2020.06.23

CS492H DL4RW Project2 Final Presentation

Team Dingbro (#3)

Boseong Kim (Presenter) | Seungsu Kim | Banseok Hwang

CS492H Deep Learning for Real-world Problems

powered by dingbro

Contents

1. Problem Definition
2. Dataset Analysis
3. Data augmentation
4. Pretrained Model & Hyper-param scheduling
5. Experiment Results
6. Conclusion & Improvements

Problem Definition

An open domain question-answering problem

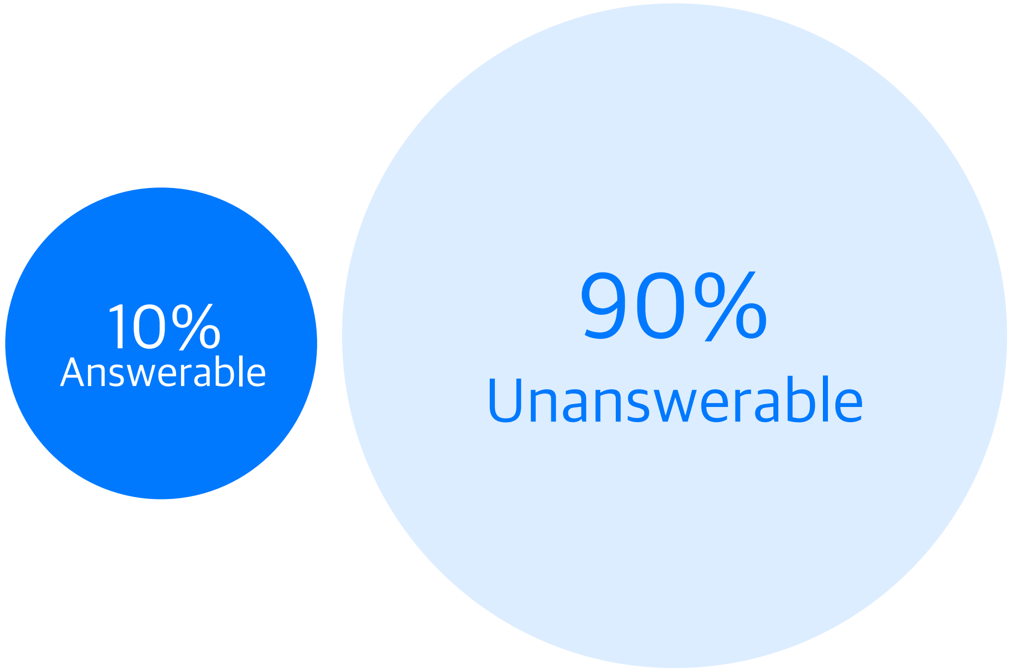
- Not include document searching

Several snippets from several sources are given for each question

- Which is different with SQuAD 1.0 or KorQuAD 1.0
- Model should find the answer from snippets

2. Dataset Analysis

Imbalanced Dataset



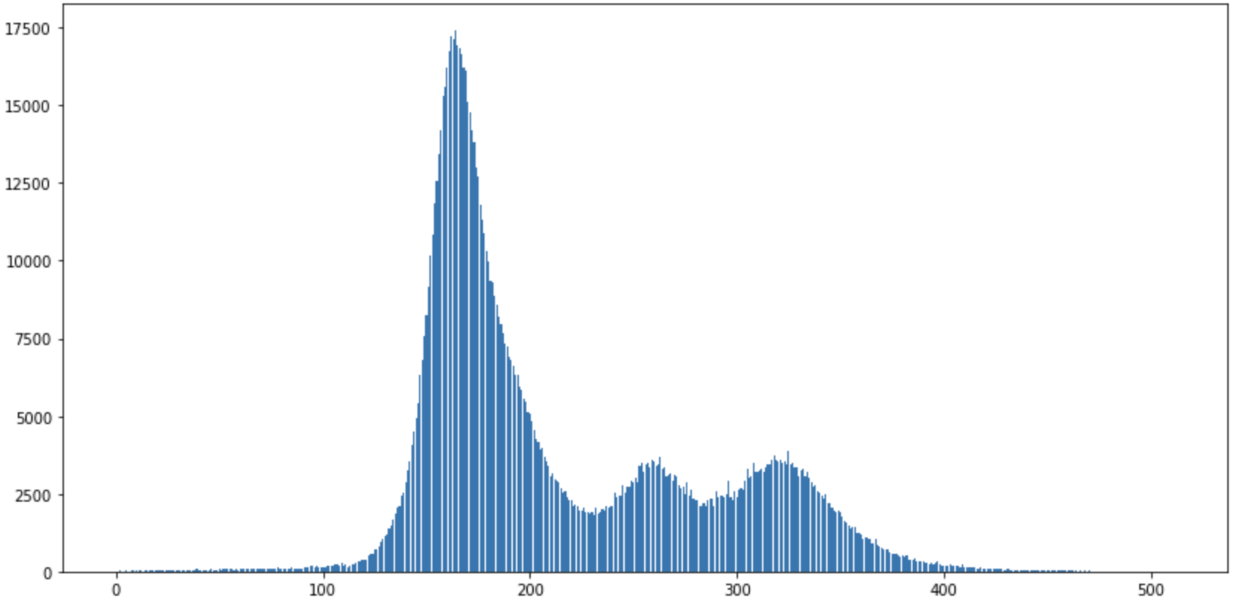
10%
Answerable

90%
Unanswerable

Imbalanced Dataset

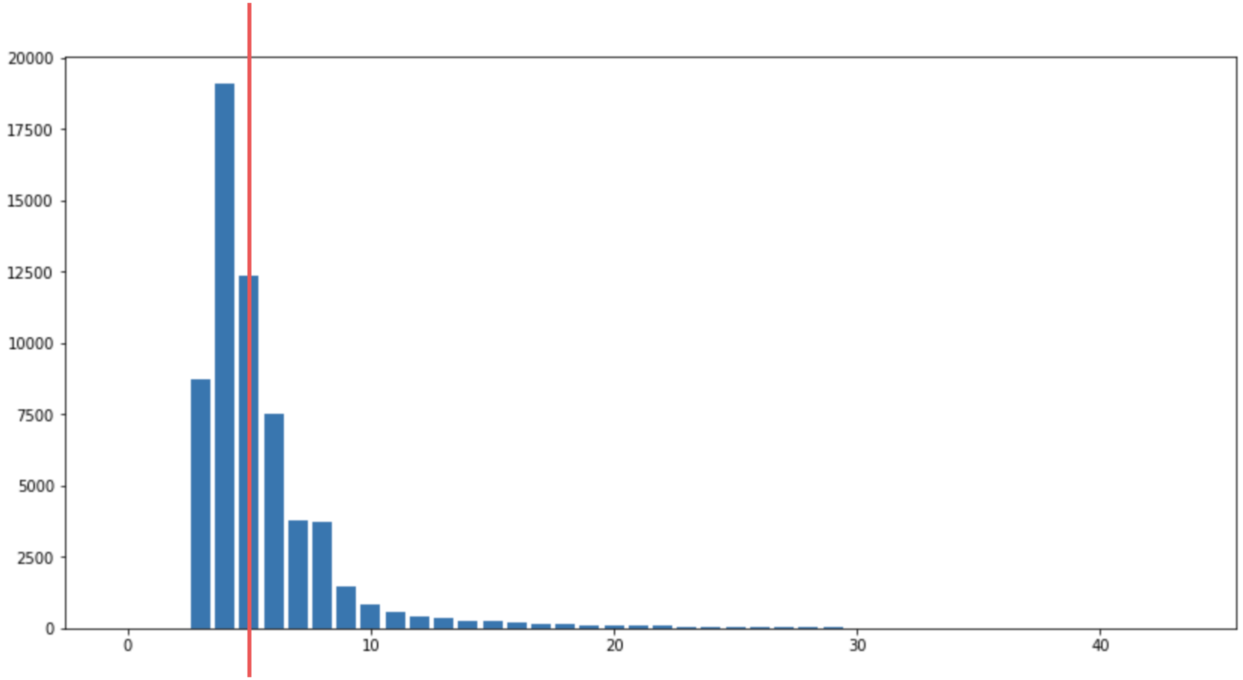
[# of Answerable Q] : [# of Unanswerable Q] = 1 : 9

Dataset Distribution Analysis



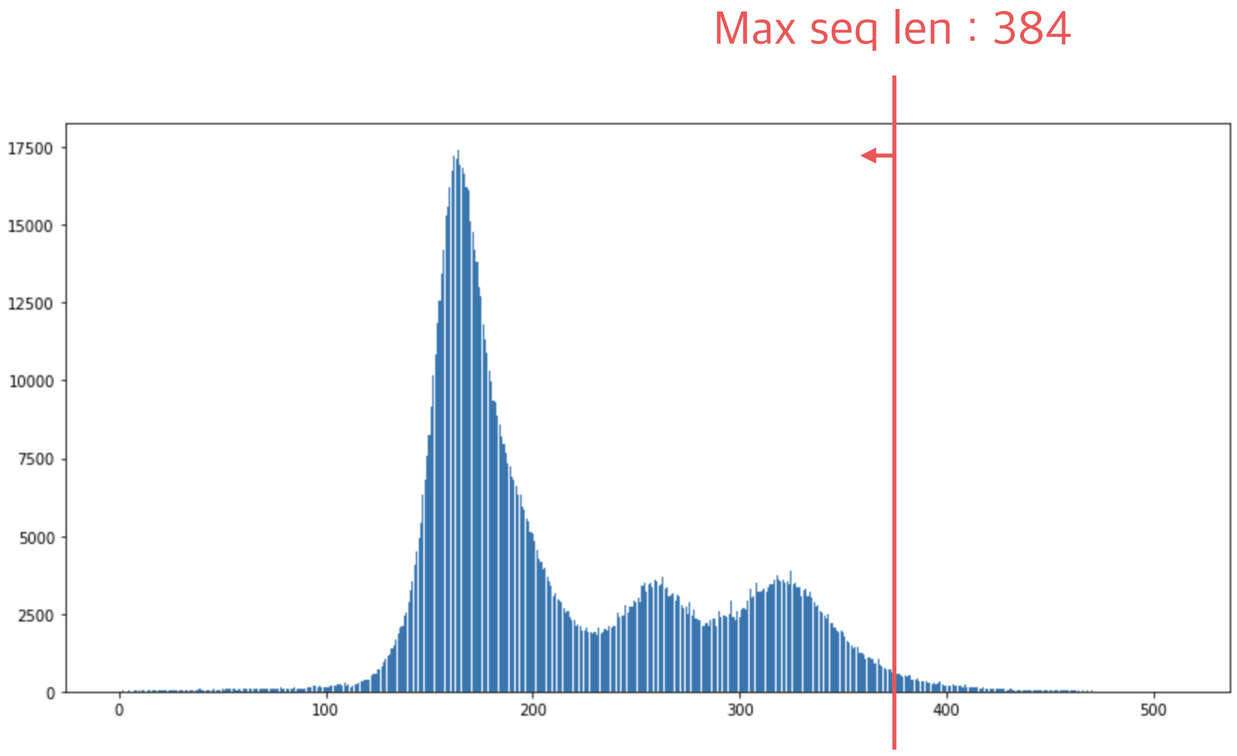
Snippet length distribution

Avg : 5.5 tokens

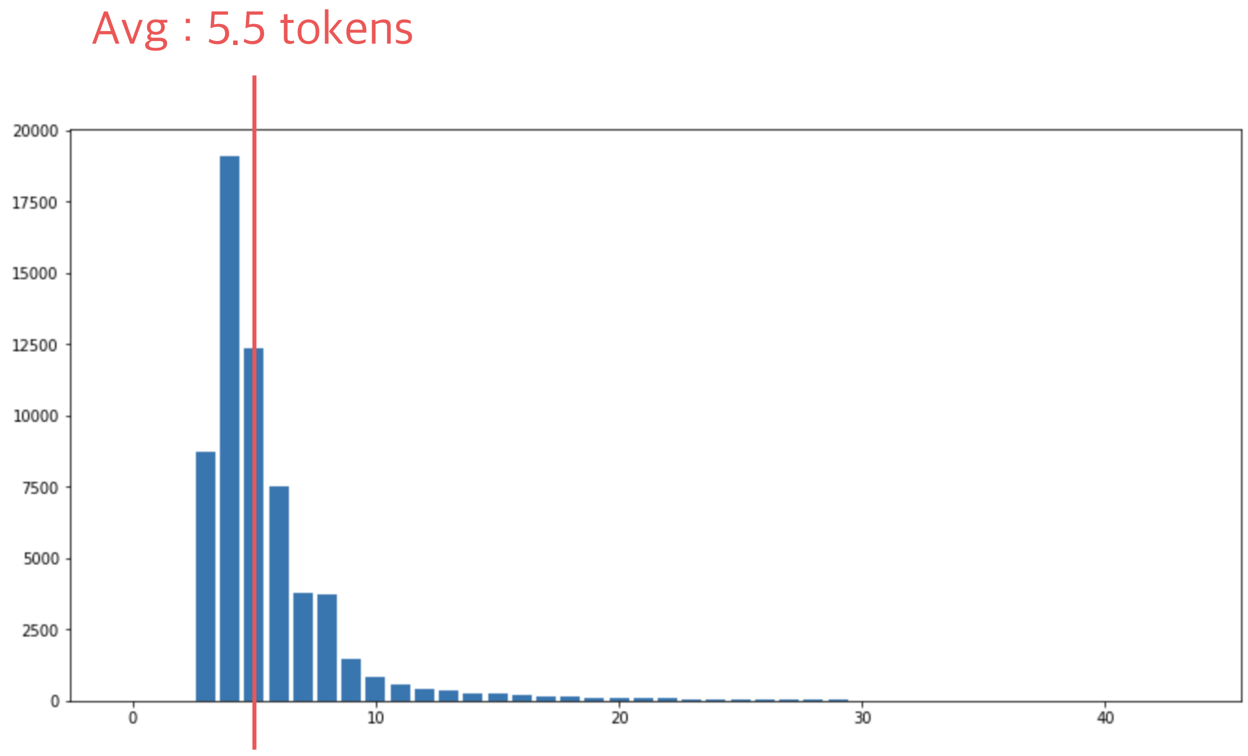


Token length distribution

Dataset Distribution Analysis 1

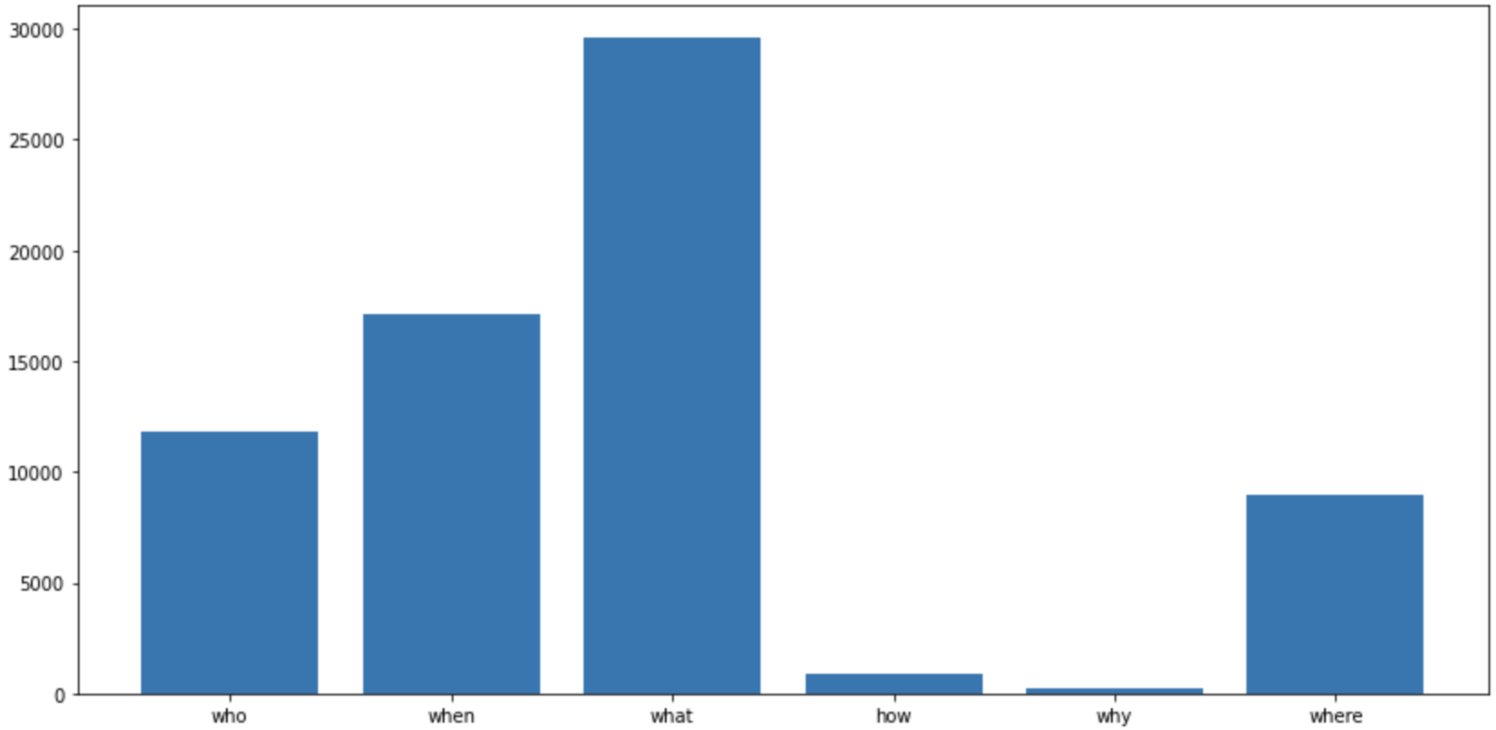


Snippet length distribution



Token length distribution

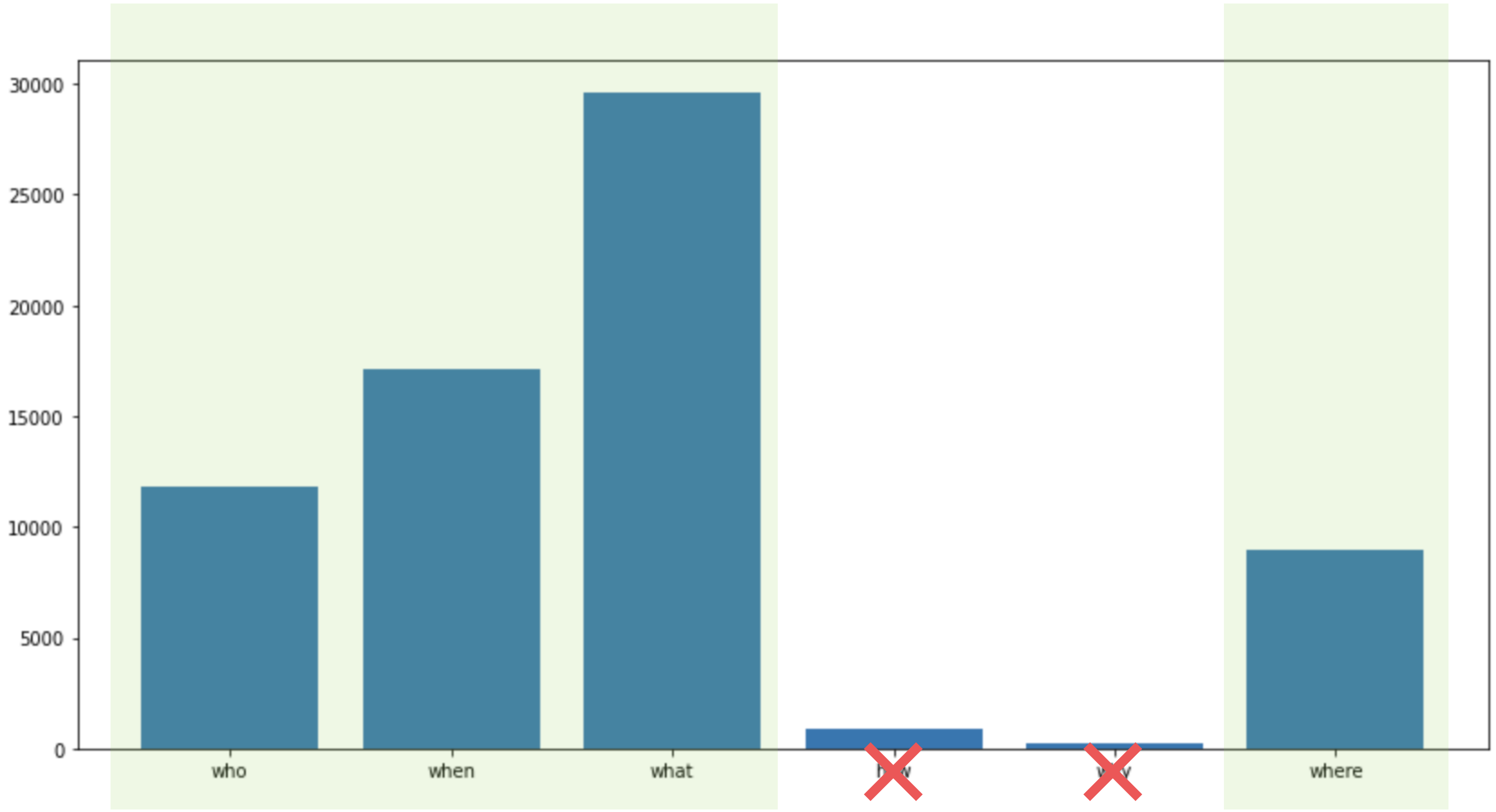
Dataset Distribution Analysis 2



Question Distribution

Majority of the questions were short-answered questions, which excludes How and Why

Dataset Distribution Analysis 2



Question Distribution

Majority of the questions were short-answered questions, which excludes How and Why

3. Data Augmentation

Problems of the Dataset

1. Too small amount of answerable questions

2. Difficult to understand the general context during the training
b/c snippets often not have a complete form of sentence

Problems of the Dataset

1. Too small amount of answerable questions

2. Difficult to understand the general context during the training b/c snippets often not have a complete form of sentence

> *“ 세종은 조선의 제4대 국왕이며 언어학자이다. 그의 업적에 대한 존경의 의미를 담은 명칭인 세종대왕으로 자주 일컬어진다...... 영문예무인성명효대왕이고, 명나라에서 받은 시호는 장헌이었다. 1397년 5월 7일 ..... 1418년에 태종이 신하들과의 회의에서 “세자의 행동이 지극히 무도(無道)하여 종사(宗社)를 이어 받을 수 없다고 대소 신료(大小臣僚)가 청(請)하였기 때문에 이미 폐(廢)하였다.”라고 하며 김한로와 연관되는 등의 심각한 비행으로 인해 왕세자에서 폐위하였다. ”*

Problems of the Dataset

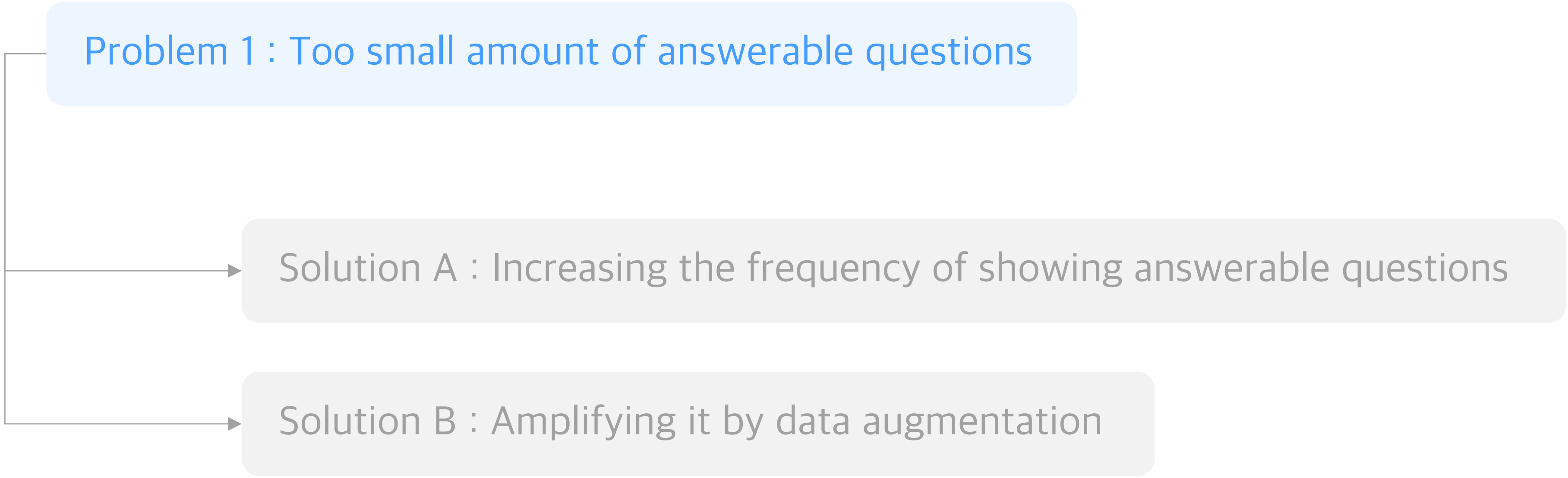
1. Too small amount of answerable questions

2. Difficult to understand the general context during the training b/c snippets often not have a complete form of sentence

> “
>
> *세종은 조선의 제4대 국왕이며 언어학자이다. 그의 업적에 대한 존경의 의미를 담은 명칭인 세종대왕으로 자주 일컬어진다...... 영문예무인성명효대왕이고, 명나라에서 받은 시호는 장헌이었다.* ***1397년 5월 7일*** *..... 1418년에 태종이 신하들과의 회의에서 “세자의 행동이 지극히 무도(無道)하여 종사(宗社)를 이어 받을 수 없다고 대소 신료(大小臣僚)가 청(請)하였기 때문에 이미 폐(廢)하였다.”라고 하며 김한로와 연관되는 등의 심각한 비행으로 인해 왕세자에서 폐위하였다.*
>
> ”

Human can’t answer this question without background knowledge

Augmentation with ELECTRA generator

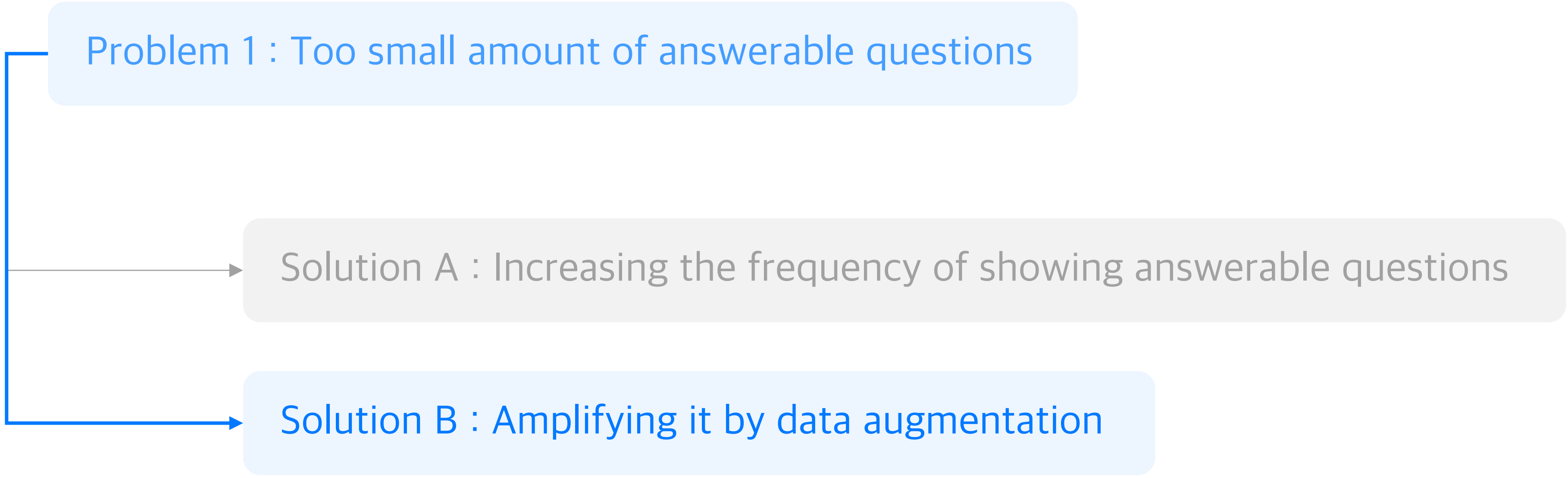


Problem 1 : Too small amount of answerable questions

Solution A : Increasing the frequency of showing answerable questions

Solution B : Amplifying it by data augmentation

Augmentation with ELECTRA generator



Problem 1 : Too small amount of answerable questions

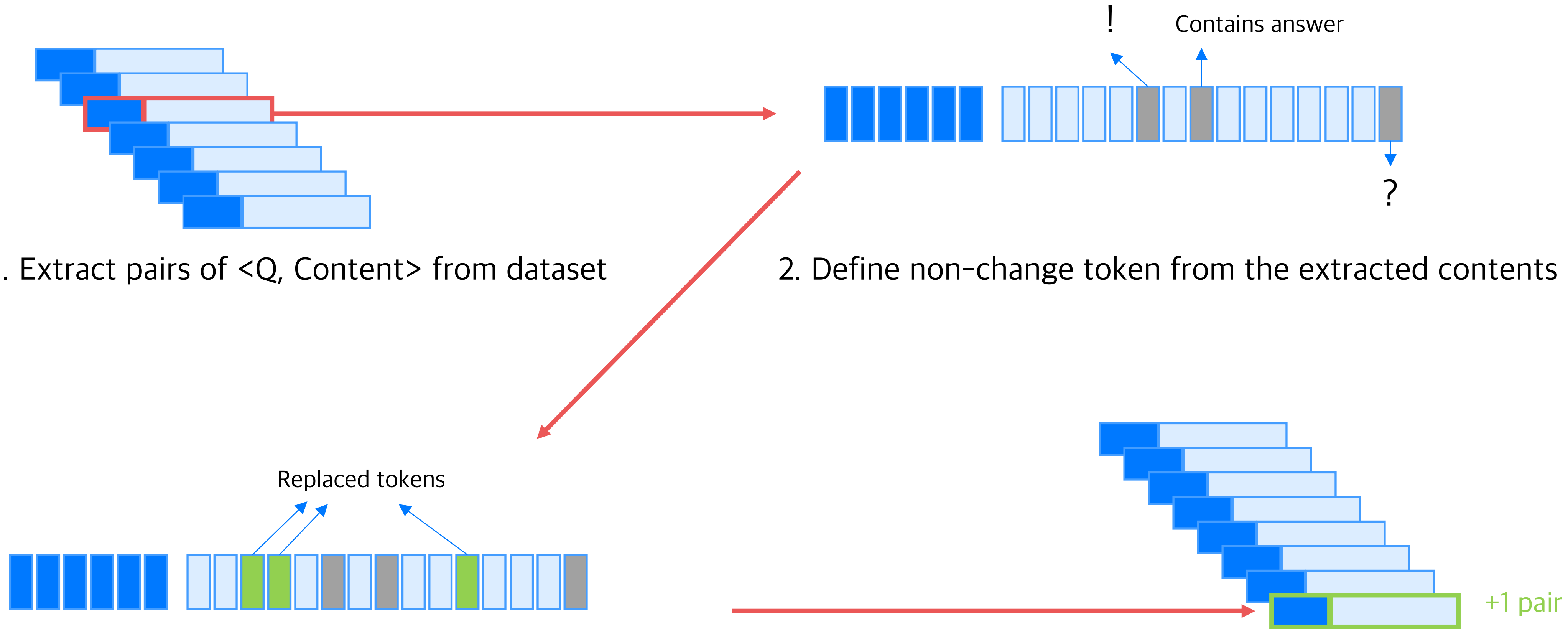
Solution A : Increasing the frequency of showing answerable questions

Solution B : Amplifying it by data augmentation

→ To prevent overfitting

Augmentation with ELECTRA generator

Detail of Augmentation



1. Extract pairs of <Q, Content> from dataset

2. Define non-change token from the extracted contents

3. Replace 25% of tokens as mask tokens except the non-change token and regenerate it as complete sentence with ELECTRA generator.

4. Add to dataset

Using Korquad 1.0 dataset

Problem 2 : Difficult to understand the general context during the training
b/c snippets often not have a complete form of sentence

Solution : Use Korquad 1.0 dataset for additional full-form contexts

→ Provides 60,000 answerable question-content pairs in a full-form.

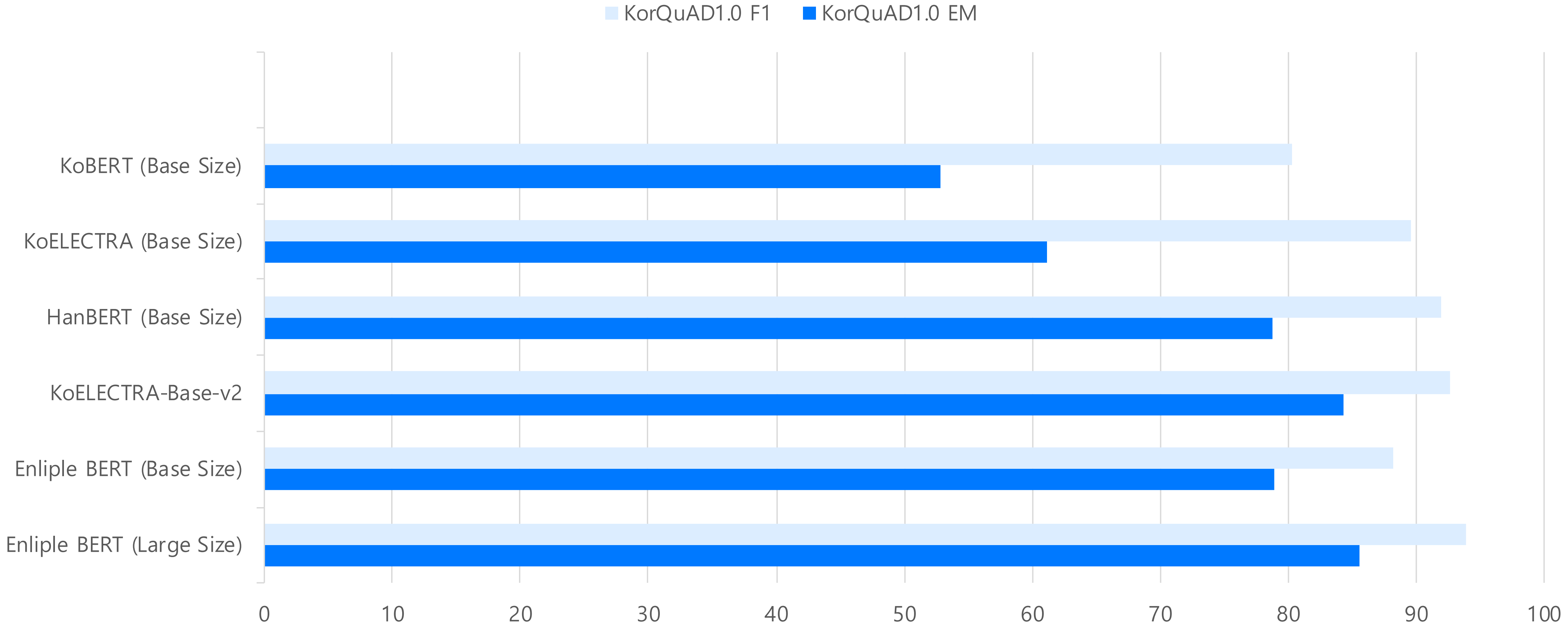
Final Dataset Config

List of Datasets in configuration base of
Korquad 1.0 dataset use/not use, Augmentation Ratio, What Dataset to Augment

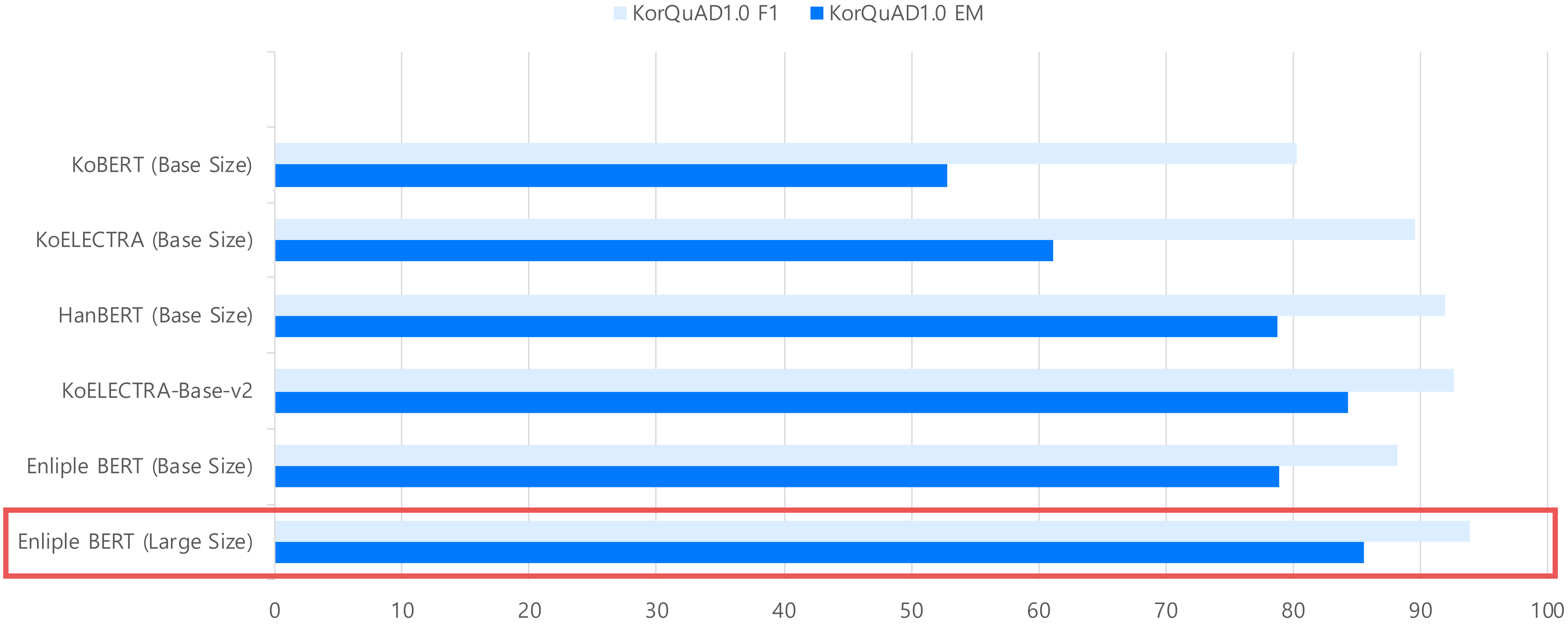
1. Only use NSML korquad open dataset
2. NSML korquad open + Korquad 1.0
3. NSML korquad open + augmented(NSML korquad open)
4. NSML korquad open + Korquad 1.0 + augmented(NSML korquad open + Korquad 1.0)
5. NSML korquad open + Korquad 1.0 + augmented(NSML korquad open)

4. Pretrained Model & Hyper-param scheduling

Pretrained Model



Pretrained Model



Hyper parameter setting

Done experiments with two hyper-parameter fields ; LR and dropout rate.

1. By using **20% of dataset**, we found the **best learning rate and dropout rate values**

2. After that progressed experiments about aforementioned 5 datasets.

5. Experiment Results

Basic Configuration

Models were evaluated with exact score and F1 score

Values were evaluated by the best answer among the N given snippets for a question, not the question-snippet pair

Model : Bert Large

Epoch : 2 epoch

Optimizer : Lamb

Weight Decay : 0

Adam epsilon : 1e-8

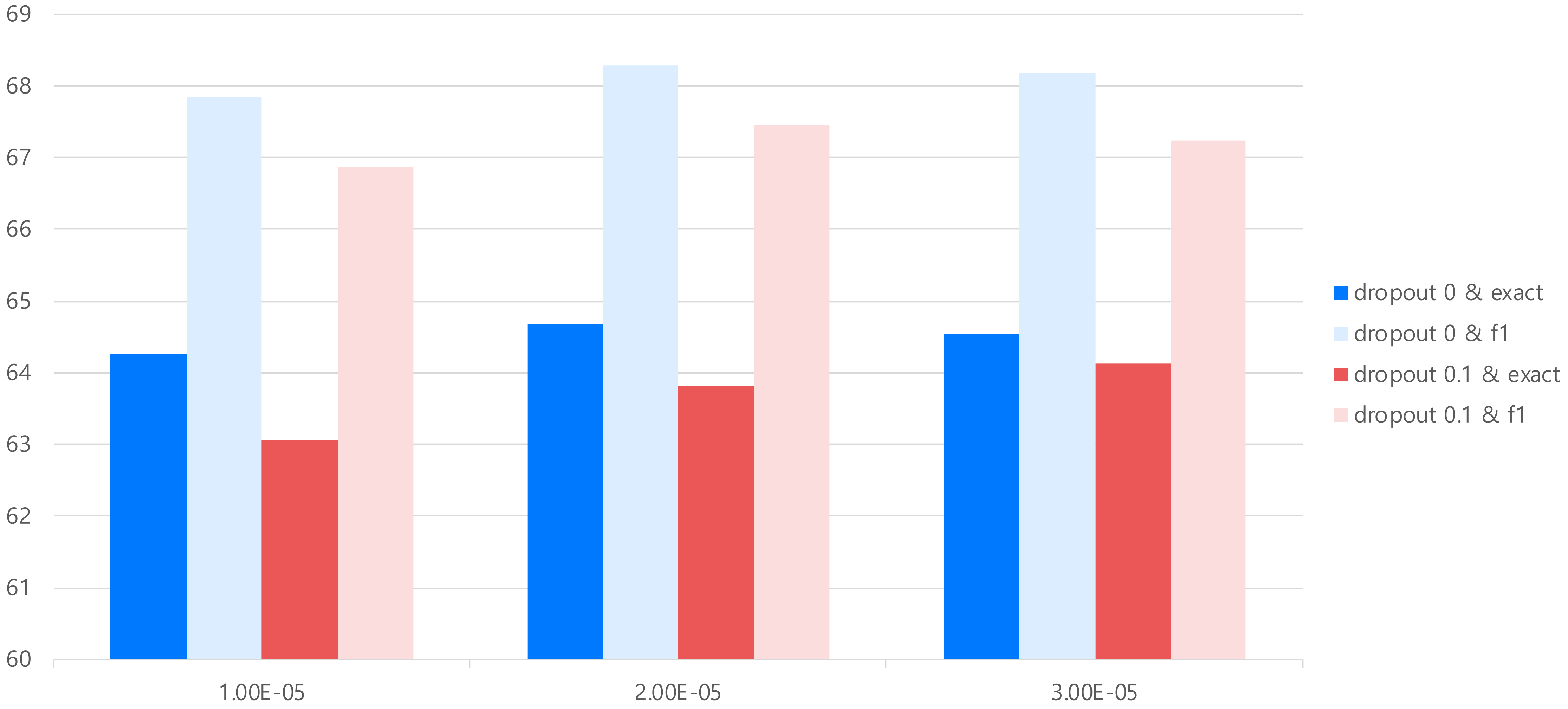
Train batch size : 20

Warm up step : 0.1 epoch

Max seq length : 384

Experiment Results

Finding LR and Dropout rate



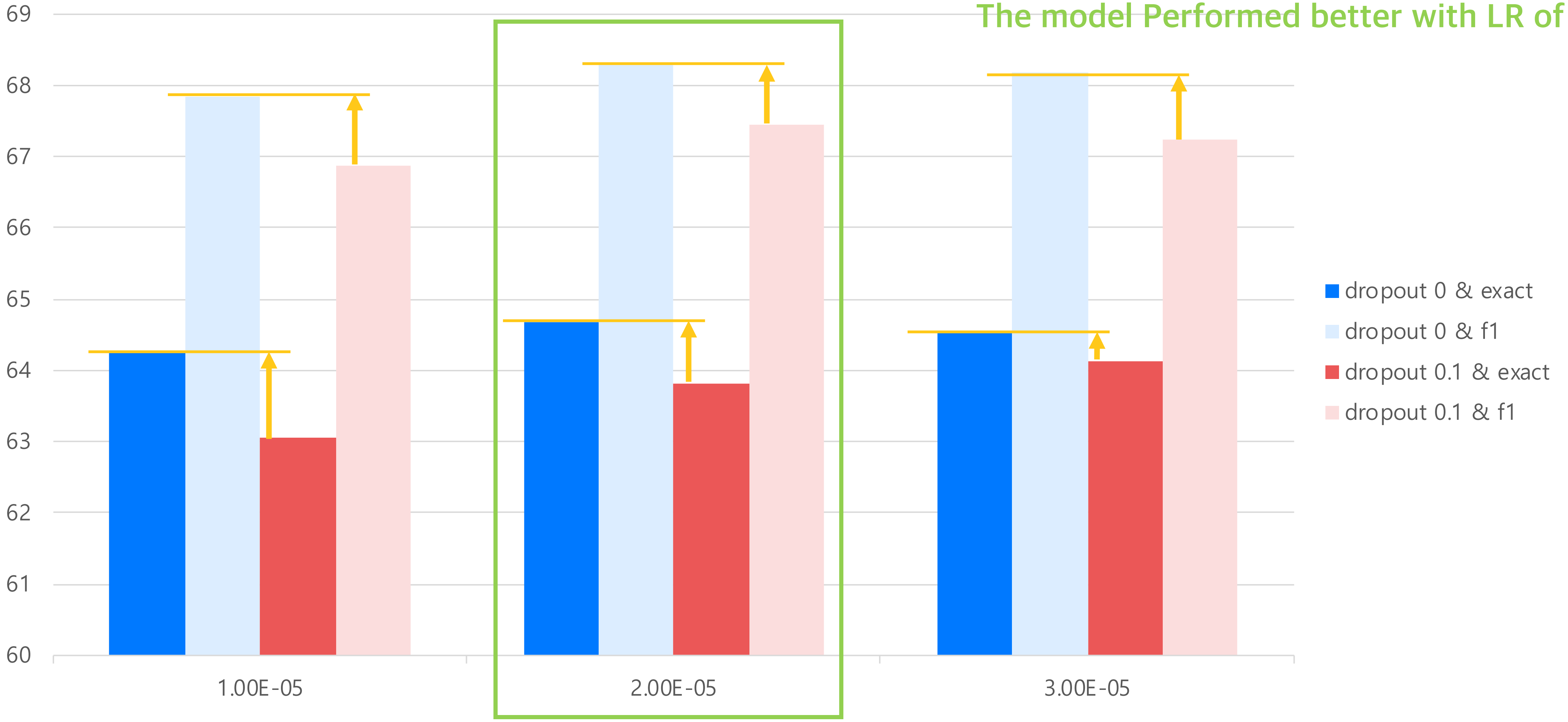
Score for each LR with the condition of [exact or f1] and dropout [0 or 1]

Experiment Results

Finding LR and Dropout rate

The model Performed better with no dropout

The model Performed better with LR of 2e-5

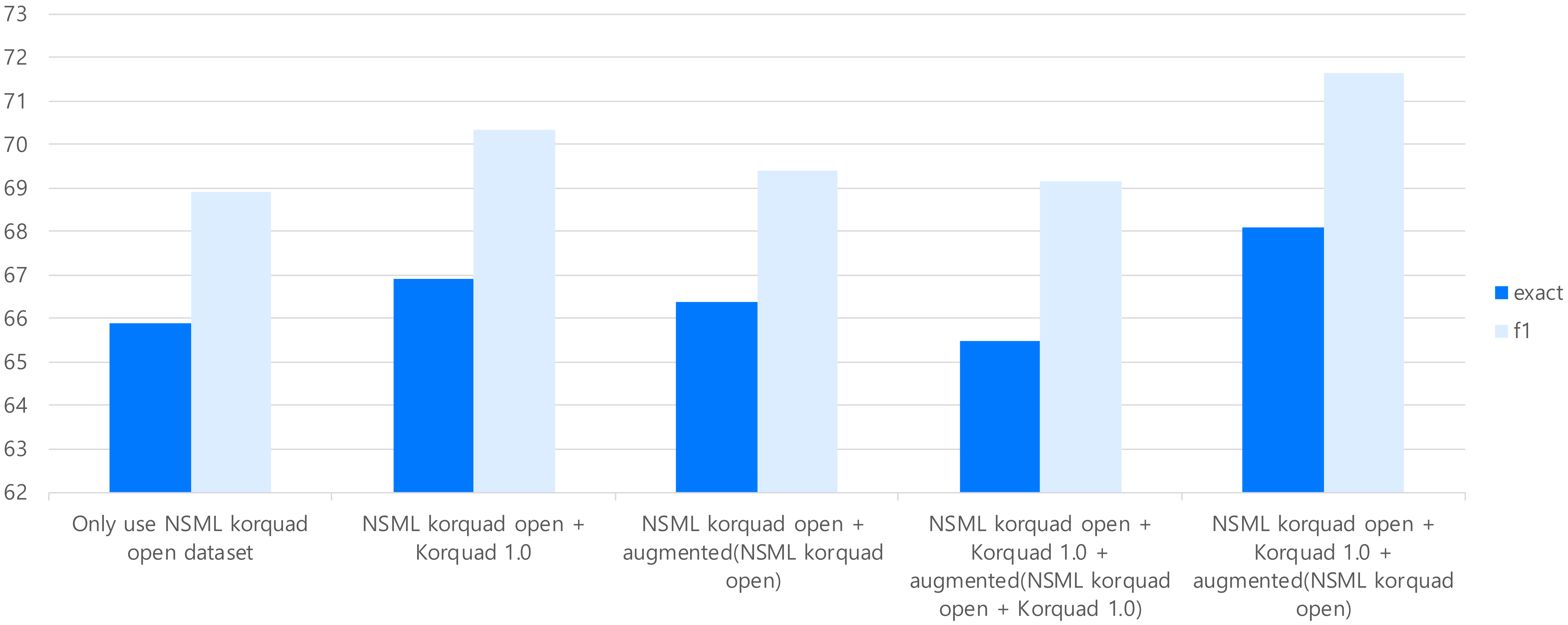


Score for each LR with the condition of [exact or f1] and dropout [0 or 1]

Experiment Results

Dataset type Experiment

With the learning rate(2e-5) and dropout value(0.0), we trained model with 5 different datasets



Score for each type of dataset with difference of exact/f1

6. Conclusion & Improvements

Conclusion

We analyzed that most of the questions in the dataset can be short-answered.

→ we hypothesize data augmentation can generate trainable data and augmented the given dataset with the generator of Electra.

We increased the diversity of the dataset by using KorQuad 1.0 dataset.

→ we achieved exact/F1 score increase of 2.26%/3.35% compared to baseline, and proved the effect of data augmentation.

Further Improvements

1. Koelectra model

2. Find best Answerable/Unanswerable pair ratio

3. NER for post processing

4. Using original context instead of snippet

5. Get more snippet data with crawling

Further Improvements

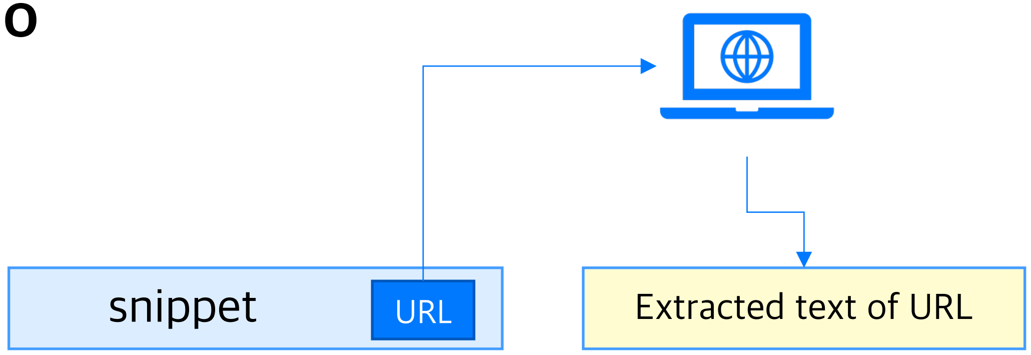
1. Koelectra model

2. Find best Answerable/Unanswerable pair ratio

3. NER for post processing

4. Using original context instead of snippet

5. Get more snippet data with crawling



Model can infer answers with more information and context

Thank You