OSA
BERLIN, 2017

Entity Classification

Conversational Intelligence

- Stefano Martina
- Fotini Simistira
- Saurabh Varshneya
- Kumar Shridhar
- Vassilis Katsouros

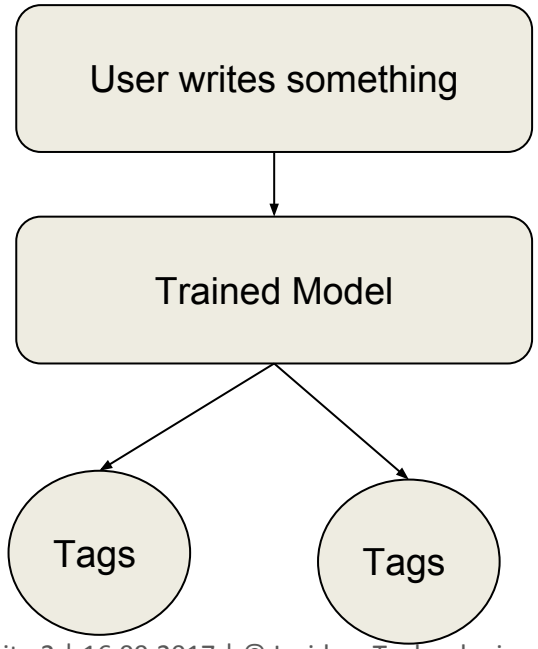
Insiders Technologies

MindGarage & insiders technologies

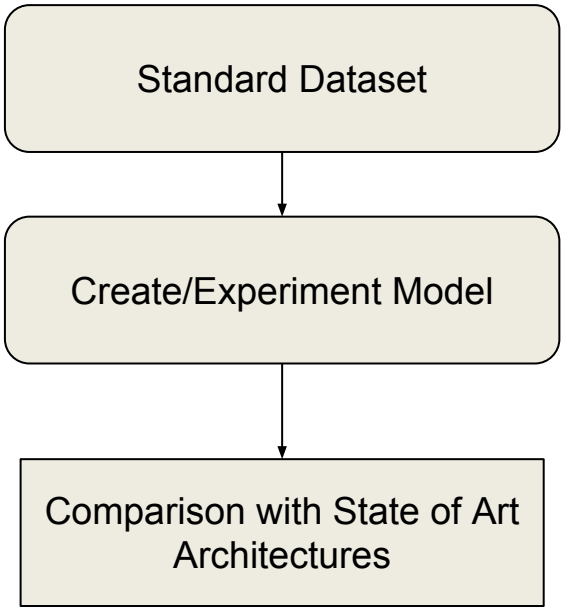
Tasks Explained

OSA BERLIN, 2017

Use Case



Scientific Experimentation



USE CASE

OSA BERLIN, 2017

Recognise Area in house size

E.g: 50 square meters

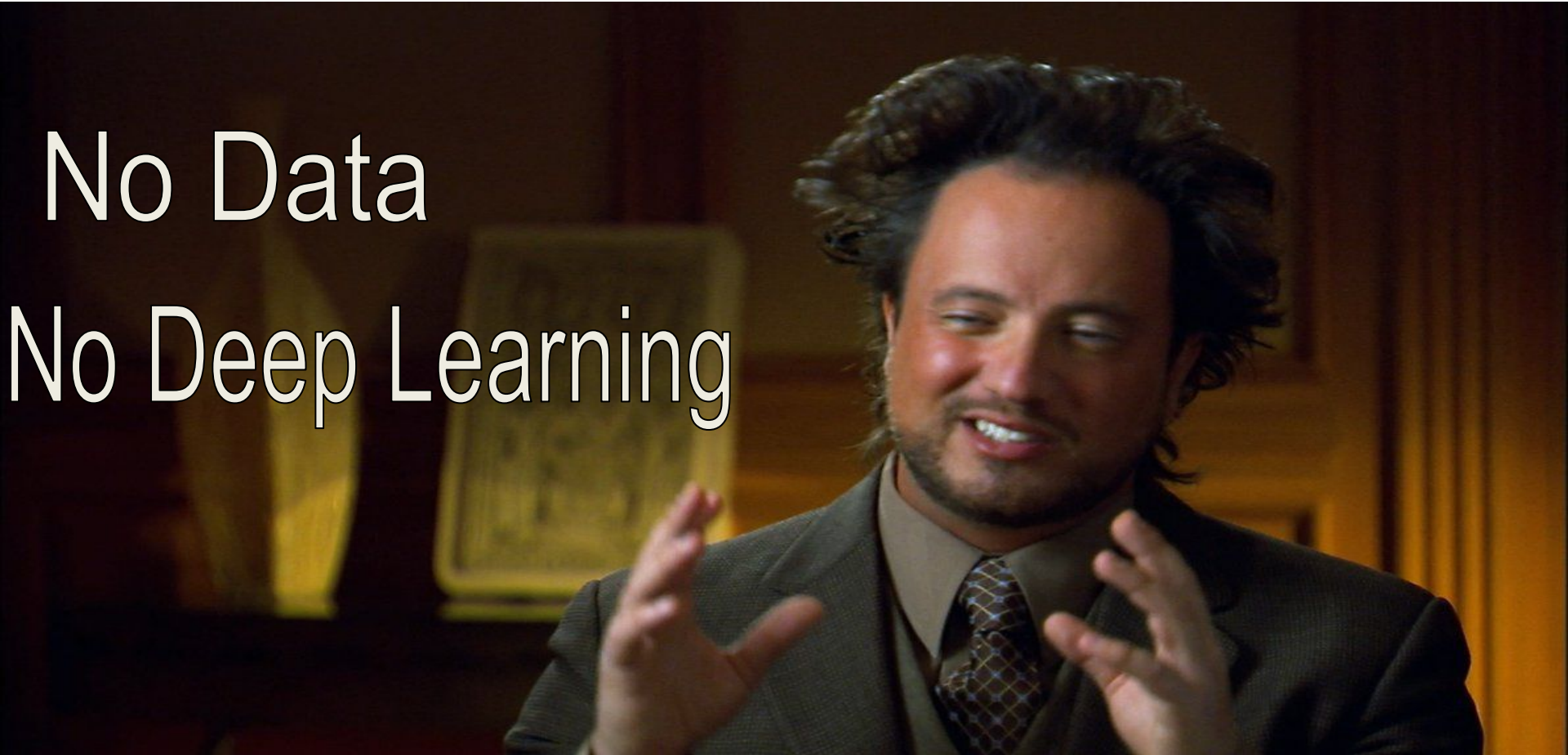
Recognize Address

-Street Address
-Street number
-Zip Code
-City (Berlin)

Recognize Date

E.g: First week of September

Problem

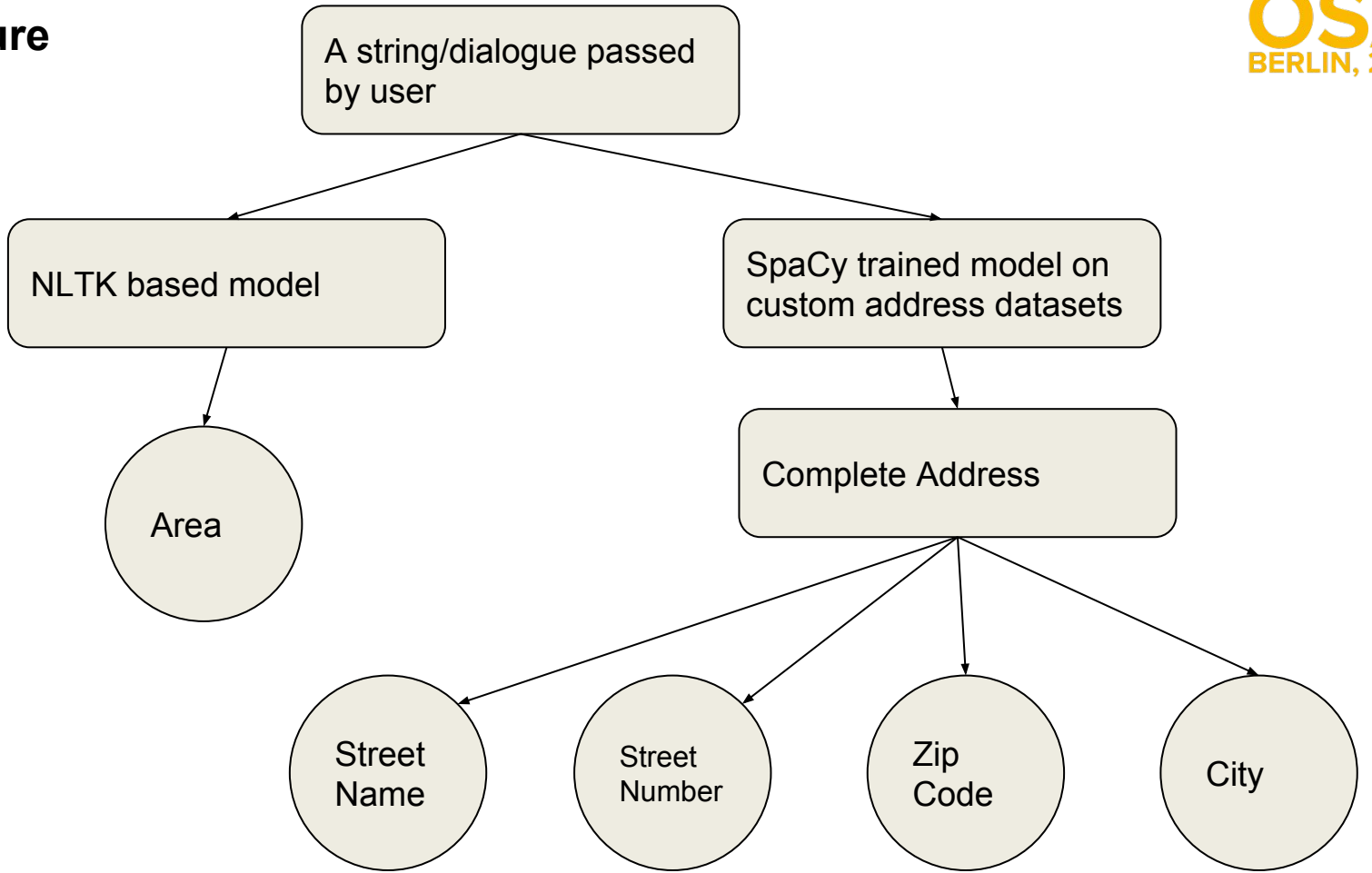


Solution 1

OSA BERLIN, 2017

1. Created Artificial Dataset (Synthetic Dataset)
2. Scraped list of all streets in Berlin and added to synthetic dataset
3. Trained Spacy Model + NLTK + Regex + Traditional Classifiers over it to identify the street addresses
4. Evaluated on test dataset (Self created test dataset)

Architecture



ADDRESS Entity Recognition using Spacy

OSA BERLIN, 2017

- Generated a synthetic corpus of 10k sentences for training
- Dialogues taken from **Maluuba fake chatbot dialogues**
- Addresses taken from **Tourpedia**

Dataset Sample

OSA BERLIN, 2017

Good afternoon, I'd <ADDRESS_start>Kantstrasse 152, Berlin, Germany<ADDRESS_end> like to book a trip with my son from August 18 to August 29. He and I would be leaving from Curitiba. What destinations can you offer?

:smile: <ADDRESS_start>Am Treptower Park 32, Berlin, Germany<ADDRESS_end>

Anything with <ADDRESS_start>Lychener Strasse 33, Berlin, Germany<ADDRESS_end> a higher rated hotel?

<ADDRESS_start>Blankenfelde, Berlin, Germany<ADDRESS_end> Try Cleveland

<ADDRESS_start>Olafstrasse 65, Berlin, Germany<ADDRESS_end> detroit

Yes, I just need <ADDRESS_start>Kantstrasse 110, Berlin, Germany<ADDRESS_end> to get back by Sepetember 12.

Evaluation: ADDRESS Entity Recognition using Spacy

OSA BERLIN, 2017

User generated corpus of 56 sentences for testing: 95.54% (54/56)

I am very upset because you didn't want to make a special price for my apartment in Jasmunder str.
After all I am your customer since January 1982, I still lived in DDR!
Thousands of demonstrators have marched through Ragower Weg 270, Berlin to protest the war in Pfarrer-Lenzel-Strasse 251 and demand the withdrawal of British troops from that country .
The Hausburgstrasse 205, Berlin march came ahead of anti-war protests today in other cities , including Eppinger Strasse 110, Berlin , Jathoweg 93 , and Fanny-Zobel-Strasse 202
Next month I will go to live in Woennichstrasse, not in Weitlingstrasse.
43sq m. Sterndann 22, Berlin.

Solution 2

OSA BERLIN, 2017

1. Took the CoNLL 2002 dataset that has B-geo and I-geo tags
   E.g: White House
   White: B-geo
   House: I-geo

2. Removed all the predefined B-geo and I-geo tags and replaced all B-geo and I-geo tags with street names
   E.g: Saarbrücker Straße
   Saarbrücker : B-geo
   Straße : I-geo

3. Trained a BLSTM over it

Deep Learning

What society thinks I do

What my friends think I do

What other computer scientists think I do

What mathematicians think I do

What I think I do

```
In [1]:
import keras
Using TensorFlow backend.
```

What I actually do

Model Architecture

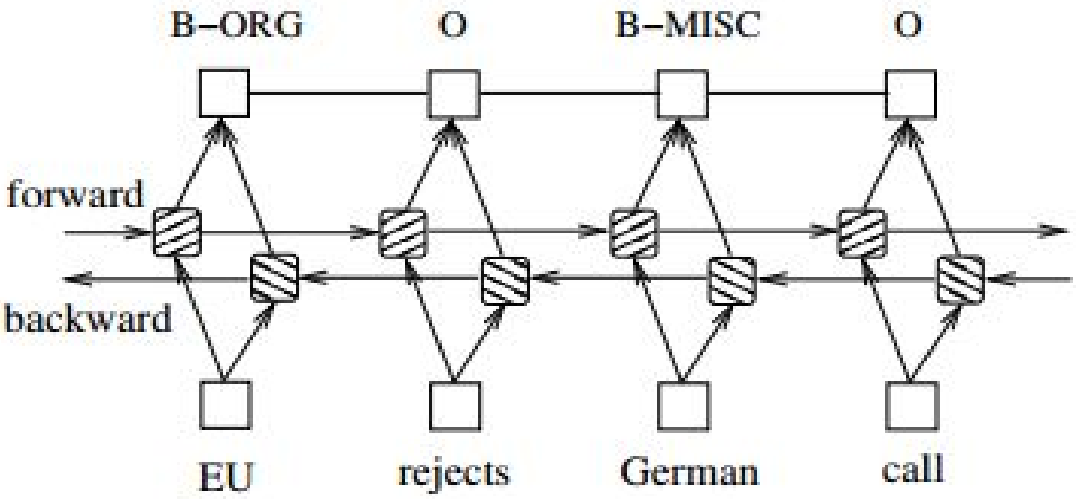


Figure 7: A BI-LSTM-CRF model.

Things we tried and where we settled

OSA BERLIN, 2017

1. Embeddings: Char vs Word vs Sentence
   Could not see any difference: Word

2. New word in embedding space
   Default value (E.g 0.25) vs training the new word: Trained it

3. BLSTM vs Seq2Seq
   BLSTM performed better on our datasets

4. Features along with Words
   with/without POS and NER tags: With POS and NER tags (Gazetteer dense knowledge space wouöd have f further improved the results)

5. CRF vs CTC : Could not try it

Scientific Experimentation

OSA
BERLIN, 2017

NER task competitions

OSA
BERLIN, 2017

2017 The 3rd Workshop on Noisy User-generated Text (W-NUT)

September 7th, Copenhagen (at EMNLP 2017)

BEST SUBMISSION:

Gustavo Aguilar, Suraj Maharjan, A. Pastor Lopez-Monroy and Thamar Solorio

Department of Computer Science

University of Houston

WNUT17 NER task

OSA BERLIN, 2017

Dataset

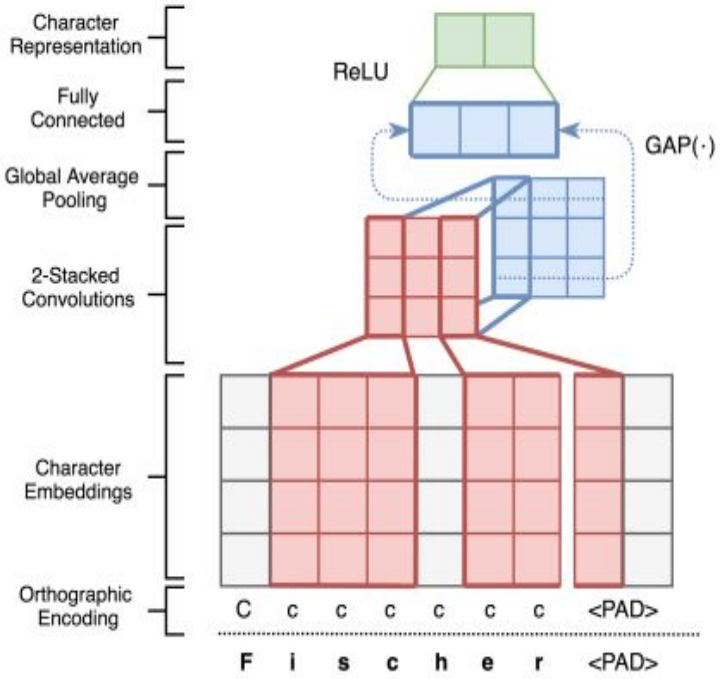
- Twitter, Reddit, YouTube, StackExchange

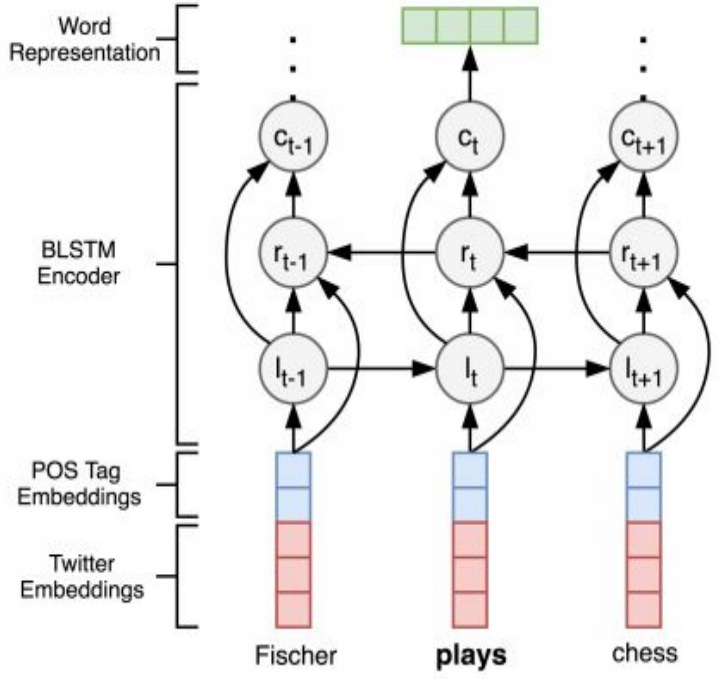
Target Entities

- person
- location (including GPE, facility)
- corporation
- product (tangible goods, or well-defined services)
- creative-work (song, movie, book and so on)
- group (subsuming music band, sports team, and non-corporate organisations)

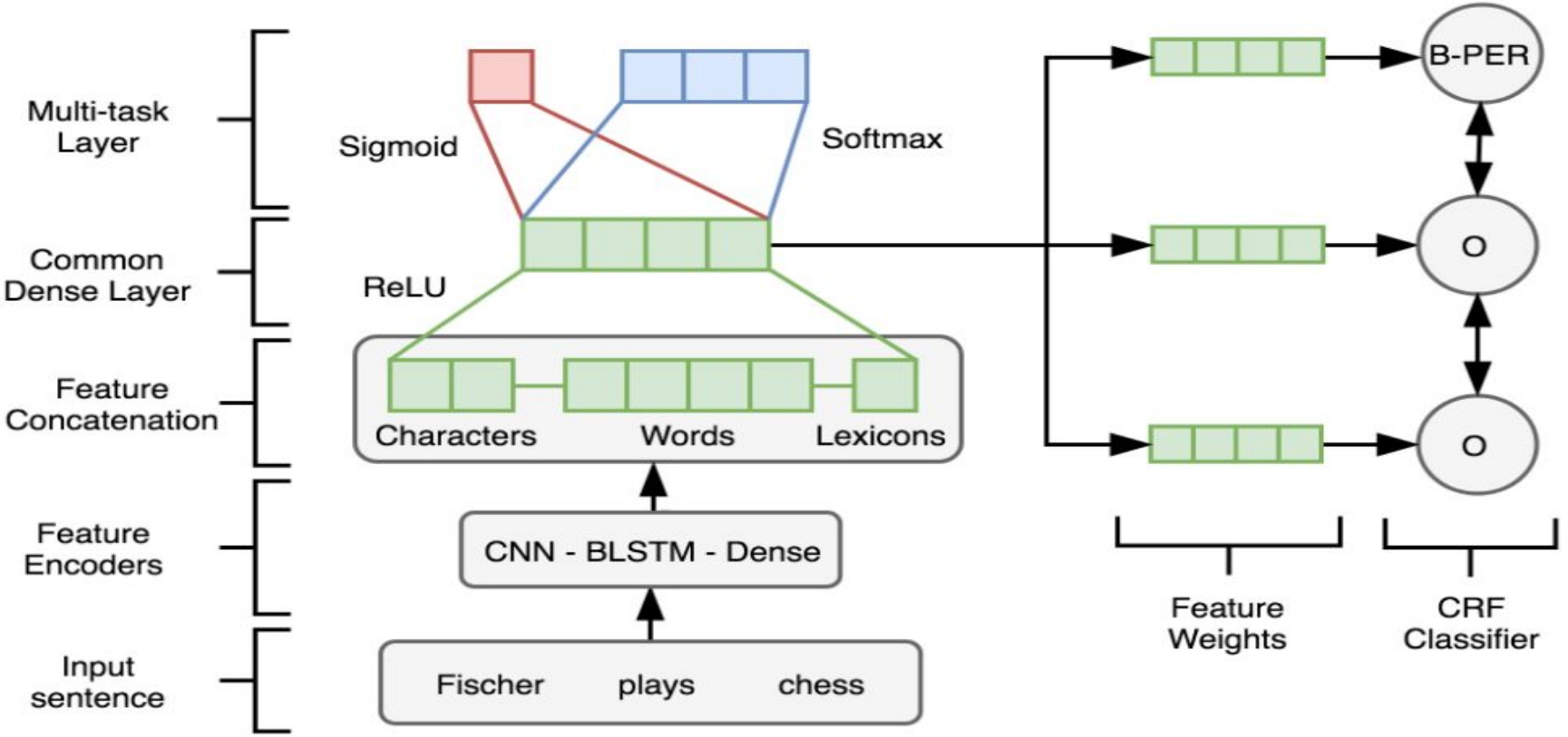
Architecture I

OSA
BERLIN, 2017





Architecture II



Experiments tried out

OSA BERLIN, 2017

- **Architectures and datasets**
  - Twitter embeddings (http://www.fredericgodin.com/software/)
    - WNUT17 dataset (SOA)
    - ACNER dataset
    - CONLL2003 dataset
  - Word embedding GloVe (https://nlp.stanford.edu/projects/glove/)
    - CONLL2003 dataset

Experimental results I

OSA BERLIN, 2017

- **ACNER** Dataset Twitter embeddings

	prec.	recall	f1-score	support
B-art	43%	7%	12%	85
B-eve	40%	35%	37%	57
B-geo	86%	91%	88%	7702
B-gpe	98%	94%	96%	3162
B-nat	55%	38%	45%	47
B-org	80%	70%	75%	3987
B-per	85%	84%	84%	3441
B-tim	92%	89%	90%	3998
I-art	43%	18%	25%	51
I-eve	23%	22%	22%	45
I-geo	82%	79%	81%	1528
I-gpe	93%	78%	85%	32
I-nat	17%	7%	10%	15
I-org	77%	80%	79%	3281
I-per	85%	90%	88%	3462
I-tim	80%	79%	80%	1292
AVG	85.20%	84.75%	84.84%	

Experimental results II

OSA BERLIN, 2017

- CONLL2003 Dataset
  - Twitter embeddings

	prec.	recall	f1-score	support
B-LOC	90%	92%	91%	1668
B-MISC	82%	84%	83%	702
B-ORG	86%	85%	85%	1661
B-PER	95%	94%	95%	1617
I-LOC	80%	81%	81%	257
I-MISC	62%	71%	67%	216
I-ORG	77%	86%	81%	835
I-PER	97%	97%	97%	1156
AVG	88.08%	89.46%	88.75%	

  - GloVe embeddings

	prec.	recall	f1-score	support
B-LOC	92%	93%	93%	1668
B-MISC	82%	82%	82%	702
B-ORG	88%	89%	89%	1661
B-PER	96%	95%	95%	1617
I-LOC	79%	88%	83%	257
I-MISC	59%	73%	65%	216
I-ORG	81%	88%	84%	835
I-PER	98%	99%	98%	1156
AVG	89.55%	91.28%	90.35%	

- **Best SOA performance (f1-score): 90.90%**

Code available

OSA BERLIN, 2017

- **Github repo**
-
  - push request to mindgarage/Ovation
  - branch team/NamedEntityRecognition

OSA
BERLIN, 2017

Thank You!

OVATION Summer Academy

MindGarage & insiders technologies