THE UNIVERSITY OF HONG KONG
DEPARTMENT OF COMPUTER SCIENCE

Named Entity and Rare Word Recognition

Wang Dezhao, Liu Wenjing, Wei Jian
Supervisor: Chan Bethany Mee Yee
Department of Computer Science, The University of Hong Kong

ABSTRACT

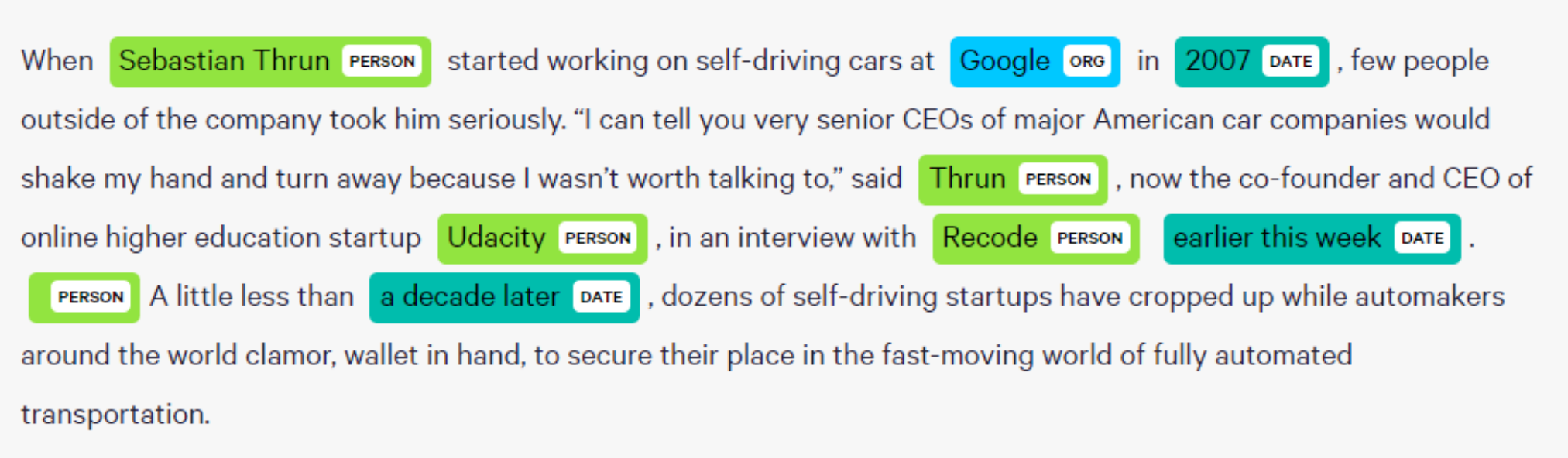
In our research, our work is to labeled the lab data and evaluated several models mentioned in: Bidirectional LSTM-CRF Models for Sequence Tagging) and made some improvement? to the word embedding stage.
Our contributions can be summarized as follows:
1. We systematically compared the performs of several existing models like LSTM-CRF.
2. we add a CNN network and a Bi-LSTM network to the word embedding stage respectively to the word embedding stage
Keywords: name entity, rare word, LSTM, word embedding

CONTACT

[Wang Dezhao]
[u3556210@connect.hku.hk]
[Liu Wenjing]
[u3556179@connect.hku.hk]
[Wei Jian]
[u3556148@connect.hku.hk]

INTRODUCTION

Named-entity recognition (NER) is a subtask of information extraction that seeks to locate and classify named entity mentions in unstructured text into pre-defined categories such as the person names, organizations, locations.
Why NER?
Question Answering
Textual Entailment
Textual Entailment


When Sebastian Thrun PERSON started working on self-driving cars at Google ORG in 2007 DATE , few people outside of the company took him seriously. "I can tell you very senior CEOs of major American car companies would shake my hand and turn away because I wasn't worth talking to," said Thrun PERSON , now the co-founder and CEO of online higher education startup Udacity PERSON , in an interview with Recode PERSON earlier this week DATE .
PERSON A little less than a decade later DATE , dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

Algorithm

In our project, the major process is followed by these steps:

- Using LSTM + CRF model to identify name entities, and divide them as 'PER' (Person), 'ORG' (Organization) and 'LOC' (Location). Others will be labeled as 'O'.
- Because the accuracy of LSTM + CRF model do not perform well, so we did some extend to this model, and we added word embedding layer for the model, here is what we called 'Char-LSTM-LSTM-CRF' and 'Char-Conv-LSTM-CRF'.
- Using word frequency to identity whether the word is rare word or not.

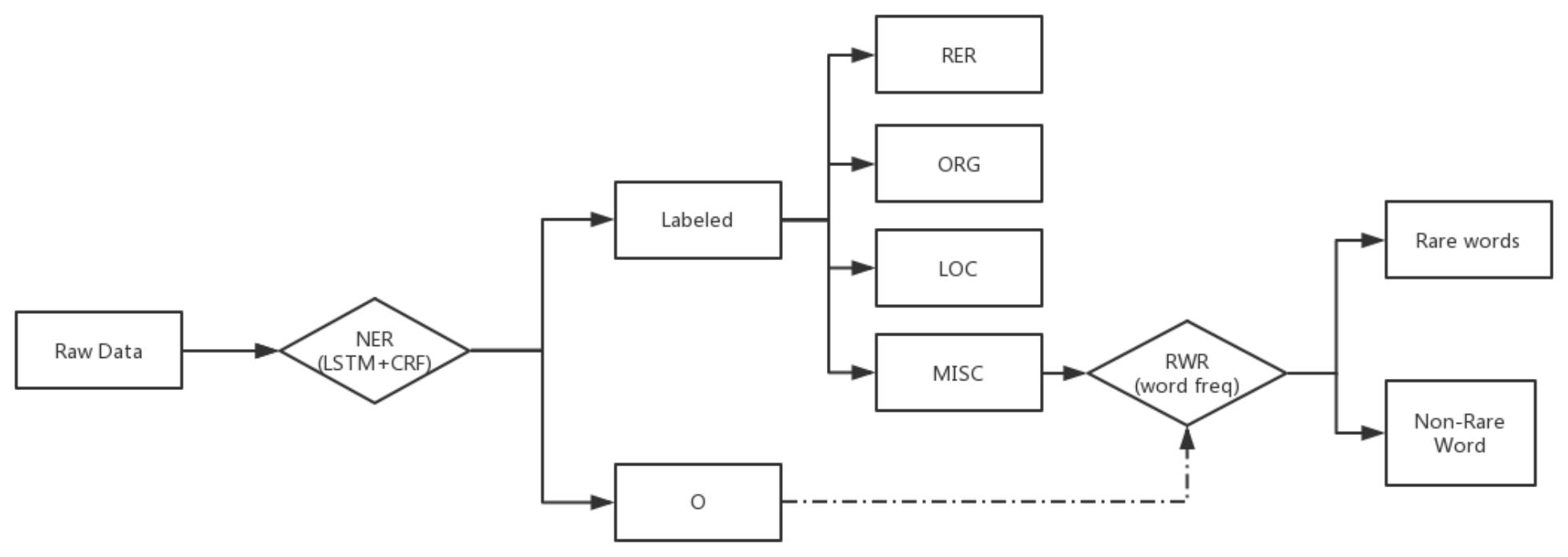


Figure 1. Process of the project

Model Construction

In this section, we will introduce several different models used in the research: word embedding, Bi-LSTM, and CRF.
a) Word embedding
Word embedding is a kind of vector. For each word, we can build or get its n-dimension word embedding.
b) Bidirectional-LSTM network
Using LSTM, we can get the left context of the sequence at every word t, but we may lose the right context of word t.
c) CRF

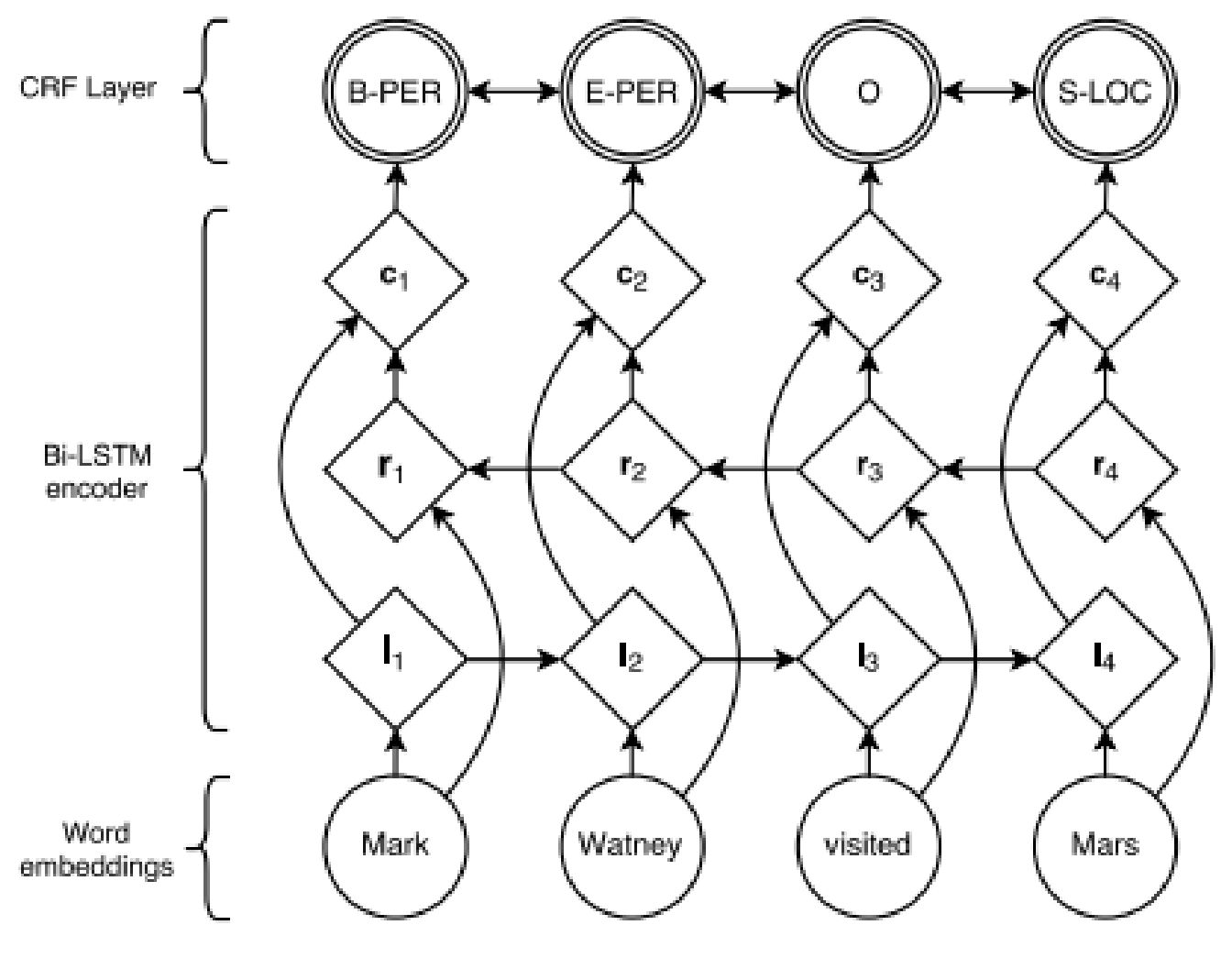


Figure 2. Model

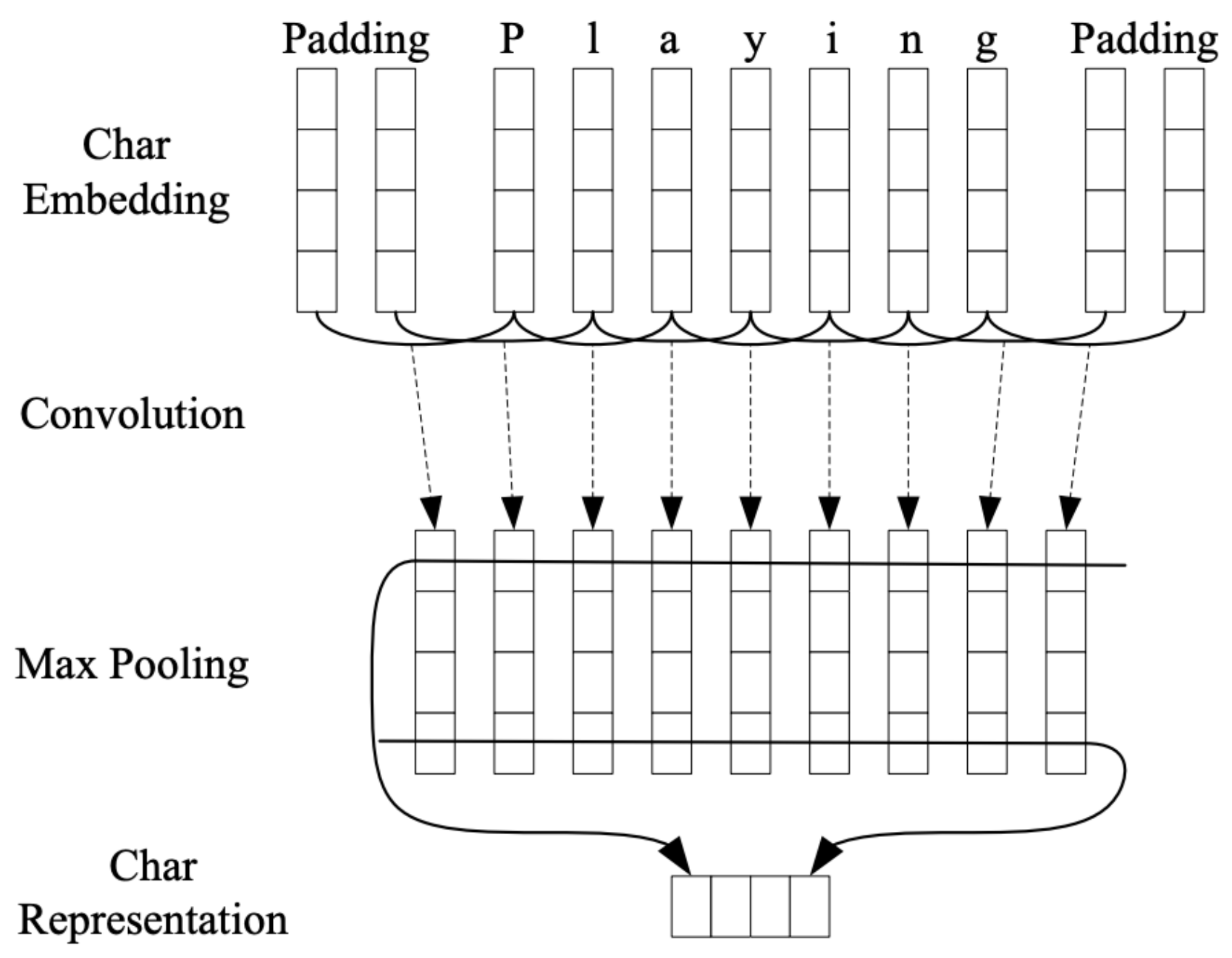


Figure 3. Word Embedding: CNN

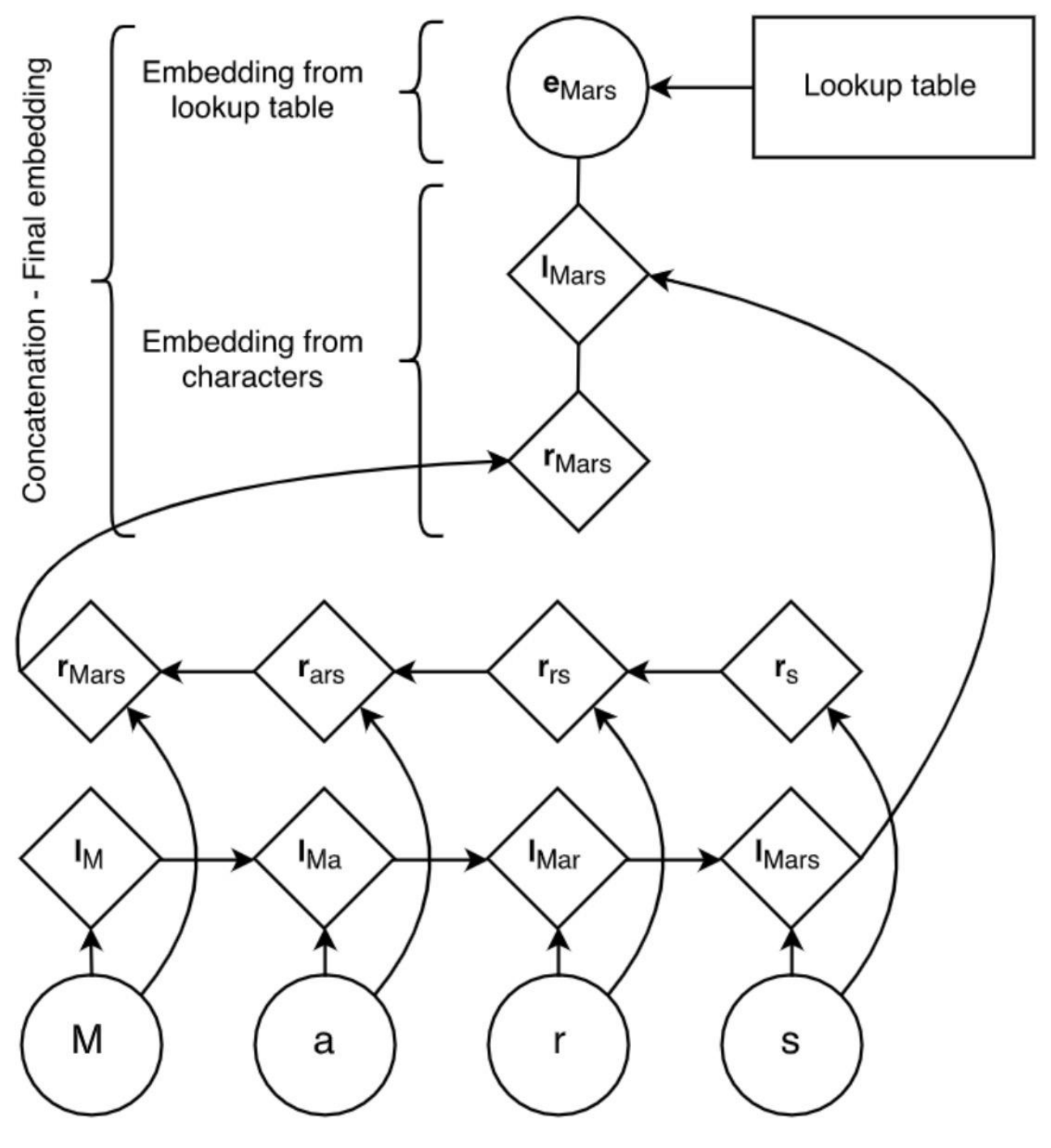


Figure 4. Word Embedding: LSTM

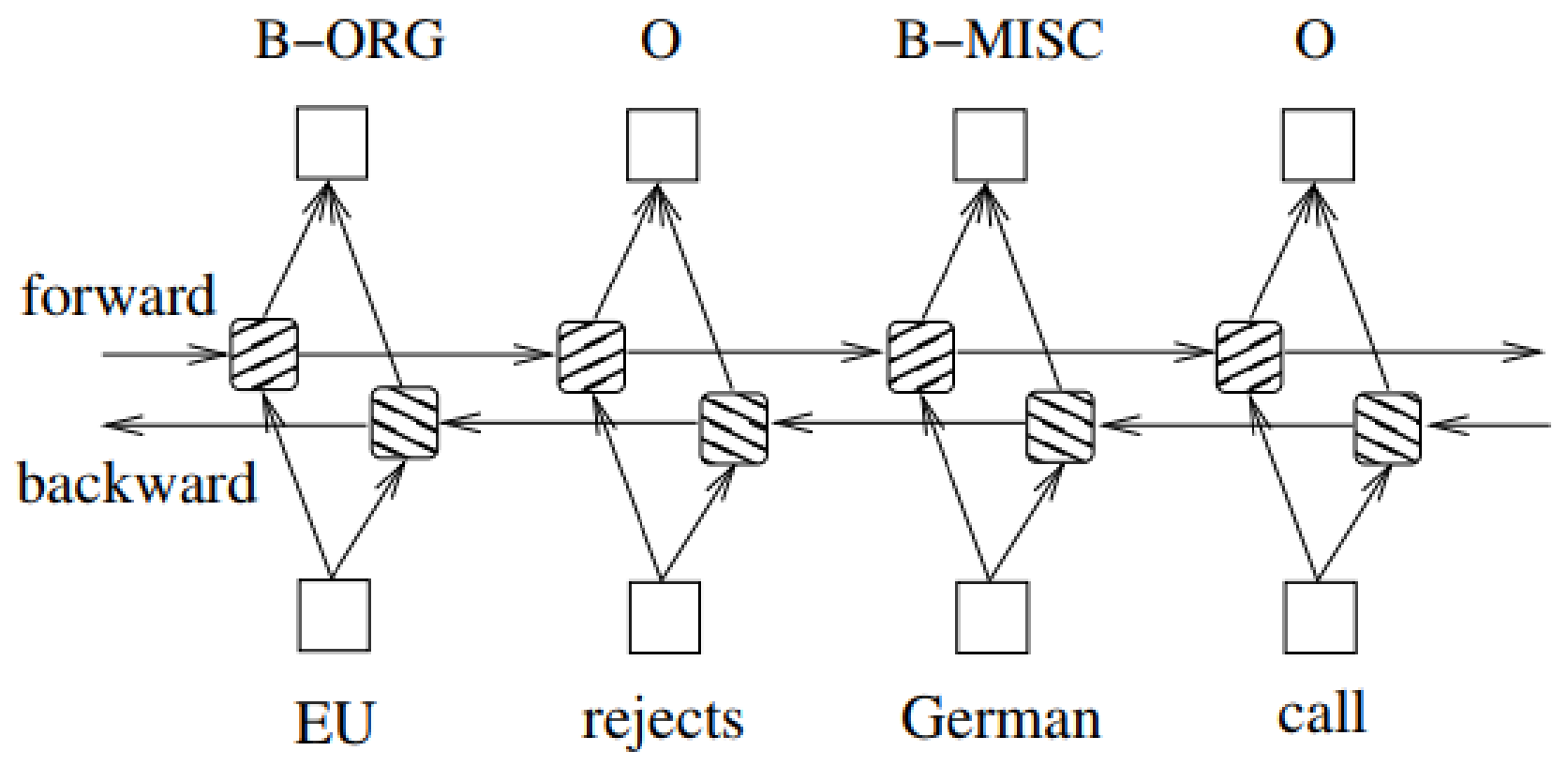
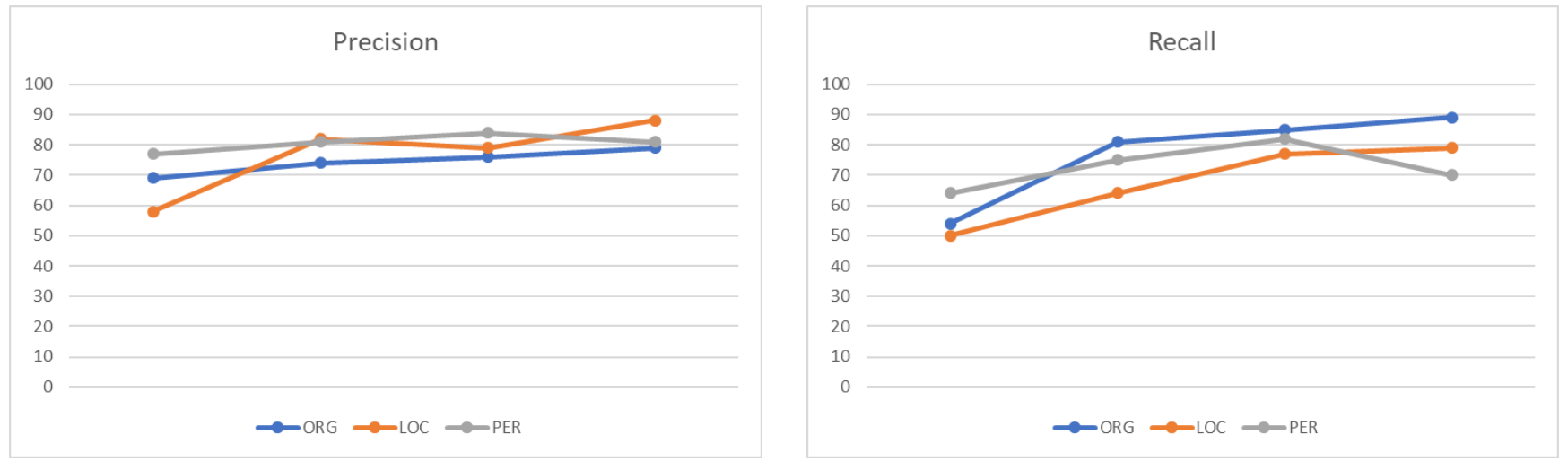


Figure 5. Bi-LSTM

Experiment_NER

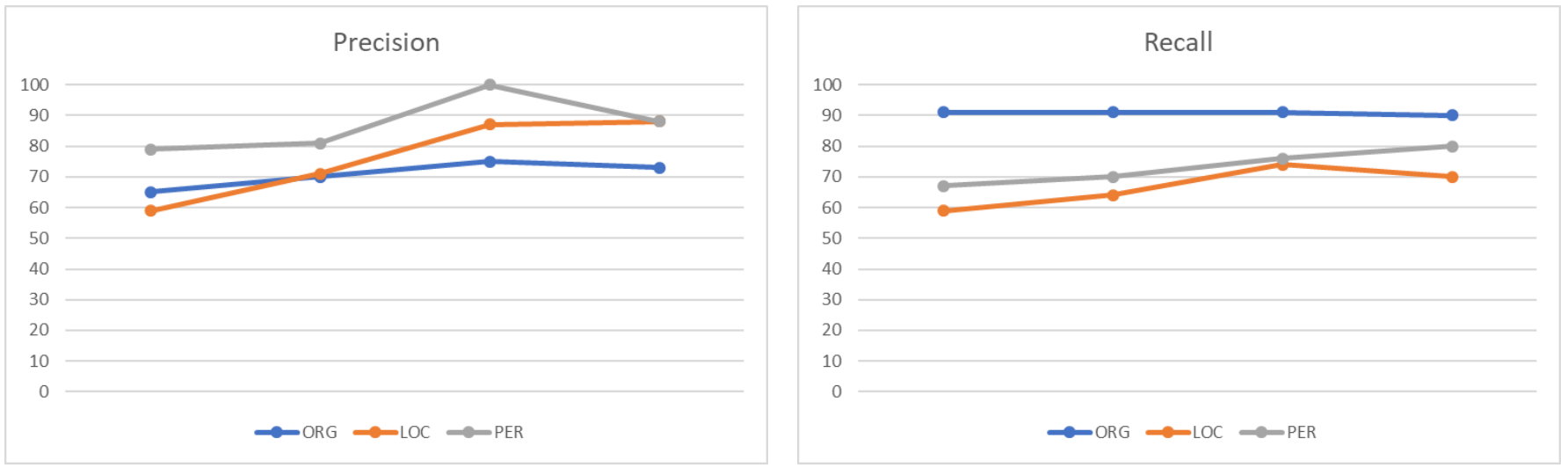
Model 1. LSTM + CRF

Train Dataset		Evaluation	ORG (%)	LOC (%)	PER (%)
Kaggle data(%)	Lab data(%)				
100	0	precision	69	58	77
		recall	54	50	64
90	10	precision	74	82	81
		recall	81	64	75
80	20	precision	76	79	84
		recall	85	77	82
70	30	precision	79	88	81
		recall	89	79	70



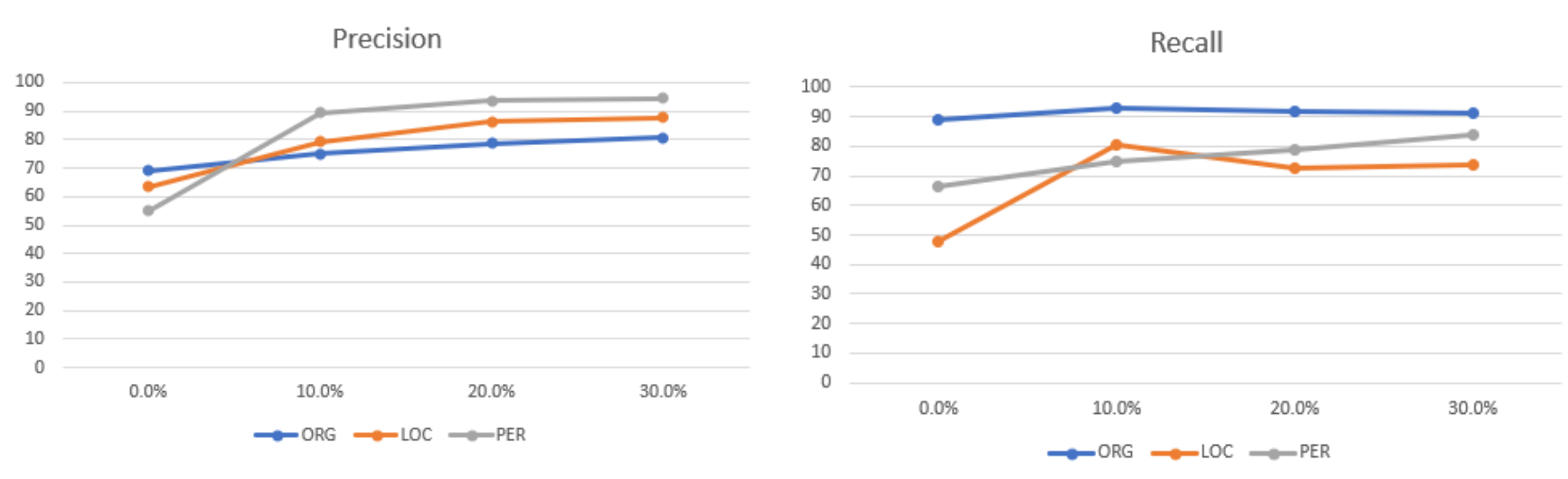
Model 2. Char-LSTM-LSTM-CRF

Train Dataset		Evaluation	ORG (%)	LOC (%)	PER (%)
Kaggle data(%)	Lab data(%)				
100	0	precision	65	59	79
		recall	91	59	67
90	10	precision	70	71	81
		recall	91	64	70
80	20	precision	75	87	100
		recall	91	74	76
70	30	precision	73	88	88
		recall	90	70	80



Model 3. Char-Conv-LSTM-CRF

Train Dataset		Evaluation	ORG (%)	LOC (%)	PER (%)
Kaggle data(%)	Lab data(%)				
100	0	precision	69	63	55
		recall	89	48	66
90	10	precision	75	79	89
		recall	93	80	75
80	20	precision	78	86	93
		recall	92	72	79
70	30	precision	80	87	94
		recall	91	74	83



Experiment_RW

Rare word recognition rules

- Word Frequency < 2.0
- Not digit(e.g., flight number, law's name), not Chinese character
- Length > 8

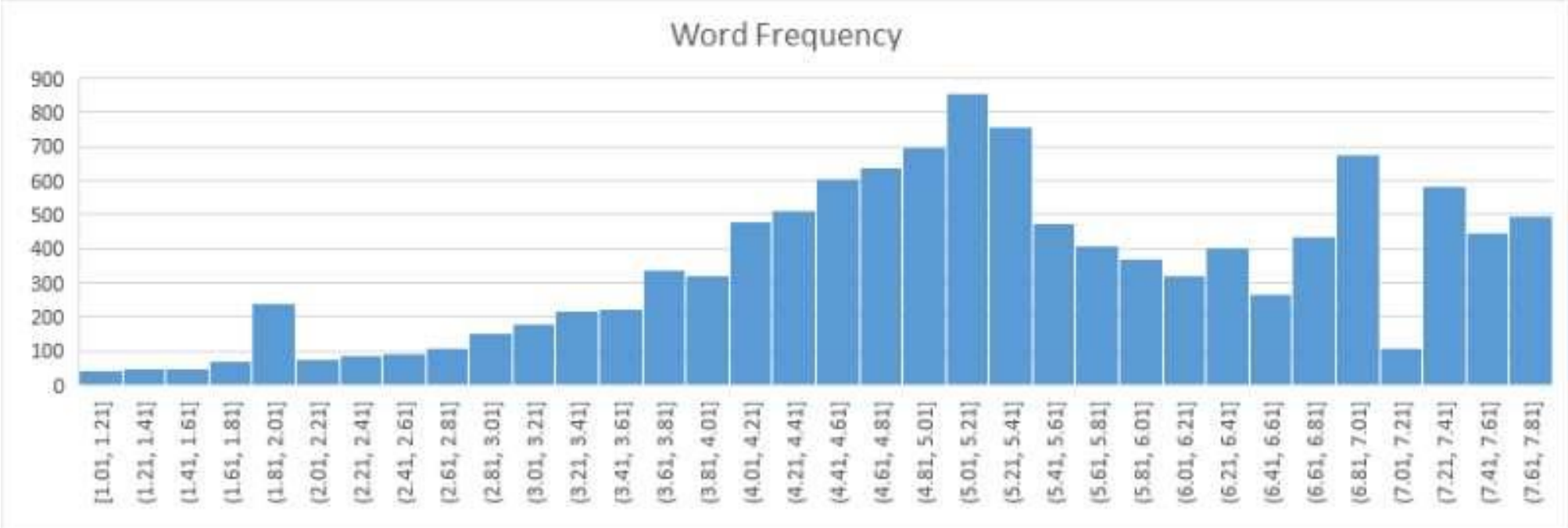


Figure 6. Words distribution

CONCLUSIONS

In our research, we systematically evaluate the performance of different LSTM-CRF models for name entity extraction and sentence tagging. We add a LSTM and a CNN layer respectively to the word embedding stage and use both pre-trained word embedding and the character-based word embedding as the final word embedding, which turns out to make a little contribution to the final accuracy.

REFERENCES

1. HUANG, Zhiheng; XU, Wei; YU, Kai. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991, 2015.
2. SANG, Erik F.; DE MEULDER, Fien. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. arXiv preprint cs/0306050, 2003.
3. CHIU, Jason PC; NICHOLS, Eric. Named entity recognition with bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics, 2016, 4: 357-370.
4. LAMPLE, Guillaume, et al. Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360, 2016.