Question Answering Using Question Generation

Information Retrievers

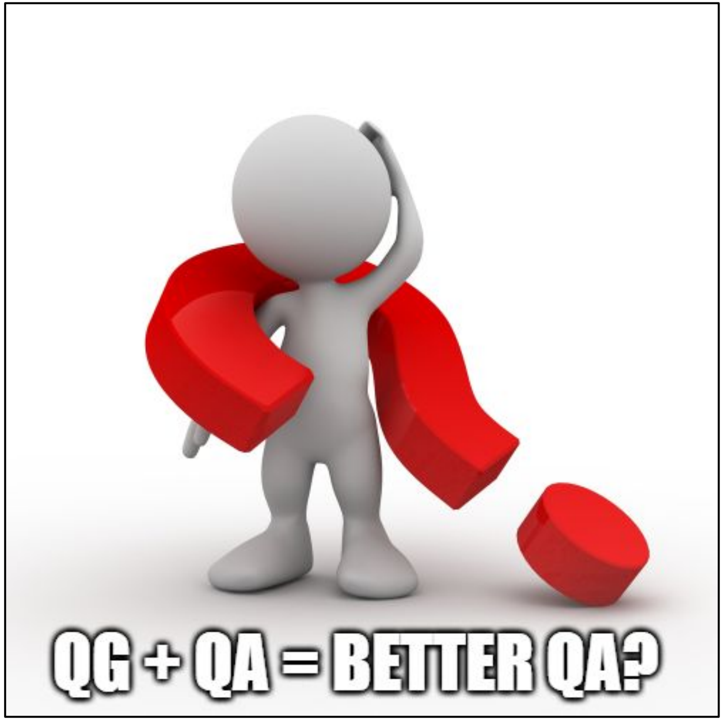
Purpose of Experiment

Question Generation:

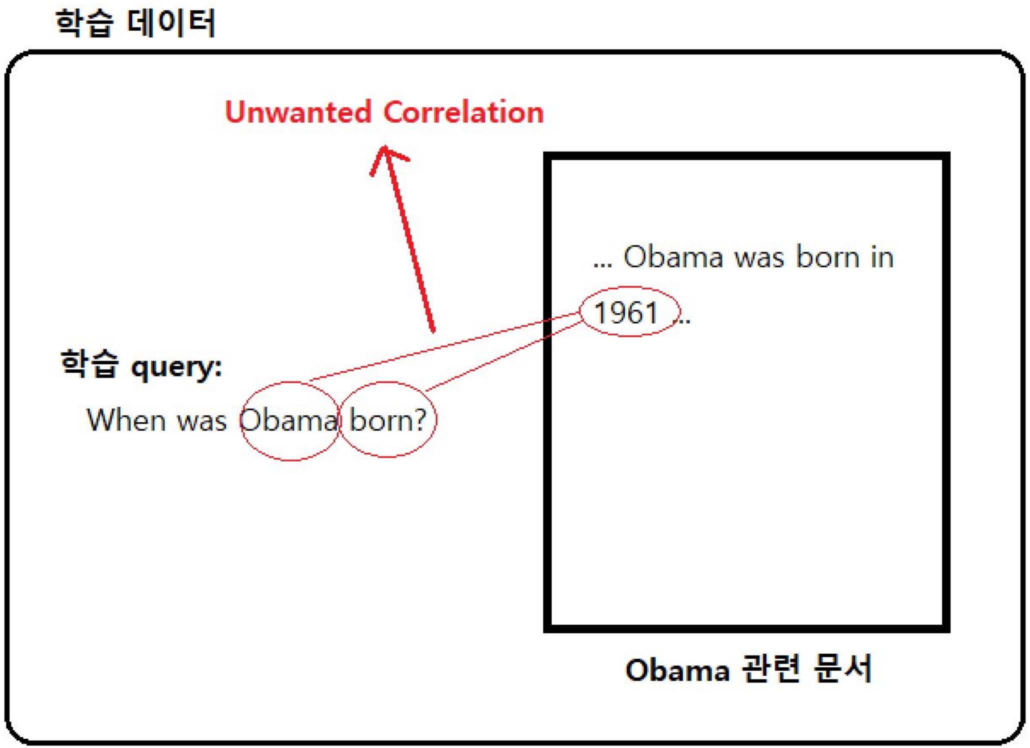
Generating a question for a given answer.

Question Answering:

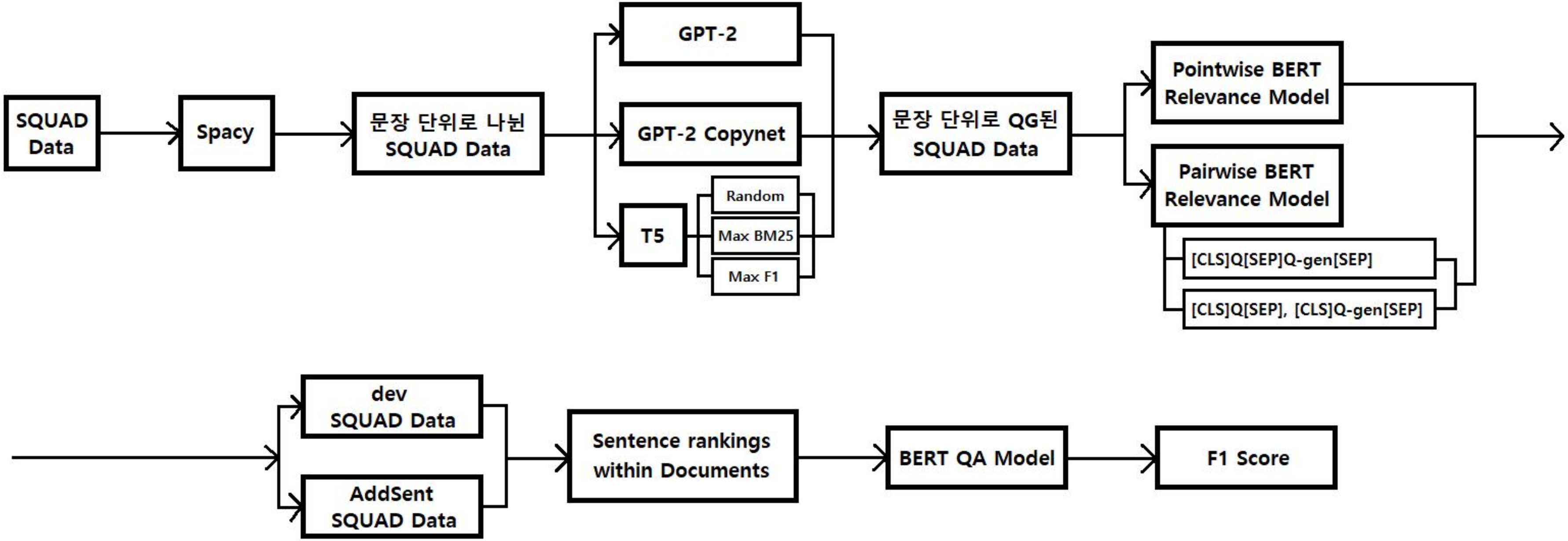
Getting an answer for a given question.



Limitations of Existing Research

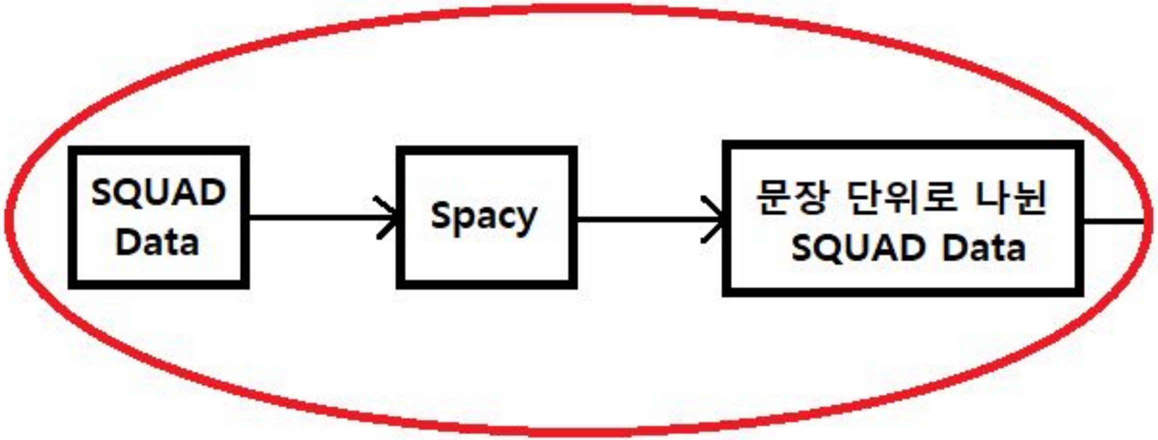


Experiment Process: An Overview



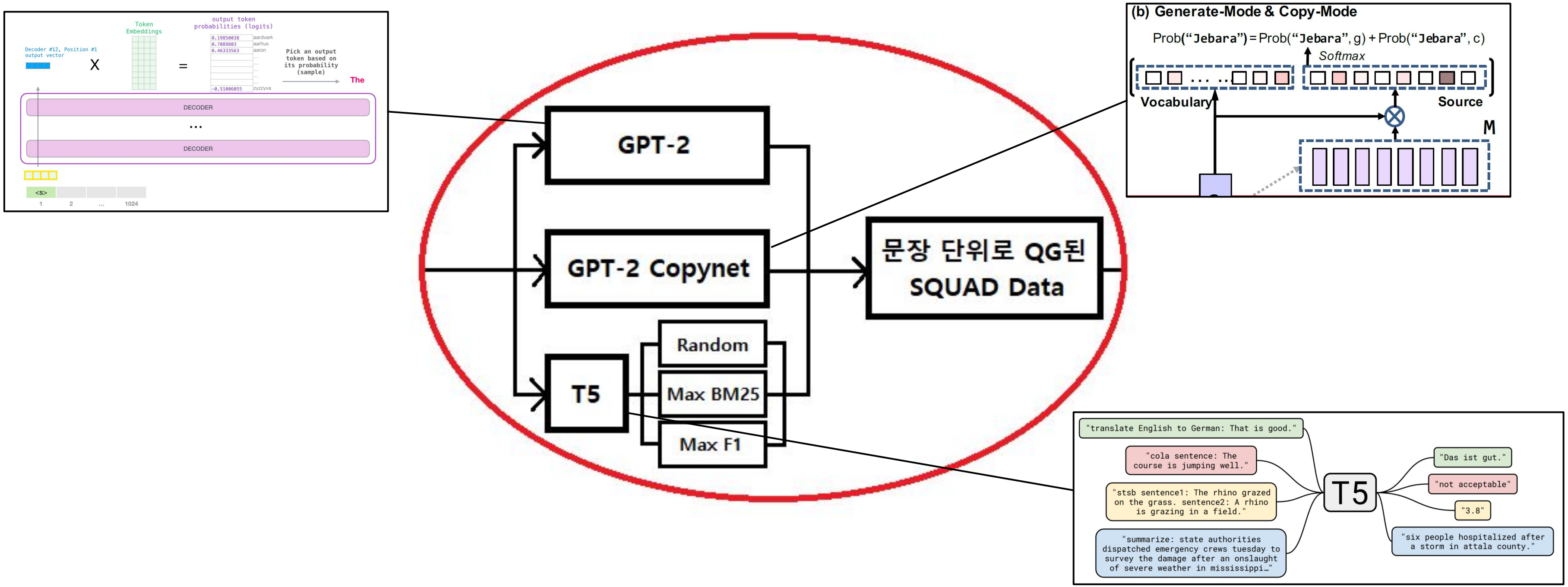
Data Processing for Question Generation Model

- Using ‘Spacy’ library, split SQUAD context data into individual sentences.
- Get sentences that includes the answers.

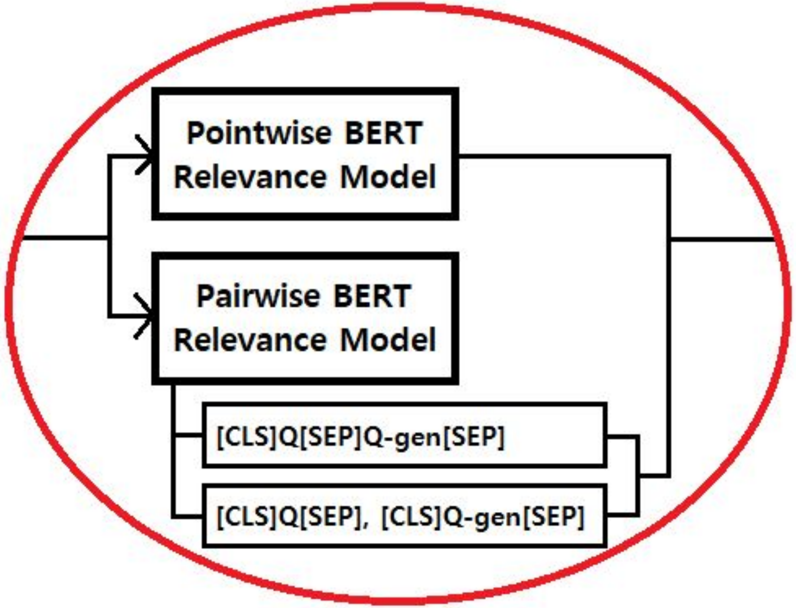


Question Generation Models

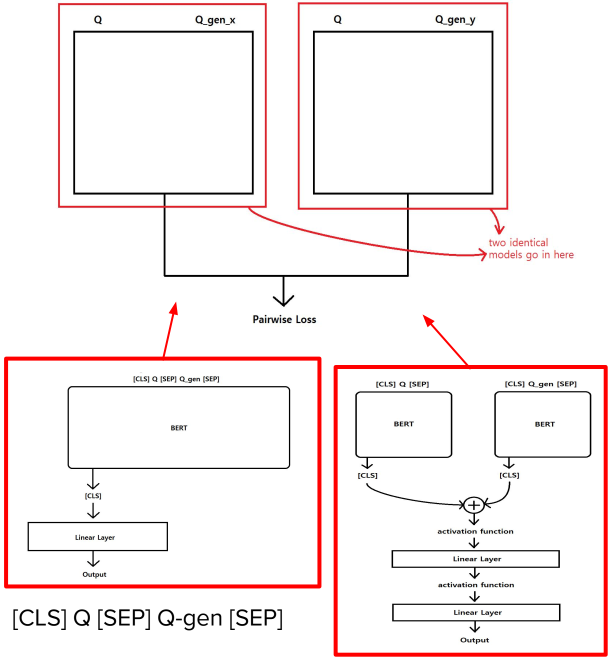
- GPT-2, GPT-2 Copynet, and T5 models for question generation



Relevance Matching Models



Pairwise Method:

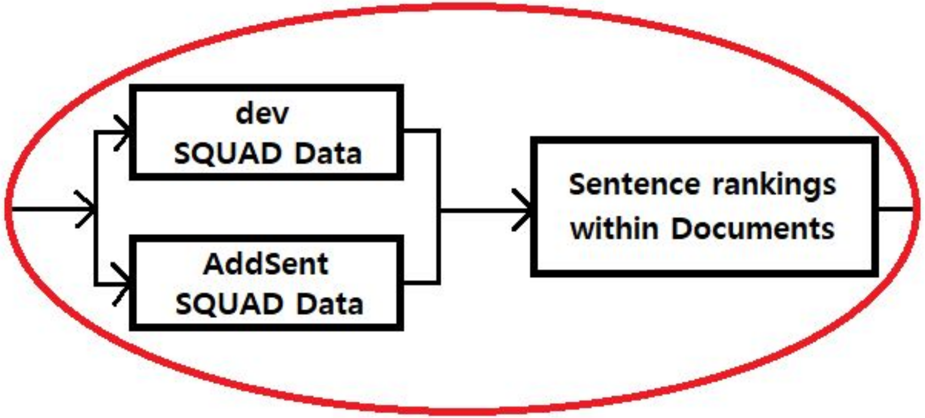


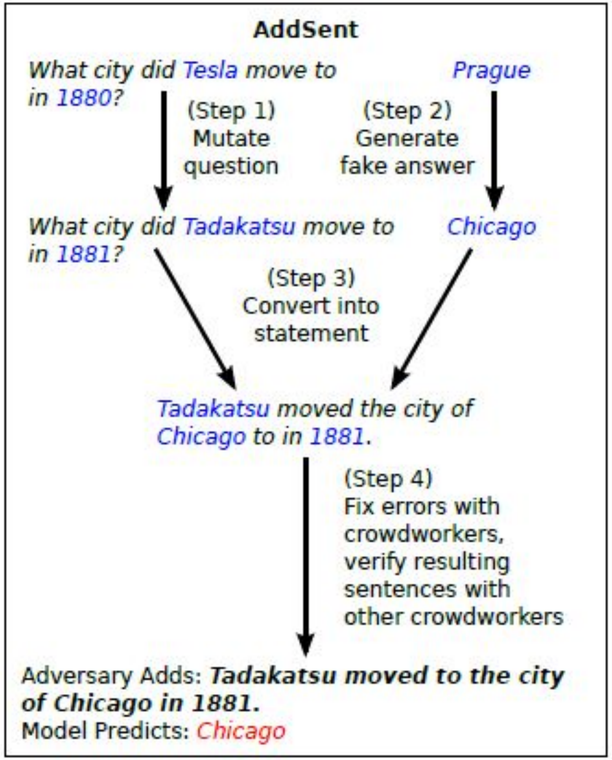
[CLS] Q [SEP] Q-gen [SEP]

[CLS] Q [SEP], [CLS] Q-gen [SEP]

Datasets

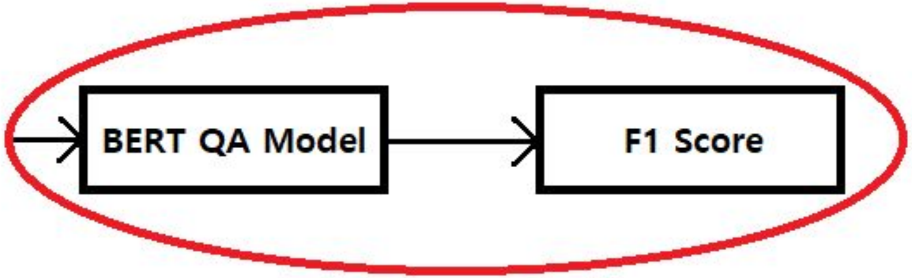
- Rank sentences from the dataset using the relevance matching model

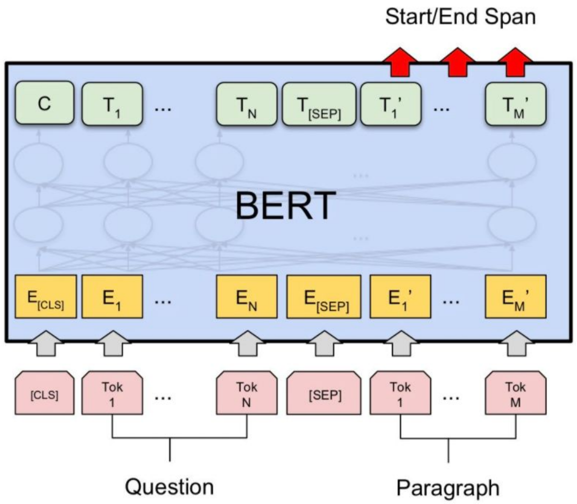




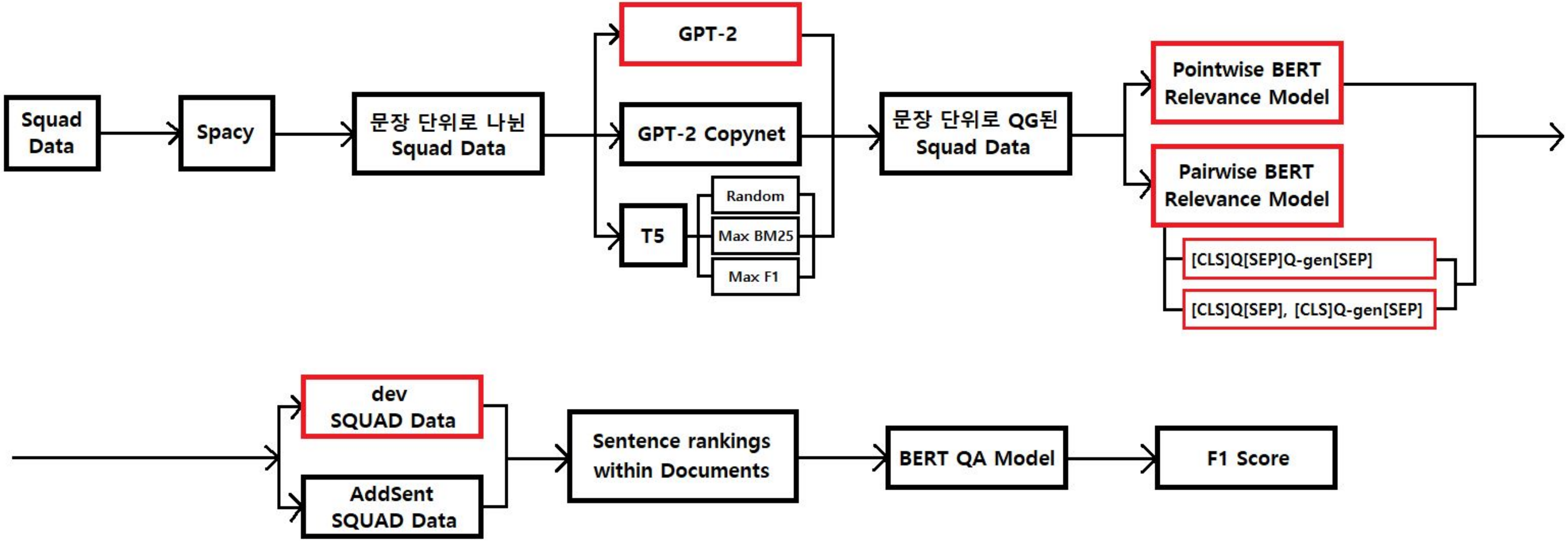
BERT Question Answering Model

- Put test dataset through QA model and extract the F1-score





Results: GPT-2 with All 3 BERT Models



Results: Relevance Matching Models

Rank results using a model trained with normal **pointwise** BERT training.

rank1	0.51
rank2	0.708
rank3	0.834
rank4	0.887

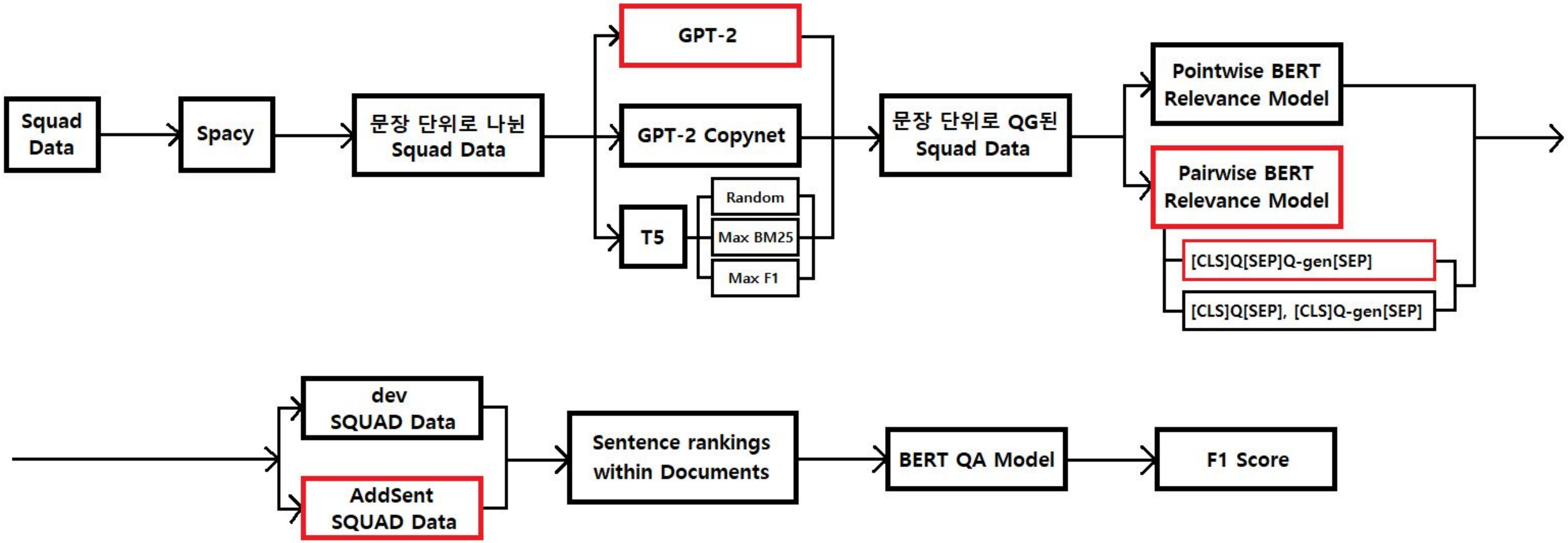
Rank results after training the model in the form **[CLS] Q [SEP] Q_gen [SEP]** with **pairwise** training (**questions generated** with GPT-2).

rank1	0.541
rank2	0.739
rank3	0.858
rank4	0.905

Rank results after training the model in the form **[CLS] Q [SEP] Q_gen [SEP]** with **pairwise** training (**original sentences**).

rank1	0.837
rank2	0.935
rank3	0.965
rank4	0.977

Results: AddSent SQUAD Dataset



Results: AddSent SQUAD Dataset

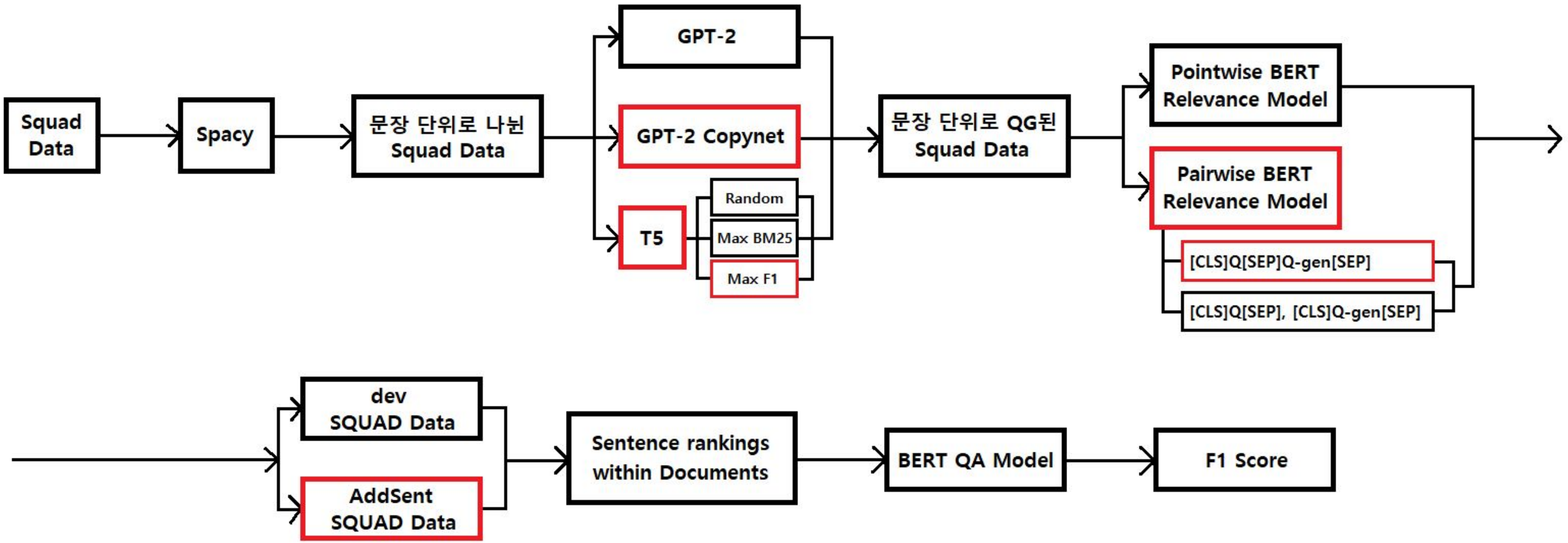
Rank results of **AddSent dataset** using model trained by pairwise BERT on **questions generated** by GPT-2.

rank1	0.356
rank2	0.67
rank3	0.817
rank4	0.904

Rank results of **AddSent dataset** using model trained by pairwise BERT on **original sentences**.

rank1	0.477
rank2	0.879
rank3	0.936
rank4	0.964

Results: GPT-2, GPT-2 CopyNet, T5



Results: GPT-2, GPT-2 CopyNet, T5

Rank results using **GPT-2**

rank1	0.356
rank2	0.67
rank3	0.817
rank4	0.904

Rank results using **Copynet**

rank1	0.369
rank2	0.692
rank3	0.84
rank4	0.918

Rank results using **T5_max F1**

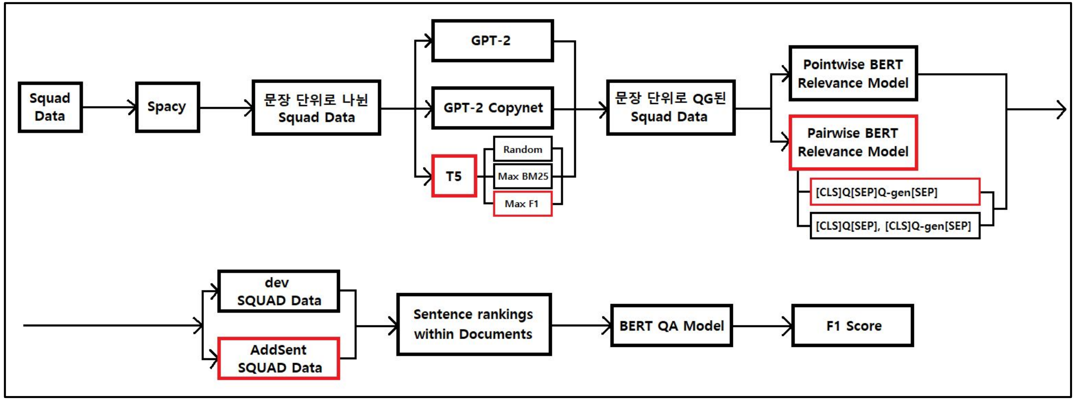
rank1	0.405
rank2	0.737
rank3	0.86
rank4	0.914

Rank results using **T5_max F1** (**hyperparameter tuning**)

rank1	0.437
rank2	0.72
rank3	0.837
rank4	0.905

Results: F1 scores

Model	Original	ADDSENT
ReasoNet-E	81.1	39.4
SEDT-E	80.1	35.0
BiDAF-E	80.0	34.2
Mnemonic-E	79.1	46.2
Ruminating	78.8	37.4
jNet	78.6	37.9
Mnemonic-S	78.5	46.6
ReasoNet-S	78.2	39.4
MPCM-S	77.0	40.3
SEDT-S	76.9	33.9
RaSOR	76.2	39.5
BiDAF-S	75.5	34.3
Match-E	75.4	29.4
Match-S	71.4	27.3
DCR	69.3	37.8
Logistic	50.4	23.2
BertQA	**73.9**	-
OG Sent. Pairwise	-	**41.9**
QG GPT-2 CopyNet	-	**35.4**
QG T5-maxF1	-	**40.4**



Results: Summary

- **Better question generation model** means **better ranking** within documents.

```
original question:  Which instruments can Madonna play?

sentence:  She learned to play drum and guitar from her then-boyfriend Dan Gilroy in the late 1970s before joining the Breakfast Club line
-up as the drummer.

generate question:  Where did Victoria start playing drum and guitar?
```

- **Better question answering model** would also **improve the F1 scores**.
- Our experiment works **better with adversarial datasets** than general datasets.

Experiment Process: An Overview