Center for Speech and Language Technologies

GROUPING

Research Proposal

(Deep Relation Extraction based on Distant Supervision)

By. Miao Fan
Dept. of C.S.
Tsinghua University
fanmiao.cslt.thu@gmail.com

Introduction to Relation Extraction(RE)

- Based on **Named Entity Recognition(NER)**, We'd like to extract the **relations** between two named entities from a certain sentence, for example,

Barack Obama(PER) is now **the president of (re: president) U.S.A(ORG)**.

Introduction to Relation Extraction(RE)

- Further, we can extract a triple from this sentence.

Barack Obama(PER) is now **the president of (re: president)** **U.S.A(ORG)**.

<Subject, Predicate, Object> =

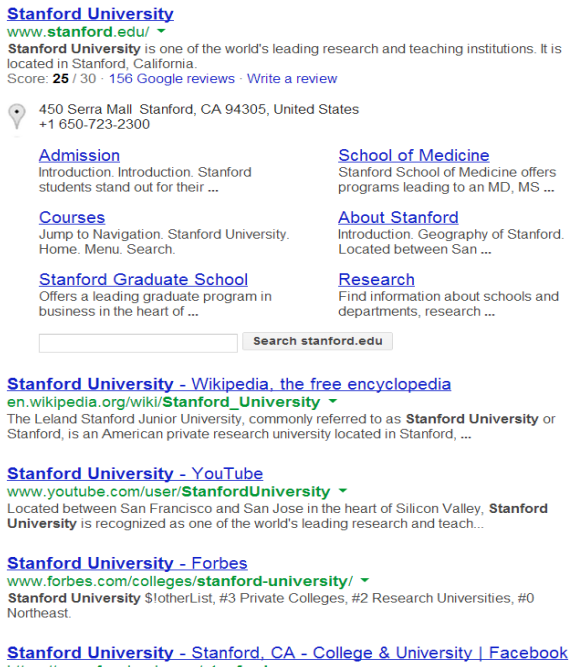
<Barack Obama, president, U.S.A>

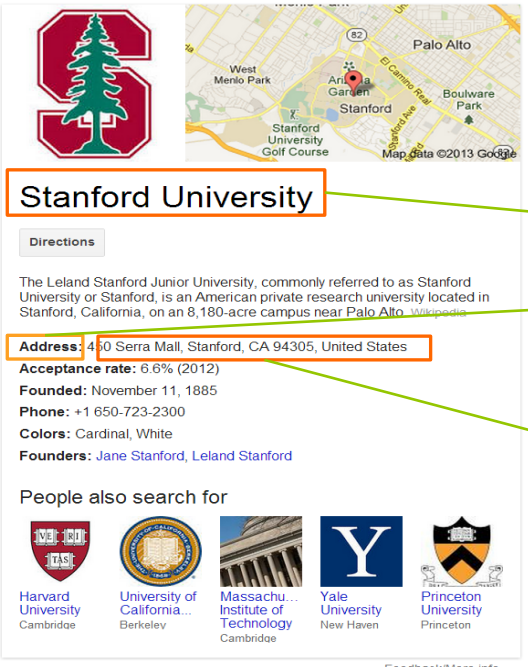
Motivation of Relation Extraction

- Extracting the Structure Data from Raw Text (Un-structure Data) .
- More specifically, mining the new cases of relation triples from texts.
- Further, we can store these structure data into database as the **knowledge**.
- Those knowledge could be used in **QA**, Entity Search Engine especially for **cellphones**.

Motivation of Relation Extraction

- Case Study:
  - Google Knowledge graph & Infobox on Wikipedia





Subject(ORG)

Predicate(Relation)

Object(LOC)

Motivation of Relation Extraction

- Case Study:
  - EntityCube @ MSRA

Related Work

- **Rule-based**
- **Learning-based(Three Paradigms)**
  - **Unsupervised Learning**
  - **Semi-supervised Learning(Bootstrapping)**
  - **Supervised Learning(Multi-class Classification)**
    - ACE (Small Corpus, Human labeled supervision)
    - **Freebase (Large Corpus, Distant weak labeled supervision)**

Related Work

- Rule-based(Pattern-based):
  - If we want to extract a relation standing for X and Y are the same thing

```
"Y such as X ((, X)* (, and|or) X)"
"such Y as X"
"X or other Y"
"X and other Y"
"Y including X"
"Y, especially X"
```

  - Not recommend!

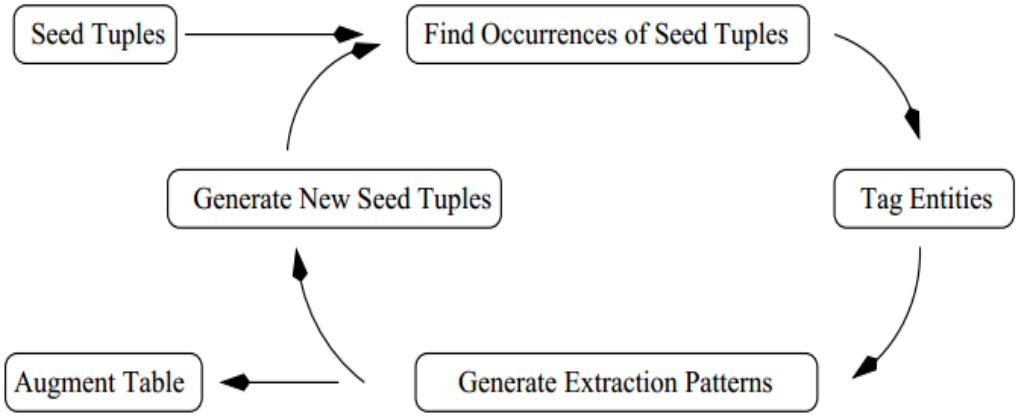
Related Work

- Learning Based
  - Unsupervised Learning
    - Extracting the string between two entities(Have been recognized by NER Tools).
    - Clustering and simplifying the words in each cluster and mapping them into given classes(Type of relations).

Related Work

- Learning Based
  - Semi-supervised Learning(Bootstrapping)

```
<ORGANIZATION>'s headquarters in <LOCATION>
<LOCATION>-based <ORGANIZATION>
<ORGANIZATION>, <LOCATION>
```



Related Work

- Learning Based
  - Supervised Learning(Multi-class Classifiers)
    - ACE corpus(Human-labeled supervision)
      - 5 to 7 major relation types (23 to 24 sub-relations)
      - 16,771 relation triples.
      - **The main issue is the lack of generalization ability(Domain Bias).**

Related Work

- Learning Based
  - Supervised Learning(Multi-class Classifiers)
    - **Freebase(Knowledge base weak-labeled distant supervision)(1)**
      - **Assumption:**

        If a sentence contains **two entities** and those entities are **an instance** of one of our **Freebase relations**, **features** are extracted from that sentence and are added to the feature **vector** for the relation. (Prepared as the **training set** for Supervised Learning).

Related Work

- Learning Based
  - Supervised Learning(Multi-class Classifiers)
    - **Freebase(Knowledge base weak-labeled distant supervision)(2)**
    - **Key Articles.**
      - (Mintz et al, ACL' 09)
      - (Riedel et al, ECML'10)
      - (Hoffmann et al, ACL'11)
      - (Takamatsu et al, ACL'12)

Related Work

- ACL'09 Distant Supervision Relation Extraction (First paper).
  - **Dataset**
    - Freebase(1.8 M instances of relations, 102 kinds of relations, 940,000 entities about PER, ORG, LOC).
    - Wikipedia(800,000 articles for training, 400,000 for testing, 14.3 sentences per article)

Related Work

- ACL'09 Distant Supervision Relation Extraction (First paper).
  - **Feature Vector**
  - **Mutli-class Logistic Regression**
  - **Evaluation**
    - **Held-out evaluation(900,000 triples for training, 900,000 for testing)**
    - **Human evaluation (Amazon Mechanical TURK)**

My Proposal

- **Deep Relation Extraction based on Distant Supervision**
  - Key Challenges:

1. How to auto-generate training corpus with high labeling confidence based on DS.
2. How to auto-select the principle features in order to reduce the dimension of the feature vector.
3. Which classifier(s) is(are) more effective?

My Proposal

- **Suggestion:**
  - doing some experiments on exploiting the **deep neural network(DNN)** learning paradigm to do the multi-class classification task.

Future Work

- Continuing surveying on state-of-the-art relation extraction methods on distant supervision and handing in **an overview draft (reviewing article)**.

- Preparing for the data resources(Wikipedia, Freebase etc.)

References

- **Bach, Nguyen, and Sameer Badaskar.** "A review of relation extraction."*Literature review for Language and Statistics II* (2007).
- **Mintz, Mike, et al.** "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*. Association for Computational Linguistics, 2009.
- **Riedel, Sebastian, Limin Yao, and Andrew McCallum.** "Modeling relations and their mentions without labeled text." *Machine Learning and Knowledge Discovery in Databases*. Springer Berlin Heidelberg, 2010. 148-163.

References

- **Hoffmann, Raphael, et al**. "Knowledge-based weak supervision for information extraction of overlapping relations." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Vol. 1. 2011.
- **Shingo Takamatsu**, Issei Sato, Hiroshi Nakagawa: Reducing Wrong Labels in Distant Supervision for Relation Extraction. ACL(1) 2012: 721-729