面向旅游领域的哈萨克语知识图谱构建

古丽拉·阿东别克 1[1,2,3] 古丽扎达·海沙 2[1,2,3] 马雅静 3[1,2,3] 陈赟 4[1,2,3]
哈依纳尔·艾尔扎提 5[4] 骆铭 6[1,2,3]

（1.新疆大学信息科学与工程学院，新疆 乌鲁木齐市 830017；2.国家语言资源监测与研究少数民族语言中心哈萨克和柯尔克孜语文基地 乌鲁木齐 830017；3. 新疆多语种信息技术实验室 乌鲁木齐 830017；4.中国科技大学计算机科学与技术学院，安徽 合肥市 230027）

摘要：知识图谱的构建是从客观世界组织和存储知识，形成一个知识体系。然而，构建高质量、大规模的知识图谱对于资源匮乏的哈萨克语来说是一个重要的挑战。本文针对旅游领域的需求，提出了一种基于命名实体识别和关系抽取的哈萨克语知识图谱构建方法。首先，使用基于深度学习方法获得哈萨克语实体特征语义信息，利用 Bi-LSTM，从不同的方向学习，获取有效的上下文信息，通过多特征融合，融合单词和单词的位置特征和实体标签，得到单词和目标实体之间的关联。其次，使用预训练语言模型捕获句子和目标实体的语义信息，以更好地适应关系分类任务。最后，利用图形数据库 neo4j 存储模式层、本体和数据层三重知识，构建了哈萨克语的中国旅游知识图谱。

关键词：知识图谱；哈萨克语；旅游；深度学习

中图分类号：TP391 **文献标识码：**A

Construction of Kazakh Knowledge Graph in Tourism

Gulila ALTENBEK[1,2,3], Gulizada HAISA[1,2,3], Yajing Ma[1,2,3], Yun Chen[1,2,3], Hayinaer Aierzhati[4], Ming Luo[1,2,3],

(1.College of Information Science and Engineering, Xinjiang University, Xinjiang, 830017, China; 2.The Base of Kazakh and Kirghiz Language of National Language Resource Monitoring and Research Center on Minority Languages, Xinjiang,830017,China; 3.Xinjiang Laboratory of Multi-language Information Technology, Xinjiang,830017,China; 4.College of computer science, University of science and technology of China, Hefei, 230027,China)

Abstract: The construction of knowledge graph is to organize and store knowledge from the objective world to form a knowledge system. However, building a high-quality, large-scale knowledge graph is an important challenge for the low-resource Kazakh language. For the needs of the tourism field, this paper proposes a method for constructing Kazakh knowledge graphs based on named entity recognition and relation extraction. First, we use the deep learning method to obtain the semantic information of the Kazakh entity features, use Bi-LSTM to learn from different directions to obtain effective context information. Through multi-feature fusion, the word and word location features and entity tags are merged to obtain the association between the word and the target entity. Second, we use R-BERT to capture the semantic information of sentences and target entities to better adapt to relation classification tasks. Finally, we use the graph database neo4j to store the triple knowledge of the model layer, the ontology and the data layer to construct the Kazakh tourism knowledge graph。

Key words: Knowledge Graph; Kazakh Language; Tourism; Deep Learning

---

基金项目：新疆大学博士基金（No. BS 180250）;国家自然科学基金（No: 62062062）

0 引言

知识图谱以结构化的形式展示了客观世界中的概念、实体以及实体间的关系。近年来，领域知识图谱成为工业界和学术界研究的热点，这类知识图谱的描述目标是特定的行业领域，通常需要依靠特定行业的数据才能构建。IMDB[1]为获取影视娱乐信息提供了便利，MusicBrainz[2]是一个结构化的音乐维基百科，Ruan et al.[3]利用了数据驱动的增量式构建方法构建了中医药知识图谱、海洋知识图谱和企业知识图谱。而针对旅游领域，Calleja et al.[4]提出了 DBtravel，这是一个由协作旅游网站 Wikitravel 生成的面向旅游的知识图谱。Zhang et al.[5]构建了中文旅游知识图谱，促进了中国旅游知识和文化的共享，Xiao et al.[6]从流行的旅游网站收集数据，建立了一个多源异构的旅游知识图谱。

目前，哈萨克语使用者通过互联网对旅游和其他一些话题的评价也越来越多，因此构建一个哈萨克语旅游资源信息和智能化服务系统具有重要的现实意义。

特定领域的知识图谱已经得到了充分研究，但是面向旅游领域的知识图谱仍然面临着巨大的挑战，尤其是面向低资源语言的领域知识图谱。因此，本文以新疆地区旅游景点为背景构建哈萨克语旅游知识图谱。本文的主要贡献如下：

（1）通过自动化方法构建了旅游领域命名实体识别和关系抽取的哈萨克语数据集。

（2）根据哈萨克语的语言特点，命名实体识别任务使用了结合 Word2vec-BiLSTM-CRF 模型，对于关系抽取任务，我们使用了 R-BERT。

（3）构建了面向旅游领域的哈萨克语知识图谱。该知识图谱的构建给其他领域的知识图谱和低资源语言知识图谱的构建提供了方法和借鉴。

本文的结构安排如下：第二部分描述了知识图谱构建的相关工作，第三部分介绍了命名实体识别的数据构建和实验实现，第四部分对关系抽取任务的数据构建和实验进行介绍，第五部分讲解了知识图谱的构建及可视化过程，最后是对本文工作的总结和展望。

1 相关工作

先前的工作一般是针对大规模的知识图谱，如 YAGO[7]、DBpedia[8]、Freebase[9]、NELL[10]等，中文通用知识图谱有 Zhishi.me[11] 和 SSCO[12]。早期主要通过人工构建的方式，形成了 WordNet[13]、ResearchCyc[14]等通用知识图谱。此后，大量知识图谱基于维百科进行构建，如 YAGO、DBpedia 等，但由于抽取的目标数据不同，它们的知识丰富度各有差异[15]。

近年来，垂直领域的知识图谱成为研究热点，垂直知识图谱面向特定领域，基于行业数据构建。Miao et al.[16]对企业知识图谱进行了研究，Zhao et al.[17]构建了工业知识图谱，Li et al.[18]构建了基于电子病历（electronic medical records）的医疗知识图谱。当前已经构建的知识图谱多针对于中文和英文主流语言[19-21]。尽管针对构建具体领域的知识图谱的研究成为热潮，旅游领域涉及较少，而且对于特定领域内哈萨克语知识图谱的研究还未开展，学术界和工业界也没有公开的面向旅游领域的哈萨克语知识图谱。目前越来越多的知识图谱基于非结构化数据构建，需要利用命名实体识别(NER)和关系抽取(RE)技术。

命名实体识别（Named entity recognition, NER）是自然语言处理中的一项基础研究任务，是构建知识图谱的重要步骤。以往的命名实体识别任务大多针对通用领域，近年来，NER 在某些特定的领域开始新的尝试，在生物医学领域，王浩畅[22]用 SVM 进行蛋白质、基因、核糖核酸等实体识别；社交媒体领域中，李源等[23]对微博中的实体进行研究；罗凌等[24]对电子病历中的实体进行研究，此外，还有一些研究较少的实体，如化学实体[25]、古籍文本中的人名[26]等。旅游领域的命名实体识别研究较少。薛征山等[27]提出基于 HMM 的旅游景点识别方法，该方法首次在旅游领域进行命名实体识别，但是没有充分考虑上下文信息。郭剑毅等[28]提出使用层叠条件随机场识别景点名的方法，但是过于依赖人工标注的特征，刘小安等[29]提出了一种基于 BiLSTM-CRF 的网络模型，避免了人工构建特征，但是该方法基于字进行识别，未能充分利用词典信息。本文使用了基于神经网络的命名实体识别模型。

关系抽取是信息抽取的重要任务，也是构建知识图谱的必要前提。近年来，专家学者们提出了较多的实体关系抽取方法，Zhou 等[30]研究了基于特征的关系抽取中词汇、句法和语义知识的融合问题，并研究不同语言特征对关系抽取性能的贡献。Kambhatla 等[31]人使用最大熵模型分类器，以结合从文本中派生的不同的词汇、句法和语义特征进行模型训练。Zhao 等人[32]在 Zhou 等的基础上，引入词义扩展特征，语法解析特征等一些新的特征，使用符号化、句法分析和深度依赖分析三个不同层次的处理信息。Zhang 等[33]提出了一种新的用于关系抽取的融合卷积树核函数和线

性核函数，研究表明该方法可以有效地捕获平面特征和结构化特征，可以扩展及包含更多的特征。然而基于特征向量和核函数方法的性能的优劣很大程度上依赖于特征集，这项任务不但繁琐耗时，而且对领域知识的需求较高。本文只针对旅游领域的关系抽取，选用的模型为 R-BERT。

2 命名实体识别

2.1 命名实体识别模型

命名实体识别方法可以分为基于分词的方法和基于字符的方法，与中文 NER 任务不同，由哈萨克语构成的句子，句子中的空格即可以达到很好地分词效果，不会产生歧义的问题，所以针对哈萨克语旅游领域命名实体识别任务，本文使用了基于词的 BiLSTM-CRF 命名实体识别模型中。BiLSTM 能够利用双向的结构，从不同的方向学习，获取有效的上下文信息；CRF 可以考虑句子级相邻标签之间的信息，解决序列偏置问题，最终获得全局最优序列。

BiLSTM-CRF 模型由 BiLSTM 和 CRF 两个模块组成，整体模型如图 1 所示。首先使用通过 BiLSTM 深度学习上下文特征信息，进行命名实体识别，最后 CRF 层对 BiLSTM 的输出序列处理，结合 CRF 中的状态转移矩阵，根据相邻之间标签得到一个全局最优序列。

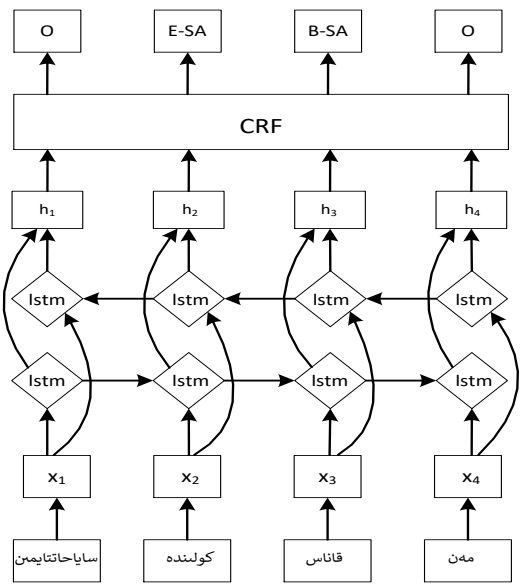

图 1 BiLSTM+CRF 命名实体识别模型

LSTM 以向量序列(x1，x2，...,xn)作为输入，并返回另一个序列(h1，h2, ...,hn）表示输入中每一步的序列的一些信息。具体计算过程如下：

给定包含 n 个字的句子

$$S = \{c_1, \ldots, c_n\} \quad (1)$$

其中c_i是第 i 个词，每个词通过查询预训练词向量，若预训练词向量表中没有，则随机生成：

$$x_i = E^c(c_i) \quad (2)$$

其中的E^c是预训练词向量表。

最后将 输入到 LSTM 结构中进行学习，LSTM 的计算过程如下：

$$f_t = \sigma(W_f \times [\hbar_{t-1}, x_t] + b_f) \quad (3)$$

$$i_t = \sigma(W_t \times [\hbar_{t-1}, x_t] + b_i) \quad (4)$$

$$C'_t = tan\hbar(W_C \times [\hbar_{t-1}, x_t] + b_C \quad (5)$$

$$C_t = f_t \times C_{t-1} + i_t \times C'_t \quad (6)$$

其中 f 表示遗忘门结构，i 表示输入门结构，o 表示输出门结构，C'表示 LSTM 的记忆单元。LSTM 通过输入门和激活函数之间相互约束来控制当前状态新信息加入的程度i_t并通过i_t生成当前状态的候选单元特征C'_t；然后与遗忘门特征 ft 相结合，生成当前状态的 LSTM 单元特征c_t。输出特征o_t用来控制当前状态 LSTM 单元特征c_t被过滤的程度。采用o_t对c_t的映射结果进行计算得到当前 LSTM 单元状态的输出特征 ht。

$$o_t = \sigma(W_o \times [\hbar_{t-1}, x_t] + b_o) \quad (7)$$

$$\hbar_t = o_t \times tan\hbar(C_t) \quad (8)$$

2.2 数据集和评价指标

2.2.1 数据集

数据集的缺乏一直是领域命名实体识别任务的研究难点之一。目前，针对旅游领域的命名实体识别研究相对较少，也缺乏公开的数据集，尤其是哈萨克语命名实体识别数据集，为了在该领域上取得较好的识别效果，也为之后的研究奠定基础。本文利用 python 爬虫技术，爬取旅游领域的相关文本数据，并且经过一系列的文本预处理、标注工作，构建了小型的哈萨克语旅游领域命名实体识别数据集。

实体标注通过给定文本中的实体和非实体相应的标签，以此区分实体的种类。该文定义了 7 类实体，分别是景点（SA），地址（LOC），组织机构（ORG），人名（PER），特色小吃（SC），民族（NA），文化（CU）。通用的标注体系包括 BIO 标注和 BIOES 标注，本文使用的是 BIOES 标注体系。命名实体数据如表 1 所示。

表 1 命名实体识别数据

Type	Train	Test	Dev
Sentence	2326	290	290
SA	2363	254	270
LOC	2201	199	209
SC	302	15	12
PER	14	33	29
NA	199	4	11

对于句子“硅化木-恐龙国家地质公园位于新疆维吾尔自治区昌吉回族自治州奇台县境内,地处天山北麓,准噶尔盆地东南缘。”标注如图 2 所示。

كرەمنيلى	B_SA	شونجى	B_LOC
-	I_SA	اۇدانىندا	E_LOC
دينوزاۋر	I_SA	،	O
مەملەكەتتىك	I_SA	تيانشانلىق	S_SA
گەولوگيالىق	I_SA	سولتۇستىك	O
باقشاسى	E_SA	باۋرايىنا	O
شينجياڭ	B_LOC	،	O
ۇيغۇر	I_LOC	جوڭعار	B_LOC
اۆتونوميالى	I_LOC	ويپاتىنلىق	E_LOC
رايونى	E_LOC	شىعىس	O
سانجى	B_LOC	وڭتۇستىك	O
حۇيزۇ	I_LOC	جيەگىنە	O
اۆتونوميالى	I_LOC	ورنالاسقان	O
وبلىسىنلىق	E_LOC	.	O

图 2 命名实体识别标注数据示例

2.2.2 **评价指标**

评价指标采使用准确率（P），召回率（R）和 F1 值，公式参数定义如下： 为正确识别的实体个数， 是识别出的不相关的实体个数， 是数据集中存在且未被识别出来的实体个数。

$$P=TP/(TP+FP)\times 100\% \quad (9)$$

$$R=TP/(TP+FN)\times 100\% \quad (10)$$

通常精确率和召回率的数值越高，代表实验的效果好，然而有时精确率越高，召回率越低。所以需要综合考量他们的加权调和平均值，也就是 值， 值定义如下：

$$F1=2PR/(P+R)\times 100\% \quad (11)$$

2.3 实验结果及分析

使用构建的数据集在 CRF、CNN+CRF、BiLSTM+CRF 模型上进行了对比实验，实验结果如表 2 所示：

表 2 实验结果

Models	P(%)	R(%)	F1(%)
CRF	76.45	62.89	69.01
CNN+CRF	76.87	64.66	70.24
BiLSTM+CRF	78.42	67.52	72.56

由表 2 可以看出，BiLSTM+CRF 模型的识别性能明显优于 CRF 模型和 CNN+CRF 模型，CNN+CRF 模型虽然取得了比 CRF 更好的成绩，但是由于 CNN 只能获取局部特征，而命名实体识别任务需要关注上下文信息，多以提升效果一般。BiLSTM 可以自动抽取观测序列的上下文相关特征，并且是从两个方向获取文本信息，而不需复杂的特征工程，添加 CRF 层后为模型添加了学习标签之间的依赖关系的能力，显著提升识别能力。

3 关系抽取

3.1 任务描述

关系抽取的任务是预测名词对之间的语义关系，给定一个文本序列 s 和一对名词 e1 和 e2，目标是识别 e1 和 e2 之间的关系，我们将该任务看作是对关系的分类。关系抽取不仅依赖整个句子的信息，也依赖具体的目标实体的信息。

针对哈萨克语旅游领域关系抽取任务，我们将 BERT 模型应用在关系分类上。我们首先在目标实体的位置前后插入特殊的标记(token)，然后将文本输入 BERT 进行 fine-tuning，以识别两个目标实体的位置并将信息传给 BERT 模型。然后，在 BERT 模型的输出 embeddings 中找到两个目标实体的位置，使用 token 嵌入作为多层神经网络分类的输入。通过这种方式，能捕获句子和两个目标实体的语义信息，以更好地适应关系分类任务。本文使用的关系抽取模型如图 3 所示。

3.2 关系抽取模型架构

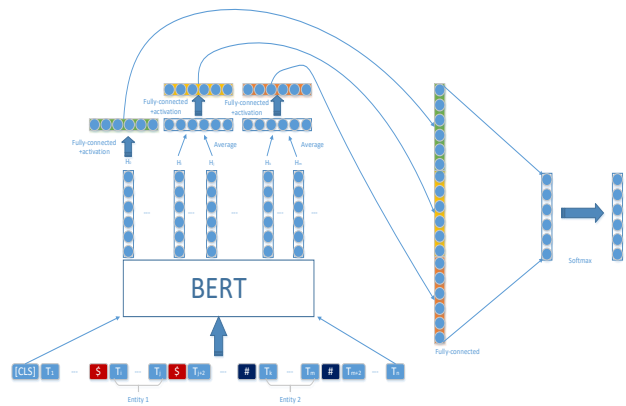

图 3 关系抽取模型

对于有两个目标实体 e1 和 e2 的语句 S 来说，为了让 BERT 模块能够获取两个实体的位置信息，在第一个位置实体的前后插入"<e1></e1>”符号，在第二个实体的前后插入“<e2></e2>”符号。举例说明，在插入一个特殊的分隔符后，一个拥有两个实体的语句将变成：

“ادامدار اراسىندداعى اسىم ‹e2›/لىتۇستىك شينجياڭدا‹/e2›سو (e1,e2)كورىنىس رايونى-ادرەسى دۇنيە ٔجۇزى سىرتىندداعى شاپتول قلانارى قوم قىستاعى، ‹e1›قاناس كولى‹/e1›، كورىنىس ساحارادداعى مەڑوزرت سايرام كولى، شۇرايلى جايلاۋ، كۇنەس، الۋان تۇسكە بويالعان ٔتۇستى قايراڭ، ..ادامدار اراسىندداعى شەكتەلگەن ورقى سياقتى ايبىندى عاجايىپ كورىنىستەر بار”

给定一个包含实体 e1 和 e2 的语句 S，假设

BERT 获取到它的最终隐藏状态为 H,实体 e1 的隐藏向量为 Hi 和 Hj，实体 e2 的隐藏向量为 Hk 和 Hm。我们对每个实体的所有向量进行求平均。然后再 tanh 函数并添加一个全连接层。

公式如下：

$$H_1^{'} = W_1[tanh(\frac{1}{j-i+1}\sum_{t=i}^{i} H_t)] + b_1 \quad (12)$$

$$H_2^{'} = W_2[tanh(\frac{1}{m-k+1}\sum_{t=k}^{i} H_t)] + b_2 \quad (13)$$

其中 W1 和 W2 共享参数，b1 和 b2 共享参数，即 W1=W2,b1=b2。对于第一个 token（[CLS]）所表示的最终隐藏状态向量，也添加一个 tanh 函数和全连接层。

$$H_0^{'} = W_0(tanh(H_0)) + b_0 \quad (14)$$

连接$H_0', H_1',$和H_2'，然后添加一个全连接的层和 softmax 层，可以表示如下:

$$H^{''} = W_3[concat(H_0', H_1', H_2')] + b_3 \quad (15)$$

$$\mathrm{P} = \mathrm{softmax}(H^{''}) \quad (16)$$

其中 W3∈ RL×3d(L 为关系类型数)，p 为输出概率。b0、b1、b2、b3 是偏置向量。我们使用交叉熵作为损失函数，在每个全连接层前使用 dropout。

3.3 数据集和评价指标

由于哈萨克语旅游语料少，本文采用远程监督方法爬取汉语旅游景点的数据来构建低资源哈萨克语进行关系抽取。本文利用 python 爬虫技术，爬取旅游领域的相关文本数据，并且经过一系列的文本预处理、标注，利用自治区哈萨克语-汉语转换器进行翻译并进行人工校对，通过统计文本中出现的关系结合新疆地区特色，定义了 30 种关系和一种不属于以上任何定义的关系（other），构建了哈萨克语旅游领域关系抽取数据集，共计 7456 条，详细数据如表 3 所示：

表 3 关系抽取详细数据

	训练集	测试集	数量
景点-地点	1262	316	1578
景点-面积	619	155	774
景点-特产	416	104	520
景点-距离	390	98	488
景点-别称	352	88	440
景点-地址	276	69	345
景点-海拔	264	66	330
景点-美食	260	65	325
景点-地区	206	52	258
景点-少数民族	198	49	247
景点-门票价格	191	48	239
民族-美食	174	43	217
景点-类型	163	41	204
民族-习俗	153	38	191
景点-级别	106	26	132
景点-温度	99	25	124
组织结构-酒店	87	22	109
组织结构-旅行社	83	21	104
地区-特产	81	20	101
景点-电话号码	74	18	92
组织机构-酒店	51	13	64
景点-所属景区	50	12	62
组织机构-旅行社	49	12	61
景点-住宿	46	11	57
景点-气候类型	42	10	52
地区-美食	30	7	37
地区-民族	29	7	36
景点-住宿价格	24	6	30
地区-温度	13	3	16
地区-气候类型	7	2	9
其他	—	—	214
总数	5794	1448	7456

实验的评价指标仍然是采使用准确率，召回率和 值，公式参数定义如 3-8，3-9，3-10 所示。

3.4 实验结果及分析

本文的实验数据集为旅游领域的 7456 条哈萨克语料,其中包含训练集 5794 条句子，1448 条测试句子，一共包含 30 种关系和一种无任何关系(other)，实验相关设置表 2 所示：

表 2 系统参数设置

序号	属性	值
1	数据集	3840 条哈萨克语料
2	最优化模型	Adam
3	学习率 rate	2e-5
4	batch_size	12
5	迭代次数 epoch	12
6	dropout	0.1
8	最大长度 maxlen	384

实验结果如表 4 和图 4 所示：

表 4 实验结果

模型	P	R	F1	ACC
R-BERT	70.69	70.85	69.64	86.56
R-BERT-NO_SEP	69.50	69.33	67.65	85.45
R-BERT-NO_ENT	66.51	71.21	68.10	86.07
R-BERT-NO_ENT-NO_SEP	63.77	66.33	63.90	84.46

这里给出三种消融对比试验：

（1）R-BERT-NO-SET-NO-NET：在模型基础上，BERT 输出部分去掉两个实体的隐含向量，去掉实体标记$和#，同时在 BERT 输出部分仅用[CLS]进行分类。

（2）R-BERT-NO-SEP：在模型基础上，去掉实体标记$和#，BERT 的输出部分保留。

（3）R-BERT-NO-ENT：在模型基础上，BERT 输出部分去掉两个实体的隐含向量，但是保留标记$和#。

对关系抽取结果的分析，由以下折线图可以发现，R-BERT 模型在未去掉实体标记和两个实体的隐含向量时效果最好，F1 值达到 69.64.

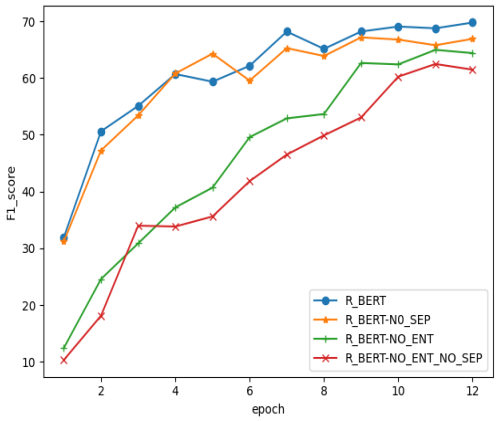

图 4 对比实验结果

4 知识图谱构建

通过关系抽取，可以获取旅游实体信息以及各实体间的 20 余种关系类型，将关系抽取获取到的结构化三元组信息用 neo4j 数据库存储，利用 python 轻量级 Flask 框架搭建可视化系统，我们使用 JS 语言中的 D3 库将处理后的节点间关系进行可视化展示，该系统主要支持国内使用的哈萨克语文字（阿拉伯字母）检索和展示，用节点文字信息、力导向图的形式将旅游信息展示给用户，给用户以直观明确的了解景点的相关信息。

新疆旅游知识图谱系统主要包含了两大功能模块：

（1）地点检索知识图谱模块

地点检索知识图谱模块，将关系抽取获得的景点及其属性关系按照地点名进行分类，其地点名包含省、市、区和县等地点信息，如下图 1，在界面左上方留有搜索框，在用户搜索框输入乌鲁木齐市，图谱展示出所有与乌鲁木齐市相关的信息，并以节点和力导向图的形式展示，并且不同的实体之间用不同的颜色进行区分，该知识图谱可以进行移动、拖拽和放缩，当鼠标点击某个实体的时候，可以突出显示与点击实体存在关系的实体，其他的实体不显示出来，这样有利于图谱的可读性以及简洁性；当鼠标移动到实体连接线时，可以显示出两个实体间存在的关系，同时隐藏其他不相干节点和关系连接线。

根据市州名可查询所属的区县信息，如图所示在搜索框输入哈萨克语乌鲁木齐市“ۇرىمجى قالاسى”，可查询乌鲁木齐市所属的县区，及其县区所包含的景点信息，如图 5 所示。

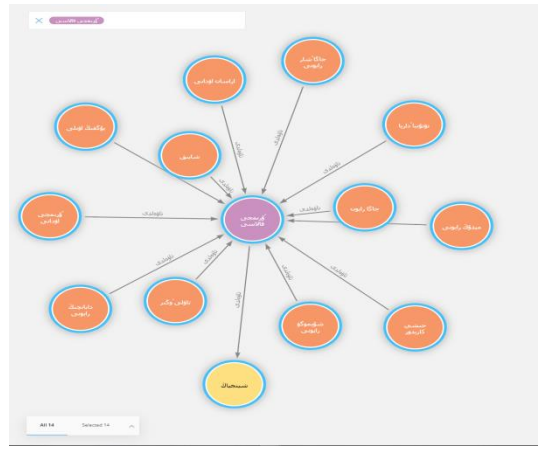

图 5 系统可视化图

在搜索框输入哈萨克语新市区“رايونى شار جاڭا”可查询的所有景点信息如图 6 所示：

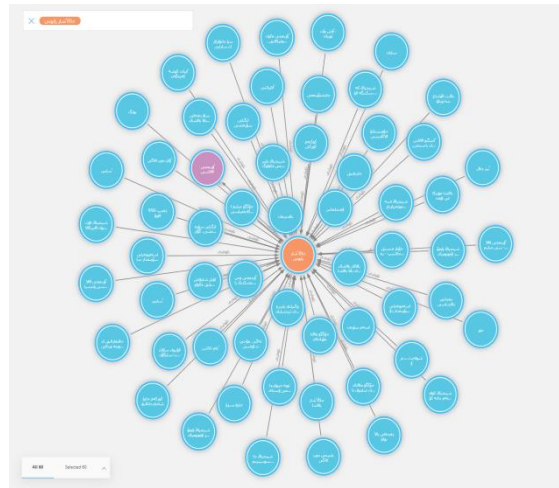

图 6 系统搜索结果图

（2）景点名称检索知识图谱模块，如下图所示，在界面左上方存在景点名信息检索框，以单个景点进行检索，在图谱的展示界面展示与该景点节点相关的节点信息以及关系导向图，用力导向图展示与景点相关的 20 余种关系属性，并且在图谱右方展示与该节点有关的所有信息作为描述性属性文本，该模块同时展示及哈语的检索和展示，输入“رايونى شار جاڭا”，可显示新市区的景点信息。

5 结论

本文构建了哈萨克语命名实体识别和关系抽取的旅游数据集，在此基础上构建了面向新疆旅游领域的哈萨克语知识图谱，为垂直领域的知识图谱的构建提供了借鉴，同时也为低资源语言的哈语自然语言处理提供了重要的资源。接下来的工作将围绕此知识图谱进行，将尝试构建更加全面的旅游领域知识图谱，或者利用构建好的知识

图谱进行知识推理、信息检索、构建一个多语言的智能问答系统。

参考文献

[1] IMDB Official. IMDB[EB/OL]. [2016-02-27]. http://www. imdb.com.

[2] MetaBrainz Foundation. Musicbrainz[EB/OL]. [2016-06-06]. http://musicbrainz.org/.

[3] RUAN T, DONG X, WANG H, et al. Evaluating and comparing web-scale extracted knowledge bases in Chinese and English[C]//Joint international semantic technology conference. Yichang: Springer International Publishing, 2015: 167-184.

[4] Calleja P, Priyatna F, Mihindukulasooriya N, et al. DBtravel: a tourism-oriented semantic graph[C]//International Conference on Web Engineering. Springer, Cham,2018: 206-212.

[5] Zhang W, Cao H, Hao F, et al. The chinese knowledge graph on domain-tourism[M]//Advanced Multimedia and Ubiquitous Engineering. Springer, Singapore, 2019: 20-27.

[6] Xiao D, Wang N, Yu J, et al. A Practice of Tourism Knowledge Graph Construction Based on Heterogeneous Information[C]//China National Conference on Chinese Computational Linguistics. Springer, Cham, 2020: 159-173.

[7] BIEGA J, KUZEY E, SUCHANEK F M. Inside YAGO2s: a transparent information extraction architecture[C]// Proceedings of the 22nd international conference on World Wide Web companion. Rio de Janeiro: International World Wide Web Conferences Steering Committee, 2013: 325-328.

[8] BIZER C, LEHMANN J, KOBILAROV G, et al. DBpedia-A crystallization point for the Web of data[J]. Web Semantics: science, services and agents on the World Wide Web, 2009, 7(3): 154-165.

[9] [9]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. Vancouver: ACM, 2008: 1247-1250.

[10] CARLSON A, BETTERIDGE J, KISIEL B, et al. Toward an architecture for never-ending language learning[C]//Proceedings of the twenty-fourth AAAI Conference on artificial intelligence. Atlanta: AAAI Press, 2010: 3.

[11] NIU X, SUN X, WANG H, et al. Zhishi. me-weaving chinese linking open data[M]. The Semantic Web–ISWC 2011. Berlin: Springer, 2011: 205-220.

[12] HU F, SHAO Z, RUAN T. Self-supervised Chinese ontology learning from online encyclopedias[J]. The scientific world journal, 2014, 2(11):1-13.

[13] MILLER G A. Word Net: a lexical database for English[J]. Communications of the ACM,1995, 38(11): 39-41.

[14] BOLLACKER K, EVAN C, PARITOSH P, etc. Freebase: a collaboratively created graph database for structuring human knowledge[C]. Proceeding of the 2008 ACM SIGMOD international conference on management of data. Vancouver: ACM, 2008:1247-1250.

[15] RUAN T, DONG X, WANG H, et al. Evaluating and comparing web-scale extracted knowledge bases in Chinese and English[C]//Joint international semantic technology conference. Yichang: Springer International Publishing, 2015: 167-184.

[16] Miao Q, Meng Y, Zhang B. Chinese enterprise knowledge graph construction based on Linked Data[C]//Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015). IEEE, 2015: 153-154.

[17] Zhao M, Wang H, Guo J, et al. Construction of an industrial knowledge graph for unstructured chinese text learning[J]. Applied Sciences, 2019, 9(13): 2720.

[18] Li L, Wang P, Yan J, et al. Real-world data medical knowledge graph: construction and applications[J]. Artificial intelligence in medicine, 2020, 103: 101817.

[19] 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.

[20] Suchanek F M, Kasneci G, Weikum G. Yago: a core of semantic knowledge[C]//Proceedings of the 16th international conference on World Wide Web. 2007: 697-706.

[21] Xu B, Xu Y, Liang J, et al. CN-DBpedia: A never-ending Chinese knowledge extraction system[C]//International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer, Cham, 2017: 428-438.

[22] Wang H, Zhao T. SVM-based biomedical name entity recognition[J]. J. Harbin Eng. Univ, 2006, 27: 570-574.

[23] LI Yuan, MA Lei, SHAO Dangguo,et al. Chinese named entity recognition for social media [J]. Chinese Information Journal,2020,34(8):61-69.

[24] LUO Ling, YANG Zhihao, SONG Yawen, LI Nan,et al. Research on naming entity recognition of Chinese electronic medical records based on stroke ELMo and multi-task learning [J]. Chinese Journal of Computers, 2020, 43(10):1943-1957.

[25] Leaman R, Wei C H, Lu Z. tmChem: a high performance approach for chemical named entity recognition and normalization[J]. Journal of cheminformatics, 2015, 7(1): 1-10.

[26] TANG Yafen. Study on automatic recognition of names in ancient Chinese classics before Qin Dynasty[J]. Modern Library and Information Technology,2013,29(7-8):63-68.

[27] XUE Zhengshan, GUO Jianyi, YU Zhengtao, et al. Recognition of Chinese tourist attractions based on HMM[J]. Journal of Kunming University of Science and Technology. 2009, 34(6):44-48.

[28] GUO Jianyi,XUE Zhengshan,YU Zhengtao,et al. Recognition of named entities in the tourism field based on stacked conditional random fields [J]. Journal of Chinese Information Processing, 2009,23(5):47-53.

[29] LIU Xiaoan, PENG Tao. Research on Chinese scenic spot recognition based on convolutional neural network[J]. Computer Engineering and Applications,2020,056(004):140-145.

[30] Zhou G D, Su J, Zhang J, et al. Exploring various knowledge in relation extraction[C]//Proceedings of the 43rd annual meeting of the association for computational linguistics (acl’05). 2005: 427-434.

[31]]Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for infor-

mation extraction[C]//Proceedings of the ACL Interactive Poster and Demonstration Sessions. 2004: 178-181.
[32] Zhao S, Grishman R. Extracting relations with integrated information using kernel methods[C]//Proceedings of the 43rd annual meeting of the association for computational linguistics (acl' 05). 2005: 419-426.
[33] Zhang M, Zhang J, Su J, et al. A composite kernel to extract relations between entities with both flat and structured features[C]//Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. 2006: 825-832.