WIKIPEDIA QA

BY MOHAMED ZAKARIA, ALI AMR & SARAH HELLA

MOTIVATION

- Large amount of information available on the web
- Information may not be reachable due to overwhelming results
- Incorrect answers may be extracted

OUR GOAL

- Create a question answering system to solve this issue
- Utilize data available on Wikipedia
- Reduce the time lost searching through countless documents

CHALLENGES

1. Gathering information from wikipedia
2. Large number of documents to search through
3. Retrieved document may not contain the answer
4. If we extract multiple answers from different documents, how to rank them?

DATASET

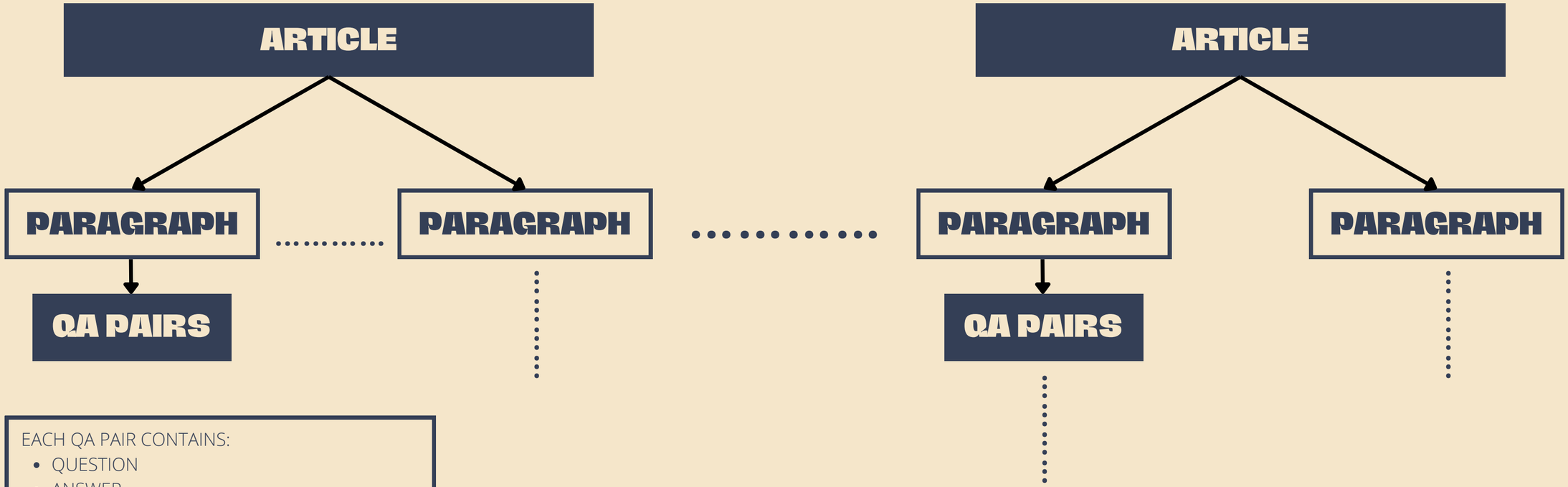
SQUAD1.1

Dataset contains 100,000+ questions that can be used for the extractive QA task. It is considered a benchmark dataset for QA models.

SQUAD2.0

Extends upon SQuAD1.1 by adding unanswerable questions to the data. This challenges the model's ability to refrain from answering when an answer is not present. This is the dataset we will use.

DATASET STRUCTURE



EACH QA PAIR CONTAINS:

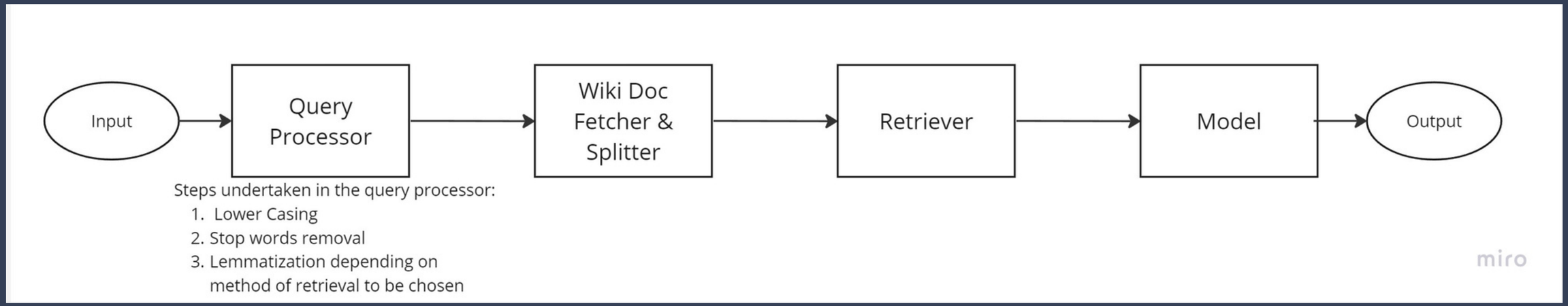
- QUESTION
- ANSWER
- ANSWER START IN CONTEXT
- IS ANSWER POSSIBLE OR NO
- PLAUSIBLE ANSWERS IF QUESTION IS UNANSWERABLE

DATA ANALYSIS

Dataset	Size	Articles	Contexts	Answerable,%	Unanswerable,%	Vocab Size	Stop Words
TRAIN	130319	442	19020	86821 , 66.6%	43498 , 33.37%	89982	623
DEV	11873	35	1204	5928 , 49.9%	5945 , 50.07%	18770	461

Table 1: Features of the SQuAD2.0 Train & Dev sets

SYSTEM ARCHITECTURE



THANK YOU

Any Questions?