question answering system based on knowledge base // NLPCC. Kunming, 2016: 722–733
- [13] Rocktäschel T, Grefenstette E, Hermann K M, et al. Reasoning about entailment with neural attention. arXiv preprint arXiv:1509.06664, 2015
- [14] Yin W, Schütze H, Xiang B, et al. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. arXiv preprint arXiv:1512.05193, 2015