RAG를 이용한 한국어 오픈 도메인 질의 응답

Rertieval-Augmented Generation for Korean Open-domain Question Answering

강대욱, 나승훈, 김태형, 류휘정, 장두성
전북대학교, KT

강대욱, 나승훈,김태형, 류휘정, 장두성
전북대학교, KT

Introduction
Related Works
Methods
Experiment
Results
Conclusion

Introduction

Introduction

오픈 도메인 질의 응답이란

- 질의를 처리하기 위해서 대량의 문서에서 정답을 찾는 태스크
- 질의와 유사한 문자의 등장 여부를 측정하는 BM25 및 TF-IDF 등의 방식을 사용

Introduction

BM25 및 TF-IDF는 질의와 문서의 문자적 유사성을 측정하므로 질의와 유사한 의미더라도 문자적 중복이 없다면 탐색이 어려움

최근 딥러닝 신경망을 사용해 질의와 문서를 각각 인코딩한 뒤 유사도를 구하는 Dense Retrieval 방식은 문자적 중복이 없더라도 의미를 통해 연관 문서를 찾을 수 있음

Introduction

Dense Retrieval 방식을 사용하는 연구 중 하나인 RAG는 검색한 문서와 질의를 사용해 인코더-디코더 모델에서 정답을 생성하여 정답이 문서에 직접적으로 등장하지 않아도 정답을 생성해낼 수 있음

본 논문에서는 RAG를 한국어 오픈 도메인 질의 응답 데이터에 실험하여 성능을 측정

Related Works

REALM

BERT 기반의 질의 및 문서 인코더와 Reader 모델을 사전학습 한 뒤 검색한 문서에서 정답 span을 예측하도록 미세조정

검색한 문서에서 정답 span을 직접 찾아내기 때문에 문서에 정답이 직접 등장하지 않는다면 적절한 답변이 어려움

RAG

질의를 질의 인코더로 입력해 인코딩된 벡터를 얻음

문서 인코더를 이용해 문서들의 벡터를 생성

질의와 문서의 벡터를 내적 연산해 Top-K 문서를 얻음

Top-K 문서와 질의를 인코더-디코더 모델에 입력해 정답을 생성

RAG

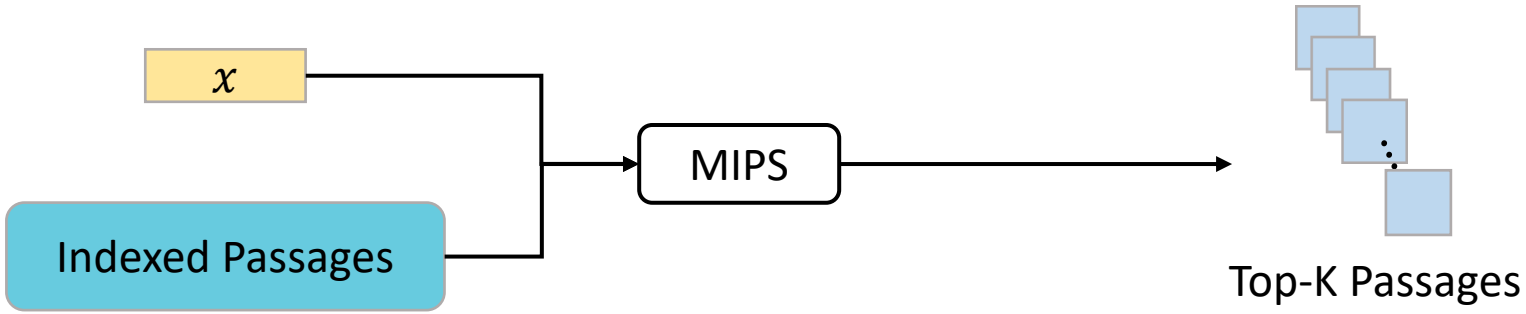
문서와 질의를 인코더-디코더 모델에 입력해 정답을 생성하기 때문에 REALM 등 문서에서 span을 탐색하는 모델들과 달리 문서에 정답이 직접 등장하지 않아도 파라미터의 지식을 이용해 정답을 생성할 수 있음

Methods

RAG Retriever

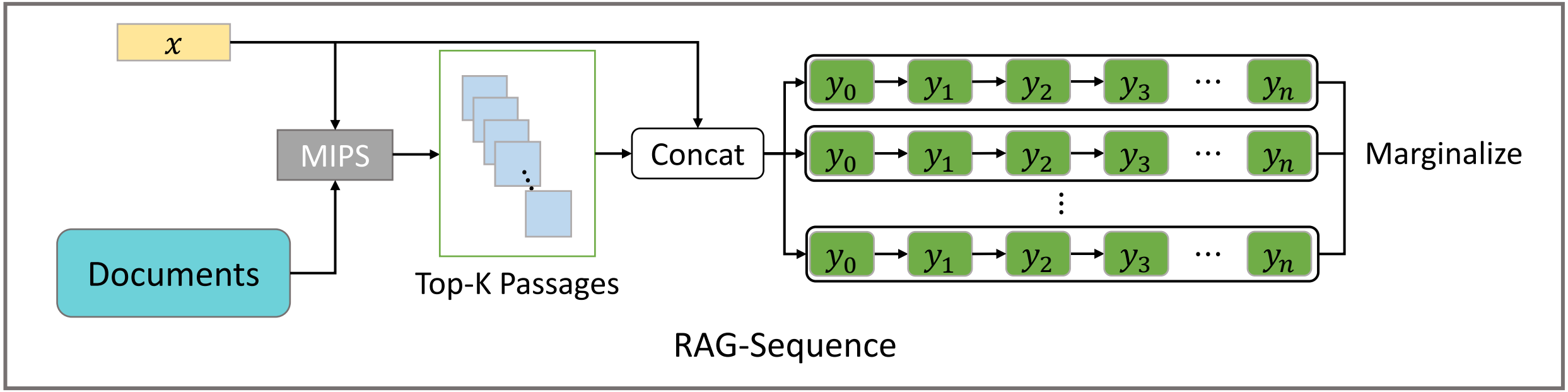
입력받은 질의 x를 인코딩한 벡터와 미리 인코딩 해 놓은 문서들의 벡터를 MIPS(Maximum Inner Product Search) 연산하여 Top-K 문서 산출

산출한 Top-K 문서들을 질의 x와 함께 인코더-디코더 모델에 입력해 정답 생성



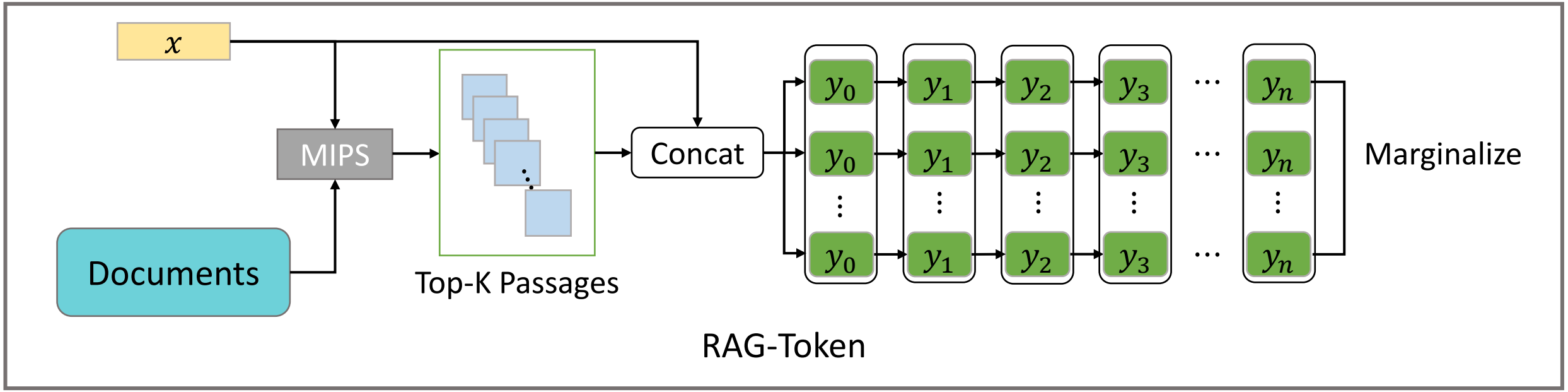
RAG-Sequence

RAG-Sequence 모델은 각각의 Top-k 문서를 사용해 문장의 모든 토큰을 생성한 뒤 취합하여 문장의 확률을 구함



RAG-Token

RAG-Token 모델은 매 토큰마다 각 Top-K 문서들로 토큰을 생성한 뒤 취합해 토큰을 확률을 구하며 이를 모두 곱해 문장의 확률을 구함



Experiments

Experiment

REALM의 인코더를 Dense Retriever로 사용

인코더-디코더 모델로 한국어 T5를 사용

한국어 위키피디아 20년 5월 1일자 덤프를
외부 지식으로 사용

KTQA 데이터에서 20 epoch 동안 미세조정 후 성능 측정

KorQuAD v1.0과 동일한 방식으로 EM(Exact Matching) 및 F1 을 측정

검색한 문서에 정답이 있는 경우만을 Has Answer로 분류해 별도로
성능을 측정

블록 수	블록 당 평균 문장 수	블록 당 평균 단어 수
87,233	4.94	67.81

표 1. 한국어 위키피디아 데이터 구성

	$\|\mathcal{D}\|$	Avg.$\|\mathcal{A}\|$
Train	15,900	1.36
Dev	900	1.35
Test	1,800	1.35

표 2. KTQA 데이터 구성

Results

Results

Model	All		Has Answer	
	EM	F1	EM	F1
REALM	50.80	63.61	**76.78**	85.06
RAG-Token	**53.14**	**66.53**	67.41	**86.08**
RAG-Sequence	50.25	62.93	64.30	81.59

Conclusion

Conclusion

한국어 오픈 도메인 질의 응답 데이터에 RAG를 적용하여 기존 REALM 등의 방식과 성능을 비교

RAG가 KTQA 데이터에서 일부 향상된 성능을 보임을 확인

Future Work

RAG 모델을 오픈 도메인 질의 응답 및 이외 다양한 자연어 처리 분야에 적용

감사합니다