DKGBuilder: An Architecture for Building a Domain Knowledge Graph from Scratch

Yan Fan[1], Chengyu Wang[1], Guomin Zhou[2], and Xiaofeng He[1(✉)]

[1] Shanghai Key Laboratory of Trustworthy Computing,
School of Computer Science and Software Engineering,
East China Normal University, Shanghai, China
eileen940531@gmail.com, chywang2013@gmail.com, xfhe@sei.ecnu.edu.cn
[2] Department of Computer and Information Technology,
Zhejiang Police College, Hangzhou, China
zhouguomin@zjjcxy.cn

Abstract. In recent years, we have witnessed the technical advances in general knowledge graph construction. However, for a specific domain, harvesting precise and fine-grained knowledge is still difficult due to the long-tail property of entities and relations, together with the lack of high-quality, wide-coverage data sources. In this paper, a domain knowledge graph construction system DKGBuilder is presented. It utilizes a template-based approach to extract seed knowledge from semi-structured data. A word embedding based projection model is proposed to extract relations from text under the framework of distant supervision. We further employ an is-a relation classifier to learn a domain taxonomy using a bottom-up strategy. For demonstration, we construct a Chinese entertainment knowledge graph from Wikipedia to support several knowledge service functionalities, containing over 0.7M facts with 93.1% accuracy.

Keywords: Knowledge graph · Taxonomy learning · Relation extraction

1 Introduction

A domain knowledge graph (DKG) is a special type of knowledge graphs that focuses on modeling relations between entities in a specific domain. It plays an important role in providing knowledge service for special-purpose applications, such as medical diagnosis, movie recommendation, etc.

Although abundant research has been conducted on general-purpose knowledge graph construction, entities and relations in a specific domain are still hard to obtain in a large quantity and a high accuracy. The difficulties mostly lie in three aspects: (i) knowledge in existing manually-built expert systems or domain relational databases usually has the low coverage; (ii) it is difficult to harvest domain facts from semi-structured/unstructured data, especially for long-tail entities and relations; and (iii) taxonomies of DKGs are often designed by experts, and constructing them is a tedious and time-consuming process.

© Springer International Publishing AG 2017
S. Candan et al. (Eds.): DASFAA 2017, Part II, LNCS 10178, pp. 663–667, 2017.
DOI: 10.1007/978-3-319-55699-4_42

In this paper, we introduce DKGBuilder, a general framework to construct a DKG solely from semi-structured and unstructured data sources. In the implementation, it takes Wikipedia pages related to a specific domain as input and extracts entities, classes, attributes and relations in a weakly supervised manner. It first constructs an initial DKG by template-based extractors over Wikipedia infoboxes and categories. We design an is-a relation classifier to build the domain taxonomy based on the Wikipedia category system in a bottom-up strategy. In order to extract long-tail relations from plain texts, a word embedding based projection model is proposed to identify new relations in the embedding space.

For demonstration, we present a transparent process of constructing a Chinese entertainment DKG from scratch. The system also supports several online tasks of knowledge service and analysis. The DKG we constructed consists of over 0.1M entities and 0.7M facts related to the entertainment industry in China. The average accuracy of these facts (i.e., attributes and relations) is 93.1%.

2 System Overview and Key Techniques

As Fig. 1 shows, DKGBuilder consists of offline and online parts. The offline system contains three modules: (i) seed knowledge graph constructor, (ii) taxonomy learner and (iii) plain-text relation extractor. The online part supports entity and class tagging, semantic search and statistical knowledge data analysis.

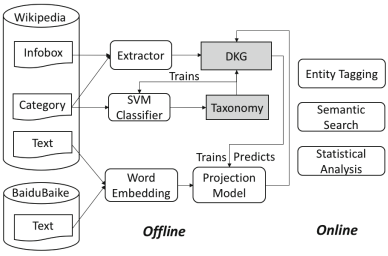

Fig. 1. Framework of DKGBuilder

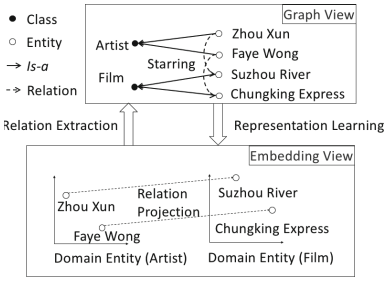

Fig. 2. Knowledge representation

The seed knowledge graph constructor builds an initial DKG with high accuracy, which employs template-based methods to extract domain entities, attributes and relations from Wikipedia infoboxes and categories with minimal human intervention. The extracted facts are later employed as training data for the latter two modules. In traditional domain databases, entities are categorized into a few coarse-grained classes (e.g. patients, diseases, symptoms and medicines in the medical domain). In DKGBuilder, the taxonomy learner classifies entities into a large number of fine-grained classes to construct the taxonomy. It employs several novel features and trains an SVM-based is-a relation classifier based on

Wikipedia category system. The plain-text relation extractor harvests long-tail knowledge by learning new relations from unstructured text. Because a limited, domain-specific corpus is usually sparse in terms of entity contexts, we propose a word embedding based projection model to learn relation representations and extract relations from text using the framework of distant supervision.

We now discuss our implementation details coping with several technical challenges in DKGBuilder.

Seed Knowledge Graph Construction. The first step is to find the most important entities in a specific domain, which minimizes the risk of extracting unrelated entities from knowledge sources. In our system, DKGBuilder takes a couple of human-defined template names from Wikipedia as input and selects entities whose infobox-template name matches one of the pre-defined names. After that, it extracts "seed" relations by mapping frequent attribute names to relations. For each attribute, we design a mapping function which converts an attribute to one/many relations. Based on the semantics of attributes, we categorize the mappings into three types: i.e., *direct*, *multiple* and *indirect*, inspired by the Wikipedia-based ontology YAGO [1]. We additionally use domain-specific filters to improve the precision of the initial DKG.

Fine-Grained Entity Categorization. The construction of domain taxonomy can be modeled as the *fine-grained entity categorization* problem. Because the template names in Wikipedia are relatively coarse-grained (e.g., actor, movie, etc.), we extend our prior work [2] to derive the entertainment taxonomy. For each entity, an SVM classifier is trained to predict whether there is a hyponymy-hypernym (i.e., "is-a") relation between the entity and each of its category names in Wikipedia. For example, "Hong Kong actor" is a hypernym of "Tony Leung Chiu-wai" while "1962 births" only provides relational facts about the actor. A set of features are engineered for accurate is-a relation prediction, such as the number of words in the category name, the POS tag of the head word of the category name, the common sequence of the entity and category names, the existence of specific language patterns, etc. Finally, the top-level of the domain taxonomy is constructed based on the rule mining algorithm proposed in [2].

Representation Learning and Relation Extraction. The limited coverage in the initial DKG prompts us to identify new relations from plain text to cover more long-tail facts. In a domain-specific corpus, especially for Chinese, robust relation extraction is challenging due to the flexibility of language expression and the text sparsity issue of entity contexts [3]. In DKGBuilder, rather than applying syntactic and/or lexical pattern matching methods, we learn entity and relation representations in the embedding space to support relation extraction in a semantic level. We first crawl a large-scale Chinese text corpus from *Baidu Baike*[1], consisting of 1.2M articles and 1.088B Chinese words after word segmentation. A skip-gram model [4] is trained over the text corpus to obtain the 100-dimensional embedding vector $\boldsymbol{v}(e)$ for each entity e.

[1] http://baike.baidu.com/.

For relation representation, similar to [6], we combine two previous relation representation approaches (i.e., vector offsets in [4] and linear projection in [5]) together in the embedding space. For an entity pair (e_i, e_j) that has a certain relation R_k, we assume there is a projection matrix M_k and an offset vector b_k such that $M_k \cdot \boldsymbol{v}(e_i) + b_k \approx \boldsymbol{v}(e_j)$. For ease of implementation, we learn relation representation using the distant supervision framework. We randomly sample relation instances from initial DKG, then learn the parameters by minimizing the following objective function via Stochastic Gradient Descent: $J(M_k, b_k; R_k) = \frac{1}{2} \sum_{(e_i,e_j) \in R_k} \|M_k \cdot \boldsymbol{v}(e_i) + b_k - \boldsymbol{v}(e_j)\|^2$.

After obtaining values of parameters, we make a single pass over the corpus. We first extract entities (e_i and e_j) that co-occur in the same sentence and pair those which have close syntax and lexical distances into candidate relation instances. For each candidate pair (e_i, e_j) and the relation R_k, the model predicts $(e_i, e_j) \in R_k$ iff $\|M_k \cdot \boldsymbol{v}(e_i) + b_k - \boldsymbol{v}(e_j)\| < \delta$ where δ is a pre-defined threshold.

In Fig. 2, we illustrate the two knowledge representations and their connections. In the graph view, entities and relations are expressed explicitly in the form of a directed graph. By representation learning, we can map the DKG into the embedding space. New relations are extracted or inferred from free text based on entity and relation representations, which are again added to the DKG.

3 Demonstration and Evaluation

We will demonstrate the knowledge graph construction process in DKGBuilder and showcase its semantic service functionality. Specifically, the Chinese entertainment DKG consists of 13K classes, 100K entities, 250K attribute-value pairs, 46 relation types and 480K relation instances. By random sampling, we analyze the overall accuracy of facts from all modules, summarized in Table 1.

The online system is developed in Java and uses Tomcat as the Web server. The knowledge data of DKGBuilder is managed by the Neo4j graph database. Besides the basic analysis tasks of entertainment knowledge data, the system supports semantic search of entities and relations, using a search engine-like interface. It also provides knowledge service for deep reading, which tags key entities and classes in documents. Screenshots are shown in Figs. 3 and 4.

Table 1. Accuracy Evaluation

Module	Accuracy
Seed Constructor	99.6%
Taxonomy Leaner	98.7%
Relation Extractor	71.4%
Overall	**93.1%**

Fig. 3. Screenshot I

Fig. 4. Screenshot II

Acknowledgements. This work is partially supported by the National Key Research and Development Program of China under Grant No. 2016YFB1000904 and NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Informatization under Grant No. U1509219.

References

1. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: WWW, pp. 697–706 (2007)
2. Li, J., Wang, C., He, X., Zhang, R., Gao, M.: User generated content oriented Chinese taxonomy construction. In: Cheng, R., Cui, B., Zhang, Z., Cai, R., Xu, J. (eds.) APWeb 2015. LNCS, vol. 9313, pp. 623–634. Springer, Cham (2015). doi:10.1007/978-3-319-25255-1_51
3. Wang, C., Gao, M., He, X., Zhang, R.: Challenges in Chinese knowledge graph construction. In: ICDE Workshops, pp. 59–61 (2015)
4. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013)
5. Fu, R., Guo, J., Qin, B., Che, W., Wang, H., Liu, T.: Learning semantic hierarchies via word embeddings. In: ACL, pp. 1199–1209 (2014)
6. Wang, C., He, X.: Chinese hypernym-hyponym extraction from user generated categories. In: COLING, pp. 1350–1361 (2016)