A Novel End-to-End Multiple Tagging Model for Knowledge Extraction

Yunhua Song
Xiangtan University
Institute of Automation, Chinese Academy of Sciences
Beijing, China
201721561958@smail.xtu.edu.cn

Hongyun Bao
Institute of Automation, Chinese Academy of Sciences
Beijing, China
hongyun.bao@ia.ac.cn

Zhineng Chen
Institute of Automation, Chinese Academy of Sciences
Beijing, China
zhineng.chen@ia.ac.cn

Jianquan Ouyang
Xiangtan University
Xiangtan, China
oyjq@xtu.edu.cn

***Abstract*—It is an emerging research topic in NLP to joint extraction of knowledge including entities and relations from unstructured text and representing them as meaningful triplets. Despite significant progresses made by recent deep neural network based solutions, these methods still confront the overlapping issue that different relational triplets may have overlapped entities in a sentence, and it is troublesome to address this issue by current solutions. In this paper, we propose a novel multiple tagging model to address the overlapping issue and extract knowledge from unstructured text. Specifically, we devise a multiple tagging scheme that transforms the problem of joint entity and relation extraction into a multiple sequence tagging problem. By using GRU as the building block for encoding-decoding, the proposed model is capable of handling the triplet overlapping problem because the decoder layer allows one entity to take part in more than one triplet. The whole network is end-to-end trianable and outputs all triplets in a sentence directly. Experimental results on the NYT and KBP benchmarks demonstrate that the proposed model siginificantly improves the recall of triplet, and consequently, achieving the new state-of-the-art in the task of triplet extraction on both datasets.**

I. Introduction

Nowadays, Knowledge Graph (KG) plays an important role in many intelligent systems and it is usually a collection of relational facts. Every fact always called knowledge,is composed of entities and a semantic relation, and can be represented as a triplet: <Entity A, Relation R, Entity B>, for instance <Beijing, Country-Capital, China>. Automatic extraction of knowledge from unstructured texts is an essential step toward constructing and refining KG. The task has gained increasing attention from NLP researchers.

So far, most previous research work utilizes a pipeline manner for extracting knowledge. Generally speaking, they decompose the task into two subtasks, one of which is Named Entity Recognition [1] and the other of which is identifying the semantic relations between two pre-assigned entities. The two subtasks can be handled independently with different models and so may propagate errors across subtasks in the process. Recent studies [3]–[5] focus on joint extraction methods, which extract entities together with relations using a single model. In order to fulfill this purpose, Yu and Lam [3], Li and Ji [4], Miwa and Sasaki [5] propose some joint learning frameworks by merging several elaborate features

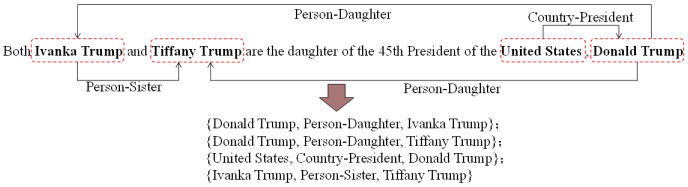

Fig. 1. A Sentence contains overlapping triplets. Entities “Donald Trump” , “Ivanka Trump” and “Tiffany Trump”, are shared in the four triplets.

simultaneously. However, they all need complicated feature engineering and heavily rely on the other NLP toolkits, which might also lead to error propagation. Fortunately, with the success of deep learning on many NLP tasks, Miwa and Bansal [6] present a neural network based method for the end-to-end entities and relations extraction. Zheng et al. [7] design a novel tagging scheme which can transform the joint extraction problem into a tagging problem and employ a LSTM-based end-to-end method to model the task. However, those models have a shortcoming on the identification of the overlapping relations from sentences.

In fact, a sentence often contains more than one relational fact, and different facts may share the same entity. For instance, shown in Fig. 1, there are at least four facts, which are represented by four triplets: <Donald Trump, Person-Daughter, Ivanka Trump>, <Donald Trump, Person-Daughter, Tiffany Trump>, <Ivanka Trump, Person-Sister, Tiffany Trump> and <United States, Country-President, Donald Trump>. Take the entity “Donald Trump” as an example. “Donald Trump” plays different roles in these facts. This phenomenon makes the existing joint models fail to extract relational triplets precisely and lose much information. Zeng et al. [8] proposes an end-to-end neural model with copy mechanism to extract relations and entities jointly. Their model can extract overlapping triplets effectively, but they should determine which relation the sentence contains firstly and then extract the entities under the constraint of the semantic relation.

Recently, the end-to-end models based on Gated Recurrent Unit (GRU), which is a simplified variant of the LSTM (Long

978-1-7281-2009-6/$31.00 ©2019 IEEE

Short-Term Memory) architecture, perform even better than those on LSTM dose in many NLP tasks, such as Arabic Named Entity Recognition [9], Sequence Chunking [10] and so on.

In this paper, we focus on the joint extraction of knowledge which are composed of two entities and one semantic relation between these two entities, especially we would like to handle the triplets overlapping issue. Thus, we propose a novel model for extracting knowledge directly, which allows one entity to freely participate in more than one triplet for dealing with the problem of triplets overlapping. Firstly, as similar as Zheng et al. [7], we utilize the tagging scheme based on a kind of novel tags that contain the information of entities and the relationships. Secondly, with the help of this tagging scheme and a multi-dimensional end-to-end model based on GRU, we transform the task into a multiple sequence tagging problem. More specifically, the model contains bi-directional Gated Recurrent Unit (Bi-GRU) layer to encode the input sentence into a sematic vector and a GRU structure as a multi-dimensional decoder to produce multiple tagging sequences. In this process, the model adds a parameter and L_2-norm to enhance the difference among multiple tagging sequences. Thirdly, according to the principles of continuity, consistency and proximity, the proposed model generates triplets. The experimental results show our multi-dimensional end-to-end model can achieve the best results on those public datasets. The major contributions of this paper are:

- We propose a multi-dimensional end-to-end model based on GRU for joint extraction of knowledge, which allows one entity with multiple tags. Therefore, the proposed model can deal with the triplets overlapping issue.
- We employ L_2-norm to measure the difference of multiple tags for the same entity, the impact of which is in the control of a parameter.
- We conduct experiments on two public datasets. Experimental results show that the proposed model outperforms the state-of-the-arts.

II. Relation Work

Automatic extraction of knowledge composed of entities and a semantic relation play an important role in knowledge graph. So far, a number of approaches to extract entities and relations have been developed, in which the extraction task have been dealt with in pipeline manner. The pipelined methods handle this task as two separated subtasks, namely Named Entity Recognition (NER) [11] and Relation Classification (RC) (e.g. [12]–[14]). Traditional NER models are made up of machine learning methods, such as HMM [15], SVM [16] or CRF [17]. With the development of deep learning, neural network integrated with machine learning to deal with NER subtask is popular (e.g. [18], [19]). The methods of relation classification can be categorized into two classes. The first contains models based on manual feature (e.g. [20], [21]), and the second is neural network based model (e.g. [22]–[24]).

While the pipeline methods neglect the information of entity extraction and relation recognition. To deal with it, joint extraction of entities and relations becomes a popular way recently. The feature engineering is adopted to joint extraction in most of researches (e.g. [4], [5], [25], [26]). However in these methods the manual feature engineering is complicated and pre-existing NLP tools is necessary. With the recent increase of interest in neural network, the joint extraction task has a new solution. For instance, the RNN-based or CNN [27]-based models (e.g. [6], [28]–[30]) was adopted to extract entities and relation jointly. They turn the entity extraction and relation extraction to two different modules, but the two modules have shared parameters. Integrating entity extraction and relation extraction into a single task is proposed for the first time by Zheng et al. [7]. They combine a novel tagging scheme with end-to-end model.

Mostly the joint extraction model consider triplet that each entity in the sentence belongs to a single valid triplet. In fact, the triplets extracting from a sentence may have same entity. These entity overlapped triplets made joint extraction models fail to extract relational triplets precisely and loss much information. Weak supervision methods [12], [25] can be used to deal with the problem, but extra informations or manual works are required. Wang et al. [31] work in supervised method, in which joint extraction is transformed into a parsing-like task , and a novel graph scheme is proposed to deal with overlap problem. The model need a lot of computing time and memory and does not conducts experiment in overlap triplets test dataset. And Zeng et al. [8] propose a parameter shared end-to-end model which extracts relations firstly and then copy entities from original sentence, which can processing overlapping triplets effectively.

In this paper, we adopted sequence tagging mechanism (e.g. [16], [32], [33]) to extract entities and relation. The tagging scheme in Zheng et al. [29] is expanded to deal with overlapping triplets. And a seq-to-seq model based on GRU (Gated Recurrent Unit) [34] is proposed in this paper.

III. Model

We propose a multi-dimensional end-to-end model based on GRU for joint extraction of knowledge. In this section, first we introduce the background for our proposed model. Second, we show some principles how the proposed model can generate triplets based on those predicted tags. Third, we illustrate the details of the proposed model.

A. Background

We introduce some notations first. We have a set of sentences $\{s_i\}$ ($i = 1, ..., N_s$, N_s is the number of sentences in the set) and a set of word vectors $\{w_j\}$ ($j = 1, ..., N_w$, N_w is the number of the different words in the set) by running word2vec. As similar as Zheng et al. [7], we also employ the BIOES (Begin, Inside, Outside, End, Single) tagging scheme, representing the word position of the entity.

In this tagging scheme, we augment the set of tags by fusing the relation type information from a predefined set of relations and the entity position information of the triplet. Every tag except the tag "O" has three parts, one is the position of an entity,

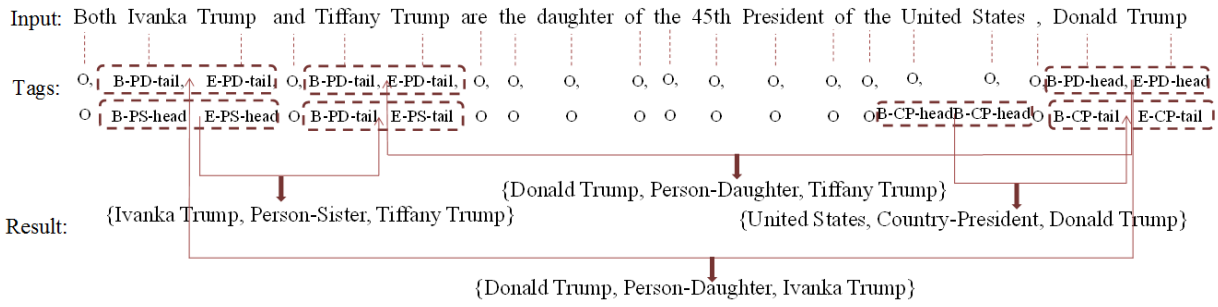

Fig. 2. Gold standard annotation for an example sentence based on the proposed tagging scheme, where "PD" means "Person-Daughter", "PS" means "Person-Sister" and "CP" means "Country-President".

the other is the type of relations and another is the position of triplet. For example, the tag "B-CP-head" means the word tagged with "B-CP-head" is the "begin" position in the "head" entity of the semantic relation "CP" (abbreviation of "Country-President"). In addition, tag "O" means the word tagged with "O" is beyond of the semantic relations from the predefined set. Thus, the total number of tags is $N_t = 2 * 4 * |R| + 1$, where R is the size of the predefined relation set. Unlike the model proposed by Zheng et al. [7], we transform the joint extraction of knowledge task into a multiple sequences tagging problem that one sentence can have more than one tagging sequence. Fig. 2 is an example of how the results are generated. As shown in Fig. 2, the input sentence contains four triplets: <Donald Trump, Person-Daughter, Ivanka Trump>, <Donald Trump, Person-Daughter, Tiffany Trump>, <Ivanka Trump , Person-Sister, Tiffany Trump> and <United States, Country-President, Donald Trump>, where "Person-Daughter" "Person-Sister" and "Country-President" are the predefined relation types. The words "Donald" "Trump" "United" "States" "Ivanka" "Trump" "Tiffany" "Trump" all have a role in generating triplets. Meanwhile, the words "Donald" "Trump" take part in three different triplets. This sentence is tagged with two diverse sequences. For example, the word of "Donald" is the first word of entity "Donald Trump" and "Donald Trump" is the head entity of the relation "Person-Daughter" as well as the tail entity of the relation "Country-President", therefore its tag is "B-PD-head" in the first sequence and "B-CP-tail" in the second sequence.

B. Principles

Based on the tagging sequences, we should define the basic principles to generate the triplets. As defined above, a word in a sentence has a tag except O with three parts : p_1-R-p_2, where $p_1 \in \{B, I, E, S\}$ and $p_2 \in \{head, tail\}$.

- **the principle of continuity**
  The principle of continuity demands that the words forming an entity should be continuous. That is to say, there are no other words with tags "O" between these words. As Fig. 2 shows, "Ivanka Trump and Tiffany Trump" is not an entity because of "and" with the tag "O" between "Ivanka Trump" and "Tiffany Trump".
- **the principle of proximity**
  (1) The words can form an entity only if their tags share the same relation type "R" and the same p_2. Then, the entity has the tag "R-p_2";
  (2) Two entities can form a triplet only if their tags share the same relation "R" and p_2 of one entity is $head$ and p_2 of the other entity is $tail$.
- **the principle of consistency**
  As similar as Zheng et al. [7], if a sentence contains two or more triplets with the same relation type, we combine every two entities nearest.

Under the principles mentioned above, we generate the triplets from the tagging sequences of the sentence.

C. Proposed Model

In this section, we introduce the details of our proposed model based on the Gate Recurrent Unit (GRU) cell (Cho et al. [34]). The framework of the proposed model is shown in Fig. 3. Firstly, the proposed model encodes a variable-length sentence into a fixed-length vector representation and then decodes this vector into multiple variable-length sequences. L_2-norm is employed to enhance the difference among multiple tagging sequences. In this paper, we call the proposed model as Multi-GRU for short.

*1) **Encoder**:* The encoder of our proposed model use a bi-directional Gated Recurrent Unit (Bi-GRU), which embedding an input sentence $s = \{w_1, ..., w_N\}$ to an output sequence $H = \{[h_1^f, h_1^b], ..., [h_N^f, ..., h_N^b]\}$. N is the number of the words in the input sentence. The process of Bi-GRU has two parallel GRU layers: forward GRU layer and backward GRU layer. In the GRU architecture shown in Fig. 3(b), for each time step t with the input w_t and previous hidden state h_{t-1}, we compute the updated hidden state $h_t^i = GRU_i(w_t, h_{t-1})$ as following:

$$r_t = \sigma\left(W_r^e h_{t-1} + U_r^e w_t + b_r^e\right) \quad (1)$$

$$z_t = \sigma\left(W_z^e h_{t-1} + U_z^e w_t + b_z^e\right) \quad (2)$$

$$\bar{h}_t = \tanh\left(W_h^e r_t * h_{t-1}^i + U_h^e w_t + b_h^e\right) \quad (3)$$

$$h_t^i = (1 - z_t) * h_{t-1}^i + z_t * \bar{h}_t \quad (4)$$

where r is the reset gate, z is the updata gate, and $\bar{h}$ is the reset hidden unit based on current input word embedding. $W^e(.) \in \mathbf{R}^{d_h \times d_h}$, $U^e(.) \in \mathbf{R}^{d_h \times d_w}$ and $b^e(.) \in \mathbf{R}^{d_h}$ are the learn-able

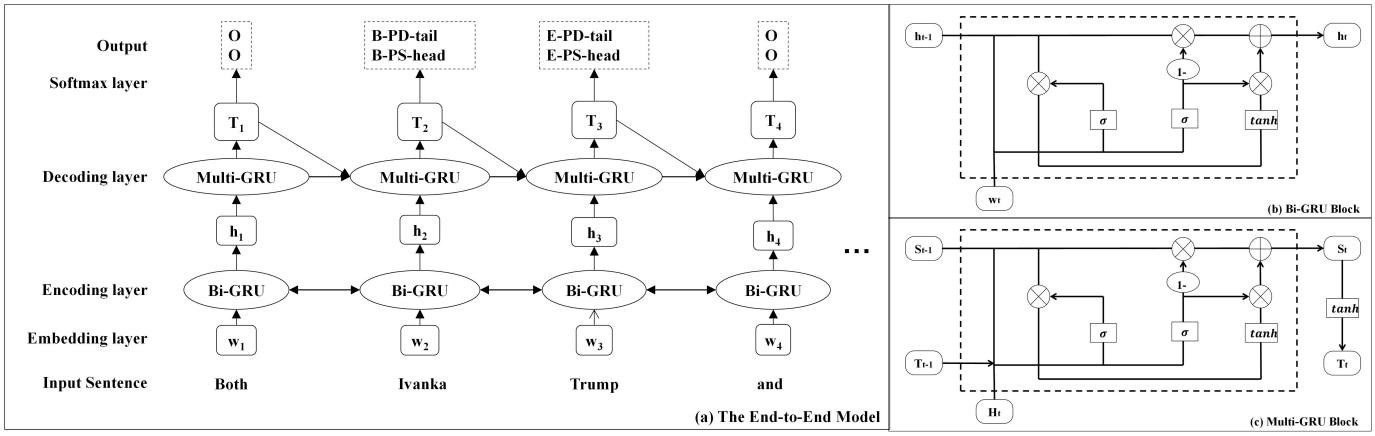

Fig. 3. An illustration of the proposed model. (a): The framework of the proposed model, (b): The GRU memory block in Bi-GRU encoding layer, (c): The GRU memory block in Multi-GRU decoding layer.

weights and bias term respectively. Thus the forward GRU output is denoted as $h_t^f = GRU_f(w_t, h_{t-1}^f)$, and the backward GRU output is denoted as $h_t^b = GRU_b(w_t, h_{t-1}^f)$. The t-th word encoding information is described as $h_t = [h_t^f, h_t^b]$.

*2) **Decoder**:* Different with Zheng et al. [7], we achieve layer, whose architecture is shown in Fig. 3(c). In this part, we utilize K as the number of the tags a word has. For generating K tags of a word, we duplicate the hidden state h_t with K times named H_t, which is transposed to the decoding layer to predict the tagging sequences. Then the multi-dimensional GRU computes the current hidden state S_t by H_t, the previous decode hidden state and the decoder output T_t:

$$R_t = \sigma\left(W_r^d S_{t-1} + U_r^d H_t + V_r^d T_{t-1} + b_r^d\right) \quad (5)$$

$$Z_t = \sigma\left(W_z^d S_{t-1} + U_z^d H_t + V_z^d T_{t-1} + b_z^d\right) \quad (6)$$

$$\bar{S}_t = \tanh\left(W_s^d R_t * S_{t-1} + U_s^d H_t + V_s^d T_{t-1} + b_s^d\right) \quad (7)$$

$$S_t = (1 - Z_t) * S_{t-1} + Z_t * \bar{S}_t \quad (8)$$

$$T_t = \tanh\left(W_T^d S_t + b_T^d\right) \quad (9)$$

where $S_t \in \mathbf{R}^{K*d_s}$ represents the current multiple hidden states of decoder, $\bar{S}_t$ is the reset hidden state, $H_t \in \mathbf{R}^{K*d_h}$ means the current encoding information denoted as concatenation of K h_t, and $T_t \in \mathbf{R}^{K*d_t}$ is the embedding of current predicted tag. $W^d(.), U^d(.), V^d(.), b^d(.)$ are the learnable weights and the bias parameters. Note that all of those parameters excepting the parameter $b^d(.)$ are quasi-diagonal matrixes, for instance:

$$W_r^d = \begin{bmatrix} W_r^{d_1} & \dots & \mathbf{0} \\ & \dots & \\ \mathbf{0} & \dots & W_r^{d_K} \end{bmatrix}, W_r^{d_i} \in \mathbf{R}^{d_s \times d_r}, i \in 1, ..., K \quad (10)$$

The tag distribution matrix Y_t is computed based on the current tag embedding T_t:

$$Y_t = W_y T_t + b_y \quad (11)$$

where W_y is the quasi-diagonal matrix as (11). Before computing the probability distribution of every tag, $Y_t = \{tag_t^1, ..., tag_t^K\}$, in which every $tag_t^i \in \mathbf{R}^{d_y}$ represent the predicted relationship distribution corresponding to the current word. A softmax layer is used to normalize the predicted distribution probabilities:

$$p_t^{i_l} = \frac{exp\left(tag_t^{i_1}\right)}{\sum_{m=1}^{d_y} exp\left(tag_t^{i_m}\right)}, i \in 1, ..., K \quad (12)$$

*3) **Objective Function**:* In the model, each input word has K predicted tag distribution probabilities corresponding to K relationship tags. The objective function of our proposed model has two sum items. The first term represents the similarity between the predicted tag and the true tag, which is achieved by maximizing the log-likelihood function. The second term denotes the diversity among the different tagging sequences, which is achieved by maximizing the second-order norm distance. Therefore, the objective function is defined as:

$$L = max \sum_{|D|} \sum_{t=1}^{N} (\sum_{i=1}^{K} H(y_t^i, p_t^i) * M(\alpha) + \sum_{1 \leqslant i < j \leqslant K} (\lambda \| p_t^i - p_t^j \|)) \quad (13)$$

where $|D|$ means the size of sample space, N is the number of words in the input sentence, y_t^i represents the true tag vector, and p_t^i,p_t^j are probability distributions of the predicted tags. And $H(y, p)$ is the log-likelihood function to evaluate the difference between the target value and the predicted value:

$$H(y, p) = \sum_{|y|} y_i log(p_i) \quad (14)$$

And $M(\alpha)$ is the biased function that makes the model inclined to predict valid relations, which is defined as follow:

$$M(\alpha) = \begin{cases} \alpha, & if tag_t^i \neq \text{`}O\text{`} \\ 1, & if tag_t^i = \text{`}O\text{`} \end{cases} \quad (15)$$

Where α is the hyper-parameter that controls the influence of the non"O" tags. The λ in objective function is the other hyper-parameter which is used to adjust the difference among multiple predicted tagging sequences.

TABLE I
THE PERCENTAGE OF THE NUMBER OF THE OVERLAPPING SENTENCES ON THE NYTT AND KBP DATASETS.

	NYT_train	NYT_test	KBP_train	KBP_test
$K_i = 1$	71.1%	100%	67.4%	97.7%
$K_i \leq 2$	89.6%	100%	95.9%	100%
$K_i \leq 3$	95.0%	100%	97.2%	100%
$K_i \leq 4$	99.6%	100%	99.6%	100%

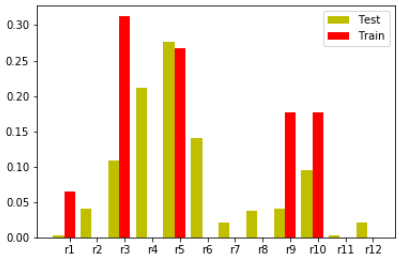

Fig. 4. The distribution of relationships in the KBP training and test datasets. r_i represents one of the predefined relationship.

IV. EXPERIMENT AND RESULT

A. Dataset

In order to evaluate the performance of our method, we conduct a series of experiments with the most powerful methods on two publicly available datasets. The first is New York Times (NYT) dataset that is developed by distant supervision method (Riedel et al. [35]). The second is KBP dataset, which contains 1.5M sentences sampling from 780K Wikipedia articles (Ling and Weld [36]).

As described in III-A, we use the novel tagging scheme to tag the training sentences. Firstly, in a greedy way, we traverse every triplet to transfer its corresponding sentence as a tagging sequence. Secondly, for every sentence, we get a candidate tagging sequence list. we chose one sequence from the list as as the baseline sequence, and use other sequences to enhance the baseline sequence until when one word should be tagged with two different tags from the set of tags except "O", and delete those used sequences from the list. If there are some sequences left, we should repeat the above process until none left in the list and increase the value of K_i. Thirdly, we have a statics about the K_i for each sequence as the baseline sequence and get the least K_i as the count of the tagging sequences for the sentence. Take the sentence of the Fig. 2 as an example, there are four triplets in the sentence which can be tagging in 2 tag sequences at least. We have the statistic about the percentage of $K_i = 1, 2, 3, 4$ sentences on the above two datasets shown in Table I.

There are 24 kinds of predefined relations in the NYT dataset. The original datasets containing 235982 sentences as the training set, and 395 sentences as the test set. Because there are three quarters sentences in the training set which have none valid triplets, and the original test set have only one sentence that has overlapped triplets. Therefore the original dataset cannot be used to verify the validity of our model in overlapping triplets. Following the previous work (Zeng et al. [8]), we filter the sentences with more than 50 words and the sentences containing no predefined relation triplets, and 63733 sentences are left. We randomly select 5% sentences from $K_i = 1, 2, 3, 4$ sentences in original training set as the new test set, and use the remaining data as the new training set.

The KBP dataset selects entities and relations matched to FreeBase. There are 12 kinds of predefined relationships remained. We found there are 6 relationships in the origin training set and 12 relationships in the test set, and the distribution of the relationships is shown in Fig. 4. Due to the relationship distribution between training set and test set is inconsistent, and the inherent supervisor nature precludes our predicting those relations that didnt appear in training set, we filter the sentences with more than 50 words and the sentences containing no predefined relation triplets, and 23400 sentences are left. And 5% sentences are randomly selected as the new KBP test set, and the rest is used as the new KBP training set.

B. Metrics and Baselines

Following Zheng et al. [7] and Zeng et al. [8], we also employ the micro Precision, Recall, and F1 score to evaluate the results. The Triplet is regarded as a correct one only if its relation and entities are both correct. We compare our models with three methods: the MultiDecoder model (Zeng et al. [8]), the NovelTagging model (Zheng et al. [7]) and the GraphTagging model (Wang et al. [31]). The MultiDecoder model is an end-to-end neural model with the copy mechanism for relational facts extraction, and it can handle the overlap relation issue. The NovelTagging model propose a novel tagging scheme to transform the joint extraction of knowledge task into a tagging problem and has good performance on NYT dataset. The GraphTagging model is a neural transition-based approach for joint entity and relation extraction, and achieves the state-of-the art F-scores on NYT dataset. We directly run the code released by the NovelTagging model and the MultiDecoder model to acquire the results. Meanwhile, we also compare our proposed framework with the classical neural network LSTM named as Multi-LSTM for short.

C. Setting

Our model is an end-to-end neural model based on GRU with a multiple tagging scheme. The input of this model is the set of word vectors initialed by running word2vec (Mikolov et al. [37]). The dimension of the word embedding is d = 300. And then, we let dropout regularize our Multi-GRU network with the dropout ratio of 0.3. The number of Bi-GRU cell units in encoding layer is 300 and the number of gru cell units in decoding layer is 600. Other parameters are set as α = 10 and λ =10 of each dataset. Meanwhile, the K means dimension of the output generated by the multi-dimensional decoder is set

TABLE II
RESULTS OF DIFFERENT MODELS IN THE NYT DATASETS AND KBP DATASETS

Model	NYT_new			NYT_rigin			KBP_new			KBP_rigin		
	Pre.	Rec.	F1 score	Pre.	Rec.	F1 score	Pre.	Rec.	F1 score	Pre.	Rec.	F1 score
MultiDecoder[8]	0.626	0.589	0.607	0.352	0.497	0.412	0.232	0.214	0.223	0.160	0.175	0.175
NovelTagging[7]	**0.826**	0.534	0.649	0.615	0.414	0.495	**0.513**	0.269	0.353	**0.396**	0.223	0.285
GraphTagging[29]	/	/	/	**0.643**	0.421	0.509	/	/	/	/	/	/
Multi-LSTM	0.795	0.603	0.692	0.608	0.477	0.535	0.453	0.331	0.382	0.332	0.264	0.294
Multi-GRU	0.812	**0.616**	**0.700**	0.597	**0.499**	**0.543**	0.480	**0.337**	**0.395**	0.301	**0.352**	**0.324**

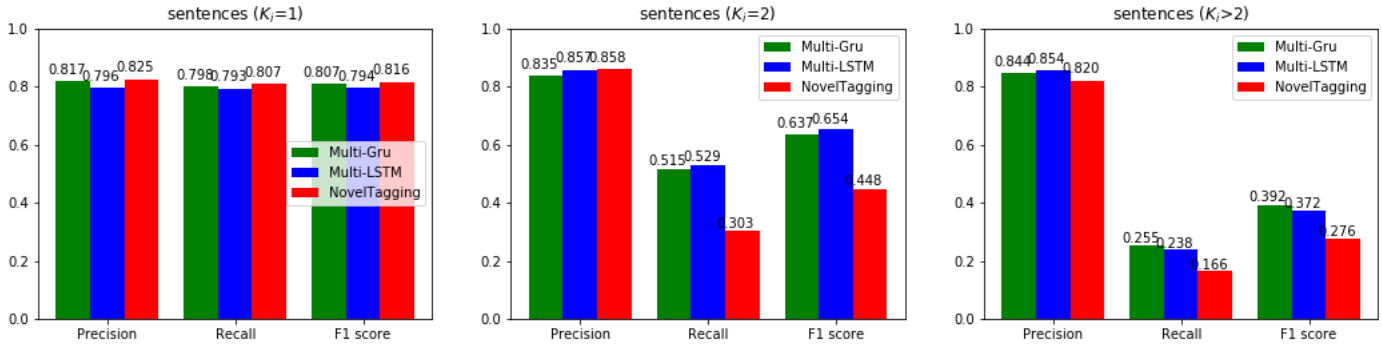

Fig. 5. Results of NovelTagging, Multi-LSTM, and Multi-GRU model with different K_i on the new NYT test set.

to 2. As Table I shows, K = 2 can deal with most of sentences. We run 10 times for every experiment then report the average results as Table II shows.

D. Experimental Results

To observe the performance of our model in extracting triplets whatever they are overlapped, we experiment on the original NYT, KBP, new NYT, new KBP test set. The results are shown in Table II, where we indicate the best performance of those models in bold type. From Table II, we can observe that on the both NYT and KBP datasets, both Multi-LSTM and Multi-GRU improve the F1 score and Recall in comparison with other methods. That is, the multiple tagging scheme is useful for joint extraction of knowledge. But the precision rate of the two proposed models are less than the Graph Taggging or the NovelTagging model. In original NYT test set, the proposed Multi-GRU model outperforms them in terms of F-scores by 13.1% , 4.8% 3.4% and 0.8% relative to MultiDecoder [8], NovelTagging [7], GraphTagging [31] and Multi-LSTM respectively. Due to the lack of the code of GraphTagging, we compare our model only with the MultiDecoder and NovelTagging model in the remaining three test sets. In the new NYT test set, the F1 score of Multi-Gru outperforms them by 9.3%5.1% and 0.8% relative to Multi-Decoder [8], NovelTagging [7] and Multi-LSTM respectively. On the original KBP test set and the new KBP test set, the proposed Multi-GRU model F1 score has increase by 4.2%, and 3.9% relative to NovelTagging respectively.

And we also can see that, in both NYT and KBP datasets, the NovelTagging model achieves the highest precision value and the MultiDecoder is more balanced in terms of micro Precision and Recall. As analyzed in Zeng et al. [8], the NovelTagging only allows one entity to participate at most one triplet. While, MultiDecoder firstly generates the relation and then applys copy mechanism to find entities so that one entity can joint in multiple triplets. However, the performance of the MultiDecoder depends on the recognition of relations, the lower the recognition rate of which is, the lower the Precision of the MultiDecoder has. Different from the NovelTagging model and the MultiDecoder model, on one hand the proposed model allows one entity with multiple tags, and on the other hand, the extra objective function has great influence of predicted influence. Finally, we compare the ability of GRU-based model with the ability of LSTM-based model. As shown in the results, the former performs better in terms of micro Precision, Recall and F1 score.

We also compare our model with NovelTagging on the results of different K_i sentence classes on the new NYT test set, shown in Fig. 5. As we can see, the NovelTagging [8] can deal with sentences with K_i=1 outstandingly. And for sentences with K_i=2, the Multi-GRU and Multi-LSTM achieve greater improvement with Recall and F1 score than NovalTagging. And when $K_i \geq 2$,the proposed model outperforms the NovelTagging in the three evaluation metrics.

E. Parameter Analysis

Parameter α in (13) is the weight that indicates how important the non "O" tags are for the objective function, and parameter λ in (13) is the weight which controls the degree of the difference among the tags one entity has at the same time.

TABLE III
OUTPUT FROM NOVELTAGGING AND THE PROPOSED MODEL. STANDARD S1 REPRESENTS THE GOLD STANDARD. THE BLUE PART IS THE CORRECT RESULT, AND THE RED ONE IS THE WRONG ONE.

Standard S1:	Thus , the Bishop of Rome has always been held by the others to be fully sovereign within his own area , as well as FirstAmong-Equals, due to the traditional belief that the Apostles [Saint Peter]E-CD-head; O and [Saint Paul]E-CD-head; O were martyred in [Rome]E-CD-tail; E-CD-tail.
NovelTagging [7]:	Thus , the Bishop of [Rome]E-CD-tail has always been held by the others to be fully sovereign within his own area , as well as "First-Among-Equals" , due to the traditional belief that the Apostles [Saint Peter]E-CD-head and [Saint Paul]E-CD-head were martyred in [Rome]E-CD-tail.
The Proposed Model:	Thus , the Bishop of Rome has always been held by the others to be fully sovereign within his own area , as well as "First Among-Equals" , due to the traditional belief that the Apostles [Saint Peter]E-CD-head; O and [Saint Paul]O; E-CD-head were martyred in [Rome]E-CD-tail; E-CD-tail.
Standard S1:	To create the cover art for "Guero" , he selected an artist , [Marcel Dzama]E-PL-head , who is a member of the Royal Art Lodge , a collective in [Winnipeg]E-PL-tail that produces childlike art .
NovelTagging [7]:	To create the cover art for "Guero" , he selected an artist , [Marcel Dzama]E-BP-head , who is a member of the Royal Art Lodge , a collective in [Winnipeg]E-DP-tail that produces childlike art .
The Proposed Model:	To create the cover art for "Guero" , he selected an artist , [Marcel Dzama]E-LP-head; O , who is a member of the Royal Art Lodge , a collective in [Winnipeg]E-LP-tail; O that produces childlike art .

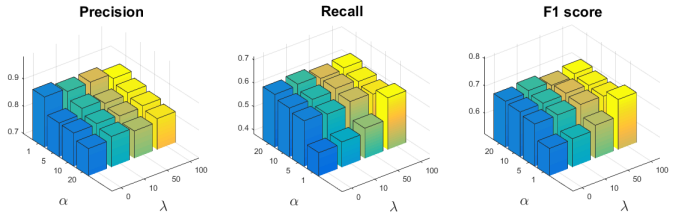

Fig. 6. Multi-GRU Results varying with diffrent parameters (i.e., α and λ) on the new NYT test set.

The different parameters of α and λ will achieve difference results. In order to further analyze the interactive effect of α and λ, we visualize the prediction results on the new NYT test set in $\alpha \in \{1, 5, 10, 20\}$ and $\lambda \in \{0, 10, 50, 100\}$ with the 3D histograms shown in Fig. 6. The smaller α is, the higher precision would be and the smaller recall would be. And when α=1 and λ=100, the Multi-GRU Model can achieve the best F1 score.

F. Case Study

In this section, we observe the prediction results of our model and the NovelTagging model. And then we select two representative examples to illustrate the advantages of our model as Table III shown. There are two cases shown in Table III and each case contains three rows, such as ground truth, the result of NovelTagging and the result of our method.

As demonstrated by sentence 1, the overlapping triplets < saint peter, Country-Person-Death, Rome> and <saint paul, Country-Person-Death, Rome> have been extracted by the two models. However, the first "Rome" in the sentence isn't related with each head entity. Every word in our model has 2 tags, and then according to some principles the results can be generated. And the nearest tail entity with each head entity is the second "Rome".

In the second example, because of the long distance of "Marcel Dzama" and "Winnipeg", the NovelTagging predicts "Marcel Dzama" as the head entity of relationship "Birth-Place" and predicts "Winnipeg" as the tail entity of relationship "Death-Place". While Our model has obtained the right relation with more detailed feature resperisentations, then, the triplet <Marcel Azama, Lived-Place, Winipeg> can be generated.

V. CONCLUSION

In this paper, with a multiple tagging scheme, we propose a end-to-end neural model based on GRU for joint entity and relation extraction. Our model can extract overlapping relations with multiple tagging sequences and extra objective function. We evaluate the effectiveness of our model on two public datasets. The experimental results show that, our model can achieve the best F1 scores compared with several recent joint extraction methods. This challenging task is far from being solved. Our future work will concentrate on how to obtain the triplets of unknown relation, which only appears in the test set.

ACKNOWLEDGMENT

This work is supported by the Natural Science Foundation of China (No. 61702514 and No. 61772526).

REFERENCES

[1] I. Hendrickx, S.N. Kim, Z. Kozareva,P. Nakov, D.O. Seaghdha, S. Pado, M. Pennacchiotti, L. Romano and S. Szpakowicz, "Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals," In Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions. Association for Computational Linguistics, 2009, pp. 94–99.

[2] B. Rink and S. Harabagiu, "UTD: Classifying Semantic Relations by Combining Lexical and Semantic Resources," In Proceedings of 5th International Workshop on Semantic Evaluation, pp. 256–259.

[3] X. Yu and W. Lam, "Jointly identifying entities and extracting relations in encyclopedia text via a graphical model approach," In Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010, pp. 1399–1407.

[4] Q. Li and H. Ji, "Incremental joint extraction of entity mentions and relations," In Proceedings of the 52rd Annual Meeting of the Association for Computational Linguistics, 2014, pp. 402–412.

[5] M. Miwa and Y. Sasaki, "Modeling joint entity and relation extraction with table representation," In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1858–1869.

[6] M. Miwa and M. Bansal, "End-to-end relation extraction using lstms on sequences and tree structures," In Proceedings of the 54rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, 2016, pp. 1105–1116.

[7] S. Zheng, F. Wang, H. Bao, Y. Hao, P. Zhou, and B. Xu, "Joint extraction of entities and relations based on a novel tagging scheme," In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, pp. 1227–1236.

[8] X. Zeng, D. Zeng, S. He, K. Liu and J. Zhao, "Extracting Relational Facts by an End-to-End Neural Model with Copy Mechanism," In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, 2018, pp. 506–514.

[9] M. Gridach and H. Haddad, "Arabic Named Entity Recognition: A Bidirectional GRU-CRF Approach," International Conference on Computational Linguistics and Intelligent Text Processing. Springer, Cham, 2017, pp. 264–275.

[10] F. Zhai, S. Potdar, B. Xiang and B. Zhou, "Neural Models for Sequence Chunking," AAAI, 2017, pp. 3365-3371.

[11] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," Lingvisticae Investigationes, vol. 30(1), 2007, pp. 3–26.

[12] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer and D.S. Weld, "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-Volume 1. Association for Computational Linguistics, 2011, pp. 541–550.

[13] M. Surdeanu, J. Tibshirani, R. Nallapati and C.D. Manning, "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, 2012, pp. 455–465.

[14] P. Zhou, J. Xu, Z. Qi, H. Bao, Z. Chen, B. Xu, "Distant supervision for relation extraction with hierarchical selective attention," Neural Networks, 108, 2018, pp. 240–247.

[15] G.D. Zhou, J. Su, "Named entity recognition using an HMM-based chunk tagger," In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002, pp. 473–480.

[16] Y. Benajiba, M. Diab and P. Rosso, "Arabic named entity recognition: An svm-based approach," In Proceedings of 2008 Arab International Conference on Information Technology (ACIT). Amman, Jordan: Association of Arab Universities, 2008, pp. 16–18.

[17] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," InProceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics, 2003, pp. 188–191.

[18] Z. Huang, W. Xu and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," arXiv preprint arXiv:1508.01991, 2015.

[19] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami and C. Dyer, "Neural Architectures for Named Entity Recognition," In Proceedings of NAACL-HLT, 2016, pp. 260–270.

[20] S. Miller, H. Fox, L. Ramshaw, and R. Weischedel, "A novel use of statistical parsing to extract information from text," In 1st Meeting of the North American Chapter of the Association for Computational Linguistics, pp. 226–233, Seattle, Washington, April 29May 4 2000.

[21] B. Rink and S. Harabagiu, "Utd: Classifying semantic relations by combining lexical and semantic resources," In Proceedings of the 5th International Workshop on Semantic Evaluation, 2010, pp. 256–259.

[22] R. Socher, B. Huval, C.D. Manning and A.Y. Ng, "Semantic compositionality through recursive matrix-vector spaces," Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Association for Computational Linguistics, 2012, pp. 1201–1211.

[23] D. Zeng, K. Liu, S. Lai, G. Zhou and J. Zhao, "Relation classification via convolutional deep neural network," Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 2335–2344.

[24] Ji G, Liu K, S. He and J. Zhao, "Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions," AAAI, 2017, pp. 3060–3066.

[25] X. Ren, Z. Wu, W. He, M. Qu, C.R. Voss, H. Ji, Tarek F Abdelzaher, and Jiawei Han, "Cotype: Joint extraction of typed entities and relations with knowledge bases," In Proceedings of the 26th WWW international conference, 2017, pp. 1015–1024.

[26] L. Liu, X. Ren, Q. Zhu, H. Gui, S. Zhi, H. Ji and J. Han, "Heterogeneous Supervision for Relation Extraction: A Representation Learning Approach," Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 4–56.

[27] Z. Chen, S. Ai, C. Jia, "Structure-Aware Deep Learning for Product Image Classification," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 15, no.1s, 2019, pp. 04:1-20.

[28] P. Gupta, H. Schtze and B. Andrassy, "Table filling multi-task recurrent neural network for joint entity and relation extraction," Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 2537–2547.

[29] S. Zheng, J. Xu, H. Bao, Z. Qi, J. Zhang, H. Hao and B. Xu, "Joint learning of entity semantics and relation pattern for relation extraction," Joint european conference on machine learning and knowledge discovery in databases, Springer, Cham, 2016, pp. 443–458.

[30] M. Zhang, Y. Zhang and G. Fu, "End-to-End Neural Relation Extraction with Global Optimization," Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1730–1740.

[31] S. Wang, Y. Zhang, W. Che and T. Liu, "Joint Extraction of Entities and Relations Based on a Novel Graph Scheme," IJCAI, 2018, pp. 4461–4467.

[32] A. Vaswani, Y. Bisk, K. Sagae and R. Musa, "Supertagging with lstms," Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016 pp. 232–237.

[33] A. Katiyar and C. Cardie, "Investigating lstms for joint extraction of opinion entities and relations," Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, 2016, pp. 919–929.

[34] K. Cho, M.B. Van, C. Gulcehre C, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724–1734.

[35] S. Riedel, L. Yao and A. McCallum, "Modeling relations and their mentions without labeled text," Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, 2010, pp. 148–163.

[36] X. Ling and D.S. Weld, "Fine-Grained Entity Recognition," AAAI, vol. 12, 2012, pp. 94–100.

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