Relation Extraction (need a better title...)

Zachariah Zhang
Department of Data Science
New York University
New York, NY 10012
zz1409@nyu.edu

Lizi Chen
Department of Computer Science
New York University
New York, NY 10012
lc3397@nyu.edu

Guangyu Zhang
Department of Computer Science
New York University
New York, NY 10012
guangyu.zhang@nyu.edu

Abstract

The ultimate goal of this project is to build a Question-Answering (QA) application based on generated Knowledge Graph (KG) from raw text data. For the following two parts, we will focus on part 1, and will explore part 2 if time permitted.

1. Knowledge Graph construction from raw text data;
2. *(Tentative, will do only if time permits)* Question-Answering application on top of the graph. May use it for metrics and evaluation.

This is a joint project for the two courses: 2/3 authors have enrolled in the "Statistical NLP" course, and 2/3 authors have enrolled in the "Inference and Representation" .

1 Problem to address

Natural language text is typically very unstructured which makes many different machine learning applications difficult. A knowledge graph can be used to model the relationships of different entities in a body of text. These graphs create a lot of value for other machine learning applications such as question answering and reading comprehension.

In recent years, deep learning has seen boom in popularity for nlp problems with its ability to model complicated and ambiguous text. Traditionally, people have adapted statistical methods, e.g. POS-tags, named entity tags, to extract entities relations, and reached an precision of 67.6

Based on revising recent papers, we will create program which consumes data into knowledge graph based on some available dataset as listed in the next sections. We will tweak the parameters, analyze the existing models, and further improve on selective ones. We may tentatively provide a Question-Answering testing application for graph demonstration and evaluation; however, given the amount of work and limited time in the remaining semester, we will only work on this part when time permits.

2 Datasets

Freebase This data set has been massively used in many knowledge graph publications.

Nell-995

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Dataset used in Riedel et al. (2010) This dataset was formed by aligning Freebase relations with the New York Times corpus (NYT). Entity mentions were found in the documents using the Stanford named entity tagger, and are further matched to the names of Freebase entities.

3 Algorithms

3.1 Distant supervision

Relation extraction plays a key role in knowledge graph. Traditionally, supervised techniques are adopted for non deep learning methods, but requires intensive human labors of annotation. Mintz et al. (2009) proposed a distant supervision method of automatically generating training data, via aligning documents with known knowledge base (Freebase). They have extracted lexical (e.g. sequence of words, POS tags) and syntactic (dependency parser) features, and performed named entity tagging by Stanford four-class named entity tagger Finkel et al. (2005). Their model is reported to extract 10,000 instances of 102 relations at a precision of 67.6%.

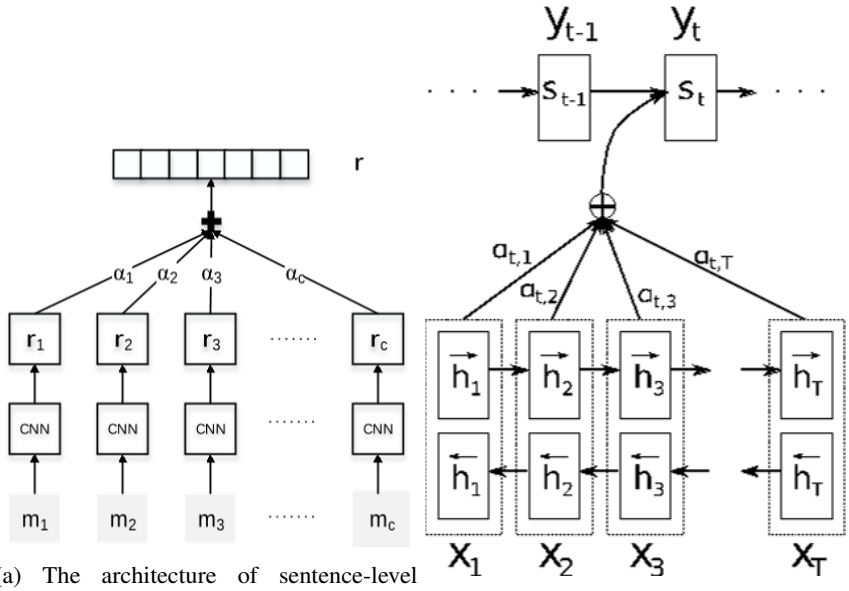
Mintz et al. (2009) has an assumption such that if two entities participate in a relation, any sentence that contain those two entities might express that relation, while in reality may result to non-trivial noises. Riedel et al. (2010) alleviates noises via deciding whether two entities are related and also mentioned in a given sentence, and then applying "constraint-driven semi-supervision" to train

3.2 Deep learning tsunami in KG

3.2.1 PCNN with attention mechanism

We will compare work developed in several different deep learning publications. We will focus on the basic CNN model (Zeng et al 2014), Piecewise CNN (Zeng et al 2015), CNN with multiple kernels (Thien and Grishman et al 2015), as well as the use of attention (Lin et al 2016). Compare the tradeoffs between these model architectures and compare with non-deep learning baselines. Lin et al. (2016)

Yankai Lin et al 2016 improved the Distant supervised relation extraction method with additional attention mechanism over its previous work. The improved version has CNN for semantic feature extraction and a sentence-level attention for the purpose of reducing wrong labeling weights.



(a) The architecture of sentence-level attention-based CNN. Used in Yankai Lin et al 2016

(b) Classical attention mechanism used in Dzmitry arXiv:1409.0473

Figure 1: Comparison between two attentions.

Input Vector Representation There are many paradigms of word embedding; converting human-readable vocabularies into numbers that can be processed by machine. Two major approaches from the past several years are the Frequency based Embedding and Prediction based Embedding.

There are two embedding steps done for the input vector representation in this paper. One is word embedding, which aims to capture syntactic and semantic meanings of the words. The other one is based on the assumption that words that close to the target entities are usually informative to determine the relation between entities. Thus, the result vector is a concatenation of the word embedding matrix and the position embedding one. The result vector:

$$w_i \in \mathbb{R}^d (d = d^a + d^b \times 2)$$

where d^a is the dimension of the word embedding and d^b is the dimension of position embedding.

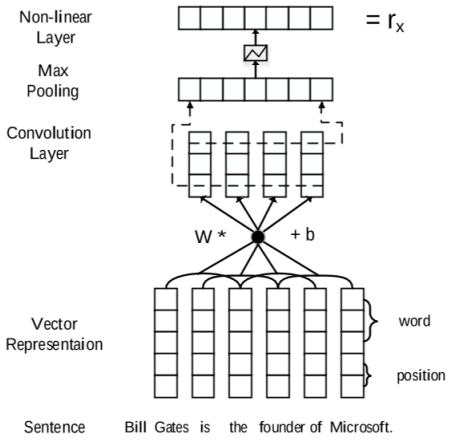


Figure 2: The architecture of CNN/PCNN used for sentence encoder

Convolution, Max-pooling, Zero-paddings, and Non-Linear Layers Due to the fact that vector representation from previous step has all different size, max-pooling is required to obtain a fixed-sized vector for the input sentence.

Selective Attention over Instances One of the core innovative part from this paper, which makes the result stood out from past works, is the attention mechanism for all entities and providing various weights to reduce wrong entities label. For each previous layer, attention attribute is added as:

$$a_i = \frac{\exp(e_i)}{\sum_k \exp(e_k)}$$

where $e_i = x_i A r$
A is a matrix, r represents the relation in matrix, which is mentioned in the previous paper as well. Therefore, e_i means the matching between the sentence and entities.

The final result conditional probability can be represented as the following through a softmax layer:

$$p(r|S,\theta) = \frac{\exp(o_r)}{\sum_{k=1}^{n_r} \exp(o_k)}$$

n_r = total number of relations
o = the final output of the neural network, which corresponds to the scores associated to all relation types.
θ = the parameter of this model

$$o = Ms + d$$

Therefore, the objective function using cross-entropy for the training is:

$$J(\theta) = \sum_{i=1}^{s} \log(p(r_i|S_i, \theta))$$

The paper uses mini-batch chosen randomly from the training set as the result converge.

3.3 Reinforcement learning

3.4 General optimization methods

The papers that we have investigated adopts general stochastic gradient descent (SGD) to minimize objective function. Dropout layer (Srivastava et al., 2014) is implemented on the output layer to prevent overfitting.

4 Evaluation Metric

We will evaluate each model on precision, recall, and F score to be consistent with the literature.

5 Questions

References

Riedel, S.; Yao, L.; McCallum, A. Modeling Relations and Their Mentions without Labeled Text. Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III. Berlin, Heidelberg, 2010; pp 148–163.

Mintz, M.; Bills, S.; Snow, R.; Jurafsky, D. Distant Supervision for Relation Extraction Without Labeled Data. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2. Stroudsburg, PA, USA, 2009; pp 1003–1011.

Finkel, J. R.; Grenager, T.; Manning, C. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA, 2005; pp 363–370.

Lin, Y.; Shen, S.; Liu, Z.; Luan, H.; Sun, M. Neural Relation Extraction with Selective Attention over Instances. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. 2016.