Deep Learning in Named Entity Recognition (NER)

Intro

Named Entity Recognition is a basic task in natural language processing that has very brand applications. Named entity usually refers to some entities that has special meaning or reference in the texts, including names, locations, originations, dates, etc. NER system extracts that information from those unstructured text, and could recognize more entities on demand, such as product name, prices, etc. The difficulties in NEW is sometimes to justify the edges of named entity, or to decide whether a word is a named entity or not. This is because named entity relies on the context, while a word could be a name of organization, but a name of person in other contexts.

Body

NER Algorithms

There are four main algorithms of NER in recent years: rules based, unsupervised, pattern based, and deep learning based. This article will mainly focus on deep learning based methods.

1. **Rule based algorithms.**

Rule based algorithms are the early attempts for the NER tasks. Usually some specialists define several rules based on their experiences in some specific fields to recognize those named entities. There methods are rarely used in modern learning tasks now.

2. **Unsupervised algorithms.**

The traditional unsupervised learning methods is clustering. labels are not necessary for the clustering based NER problem. It is recognizing the entities with large scale of expectation based on statistical computations and text similarity.

3. **Pattern based supervised algorithms.**

This method treats NER as a multi-class classification problem on a text sequence. The mostly used machine learning algorithms are HMM, CFR and ME algorithms. SVM methods could work as well, however, the only problem is the context is usually not considered in SVM. Within those methods, CRF usually performs best.

4. **Deep learning based algorithms.**

There are three main reasons that deep learning is a good way of doing NER: a) compared with HMM and CRF, the non-linear activation function can learn complicated patterns from the raw data; b) traditional pattern based algorithms requires lots of engineering techniques and fields knowledge to learn the pattern, while deep learning based algorithms doesn't. c) Deep learning is an end-to-end format, so more complicated NER methods could be designed correspondingly. There are 3 layers of the network structure in the deep learning algorithm in NER:

4.1 Distributed Representation for Input

As vector inputs are required in neural networks, the input string needs to transferred to vector. There are two main ways: one is one-hot module, that each vector only has one dimension with 1, and 0 elsewhere; the other method is to do word embedding, which is a distributed way that represents a lantern pattern in a dimension. There are three levels of doing distributed representation: word-level, character-level, and hybrid representations.

4.2 Context Encoder Architectures

When the distributed representation vectors are obtained, the next step is to extract the encoding pattern based on the context. The most popular methods are CNN, RNN and transformer:

Convolutional Neural Networks (CNN): A typical structure of CNN starts from input string sequence. The distributed representation for input layer looks up the density vector of each word, and make a convolution in CNN and output finally.

Recurrent Neural Networks (RNN): RNN uses biLSTM to get pattern from the input and concatenates the two information. Similar to CNN, there's a distributed representation for input layer that gets the density vector and decode forward and backward separately, then summarize in the tag decoder layer.

Recursive Neural Networks: unlike the RNN, recursive neural network is a typological structure that considers the structure of the whole sentence. It calculates the semantic part from the bottom to the root.

4.3 Tag Decoder Architectures

This layer is used to generate a tag sequence based on the input. The most popular methods are MLP+Softmax, CRF, RNN and Pointer Networks.

Public Datasets of NER

Corpus	Year	Text Source	#Tags	URL
MUC-6	1995	Wall Street Journal texts	7	https://catalog.ldc.upenn.edu/LDC2003T13
MUC-6 Plus	1995	Additional news to MUC-6	7	https://catalog.ldc.upenn.edu/LDC96T10
MUC-7	1997	New York Times news	7	https://catalog.ldc.upenn.edu/LDC2001T02
CoNLL03	2003	Reuters news	4	https://www.clips.uantwerpen.be/conll2003/ner/
ACE	2000 - 2008	Transcripts, news	7	https://www.ldc.upenn.edu/collaborations/past-projects/ace
OntoNotes	2007 - 2012	Magazine, news, conversation, web	89	https://catalog.ldc.upenn.edu/LDC2013T19
W-NUT	2015 - 2018	User-generated text	18	http://noisy-text.github.io
BBN	2005	Wall Street Journal texts	64	https://catalog.ldc.upenn.edu/ldc2005t33
NYT	2008	New York Times texts	5	https://catalog.ldc.upenn.edu/LDC2008T19
WikiGold	2009	Wikipedia	4	https://figshare.com/articles/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500
WiNER	2012	Wikipedia	4	http://rali.iro.umontreal.ca/rali/en/winer-wikipedia-for-ner
WikiFiger	2012	Wikipedia	113	https://github.com/xiaoling/figer
N^3	2014	News	3	http://aksw.org/Projects/N3NERNEDNIF.html
GENIA	2004	Biology and clinical texts	36	http://www.geniaproject.org/home
GENETAG	2005	MEDLINE	2	https://sourceforge.net/projects/bioc/files/
FSU-PRGE	2010	PubMed and MEDLINE	5	https://julielab.de/Resources/FSU_PRGE.html
NCBI-Disease	2014	PubMed	790	https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/
BC5CDR	2015	PubMed	3	http://bioc.sourceforge.net/
DFKI	2018	Business news and social media	7	https://dfki-lt-re-group.bitbucket.io/product-corpus/

Open source Toolkits of NER

NER System	URL
StanfordCoreNLP	https://stanfordnlp.github.io/CoreNLP/
OSU Twitter NLP	https://github.com/aritter/twitter_nlp
Illinois NLP	http://cogcomp.org/page/software/
NeuroNER	http://neuroner.com/
NERsuite	http://nersuite.nlplab.org/
Polyglot	https://polyglot.readthedocs.io
Gimli	http://bioinformatics.ua.pt/gimli
spaCy	https://spacy.io/
NLTK	https://www.nltk.org
OpenNLP	https://opennlp.apache.org/
LingPipe	http://alias-i.com/lingpipe-3.9.3/
AllenNLP	https://allennlp.org/models
IBM Watson	https://www.ibm.com/watson/

Conclusion

This article goes over the definition of named entity recognition, the public datasets, open source toolkits, and four categories of algorithms in NER. It focuses on the deep learning algorithms in NER and explained the three-layers representation of the network structure. Within each layer, several popular methodologies and representations are also introduced.

References

1. Deep active learning for named entity recognition
2. Toward Mention Detection Robustness with Recurrent Neural Networks
3. A Neural Layered Model for Nested Named Entity Recognition