Neural Keyphrase Generation for Social Media Understanding

WANG, Yue

Ph.D. Oral Defense

Supervisor: Prof. Michael R. Lyu & Prof. Irwin King

2020/12/01

香港中文大學

The Chinese University of Hong Kong

How Social Media Change Our Life?

Kitchen table conversation

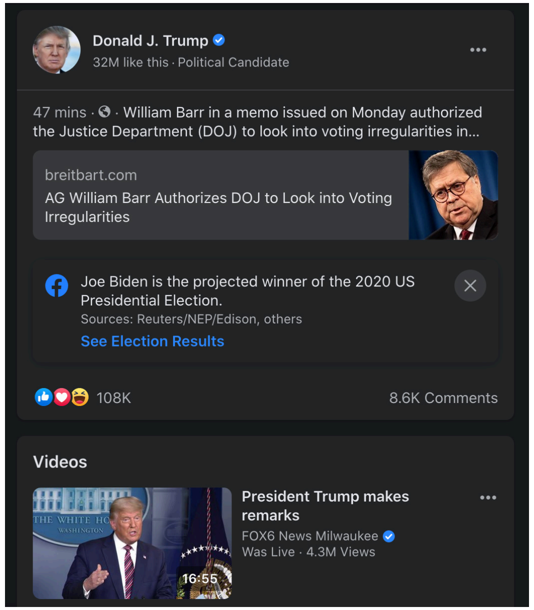
Online social networking

Social Media is Connecting the World

3.5 billion users (45% of the population)

- 3 hours per day

Twitter



Facebook

Sina Weibo

https://www.oberlo.com/blog/social-media-marketing-statistics

Social Media is Everywhere

How to automatically understand the massive amount of social media content?

Information sharing

Entertainment

Marketing

How to Understand Social Media Content?

Big Data

Keyphras Prediction

Microblog search

Trending topic discovery

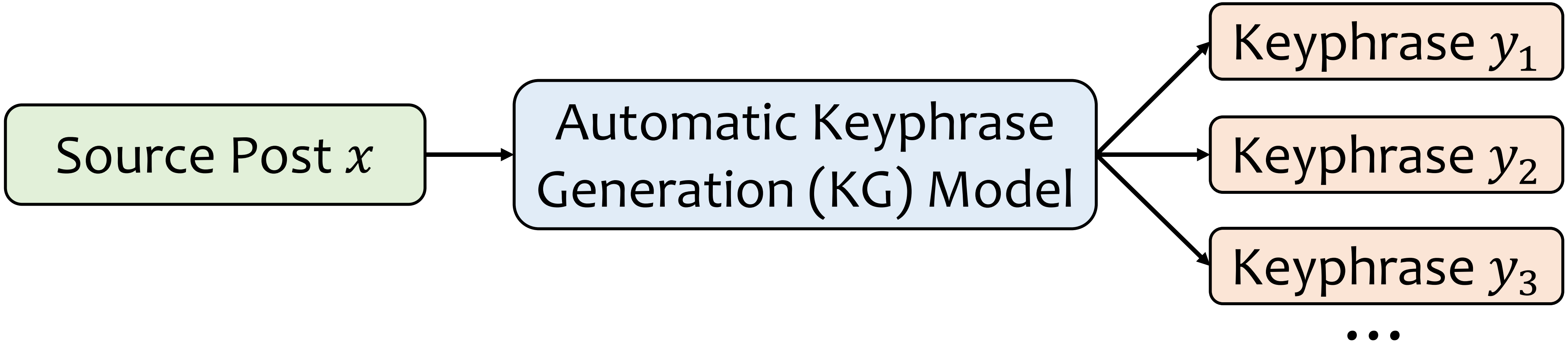
Sentiment analysis

Problem Definition

- Hashtags → **Keyphrases**

- Pressing need: there are only 15% of tweets containing hashtags
- **Keyphrase generation:**

E.g., "#ACL2019nlp" → {"ACL", "2019", "nlp"}

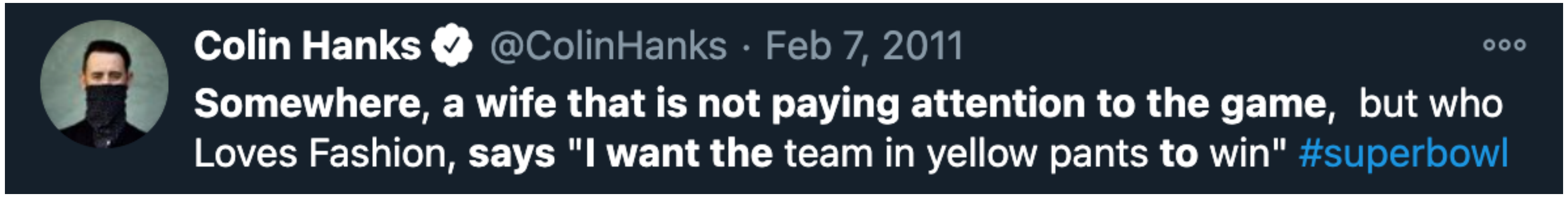


Present and Absent Keyphrase

- Present keyphrase

Congratulations to all authors who have a paper accepted at ACL 2019 nlp! ... → KG Model → ACL 2019 nlp

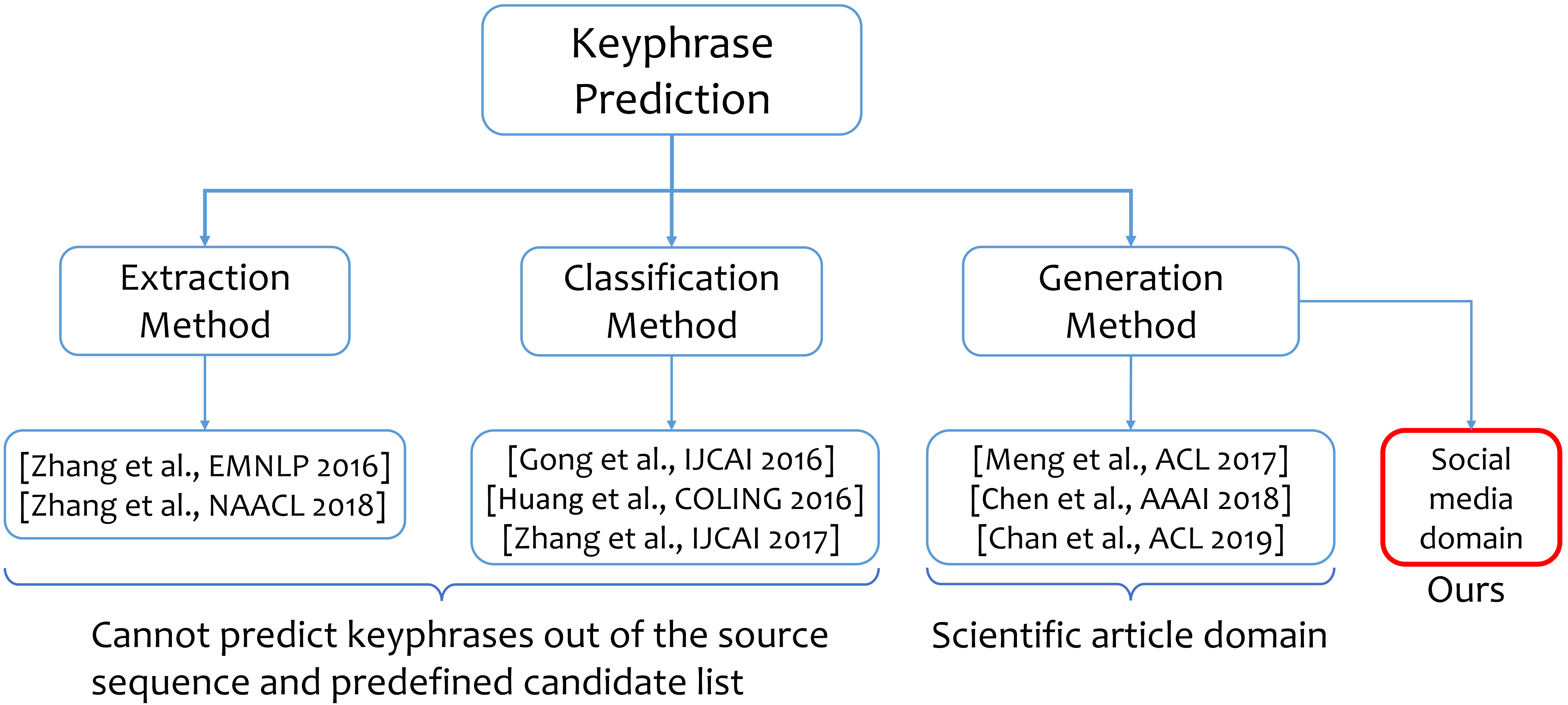
- Absent keyphrase



Somewhere, a wife …, says "I want the team in yellow pants to win." → KG Model → Super bowl

More difficult!

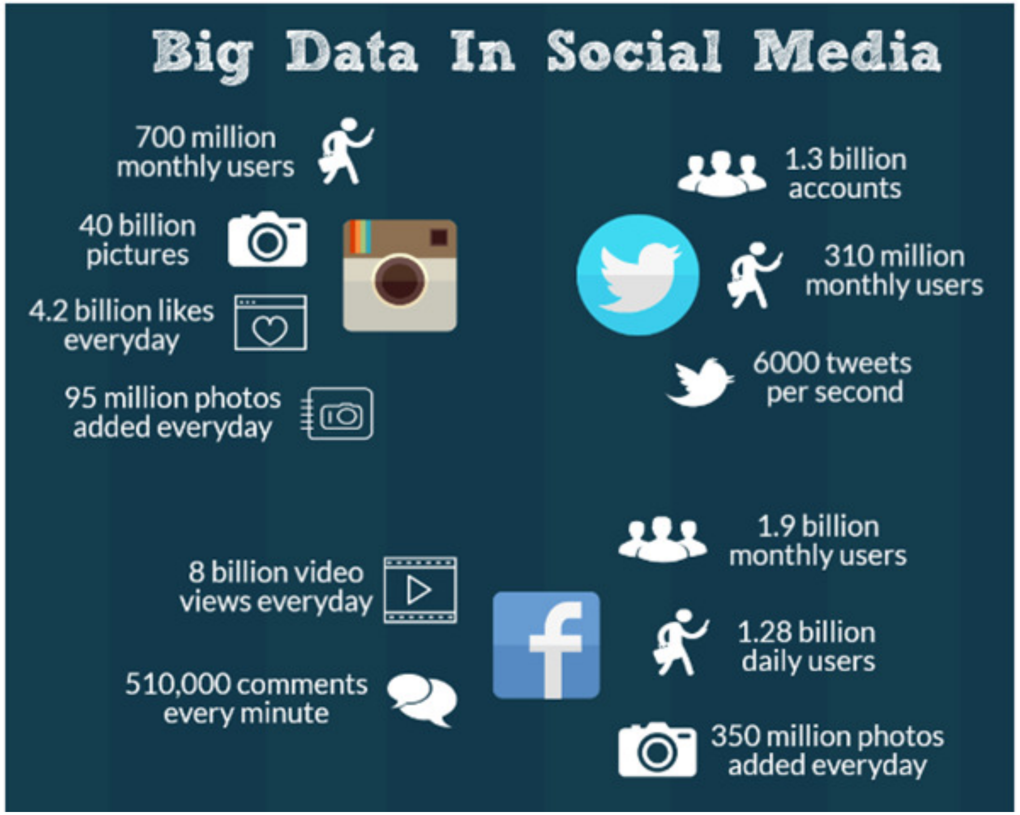
Previous Method



Keyphrase prediction in social media is challenging!

Challenge – Huge Volume

- Facebook: 4 million posts per minute
- Twitter: 21 million posts per hour
- Weibo: 130 million posts per day



Challenge – Data Sparsity

- Informal style
- Short in length
- Syntax errors

Example Tweet

lol~~
fearless man we r :-)
keep fffffffighting @StephenCurry30

Challenge – Multimedia Data

- What's the largest difference in Twitter content in 2010 and 2020?
  - Many more tweets contain multimedia data!

- Approximately 12% tweets are accompanied by images

2010

2020

Our Solution

?

Example

Thank you fox for showing the good sposmanship segment! That's what it should always be like. ***#SuperBowl***

Implicit topic

Sports

Explicit conversation

Replying messages forming a conversation

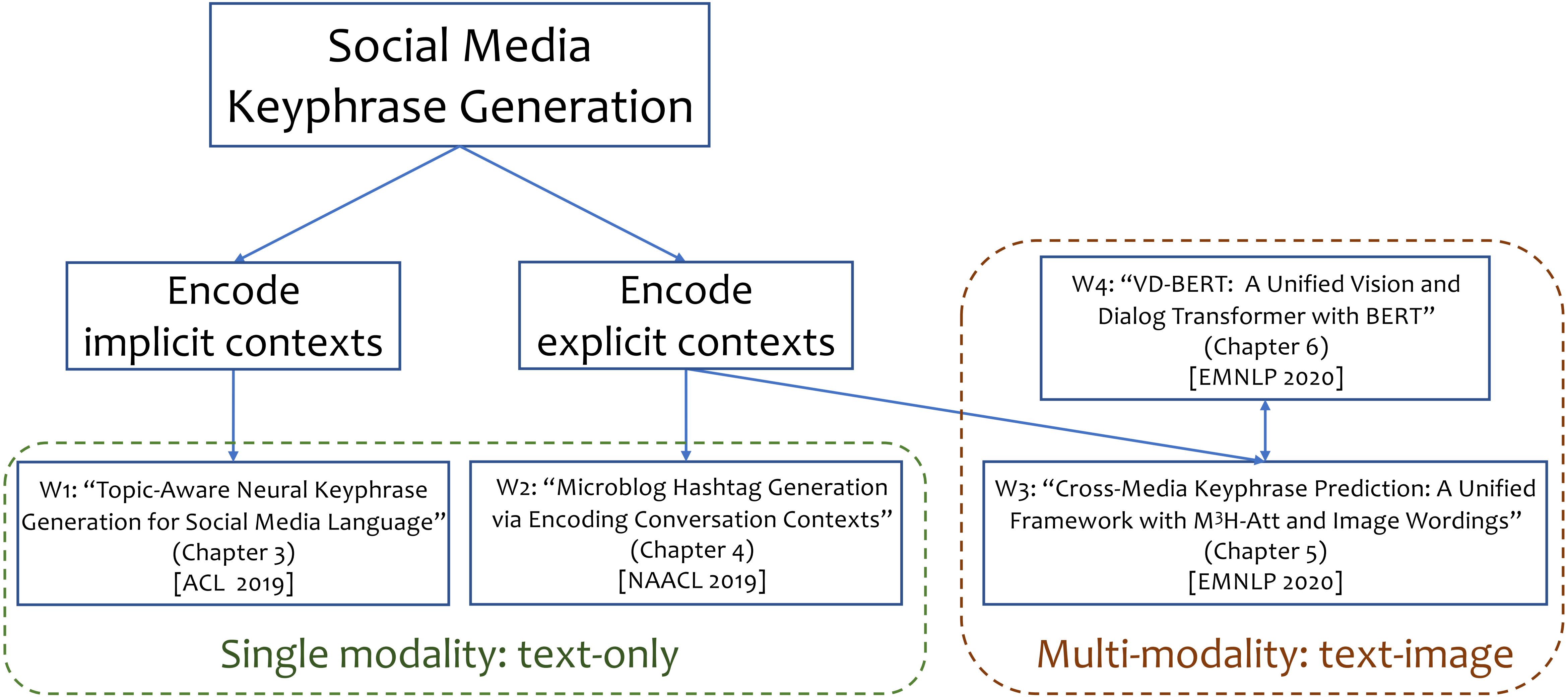
[T1] Bet you are happy dancing right about now lol! You are the biggest Steelers fan I know, so I have been thinking of you tonight.

[T2] Thank you! That's a huge compliment. They have won a lot this season. It would have been poetic to end the season that way.

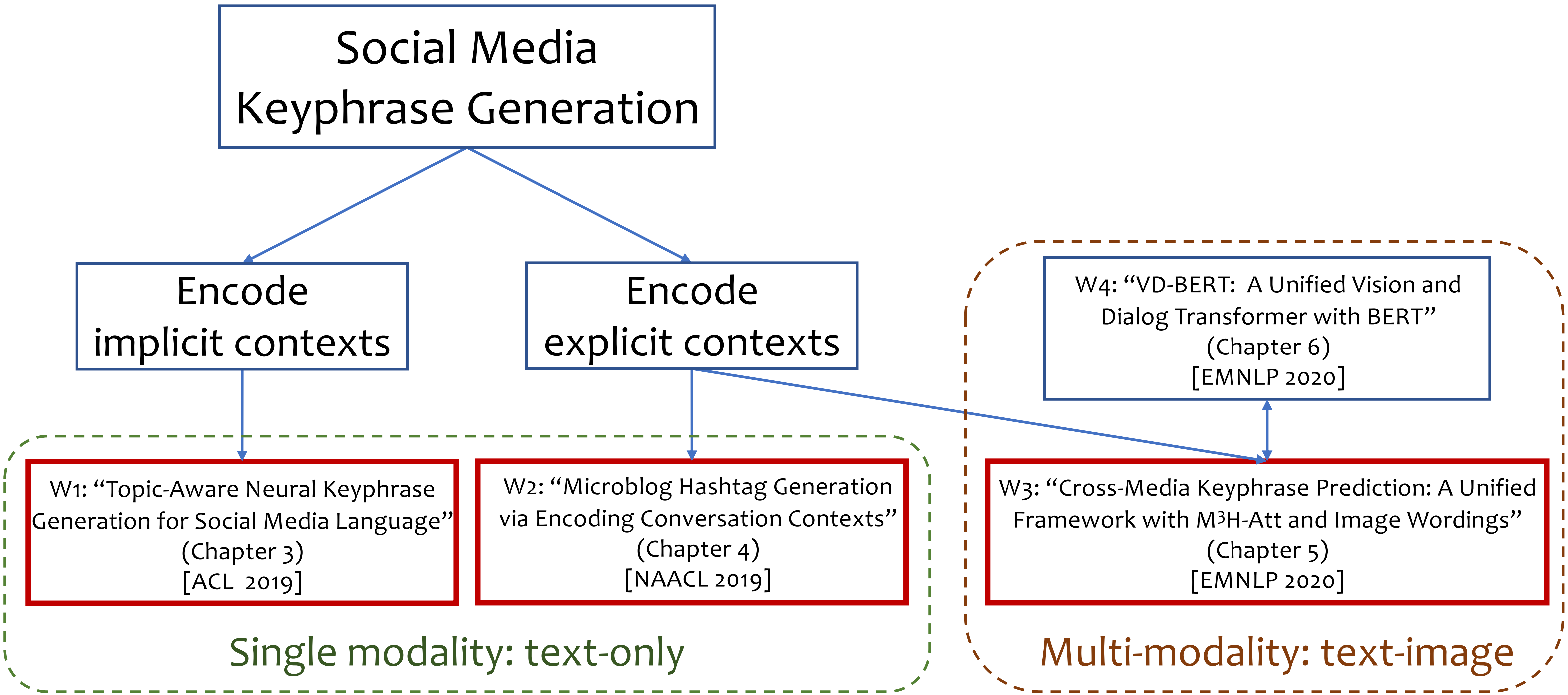
[T3] Yes, just think of all the money you will save, not having to buy all the **SuperBowl** champions gear.

Explicit image

Thesis Contributions



Thesis Contributions



Outline

- Topic 1: Topic-aware Keyphrase Generation
- Topic 2: Conversation-aware Keyphrase Generation
- Topic 3: Unified Cross-media Keyphrase Prediction
- Conclusion and Future Work

Outline

- Topic 1: Topic-aware Keyphrase Generation
- Topic 2: Conversation-aware Keyphrase Generation
- Topic 3: Unified Cross-media Keyphrase Prediction
- Conclusion and Future Work

Motivation

Example

Somewhere, a wife that is not paying attention to the game, says "I want the team in yellow pants to win."

Relevant tweets

[T1] I been a steelers fan way before black & yellow and this super bowl!

[T2] I will bet you the team with yellow pants wins.

[T3] Wiz Khalifa song "black and yellow" to spur the pittsburgh steelers and Lil Wayne is to sing "green and yellow" for the packers.

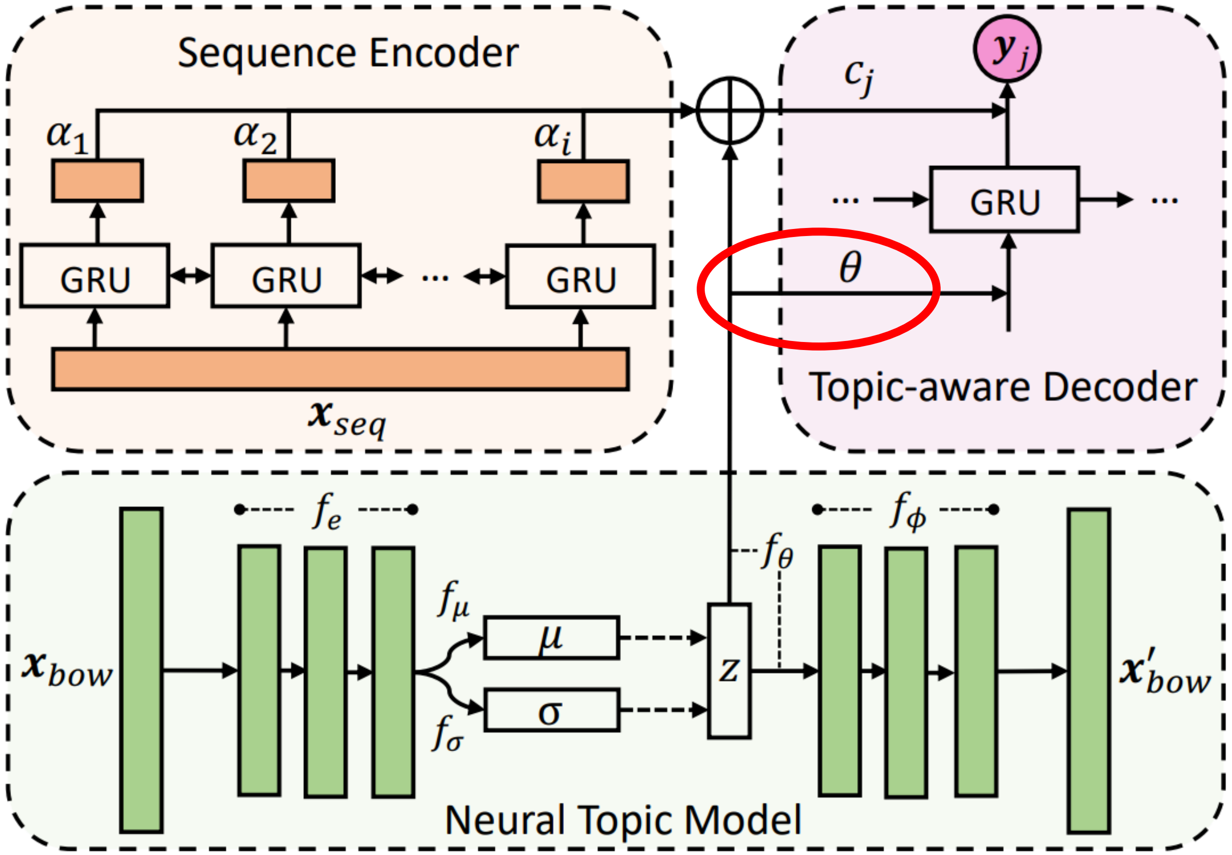
- By looking at other tweets with a similar topic, we can infer "***Superbowl***"
- ***Latent topics*** learned from the corpus can alleviate the data sparsity

Methodology

Seq2seq

Jointly train

NTM



Input

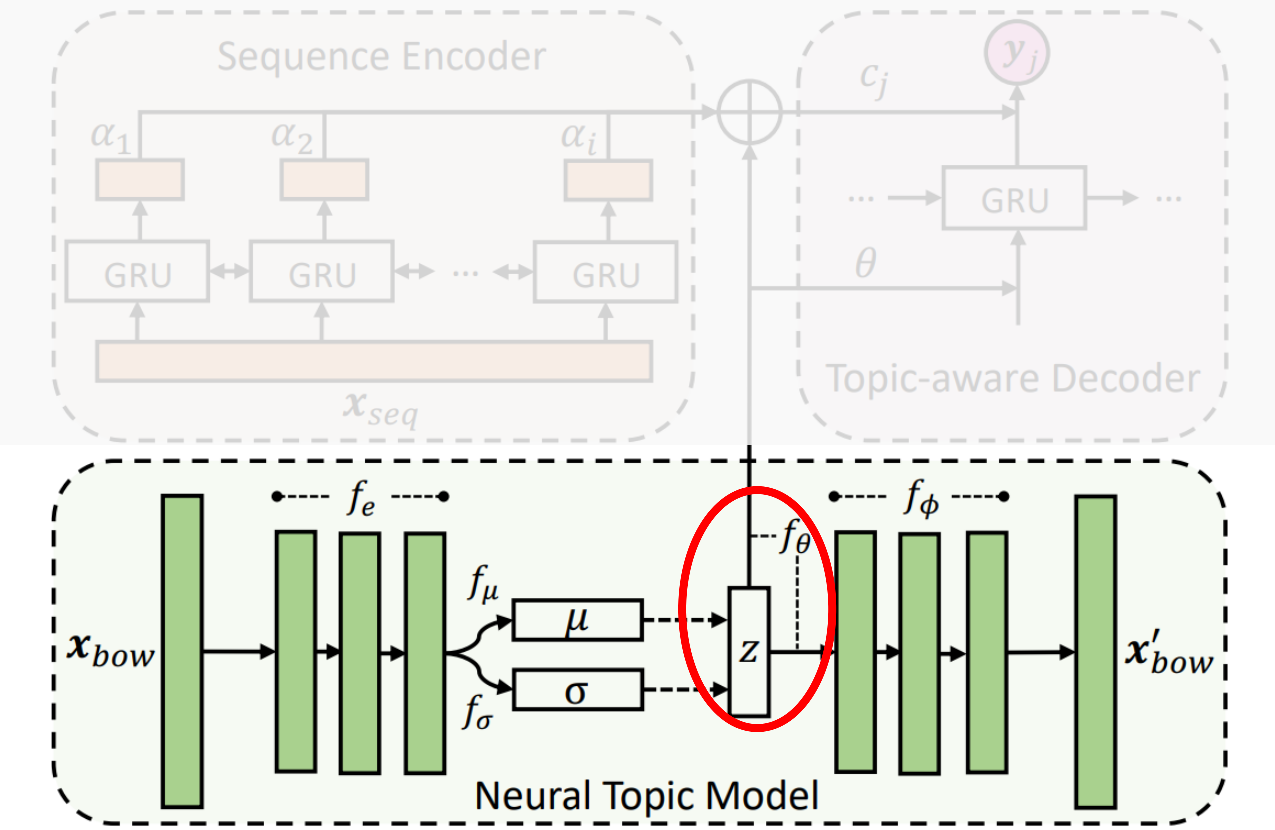
Post in word index $\boldsymbol{x}_{seq}$

Post in bag of words $\boldsymbol{x}_{bow}$

Keyphrase: $\langle \boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_{|\boldsymbol{y}|} \rangle$

Output

Methodology



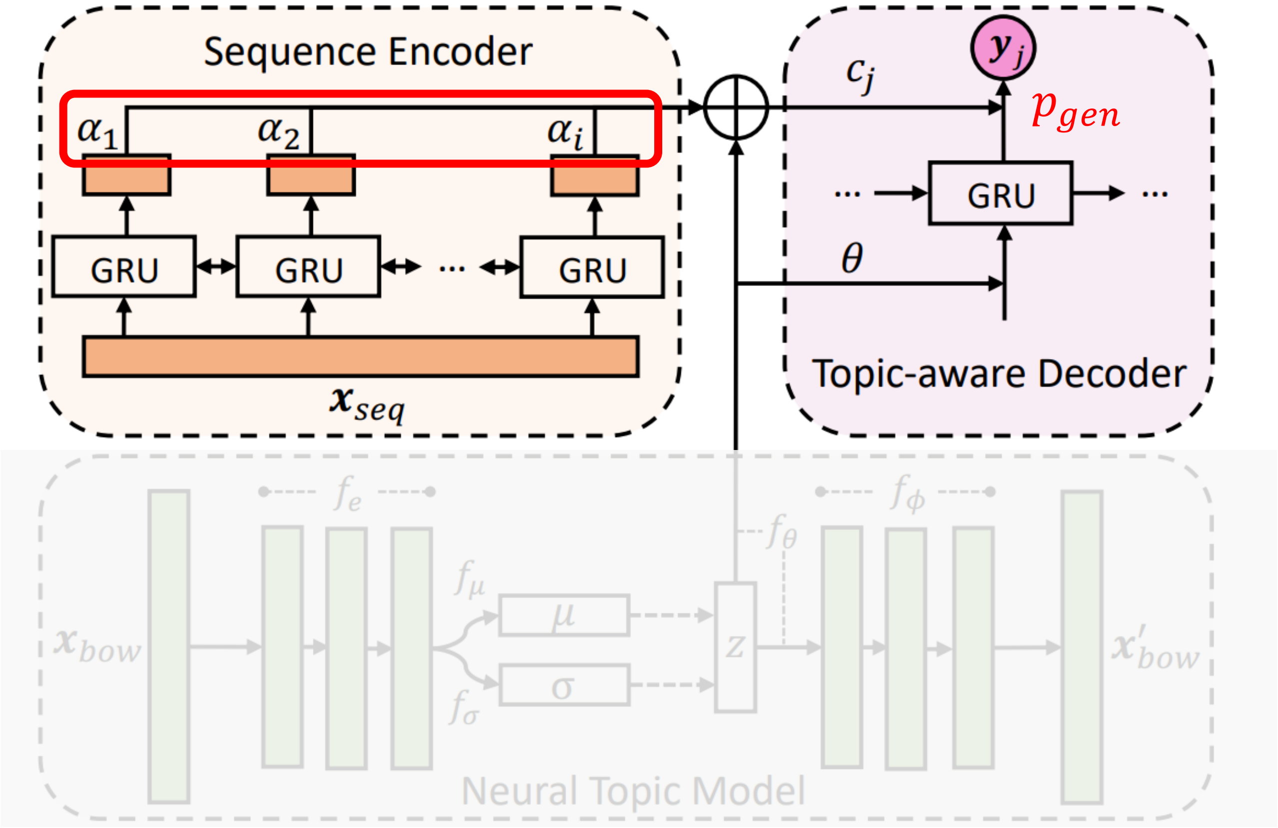
BoW Encoder BoW Decoder

Neural Topic Model (NTM)

- Proposed by [Miao et al., ICML 2017]

- BoW Encoder
  - Prior latent variables
  - $\boldsymbol{\mu} = f_\mu(f_e(\boldsymbol{x}_{bow}))$
  - $\log \boldsymbol{\sigma} = f_\sigma(f_e(\boldsymbol{x}_{bow}))$
- BoW Decoder
  - Draw latent variable $\boldsymbol{z} \sim N(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$
  - Topic mixture $\boldsymbol{\theta} = softmax(f_\theta(\boldsymbol{z}))$
  - For each word $w \in \boldsymbol{x}$:
    - Draw word $w \sim softmax(f_\varphi(\boldsymbol{\theta}))$

Methodology



Seq2Seq keyphrase generation model

- Global vocabulary:

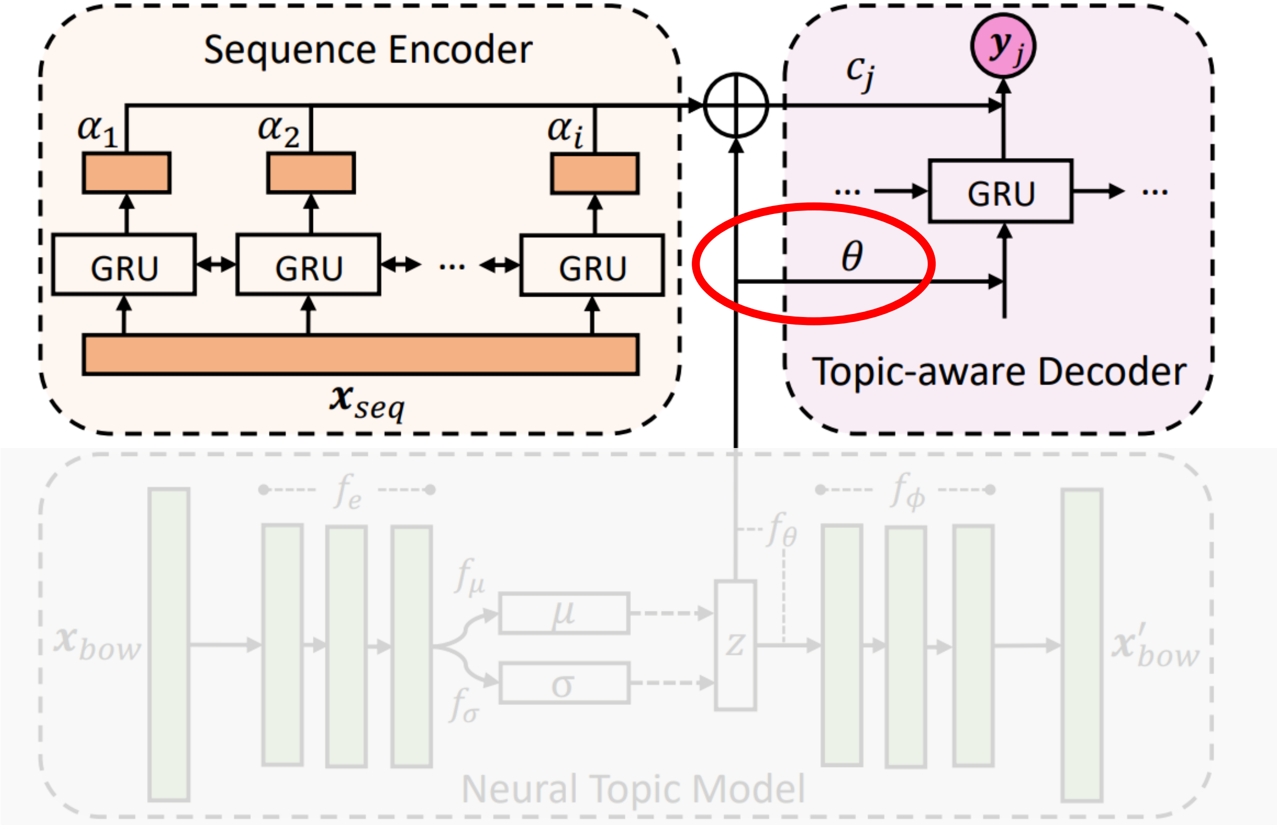
$$p_{gen} = softmax(\mathbf{W}_{gen}[\mathbf{s}_j; \mathbf{c}_j] + \mathbf{b}_{gen})$$

- Local extractive distribution: $\{\alpha_{ij}\}_{i=1}^{|\mathbf{x}|}$

- Generation with copy mechanism:
  - Proposed by [See et al., ACL 2017]

$$p_j = \lambda_j \cdot p_{gen} + (1 - \lambda_j) \cdot \sum_{i=1}^{|\mathbf{x}|} \alpha_{ij},$$

Methodology



How to feed the topic θ into the keyphrase generation model?

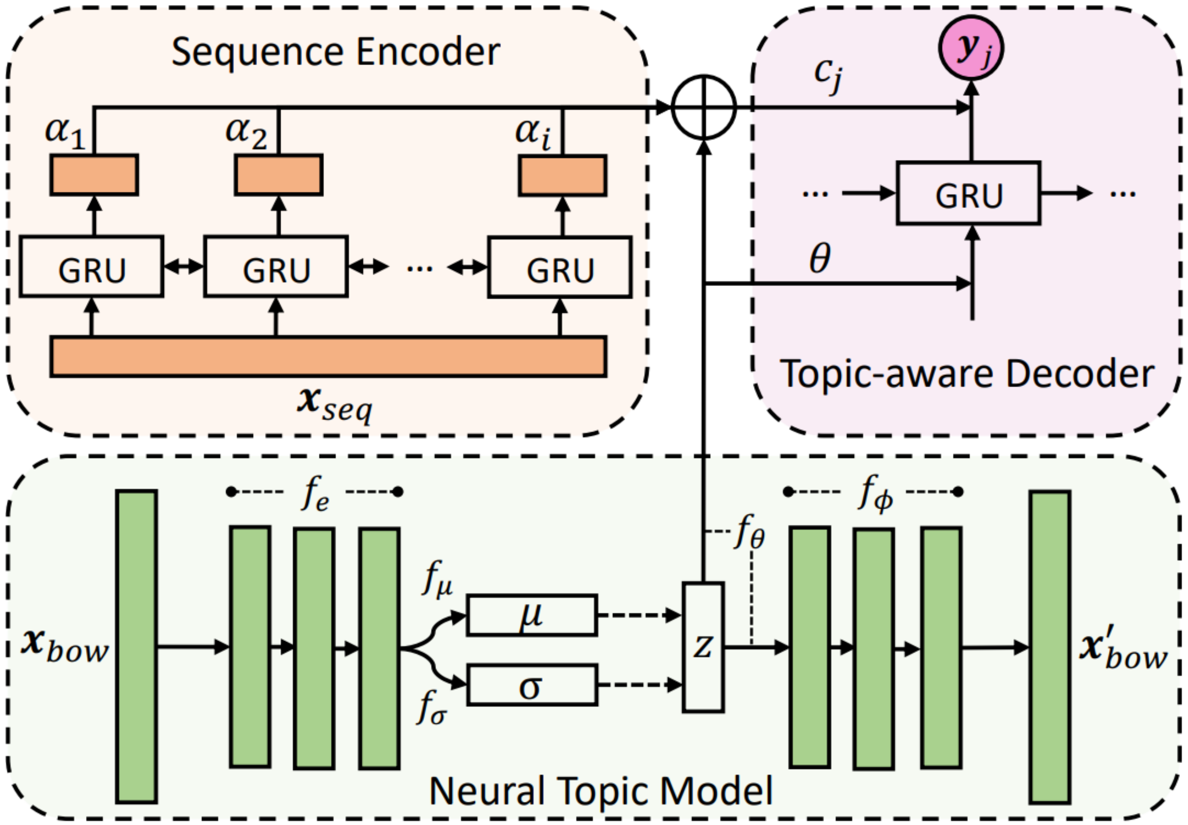
- Three paths

Decoder state: $s_j = f_{GRU}([u_j; \theta], s_{j-1})$

Attention: $f_\alpha(\cdot) = \boldsymbol{v}_\alpha^T tanh(W_\alpha[h_i; s_j; \theta] + b_\alpha)$

Copy switch: $\lambda_j = \sigma(W_\lambda[u_j; s_j; c_j; \theta] + b_\lambda)$

Methodology



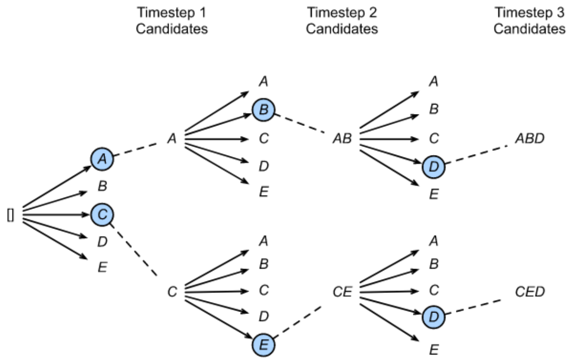
- End-to-end joint training

$$\mathcal{L}_{NTM} = D_{KL}(p(\mathbf{z}) \,||\, q(\mathbf{z} \,|\, \mathbf{x})) - \mathbb{E}_{q(\mathbf{z} \,|\, \mathbf{x})}[p(\mathbf{x} \,|\, \mathbf{z})],$$

$$\mathcal{L}_{KG} = -\sum_{n=1}^{N} \log(Pr(\mathbf{y}_n \,|\, \mathbf{x}_n, \theta_n)),$$

$$\mathcal{L} = \mathcal{L}_{NTM} + \gamma \cdot \mathcal{L}_{KG}$$

- Inference
  - Beam search



Datasets

- We newly construct three datasets in both English and Chinese

Source posts	# of posts	Avg len per post	# of KP per post	Source vocab
Twitter	44,113	19.52	1.13	34,010
Weibo	46,296	33.07	1.06	98,310
StackExchange	49,447	87.94	2.43	99,775
Target KP	$\|KP\|$	**Avg len per KP**	**% of abs KP**	**Target vocab**
Twitter	4,347	1.92	71.35	4,171
Weibo	2,136	2.55	75.74	2,833
StackExchange	12,114	1.41	54.32	10,852

- KP→Keyphrase

Datasets

Source posts	# of posts	Avg len per post	# of KP per post	Source vocab
Twitter	44,113	19.52	1.13	34,010
Weibo	46,296	33.07	1.06	98,310
StackExchange	49,447	87.94	2.43	99,775

Target KP	$\|KP\|$	Avg len per KP	% of abs KP	Target vocab
Twitter	4,347	1.92	71.35	4,171
Weibo	2,136	2.55	75.74	2,833
StackExchange	12,114	1.41	54.32	10,852

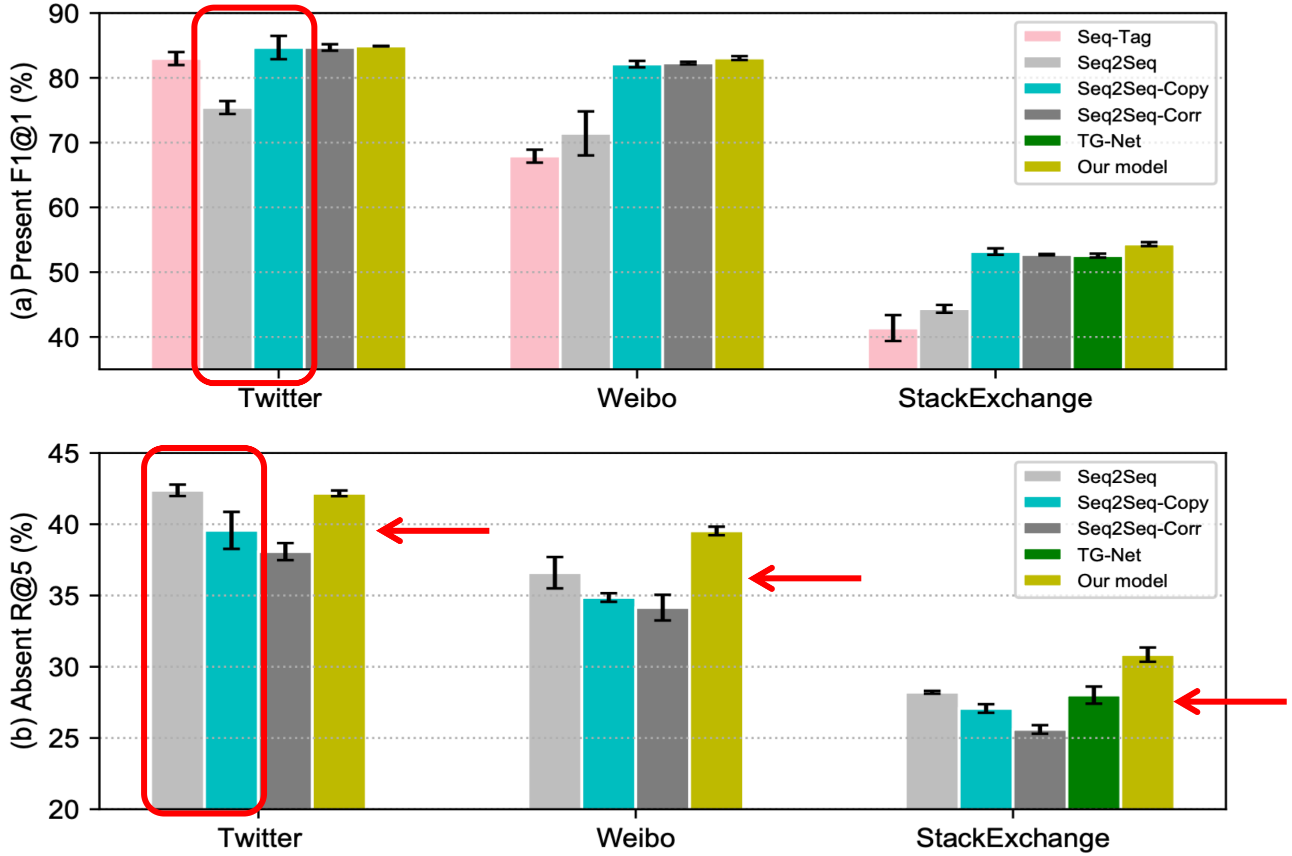
- StackExchange has much longer text and more unique keyphrases
- High absent keyphrase rates (over 50%)

Main Results

Model	Twitter			Weibo			StackExchange		
	F1@1	F1@3	MAP	F1@1	F1@3	MAP	F1@3	F1@5	MAP
Baselines									
MAJORITY	9.36	11.85	15.22	4.16	3.31	5.47	1.79	1.89	1.59
TF-IDF	1.16	1.14	1.89	1.90	1.51	2.46	13.50	12.74	12.61
TEXTRANK	1.73	1.94	1.89	0.18	0.49	0.57	6.03	8.28	4.76
KEA	0.50	0.56	0.50	0.20	0.20	0.20	15.80	15.23	14.25
State of the arts									
SEQ-TAG	22.79±0.3	12.27±0.2	22.44±0.3	16.34±0.2	8.99±0.1	16.53±0.3	17.58±1.6	12.82±1.2	19.03±1.3
SEQ2SEQ	34.10±0.5	26.01±0.3	41.11±0.3	28.17±1.7	20.59±0.9	34.19±1.7	22.99±0.3	20.65±0.2	23.95±0.3
SEQ2SEQ-COPY	36.60±1.1	26.79±0.5	43.12±1.2	32.01±0.3	22.69±0.2	38.01±0.1	31.53±0.1	27.41±0.2	33.45±0.1
SEQ2SEQ-CORR	34.97±0.8	26.13±0.4	41.64±0.5	31.64±0.7	22.24±0.5	37.47±0.8	30.89±0.3	26.97±0.2	32.87±0.6
TG-NET	-	-	-	-	-	-	32.02±0.3	27.84±0.3	34.05±0.4
Our model	**38.49**±0.3	**27.84**±0.0	**45.12**±0.2	**34.99**±0.3	**24.42**±0.2	**41.29**±0.4	**33.41**±0.2	**29.16**±0.1	**35.52**±0.1

- Social media keyphrase prediction is challenging
- Seq2seq-based keyphrase generation models are effective
- Latent topics are consistently helpful for indicating keyphrases

Present and Absent Keyphrase Prediction



- Our model achieves comparable or better performance in both settings
- Copy mechanism sacrifice the absent keyphrase prediction performance for better predicting the present ones .
  - ➔ Latent topics help to alleviate such side effect

Latent Topic Analysis

- Topic coherence (C_V scores)

Datasets	Twitter	StackExchange
LDA	41.12	35.13
BTM	43.12	43.52
NTM	43.82	43.04
Our model	**46.28**	**45.12**

- Top words for "super bowl" topic

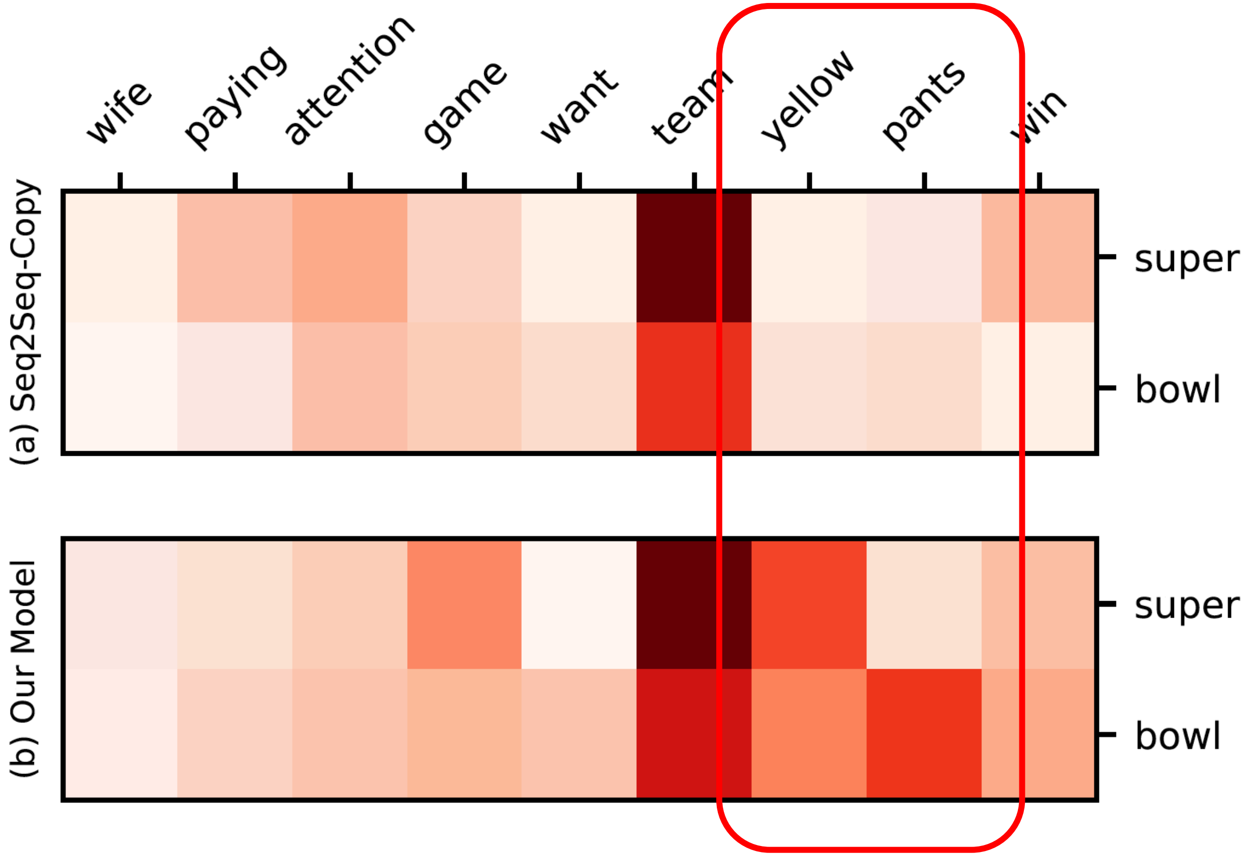
LDA	bowl super quote steeler jan watching egypt playing glee girl
BTM	bowl super anthem national christina aguilera fail word brand playing
NTM	super bowl eye protester winning watch halftime ship sport mena
Our model	bowl super yellow green packer steeler nom commercial win winner

Red and underlined words indicate non-topic words

Case Study

Somewhere, a wife that is not paying attention to the game, says "I want the team in yellow pants to win."

Our model correctly predicts *"super bowl"*, while seq2seq-copy without topic guidance wrongly predicts *"team follow back"*



Why? Visualize attention!

1^{st} Topic	steeler national team packer win

Summary

- We propose the first topic-aware keyphrase generation model that allows end-to-end training with latent topics
- We newly construct three large-scale social media datasets in both English and Chinese for this task
- Extensive experiments demonstrate the effectiveness of our proposed model for understanding social media language

(96 stars)

https://github.com/yuewang-cuhk/TAKG

Outline

- Topic 1: Topic-aware Keyphrase Generation
- Topic 2: Conversation-aware Keyphrase Generation
- Topic 3: Unified Cross-media Keyphrase Prediction
- Conclusion and Future Work

Motivation

Example

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

[R1] How annoying is she. I just worked out what she sounds like one of those turbo charged cars when they change gear or speed.

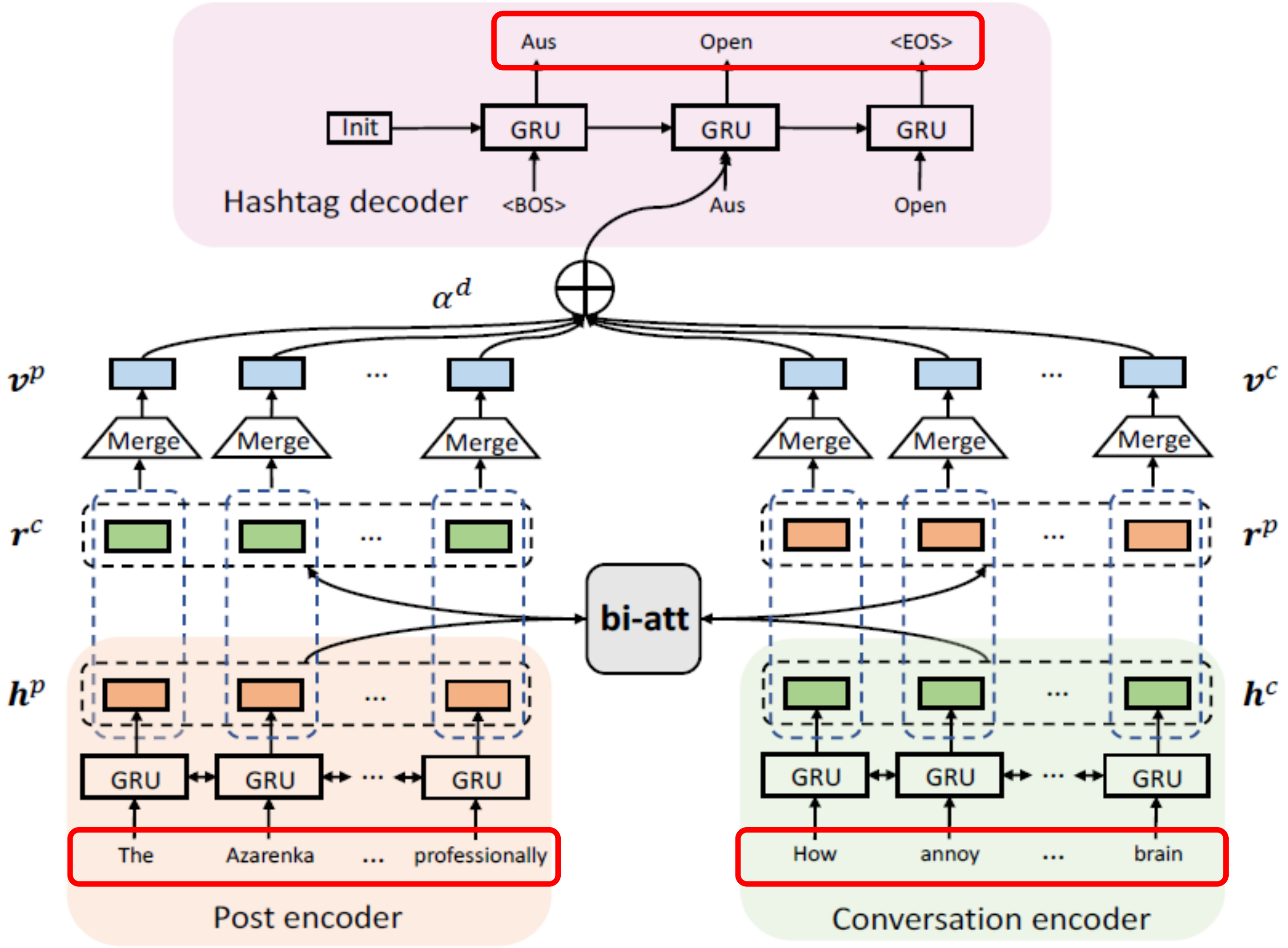
[R2] On the topic of noises, I was at the ***Nadal-Tomic*** game last night and I loved how quiet ***Tomic*** was compared to ***Nadal***.

[R3] He seems to have a shitload of talent and the ***postmatch*** press conf. He showed a lot of maturity and he seems nice.

[R4] ***Tomic*** has a fantastic ***tennis*** brain...

- From the **user conversation**, we can imply its keyphrase: ***AusOpen***

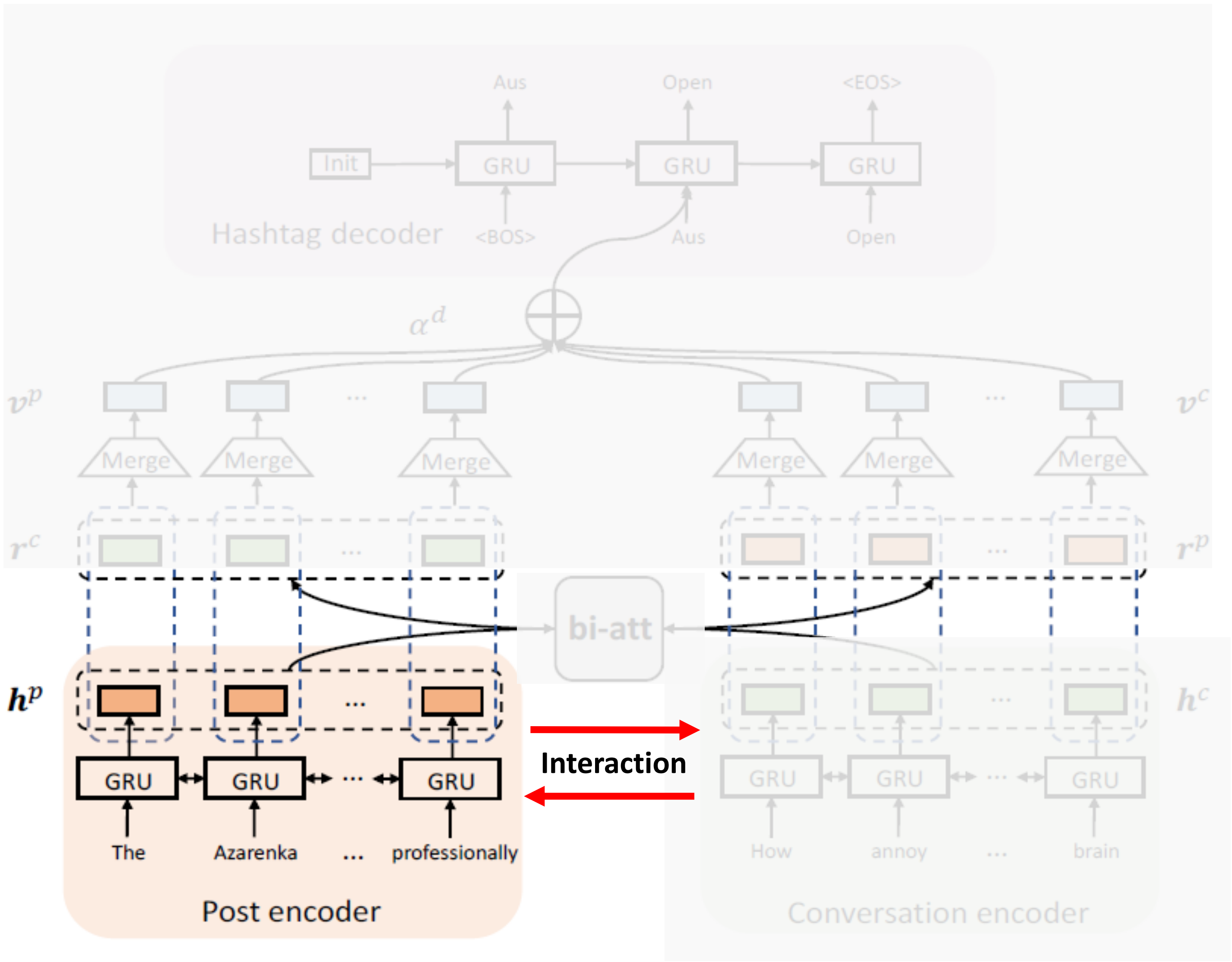
Methodology



- **Input**
  - Target post: $\langle x_1^p, x_2^p, \ldots, x_{|\boldsymbol{x}^p|}^p \rangle$
  - Conversation: $\langle x_1^c, x_2^c, \ldots, x_{|\boldsymbol{x}^c|}^c \rangle$
    - *Combine user replies sequentially*

- **Output**
  - Keyphrase: $\langle y_1, y_2, \ldots, y_{|y|} \rangle$
  - "**A**us**O**pen"→ "**A**us **O**pen"

Methodology



Post encoder

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

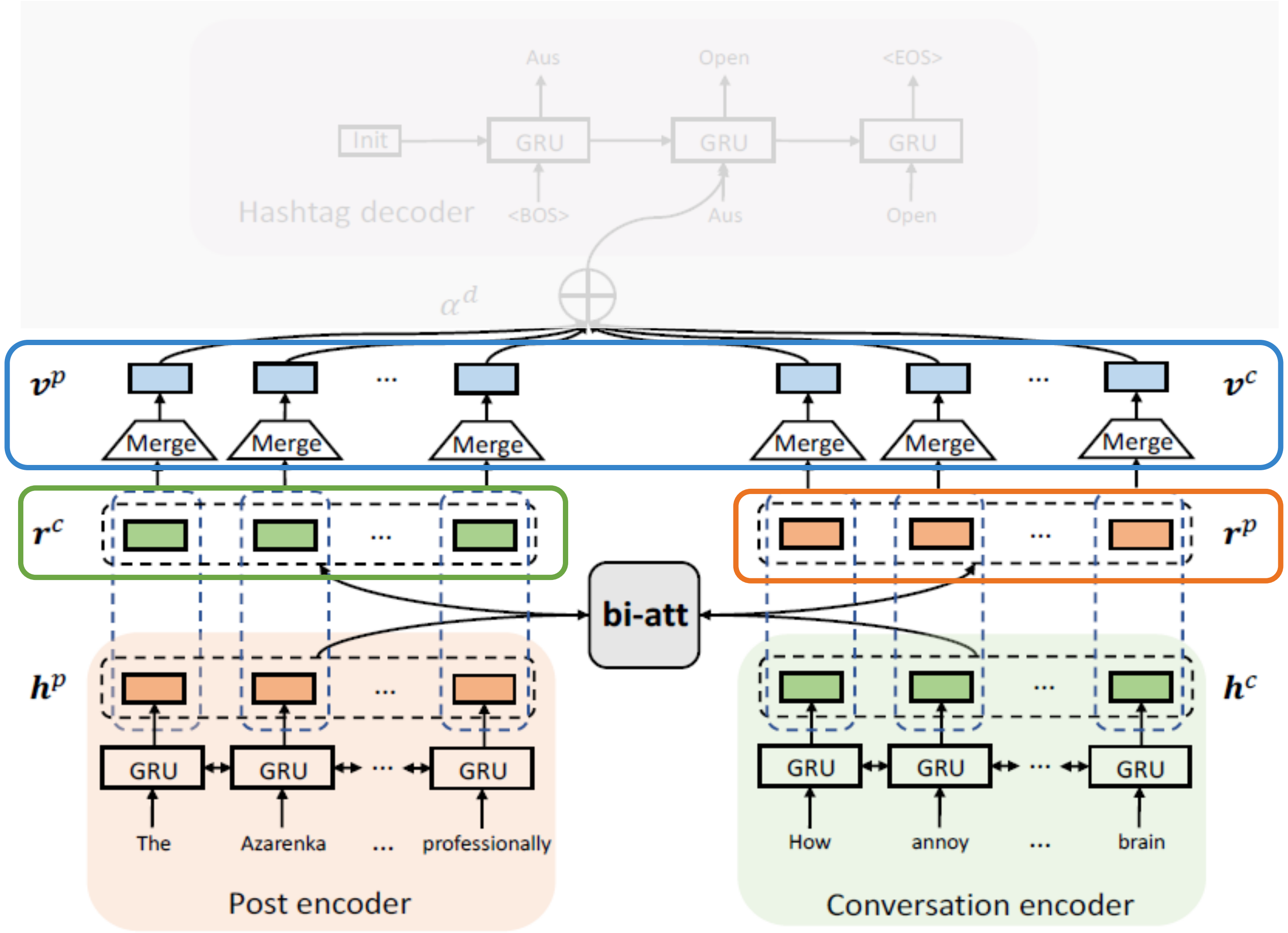
Conversation encoder

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

Bi-attention (bi-att)

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

Methodology



Conversation-attentive vector

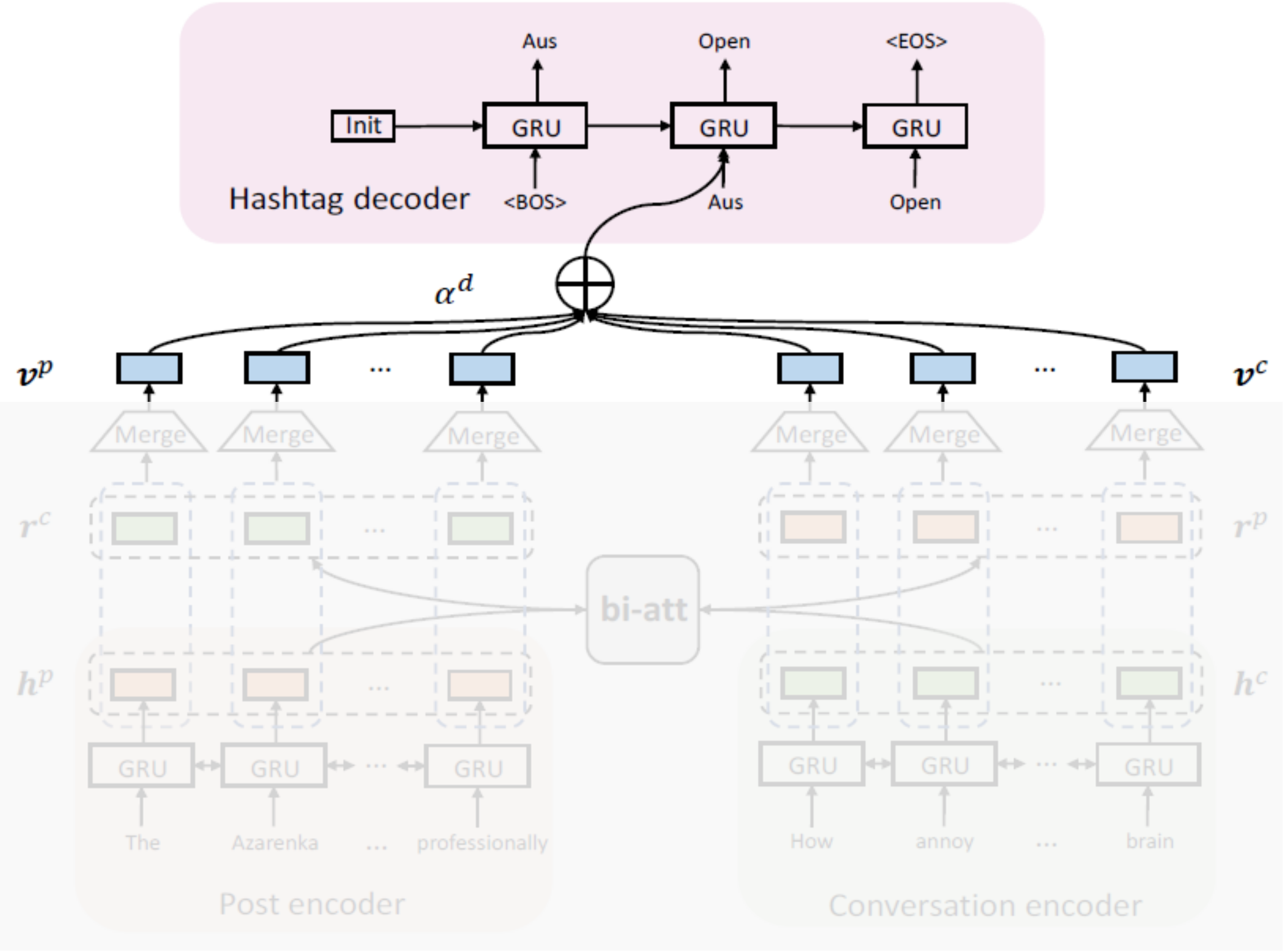
- $r_i^c = \sum_{j=1}^{|x^c|} \alpha_{ij}^c \, h_j^c$

Post-attentive vector

- $r_j^p = \sum_{i=1}^{|x^p|} \alpha_{ij}^p \, h_i^p$

Merge layer

- $v^p = \tanh(W_p[h^p; r^c] + b_p)$,
- $v^c = \tanh(W_c[h^c; r^p] + b_c)$,
- $v = [v^p; v^c]$,



Keyphrase 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

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

Inference: beam search

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

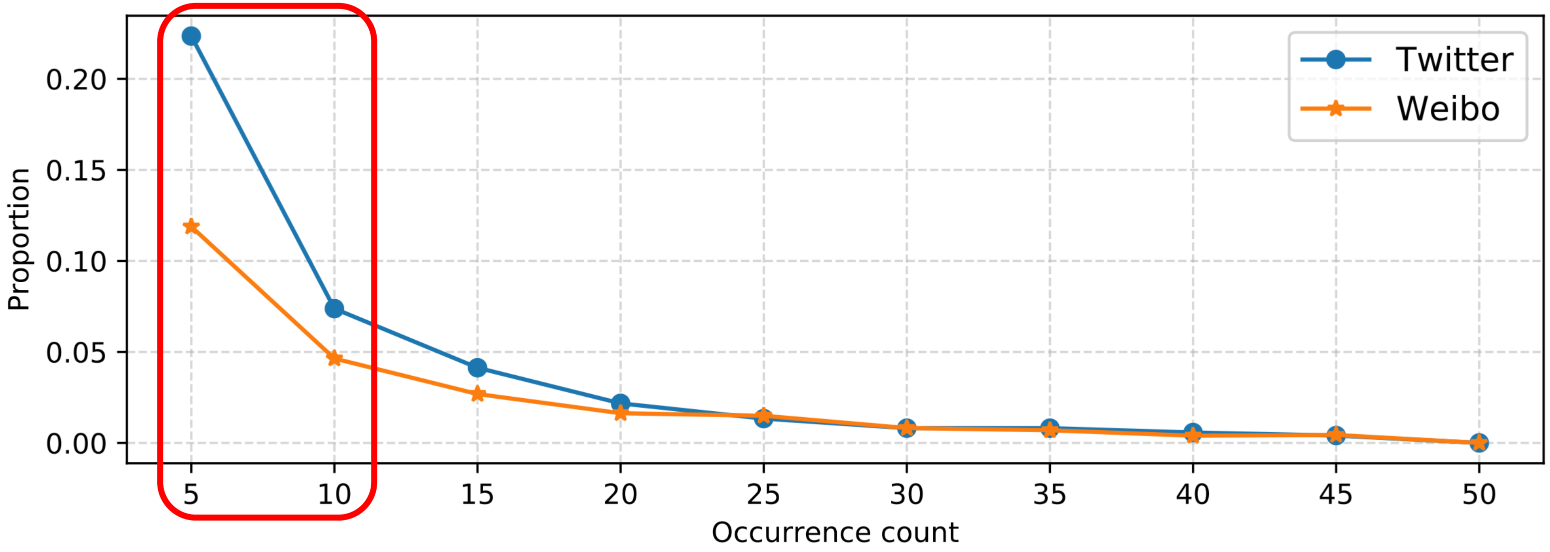
- Keyphrase statistics (present ratio)

Datasets	$\lvert$Tagset$\rvert$	$\mathcal{P}$	$\mathcal{C}$	$\mathcal{P} \cup \mathcal{C}$
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

- Keyphrase frequency distribution



Large and **imbalanced** keyphrase space!

Main Results

Exact match Partial match

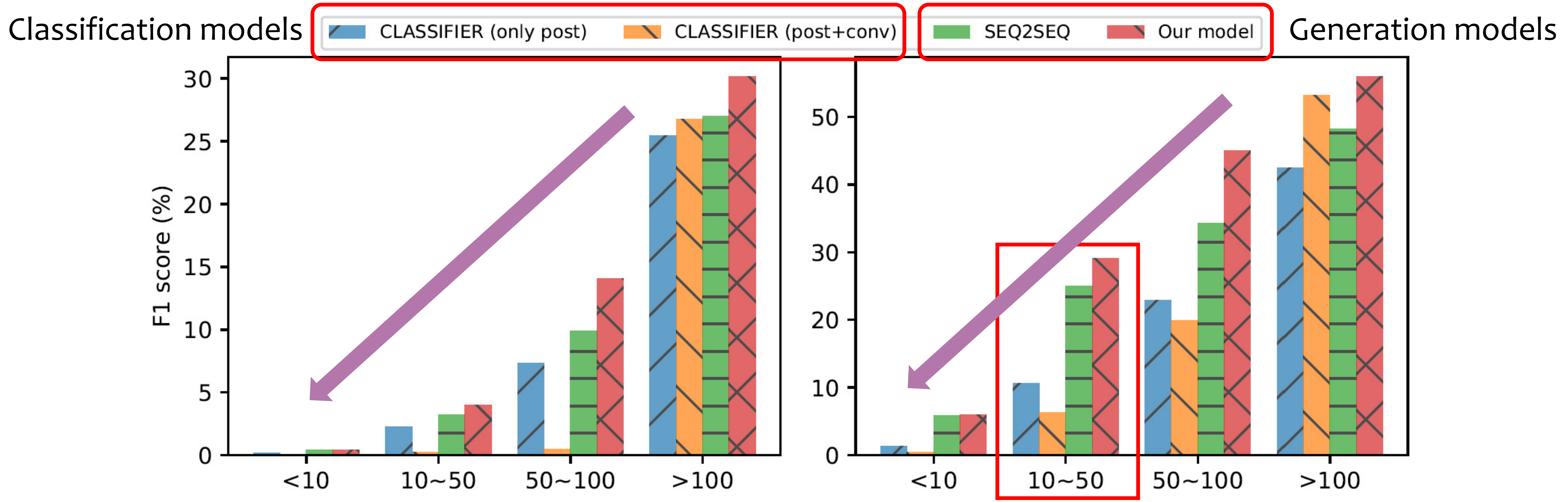
Model	Twitter F1@1	Twitter F1@5	Twitter MAP	Twitter RG-1	Twitter RG-4	Weibo F1@1	Weibo F1@5	Weibo MAP	Weibo RG-1	Weibo RG-4
Baselines										
RANDOM	0.37	0.63	0.89	0.56	0.16	0.43	0.67	0.97	2.14	1.13
LDA	0.13	0.25	0.35	0.60	-	0.10	0.86	0.94	3.89	-
TF-IDF	0.02	0.02	0.03	0.54	0.14	0.85	0.73	1.30	8.04	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
- Our model significantly outperforms all the comparison models
- Generation models are better than classification models

Classification vs. Generation



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

- The keyphrase 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 keyphrases (ROUGE-1 in %)

- It is difficult to generate new keyphrases
- 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
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
OUR MODEL (*full*)	**12.29**	**31.96**

w/o bi-att — SEQ2SEQ (*post + conv*)

w/ bi-att — OUR MODEL (*full*)

Post is more important!

Bi-attention is helpful!

Ablation results (F1 in %)

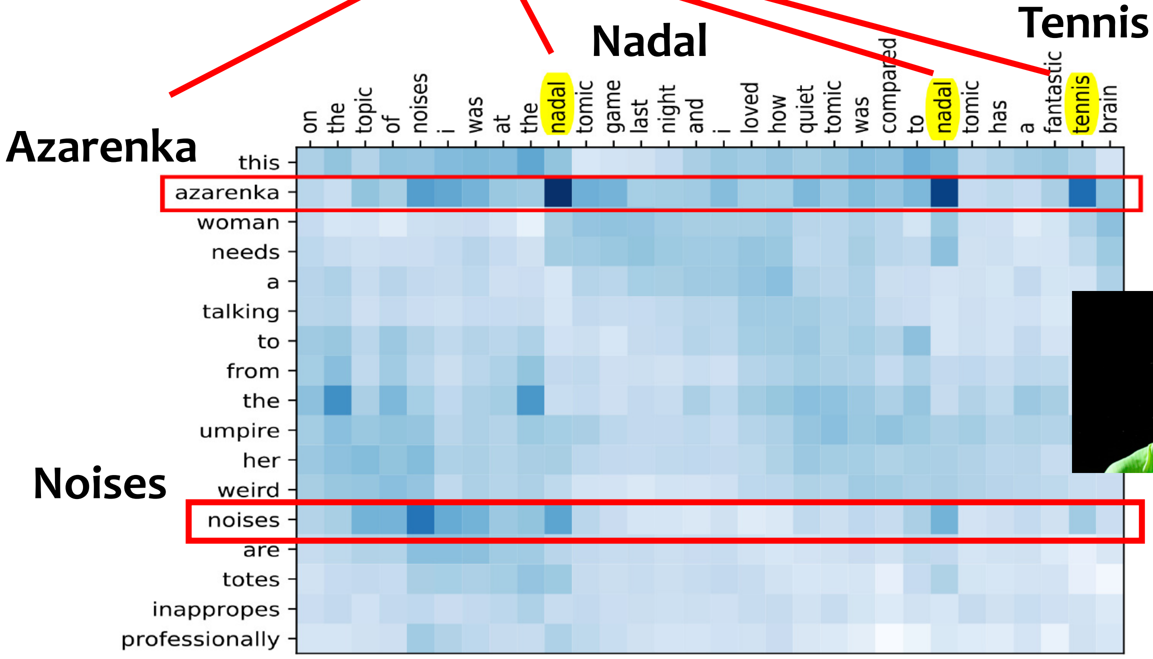
Case Study

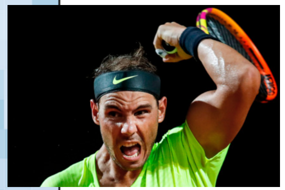
Case post

"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) Bi-attention heatmap visualization

Summary

- We are the first to approach microblog keyphrase 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 new state-of-the-art results on two datasets

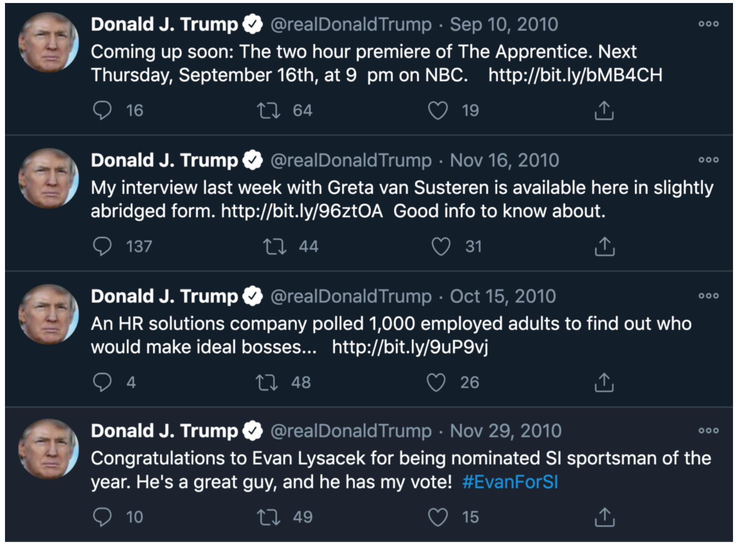
https://github.com/yuewang-cuhk/HashtagGeneration

Outline

- Topic 1: Topic-aware Keyphrase Generation
- Topic 2: Conversation-aware Keyphrase Generation
- Topic 3: Unified Cross-media Keyphrase Prediction
- Conclusion and Future Work

Motivation

- With the development of mobile Internet…



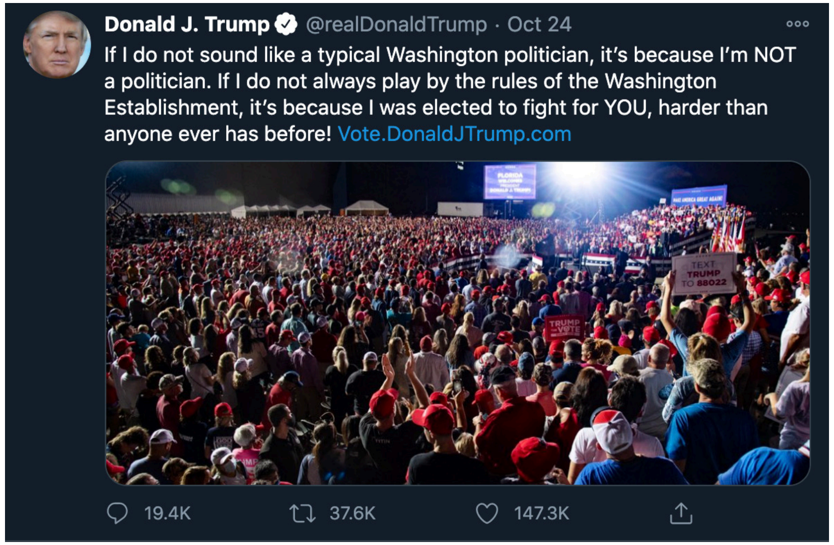
Donald J. Trump @realDonaldTrump · Sep 10, 2010
Coming up soon: The two hour premiere of The Apprentice. Next Thursday, September 16th, at 9 pm on NBC. http://bit.ly/bMB4CH
16 64 19

Donald J. Trump @realDonaldTrump · Nov 16, 2010
My interview last week with Greta van Susteren is available here in slightly abridged form. http://bit.ly/96ztOA Good info to know about.
137 44 31

Donald J. Trump @realDonaldTrump · Oct 15, 2010
An HR solutions company polled 1,000 employed adults to find out who would make ideal bosses... http://bit.ly/9uP9vj
4 48 26

Donald J. Trump @realDonaldTrump · Nov 29, 2010
Congratulations to Evan Lysacek for being nominated SI sportsman of the year. He's a great guy, and he has my vote! #EvanForSI
10 49 15

2010



Donald J. Trump @realDonaldTrump · Oct 24
If I do not sound like a typical Washington politician, it's because I'm NOT a politician. If I do not always play by the rules of the Washington Establishment, it's because I was elected to fight for YOU, harder than anyone ever has before! Vote.DonaldJTrump.com
19.4K 37.6K 147.3K

2020

Motivation

- How to predict keyphrases for cross-media posts?

- Limited text features

Motivation

- How to predict keyphrases for cross-media posts?

- Limited text features

Image could provide essential clues!

Challenge

- Unique challenges compared to conventional multi-modal tasks

Semantics shared in both modalities

Caption: a man talking to a giraffe in an enclosure

Q: what color is the giraffe?
A: brown and tan

Complex text-image relationship

?

Tweet: Contemplating the mysteries of life from inside my egg carton...

Challenge

- Complex text-image relationship in social media
  - Four diverse semantic relations [Vempala and Preotiuc-Pietro, ACL 2019]

Post (a): Sharing is caring. Good girl Kit, cause I know how much you love your bed. *#Dogs #Kindness*

Post (b): Waves crash against the North Pier this evening at Tynemouth, River Tyne in the UK @david1hirst *#StormHour*

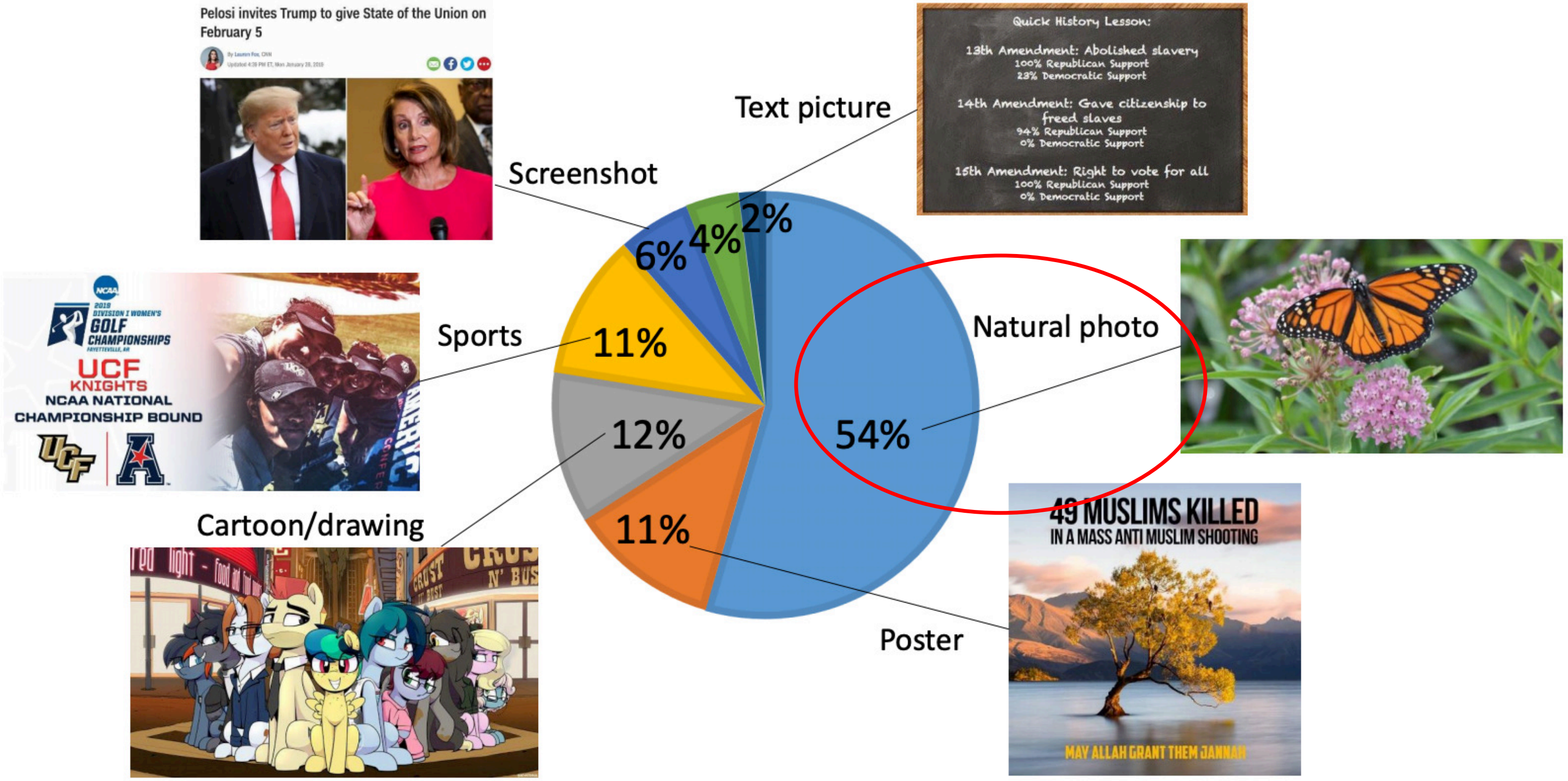
Post (c): "I am declaring an emergency that only i can fix" *#BoycottTrumpPrimeTime*

Post (d): The whole of the uk when armadillo and danny say anything *#LoveIsland*

(a) text is represented and image adds to. (b) text is represented and image does not add to.
(c) text is not represented and image adds to. (d): text is not represented and image does not add to.

Challenge

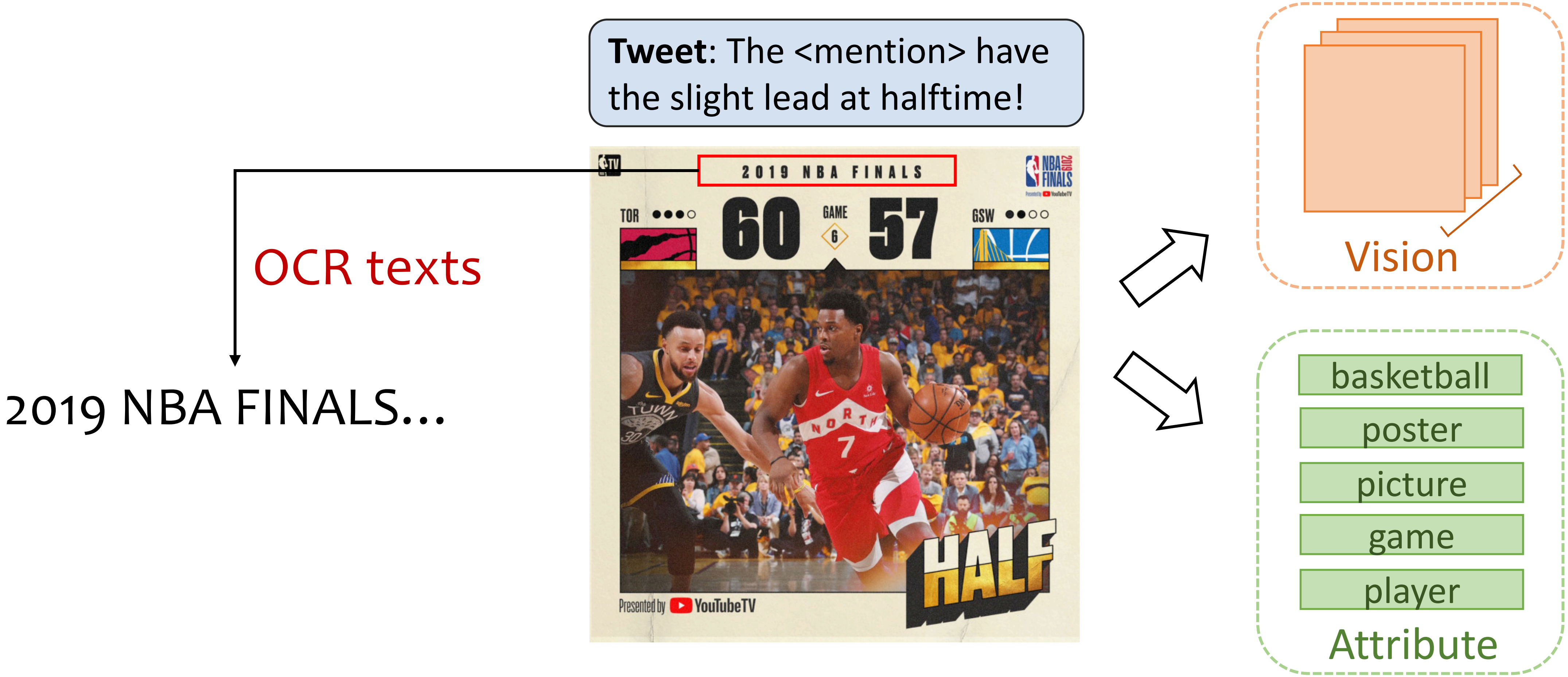
- Diverse image category
  - Category distribution of 200 tweet image samples



Many images contain texts!

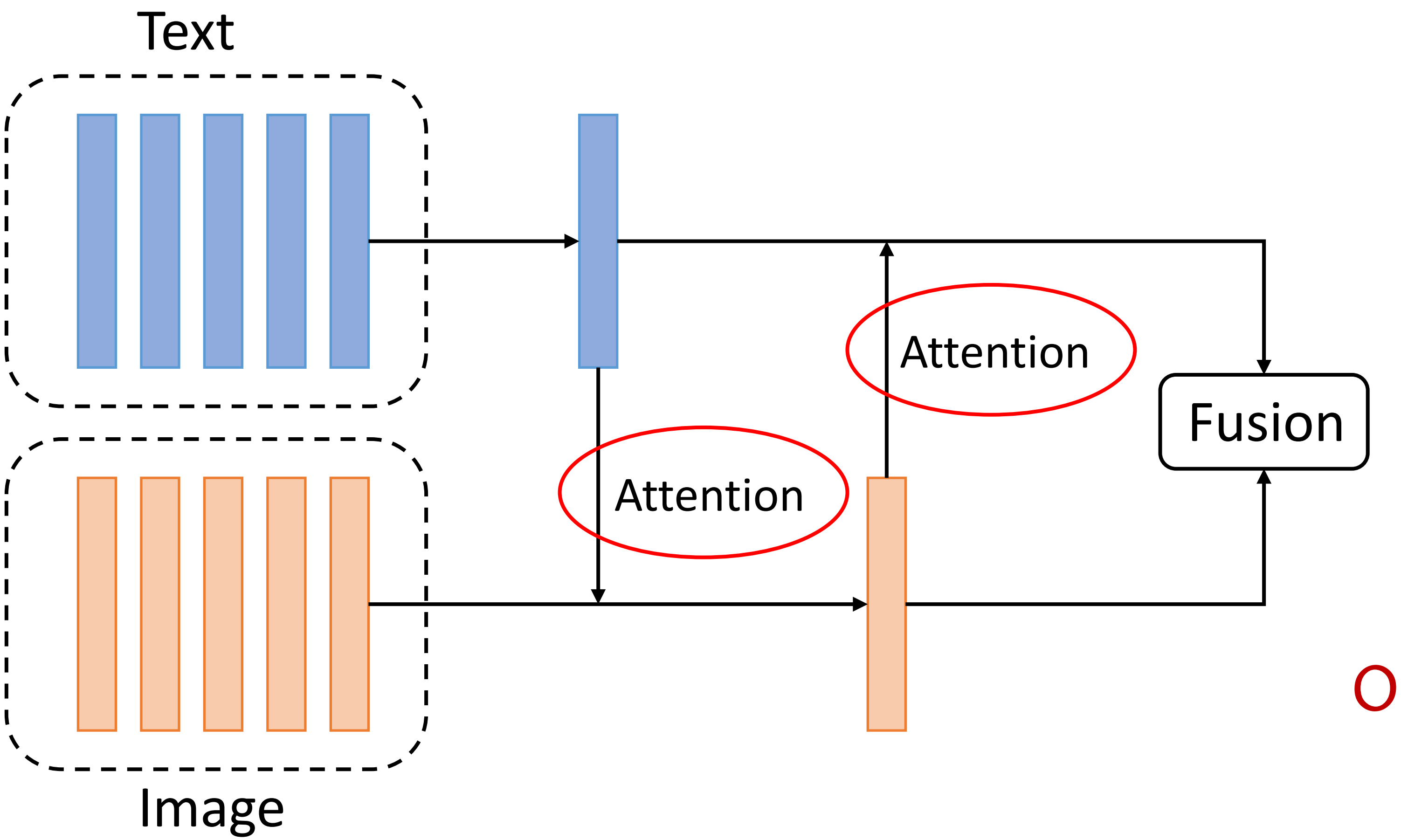
Our Solution

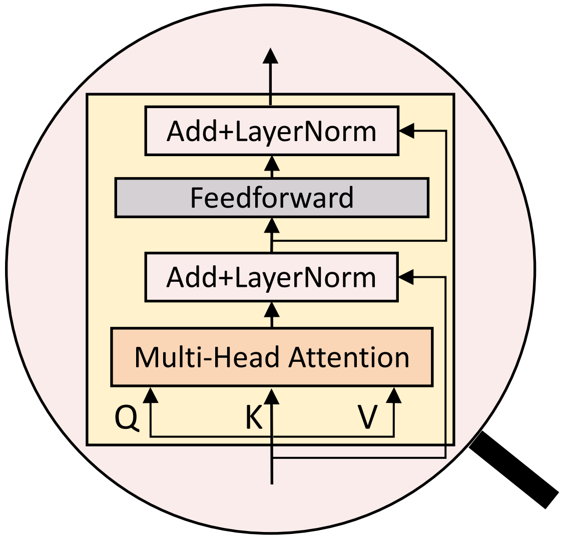
- Encode more indicative features from the images
  - Image wordings: *image attributes* and OCR (Optical Character Recognition) *texts*



Our Solution

- Better attention mechanism to model complex text-image interactions
  - Traditional co-attention network is suboptimal [Zhang et al., IJCAI 2017]





Ours: multi-head attention

Our Solution

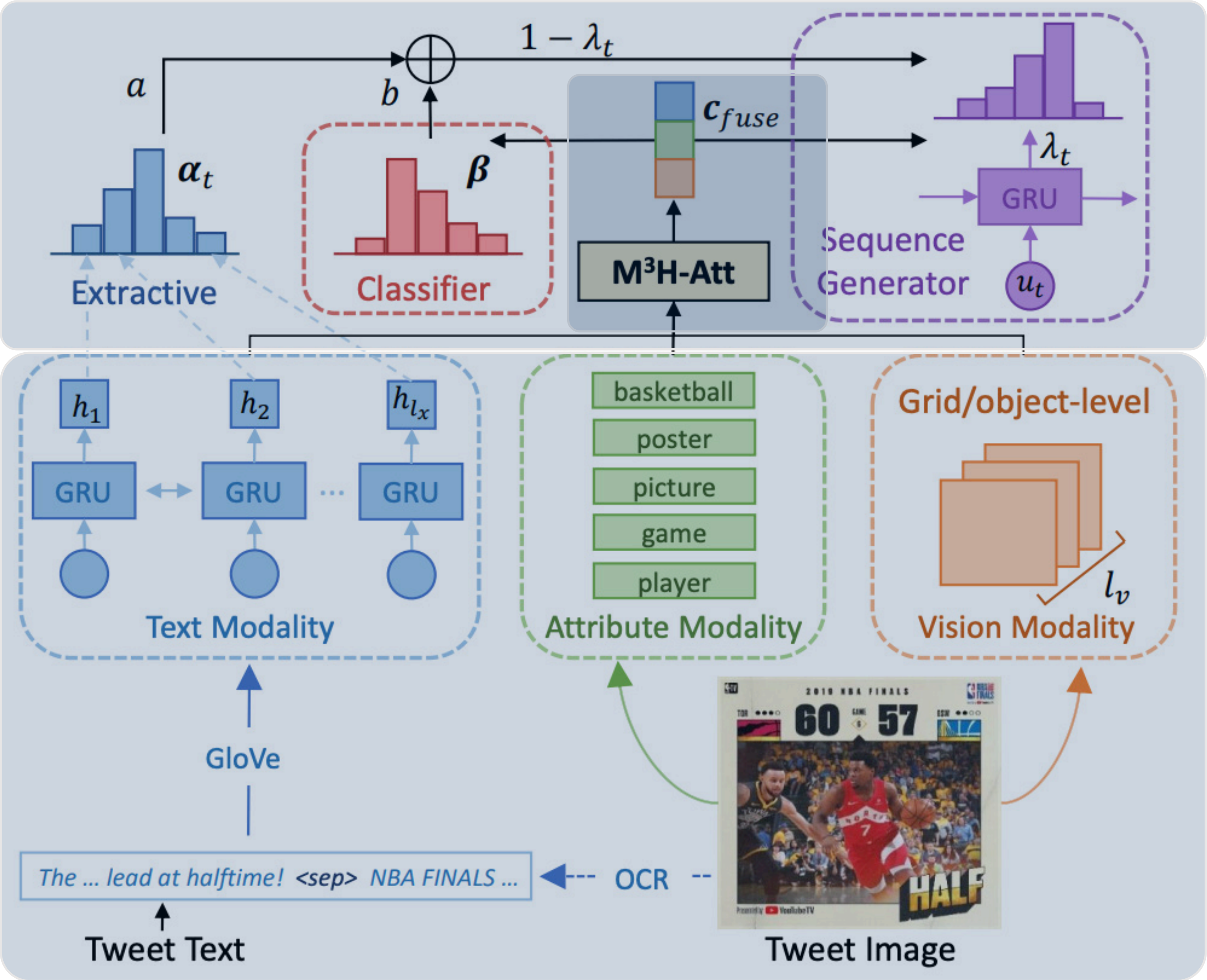
- Previous methods
  - Keyphrase classification for text-image posts
    - [Zhang et al., IJCAI 2017] and [Zhang et al., AAAI 2019]
    - Cannot produce keyphrases out of the predefined candidate list
  - Keyphrase generation for text-only posts
    - [Wang et al., NAACL 2019] and [Wang et al., ACL 2019]
    - Poor performance in predicting absent keyphrases

A unified model to combine both

Methodology

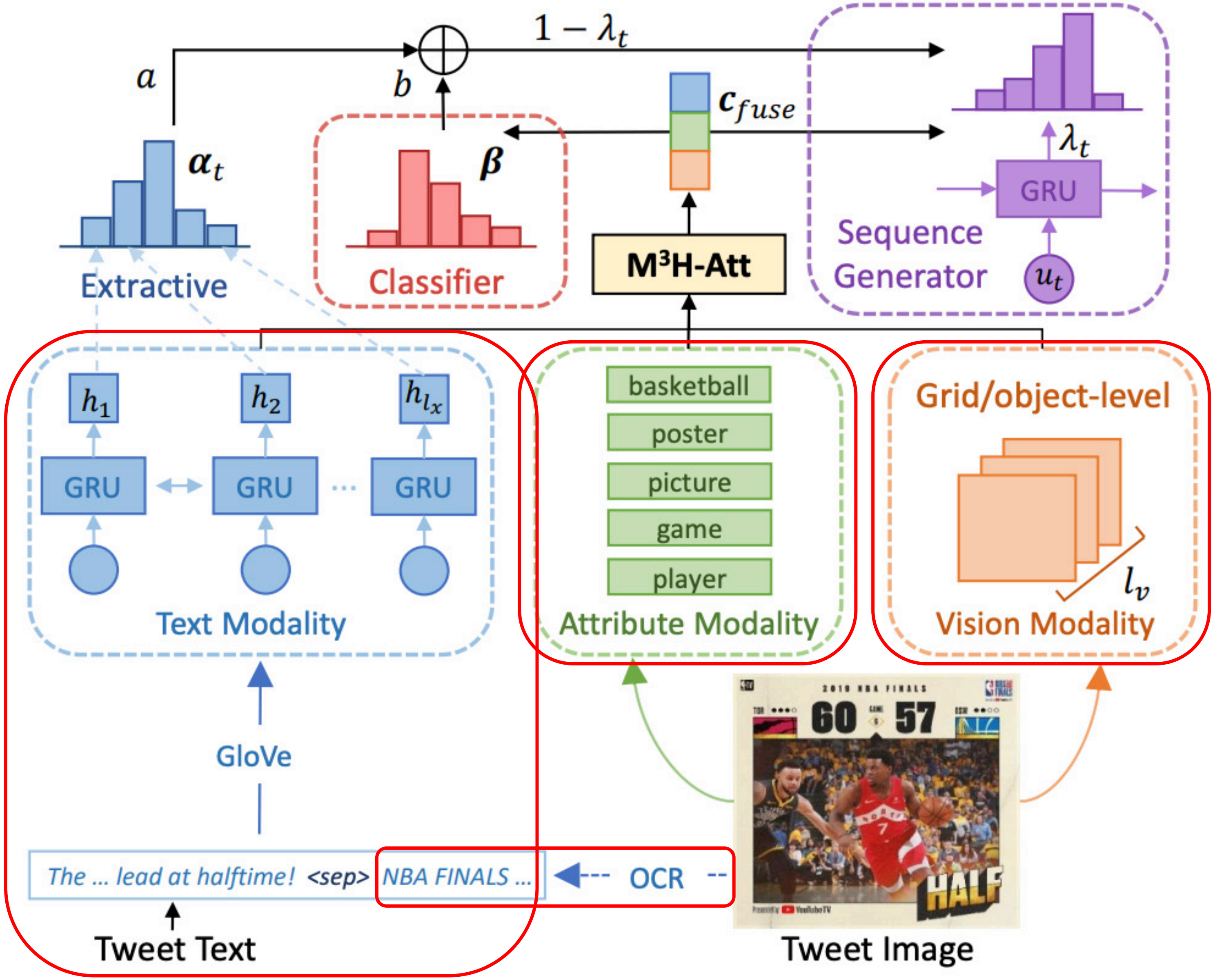
- Input
  - Image I
  - Target post: $\langle x_1, \ldots, x_{l_x} \rangle$
- Output
  - Keyphrase: $\langle y_1, \ldots, y_{l_y} \rangle$
  - “NBAFINALS”→ “NBA FINALS”

- Encoding text and image
- Multi-modal fusion
- Unified prediction



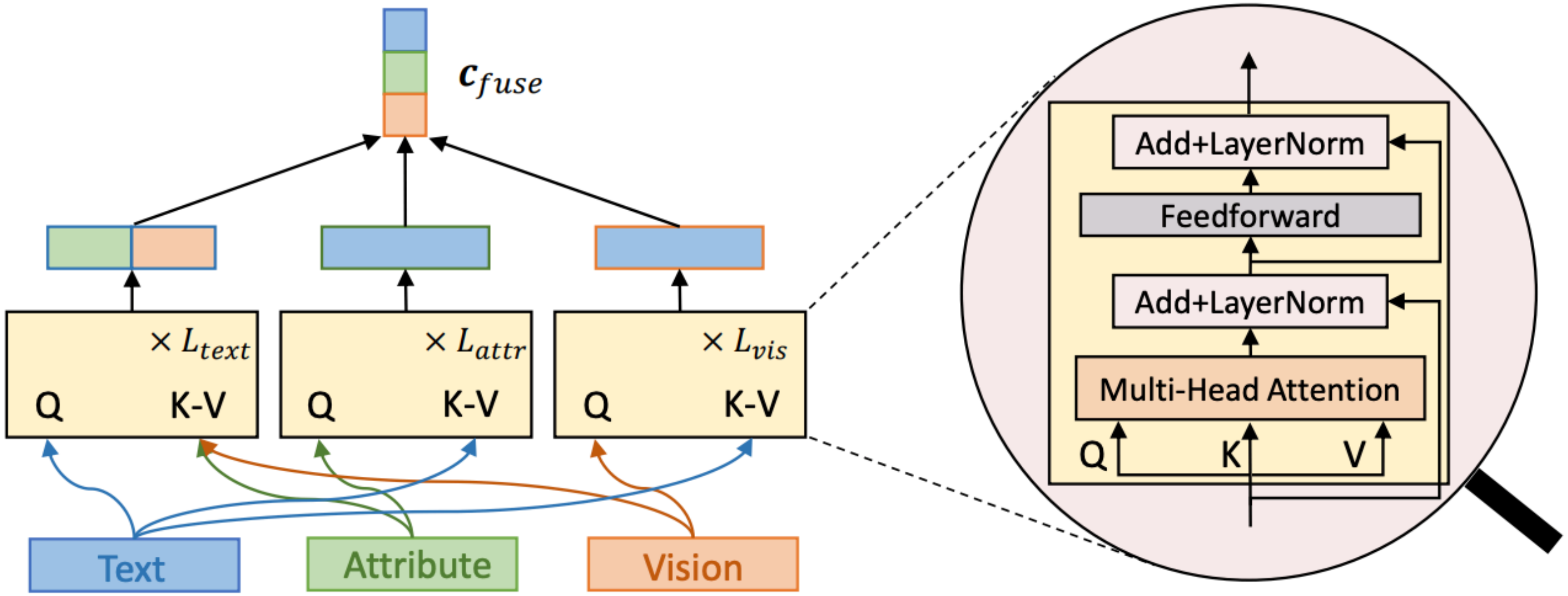
Encoding Text and Image

- Textual features
  - Bi-GRU encoder
- Visual features
  - Grid-level or object-level
- Image attributes
  - Pretrained attribute predictor using COCO-caption data
- OCR texts
  - Detected from Tesserocr
  - Append to the tweet text



Multi-modal Fusion

- Multi-Modality Multi-Head Attention (M3H-Att)
  - Capture the interactions among three modalities: {text, attribute, vision}



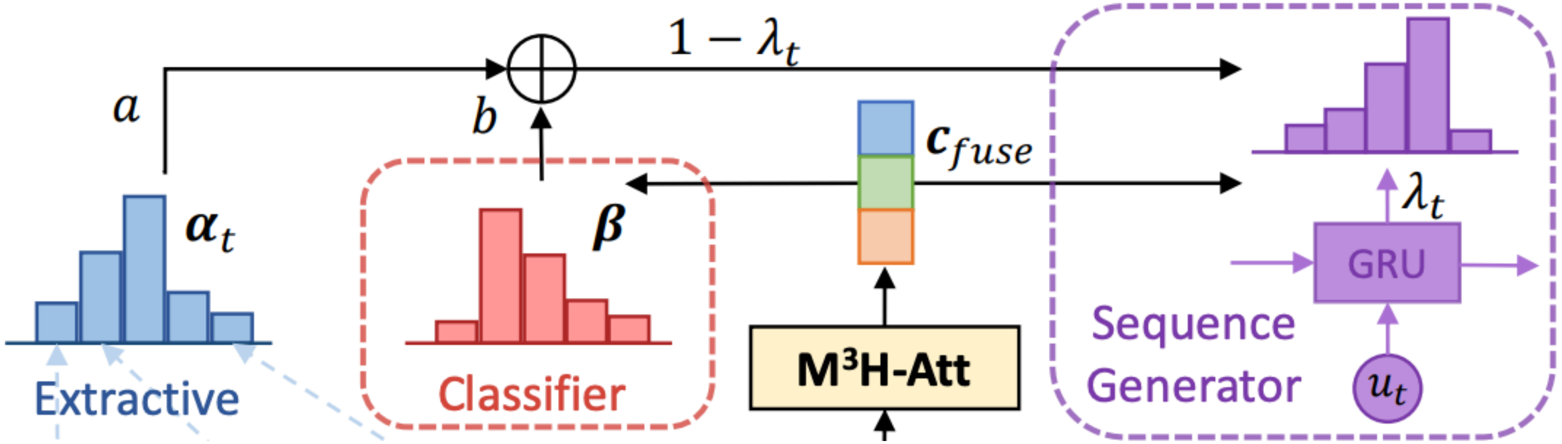
$$\mathcal{A}(Q,K,V) = \mathtt{softmax}(\frac{QK^T}{\sqrt{d_K}})V,$$

$$\mathcal{A}^M(Q,K,V) = [head_1; \ldots; head_H]W^O,$$

$$\text{where} \quad head_h = \mathcal{A}(QW_h^Q, KW_h^K, VW_h^V)$$

Unified Prediction

- Combine keyphrase classification and generation



Classification output aggregator

$$P_{unf}(y_t) = \lambda_t \cdot P_{gen}(y_t) \ + \tag{11}$$

$$(1-\lambda_t)\cdot(a \cdot \sum_{i:x_i=y_t}^{l_x} \alpha_{t,i} + b \cdot \sum_{j:w_j=y_t}^{l_w} \beta_j), \tag{12}$$

Joint training

$$\mathcal{L}(\theta) = -\sum_{n=1}^{N}[\underbrace{\log P_{cls}(\mathbf{y}^n)}_{\text{Classification}} + \gamma \cdot \sum_{t=1}^{l_y^n} \underbrace{\log P_{unf}(y_t^n)}_{\text{Unified}}], \tag{13}$$

Dataset

- Experiment dataset: 53,701 text-image tweets from Twitter

Split	#Post	Post Len	#KP /Post	|KP|	KP Len	% of occ. KP	Vocab
Train	42,959	27.26	1.33	4,261	1.85	37.14	48,019
Val	5,370	26.81	1.34	2,544	1.85	36.01	16,892
Test	5,372	27.05	1.32	2,534	1.86	37.45	17,021

Table 1: Data split statistics. KP: keyphrase; |KP|: the size of unique keyphrase; % of occ. KP: percentage of keyphrases occurring in the source post.

Low present rate!

Dataset

- Top five image attributes: {*man, shirt, woman, sign, white*}

Word cloud visualization

Main Results

- Observations
  - Textual features are more important than visual signals

	Models	F1@1	F1@3	MAP@5
	EXT-ORACLE	39.50	23.20	39.26
Image-only	CLS-VGG-MAX	14.20_{35}	12.20_{24}	17.68_{31}
	CLS-VGG-AVG	15.69_{21}	13.67_{06}	19.70_{20}
	CLS-BUTD-MAX	17.65_{32}	15.00_{21}	21.77_{29}
	CLS-BUTD-AVG	20.02_{27}	16.97_{06}	24.73_{11}
Text-only	CLS-AVG	35.96_{11}	27.59_{05}	41.84_{14}
	CLS-MAX	38.33_{47}	28.84_{09}	44.15_{34}
	CLS-TMN	40.33_{39}	30.07_{28}	46.28_{27}
	GEN-ATT	38.36_{28}	27.83_{15}	43.35_{20}
	GEN-COPY	42.10_{19}	29.91_{30}	46.94_{35}
	GEN-TOPIC	43.17_{24}	30.73_{13}	48.07_{23}
Text-Image	CLS-BAN	38.73_{18}	29.68_{23}	45.03_{15}
	CLS-IMG-ATT	41.48_{33}	31.22_{14}	47.93_{34}
	CLS-CO-ATT	42.12_{38}	31.55_{33}	48.39_{34}
	CLS-M^3H-ATT (ours)	44.11_{17}	31.47_{14}	49.45_{11}
	+ image wording	44.46_{12}	32.82_{24}	50.39_{15}
	+ joint-train	45.16_{09}	33.27_{10}	51.48_{11}
	GEN-M^3H-ATT (ours)	44.25_{05}	31.58_{13}	49.35_{10}
	+ image wording	44.56_{09}	31.77_{23}	49.95_{22}
	+ joint-train	45.69_{17}	32.78_{09}	51.37_{12}
	GEN-CLS-M^3H-ATT (ours)	**47.06_{04}**	**33.11_{01}**	**52.07_{03}**

Average scores from 5 random seeds. Subscripts denote the standard deviation, e.g., 47.06_{04} denotes 47.06 ± 0.04

Main Results

- Observations
  - Textual features are more important than visual signals
  - Vision can provide complementary information to the text

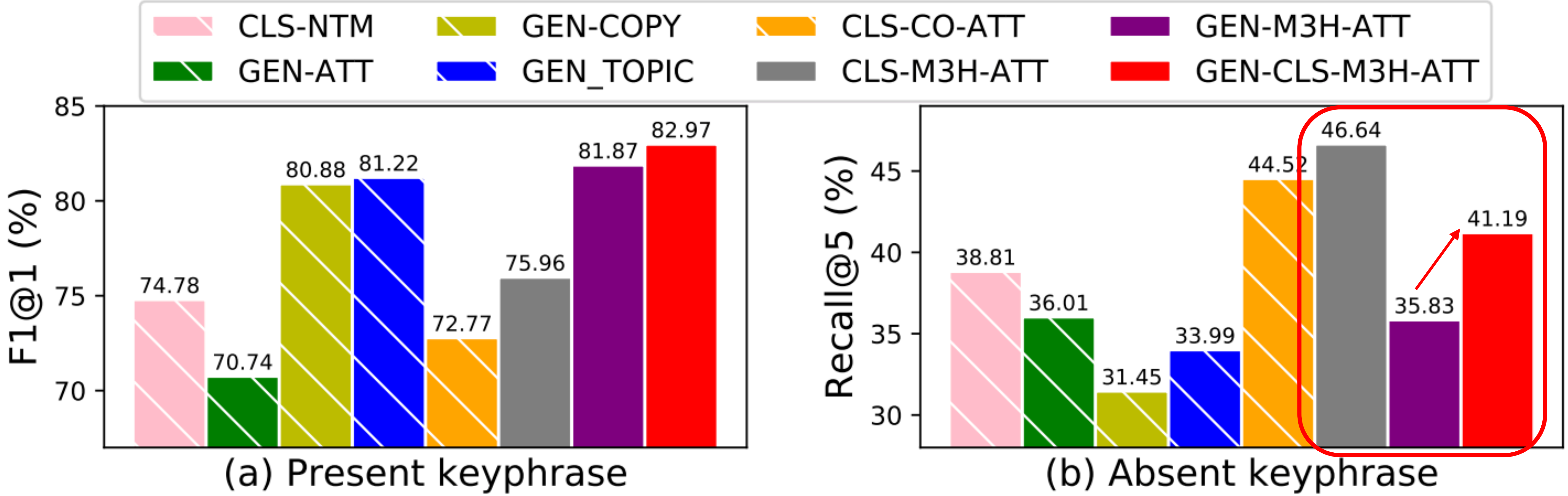
	Models	F1@1	F1@3	MAP@5
	EXT-ORACLE	39.50	23.20	39.26
Image-only	CLS-VGG-MAX	14.20_{35}	12.20_{24}	17.68_{31}
	CLS-VGG-AVG	15.69_{21}	13.67_{06}	19.70_{20}
	CLS-BUTD-MAX	17.65_{32}	15.00_{21}	21.77_{29}
	CLS-BUTD-AVG	20.02_{27}	16.97_{06}	24.73_{11}
Text-only	CLS-AVG	35.96_{11}	27.59_{05}	41.84_{14}
	CLS-MAX	38.33_{47}	28.84_{09}	44.15_{34}
	CLS-TMN	40.33_{39}	30.07_{28}	46.28_{27}
	GEN-ATT	38.36_{28}	27.83_{15}	43.35_{20}
	GEN-COPY	42.10_{19}	29.91_{30}	46.94_{35}
	GEN-TOPIC	43.17_{24}	30.73_{13}	48.07_{23}
Text-Image	CLS-BAN	38.73_{18}	29.68_{23}	45.03_{15}
	CLS-IMG-ATT	41.48_{33}	31.22_{14}	47.93_{34}
	CLS-CO-ATT	42.12_{38}	31.55_{33}	48.39_{34}
	CLS-M^{3}H-ATT (ours)	44.11_{17}	31.47_{14}	49.45_{11}
	+ image wording	44.46_{12}	32.82_{24}	50.39_{15}
	+ joint-train	45.16_{09}	33.27_{10}	51.48_{11}
	GEN-M^{3}H-ATT (ours)	44.25_{05}	31.58_{13}	49.35_{10}
	+ image wording	44.56_{09}	31.77_{23}	49.95_{22}
	+ joint-train	45.69_{17}	32.78_{09}	51.37_{12}
	GEN-CLS-M^{3}H-ATT (ours)	$\mathbf{47.06}_{04}$	$\mathbf{33.11}_{01}$	$\mathbf{52.07}_{03}$

Main Results

- Observations
  - Textual features are more important than visual signals
  - Vision can provide complementary information to the text
  - Our unified model M3H-Att and image wordings achieves the best results

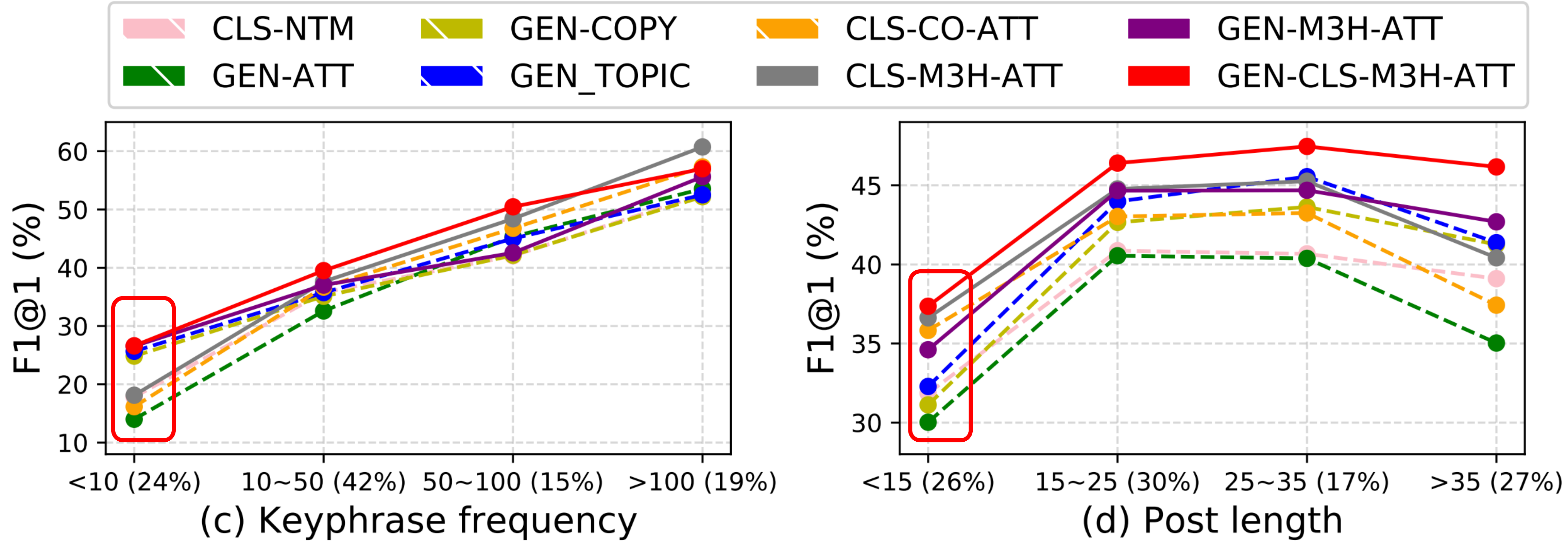
	Models	F1@1	F1@3	MAP@5
	EXT-ORACLE	39.50	23.20	39.26
Image-only	CLS-VGG-MAX	14.20_{35}	12.20_{24}	17.68_{31}
	CLS-VGG-AVG	15.69_{21}	13.67_{06}	19.70_{20}
	CLS-BUTD-MAX	17.65_{32}	15.00_{21}	21.77_{29}
	CLS-BUTD-AVG	20.02_{27}	16.97_{06}	24.73_{11}
Text-only	CLS-AVG	35.96_{11}	27.59_{05}	41.84_{14}
	CLS-MAX	38.33_{47}	28.84_{09}	44.15_{34}
	CLS-TMN	40.33_{39}	30.07_{28}	46.28_{27}
	GEN-ATT	38.36_{28}	27.83_{15}	43.35_{20}
	GEN-COPY	42.10_{19}	29.91_{30}	46.94_{35}
	GEN-TOPIC	43.17_{24}	30.73_{13}	48.07_{23}
Text-Image	CLS-BAN	38.73_{18}	29.68_{23}	45.03_{15}
	CLS-IMG-ATT	41.48_{33}	31.22_{14}	47.93_{34}
	CLS-CO-ATT	42.12_{38}	31.55_{33}	48.39_{34}
	CLS-M^{3}H-ATT (ours)	44.11_{17}	31.47_{14}	49.45_{11}
	+ image wording	44.46_{12}	32.82_{24}	50.39_{15}
	+ joint-train	45.16_{09}	33.27_{10}	51.48_{11}
	GEN-M^{3}H-ATT (ours)	44.25_{05}	31.58_{13}	49.35_{10}
	+ image wording	44.56_{09}	31.77_{23}	49.95_{22}
	+ joint-train	45.69_{17}	32.78_{09}	51.37_{12}
	GEN-CLS-M^{3}H-ATT (ours)	$\mathbf{47.06}_{04}$	$\mathbf{33.11}_{01}$	$\mathbf{52.07}_{03}$

Present and Absent Keyphrase



- Generation models are better for present keyphrases while classification models are better for absent ones
- Our output aggregation strategy can cover generation models' weakness for absent keyphrases

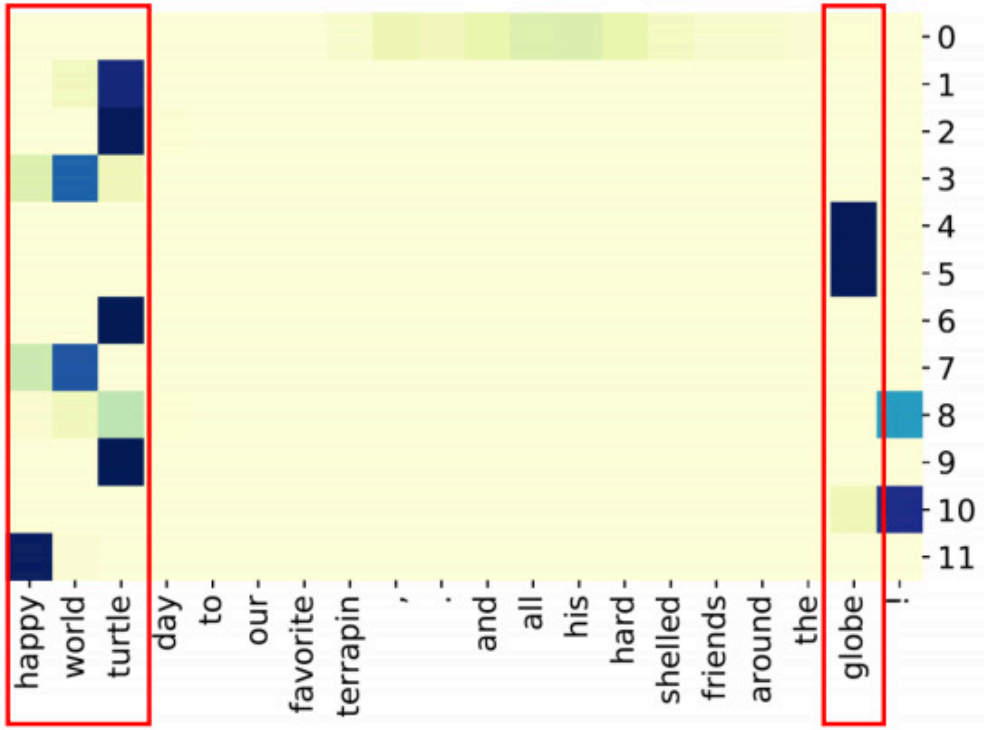
Keyphrase Frequency and Post Length



- Generation models with copy mechanism are better for predicting low-frequent keyphrases than classification models
- Image modality plays a more important role when texts contain limited features (<15 tokens)

What our model learns?

- Image-to-text attention visualization for all 12 heads



Happy world turtle **globe**

What our model learns?

- Text-to-image attention visualization

Text: The <mention> have the slight lead at halftime!

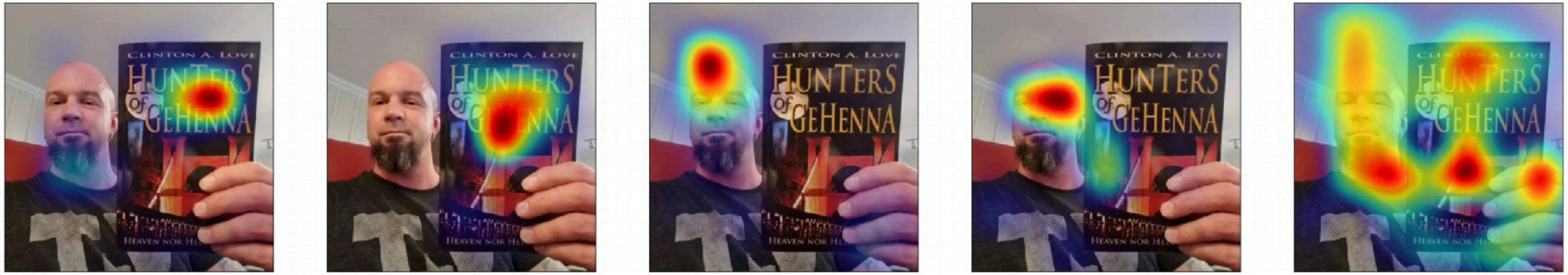
Head 0 | Head 5 | Head 9 | Head 11

Main objects: two players | Textual region | Global view

What our model learns?

- More examples for text-to-image attention

Post (c): Yeah! It' s here! There is nothing like holding your work in your own hand



Head 2 **Head 5** **Head 6** **Head 8** **Head 9**

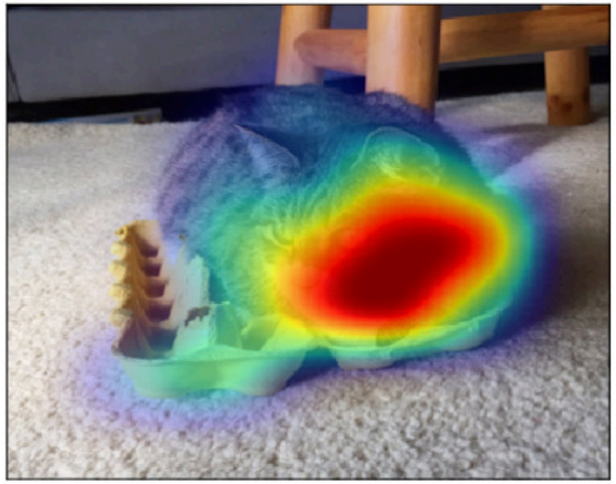
Post (e): So excited to hear her new song never really over every hour all day

Head 0 **Head 1** **Head 2** **Head 9** **Head 11**

What our model predicts?

- Blue tokens are the top four attributes and purple ones are OCR tokens

Post (a): Contemplating the mysteries of life from inside my egg carton...☺



(cat yellow grey bananas)
GEN-COPY: star wars
CLS-CO-ATT: **cats of twitter**
Our: **cats of twitter**

Post (b): Epic Texas *#sunset* from NNE Bastrop County TX. @TxStormChasers

(sky sun sunset field)
GEN-COPY: storm hour
CLS-CO-ATT: storm hour
Our: **sunset**

Post (c): Your plastic bag ends up somewhere, and sometimes, it goes to the ocean. *#WorldOceansDay*

(world oceans day June 8)
GEN-COPY: plastic fandom
CLS-CO-ATT: plastic
Our: **world oceans day**

Summary

- We design a novel *Multi-Modality Multi-Head Attention* (*M³H-Att*) to capture the complex text-image interaction for cross-media keyphrase prediction

- We propose to encode *image wordings* to bridge their semantic gap

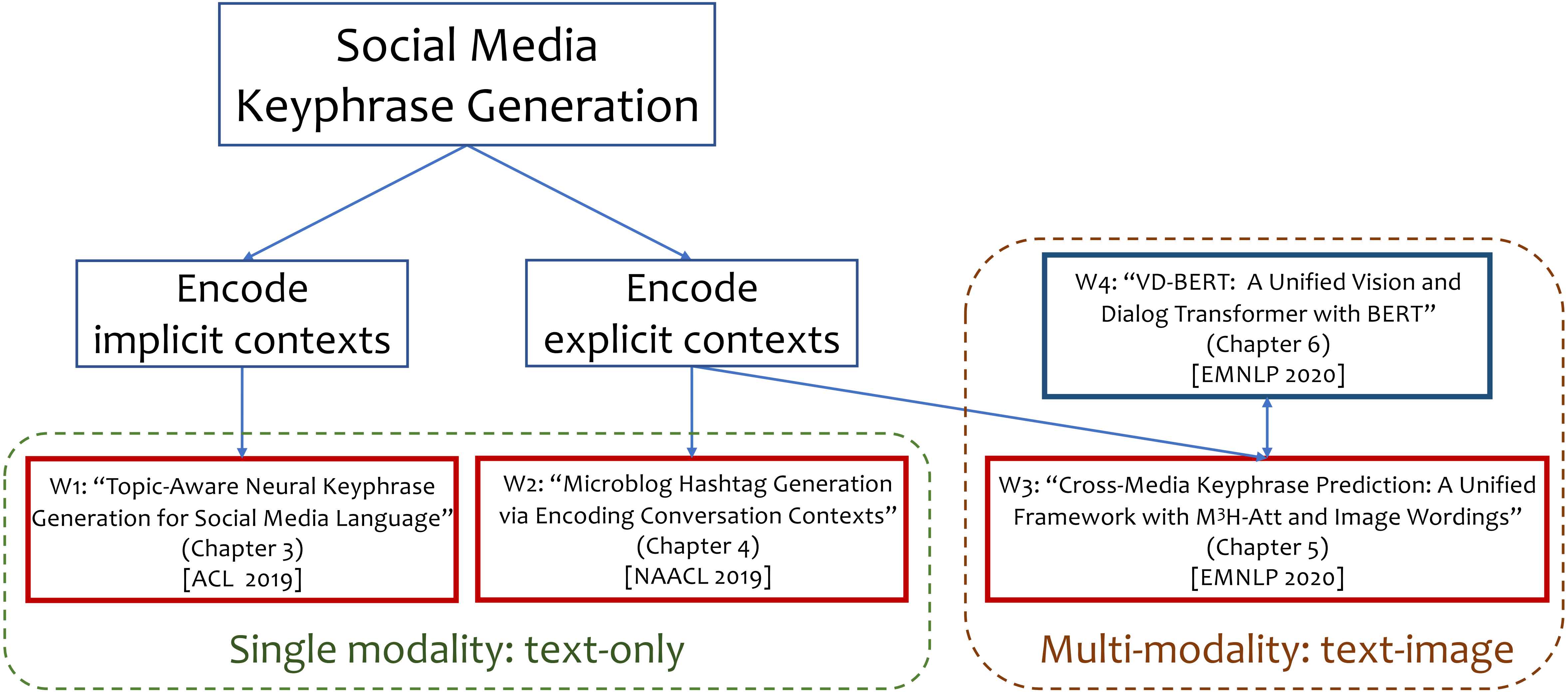
- We are the first to propose a *unified* framework coupling classification and generation models for better keyphrase prediction

https://github.com/yuewang-cuhk/CMKP

Outline

- Topic 1: Topic-aware Keyphrase Generation
- Topic 2: Conversation-aware Keyphrase Generation
- Topic 3: Unified Cross-media Keyphrase Prediction
- Conclusion and Future Work

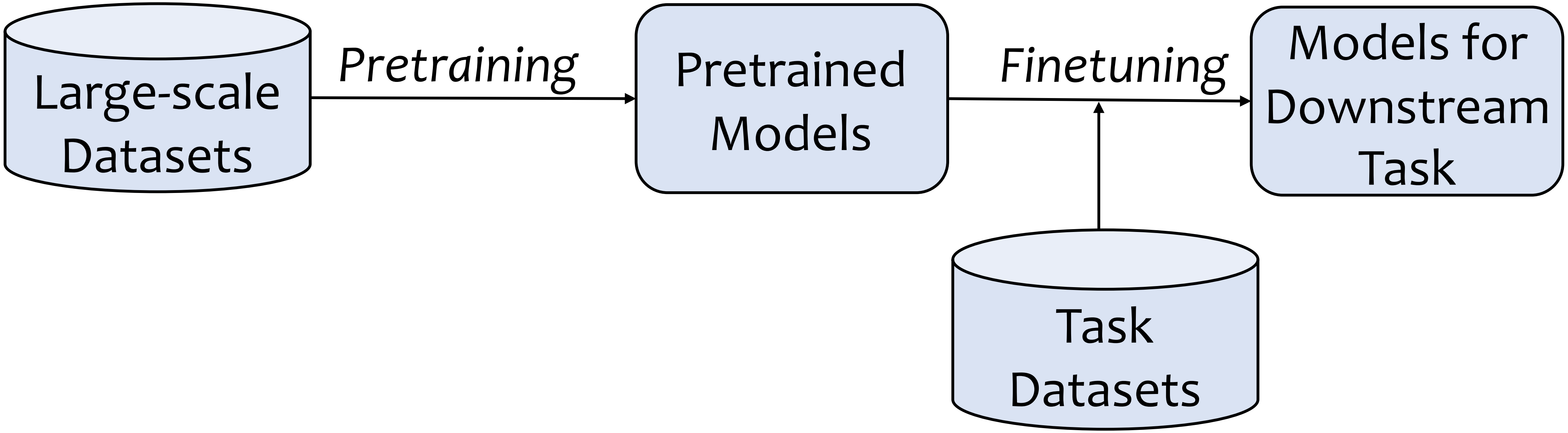
Conclusion



Social Media Keyphrase Generation
Encode implicit contexts
Encode explicit contexts
W4: “VD-BERT: A Unified Vision and Dialog Transformer with BERT” (Chapter 6) [EMNLP 2020]
W1: “Topic-Aware Neural Keyphrase Generation for Social Media Language” (Chapter 3) [ACL 2019]
W2: “Microblog Hashtag Generation via Encoding Conversation Contexts” (Chapter 4) [NAACL 2019]
W3: “Cross-Media Keyphrase Prediction: A Unified Framework with M3H-Att and Image Wordings” (Chapter 5) [EMNLP 2020]
Single modality: text-only
Multi-modality: text-image

Future Work (1)

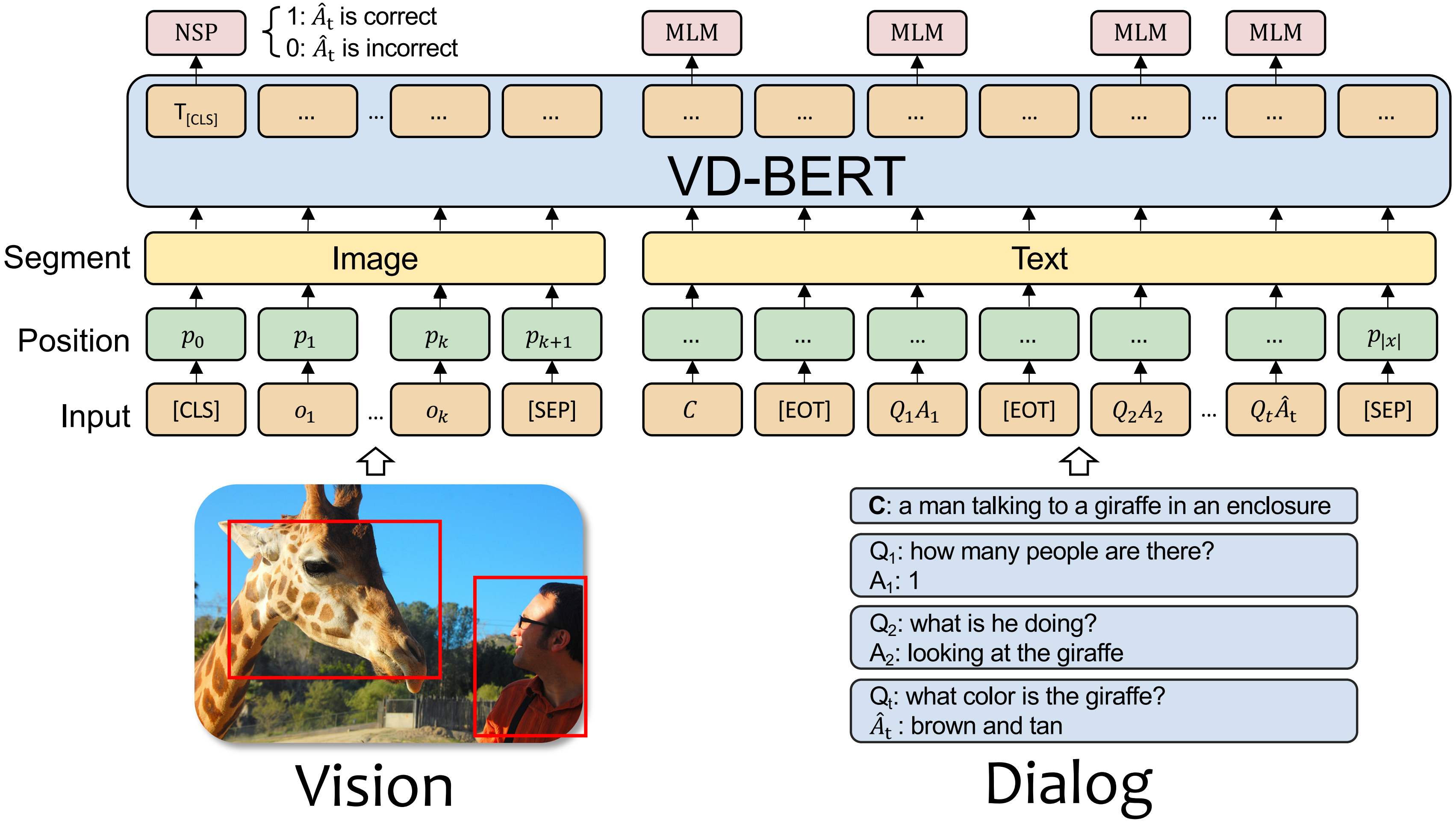
- Extend vision-language pretraining to benefit cross-media understanding



Pretrain-then-finetune paradigm

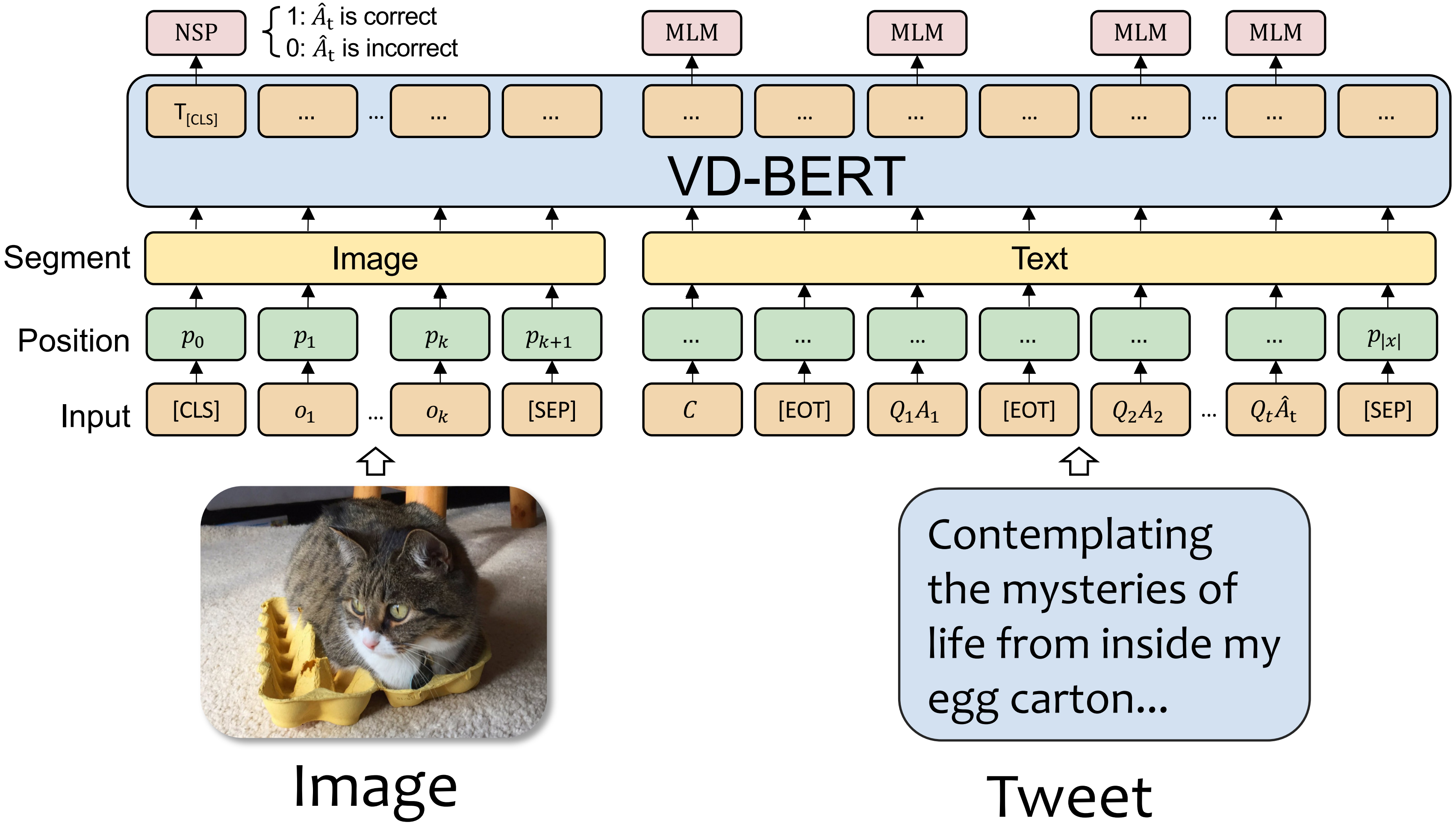
Future Work (1)

- Vision-language pretraining can achieve effective vision and dialog fusion



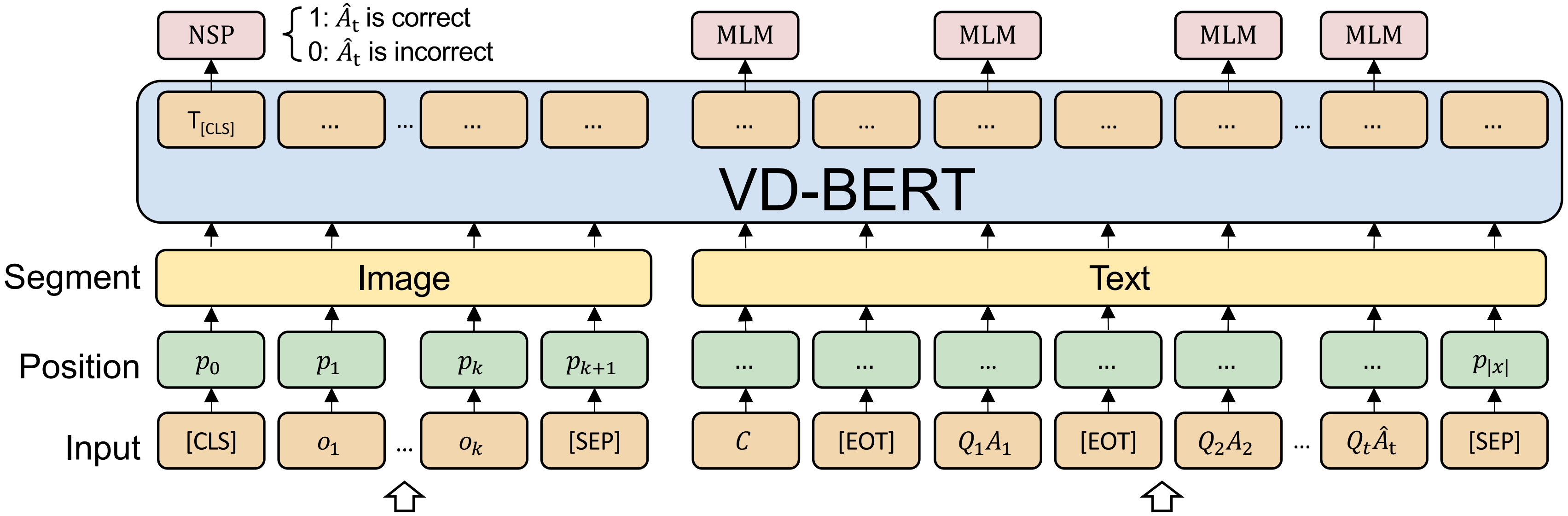
Future Work (1)

- Whether it can encourage fusion of vision and social media post?

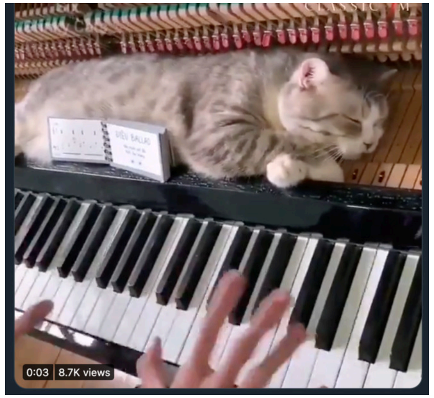


Future Work (1)

- Extend to video-text understanding…



Video

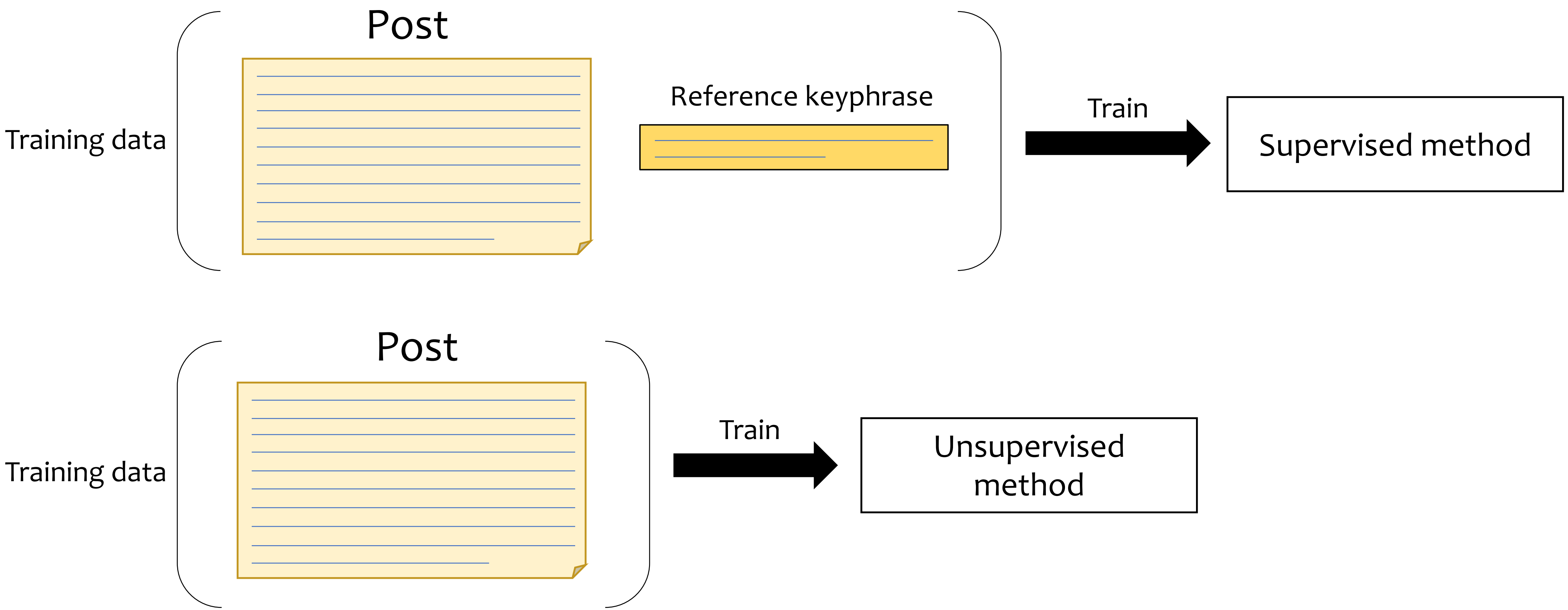


Let the keys fly hooman, I love it when you play the piano…

Tweet

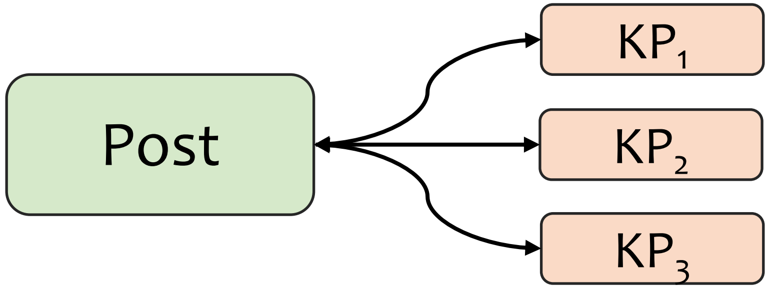
Future Work (2)

- Unsupervised learning for keyphrase prediction



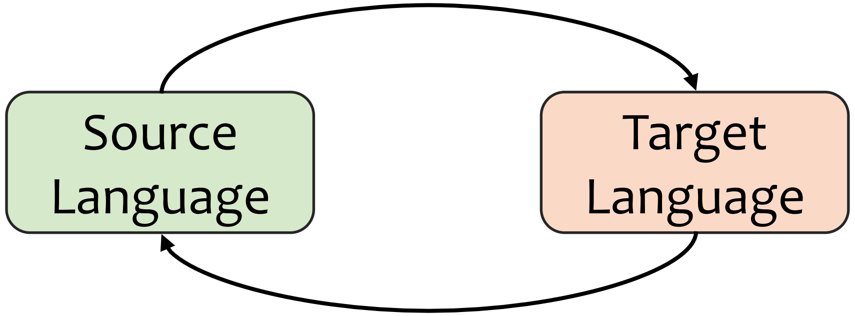
Future Work (2)

- Unsupervised keyphrase extraction
  - [Bennani-Smires et al., CoNLL 2018]



Similarity

- Unsupervised machine translation
  - [Lample et al., EMNLP 2018]



Back translation

Unsupervised learning for keyphrase generation

Publications

1. **Yue Wang**, Shafiq Joty, Michael R. Lyu, Irwin King, Caiming Xiong, and Steven C.H. Hoi. *VD-BERT: A Unified Vision and Dialog Transformer with BERT.* In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (**EMNLP**), Long Paper, 2020.
2. **Yue Wang**, Jing Li, Michael Lyu and Irwin King. *Cross-Media Keyphrase Prediction: A Unified Framework with Multi-Modality Multi-Head Attention and Image Wordings*. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (**EMNLP**), Long Paper, 2020.
3. **Yue Wang**, Jing Li, Hou Pong Chan, Irwin King, Michael R. Lyu, Shuming Shi. *Topic-Aware Neural Keyphrase Generation for Social Media Language*. In Proceedings of the 57th Conference of the Association for Computational Linguistics (**ACL**), Long Paper, 2019.
4. **Yue Wang**, Jing Li, Irwin King, Michael R. Lyu, Shuming Shi. *Microblog Hashtag Generation via Encoding Conversation Contexts*. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (**NAACL-HLT**), Long Paper, 2019.
5. Jian Li, **Yue Wang**, Michael R. Lyu, Irwin King. *Code Completion with Neural Attention and Pointer Networks*. In Proceedings of the 27th International Joint Conference on Artificial Intelligence (**IJCAI**), Long Paper, 2018.
6. ***Yue Wang**, Jing Li, Irwin King, Michael Lyu. *Encoding Explicit and Implicit Contexts for Social Media Keyphrase Generation*. Target at Journal of **Neurocomputing**.
7. ***Yue Wang**, Michael Lyu, Irwin King. *A Survey on Recent Advances in Vision and Language Representation Learning*. Target at IEEE Transactions on Knowledge and Data Engineering (**TKDE**).

Thanks!