Joint Passage Ranking for Diverse Multi-Answer Retrieval

Sewon Min, Kenton Lee, Ming-Wei Chang, Kristina Toutanova, Hannaneh Hajishirzi

EMNLP 2020

인공지능학과 김수형

저자 소개

First Author
Sewon Min (Ph.D. student in Washington Univ.)

Interests - NLU, QA

Publications
- Joint Passage Ranking for Diverse Multi-Answer Retrieval, EMNLP(2021)
- NeurIPS 2020 EfficientQA Competition: Systems, Analyses and Lessons Learned, PMLR(2021)

Co-Author
Hannaneh Hajishirzi(Assistant Professor Washington Univ.)

Interests - QA, reading comprehension, representation learning

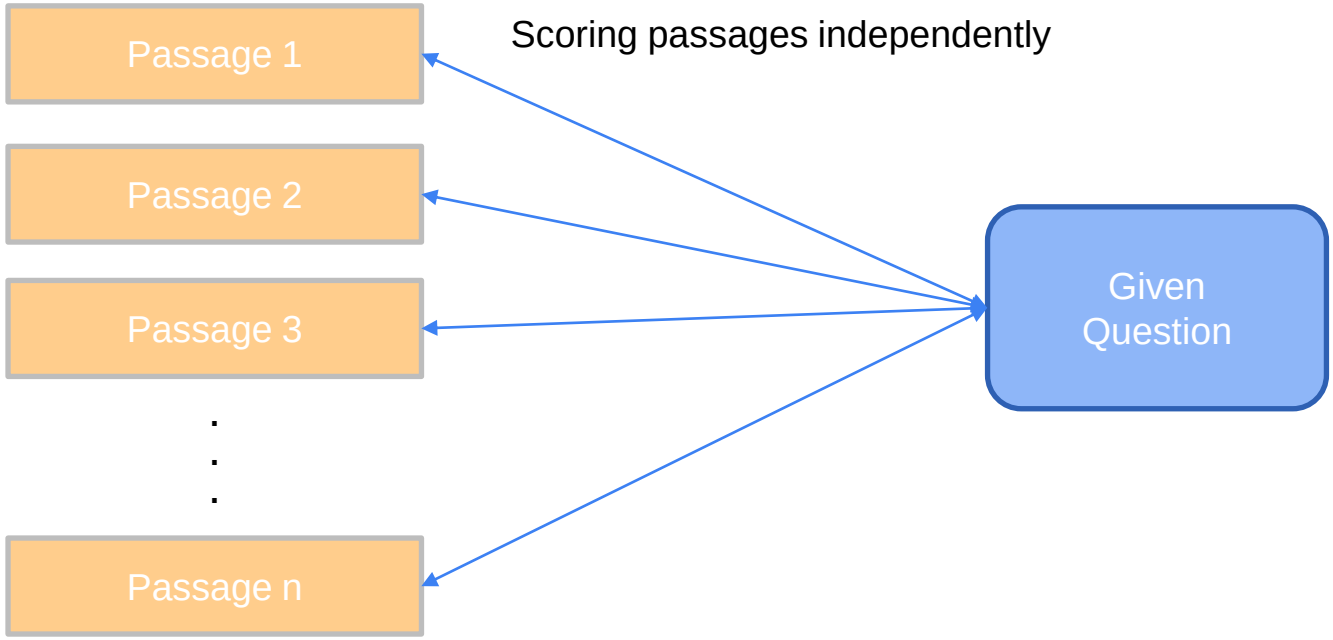
Publications
- One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval, NeurIPS(2021)
- DIALKI: Knowledge Identification in Conversational Systems through Dialogue-Document Contextualization, EMNLP(2021)

목차

- Motivation
- Introduction
  - AMBIGQA task
  - Evaluation metrics
  - AMBIGNQ dataset
- Model Overview
- Experiments
- Ablation Study
- Summary

Motivation

기존의 QA method : one question -> single answer



Motivation

But Questions are open-ended & ambiguous.

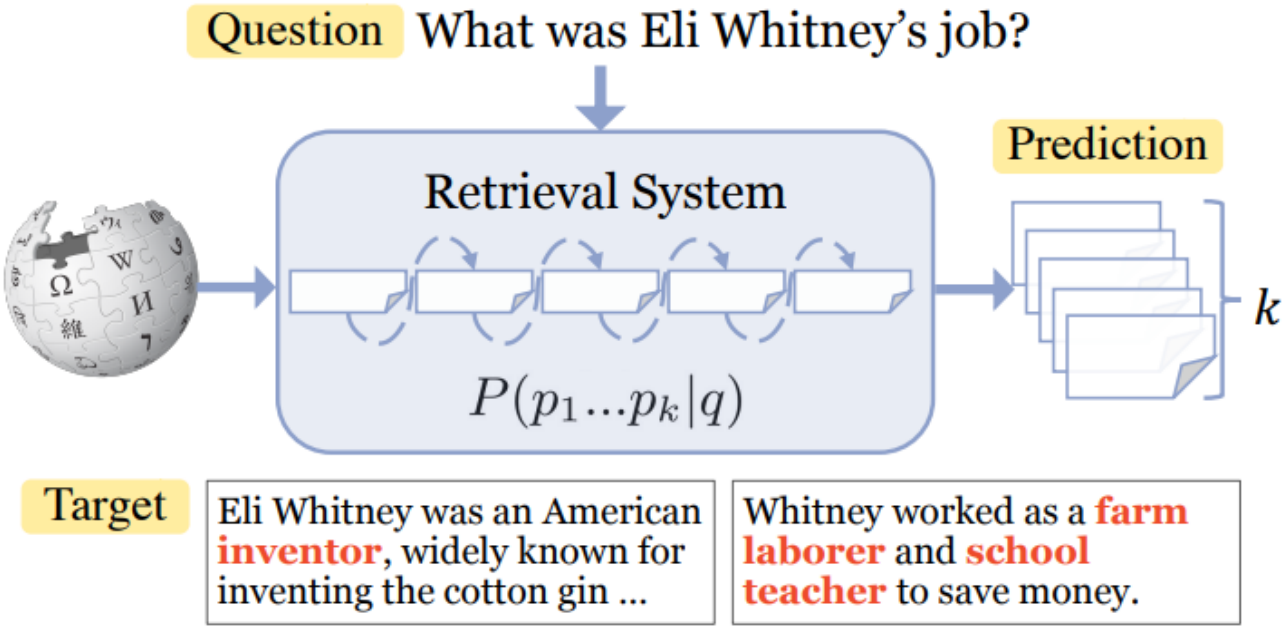
Question What was Eli Whitney's job?

Target

Eli Whitney was an American **inventor**, widely known for inventing the cotton gin ...	Whitney worked as a **farm laborer** and **school teacher** to save money.

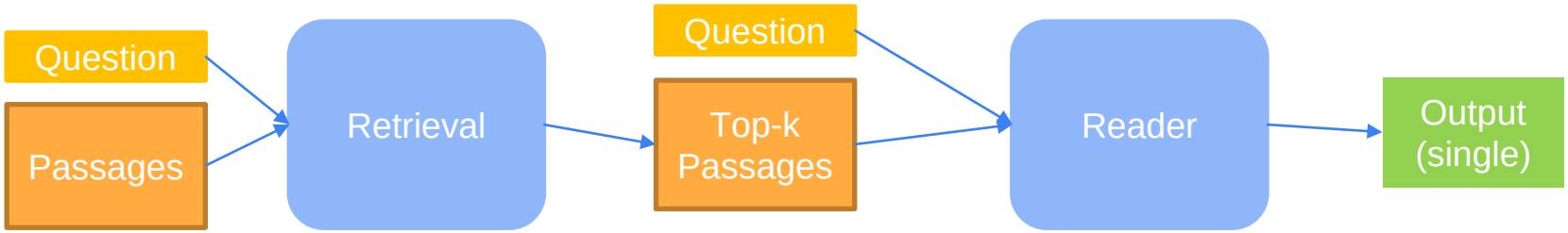
Motivation

But Questions are open-ended & ambiguous.

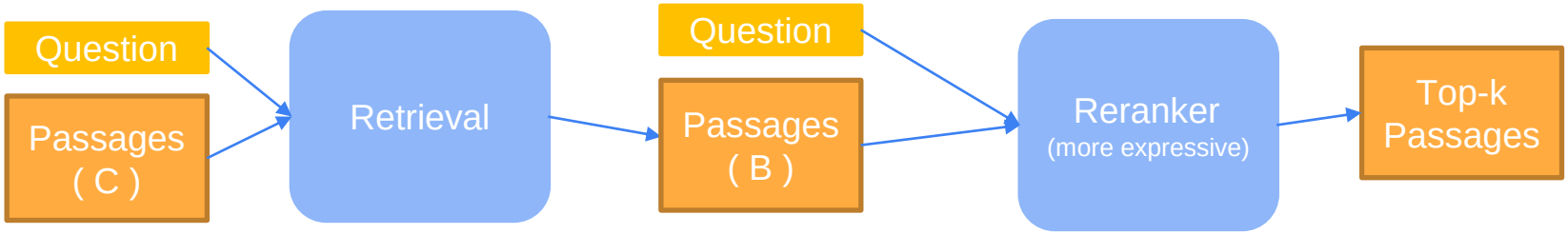


Background

Single-Answer Retrieval

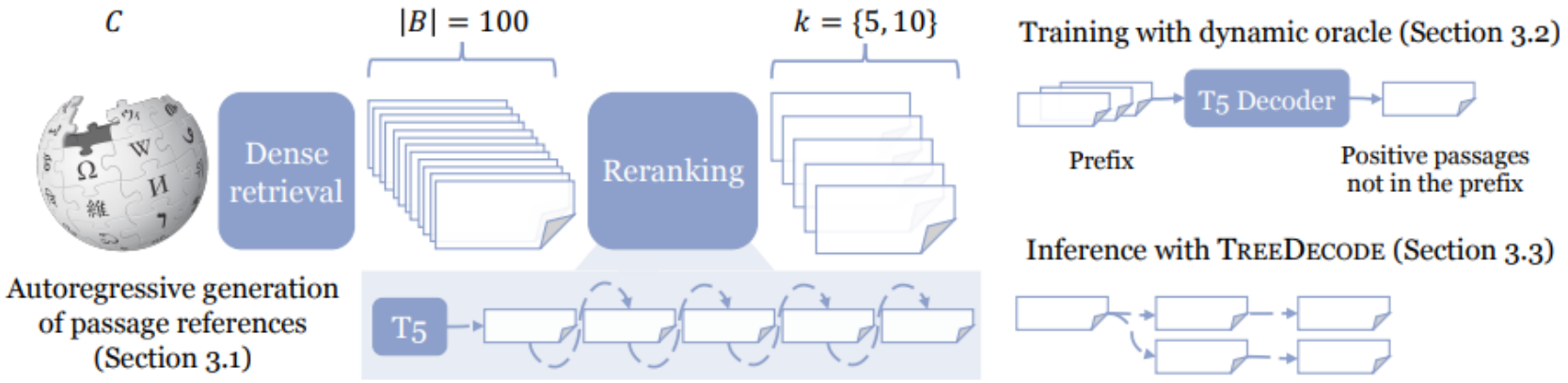


Reranking

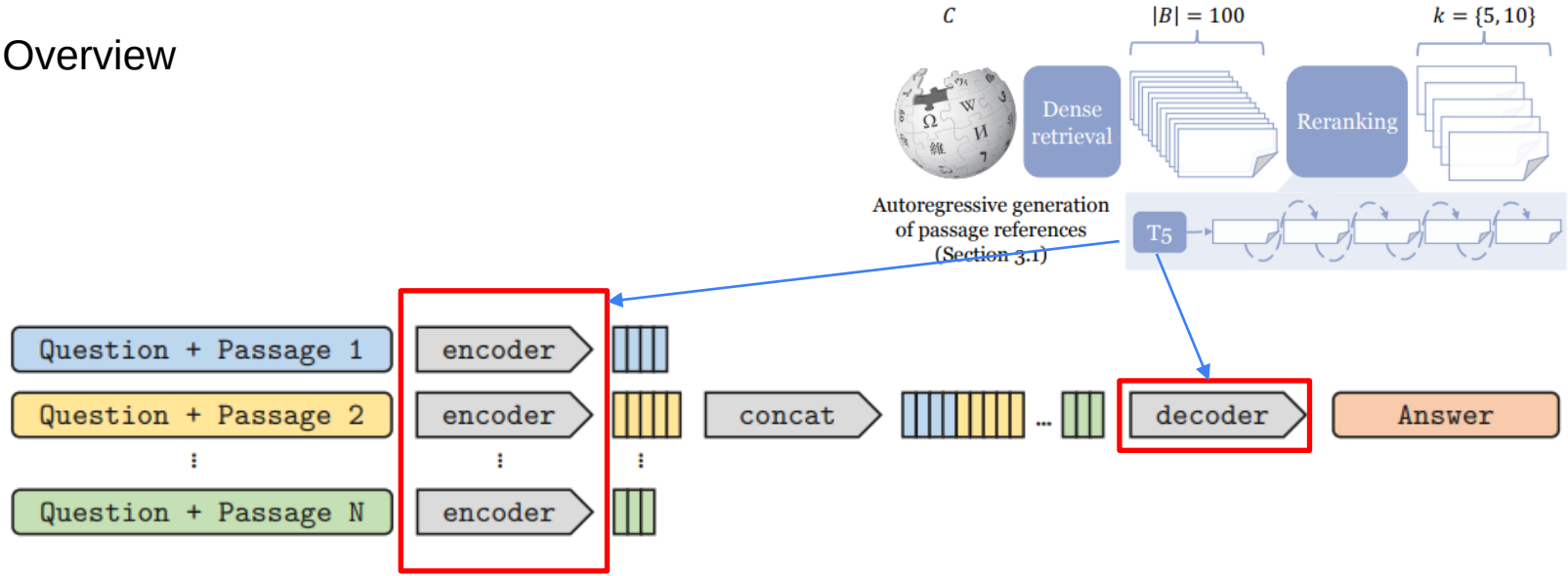


$$(k < |\mathcal{B}| \ll |\mathcal{C}|)$$

Model Overview

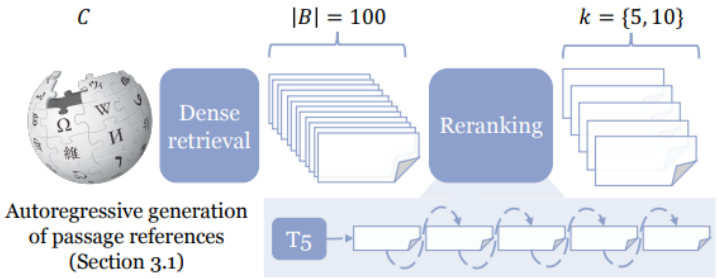


Model Overview



Izacard, Gautier, and Edouard Grave. "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering." *EACL 2021-16th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2021.

Model Overview



Passage - 1

Question

1 (index)

Encoder

En Output -1

En Output -2

.
.
.

En Output -B

Decoder

$\tilde{p}_k$ = index

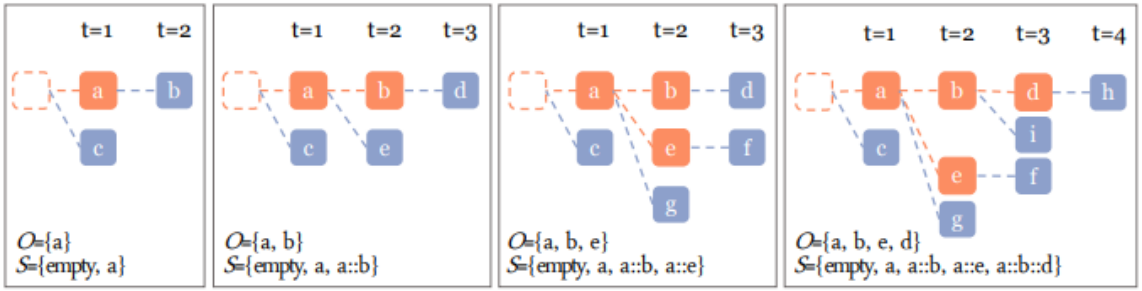
$$\tilde{p}_1 \ldots \tilde{p}_k$$

$$\sum_{1 \le t \le k} \sum_{o \in \tilde{\mathcal{O}} - \mathcal{P}_{t-1}} -\log P(o|q, \mathcal{B}, \mathcal{P}_{t-1})$$

$\tilde{\mathcal{O}}$ = positive passages

$\tilde{p}_1 \ldots \tilde{p}_t$ (denoted as $\mathcal{P}_t$)

Model Overview



Inference with TREEDECODE (Section 3.3)

Algorithm 1 Decoding algorithms for JPR.

procedure SEQDECODE(k, $P(p|p_1...p_n)$)
  $O \leftarrow []$ *// a list of selected passages*
  while $i = 1, \ldots, k$ **do**
    $\hat{p} \leftarrow \text{argmax}_{p \in \mathcal{B} - O.\text{toSet}()} \log P(p|O)$
    $O \leftarrow O :: \hat{p}$
  return Set(O)
procedure TREEDECODE(k, $P(p|p_1...p_n)$, l)
  $\mathcal{O} \leftarrow \emptyset$ *// a set of selected passages*
  $\mathcal{S} \leftarrow$ [Empty] *// a tree*
  while $|\mathcal{O}| < k$ **do**
    $P'(p|s) \leftarrow P(p|s)\mathbb{I}[s :: p \notin \mathcal{S}]$
    $(\hat{s}, \hat{p}) \leftarrow \text{argmax}_{p \in \mathcal{B}, s \in \mathcal{S}} l(|s|+1) \log P'(p|s)$
    $\mathcal{O} \leftarrow \mathcal{O} \cup \{\hat{p}\}, \mathcal{S} \leftarrow \mathcal{S}.\text{append}(\hat{s} :: \hat{p})$
  return $\mathcal{O}$

$$f(\mathcal{S}) = \sum_{p_1 \ldots p_{t_i} \in \mathcal{S}} \log P(p_{t_i} | p_1 \ldots p_{t_{i-1}})$$

Intuitively, if the model believes that there are many distinct answers covered by different passages, it will choose to take the next step, being closer to SEQDECODE. On the other hand, if the model believes that there are very few distinct answers, it will choose to take more predictions from the same step, resulting in behavior closer to independent scoring.

Experiments - Results

k	Models	WEBQSP		AMBIGQA	TREC	
		Dev	Test	Dev	Dev	Test
5	DPR$^+$ only	56.4/37.8	57.0/38.9	55.2/36.3	53.8/**29.9**	57.8/36.6
	DPR$^+$+Nogueira et al. (2020)	60.2/40.9	60.2/39.9	63.4/43.1	53.8/28.4	61.0/39.5
	INDEPPR	60.6/40.2	62.9/45.2	63.7/43.7	53.8/28.4	**62.4/41.1**
	JPR	**68.5/56.7**	**64.9/50.6**	**64.8/45.2**	**55.5/29.9**	**62.4/41.1**
10	DPR$^+$ only	61.4/42.5	59.0/38.6	59.3/39.6	55.5/28.4	60.1/38.4
	DPR$^+$+Nogueira et al. (2020)	64.7/45.7	62.9/41.5	65.8/46.4	55.5/28.4	**64.8**/43.0
	INDEPPR	65.6/47.2	63.3/43.1	65.5/46.2	53.8/26.9	63.8/42.2
	JPR	**68.9/55.1**	**65.7/48.9**	**67.1/48.2**	**56.3/29.9**	64.5/**43.3**

Training method	MRECALL
Dynamic oracle	**67.6/56.7**
Dynamic oracle w/o negatives	65.1/52.0
Teacher forcing	66.4/51.2

Table 4: Ablations in training methods for JPR. Results on WEBQSP ($k = 5$). All rows use SEQDECODE (instead of TREEDECODE).

k	Decoding	WEBQSP		AMBIGQA	
		d	MRECALL	d	MRECALL
5	SEQDECODE	5.0	67.6/**56.7**	5.0	63.1/42.5
	TREEDECODE	3.0	**68.5/56.7**	2.1	**64.8/45.2**
10	SEQDECODE	10.0	68.0/54.3	10.0	65.0/45.9
	TREEDECODE	5.4	**68.9/55.1**	2.9	**67.1/48.2**

Table 5: Ablations in decoding methods for JPR. d refers to the average depth of the tree ($\max_{s \in \mathcal{S}} |s|$ in Algorithm 1).

Experiments - Results

Q: Who play Mark on the TV show Roseanne?	
INDEPPR	JPR
#1 Glenn Quinn ... He was best known for his portrayal of Mark Healy on the popular '90s family sitcom Roseanne.	**#1** Glenn Quinn ... He was best known for his portrayal of Mark Healy on the popular '90s family sitcom Roseanne.
#2 Glenn Quinn, who played Becky's husband, Mark, died in December 2002 of a heroin overdose at the age of 32 ...	**#2** Becky begins dating Mark Healy (Glenn Quinn) ...
#3 Becky begins dating Mark Healy (Glenn Quinn) ...	**#3** Glenn Quinn, who played Becky's husband, Mark, died in December 2002 of a heroin overdose at the age of 32 ...
#4 Johnny Galecki ... on the hit ABC sitcom Roseanne as the younger brother of Mark Healy (Glenn Quinn) ...	**#4** Roseanne (season 10) ... In September 2017, Ames McNamara was announced to be cast as Mark Conner-Healy.

Table 6: An example prediction from INDEPPR and JPR; answers to the input question highlighted. While INDEPPR repeatedly retrieves passages supporting the same answer *Glenn Quinn* and fails to cover other answers, JPR successfully retrieves a passage covering a novel answer *Ames McNamara*.

Ablation study - disambiguation

Retrieval	QA Model	k	Mem	WEBQSP		AMBIGQA		TREC	
				Dev	Test	Dev	Test	Dev	Test
DPR$^+$ only	T5-3B	10	x1	50.7/45.3	51.9/45.0	43.5/34.6	39.6/31.4	46.2/32.2	44.7/32.1
INDEPPR	T5-3B	10	x1	51.8/46.9	51.8/45.0	47.6/36.2	42.3/32.0	44.6/32.8	45.9/31.8
JPR	T5-3B	10	x1	**53.6/49.5**	**53.1/47.2**	**48.5/37.6**	**43.5/34.2**	**48.6/32.8**	**46.8/33.3**
DPR$^+$ only	T5-large	40	x1	51.4/47.0	52.4/45.8	45.5/34.9	41.1/30.9	40.1/32.8	42.5/32.2
Gao et al. (2021)	BART-large	100	x3	-	-	48.3/37.3	42.1/33.3	-	-

Table 7: Question Answering results on multi-answer datasets. The two values in each cell indicate F1 on all questions and F1 on questions with multiple answers only, respectively. *Mem* compares the required hardware memory during training. Note that Gao et al. (2021) reranks 1000 passages instead of 100, and trains an answer generation model using x3 more memory than ours. **Better retrieval enables using larger answer generation models on fewer retrieved passages.**

Summary

- Reranking을 활용한 Multi-answer Retrieval System
- Transformer decoder를 활용하여 passage 에 조건부 수식을 적용하였다.
- 기존의 single answer 모델의 약점을 잘 파고들었고 성능 또한 좋게 나왔다.

Q&A

Appendix