Knowledge and Pre-trained Language Models Inside and Out: a deep-dive into datasets and external knowledge

Chenyang Lyu
ML-Labs, School of Computing, DCU
5th May 2022

Outline

- Introduction
- Research Questions
- Progress
  - Sentiment Analysis with User and Product context
  - Unsupervised QA via Summarization-Informed QG
  - Analysing Extractive QA Dataset
- Future Plans

1. Introduction - the success of Pre-trained Language Models

- Pre-trained Language Models (PLMs) such as BERT(Devlin et al), GPT(Radford et al), BART(Lewis et al), etc. have proven successful on many NLP tasks even surpassing human performance on some tasks.

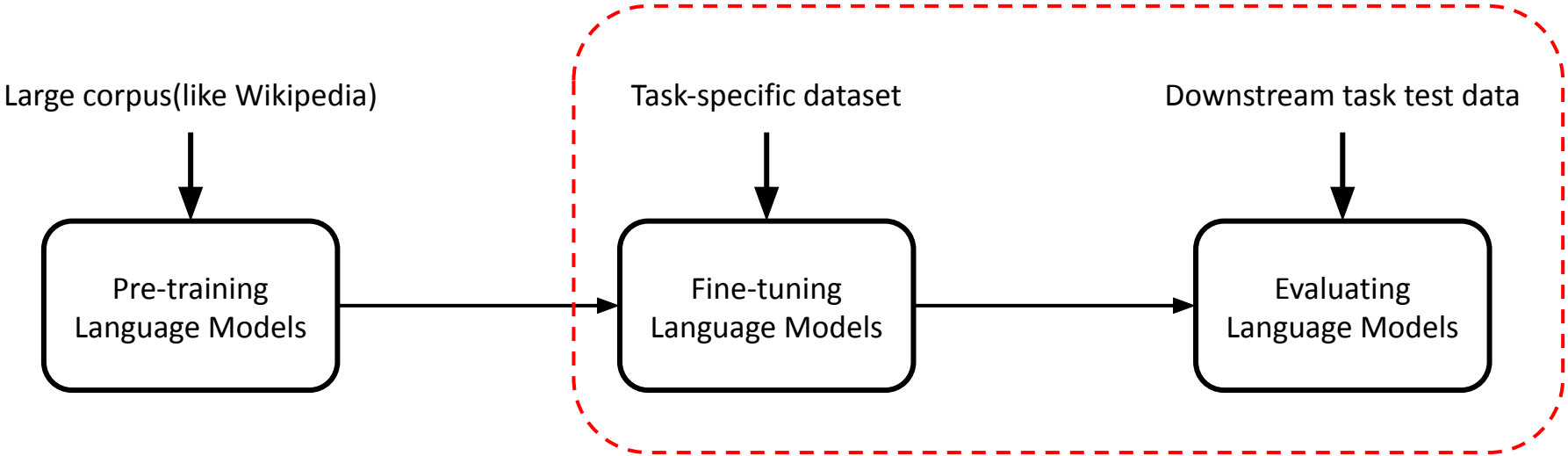
SQuAD1.1 Leaderboard

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490

http://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

1. Introduction - the success of Pre-trained Language Models

- New paradigm of NLP:



I focus on fine-tuning phase of PLMs

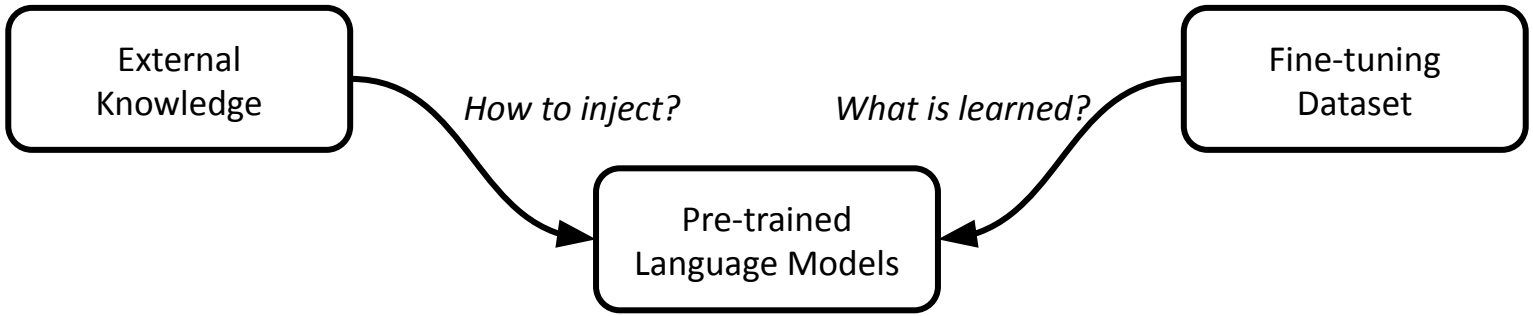
1. Introduction

- Do PLMs need knowledge?
    - Incorporating knowledge is a way to extend the ***context*** of words.
- How PLMs acquire knowledge from fine-tuning data?
    - A better understanding of the data has the potential to improve the generalizability of models.

awareness, understanding, or information that has been obtained by experience or study, and that is either in a person's mind or possessed by people generally
- Knowledge, Cambridge Dictionary

1. Introduction

- We have two major goals focusing on the learning of knowledge in this project:
  - How to extend the ***context*** of PLMs using external ***knowledge***?
  - What ***knowledge*** do PLMs learn from data to extend their ***context***?



2. Research Questions

- Two major research questions and four specific research questions :
  - RQ1: How to inject external knowledge into PLMs?
    - How can we utilize the extra information in the metadata of product reviews to improve document-level sentiment analysis?
    - How can we leverage linguistic knowledge and summarization data to improve Unsupervised QA?
  - RQ2: How do PLMs learn knowledge from fine-tuning data?
    - How does a machine learning model (typically a neural model) learn from sentiment analysis and QA data - which part of the data accounts for the model's performance on dev/test set?
    - Exploring multilingual representations and analysing their role in learning from multilingual corpora for PLMs.

2. Research Questions

- The main idea of RQ1 is to encode information beyond texts.
- Focus on two tasks:
  - Sentiment analysis
  - Question answering

3. Progress

- Sentiment Analysis with User and Product Context (COLING 2020)
- Unsupervised QA with Summarisation-Informed QG (EMNLP 2021)
- Analysing Extractive QA Dataset (Insights Workshop at ACL 2022)

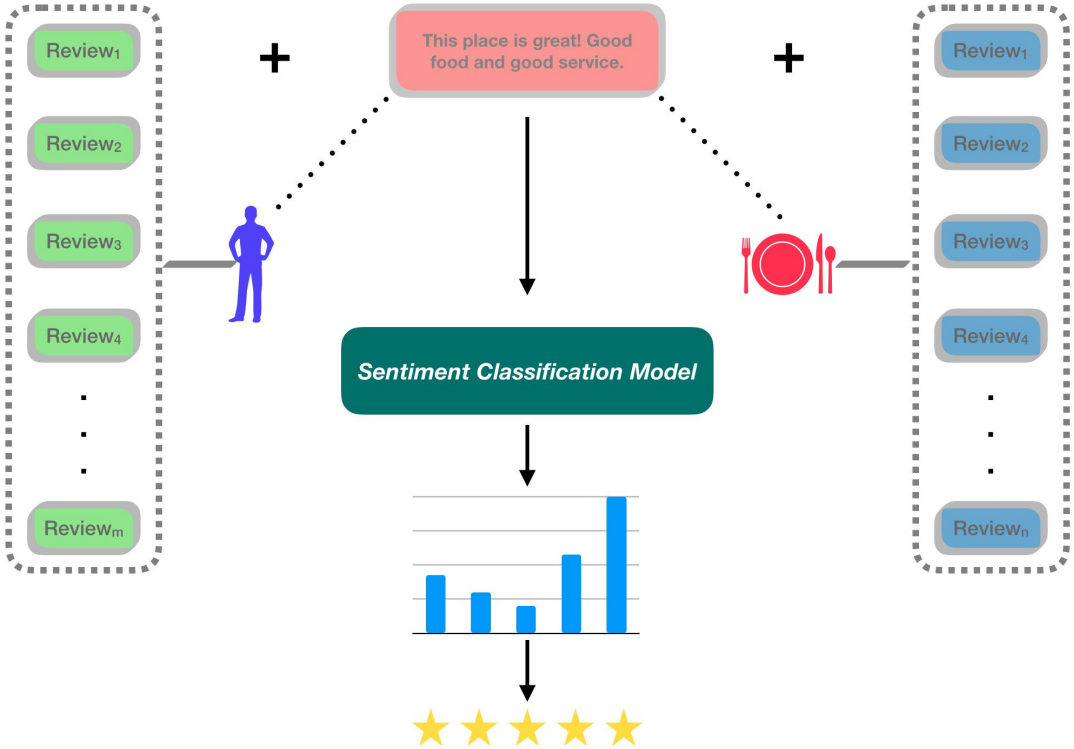
3. Progress - Sentiment Analysis with User and Product Context

Sentiment Analysis with User and Product Information

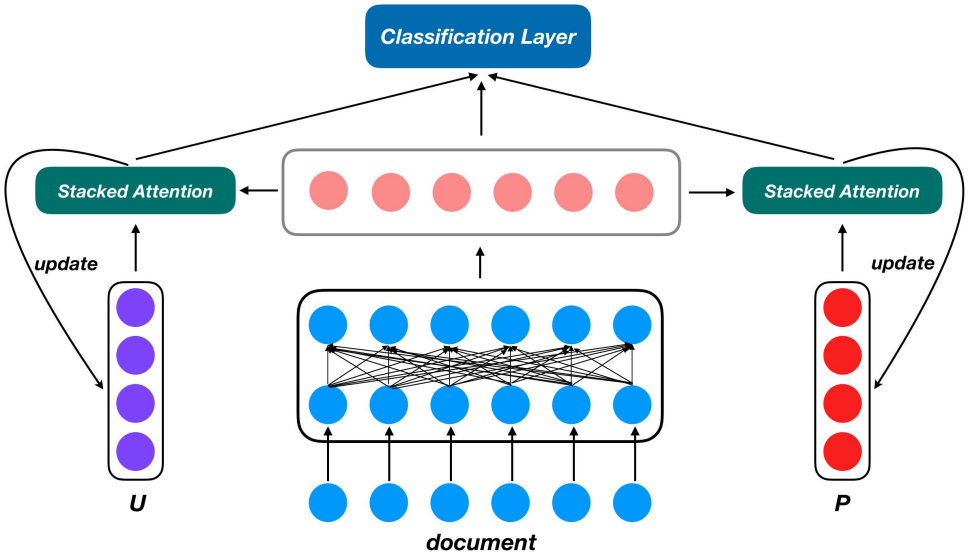
- Not only we have the review text, but also the user and product IDs.
- Modeling the user, who has written the review, and the product being reviewed is worthwhile for polarity prediction.

Chenyang Lyu, Jennifer Foster, Yvette Graham, Improving Document-Level Sentiment Analysis with User and Product Context, COLING 2020

Introduction



Methodology - Incorporating User and Product Context



Step 1:

- Obtain document representation.

Step 2:

- Use user and product embedding vectors to gather information from document representation through attention function.

Step 3:

- Fuse user-biased and product-biased information to obtain a final review representation, then pass it to a classification layer to get sentiment label.

Step 4:

- Incrementally add current biased representation to corresponding user and product embeddings.

Methodology - Incorporating User and Product Context

Get document representation:

$$H_d = BERT\_encoder(d)\ (1)$$

Inject user and product preferences:

$$C_u^t = stacked - attention(E_u, H_d)\ (2)$$

$$C_p^t = stacked - attention(E_p, H_d)\ (3)$$

Gating mechanism:

$$z_u = \sigma(W_{zu} C_u^t + W_{zh} H_d + b_u)\ (4)$$

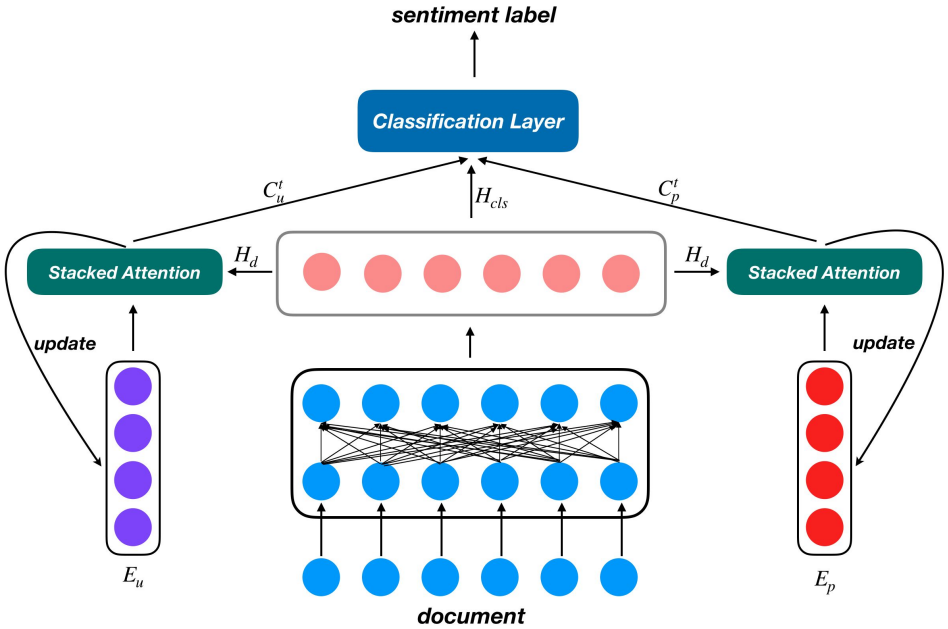
$$z_p = \sigma(W_{zp} C_p^t + W_{zh} H_d + b_p)\ (5)$$

Final representation:

$$H_{biased} = H_{cls} + z_u \odot C_u^t + z_p \odot C_p^t\ (6)$$

Update user and product matrix:

$$E_u' = \sigma(E_u + \lambda_u C_u^t) \quad E_p' = \sigma(E_p + \lambda_p C_p^t)(7)$$



Experiments and Analysis

- Datasets:
  - Our experiments are conducted on the IMDB, Yelp-13 and Yelp-14 benchmark datasets.

Datasets	Classes	Documents	Users	Products	Docs/User	Docs/Product	Words/Doc
IMDB	1–10	84,919	1,310	1,635	64.82	51.94	394.6
Yelp-2013	1–5	78,966	1,631	1,633	48.42	48.36	189.3
Yelp-2014	1–5	231,163	4,818	4,194	47.97	55.11	196.9

Table 1: Statistics of IMDB, Yelp-2013 and Yelp-2014.

- Experimental setup:
  - Learning rate: {8e-6, 3e-5, 5e-5}, weight decay: {0, 1e-1, 1e-2, 1e-3}
  - Warmup ratio: 0.1, linear decay.
  - Maximum length to BERT: 512 wordpiece tokens.
  - Optimizer: AdamW.
  - Batch size:{8, 16}
  - Epochs:{2, 3}

Experiments

- Experimental results:
  - Our proposed model achieves the best accuracy and RMSE on Yelp-2013 and Yelp-2014, and the best RMSE on IMDB.

	IMDB		Yelp-2013		Yelp-2014	
	Acc. (%)	RMSE	Acc. (%)	RMSE	Acc. (%)	RMSE
BERT VANILLA	$47.9_{0.46}$	$1.243_{0.019}$	$67.2_{0.46}$	$0.647_{0.011}$	$67.5_{0.71}$	$0.621_{0.012}$
IUPC W/O UPDATE	$52.1_{0.31}$	$1.194_{0.010}$	$69.7_{0.37}$	$0.605_{0.007}$	$70.0_{0.29}$	$0.601_{0.007}$
IUPC (our model)	$53.8_{0.57}$	$\mathbf{1.151_{0.013}}$	$\mathbf{70.5_{0.29}}$	$\mathbf{0.589_{0.004}}$	$\mathbf{71.2_{0.26}}$	$\mathbf{0.592_{0.008}}$
UPNN	43.5	1.602	59.6	0.784	60.8	0.764
UPDMN	46.5	1.351	63.9	0.662	61.3	0.720
NSC	53.3	1.281	65.0	0.692	66.7	0.654
CMA	54.0	1.191	66.3	0.677	67.6	0.637
DUPMN	53.9	1.279	66.2	0.667	67.6	0.639
HCSC	54.2	1.213	65.7	0.660	67.6	0.639
HUAPA	55.0	1.185	68.3	0.628	68.6	0.626
CHIM	**56.4**	1.161	67.8	0.641	69.2	0.622
RRP-UPM	56.2	1.174	69.0	0.629	69.1	0.621

3. Progress - Unsupervised QA via Summarization-Informed QG

- Template-based QG
  - Use hand-crafted rules induced from linguistic knowledge
  - *Shortcoming:*
    - *Generated questions have high lexical overlap with source text*
- Supervised QG
  - Use existing QA datasets to train a QG system (typically a neural model).
  - *Shortcoming:*
    - *Rely on the availability of QA dataset which is expensive to obtain and heavily tied to a certain domain and language.*

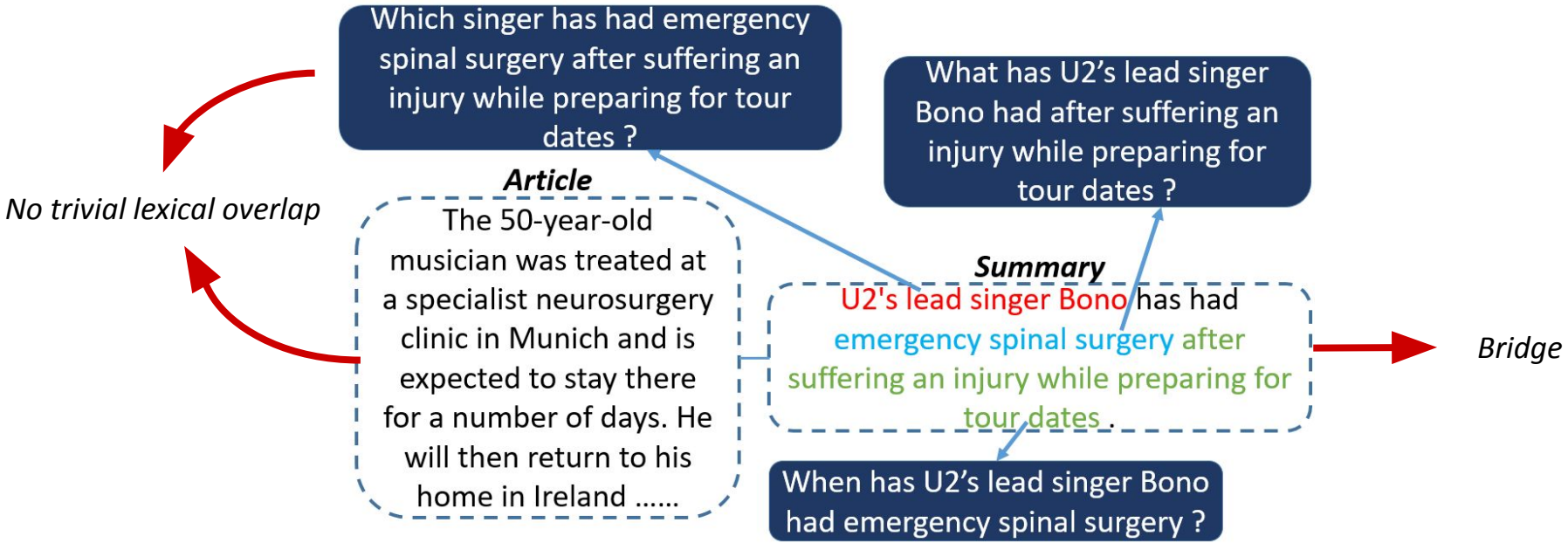
Chenyang Lyu, Lifeng Shang, Yvette Graham, Jennifer Foster, Xin Jiang and Qun Liu, Improving Unsupervised Question Answering via Summarization-Informed Question Generation, EMNLP 2021

3. Progress - Unsupervised QA via Summarization-Informed QG

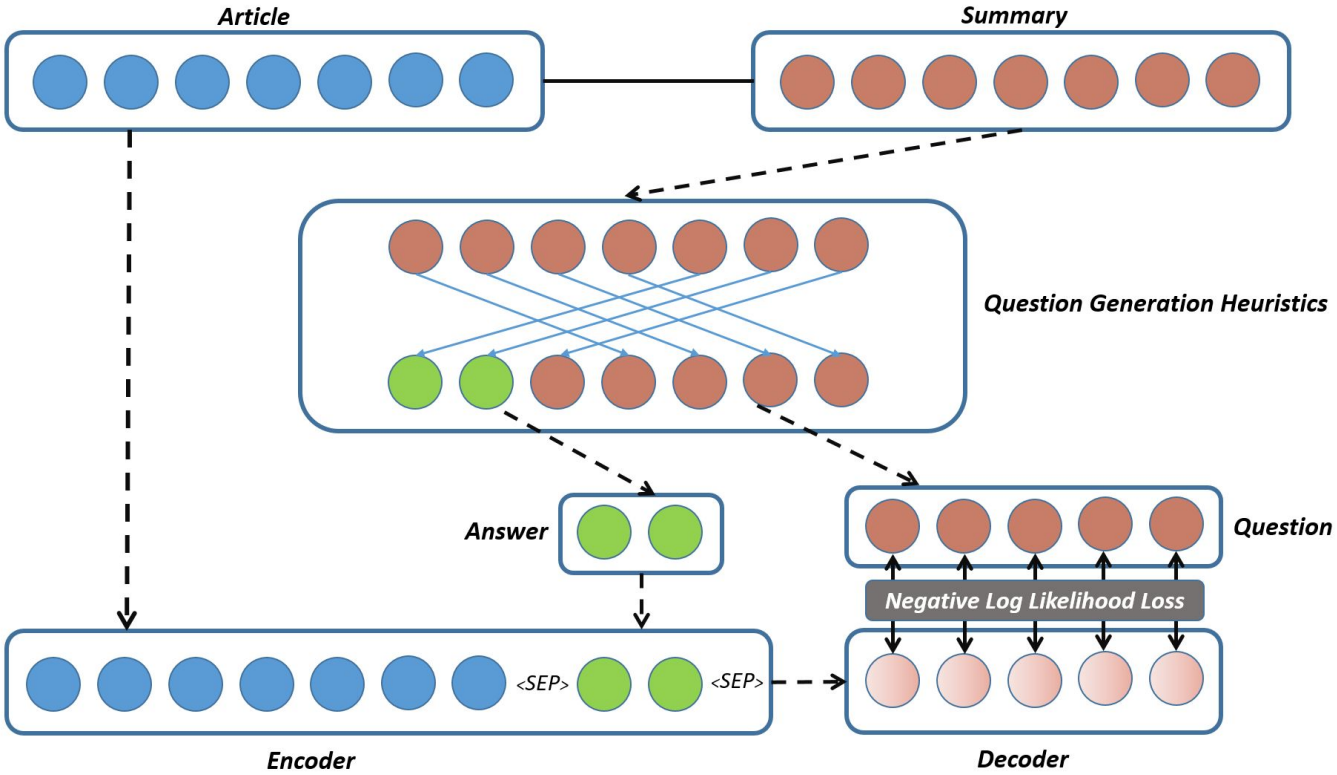
- We propose an unsupervised QG approach:
  - Employ summary data as a bridge between passage and question
  - Generate questions based on summaries using heuristics
  - Train a QG system using data created above

Chenyang Lyu, Lifeng Shang, Yvette Graham, Jennifer Foster, Xin Jiang and Qun Liu, Improving Unsupervised Question Answering via Summarization-Informed Question Generation, EMNLP 2021

Methodology - Unsupervised QG

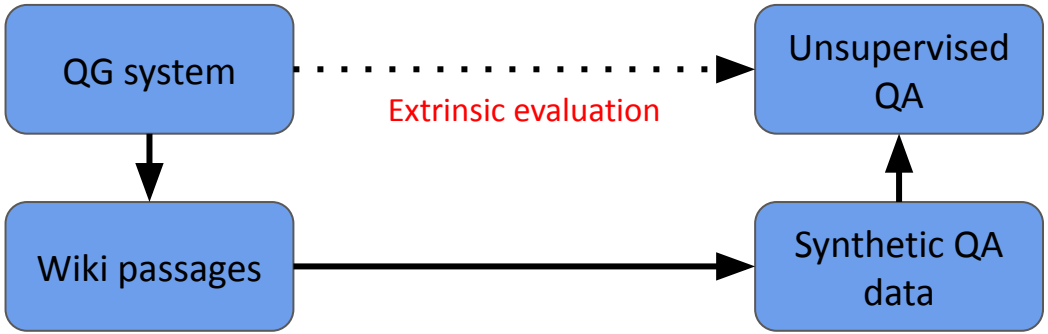


Methodology - Unsupervised QG



Experiments and Analysis

- How to evaluate our QG system?
  - BLEU, ROUGE and Meteor are not suitable for evaluating QG since with a <passage, answer> pair there could be multiple plausible questions
- We **extrinsically** evaluate our QG system using unsupervised QA
  - We train a QG system using the data generated by QG heuristics
  - Then we use wikipedia passages to generate synthetic QA data using the QG system



Experimental Results

We employ our synthetic QA dataset (20k samples) to fine-tune a BERT-large model then evaluate it on SQuAD, Natural Questions and TriviaQA (in-domain)

Models	SQuAD1.1 EM	SQuAD1.1 F-1
SUPERVISED MODELS		
Match-LSTM	64.1	73.9
BiDAF	66.7	77.3
BERT-base	81.2	88.5
BERT-large	84.2	91.1
UNSUPERVISED MODELS		
Lewis et al. (2019)	44.2	54.7
Li et al. (2020)	62.5	72.6
Our Method	**65.6**	**74.5**

Models	NQ EM	NQ F-1	TriviaQA EM	TriviaQA F-1
SUPERVISED MODELS				
BERT-base	66.1	78.5	65.1	71.2
BERT-large	69.7	81.3	67.9	74.8
UNSUPERVISED MODELS				
Lewis et al. (2019)	27.5	35.1	19.1	23.8
Li et al. (2020)	31.3	48.8	27.4	38.4
Our Method	**46.0**	**53.5**	**36.7**	**43.0**

Experimental Results

To investigate the transferability and generalizability of our synthetic QA data, we further apply it on three **out-of-domain** QA datasets, NewsQA, DuoRC and BioASQ

	NewsQA		BioASQ		DuoRC	
	EM	F-1	EM	F-1	EM	F-1
Lewis et al. (2019)	19.6	28.5	18.9	27.0	26.0	32.6
Li et al. (2020)	33.6	46.3	30.3	38.7	32.7	41.1
Our Method	**37.5**	**50.1**	**32.0**	**43.2**	**38.8**	**46.5**

Summarising the progress for RQ1

- Sentiment analysis paper: encoding knowledge into the model
- Unsupervised QA paper: encoding knowledge into the data

3. Progress - Analysing Extractive QA Dataset

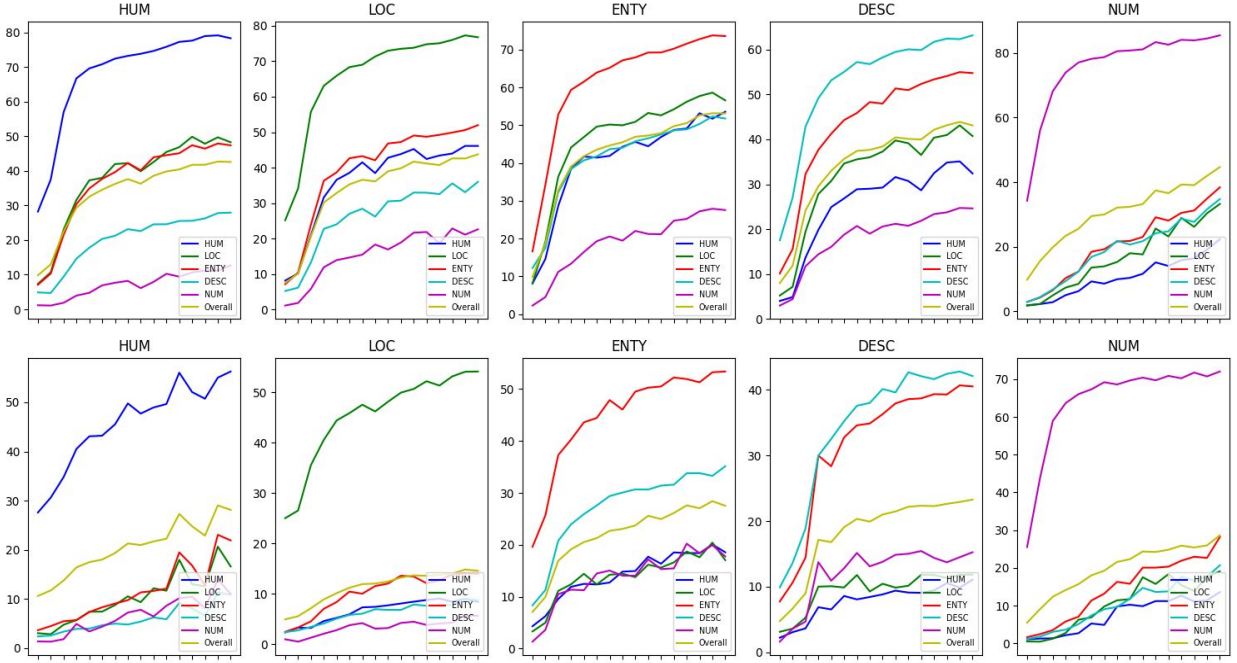
- We design several probing tasks for the purpose of investigating the internal characteristics of QA fine-tuning datasets.
- The dimensions of probing tasks we used include:
  - Question type
  - Difficulty level (defined by question-context lexical overlap)
  - Answer corruption

Experiments

- Question type:
  - We use a question classifier to divide the QA examples according to their question types.(*HUM, LOC, ENTY, DESC, NUM*),
  - Then we use the QA examples in each question type to train a BERT model with increasing sample size.
  - We evaluate the performance on dev sets of SQuAD1.1 and NewsQA.
  - The dev sets are also categorized according to their question type.

Experiments

- Visualization of **F-1 learning curves** for the QA systems trained on the examples of five question types (*HUM, LOC, ENTY, DESC, NUM*), tested on the examples for each question type and the original dev set of SQuAD1.1 (top) and NewsQA (bottom).

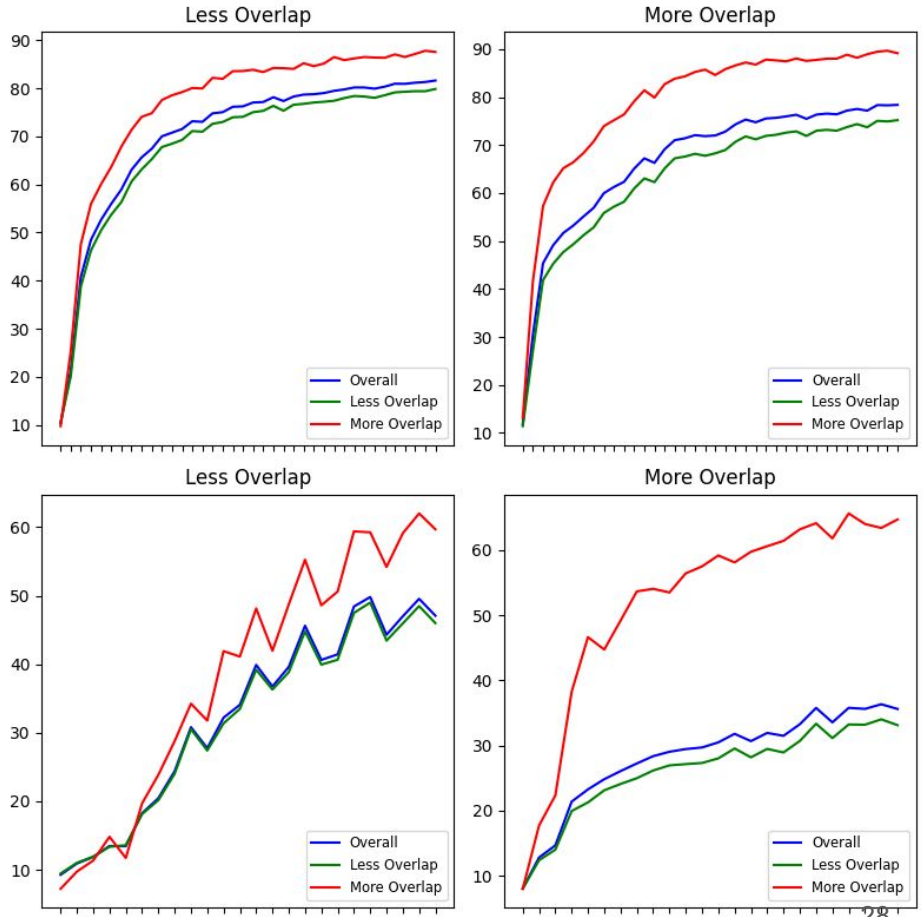


Experiments

- Difficulty level:
  - We use question-context lexical overlap to define difficulty.
  - QA examples with high lexical overlap are defined as ***easy*** examples.
  - QA examples with low lexical overlap are defined as ***difficult*** examples.

Experiments

- Visualization of **F-1 score curve** with more or less lexical overlap on SQuAD1.1 (top) and NewsQA (bottom).
- Results show that QA examples with **low question-context lexical overlap are more effective**.



Experiments

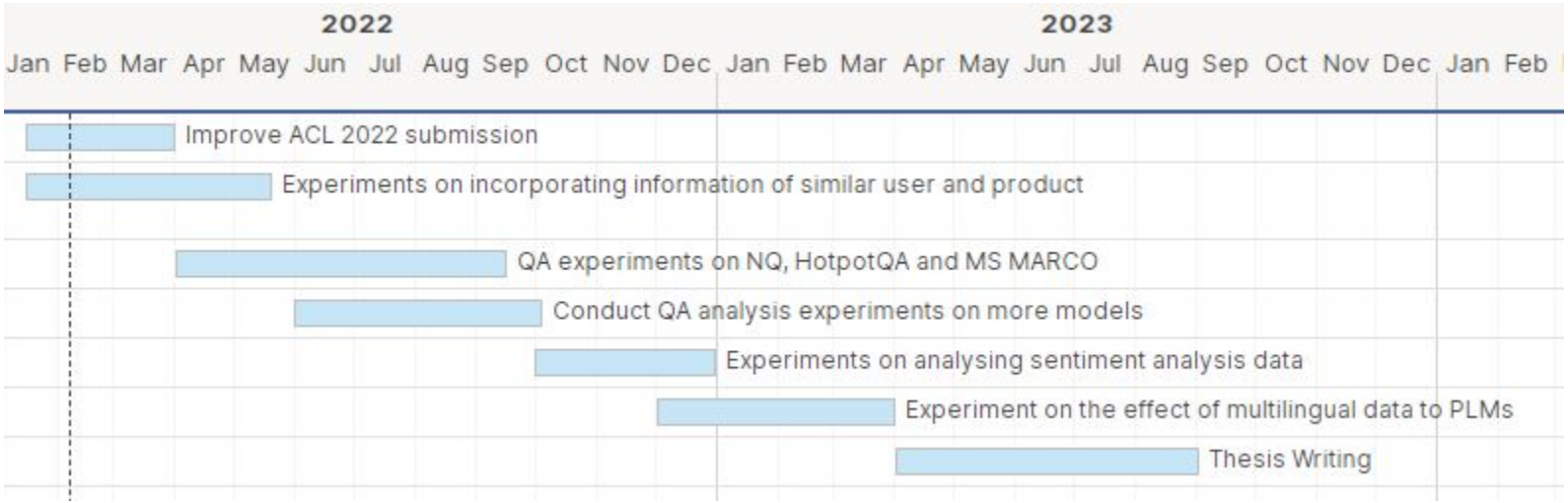
- Answer corruption:
  - We replace answers in dev set with random tokens:
    - Context: The American Football Conference (AFC) champion ***jysbdefziqvzbi*** defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title.
    - Question: Which NFL team won Super Bowl 50?
    - Original answer: Denver Broncos
    - Corrupted answer: ***jysbdefziqvzbi***

Experiments

- Evaluation results **(EM/F-1)** on dev sets of SQuAD1.1 and NewsQA with corrupted answers.

		Overall
SQuAD1.1	Original	66.97/80.96
	Random tokens	55.99/61.40
NewsQA	Original	49.22/64.53
	Random tokens	31.72/35.91

4. Future plans



4. Future plans

- Analysing how PLMs learn from fine-tuning data.
    - Include more QA datasets and models.
    - Apply the analytical methods to sentiment analysis.
- Exploring multilingual language learning for PLMs.
    - Focus on QA and sentiment analysis.

Publications

Extending the Scope of Out-of-Domain: Examining QA models in multiple subdomains
***Chenyang Lyu**, Jennifer Foster and Yvette Graham*
The 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022 Workshop on Insights from Negative Results in NLP.

Achieving Reliable Human Assessment of Open-Domain Dialogue Systems
*Tianbo Ji, Yvette Graham, Gareth J. F. Jones, **Chenyang Lyu** and Qun Liu*
The 60th Annual Meeting of the Association for Computational Linguistics, ***ACL 2022***.

Improving Unsupervised Question Answering via Summarization-Informed Question Generation
***Chenyang Lyu**, Lifeng Shang, Yvette Graham, Jennifer Foster, Xin Jiang and Qun Liu*
The 2021 Conference on Empirical Methods in Natural Language Processing, ***EMNLP 2021***.

Improving Document-Level Sentiment Analysis with User and Product Context
***Chenyang Lyu**, Jennifer Foster and Yvette Graham*
The 28th International Conference on Computational Linguistics, ***COLING 2020***.

Incorporating Context and Knowledge for Better Sentiment Analysis of Narrative Text
***Chenyang Lyu**, Tianbo Ji and Yvette Graham*
The Third International Workshop on Narrative Extraction from Texts held in conjunction with the 42nd European Conference on Information Retrieval, ECIR 2020 workshop.

References

Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019.

Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).

Lewis, Mike, et al. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.

Lyu, Chenyang, Jennifer Foster, and Yvette Graham. "Improving Document-Level Sentiment Analysis with User and Product Context." *Proceedings of the 28th International Conference on Computational Linguistics*. 2020.

Lyu, Chenyang, et al. "Improving Unsupervised Question Answering via Summarization-Informed Question Generation." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021.

Thanks !