Data Set and Evaluation of Automated Construction of Financial Knowledge Graph Postprint

Authors: Wenguang Wang, Yonglin, Xu, Chunhui, Du, Yunwen, Chen, Yijie, Wang, Hui, Wen, Wang Wenguang

Date: 2022-11-27T00:00:00+00:00

Abstract

With the technological development of entity extraction, relationship extraction, knowledge reasoning, and entity linking, the research on knowledge graph has been carried out in full swing in recent years. To better promote the development of knowledge graph, especially in the Chinese language and in the financial industry, we built a high-quality data set, named financial research report knowledge graph (FR2KG), and organized the automated construction of financial knowledge graph evaluation at the 2020 China Knowledge Graph and Semantic Computing Conference (CCKS2020). FR2KG consists of 17,799 entities, 26,798 relationship triples, and 1,328 attribute triples covering 10 entity types, 19 relationship types, and 6 attributes. Participants are required to develop a constructor that will automatically construct a financial knowledge graph based on the FR2KG. In addition, we summarized the technologies for automatically constructing knowledge graphs, and introduced the methods used by the winners and the results of this evaluation.

Full Text

Preamble

DATA PAPER

ChinaXiv Partner Journal

Data Set and Evaluation of Automated Construction of Financial Knowledge Graph

Wenguang Wang†, Yonglin Xu, Chunhui Du, Yunwen Chen, Yijie Wang & Hui Wen

DataGrand Inc., Shanghai 201203, China

Keywords: Knowledge graph; Entity extraction; Relation extraction; FR2KG data set; CCKS

Citation: Wang, W.G., et al.: Data set and evaluation of automated construction of financial knowledge graph. Data Intelligence 3(3), 418-443 (2021). doi: 10.1162/dint_a_{00108}

Received: February 11, 2021; **Revised:** March 20, 2021; **Accepted:** April 22, 2021

ABSTRACT

With recent technological advances in entity extraction, relationship extraction, knowledge reasoning, and entity linking, research on knowledge graphs has gained significant momentum. To further promote the development of knowledge graphs, particularly for the Chinese language and financial industry, we have constructed a high-quality dataset named Financial Research Report Knowledge Graph (FR2KG) and organized an evaluation task on automated construction of financial knowledge graphs at the 2020 China Knowledge Graph and Semantic Computing Conference (CCKS2020). FR2KG comprises 17,799 entities, 26,798 relationship triples, and 1,328 attribute triples, covering 10 entity types, 19 relationship types, and 6 attributes.

Participants were required to develop a constructor that automatically builds a financial knowledge graph based on FR2KG. We also summarized technologies for automated knowledge graph construction and introduced the methods employed by the winning teams along with the evaluation results.

1. INTRODUCTION

Driven by advancements in entity extraction, relation extraction, knowledge reasoning, and entity linking technologies, knowledge graph research has flourished in recent years. However, systematic and automated technologies for knowledge graph construction remain relatively underdeveloped, and the ability to automatically construct knowledge graphs ultimately determines the widespread adoption of knowledge graph technologies. To advance this field, high-quality datasets and evaluation systems are essential. In this regard, ImageNet [1] serves as a classic example that has significantly propelled computer vision research. Similarly, in the knowledge graph domain, the Text Analysis Conference (TAC) [2] has released multiple datasets and organized corresponding evaluations to continuously drive progress. Nevertheless, no comparable dataset or evaluation exists for the Chinese language.

Among domain-specific knowledge graphs, financial knowledge graphs are among the most widely applied, finding use in investment research, risk tracking and control, corporate public opinion management, intelligent question answering, and industrial analysis. In the financial domain, constructing knowledge graphs from various unstructured texts represents a fundamental

task of substantial value. However, existing studies on financial knowledge graph construction [3, 4] remain limited, and there is currently no dataset specifically designed for automated construction of financial knowledge graphs with corresponding evaluation benchmarks.

To promote knowledge graph development, particularly in Chinese and for the financial industry, we organized an evaluation task on automated construction of financial knowledge graphs at the 2020 China Knowledge Graph and Semantic Computing Conference (CCKS2020). The data source consisted of Chinese financial research reports on macroeconomics, industries, and companies from professional financial institutions. These reports are characterized by their comprehensiveness, high reliability, depth, and quality, covering financial indicators, policies, and rich data that are highly suitable for building financial knowledge graphs to support in-depth analysis and intelligent decision-making for financial institutions, governments, and research institutes. However, because financial research reports encompass extensive domains and specialized knowledge, and different authors express the same content differently, constructing knowledge graphs from these reports presents considerable challenges. Addressing these problems can significantly advance automated knowledge graph construction and holds substantial academic value.

We collected 1,200 financial research reports and designed a knowledge graph schema comprising 10 entity types, 19 relationship types, and 6 attributes. Financial industry experts annotated 17,799 entities, 26,798 relationship triples <entity, relationship, entity>, and 1,328 attribute triples <entity, attribute key, attribute value> based on this schema, creating the largest published Chinese dataset for automated knowledge graph construction and evaluation. Following the TAC 2016 Cold Start KBP Track plan [5], the evaluation begins with a predefined knowledge graph schema and seed knowledge graph, then automatically extracts entities, relationship triples, and attribute triples from unstructured financial research report texts. The evaluation imposes no restrictions on algorithms and models; participants may utilize open pre-trained models such as BERT [6] and ERNIE [7], as well as third-party open knowledge graphs like those from OpenKG. Participants are also encouraged to employ various methods, including unsupervised, weakly supervised, and distantly supervised approaches, to achieve automated knowledge graph construction.

The remainder of this paper is organized as follows. Section 2 presents the evaluation tasks and methods, along with the release of the professional Financial Research Report Knowledge Graph (FR2KG) dataset. Section 3 describes the properties of FR2KG. Section 4 surveys current knowledge graph construction technologies, and Section 5 introduces the methods used by the winning teams and the evaluation results. Finally, Section 6 discusses challenges and future prospects for automated construction of domain knowledge graphs.

2.1 Task Definition

The evaluation task involves constructing a financial knowledge graph from unstructured financial research report texts based on a given knowledge graph schema:

- **Given:** Unstructured text from financial research reports
- **Given:** Knowledge graph schema
- **Given:** Seed knowledge graph

Participants must develop a constructor that extracts entities, attribute triples, and relationship triples conforming to the schema from the provided unstructured text.

The dataset contains 1,200 financial research reports. After removing tables, images, headers, footers, and other redundant information, the remaining content was converted to plain text format. We collaborated with financial research experts to analyze these reports and designed a financial knowledge graph schema based on financial research characteristics and evaluation requirements. The unstructured text was then annotated by trained annotators, with results reviewed by financial research experts, creating an annotated knowledge graph (annotated KG) composed of verified data. This annotated KG was randomly divided into seed knowledge graph (seed KG) and evaluation knowledge graph (evaluation KG) using the following method: 1) Randomly select 200 of the 1,200 text files; 2) Extract entities, relationship triples, and attribute triples corresponding to these 200 files as seed KG; 3) Remove all seed KG data from the annotated KG, using the remaining data as evaluation KG.

This completes the data processing and annotation pipeline, yielding the full FR2KG dataset, including knowledge graph schema, unstructured financial research report texts, seed KG, and evaluation KG. The evaluation goal is to develop a financial knowledge graph constructor using FR2KG that automatically extracts entities, attribute triples, and relationship triples from unstructured text. The constructed financial knowledge graph excludes data already present in the seed KG, and evaluation employs the metrics described in the next section to measure quality. The entire process is illustrated in Figure 1 [Figure 1: see original paper].

Given the FR2KG dataset, participants aim to develop the most effective financial knowledge graph constructor for automatically extracting entities, attribute triples, and relationship triples from unstructured financial research report texts, constructing a knowledge graph as consistent as possible with expert annotations. To approximate real application scenarios and ensure fairness for all participants, the evaluation permits use of various open or public data, including but not limited to pre-trained models, open knowledge graphs from OpenKG, and other sources. If participants wish to use private data, such data must be publicly available for use by other participants.

2.2 Evaluation Metrics

This evaluation employs the F1 score defined below to assess knowledge graph constructor performance, where higher scores indicate better performance. Knowledge graph data is divided into three types: entities, attribute triples <entity, attribute key, attribute value>, and relationship triples <entity, relationship, entity>. Precision (p), recall (r), and F1 score (F1) are defined as follows.

First, we define the following variables:

- y_E: The set of all <entity, entity type> pairs extracted by the constructor; $|y_E|$ represents the number of entities.
- $\hat{y}_E$: The set of all <entity, entity type> pairs annotated by experts; $|\hat{y}_E|$ represents the number of entities.
- y_A: The set of all attribute triples <entity, attribute key, attribute value> extracted by the constructor; $|y_A|$ represents the number of triples.
- $\hat{y}_A$: The set of all attribute triples <entity, attribute, attribute type> annotated by experts; $|\hat{y}_A|$ represents the number of triples.
- y_R: The set of all relation triples <entity, relationship, entity> extracted by the constructor; $|y_R|$ represents the number of triples.
- $\hat{y}_R$: The set of all relation triples <entity, relation, entity> annotated by experts; $|\hat{y}_R|$ represents the number of triples.
- $y \cap \hat{y}$: The intersection of y and $\hat{y}$, representing the portion extracted by the constructor that matches expert annotations, i.e., correctly extracted entities, attribute triples, or relationship triples; $|y \cap \hat{y}|$ represents the count.

Thus, we have:

Entities:

$$p_E = \frac{|y_E \cap \hat{y}_E|}{|y_E|}, \quad r_E = \frac{|y_E \cap \hat{y}_E|}{|\hat{y}_E|}, \quad F1_E = \frac{2p_{Er}E}{p_E + r_E}$$

Attribute triples:

$$p_A = \frac{|y_A \cap \hat{y}_A|}{|y_A|}, \quad r_A = \frac{|y_A \cap \hat{y}_A|}{|\hat{y}_A|}, \quad F1_A = \frac{2p_{Ar}A}{p_A + r_A}$$

Relationship triples:

$$p_R = \frac{|y_R \cap \hat{y}_R|}{|y_R|}, \quad r_R = \frac{|y_R \cap \hat{y}_R|}{|\hat{y}_R|}, \quad F1_R = \frac{2p_{Rr}R}{p_R + r_R}$$

Finally, we define the overall evaluation F1-score for the entire knowledge graph as in Equation (10):

$$F1 = \frac{F1_E + 2(F1_A + F1_R)}{5}$$

We use a weighted average of the F1-scores for three data types, with attribute triples and relation triples weighted twice as heavily as entities, because we consider extracting attribute and relation triples to be twice as difficult as extracting entities. For an attribute triple, both the entity and attribute value must be correctly extracted, and the attribute value should match the attribute key, which is analogous to identifying an entity and matching its type. Therefore, we determined that attribute extraction should carry double the weight of entity extraction. For a relation triple, two entities must be correctly extracted and the corresponding relationship type must be matched. While detailed study of the relative difficulty of entity, attribute, and relationship extraction would be valuable, our survey of existing literature found no corresponding research, so we determined the weight of 2 based on our experience.

3.1 FR2KG Overview

Among existing datasets, some are manually annotated [8, 9], some are collaboratively annotated by humans and algorithms, and others achieve higher precision through improved algorithms. However, most datasets focus on general domains with news and common-sense articles (such as Wikipedia) or specific domains like biomedical and scientific literature. Financial knowledge graph datasets are rare, and Chinese financial knowledge graph datasets are even rarer. This evaluation marks the first publication of the FR2KG dataset with corresponding unstructured texts, aiming to promote development of distant supervision and weak supervision technologies for automated domain knowledge graph construction.

The FR2KG dataset construction process, described previously and illustrated in Figure 1, proceeded as follows. First, we collected 1,200 financial research reports. Financial domain experts analyzed these reports, extracted plain text from the main body, and saved it in TXT format as the basic unstructured text corpus. Experts and the knowledge graph team then jointly studied these corpora, designing the knowledge graph schema from a financial business perspective and iteratively optimizing it based on evaluation characteristics, ultimately determining a schema containing 10 entity types, 6 entity attributes, and 19 inter-entity relationships. Subsequently, we annotated these corpora using the annotation system of the Yuanhai Knowledge Graph Platform, a DataGrand Inc. product specifically designed for knowledge graph annotation supporting entities, entity attributes, and inter-entity relationships. Before annotation, all annotators received training from financial experts to align their understanding of the schema. All annotated data were reviewed by experts and then divided into seed KG and evaluation KG as described previously. Examples of the FR2KG dataset are shown in Figures 2 and 3 [Figure 2: see original paper][Figure 3: see original paper].

A summary of FR2KG is presented in Table 1 . Currently the largest dataset for automated construction of Chinese financial knowledge graphs, Table 1 describes the data, with detailed sections on FR2KG following below.

3.2 Financial Research Reports

Financial research reports vary significantly in length. As shown in Table 2 , the longest text contains 13,857 characters while the shortest has only 242 characters—a nearly 60-fold difference. However, most text lengths cluster in the 1,000–3,000 character range, accounting for 70% of all reports. In terms of paragraphs, the shortest report has 4 paragraphs while the longest has 74, with most reports containing 10–30 paragraphs (82% of the total).

3.3 Schema

Figure 4 [Figure 4: see original paper] illustrates the FR2KG schema, featuring 10 entity types (represented as ellipses) and 19 relationships between entity types (represented as directed arrows). For example, in the relationship <Person, Invest, Organization>, the directed arrow points from "Person" to "Organization." Among these entity types, three have attributes (Table 3). Notably, time-type attribute values are normalized to the "YYYY-mm-dd" format during annotation, and participants are required to similarly normalize time data in their knowledge graph construction.

The FR2KG schema, shown in Figure 4, supports rich applications in investment research, financial risk assessment and control, product analysis, industrial chain analysis, and other domains. For instance, relationships <Person, Invest, Organization> and <Organization, Invest, Organization> enable in-depth investment and financing analysis. Similarly, relationships <Organization, Sell, Product> and <Organization, Buy, Product> facilitate supply chain analysis, helping identify corporate advantages in the supply chain and assess supply chain risks.

3.4 Entities and Attributes

Tables 4 and 5 summarize the entities and their attribute triples in FR2KG. Table 4 shows entity statistics, while Table 5 details attribute triple statistics across entity types including Research Report, Organization, and Article.

3.5 Relationships

Table 6 summarizes relationship triple statistics in FR2KG, detailing head entity types, relationships, tail entity types, and counts for both seed and evaluation KGs across 19 relationship types.

3.6 Properties of FR2KG

As a comprehensive domain knowledge graph dataset, FR2KG is currently the largest Chinese financial knowledge graph dataset dedicated to automated knowledge graph construction. We plan to continue expanding and enriching its content in the future.

Scale: FR2KG is committed to advancing automated domain knowledge graph construction, offering rich and diverse data types at the largest current scale. The content is extensive, encompassing common stock research reports, industry research reports, and macroeconomic research reports from the financial sector.

Relationships: The dataset provides 19 common relationships in the financial domain (Figure 4) that enable diverse financial industry analyses. For example, relationships <Organization, Has, Risk> and <Industry, Has, Risk> support improved industry or enterprise risk analysis, assessment, and early warning. Relationships <Person, Work for, Organization>, <Person, Invest, Organization>, and <Organization, Invest, Organization> facilitate deep-level equity relationship mining, offering significant value for bank loans and investment analysis.

Professionalism: FR2KG' s schema was jointly designed by financial industry experts and knowledge graph experts. Annotators received training from financial experts before annotation, and all results were reviewed by financial experts to ensure high professional quality.

Diversity: While FR2KG aims to evaluate automated financial knowledge graph construction performance, its applications extend beyond this objective. The dataset can evaluate various other knowledge graph technologies, including link prediction and node classification in graph neural networks, various graph algorithm tasks, and deep learning-based implementations of traditional graph algorithms.

4.1 Entity Extraction

Entity extraction, or named entity recognition (NER), aims to identify mentions of rigid designators from text belonging to predefined semantic types such as person, location, and organization. Popular recent datasets include CoNLL03 [14] and OntoNotes5.0. CoNLL03 contains annotations for Reuters news in English and German, with the English dataset focusing heavily on sports news and annotating four entity types: person, location, organization, and miscellaneous. OntoNotes5.0 annotates a large corpus across various genres with structural information and shallow semantics using 18 entity types. BOSON [15], People' s Daily [16], and MSRA [17] are Chinese entity extraction datasets for general domains, while FR2KG focuses specifically on the Chinese financial domain.

Recent supervised entity extraction research has primarily focused on input representation design and neural model architectures, including context encoders and tag decoders. Additionally, unsupervised and semi-supervised entity extraction has achieved remarkable progress.

4.1.1 Supervised Entity Extraction

Input representation constitutes the first step in entity extraction. This subsection summarizes word-level representation, character-level representation, lan-

guage models, and other representations. Since Mikolov et al. [18] proposed word2vec, many NER studies have used the word2vec toolkit to train word-level representations on various corpora such as PubMed [19], Gigaword [20], NYT [21], and SENNA [22]. GloVe [23] and FastText [24] are also widely used. Character-level representation has proven useful for exploiting explicit sub-word-level information and naturally handling out-of-vocabulary cases [21, 25, 26]. However, both word-level and character-level representations contain only lexical meaning without context.

Consequently, many studies have incorporated context-dependent language model representations into input representation. Peters et al. [27] proposed TagLM, a language model-augmented sequence tagger that considers both pre-trained word embeddings and bidirectional language model embeddings for each token. Building on TagLM, Peters et al. [26] introduced the famous pre-trained bidirectional language model ELMo, which differs by allowing the task model to learn a weighted average of all bidirectional LM layers rather than using only the top layer. Unlike CNNs and RNNs, transformers [28] utilize stacked self-attention and point-wise fully connected layers as basic building blocks for encoders and decoders. Based on transformers, BERT [6] pre-trains a deep bidirectional transformer by jointly conditioning on both left and right contexts in all layers. Combining pre-trained language model embeddings with traditional embeddings has become a de facto standard [29, 30, 31, 32, 33]. Novel input representations continue to be explored, including external knowledge from Wikidata [30], dependency trees [31], and global contextual embeddings [32, 33].

After converting input sentences into representations, context encoders capture contextual dependencies and tag decoders predict token labels. Collobert et al. [34] used CNNs to produce local features around each word, applying max or average pooling to extract global features. Strubell et al. [35] proposed Iterated Dilated CNN (ID-CNN), where four stacked dilated convolutions with width three obtain richer contextual information. Compared to CNNs, bidirectional RNNs leverage both forward and backward information in sentences, effectively extracting whole-sentence features, making them the most popular encoder for NER tasks. While directly connecting bidirectional RNN hidden layers to a softmax layer is possible, adding a CRF layer as the tag decoder helps model sentence-level constraints to ensure prediction validity. Huang et al. [22] first utilized the BiLSTM-CRF architecture for sequence tagging tasks including POS, chunking, and NER. Many subsequent studies also employed BiLSTM encoders with CRF decoders [21, 36, 37]. Beyond CRF, RNN [38] and pointer networks (PtrNet) [39] have been explored as tag decoders. Shen et al. [40] reported that RNN tag decoders outperform CRF and train faster when many entity types exist. However, RNN and PtrNet decoders suffer from greedy decoding, where current step input requires previous step output. As pre-trained models capture sufficient semantic information, some studies use only BERT without BiLSTM-CRF. Notably, Li et al. [41] framed NER as a machine reading comprehension (MRC) problem solvable by fine-tuning BERT.

4.1.2 Semi-supervised Entity Extraction

Semi-supervised entity extraction aims to manually add a small number of appropriate entities as training corpus according to predefined entity types, then use pattern learning methods for continuous iterative learning and manual adjustment to finally generate a named entity dataset, reducing dependence on manually annotated corpora. Liu et al. [42] used small amounts of existing labeled data to train initial KNN and CRF models, performed semi-supervised learning on tweet data, and improved training data by learning from large amounts of unlabeled text. Etzioni et al. [43] proposed the KNOWITALL system, which uses pattern learning, subclass extraction, list extraction, and other modes to perform NER on unlabeled data based on a set of predicate inputs. Zhang and Elhadad [44] proposed a method for extracting named entities from biomedical text based on terminology, corpus statistics (such as inverse document frequency and context vectors), and shallow syntactic knowledge (such as noun phrase chunks), verifying the method on two mainstream biomedical datasets.

4.1.3 Unsupervised Entity Extraction Methods

Based on vocabulary resources, vocabulary patterns, and statistical data calculated from large corpora, named entities can be inferred through clustering and combining similarity in sentence context. Nadeau et al. [45] proposed an unsupervised system for constructing geographical name dictionaries and resolving named entity ambiguities. Zhang and Elhadad [44] proposed an unsupervised method using terminology, corpus statistics, and shallow grammatical knowledge to extract named entities from biomedical texts, demonstrating effectiveness and versatility. Brooke et al. [46] performed Brown clustering based on pre-segmented expectations, combined with rank values for each class after clustering, and constructed bootstrap seeds for training to extract domain-specific entities. Jia et al. [47] used cross-domain language modeling to obtain task and domain vectors for NER entity extraction in unsupervised and supervised fields. Collins and Singer [48] used only seven simple "seed" rules to implement NER on raw data and proposed two unsupervised named entity classification algorithms.

4.2 Relation Extraction

Relation extraction is typically formulated as a classification task that predicts semantic relationships between pairs of nominals. Given sentence S with annotated nominal pairs e_1 and e_2, the goal is to identify the relationship between e_1 and e_2. Relation extraction is generally divided into supervised, unsupervised, and distantly supervised approaches. End-to-end entity and relation extraction has also gained popularity. Supervised datasets are high-quality with minimal noise but often small in scale. SemEval2010 Task 8 [49] contained nine directed relation types with 10,717 samples (8,000 training, 2,717 testing). ACE2005 contains 599 news and email-related documents with seven main relation types, averaging 700 instances per type for training and testing. Besides ACE2005' s

Chinese corpus, DuIE [50] is another large-scale Chinese information extraction dataset. The FR2KG proposed in this study focuses on relation extraction in the Chinese financial domain. For distantly supervised relation extraction, the New York Times (NYT) dataset aligns relations with Freebase, containing 52 possible relation categories plus a special NA category (indicating no relation). The training data comprises 522,611 sentences, 281,270 entity pairs, and 18,252 relations.

4.2.1 Supervised Relation Extraction

Zeng et al. [51] used CNNs to extract lexical and sentence-level features for relation extraction tasks. Lexical-level feature vectors were concatenated from word vectors of labeled entities, context, and WordNet semantic category features. Sentence-level features were automatically extracted using max-pooling CNNs. To eliminate artificial class impact, Santos et al. [52] used pairwise ranking loss instead of cross-entropy for training. Since sentences contain much irrelevant information, sequence feature extraction methods cannot accurately predict inter-entity relationships. Xu et al. [53] noted that the shortest dependency path (SDP) helps determine relationships between entities, successively taking the SDP from subject to object as input, passing it through a lookup table layer to produce local features around each dependency path node, then combining these into a global feature vector via CNN for softmax classification. Similarly, Xu et al. [54] used a four-channel LSTM to extract words, parts of speech, grammatical relations, and WordNet semantic features along the SDP. However, SDP-based studies may neglect crucial information. Zhang et al. [55] encoded complete dependency structures over input sentences using efficient graph convolutional networks (GCN), then extracted entity-centric representations for robust relation predictions. To avoid introducing irrelevant information between entities in complete dependency trees, Guo et al. [56] proposed AGGCN to automatically generate substructures for relation extraction tasks.

4.2.2 Distant Supervision Relation Extraction

Supervised relation extraction requires large amounts of expert-labeled data, limiting its application. Mintz et al. [57] proposed the distant supervision hypothesis: if two entities have a relationship in a known knowledge base, then all sentences mentioning these entities express that relationship in some way. They applied this assumption to align documents with existing databases and automatically generate large training datasets. Zeng et al. [58] proposed a piecewise convolutional neural network (PCNN) for feature extraction and used multi-instance learning to alleviate data noise. However, this approach failed to fully utilize information across different sentences and ignored the possibility of multiple relationships between the same entity pair. Therefore, after using PCNN to extract features from each sentence in a bag, Jiang et al. [59] applied cross-sentence max pooling to select features from different sentences, then aggregated the most important features into representations for each entity pair, finally ap-

plying sigmoid instead of softmax to judge multiple label possibilities. Since different sentences contribute differently, [60, 61] focused on using attention mechanisms for sentence selection. Inspired by the transE model [62], for entity pairs e_1 and e_2 in each bag, $e_1 - e_2$ represents the relation between them. PCNN-extracted features and relation $e_1 - e_2$ are concatenated to obtain each sentence' s weight, with bag features being the weighted sum of all sentence feature vectors. Du et al. [63] proposed a new multi-layer structured self-attention model based on BiLSTM, where a two-dimensional matrix-based word-level attention mechanism focuses on different sentence aspects to better learn contextual representations, and a two-dimensional sentence-level attention mechanism for multi-instance learning focuses on different effective examples to better select sentences. Many studies use existing knowledge bases to add information and alleviate mislabeling in distant supervision. Vashishth et al. [64] added entity types and relations as knowledge base (KB) information to improve prediction performance. Wang et al. [65] proposed a label-free distant supervision method that avoids using relation labels under inadequate assumptions, instead using only prior knowledge derived from the KB to directly and softly supervise classifier learning.

4.2.3 Unsupervised Relation Extraction

Unsupervised learning methods assume that entity pairs with the same semantic relationship have similar contextual information, and each entity pair' s corresponding context can represent its semantic relationship. Hasegawa et al. [66] clustered entity pairs with similar contextual semantics, then selected core vocabulary to mark semantic relationships between categories. [67] improved Hasegawa' s hypothesis by eliminating candidate entity pairs with multiple relationships or performing multi-level clustering to extract relationships. Davidov et al. [68] used Google search as background knowledge to define concept words, automatically extracting related entities and semantic relationships without pre-defining any relation types. Yan et al. [69] combined dependency features with shallow grammatical templates and used clustering methods to extract all semantic relationships of entities in Wikipedia entries from large-scale corpora. Bollegala et al. [70] analyzed post-clustering templates, found implicit semantic relationships between entity pairs, and selected suitable extraction templates from candidate relationship templates, expanding entity relationship scope and improving accuracy and recall.

4.2.4 End-to-end Entity and Relation Extraction

Entity-relation extraction can be pipelined through separate NER and relation classification stages. This independent framework offers flexibility but ignores correlations between tasks, where entity recognition errors may propagate to relation classification. In contrast, joint learning frameworks use a single model to extract entities and relations, effectively integrating entity and relation information.

An end-to-end approach shares model parameters between entity recognition and relation classification tasks. Miwa et al. [71] proposed an end-to-end model capturing both word sequence and dependency tree substructure information by stacking tree-structured LSTM on BiLSTM. Zheng et al. [21] designed novel tags containing entity and relationship information, transforming joint extraction into a tagging problem. However, this scheme struggles with overlapping triples of different relations in sentences. Zeng et al. [72] classified sentences into three types based on triplet overlap degree: Normal, EntityPairOverlap, and SingleEntityOverlap, proposing an end-to-end sequence-to-sequence model with a copy mechanism. The encoder converts natural language sentences into fixed-length semantic vectors, and the decoder generates multiple triplets from these vectors.

To better consider interactions between different relations in sentences, especially overlapping relations, Fu et al. [73] proposed GraphRel, an end-to-end model that in its first stage learns to automatically extract hidden features for each word by stacking a BiLSTM sentence encoder and GCN dependency tree encoder to tag entity mention words and predict relation triplets. In the second stage, GraphRel uses a novel relation-weighted GCN to better predict interactions between triples.

While shared model parameters eliminate the need for subtask constraints, independent sub-models cannot fully exploit relationships between tasks, necessitating global optimization for joint extraction. Building on parameter sharing, Sun et al. [74] proposed a global loss function to explore mutual influence between entity and relation models. Since most existing methods determine relation types only after recognizing all entities, they do not fully model interactions between relation types and entity mentions. Takanobu et al. [75] applied a hierarchical reinforcement learning framework to enhance interaction, where a high-level process detects relationship indicators at specific locations, triggering a low-level process to identify corresponding entities when a relationship is determined. After low-level tasks complete, the high-level process continues searching for the next relationship. Li et al. [76] transformed entity-relation extraction into multiple rounds of question answering, converting the task into answer determination from context, better capturing label hierarchy dependencies. However, this intermediate method is computationally inefficient as it must scan all entity template questions and related relationship template questions in each sentence.

5. PARTICIPANTS OVERVIEW AND EVALUATION RESULTS

A total of 740 teams participated in the evaluation. Among the top 18 teams, three were from companies, ten from universities, three were university-company collaborations, and two did not disclose affiliation information. Table 7 summarizes the prize-winning teams.

The top five teams submitted brief method descriptions, which we analyze and summarize below. All teams used rule-based methods or labeling functions to generate training corpora; only one team manually labeled 20 research reports as supplementary validation samples in addition to automatically generated samples. All teams employed BERT-based models for entity extraction, supplemented by rule-based methods for specific entity types. One team used the BERT-softmax model, three used BERT-CRF architecture, and another used BERT-MRC [38] architecture.

For relationship and attribute extraction, all teams used co-occurrence-based methods. Co-occurrence is the fundamental assumption of distant supervision: when two entities appear together in short text, they likely share a corresponding relationship. Based on this assumption, three teams used rule-based methods to determine relationship existence, while two used BERT-based models for relationship classification. One team employed clustering to group research reports on similar topics.

In summary, for knowledge graph entity, relationship, and attribute extraction tasks, BERT-based pre-trained models remain the best and most popular current approach, widely adopted across teams. As this evaluation closely approximates real industry application scenarios, rule-based methods remain highly effective in some cases and serve as valuable complements to algorithmic approaches.

6. CHALLENGES AND LOOKING AHEAD

The highest F1 score achieved in this evaluation is approximately 0.5 for automatically constructing financial knowledge graphs based on predefined schemas, far from real application requirements. This reveals challenging topics and new research directions for knowledge graphs.

Existing methods for automatically constructing knowledge graphs with given schemas and seed knowledge graphs are not highly effective. Developing end-to-end methods or multi-step frameworks for automated knowledge graph construction remains difficult.

Given a knowledge graph schema, automatically annotating training data for entity, attribute, and relationship extraction is a worthwhile research problem. Furthermore, building excellent models to construct high-precision, high-recall knowledge graphs from automatically annotated training data remains challenging.

For entity extraction, winning participants used BERT-based models with rule-based methods. Further research should explore end-to-end models and unified frameworks for real-world scenarios.

Current relationship and attribute extraction relies heavily on co-occurrence with rule-based or model-based filtering, which depends critically on entity extraction performance. High-precision, high-recall entity extraction yields

good relationship and attribute extraction. However, developing methods that achieve effective relationship and attribute extraction despite considerable entity noise is valuable.

This evaluation did not consider end-to-end models for joint entity and relationship extraction, possibly because numerous entity and relation types exist with limited training corpora. Developing end-to-end models under these conditions is challenging.

Research should extend knowledge graph schemas to include, for example, 50 entity types, hundreds of attributes, and inter-entity relationships.

Further research on automatic construction of multilingual knowledge graphs is needed. This evaluation did not address multilingualism, as our goal was a Chinese-language financial research report knowledge graph (FR2KG). The FR2KG dataset did not involve entity fusion across multiple languages. Constructing knowledge graphs from multilingual corpora involves new topics including entity alignment and cross-language relationship extraction, making evaluation of multilingual knowledge graph construction meaningful.

This evaluation implicitly involved limited entity disambiguation and fusion without explicit evaluation in this area. Future knowledge disambiguation and fusion evaluations will likely become increasingly active.

Detailed study of the relative difficulty of entity, attribute, and relationship extraction is significant, as is establishing reasonable metrics for automated knowledge graph construction.

7. CONCLUSIONS

This paper introduces FR2KG, a high-quality dataset comprising 17,799 entities, 26,798 relationship triples, and 1,328 attribute triples covering 10 entity types, 19 relationship types, and 6 attributes. We present an overview of the automated financial knowledge graph construction evaluation at CCKS2020, summarize technologies for automated knowledge graph construction, and identify challenging topics and new research directions in knowledge graphs.

AUTHOR CONTRIBUTIONS

W.G. Wang (wangwenguang@datagrand.com) served as project team leader, providing overall technical leadership, designing the data schema, performing curation, and contributing to manuscript writing and editing. Y.L. Xu (xuyonglin@datagrand.com) investigated the latest relation extraction progress and wrote relevant chapters. C.H. Du (duchunhui@datagrand.com) investigated the latest entity extraction progress and wrote relevant chapters. Y.W. Chen (chenyunwen@datagrand.com) organized data annotation and contributed as senior author to writing and editing. Y.Y. Wang (wangyijie@datagrand.com) and H. Wen (wenhui@datagrand.com) contributed to evaluation result statistics and

review. All authors made meaningful and valuable contributions to manuscript revision and proofreading.

DATA AVAILABILITY STATEMENT

All data are available at Data Intelligence' s data repository at the Science Data Bank, https://doi.org/10.11922/sciencedb.01060, under an Attribution 4.0 International (CC BY 4.0) license.

REFERENCES

[1] Jia, D., et al.: ImageNet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)

[2] Ji, H., Nothman, J.: Overview of TAC-KBP2016 tri-lingual EDL and its impact on end-to-end KBP. In: Proceedings of Text Analysis Conference, pp. 1–15 (2016)

[3] Elhammadi, S., et al.: A high precision pipeline for financial knowledge graph construction. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 967–977 (2020)

[4] Ein-Dor, L., et al.: Financial event extraction using Wikipedia-based weak supervision. In: Proceedings of the Second Workshop on Economics and Natural Language Processing, pp. 10–15 (2019)

[5] TAC KBP 2016 Cold Start Track. Available at: https://tac.nist.gov/2016/KBP/ColdStart/index.html. Accessed 30 July 2021

[6] Devlin, J., et al.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 4171–4186 (2019)

[7] Zhang, Z., et al.: ERNIE: Enhanced language representation with informative entities. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1441–1451 (2019)

[8] Ringland, N., et al.: NNE: A data set for nested named entity recognition in English newswire. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5176–5181 (2019)

[9] Thorne, J., et al. FEVER: A large-scale data set for fact extraction and VERification. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 809–819 (2018)

[10] Yao, Y., et al: DocRED: A large-scale document-level relation extraction data set. In: Proceedings of the 57th Annual Meeting of the Association for

Computational Linguistics, pp. 764–777 (2019)

[11] Zhu, T., et al.: Towards accurate and consistent evaluation: A data set for distantly-supervised relation extraction. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 6436–6447 (2020)

[12] Wang, L., et al.: CORD-19: The Covid-19 open research data set. arXiv preprint arXiv:2004.10706v2 (2020)

[13] D' Souza, J., et al.: The STEM-ECR data set: Grounding scientific entity references in STEM scholarly content to authoritative encyclopedic and lexicographic sources. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 2192–2203 (2020)

[14] Sang, E.F., Meulder, F.D.: Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 142–147 (2003)

[15] BOSON data set. Available at: https://github.com/InsaneLife/ChineseNLPCorpus/tree/master/NER/bos
Accessed 30 June 2021

[16] People's Daily data set. Available at: https://github.com/InsaneLife/ChineseNLPCorpus/tree/master/NE
Accessed 30 June 2021

[17] Levow, G.A.: The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pp. 108–117 (2006)

[18] Mikolov, T., et al.: Efficient estimation of word representations in vector space. In: International Conference on Learning Representations, pp. 1–12 (2013)

[19] Yao, L., et al.: Biomedical named entity recognition based on deep neutral network. International Journal Hybrid Information Technitology 8(8), 279–288 (2015)

[20] Nguyen, T.H., et al.: Toward mention detection robustness with recurrent neural networks. arXiv preprint arXiv:1602.07749 (2016)

[21] Zheng, S., et al.: Joint extraction of entities and relations based on a novel tagging scheme. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 1227–1236 (2017)

[22] Huang, Z., et al.: Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)

[23] Li, P.H., et al.: Leveraging linguistic structures for named entity recognition with bidirectional recursive neural networks. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2664–2669 (2017)

[24] Wang, C., et al.: Code-switched named entity recognition with embedding attention. In: Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching (CALCS), pp. 154–158 (2018)

[25] Kuru, O., et al.: CharNER: Character-level named entity recognition. In: Proceedings of the 26th International Conference on Computational Linguistics, pp. 911–921 (2016)

[26] Peters, M.E., et al.: Deep contextualized word representations. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 2227–2237 (2018)

[27] Peters, M.E., et al.: Semi-supervised sequence tagging with bidirectional language models. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 1756–1765 (2017)

[28] Vaswani, A., et al.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010 (2017)

[29] Liu, T., et al.: Towards improving neural named entity recognition with gazetteers. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5301–5307 (2019)

[30] Song, C.H., et al.: Improving neural named entity recognition with gazetteers. arXiv preprint arXiv:2003.03072 (2020)

[31] Jie, Z., Lu, W.: Dependency-guided LSTM-CRF for named entity recognition. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 3860–3870 (2019)

[32] Liu, Y., et al.: GCDT: A global context enhanced deep transition architecture for sequence labeling. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2431–2341 (2019)

[33] Luo, Y., et al.: Hierarchical contextualized representation for named entity recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 8441–8448 (2019)

[34] Collobert, R., et al.: Natural language processing (Almost) from scratch. Journal of Machine Learning Research 12, 1462–1467 (2011)

[35] Strubell, E., et al.: Fast and accurate entity recognition with Iterated Dilated Convolutions. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2670–2680 (2017)

[36] Lample, G., et al.: Neural architectures for named entity recognition. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 260–270 (2016)

[37] Chiu, J.P.C., et al.: Named entity recognition with Bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics 4, 357–370 (2016)

[38] Chaudhary, A., et al.: A little annotation does a lot of good: A study in bootstrapping low-resource named entity recognizers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 5163–5173 (2019)

[39] Zhai, F., et al.: Neural models for sequence chunking. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3365–3371 (2017)

[40] Shen, Y., et al.: Deep active learning for named entity recognition. In: Proceedings of the 2nd Workshop on Representation Learning for NLP, pp. 252–256 (2017)

[41] Li, X., et al.: Dice loss for data-imbalanced NLP tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 465–476 (2020)

[42] Liu, X., et al.: Recognizing named entities in tweets. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 359–367 (2011)

[43] Etzioni, O., et al.: Unsupervised named-entity extraction from the Web: An experimental study. Artificial Intelligence 165(1), 91–134 (2005)

[44] Zhang, S., Elhadad, N.: Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. Journal of Biomedical Informatics 46(6), 1088–1098 (2013)

[45] Nadeau, D., et al.: Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. In: Conference of the Canadian Society for Computational Studies of Intelligence, pp. 266–277 (2006)

[46] Brooke, J., et al.: Bootstrapped text-level named entity recognition for literature. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 344–350 (2016)

[47] Jia, C., et al.: Cross-domain NER using cross-domain language modeling. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2464–2474 (2019)

[48] Collins, M., Singer, Y.: Unsupervised models for named entity classification. In: 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 100–110 (1999)

[49] Hendrickx, I., et al.: SemEval-2010 Task 8: Multi-way classification of semantic relations between pairs of nominals. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 33–38 (2010)

[50] Li, S., et al.: DuIE: A large-scale Chinese data set for information extraction. In: The CCF International Conference on Natural Language Processing and Chinese Computing, pp. 791–800 (2019)

[51] Zeng, D., et al.: Relation classification via Convolutional Deep Neural Network. In: Proceedings of the 25th International Conference on Computational Linguistics, pp. 2335–2344 (2014)

[52] Santos, C.N., et al.: Classifying relations by ranking with Convolutional Neural Networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, pp. 626–634 (2015)

[53] Xu, K., et al.: Semantic relation classification via Convolutional Neural Networks with simple negative sampling. In: EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, pp. 536–540 (2015)

[54] Xu, Y., et al.: Classifying relations via long short term memory networks along shortest dependency paths. In: EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, pp. 1785–1794 (2015)

[55] Zhang, Y., et al.: Graph convolution over pruned dependency trees improves relation extraction. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2205–2215 (2018)

[56] Guo, Z., et al.: Attention guided graph convolutional networks for relation extraction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 241–251 (2020)

[57] Mintz, M., et al.: Distant supervision for relation extraction without labeled data. In: Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pp. 1003–1011 (2009)

[58] Zeng, D., et al.: Distant supervision for relation extraction via piecewise convolutional neural networks. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1753–1762 (2015)

[59] Jiang, X., et al.: Relation extraction with multi-instance multi-label convolutional neural networks. In: Proceedings of the 26th International Conference on Computational Linguistics, pp. 1471–1480 (2016)

[60] Ji, G., et al.: Distant supervision for relation extraction with sentence-level attention and entity descriptions. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence, pp. 3060–3066 (2017)

[61] Lin, Y., et al.: Neural relation extraction with selective attention over instances. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 2124–2233 (2016)

[62] Bordes, A., et al.: Translating embeddings for modeling multi-relational data. In: Advances in Neural Information Processing Systems, pp. 2787–2795

(2013)

[63] Du, J., et al.: Multi-level structured self-attentions for distantly supervised relation extraction. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2216–2225 (2018)

[64] Vashishth, S., et al.: RESIDE: Improving distantly-supervised neural relation extraction using side information. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 1257–1266 (2018)

[65] Wang, G., et al.: Label-free distant supervision for relation extraction via knowledge graph embedding. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2246–2255 (2018)

[66] Hasegawa, T., et al.: Discovering relations among named entities from large corpora. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, pp. 415–422 (2004)

[67] Rozenfel, B., Ronen, F.: High-performance unsupervised relation extraction from large corpora. In: The Sixth International Conference on Data Mining, pp. 1032–1037 (2006)

[68] Davidov, D., et al.: Fully unsupervised discovery of concept-specific relationships by Web mining. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 232–239 (2007)

[69] Yan, Y., et al.: Unsupervised relation extraction by mining Wikipedia texts using information from the Web. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 1021–1029 (2009)

[70] Bollegala, D.T., et al.: Measuring the similarity between implicit semantic relations from the Web. In: Proceedings of the 18th International Conference on World Wide Web, pp. 651–660 (2009)

[71] Miwa, M., et al.: End-to-end relation extraction using LSTMs on sequences and tree structures. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1105–1116 (2016)

[72] Zeng, X., et al.: Extracting relational facts by an end-to-end neural model with copy mechanism. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 506–514 (2018)

[73] Fu, T.J., et al.: GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1409–1418 (2019)

[74] Sun, C., et al.: Extracting entities and relations with joint minimum risk training. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 2256–2265 (2018)

[75] Takanobu, R., et al: A hierarchical framework for relation extraction with reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 7072–7079 (2019)

[76] Li, X., et al.: Entity-relation extraction as multi-turn question answering. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1340–1350 (2019)

AUTHOR BIOGRAPHY

Wenguang Wang received his M.S. degree from Zhejiang University, China. He is currently Vice President of DataGrand Inc. His research interests include knowledge graphs, natural language processing, computer vision, deep learning, reinforcement learning, and content generation. He holds ten patents and has published several academic papers in artificial intelligence. He is a member of the China Computer Federation (CCF), Chinese Association for Artificial Intelligence (CAAI), and Chinese Information Processing Society of China (CIPS). ORCID: 0000-0002-9617-0818

Yonglin Xu is currently an algorithm engineer at DataGrand Inc. He received his Master' s degree from Shanghai University in 2019. His research interests include information extraction and knowledge graphs. ORCID: 0000-0002-7716-0841

Chunhui Du received his B.S. degree from the University of Electronic Science and Technology of China in 2018. He is currently pursuing a PhD degree from the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China. His research interests include federated learning and natural language processing. ORCID: 0000-0002-9086-6021

Yunwen Chen received his PhD degree from Fudan University, China. He is the founder and CEO of DataGrand Inc., Shanghai, a leading AI company in China. He previously served as Chief Data Officer of Shanda, Inc., Burlington, IA, USA, Senior Director at Tencent, Inc., Shenzhen, China, and Researcher at Baidu, Inc., Beijing, China. He holds 32 patents and has published several academic papers. His current research interests include data mining, natural language processing, search and recommendation systems, and knowledge graphs. Dr. Chen received the Distinguished Graduate Student award in 2008. He is a Senior Member of the China Computer Federation (CCF) and a member of ACM. ORCID: 0000-0003-4513-9439

Yijie Wang is an algorithm engineer at DataGrand Inc. He received his Master' s degree from Northeastern University in 2018. His research interests include data mining and knowledge graphs. ORCID: 0000-0003-1310-7467

Hui Wen graduated with a Master' s degree in Computer Application Technology from Tongji University in 2014. He is a co-founder and senior scientist at DataGrand Inc., responsible for research and development of knowledge graphs

and data mining. He has rich R&D experience and strong interests in knowledge graphs, recommendation systems, search systems, and distributed systems. ORCID: 0000-0001-7593-8544

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