Microblog Hashtag Generation via Encoding Conversation Contexts

Yue Wang[1], Jing Li[2], Irwin King[1], Michael R. Lyu[1], Shuming Shi[2]

[1] *The Chinese University of Hong Kong*

[2] *Tencent AI Lab*

NAACL-HLT 2019

Tencent AI Lab

Outline

- Background
- The Framework of Our Model
- Experiments
- Conclusions

Outline

- **Background**
- The Framework of Our Model
- Experiments
- Conclusions

Background


NAACL HLT
@NAACLHLT
Following

#nlproc Twitter! Help communicate #naacl2019 as it happens by becoming an official livetweeter! You'll even get mentioned on the program! Signup form here:

- Hashtags can reflect **keyphrases** or **topics**

Microblog Search

Text Summarization

Sentiment Analysis

Background

- **Large volume**: 500 million tweets per day!

- Only less than **15%** tweets contain at least one hashtag.

- There is **a pressing need** for automatic hashtag annotation!

Challenge

Why automatic hashtag annotation is **challenging**?

- **Data sparsity**
  - ✓Informal style
  - ✓Short in length
  - ✓Syntax errors

Example

"This Azarenka woman needs a talking to from the umpire her weird noises are totes inappropes professionally."

Intuition

Example

"This Azarenka woman needs a talking to from the umpire her weird noises are totes inappropes professionally."

[T1] How annoying is she. I just worked out what she sounds like one of those turbo charged cars when they change gear or speed.

[T2] On the topic of noises, I was at the ***Nadal-Tomic*** game last night and I loved how quiet ***Tomic*** was compared to ***Nadal***.

[T3] He seems to have a shitload of talent and the ***postmatch*** press conf. He showed a lot of maturity and he seems nice.

[T4] ***Tomic*** has a fantastic ***tennis*** brain...

- From the **user conversation**, we can imply the hashtag: ***#AusOpen***

Related Work

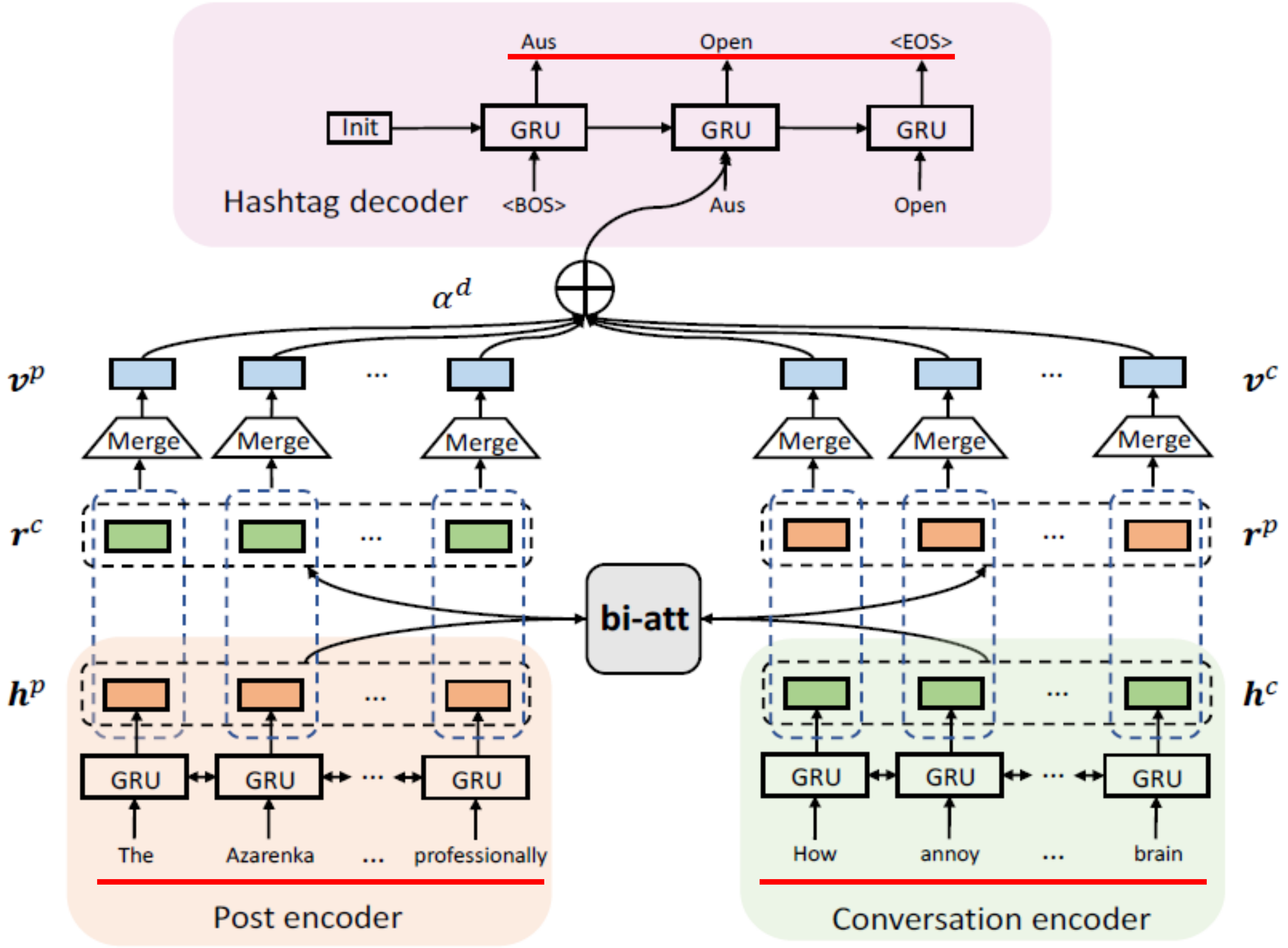
- **Microblog hashtag annotation**
  - Extraction (Zhang et al. 2016 EMNLP, 2018 NAACL-HLT)
    - ☐ Cannot produce hashtags absent in the post
  - Classification (Gong and Zhang, 2016 IJCAI)
    - ☐ Cannot produce hashtags absent in the predefined list
  - Topic models (Gong et al., 2015 EMNLP)
    - ☐ Cannot generate phrase-level hashtags
- **Neural language generation**
  - Encoder-Decoder framework (Sutskever et al., 2014 NeurIPS)
  - Keyphrase generation (Meng et al., 2017 ACL)
    - Performance compromised due to the sparsity of social media language

Keyphrase Generation + Conversation Contexts = Our Model

Outline

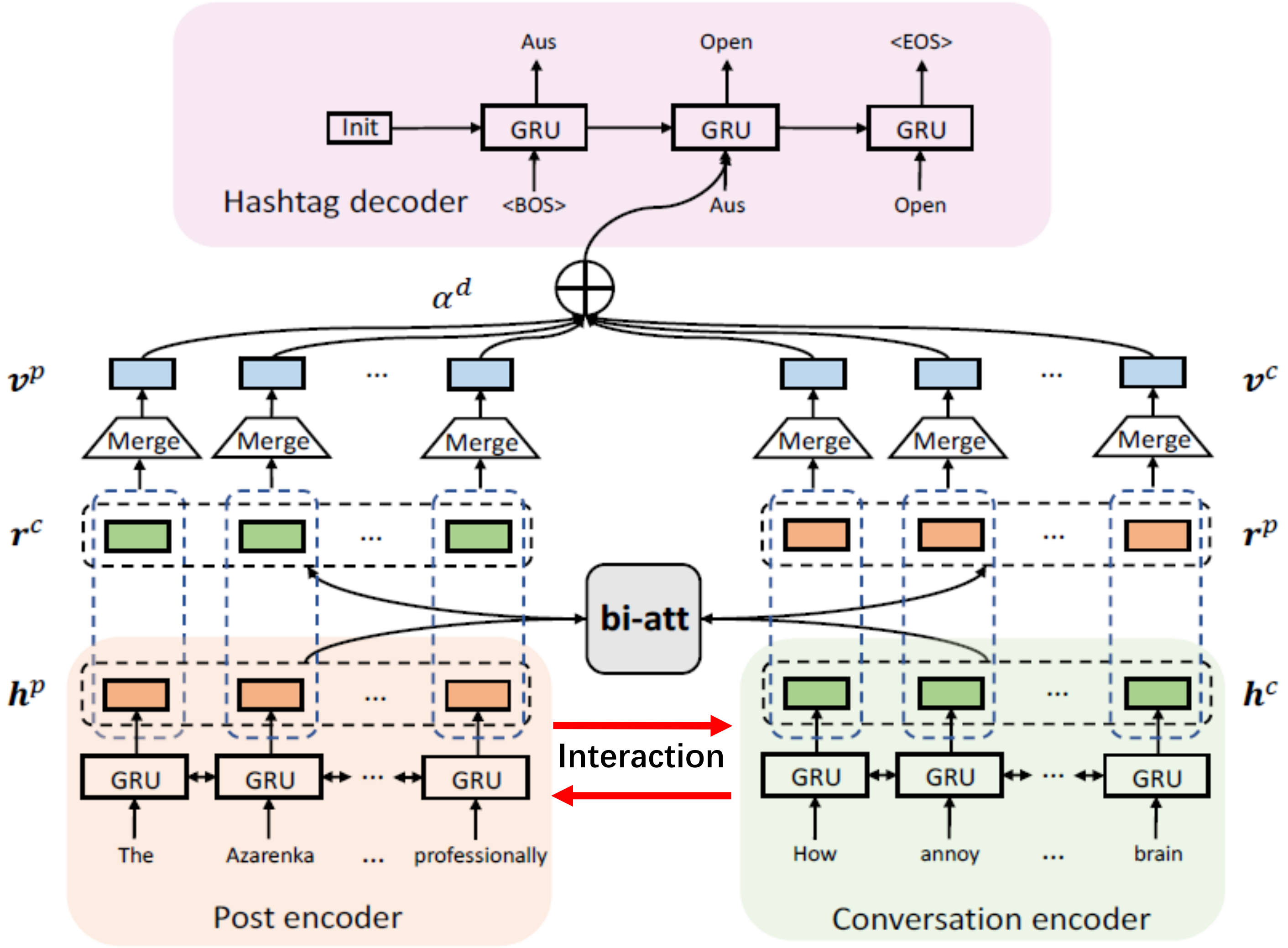
- Background
- **The Framework of Our Model**
- Experiments
- Conclusions

The Framework of Our Model



- **Input**
  - Target post: $\langle x_1^p, x_2^p, \dots, x_{|x^p|}^p \rangle$
  - Conversation: $\langle x_1^c, x_2^c, \dots, x_{|x^c|}^c \rangle$
- **Output**
  - Hashtag: $\langle y_1, y_2, \dots, y_{|y|} \rangle$
  - "**A**us**O**pen" → "**A**us **O**pen"

The Framework of Our Model



Post encoder

- $\boldsymbol{h}^p = \mathrm{BiGRU}(\boldsymbol{x}^p)$

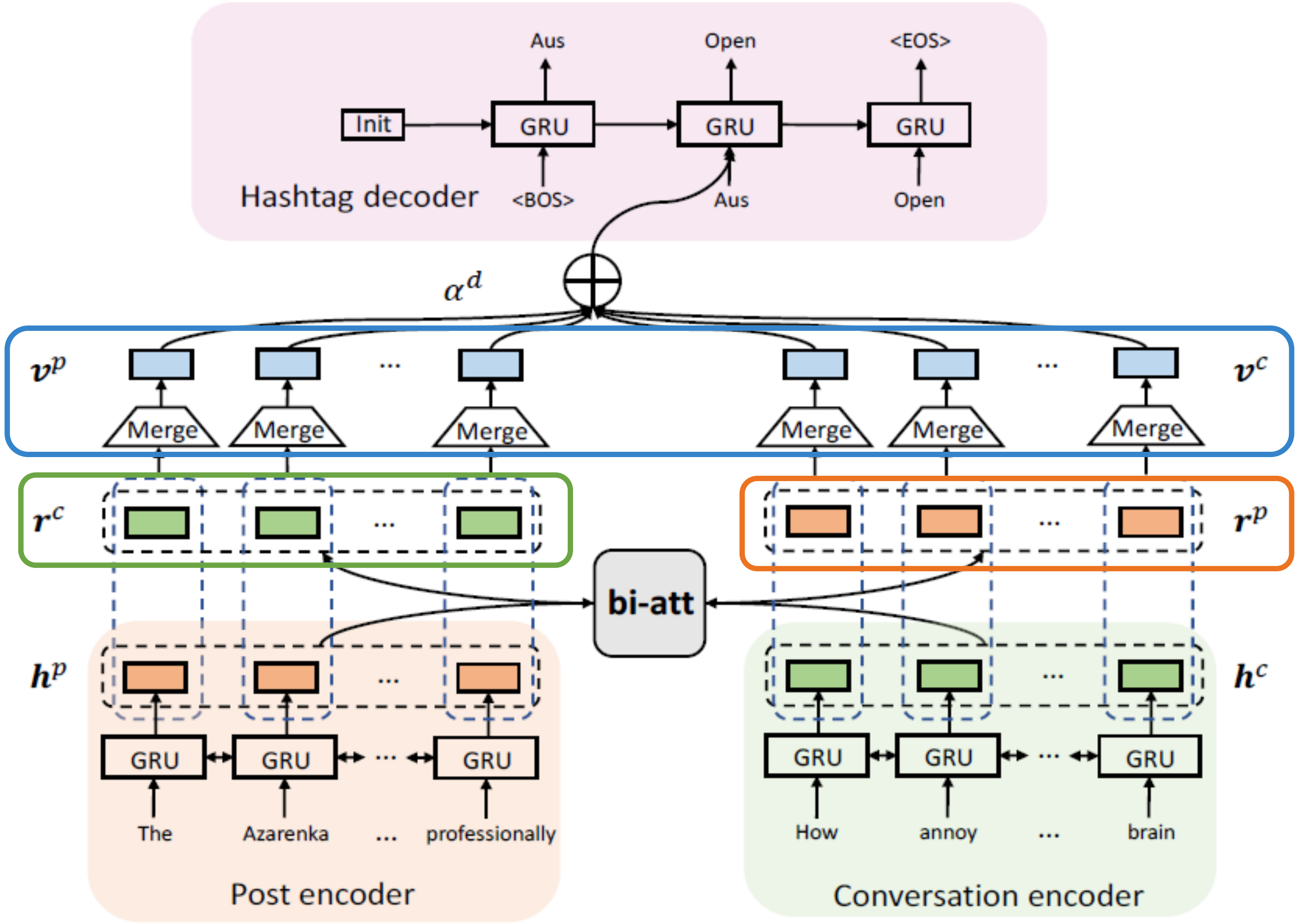
Conversation encoder

- $\boldsymbol{h}^c = \mathrm{BiGRU}(\boldsymbol{x}^c)$

Bi-attention (bi-att)

- $\alpha_{ij}^c = \frac{\exp(f_{score}(\boldsymbol{h}_i^p, \boldsymbol{h}_j^c))}{\sum_{j'=1}^{|x^c|} \exp(f_{score}(\boldsymbol{h}_i^p, \boldsymbol{h}_{j'}^c))}$,
- $\alpha_{ij}^p = \frac{\exp(f_{score}(\boldsymbol{h}_i^p, \boldsymbol{h}_j^c))}{\sum_{i'=1}^{|x^p|} \exp(f_{score}(\boldsymbol{h}_{i'}^p, \boldsymbol{h}_j^c))}$,
- $f_{score}(\boldsymbol{h}_i^p, \boldsymbol{h}_j^c) = \boldsymbol{h}_i^p \, \boldsymbol{W}_{bi-att} \boldsymbol{h}_j^c$

The Framework of Our Model



Conversation-attentive vector

- $\boldsymbol{r}_i^c = \sum_{j=1}^{|x^c|} \alpha_{ij}^c \, \boldsymbol{h}_j^c$

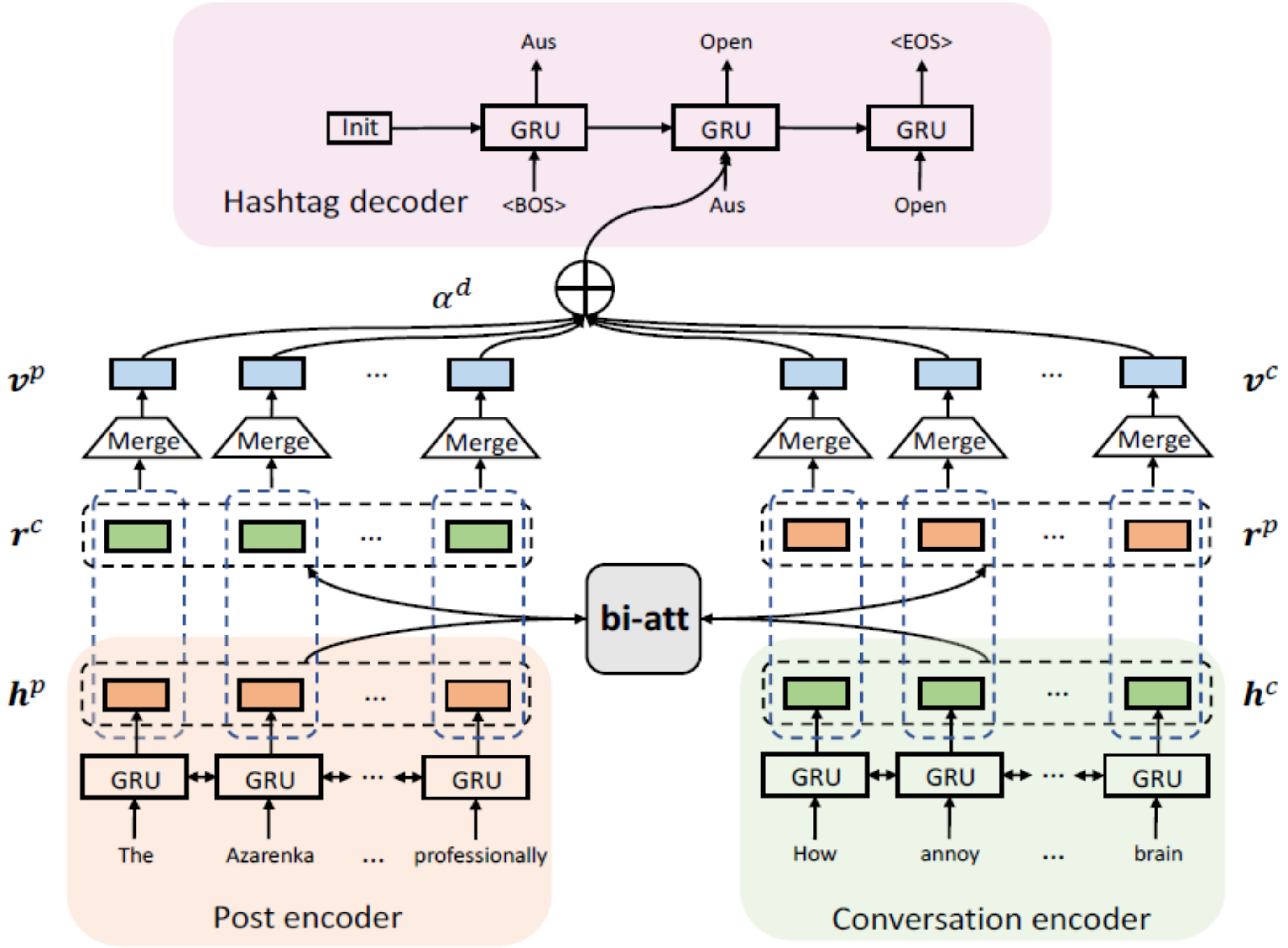
Post-attentive vector

- $\boldsymbol{r}_j^p = \sum_{i=1}^{|x^p|} \alpha_{ij}^p \, \boldsymbol{h}_i^p$

Merge layer

- $\boldsymbol{v}^p = \tanh(\boldsymbol{W}_p[\boldsymbol{h}^p; \boldsymbol{r}^c] + \boldsymbol{b}_p)$,
- $\boldsymbol{v}^c = \tanh(\boldsymbol{W}_c[\boldsymbol{h}^c; \boldsymbol{r}^p] + \boldsymbol{b}_c)$,
- $\boldsymbol{v} = [\boldsymbol{v}^p; \boldsymbol{v}^c]$,

The Framework of Our Model



Hashtag decoder

- $\Pr(\boldsymbol{y}_t) = softmax(\boldsymbol{W}_v[\boldsymbol{s}_t; \boldsymbol{c}_t] + \boldsymbol{b}_v)$,
- $\mathbf{c}_\mathrm{t} = \sum_{i=1}^{|\boldsymbol{x}^p|+|\boldsymbol{x}^c|} \alpha_{ij}^d \boldsymbol{v}_i$,
- $\alpha_{ti}^d = \frac{\exp(g_{score}(\boldsymbol{s}_t, \boldsymbol{v}_i))}{\sum_{i'=1}^{|\boldsymbol{x}^p|+|\boldsymbol{x}^c|} \exp(g_{score}(\boldsymbol{s}_t, \boldsymbol{v}_{i'}))}$,
- $g_{score}(\boldsymbol{s}_t, \boldsymbol{v}_i) = \boldsymbol{s}_t \boldsymbol{W}_{att} \boldsymbol{v}_i$

Loss function

- $L(\theta) = -\sum_{n=1}^{N} \log(\Pr(y_n | x_n^p, x_n^c; \theta))$.

Inference: beam search

Outline

- Background
- The Framework of Our Model
- **Experiments**
- Conclusions

Dataset

- **Twitter**: English dataset from TREC 2011 Twitter
- **Weibo**: Chinese dataset crawled from Sina Weibo

Datasets	# of posts	Avg len of posts	Avg len of convs	Avg len of tags	# of tags per post
Twitter	44,793	13.27	29.94	1.69	1.14
Weibo	40,171	32.64	70.61	2.70	1.11

- 80% training, 10% validation, 10% testing
- Gold standards : hashtags appearing before or after the post

Dataset

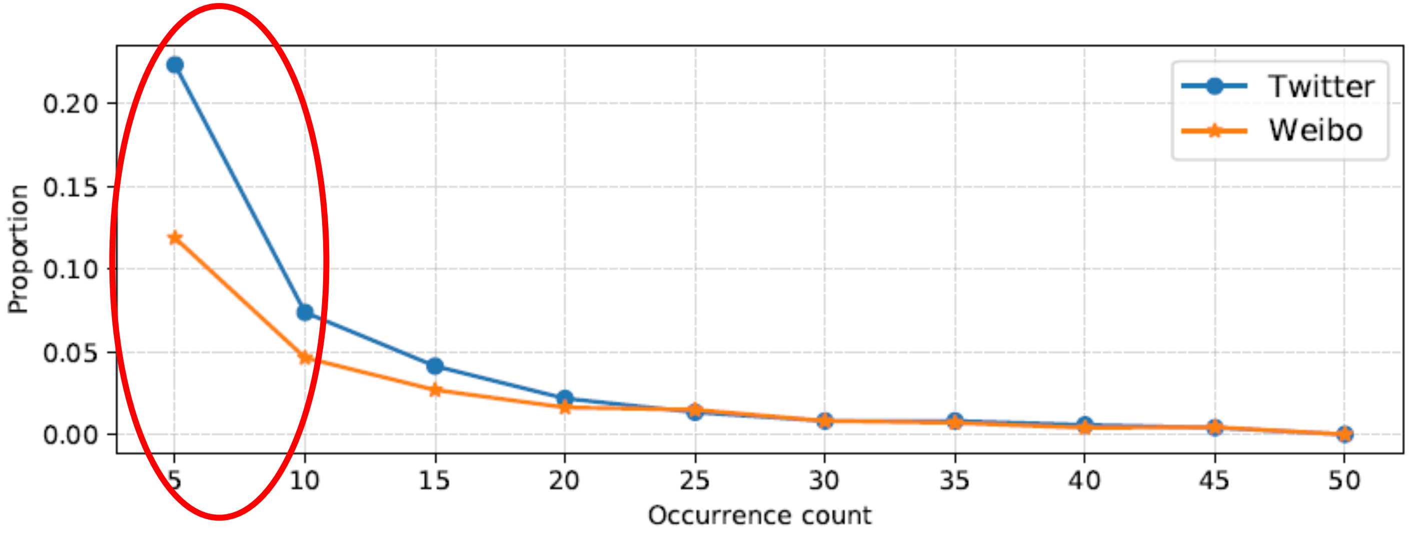
- Hashtag statistics (present ratio)

Datasets	|Tagset|	$\mathcal{P}$	$\mathcal{C}$	$\mathcal{P} \cup \mathcal{C}$
Twitter	4,188	2.72%	5.58%	7.69%
Weibo	5,027	8.29%	6.21%	12.52%

P: target post
C: conversation

Low present ratio

- Hashtag frequency distribution



Large and **imbalanced** hashtag space!

Main Experiment

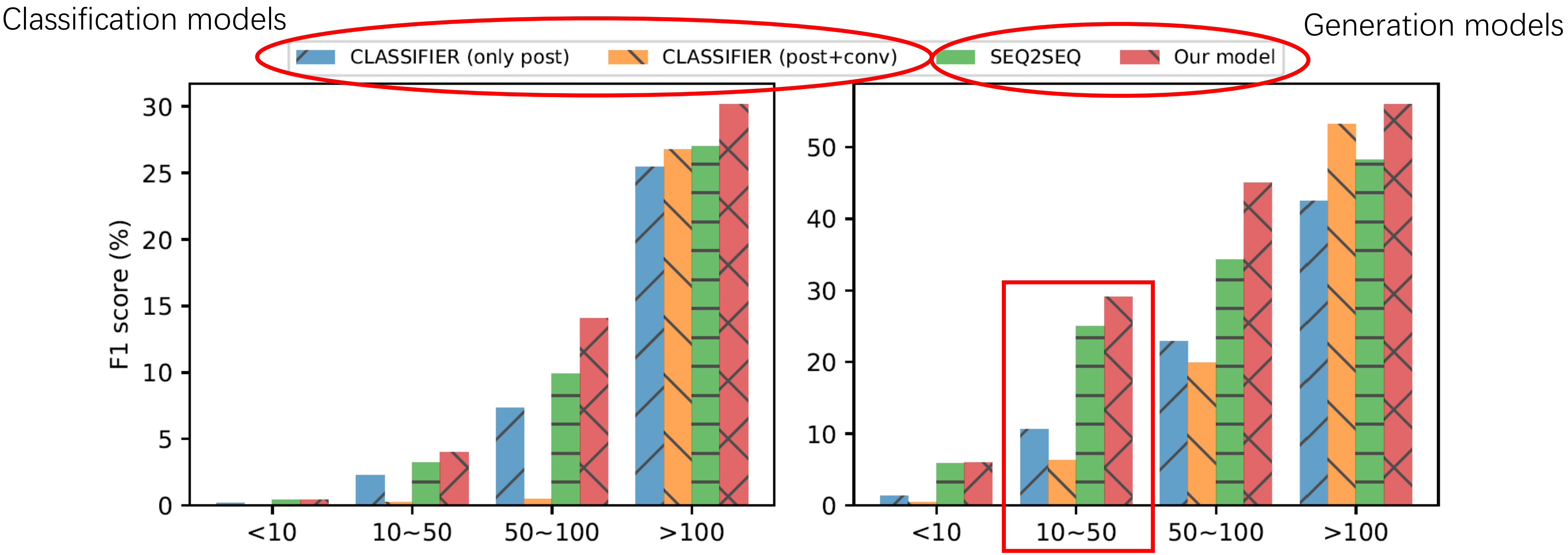
Model	Twitter					Weibo				
	F1@1	F1@5	MAP	RG-1	RG-4	F1@1	F1@5	MAP	RG-1	RG-4
Baselines										
RANDOM	0.37	0.63	0.89	0.56	0.16	0.43	0.67	0.97	2.14	1.13
LDA	0.13	0.25	0.35	0.60	-	0.10	0.86	0.94	3.89	-
TF-IDF	0.02	0.02	0.03	0.54	0.14	0.85	0.73	1.30	8.04	4.29
EXTRACTOR	0.44	-	-	1.14	0.14	2.53	-	-	7.64	5.20
State of the arts										
CLASSIFIER (*post only*)	9.44	6.36	12.71	10.75	4.00	16.92	10.48	22.29	25.34	21.95
CLASSIFIER (*post+conv*)	8.54	6.28	12.10	10.00	2.47	17.25	11.03	23.11	25.16	22.09
GENERATORS										
SEQ2SEQ	10.44	6.73	14.00	10.52	4.08	26.00	14.43	32.74	37.37	32.67
SEQ2SEQ-COPY	10.63	6.87	14.21	12.05	4.36	25.29	14.10	31.63	37.58	32.69
OUR MODEL	**12.29***	**8.29***	**15.94***	**13.73***	**4.45**	**31.96***	**17.39***	**38.79***	**45.03***	**39.73***

Why?

- The "*" indicates significantly better than other models ($p < 0.05$, paired t-test).

- The task is very challenging, especially for Twitter dataset
- Our model significantly outperforms all the comparison models

Classification vs. Generation



Varying hashtag frequency: Twitter (left) and Weibo (right).

- The hashtag frequency ↓ , the performance ↓
- Generation models consistently outperform classification models
- Generation models perform more robustly

Classification vs. Generation

Model	**Twitter**	**Weibo**
CLASSIFIER (*post only*)	1.15	1.65
CLASSIFIER (*post+conv*)	1.13	1.52
SEQ2SEQ	1.33	10.84
OUR MODEL	**1.48**	**12.55**

Unseen hashtags (ROUGE-1 in %)

- It is difficult to generate new hashtags
- At least 6.5x improvements over classification models on Weibo

Ablation Study

	Model	Twitter	Weibo
	SEQ2SEQ (*post only*)	10.44	26.00
	SEQ2SEQ (*conv only*)	6.27	18.57
w/o bi-att	SEQ2SEQ (*post + conv*)	11.24	29.85
	OUR MODEL (*post-att only*)	11.18	28.67
	OUR MODEL (*conv-att only*)	10.61	28.06
w/ bi-att	OUR MODEL (*full*)	**12.29**	**31.96**

Ablation result (F1 in %)

- Post is more important!
- Bi-attention is helpful!

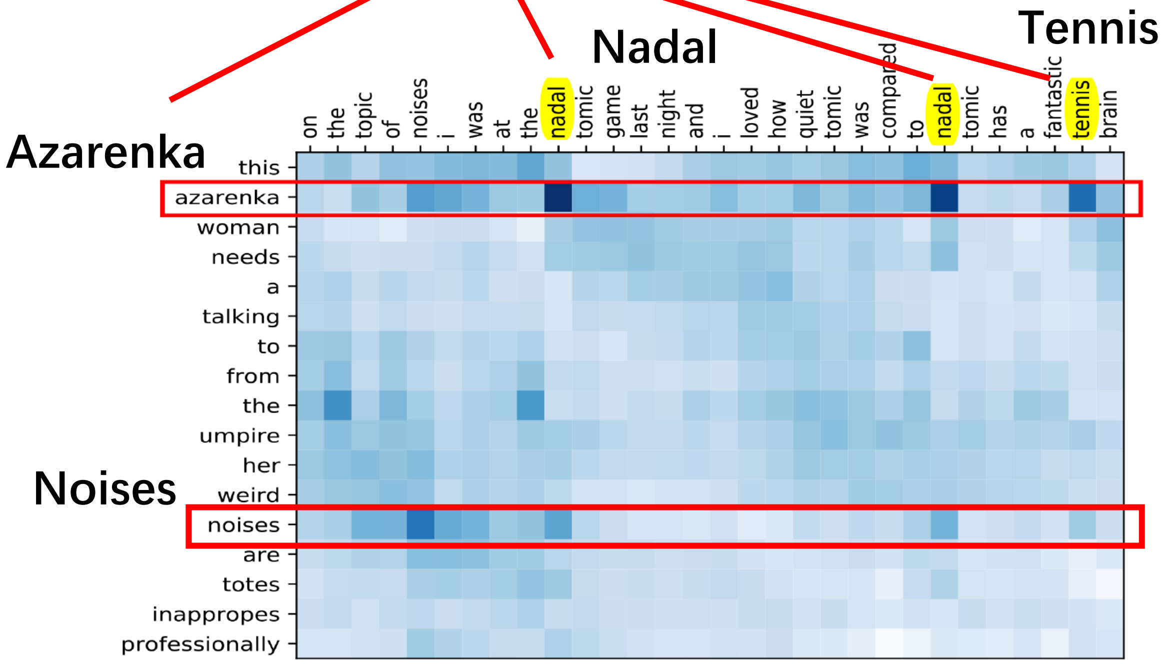
Case Study

Case post

“This Azarenka woman needs a talking to from the umpire her weird noises are totes inappropes professionally.” ***#AusOpen***

Model	Top five outputs
LDA	found; stated; excited; card; apparently
TF-IDF	inappropes; umpire; woman need; azarenka woman; the umpire
CLASSIFIER	fail; facebook; just saying; quote; pro choice
SEQ2SEQ	fail; jan 25; yr; eastenders; facebook
OUR MODEL	***aus open*** ; bbc football ; bbc aus ; arsenal ; murray

(a) Model outputs for the case post



(b) Heatmap visualization of bi-attention

Outline

- Background
- The Framework of Our Model
- Experiments
- **Conclusions**

Conclusions

- We are the first to approach microblog hashtag annotation with sequence generation architecture.
- To alleviate data sparsity, we enrich context for short target posts with their conversations using a bi-attention mechanism.
- Our model establishes a new state-of-the-art results on two datasets.

- **Future work**
  - Extend to other scenarios, e.g., dialogue
  - Deal with the data sparsity
    - Other external knowledge, e.g., multimodal
    - When external knowledge is unavailable (our ACL 19 work, to appear)

Thanks!

Code: https://github.com/yuewang-cuhk/HashtagGeneration

Contact: yuewang-cuhk.github.io

Reference

1. Yuyun Gong and Qi Zhang. 2016. Hashtag recommendation using attention-based convolutional neural network. In International Joint Conference on Artificial Intelligence.
2. Yeyun Gong, Qi Zhang, and Xuanjing Huang. 2015. Hashtag recommendation using dirichlet process mixture models incorporating types of hashtags. In Empirical Methods in Natural Language Processing.
3. Qi Zhang, Yang Wang, Yeyun Gong, and Xuanjing Huang. 2016. Keyphrase extraction using deep recurrent neural networks on twitter. In Empirical Methods in Natural Language Processing.
4. Yingyi Zhang, Jing Li, Yan Song, and Chengzhi Zhang. 2018. Encoding conversation context for neural keyphrase extraction from microblog posts. In North American Chapter of the association for Computational Linguistics: Human Language Technologies.
5. Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural etworks. In Neural Information Processing Systems.
6. Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. 2017. Deep keyphrase generation. In Association for Computational Linguistics.