Unsupervised Learning for Relation Extraction

Cheng Chen u5969643
Supervision: Lizhen Qu
October 28, 2016
Abstract

In many specific domain such as the biomedical with insufficient data set for relation extraction, we introduce the unsupervised pre-train method for two type of the deep neural network relation extraction models (including BiLSTM and PCNN), to generate a better performance model on the Bioinfer data set.
1. Introduction

The amount of textual information is greatly increasing with the widely used of digital textual technology. Therefore, there is a growing interest in many business domains to develop techniques to automatically extract information from specific document collections, store them in structured format which could be useful for future applications, such as document search and indexing. Relation extraction (RE) is one kinds of these techniques. it aims to detect and classify the semantic relations between mentions of entities in text. For example, sentence one’s content is “ANU is a national research university”, and sentence two is about The ANU and UC are in Canberra”, so we can find relation between these two sentences that “ANU is an university in Canberra”.

Recently, Machine learning is a popular way to solve difficult tasks like relation extraction. RE models using supervised machine learning method are trained on hand-labeled corpus. One of such corpus is The NIST Automatic Content Extraction (ACE) RDC 2003 and 2004 corpora (Zelenko et al, 2003). The ACE porpora incude more than 1,000 handcrafted documents, which including 16,771 relation instances. Supervised methods is usually divided into two types: feature based-methods and kernel-based methods. Feature-based approache give each features gathered from the labeled corpus a specific symbolic ID after the textual analusis, then the context information in the corpus is saved into a liner vector consist of the IDs, and store in a paradigm which is like the bag-of-words model (Wang et al., 2011). The other type of methods, kernel-based methods corpora (Zelenko et al, 2003), require pre-processed input data which saved in the parse trees, such as dependency parse trees. These approaches considered more effective when they are working on the bigger size of linguistic knowledge dataset.

After many years study about the RE techniques, the deep neural network (DNN) is are been widely applied and also got better peformance than other RE models (Xu et al., 2015). Many RE techniques are rely on large amount of manually work. The origin text need to be collectted and classified into the particular set that needed, after human analyzing, the manual craft patterns is setted to each text, then the RE model can wrod on this corpus and the patterns to extracting the realtions. The model based on hand crafted information haa three problems: First, the text need to be analyzed and setted new patterns if the
corpus used for different relation extract tasks. Second, the model's executing quality highly influenced by the quality of manual works which could have unpredictable potential risks. Besides, in many particular research and application domains, such as biomedical, the good quality knowledge set are scarce. So these questions lead us to consider about the possibility of training DNN RE with the help of unlabelled small data set?

2. Relation Extraction

Recently deep learning have provide great contribution in the relation extraction domain when compare with the traditional methods (Jiang and Zhai, 2007) on the NIST Automatic Content Extraction (ACE) RDC 2003 and 2004 corpora. As the result of the related work collection and analysis, it is found that most of the deep relation extraction models are mainly using the variants of convolutional neural networks (CNN) (Zeng et al., 2014; Xu et al., 2015) as well as recurrent neural networks (RNN) (Socher et al., 2012; Ebrahimi and Dou, 2015)

Some of the relation extraction models are working with the semi-supervised learning algorithm label propagation which used unlabelled data for model training. By developing these models on the manifold structure, which also noticed as cluster stucture, in the corpus data set, the semi-supervised methods can show effective performance on combining unlabeled data with the labeled data (Chen et al 2006). More specific, these semi-supervised methods are graph-based methods, which means they are built on a similarity graph between the relation mentions as a pair. This brings some limition for these type of algorithm. which restrict the labeling function which must be satisfied at the same time. The constraint condition are: First, the unlabeled input size should be designed as same as the labeled nodes. this condition is usually considered as a loss function in the regularization framework. Second, the whole graph should be smooth enough. Correspondently, this condition is defined as a regularizer in the regularization framework. Therefore, it can be seen that deep relation extraction models are facing many challenge to use this type of semi-supervised learning algorithm. To overcome these question, the methods used in this experiment are encoding the input data and provide the well structured medial data, which will used for traing the relation extraction models.
3. Deep Relation Extraction Models

Generally, relation extraction model are used to assign the appropriate type of relations between two entities based on the context information. More Specific, in a document D, a relation can be defined as a tuple \( r_i = (e_1, e_2, \ldots e_n) \), the \( e_i \) denoteds the entities in a finite named entity set \( E \), and a relation \( r_i \) are a specific relation in the finite relationship set \( R \). Many researchs about relation extraction are foucing on binary relations. For example, \( \text{CONTAIN(tank, water)} \) is a typical binary relation. Therefore, a relation extraction model can be defined as \( f: E \rightarrow R \). To train a relation extraction model on a training set \( (e_1, e_2) \rightarrow r_1 \in E \times R \), by minimizing a loss function \( L_s : E \times R \rightarrow R \). It is found in researches that the relation between entities are tend to be the same type when the entity mentions are holding high similarity context (Chen et al, 2006).

However, when generating new relation extraction model for some specific like biomedeical, it is usually restricted by the insufficient training data, so there will be insufficient instance have this kind of context similarity. Therefore, in order to generate a better performance relation extraction model on these dataset, we applied pre-train approach, which will be discussed in next section with more detail. Next, the pre-trained model will apply for a supervised Loss, which be fine tuned on the hand-crafted labelled dataset.

3.2 General structure of deep relation extraction models

The two type of deep relation extraction models in this experiment mainly include four parts. As illustrated in Figure 1, either the entity pair as well as the word content between them, or the shortest dependency path between two entity mentions, which generated by the denpendcy tree, are the input data set. The second part is a entity match table which map the input string into a entity embeddings. The third part is a hidden representation \( h(x) \) generated by a feature learning algorithm. Thr fourth part is a multi-class logistic regression classifier for relations, for the supervised baseline models, the \( h(x) \) will be input with the handcrafted features to the classifier, and for the models including unsupervised pre-training, \( h(x) \) will be input for a classifier work for unsupervised predictive task.
3.2 Bidirectional LSTM

The Long Short Term Memory (LSTM) networks is a type of Recurrent neural networks (RNNs) which introduced by Hochreiter and Schmidhuber (1997). The LSTM is usually used for solving the vanishing and exploding gradients problems. When compared with backpropagated architectures, which are unable to access the long-term information, the LSTM models keep these information by applying the memory blocks (Gers et al., 2002). The LSTM memory blocks are consisted by input gate, output gate, forget gate and the cell. The cell holds the short-term information, and encode the current time step’s input information. The other three gates are able to decide whether the information should be keep or drop from the long-time information by applying elementwise multiplication. More specifically, the input gate $i$ with a trained activation function will decide which input values should be sent to the cell. The forget gate $f$ will control the cell to forget the certain value, and the output gate $o$ is designed to choose which values in the cell should be added to the output and send to the next layer. As shown in the formula 1, the $W$ is the parameter matrices, and the $b$ is the bias for each gate and update function, where the $i$ in subscript represent the input gate, $f$ represent forget gate and the $o$ represent output gate, $t$ represent which the time step is. The $\cdot$ denotes elementwise multiplication and $\sigma$ represents the sigmoid function.
The essential idea of the bidirectional LSTM (BiLSTM) is to provide the training input in a different way. Specifically, the training data is put in both forwards direction and backwards direction, so the network layers will get more sequential information about the data, and keep information by delivering it to the same output layer.

The BiLSTM model is considered that it can represent effectively the long-term information, which is useful feature for the relation extration model to keep the word sequence information within the context between two entities. As illustrated in Figure 2, the BiLSTM models including two LSTM models, which share the parameters. The BiLSTM model is considered as the feature learning component mentioned before in the general structure relation extraction models. It get the word sequence contain the entity pair and the tokens between them as input, the key is the input is send both in the forward and backward directions. Therefore, the output of this BiLSTM is two hidden representations $h>$ and $h<$ generated respectively by the two LSTM models, and the two hidden representations will be send to the classifier as input.

\[
\begin{align*}
    i_t &= \sigma(W_{iz}x_t + W_{ih}h_{t-1} + b_i) \\
    f_t &= \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \\
    o_t &= \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \\
    u_t &= \tanh(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \\
    c_t &= f_t \cdot c_{t-1} + i_t \cdot \tanh(u_t) \\
    h_t &= o_t \cdot c_t
\end{align*}
\]

Formular 1
3.3 PCNN

As shown in Figure 3 PCNN accept the word sequence between entities as the input data. The input word sequence is will be mapped to the word embeddings, and the word embeddings is also connected to its position embeddings, which encodes the relative distance of the each word in the word sequence to each entity mention. However, the input word sequences are in different length, so the resulted embeddings are be generated with the word sequence added with padding tokens. Next, the convolutional layer are applied as the input for the classifier.

alpha-catenin inhibits beta-catenin signaling by preventing formation of a beta-catenin DNA complex.

Figure 3
4. Unsupervised Pre-training

The unsupervised pre-train methods in this paper is inspired by the Skip-gram model (Mikolov et al., 2013). Briefly, there are two stage in this pre-train method. First, the average entity embeddings of entities in Bioinfer corpus are generated. Second, the seq2seq decoding loss is applied as pre-training loss, which is used for the two kinds of deep RE models feature learning components’ training.

4.1 Seq2seq Decoding Loss

In this paper, the loss of unsupervised pre-train methods is extended from the semi-supervised sequence-to-sequence (seq2seq) model (Dai and Le, 2015). Seq2seq is a widely used model for machine translation, text parsing and speech recognition (Dai and Le, 2015). The seq2seq model can be consider as two parts. First, an LSTM encoder that read the input sequence, then reform the input sentence into a hidden state, another part is an LSTM decoder that is used to generate the predicted output sequence. As the parameters are same for encode and decode parts in the seq2seq model, the parameters will be used for the DNN RE model to initialize the classifier. With the parameters set generated in this pre-train stage, the RE models are possible to get better performance on the limited data set with less training time.

In this experiment, the decoder that using LSTM is kepted in the seq2seq model, however, the encoder is change to the different type of DNN, which is PCNN. so more generally, the seq2seq encoder is replaced by the the feature learning component h(x) discussed in the previous section.

In this case, the LSTM decoder get the current sentences as part of the input, combine the previous hidden state’s output u (t-1), and the previous time step cell ct-1’s output as the whole input data for current layer. As for the special case, when the decoder started, the input for the first layer are u (0) = h(x), and c (0) = 0. Then at each step, we use ut to predict the next word in the sequence. Instead of using the standard softmax function, as In (Dai and Le, 2015), for high computational efficiency, we apply negative sampling and obtain -log σ(vT T+1uT) - ∑k i=1 Ew(i) ~Pn(w(i)) log σ(-vw T iuT)
5. Experimental Setup

5.1 Setup and preprocessing

The re-Bio Information Extraction Resource (reBioinfer) (Hanchey et al, 2011) are be used for training the pre-training, RE model training, testing and developing. When start setting up the experiment, It is found that the rebioinfer data set including four type of XML files, three of them only save the dependency information, the extended attitute of the entities in each sentences and the sentence token information, the last type, is the file keep all the inforamtion needed for this experiment, but it also contain many features that will be redundant for the relation extraction tasks. Therefore, the first step is to filter the data. The functions of the program are: First, reading and decoding the XML file one by one, create the compele sentence data because there are no whole sentence saved in the reBioinfer data set as it is save in a separated words sequence. Second, the entities in each sentences are saved in with entity ids, it is not obvious and lost the textual inforamtion Therefore, the entity ID get from the reBioinfer is been rewrite in the form as an ENTITY_ with the corresponding metion. For instance, the e4, which means the fourth entity in the
The reBioinfer dataset, is “beta-catenin*T-cell factor*DNA complex”, as the is a textual information is generate from the sentence build in first step should be keep for the relation extraction tasks needs, the entity mention of e4 will records as “ENTITY_formation_of_a_beta-catenin_*t-cell_factor_*dna_complex”. Third, position of entities are also not directly saved in the reBioinfer data. For example, the description for entity position is recorded like “e4” (entity No. 4) start from “w10” (word No. 10) to “w15” (word No. 15). Therefore, to generate the entity embeddings, the position of entities is transferred into the actual token position and match on the sentences got from the first step.

As for the relations between entities are clear saved in the reBioinfer data, it can be easily pick out and saved. However, the reBioinfer data set, as same as the origin Bioinfer data set contain n-ary relation, these relations are been decomposed into multi binary relations. For instance, in this sentence “Further, PRP incubated with IL-6 showed a dose dependent increase in TXB2 and BTG secretion as measured by RIA and an increased incorporation of actin binding protein, talin, and myosin into the cytoskeletal core (triton insoluble and residue) as shown by SDS-PAGE”, the entity “actin binding protein” has the relation “EXTENT” with the entity “talin” and the entity “myosin”, so the relation separate save as the entity “actin binding protein” EXTENT “talin”, and “actin binding protein” EXTENT “myosin”.

In addition, it is found that the reBioinfer dataset also contain the sentences which have no relations between the entities inside, the information about the entity, the sentence are still been keep into the data set and been used to generate the entity embeddings. In summary, the reBioinfer, as well as the origin Bioinfer data set, contain 2200 sentences, after the filtrate step, the 2121 sentences save the relation information are keep for this experiment, with 62 types of relations and 1061 kinds of entities. These file are saved in xmi format files separately.

As the sentences including the tokens, entities, dependency and relation information are dispersedly save in xmi format files in the reBioinfer dataset, Therefore it is not efficient for the DNN RE models to gather the information file by file. The essential information is picked out by the program designed for the specific reBioinfer data set, and save as a json file contain all these information
which could be generally used for this experiment and future research. Based on this file, the entity embedding is more efficiently generated only in minutes.

In this experiment, the relation mentions are been split into three sets, that 70% as training, 20% relabeled relation mentions as testing set, 10% for development. Because the reBioinfer data set is a small data set, for the training and testing set, they might missing some specific type of relation mentions. To solve this problem, the split program firstly collection the number of each type of relation mentions, and generated the split data set contain appropriate number of relation mentions. Besides, to make the training result meaningful on the small Bioinfer data set, the split operation is randomly pick up the relations in same type and put into training, testing and developing data set

5.2 Baseline

The two type of deep RE models with BiLSTM and PCNN with random initialisation, and a logistic regression trained on the reBioinfer data set keeps handcrafted features as baselines. As the randomly initialised parameters is influential to the performance of these RE models, the models were generated by servel times and the averaged micro-F1 was picked as the expected performance result fot these models.

5.3 Results and Discussions

As can be seen is the Figure 5, the deep relation extraction models generated with unsupervised pre-train perform 10% higher than the baseline, This result could be evaluated as a good improvement for the relation extraction tasks on the biomedical domain. Among all the result on training and developing, the pre-trained BiLSTM models reach the highest result with 24% F1 score, The pre-trained PCNN RE model performed a little bit lower, and still reached to the 19% F1 score.

Relatively, for the both BiLSTM and PCNN baseline RE models generated with handcrafted data set, they are generally only got 10%~12% F1 score. As for the baseline RE models with random initialisation, the result is just a little bit better,
which between 11% to 14% F1 score. As discussed before, these models was trained for several times to reduce the influence of randomly initialisation, so the result for randomly initialized model is meaningful.

Figure 5

However, after analysis of the result and back viewed the Bioinfer data set, The reason why the overall models got low F1 score might include the lack of certain relation instance. More specific, the 68% types of relation only have less than 25 sentences which contain the corresponding realtion (see Table 1). as these 25 sentences will also been split to three part for training, testing and developing, so the RE model generateing really suffered from the insufficient of the training data. Therefore, the specific relation with more than 100 sentence instances could be consider could be will figure by the models. Some of these relations, such as “BIND”, “COLOCALIZE”, is picked out for advance evaluation. As shown in the Figure 6, the F1 score got from the pre-trained BiLSTM and PCNN RE models, for “BIND” relation is reach the 42% and 39%. In constrast, the baseline RE models also have getter result, for “BIND” it got 22% F1 score with handcrafted based RE models, and 28% F1 score with the random initialized RE models. Obviously, for these perticular relations have
enough instance for the RE model, the result for all models have significantly improved, and the pre-trained models are 20% better than the baseline models.

<table>
<thead>
<tr>
<th>0-25 sentences</th>
<th>25-50 sentences</th>
<th>50-100 sentences</th>
<th>100-200 sentences</th>
<th>&gt;200 sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>68%</td>
<td>0.06%</td>
<td>0.09%</td>
<td>0.15%</td>
<td>0.02%</td>
</tr>
</tbody>
</table>

Table 1

Even the relation extraction models is highly influenced by the small size of the unlabelled data, it still shows the advantage could be implemented to generated a better RE models on specific domain which might difficult to get enough corpus for training.

6. Conclusion
In the particular Bioinfer domain that lack of large amount labelled textual dataset, this experiment have applied two type of deep relation extraction models with the unsupervised pre-training methods, and achieved 10% better performance than the RE models without pre-training even influenced by the insufficient data, and achieved 20% better for some particular relation with a little bit large training data.

Reference:


Appendix 1

INDEPENDENT STUDY CONTRACT

Note: Enrolment is subject to approval by the projects co-ordinator

SECTION A (Students and Supervisors)

UniID: u5969643

SURNAME: Chen  FIRST NAMES: Cheng

PROJECT SUPERVISOR (may be external): Lizhen Qu - NICTA/DATA61

COURSE SUPERVISOR (a RSCS academic): _____________________________________________

COURSE CODE, TITLE AND UNIT: _Individual Project__ COMP6470 _6 units________________

☒ S2 YEAR: 2016__________

PROJECT TITLE:

Unsupervised Learning for Relation Extraction

LEARNING OBJECTIVES:

Deep learning, Recurrent Neural Networks, Convolutional Neural Networks, Natural Language processing, Relation Extraction.

PROJECT DESCRIPTION:

Recent advance of Information Extraction techniques shows that deep learning techniques have an edge over the traditional methods by the capability of learning rich feature representations from text. Pure supervised learning of deep learning models still requires substantial training data. Despite of that, unsupervised pre-training is the technique with large potential of reducing training data. Therefore, this project aims to explore novel deep
learning techniques for relation extraction by using freely available unlabelled corpora. The resulted models will be compared with the state-of-the-art methods on benchmark datasets.

The learning objectives are as follows:

a) Have an overarching understanding of deep learning for relation extraction.

b) Understand the core algorithms and models for relation extraction.

c) Be able to read current research papers in the field of relation extraction.

d) Be able to design and implement the models with the state-of-the-art art toolkit Deep learning framework Torch.

e) Be able to design experiments to compare new models with the state-of-art methods in a systematic way.

ASSESSMENT (as per course’s project rules web page, with the differences noted below):

<table>
<thead>
<tr>
<th>Assessed project components:</th>
<th>% of mark</th>
<th>Due date</th>
<th>Evaluated by:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Report: name style: <em><strong><strong>Research report</strong></strong></em>____ (e.g. research report, software description...)</td>
<td><strong>60</strong> (60%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Artefact: name kind: __ Working Algorithm_________ (e.g. software, user interface, robot...)</td>
<td><em>30</em> (30%)</td>
<td></td>
<td>Lizhen Qu</td>
</tr>
<tr>
<td>Presentation:</td>
<td><strong>10</strong> (10%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MEETING DATES (IF KNOWN):

Weekly or bi-weekly upon availability.
STUDENT DECLARATION: I agree to fulfil the above defined contract:

……Chen Cheng…………………………. ……20th July 2016……….
Signature Date

SECTION B (Supervisor):

I am willing to supervise and support this project. I have checked the student's academic record and believe this student can complete the project.

……Lizhen Qu…………………………….. ……20 July 2016……….
Signature Date

REQUIRED DEPARTMENT RESOURCES:

SECTION C (Course coordinator approval)

………………………………………………….. ……
Signature Date

SECTION D (Projects coordinator approval)

………………………………………………….. ……
Signature Date