Knowledge and Information Systems (2020) 62:317–336
https://doi.org/10.1007/s10115-019-01351-4

REGULAR PAPER

Constructing biomedical domain-specific knowledge graph with minimum supervision

Jianbo Yuan[1] · Zhiwei Jin[2] · Han Guo[2] · Hongxia Jin[3] · Xianchao Zhang[4] · Tristram Smith[5] · Jiebo Luo[1]

Received: 4 January 2018 / Revised: 27 February 2019 / Accepted: 8 March 2019 / Published online: 23 March 2019

Abstract
Domain-specific knowledge graph is an effective way to represent complex domain knowledge in a structured format and has shown great success in real-world applications. Most existing work on knowledge graph construction and completion shares several limitations in that sufficient external resources such as large-scale knowledge graphs and concept ontologies are required as the starting point. However, such extensive domain-specific labeling is highly time-consuming and requires special expertise, especially in biomedical domains. Therefore, knowledge extraction from unstructured contexts with minimum supervision is crucial in biomedical fields. In this paper, we propose a versatile approach for knowledge graph construction with minimum supervision based on unstructured biomedical domain-specific contexts including the steps of entity recognition, unsupervised entity and relation embedding, latent relation generation via clustering, relation refinement and relation assignment to assign *cluster-level* labels. The experimental results based on 24,687 unstructured biomedical science abstracts show that the proposed framework can effectively extract 16,192 structured facts with high precision. Moreover, we demonstrate that the constructed knowledge graph is a sufficient resource for the task of knowledge graph completion and new knowledge inference from unseen contexts.

Keywords Knowledge graph construction · Biomedical · Domain-specific · Minimum supervision

✉ Jianbo Yuan
jyuan10@cs.rochester.edu

1 Department of Computer Science, University of Rochester, Rochester, NY 14623, USA

2 Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

3 Samsung Research America, Mountain View, CA, USA

4 School of Software Technology, Dalian University of Technology, Dalian, China

5 Department of Pediatrics, University of Rochester Medical Center, Rochester, NY, USA

1 Introduction

Knowledge graph is an effective way for representing semantical real-world facts in a structured form. A knowledge graph is a collection of triplets among which each indicates a piece of fact in the form of "head entity–relation–tail entity". Large-scale knowledge graphs are built by leading institutes containing enormous facts such as YAGO3 [21], DBpedia [19] and Freebase [8]. Knowledge graphs have been widely adopted and shown promising benefits in a wide range of applications such as information retrieval, question & answering and knowledge reasoning. In addition, facts extracted from specific domains convey domain knowledge which is usually only accessible to experts in such areas. For example, biomedical knowledge graphs have been beneficial to decision making and information inference in the areas of biomedical domain and healthcare by adding the domain knowledge [5,13] and improving the end-to-end healthcare applications [5,38] by embedding the domain knowledge as external constrains into existing systems. Therefore, the construction of knowledge graphs conveying domain knowledge, especially for biomedical science and healthcare, is of great significance to the success of many domain-specific real-world applications.

Conventional methods of knowledge graph construction consist of two perspectives: manual construction and automatic or semiautomatic methods. The manually curated or collaborative constructions of knowledge graphs such as WordNet[1] and concept ontologies such as Unified Medical Language System (UMLS) Metathesaurus,[2] are highly time-consuming. On the other hand, most of the automatic construction approaches are pipeline-based processes which include entity recognition and relation extraction, among which the entity recognition in general is a well-explored task. Biomedical entity recognition is special in that conventional NLP approaches yield inaccurate performance without the help of external resources such as UMLS compared with the supervised approaches which consider the task as a classification problem, distant supervision approaches generate the training labels heuristically before applying the classifications to avoid the cost of extensive human annotation. Both supervised and distant supervised approaches require sufficient labeled training samples or an external knowledge base to begin with, and highly depend on the hand-crafted language patterns and features predefined by the domain experts. Additionally, biomedical domain-specific knowledge graph construction is different in that: (a) biomedical entity recognition is difficult for conventional Name Entity Recognition (NER) algorithms without adopting domain-specific resources such as UMLS, (b) the learning of semantic embeddings for biomedical entities has to be generated from domain-specific corpus, (c) none large-scale biomedical knowledge graph exists based on which knowledge graph completion or distant supervision can be performed and (d) large-scale labeling for a manual construction is time-consuming and requires extensive expertise.

To overcome these limitations, we propose a versatile framework with minimum supervision for domain-specific knowledge graph construction with open-ended relations extracted from unstructured biomedical science articles. No extensive labeling, predefined relations or language patterns are required by the workflow. The proposed process includes biomedical entity recognition, unsupervised entity and relation embedding based on skip-gram [23], latent relation generation by clustering based on relation embeddings, and a relation refinement on the automatic generated latent (noisy) labels based on a convolutional neural network (CNN) with attention model. All the previous steps are conducted without any human annotations. In the end, the semantical relations are assigned manually only on a *cluster-level*

[1] https://wordnet.princeton.edu/.

[2] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/.

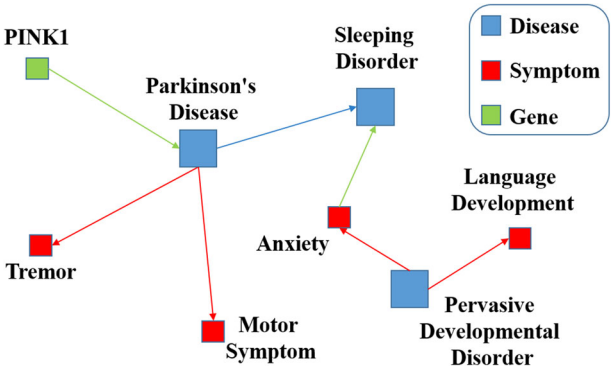


Fig. 1 An example of the constructed biomedical knowledge graph

(per cluster) instead of being evaluated per sample. In this way, we achieve the objective of constructing a domain-specific knowledge graph with minimum supervision. Our approach should certainly work in friendlier and none domain-specific cases when the labels are easier to obtain; nonetheless, we focus on the biomedical domain because the limited resources in the specific domain motivate approaches with minimum supervision and would lead to a higher real-world impact.

For evaluation purposes, we have collected 24,687 autism spectrum disorder-related article abstracts from PubMed[3] related to the areas of medical, healthcare, etc. In our experiments, 6827 entities are extracted from the collected dataset and form into **entity pairs** based on their co-occurrence as **candidate facts**. All the candidates are then categorized into 20 clusters with kmeans++ [2]. In the end, 16,192 facts under six strong relations are extracted from the collected dataset, with a precision of 83.3% for the top-10 and 59.3% for the top-50 most confident facts averaged over all relation categories on a testing set consisting of 1000 manually evaluated samples (50 per cluster). Selected examples in the constructed knowledge graph are shown in Fig. 1. Additionally, we demonstrate the capability of inferring new facts from unseen texts based on the constructed knowledge graph with very promising results.

Our contributions are summarized as follows:

- We propose a novel approach to constructing biomedical domain-specific knowledge graphs. It is capable of capturing open-ended relations without extensive labeling, requiring special expertise, or the help of existing knowledge graphs. This versatile approach can be potentially applied to other domains that share the similar limitations.
- We evaluate the correctness of the constructed biomedical knowledge graphs and their ability to extract new knowledge from unseen contents.
- We construct knowledge graphs efficiently from a large-scale set of unstructured biomedical abstracts with high precision. The domain knowledge implied in the knowledge graph is beneficial to real-world applications such as disease diagnosis and medical information retrieval.

2 Related work

Knowledge graph construction approaches are generally categorized as follows: manual construction in a curated or collaborative approach and supervised or semisupervised modeling

[3] https://www.ncbi.nlm.nih.gov/pubmed.

[24]. Manual knowledge graph construction in a curated way highly depends on experts and can result in a high accuracy, but lacks in scalability and velocity, which is usually adopted for domain-specific problems such as biomedical-related applications, for example, UMLS. Knowledge graphs constructed and evaluated collaboratively by open groups have a better velocity but are still far from satisfaction, such as Freebase [8]. Consequently, probabilistic approaches are drawing more attention in order to build large-scale and highly accurate knowledge graphs. For example, YAGO3 [21] is evaluated to achieve a precision of more than 95% constructed from Wikipedia Infoboxes. However, as opposed to unstructured formats, semistructured data only cover a limited fraction of the knowledge despite how large the number is of the facts extracted from them. Probabilistic approaches such as DeepDive [25] are designed to work on unstructured data (Internet data) via learned linguistic features and self-defined statistical rules [1].

More specifically, automatic and semiautomatic knowledge graph construction workflow includes entity recognition and relation extraction. For relation extraction task, recent approaches take the advantages of the probabilistic or deep learning algorithms in order to remove the limitations of applying hand-crafted features. For example, previous work such as TransE and its extensions TransH and TransR considered the relations between entities as a translation in the vector space [9,34]. More recently, [39] and [20] constructed a CNN with piecewised max-pooling and enhanced the performance by adding a sentence-level attention model to reduce the side effects raised by noisy labels. Xie et al. [35] utilized entity type information and the hierarchies of entity types for knowledge representation learning. In cases of lacking labeled data, distant supervision takes advantages of the existing knowledge graphs to heuristically generate noisy training data [4,16,27,28]. Distant supervision incorporated with multi-instance learning has shown to be effective for relation extraction [20,33] to reduce the side effects of applying noisy labels.

Domain knowledge acquisition in biomedical domain from texts has been an active field where most previous studies on conceptual ontologies rely highly on the annotations from domain experts entirely or as a starting point. For example, UniProt is a large-scale dataset of protein sequences and annotations constructed based on manual annotation and collaboratively rule-based completion relying on the expert-curated knowledge [11]. Gene Ontology is constructed in a similar collaborative way [3] and contains logical structure of the biological functions and their relationships.[4] Compared with studies on ontologies such as Gene Ontology, knowledge graph is more general purposed, with open-ended relations, focusing less on the logical structure and more on the contents and rich semantics. For biomedical knowledge acquisition, previous studies such as Bio2RDF [6] aimed to link related concepts and articles. More similar work to ours is SemRep which extracts hypernymic propositions using linguistic feature (sentence structure) under predefined categories such as drugs and chemicals where existing ontology is also required [29]. Knowlife applied seed facts of 13 relations to extract sentence-level and document-structure patterns for knowledge graph construction and achieved high precisions with typing and mutual-exclusion constrains for pruning out invalid candidate facts [13]. Compared with the studies discussed above, we are aiming to extract open-ended relations between entities without predefining relations or requiring large-scale manual labeling.

Since a domain-specific knowledge graph conveys domain knowledge in a structured form which can fill in the gap between expertise and the crowds and is essential for the success of a wide range of real-world applications. Google powers its health-related searches with

[4] http://www.geneontology.org/page/introduction-go-resource.

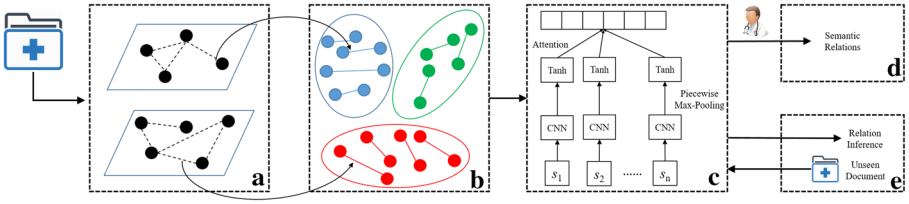


Fig. 2 The framework of minimally supervised biomedical knowledge graph construction: **a** Entity recognition and entity pairing from biomedical documents, **b** entity & relation embedding, and kmeans++ clustering for latent label generation, **c** relation refinement on multi-instance attention-based CNN, **d** cluster-level relation assignment and **e** relation inference on unseen documents as an application of the constructed knowledge graph

carefully curated medical knowledge graphs by a team of experts,[5] and successful examples also have been shown in discovering miss-diagnosis [5], linking implicitly related biomedical entities and more accurate domain-driven information retrieval in academia [5,12].

3 Methodology

In this work, we propose to construct a domain-specific knowledge graph with minimum supervision which can generate open-ended relations from unstructured biomedical science articles. A domain-specific knowledge graph construction shares the limitations with conventional methods as well as its own constrains. Therefore, we argue that a minimally supervised construction of biomedical domain-specific knowledge graph which is capable to extract open-ended relations is desired.

3.1 Framework

Unstructured biomedical articles are collected from PubMed and are input to the proposed knowledge graph construction framework (Fig. 2) including the following steps:

Preprocessing Non-English characters and punctuations are removed from the unstructured text contents. To prepare data for later steps, we applied Standford CoreNLP to parse the texts into sentences [22].
Entity recognition Biomedical entities are extracted from PubMed documents by MinHash lookup [10] method based on the UMLS [32]. Two entities are assumed to have explicit or implicit relations if they frequently co-occur in one sentence. An **entity pair** is defined by linking such entities and considered as a **candidate fact** (Fig. 2a).
Entity and relation embedding Skip-gram [23] is applied for entity embedding, and the subtraction of the two entity vectors for each entity pair is used as the vector representation of their relation (Fig. 2b).
Latent label generation Given the relation vector for each entity pair, kmeans++ [2] is then used for categorizing the entity pairs with similar relations to the same cluster where the latent labels are used as weak labels for relation refinement (Fig. 2b).

[5] https://googleblog.blogspot.com/2015/02/health-info-knowledge-graph.html.

Relation refinement The latent labels generated are considered to be noisy. We apply a sentence-level attention model over multi-instance learning on a piecewise CNN [20] for relation refinement by reducing the attentions on the noisy instances (Fig. 2c).
Relation assignment The learning and construction process is unsupervised till the label assignment. Semantical relations are obtained by a *cluster-by-cluster* manual labeling, which is considered to be far less costly than evaluating every single extracted entity pair (Fig. 2d).
Relation inference The constructed biomedical domain-specific knowledge graph conveys domain knowledge which can be applied for relation inference on unseen documents by using the same step of relation refinement with the labels generated in relation assignment (Fig. 2e).

3.2 Entity recognition

Biomedical entity recognition is an essential component in biomediscal knowledge graph construction which is adopted early in the construction workflow, and special in that conventional NLP approaches yield inaccurate performance without the help of external recourses such as UMLS because it provides the largest thesaurus and the semantical ontology of biomedical concepts (referred as entities in knowledge graph). The entities which are no longer than five words and one hundred characters are selected to build the vocabulary from the latest version of UMLS knowledge sources including lexical variations of the concepts such as verb conjugations and different word orders. A MinHash [10] is applied [32] for fast lookup where an entity e is expressed in the form of a set s_e of character trigrams, and the probability of two entities falling into the same bucket equals to their Jaccard similarity. Each trigram set representing a possible entity or an entity in the vocabulary is hashed by concatenating the minimums of the hashing results by a set of functions $\{\pi_1, \pi_2, \ldots, \pi_n\}$ among which each π maps the trigram set into an integer set. The MinHash [10] of an entity e is $[\min(\pi_1(s_e)), \min(\pi_2(s_e)), \ldots, \min(\pi_n(s_e))]$. For each hashing function $\pi_i, i \in [1, 2, \ldots, n]$, the probability of two entities s_{e_1} and s_{e_2} falling into the same bucket is described in Eq. 1 [32]:

$$p\left[\min(\pi_i(s_{e_1})) = \min(\pi_i(s_{e_2}))\right] = J(s_{e_1}, s_{e_2}) \tag{1}$$

where $J(s_{e_1}, s_{e_2})$ denotes the Jaccard similarity between the two entity trigram sets. Entity disambiguation is handled in two ways: First, entity mentions (namely entity candidates) are generated from consecutive words with length from one to five, and undesirable candidates are discarded heuristically; Second, a set of unambiguous entities is constructed consisting of Medical Subject Headings (MeSH) terms[6] (which is considered unambiguous) and entities with only one UMLS match, based on which heuristics such as singular/plural forms, linguistic semantic patterns and co-occurring semantic types are applied remove ambiguities based on the anchors [31]. A recognized entity can be a single word, a multi-word phrase, or a phrase abbreviation. After the entity recognition, we consider each multi-word phrase as one single-word entity by tagging it with underlines in between and reprocess the input documents.

[6] https://www.nlm.nih.gov/mesh/.

3.3 Entity and relation embedding

Previous word embedding approaches such as skip-gram [23] and knowledge graph embedding algorithms [9] consider the relation representations of entity pairs as a translation between entities in the vector space. For example, the calculation stands as $vec(\text{“}Madrid\text{”}) - vec(\text{“}Spain\text{”}) + vec(\text{“}France\text{”}) \approx vec(\text{“}Paris\text{”})$. Intuitively, the relation vector is calculated as a subtraction between the entity vectors for each entity pair to represent its relation. Other word embedding approaches follow the similar ideas such as [7,26]. Since our focus is not on evaluating word embedding models, the discussion between different word embedding models is beyond the scope of this paper and we selected the commonly used skip-gram. Given a word set T including words $\{w_1, w_2, \ldots, w_T\}$, the vector $\mathbf{r}_{i,j}$ representing the relation between w_i and w_j is defined as:

$$\mathbf{r}_{ij} = \mathbf{w}_i - \mathbf{w}_j \tag{2}$$

where $\mathbf{w}_i$ and $\mathbf{w}_j$ denote the word vectors for word w_i and w_j, respectively. The objective of the skip-gram model is to predict the surrounding words of the current word by learning a word vector. Given a window size c, the objective function is to maximize the average log probability within the window size of the current word:

$$\frac{1}{|T|} \sum_{t=1}^{|T|} \sum_{-c \le j \le c, j \ne 0} \log p\left(w_{t+j} | w_t\right) \tag{3}$$

where the conditional probability is defined by the softmax function. Alternatively, hierarchical softmax and negative sampling are applied instead of the softmax function for better computational efficiency.

3.4 Latent relation label generation

Since relations of the same category are expected to share similar embedding vectors, we apply kmeans++ [2], a variation of kmeans algorithm with careful center seedings, to categorize the relations into latent labels (latent relation types). Since at this stage, we have no knowledge of how many relation categories exit, we run the kmeans++ with different configurations and calculate the Silhouette Coefficient [30] to validate the parameter settings. After clustering, each entity pair is assigned with a latent label (relation category) r_i indicating that it belongs to the i-th cluster.

Entity ordering Relations can be asymmetric (only one-direction is considered correct) or symmetric. For asymmetric relations, we expect two or more clusters formed by each asymmetric relation because no fixed order is preserved to tell whether an entity is the head entity or the tail entity based only on the contexts and the relation vectors under the same latent label may have opposite-directed embeddings in the vector space. For example, the entity pair $(autism, poor\_social\_interaction)$ and $(learning\_disability, autism)$ both represents the relation *is_Symptom_of*, however, the relation vectors are likely to be negative with each other rather than equal due to the order issue. Consequently, we intentionally increase the number of clusters by an empirical choice based on the Silhouette Coefficient, and later during the manual relation assignment for each cluster, we assign them as one relation and obtain the correct ordering for each cluster. The manual work is per cluster which is not difficult. Additionally, since the number of clusters is set more than the number

of relations, we assume there are clusters contains nonexistent relations and discard these clusters during the label assignment.

Mining strong relevance Discovering entities with strong relevance is a related area to building concept ontologies and knowledge graph. Existing work needs predefined entity types or relations. For example, entities fall into the "Disease–Drug" category potentially carries the relation of "treat" [5,17]. In our approach, the latent labels are generated by unsupervised clustering instead of assigned with explicit categories, which provided us with the flexibility of discovering open-ended relation, or implicit relevance regardless of the entity categories. The entities are considered to have strong relevance if the corresponding entity pair is categorized into the cluster which are not discarded in the later process with a high confidence.

3.5 Relation refinement and inference

Although the latent relation labels have been generated through clustering, the labels are considered to be noisy because the process of entity relation embedding and latent label generation are both based on unsupervised learning methods. Therefore, we perform a pseudo-supervised classification based on the piecewise CNN model [20] for a relation refinement. In order to avoid the "garbage-in garbage-out" situation, a sentence-level attention model on multi-instance learning is used to reduce the weights of noisy labels [20]. Since studies have shown that the corresponding contexts to the entities of interests contain useful information to reduce the semantic ambiguity [37,39], we embed both of the targeted entities and their contexts in order to achieve a context-aware embedding. The inputs of the model include the entity pair, the contexts where they co-occur and the latent relation label with no semantics. After training the multi-instance model, the noisy latent labels are refined by using the trained model to generate more robust labels on the same share of data. Human supervision is only involved when assigning the cluster-level semantical relations. Entity pairs are considered to have no relation if no consensus is reached for its cluster, or the confidence of relation prediction output by the CNN model is below the threshold. Additionally, the trained model can be used to infer new knowledge from unseen data.

The model structure is shown in Fig. 2c. To deal with the wrong label issues, the model performs on multi-instance instead of a single instance. Given a set S of sentences $\{s_1, s_2, \ldots, s_{|S|}\}$ and the corresponding entity pairs, the model predicts the probability of the relation r_i in a relation set $R = \{r_1, r_2, \ldots, r_{|R|}\}$ for these entity pairs. Each sentence s_i is embedded into a fix-length vector representation v_i by a sentence encoding process including word and position embedding, a convolution layer, a piecewise max-pooling and a nonlinear layer. After the vector representation of each sentence is learned, an attention model over the set S is applied to selectively extract information from the sentences with the correct labels.

Sentence embedding In addition to the entity embedding, the embedding of the corresponding contexts and the entity positions contains complementary semantic information and potentially alleviates the entity and relation ambiguity by achieving a *context-aware* semantic embedding. Each word in a sentence s is represented in a vector consists of the corresponding word vector learned from the skip-gram model and the relative distance between the current word and the two target entities as the position information. The vectors are then concatenated and fed into a convolutional layer followed by a piecewise max-pooling and a nonlinear layer ($\tan h$). Max-pooling ensures a fixed-length vector learning for each sentence.

The piecewise max-pooling is a variation of max-pooling in that for each convolutional filter, it returns three maximums of the sentence segments determined by the two target entities other than returning a single maximum for the whole sentence [20]. The output features from all the filters with piecewise max-pooling are concatenated as a sentence vector **v** in the end.

Attention on multi-instance A vector **V** denotes the learned representation for the sentence set S:

$$\mathbf{V} = \sum_{i=1}^{|S|} \alpha_i \mathbf{v}_i \tag{4}$$

where α_i indicates the attention weights calculated by the softmax over e_i defined by the product of $\mathbf{v_i}$, a weighted diagonal matrix A and relation vector $\mathbf{r}$:

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{|R|} \exp(e_j)} \tag{5}$$

Let r denotes the relation between the target entities, θ denotes all parameters and o denotes the final output of the neural network. The conditional probability is thus defined as:

$$p\left(r_i|S,\theta\right) = \frac{\exp(o_i)}{\sum_{j=1}^{|R|} \exp(o_j)} \tag{6}$$

Therefore, given all sentence sets $\{S_1, S_2, \ldots, S_n\}$, the objective function is defined using cross-entropy:

$$J(\theta) = \sum_{i=1}^{n} \log p\left(r_i|S_i,\theta\right) \tag{7}$$

4 Experiments

We evaluate our approach of constructing biomedical knowledge graph on a selective subset of PubMed science article abstracts and validate the knowledge graph in terms of the quality of associating entities with strong relevance, the integrity of open-ended relation extraction and the capability of new knowledge inference on unseen contexts.

4.1 Data collection and annotation

Autism Spectrum Disorder (ASD) is a general classification for a broad range of disorders with a variety of issues stemming from complications with neurological development with a high incidence of 1 out of 68 individuals. No consensus has been reached about the exact cause of ASD, and the fact that the ASD diagnosis procedure is extremely complex and time-consuming makes it a topic worth more investigation. Aiming to extract domain knowledge on ASD related abstracts and further utilize it for practical applications such as automatic ASD diagnosis, we worked with doctors and gathered ASD related queries such as *autism, autistic, pervasive developmental disorder, pdd, asperger, kanner syndrome*. Pubmed articles were collected based on the queries, and we kept those with available abstracts published during the recent 10 years (2007–2017) and removed duplicates, bad links or non-English articles. In the end, we collected 24,687 English biomedical article abstracts about ASD. After data preprocessing, entity extraction and entity pairing based on co-occurrence, we

Table 1 Relevance discovery for top-*n* entity pairs

Tasks	Hits@10	Hits@20	Hits@50
Relevance	0.908	0.846	0.817

empirically selected the minimum co-occurrence to be 5, resulting in a total number of 63,330 candidate entity pair out of 6827 entities for further analysis. The data annotation for evaluating the constructed knowledge graph (relations between entities) was done by three graduate students under the supervision of therapists specialized in behavioral pediatrics and clinical psychology.

4.2 Mining strong relevance

During the clustering process, we selected a relatively large number of clusters (20 in our case) based on the Silhouette Coefficient. With more clusters than the actual types of relations, it is reasonable to assume that more than one clusters can share the same relation and one or more clusters consist of entity pairs without any relation. After the step of assigning semantical relations to the output classes by the CNN model, we discarded the entity pairs if they are with low confidence scores, or under the clusters where no explicit semantical relations could be assigned. The remainder of the entity pairs are then considered to have strong relevance between its entity components. We evaluated the correctness of relevance discovered between entities in Table 1 by calculating the precisions on top-*n* samples from each cluster ranked by their closeness to the cluster centers denoted as *Hits@n*, resulting in a total of 1000 labeled entity pairs for evaluation. We achieved a high precision on entity relevance mining averaged for all clusters on the top-n entity pairs. We are not able to provide recall due to the same issue as recommendation system evaluations that it is not feasible to measure how many facts or good recommendations exist.

4.3 Relation extraction and refinement

Based on the testing results, we assigned 6 strong semantical relations to the 20 clusters generated from kmeans++ among which seven of them are discarded because of the nonexistent relations and lack of consensuses on one specific relation. The 6 relations we extracted are as follows: *is_Symptom_of, Experimental, Causes, Affects, is_Related_to,* and *Belongs_to*. The *is_Related_to* relation is assigned when an implicit relation exists instead of an explicit one. For example, *repetitive_behaviors* and *hypersensitivity* are assigned as *is_Related_to* because both entities are considered as a symptom or an indicator of *Obsessive_Compulsive_Disorder*. The *is_Experimental* relation is between an approach or analysis and a subject, such as *western_blot* and *immunogenetics*. The *Affects* relation occurs between a type of disease and organs or genes such as *ASD* and *RAI1*. The *Belongs_to* is defined as the combination of *is_a* and *is_part_of* because in our experiments the entity pairs with these two relations are always embedded into the same cluster. To evaluate the knowledge graph we constructed, we validated the performance of our approach on the same 1000 manually labeled entity pairs we obtained on Sect. 4.2.

Table 2 Relation extraction on top-n relations

Methods	Hits@10	Hits@20	Hits@50	Hits@100
MCNN & ONE	0.90	0.85	0.80	0.72
MCNN & ATT	1.00	0.75	0.70	0.73
PCNN & ONE	1.00	0.85	0.76	0.70
PCNN & ATT	**1.00**	**1.00**	**0.82**	**0.75**

4.3.1 Baselines and model selection

The baseline selections are subject to the constrains on the applications and the resources required by the selected methods. From the applicationwised perspective, methods that extract predefined relations are not suitable to compare directly with the proposed approach which features in open-ended relation extraction. From the resourcewised perspective, collaborative construction methods which require external ontologies as inputs such as UniPro and Gene Ontology do not align with the objective of this work either because the proposed framework does not imply such requirement, which is more practical that not all biomedical-related subfields have similar resources like genes and proteins. Additionally, we only have the resources to label relations on a subset (1000 in total) of the samples since manual annotation is highly time-consuming and requires expertise. The limited sample size and the lack of external resources make it unfeasible train supervised and distantly supervised approaches. Such limitations motivate our work to solve the problem with minimum supervision.

Therefore, we evaluated the performance of four different models: **MCNN** denotes the CNN model with normal max-pooling and **PCNN** is the piecewise CNN; **ATT** denotes the attention model and **ONE** is a special case where the attention weights α are binary as described in Eq. 4. We trained the models with our entire dataset and tested their robustness to noisy labels on the same 1000 labeled entity pairs. To be clear, the input labels to these models are latent labels such as r_1, r_2, etc., and the models are expected to refine the noisy labels although the entity pairs may be tagged with wrong labels with selective attention on multi-instance learning. We measured the precisions on the top-100 most confident relations extracted from each model in Table 2. The PCNN model with selective attention consistently outperformed the other models indicating the attention model with piecewise max-pooling can benefit the classification performance. Therefore, we selected the PCNN&ATT model for further experiments to achieve the optimal performance.

4.3.2 Relation refinement

The kmeans++ is used to evaluate the performance on the relation embedding and the effectiveness of relation refinement as a component analysis. We calculated the precisions for the top-10, 20 and 50 entity pairs of each cluster ranked by their closeness to the cluster center in the embedding space. For the clusters which are assigned with the same semantical relation category, we averaged the numbers and results are shown in Table 3. The precisions generally decreased slowly with the number of entity pairs evaluated increased, indicating a consistent performance over more entity pairs. The precisions of *Belongs_to* and *is_Related_to* increased on top-20 or top-50 because these two relations are difficult to separate using kmeans++ and some of the top entity pairs are false positives while true positives are ranked low. The performance of *Belongs_to* is a sign of insufficient domain knowledge extracted

Table 3 Relation clustering on top-*n* entity pairs

Relations	Hits@10	Hits@20	Hits@50
is_Symptom_of	0.80	0.65	0.50
is_Experimental	0.90	0.95	0.88
Causes	0.50	0.48	0.42
Affects	0.57	0.58	0.52
is_Related_to	0.33	0.32	0.49
Belongs_to	0.15	0.33	0.26

Table 4 Relation refinement on top-*n* entity pairs

Relations	Hits@10	Hits@20	Hits@50
is_Symptom_of	0.90	0.80	0.50
is_Experimental	0.90	0.90	0.70
Causes	0.85	0.78	0.66
Affects	0.93	0.95	0.74
is_Related_to	0.87	0.68	0.66
Belongs_to	0.55	0.35	0.30
Average	0.83	0.74	0.59

under such relation which results in an inconsistent performance on the knowledge graph construction and further completion tasks.

As observed in Table 3, the performance of relation clustering is non-perfection and indicates that the latent relations generated by kmeans++ are noisy labels. Table 4 shows the precisions of the 6 semantical relations following the same experiment setting based on the PCNN&ATT model. Compared with the performance before relation refinement (Table 3), a significant gain is observed on the overall performance indicating the necessity of relation refinement. The precisions of *is_Related_to* and *Belongs_to* improved showing that the PCNN&ATT model is capable of extracting expressive information from the noisy labels. The performance of *is_Experimental* dropped on the top-20 and top-50 results, but still yielded a decent performance. The average precisions on top-10, top-20 and top-50 are 83.3%, 74.3% and 59.3%, respectively. Additionally, entity pairs under symmetric relations tend to form more clusters while entity pairs under asymmetric relations tends to form less which is as discussed in previous sections. For example, entity pairs under *is_Experimental* and *is_Symptom_of* form only one cluster for each relation implying that entities belong to such relations usually occur in a preserved order in biomedical articles, while entity pairs of the other relations preserve no such order which makes it difficult for rule-based knowledge extraction of such relations.

4.3.3 Qualitative analysis

Since it is not feasible to compare our algorithm directly with other approaches as discussed in Sect. 4.3.1, we conducted qualitative analysis with the existing relation extraction framework SemRep [29] at a volume level. Although in the initial work, SemRep was proposed to extract one relation (i.e., hypernymic proposition), we were able to extract multiple relations with its current implementation.[7] We used the collected PubMed abstracts as input corpus, and

[7] https://skr3.nlm.nih.gov/index.html.

Table 5 Volume level comparison with SemRep

Algorithms	Entities	Relations	Triplets
Semrep	6720	90	21,764
Semrep_Com	–	11	17,263
Semrep_Uncom	–	63	1259
MinSup (Ours)	6827	6	16,192

the statistics of obtained results are shown in Table 5 where our algorithm is denoted as minimum supervision (MinSup). We further broke down the relations extracted by SemRep into *common relations* each of which contains over 500 triplets and *uncommon relations* which contains only less than 100 triplets each, denoted as SemRep_Com and SemRep_Uncom, respectively. We further observe the following differences:

Entity recognition The number of entities extracted by the proposed framework based on MinHash [31,32] are on pair with the number generated by SemRep. SemRep tends to extract both biomedical entities and general entities, while our approach focuses more on the biomedical entities. For example, triplets such as "*is_a*(Israel, country)" and "*is_a*(length, size)" are extracted by SemRep, which are certainly considered as valid facts though not biomedical domain-specific. There are 837 different entities among the two results, most of which are caused by errors in entity normalization and word segmentation, but overall both approaches yield decent results.

Negation Relation could be changed due to one negation in the context which makes negation an important factor to consider in constructing knowledge graphs. In SemRep, negations are aggregated as new relations such as "*NEG_process_of*" and "*NEG_associated_with*". Admittedly, negated relations are very important since such relations sometimes contain useful facts. Knowlife considers negation as an additional linguistic pattern [13] and not included as types of relations. Since we apply contextual semantics during relation extraction (sentence embedding), negation information is embedded for the later classification. In this way, the negated samples are thus categorized as "*none*" relation with other candidate facts that fall below the threshold.

Relation extraction As shown in Table 5, SemRep is able to extract more fine-grained and less common relations from the corpus, and our proposed framework extracts coarse-grained common relations only. Such limitation of our work and potential solution is discussed in details in Sect. 4.5. The top-5 most common relations extracted by SemRep are: *Process_of*, *Location_of*, *Affects*, *Coexist_with* and *Part_of*. Among all the 90 relations, 63 relations occur in less then 100 triplets (1259 in total), and 55 relations occur in less than 50 triplets (630 in total), which suggests that approaches with human supervision (rule-based or supervised) such as SemRep is able to detect weak signals for fine-grained relation extraction. On the other hand, the proposed algorithm is only able to detect common relations if trained with minimum supervision. The number of extracted triplets under common relations by SemRep is in a similar scale with our results. Additionally, we compare the semantical relations extracted from both systems. In our approach and also in Knowlife [13], we consider *is_Symptom_of* as one relation, but it is categorized as one kind of *is_a* relation. The *Causes* and *Affects* relations exit in both approaches. The *is_Related_to* relations in our approach are comparable to the *Associated_with* relations from SemRep. We thus randomly selected 50 triplets from

Table 6 Hits@50 of relation extraction on comparable relations

Relations		Hits@50	
SemRep	MinSup (Ours)	SemRep	MinSup (Ours)
Causes	Causes	0.72	0.66
Affects	Affects	0.80	0.74
Associated_with	is_Related_to	0.84	0.66

each comparable relation in SemRep and manually evaluated the precisions (Hits@50) as demonstrated in Table 6. Our evaluation on SemRep is similar to the results (an averaged precision of 0.77) reported on SemMed [18] on multi-relation extractions. The precisions on SemRep are very close to the precisions on top-20 of our configuration and higher than the precisions on top-50 of our proposed approach. In general, our minimally supervised approach demonstrated comparable and promising results. One thing to note is that we excluded the 1000 testing samples from training the PCNN model, and the performance is expected to improve on other test sets if these samples are included in the training since these samples are closest to the cluster centers and contain strong relation information. However, we need to use them as test set in order to conduct a fair comparison and show the effectiveness of relation refinement process.

4.4 Knowledge inference

An important perspective of the knowledge graph evaluation is its capability of conveying sufficient domain knowledge for new knowledge inference, i.e., knowledge graph completion. We constructed a biomedical knowledge graph with a high precision and explored its application to infer new knowledge with the PCNN&ATT model. In this experiment, we left out all the documents containing the manually labeled entity pairs as the testing set and fed the kmeans++ clustering and PCNN&ATT model with the remainder of the entity pairs and their contexts as training data with only latent labels. Testing results are shown in Table 7. The numbers are averaged for the relations over the number of clusters associated each relation. The overall performance for inferring relations including *is_Experimental*, *Causes*, *Affect* and *is_Related_to* is decent with high precisions testing on the top-50 samples, suggesting that the constructed knowledge graph is a sufficient source for knowledge inference of certain types of relations and can be enriched after the initial construction. The precision of *is_Symptom_of* dropped on top-50 and the overall performance of *Belongs_to* was low which indicates insufficient information is left in the training set when we removed the top-50 entity pairs from each cluster for testing which are assumed to contain the distinguishable information for accurate relation extraction.

4.5 Discussion

As our previous experiments demonstrate promising performances in tasks of relation classification and inference, we further analyze the proposed pipeline system in the following aspects:

Table 7 Relation inference on top-n entity pairs

Relations	Hits@10	Hits@20	Hits@50
is_Symptom_of	0.70	0.70	0.38
is_Experimental	0.80	0.70	0.54
Causes	0.75	0.65	0.52
Affects	0.73	0.85	0.72
is_Related_to	0.77	0.78	0.66
Belongs_to	0.20	0.15	0.28

Domain-specific characteristics Although our approach should certainly work in general purposed knowledge extraction where less limitations exist, the two main steps involving domain-specific characteristics, which are entity recognition and entity & relation embedding, contribute to solving this more challenging problem in biomedical domain. Empirically we have compared the general-purposed NER tools such as Standford NER [14] with domain-specific tools [13,32], and the results indicate that domain-specific knowledge is necessary (such as UMLS). Additionally, the entity embeddings and relation embeddings are extracted and derived from domain-specific corpus, which are scientific articles in ASD in our case. The corpus conveys domain knowledge and generates semantical meaning for the entities and relations through embedding vectors. As a demonstration, we further cluster the entities based on the embeddings learned from skip-gram and observe that clusters mostly consist of entities with same types. For example, we find clusters consist of symptoms such as erythema, social phobias, hyperphagia and agrammatism; proteins related entities such as P38 mitogen-activated protein kinases, Granulocyte-macrophage colony-stimulating factor and anandamide; and organ related entities such as thalamic, cerebrum and ventricles.
Component analysis According to the experimental results in Sect. 4.3, improvements are obtained incrementally by adding model schemes in relation extraction and denoising through relation refinement. Because there is no supervision (manual annotation) involved during the training process, results generated by skip-gram embedding and clustering are noisy and unreliable. Therefore, a step of relation refinement is necessary and multi-instance learning scheme is appropriate to work with weak labels for that purpose. Additionally, we obtain the optimized neural network during our model selection (Table 2) by adding piecewise max-pooling. The combination of multi-instance learning with attention mechanism demonstrates significant improvements and obtains a robust performance by comparing Tables 3 and 4.
Limitations Despite we show promising results in both relation extraction (after refinement) and relation inference tasks, there are mainly two limitations we observe both results from the unsupervised training process: (1) the proposed framework is only capable of generating coarse-grained relations and (2) relations with similar semantics are difficult to distinguish. Since all relations are initialized based on the clustering results, the only factor we control that has an impact on generating relations is the number of clusters. However, weak signals from the relations with only a few observed samples (*uncommon relations*) cannot form a cluster of their own and are overwhelmed by the relations with much more samples. Additionally, it is difficult to distinguish closely related relations because the relation embeddings are also derived from unsupervised learning. As we have discussed in Sect. 4.3 that the relation *is_a* is too semantically similar to *is_part_of* in the embedding space that the relation vectors from both rela-

tions are always mixed within clusters and cannot be distinguished. For example, *right_fusiform_gyrus* is a *brain_region* while *insular_cortex* is in the *brain* but they are embedded into the same cluster. A potential solution to these limitations caused by the lack of supervision is to integrate one-shot learning relation extraction [37] where only one or a few samples of the uncommon relations are needed in order to capture the weak signals and obtain fine-grained relations, and the framework remains in a minimally supervised fashion.

Potential improvements We propose this approach to work under the extreme case where neither large-scale relation annotations nor external knowledge such as ontology or knowledge graph exist. On the other hand, the framework can potentially benefit from external domain knowledge if available. For example, if the targeted domain is protein and the targeted corpus matches the domain knowledge, UniProt ontology in this case, then it can be used either as a source of distant supervision [4], or to provide relation anchors for relation linking [15]. If large-scale relation annotations are available, it is more intuitive to formulate the problem in a supervised fashion where knowledge graph embedding approaches can be used to extract initial features. However, this is not likely the case in practical especially in biomedical domain. Additionally, progressive process has shown benefits in handling classification tasks with weak labels [36], which indicates potential improvements in our approach by adopting statistical drop-out mechanism and iteratively training the model with refined labels. Similarly, consistency constrains such as entity and relation typing check are effective to filter out candidate facts with mismatched entities as discussed in [13], which is potentially helpful in our case to further refine the false predictions caused by weakly supervision.

5 Conclusion and future work

In this paper, we have proposed a minimally supervised approach for biomedical knowledge graph construction. It is capable of extracting open-ended relations with high precisions and can be extended to other domains such as psychology. A relation refinement process based on a piecewise CNN with selective attention and multi-instance learning shows a significant improvement in the overall performance over the noisy labels generated from kmeans++ clustering. The proposed approach is shown to be accurate and effective for knowledge graph construction and the constructed knowledge graph is sufficient for further knowledge graph completion supported by the experimental results. In the future, we are interested in taking advantage of the domain knowledge embedded in the constructed knowledge graph and exploring its real-world applications, such as fine-grained medical retrieval enrichment, robust disease diagnosis, and a more interpretable representation learning on electronic health records (EHR) by a joint embedding with the constructed knowledge graph.

Acknowledgements This work is supported in part by the New York State through the Goergen Institute for Data Science and our corporate sponsors, Carestream Health and NSF awards #1704309 and #1722847.

References

1. Angeli G, Premkumar MJJ, Manning CD (2015) Leveraging linguistic structure for open domain information extraction. 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, July 26–31, 2015, vol 1. Long Papers, Beijing, China, pp 344–354

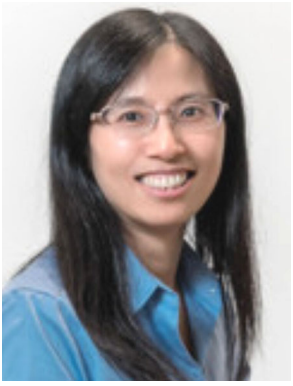
2. Arthur D, Vassilvitskii S (2007) k-means++: the advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms. Society for Industrial and Applied Mathematics, Philadelphia, pp 1027–1035
3. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al (2000) Gene ontology: tool for the unification of biology. Nat Genet 25(1):25
4. Augenstein I, Vlachos A, Maynard D (2015) Extracting relations between non-standard entities using distant supervision and imitation learning. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Florence, pp 747–757
5. Bai T, Gong L, Wang Y, Wang Y, Kulikowski CA, Huang L (2016) A method for exploring implicit concept relatedness in biomedical knowledge network. BMC Bioinform 17(9):265
6. Belleau F, Nolin M-A, Tourigny N, Rigault P, Morissette J (2008) Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform 41(5):706–716
7. Bojanowski P, Grave E, Joulin A, Mikolov T (2016) Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606
8. Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J (2008) Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD international conference on management of data. ACM, New York City, pp 1247–1250
9. Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O (2013) Translating embeddings for modeling multi-relational data. In: Advances in neural information processing systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5–8, 2013, Lake Tahoe, Nevada, United States, pp 2787–2795
10. Broder AZ (1997) On the resemblance and containment of documents. In: Proceedings compression and complexity of sequences. IEEE, Piscataway, pp 21–29
11. Consortium U (2016) Uniprot: the universal protein knowledgebase. Nucleic Acids Res 45(D1):D158–D169
12. Ernst P, Siu A, Milchevski D, Hoffart J, Weikum G (2016) Deeplife: an entity-aware search, analytics and exploration platform for health and life sciences. ACL, Vancouver, p 19
13. Ernst P, Siu A, Weikum G (2015) Knowlife: a versatile approach for constructing a large knowledge graph for biomedical sciences. BMC Bioinform 16(1):157
14. Finkel JR, Grenager T, Manning C (2005) Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics, Florence, pp 363–370
15. Galárraga L, Heitz G, Murphy K, Suchanek FM (2014) Canonicalizing open knowledge bases. In: Proceedings of the 23rd ACM international conference on information and knowledge management. ACM, New York City, pp 1679–1688
16. Hoffmann R, Zhang C, Ling X, Zettlemoyer L, Weld DS (2011) Knowledge-based weak supervision for information extraction of overlapping relations. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, vol 1. Association for Computational Linguistics, Florence, pp 541–550
17. Ji M, He Q, Han J, Spangler S (2015) Mining strong relevance between heterogeneous entities from unstructured biomedical data. Data Min Knowl Discov 29(4):976–998
18. Kilicoglu H, Fiszman M, Rodriguez A, Shin D, Ripple A, Rindflesch TC (2008) Semantic medline: a web application for managing the results of pubmed searches. In: Proceedings of the third international symposium for semantic mining in biomedicine, vol 2008. Citeseer, Princeton, pp 69–76
19. Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes PN, Hellmann S, Morsey M, Van Kleef P, Auer S et al (2015) Dbpedia-a large-scale, multilingual knowledge base extracted from wikipedia. Semant Web 6(2):167–195
20. Lin Y, Shen S, Liu Z, Luan H, Sun M (2016) Neural relation extraction with selective attention over instances. In: Proceedings of ACL, vol 1, pp 2124–2133
21. Mahdisoltani F, Biega J, Suchanek F (2014) Yago3: a knowledge base from multilingual wikipedias. In: CIDR conference 7th Biennial conference on innovative data systems research
22. Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D (2014) The stanford coreNLP natural language processing toolkit. ACL, Florence, p 55
23. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5–8, 2013, Lake Tahoe, Nevada, United States, pp 3111–3119
24. Nickel M, Murphy K, Tresp V, Gabrilovich E (2016) A review of relational machine learning for knowledge graphs. Proc IEEE 104(1):11–33

25. Niu F, Zhang C, Ré C, Shavlik JW (2012) Deepdive: web-scale knowledge-base construction using statistical learning and inference. VLDS 12:25–28
26. Pennington J, Socher R, Manning C (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
27. Ren X, Wu Z, He W, Qu M, Voss CR, Ji H, Abdelzaher TF, Han J (2016) Cotype: joint extraction of typed entities and relations with knowledge bases. arXiv preprint arXiv:1610.08763
28. Riedel S, Yao L, McCallum A (2010) Modeling relations and their mentions without labeled text. In: Machine Learning and Knowledge Discovery in Databases, European Conference, Barcelona, Spain, September 20–24, 2010, Proceedings, Part III, pp 148–163. https://doi.org/10.1007/978-3-642-15939-8_10
29. Rindflesch TC, Fiszman M (2003) The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform 36(6):462–477
30. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65
31. Siu A, Ernst P, Weikum G (2016) Disambiguation of entities in medline abstracts by combining mesh terms with knowledge. ACL, Florence, p 72
32. Siu A, Nguyen DB, Weikum G (2013) Fast entity recognition in biomedical. In: Proceedings of workshop on data mining for healthcare (DMH) at conference on knowledge discovery and data mining (KDD). ACM Press, New York
33. Surdeanu M, Tibshirani J, Nallapati R, Manning CD (2012) Multi-instance multi-label learning for relation extraction. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Association for Computational Linguistics, Florence, pp 455–465
34. Wang Z, Zhang J, Feng J, Chen Z (2014) Knowledge graph embedding by translating on hyperplanes. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27–31, 2014, Québec City, Québec, Canada, pp 1112–1119
35. Xie R, Liu Z, Sun M (2016) Representation learning of knowledge graphs with hierarchical types. In: Proceedings of the twenty-fifth international joint conference on artificial intelligence, pp 2965–2971
36. You Q, Luo J, Jin H, Yang J (2015) Robust image sentiment analysis using progressively trained and domain transferred deep networks. In: Proceedings of the twenty-ninth AAAI conference on artificial intelligence. AAAI Press, Palo Alto, pp 381–388
37. Yuan J, Guo H, Jin Z, Jin H, Zhang X, Luo J (2017) One-shot learning for fine-grained relation extraction via convolutional siamese neural network. In: IEEE international conference on big data. IEEE, Piscataway, pp 2194–2199
38. Yuan J, Holtz C, Smith T, Luo J (2016) Autism spectrum disorder detection from semi-structured and unstructured medical data. EURASIP J Bioinform Syst Biol 2017(1):3
39. Zeng D, Liu K, Chen Y, Zhao J (2015) Distant supervision for relation extraction via piecewise convolutional neural networks. In: EMNLP, pp 1753–1762

Jianbo Yuan received a B.S. degree from Harbin Institute of Technology, Harbin, China, in 2012 and two M.S. degrees from University of Rochester, USA, in 2014 and 2016. He is currently a Ph.D. student at the Department of Computer Science, University of Rochester, USA. His research interests include knowledge extraction, image and text understanding in healthcare. He has served as reviewer for IEEE Transitions on Multimedia, IEEE Transactions on Big Data and IEEE Transactions on Knowledge and Data Engineering.

Zhiwei Jin received his B.S. degree in Software Engineering from Wuhan University, China, in 2012. He obtained his Ph.D. degree in Institute of Computing Technology, Chinese Academy of Sciences, in 2018, under the supervision of Professor Yongdong Zhang. His research interests include multimedia content analysis and data mining.

Han Guo received his B.S. degree in software engineering from Shandong University, Jinan, China, in 2015, and the Master degree at the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2018. His research interests include multimedia content analysis and data mining.



Hongxia Jin is senior director of Samsung Research America Artificial Intelligence Center, heading Natural Language Processing and Data Mining research. The main areas of her current work include Natural Language Understanding, Natural Language Generation, Question Answering and Knowledge Representation for Conversational AI to help virtual assistants become more intelligent. Dr. Jin brings many years of experience in AI to Samsung, ranging from traditional symbolic AI and compiler design for fifth generation AI machine language Prolog to current data driven deep learning based approaches. Prior to Samsung, she worked at IBM research center for more than a dozen years using NLP linguistic features and modern machine learning to analyze human behavior. She has published 100+ peer-reviewed research papers and 100+ patents. She obtained her B.S in Computer Science from University of Science and Technology of China and Ph.D in Computer Science from the Johns Hopkins University in 1999.

Xianchao Zhang is Professor in the School of Software Technology, Dalian University of Technology, China, and is Participant of the Program for New Century Excellent Talents of Ministry of Education of China. He received a B.S. and M.S. degree from National University of Defense Technology majoring computational mathematics and applied mathematics and a Ph.D. degree from the School of Computer Science and Technology, University of Science and Technology of China. His research focuses on big data analysis, machine learning and data mining. Dr. Zhang is a member of ACM, IEEE and CCF.

Tristram Smith is the Haggerty-Friedman Professor of Developmental/Behavioral Pediatric Research at the University of Rochester Medical Center (URMC), where he leads federally funded studies comparing the efficacy of different interventions for children with autism spectrum disorder (ASD). He is also a clinician in URMC's Community Consultation Program. Priorly, he directed clinics for children with autism and their families in the states of California, Iowa and Washington.

Jiebo Luo joined the Department of Computer Science at the University of Rochester in 2011, after a prolific career of over 15 years with Kodak Research. He has authored over 400 peer-reviewed technical papers and holds over 90 U.S. patents. His research interests include computer vision, machine learning, data mining, social media and biomedical informatics. He has served as the Program Chair of ACM Multimedia 2010, IEEE CVPR 2012, ACM ICMR 2016 and IEEE ICIP 2017, and on the Editorial Boards of the IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Multimedia, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Big Data, Pattern Recognition, Machine Vision and Applications and ACM Transactions on Intelligent Systems and Technology. He is a Fellow of the SPIE, IAPR, IEEE, ACM and AAAI.