One-shot Learning for Fine-grained Relation Extraction via Convolutional Siamese Neural Network

Jianbo Yuan*, Han Guo†, Zhiwei Jin†, Hongxia Jin‡, Xianchao Zhang§, and Jiebo Luo*
* *Department of Computer Science, University of Rochester, Rochester, NY, USA*
† *Institute of Computing Technology, Chinese Academy of Science, Beijing, China*
‡ *Samsung Research America, Mountain View, CA, USA*
§ *School of Software, Dalian University of Technology, Dalian, China*
{*jyuan10, jluo*}*@cs.rochester.edu,* {*hanguo, zhiweijin*}*@ict.ac.cn, hongxia.jin@samsung.com, xczhang@dlut.edu.cn*

***Abstract*—Extracting fine-grained relations between entities of interest is of great importance to information extraction and large-scale knowledge graph construction. Conventional approaches on relation extraction require an existing knowledge graph to start with or sufficient observed samples from each relation type in the training process. However, such resources are not always available, and fine-grained manual labeling is extremely time-consuming and requires extensive expertise for specific domains such as healthcare and bioinformatics. Additionally, the distribution of fine-grained relations is often highly imbalanced in practice. We tackle this label scarcity and distribution imbalance issue from a one-shot classification perspective via a convolutional siamese neural network which extracts discriminative semantic-aware features to verify the relations between a pair of input samples. The proposed siamese network effectively extracts uncommon relations with only limited observed samples on the tasks of 1-shot and few-shot classification, demonstrating significant benefits to domain-specific information extraction in practical applications.**

***Keywords*-relation extraction; fine-grained; one-shot learning; siamese neural network;**

I. Introduction

Knowledge graph provides an effective way to represent real world factual information in a structured form and has shown great success in applications such as question answering and intelligent chatbot. Specifically, a *fact* in a knowledge graph is denoted as a *triplet* including a head entity, a tail entity and the semantic relation between the two entities. Extensive efforts on collaboratively constructing knowledge graphs have been made and large-scale knowledge graphs are available including YAGO [1], DBpedia [2], Freebase [3], etc. However, the knowledge coverages are not yet a satisfactory despite of the large scales of such knowledge graphs. Therefore, the automatic knowledge graph construction and completion have been motivated for alleviating the intensive labor of manual labeling and enhancing the knowledge graph's coverage and scalability. The automatic process involves entity recognition and relation extraction, among which the entity recognition is a well-investigated task. For example, the general-purpose name entity recognition (NER) [4], entities extraction based on distant supervision [5], [6], and domain-specific entity recognition based on external knowledge resources [7], [8]. Therefore, we focus mainly on the relation extraction task as proposed in previous studies [9]–[13].

Relation extraction is the process of extracting relational information from textual data and is specifically referred as extracting the relations between targeted entities for knowledge graph construction in our case [11], [14], [15]. The approaches of relation extraction are categorized as: *weakly supervised*, *supervised* and *distant supervised*. For *weakly supervised* approaches, a small set of manually selected seed samples or patterns are used for bootstrapping to extract relations [15]. The performances of such approaches depend highly on the quality of selected seed samples or patterns and have difficulties in identifying the relations which appear not frequently enough in the data. Compared with weakly supervised relation extraction, *supervised* approaches yield better performance [14] but are subject to the scale limitation of sufficient manually annotated data for *every* relation type. To address this issue, *distant supervision* is proposed [9] which heuristically aligns plain texts with existing knowledge graphs, and obtains labeled data in large scales based on which supervised relation extraction can be performed for new recognized entity pairs [6], [11], [12], [16]. The adoption of distant supervision shares the same limitation with the supervised approaches that manually crafted features are extracted and impact the overall performance significantly. For example, sentence level features is commonly used in biomedical domain [17], [18], as well as a family of embedding features such as TransE [19]. Another issue is that, the relation distribution is highly imbalanced in practice especially for *fine-grained* relation types [20], where exist only a limited number of observed samples since some relations occur not as frequently and can be hardly typed accurately using conventional classification approaches.

To address these two issues, we consider the fine-grained relation extraction task as a one-shot classification problem of which the objective is to accurately predict uncommon relations with only one or a few observed samples [21]–[23]. A siamese neural network [24], [25] is proposed to rank the similarity between pair-wised inputs based on the features

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

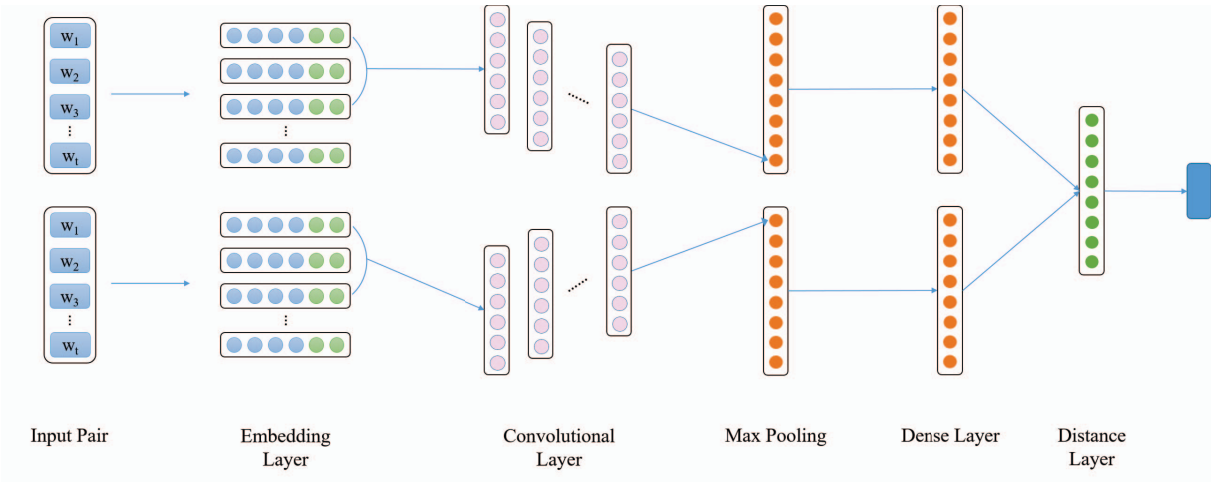

Figure 1: Convolutional Siamese Neural Network

extracted from a dual convolutional neural network (CNN) sharing the same weights. With the adoption of the CNNs, we avoid potential issues resulting from manually crafted features by mapping the targeted entities into an embedding space containing discriminative relation-related information, together with their corresponding contexts for the purpose of reducing the potential for relation ambiguity and achieving a *context-aware semantic* embedding [6], [26], [27]. Our experiments demonstrate that the proposed siamese network effectively learns a semantic embedding of the recognized entities and the corresponding contexts, and extracts relations accurately in cases of one-shot and ten-shots, and even the extreme case of zero-shot classification tasks. To the best of our knowledge, we are the first to apply one-shot learning strategy to extract fine-grained relations without requiring extensive training samples for *every* targeted relation types. Additionally, the proposed siamese network is of significant benefit to domain-specific information extraction tasks for practical applications by providing a trade-off balancing the performance against the cost of manual labeling, such as in bioinformatics where the fine-grained labeling is time-consuming and requires extensive expertise [28], [29].

II. Methodology

The network structure of the proposed convolutional siamese network is shown in Figure 1. The input to the siamese network is a pair of samples (text sequences) consisting of the targeted entities and their corresponding contexts. 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 the proposed *context-aware* semantic embedding.

A. Entity and Context Embedding

We apply an unsupervised word embedding step to convert the entities and their corresponding contexts into vectors based on the skip-gram model [30] which is effective in learning a semantic embedding from word sequences. The semantic information conveyed by each word is assumed to distribute along a window in the distributed representations. Given a window size c and the current word w_t in a word sequence $T = \{w_1, w_2, ..., w_N\}$ containing the targeted entities, the objective is to maximize the log likelihood for the surrounding words of the current word as shown in Equation 1, where the conditional probability is defined by the softmax function. Note that the input word sequence T denotes one data sample based on which entity relations are extracted and is fed into one half of the convolutional siamese network as shown in Figure 1, and our proposed siamese network takes a *pair* of the word sequences in the training and testing processes to verify whether the two input data samples contain the same relation type.

$$\frac{1}{N}\sum_{t=1}^{N}\sum_{-c\leq q\leq c, q\neq 0}\log p\left(w_{t+q}|w_t\right) \tag{1}$$

In addition to using the embeddings of the targeted entities, we extract embeddings from the whole word sequence in order to preserve the semantic information to prevent potential entity and relation ambiguities. A word w_t in the sequence T are mapped into a word vector $\mathbf{w}_t \in \mathbb{R}^{1\times n}$, and then concatenated into a matrix $\mathbf{T}$ to represent the input sequence where:

$$\mathbf{T} = \left[\mathbf{w}_1^\top, \mathbf{w}_2^\top, ..., \mathbf{w}_N^\top\right]^\top \in \mathbb{R}^{N\times n} \tag{2}$$

The number of rows in $\mathbf{T}$ are fixed to be the maximum length among all the input sequences, and we pad the rows with zero vectors if the input sequence is not long enough.

B. Position Embedding

In addition to the word level features discussed above, the sentence level features are widely used as well in the relation extraction tasks to capture the structural semantic information such as the shortest path of dependency graph [17] between the targeted entities in their corresponding contexts. Similarly to [6], [10], we compute the relative distances referred as d_1^t and d_2^t denoting the distances (namely the number of words) between the current word w_t and the two targeted entities e_1 and e_2 between which we want to extract the relation. We then map the relative distances with a *position embedding* $\mathbf{p}_t$ which is initialized randomly and updated during the training processing as defined in Equation 3:

$$\mathbf{p}_t = \left[\mathbf{d}_1^t, \mathbf{d}_2^t\right] \in \mathbb{R}^{1\times \mathrm{m}} \tag{3}$$

where $\mathbf{d}_1^t$ and $\mathbf{d}_2^t$ indicate the position embeddings of current word w_t. The position embeddings are then concatenated with the word embedding $\mathbf{w}_t$ to form the outputs of the embedding layer. Therefore, the output vector x_T given a text sequence T after the embedding layer is rewritten as:

$$\begin{gathered} x_T = [(\mathbf{w}_1, \mathbf{p}_1)^\top, (\mathbf{w}_2, \mathbf{p}_2)^\top, ..., (\mathbf{w}_N, \mathbf{p}_N)^\top]^\top, \\ x_T \in \mathbb{R}^{\mathrm{N}\times(\mathrm{n}+\mathrm{m})} \end{gathered} \tag{4}$$

C. Convolutional Siamese Neural Network

After the embedding process as shown in Figure 1, the embedded vectors are then fed into a convolutional layer, max-pooling and a dense (fully-connected) layer. Since the input text sequences are in one dimension, the convolution operation is in one dimension which is different from the normal cases as in most computer vision tasks. We select 230 convolutional filters with ReLU activation and set the window size to be 3. After the global max-pooling, we map the concatenated features into a 256×1 vector with the dense layer. The siamese neural network consists of a dual convolutional structure sharing the same hyper parameters as shown in Figure 1. Let x_i and x_j denote a pair of input vectors initialized by the embedding layer, namely the entity embedding and the position embedding, $f_w(x)$ be the convolutional and max pooling functions where w indicate the weights, and $g(\cdot)$ be the full-connected mapping from dense layer. The objective of the siamese network is to map the pair of input vectors into a small metrical distance in the feature space if they belong to the same relation type (class), or into a large distance if the inputs belong to two different relations. Given an input embedding x, the convolutional function with max pooling operation $f_w(\cdot)$ outputs a flattened vector which is used to go through the dense layer $g(\cdot)$ and compute the distance layer $dist(\cdot)$ defined by the absolute distance in Equation 5.

$$dist\left(x_i, x_j\right) = \left|g\left(f_w(x_i)\right) - g\left(f_w(x_j)\right)\right| \tag{5}$$

The output of the distance layer is then fed into a classifier layer and mapped into the range of $[0, 1]$ to predict whether the pair of inputs belong to the same category. We follow Koch *et al.* [25] and choose a sigmoid activation $\sigma(\cdot)$ (Equation 6) in the final classification.

$$p\left(x_i, x_j\right) = \sigma\left(W \cdot dist\left(x_i, x_j\right)\right) \tag{6}$$

where W denotes the weights of the classifier layer. Let $y_{i,j}$ be the relation label of input sample i and j, we optimize the binary cross-entropy loss $\mathcal{L}$ defined in Equation 7 with a $L2$ regularization to enforce the weights to learn smaller values. In the testing phase, an observed sample is paired with an unseen sample to form one *input pair* for predicting whether the two samples are from the same relation.

$$\begin{aligned} \mathcal{L}_{i,j} =& (1 - y_{i,j}) \log\left(1 - p\left(x_i, x_j\right)\right) \\ &+ y_{i,j} \log p\left(x_i, x_j\right) + \lambda \left\|W\right\|_2 \end{aligned} \tag{7}$$

When setting up the baselines (classifications without siamese network and k-nearest neighbors based on the convolutional features), we use the same convolutional network structure (shown as a half of the siamese neural network in Figure 1) to extract the same word level and sentence level features for fair comparisons. After the convolutional layer, max pooling and dense layer, the output vector $g\left(f_w(x)\right)$ is then directly applied to a sigmoid activation for the classification purposes with the distance layer removed, since in this case the input is a single word sequence other than a pair of input samples. The same cross-entropy loss in Equation 7 is used to train the deep neural network for the baseline algorithms.

III. Experiments

We train the deep neural networks on NYT dataset [20] and evaluate the fine-grained relation extraction with and without the one-shot learning strategy (siamese network) in Section III-B and III-C, respectively. For the configuration without applying siamese network, we remove the distance layer before classification since in this case we only use a single sample instead of a pair of samples as one network input. All experiments are implemented using Keras[1] with TensorFlow[2] as backend, and run on *NVIDIA Titan X* GPUs with 12 GB memory.

A. Data Preparation

The NYT dataset [20] contains large-scale text corpus subsampled from the New York Times articles between 1987-2007. In order to simulate a real world scenario and to obtain the labels for the entities and relations in the dataset, we use the labels generated by distant supervision provided by [16]. A highly imbalanced relation distribution is observed ranging from only a few to thousands of samples per

[1]https://keras.io/
[2]https://www.tensorflow.org/

relation which justifies our motivation of adopting the one-shot learning strategy for the fine-grained relation extraction task in practical use. We then mix the training and testing set provided by Ren *et al.* [16] and conduct a new partition. For the classification task, we follow [16] to remove the *None* relation [16] and remove the relations containing insufficient data with a minimum of 20 samples in order to guarantee sufficient testing samples. In the end, we obtain $\sim$108k unique relation mentions which are categorized in 22 relation types, among which 12 relation types are *common relations* (CR) and 10 relation types are *uncommon relations* (UR) with a threshold of 500 samples. For the data partition, we use 80% ($\sim$85k) of the common relations for training and the left 20% ($\sim$21k) for testing, while we hold out 10 samples from each uncommon relations (100 samples in total) for training purposes. All the other uncommon relation samples (1764 samples in total) are left out for testing. The detailed integration of the partitioned data for further experiments are discussed in Section III-B and III-C.

Table I: Relation Extraction via CNN on All Relations (AR), Common Relations (CR) and Uncommon Relations (UR).

Number of UR Samples	AR	CR	UR
1-shot	0.738	0.799	0.001
10-shot	0.743	0.804	0.002

B. Relation Extraction: Baselines

To motivate the adoption of one-shot classification, we first evaluate the performance of relation extraction without applying the siamese network. In other words, we use the initial features output by the embedding layer as inputs to feed into the same CNN structure as we proposed in Section II-C and remove the distance layer before classification. The 10 samples from each uncommon relations are used to construct two *training* configurations: (a) ***1-shot*** training which involves the common relation samples and only *one* sample random selected from the 10 held-outs for each uncommon relation, and (b) ***10-shot*** training which involves the common relations and *all* of the 10 held-out uncommon relation samples. As the results shown in Table I, the overall accuracy on extracting *all* relations are decent which means the CNN model is efficient for relation extraction. However, if we tested on the common relations and the uncommon relations separately, we observe a dramatical decrease on the performance of extracting uncommon relations. Such difference indicates that the CNN model is dominated by the common relations and hardly learns any discriminative features to extract uncommon relations. Such extreme imbalanced nature of the dataset demonstrates the necessity of more carefully curated learning strategy in order to extract the relations when the observed samples are limited. Since we are aiming to demonstrate the effectiveness of extracting uncommon relations with the proposed siamese network, the remainder of the experiments will be tested only on uncommon relation samples and all the relation samples, respectively.

Table II: Relation Extraction via k-NN on Uncommon Relations (UR) and All Relations (AR).

Number of	UR&k-NN			AR&k-NN		
UR Samples	$k=1$	$k=3$	$k=5$	$k=1$	$k=3$	$k=5$
1-shot	0.035	0.076	0.320	0.035	0.042	0.076
10-shot	0.282	0.721	0.819	0.207	0.474	0.684

Additionally, we evaluate the relation extraction performance using the k-nearest neighbors algorithm (k-NN) with $k = 1, 3, 5$ based on the same convolutional deep features and under the same experimental settings. Compared with classification sorely on the sigmoid activation, the k-NN algorithm is expected to marginally better extract uncommon relations with a reasonable. Similarly, the outputs from the dense layer are used as features for a nearest neighbor (NN) classification tested on all relations (referred as AR&k-NN) and uncommon relations only (UR&k-NN) in Table II as a baseline for the proposed siamese network. We observe a gain on performance over the results in Table I as the number of chose neighbors and the number of training samples belonging to the uncommon relations rise. By applying k-NN on the CNN features we are able to extract uncommon relations more accurately.

C. One-shot Relation Extraction

The proposed siamese network takes pairs of samples as inputs, thus we pair the training samples without any duplications. For each training sample following the partition we introduced in Section III-A, we randomly select five samples from the same class and five from different classes to pair with this sample in order to construct a balanced training set. According to our experiments, a randomly selected subset of the whole constructed training set is sufficient for learning a discriminative model and the performance is consistent despite of the scales as long as there are sufficient samples involved in the training and all selected training uncommon relation samples are included. In this experiment we have the same settings of *1-shot* and *10-shot* training which denotes the number of sample(s) from each uncommon relations included in the pairing and training process. Additionally, we add an extreme case of *0-shot* training following [25] where *none* of the uncommon relation samples are involved in the training. To be clear, the held-out samples are still in need during the testing used as the observed uncommon relation samples since the siamese network requires pairs of input samples in both training and testing. As a result, we obtain about 200k input pairs for the training process.

Since our objective is to enhance the performance of extracting the uncommon relations, we only evaluate on

Table III: Relation Extraction via Siamese Neural Network on Uncommon Relations Only.

Number of	Paired with UR			Paired with AR		
UR Samples	Top 1	Top 3	Top 5	Top 1	Top 3	Top 5
0-shot	0.334	0.766	0.928	0.196	0.604	0.760
1-shot	0.453	0.825	0.914	0.339	0.701	0.844
10-shot	0.540	0.855	0.938	0.393	0.740	0.859

the **uncommon relation samples only**. We further partition the 1764 test samples into a 585 validation/1179 testing partition to obtain the optimal parameter settings. For the testing, each validation and test sample is paired with a set of selected observed samples which are included in the training process consisting of: (a) one or ten randomly selected common relation samples, and correspondingly one random selected or all held-out uncommon relation samples for a *22-way* classification (AR observed-UR unobserved pairing for testing), respectively; and (b) one or ten uncommon relation samples only following the experiments in [25] for a *10-way* classification (UR observed-UR unobserved pairing).

For 1-shot training, only one observed sample of each uncommon relation is involved in the training, thus each testing sample should not be paired with all the 10 held-out samples. We rank the network outputs between each test sample and each selected observed sample and evaluate the accuracy of whether the top 1, 3 and 5 similarities contain the correct prediction. In Table III, each test sample is paired with only one observed sample when tested with the model trained under the 1-shot configuration, and ten observed samples are used and *averaged* for 0-shot and 10-shot settings. The top-1 accuracies indicate that the proposed siamese network is capable of extracting uncommon relations accurately with only a few observed samples required which significantly outperforms the baseline methods, and we obtain a 54% accuracy in the 10-way classification task with only 10 training samples of each uncommon relation. The proposed siamese network yields decent accuracy even on the extreme 0-shot case where none of the uncommon relation samples were used in the training phase. As shown in Table III, the accuracies of relation extraction when *paired with uncommon relations* only in the testing are generally better than the performance when all the testing samples were *paired with all relations*, which is as expected since the previous task is a relatively easier 10-way classification problem. We observe a gain on the overall performance as more of the uncommon relation samples involved in the training. The 10-shot training yields a consistently better accuracy than the 0-shot and 1-shot cases. Such improvements demonstrate that more informative labeled data can potentially benefit the performance. Therefore, the proposed fine-grained relation extraction algorithm provides a solution to balancing the trade-offs between the cost of extensive fine-grained labeling and the overall relation extraction performance.

IV. Conclusions

Relation extraction is of great importance to information extraction in practice where the fine-grained relation distribution is always highly imbalanced and obtaining sufficient labeled data for all relations is difficult. In this paper, we propose a convolutional siamese neural network which learns discriminative features for relation extraction and requires only a limited number of observed samples for uncommon relations. The proposed siamese network achieves promising results on 1-shot and 10-shot learning cases, which provides a solution to balance the cost of manual labeling and the accuracy for fine-grained relation extraction and is beneficial to practical applications such of domain-specific information extraction in healthcare and bioinformatics.

V. Acknowledgement

This work is in part supported by the New York State through the Goergen Institute for Data Science and our corporate sponsors.

References

[1] F. Mahdisoltani, J. Biega, and F. M. Suchanek, "YAGO3: A knowledge base from multilingual wikipedias," in *Seventh Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, Online Proceedings*, 2015.

[2] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer *et al.*, "Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia," *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015.

[3] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, 2008, pp. 1247–1250.

[4] J. R. Finkel, T. Grenager, and C. Manning, "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, 2005, pp. 363–370.

[5] X. Ren, A. El-Kishky, C. Wang, and J. Han, "Automatic entity recognition and typing in massive text corpora," in *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 1025–1028.

[6] D. Zeng, K. Liu, Y. Chen, and J. Zhao, "Distant supervision for relation extraction via piecewise convolutional neural networks," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal*, 2015, pp. 1753–1762.

[7] A. Siu, D. B. Nguyen, and G. Weikum, "Fast entity recognition in biomedical text," in *Proceedings of Workshop on Data Mining for Healthcare (DMH) at Conference on Knowledge Discovery and Data Mining (KDD). New York, NY, USA: ACM Press*, 2013.

[8] A. Siu, P. Ernst, and G. Weikum, "Disambiguation of entities in MEDLINE abstracts by combining mesh terms with knowledge," in *Proceedings of the 15th Workshop on Biomedical Natural Language Processing, BioNLP, Berlin, Germany*, 2016, pp. 72–76.

[9] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on the AFNLP*. Association for Computational Linguistics, 2009, pp. 1003–1011.

[10] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, "Relation classification via convolutional deep neural network," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, August 2014, pp. 2335–2344.

[11] A. Nagesh, G. Haffari, and G. Ramakrishnan, "Noisy or-based model for relation extraction using distant supervision." in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing EMNLP, Doha, Qatar*, 2014, pp. 1937–1941.

[12] I. Augenstein, A. Vlachos, and D. Maynard, "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, EMNLP 2015, Lisbon, Portugal*, 2015, pp. 747–757.

[13] Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun, "Neural relation extraction with selective attention over instances," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL, Berlin, Germany*, vol. 1, 2016, pp. 2124–2133.

[14] Z. GuoDong, S. Jian, Z. Jie, and Z. Min, "Exploring various knowledge in relation extraction," in *Proceedings of the 43rd annual meeting on association for computational linguistics*. ACL, 2005, pp. 427–434.

[15] R. Bunescu and R. Mooney, "Learning to extract relations from the web using minimal supervision," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, ACL, Prague, Czech Republic*, 2007.

[16] X. Ren, Z. Wu, W. He, M. Qu, C. R. Voss, H. Ji, T. F. Abdelzaher, and J. Han, "Cotype: Joint extraction of typed entities and relations with knowledge bases," *arXiv preprint arXiv:1610.08763*, 2016.

[17] P. Ernst, A. Siu, and G. Weikum, "Knowlife: a versatile approach for constructing a large knowledge graph for biomedical sciences," *BMC bioinformatics*, vol. 16, no. 1, p. 157, 2015.

[18] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, "Bio2rdf: towards a mashup to build bioinformatics knowledge systems," *Journal of biomedical informatics*, vol. 41, no. 5, pp. 706–716, 2008.

[19] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in neural information processing systems*, 2013, pp. 2787–2795.

[20] S. Riedel, L. Yao, and A. McCallum, "Modeling relations and their mentions without labeled text," *Machine learning and knowledge discovery in databases*, pp. 148–163, 2010.

[21] L. Fei-Fei, R. Fergus, and P. Perona, "A bayesian approach to unsupervised one-shot learning of object categories," in *9th IEEE International Conference on Computer Vision, ICCV, Nice, France*. IEEE, 2003, pp. 1134–1141.

[22] O. Vinyals, C. Blundell, T. P. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," *CoRR*, vol. abs/1606.04080, 2016.

[23] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "One-shot learning with memory-augmented neural networks," *arXiv preprint arXiv:1605.06065*, 2016.

[24] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *In Proceedings of the IEEE Conference on Computer Vision and Patten Recognition*, vol. 1. IEEE, 2005, pp. 539–546.

[25] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML Deep Learning Workshop*, 2015.

[26] Z. Wang and J. Li, "Text-enhanced representation learning for knowledge graph," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI*, 2016, pp. 1293–1299.

[27] H. Xiao, M. Huang, L. Meng, and X. Zhu, "SSP: semantic space projection for knowledge graph embedding with text descriptions," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, California, USA.*, 2017, pp. 3104–3110.

[28] T. Bai, L. Gong, Y. Wang, Y. Wang, C. A. Kulikowski, and L. Huang, "A method for exploring implicit concept relatedness in biomedical knowledge network," *BMC bioinformatics*, vol. 17, no. 9, p. 265, 2016.

[29] J. Yuan, C. Holtz, T. Smith, and J. Luo, "Autism spectrum disorder detection from semi-structured and unstructured medical data," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2017, no. 1, p. 3, 2016.

[30] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.