The implementation of entity extraction system based on semantic role labeling

IST664_M001 NLP Project, Instructor: Lu Xiao, Presented by Haiyang Sun, Hao Zhang, Xuehan Chen, Yanqi Yao, Yirong Wang

INCRODUCTION

Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents

1. Find and understand limited relevant parts of texts
2. Gather information from many pieces of text

Named Entity Recognition (NER) is the process of finding entities (people, cities, organizations, dates, ...) in a text.

NER is a subtask under Information Extraction.

Target:
Identify named entities
Classify named entities

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
6 We do n't support any such recommendation
because we do n't see any grounds for it the
Commission 's chief spokesman Nikolaus van
der Pas told a news briefing
7 He said further scientific study was
required and if it was found that action was
needed it should be taken by the European
Union
8 He said a proposal last month by EU Farm
Commissioner Franz Fischler to ban sheep
brains spleens and spinal cords from the
human and animal food chains was a highly
specific and precautionary move to protect
human health
```

CoNLL 2003 dataset provides sentences from newspaper/magazine.

Each word are correctly tagged.

Use python to get normal sentences from dataset.

```
-DOCSTART- -X- -X- O

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
it PRP B-NP O
```

Example:

EU NNP B-NP B-ORG

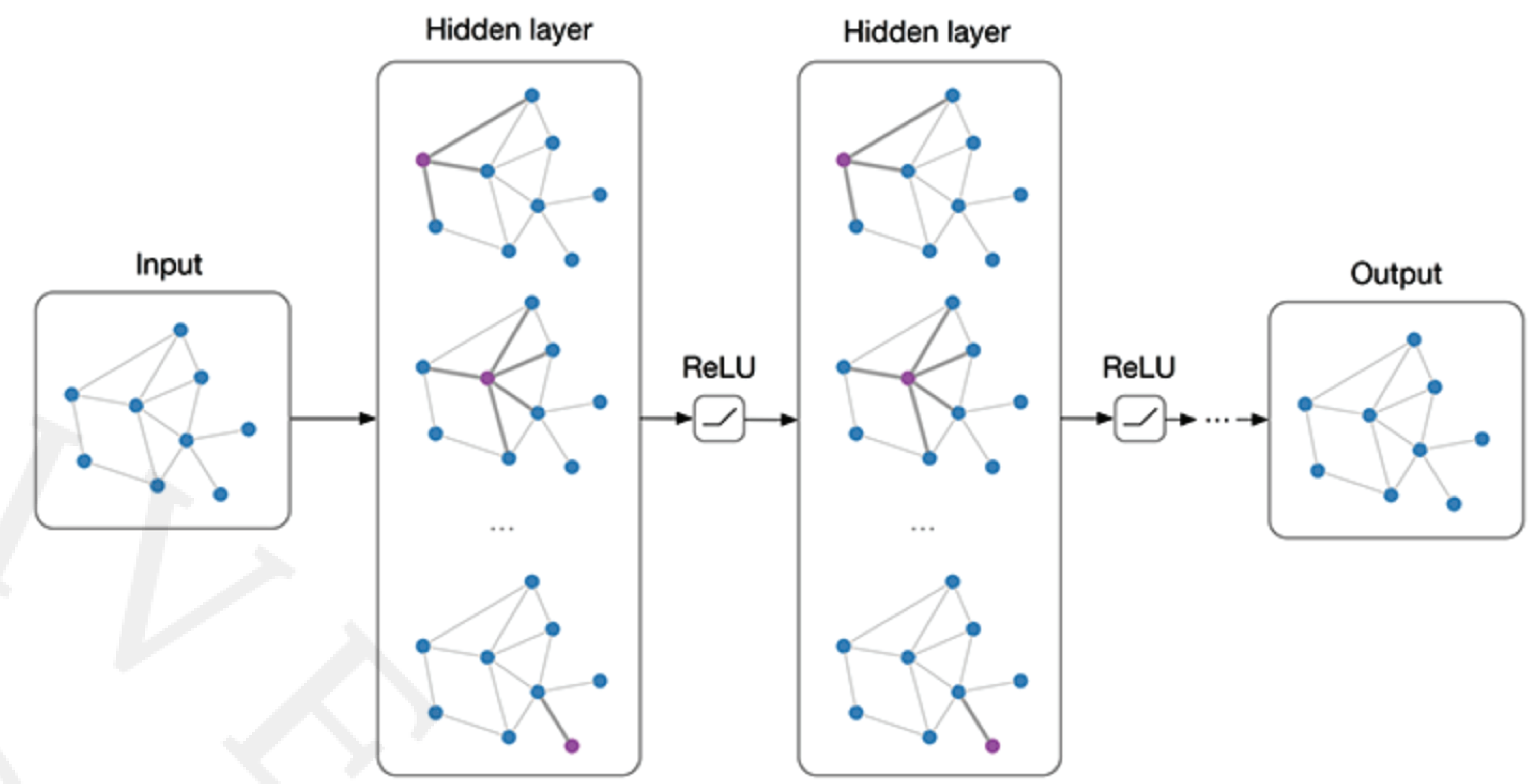
EU : word
NNP : part-of-speech (POS) tag
B-NP: syntactic chunk tag
B-ORG: named entity tag

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)

GCN

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

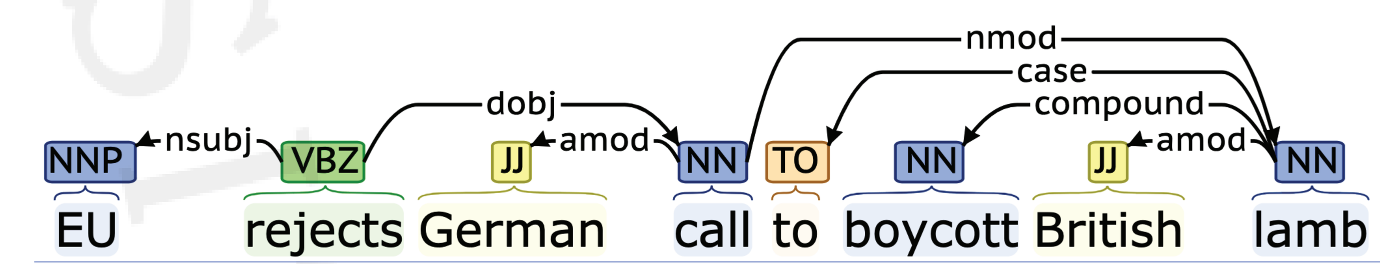


PROCESS

Example:

EU	rejects	German	call	to	boycott	British	lamb
1	2	3	4	5	6	7	8

With help of ***Stanford Parser***, we write python code to generate every sentence's dependencies



N*N matrix represent the dependencies between words in a sentence.

(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)

	1	2	3	4	5	6	7	8
1	0	0	0	0	0	0	0	0
2	1	0	0	1	0	0	0	1
3	0	0	0	0	0	0	0	0
4	0	0	1	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0
8	0	0	0	0	1	1	1	0

N * N matrix

We decided to use 300D ***word2vec*** to represent a word.

Word2vec is a two-layer neural net that processes text. Its input is a text corpus and its output is a set of vectors:

N*D matrix represent features of each word

A feature description xi for every node i; summarized in a N×D feature matrix X
(N: number of nodes, D: number of input features)

0.037	-0.203	0.212	0.244	...
-0.034	0.066	-0.187	-0.039	...
-0.289	-0.083	0.091	-0.273	...
-0.106	-0.065	-0.009	-0.054	...
-0.041	0.264	0.079	0.150	...
0.194	0.212	0.098	-0.336	...
0.158	0.283	0.233	-0.119	...
...	...	...	...	...

N * D matrix

N*E matrix represent class of each produces a node-level output Z (an N * E feature matrix, where E is the number of output features

	PER	LOC	ORG	MISC	O
1	0	0	1	0	0
2	0	0	0	0	1
3	0	0	0	1	0
4	0	0	0	0	1
5	0	0	0	0	1
6	0	0	0	0	1
7	0	0	0	1	0
8	0	0	0	0	1

N * E matrix

CONCLUSION

In this project, we implemented an sentence entity extraction system based on deep learning and symantic role labeling. After training, the classifier reaches the accuracy of 86.02%. Although the accuracy is still a little bit lower than the state-of-art approach, we believe that there are still many things we can do to improve the accuracy.

Syracuse University

Reference

[1] James H Martin and Daniel Jurafsky. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Pearson/Prentice Hall Upper Saddle River, 2009.
[2] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet project. COLING/ACL-98, pages 86-90, 1998.
[3] C. J Fillmore. Frames and the semantics of understanding. Quaderni di Semantica, page 222-254, 1985.
[4] P. N Johnson-Laird. Mental models. Harvard University Press, Cambridge, MA.
[5] R. C. Schank and R. P Abelson. Scripts, plans, and knowledge. Proceedings of IJCAI-75, page 151-157, 1975.
[6] Raymond R Smullyan. First-order logic, volume 43. Springer Science & Business Media, 2012.
[7] James Allen. Natural language understanding. Pearson, 1995.