TBSI-WOLT
The 3rd TBSI Workshop on Learning Theory

Chinese NER for Processing Resume Text

Yongchi Zhu & Qiqi Chen Advisor: Yang Li

TBSI 清华-伯克利深圳学院 Tsinghua-Berkeley Shenzhen Institute

Motivation

➢ Fine-Grained Chinese Resume Processing: Extract semantic information from semi-structured Chinese text

深圳市统计局行政处主任、办公室主任
深圳市统计局行政处科员 ➔

深圳市
└统计局
└行政处
|— 科员
└ 主任
└办公室
└ 主任

➢ Named Entity Recognition (NER) : A word-based sequential tagging task

深圳市L 统计局O 行政处S 主任P										
深	圳	市	统	计	局	行	政	处	主	任
B-LOC	B-ORG	M-LOC	E-LOC	B-ORG	M-ORG	E-ORG	M-SUB	E-SUB	B-POS	E-POS

➢ Challenge: a) Solve the problem of joint appointment separation in the job experience entry
b) Scalability of data processing

Background on NER

➢ Probability graph approaches

	Advantage	Limitations
HMM	Simple and easy to implement	Depend only on each state and corresponding observer
CRF	No independence assumptions and can accommodate arbitrary context information	High training cost and complexity

➢ Neural network approaches
- LSTM: Better to obtain dependencies over long distances but can not encode information from back to front
- Bi-LSTM: Can combine bidirectional semantics

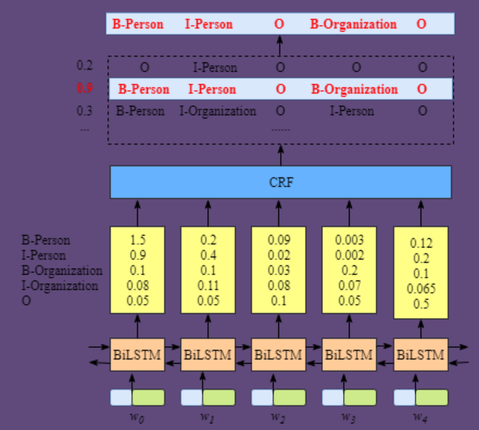
➢ A Bi-LSTM-CRF model
- In addition to the above advantages, Less word vector dependence is needed

Dataset

➢ 7732 manually labeled experience entries
➢ Ratio of training data, validation data and test data: 8:1:1
➢ 12 NER labels

Methodology

➢ BiLSTM-CRF model



➢ CRF Training&Prediction

$$P_{\text{total}} = P_1 + P_2 + \cdots + P_N = e^{S_1} + e^{S_2} + \cdots + e^{S_N}$$

$$\text{LogLossFunction} = -\log \frac{P_{\text{RealPath}}}{P_1 + P_2 + \cdots + P_N}$$

$$\text{Score of Path} = \text{Emission Score} + \text{Transition Score}$$

Emission	START	B-Person
word0	x0,start	x0,b-person
word1	x1,start	x1,b-person
word2	x2,start	x2,b-person

Transition	START	B-Person
START	T start,start	T start,b-person
B-Person	T b-person,start	T b-person,b-person

Improvements

➢ Transition Matrix Initialization

Impossible Transition i->j: C(j) cannot follow C(i)

(for all j
D(i, j) = -10000)

Transition	EO	BP	BS
EO	-10000	0.4	0.1

Unique Transition i->j: only C(j) can follow C(i)

(for all j'!= j
D(i, j') = -10000)

Transition	BO	BS	BP
EL	0.5	-10000	-10000

➢ Preprocessing of Input Data
- Multiple Positions at the Same Time “兼”, “兼任”
- Remove stop words such as adverbs

Results

Methods	Recall	Precision	F1 score
HMM	91.22%	91.49%	91.30%
CRF	95.43%	95.43%	95.42%
BiLSTM	95.32%	95.37%	95.32%
BiLSTM-CRF	96.33%	96.29%	96.29%

	precision	recall	F1-score	support
BL	0.9876	0.9922	0.9899	1531
M0	0.9732	0.9854	0.9793	9401
MP	0.9035	0.8183	0.8588	1152
ML	0.9816	0.9833	0.9825	1140
ES	0.9525	0.9669	0.9597	1452
EP	0.9994	0.9968	0.9981	2064
E0	0.9682	0.9819	0.9750	1551
BS	0.9459	0.9628	0.9543	1452
EL	0.9863	0.9902	0.9883	1531
BP	0.9330	0.9258	0.9294	2064
MS	0.9558	0.9509	0.9534	3892
B0	0.9527	0.9749	0.9637	1551
avg/total	0.9629	0.9633	0.9629	28781

Total Accuracy is above 0.96, while the accuracy of every element of the sentence is above 0.9. The precision of MP is the lowest for the lack of training data (most of the position is 2-character word)

Application

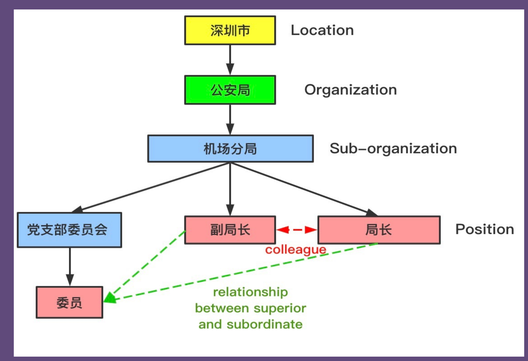
➢ Joint Appointment Separation

Example 1: L O P1 P2 ➡ L O P1+L O P2
深圳市 A公司 董事 、副总经理
(Shenzhen Company-A Director & Vice General Manager)
➡ 深圳市 A公司 董事 & 深圳市 A公司 副总经理

Example 2: L O P1 S P2 ➡ L O P1+L O S P2
深圳市 A公司 董事 兼 财务部 经理
(Shenzhen Company-A Director & Finance Department Manager) ➡
深圳市 A公司 董事 & 深圳市 A公司 财务部 经理

Example 3: L O S1 P1 S2 P2 ➡ L O S1 P1+L O S2 P2
深圳市 A公司 财务部 经理 兼 项目部 总监
(Shenzhen Company-A Finance Department Manager & Project Department Director) ➡
深圳市 A公司 财务部 经理 & 深圳市 A公司 项目部 总监

➢ Organizational affiliation



➢ Future works
- Acronyms & Missing entities
- Rare location recognition
- Complicated position&organization recognition

Reference

[1] Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.
[2] Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
[3] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. Neural Computation, 9(8):1735-1780.
[4] Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.