Knowledge Graph Construction and Applications for Web Search and Beyond: Postprint

Authors: Wang, Peilu, Jiang, Hao, Xu, Jingfang, Zhang, Qi, Zhang, Qi

Date: 2022-11-29T00:00:00+00:00

Abstract

Knowledge graph (KG) has played an important role in enhancing the performance of many intelligent systems. In this paper, we present the solution for constructing a large-scale multi-source knowledge graph from scratch at Sogou Inc., including its architecture, technical implementation and applications. Unlike prior work that constructs knowledge graphs using graph databases, we construct the knowledge graph atop SogouQdb, a distributed search engine developed by Sogou Web Search Department, which can be readily scaled to support petabytes of data. To complement the search engine, we additionally introduce a suite of models to support inference and graph-based querying. Currently, the Sogou knowledge graph comprises data collected from 136 distinct websites, continuously updated, consisting of 54 million entities and over 600 million entity links. We further introduce three applications of the knowledge graph at Sogou Inc.: entity detection and linking, knowledge-based question answering, and knowledge-based dialogue systems. These applications have been deployed in Web search products to assist users in acquiring information more efficiently.

Full Text

Preamble

DATA PAPER
ChinaXiv Cooperative Journal

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 graphs (KGs) have played an important role in enhancing the performance of many intelligent systems. In this paper, we introduce Sogou Inc.' s solution for building a large-scale, multi-source knowledge graph from scratch, including its architecture, technical implementation, and applications. Unlike previous approaches that rely on graph databases, we build our knowledge graph on top of SogouQdb, a distributed search engine developed by Sogou' s Web Search Department that can easily scale 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 Sogou knowledge graph contains data collected from 136 different websites that are constantly updated, consisting of 54 million entities and over 600 million entity links. We also introduce three applications of the knowledge graph at Sogou Inc.: entity detection and linking, knowledge-based question answering, and knowledge-based dialogue systems. These applications have been integrated into Web search products to help users acquire information more efficiently.

1. INTRODUCTION

A knowledge graph (KG) is a specialized database that integrates information into an ontology. As an effective method for storing and searching knowledge, knowledge graphs have been applied in many intelligent systems and have attracted significant 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 fully meets the application requirements of Sogou Inc. The main challenges are listed below:

Lack of data: Although the largest published knowledge graph (Wikidata) reportedly contains millions of entities and billions of triples, most of its data are extracted from Wikipedia and remain insufficient for Web search applications such as general-purpose question answering and recommendations. For example, none of the existing knowledge graphs contains information about the latest Chinese songs, which can only be obtained from specific websites.

Uncertainty of scalability: None of the existing works explicitly reports their system' s capability to handle large-scale data or discusses how the knowledge graph could be expanded on a server cluster. This problem might not be critical for academic research since even the largest knowledge graph can still be stored on a single server with a large hard disk drive. However, for search engines, the potential data requirements of a knowledge graph are much larger, making distributed storage unavoidable.

To address these challenges, we propose a novel solution for building a large-scale knowledge graph. We use SogouQdb—a distributed search engine developed by Sogou' s Web Search Department for internal use—as the core storage engine to achieve scalability, and we develop a series of models to provide inference and graph-based querying functions that make the system compatible with other knowledge graph applications. The inference is conducted on HDFS with Spark, enabling the inference procedure to handle big data. The Sogou knowledge graph has been built using this solution and deployed to support online products. Currently, it consists of 54 million entities and over 600 million entity links extracted from 136 different websites that are constantly updated.

We also introduce three applications of knowledge graphs at Sogou Inc.: entity detection and linking, knowledge-based question answering, and knowledge-based dialogue systems. These applications have been used as infrastructure services 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 review related work on widely known published knowledge graphs. In Section 3, we elaborate on our solution for constructing a knowledge graph from scratch. Section 4 presents applications of knowledge graphs, particularly at Sogou Inc. Finally, we conclude in Section 5.

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

2. RELATED WORK

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

Freebase [**1**] was published as an open shared database in 2007 and was shut down in 2016 after all its data were transferred to Wikidata. Freebase data were collected from Wikipedia , NNDB , Fashion Model Directory , and MusicBrainz , and were also contributed by its users . Freebase contains more than 1.9 billion triples , 4,000 types, and 7,000 properties [1].

Wikidata [**2**] was first published by Wikidata in 2012 and has been publicly maintained since then. Wikidata contains more than 55 million entities, with data primarily coming from its Wikipedia sister projects including Wikipedia, Wikivoyage, Wikisource, and other websites .

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

YAGO [4] is an open-source semantic knowledge graph derived from Wikipedia, WordNet, and GeoNames. YAGO contains more than 10 million entities and 120 million facts about entities .

ConceptNet [5] originated from the Open Mind Common Sense project launched in 1999 and has grown into an open multilingual knowledge graph containing 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 contains more than 16.8 million entities and 223 million entity links .

---

3. CONSTRUCTION

An overview of the Sogou knowledge graph construction framework is shown in Figure 1 [Figure 1: see original paper]. Data are collected from various websites that allow their data to be downloaded or crawled, such as Wikipedia and SogouBaike. The extracted data are stored in a distributed database in the form of JSON-LD (JavaScript Object Notation for Linked Data), a commonly used concrete RDF syntax. As an additional method for supplying data, we introduce an inference model that infers new relationships between entities. To search and browse the knowledge graph, we developed a SPARQL query engine that provides RESTful API services. To support search engine products like question answering and recommendation, the knowledge graph data are processed to adapt to the data format required by specific tasks. In this section, we introduce each component of the construction framework.

3.1 Data Extraction

The role of data extraction is to transform data from various input sources into a predefined format. Specifically, the input and output of data extraction are defined as follows:

Input: Data downloaded or crawled from the Internet, such as Web pages, XML data, or JSON data obtained via APIs. While the input data mostly comprise free text, many sources 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 records the knowledge information extracted from the input data.

Data extraction operations can be classified into two categories: **Structured data extraction** deals only with input data containing structured information –specifically, data with recognizable markup. **Free text extraction** detects entity mentions from plain text and extracts property information about specific entities.

3.1.1 Structured Data Extraction

Since structured information has recognizable markups, we use rule-based methods to build extractors. The extractors first parse Web pages into a unified DOM tree, then locate target information according to manually written rules and save the extracted data in JSON-LD format. For each website, we build specialized extractors to handle its data, enabling independent updates for different websites. As of March 2019, the 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 combines a series of sub-tasks, including extracting named entity mentions from plain text, linking mentions to entities in the knowledge graph, and extracting entity properties or relationships between extracted entities. Since training a model to handle all entity types is quite time-consuming, we currently focus on limited entity types: 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, with feature and parameter selection following the work of [7], which achieved the best performance in the TAC KBP 2017 competition [8]. The training data are constructed from SogouBaike and Wikipedia Web pages containing anchor markups. More details about the model and training data can be found in Section 4.1.

3.2 Normalization

This component normalizes property values of extracted entities and maps entity classes and properties to terms in the Sogou knowledge graph ontology. Additionally, data types of properties are specified to ensure high-quality processed data. The input and output of this component are defined as follows:

Input: Output from data extraction–structured data in the form of JSON-LD.

Output: Structured data in the form of JSON-LD with normalized property names and property values.

The property value types follow the definition of the 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 the Sogou knowledge base is compatible with http://schema.org. Currently, we maintain only one knowledge graph at http://kg.sogou.com, though the framework can support multiple knowledge graphs by setting different values.

3.3 Merging

The merging component serves as the entry point to the KG storage, which is a distributed database storing the entire knowledge graph. Any operation that modifies the KG database—including adding new data, updating, or deleting data—must be transformed into unit operations following a predefined interface (including "add", "update", and "delete") in the merging component. All unit operations are executed with logs that can be used to roll back to any historical version.

For adding entities, the merging component checks whether the entity already exists in the KG database. If the entity to be added is found, the old entity' s property values are updated to match those of the added entity. Otherwise, the entity is added to the database as a new entity. To distinguish 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 is executed on entities with the given IDs.

3.4 Inference

As an additional method for supplying data, the inference component infers new relationships between entities based on existing relations. For example, knowing that A is B' s son allows us to infer that B is A' s father. In the construction framework, inference is conducted on the entire dataset dumped from the KG database, and the inference results are added back to the KG through the merging component. Currently, all 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 models' results is not yet sufficient for production use.

3.5 Knowledge Graph Storage

The Sogou knowledge graph storage is developed on top of SogouQdb, an open-source search engine. Figure 2 [Figure 2: see original paper] provides an overview of the KG storage architecture. SogouQdb is used as a distributed database to store data and provide search services. The KG Storage Service wraps SogouQdb to provide storing and querying APIs more suitable for knowledge graph-based applications. In practice, we find that querying requests far outnumber storing requests and consume more computational resources. To reduce costs and improve querying speed, we add a cache layer between the querying API and the KG storage service.

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

---

4. APPLICATIONS

4.1 Entity Linking

The entity linking task identifies character strings representing entities from natural language text and maps them to specific entities in the knowledge base. For example, Wiki editors manually add hyperlinks to phrases representing entities in the text, linking them to corresponding Wikipedia pages. A phrase with Wikipedia internal hyperlinks is called Anchor Text. Traditional entity linking methods rely on feature engineering, calculating link matching degrees through features between candidate entities and their context. Features typically include prior entity information, contextual semantic features, and entity-associated features. Commonly used models include Ranking SVM [9], CRF [10], and S-MART [11]. With the development of neural networks, feature learning is gradually replacing feature engineering. These methods calculate context representations of entity phrases and representations of candidate entities through specific neural networks, defining matching scores as vector similarities. Deep learning-based entity linking models include [12, 13, 14, 15]. Additionally, knowledge graph embedding is applied to entity linking tasks, learning vector representations for each entity through large-scale knowledge base triplets as training data, so that similar entities have similar vector representations. Knowledge base-based vector learning methods include [12, 16, 17, 18, 19].

The key challenge in entity linking is finding the correct entity from multiple candidates and eliminating ambiguity. For example, "Li Na"has multiple possible candidate entities: a tennis star, a pop singer, a football baby from Sogou, or even a movie with the same name. Without context information, accurate entity linking is difficult. A well-designed entity linking service must consider many factors, including the entity' s prior knowledge, the matching degree between the entity and phrase, and the fit degree between the contexts of the entity and phrase.

At Sogou, we treat the entity linking problem as a ranking problem, considering entity priors, similarity between entity descriptions and context, and coherence between entities in the same paragraph. Based on the Sogou knowledge graph, we have developed a set of entity linking APIs that provide short text linking, long text linking, and table linking services. These services link entities contained in text to the Sogou Knowledge Graph.

The **short text entity linking service** is primarily used for entity linking of query text in our search engine. After entity linking, structural information from the knowledge graph related to the entities is displayed to users, along with illustrations and pictures (Figure 3 [Figure 3: see original paper]). Simultaneously, based on entity types and relationships in the knowledge graph, recommendations for relevant entities are provided (Figure 4 [Figure 4: see original paper]). These richer results enable users to obtain what they want and what interests them more quickly. Additionally, entity linking forms the basis for automatic question answering, particularly for knowledge-based question answering, as existing entities in questions must be accurately linked to limit the scope of semantic search.

The **long text entity linking service** is mainly used for Anchor Text generation in Web pages, as shown in Figure 5 [Figure 5: see original paper]. To help readers quickly access introductory information about entities in Sogou Encyclopedia pages, these entities contain hyperlinks to their own pages (i.e., Anchor Text). Our automated entity linking service greatly improves manual editing efficiency. Furthermore, the long text entity linking service is applied to Sogou' s news feed with personalized recommendations. Combined with the entity linking process, similar or related entities are extended in the knowledge graph, allowing us to provide personalized recommendations along with interpretable reasons.

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

4.2 Knowledge-Based Question Answering

Knowledge graphs typically come with descriptive languages, 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 an interface to provide a more user-friendly way for knowledge querying. On one hand, natural language has very strong expressive power. On the other hand, this method does not require users to have any professional training. Due to its broad application prospects, knowledge-base question answering (KBQA) has become a research hotspot 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 questions containing only simple semantics, the automatic question answering process is equivalent to converting the question into a fact triplet in the knowledge base. However, human-posed problems are not always presented in simple forms; more restrictions are added. For example, multiple entities and types related to the answer may appear in the question. In complex semantic scenarios, KBQA faces the following challenges: 1) How to find multiple relationships from questions and combine them into candidate semantic structures; 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. Semantic parsing-based methods convert questions into formal query statements of a standard knowledge base, i.e., finding the optimal (question, semantic query) pair instead of a simple answer entity. Related work includes generating semantic parsing trees using Combinatory Categorial Grammar (CCG) [20, 21, 22] and λ-DCS [23, 24, 25]. Typical application projects include ATIS [26] for air travel information question answering, CLANG [27] for robot soccer games, GeoQuery [28] for U.S. geographic knowledge question answering, and the open-source question answering system SEMPRE [23]. Ranking methods do not require formal question representation but directly rank candidate entities or answers in the knowledge base. These methods follow the representation-comparison framework, with traditional feature-based engineering methods including [29] and deep learning-based methods including [30, 31, 32].

We have implemented a KBQA system and integrated it into Sogou Search Engine (Figure 6 [Figure 6: see original paper]) and Sogou' s dialogue service. Sogou' s KBQA relies mainly on the combination of manual templates and models. Using templates, user queries are directly converted into structured KB queries. In the model-based approach, 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 obtained by sorting results using nodes and edges in the subgraph as candidate paths and answers.

4.3 Knowledge-Based Dialogue System

Knowledge-based dialogue is a more natural and friendly knowledge service that can satisfy users' needs and complete specific knowledge acquisition tasks through multiple rounds of human-agent interaction. The latest developments in dialogue systems are based on deep learning techniques, using encoder-decoder models 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 dialogue systems and humans. Using memory networks, [37, 38] have achieved good results in open-domain dialogue. By 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 obtain the input topic and adds topic information and input representation to a joint attention module to generate topic-related responses. [41] classifies each discourse in the conversation into a field and uses it to generate the domain and content of the next discourse. Dialogue systems also need personality and emotion to appear more human-like. [42] applies emotion embedding into the generative model, while [43, 44] both consider user information in creating more realistic chatbots.

With the large-scale growth of knowledge graph resources and the rapid development of machine learning models, dialogue systems are gradually moving from limited domains to open domains. Sogou Wang Zai Robot is an automatic question-answering robot developed by Sogou, as shown in Figure 7 [Figure 7: see original paper]. It combines Sogou' s knowledge graph, dialogue technology, and 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 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 approach. Users prefer to receive "The height of Andy Lau, actor of Hong Kong, China, is 174 cm." This approach provides more background information related to the answer (e.g., actor of Hong Kong, China). Additionally, this complete natural language sentence better supports follow-up tasks such as answer verification and speech synthesis. To generate natural language answers, we use the encoder-decoder framework. Copy and retrieval mechanisms are also introduced for complex questions requiring facts from the knowledge graph. Different types of words are obtained from different sources using different semantic unit acquisition methods such as copy, retrieval, or prediction, thus generating natural answers for complex questions.

Another problem that needs to be solved in dialogue agents is dialogue consistency—i.e., the stability of the agent' s persona. This also requires integration of external knowledge, such as 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 remain the same. When asked "Where were you born?" or "Are you from Beijing?" , the answer should always be consistent. We model Sogou Wang Zai' s information and import it into an encoder-decoder model via embeddings. Thus, when the question relates to personal information, the system generates responses from identity information vectors, achieving good consistency.

---

5. CONCLUSION

In this paper, we propose a novel solution used at Sogou Inc. for building knowledge graphs on top of a distributed search engine—specifically, SogouQdb. Our solution enhances SogouQdb by introducing data inference and a graph-based query engine, making it compatible with commonly used knowledge graph applications. Additionally, benefiting from SogouQdb, the Sogou knowledge graph can be easily scaled to store petabytes of data. We also introduce three applications of knowledge graphs at Sogou Inc.: entity detection and linking, knowledge-based question answering, and knowledge-based dialogue systems, which have been integrated into Web search products to make knowledge acquisition more efficient.

AUTHOR CONTRIBUTIONS

All authors contributed equally to this work. J. Xu (xujingfang@sogou-inc.com) is the leader of Sogou Knowledge Graph, who drew the blueprint for the entire system. Q. Zhang (qizhang@sogou-inc.com) provided valuable insights and information for the construction and applications of the knowledge graph. P. Wang (wangpeilu@sogou-inc.com) and H. Jiang (jianghao216568@sogou-inc.com) primarily drafted the paper, with P. Wang summarizing the construction portion and H. Jiang summarizing the application portion. All authors made meaningful and valuable contributions to revising and proofreading the final 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. Vrandecˇic´, & 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/15f3/7711fe63a80dcb09f85c

[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/D10-1148.

[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. Investigation of recurrent-neural-network 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 graphs, 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 graphs, and chatbots.

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 natural language processing and information retrieval. He has published more than 70 papers in related areas. He received his Bachelor' s degree in Computer Science and Technology from Shandong University in 2003 and PhD degree in Computer Science from Fudan University in 2009.

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

Profile key	Profile value
Gender	Wang Zai
Hobbies	Three Cartoon
Speciality	Piano

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv —Machine translation. Verify with original.