DATA PAPER

Knowledge Graph Construction and Applications for Web Search and Beyond

Peilu Wang, Hao Jiang, Jingfang Xu & Qi Zhang[†]

Sogou Inc., Beijing 100084, China

Keywords: Knowledge graph; Search engine; Question answering

Citation: P. Wang, H. Jiang, J. Xu & Q. Zhang. Knowledge graph construction and applications for Web search and beyond. Data Intelligence 1(2019), 345-361. doi: 10.1162/dint_a_00019

Received: May 27, 2019; Revised: August 18, 2019; Accepted: August 22, 2019

ABSTRACT

Knowledge graph (KG) has played an important role in enhancing the performance of many intelligent systems. In this paper, we introduce the solution of building a large-scale multi-source knowledge graph from scratch in Sogou Inc., including its architecture, technical implementation and applications. Unlike previous works that build knowledge graph with graph databases, we build the knowledge graph on top of SogouQdb, a distributed search engine developed by Sogou Web Search Department, which can be easily scaled to support petabytes of data. As a supplement to the search engine, we also introduce a series of models to support inference and graph based querying. Currently, the data of Sogou knowledge graph that are collected from 136 different websites and constantly updated consist of 54 million entities and over 600 million entity links. We also introduce three applications of knowledge graph in Sogou Inc.: entity detection and linking, knowledge based question answering and knowledge based dialogue system. These applications have been used in Web search products to help user acquire information more efficiently.

1. INTRODUCTION

A knowledge graph (KG) is a kind of special database which integrates information into an ontology. As an effective way to store and search knowledge, knowledge graph has been applied in many intelligent systems and drawn a lot of research interest. While many knowledge graphs have been constructed and published, such as Freebase [1], Wikidata [2], DBpedia [3] and YAGO [4], none of these works could completely fulfill the application requirement of Sogou Inc. The main challenges are listed below:

[†] Corresponding author: Qi Zhang (Email: qizhang@sogou-inc.com; ORCID: 0000-0003-0947-4942).

Lack of data: Though the biggest published knowledge graph (Wikidata) is reported to contain millions of entities and billions of triples, most of their data are extracted from Wikipedia and are still far less than fulfilling the requirements of Web search applications such as general purpose question answering and recommendations. For example, none of the existing knowledge graphs contains the latest Chinese songs' information which can only be obtained from specific websites.

Uncertainty of scalability: None of the existing works explicitly report their systems' capability to deal with large-scale data or discuss how the knowledge graph could be expanded on server cluster. This problem might not be very important for academic research since even the biggest knowledge graph's data can still be held by single server with a large hard disk drive. In the case of search engines, the potential data requirement of a knowledge graph is much larger and using distributed storage is unavoidable.

To solve these challenges, we propose a novel solution of building a large-scale knowledge graph. We use a distributed search engine called SogouQdb that is developed by Sogou Web Search Department for inner use as the core storage engine to obtain the capability of scalability, and develop a series of models to supply inference and graph-based querying functions which make the system compatible with the other knowledge graph applications. The inference is conducted on HDFS with Spark which makes the inference procedure capable of dealing with big data. The Sogou knowledge graph is built with this solution and has been published to support online products. Currently, the Sogou knowledge graph consists of 54 million entities and over 600 million entity links. The data are extracted from 136 different websites and constantly updated.

We also introduce three applications of knowledge graphs in Sogou Inc.: entity detection and linking, knowledge-based question answering and knowledge-based dialogue systems. These applications have been used as an infrastructural service in Web search products to help users find the information they want more efficiently.

The rest of this paper is organized as follows: In Section 2, we introduce the related works of widely known published knowledge graphs. In Section 3, we elaborate our solution to construct a knowledge graph from scratch. Section 4 presents the application of knowledge graphs, especially in Sogou Inc. Finally, we draw a conclusion in Section 6.

2. RELATED WORK

While many works about building a domain-specific knowledge graph have been published, we focus on works of building large-scale multi-domain knowledge graphs and list the most widely known works in this section.

Freebase [1] was published as an open shared database in 2007 and was shut down in 2016 after all of its data are transferred to Wikidata. The data of Freebase were collected from Wikipedia[①], NNDB[②], Fashion

[①] https://www.wikipedia.org/

[②] http://www.nndb.com

Model Directory[3] and MusicBrainz[4], and were also contributed by its users[5]. Freebase has more than 1.9 billion triples[6], 4,000 types and 7,000 properties [1].

Wikidata [2] was firstly published by Wikidata in 2012 and has been publicly maintained until now. The data of Wikidata[7] that contain more than 55 million entities mainly come from its Wikipedia sister projects including Wikipedia, Wikivoyage, Wikisource and other websites.

DBpedia [3] is a large-scale multilingual knowledge graph and its data were extracted from Wikipedia and collaboratively edited by the community. The English version of DBpedia[8] contains more than 4.58 million entities and data of DBpedia in 125 languages have 38.3 million entities [5, 6, 7].

YAGO [4] is an open-sourced semantic knowledge graph derived from Wikipedia, WordNet and GeoName. YAGO has more than 10 million entities and 120 million entities' facts[9].

ConceptNet [5] that originated from the Open Mind Common Sense project which was launched in 1999 has grown to be an open multilingual knowledge graph. ConceptNet contains more than 8 million entities and 21 million entity links.

CN-DBpedia [6] is a Chinese KG published in 2017 that specifically focuses on extracting knowledge from Chinese encyclopedias. CN-DBpedia has more than 16.8 million entities and 223 million entity links[10].

3. CONSTRUCTION

An overview of the construction framework of Sogou knowledge graph is shown in Figure 1. The data of Sogou knowledge graph are collected from various websites which allow their data to be downloaded or crawled, e.g., Wikipedia and SogouBaike. The extracted data are stored in a distributed database in the form of JSON-LD (JavaScript Object Notation for Linked Data) which is a commonly used concrete RDF syntax. As an additional way to supply data, we introduce inference model which infers new relationships between entities. To search and browse the knowledge graph, a SPARQL query engine is developed that provides RESTful APIs services. For supporting a search engine's products like question answering and recommendation, the knowledge graph data are processed to adapt to the data form of specific tasks. In this section, we give an introduction of each part of the construction framework.

[3] https://www.fashionmodeldirectory.com/

[4] https://musicbrainz.org

[5] https://en.wikipedia.org/wiki/Freebase

[6] https://developers.google.com/freebase/

[7] https://www.wikidata.org/wiki/Wikidata

[8] https://wiki.dbpedia.org/about.

[9] https://datahub.io/collections/yago

[10] http://kw.fudan.edu.cn/

3.1 Data Extraction

The role of data extraction is extracting data into pre-defined form from various input data. Specifically, the input and output of data extraction are defined as follows:

> Input: Data downloaded or crawled from the Internet, e.g., the Web pages, XML data or JSON data downloaded by APIs. While the input data comprise mostly of free text, many data contain structured information such as: images, geo-coordinates, links to external Web pages and disambiguation pages.
> Output: Structured data in the form of JSON-LD that record the knowledge information extracted from the input data.

Data extraction operations can be classified into two categories: Structured data extraction only deals with the input data with structured information, specifically, the data that contain recognizable markup. Free text extraction detects entities and extracts the property information of specific entities from free text.

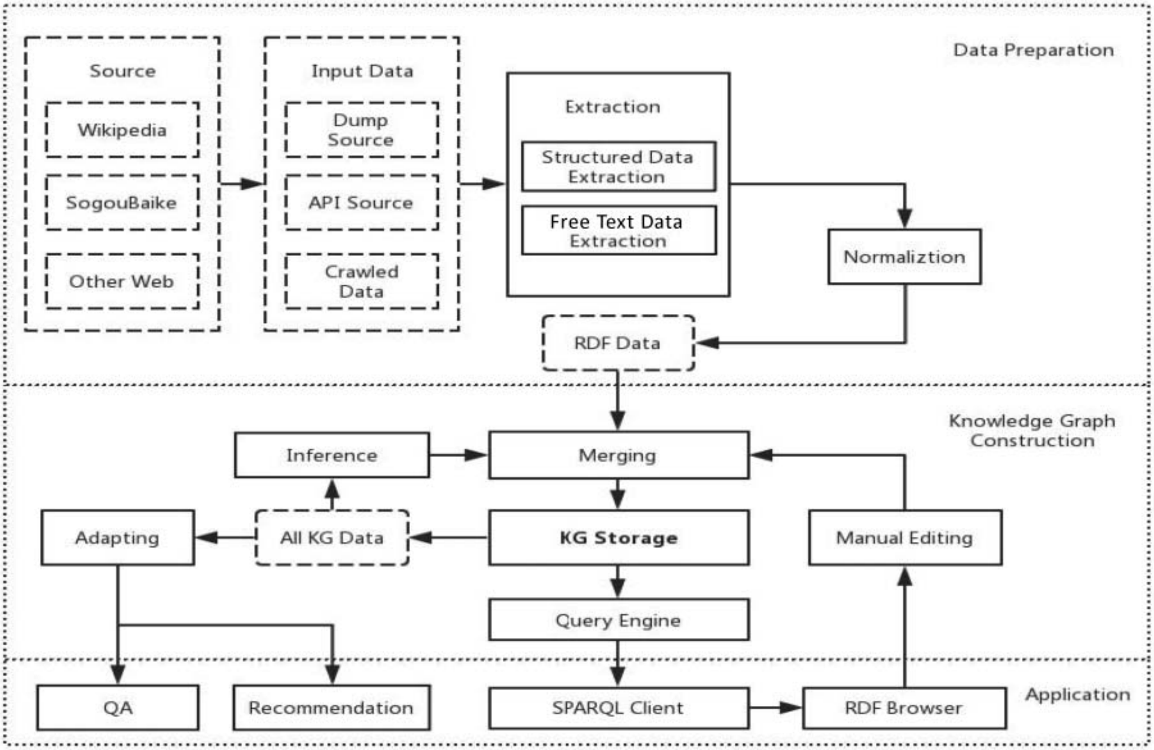


Figure 1. Overview of Sogou knowledge graph construction framework. The framework could be divided into three parts: Data Preparation contains operations including collecting data from various sources, extracting data from both structured source and free text and normalizing data; Knowledge graph construction contains all models to build a knowledge graph based on the extracted and normalized data; Application is composed of applications or services of a knowledge graph. A box with solid line represents an operation or model to process data while a box with dashed line represents the intermediate data.

3.1.1 Structured Data Extraction

As the structured information has recognizable markups, we use rule-based method to build the extractors. The extractors firstly parse the Web page to unified DOM-tree, then find the target information according to the manually written rules and save the extracted data in JSON-LD form. For each website, we build specialized extractors to deal with its data to make it independently update the data of different websites. Currently, in March 2019, Sogou knowledge graph system has 45 websites as data sources and 77 rule-based extractors.

3.1.2 Free Text Extraction

The task of free text extraction is combined with a series of sub-tasks including extracting named entity mentions from plain text, linking the mentions to the entities in knowledge graphs and extracting entities' properties or the relationships between extracted entities. Since training a model that could deal with all entity types is quite time consuming, we currently just focus on limited types of entities including: Person (PER), Geo-political Entity (GPE), Organization (ORG), Facility (FAC) and Location (LOC). For named entity recognition and linking tasks, we train a Bi-LSTM-CRF model and the feature and parameter selection follows work of [7] which got the best performance in TAC KBP 2017 competition [8]. The training data are constructed by the SogouBaike and Wikipedia Web pages that contain anchor markups. More details of the model and the training data can be found in Section 4.1.

3.2 Normalization

This part normalizes property values of extracted entities and maps entities' class and property to terms in the Sogou knowledge graph's ontology. Besides, data types of property are also specified, which ensures the high quality of processed data. The input and output of this part are defined as follows:

Input: Output of data extraction: Structured data in the form of JSON-LD.
Output: Structured data in the form of JSON-LD with normalized property name and property value.
The type of property value follows the definition of Sogou knowledge graph schema. A simplified example is given below:

```
{
    “@context”: {“@vocab”: “http://schema.sogou.com” “kg”: “http://kg.sogou.com”
    }
    “@id”: “4962641”, “@type”: [“Person”], “name”: “Dehua Liu”, “birthDate”: “1961–09–27”,
    “hasOccupation” [“Singer”, “Actor”] “sogouBaikeUrl”: “https://baike.sogou.com/v4962641.htm”
}
```

The schema http://schema.sogou.com used in Sogou knowledge base is compatible with http://schema.org. Currently, we maintain only one knowledge graph at http://kg.sogou.com while the framework can support more knowledge graphs by setting different values.

3.3 Merging

The merging section is the entrance of KG storage which is a distributed database storing the whole knowledge graph. Any operations aiming to change the KG database including adding new data, updating or deleting data have to be transformed into unit operations following a pre-defined interface (including "add", "update" and "delete") in the merging section. All unit operations are executed with logs which can be used to roll back to any historical version.

For adding entities, the merging section checks whether the entity already exists in the KG database. If the entity to be added is found in database, the old entity's property value will be updated to the value of added entity's same properties. Otherwise, the entity will be added into the database as a new entity. To distinguish the entities with the same name, we develop a heuristic model that also compares the entities' property values. For updating and deleting data, the @id property is required and the operation will be executed to the entities with given ids.

3.4 Inference

As an additional way to supply data, the inference section infers new relationships of entities based on the existing relations. For example, when we know A is B's son, we could infer a new relation that B is A's father. In the construction framework, the inference is conducted on the whole data that are dumped from KG database and the inference result is added back to the KG through the merging part. Currently, all of our inference models are rule-based. While neural network based inference methods (such as TransE and TransR) can infer more potential relations, the accuracy of these inference models' result is not good enough to be applied to products.

3.5 Knowledge Graph Storage

The Sogou knowledge graph storage is developed on top of SogouQdb which is an open source search engine. Figure 2 gives an overview of the architecture of the KG storage. SogouQdb is used as a distributed database to store data and provide search services. KG Storage Service wraps up SogouQdb to provide

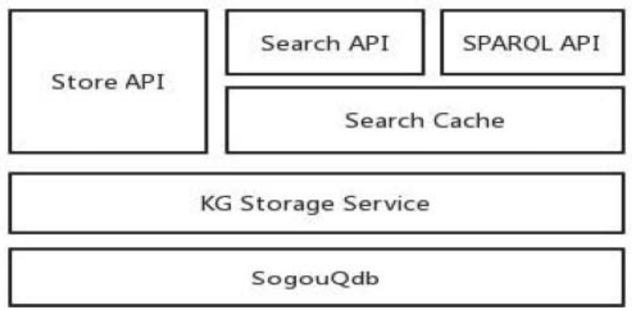


Figure 2. Overview of Sogou knowledge graph storage architecture.

storing and querying APIs that are more proper for applications of knowledge graph based cases. In practice, we find the querying requests are much more than storing requests and cost more computation resources. To reduce cost and improve querying speed, a cache layer is added between querying API and the KG storage service.

Compared with graph databases such as Neo4j and OrientDB which is commonly used in knowledge graph storage, using SogouQdb has more advantages on querying speed, scalability and more engineering optimizations. One disadvantage of SogouQdb is that it does not natively support knowledge graph query languages such as SPARQL. To solve this problem, we introduce the KG storage service to parse SPARQL to SogouQdb's APIs. Another disadvantage is that SogouQdb is relatively inefficient for conducting data inference. To solve this problem, we separate the inference part from KG storage and conduct the inference on HDFS using Spark. The data to be inferred are dumped from SogouQdb using Qdb-Hadoop tools.

4. APPLICATION

4.1 Entity Linking

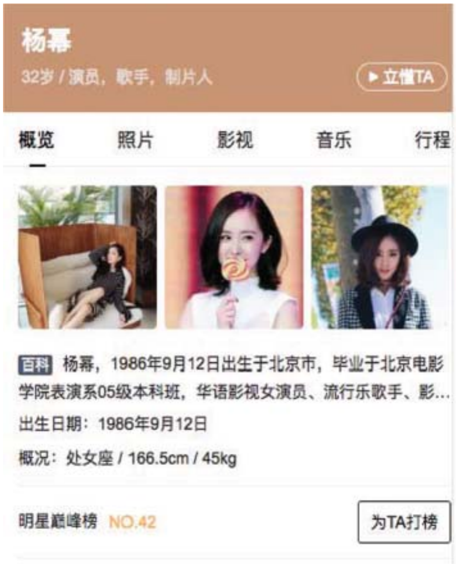
The entity linking task identifies the character string representing the entity from the natural language text and maps it to a specific entity in the knowledge base. For example, Wiki editors manually add hyperlinks to phrases representing entities in the text to the corresponding Wikipedia pages. This phrase with Wiki internal hyperlinks is called Anchor Text. Traditional entity linking method is based on feature engineering. This kind of method calculates the link matching degree through the features between the candidate entity and its context. Features usually include prior information of entities, contextual semantic features, and features associated with entities. Commonly used models include Ranking SVM [9], CRF [10] and S-MART [11]. With the development of neural networks, feature learning is gradually replacing the original method based on feature engineering. This kind of method calculates the context representation of the entity phrase and the representation of the candidate entity through a specific neural network. The matching score is defined as the similarity between vectors. The entity linking models based on deep learning include [12, 13, 14, 15]. In addition, knowledge graph embedding is also applied to entity linking tasks. The vector representation of each entity is learned through a large number of knowledge base triplets as training data, so that similar entities have similar vector representations. The methods of vector learning based on knowledge base include [12, 16, 17, 18, 19].

The focus of entity linking is to find the correct entity from multiple candidates and eliminate ambiguity. For example, "Li Na" has multiple possible candidate entities, which may represent a tennis star, pop singer, football baby of Sogou, or even a movie with the same name. In the absence of context information, it is difficult to link entities accurately. A well-designed entity linking service needs to consider many factors, including the prior knowledge of the entity itself, the matching degree between the entity and the phrase, and the fit degree between the context in which the entity and the phrase are located.

In Sogou, the entity linking problem is treated as a ranking problem. We take into consideration the entity prior, the similarity between the entity description and the context, and the coherence between entities and entities in the same paragraph. Based on the knowledge graph of Sogou, we have developed a set of entities linking APIs, which provides short text linking service, long text linking service and table linking service. These services link the entities contained in the text to the Sogou Knowledge Graph.

The short text entity linking service is mainly used for entity linking of query text in our search engine. After entity linking, the structural information in the knowledge graph related to the entities is shown to the user, along with illustrations and pictures (Figure 3). At the same time, based on the type of entities and the relationship between entities in the knowledge graph, recommendations of relevant entities are given (Figure 4). These richer results make it quicker for users to obtain what they want and what they are interested in. Also, entity linking is the basis for automatic question answering, especially for the task of knowledge-based question answering. The existing entities in the question need to be accurately linked, to limit the scope of semantic search.

Long text entity service is mainly used for Anchor Text generation in Web pages, as shown in Figure 5. In order to help readers quickly access the introduction information of entities in Sogou Encyclopedia's pages, these entities contain hyperlinks to their own pages, i.e., Anchor Text. Our automated entity linking service greatly improves the manual editing efficiency. In addition, the long text entity linking service is also applied to Sogou's news feed with personalized recommendations. In combination with the entity linking process, similar or related entities are extended in the knowledge graph. Thus we can provide personalized recommendations, along with very interpretable reasons.



百科 杨幂，1986年9月12日出生于北京市，毕业于北京电影学院表演系05级本科班，华语影视女演员、流行乐歌手、影...

出生日期：1986年9月12日

概况：处女座 / 166.5cm / 45kg

Figure 3. After query entity linking, entity-related information and pictures are shown in search results.

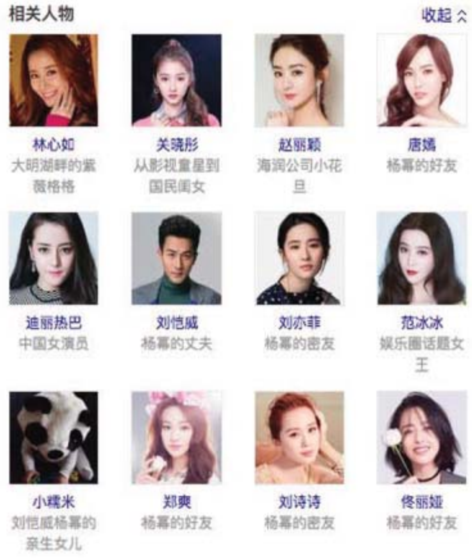
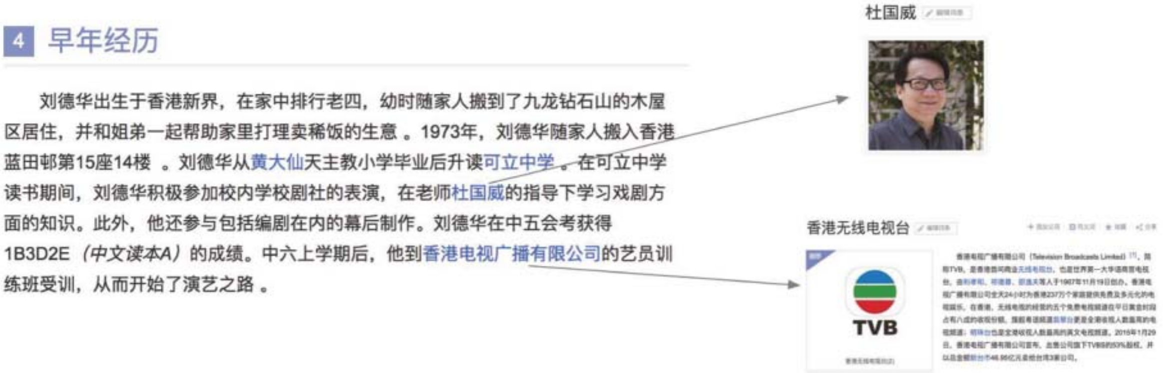


Figure 4. Search engine also gives related entity recommendations, due to entity relations in our KG.



4 早年经历

刘德华出生于香港新界，在家中排行老四，幼时随家人搬到了九龙钻石山的木屋区居住，并和姐弟一起帮助家里打理卖稀饭的生意 。1973年，刘德华随家人搬入香港蓝田邨第15座14楼 。刘德华从黄大仙天主教小学毕业后升读可立中学 。在可立中学读书期间，刘德华积极参加校内学校剧社的表演，在老师杜国威的指导下学习戏剧方面的知识。此外，他还参与包括编剧在内的幕后制作。刘德华在中五会考获得1B3D2E *（中文读本A）*的成绩。中六上学期后，他到香港电视广播有限公司的艺员训练班受训，从而开始了演艺之路 。

杜国威

香港无线电视台

香港电视广播有限公司（Television Broadcasts Limited）[1]，简称TVB，是香港首间商业无线电视台，也是世界第一大华语商营电视台，由利孝和、祁德尊、邵逸夫等人于1967年11月19日创办。香港电视广播有限公司全天24小时为香港237万个家庭提供免费及多元化的电视娱乐。在香港，无线电视的经营的五个免费电视频道在平日黄金时段占有八成的收视份额，旗舰电视频道翡翠台更是全港收视人数最高的电视频道；明珠台也是全港收视人数最高的英文电视频道。2015年1月29日，香港电视广播有限公司宣布，出售公司旗下TVBS的53%股权，并以总金额新台币46.95亿元卖给台湾3家公司。

Figure 5. Long text entity service used for Anchor Text generation.

Table entity linking is also used to generate entity Anchor Text in tables online, such as entities in tables of Sogou Encyclopedia. Meanwhile, tables provide rich entity type information, entity relationship information, etc. After entity linking, these tables can also supply a large amount of high confidence triplet information to our knowledge graph.

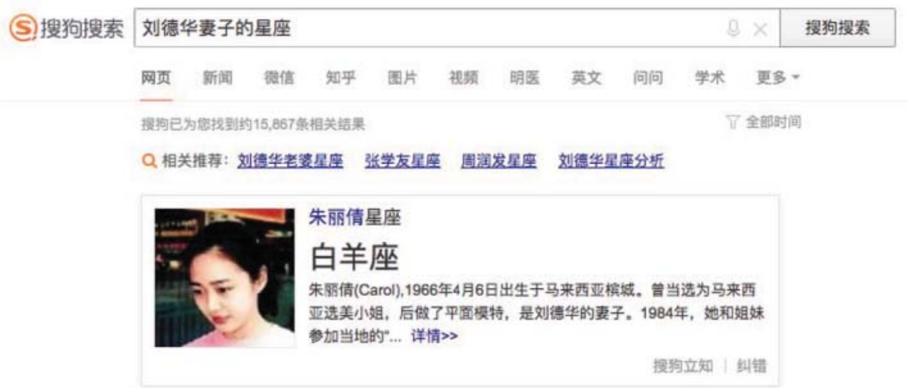
4.2 Knowledge-Based Question Answering

Knowledge graph usually comes with a descriptive language, such as MQL provided by Freebase, SPARQL formulated by W3C, and CycL provided by Cyc. However, for ordinary users, this structured query syntax has a high usage threshold. A knowledge-based question answering system uses natural language as interface to provide a more friendly way for knowledge querying. On the one hand, natural language has very strong expressive power. On the other hand, this method does not require users to receive any professional training. Due to its broad application prospect, knowledge-base question answering (KBQA) has become a research hot-spot in both academia and industry.

For question understanding, we focus on the automatic question answering task based on a knowledge graph. The task is to find one or more corresponding answer entities from the knowledge graph for questions describing objective facts. For a question that contains only simple semantics, the process of automatic question answering is equivalent to converting the question into a fact triplet on the knowledge base. However, the problems raised by human beings are not always presented in simple forms. More restrictions will be added to them. For example, there are multiple entities and types related to the answer in the question. In complex semantic scenarios, the KBQA has the following challenges: 1) How to find multiple relationships from questions and combine them into a candidate semantic structure; 2) How to calculate the matching degree between natural language questions and complex semantic structures.

Commonly used methods are based on semantic parsing or ranking. The method based on semantic parsing is to convert the question into a formal query statement of a certain standard knowledge base, i.e., finding the optimal (question, semantic query) pair instead of a simple answer entity. Related work includes the generation of semantic parsing trees using the Combinatory Categorial Grammar(CCG) [20, 21, 22], and λ-DCS [23, 24, 25]. Typical application projects include ATIS [26] in the air travel information question and answer system, CLANG [27] in the robot soccer game, GeoQuery [28] in the US geographic knowledge question and answer system, and an open source question and answer system SEMPRE [23]. The ranking method does not need formal representation of questions, but directly ranks candidate entities or answers in the knowledge base. This kind of method follows the representation-comparison framework, in which the traditional feature-based engineering methods include [29] and deep learning based methods include [30, 31, 32].

We have implemented a KBQA system and integrated it into Sogou Search Engine (Figure 6) and Sogou's dialogue service. Sogou's KBQA relies mainly on the combination of manual templates and models. By using templates, the user's query is directly converted into structural KB query. In the model approach, the entities in the query are first linked to the knowledge graph, and then a subgraph is constructed with the entity as the center. The final answer is to sort the results by using the nodes and edges in the subgraph as candidate paths and answers.



刘德华妻子的星座

搜狗已为您找到约15,867条相关结果

相关推荐：刘德华老婆星座 张学友星座 周润发星座 刘德华星座分析

朱丽倩星座

白羊座

朱丽倩(Carol),1966年4月6日出生于马来西亚槟城。曾当选为马来西亚选美小姐，后做了平面模特，是刘德华的妻子。1984年，她和姐妹参加当地的"... 详情>>

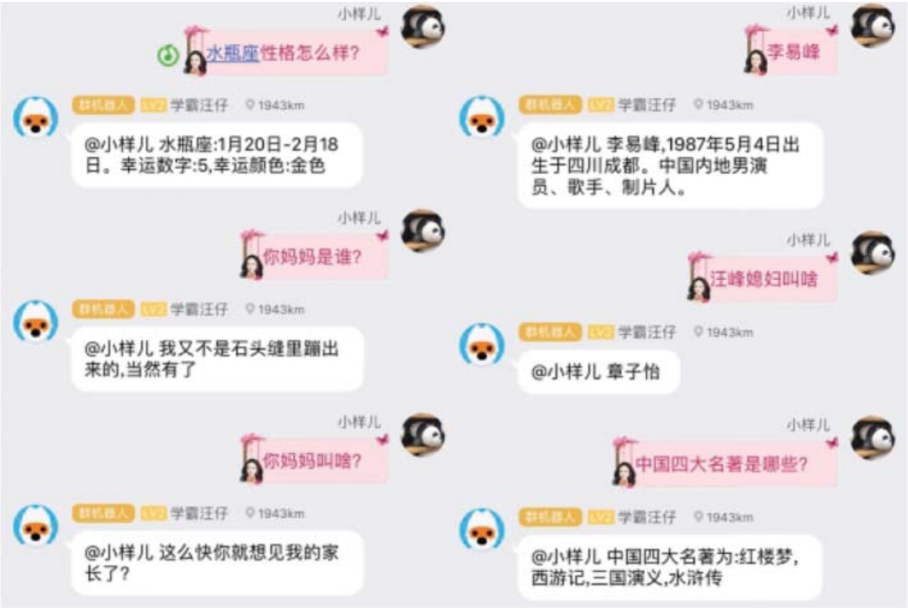
Figure 6. KBQA in Sogou Search Engine, which answers the query directly in the first search result.

4.3 Knowledge-Based Dialogue System

Knowledge based dialogue is a more natural and friendly knowledge service, which can satisfy users' needs and complete specific knowledge acquisition tasks through multiple rounds of human-agent interaction. The latest development in dialogue systems is based on deep learning techniques, using the encoder-decoder model to train the entire system. Related work includes [33, 34, 35, 36]. Combining an external knowledge base is a way to bridge the gap between the dialogue system and humans. Using memory network, [37, 38] have achieved good results in the open domain dialogue. Combining words in the generation process with common words in the knowledge base, [39] produces natural and correct answers. [40] uses Twitter's LDA model to get the input topic, and add the topic information and input representation to the joint attention module to generate a topic-related response. [41] classifies each discourse in the conversation into a field and uses it to generate the domain and content of the next discourse. Dialogue system also needs personality and emotion to look more like humans. [42] applies emotion embedding into the generative model. [43, 44] both consider the user's information in creating a more realistic chat bot.

With the large-scale growth of knowledge graph resources and the rapid development of machine learning models, dialogue systems are gradually moving from limited areas to open areas. Sogou Wang Zai Robot is an automatic question-answering robot developed by Sogou, as shown in Figure 7. It combines Sogou's knowledge graph, Sogou's dialogue technology and Sogou's intelligent voice technology to provide accurate answers in daily conversations.

Dialogue generation based on knowledge graphs is a key technology in knowledge-based dialogues. Traditional KBQA provides only accurate answers to all questions. For example, when asked "How tall is Andy Lau?", the system only returns "174 cm". However, merely providing this kind of answer is not a friendly interactive way. Users prefer to receive "The height of Andy Lau, actor of Hong Kong, China, is 174 cm". This way provides more background information related to the answer (for example, actor of Hong Kong, China). In addition, this complete natural language sentence can better support the follow-up tasks such as answer verification and speech synthesis. In order to generate natural language answers, we use the encoder-decoder framework. Copy and retrieval mechanism is also introduced for complex questions that require facts in the knowledge graph. Different types of words are obtained from different sources by using different semantic unit acquisition methods such as copy, retrieval or prediction. Thus natural answers are generated for complex questions.



小样儿

水瓶座性格怎么样?

群机器人 LV2 学霸汪仔 1943km

@小样儿 水瓶座:1月20日-2月18日。幸运数字:5,幸运颜色:金色

小样儿

你妈妈是谁?

群机器人 LV2 学霸汪仔 1943km

@小样儿 我又不是石头缝里蹦出来的,当然有了

小样儿

你妈妈叫啥?

群机器人 LV2 学霸汪仔 1943km

@小样儿 这么快你就想见我的家长了?

小样儿

李易峰

群机器人 LV2 学霸汪仔 1943km

@小样儿 李易峰,1987年5月4日出生于四川成都。中国内地男演员、歌手、制片人。

小样儿

汪峰媳妇叫啥

群机器人 LV2 学霸汪仔 1943km

@小样儿 章子怡

小样儿

中国四大名著是哪些?

群机器人 LV2 学霸汪仔 1943km

@小样儿 中国四大名著为:红楼梦,西游记,三国演义,水浒传

Figure 7. A user is having a dialogue with our conversational assistant, Wang Zai.

Another problem that needs to be solved in the dialogue agent is the consistency of the dialogue, i.e., the stability of the agent's portrait. It also requires the integration of external knowledge, e.g., personal information in Table 1. Although the agent is a robot, it needs to have a unified personality. Its gender, age, native place and hobbies should always be the same. When asked "where were you born?" or "Are you from Beijing?", the answer will always be consistent. We model Sogou Wang Zai's information and import it into an encoder-decoder model in embeddings. Thus when the question is related to personal information, it will generate responses from vectors of the identity information, which achieves good consistency effect.

Table 1. Wang Zai's personal information for more consistent question answering.

Profile key	Profile value
Name	Wang Zai
Age	Three
Gender	Boy
Hobbies	Cartoon
Speciality	Piano

5. CONCLUSION

In this paper, we propose a novel solution that is used in Sogou Inc. in building knowledge graphs on top of a distributed search engine, specifically, SogouQdb. Our solution supplies SogouQdb by introducing data inference and graph-based query engine which makes the solution compatible with commonly used knowledge graph applications. Besides, benefited from SogouQdb, the Sogou knowledge graph can be easily scaled to store petabytes of data. We also introduce three applications of a knowledge graph in Sogou Inc.: entity detection and linking, knowledge-based question answering and knowledge-based dialogue system which have been used as the Web search products to make knowledge acquisition more efficient.

AUTHOR CONTRIBUTIONS

All of the authors contributed equally to the work. J. Xu (xujingfang@sogou-inc.com) is the leader of Sogou Knowledge Graph, who drew the blueprint of the whole system. Q. Zhang (qizhang@sogou-inc.com) brought valuable insights and information to the construction and applications of the knowledge graph. P. Wang (wangpeilu@sogou-inc.com) and H. Jiang (jianghao216568@sogou-inc.com) mainly drafted the paper, while P. Wang summarized the construction part and H. Jiang summarized the application part. All of the authors have made meaningful and valuable contributions in revising and proofreading the resulting manuscript.

REFERENCES

[1] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, & J. Taylor: Freebase: A collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, 2008, pp. 1247–1250. doi: 10.1145/1376616.1376746.

[2] D. Vrandečić, & M. Krötzsch. Wikidata: A free collaborative knowledgebase. Communications of the ACM 57(10)(2014), 78–85. doi: 10.1145/2629489.

[3] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P.N. Mendes, S. Hellmann, … & S. Auer: DBpedia—A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web 6(2)(2015), 167–195. doi: 10.3233/SW-140134.

[4] F.M. Suchanek, G. Kasneci, & G. Weikum. YAGO: A core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web, 2017, pp. 697–706. Available at: http://www2007.wwwconference.org/papers/paper391.pdf.

[5] R. Speer, & C. Havasi. Representing general relational knowledge in ConceptNet 5. In: Proceedings of the 8th Conference on Language Resources and Evaluation (LREC'12), 2012, pp. 3679–3686.

[6] B. Xu, Y. Xu, J. Liang, C. Xie, B. Liang, W. Cui, & Y. Xiao. CN-DBpedia: A never-ending Chinese knowledge extraction system. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, 2017, pp. 428–438. doi: 10.1007/978-3-319-60045-1_44.

[7] T. Yang, D. Du, & F. Zhang. The tai system for trilingual entity discovery and linking track in TAC-KBP2017. Available at: https://pdfs.semanticscholar.org/15f3/7711fe63a80dcb09f85ce597ddbc712bd767.pdf.

[8] H. Ji, X. Pan, B. Zhang, J. Nothman, J. Mayfield, P. McNamee, & C. Costello. Overview of TAC-KBP2017 13 languages entity discovery and linking. Available at: http://nlp.cs.rpi.edu/paper/kbp2017.pdf.

[9] T. Joachims. Optimizing search engines using clickthrough data. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 133–142. doi: 10.1145/775047.775067.

[10] G. Luo, X. Huang, C.Y. Lin, & Z. Nie. Joint named entity recognition and disambiguation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 879–888. Available at: https://www.aclweb.org/anthology/D15-1104.

[11] Y. Yang, & M.W. Chang. S-mart: Novel tree-based structured learning algorithms applied to tweet entity linking. arXiv preprint. arXiv:1609.08075, 2016.

[12] W. Fang, J. Zhang, D. Wang, Z. Chen, & M. Li. Entity disambiguation by knowledge and text jointly embedding. In: Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, 2016, pp. 260–269. Available at: https://www.aclweb.org/anthology/K16-1026.

[13] M. Francis-Landau, G. Durrett, & D. Klein. Capturing semantic similarity for entity linking with convolutional neural networks. arXiv preprint. arXiv:1604.00734, 2016.

[14] N. Gupta, S. Singh, & D. Roth. Entity linking via joint encoding of types, descriptions, and context. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2681–2690. doi: 10.18653/v1/D17-1284.

[15] Y. Sun, L. Lin, D. Tang, N. Yang, Z. Ji, & X. Wang. Modeling mention, context and entity with neural networks for entity disambiguation. In: Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15), 2015, pp. 1333–1339. Available at: https://aaai.org/ocs/index.php/IJCAI/IJCAI15/paper/view/11048/10848.

[16] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, & O. Yakhnenko. Translating embeddings for modeling multi-relational data. In: Advances in Neural Information Processing Systems 26 (NIPS 2013) 2013, pp. 1–9. Available at: http://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data.pdf.

[17] A. Bordes, J. Weston, R. Collobert, & Y. Bengio. Learning structured embeddings of knowledge bases. In: Proceedings of the 25th AAAI Conference on Artificial Intelligence, 2011, pp. 301–306. Available at: https://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/view/3659/3898.

[18] T. Mikolov, Q.V. Le, & I. Sutskever. Exploiting similarities among languages for machine translation. arXiv preprint. arXiv:1309.4168, 2013.

[19] Z. Wang, J. Zhang, J. Feng, & Z. Chen. Knowledge graph embedding by translating on hyperplanes. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2014, pp. 1112–1119. Available at: https://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8531/8546.

[20] Q. Cai, & A. Yates. Large-scale semantic parsing via schema matching and lexicon extension. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013, pp. 423–433. Available at: https://www.aclweb.org/anthology/P13-1042.

[21] T. Kwiatkowski, L. Zettlemoyer, S. Goldwater, & M. Steedman. Inducing probabilistic CCG grammars from logical form with higher-order unification. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, pp. 1223–1233. Available at: https://www.aclweb.org/anthology/D10-1119.

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

[23] J. Berant, A. Chou, R. Frostig, & P. Liang. Semantic parsing on Freebase from question-answer pairs. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1533–1544. Available at: https://www.aclweb.org/anthology/D13-1160.

[24] J. Berant, & P. Liang. Semantic parsing via paraphrasing. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, pp. 1415–1425. Available at: https://www.aclweb.org/anthology/P14-1133.

[25] P. Liang. Lambda dependency-based compositional semantics. arXiv preprint. arXiv:1309.4408, 2013.

[26] P.J. Price. Evaluation of spoken language systems: The atis domain. In: HLT '90 Proceedings of the Workshop on Speech and Natural Language, 1990, pp. 91–95. doi: 10.3115/116580.116612.

[27] M. Chen, K. Dorer, E. Foroughi, F. Heintz, Z.X. Huang, S. Kapetanakis, K. Kostiadis, ... & X. Yin. RoboCup soccer server: For soccer server version 7.07 and later. (February 11, 2003). Available at: https://rcsoccersim.github.io/rcssserver-manual-20030211.pdf.

[28] J.M. Zelle, & R.J. Mooney. Learning to parse database queries using inductive logic programming. In: Proceedings of the National Conference on Artificial Intelligence, 1996, pp. 1050–1055. Available at: http://aaai.org/Papers/AAAI/1996/AAAI96-156.pdf.

[29] X. Yao, & B. van Durme. Information extraction over structured data: Question answering with Freebase. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014, pp. 956–966.

[30] D. Lukovnikov, A. Fischer, J. Lehmann, & S. Auer. Neural network-based question answering over knowledge graphs on word and character level. In: Proceedings of the 26th international conference on World Wide Web, 2017, pp. 1211–1220. doi: 10.1145/3038912.3052675.

[31] W. Yin, M. Yu, B. Xiang, B. Zhou, & H. Schütze. Simple question answering by attentive convolutional neural network. arXiv preprint. arXiv:1606.03391, 2016.

[32] M. Yu, W. Yin, K.S. Hasan, C.D. Santos, B. Xiang, & B. Zhou. Improved neural relation detection for knowledge base question answering. arXiv preprint. arXiv:1704.06194, 2017.

[33] N. Asghar, P. Poupart, X. Jiang, & H. Li. Deep active learning for dialogue generation. arXiv preprint. arXiv:1612.03929, 2016.

[34] S. Lee, & M. Eskenazi. Recipe for building robust spoken dialog state trackers: Dialog state tracking challenge system description. In: Proceedings of the SIGDIAL 2013 Conference, 2013, pp. 414–422. Available at: https://sigdial.org/files/workshops/conference14/proceedings/pdf/SIGDIAL66.pdf.

[35] G. Tur, L. Deng, D. Hakkani-Tür, & X. He. Towards deeper understanding: Deep convex networks for semantic utterance classification. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 5045–5048. doi: 10.1109/ICASSP.2012.6289054.

[36] Y. Wu, W. Wu, Z. Li, & M. Zhou. Topic augmented neural network for short text conversation. arXiv preprint. arXiv: 1605.00090v2, 2016.

[37] E. Denton, S. Chintala, A. Szlam, & R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In: Advances in Neural Information Processing Systems, 2015, pp. 1486–1494. Available at: http://papers.nips.cc/paper/5773-deep-generative-image-models-using-a-laplacian-pyramid-of-adversarial-networks.pdf.

[38] K. Sohn, H. Lee, & X. Yan. Learning structured output representation using deep conditional generative models. In: Advances in Neural Information Processing Systems, 2015, pp. 3483–3491. Available at: http://papers.nips.cc/paper/5775-learning-structured-output-representation-using-deep-conditional-generative-models.pdf.

[39] J.D. Williams, & G. Zweig. End-to-end LSTM-based dialog control optimized with supervised and reinforcement learning. arXiv preprint. arXiv:1606.01269, 2016.

[40] T.H. Wen, M. Gasic, N. Mrksic, P.H. Su, D. Vandyke, & S. Young. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. arXiv preprint. arXiv:1508.01745, 2015.

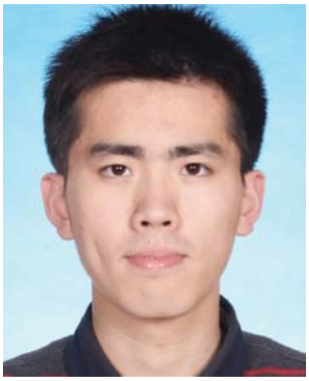
[41] S.R. Bowman, L. Vilnis, O. Vinyals, A.M. Dai, R. Jozefowicz, & S. Bengio. Generating sentences from a continuous space. arXiv preprint. arXiv:1511.06349, 2015.

[42] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, & E. Hovy. Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1480–1489.

[43] C. Kamm. User interfaces for voice applications. Proceedings of the National Academy of Sciences 92(22), 1995, 10031–10037. doi: 10.1073/pnas.92.22.10031.

[44] G. Mesnil, X. He, L. Deng, & Y. Bengio. Investig ation of recurrent-neuralnetwork architectures and learning methods for spoken language understanding. In: Interspeech, 2013, pp. 3771–3775.

AUTHOR BIOGRAPHY



Peilu Wang is a researcher at Sogou. He received both his Master's degree and Bachelor's degree from Shanghai Jiao Tong University. He currently works on explainable entity recommendation and his main research interests include natural language processing, knowledge graph and information retrieval.

Hao Jiang is a researcher at Sogou. He currently works on question answering and entity linking. He received his Master's degree from Nanjing University in 2014, and Bachelor's degree from Nanjing University in 2011. His main research interests include natural language processing, knowledge graph and chatbot.

Jingfang Xu is Vice President at Sogou. She received her PhD degree from Tsinghua University. She has rich experience in information retrieval, data mining and natural language processing. As the head of the search engine division, Dr. Xu has made outstanding achievements in Web search, question answering, computer vision, machine translation and other related fields.

Qi Zhang is the chief researcher at Sogou. His major interests include nature language processing and information retrieval. He has published more than 70 papers in the related areas. He received his Bachelor's degree in Computer Science and Technology from Shandong Univeristy in 2003 and PhD degree in Computer Science from Fudan Univerisity in 2009.