Making Multilingual Summarization More Accessible

Nick Schoelkopf,[1] Ansong Ni MS,[1] and Dragomir Radev PhD[1]

[1]Department of Computer Science, Yale University, New Haven, CT

LILY Lab

Introduction

Text Summarization is an important task in NLP in which given a longer source document, a model must generate a shorter and coherent summary that conveys the important information in the longer document. This task is extremely important and has many practical uses, but research in this field has been decidedly English-focused, with the vast majority of summarization datasets containing only English document-summary pairs. Current state-of-the-art summarization models are only trained on this English data and only tested on English data, causing text summarization in other languages to be pushed to the sidelines. Indeed, in our own SummerTime toolkit for text summarization, prior to this semester we included no models that could handle non-English data, and only one dataset containing non-English languages (MLSum). In order to combat this state of affairs, we add a number of tools that can be used to perform text summarization in many different non-English languages, in order to make these systems and data more accessible to users.

Materials and Resources

For this project, we build on our existing Python toolkit for text summarization, SummerTime: A Text Summarization Toolkit for Non-Experts. We primarily use the PyTorch and Huggingface libraries to build this toolkit.

We use pretrained model checkpoints from Huggingface and pretrained translation systems from the University of Helsinki.

The data in XLSum were made available by the authors via Huggingface, and the authors of Massivesumm made article URLs and extraction methods available. Both datasets are crawled from news site articles and their associated embedded in-page summaries, so the data may be subject to the biases of the news sources, and may also have additional noise due to nonuniformly formatted web pages.

Table 1. Models

Model Name	Translate First?	Languages
mT5	No	101
mBART	No	50
Pipeline	Yes	~70

Table 2. Datasets

Dataset Name	# Examples	Languages
XLSum	1.35M	45
MassiveSumm	28M	92

Figure 3. Example Code

```
mt5_model = st_model.MT5Model()

# load Spanish portion of MLSum dataset
mlsum = datasets.MlsumDataset(["es"])

corpus = itertools.islice(mlsum.train_set, 5)
corpus = [instance.source for instance in train_set]

# mt5 model will automatically detect Spanish as the language and indicate that this is supported!
mt5_model.summarize(corpus)
```

Methods and Results

We include three different model architectures that support multilingual summarization in SummerTime, namely mBART-50, mT5, and a translation pipeline model. We also add two new multilingual datasets to the SummerTime library. The first, XLSum, includes data from 1.35 million article-summary pairs from BBC in 45 different languages, while the other, MassiveSumm, contains upwards of 28 million articles and summaries in 92 different languages and more than 35 different scripts.

The three models implemented cover a wide range of languages as well, with mT5 alone covering over 100 languages (including but not limited to all 45 of the languages in XLSum).

Conclusion and Future Directions

By adding these models and datasets for multilingual summarization to SummerTime, we attempt to push back against the centering of English in text summarization research and provide the tools for more people to easily perform summarization experiments on other languages.

Future directions will include using these created tools to perform experiments comparing performance across languages in low and high resource settings, as well as comparing performance disparities between end-to-end models such as mT5 and mBART and a pipeline model that translates into English, performs summarization, and translates back to the foreign language. Using these tools to perform such tests will grant a much better sense of the state of multilingual text summarization and provide directions for further research.

Acknowledgement

I would like to thank Professor Dragomir Radev for his assistance and guidance, Ansong Ni for his advice and for leading this project, and the other SummerTime contributors for their input and assistance.