Baidu百度

第七届
百度奖学金
BAIDU SCHOLARSHIPS
颁奖典礼
暨AI未来说博士生专场

当 NLP 邂逅 Social Media

构建计算机与网络语言的桥梁

汇报人：桂韬

导师：张奇、黄萱菁教授

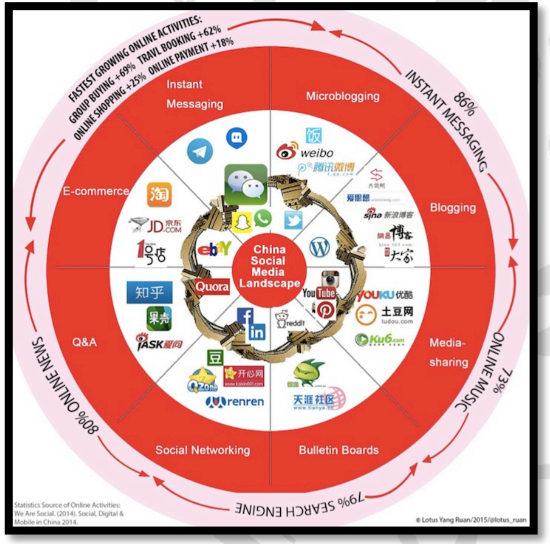
FUDAN UNIVERSITY 1905

目录

社交媒体

自发传播

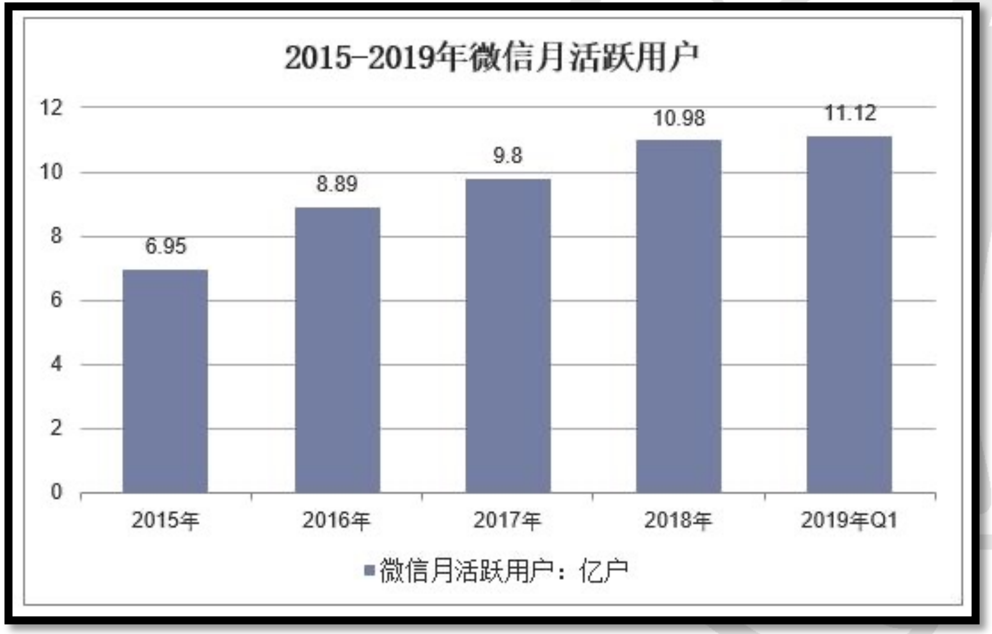
“社会化”属性



表现形式多样

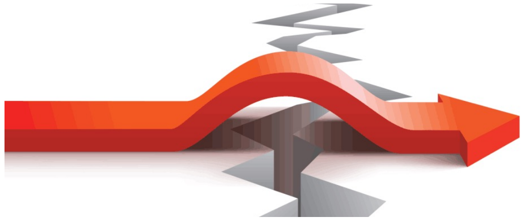
网络语言

近年来，以微博、微信以及社交网站等为代表的社会媒体在我国发展迅速。据2018年《微信数据报告》显示，微信月活跃用户突破十亿，每天产生450亿条消息。随着网络的不断普及，人们越来越多的交流也通过网络实现，也因此诞生一种网络上的自然交际语言。



非规范性

明年他要 C 位出道
这是神马规矩
I 服了 U！
皮一下，很开心



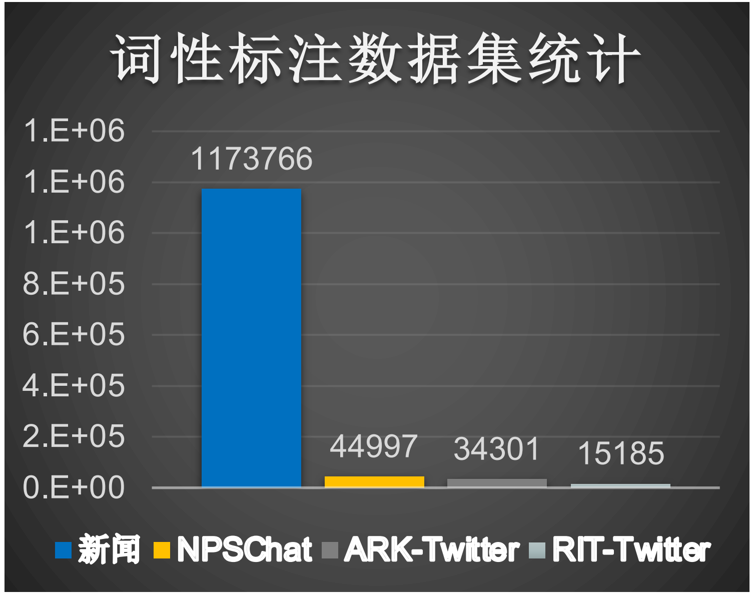
热点追踪

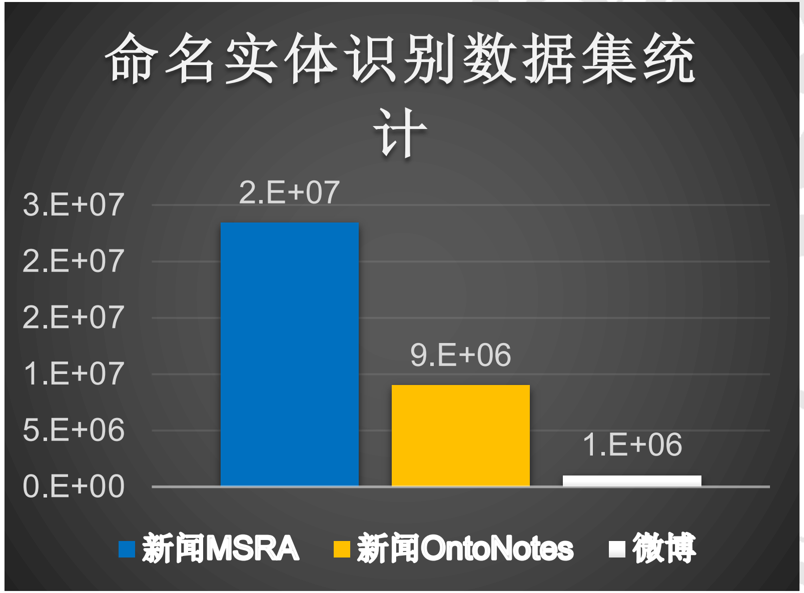
网络语言困境

网络语言非规范化问题研究

网络语言困境

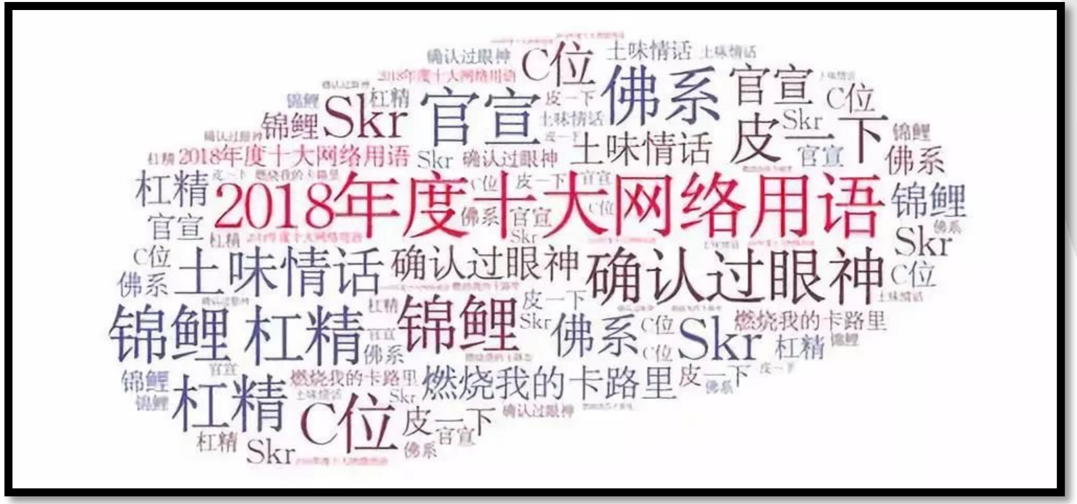
■标注数据少　■旧词新意、另造新词　■语法、语用不规范





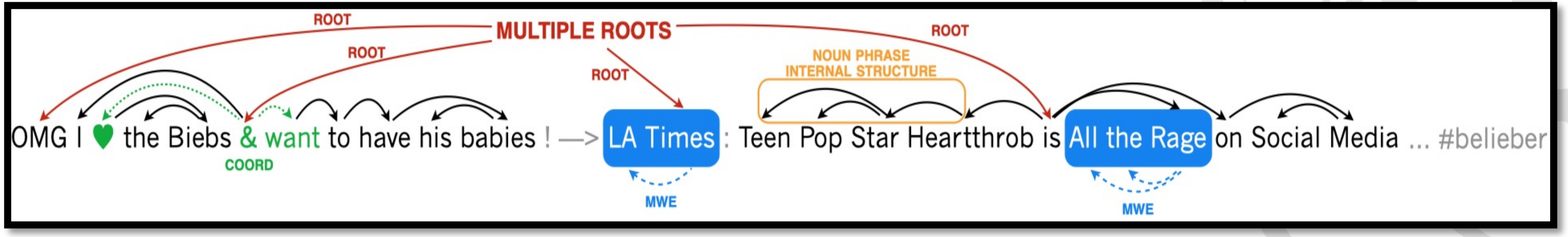
网络语言困境

■标注数据少　■旧词新意、另造新词　■语法、语用不规范



网络语言困境

■标注数据少 ■旧词新意、另造新词 ■语法、语用不规范



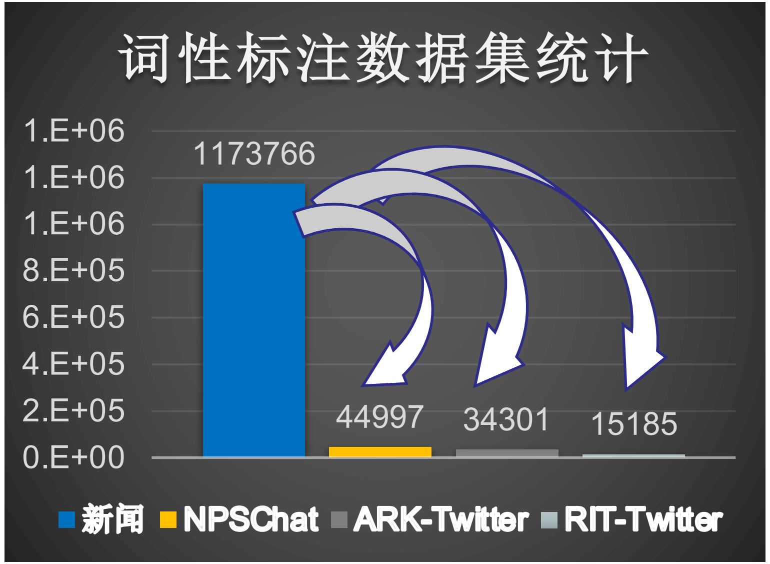
推特句法分析树

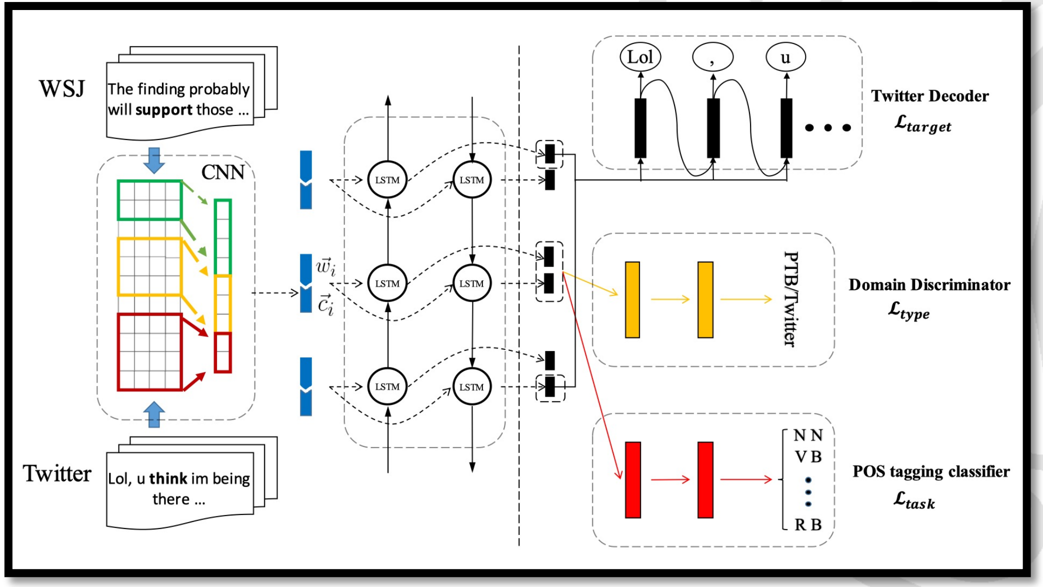
网络语言脱困

迁移学习　外部知识　全局语义　动态建模

网络语言脱困

■ 标注数据少 ⟶ 利用新闻语料、无标注语料





Gui, Tao, et al. "Part-of-speech tagging for twitter with adversarial neural networks." EMNLP 2017.

网络语言脱困

■ 标注数据少 ➡ 利用新闻语料、无标注语料

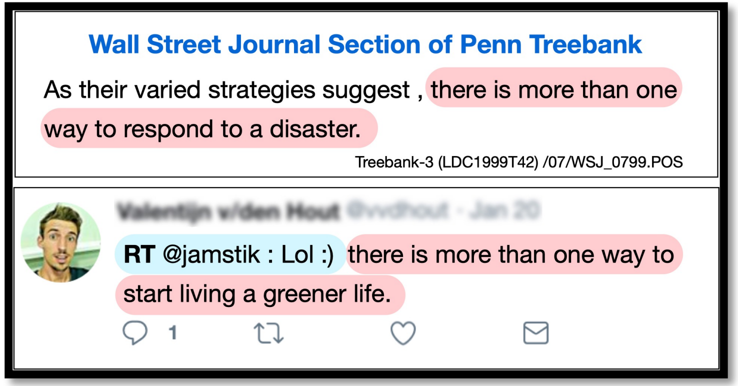
Methods	RIT-Test	RIT-Dev
Stanford-WSJ (Toutanova et al., 2003)	73.37%	83.29%
Stanford-MIX	83.14%	84.19%
T-POS (Ritter et al., 2011)	84.55%	84.83%
GATE Tagger (Derczynski et al., 2013)	88.69%	89.37%
ARK Tagger (Owoputi et al., 2013)	90.40%	-
bi-LSTM (word level)	75.91%	76.94%
bi-LSTM (word level pretrain)	85.99%	86.93%
bi-LSTM (character level)	82.85%	84.30%
bi-LSTM (combine)	89.48%	89.30%
bi-LSTM (combine + WSJ)	83.54%	83.64%
bi-LSTM (combine + WSJ + adversarial)	83.76%	84.45%
bi-LSTM (combine + WSJ + fine-tune)	89.87%	90.23%
bi-LSTM (combine + WSJ + adversarial + fine-tune)	90.60%	90.73%
TPANN (combine + WSJ + adversarial + fine-tune + autoencoder)	**90.92%**	**91.08%**

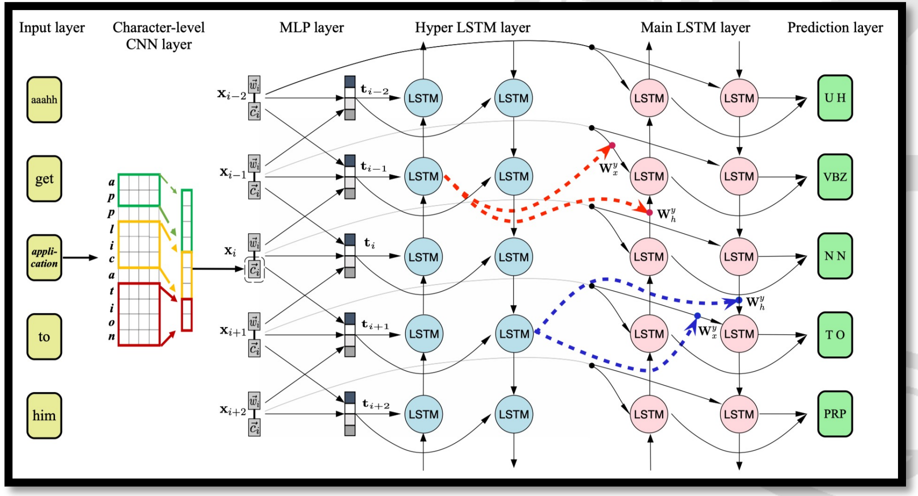
Gui, Tao, et al. "Part-of-speech tagging for twitter with adversarial neural networks." EMNLP 2017.

网络语言脱困

■ 标注数据少 ➡ 利用新闻语料、无标注语料

+ 保留网络语言特性





Gui, Tao, et al. "Transferring from Formal Newswire Domain with Hypernet for Twitter POS Tagging." EMNLP 2018.

网络语言脱困

- 标注数据少 → 利用新闻语料、无标注语料 + 保留网络语言特性

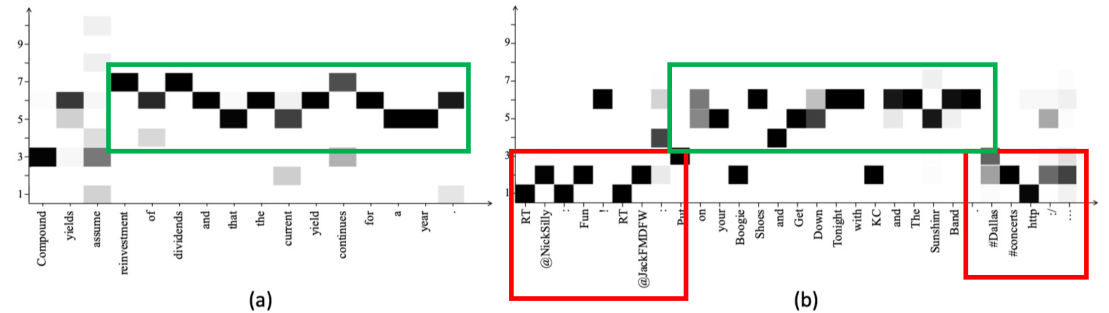
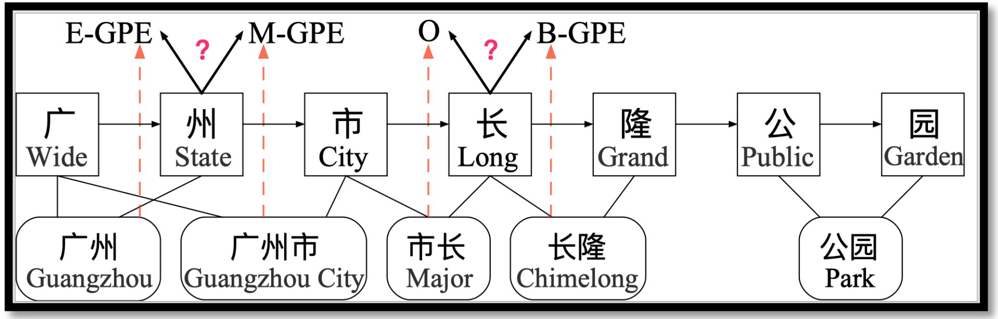


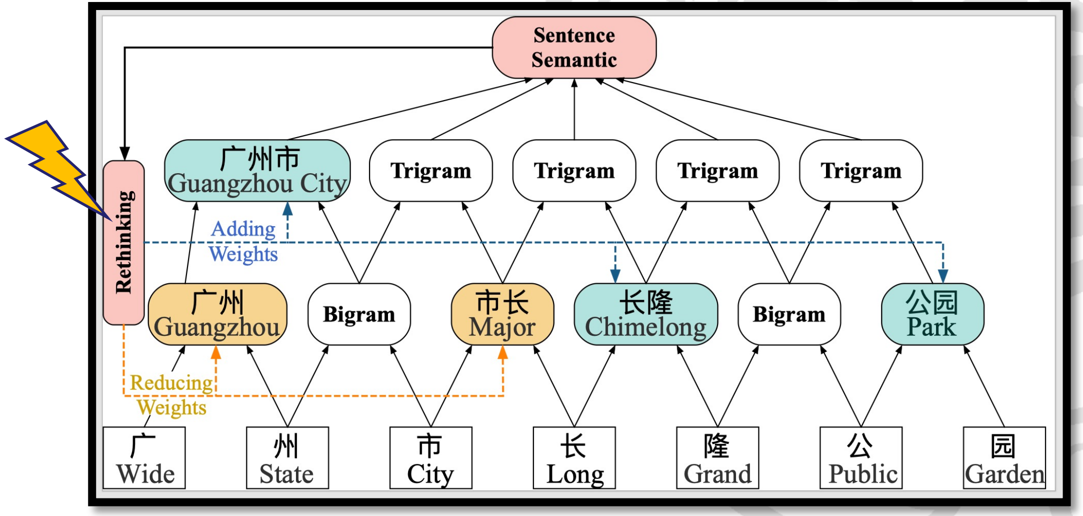
Figure 3: Visualization of context distribution. The left sentence comes from WSJ, and the right one comes from Twitter.

Gui, Tao, et al. "Transferring from Formal Newswire Domain with Hypernet for Twitter POS Tagging." EMNLP 2018.

网络语言脱困

■ 旧词新意、另造新词 ⟶ 外部知识 + 反思机制

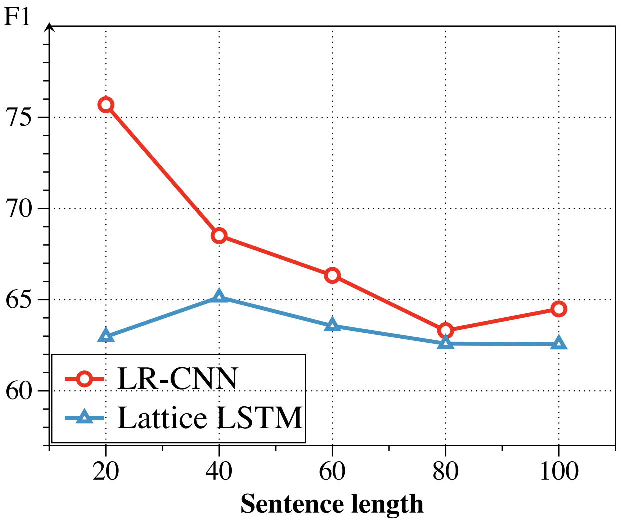




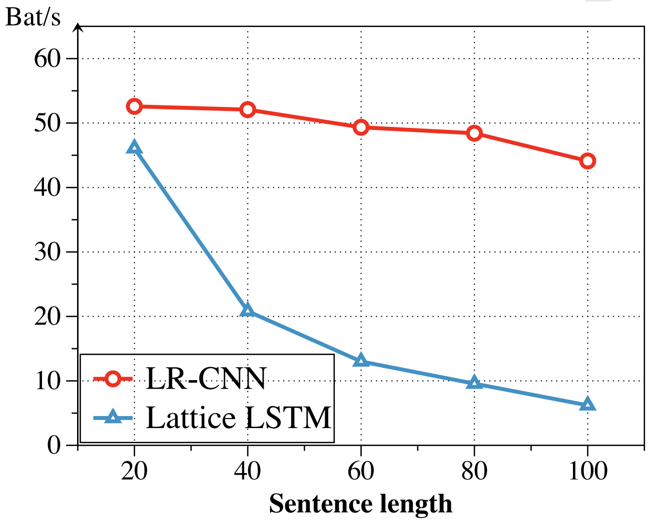
Gui, Tao, et al. "Cnn-based chinese ner with lexicon rethinking." *IJCAI* 2019.

网络语言脱困

■ 旧词新意、另造新词 → 外部知识 + 反思机制



F1 值



速度

Gui, Tao, et al. "Cnn-based chinese ner with lexicon rethinking." *IJCAI* 2019.

网络语言脱困

■ 旧词新意、另造新词 ⟶ 外部知识 + 全局语义

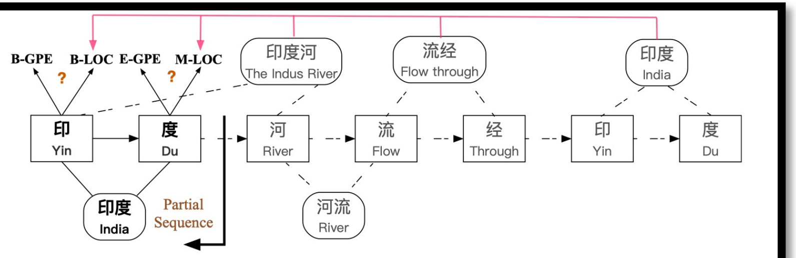
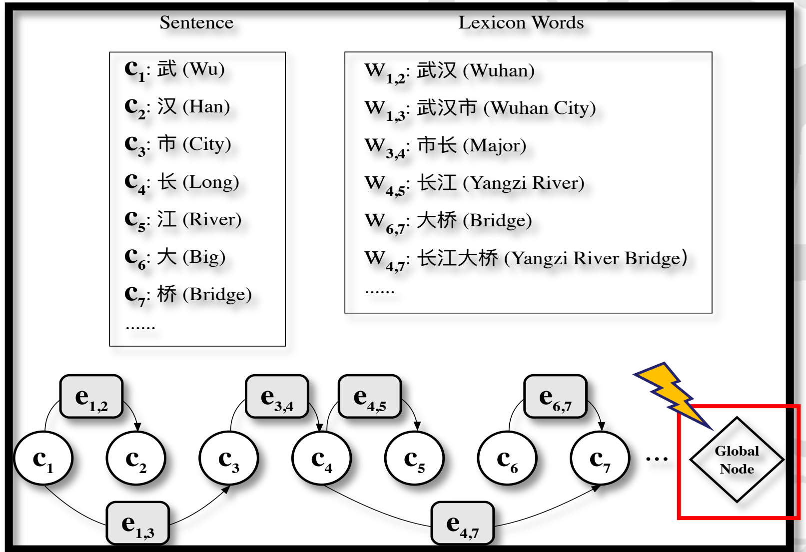


Figure 1: Example of word character lattice with partial input. Because of the characteristic of chain structure, RNN-based methods must predict the label “度” using only previous partial sequences “印度 (India)”, which may suffer from word ambiguities without global sentence semantics.



Gui, Tao, et al. “A Lexicon-Based Graph Neural Network for Chinese NER.” EMNLP-IJCNLP 2019.

网络语言脱困

■ 旧词新意、另造新词 ⟶ 外部知识 + 全局语义

Input	Models	P	R	F1
Gold seg.	Yang et al. (2016)	65.59	71.84	68.57
	Yang et al. (2016)*†	72.98	**80.15**	**76.40**
	Che et al. (2013)*	77.71	72.51	75.02
	Wang et al. (2013)*	76.43	72.32	74.32
	Word-level LSTM	76.66	63.60	69.52
	+char+bichar	**78.62**	73.13	75.77
	Word-level CNN	66.84	62.99	64.86
	+char+bichar	68.22	72.37	70.24
Auto seg.	Word-level LSTM	72.84	59.72	65.63
	+char+bichar	73.36	70.12	71.70
	Word-level CNN	54.62	55.20	54.91
	+char+bichar	64.69	65.09	64.89
No seg.	Char-level LSTM	68.79	60.35	64.30
	+bichar+softword	74.36	69.43	71.89
	Char-level CNN	56.78	60.99	58.81
	+bichar+softword	59.60	65.14	62.25
	Lattice LSTM	**76.35**	71.56	73.88
	LGN	76.13	**73.68**	**74.89**

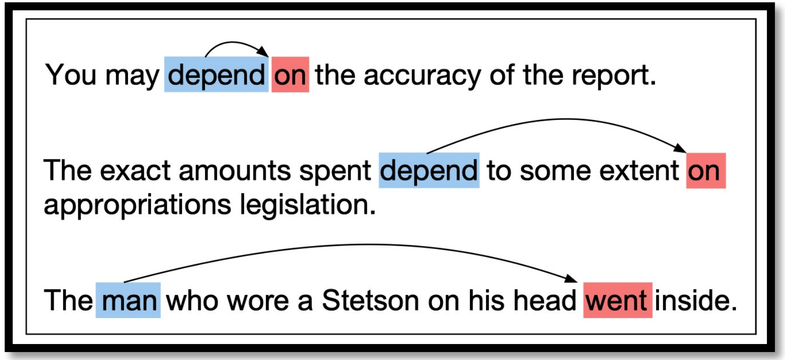
Table 2: Main results on OntoNotes.

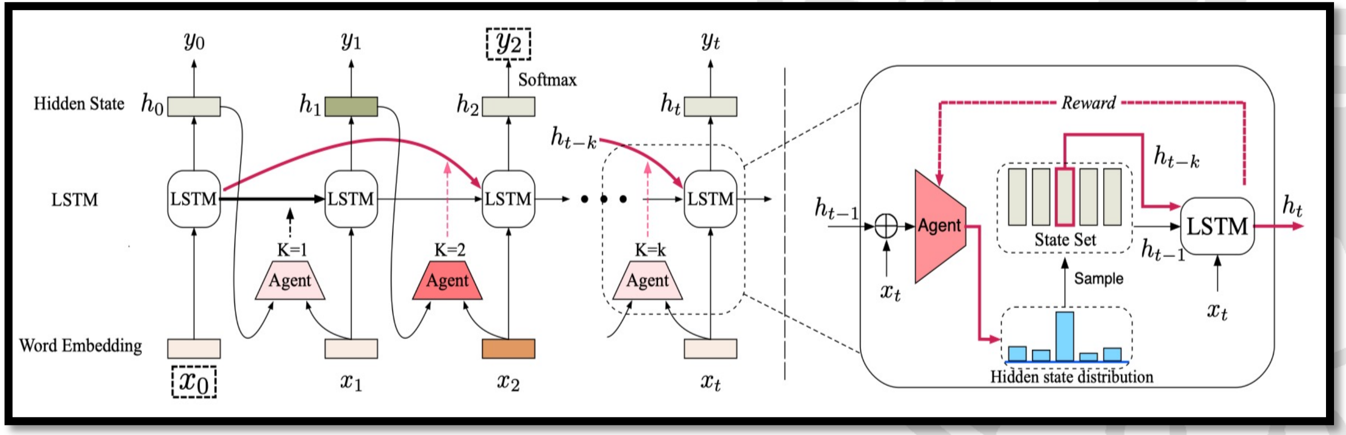
Sentence	印度河流经巴基斯坦 The Indus River flows through Pakistan.
Gold seg	印度河 流经 巴基斯坦 The Indus River, flow through, Pakistan
Lexicon words	印度 河流 印度河 流经 巴基斯坦 India, river, The Indus River, flow through, Pakistan
Lattice LSTM	B E (GPE) O O O B M M E (GPE) 印度 (GPE) 河流经 巴基斯坦 (GPE) India (GPE) ... Pakistan (GPE).
LGN -global	B E (GPE) O O O B M M E (GPE) 印度 (GPE) 河流经 巴基斯坦 (GPE) India (GPE) ... Pakistan (GPE).
LGN (one step)	B M E (GPE) O O B M M E (GPE) 印度河 (GPE) 流经 巴基斯坦 (GPE) The Indus River (GPE) flows through Pakistan (GPE).
LGN	B M E (LOC) O O B M M E (GPE) 印度河 (LOC) 流经 巴基斯坦 (GPE) The Indus River (LOC) flows through Pakistan (GPE).

Gui, Tao, et al. "A Lexicon-Based Graph Neural Network for Chinese NER." EMNLP-IJCNLP 2019.

网络语言脱困

- 语法、语用不规范 ⟶ 动态建模依赖关系

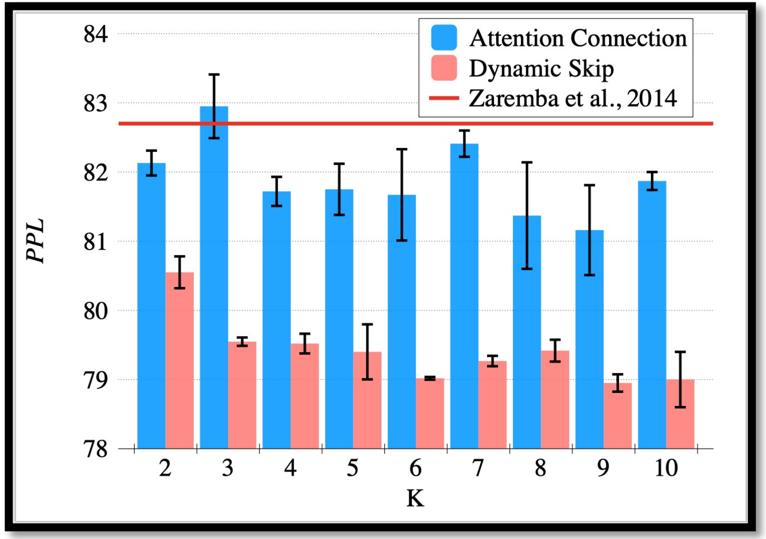


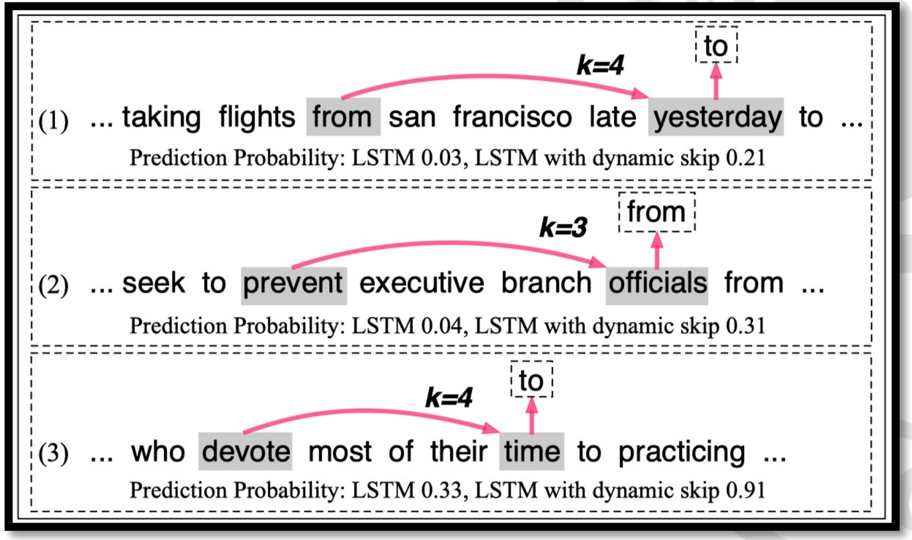


Gui, Tao, et al. "Long Short-Term Memory with Dynamic Skip Connections." AAAI 2019.

网络语言脱困

- 语法、语用不规范 ➡ 动态建模依赖关系





Gui, Tao, et al. "Long Short-Term Memory with Dynamic Skip Connections." AAAI 2019.

网络语言价值

心理疾病早期发现　用户行为预测

网络语言价值

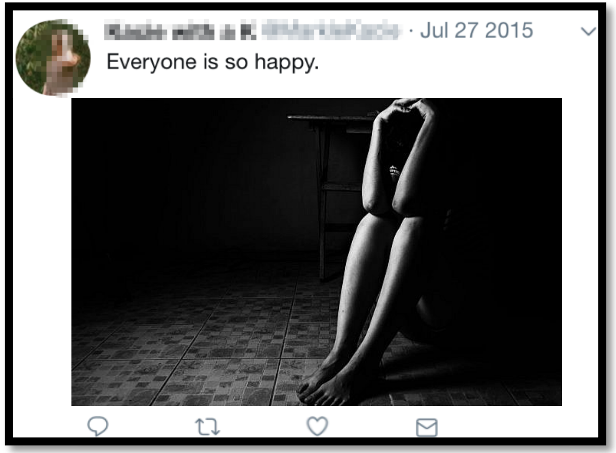
Stock Prediction

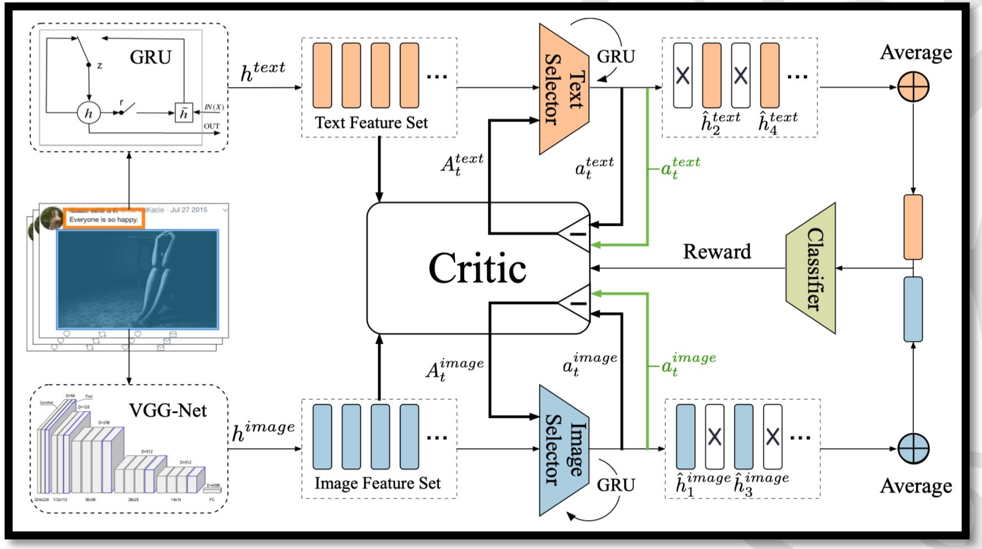
Public Health Analysis

Real-time Event Detection

网络语言价值

- 多模态网络语言 → 早期抑郁症发现





Gui, Tao, et al. “Cooperative Multimodal Approach to Depression Detection in Twitter.” AAAI 2019.

网络语言价值

■ 多模态网络语言 ➡ 早期抑郁症发现

Dataset	Top words (by frequency)
Selected data of depressed users	bad, cancer, insurance, hate, medical, pain, cost, mental, ...
Unselected data of depressed users	people, online, time, know, life, free, school, weight, work, ...
Original data of non-depressed users	wow, idk, like, party, gotta, funny, ◡̈, honestly, team, :) ...

Table 3: Example words arranged in descending order of word frequency.

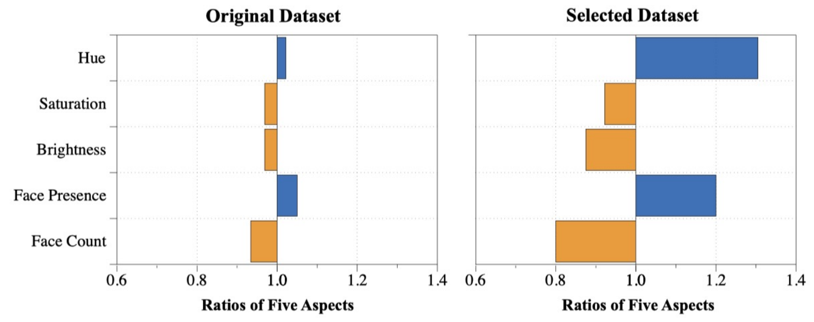
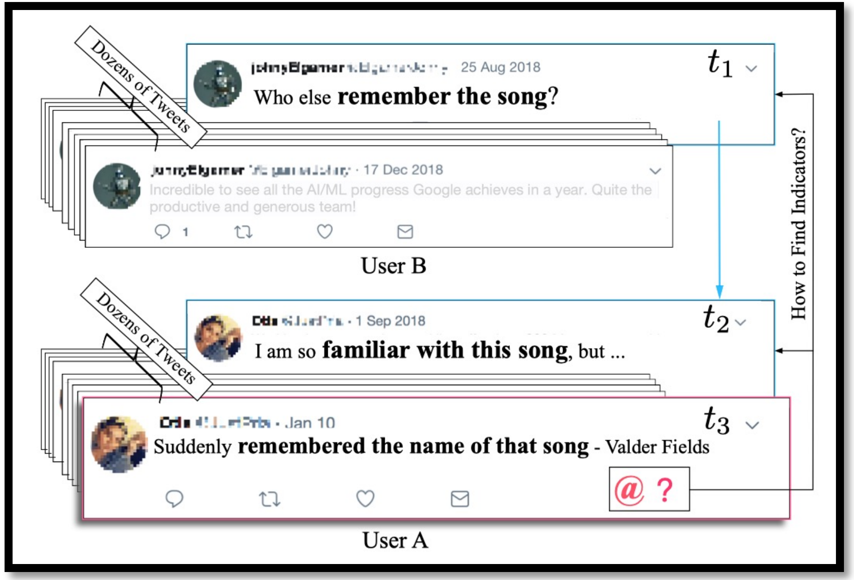


Figure 5: Comparison of original and selected posts. The y-axis values show the five aspects of each image, and the x-axis values are the ratios of these five aspect values of depressed users to those of non-depressed users.

Gui, Tao, et al. “Cooperative Multimodal Approach to Depression Detection in Twitter.” AAAI 2019.

网络语言价值

■ 网络语言交互 ➡ 用户行为预测



Gui, Tao, et al. "Mention Recommendation in Twitter with Cooperative Multi-Agent Reinforcement Learning." ACM SIGIR, 2019.

网络语言价值

■ 网络语言交互 ➡ 用户行为预测

	Method	Precision	Recall	F-Score	MRR	Hits@3	Hits@5
I	**NB**, Pedregosa et al. 2011 [24]	51.42	50.37	50.89	63.09	67.09	78.73
	PMPR, Li et al. 2011 [24]	58.10	57.39	57.74	69.85	73.42	86.36
	CAR, Tang et al. 2015 [29]	59.74	58.62	59.17	70.57	74.68	87.34
II	**LSTM**, Hochreiter and Schmidhuber 1997 [9]	65.54	64.60	65.07	74.53	78.48	90.31
	CAN, Lu et al. 2016 [16]	63.29	62.66	62.97	71.38	76.52	90.58
	MLAN, Yu et al. 2017 [37]	60.16	59.53	59.84	71.37	77.22	91.14
	DAN, Nam et al. 2017 [23]	73.42	72.78	73.10	80.94	82.28	91.37
	MAN, Moon et al. 2018 [21]	68.35	67.72	68.03	75.18	77.22	88.61
	AU-HMNN, Huang et al. 2017 [10]	74.23	73.05	73.64	81.16	83.54	92.41
III	**Random Sampling**	70.94	69.72	70.32	77.70	82.88	93.67
	IQL, Tampuu et al. 2017 [28]	71.04	70.26	70.65	79.01	82.13	92.16
	CROMA	**74.55**	**74.09**	**74.32**	**81.85**	**86.36**	**95.00**

Table 3: Comparison of different methods between adding CROMA RL and without CROMA RL for F1, Hit@3, and Hit@5 scores. Results annotated with * are obtained when the number of historical tweets per user is restricted to five, others are trained with all 50 historical tweets.

Method	F1		Hit@3		Hit@5		
	w/o RL	w/ RL	w/o RL	w/ RL	w/o RL	w/o RL *	w/ RL
LSTM	65.07	**+0.87**	78.48	**+1.33**	90.31	-0.44	**+0.95**
CAN	62.97	**+1.23**	76.52	**+1.83**	90.58	-1.85	**+0.39**
MLAN	59.84	**+1.05**	77.22	**+1.62**	91.14	-0.81	**+1.17**
DAN	73.10	**+0.71**	82.28	**+0.86**	91.37	+1.04	**+1.20**
MAN	68.03	**+0.94**	77.22	**+1.91**	88.61	-6.33	**+1.81**
AU-HMNN	73.64	**+0.68**	83.54	**+2.82**	92.41	+0.00	**+2.59**

Gui, Tao, et al. “Mention Recommendation in Twitter with Cooperative Multi-Agent Reinforcement Learning.” ACM SIGIR, 2019.

THANK YOU

学术主页

代码地址