need a better title...

Zachariah Zhang
Department of Data Science
New York University
New York, NY 10012

Lizi Chen
Department of Computer Science
New York University
New York, NY 10012
zz1409@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 ? 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

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. **?**

3.3 Reinforcement learning

3.4 General optimization methods

4 Evaluation Metric

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

5 Questions

References