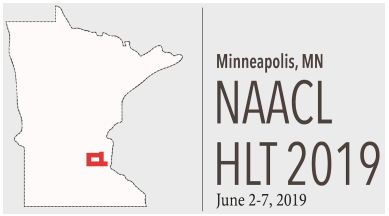
Predicting Helpful Posts in Open-Ended Discussion Forums: A Neural Architecture

Kishaloy Halder, Min-Yen Kan and Kazunari Sugiyama

June 5, 2019

NUS National University of Singapore
School of Computing



WING
Web Information Retrieval
Natural Language Processing Group

Photo Credits: https://www.pexels.com/

Online Discussion Forums

Learning from the community's collective wisdom

Users ask questions, share anecdotal observations

Others reply with relevant information or personal opinions

reddit

HealthBoards HEALTH MESSAGE BOARDS

Patient

coursera education for everyone

Photo Credits: https://www.pexels.com/

Discussion Forum
≠
Community Question Answering

CQA mostly receives factoid based questions

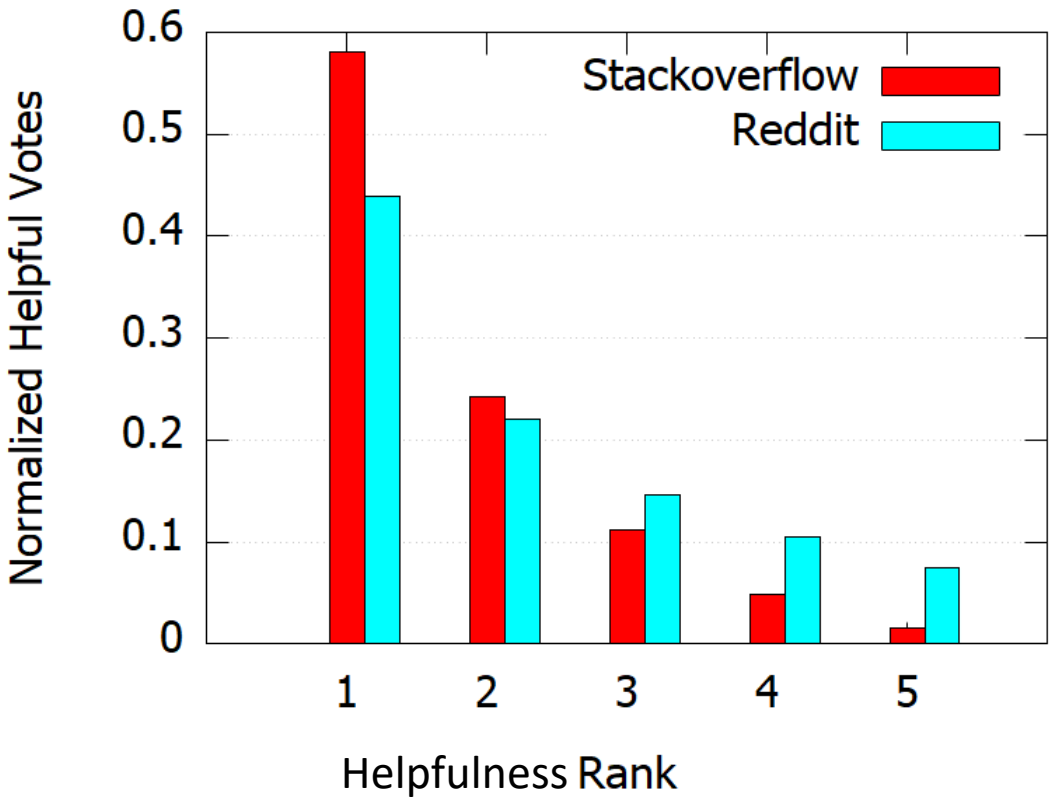
- Single correct answer

In contrast, in discussion forums, the thread opening post is not always a question

- Personal anecdotes, asking for recommendations
- Multiple "correct" answers

Threads are more subjective (open-ended) in discussion forums

Discussion Forum
≠
Community Question Answering



CQA mostly receives factoid based questions

- Single correct answer

In contrast, in discussion forums, the thread opening post is not always a question

- Personal anecdotes, asking for recommendations
- Multiple "correct" answers

Threads are more subjective (open-ended) in discussion forums

Predicting Helpful Posts from Discussion Threads

Task

Given post text, identify whether it is helpful to users

- Interested in the textual content of the posts only (not social media-style features, *e.g.*, followers etc)

Helpfulness: decided by user feedback

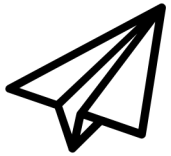
- "upvote", "like", "mark as helpful", "highlight"

Motivation:

- Early detection of helpful posts can aid to the recommendation process
- Can also help in summarizing long running threads

Discussion Thread

Order	Post Text	Helpful?
1	How to do X?	
2	Do you really need X?	No
3	Sorry, new here.	No
4	Sure, follow these steps…	Yes!
5	I can tell you about Y.	No



Notify users interested in X

Our Approach

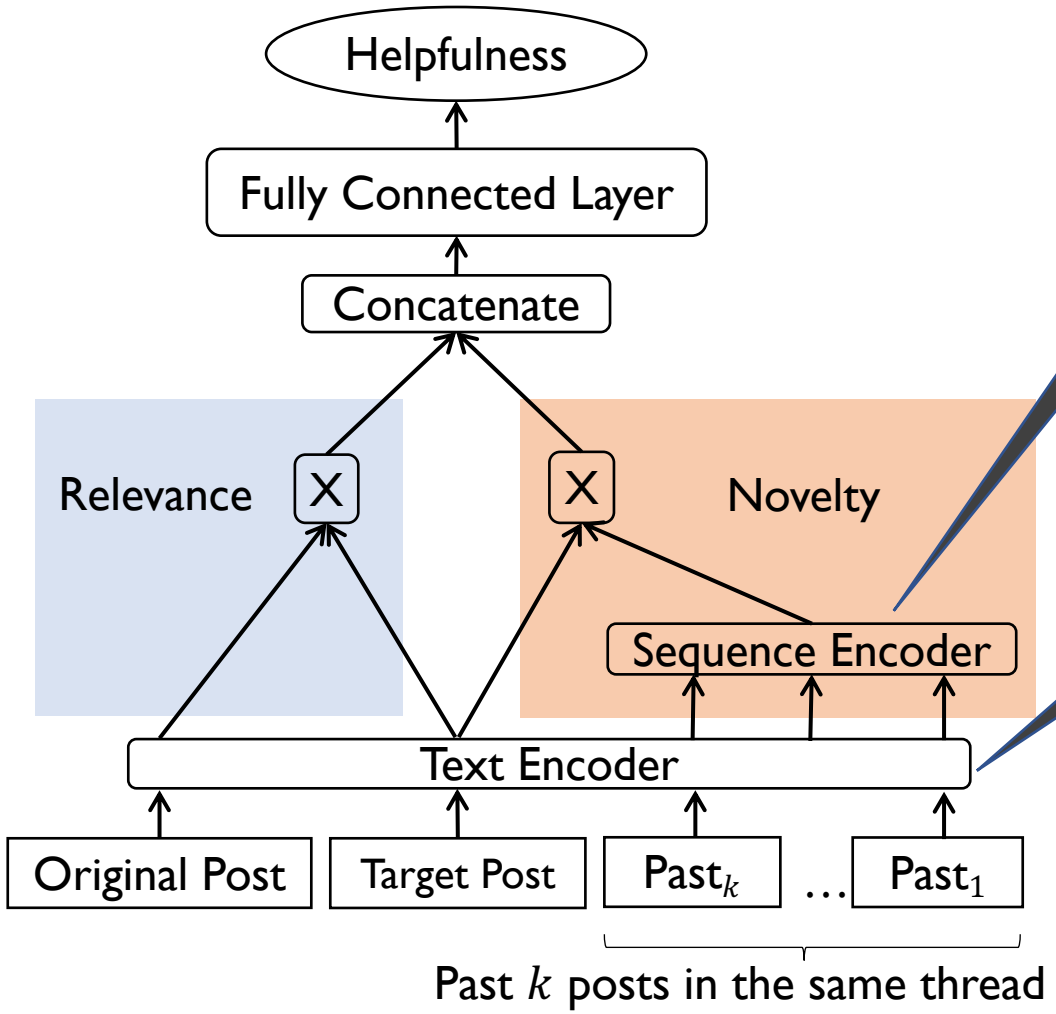
Sample thread from Reddit (r/HealthAnxiety)

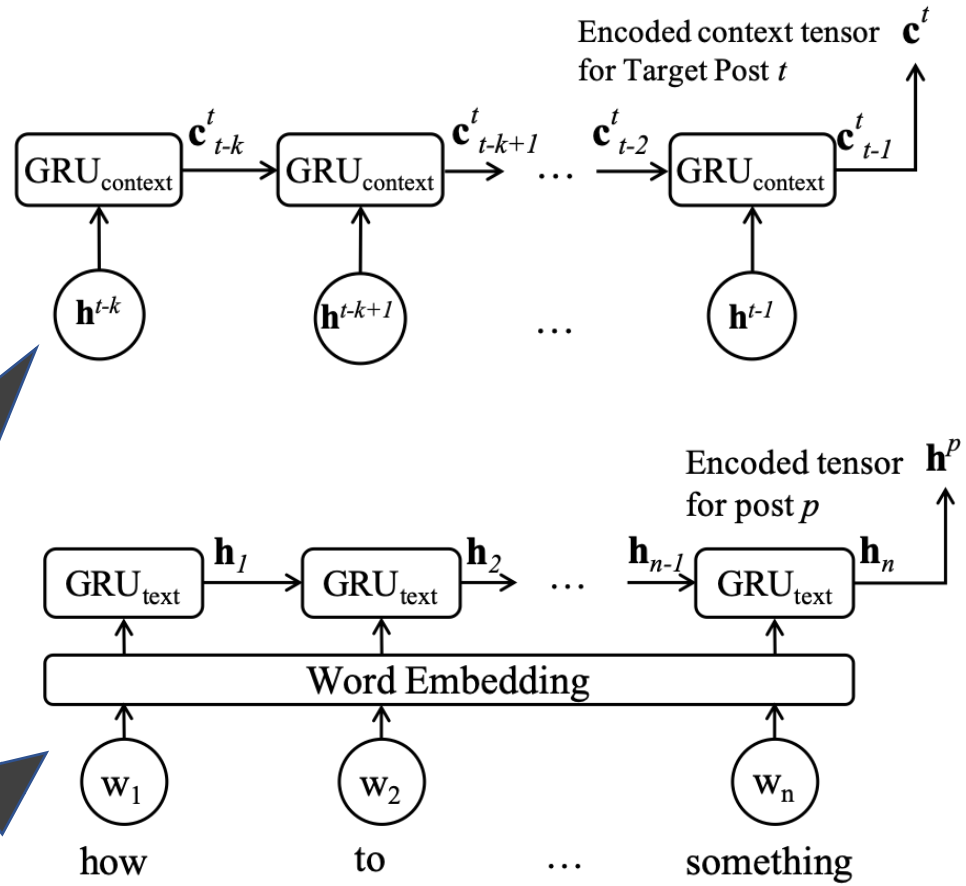
Order	Post Text	Relevant?	Novel?	Helpful?
Original Post	I was working yesterday .. and my back was bent over and when I got up I felt like I strained my back but now my mind is linking it to my kidney..			
1	I have this and my doc has told me it's muscular and physio might help..	Yes	Yes	Yes
2	Kidney pain is usually constant and doesn't change when you move, or get better when you change position, from how I understand it .. you'll be fine :)	Yes	Yes	Yes
3	If it happens only when you move there is a big chance it's a muscle spasm, this happens after some physical activities.	Yes	No	No

Hypothesis: A post would be helpful if it is

- *relevant* to the original post and
- introduces some *novel* information compared to past posts in the same thread

Neural Architecture





Post Helpfulness Prediction Model

Post content is not used directly to avoid popularity bias

Trained with binary cross-entropy loss

End-to-end trainable

Experiments - Datasets

Dataset	# Posts	#Threads	Avg. # Posts / Thread	Avg. #Words / Post
Reddit_10+	200,006	9,744	20.52	29.45
Reddit_3+	200,016	28,763	6.95	30.58
Android Apps	11,643	2,077	5.60	56.53
Matrix	19,159	2,484	4.08	65.30
Travel	30,116	10,250	2.93	163.43

Reddit: a generic discussion forum

- Public dumps available
- Created two datasets to understand modeling capabilities
  - Reddit_10+: with threads having more than 10 posts
  - Reddit_3+: threads w/ $\geq$ 3 posts

Coursera: MOOC discussion forum on online lectures

- Android Apps
- Matrix

Travel Stack Exchange

- Questions are mainly subjective

Data splitting: 80-10-10 for train, dev, and test sets

Experiments - Baselines

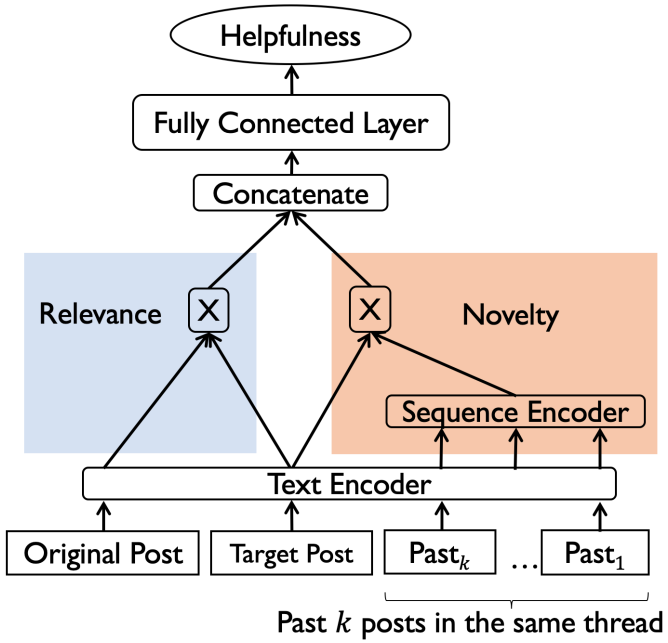
- BiLSTM (Sun *et al.*, '17): Bidirectional LSTM encoders on post text.
- Stacked LSTM (Liu *et al.*, '16): a stack of 2 LSTM layer encoders on the post text.
- LSTM with Attention (Rocktäschel et al., '16): LSTM with hierarchical attention.
- Answer Sentence Selection (Yu et al., '14): a CNN model pioneered in TREC QA.

Ablation Study

- Only the relevance component
- Only the novelty component

Ground Truth Label for Helpfulness

User feedback in forms of "upvote", "like", "mark as helpful"
80th percentile vote count as the threshold



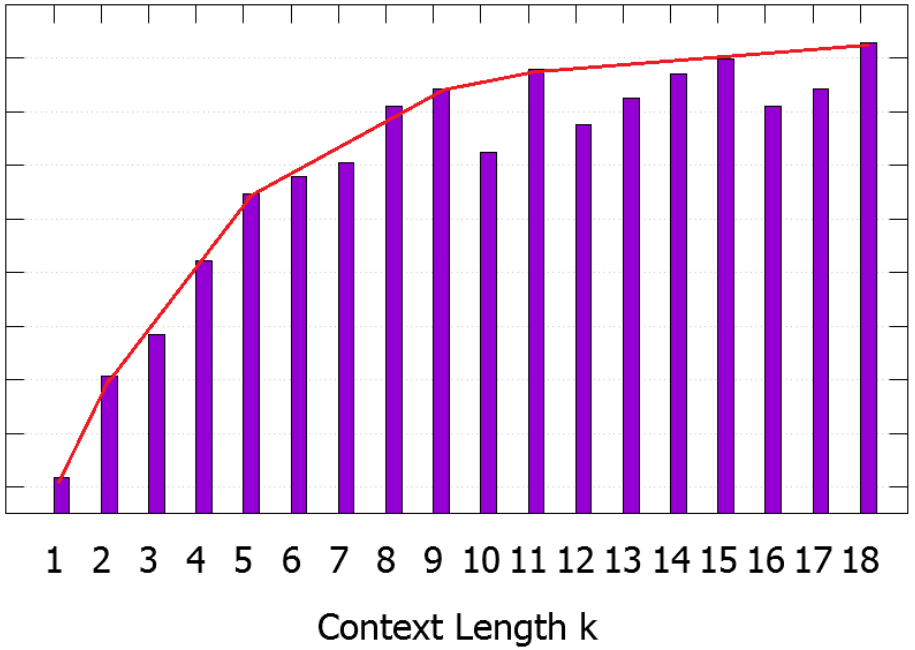
Results

Model	Reddit_10+			Reddit_3+			Android Apps			Matrix			Travel		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BiLSTM	0.23	0.23	0.23	0.23	0.22	0.22	**0.36**	0.32	0.34	0.29	0.35	0.32	0.28	**0.31**	0.29
Stacked LSTM	0.24	0.21	0.22	0.23	0.20	0.21	0.34	0.29	0.31	0.32	0.29	0.31	0.23	0.26	0.25
LSTM with Attention	0.24	0.21	0.23	0.24	0.21	0.22	0.34	0.27	0.30	0.30	0.36	0.33	0.25	0.26	0.25
Answer Sentence Selection	0.28	0.27	0.27	0.31	0.32	0.32	0.28	0.21	0.24	0.33	0.34	0.33	0.30	**0.31**	0.31
Our Model (Relevance only)	0.30	0.30	0.30	0.32	0.34	0.33	0.31	0.35	0.33	0.38	0.31	0.34	0.35	0.30	0.32
Our Model (Novelty only)	**0.53**	0.38	0.44	**0.42**	0.27	0.33	0.33	0.24	0.28	**0.43**	0.27	0.33	**0.47**	0.27	**0.34**
Our Model (full)	0.48	**0.53**	**0.51**	0.41	**0.39**	**0.40**	0.35	**0.40**	**0.38**	0.37	**0.37**	**0.37**	0.37	**0.31**	**0.34**

- A challenging task from text-only perspective
- Our model outperforms the state-of-the-art text classification models
- Ablation study shows that considering original post or past posts help compared to the vanilla models

Research Questions

Effect of Context Length



Longer context improves accuracy in general
The accuracy improves sharply from context length 1-11
From length 11-18, the improvement is positive but the rate is lower
A trade-off exists between training time and accuracy

Need of Encoding Order of Past Posts

What happens if we ignore the order of past posts in a thread?

Let's replace $GRU_{context}$ with simple averaging of past posts

Context Modeling	Reddit_10+	Reddit_3+	Android Apps	Matrix	Travel
Average	0.40	0.35	0.36	0.36	0.33
$GRU_{context}$	**0.53**	**0.40**	**0.38**	**0.37**	**0.34**

$GRU_{context}$ outperforms averaging based model consistently across all datasets

Encoding the sequence of past posts is important to improve accuracy

Does open-endedness (really) matter?

Differential analysis between our model and best text classifier to understand modeling differences

Considered each test point where only one of the models was correct, not both

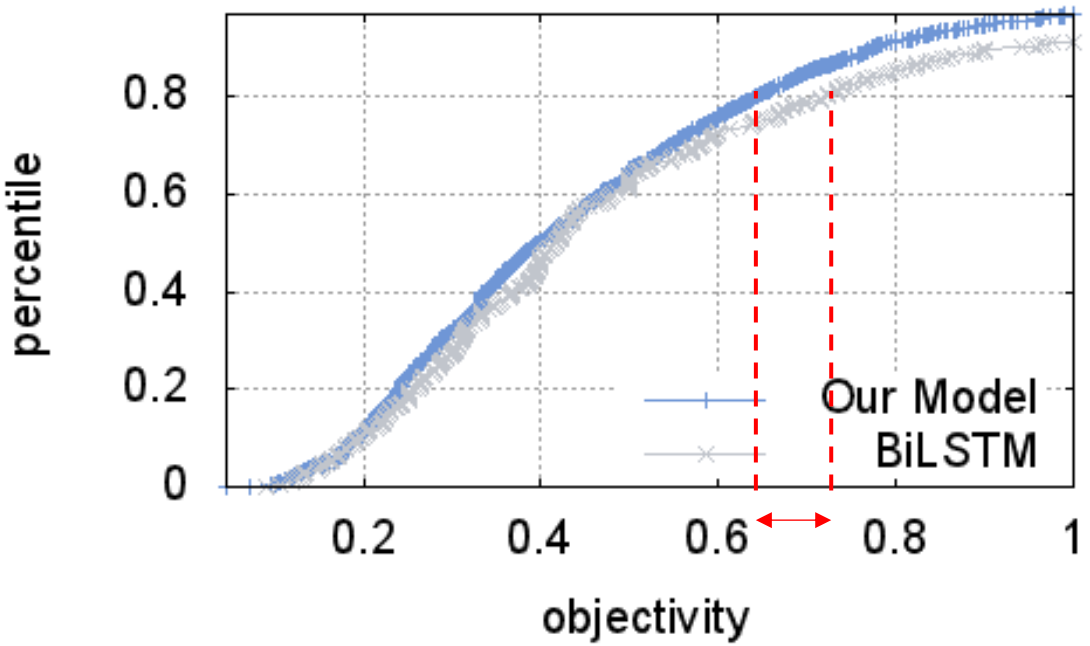
Defined a simple function for thread *objectivity*

$$objectivity = \frac{max(vote(x)) - min(vote(x))}{\sum vote(x)}$$

Open-ended → low *objectivity* score

Objectivity CDF shows that the objectivity scores for
- Blue: threads where only our model was correct
- Grey: threads where only BiLSTM was correct

Objectivity scores for Blue threads are lower



Our Model tends to do better when a thread is more open-ended in nature

Conclusion

Presented a key difference between Discussion Forums and CQA

- One is open-ended, the other one is not

Posts in threads are found to be helpful when they provide relevant but novel information

Proposed a novel neural architecture to encode relevance and novelty

The model outperforms competitive text classifiers across 5 datasets from 3 different domains

Thanks for listening!
Questions? Comments?
Email: kishaloy@comp.nus.edu.sg