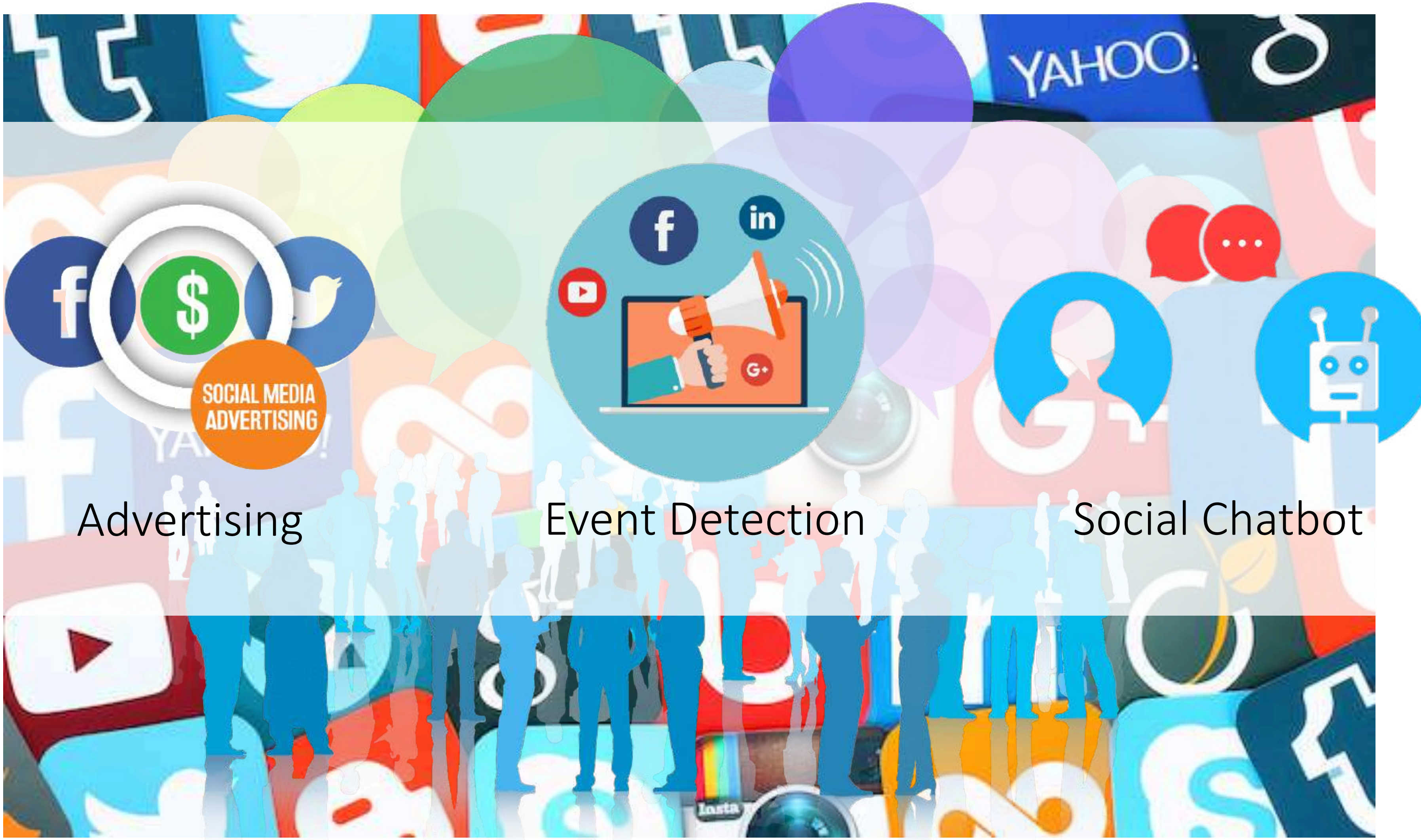
香港中文大學
The Chinese University of Hong Kong

Latent Variable Modeling for Natural Language Understanding

ZENG, Jichuan

Supervisors: Prof. Irwin King and Prof. Michael R. Lyu

2019/09/06


YAHOO!
SOCIAL MEDIA ADVERTISING
Advertising
Event Detection
Social Chatbot

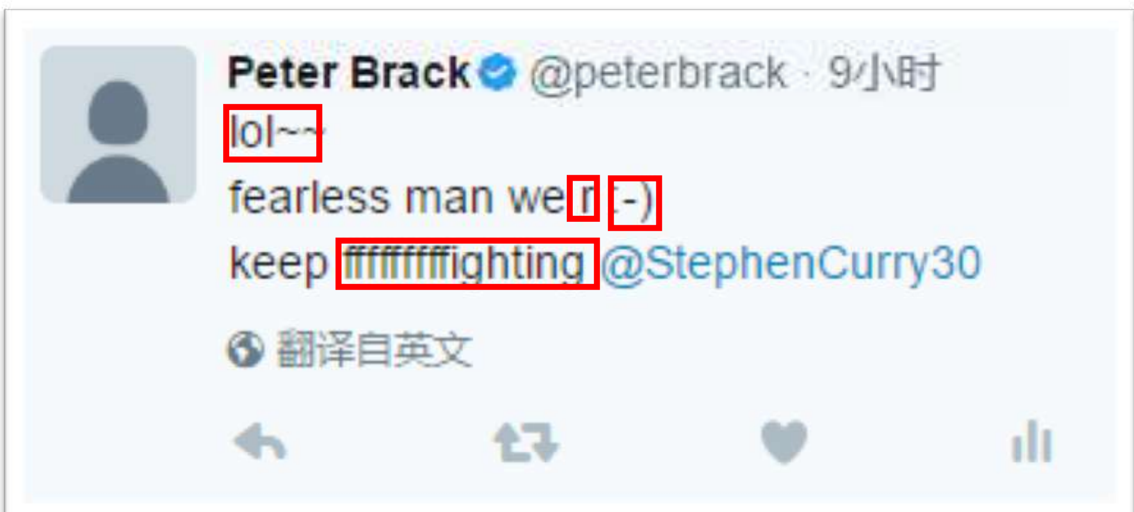
Understanding social media text is **important**, but challenging!

Challenge - Huge Volume without Label

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

Challenge - Data Sparsity

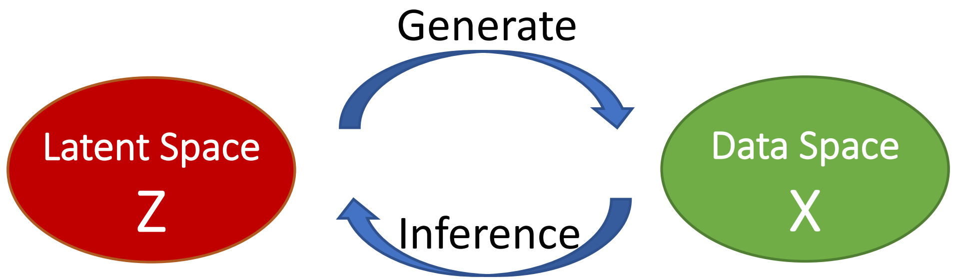
- Short in length
- Informal style
- Syntax errors



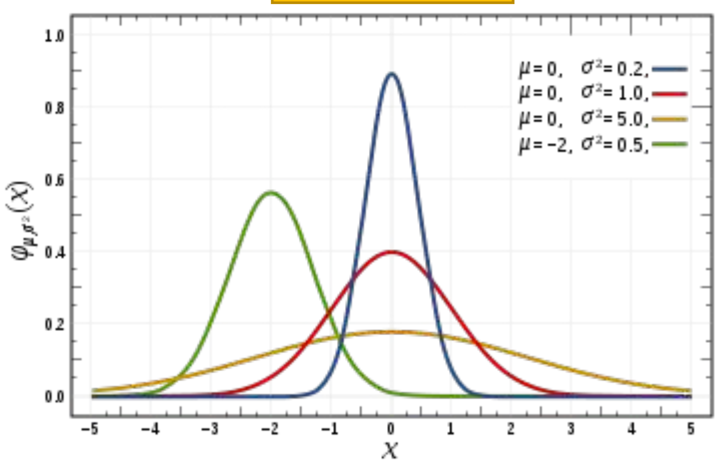
Challenge - Open Domain

- Wide variety of topics
- No pre-defined task-specific scheme
- Limited external resources

Latent Variable Modeling



Example 1



Latent variables: μ, σ^2

Example 2

Item

User

User/Item	A	B	C	D
Ted		4		3
Carol	3		2	
Bob		5		2
Alice			4	

Ted	1.4	.9
Carol	1.2	1
Bob	1.5	.9
Alice	1.2	.8

User

*

A	B	C	D
1.4	1.3	.9	1.2
.8	1.1	2	.8

Item

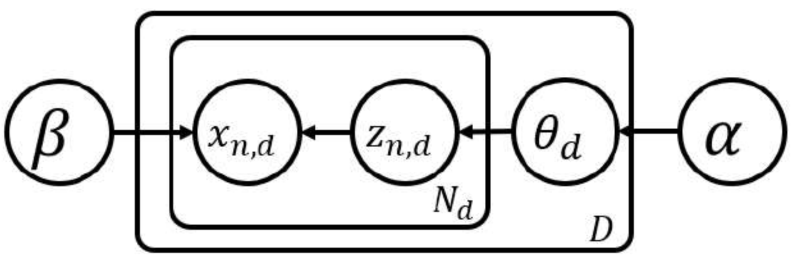
$$\mathrm{R} = \mathrm{U}V^T$$

Latent variables: u, v

Latent Dirichlet Allocation (LDA)

- Each **topic** is a distribution of words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

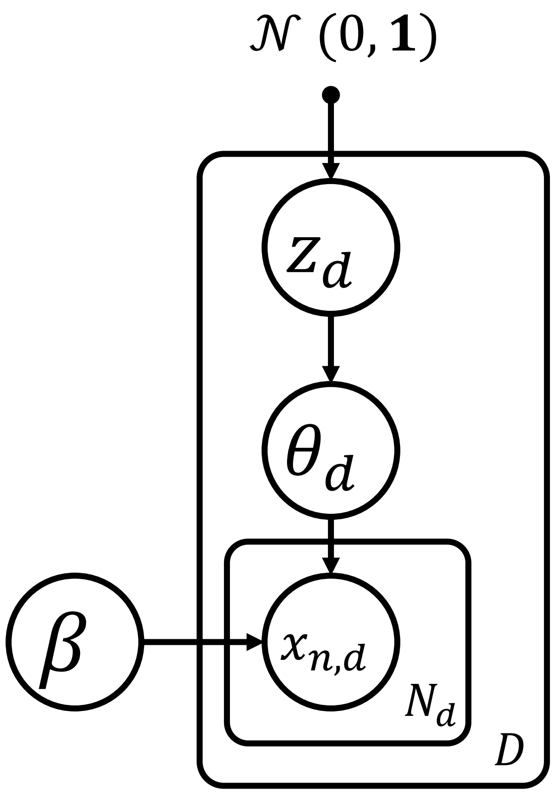
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI



The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

[Blei et al., 2006]

Neural Topic Model (NTM)



Generative Model of **NTM**

Generative process:

- For each document x_d:

$$z_d \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$$
$$\theta_d = Softmax(f_\theta(z_d))$$

- For each word in x_d :

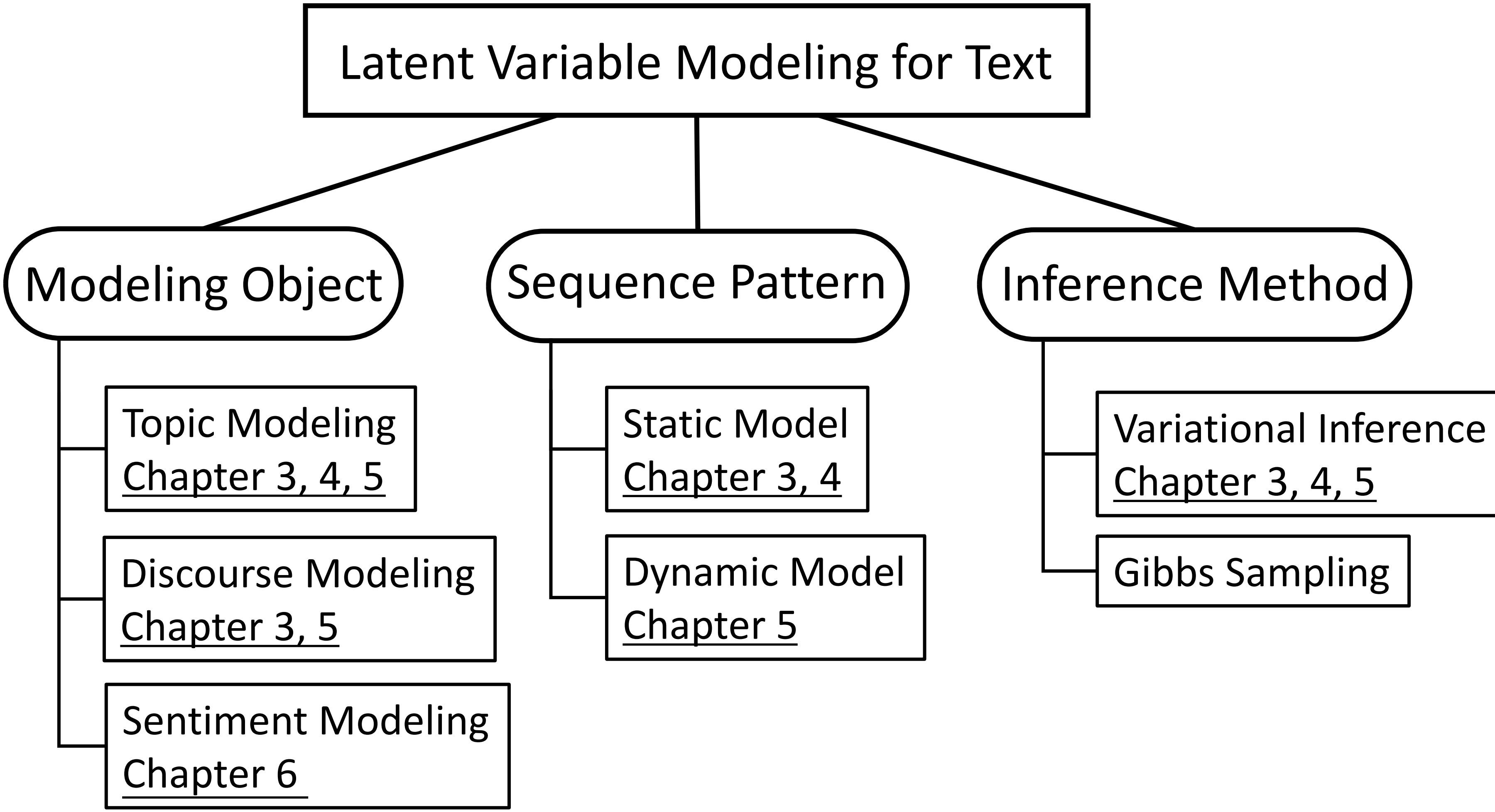
$$w_d = Softmax(f_\beta(\theta_d))$$
$$x_{d,n} \sim Multi(w_\mathrm{d})$$

Inference Process:

$$\boldsymbol{\mu} = f_\mu\big(f_e(x_d)\big),$$
$$log\boldsymbol{\sigma} = f_\sigma\big(f_e(x_d)\big)$$

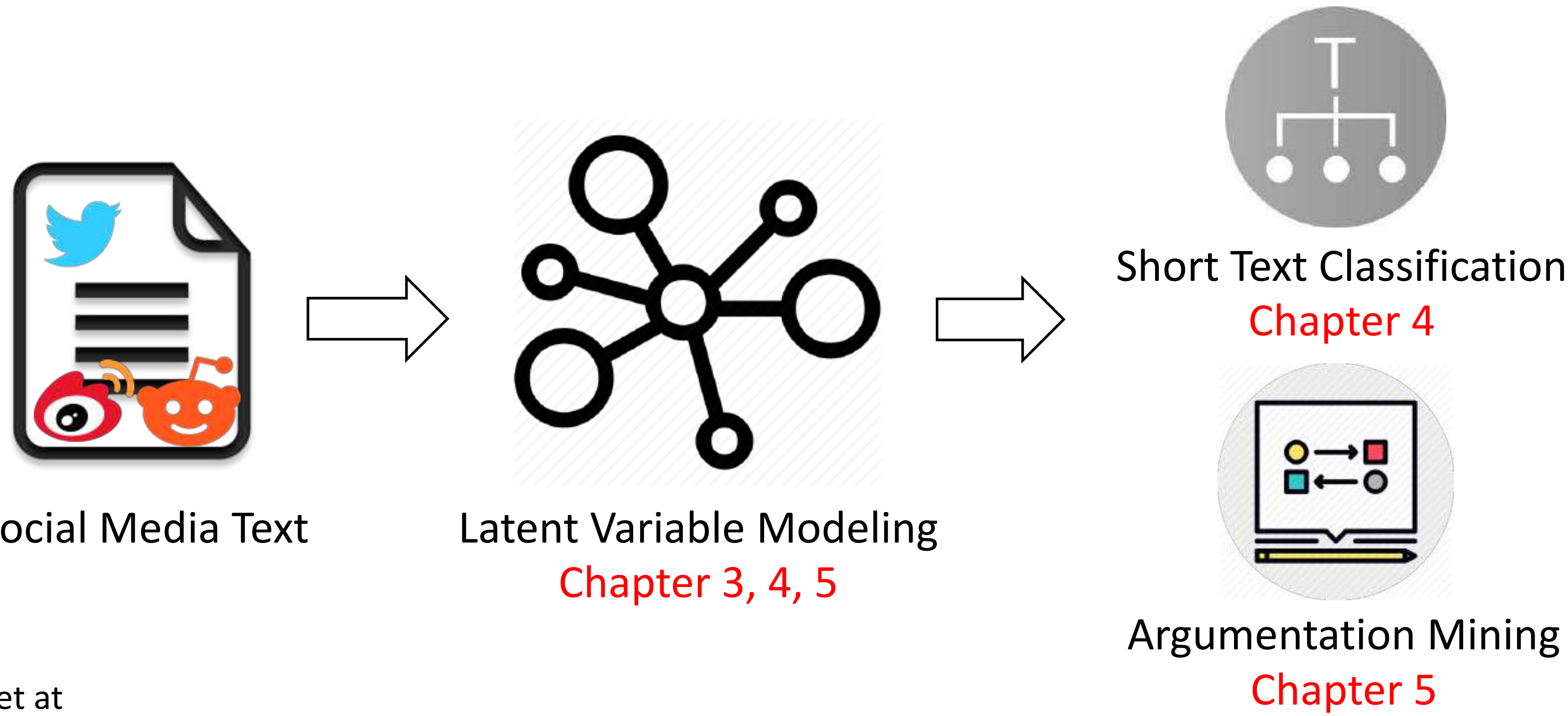
[Srivastava and Sutton, 2017; Miao et al., 2017]

A Taxonomy


Latent Variable Modeling for Text
Modeling Object
Sequence Pattern
Inference Method
Topic Modeling
Chapter 3, 4, 5
Discourse Modeling
Chapter 3, 5
Sentiment Modeling
Chapter 6
Static Model
Chapter 3, 4
Dynamic Model
Chapter 5
Variational Inference
Chapter 3, 4, 5
Gibbs Sampling

Thesis Contributions

- **Microblog Conversation Modeling** [TACL’19] (Chapter 3)
- **Short Text Classification** [EMNLP’18] (Chapter 4)
- **Argumentation Mining** [*WWW’20](Chapter 5)



* Target at

Thesis Contributions

- **Microblog Conversation Modeling** [TACL'19] (**Chapter 3**)
  - Joint modeling topics and discourse
  - Produce coherent topics and meaningful discourse
  - Extensible with other NN framework
- **Short Text Classification** [EMNLP'18] (**Chapter 4**)
  - Jointly explore topic modeling and text classification
  - Alleviate data sparsity issue
  - 0.5%-3.5% abs accuracy increase in 4 datasets
- **Argumentation Mining** [*WWW'20](**Chapter 5**)
  - Modeling dynamic topics and discourse in argumentation Process
  - Substantial improvement in persuasiveness prediction
  - Reveal the key factors of persuasiveness

Outline

- Topic 1: Microblog Conversation Modeling
- Topic 2: Short Text Classification
- Topic 3: Argumentation Persuasiveness Analysis
- Conclusion and Future Work

Outline

- Topic 1: Microblog Conversation Modeling
- Topic 2: Short Text Classification
- Topic 3: Argumentation Persuasiveness Analysis
- Conclusion and Future Work

Motivation

How shall we automatically extract the critical points and make sense of these microblog conversations?

Motivation

What	What is the message talking about / *Donald Trump, Louisiana, flood victims*
How	How the opinion is voiced / *Statement, Reference, Sarcasm*

Microblog Search

Social Chatbot

Sentiment Analysis

Challenges

- The volume of microblog is growing quickly
  - ➢ Need to design an effective and efficient method.
- Most of the text data are unannotated
  - ➢ Difficult to build a supervised model to predict *What* and *How.*
- Severe data sparsity issue
  - ➢ Difficult to understand the microblog message without the context.

Example

Just watched HRC openly endorse a gun-control measure which will fail in front of the Supreme Court.

Statement

Comment

People said the same thing about Obama, and nothing took place. Gun laws just aren't being enforced like they should be. :/

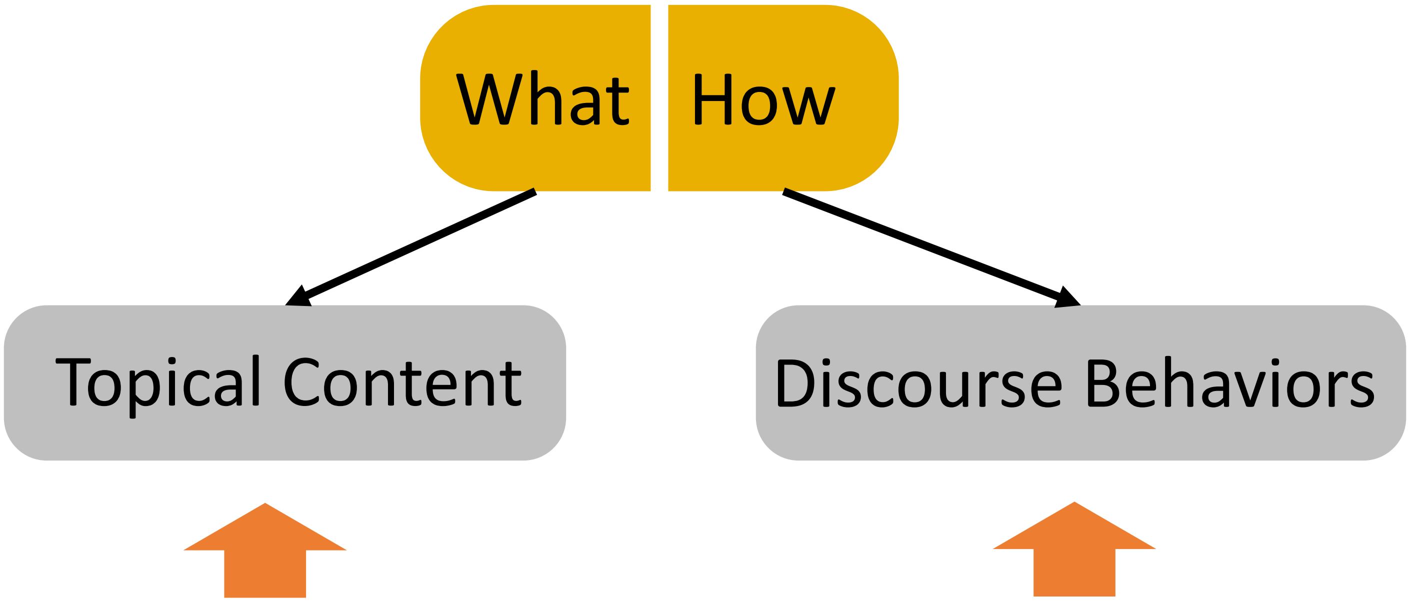
Okay, hold up. What do you think I'm referencing here? It's not what you're talking about.

Question

Agreement

Thought it was about gun control. I'm in agreement that gun rights shouldn't be stripped.

Observations



<u>Topic words:</u>

Gun-control
Supreme Court
Obama
HRC
...

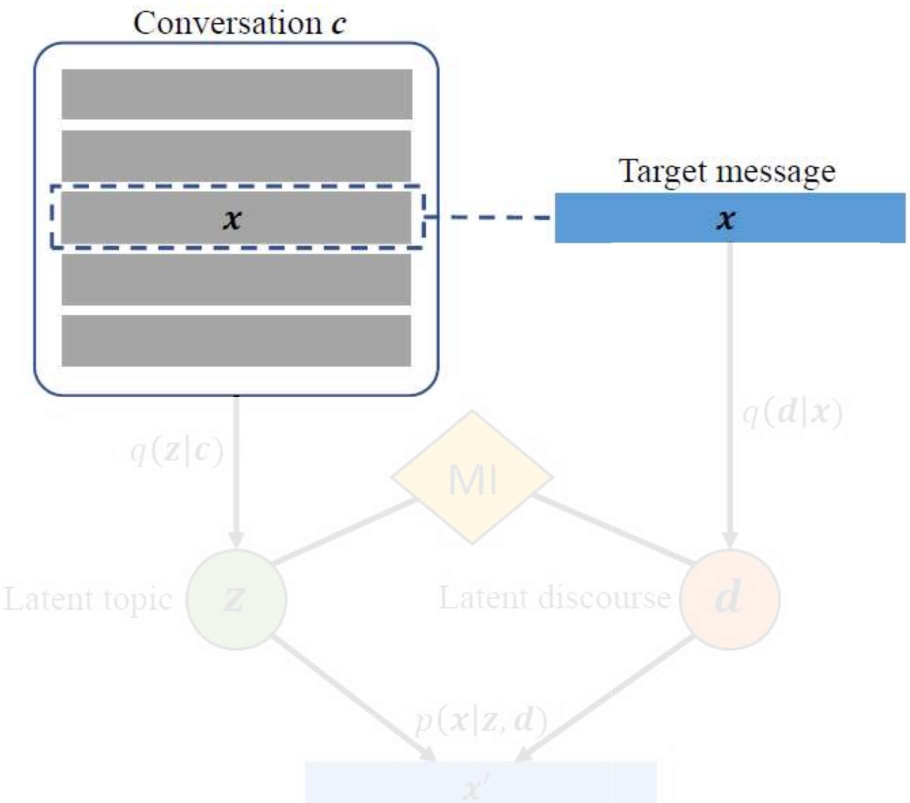
<u>Discourse words:</u>

Will
Should
What
?
...

Existing Work

- Topic Modeling for Social Media
  - Not work well on short text messages[Blei et al., 2003]
  - Cannot use the richer context information in a conversation [Yan et al., 2013, Nguyen et al., 2015]
  - The heuristically aggregation strategies are unnatural [Hong et al., 2010, Ramage et al., 2010]
- Conversation Discourse
  - Require High-quality labeled data [Stolcke et al., 2000, Ji et al., 2016]
  - Did not consider the effect of conversation topics [Ritter et al., 2010, Jotty et al., 2011, Zhao et al., 2018]
  - Sampling based, low efficiency, hard to extend [Li et al., 2016]

Framework



Generative Process

- For each conversation c:

$$z \sim \mathcal{N}(\mu, \sigma^2)$$
$$\theta = Softmax(f_\theta(z))$$

- For each message x in c:

$$d \sim Multi(\pi)$$

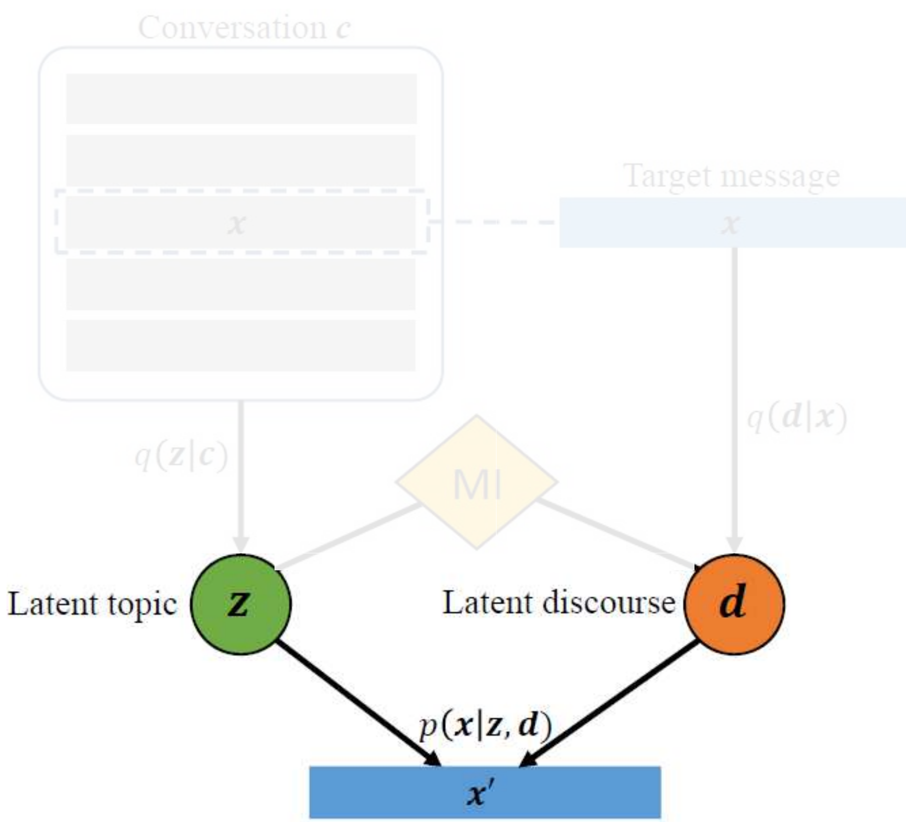
- For each word in x:

$$w_n = Softmax(f_{\phi^T}(\theta) + f_{\phi^D}(d))$$
$$x_n \sim Multi(w_n)$$

Inference Process

$$\mu = f_\mu(f_e(c_{BoW}))$$
$$log\sigma = f_\sigma(f_e(c_{BoW}))$$
$$\pi = Softmax(f_\pi(c_{BoW}))$$

Framework



Generative Process

- For each conversation $\boldsymbol{c}$:

  $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$

  $\theta = Softmax(f_\theta(\boldsymbol{z}))$
- For each message $\boldsymbol{x}$ in $\boldsymbol{c}$:

  $d \sim Multi(\boldsymbol{\pi})$
- For each word in $\boldsymbol{x}$:

  $w_n = Softmax(f_{\phi^T}(\theta) + f_{\phi^D}(d))$

  $x_n \sim Multi(w_n)$

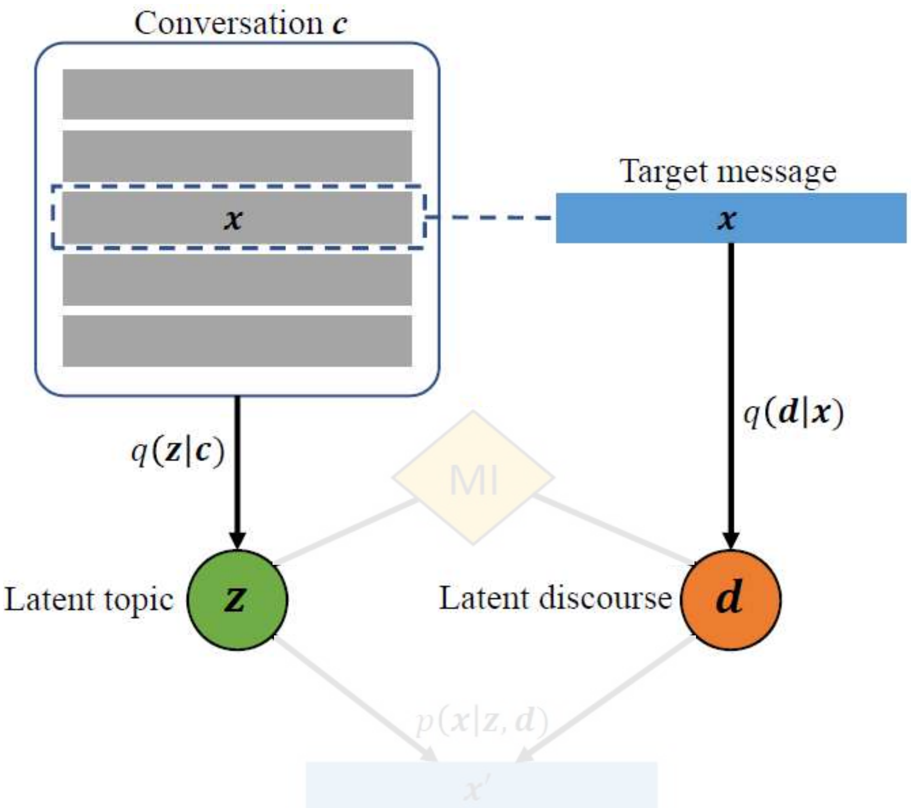
Inference Process

$\boldsymbol{\mu} = f_\mu(f_e(\boldsymbol{c}_{BoW}))$

$log\boldsymbol{\sigma} = f_\sigma(f_e(\boldsymbol{c}_{BoW}))$

$\boldsymbol{\pi} = Softmax(f_\pi(\boldsymbol{c}_{BoW}))$

Framework



Generative Process

- For each conversation $\boldsymbol{c}$:

  $$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$$
  $$\theta = Softmax(f_\theta(\boldsymbol{z}))$$

- For each message $\boldsymbol{x}$ in $\boldsymbol{c}$:

  $$d \sim Multi(\boldsymbol{\pi})$$

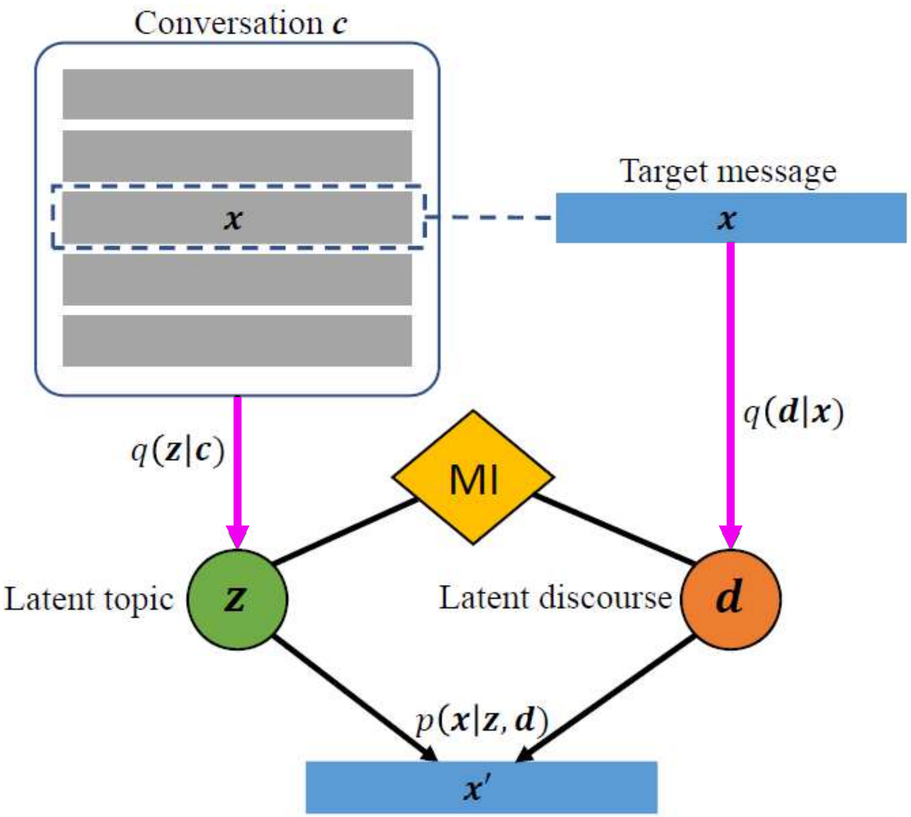
- For each word in $\boldsymbol{x}$:

  $$w_n = Softmax(f_{\phi^T}(\theta) + f_{\phi^D}(d))$$
  $$x_n \sim Multi(w_n)$$

Inference Process

$$\boldsymbol{\mu} = f_\mu(f_e(\boldsymbol{c}_{BoW}))$$
$$log\boldsymbol{\sigma} = f_\sigma(f_e(\boldsymbol{c}_{BoW}))$$
$$\boldsymbol{\pi} = Softmax(f_\pi(\boldsymbol{x}_{BoW}))$$

Framework



Training Losses

- Evidence lower bound (ELBO) losses

$$\mathcal{L}_z = -D_{KL}(q(\boldsymbol{z}|\boldsymbol{c})||p(\boldsymbol{z})) + \mathbb{E}_{q(z|c)}[\log p(\boldsymbol{c}|\boldsymbol{z})]$$

$$\mathcal{L}_d = -D_{KL}(q(\boldsymbol{d}|\boldsymbol{x})||p(\boldsymbol{d})) + \mathbb{E}_{q(d|x)}[\log p(\boldsymbol{x}|\boldsymbol{d})]$$

- Reconstruction loss

$$\mathcal{L}_x = \mathbb{E}_{q(z|c)q(d|x)}[\log p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{d})]$$

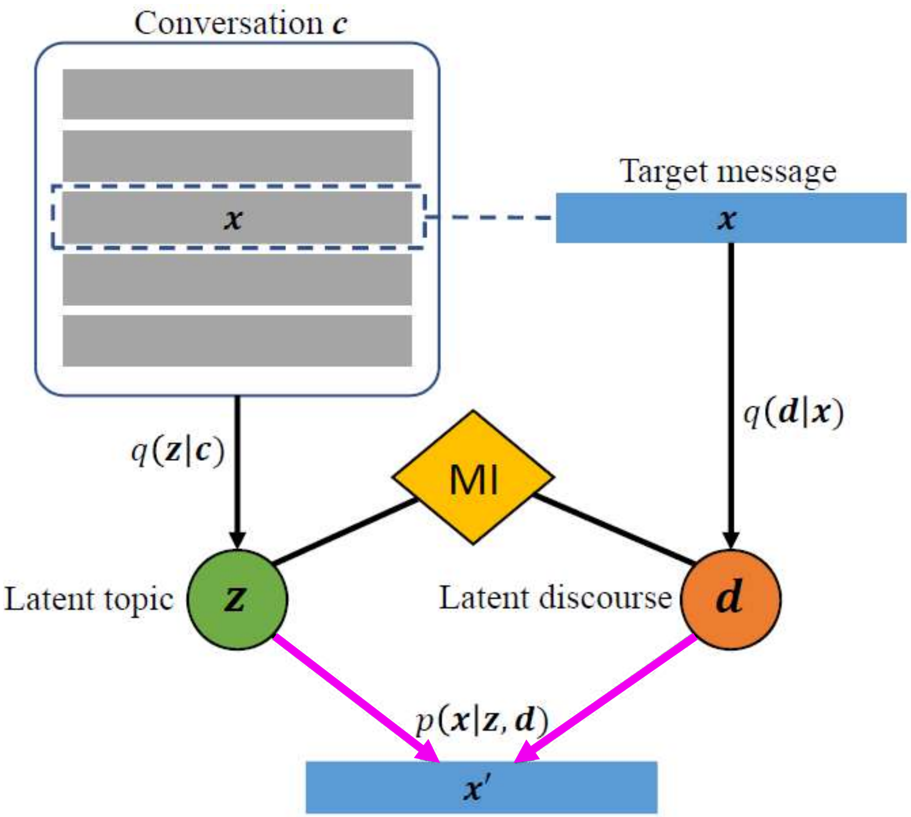
- Mutual information penalty

$$\mathcal{L}_{MI} = \mathbb{E}_{q(\boldsymbol{z})}[D_{KL}(\log p(\boldsymbol{d}|\boldsymbol{z})\,||p(\boldsymbol{d}))]$$

Final Objective

$$\mathcal{L} = \mathcal{L}_z + \mathcal{L}_d + \mathcal{L}_x - \lambda\mathcal{L}_{MI}$$

Framework



Training Losses

- Evidence lower bound (ELBO) losses

$$\mathcal{L}_z = -D_{KL}(q(\boldsymbol{z}|\boldsymbol{c})||p(\boldsymbol{z})) + \mathbb{E}_{q(z|c)}[\log p(\boldsymbol{c}|\boldsymbol{z})]$$

$$\mathcal{L}_d = -D_{KL}(q(\boldsymbol{d}|\boldsymbol{x})||p(\boldsymbol{d})) + \mathbb{E}_{q(d|x)}[\log p(\boldsymbol{x}|\boldsymbol{d})]$$

- Reconstruction loss

$$\mathcal{L}_x = \mathbb{E}_{q(z|c)q(d|x)}[\log p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{d})]$$

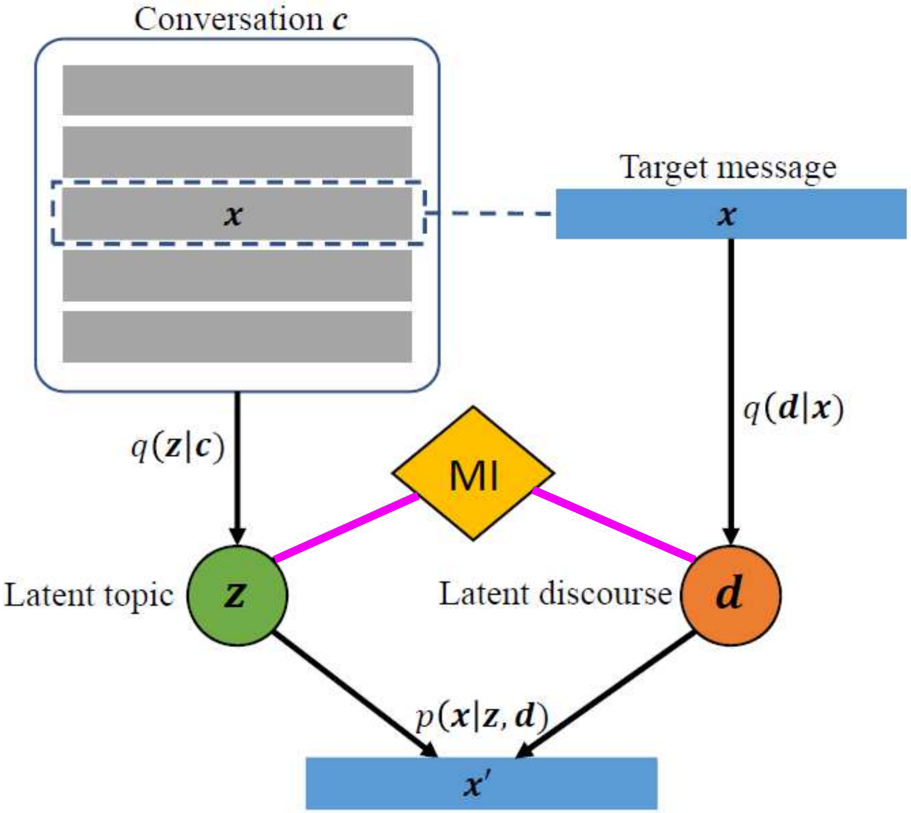
- Mutual information penalty

$$\mathcal{L}_{MI} = \mathbb{E}_{q(z)}[D_{KL}(\log p(\boldsymbol{d}|\boldsymbol{z})\,||p(\boldsymbol{d}))]$$

Final Objective

$$\mathcal{L} = \mathcal{L}_z + \mathcal{L}_d + \mathcal{L}_x - \lambda\mathcal{L}_{MI}$$

Framework



Training Losses

- Evidence lower bound (ELBO) losses

$$\mathcal{L}_z = -D_{KL}(q(\boldsymbol{z}|\boldsymbol{c})||p(\boldsymbol{z})) + \mathbb{E}_{q(z|c)}[\log p(\boldsymbol{c}|\boldsymbol{z})]$$

$$\mathcal{L}_d = -D_{KL}(q(\boldsymbol{d}|\boldsymbol{x})||p(\boldsymbol{d})) + \mathbb{E}_{q(d|x)}[\log p(\boldsymbol{x}|\boldsymbol{d})]$$

- Reconstruction loss

$$\mathcal{L}_x = \mathbb{E}_{q(z|c)q(d|x)}[\log p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{d})]$$

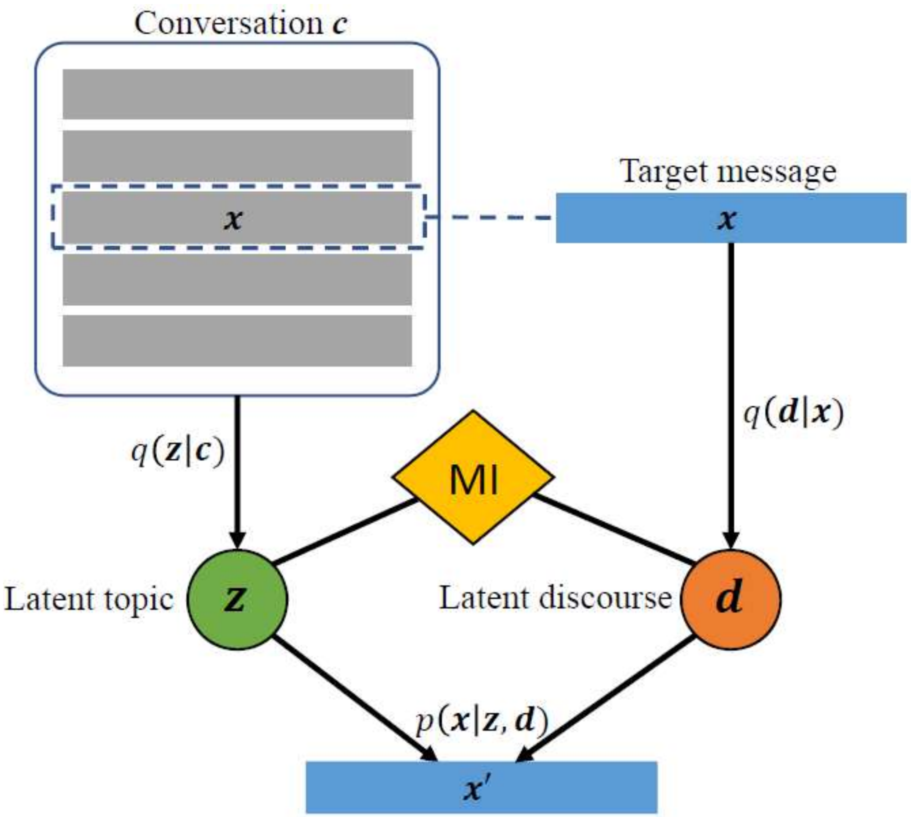
- Mutual information penalty

$$\mathcal{L}_{MI} = \mathbb{E}_{q(\boldsymbol{z})}[D_{KL}(\log p(\boldsymbol{d}|\boldsymbol{z})\,||p(\boldsymbol{d}))]$$

Final Objective

$$\mathcal{L} = \mathcal{L}_z + \mathcal{L}_d + \mathcal{L}_x - \lambda\mathcal{L}_{MI}$$

Framework



Training Losses

- Evidence lower bound (ELBO) losses

$$\mathcal{L}_z = -D_{KL}(q(\boldsymbol{z}|\boldsymbol{c})||p(\boldsymbol{z})) + \mathbb{E}_{q(z|c)}[\log p(\boldsymbol{c}|\boldsymbol{z})]$$

$$\mathcal{L}_d = -D_{KL}(q(\boldsymbol{d}|\boldsymbol{x})||p(\boldsymbol{d})) + \mathbb{E}_{q(d|x)}[\log p(\boldsymbol{x}|\boldsymbol{d})]$$

- Reconstruction loss

$$\mathcal{L}_x = \mathbb{E}_{q(z|c)q(d|x)}[\log p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{d})]$$

- Mutual information penalty

$$\mathcal{L}_{MI} = \mathbb{E}_{q(\boldsymbol{z})}[D_{KL}(\log p(\boldsymbol{d}|\boldsymbol{z})\,||p(\boldsymbol{d}))]$$

Final Objective

$$\mathcal{L} = \mathcal{L}_z + \mathcal{L}_d + \mathcal{L}_x - \lambda\mathcal{L}_{MI}$$

Dataset

- TREC2011: Microblog conversations concerning a wide rage of topics.
- TWT16: Conversations centered around U.S. presidential election in 2016.

Datasets	# of convs	Avg msgs per conv	Avg words per msg	|Vocab|
TREC	116,612	3.95	11.38	9,463
TWT16	29,502	8.67	14.70	7,544

80% training, 10% development, 10% testing

Topic Coherence

Models	K = 50 TREC	K = 50 TWT16	K = 100 TREC	K = 100 TWT16
Baselines				
LDA	0.467	0.454	0.467	0.454
BTM	0.460	0.461	0.466	0.463
LF-DMM	0.456	0.448	0.463	0.466
LF-LDA	0.470	0.456	0.467	0.453
NTM	0.478	0.479	0.482	0.443
Li et al. (2018)	0.463	0.433	0.464	0.435
Our models				
TOPIC ONLY	0.478	0.482	0.481	0.471
TOPIC+DISC	**0.485**	**0.487**	**0.496**	**0.480**

C_V coherence scores

LDA	people trump police violence gun death protest guns flag shot
BTM	gun guns people police wrong right think law agree black
LF-DMM	gun police black said people guns killing ppl amendment laws
Li et al. (2018)	wrong don trump gun understand laws agree guns doesn make
NTM	gun understand yes guns world dead real discrimination trump silence
TOPIC ONLY	shootings gun guns cops charges control mass commit know agreed
TOPIC+DISC	guns gun shootings chicago shooting cops firearm criminals commit laws

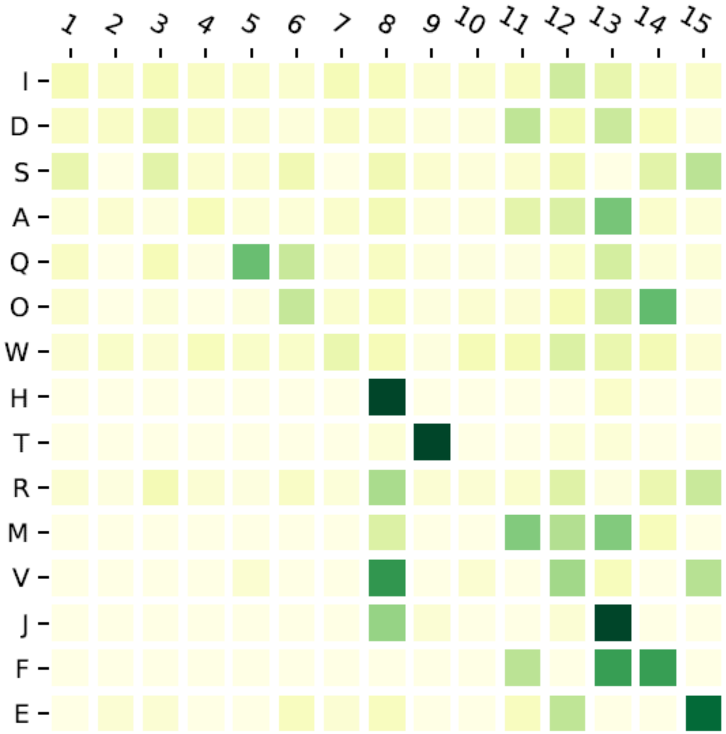
Top 10 representative terms of "gun control". *Non-topic words* are italic and blue, and off-topic words are underlined and red.

Discourse Interpretability

Mastodon dataset [Cerisara et al. 2018]
2,217 microblog messages forming 505 conversations, 15 discourses

Models	Purity	Homogeneity	VI
Baselines			
LAED	0.505	0.022	6.418
Li et al. (2018)	0.511	0.096	5.540
Our models			
DISC ONLY	0.510	0.112	5.532
TOPIC+DISC	**0.521**	**0.142**	**5.097**

The purity, homogeneity, and variation of information (VI) scores for the latent discourse roles



I: statement, D: disagreement, S: suggest, A: agreement, Q: yes/no question, O: wh*/open question, W: open+choice answer, H: initial greetings, T: thanking, R: request, M: sympathy, V: explicit performance, J: exclamation, F: acknowledge, and E: offer.

Discourse Interpretability

Discourse Roles	**TREC**	**TWT16**
Question	was what why is how that like ? ?? you	? why what MENT do does it the to did
Response	! love ha !! you saw lmao lol awesome !!!	doin uhhh ❗ awards yay joseph 😖 👋 muted
Agreement	okaay thankss wateva okayy txtd twitcam entertained havee goooood darlin	! you are agree re to they we with their
Quotation	& ' < > (... feat " " ")	» « (< ¦ ¦ < MENT .< ,- - ?< "
Statement	to will ! the be rt my in on and	will have if do be can want vote should ?
Argument	f∗∗k damn rt lmfao hair girl thing lmao ass bit∗h	😂 he said him she her but wrong did never

Top 10 representative terms of example discourse roles learned from TREC and TWT16.

Case Study

just watched hrc openly endorse gun control measure which will fail in front of the supreme court this is train wreck

people said the same thing about obama and nothing took place gun laws just aren being enforced like they should be :/

Visualization of the topic-discourse assignment of a twitter conversion from TWT16

Model Extensiblity

Models	TREC		TWT16	
	Acc	Avg F1	Acc	Avg F1
CNN only	0.199	0.167	0.334	0.311
Separate-Train	0.284	0.270	0.391	0.390
Joint-Train	**0.297**	**0.286**	**0.428**	**0.413**

Joint training with other NN architectures can bring benefits.

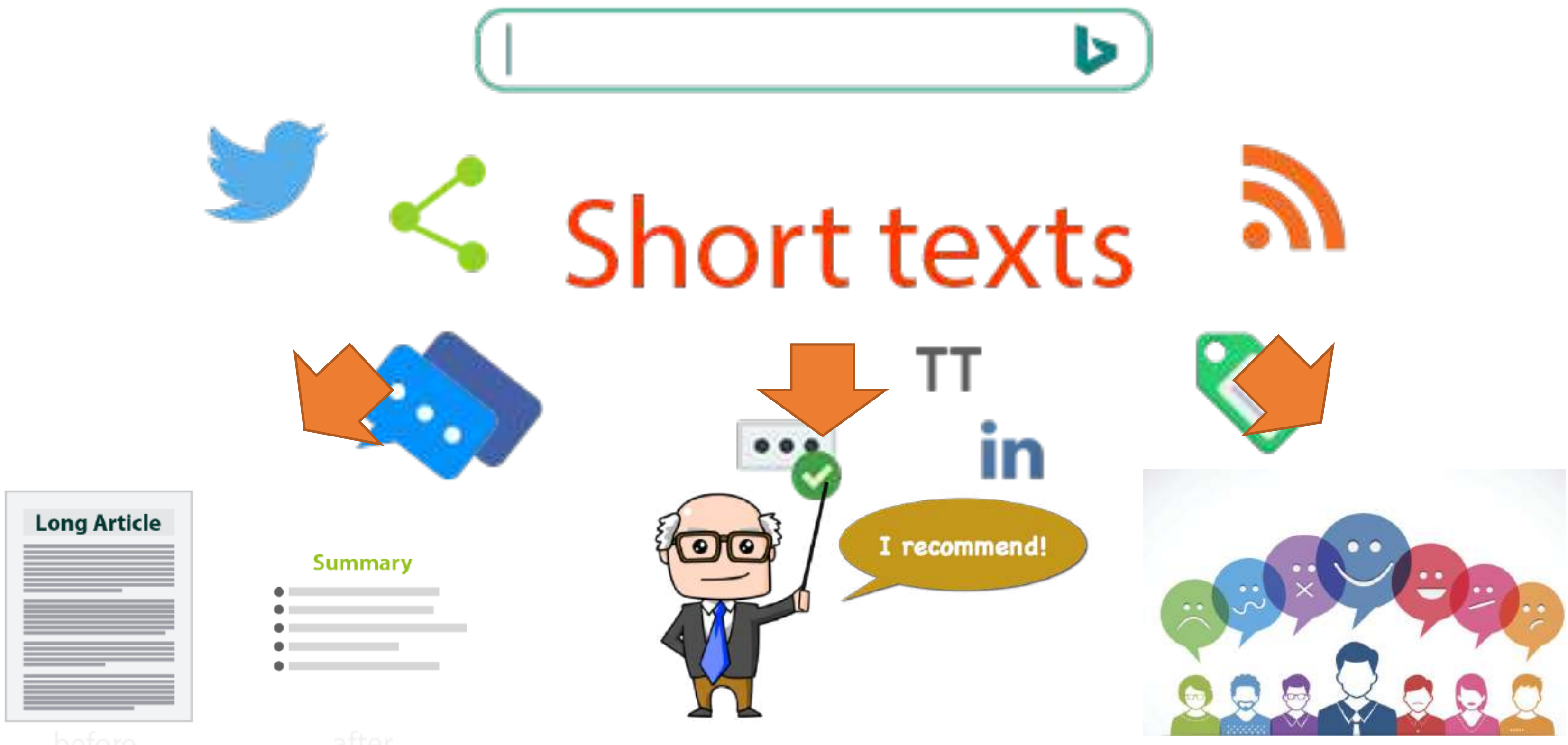
Summary

- We propose an unsupervised neural network framework that jointly explores topic and discourse from microblog conversation.
- Extensive experiments show that our model can generate coherent topics and meaningful discourse roles.
- Our model can be easily extended with other neural network architectures (such as CNN) to present better performance.

Outline

- Topic 1: Microblog Conversation Modeling
- Topic 2: Short Text Classification
- Topic 3: Argumentation Persuasiveness Analysis
- Conclusion and Future Work

Motivation



Text Summarization Recommendation Sentiment Analysis

Challenge - Data Sparsity

- Short and noisy
- Informal style
- Lack contextual information

Motivative Example

Training instances

R_1: [SuperBowl] I'll do anything to see the **Steelers** win.

R_2: [New.Music.Live] Please give **wristband**, she have major **Bieber** Fever.

Test instance

S: [New.Music.Live] I will do anything for **wristbands** gonna tweet till I win.

Justin Bieber
Bieber
Music Live
Purple Glass
…

Existing Work

- External Knowledge
  - Wikipedia, knowledge base [Jin et al. 2011, Lucia and Ferrari 2014, Wang et al. 2017]
  - Manually-crafted features [Pak and Paroubek 2010, Jiang et al. 2011]
  - Domain-specific, task-specific, not work well in social media
- Word Collocation Patterns
  - Word embeddings [Bowman et al. 2016, Krisknamurthy et al. 2017]
  - Topic models [Phan et al. 2008, Chen et al. 2011, Ren et al. 2016]
  - Need pre-trained, without joint modeling

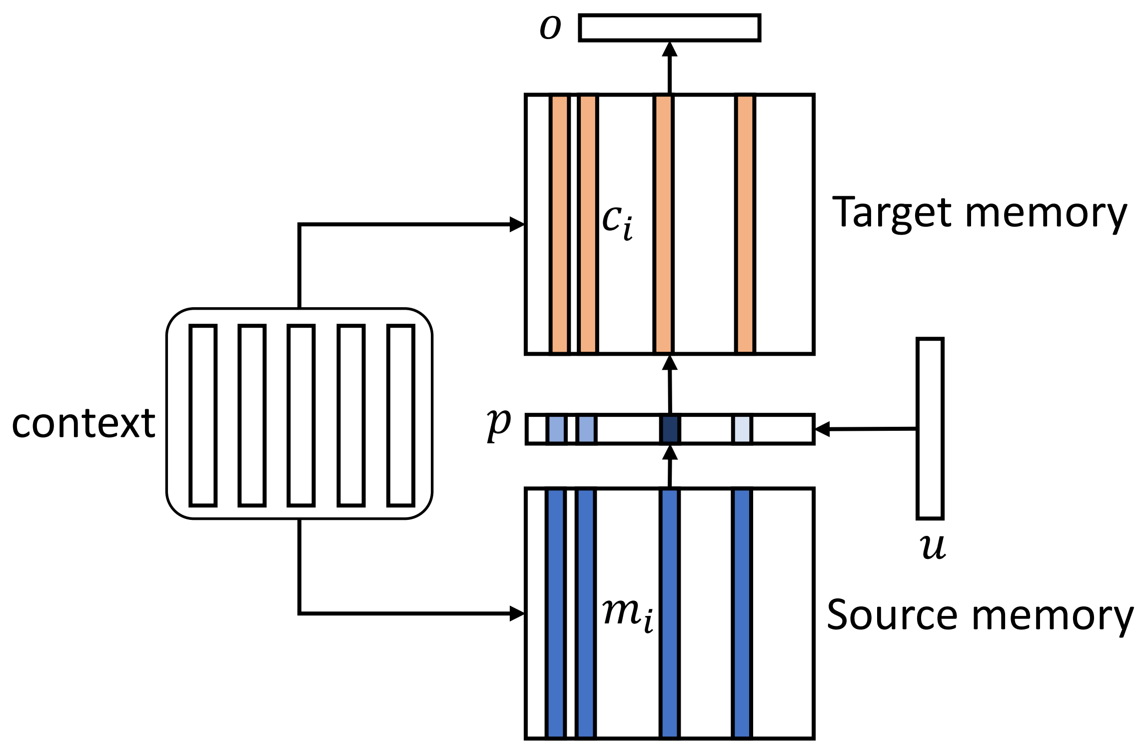
Model Intuition

S	I	will	do	anything	for	wristbands
Top-5 topic words	think	win	good	will	will	bieber
	win	think	like	for	for	newmusiclive
	play	watch	fan	thing	thing	justin
	score	day	look	know	know	tuesday
	carroll	big	for	na	na	glass

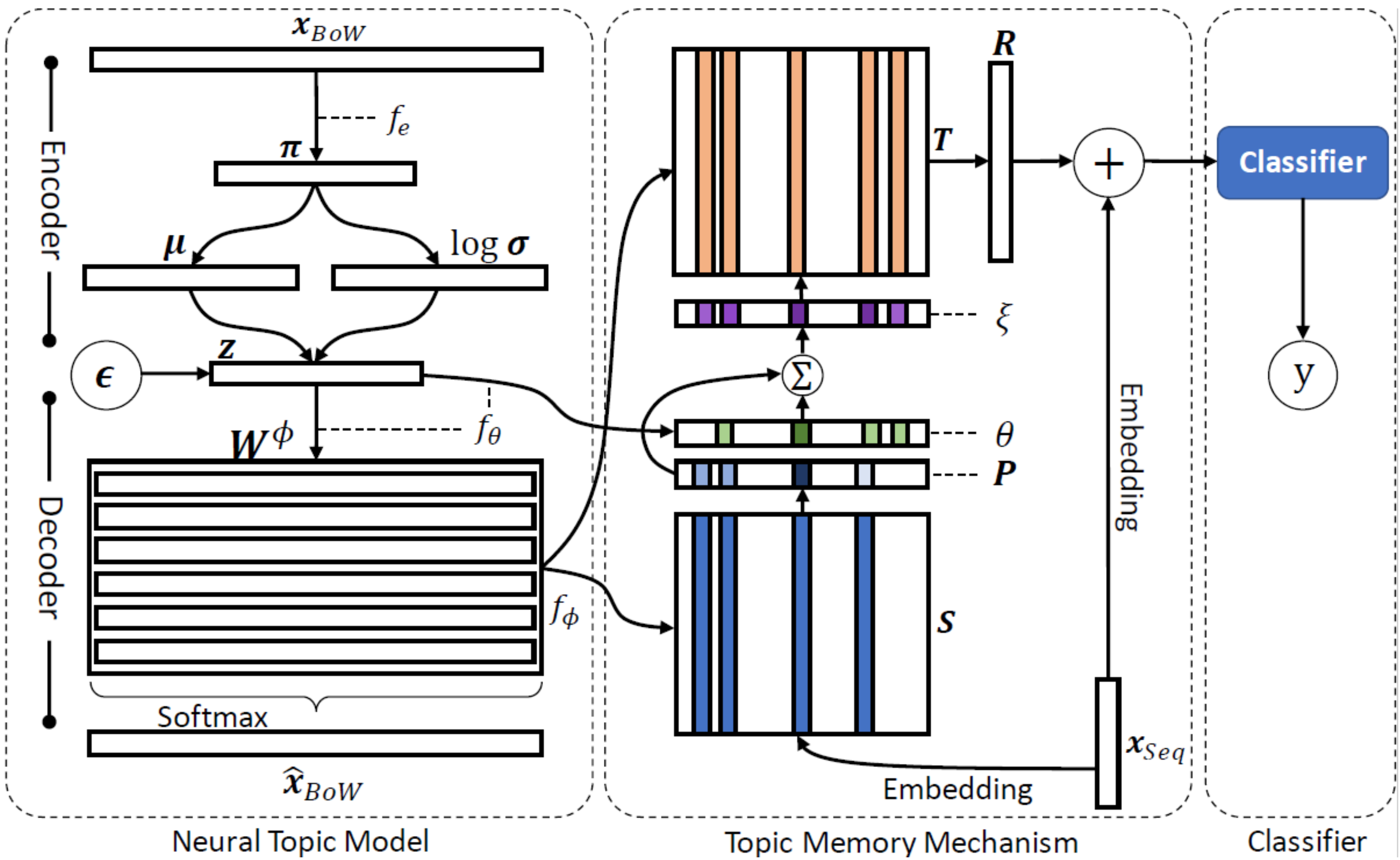
- Integrate “context” information (topic words)
- Pay attention to the indicative words (e.g., **wristbands**)

Memory Networks

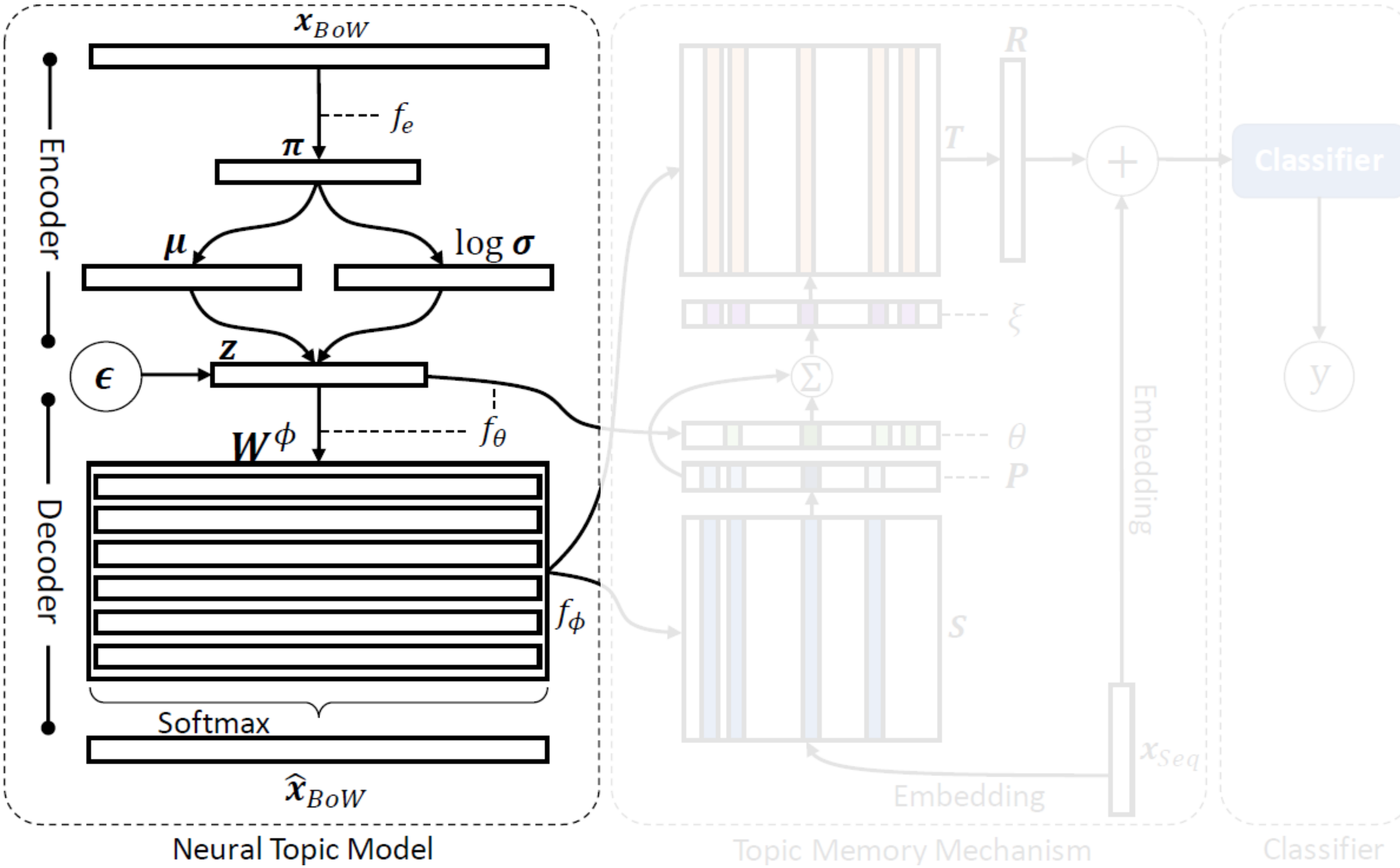
- Source memory
- Memory weight
- Target memory



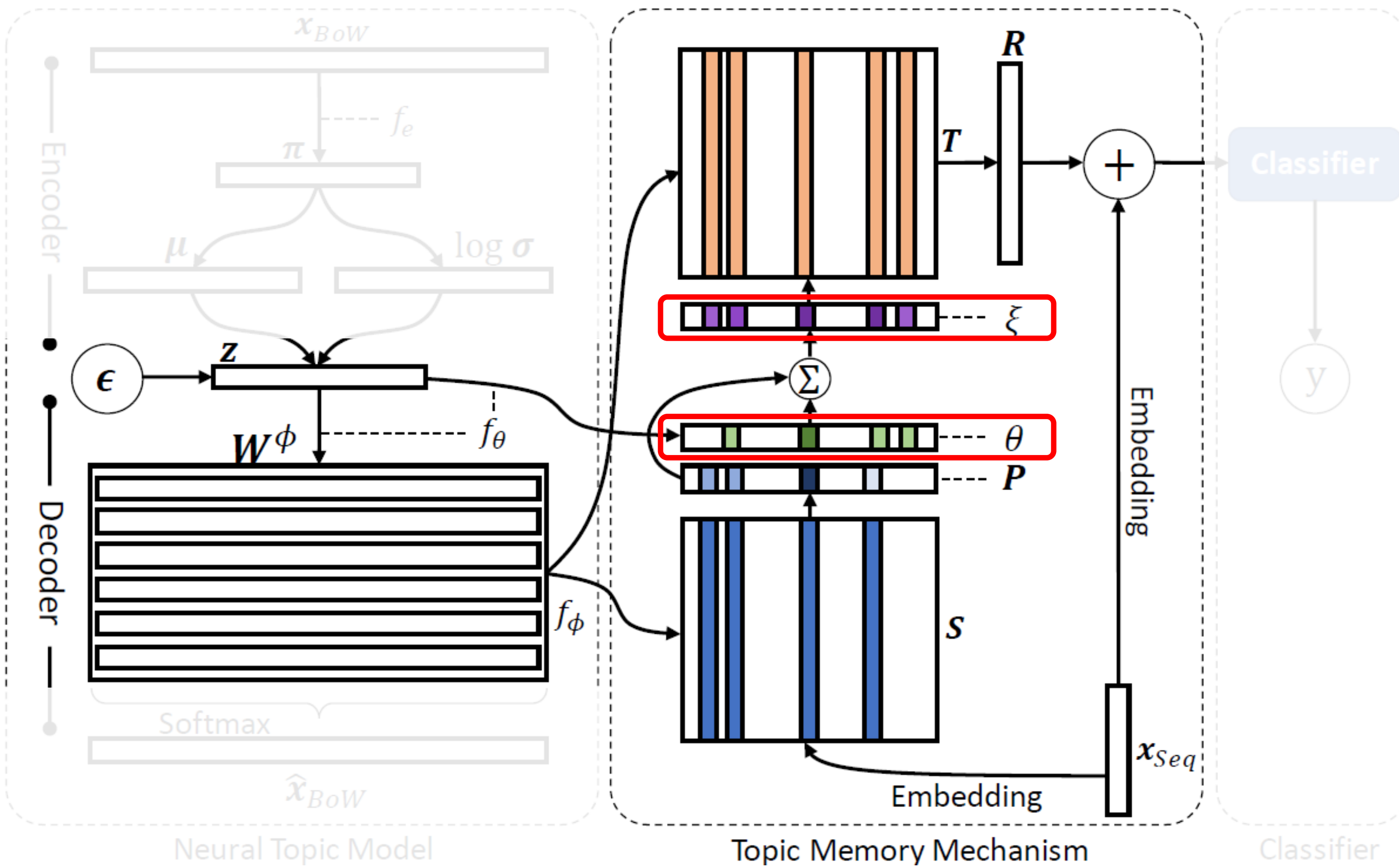
Topic Memory Network



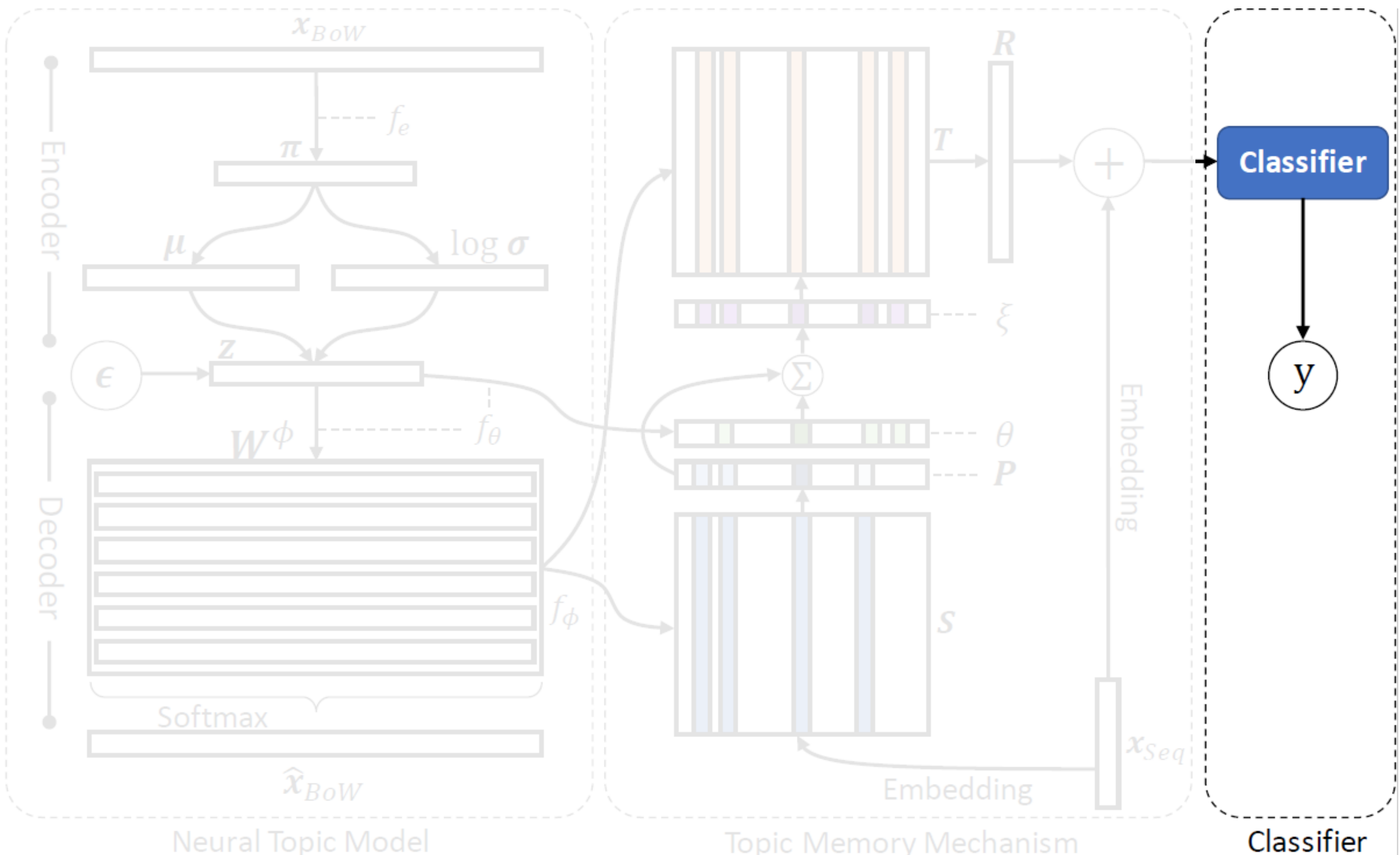
Topic Memory Network



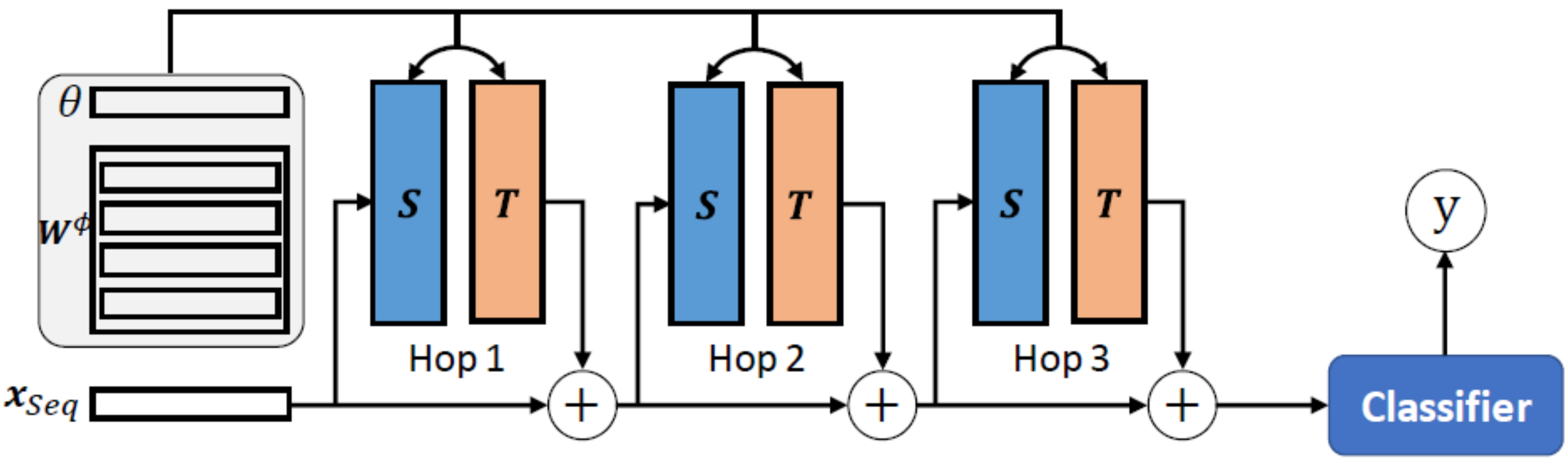
Topic Memory Network



Topic Memory Network



Multi-hop Topic Memory Networks



Topic memory networks with three hops

Learning Objective

- Loss function of NTM (ELBO loss)

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

- Loss function of classification (cross entropy)

$$\mathcal{L}_{CLS} = -\sum_{c} y_c log\big(p(y_c|x)\big)$$

- Overall loss function

$$\mathcal{L} = \mathcal{L}_{NTM} + \lambda\mathcal{L}_{CLS}$$

Dataset

Dataset	# of classes	# of docs	Avg len per doc	Vocab size
TagMyNews	7	32,567	8	9,433
Snippets	8	12,332	17	7,334
Twitter	50	15,056	5	6,962
Weibo	50	21,944	6	10,121

80% training, 10% development, 10% testing.

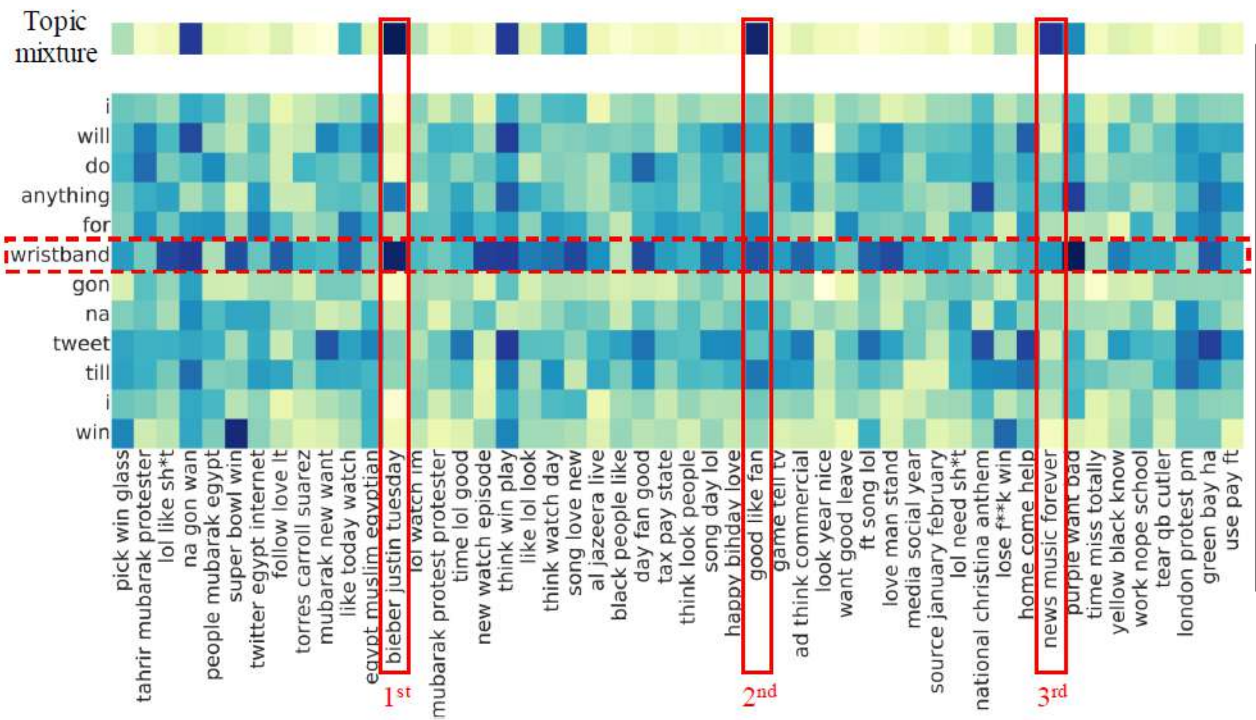
We use **hashtags (#)** as the classification labels for Twitter and Weibo dataset.

Classification Results

Models	Snippets		TagMyNews		Twitter		Weibo	
	Acc	Avg F1	Acc	Avg F1	Acc	Avg F1	Acc	Avg F1
Comparison models								
Majority Vote	0.202	0.068	0.247	0.098	0.073	0.010	0.102	0.019
SVM+BOW (Wang and Manning, 2012)	0.210	0.080	0.259	0.058	0.070	0.009	0.116	0.039
SVM+LDA (Blei et al., 2003)	0.689	0.694	0.616	0.593	0.159	0.111	0.192	0.147
SVM+BTM (Yan et al., 2013)	0.772	0.772	0.686	0.677	0.232	0.164	0.331	0.277
SVM+NTM (Miao et al., 2017)	0.779	0.776	0.664	0.654	0.261	0.177	0.379	0.348
AttBiLSTM (Zhang and Wang, 2015)	0.943	0.943	0.838	0.828	0.375	0.348	0.547	0.547
CNN (Kim, 2014)	0.944	0.944	0.843	0.843	0.381	0.362	0.553	0.550
CNN+TEWE (Ren et al., 2016)	0.944	0.944	0.846	0.846	0.385	0.368	0.537	0.532
CNN+NTM	0.945	0.945	0.844	0.844	0.382	0.365	0.556	0.556
Our models								
TMN (*Separate TM Inference*)	0.961	0.961	0.848	0.847	0.394	**0.386**	0.568	0.569
TMN (*Joint TM Inference*)	**0.964**	**0.964**	**0.851**	**0.851**	**0.397**	0.375	**0.591**	**0.589**

Case Study

S: [New.Music.Live] I will do anything for wristbands gonna twitter till I win.



Topic memory visualization for the test instance

1st Topic	2nd Topic	3rd Topic
bieber	good	news
justin	like	music
tuesday	fan	forever
newmusiclive	look	change
think	tweet	right
thought	thing	attack
jb	know	time
new	rumble	wing
music	na	aistdirect
live	right	films

Top-10 words of these topics indicated by topic-word weights ϕ

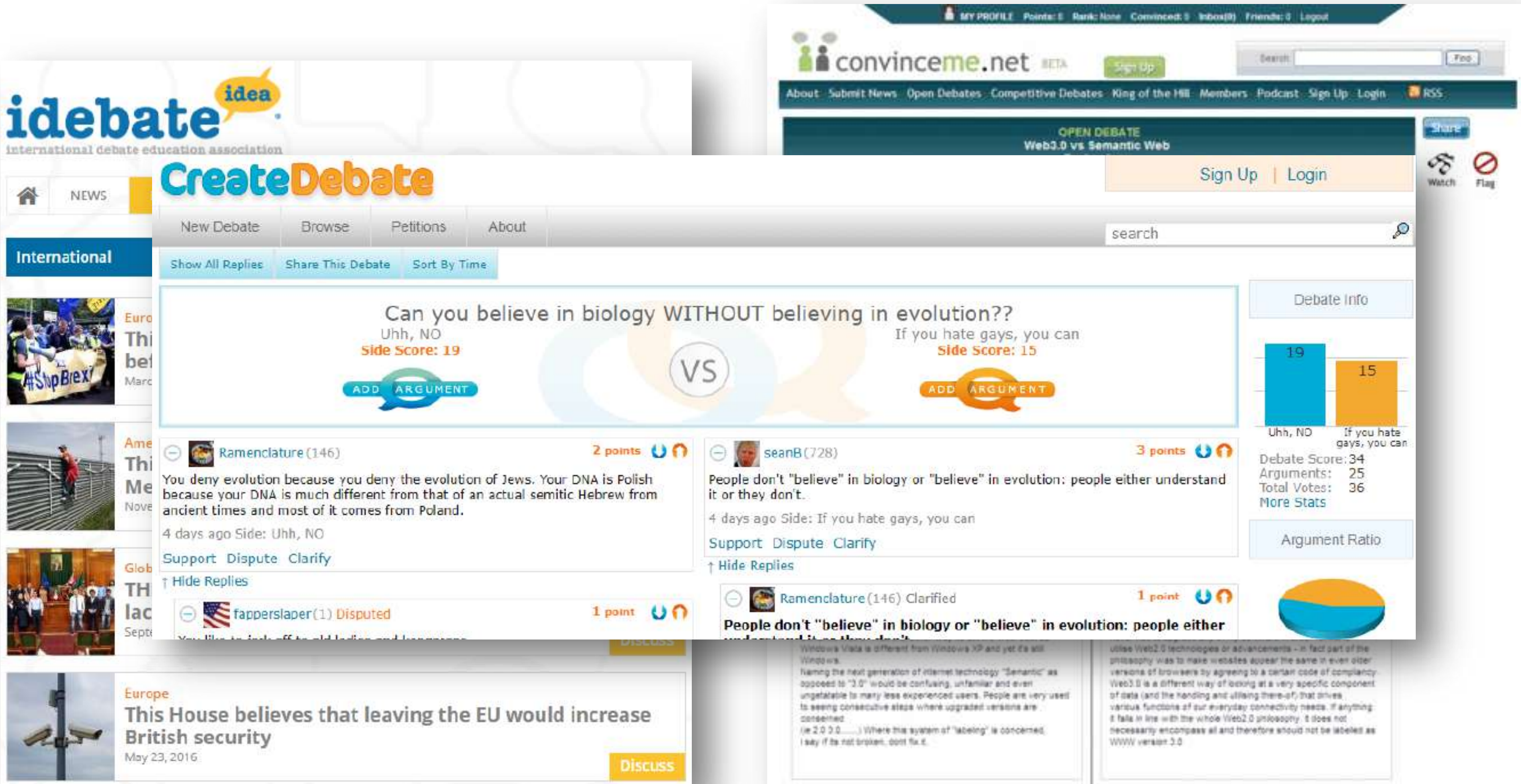
Summary

- We propose topic memory network framework for short text classification which can alleviate data sparsity issue for short text.
- We evaluate our model in 4 benchmark datasets 0.5%-3.5% abs accuracy increasement.
- To the best of our knowledge, we are the first to jointly explore topic modeling and classification in a deep learning framework.

Outline

- Topic 1: Microblog Conversation Modeling
- Topic 2: Short text classification
- Topic 3: Argumentation Persuasiveness Analysis
- Conclusion and Future Work

Online Argumentation



Persuasiveness Analysis is Challenging

Prompt: Is the school uniform a good or bad idea?
Stance: Good!

Argument 1	**Argument 2**
I think it's good within certain limits. I went to a school with a uniform, and it was far less stressful than non-uniform college.	Student victimization is likely to be lowered and fights and gang activity should be decreased.

- Evidences, facts
- Syntax, rhetoric
- Emotional aspects
- …

*"The aim of argument, or of discussion, should not be victory, but **progress**."*

—— Joseph Joubert, French essayist
1754 - 1824

Argumentation Process

...

A_1 [*Evidence*]: ... There is research that indicates "that *those who spoke two or more languages had significantly better cognitive abilities compared to what would have been expected from their baseline test*." ⟨url⟩. ... Another study found that " *the language-learning participants ended up with increased density in their grey matter and that their white matter tissue had been strengthened.* " ⟨url⟩

A_2 [*Metaphor*]: The common comparison is made to learning music, as /u/awesomeosprey has pointed out. I did some research into the matter. It seems that *learning a musical instrument does have long-lasting benefits* (⟨url⟩) *that relate to "higher-order aspects of cognition."*

...

A_4 [*Reference*] ... But a quick search and I have other sources: ⟨digit⟩ ⟨url⟩, ⟨digit⟩ ⟨url⟩, ⟨digit⟩ ⟨url⟩. The most interesting study is this one (⟨url⟩), but I can't find a complete version of it, sorry. /n/nNote: Study ⟨digit⟩ has an exceptionally small sample size. It's still interesting reading.

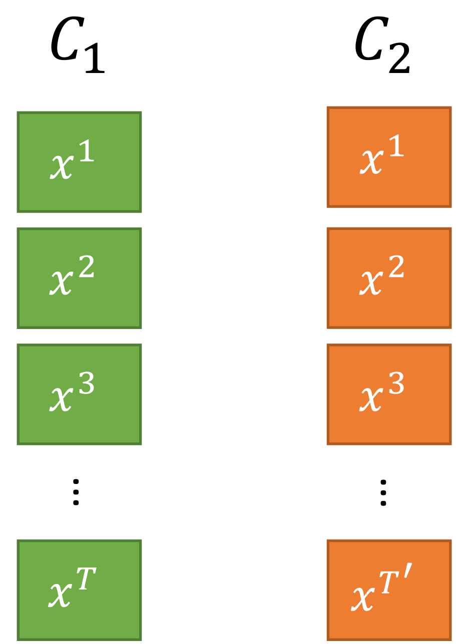
Against "Learning a second language isn't worth it for most people anymore"

Existing work

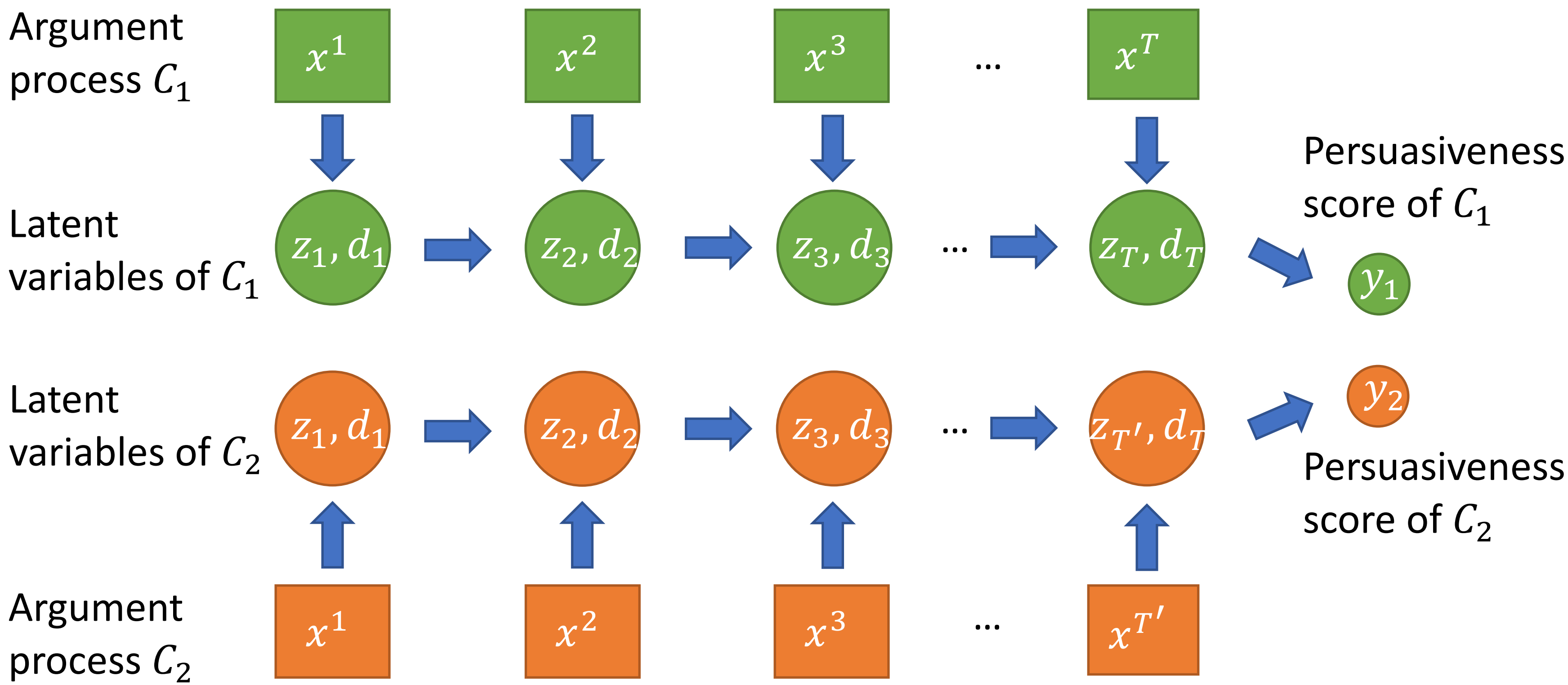
- Argumentation Persuasiveness
  - Without considering argumentation process [Wei et al., 2016, Habernal et al., 2016, Jo et al., 2018]
  - Crafting hand-made features, require labor-intensive feature engineering [Tan et al., 2016, Hidey et al., 2017, Niculae et al., 2017]
  - Without deep understanding the argumentation process [Zhang et al., 2016, Hidey et al., 2018]
- Conversation Process Understanding
  - Unsupervised modeling conversation, did not focus on argument persuasiveness [Ritter et al., 2010, Joty et al., 2011, Qin et al., 2017, Zeng et al., 2019]
  - Without considering the latent key factors [Kumar et al., 2016, Zhang et al., 2017]

Problem Setup

- Given two conversational argument processes (C_1 and C_2) from the same debate D, each one is consisted of a sequence of argumentative turns ($C = \{x^1, \dots x^T\}$).
- Goals:
  - predict which one is more convincing/persuasive.
  - Extract the key factors of persuasiveness and their changes in the argument process.

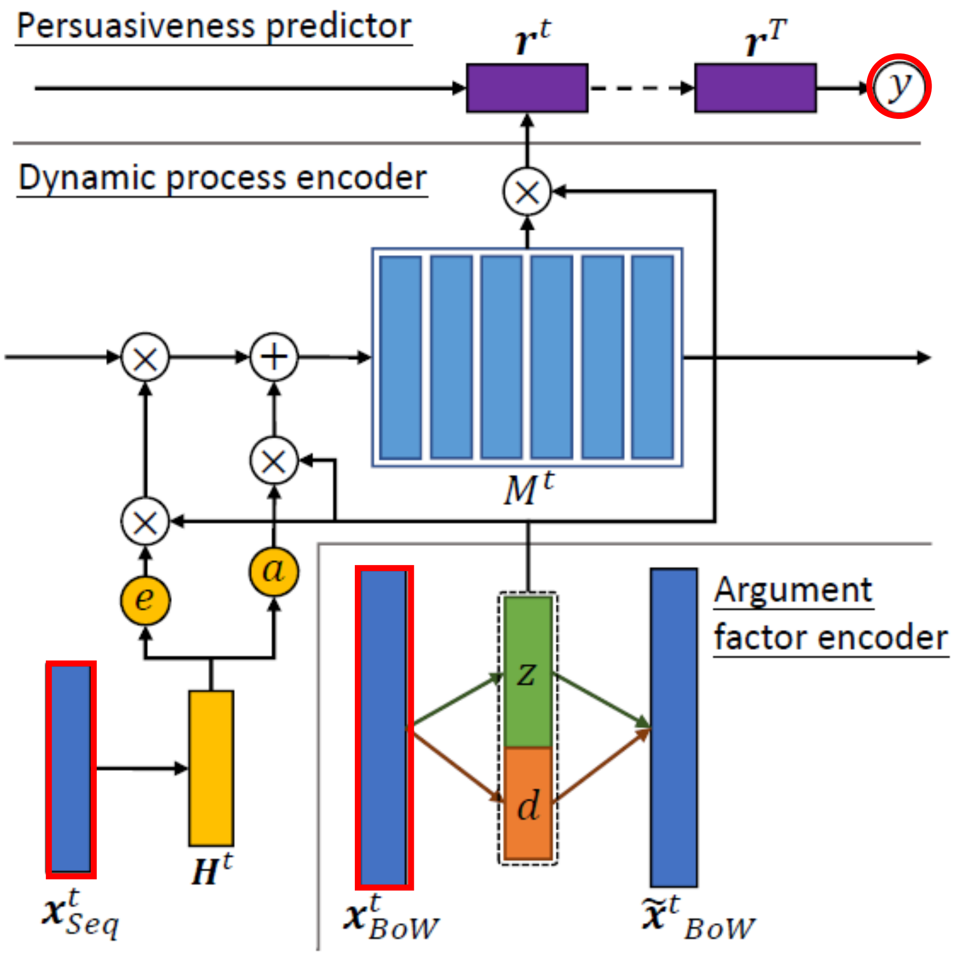


Model Intuition



If C_1 is more persuasive than C_2, we have $y_1 > y_2$, else $y_1 < y_2$.

Dynamic Topic/Discourse Memory Network



- Argument Factor Encoder

$$z^t \leftarrow x^t_{BoW}, d^t \leftarrow x^t_{BoW}$$

- Dynamic Process Encoder
  - Memory weight

$$w^t = [z^t; d^t],$$

  - Memory state update

$$e^t = sigmoid\left(f_e(H^t)\right),$$
$$a^t = \tanh\left(f_a(H^t)\right),$$
$$M_i^t = M_i^{t-1}\left[1 - w_i^t e^t\right] + w_i^t a^t$$

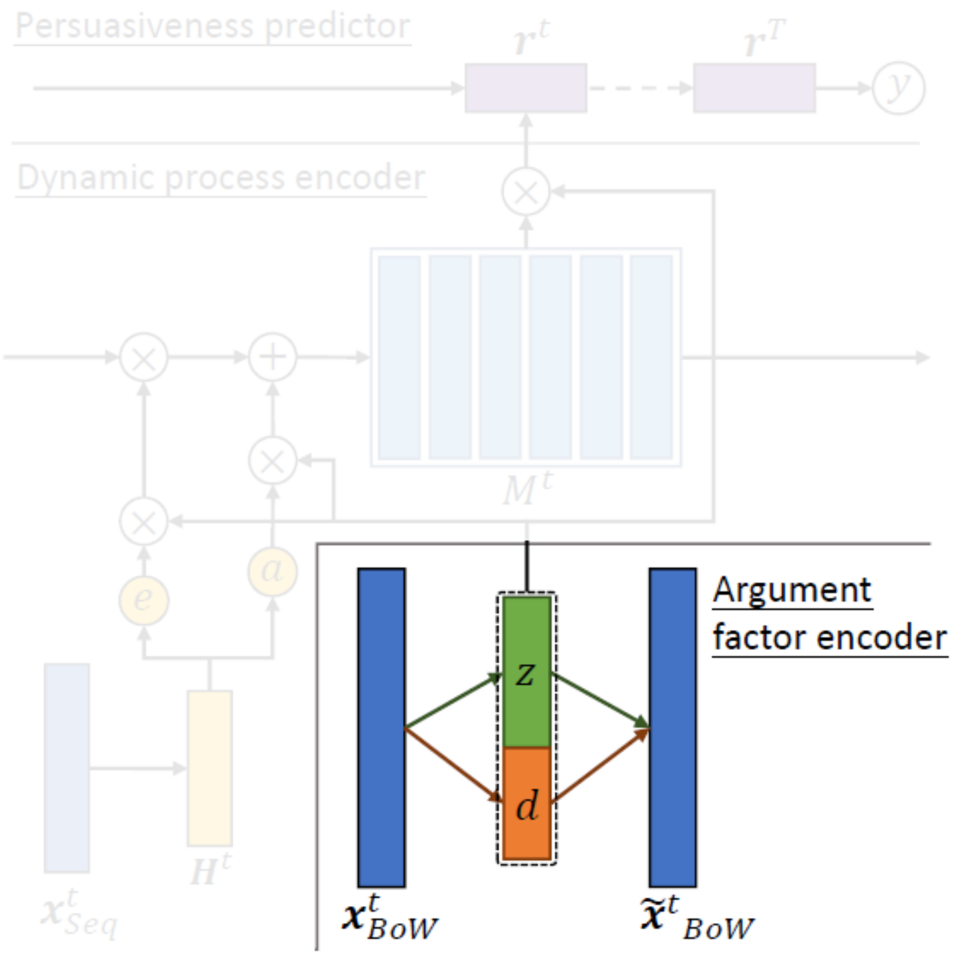
  - Memory read content

$$r^t = \sum_{i=1} w_i^t M_i^t$$

- Persuasiveness Predictor

$$r = RNN(\{r^t\})$$
$$y = f_y(r)$$

Dynamic Topic/Discourse Memory Network



- Argument Factor Encoder

  $$z^t \leftarrow x^t_{BoW}, d^t \leftarrow x^t_{BoW}$$

- Dynamic Process Encoder
  - Memory weight

    $$w^t = [z^t; d^t],$$

  - Memory state update

    $$e^t = sigmoid\left(f_e(H^t)\right),$$
    $$a^t = \tanh\left(f_a(H^t)\right),$$
    $$M_i^t = M_i^{t-1}\left[1 - w_i^t e^t\right] + w_i^t a^t$$

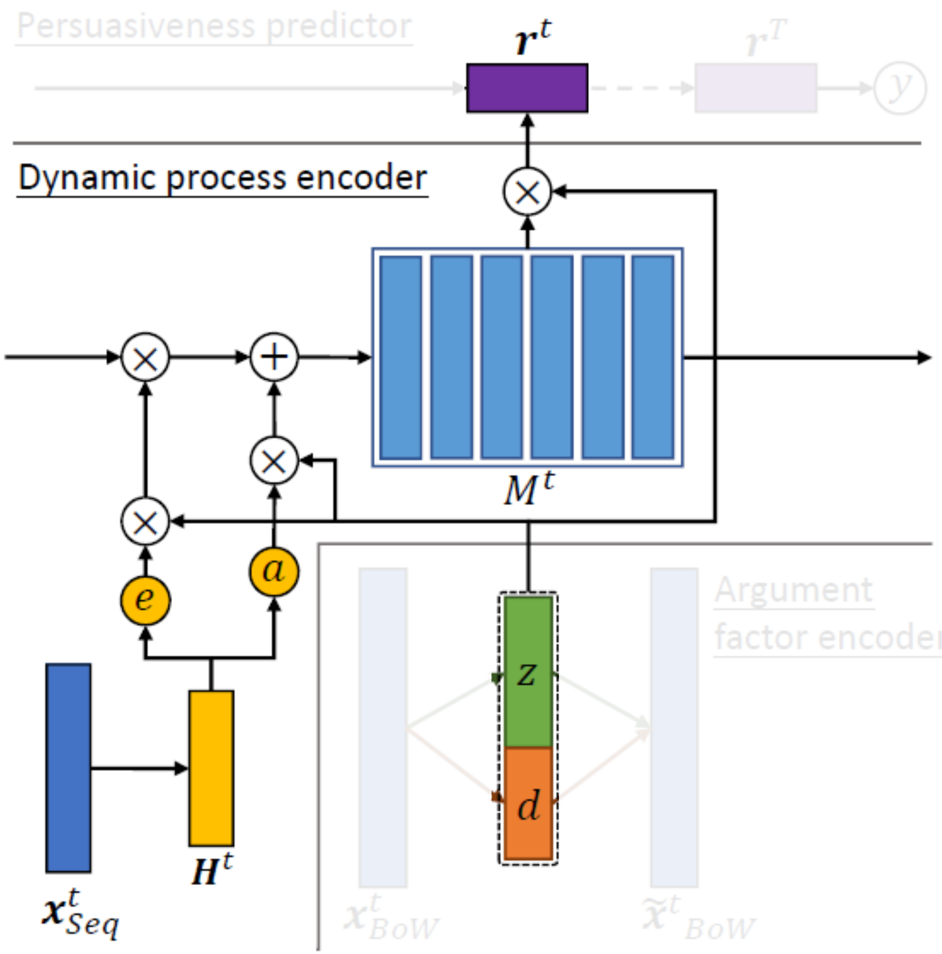
  - Memory read content

    $$r^t = \sum_{i=1} w_i^t M_i^t$$

- Persuasiveness Predictor

  $$r = RNN(\{r^t\})$$
  $$y = f_y(r)$$

Dynamic Topic/Discourse Memory Network



- **Argument Factor Encoder**

  $$z^t \leftarrow x^t_{BoW}, d^t \leftarrow x^t_{BoW}$$

- **Dynamic Process Encoder**
  - Memory weight

    $$w^t = [z^t; d^t],$$

  - Memory state update

    $$e^t = sigmoid\left(f_e(H^t)\right),$$
    $$a^t = \tanh\left(f_a(H^t)\right),$$
    $$M_i^t = M_i^{t-1}\left[1 - w_i^t e^t\right] + w_i^t a^t$$

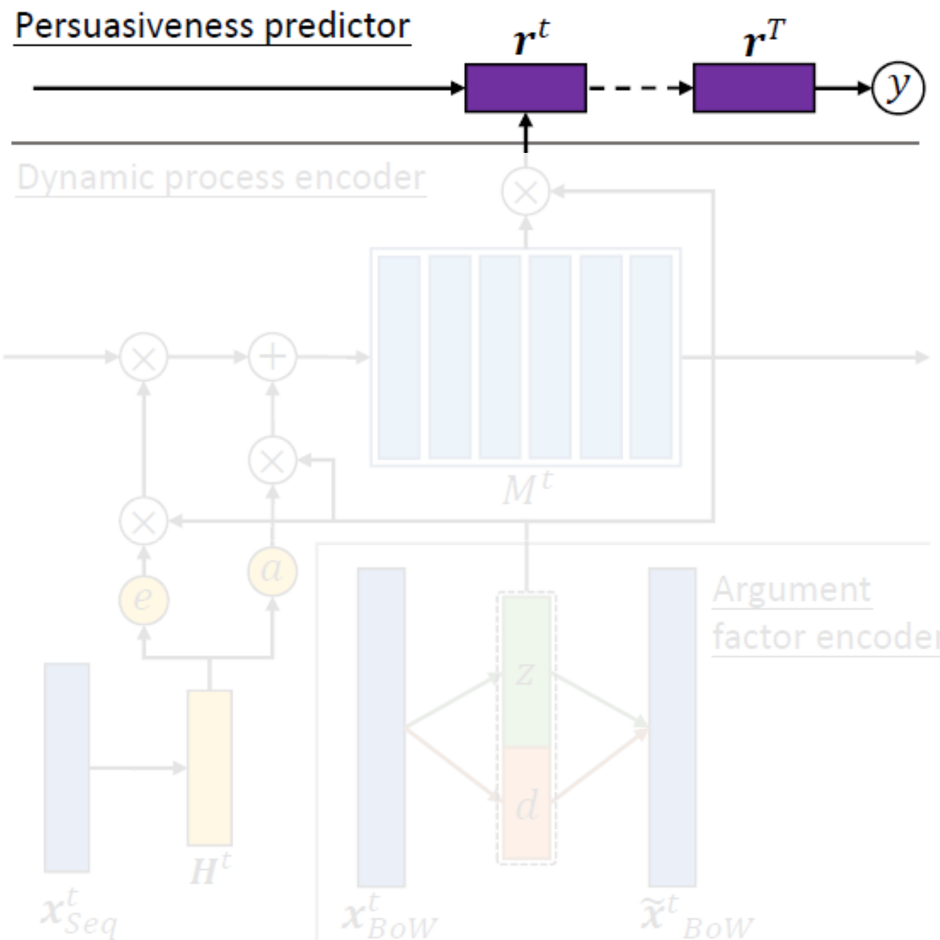
  - Memory read content

    $$r^t = \sum_{i=1} w_i^t M_i^t$$

- Persuasiveness Predictor

  $$r = RNN(\{r^t\})$$
  $$y = f_y(r)$$

Dynamic Topic/Discourse Memory Network



- **Argument Factor Encoder**

  $$z^t \leftarrow x^t_{BoW}, d^t \leftarrow x^t_{BoW}$$

- **Dynamic Process Encoder**
  - Memory weight

    $$w^t = [z^t; d^t],$$

  - Memory state update

    $$e^t = sigmoid\big(f_e(H^t)\big),$$
    $$a^t = \tanh\big(f_a(H^t)\big),$$
    $$M^t_i = M^{t-1}_i\big[1 - w^t_i e^t\big] + w^t_i a^t$$

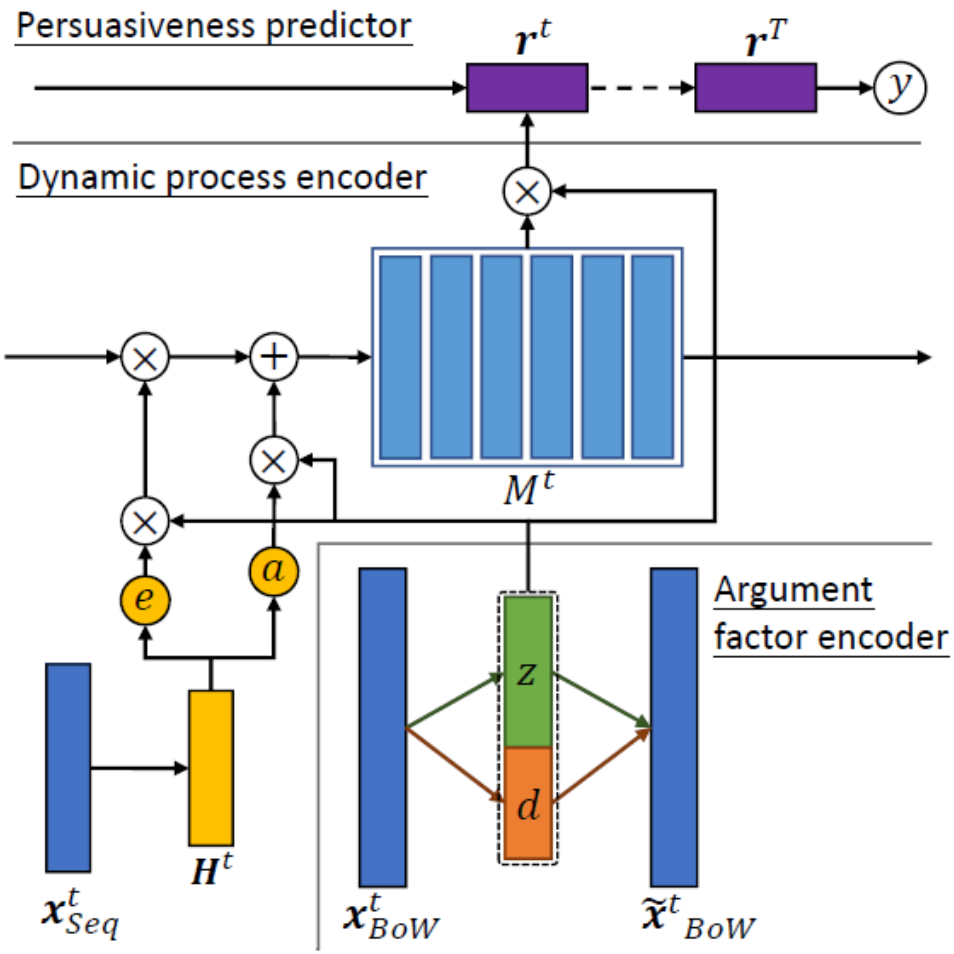
  - Memory read content

    $$r^t = \sum_{i=1} w^t_i M^t_i$$

- **Persuasiveness Predictor**

  $$r = RNN(\{r^t\})$$
  $$y = f_y(r)$$

Dynamic Topic/Discourse Memory Network



- **Training Losses**
  - ➢ Argument factor loss

$$\mathcal{L}_{Factor} = \mathcal{L}_z + \mathcal{L}_d + \mathcal{L}_x - \lambda \mathcal{L}_{MI}$$

  - ➢ Persuasiveness prediction loss

$$\mathcal{L}_{Pred} = \log(1 + \exp(y^- - y^+))$$

- **Final Objective**

$$\mathcal{L} = \mathcal{L}_{Pred} - \gamma \sum_t \mathcal{L}^t_{Factor}$$

Dataset

Datasets	# of moots	# of convs	# of turns	Avg. turns per conv	Avg. words per turn	|Vocab|
CMV	2,396	10,341	39,644	3.8	96.2	13,541
Court	204	655	17,599	26.9	46.1	6,260

80% training, 10% development, 10% testing.

Persuasiveness Prediction

Models	CMV		Court	
	Acc.	F1	Acc.	F1
Baselines				
LR-TFIDF	0.571	0.727	0.631	0.773
JTDM	0.615	0.762	0.642	0.782
HATT-RNN	0.828	0.890	0.559	0.717
DMN	0.858	0.893	0.662	0.755
DKVMN	0.896	0.911	0.726	0.841
Our models				
W/O TOPIC	0.871	0.931	0.797	0.887
W/O DISCOURSE	0.922	0.959	0.821	0.902
W/O MEMORY	0.885	0.918	0.761	0.864
FULL MODEL	**0.939**	**0.968**	**0.833**	**0.909**

Pairwise classification results on persuasiveness prediction.

Case Study

...

A_1 [*Evidence*]: ... There is research that indicates "that *those who spoke two or more languages had significantly better cognitive abilities compared to what would have been expected from their baseline test.*" ⟨url⟩. ... Another study found that " *the language-learning participants ended up with increased density in their grey matter and that their white matter tissue had been strengthened.* " ⟨url⟩

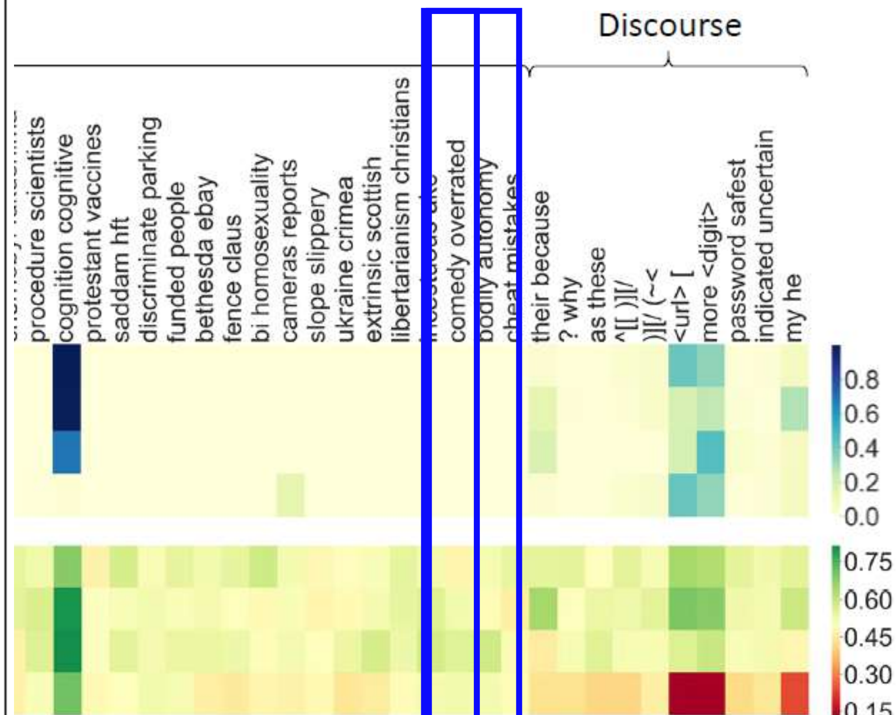
A_2 [*Metaphor*]: The common comparison is made to learning music, as /u/awesomeosprey has pointed out. I did some research into the matter. It seems that *learning a musical instrument does have long-lasting benefits (⟨url⟩) that relate to "higher-order aspects of cognition."*

...

A_4 [*Reference*] ... But a quick search and I have other sources: ⟨digit⟩ ⟨url⟩, ⟨digit⟩ ⟨url⟩, ⟨digit⟩ ⟨url⟩. The most interesting study is this one (⟨url⟩), but I can't find a complete version of it, sorry. /n/nNote: Study ⟨digit⟩ has an exceptionally small sample size. It's still interesting reading.

(a)

(b)



ndicate topics shift and discourse flow.

c or discourse.

Summary

- We propose to dynamically track both topics and discourse factors in conversational argumentation for persuasiveness analysis.
- We achieve substantial improvement in persuasiveness prediction.

Outline

- Topic 1: Modeling Microblog Conversation
- Topic 2: Short Text Classification
- Topic 3: Argumentation Persuasiveness Analysis
- Conclusion and Future Work

Conclusion

Contributions

— Microblog Conversation Modeling

- ➢ Unsupervised neural framework for modeling topics and discourse

— Short Text Classification

- ➢ Topic memory mechanism to alleviate data sparsity issue

— Argumentation Persuasiveness Analysis

- ➢ Reveal the key factors of persuasiveness in argumentation process

Future Work

- Topic, Discourse and Sentiment-Aware Social Chatbot

Future Work

- Conversational Text-to-SQL

D_1 :	Database about student dormitories containing 5 tables	
Q_1 :	What are the names of all the dorms?	INFORM SQL
S_1 :	`SELECT dorm_name FROM dorm`	
A_1 :	(Result table with many entries)	
R_1 :	This is the list of the names of all the dorms.	CONFIRM SQL
Q_2 :	Which of those dorms have a TV lounge?	INFORM_SQL
S_2 :	`SELECT T1.dorm_name FROM dorm AS T1 JOIN has_amenity AS T2 ON T1.dormid = T2.dormid JOIN dorm_amenity AS T3 ON T2.amenid = T3.amenid WHERE T3.amenity_name = 'TV Lounge'`	
A_2 :	(Result table with many entries)	
R_2 :	This shows the names of dorms with TV lounges.	CONFIRM_SQL
Q_3 :	What dorms have no study rooms as amenities?	AMBIGUOUS
R_3 :	Do you mean among those with TV Lounges?	CLARIFY
Q_4 :	Yes.	AFFIRM

Publications

1. Jichuan Zeng, Jing Li, Yulan He, Cuiyun Gao, Michael R. Lyu, and Irwin King. What You Say and How You Say it: Joint Modeling of Topics and Discourse in Microblog Conversations Proceedings of the Transactions of the Association for Computational Linguistics (TACL), 2019 (oral presented in ACL 2019).
2. Cuiyun Gao, Jichuan Zeng, Xin Xia, David Lo, Michael R. Lyu, and Irwin King. RRGen: Automating App Review Response Generation. Proceedings of the 34th IEEE/ACM International Conference on Automated Software Engineering (ASE), 2019.
3. Cuiyun Gao, Wujie Zheng, Yutang Deng, David Lo, Jichuan Zeng, Michael R. Lyu, and Irwin King. Emerging App Issue Identification from User Feedback: Experience on WeChat. Proceedings of the 41th International Conference on Software Engineering (ICSE), 2019.
4. Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R. Lyu, and Irwin King. Topic Memory Networks for Short Text Classification. Proceedings of the 28th International Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019 (oral).
5. Cuiyun Gao, Jichuan Zeng, David Lo, Chin-Yew Lin, Michael R. Lyu, and Irwin King. INFAR: Insight Extraction from App Reviews. Proceedings of the 26th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE), demo track, 2018.
6. Cuiyun Gao, Jichuan Zeng, Federica Sarro, Michael R. Lyu, and Irwin King. Exploring the Effects of Ad Schemes on the Performance Cost of Mobile Phones. Proceedings of the 1st International Workshop on Advances in Mobile App Analysis (A-Mobile), co-located with (ASE), 2018.

Publications

7. Cuiyun Gao, Jichuan Zeng, Michael R. Lyu, and Irwin King. Online App Review Analysis for Identifying Emerging Issues. Proceedings of the 40th International Conference on Software Engineering (ICSE), 2018.

8. Jichuan Zeng, Haiqin Yang, Irwin King and Michael R. Lyu. A Comparison of Lasso-type Algorithms on Distributed Parallel Machine Learning Platforms. Distributed Machine Learning and Matrix Computations Workshop, 28th Annual Annual Conference on Neural Information Processing Systems (NIPS), workshop, 2014.

In preparation

1. Jichuan Zeng, Jing Li, Yulan He, Cuiyun Gao, Michael R. Lyu, Irwin King. What Change Your Mind: The Roles of Dynamic Topics and Discourse in Argumentation Process. Target at International World Wide Web Conferences (WWW), 2020.
2. Jichuan Zeng*, Cuiyun Gao*, David Lo, Zhiyuan Wen, Michael R. Lyu, Irwin King. Real-Time App Review Analysis via Online Joint Sentiment-Topic Tracing. Target at IEEE Transactions on Software Engineering (TSE).
3. Cuiyun Gao, Jichuan Zeng, David Lo, Xin Xia, Michael R. Lyu, and Irwin King. What Are Users Complaining about Mobile In-App Ads? An Empirical Study on In-App Ad Reviews. Target at IEEE Transactions on Software Engineering (TSE).

Thank you!

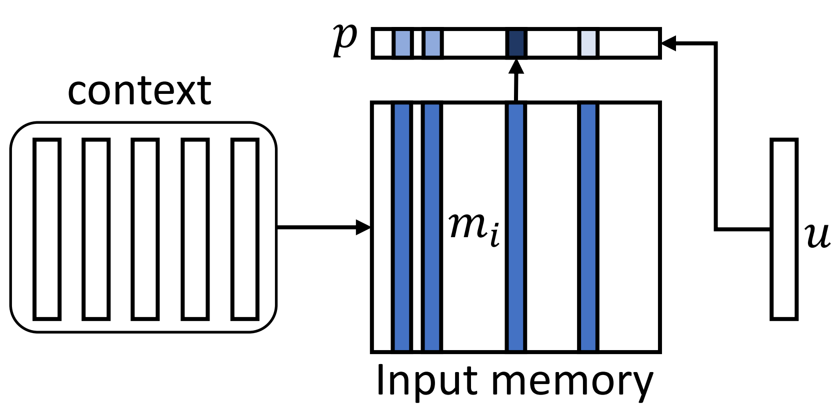
Q&A

Memory Networks

- Input memory

The match p between input embedding u and each memory slot m_i:

$$p_i = Softmax(u^T m_i)$$

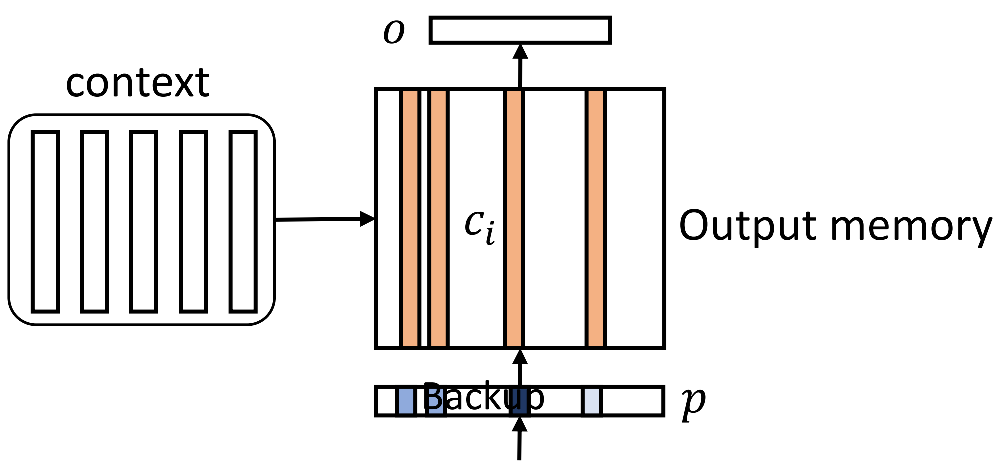


Memory Networks

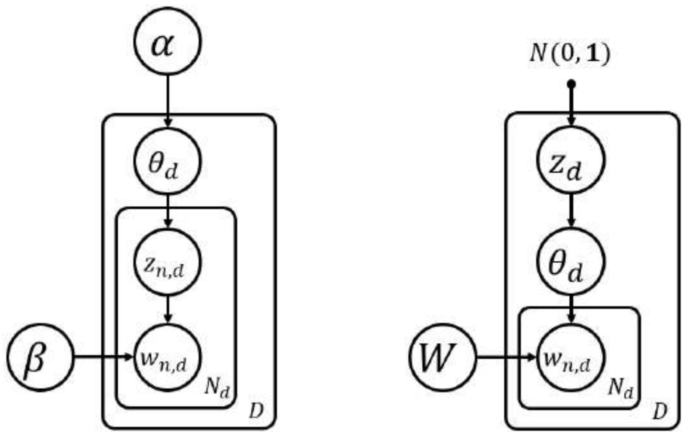
- Output memory

Compute the output vector o by the p weighted sum over the transformed input c_i:

$$o_p = \sum p_i c_i$$



LDA V.S. NTM

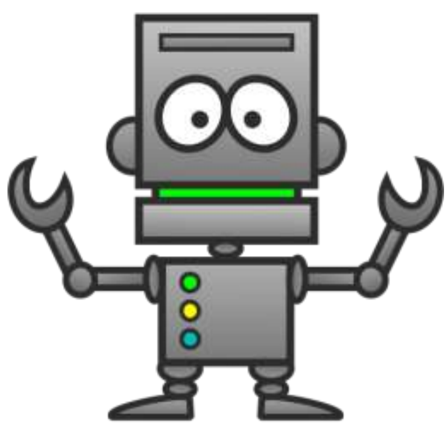


	LDA	NTM
Probabilistic model	Yes	Yes
Inference	Hard?	Easy
Discrete topic	Yes	Yes
Extensible	Hard	Yes

Natural Language Understanding

To understand a human language is to:

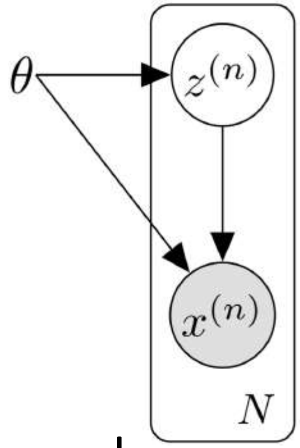
- Determine its category → Text categorization
- Give answer for a question → Question answering
- Transduce into another form → Sematic parsing
- …



Variational Inference

- Objective: Find model parameters θ that maximize the likelihood of the data.

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{n=1}^{N} \log p(x^n; \theta)$$



- Likelihood can be decomposed into lower bound and gap.

$$L(\theta) = \mathbb{E}_q \left[log \frac{p(x,z;\theta)}{q(z|x)} \right] + D_{KL}[q(z)|p(z|x)]$$

$$L_{ELBO} = \log p(x;\theta) - D_{KL}[q(z)|p(z|x)]$$

Gibbs Sampling

- A Markov-chain Monte Carlo (MCMC) approach to generate a sample from a joint distribution.
- MCMC methods get samples from a probability distribution based on constructing a Markov chain that has the desired distribution as its stationary distribution.
- Gibbs sampling is special case of Metropolis-Hastings.
- To be more efficient, LDA use collapsed Gibbs sampling:

$$p(z|\alpha, w, \beta) \propto p(z|\alpha)p(w|\beta, z)$$
$$= \int p(z|\theta)p(\theta|\alpha)d\theta \cdot \int p(w|B, z)p(B|\beta)dB$$

$$p(z_i = j|z_{-i}, w) \propto \frac{n_{-ij}^{w_i} + \beta}{n_{-ij}^{\cdot} + V\beta} \cdot \frac{n_{-ij}^{d} + \alpha}{n_{-i,\cdot}^{d} + K\alpha}$$

C_v Score

- Given a single pair $S_i\ = (W', W^*)$ of words or word subsets, C_v score measure how strong the W' and W^* are correlated.
- C_v Score is the aggregation indirect cosine measure with the NPMI.

$$C_v(W', W^*) = \frac{\log \frac{P(W', W^*) + \epsilon}{P(W') \cdot P(W^*)}}{-\log(P(W', W^*) + \epsilon)} + s_{cos}(\vec{v}_{m,r}(W'), \vec{v}_{m,r}(W^*))$$

[Röder et al., 2015]

Clustering Metrics

- Purity [Zhao and Karypis, 2001]

$$Purity = \sum_i \frac{|C_i|}{N} \max_j \frac{|C_i \cap L_j|}{|C_i|}$$

- Homogeneity [Rosenberg and Hirschberg, 2007]

$$Homogeity = 1 - \frac{H(L|C)}{H(L)}$$

where $H(L|C) = -\sum_c \sum_l \frac{a_{l,c}}{N} log \frac{a_{l,c}}{\sum_l a_{l,c}}$

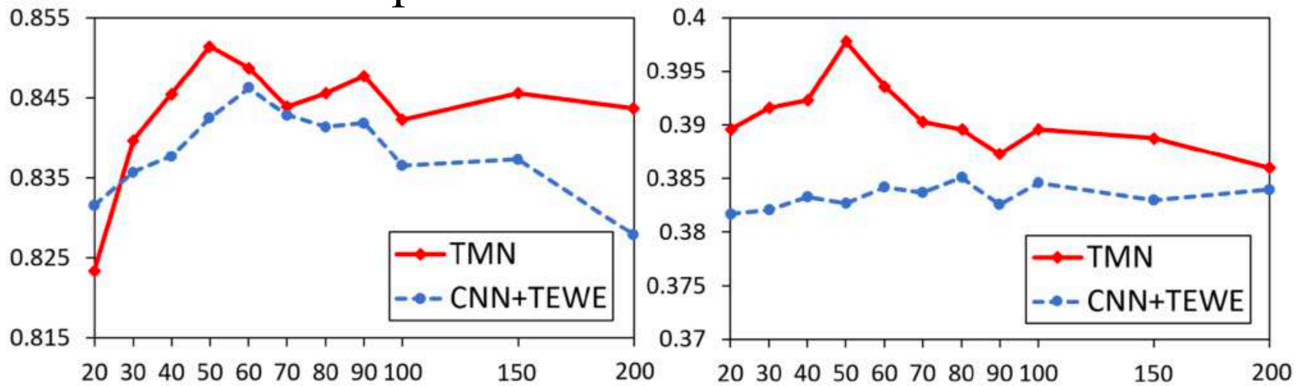
$$H(L) = -\sum_l \frac{\sum_c a_{l,c}}{N} log \frac{\sum_c a_{l,c}}{N}$$

- Variation of Information [Goldwater and Griffiths, 2007]

$$VI = H(L|C) + H(C|L)$$

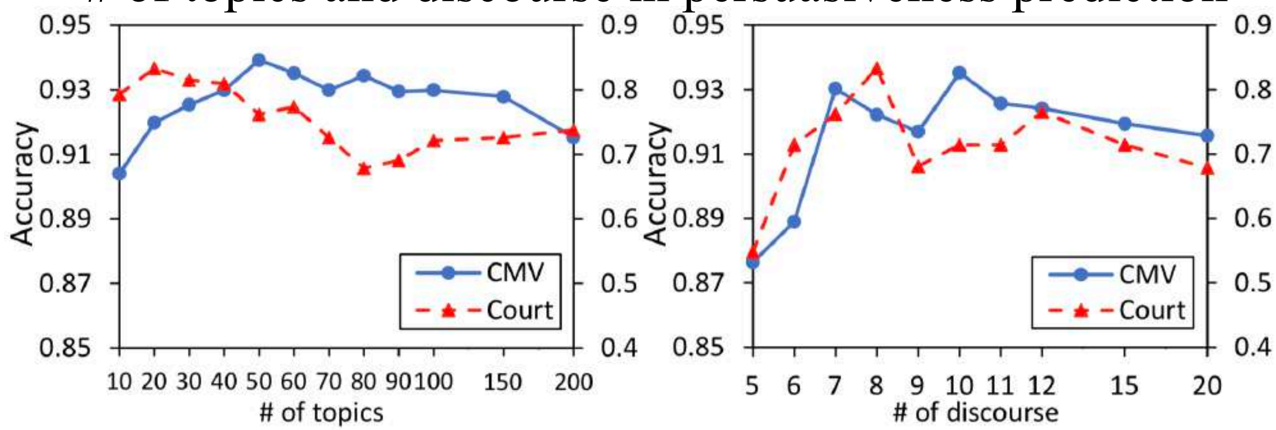
Hyper-parameters for T2, T3

of topics in short text classification



(a) TagMyNews (b) Twitter

# of topics and discourse in persuasiveness prediction

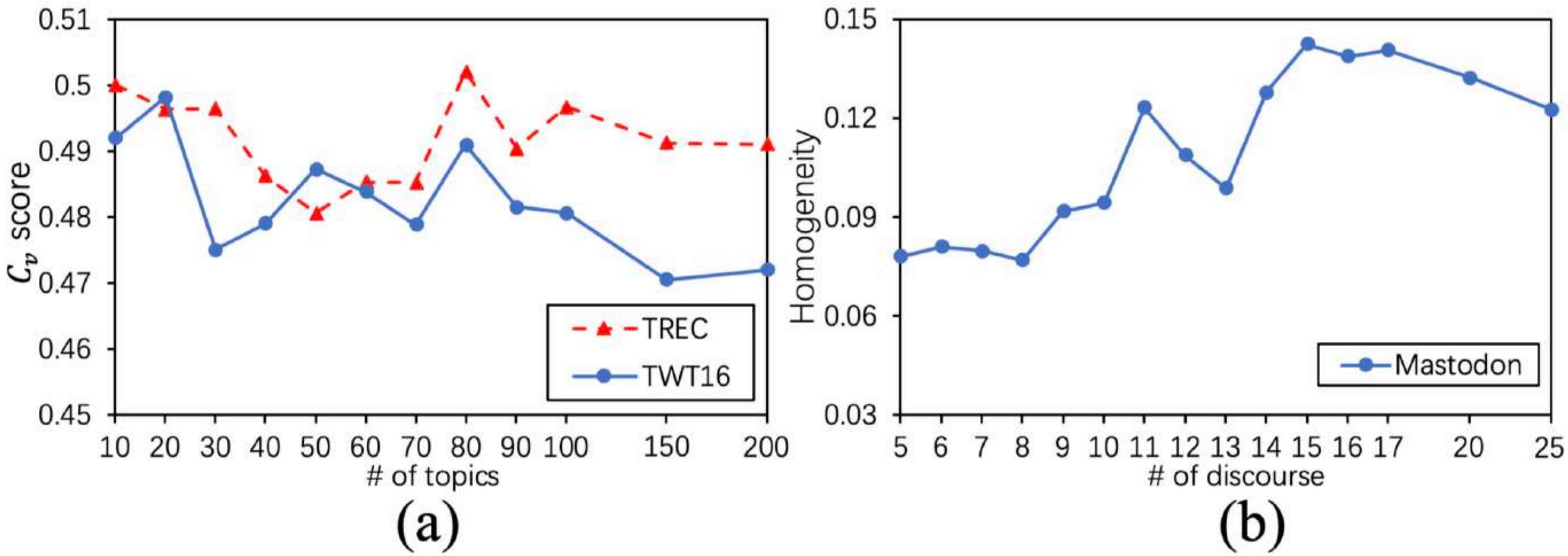


(a) (b)

Existing Work

- Topic Modeling for Social Media
  - Latent Dirichlet allocation (LDA) [Blei et al., 2003]
    - Not work well on short text messages
  - Short text topic modeling (BTM, LFDMM) [Yan et al., 2013, Nguyen et al., 2015]
    - Cannot use the richer context information in a conversation
  - Exploring heuristically aggregation [Hong et al., 2010, Ramage et al., 2010]
    - Manual defined aggregation strategies are unnatural
- Conversation Discourse
  - Discourse prediction & parsing [Stolcke et al., 2000, Ji et al., 2016]
    - High-quality labeled data are needed
  - Unsupervised discourse modeling [Ritter et al., 2010, Jotty et al., 2011, Zhao et al., 2018]
    - Did not consider the effect of conversation topics
  - Exploiting interactional structure and topics [Li et al., 2016]
    - Low efficiency, non-neural framework

Hyper-parameters for T1



- Relatively larger topic numbers are better for TREC (K=80).
- Small topic numbers are better for TWT16 (K=20).
- The optimum discourse number is the same with manually annotated benchmark.

Message Representations

Models	TREC		TWT16	
	Acc	Avg F1	Acc	Avg F1
Baselines				
BoW	0.120	0.026	0.132	0.030
TF-IDF	0.116	0.024	0.153	0.041
LDA	0.128	0.041	0.146	0.046
BTM	0.123	0.035	0.167	0.054
LF-DMM	0.158	0.072	0.162	0.052
NTM	0.138	0.042	0.186	0.068
Our model	**0.259**	**0.180**	**0.341**	**0.269**

Evaluation of tweet classification results of SVM, we use hashtags (#) as the classification labels.

Existing Work

- External Knowledge
  - Wikipedia, knowledge base [Jin et al. 2011, Lucia and Ferrari 2014, Wang et al. 2017]
    - ➢ Domain-specific, not work well in social media
  - Manually-crafted features [Pak and Paroubek 2010, Jiang et al. 2011]
    - ➢ Task-specific, not work well in general-purpose classification tasks
- Word Collocation Patterns
  - Word embeddings [Bowman et al. 2016, Krisknamurthy et al. 2017]
    - ➢ Word-level lexico-semantic, not for corpus
  - Topic models [Phan et al. 2008, Chen et al. 2011, Ren et al. 2016]
    - ➢ Need pre-trained topic model

Topic Coherence

- Quantitative analysis

Model	TagMyNews	Snippets	Twitter
LDA	0.449	0.436	0.436
BTM	0.463	0.435	0.435
NTM	0.468	0.463	0.463
TMN	**0.499**	**0.487**	**0.468**

C_V coherence scores

- Qualitative analysis

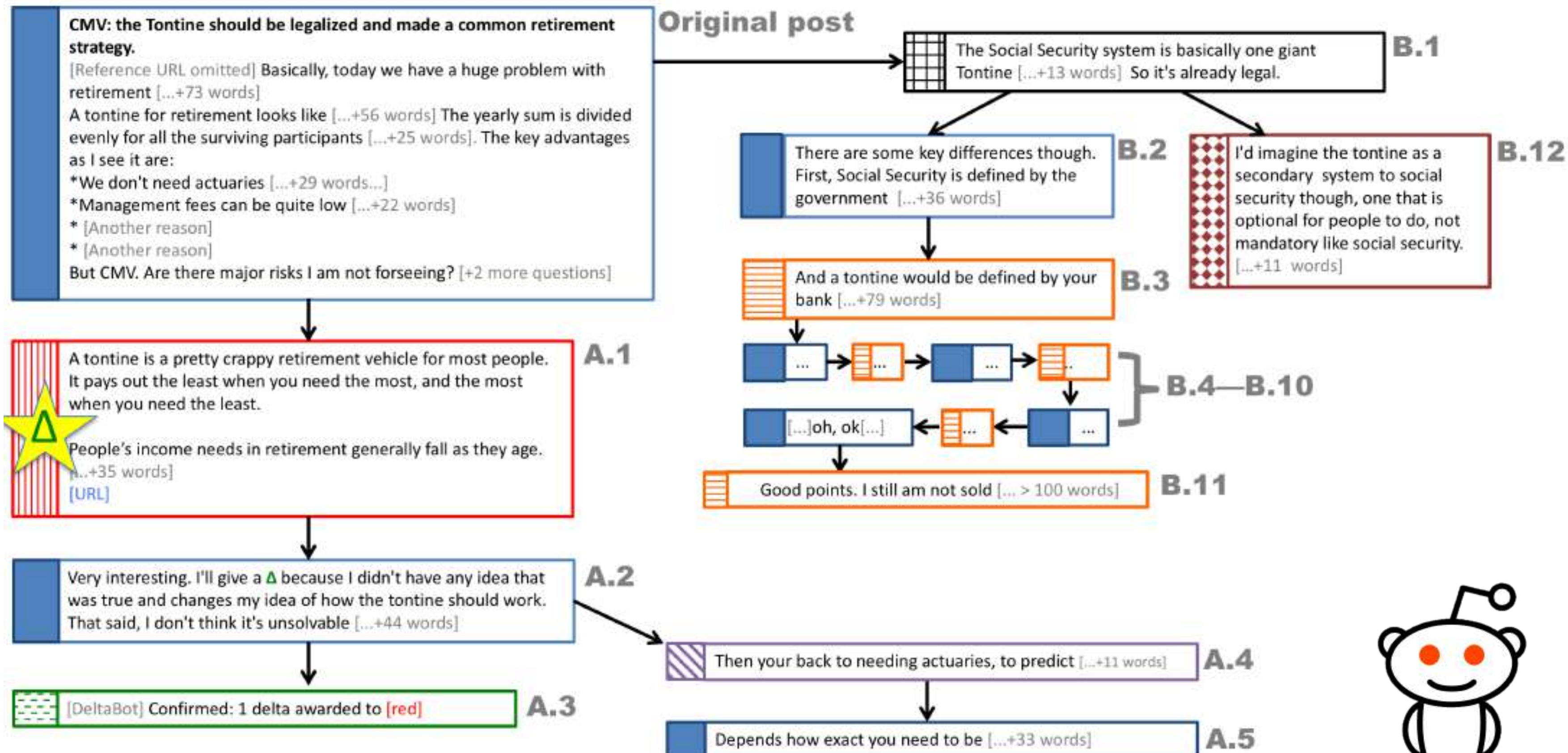
LDA	mubarak *bring run* obama democracy speech *believe* regime power <u>bowl</u>
BTM	mubarak egypt push internet people government *phone* hosni *need* son
NTM	mubarak people egyptian egypt *stay tomorrow* protest news *phone* protester
TMN	mubarak protest protester tahrir square egyptian al jazeera repo cairo

Top 10 representative terms of "Egyptian revolution of 2011". *Non-topic words* are italic and blue, and <u>off-topic words</u> are underlined and red.

Existing work

- Argument persuasiveness
  - Identifying convincing arguments or viewpoints[Wei et al., 2016, Habernal et al., 2016, Jo et al., 2018]
    - ➢ Without considering argumentation process.
  - Crafting hand-made features [Tan et al., 2016, Hidey et al., 2017, Niculae et al., 2017]
    - ➢ Require labor-intensive feature engineering, limited generalization ability
  - Argument sequence influence [Zhang et al., 2016, Hidey et al., 2018]
    - ➢ Without deep understanding the argumentation process
- Conversation process understanding
  - Modeling dynamic conversation [Ritter et al., 2010, Joty et al., 2011, Qin et al., 2017, Zeng et al., 2019]
    - ➢ Unsupervised model, not related to argument persuasiveness
  - Dynamic memory network [Kumar et al., 2016, Zhang et al., 2017]
    - ➢ Without considering the key factors

Reddit/ChangeMyView



[Tan et al., WWW16]

Supreme Court

Turns from *S. D. Warren Co. v. Maine Bd. of Environmental Protection* (04-1527)

JUSTICE SOUTER: -- "reinforcing," and maybe it's "changing." I mean, you're characterizing it one way. We start with a different canon of meaning, and that is that we look to the words around which, in connection with which, the word is used. In here, it's being used without certain modifiers or descriptive conditions. In other cases, it is being used with them. And that's a good reason to think that probably the word is intended to mean something different in those situations.

MR. KAYATTA: Well, I would -- I would hesitate, Justice Souter, to go from taking a specific word, like "discharge," and, therefore, saying that it meant something that is both more general and much more easily set.

JUSTICE SOUTER: No, but your argument, I thought, was simply this, that it uses "discharge" in, you know, X number -- I forget how many you had -- and it's perfectly clear that in most of those instances it requires an addition; and, therefore, it should be construed as requiring it here. My point was that in a great many of those instances, the statute is not merely using the word in isolation; it's using it in connection with a couple of other words, like "discharge a pollutant." And it, therefore, number one, makes sense to construe "discharge of a pollutant" differently from "discharge." That's the -- that's the only point.

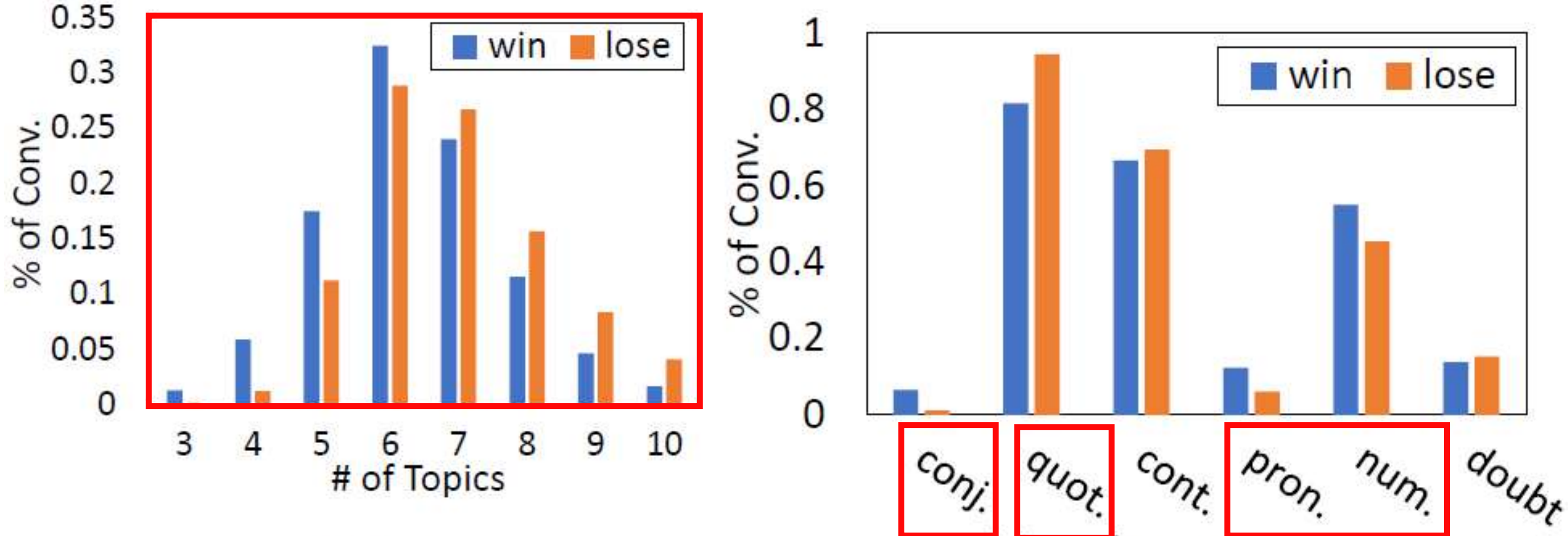
SUPREME COURT
OF THE UNITED STATES

[Danescu-Niculescu-Mizil et al., 2012]

Dataset Construction

- CMV [Tan et al. 2016]
  - Filter out threads without Δ or with too few participants.
  - Flatten the discussion threads into conversation paths.
  - Exclude messages from opinion holder.
  - For each Δ awarded conversation path, we randomly pick N non-Δ conversation paths in the same thread, forming N conversation pairs.
- Court [Danescu-Niculescu-Mizil et al., 2012]
  - Break the case into conversation paths.
  - Exclude messages from justices.
  - For each conversation path that win the justices' favor, we randomly pick N negative conversation paths in the same case, forming N conversation pairs.

Effects of Topics and Discourse



(a) Topic (b) Discourse

- Strong and focused argument points are better than diverse topics.
- Personal pronoun and numbers are more likely to appear in the winning side than the losing side.
- Conjunction words, though not widely used, is obviously more endorsed by winning sides.
- Losing sides are more in favor of the quotation discourse.

Suggestions for Better Persuasions

- Topics in argumentation are more important than discourse styles.
- Strong and focused argument points are better than diverse topics.
- When delivering arguments, well organize the points and address them in a modest and concrete way.

Posterior Collapse

- We say a posterior is collapsing, when signal from input x to posterior parameters is either too weak or too noisy, and as a result, decoder starts ignoring z samples drawn from the posterior $q_\phi(z|x)$.

RetardedCatfish 3Δ Score hidden · 7 hours ago

On the other hand perfect attendance encourages positive traits like stoicism, steadfastness, dedication, endurance and willpower while discouraging negative traits like weakness, feebleness, defeatism and abandonment

Ulysses S Grant is not remembered as a not a great man because he backed down and surrendered and stayed home when things were difficult and uncertain. He is remembered because he was bullheaded and stubborn and he kept going even when everyone doubted him and when it was dark and when his work was painful

Reply Share Report Save

kanyeBest11 Score hidden · 7 hours ago

While I do agree that it may promote certain good personality traits, I think there is other ways to promote those traits.

Promoting people to do certain clubs, take harder classes and working for better grades can also arguably promote those traits. It not only promotes those traits, it also teaches you that things in life can be difficult and it does it in a healthier way

Reply Share Report Save

RetardedCatfish Score hidden · 7 hours ago (13 children)

AseRayAes 3Δ Score hidden · 7 hours ago

Do you think it is bad for schools to use attendance as a motivator for students?

Reply Share Report Save

kanyeBest11 Score hidden · 7 hours ago

Depends on the situation , I think using perfect attendance promotes kids showing up to school ill. It isn't healthy for anyone, because it gets other kids sick and it teaches the perfect attendance kids that pandering to some stupid award is more important than your health

Reply Share Report Save