Towards multilingual sequence labelling using cross-lingual word embeddings

Chao Li, Carsten van Weelden, Luigi Lorato, Carsten L. Hansen, Mihai Rotaru and Jakub Zavrel

Today

- Textkernel
- BI-LSTM Sequence labeling Model
- Embedding Alignment and Multilingual model
- Conclusion and Future work

Textkernel

- Focus: NLP applications in HR domain
  - CV Parsing for 17 languages
  - Job parsing for 7 languages
  - Semantic search
  - Recommended jobs and candidates
- Started in 2001, located in Amsterdam
- 100+ people, +30 nationalities

Outline

- Textkernel
- BI-LSTM Model for Sequence Labeling
- Embedding Alignment and Multilingual model
- Conclusion and Future work

CV

Fiona Lee City: Paris

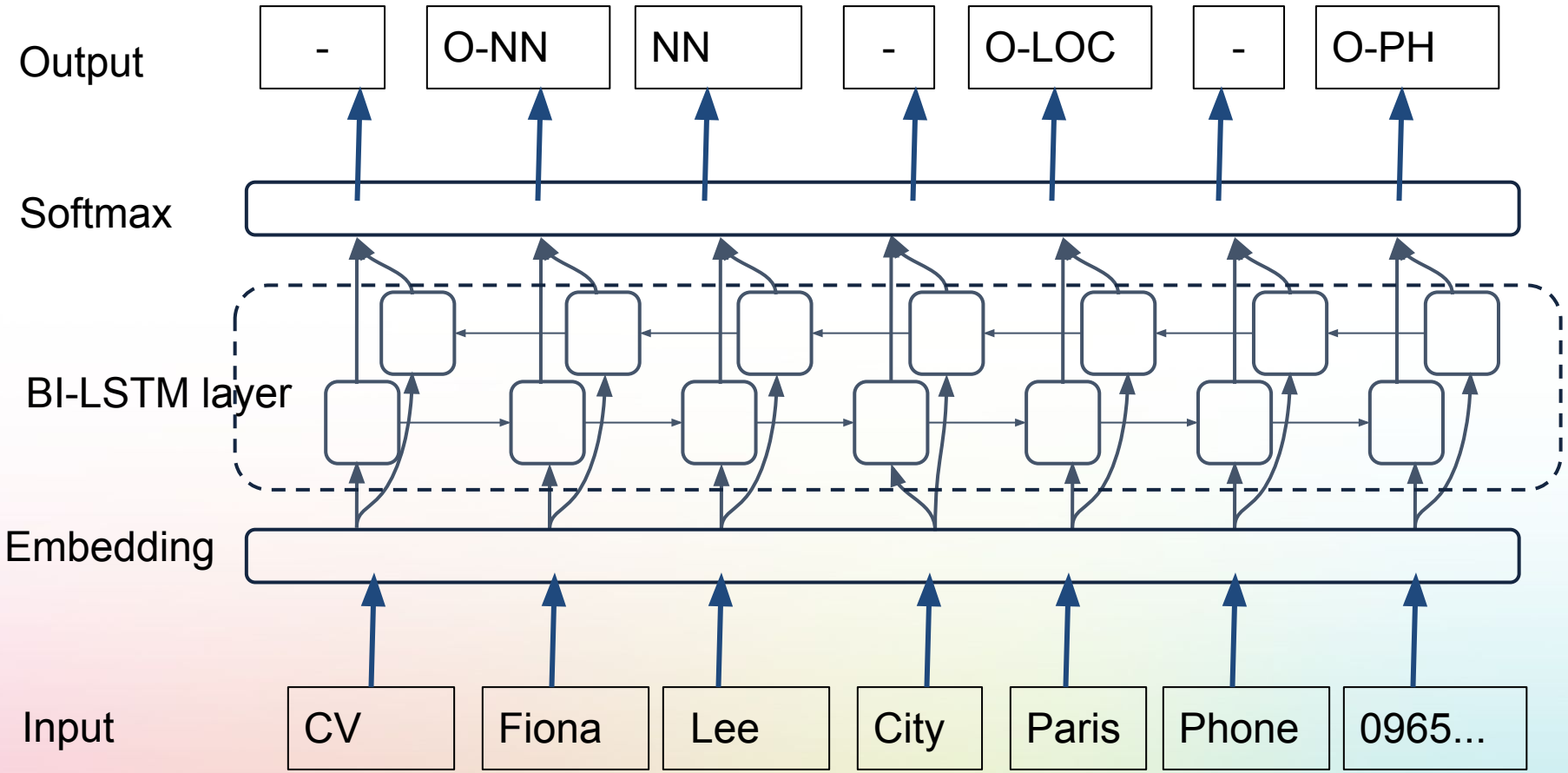
Phone: 0965…..

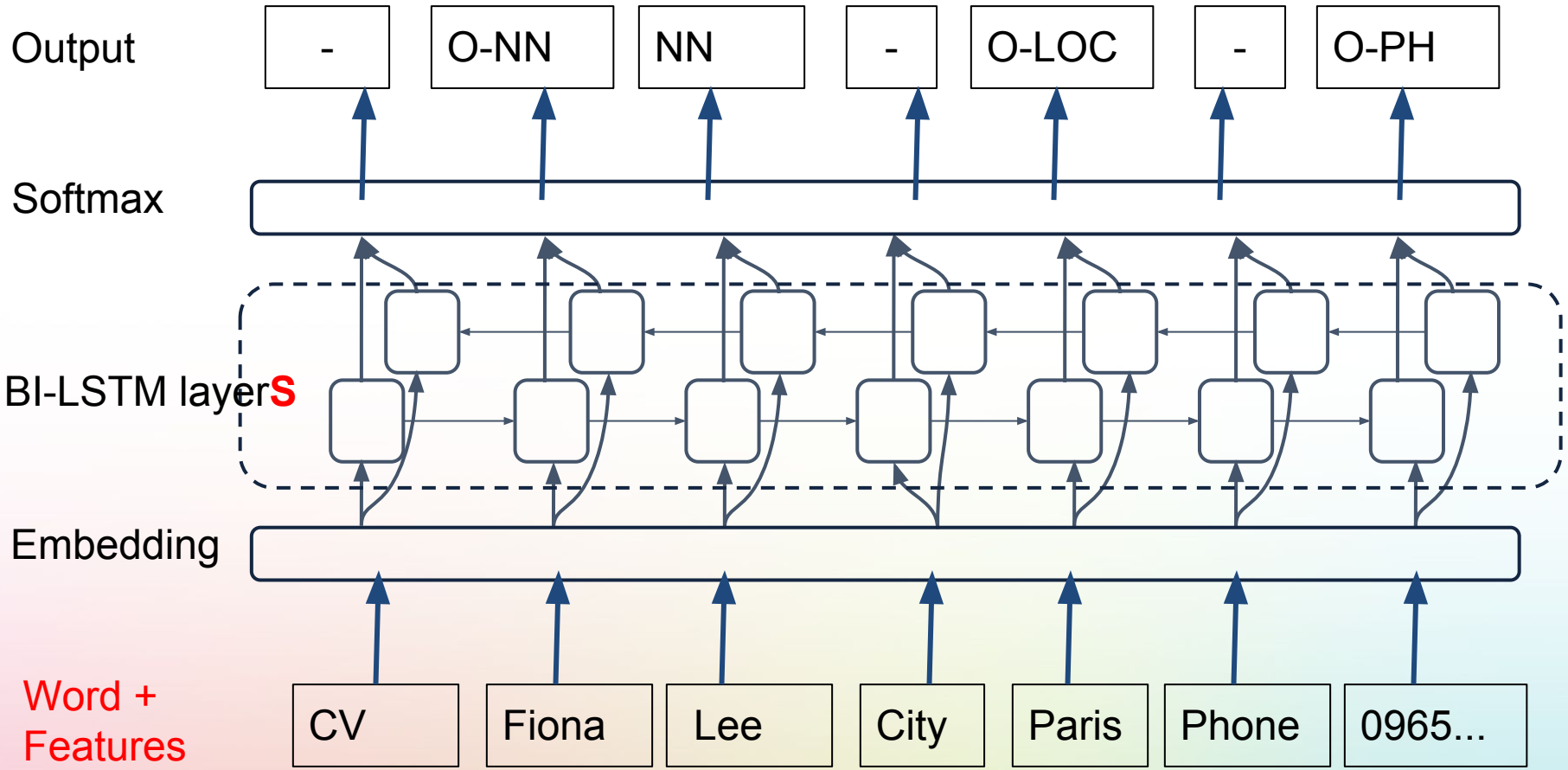
Work Experience

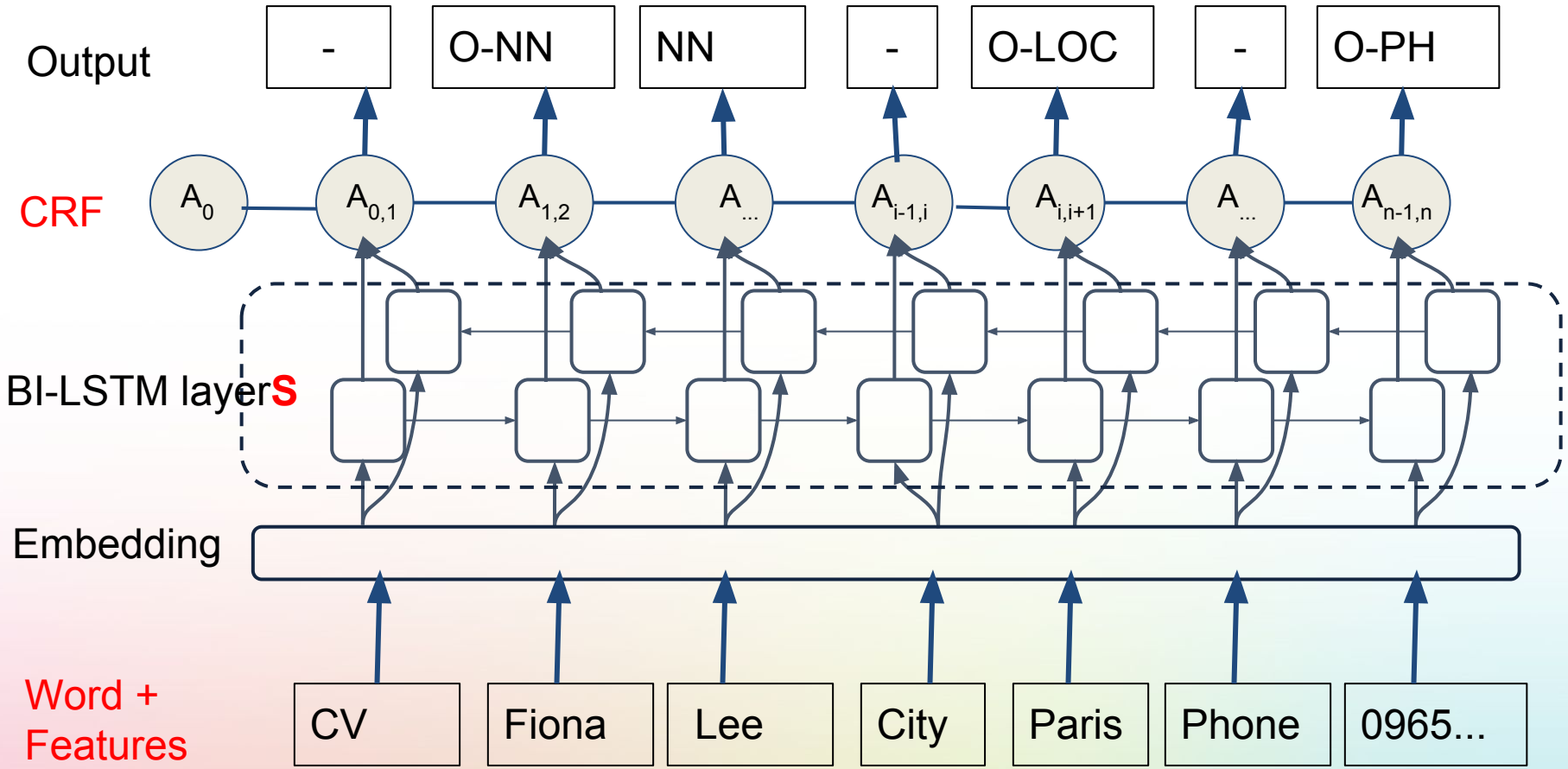
1996-2000 Java Engineer IBM, Amsterdam

…….

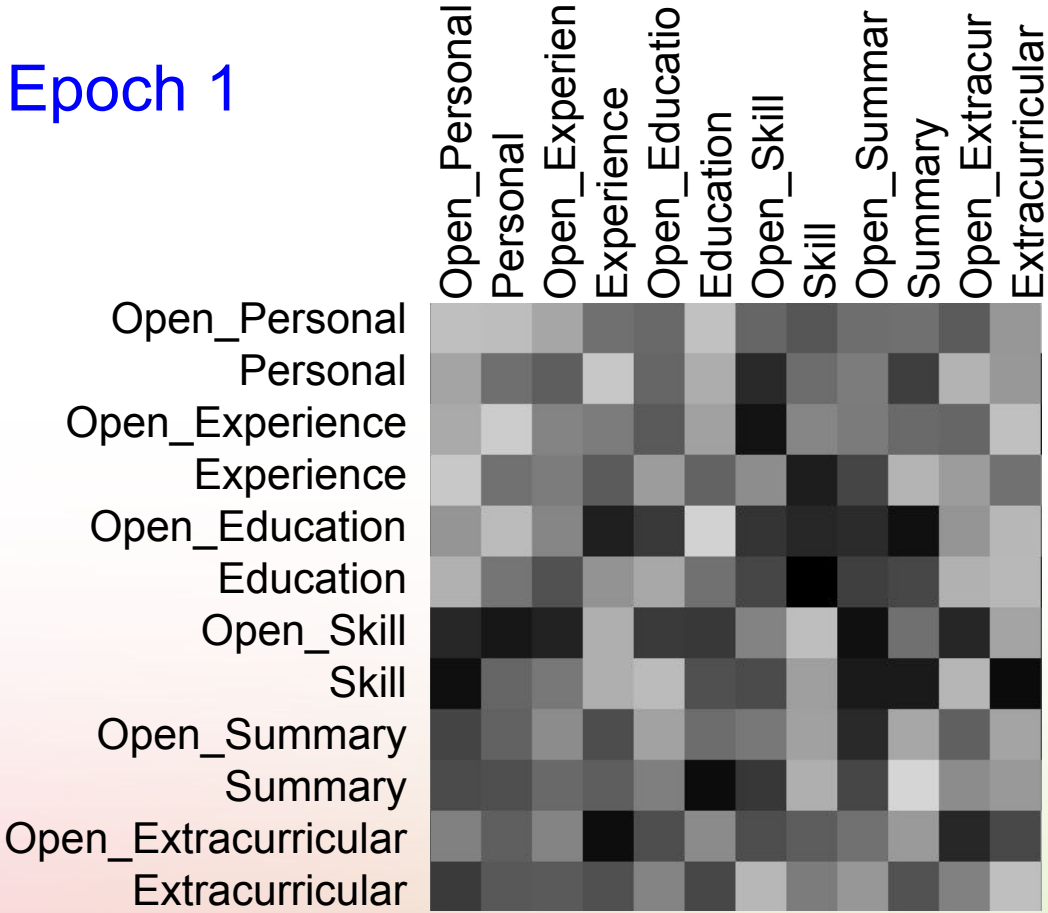
CV	Fiona	Lee	City	Paris	Phone	0965...


Output
-
O-NN
NN
-
O-LOC
-
O-PH
Softmax
BI-LSTM layer
Embedding
Input
CV
Fiona
Lee
City
Paris
Phone
0965...

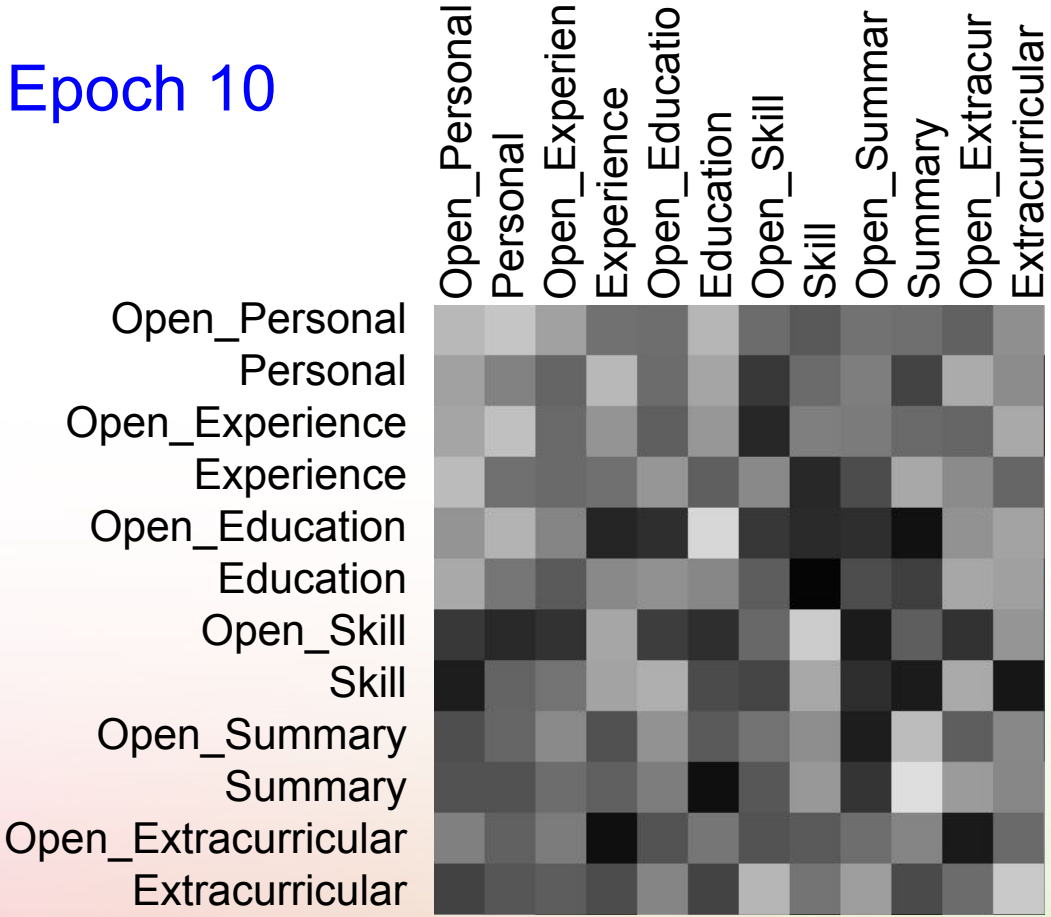

Output
-
O-NN
NN
-
O-LOC
-
O-PH
Softmax
BI-LSTM layerS
Embedding
Word +
Features
CV
Fiona
Lee
City
Paris
Phone
0965...


Output
-
O-NN
NN
-
O-LOC
-
O-PH
CRF
A0
A0,1
A1,2
A...
Ai-1,i
Ai,i+1
A...
An-1,n
BI-LSTM layerS
Embedding
Word + Features
CV
Fiona
Lee
City
Paris
Phone
0965...

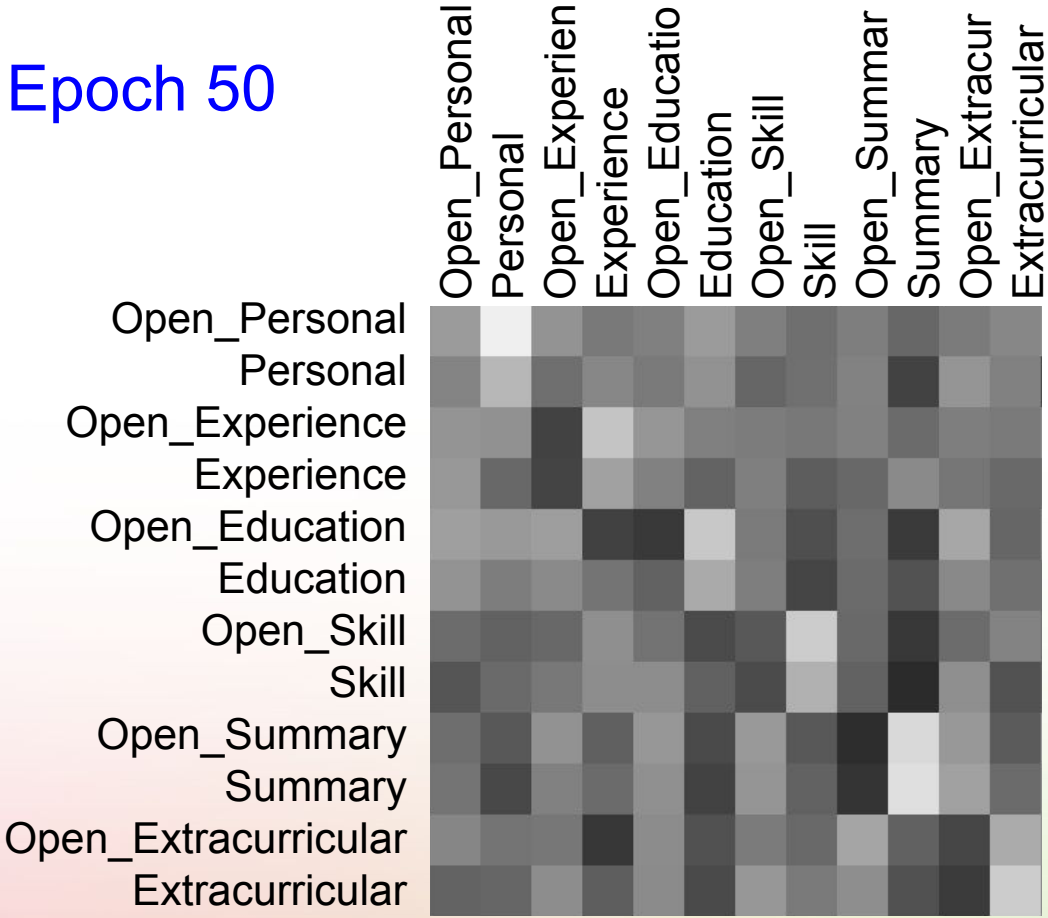
Train: Epoch 1



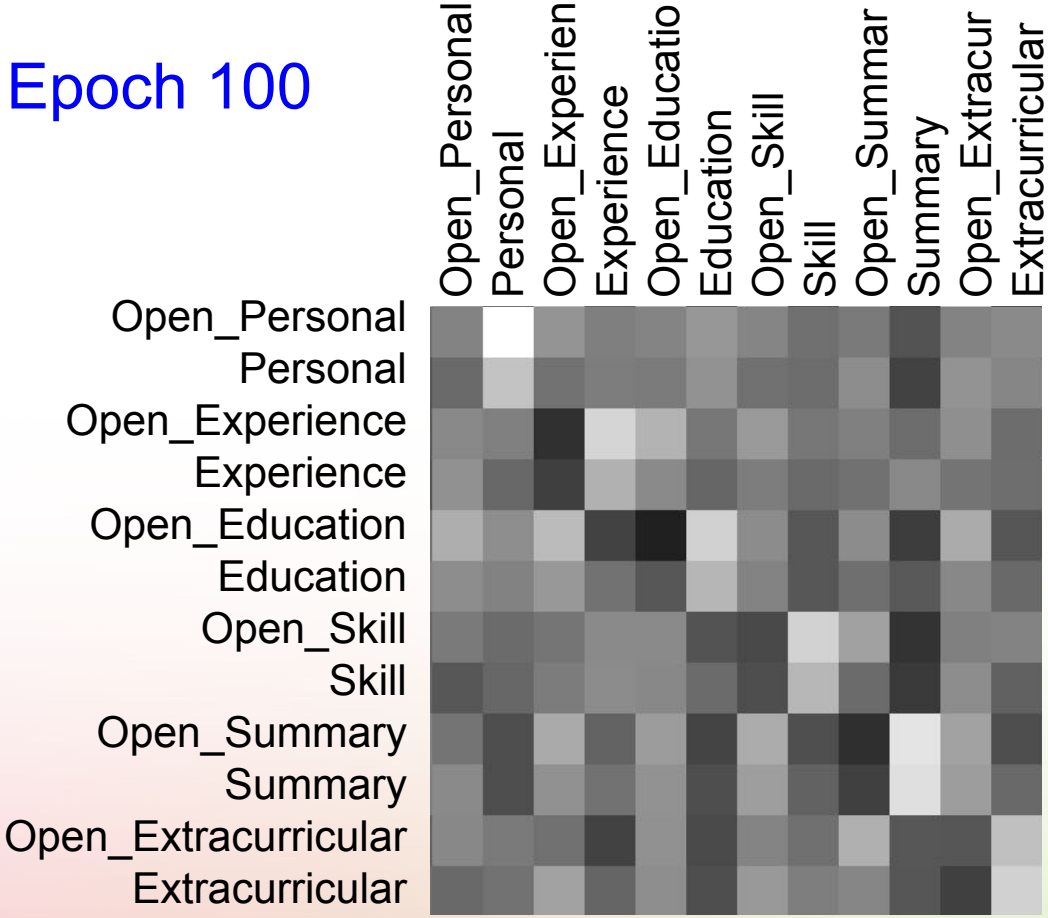
Train: Epoch 10

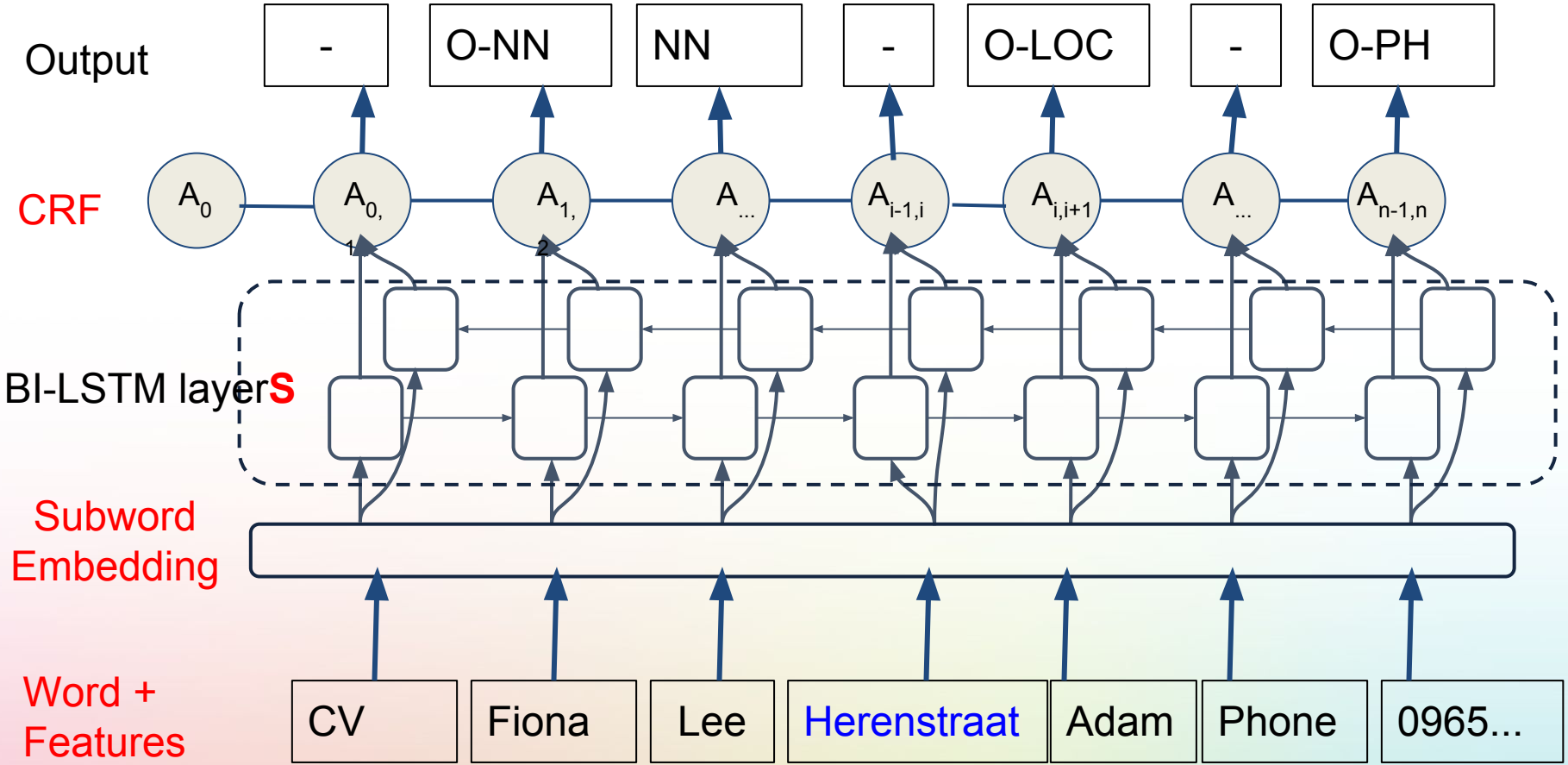


Train: Epoch 50



Train: Epoch 100

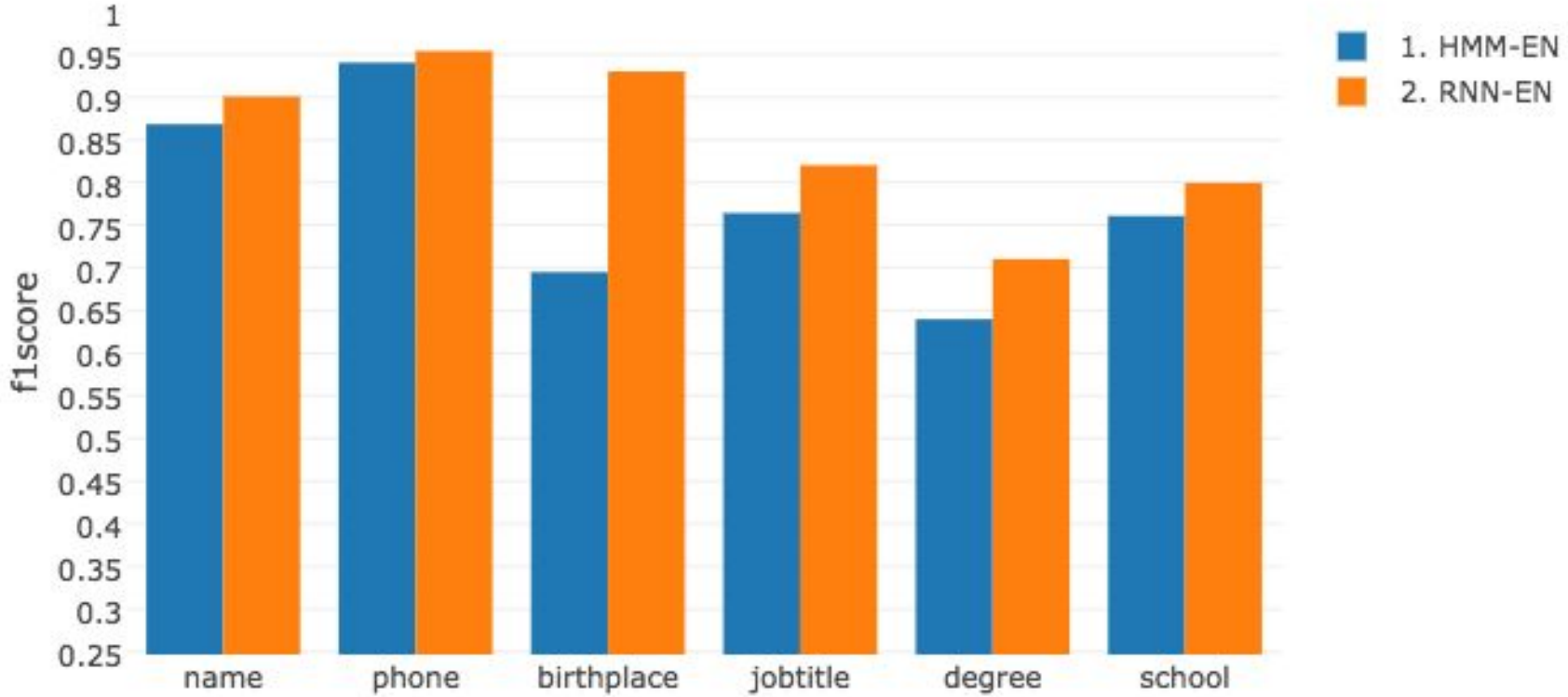



Output
-
O-NN
NN
-
O-LOC
-
O-PH
CRF
A0
A0,1
A1,2
A...
Ai-1,i
Ai,i+1
A...
An-1,n
BI-LSTM layerS
Subword Embedding
Word + Features
CV
Fiona
Lee
Herenstraat
Adam
Phone
0965...

NN Sequence labeling Model

- word based
  - Multilayer BI-LSTM (dropout, batchNorm)
  - CRF layer to improve the label consistency
  - Subword embedding to eliminate unknown words

- Phrase based
  - CNN to compose phrase embeddings


1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
f1score
name
phone
birthplace
jobtitle
degree
school
1. HMM-EN
2. RNN-EN

Outline

- Textkernel
- BI-LSTM Model for Sequence Labeling
- Embedding Alignment and Multilingual model
- Conclusion and Future work

Multilingual CV parsing model

- Hard to get training data for new languages
- CVs written similarly across languages/countries

简历

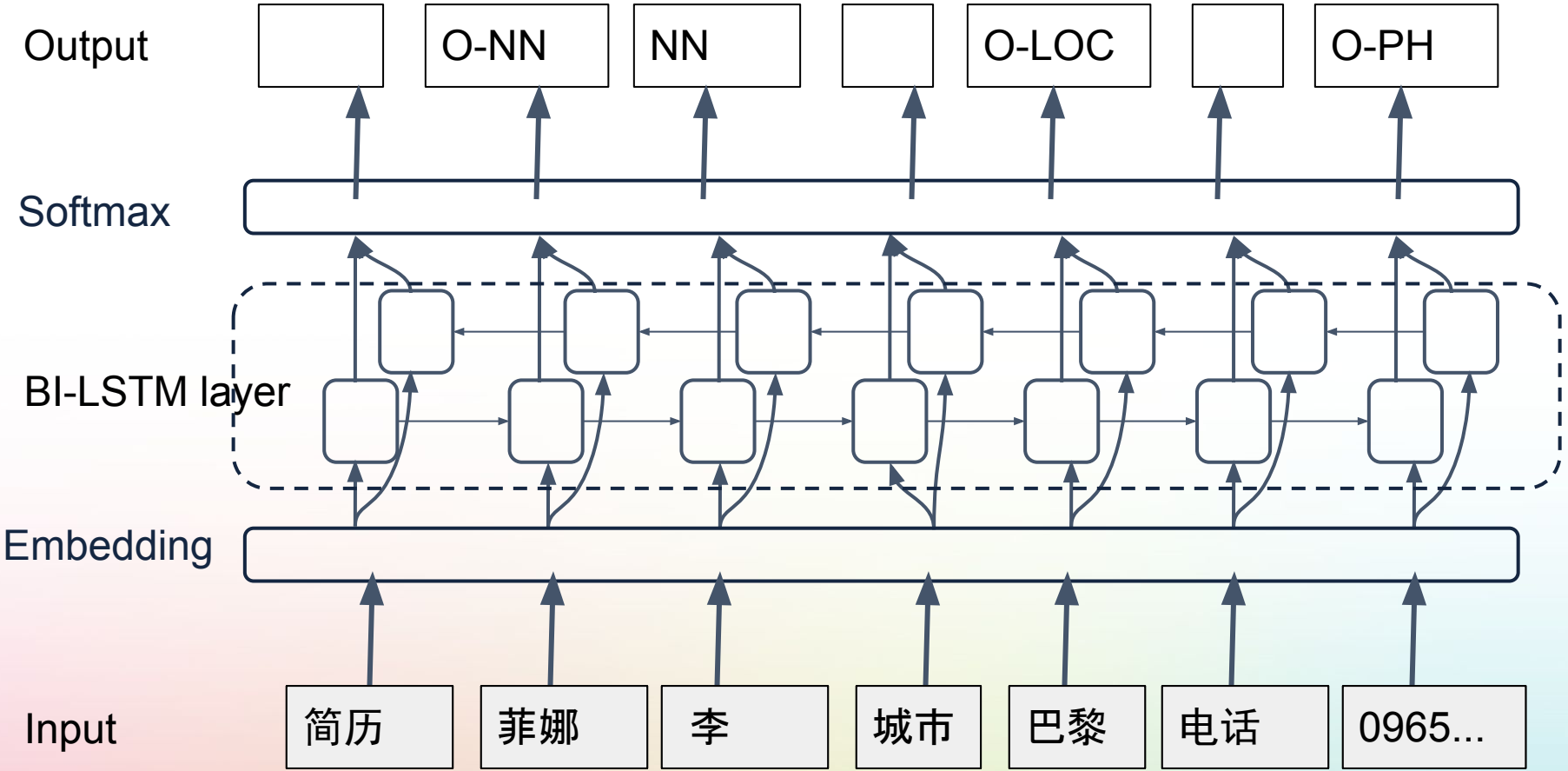
菲娜 李 城市: 巴黎

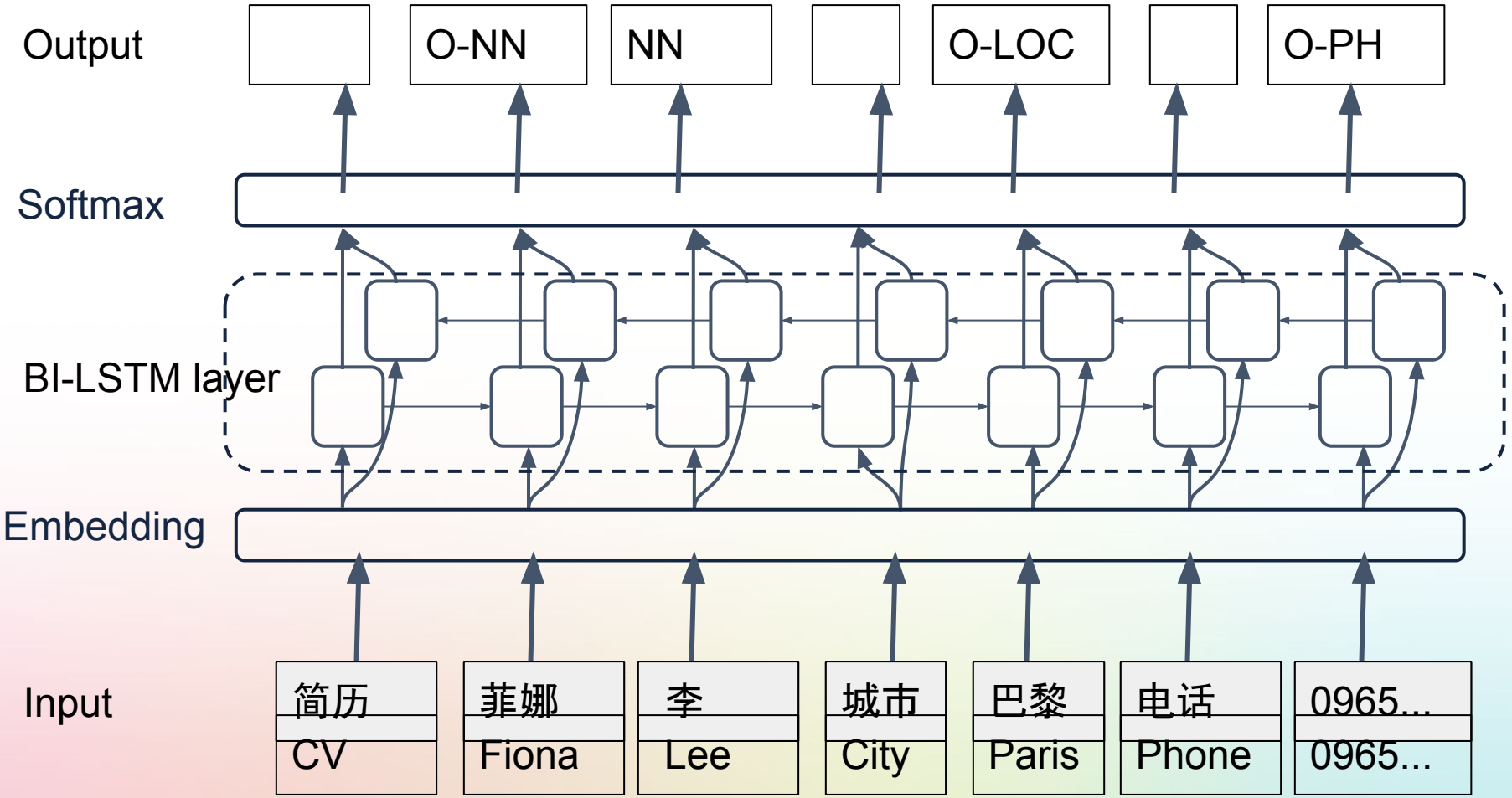
电话: 0965…..

工作经历

1996-2000 Java 工程师 IBM, 阿姆斯特丹

…….


Output
O-NN
NN
O-LOC
O-PH
Softmax
BI-LSTM layer
Embedding
Input
简历
菲娜
李
城市
巴黎
电话
0965...


Output
O-NN
NN
O-LOC
O-PH
Softmax
BI-LSTM layer
Embedding
Input
简历
CV
菲娜
Fiona
李
Lee
城市
City
巴黎
Paris
电话
Phone
0965...
0965...

Multilingual CV parsing model

- CVs written similarly across languages/countries

- Align embeddings across languages:

CN to EN: Find **U** and **V** such that S_{en} and S_{cn} are maximum correlated:

- $S_{en} = U * WV_{en}$
- $S_{cn} = V * WV_{cn}$

Multilingual CV parsing model

- CVs written similarly across languages/countries

- Align embeddings across languages

CN to EN: mapping Chinese embedding to English vector space

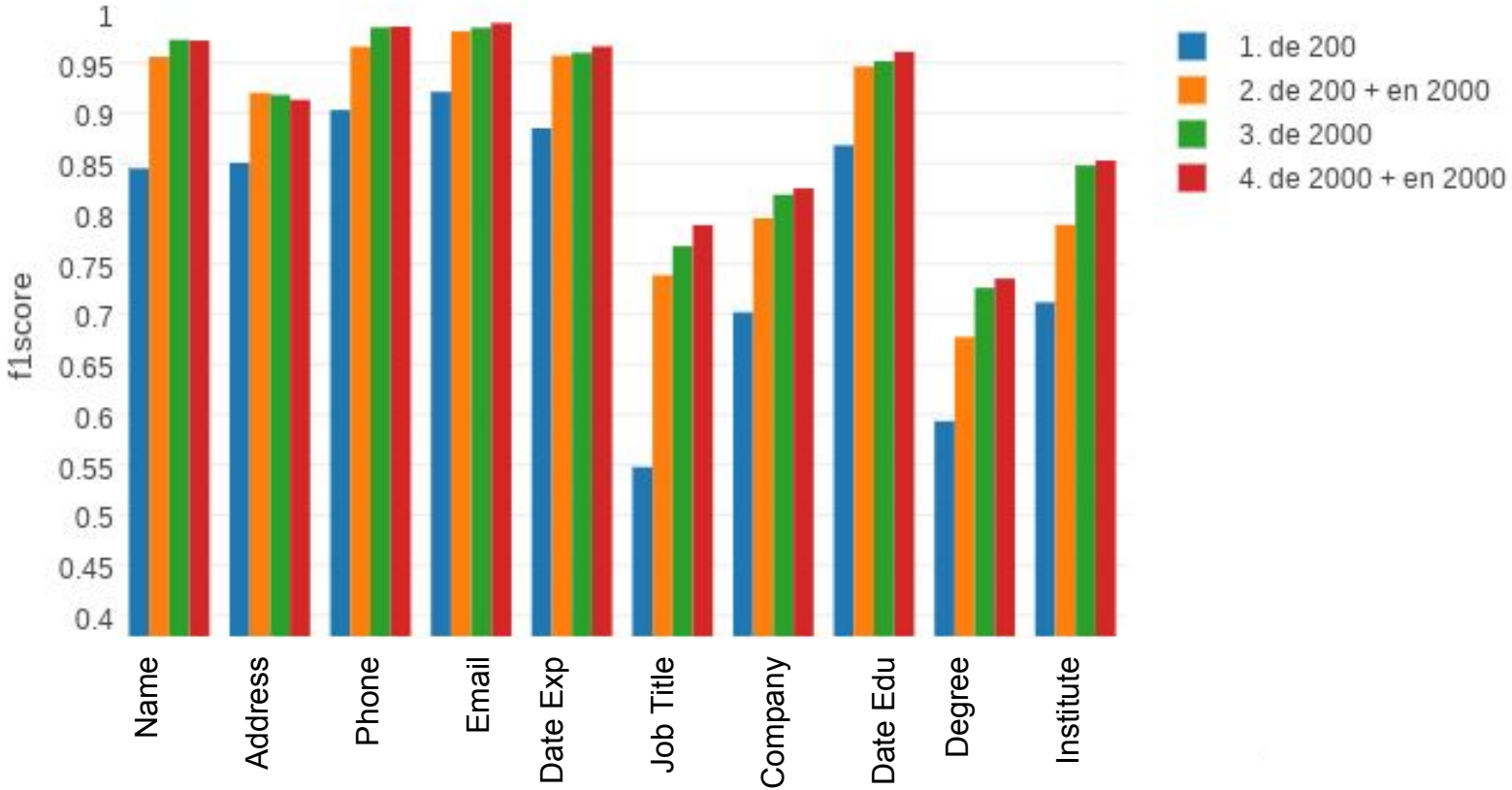
- $S^*_{en} = U^{-1} * U * WV_{en} = WV_{en}$
- $S^*_{cn} = U^{-1} * V * WV_{cn}$

Multilingual CV parsing model

- CVs written similarly across languages/countries

- Align embeddings across languages

  Embedding Alignment


f1score
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
Name
Address
Phone
Email
Date Exp
Job Title
Company
Date Edu
Degree
Institute
1. de 200
2. de 200 + en 2000
3. de 2000
4. de 2000 + en 2000

Summary

Conclusion:

- RNN sequence labeling model is better than baseline models
- Aligned embedding allow cross lingual training
- Pave the way for a multilingual model

Outlook

- Improve the alignment of embeddings
- Better evaluation of multilingual embeddings
- Train a multilingual model and test on an unseen language

Thank you?