Paper Review

TASER
: Task-Aware Specialization for Efficient and Robust Dense Retrieval for Open-Domain Question Answering

Hao Cheng, Hao Fang, Xiaodong Liu, Jianfeng Gao

Microsoft, 2022

Yejin Yoon

HYU NLP Lab.

Dept. of Artificial Intelligence Application

Hanyang University

stillwithyou@hanyang.ac.kr

Review of TASER:

http://t2m.kr/yrcEh

Before starting the presentation,
please check the link above

What Are Covered in This Presentation

- **Details of TASER**
  - **TASER :** Cheng, Hao, et al. "Task-Aware Specialization for Efficient and Robust Dense Retrieval for Open-Domain Question Answering." arXiv preprint arXiv:2210.05156 (2022).

- **Some Pre-Requisites**
  - ODQA(Open-domain Question Answering)
  - Dense Retriever
  - Bi-Encoder
  - MoE w/ Switch Transformers

What Are NOT Covered in This Presentation

- **Details of Predecessors**
  - **Gumbel-Softmax**: Jang, Eric, Shixiang Gu, and Ben Poole. "Categorical reparameterization with gumbel-softmax." arXiv preprint arXiv:1611.01144 (2016).
  - **Transformer**: Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
  - **DPR**: Karpukhin, Vladimir, et al. "Dense passage retrieval for open-domain question answering." arXiv preprint arXiv:2004.04906 (2020).
  - **GShard**: Lepikhin, Dmitry, et al. "Gshard: Scaling giant models with conditional computation and automatic sharding." arXiv preprint arXiv:2006.16668 (2020).
  - **Switch Transformers**: Fedus, William, Barret Zoph, and Noam Shazeer. "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity." (2021).

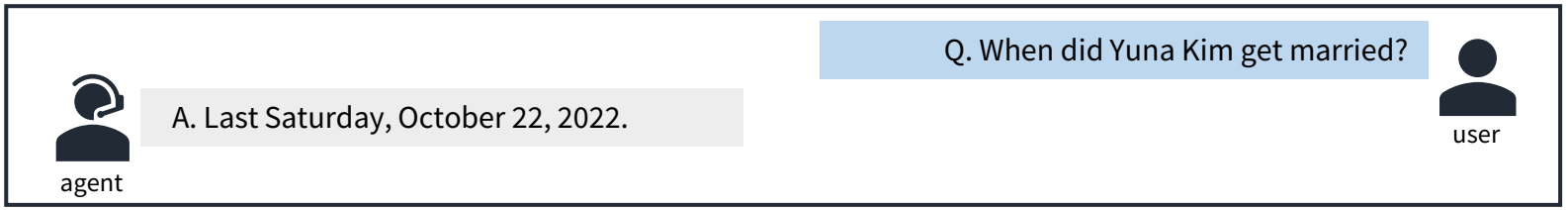
Pre-Requisites

- What is **ODQA**(Open-domain Question Answering)?
- What is **Dense Retriever**? (Bi-Encoder)
- What is **MoE**? (Switch Transformers) – *Later!*

Pre-Requisites : What is ODQA?

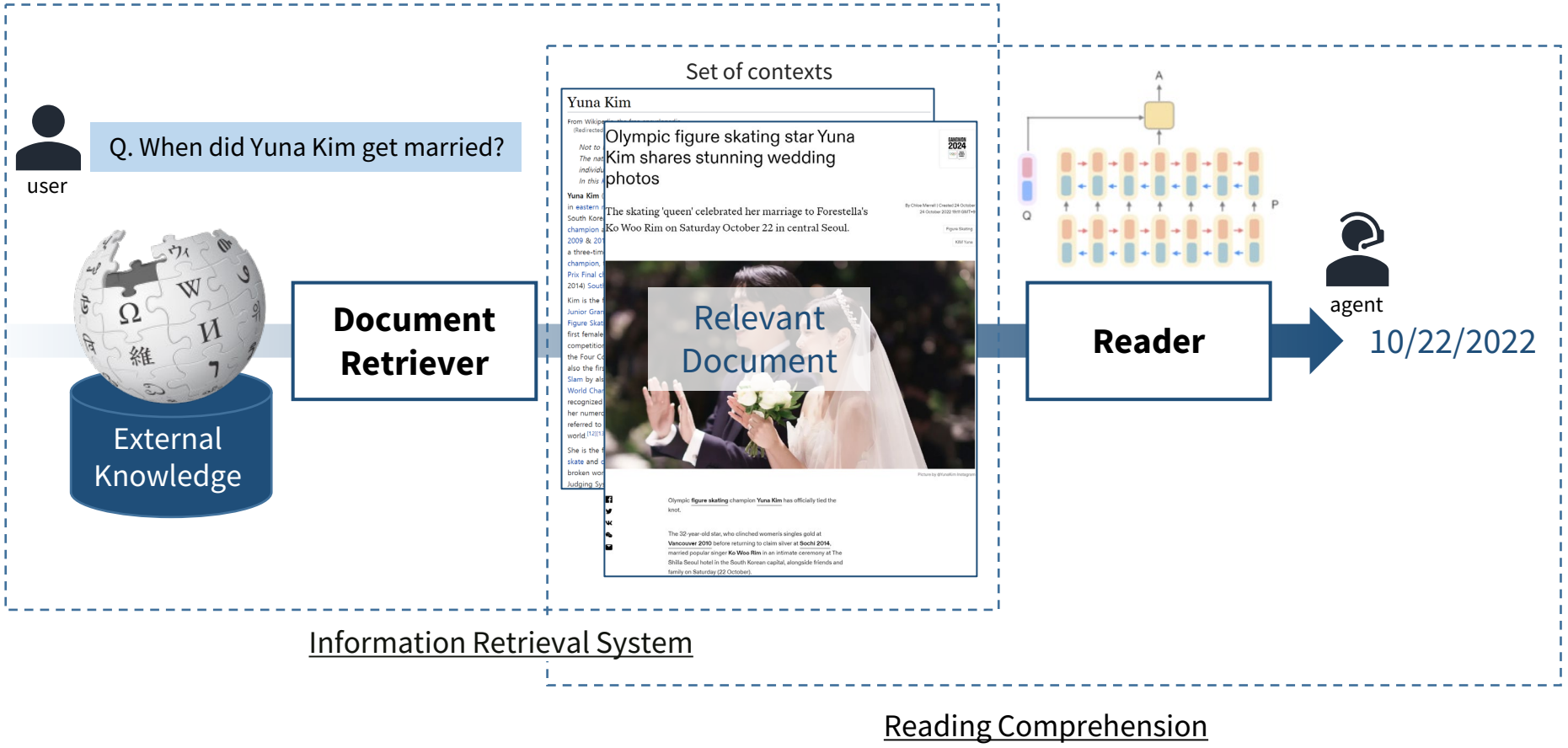
A model that can **answer any question** with regard to **factual knowledge** can lead to many useful and practical applications, such as working as a **chatbot** or an **AI assistant**.

- **Open-domain Question Answering**
  - A type of language tasks, asking a model to produce answers to factoid questions in natural language.
  - Given only a question, the model outputs the best answer it can find.
    - Questions could be about nearly **anything** relying on world knowledge
    - The challenge is that the context containing relevant information about the question is **not** provided.
  - Usually, in ODQA, factoid questions are considered that have **short and concise** answers unlike long-form or non-factoid questions.
    - Therefore, it is simple to evaluate model performance



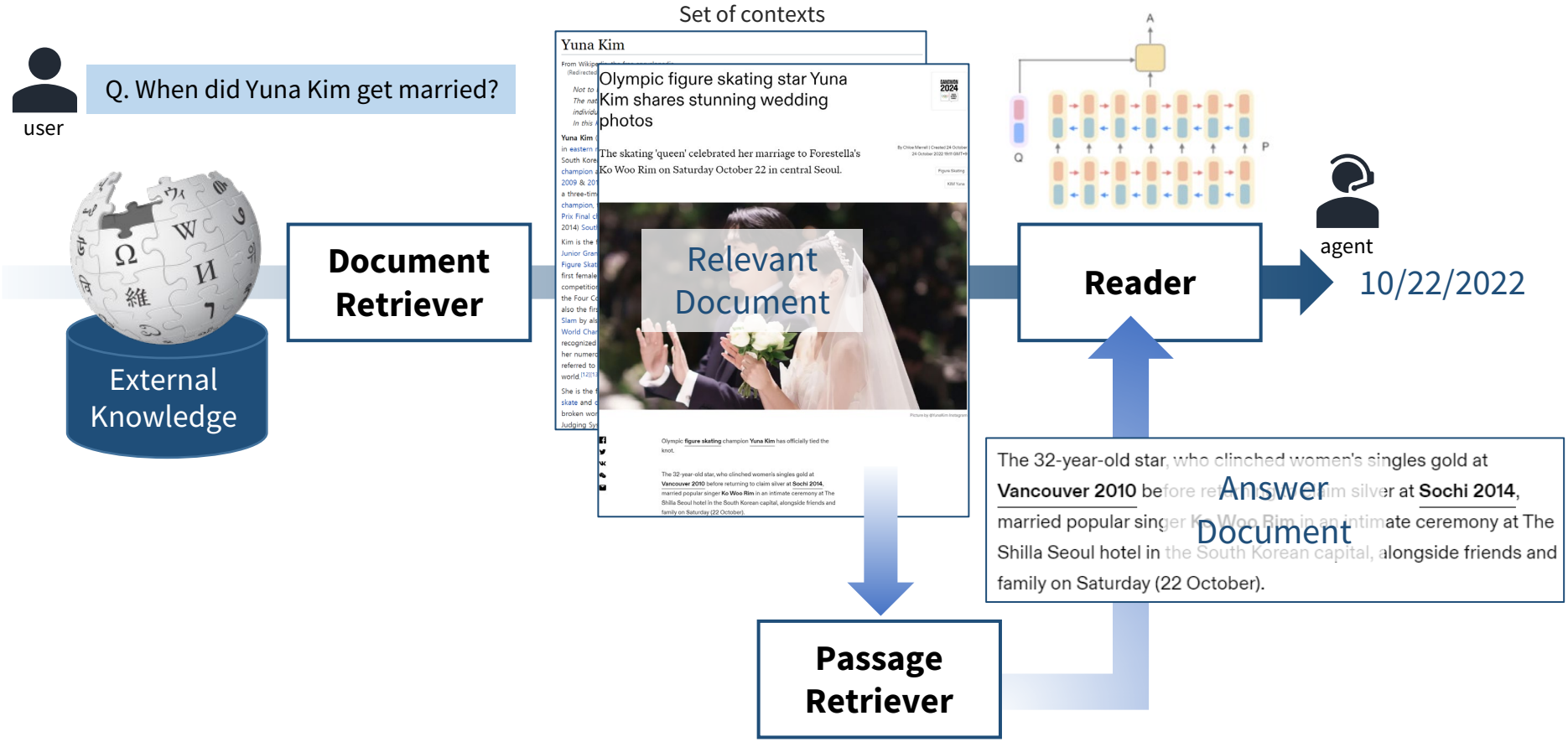
Pre-Requisites : What is ODQA?

- **Pipeline approaches (2-stage retriever-reader system)**



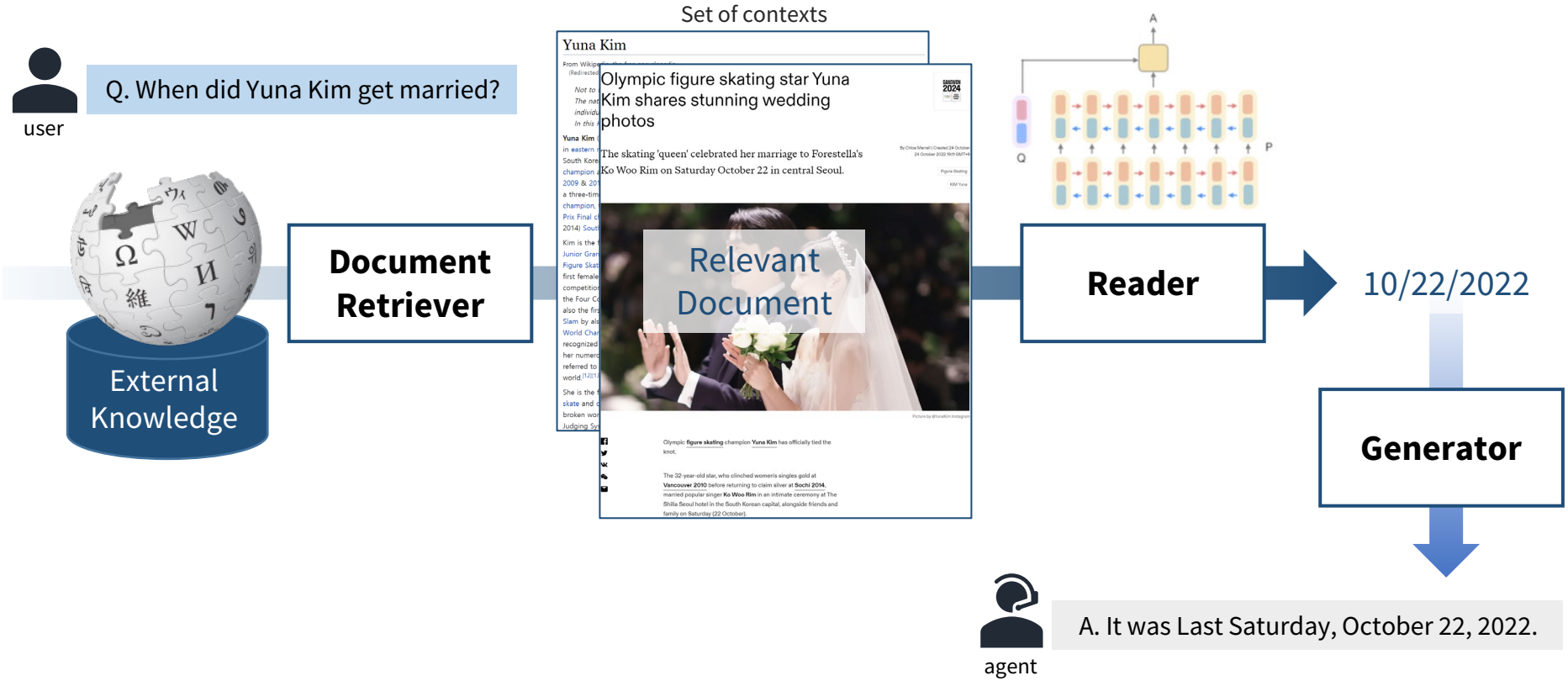
Pre-Requisites : What is ODQA?

- **Pipeline approaches (2-stage retriever-retriever-reader system)**



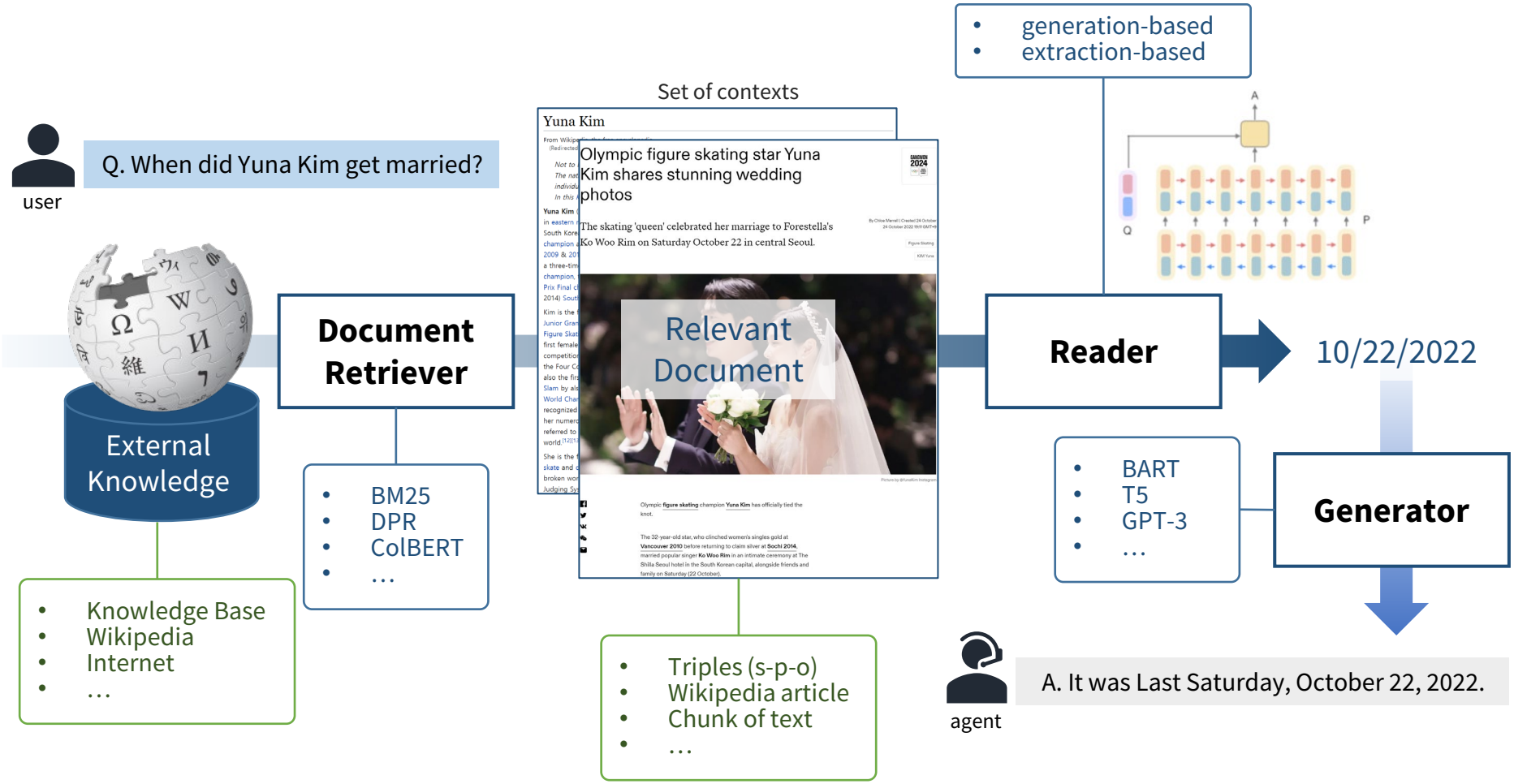
Pre-Requisites : What is ODQA?

- **Pipeline approaches (3-stage retriever-reader-generator system)**



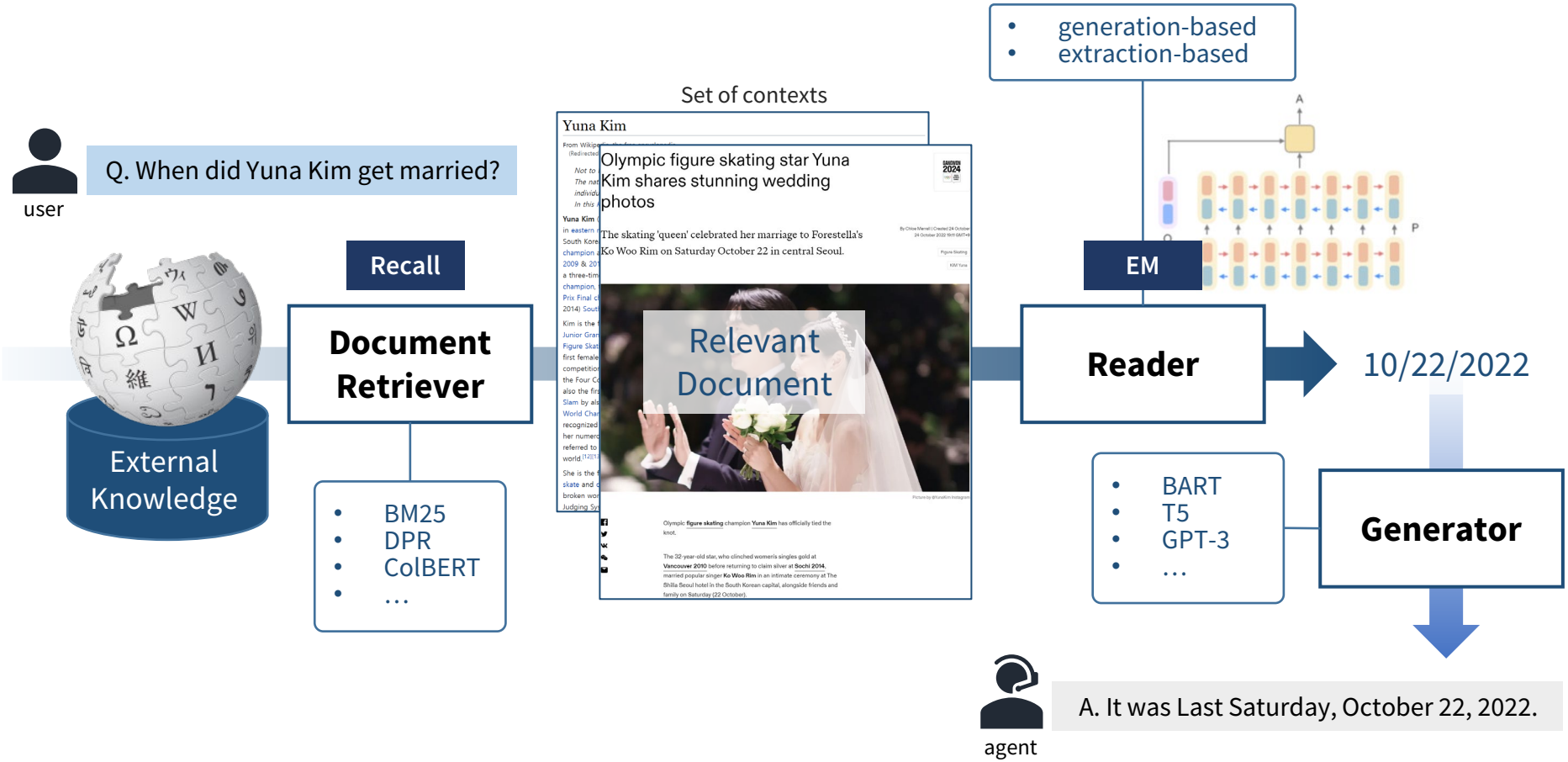
Pre-Requisites : What is ODQA?

- **Pipeline approaches (3-stage retriever-reader-generator system)**



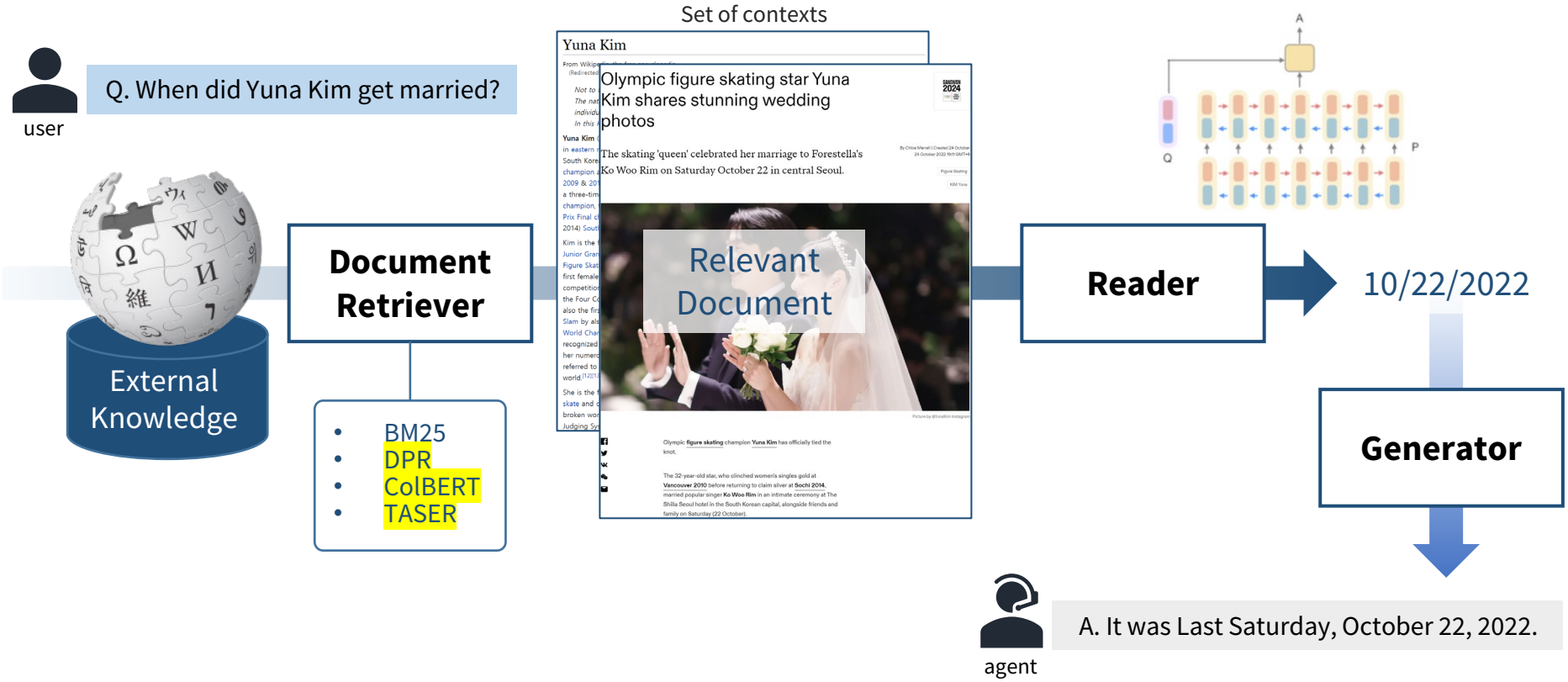
Pre-Requisites : What is ODQA?

- **Pipeline approaches (3-stage retriever-reader-generator system)**



Pre-Requisites : What is Dense Retriever?

- **Pipeline approaches (3-stage retriever-reader-generator system)**



Pre-Requisites : What is Dense Retriever?

- **Sparse vector Retriever vs. Dense vector Retriever**

BM25

- A variation of TF-IDF
- TF score is dampened after returning large numbers of matches btw the query and contexts.
- consider the document length
- Search Engine, Recommendation System, …

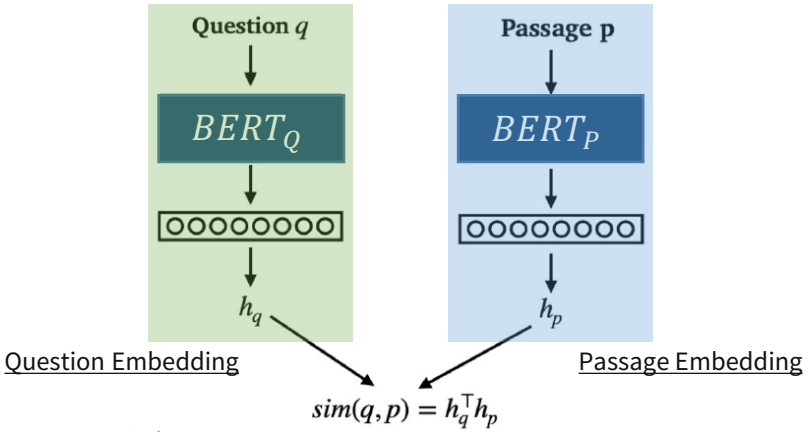
$$\sum_{term \in Q} IDF(term) \cdot \frac{TFIDF(term, D) \cdot (k_1 + 1)}{TFIDF(term, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{avgdl}\right)} + \delta$$

Training	Retriever	Top-20 NQ	TriviaQA	WQ	TREC	SQuAD
None	BM25	59.1	66.9	55.0	70.9	68.8
Single	DPR	78.4	79.4	73.2	79.8	63.2
	BM25 + DPR	76.6	79.8	71.0	85.2	**71.5**
Multi	DPR	**79.4**	78.8	**75.0**	**89.1**	51.6
	BM25 + DPR	78.0	**79.9**	74.7	88.5	66.2

DPR

- alternative to the traditional TF-IDF techniques for passage retrieval
- use **dense vectors** encoded with semantic meaning
- can train and fine-tune for specific tasks

$$sim(q, p) = E_Q(q)^T E_P(p)$$



initialized from the same pretrained model but separately parameterized for q & p

TASER

: Task-Aware Specialization for Efficient and Robust Dense Retrieval for Open-Domain Question Answering

Hao Cheng, Hao Fang, Xiaodong Liu, Jianfeng Gao

Microsoft, 2022

Yejin Yoon

Contents

Summary(Conclusion)

Summary (Conclusion)

- **TASER** can achieve superior accuracy, surpassing BM25,
  while using about 60% of the parameters as bi-encoder dense retrievers.

- In out-of-domain evaluations,
  TASER is also empirically more robust than bi-encoder dense retrievers.

Background & Problem States

Background

- **Background** : #DPR, #MoE

- **Dense retrieval models** become increasingly popular because of its effectiveness on knowledge-intensive NLP tasks
  - the *de-facto* architecture for ODQA uses two isomorphic encoders
- Dense retrieval Models is **inefficient**!
  - parameter-inefficient in that there is no parameter sharing between encoders
- Dense retrievers underperform BM25 in various settings

Problem States

- **More parameter-efficient and robust architecture**

  → Can we use only **1 encoder architecture** for question-passage representation?

Suggestion

- Model Architecture
- Pre-Requisites: What is MoE?
- Training

Model Architecture

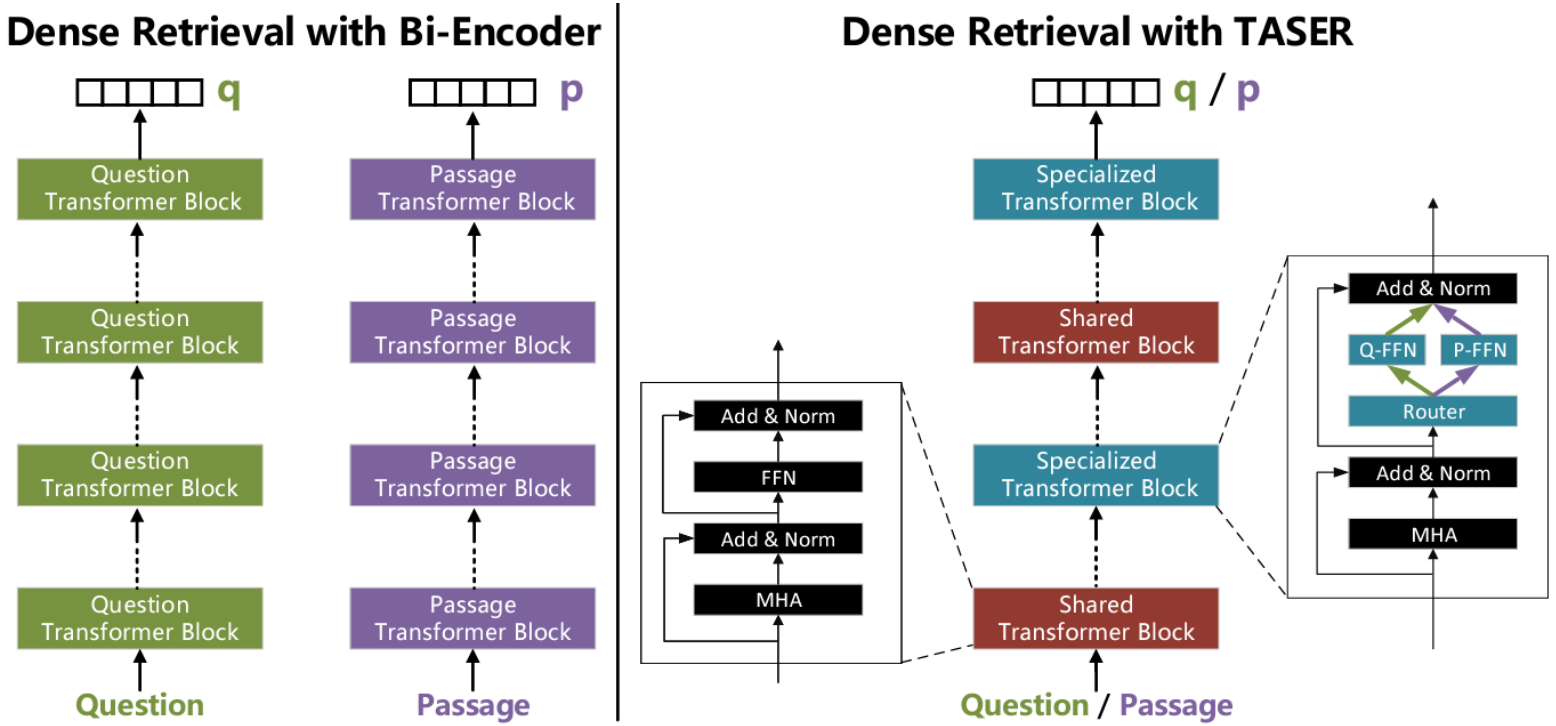
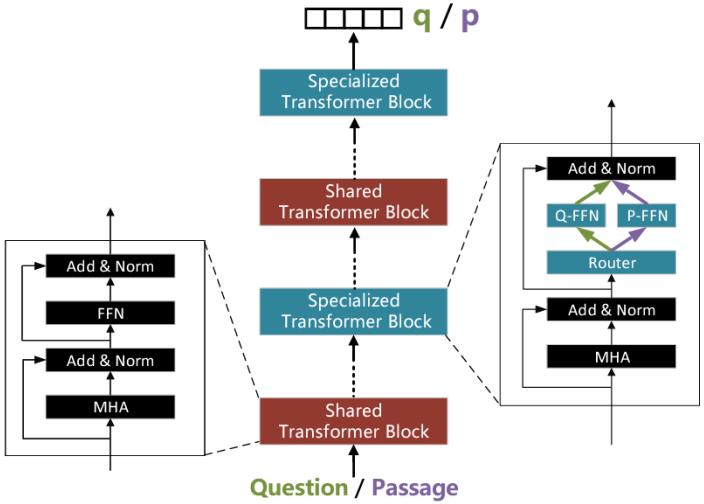


Figure 1: The dense retrieval architectures using a bi-encoder (left) and task-aware specialization (right). The question and passage transformer blocks in the bi-encoder are isomorphic to the shared transformer blocks in TASER. A specialized transformer block consists of several expert FFN sub-layers and a router. The router is used to choose among expert FFN sub-layers based on input. Only the deterministic routing Det-R is shown in the figure, which has two two expert FFN sub-layers (a Q-FFN for questions and a P-FFN for passages).

Model Architecture

- **TASER** interleaves **shared Transformer blocks** with **specialized Transformer blocks**
  - **shared Transformer block**
    : identical to the Transformer block used in the bi-encoder architecture
    : the entire block is shared for both questions and passages
  - **specialized Transformer block**
    : **MoE-style** task-aware specialization to the FFN sub-layer
    : use multiple expert FFN sub-layers in parallel, each with its own set of parameters
    : a *router* is used to choose among these expert FFN sub-layers



Pre-Requisites : What is MoE & Switch Transformer?

- **GShard**

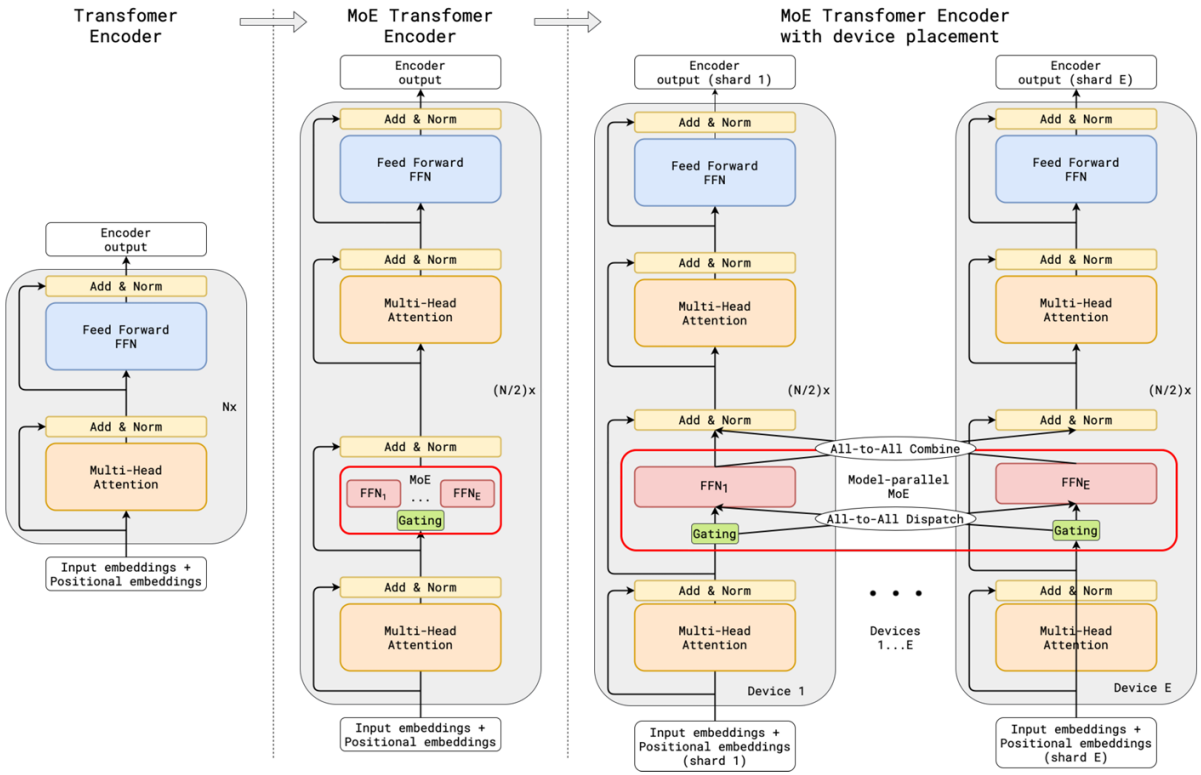
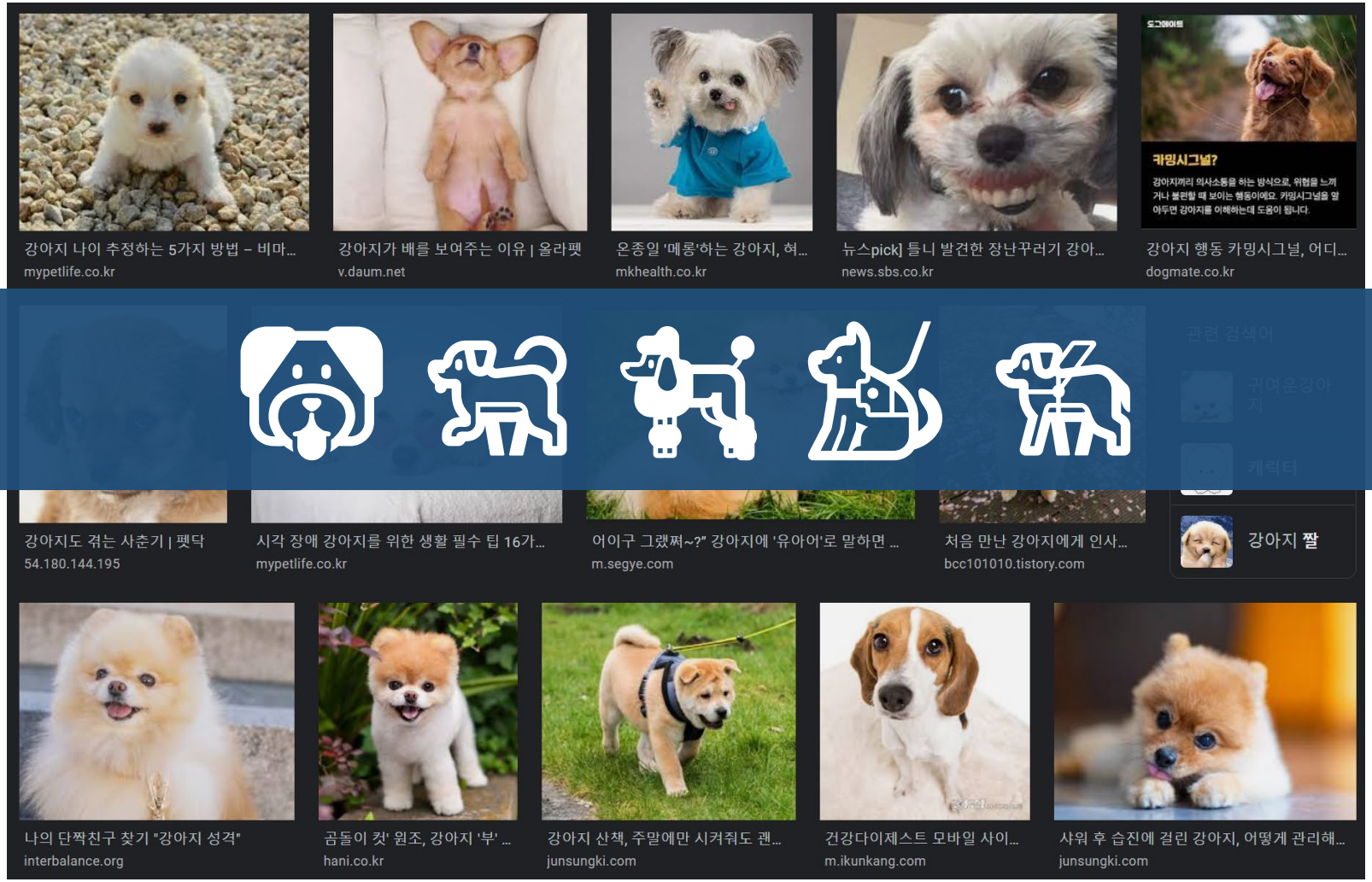


Figure 3: Illustration of scaling of Transformer Encoder with MoE Layers. The MoE layer replaces the every other Transformer feed-forward layer. Decoder modification is similar. (a) The encoder of a standard Transformer model is a stack of self-attention and feed forward layers interleaved with residual connections and layer normalization. (b) By replacing every other feed forward layer with a MoE layer, we get the model structure of the MoE Transformer Encoder. (c) When scaling to multiple devices, the MoE layer is sharded across devices, while all other layers are replicated.

Pre-Requisites : What is MoE & Switch Transformer?

- **Simple example of MoE**



Pre-Requisites : What is MoE & Switch Transformer?

- **MoE with routing mechanism**

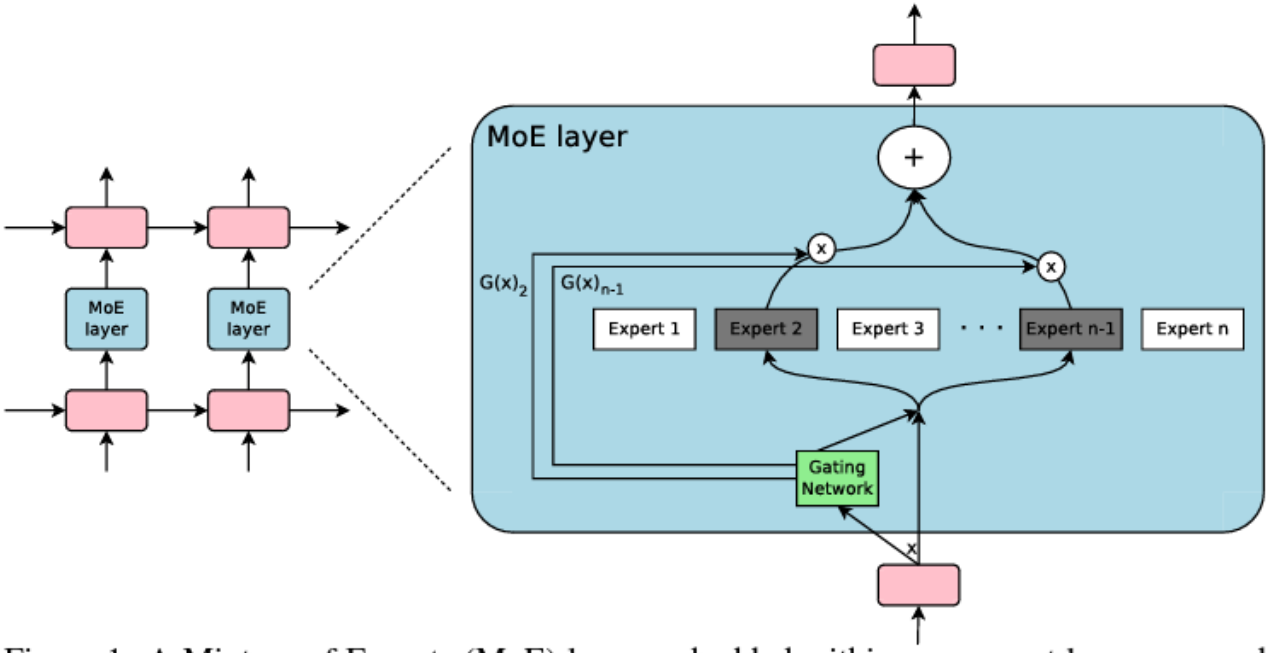


Figure 1: A Mixture of Experts (MoE) layer embedded within a recurrent language model. In this case, the sparse gating function selects two experts to perform computations. Their outputs are modulated by the outputs of the gating network.

Pre-Requisites : What is MoE & Switch Transformer?

- **MoE with routing mechanism**

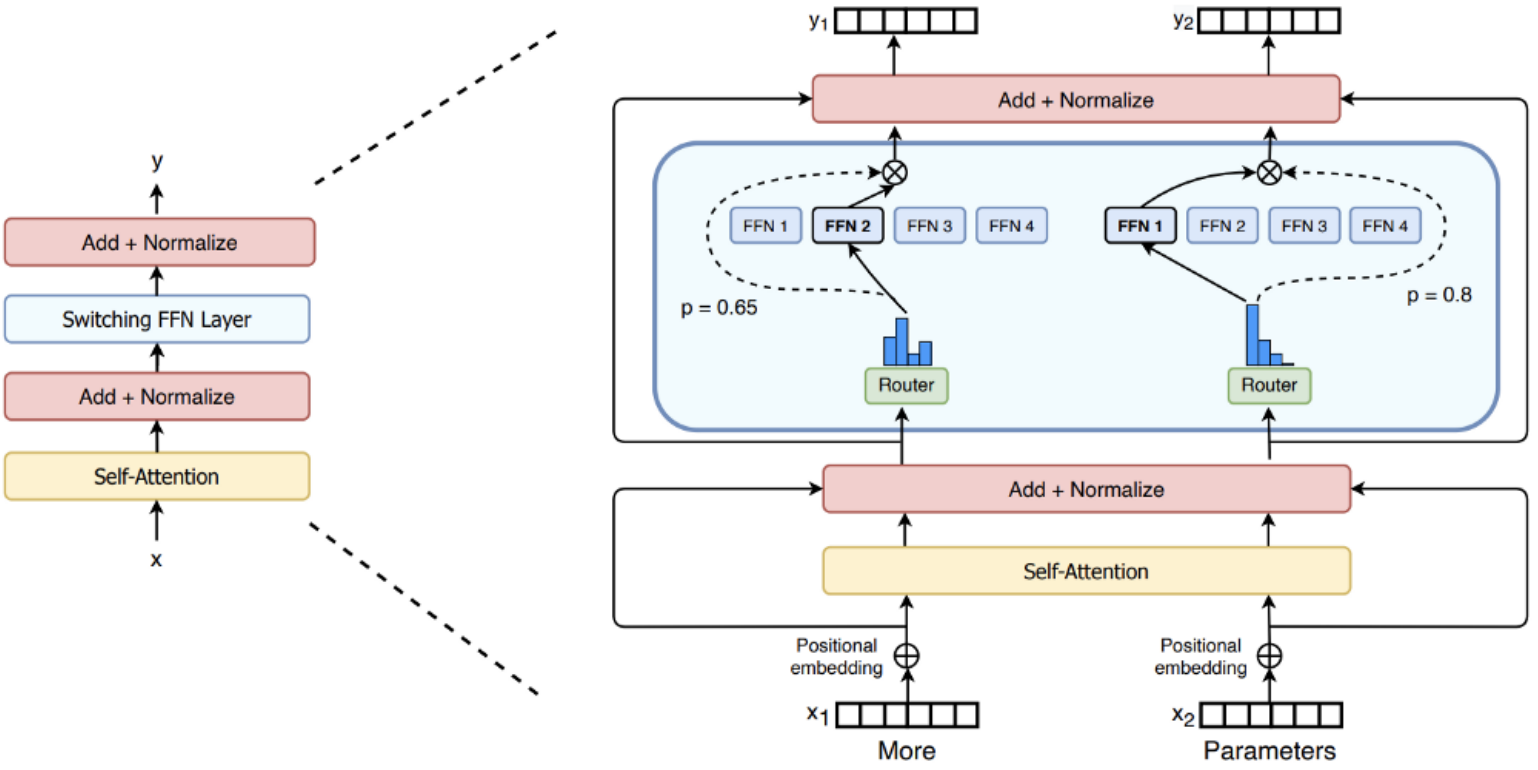
$$\mathcal{G}_{s,E} = GATE(x_s)$$
$$FFN_e(x_s) = wo_e \cdot ReLU(wi_e \cdot x_s)$$
$$y_s = \sum_{e=1}^{E} \mathcal{G}_{s,e} \cdot FFN_e(x_s)$$

$$\mathrm{h(x)} = \mathrm{W_r} \cdot \mathrm{x}$$
$$p = \mathrm{softmax(h(x))}$$

Select top-k experts !

Pre-Requisites : What is MoE & Switch Transformer?

- **Switch Transformer**



- As each token only passes through a single FFN, the computation does not increase, but the number of parameters increases with the number of experts.

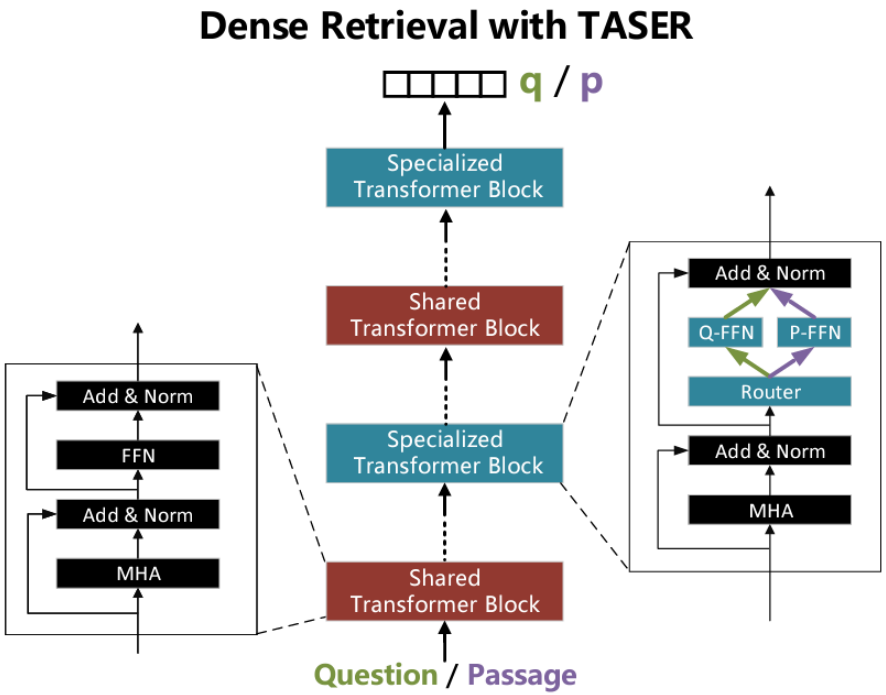
Model Architecture

- **TASER** uses one specialized Transformer block after every T shared Transformer blocks in the stack starting with a shared one at the bottom.

- ***Det-R***: the deterministic routing
  - only 2 expert FFN sub-layers are needed for ODQA retrieval: Query & Passage
  - router determines the expert FFN sub-layer based on whether the input us a question or a passage
  - using the contrastive learning objective L_{sim}

$$L_{sim} = -\frac{\exp(sim(q, p^+)}{\sum_{p' \in \mathcal{P} \cup \{p^+\}} \exp(sim(q, p')}$$

→similar to the bi-encoder architecture



Model Architecture

- **Seq-R & Tok-R**
  - Both are learned jointly with the task-specific objective
  - The router uses a parameterized routing function

$$R(\boldsymbol{u}) = GumbelSoftmax(\boldsymbol{A}\boldsymbol{u} + \boldsymbol{c})$$

    * GumbelSoftmax: a continuous distribution that approximates samples from a categorical distribution and also works with backpropagation
  - The routing function is jointly learned with all other parameters using the discrete reparameterization trick
  - Seq-R: routing is performed at the sequence level
    - All tokens in a sequence share the same $\boldsymbol{u} = \boldsymbol{h}_{[\boldsymbol{CLS}]}$
  - Tok-R: router independently routes each token $\boldsymbol{u} = \boldsymbol{h}_j$
  - Using the entropic regularization
    - To avoid routing all inputs to the same expert FNN sub-layer

$$L_{ent} = -\sum_{\{i=1\}}^{I} P(i) log P(i)$$

$$P(i) = \mathrm{Softmax}(\mathrm{Ah} + \mathrm{c})_{\mathrm{i}}$$

Model Architecture

- **Seq-R & Tok-R**
  - Both are learned jointly with the task-specific objective
  - The router uses a parameterized routing function

$$R(\boldsymbol{u}) = GumbelSoftmax(\boldsymbol{A}\boldsymbol{u} + \boldsymbol{c})$$

    * GumbelSoftmax: a continuous distribution that approximates samples from a categorical distribution and also works with backpropagation

  - The routing function is jointly learned with all other parameters using the discrete reparameterization trick

$$L_{joint} = L_{sim} + \beta L_{\text{ent}}$$

$$*hyperparameter\ \beta = 0.01$$

  - Seq-R: routing is performed at the sequence level
    - All tokens in a sequence share the same $\boldsymbol{u} = \boldsymbol{h}_{[CLS]}$
  - Tok-R: router independently routes each token $\boldsymbol{u} = \boldsymbol{h}_j$

$$L_{sim} = -\frac{\exp(sim(q, p^+)}{\sum_{p' \in \mathcal{P} \cup \{p^+\}} \exp(sim(q, p')} \cdots Det - R$$

  - Using the entropic regularization
    - To avoid routing all inputs to the same expert FNN sub-layer

$$L_{ent} = -\sum_{\{i=1\}}^{I} P(i) log P(i) \quad \cdots Seq - R, Tok - R$$

$$L_{ent} = -\sum_{\{i=1\}}^{I} P(i) log P(i)$$

$$P(i) = \text{Softmax}(\text{Ah} + \text{c})_{\text{i}}$$

Training

- **2-training paradigms: *single-set, multi-set***
  - *single-set* training: a model is trained using only a single dataset, evaluated on the same dataset
    - Model might be dataset-specific and NOT perform well in other datasets
  - *multi-set* training: a model trained by combining training data from multiple datasets to obtain a model that works well across the board (robust 😎)

- **Hard Negative Mining**
  - Recall that L_{sim} needs to use set of negative passage for each question!

  Step 1. train model with DPR negative sampling setting $\mathcal{P}_1$

  Step 2. conduct $\mathcal{P}_2$ by retrieving top-100 ranked passages for each question excluding the gold passage
  - *single-set* training: combine $\mathcal{P}_1$ & $\mathcal{P}_2$ to train final model
  - *multi-set* training: only use $\mathcal{P}_2$ to train the final model for efficiency consideration

Evaluation

- Datasets & Metrics
- Comparing TASER Variants
- In-Domain & Out-of-Domain Evaluation

Datasets & Metrics

- **In-domain evaluation**
  - NQ(Natural Questions)
  - TriviaQA
  - SQuAD
  - WebQ(Web Questions)
  - TREC(Curated Trec)
  - All data splits and the Wikipedia collection for retrieval used in our experiments are the same as DPR
  - Metric: top-K retrieval accuracy (R@K)
    - whether any gold answer string is contained in the top K retrieved passages

Datasets & Metrics

- **Out-domain evaluation**
  - **EntityQuestions**: consists of *entity-centric* questions with a broader set of entities which have different frequencies in Wikipedia.
  - **BEIR 😂 (NeurIPS 2021)**: a benchmark for zero-shot evaluation of retrieval systems, constructed from a diverse set of text retrieval datasets. (* license issue!)
    - ArguAna: long arguments from ideate.org
    - DBPedia: single-hop questions, articles from DBPedia
    - FEVER: short claims
    - HotpotQA : multi-hop questions

    } reflect the *model generalization performance* wrt richer query type and document index shifts
  - Metric
    - EntityQuestions: top-K retrieval accuracy (R@K)
    - BEIR: top-10 hits (nDCG@10)

Comparing TASER Variants

- ***single-set* training**: evaluating performance on NQ

- **Baseline: DPR** initialized from the **BERT-base**

- Finetuned up to **40** epochs with **Adam** using a learning rate chosen from $\{3e-5, 5e-5\}$

Model	I	# Params	Dev	Test
DPR	-	218M	-	78.4
$\text{TASER}_{\text{Shared}}$	1	109M	78.2	79.3
$\text{TASER}_{\text{Det-R}}$	2	128M	**79.2**	**80.7**
$\text{TASER}_{\text{Seq-R}}$	2	128M	**79.2**	80.6
$\text{TASER}_{\text{Seq-R}}$	4	166M	78.4	80.1
$\text{TASER}_{\text{ToK-R}}$	2	128M	78.5	79.8
$\text{TASER}_{\text{ToK-R}}$	4	166M	78.5	79.8
DPR†	-	218M	-	81.3
$\text{TASER}_{\text{Det-R}}$†	2	128M	**82.4**	**83.7**

Table 1: R@20 on NQ dev and test sets under the single-set training setting. I is the number of expert FFNs. The # params column shows the number of parameters in the model. † means the model is trained with hard negatives mining described in §3.3. The results for DPR and DPR† are reported in (Karpukhin et al., 2020) and https://tinyurl.com/yckar3f6, respectively.

In-Domain Evaluation

- Initialization
  based on **BERT-base**$\diamond$, **coCondenser-Wiki**$*$

Model	Num. Parameters
DPR	218M
coCodenser	218M
xMoCo	218M
SPAR-Wiki; SPAR-PAQ	436M
DPR-PAQ	710M
TASER$^{\diamond}$; TASER*	128M

- All TASER models use hard negatives mined from **NQ**, **TriviaQA**, **WebQ**.
  - NQ+TriviaQA for model selection
  - Other training details are the same

- **Linearly combined score**

$$sim(\mathbf{q}, \mathbf{p}) + \alpha \cdot BM25(\mathbf{q}, \mathbf{p})$$

  - α in the range [0.5, 2.0] with an interval of 0.1
  - Use single α for multi-set training instead of dataset-specified weights
  - Separately retrieve $K' \leq 100$ candidates from TASER and BM25, then retain the top K based on the hybrid scores

In-Domain Evaluation

Model	NQ @20	NQ @100	TriviaQA @20	TriviaQA @100	WebQ @20	WebQ @100	TREC @20	TREC @100	SQuAD @20	SQuAD @100
BM25[1]	62.9	78.3	76.4	83.2	62.4	75.5	80.7	89.9	71.1	81.8
Single-Set Training										
DPR[2]	78.4	85.4	79.4	85.0	73.2	81.4	79.8	89.1	63.2	77.2
DPR-PAQ[3]	84.7	**89.2**	-	-	-	-	-	-	-	-
coCondenser[4]	84.3	89.0	83.2	87.3	-	-	-	-	-	-
Multi-Set Training (without SQuAD)										
DPR[1]	79.5	86.1	78.9	84.8	75.0	83.0	88.8	93.4	52.0	67.7
DPR + BM25[1]	82.6	88.6	82.6	86.5	77.3	84.7	90.1	95.0	75.1	84.4
xMoCo[5]	82.5	86.3	80.1	85.7	78.2	84.8	89.4	94.1	55.9	70.1
SPAR-Wiki[6]	83.0	88.8	82.6	86.7	76.0	84.4	89.9	95.2	73.0	83.6
SPAR-PAQ[6]	82.7	88.6	82.5	86.9	76.3	85.2	90.3	95.4	72.9	83.7
Multi-Set Training (with SQuAD)										
DPR	80.9	86.8	79.6	85.0	74.0	83.4	88.0	94.1	63.1	77.2
TASER$^{\diamond}$	83.6	88.6	82.0	86.6	77.9	85.4	91.1	95.7	69.7	81.2
TASER$^{\diamond}$ + BM25	83.8	88.6	83.3	87.1	78.7	85.7	91.6	95.8	77.2	86.0
TASER*	84.9	**89.2**	83.4	87.1	78.9	85.4	90.8	**96.0**	72.9	83.4
TASER* + BM25	**85.0**	**89.2**	**84.0**	**87.5**	**79.6**	**85.8**	**92.1**	**96.0**	**78.0**	**87.0**

Model	Num. Parameters
DPR	218M
coCodenser	218M
xMoCo	218M
SPAR-Wiki; SPAR-PAQ	436M
DPR-PAQ	710M
TASER$^{\diamond}$; TASER*	128M

Table 2: In-domain evaluation results. Test set R@20 and R100 are reported. $^{\diamond}$ and * indicate TASER models are initialized using BERT-base and coCondenser-Wiki, respectively. [1]: (Ma et al., 2021). [2]: (Karpukhin et al., 2020). [3]: (Oğuz et al., 2021) [4]: (Gao and Callan, 2022). [5]: (Yang et al., 2021). [6]: (Chen et al., 2022).

Out-of-Domain Evaluation

	Macro R@20	Micro R@20	Micro R@100
BM25	71.2	70.8	79.2
DPR_{Multi}	56.7	56.6	70.1
TASER$^{\diamond}$	64.7	64.3	76.2
TASER*	66.7	66.2	77.9

Table 4: Out-of-domain evaluation results on EntityQuestions. Results for BM25 and DPR_{Multi} are from (Sciavolino et al., 2021) and (Chen et al., 2022).

	ArguAna	DBPedia	FEVER	HotpotQA	NQ
BM25	31.5	31.3	75.3	60.3	32.9
DPR_{Multi}	17.5	26.3	56.2	39.1	47.4
TASER$^{\diamond}$	32.8	31.4	59.6	50.7	51.3
TASER*	30.5	31.6	58.8	54.5	49.9

Table 5: Out-of-domain evaluation results on BEIR. nDCG@10 scores are reported. BM25 and DPR_{Multi} results are from (Thakur et al., 2021). Results on NQ reflect the in-domain performance.

Conclusion

Conclusion

- **TASER** improve the efficiency and robustness of dense retrieval for ODQA.
  - parameter efficiency: interleaves shared encoder blocks with specialized ones in a single encoder where some sub-networks are task-specific (almost no additional computation cost!)
  - robustness: outperforms 5 in-domain datasets & 2 OOD benchmarks

- Similar to bi-encoder models, advanced techniques can be applied to **TASER** to achieve further improvement in retrieval performance
  - Hard negatives mining
  - Ensembling with BM25

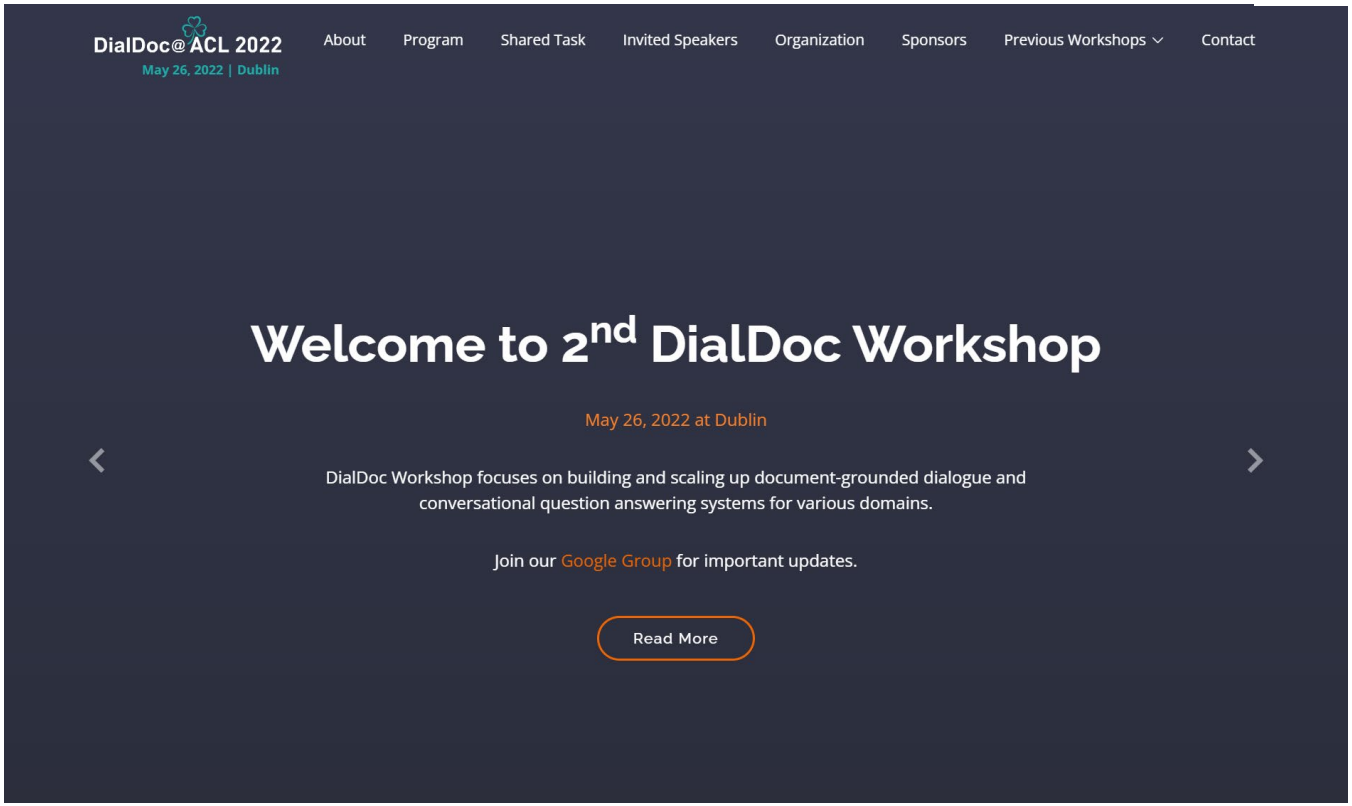
Limitation

- In-domain evaluation focus on **passage retrieval** for ODQA
  - It can be also used in other types of retrieval tasks which may have different input and output format 😎
    - KILT benchmark, …
- The cost of training dense vector model
  - Although **TASER** significantly reduce the number of model parameters, the training cost is still high 📈
- The learned routing(Seq-R, Tok-R) does not outperform the Det-R
- There is still a **gap** between **TASER** and **BM25** in OOD evaluation 😢

Personal Opinions

- **Isn't it relatively worth a try?** compared to BB3…👀

- **SHARED TASK at $\mathbb{D}ial\mathbb{D}oc$**
  - DialDoc Workshop @ACL: 🔗 https://doc2dial.github.io/



{ End Page }

Thank you :D

Yejin Yoon

HYU NLP Lab.
Dept. of Artificial Intelligence Application,
Hanyang University

stillwithyou@hanyang.ac.kr

HYU 한양대학교 HANYANG UNIVERSITY