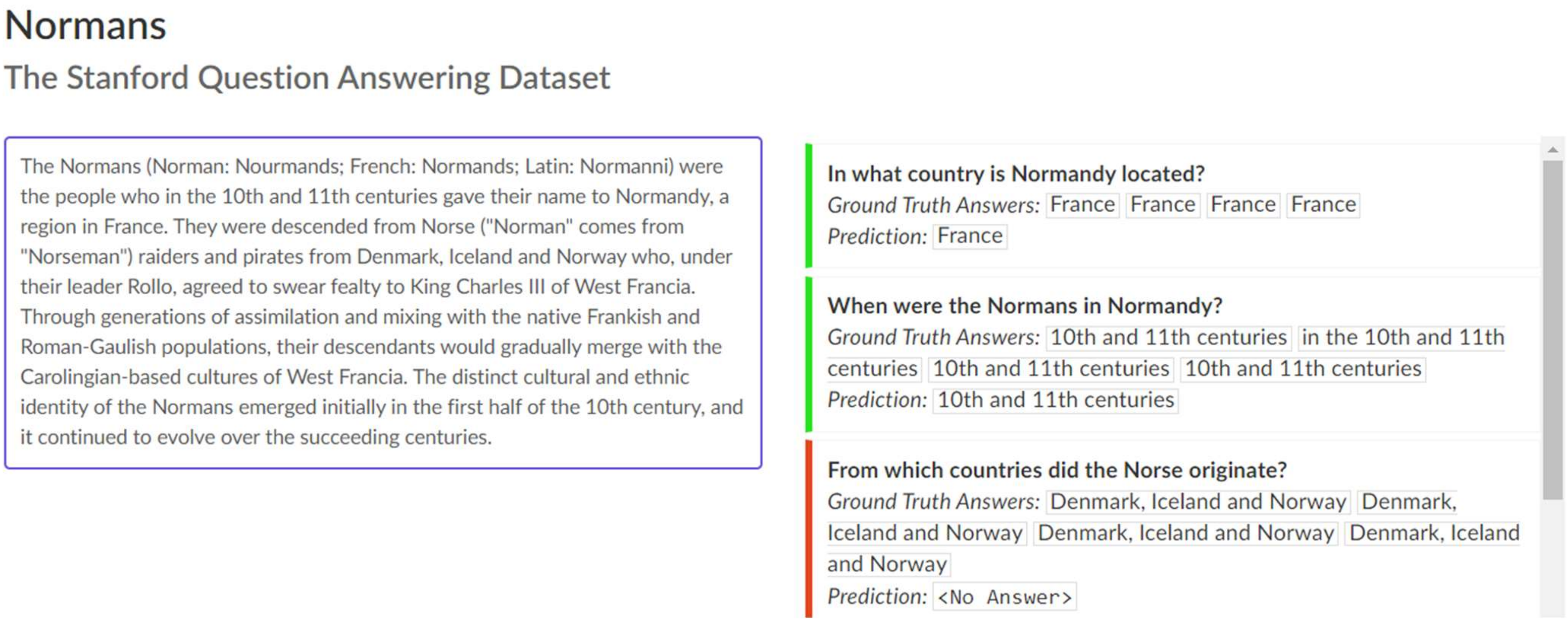
Week 10 – QA and MT

EGCO467 Natural Language and Speech Processing

Project submit

- Give presentation on Nov. 30
  - demo the model
- Github link
- PPT file

Squad 2.0



Normans

The Stanford Question Answering Dataset

The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia. Through generations of assimilation and mixing with the native Frankish and Roman-Gaulish populations, their descendants would gradually merge with the Carolingian-based cultures of West Francia. The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries.

In what country is Normandy located?

Ground Truth Answers: France France France France

Prediction: France

When were the Normans in Normandy?

Ground Truth Answers: 10th and 11th centuries in the 10th and 11th centuries 10th and 11th centuries 10th and 11th centuries

Prediction: 10th and 11th centuries

From which countries did the Norse originate?

Ground Truth Answers: Denmark, Iceland and Norway Denmark, Iceland and Norway Denmark, Iceland and Norway Denmark, Iceland and Norway

Prediction: <No Answer>

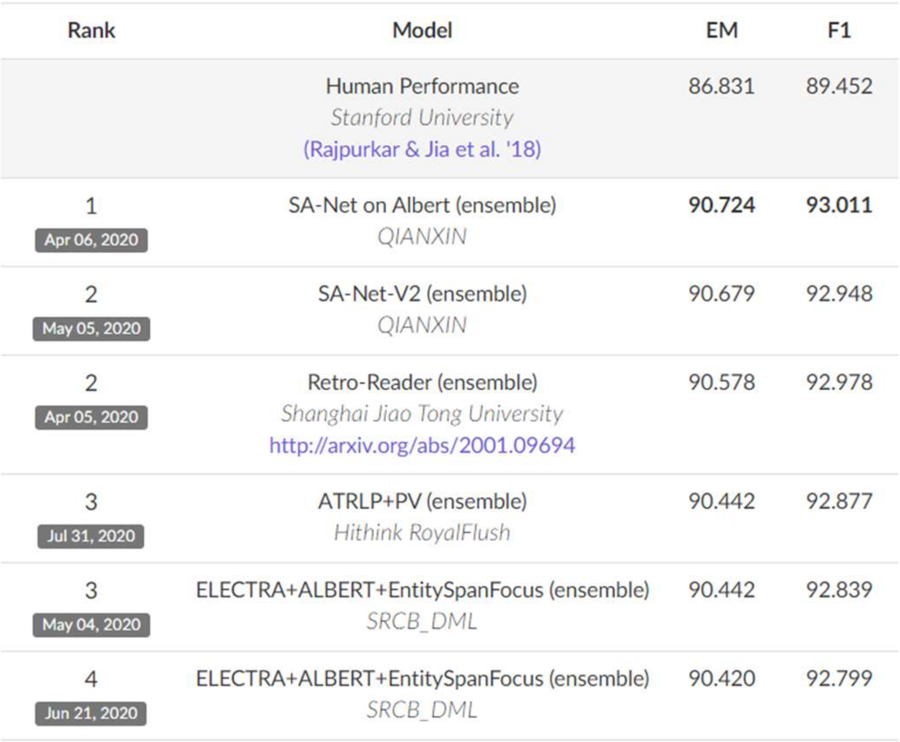
Squad 2.0

- SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones.

Evaluation Criterion

- Exact Match is a binary measure (i.e. true/false) of whether the system output matches the ground truth answer exactly.
- if answer is "**Einstein**" but the ground truth answer was "**Albert Einstein**", => EM score is 0
- F1: the system would have 100% precision (its answer is a subset of the ground truth answer) and 50% recall (it only included one out of the two words in the ground truth output)
- When a question has no answer, both the F1 and EM score are 1 if the model predicts no-answer, and 0 otherwise.

Leaderboard

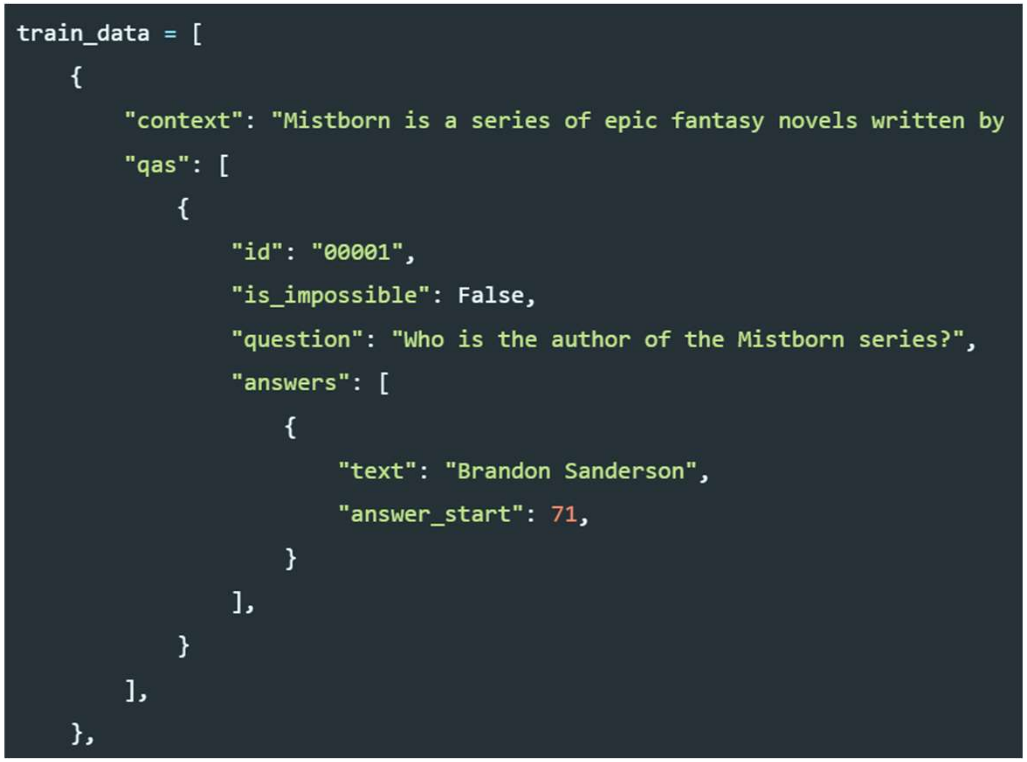


Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	**90.724**	**93.011**
2 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
2 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
3 Jul 31, 2020	ATRLP+PV (ensemble) *Hithink RoyalFlush*	90.442	92.877
3 May 04, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.442	92.839
4 Jun 21, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.420	92.799

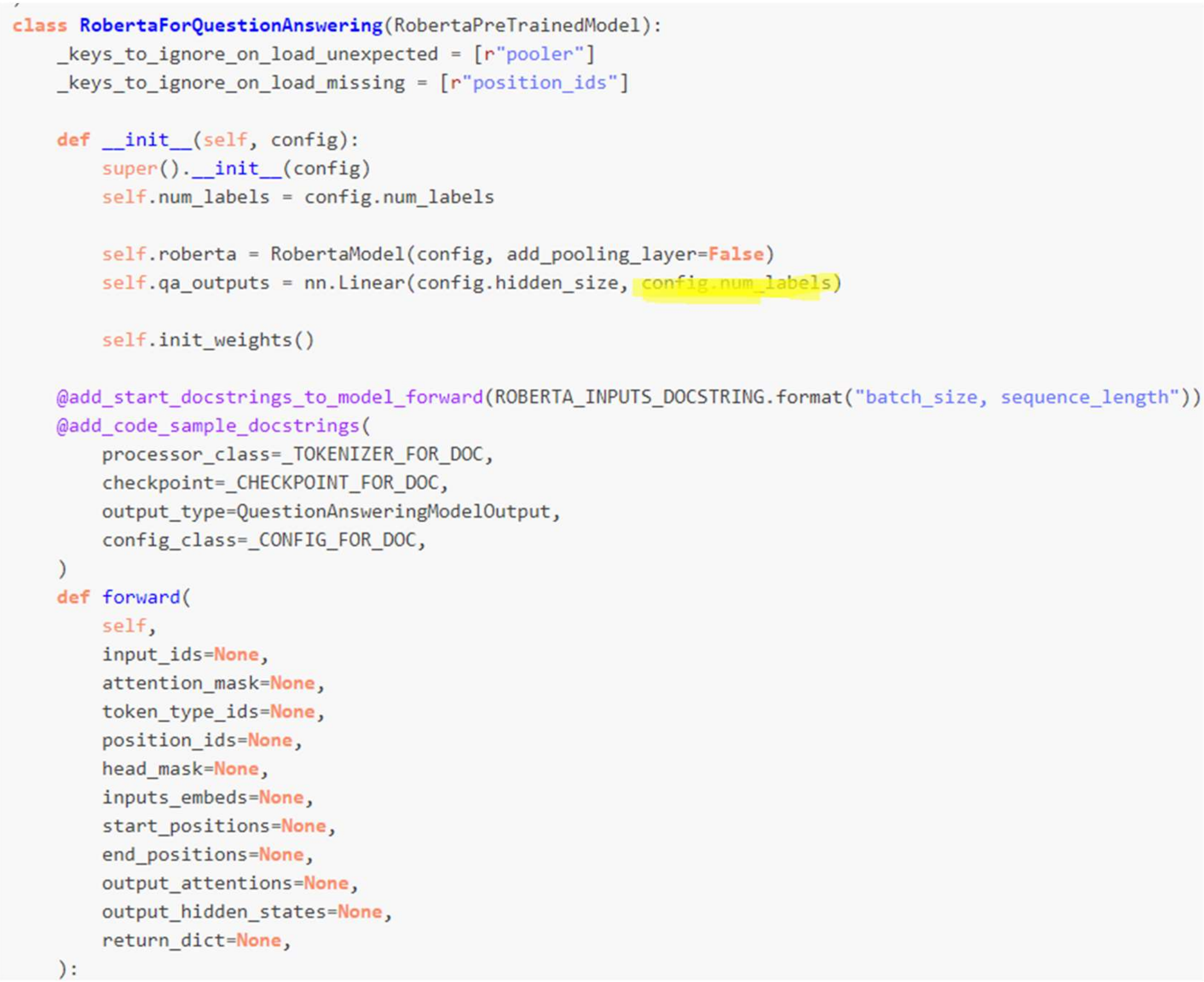
Demo

- https://huggingface.co/deepset/roberta-base-squad2

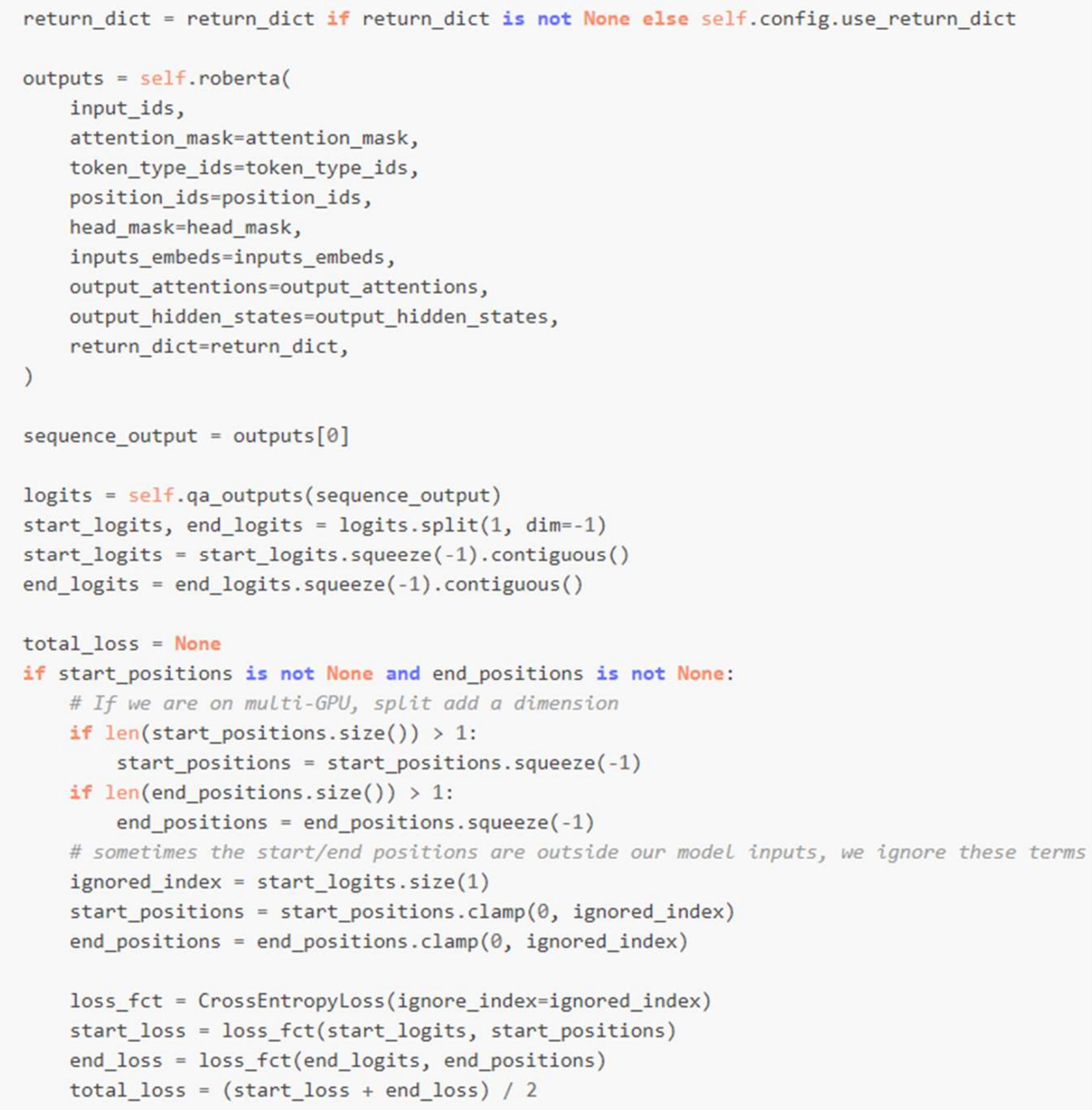
Squad Data



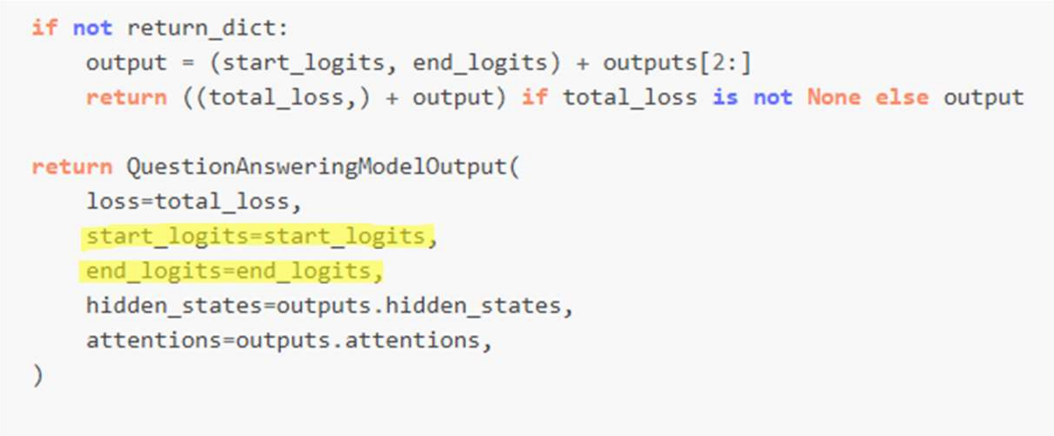
```
train_data = [
    {
        "context": "Mistborn is a series of epic fantasy novels written by
        "qas": [
            {
                "id": "00001",
                "is_impossible": False,
                "question": "Who is the author of the Mistborn series?",
                "answers": [
                    {
                        "text": "Brandon Sanderson",
                        "answer_start": 71,
                    }
                ],
            }
        ],
    },
```



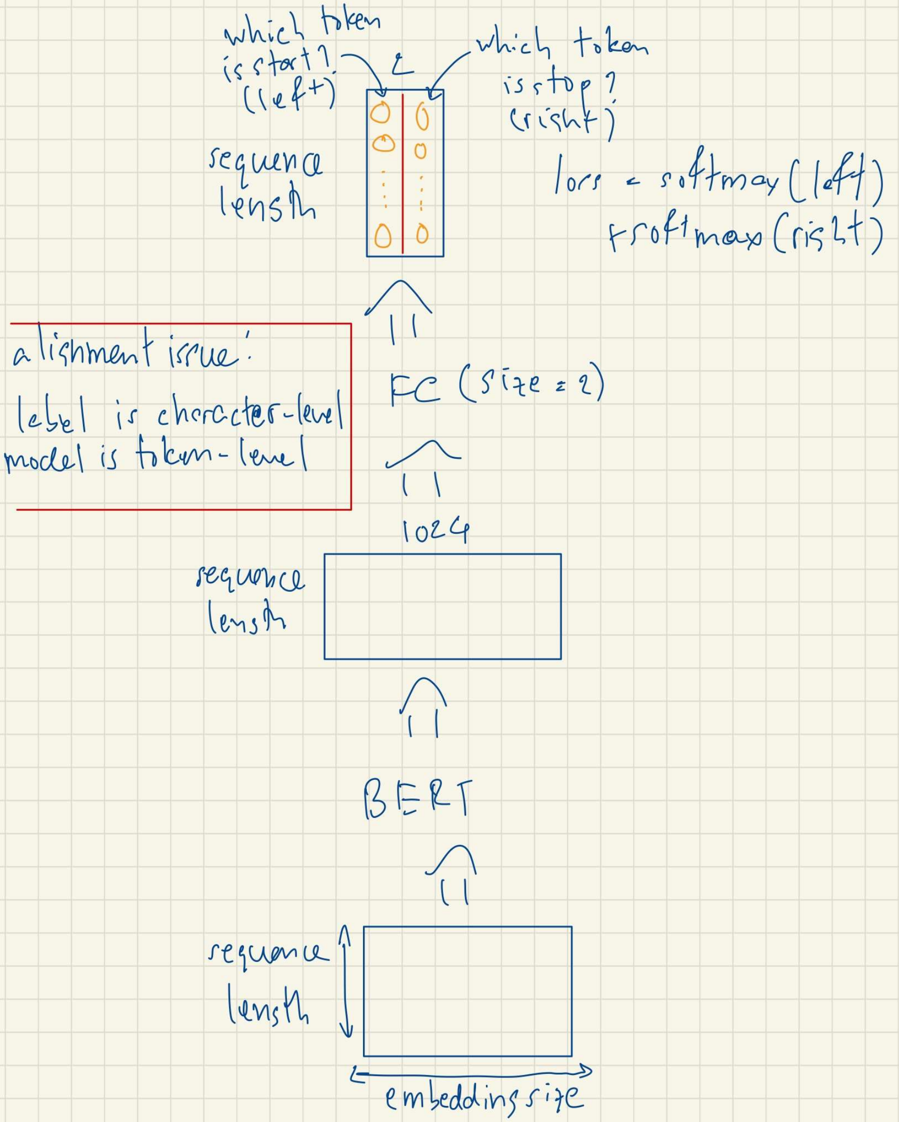
```
,
class RobertaForQuestionAnswering(RobertaPreTrainedModel):
    _keys_to_ignore_on_load_unexpected = [r"pooler"]
    _keys_to_ignore_on_load_missing = [r"position_ids"]

    def __init__(self, config):
        super().__init__(config)
        self.num_labels = config.num_labels

        self.roberta = RobertaModel(config, add_pooling_layer=False)
        self.qa_outputs = nn.Linear(config.hidden_size, config.num_labels)

        self.init_weights()

    @add_start_docstrings_to_model_forward(ROBERTA_INPUTS_DOCSTRING.format("batch_size, sequence_length"))
    @add_code_sample_docstrings(
        processor_class=_TOKENIZER_FOR_DOC,
        checkpoint=_CHECKPOINT_FOR_DOC,
        output_type=QuestionAnsweringModelOutput,
        config_class=_CONFIG_FOR_DOC,
    )
    def forward(
        self,
        input_ids=None,
        attention_mask=None,
        token_type_ids=None,
        position_ids=None,
        head_mask=None,
        inputs_embeds=None,
        start_positions=None,
        end_positions=None,
        output_attentions=None,
        output_hidden_states=None,
        return_dict=None,
    ):
```



```
return_dict = return_dict if return_dict is not None else self.config.use_return_dict

outputs = self.roberta(
    input_ids,
    attention_mask=attention_mask,
    token_type_ids=token_type_ids,
    position_ids=position_ids,
    head_mask=head_mask,
    inputs_embeds=inputs_embeds,
    output_attentions=output_attentions,
    output_hidden_states=output_hidden_states,
    return_dict=return_dict,
)

sequence_output = outputs[0]

logits = self.qa_outputs(sequence_output)
start_logits, end_logits = logits.split(1, dim=-1)
start_logits = start_logits.squeeze(-1).contiguous()
end_logits = end_logits.squeeze(-1).contiguous()

total_loss = None
if start_positions is not None and end_positions is not None:
    # If we are on multi-GPU, split add a dimension
    if len(start_positions.size()) > 1:
        start_positions = start_positions.squeeze(-1)
    if len(end_positions.size()) > 1:
        end_positions = end_positions.squeeze(-1)
    # sometimes the start/end positions are outside our model inputs, we ignore these terms
    ignored_index = start_logits.size(1)
    start_positions = start_positions.clamp(0, ignored_index)
    end_positions = end_positions.clamp(0, ignored_index)

    loss_fct = CrossEntropyLoss(ignore_index=ignored_index)
    start_loss = loss_fct(start_logits, start_positions)
    end_loss = loss_fct(end_logits, end_positions)
    total_loss = (start_loss + end_loss) / 2
```



```
if not return_dict:
    output = (start_logits, end_logits) + outputs[2:]
    return ((total_loss,) + output) if total_loss is not None else output

return QuestionAnsweringModelOutput(
    loss=total_loss,
    start_logits=start_logits,
    end_logits=end_logits,
    hidden_states=outputs.hidden_states,
    attentions=outputs.attentions,
)
```

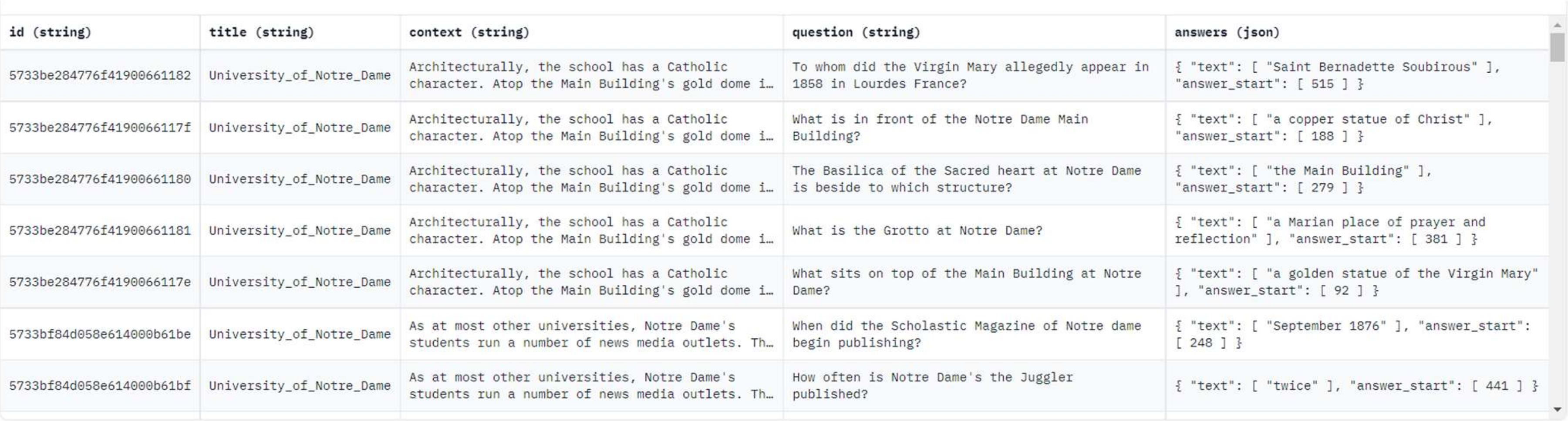


loss = softmax(left) + softmax(right)

alignment issue:

label is character-level
model is token-level

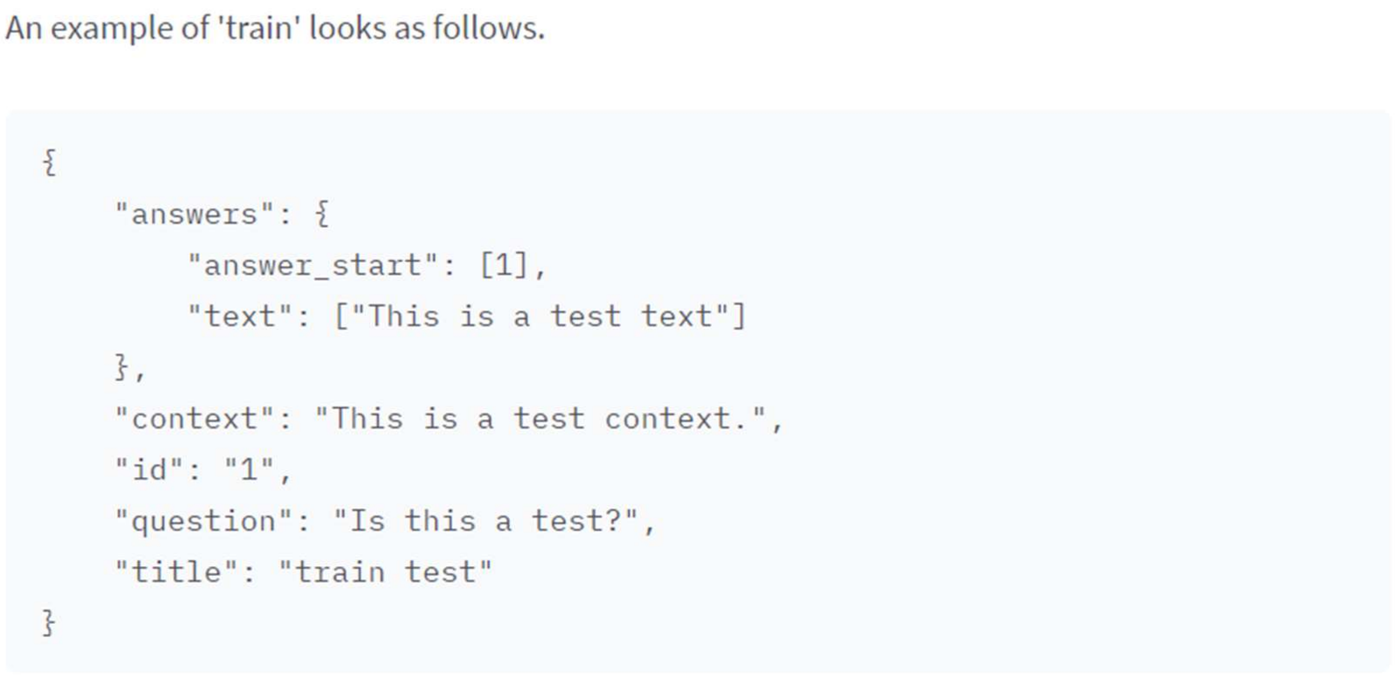
dataset format



id (string)	title (string)	context (string)	question (string)	answers (json)
5733be284776f41900661182	University_of_Notre_Dame	Architecturally, the school has a Catholic character. Atop the Main Building's gold dome i…	To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?	{ "text": ["Saint Bernadette Soubirous"], "answer_start": [515] }
5733be284776f4190066117f	University_of_Notre_Dame	Architecturally, the school has a Catholic character. Atop the Main Building's gold dome i…	What is in front of the Notre Dame Main Building?	{ "text": ["a copper statue of Christ"], "answer_start": [188] }
5733be284776f41900661180	University_of_Notre_Dame	Architecturally, the school has a Catholic character. Atop the Main Building's gold dome i…	The Basilica of the Sacred heart at Notre Dame is beside to which structure?	{ "text": ["the Main Building"], "answer_start": [279] }
5733be284776f41900661181	University_of_Notre_Dame	Architecturally, the school has a Catholic character. Atop the Main Building's gold dome i…	What is the Grotto at Notre Dame?	{ "text": ["a Marian place of prayer and reflection"], "answer_start": [381] }
5733be284776f4190066117e	University_of_Notre_Dame	Architecturally, the school has a Catholic character. Atop the Main Building's gold dome i…	What sits on top of the Main Building at Notre Dame?	{ "text": ["a golden statue of the Virgin Mary"], "answer_start": [92] }
5733bf84d058e614000b61be	University_of_Notre_Dame	As at most other universities, Notre Dame's students run a number of news media outlets. Th…	When did the Scholastic Magazine of Notre dame begin publishing?	{ "text": ["September 1876"], "answer_start": [248] }
5733bf84d058e614000b61bf	University_of_Notre_Dame	As at most other universities, Notre Dame's students run a number of news media outlets. Th…	How often is Notre Dame's the Juggler published?	{ "text": ["twice"], "answer_start": [441] }

https://huggingface.co/datasets/squad/viewer/plain_text/train

dataset format



An example of 'train' looks as follows.

```
{
    "answers": {
        "answer_start": [1],
        "text": ["This is a test text"]
    },
    "context": "This is a test context.",
    "id": "1",
    "question": "Is this a test?",
    "title": "train test"
}
```

Simple Transformers

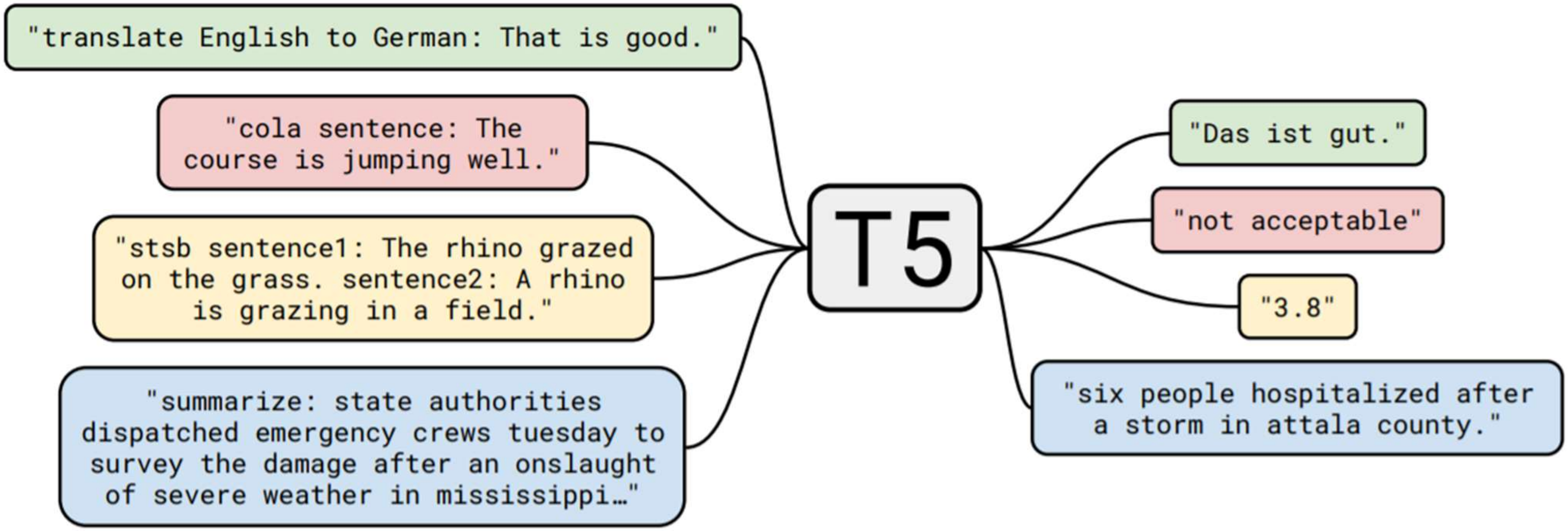
- https://simpletransformers.ai

Machine Translation

Data Source

- https://github.com/Helsinki-NLP/Tatoeba-Challenge/blob/master/data/README-v2021-08-07.md

T5 (Text-to-Text Transfer Transformer)



"translate English to German: That is good."

"cola sentence: The course is jumping well."

"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."

"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"

T5

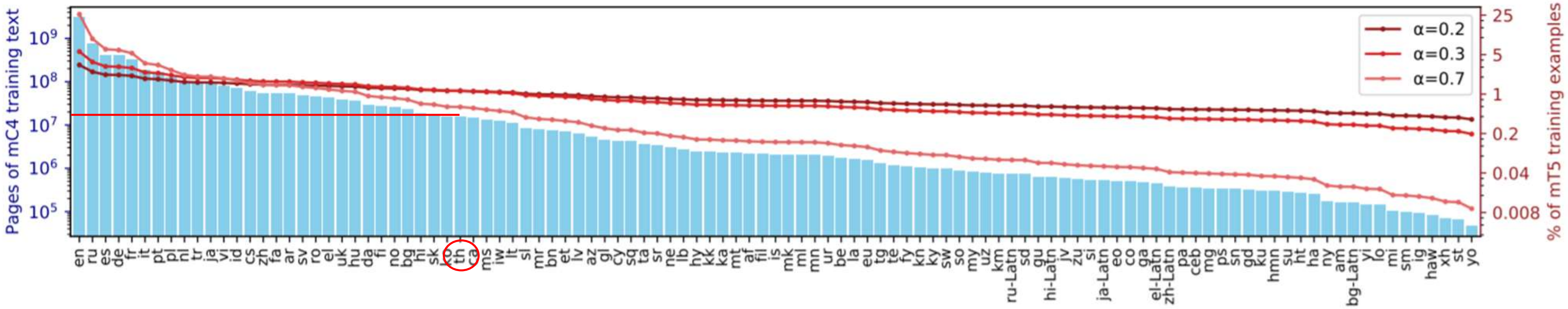
"Das ist gut."

"not acceptable"

"3.8"

"six people hospitalized after a storm in attala county."

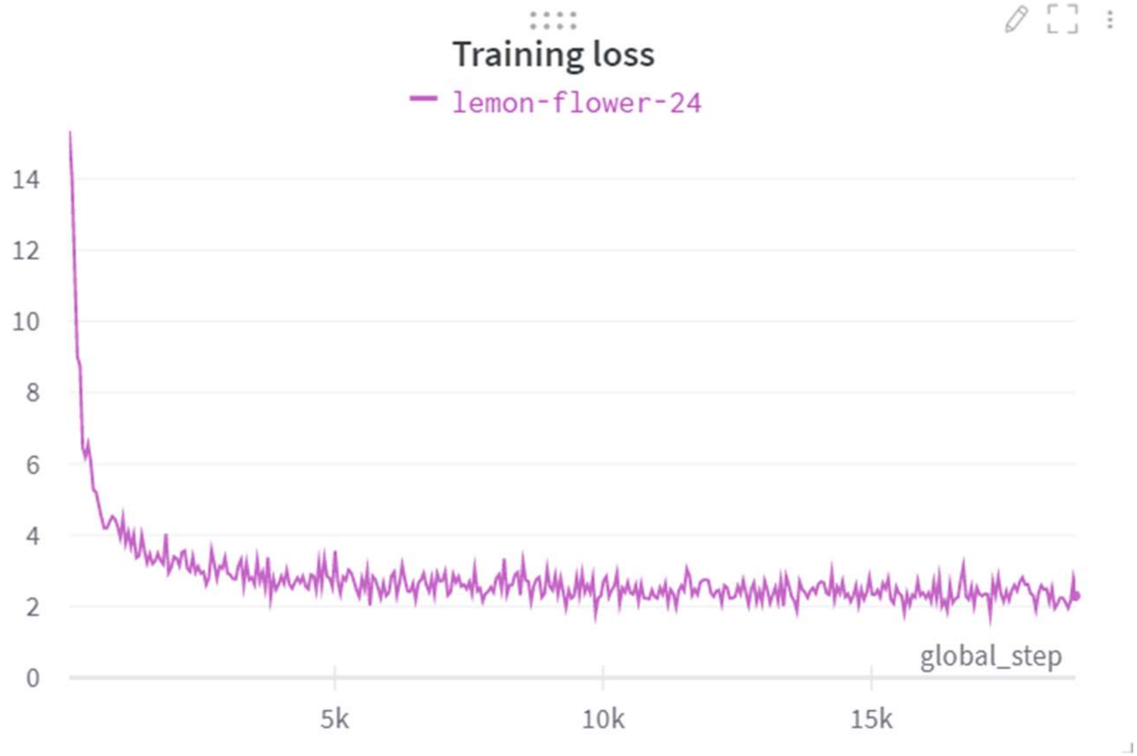
mT5 training data



one page of text is about 2 KB
20M pages of text is about 40 GB

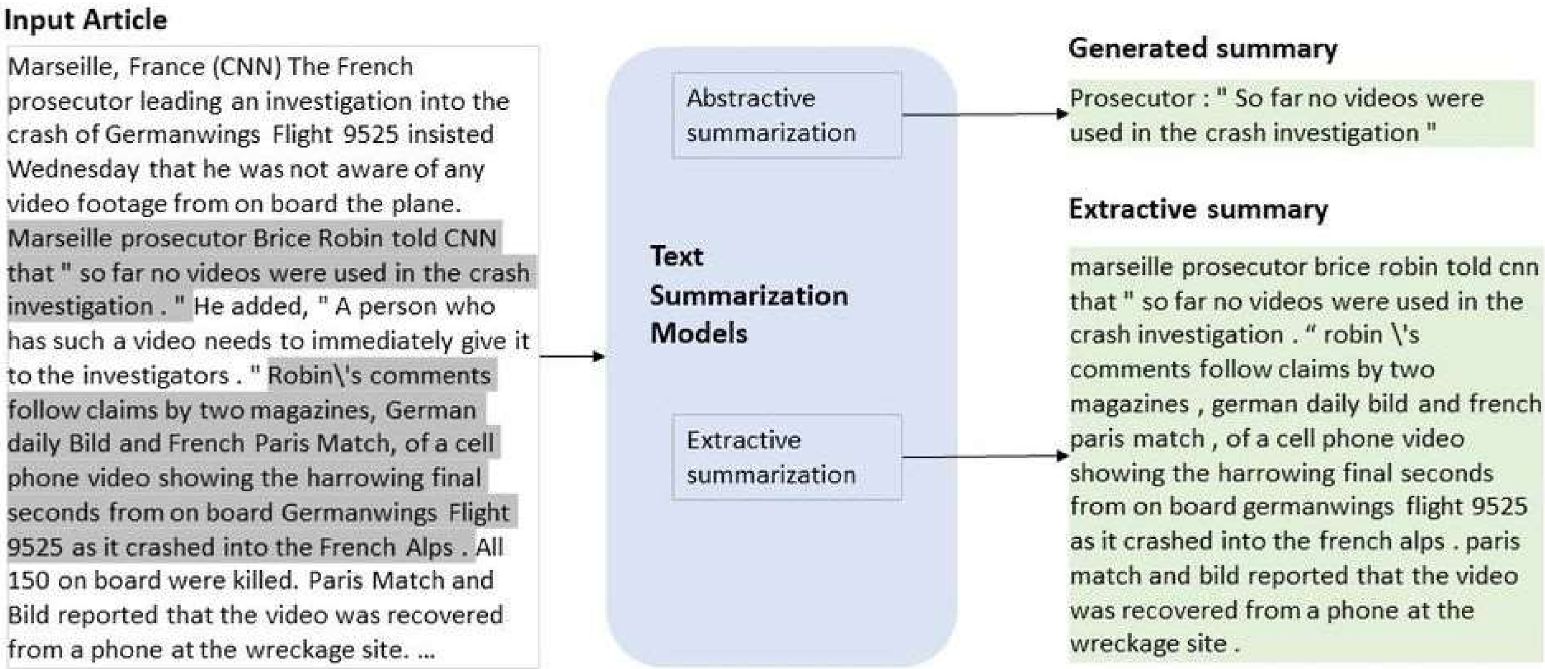
MT example

- 10 - MT.ipynb



Other Tasks

Summarization



Input Article

Marseille, France (CNN) The French prosecutor leading an investigation into the crash of Germanwings Flight 9525 insisted Wednesday that he was not aware of any video footage from on board the plane. Marseille prosecutor Brice Robin told CNN that " so far no videos were used in the crash investigation . " He added, " A person who has such a video needs to immediately give it to the investigators . " Robin\'s comments follow claims by two magazines, German daily Bild and French Paris Match, of a cell phone video showing the harrowing final seconds from on board Germanwings Flight 9525 as it crashed into the French Alps . All 150 on board were killed. Paris Match and Bild reported that the video was recovered from a phone at the wreckage site. ...

Text Summarization Models

Abstractive summarization

Extractive summarization

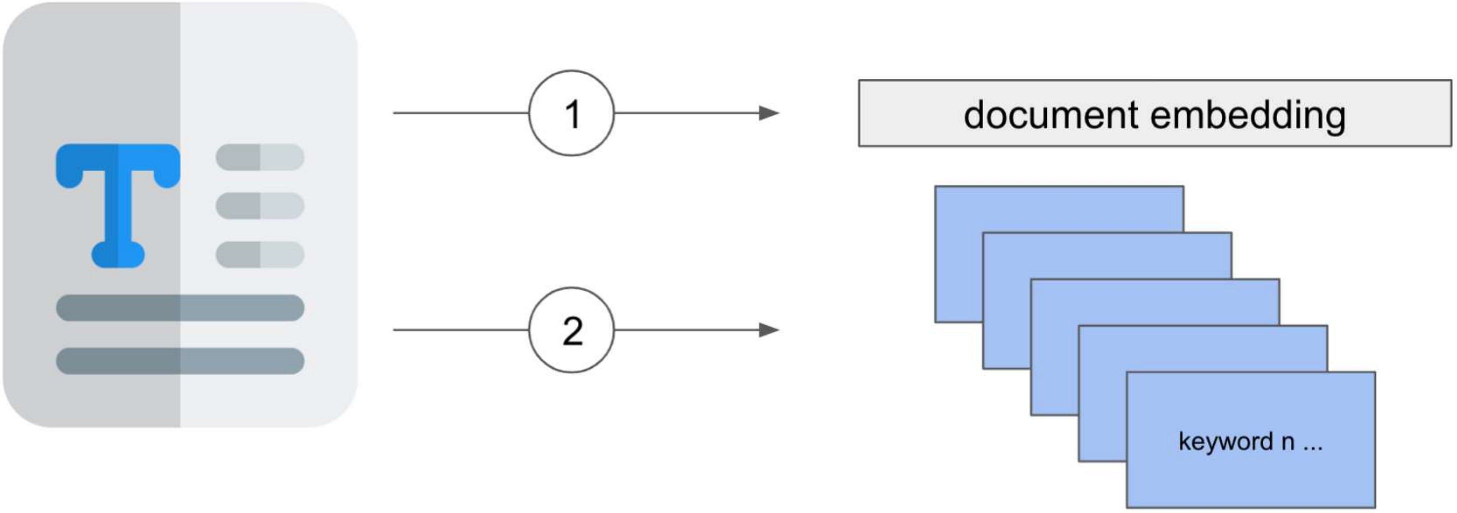
Generated summary

Prosecutor : " So far no videos were used in the crash investigation "

Extractive summary

marseille prosecutor brice robin told cnn that " so far no videos were used in the crash investigation . " robin \'s comments follow claims by two magazines , german daily bild and french paris match , of a cell phone video showing the harrowing final seconds from on board germanwings flight 9525 as it crashed into the french alps . paris match and bild reported that the video was recovered from a phone at the wreckage site .

Keyword Extraction



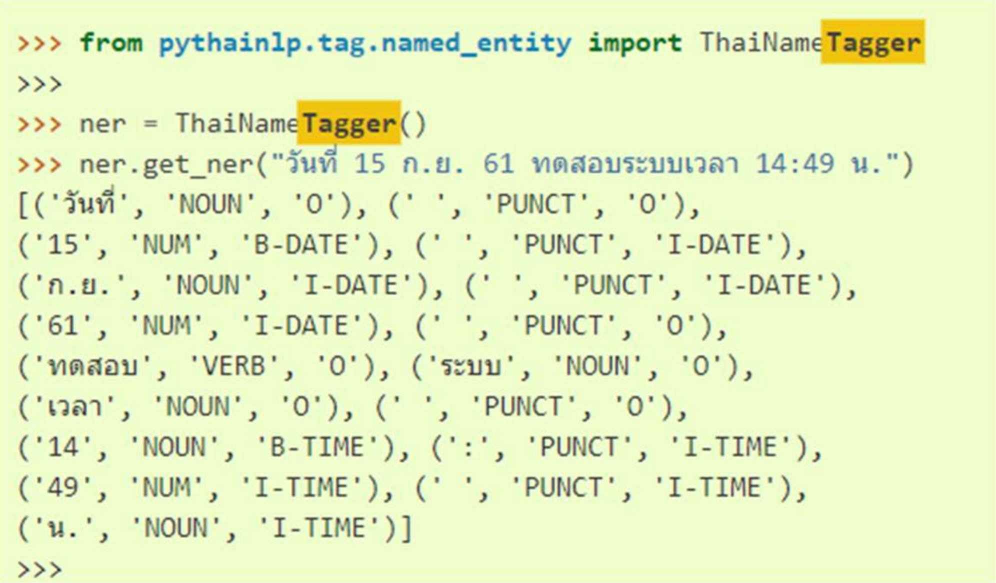
Project Topic 1

- Extract location from tweets

1. Mine Twitter for a certain hashtag.
2. Extract location from tweets.
3. (Extra) visualize on map.

NER tagger

- https://pythainlp.github.io/docs/3.0/api/tag.html?highlight=tagger



```
>>> from pythainlp.tag.named_entity import ThaiNameTagger
>>>
>>> ner = ThaiNameTagger()
>>> ner.get_ner("วันที่ 15 ก.ย. 61 ทดสอบระบบเวลา 14:49 น.")
[('วันที่', 'NOUN', 'O'), (' ', 'PUNCT', 'O'),
('15', 'NUM', 'B-DATE'), (' ', 'PUNCT', 'I-DATE'),
('ก.ย.', 'NOUN', 'I-DATE'), (' ', 'PUNCT', 'I-DATE'),
('61', 'NUM', 'I-DATE'), (' ', 'PUNCT', 'O'),
('ทดสอบ', 'VERB', 'O'), ('ระบบ', 'NOUN', 'O'),
('เวลา', 'NOUN', 'O'), (' ', 'PUNCT', 'O'),
('14', 'NOUN', 'B-TIME'), (':', 'PUNCT', 'I-TIME'),
('49', 'NUM', 'I-TIME'), (' ', 'PUNCT', 'I-TIME'),
('น.', 'NOUN', 'I-TIME')]
>>>
```

Project Topic 2

- Extract sentiments from tweets

1. Extract top k most common positive words
2. Extract top k most common negative words
3. Overall sentiment (positive, negative, neutral) of tweets with this hashtag

Sentiment Words List

- https://github.com/PyThaiNLP/lexicon-thai/tree/master/sentiment