Multilingual Question Answering Extraction

Nojus Mickus (s174447), Simon Nørby Knudsen (s174479), Harald Skat-Rørdam (s175393), Mia Knudsen (s183998)
Supervisor: Ole Winther
DTU Compute, Technical University of Denmark

Introduction

Question and answering (QA) is an essential part of the information age - think of asking a machine a question, and getting a paragraph of the answer in return. This greatly helps in accumulating accurate information. However, many questions are hard to answer as they relate to less frequently used fields such as biomechanics, quantum mechanics, or - which is the focus of this project - languages other than English. Thus we are interested in researching to what extent the performance of the deep learning NLP models varies between different languages.

This will be done, by fine-tuning the following 4 pre-trained models from Hugging Face model hub[a]:

- English BERT: `bert-base-cased` [1]
- English RoBERTa: `roberta-base` [2]
- Multilingual BERT for English and Korean: `bert-base-multilingual-cased` [1]
- Multilingual RoBERTa for English and Korean: `xlm-roberta-base` [3]

Then comparing performance using Exact Match and F1-Score. English and Korean are two very different languages, they have different appearances, grammar, language families, and lengths. Importantly, they also have different levels of training and data sets available [4][5].

For the analysis we will use the following data set:

- TyDi QA [6]

[a] https://huggingface.co/models

Key points

- We pre-process our data to transform it into SQuAD format.
- We fine-tune the RoBERTa and BERT models to our specific data sets.
- We compare the model performances in English and Korean respectively.

English specific RoBERTa model

RoBERTa: **R**obustly **o**ptimized **BERT a**pproach

The RoBERTa model is an improvement by Facebook to the BERT model. The main modifications are[7]:

1. More training with more data (16Gb vs 160Gb)
2. Removing NSP (Next Sentence Prediction)
3. Training on longer sequences (256 sequences to 8000 sequences)
4. Changing from static masking to dynamic masking - Generating new masking pattern for each new sequence

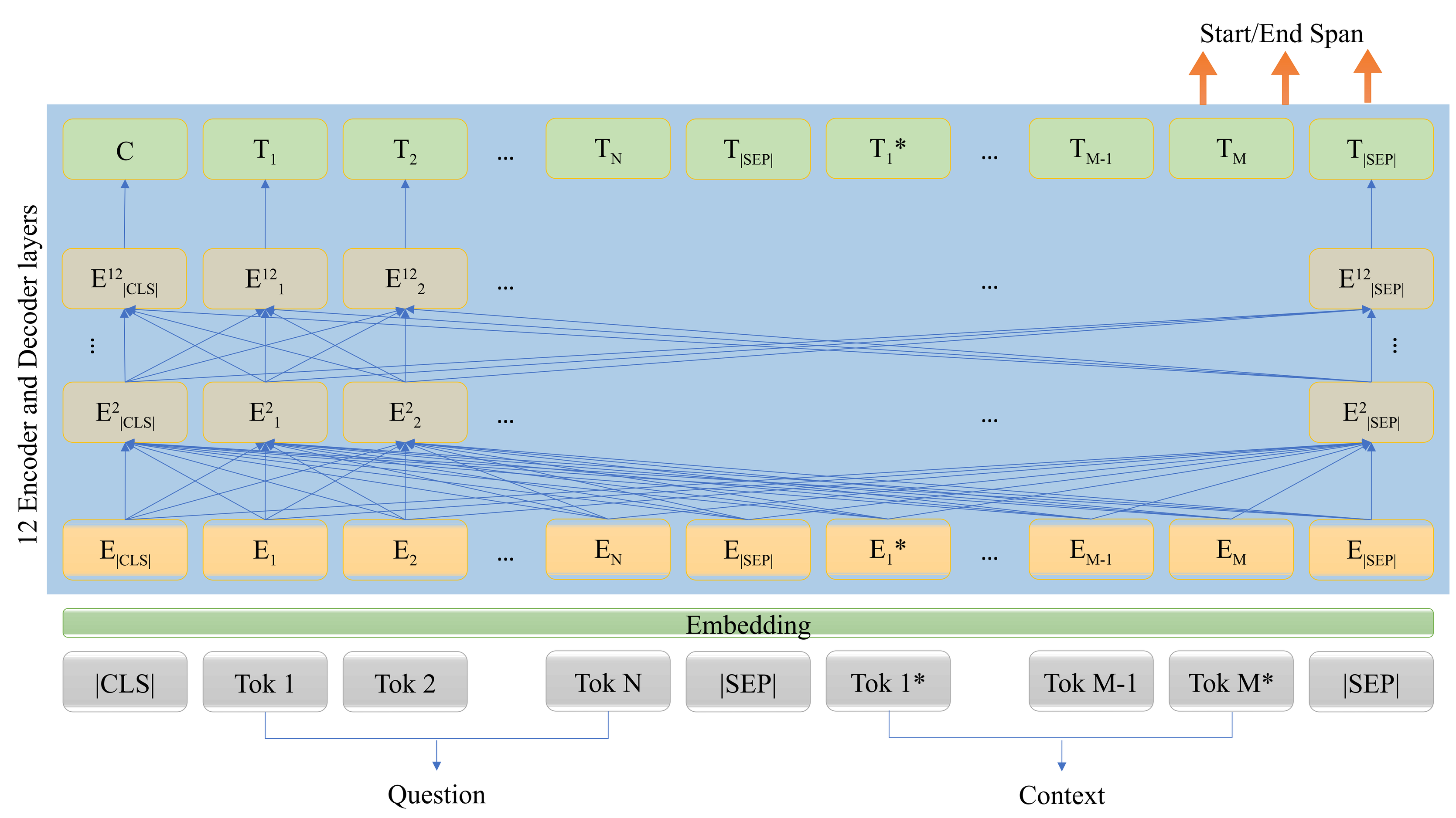

Figure 1: Model architecture for RoBERTa-base.

Data set

For the task of extractive question and answering we will be using the TyDi QA data set. The data set contains the answerable and unanswerable questions, context, and answer pairs for 11 different languages. The focus of this poster will be on English and Korean. In figure 2 an example of a row is shown.

Table 1: Distribution of training and validation data for English and Korean in TyDi QA, where 50% are answerable and 50% unanswerable.

	Training	Validation
English	7389	990
Korean	3249	617

Question: When was Constantinople established?

Context: Constantinople (Greek: Κωνσταντινούπολις, translit.Kōnstantinoúpolis; Latin: Cōnstantīnopolis) was the capital city of the Roman/Byzantine Empire (330–1204 and 1261–1453), and also of the brief Crusader state known as the Latin Empire (1204–1261), until finally falling to the Ottoman Empire (1453–1923). It was reinaugurated in 324 from ancient Byzantium as the new capital of the Roman Empire by Emperor Constantine the Great, after whom it was named, and dedicated on 11 May 330.[5] The city was located in what is now the European side and the core of modern Istanbul.

Answers: 11 May 330

Figure 2: An example of a question, context, and answer.

Procedure

The TyDi QA data set consists of both answerable and unanswerable questions. When an extraction model is used it will always extract an answer to the given question regardless of the answer existing in the associated context. This would result in many incorrectly answered questions, and in a real-life setting, would create the wrong impression for users, thus first determining if a question has an answer is important. In figure 3 the procedure for obtaining the answers for the input question and context pair is displayed.

We need to predict if the question is answered based on the given context, this can be done by fine-tuning one of our models using hugging face's *AutoModelForSequenceClassification* architecture which will produce a binary output indicating answerable or unanswerable. The extraction of the answer for the question based on the context uses hugging face's *AutoModelForQuestionAnswering* architecture. By combining the output from the two models we will be able to create more reliable results.

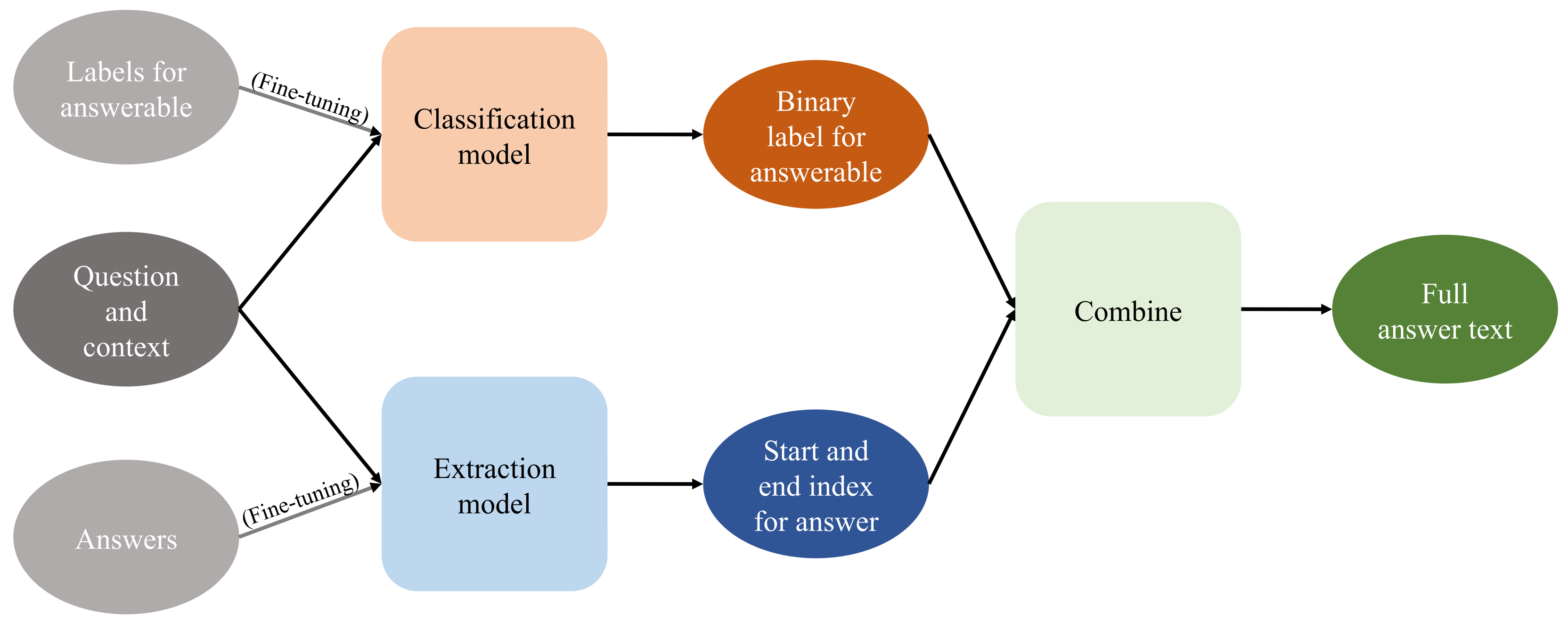

Figure 3: Procedure for creating the final output. The dashed lines symbolize data used for training. Squares symbolize model/code blocks and ellipses symbolize data.

Classification

The validation subsets consist of 50% answerable question and 50% unanswerable questions for both languages.

	English Accuracy	**Korean** Accuracy
Multilingual BERT	84.85	85.09
Multilingual RoBERTa	82.42	*86.39*
BERT	82.83	-
RoBERTa	**86.06**	-

Table 2: Classification accuracy for different models in English and Korean (No BERT and RoBERTa models in Korean, since these models on huggingface are fine-tuned versions of the multilingual models).

It's unexpected to see the best performance for Korean. This is probably due to the limited training resources available, and thus with more epochs, we would see an overall better performance for English.

Model performance

	Before Classification		After Classification	
English	F1	Exact	F1	Exact
Multilingual BERT	33.02	26.26	67.93	62.12
Multilingual RoBERTa	33.60	26.36	70.67	64.14
BERT	33.12	25.96	68.21	62.12
RoBERTa	35.64	27.37	**72.77**	**65.35**

Table 3: Extraction performance (F1-score and Exact matches) for different models in English.

	Before Classification		After Classification	
Korean	F1	Exact	F1	Exact
Multilingual BERT	28.94	23.99	67.62	63.05
Multilingual RoBERTa	29.57	25.93	**69.35**	**66.29**

Table 4: Extraction performance (F1-score and Exact matches) for different models in Korean.

We see a significant improvement when we use the classification model. Additionally, English RoBERTa performs the best which is expected both because of its optimization and training data size.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019.

[4] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116:8450, 2019.

[5] Wikimedia. List of wikipedias, 2022.

[6] Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages, March 2020. arXiv:2003.05002 [cs].

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

s174447@student.dtu.dk, s174479@student.dtu.dk, s175393@student.dtu.dk, s183998@student.dtu.dk