Semantic based Graph Convolutional Neural Network for Entity Extraction

Group Members（sort by first name）:

Haiyang Sun

Hao Zhang

Xuehan Chen

Yanqi Yao

Yirong Wang

Information Extraction

- Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents
  - Find and understand limited relevant parts of texts
  - Gather information from many pieces of text
  - Produce a structured representation of relevant information:
    - relations (in the database sense), a.k.a.,
    - a knowledge base

Named Entity Recognition

- Named Entity Recognition (NER) is the process of finding entities (people, cities, organizations, dates, ...) in a text.
- Is a subtask under Information Extraction.
- Target:
  - Identify named entities
  - Classify named entities

A very important sub-task: find and classify names in text, for example:

The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Person Date Organization

Rule-Based Methods for Entity Extraction

- Many real-life extraction tasks can be conveniently handled through a collection of rules, which are either hand-coded or learnt from examples.
- Basic rules
  - *Contextual Pattern -> Action*

Rules to Identify a Single Entity

An example of a pattern for identifying person names of the form "Dr. Yair Weiss" consisting of a title token as listed in a dictionary of titles (containing entries like: "Prof", "Dr", "Mr"), a dot, and two capitalized words is

({DictionaryLookup = Titles} {String = "."} {Orthography type = capitalized word}{2}) ->Person Names.

Rules for Multiple Entities

Some rules take the form of regular expressions with multiple slots, each representing a different entity so that this rule results in the recognition of multiple entities simultaneously.

({Orthography type = Digit}):Bedrooms ({String ="BR"}) ({}) ({String ="$"}) ({Orthography type = Number}):Price -> Number of Bedrooms = :Bedroom, Rent =: Price*

Dataset Preprocessing

- To feed GCN model and generate the neural network, we must provide three matrices.
- N*N matrix represent the dependencies between words in a sentence
- N*D matrix represent features of each word
- N*E matrix represent class of each word

Dataset Preprocessing

```
EU NNP B-NP B-ORG
rejects VBZ B-VP O
German JJ B-NP B-MISC
call NN I-NP O
to TO B-VP O
boycott VB I-VP O
British JJ B-NP B-MISC
lamb NN I-NP O
. . O O

Peter NNP B-NP B-PER
Blackburn NNP I-NP I-PER

BRUSSELS NNP B-NP B-LOC
1996-08-22 CD I-NP O

The DT B-NP O
European NNP I-NP B-ORG
Commission NNP I-NP I-ORG
said VBD B-VP O
on IN B-PP O
Thursday NNP B-NP O
```

EU NNP B-NP B-ORG

EU : word

NNP : part-of-speech (POS) tag

B-NP: syntactic chunk tag

B-ORG: named entity tag

Four types of named entities:

-persons（B-PER）
-locations（B-LOC)
-organizations (B-ORG)
-names of miscellaneous entities that do not belong to the previous three groups (B-MISC)

-not part of a phrase (O)

- *https://www.clips.uantwerpen.be/conll2003/ner/*

Dataset Preprocessing

- CoNLL 2003 dataset provides sentences from newspaper/magazine
- Each word are correctly tagged
- Use python to get normal sentences from dataset

```
1 EU rejects German call to boycott British lamb
2 Peter Blackburn
3 BRUSSELS 1996-08-22
4 The European Commission said on Thursday it disagreed with German advice to consumers
to shun British lamb until scientists determine whether mad cow disease can be
transmitted to sheep
5 Germany 's representative to the European Union 's veterinary committee Werner
Zwingmann said on Wednesday consumers should buy sheepmeat from countries other than
Britain until the scientific advice was clearer
```

Dataset Preprocessing

- With help of **Stanford Parser**, we write python code to generate every sentence's dependencies in the dataset
- From this generated file we can get dependencies and generate N*N matrix for each sentence

```
Index 1 word pairs:
 (rejects-2,call-4) (rejects-2,lamb-8) (rejects-2,EU-1) (call-4,German-3) (lamb-8,to-5)
(lamb-8,British-7) (lamb-8,boycott-6)
```

```
Index 4 word pairs:
 (Commission-3,The-1) (Commission-3,European-2) (said-4,disagreed-8) (said-4,Commission-3)
(Thursday-6,on-5) (disagreed-8,shun-15) (disagreed-8,Thursday-6) (disagreed-8,advice-11)
(disagreed-8,consumers-13) (disagreed-8,it-7) (advice-11,with-9) (advice-11,German-10)
(consumers-13,to-12) (shun-15,determine-20) (shun-15,lamb-17) (shun-15,to-14)
(lamb-17,British-16) (determine-20,scientists-19) (determine-20,transmitted-27)
(determine-20,until-18) (disease-24,cow-23) (disease-24,mad-22) (transmitted-27,can-25)
(transmitted-27,be-26) (transmitted-27,disease-24) (transmitted-27,sheep-29)
(transmitted-27,whether-21) (sheep-29,to-28)
```

Dataset Preprocessing

- We decided to use 300D word2vec to represent a word.
- This program is still in debugging…

```
EU:  0.037353516 -0.203125 0.21289062 0.24414062 -0.28515625 -0.034423828 0.06689453
-0.1875 -0.0390625 0.008483887 -0.2890625 -0.083496094 0.09082031 -0.2734375 -0.39257812
-0.10644531 -0.06591797 -0.0099487305 -0.05419922 -0.041748047 0.26367188 0.079589844
0.15039062 0.19433594 0.21289062 0.09863281 -0.3359375 0.15820312 0.28320312 0.23339844
-0.119140625 -0.23046875 0.26171875 0.059570312 0.026123047 -0.34179688 -0.15429688
0.13769531 0.09863281 0.055664062 0.31445312 0.09814453 0.15820312 0.19726562 0.022705078
-0.076171875 -0.296875 0.21875 -0.359375 0.18847656 -0.10839844 0.0031585693 -0.05834961
0.19628906 0.12890625 -0.23144531 -0.39257812 0.01361084 -0.29492188 -0.07763672
-0.18554688 -0.29882812 0.014099121 0.021728516 0.12988281 -0.18066406 -0.015625
0.11816406 -0.26757812 -0.16210938 -0.12060547 0.21484375 0.18847656 0.13671875
-0.29882812 -0.07128906 0.21289062 0.18359375 0.022827148 0.34960938 -0.3828125
-0.41601562 0.03149414 0.06982422 0.07910156 0.19335938 -0.05053711 -0.30078125 0.140625
0.26953125 -0.048095703 -0.29882812 -0.25976562 0.15429688 -0.076660156 -0.20214844
-0.05493164 -0.35742188 0.421875 -0.10595703 -0.057861328 -0.040283203 -0.13574219
0.06225586 0.07519531 0.19140625 -0.14355469 -0.20019531 0.15527344 -0.24609375 0.20996094
-0.16308594 0.14257812 0.31640625 0.23535156 0.19824219 -0.13574219 0.036132812 0.29882812
0.20703125 0.07763672 -0.04272461 -0.24609375 -0.171875 0.045166016 -0.2421875 0.039794922
-0.0017166138 -0.5390625 -0.02734375 0.14453125 0.20019531 -0.18554688 0.059570312
-0.21191406 -0.2265625 -0.050048828 -0.22363281 0.28515625 -0.30664062 0.2265625
```

To do list

- Continue to do research on GCN paper
- Collect entity information from dataset and generate N*E matrix
- Feed three matrices into GCN model, calculate the accuracy
- Calculate accuracy/call back rate
- Compare accuracy/call back rate with latest paper

Reference

- Information Extraction and Named Entity Recognition, Lecture PPT, Stanford https://web.stanford.edu/class/cs124/lec/Information_Extraction_and_Named_Entity_Recognition.pdf
- SEMI-SUPERVISED CLASSIFICATION WITH GRAPH CONVOLUTIONAL NETWORKS https://arxiv.org/pdf/1609.02907.pdf
- Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering https://arxiv.org/pdf/1606.09375.pdf
- Graph Convolutional Network https://tkipf.github.io/graph-convolutional-networks/
- Information Extraction CIS, LMU München Winter Semester 2015-2016 Dr. Alexander Fraser, CIS http://www.cis.uni-muenchen.de/~fraser/information_extraction_2015_lecture/03_rule_based_NER.pdf