Comparing Conditional Random Fields and LSTM Networks for Named Entity Recognition

Josef Gugglberger (student)   Clemens Sauerwein (supvervisor)

December 23, 2019

LV 703605-6 Masterseminar

universität innsbruck

Motivation

test

Overview

1. Background & Related Work
2. Implementation Details
3. Evaluation and Comparison
4. Conclusion

Background & Related Work

Named Entity Recognition

Definition: NER

Named Entity Recognition is the task of locating and classifying named entities in unstructured text. A named entity is classified into a predefined set of categories.

Named Entity Recognition

Definition: NER

Named Entity Recognition is the task of locating and classifying named entities in unstructured text. A named entity is classified into a predefined set of categories.

James visited the Eiffel Tower in 2012.

↓

James [PERSON] visited the Eiffel [LOCATION] Tower [LOCATION] in 2012 [TIME].

Conditional Random Fields

Definition: CRF

A *Conditional Random Field* is a discriminative probabilistic classifier. It makes its prediction not just based on the input sample, but also based on the context of the input sample.

Conditional Random Fields

Definition: CRF

A *Conditional Random Field* is a discriminative probabilistic classifier. It makes its prediction not just based on the input sample, but also based on the context of the input sample.

$$p(y|x) = \frac{1}{Z(x)} \prod_{t=1}^{T} exp(\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t)) \tag{1}$$

where $Z(x)$ is an normalization function:

$$Z(x) = \sum_{y} \prod_{t=1}^{T} exp(\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t)) \tag{2}$$

Recurrent Neural Networks

Definition: RNN

RNNs are a special type of artificial neural networks that have a feedback loop feeding the hidden layers back into themselves. The loop provides a kind of memory that allow the network to better recognize patterns.

Recurrent Neural Networks

Definition: RNN

RNNs are a special type of artificial neural networks that have a feedback loop feeding the hidden layers back into themselves. The loop provides a kind of memory that allow the network to better recognize patterns.

- Suited for sequence labeling
- Problems with long term dependencies
- Vanishing and exploding gradient

Long-Short-Term-Memory Networks

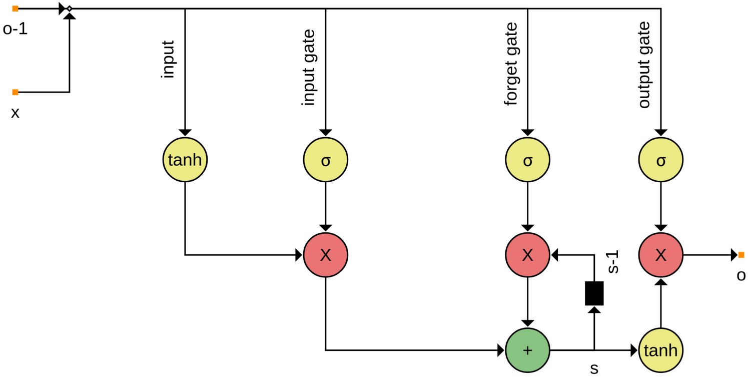
Definition: LSTM networks

LSTM networks are a special case of RNNs, which where designed to overcome issues with the vanishing gradient on long term relationships.

Long-Short-Term-Memory Networks

Definition: LSTM networks

LSTM networks are a special case of RNNs, which where designed to overcome issues with the vanishing gradient on long term relationships.



Implementation Details

Dataset CoNLL

Conference on Computational NL Learning

CoNLL 2003 was a shared task on language independent named entity recognition.

Four types of Named Entities:

- Person
- Location
- Organization
- Miscellaneous

Dataset W-NUT

Workshop on Noisy User-generated Text

W-NUT 17 was a workshop that focused on NLP on noisy and informal text, such as comments from social media.

Four types of Named Entities:

- Person
- Location
- Corporation
- Consumer good
- Creative work
- Group

Dataset Syntax

Word	POS	Syntax Chunk	NE
U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

Conditional Random Fields

Libraries:

- pycrfsuite
- nltk
- gensim

Features should describe characteristics of named entities.

- Word Features
- Sentence & Collection Features
- Dictionary Features
- Features from unsupervised ML algorithms

Word Features

- length of word
- the word starts with an upper-case letter
- the word contains an upper-case letter
- the word contains a digit
- the word contains a special character (-, /, etc.)
- word shape: 'Word' $\rightarrow$ 'Aa+', 'WORD' $\rightarrow$ 'A+', '2019-12-12' $\rightarrow$ '9999#99#99'

Sentence & Collection Features

- position of word in sentence
- number of occurrences in collection

Dictionary Features

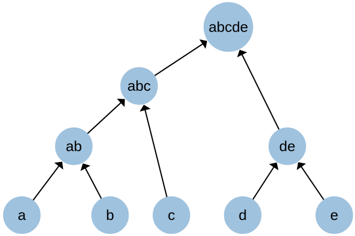
The word is contained in:

- stop-words list
  - is, as, the, are, has, that, etc.
  - Problems: 'The Who', 'Take That'
- name list
  - 7579 person names form nltk corpus
- word list
  - dictionary of 235892 words from nltk corpus
- wordnet
  - dictionary and thesaurus
  - provides hypernyms, synonyms, etc.

Features from unsupervised ML algorithms

The cluster of each word is used as a feature.

- brown cluster
  - hierarchical clustering algorithm



Features from unsupervised ML algorithms

The cluster of each word is used as a feature.

- brown cluster
  - hierarchical clustering algorithm
- Latent Dirichlet Allocation (LDA) topic
  - modelling the abstract topics of document
  - example: document A is 20% topic 1, 60% topic 2 and 20% topic 3

Features from unsupervised ML algorithms

The cluster of each word is used as a feature.

- brown cluster
  - hierarchical clustering algorithm
- Latent Dirichlet Allocation (LDA) topic
  - modelling the abstract topics of document
  - example: document A is 20% topic 1, 60% topic 2 and 20% topic 3
- gensim implementation of w2v cluster
  - maps similar words to similar vectors
  - $w2v(king) - w2v(man) + w2v(woman) =\sim w2v(queen)$

LSTM Network

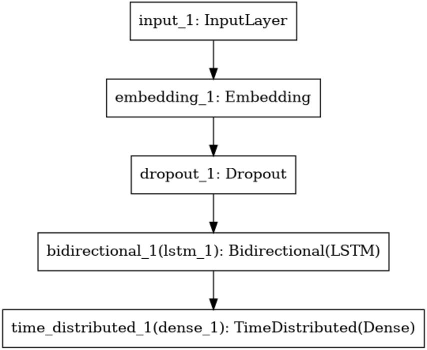
Libraries:

- Keras functional API
- Tensorflow as backend

LSTM Network

Libraries:

- Keras functional API
- Tensorflow as backend



Bidirectional LSTM Layer

- idea is to duplicate LSTM layer
  - input as-is is feed into first LSTM layer
  - input reversed is feed into second LSTM layer
- speech depends on context past and future

Bidirectional LSTM Layer

- idea is to duplicate LSTM layer
    - input as-is is feed into first LSTM layer
    - input reversed is feed into second LSTM layer
- speech depends on context past and future

Example:

The other day we saw Paris .

Bidirectional LSTM Layer

- idea is to duplicate LSTM layer
  - input as-is is feed into first LSTM layer
  - input reversed is feed into second LSTM layer
- speech depends on context past and future

Example:

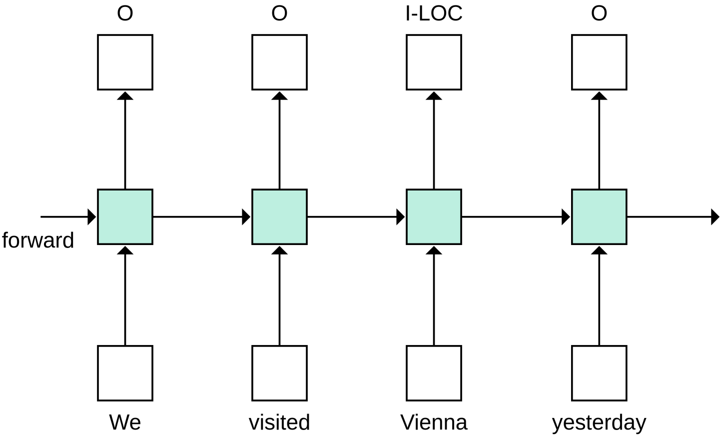
The other day we saw Paris Hilton.

Time Distributed Dense Layer

- adds the same dense layer to every timestep

Time Distributed Dense Layer

- adds the same dense layer to every timestep



The best of both worlds?

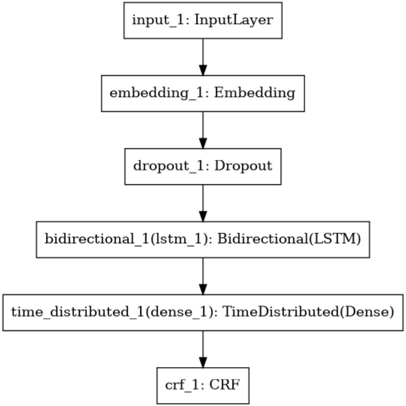
combine the LSTM approach with CRF by adding a CRF layer at bottom:

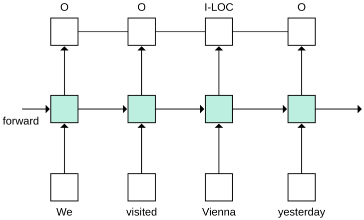
- use past input features via LSTM layer
- use sentence level tag information via CRF layer

The best of both worlds?

combine the LSTM approach with CRF by adding a CRF layer at bottom:

- use past input features via LSTM layer
- use sentence level tag information via CRF layer





Evaluation and Comparison

Evaluation

Evaluation of the implemented NER systems with metrics:

- accuracy
- recall
- F1-score

Evaluation

Evaluation of the implemented NER systems with metrics:

- accuracy
- recall
- F1-score

Evaluation performed based on:

- token level
- named entity level (CoNLL standard)

Results CoNLL Dataset

token based:

Method	Accuracy	Recall	F1-score
CRF	0.873	0.958	**0.914**
Bi-LSTM	**0.934**	0.893	0.913
Bi-LSTM-CRF	0.853	**0.963**	0.904

named entity based:

Method	Accuracy	Recall	F1-score
CRF	0.867	0.947	**0.905**
Bi-LSTM	0.851	**0.959**	0.902
Bi-LSTM-CRF	**0.906**	0.864	0.884

Results WNUT Dataset

token based:

Method	Accuracy	Recall	F1-score
CRF			
Bi-LSTM			
Bi-LSTM-CRF			

named entity based:

Method	Accuracy	Recall	F1-score
CRF			
Bi-LSTM			
Bi-LSTM-CRF			

Conclusion