Utterance end detection

Tomáš Chobola

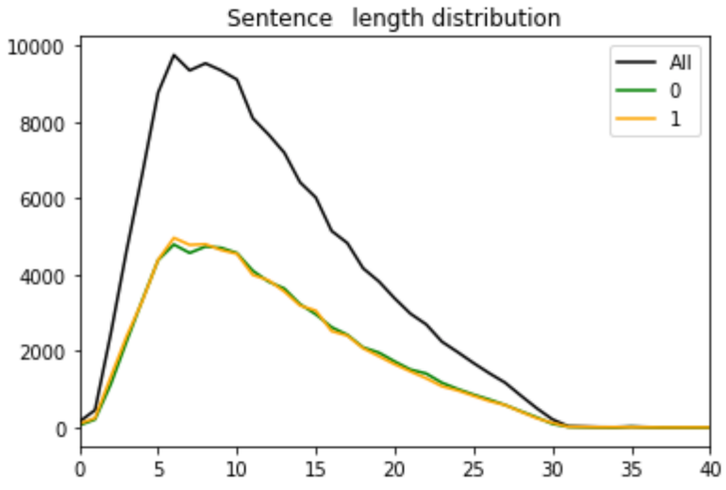
Task

- Classify an utterance as finished or not-finished based on the already spoken words
- **Binary classification**
    - “I would like to know which way to go.” = finished
    - “I would like to know which” = not-finished

Datasets

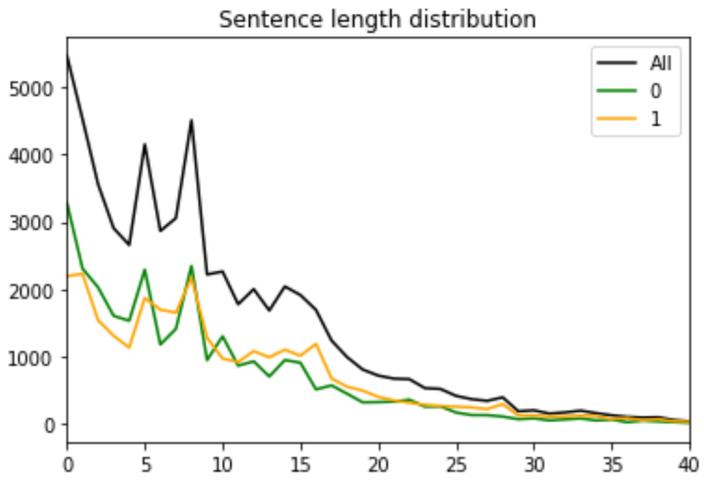
MultiWOZ
Human-human written conversations spanning over multiple domains and topics

142,898 utterances



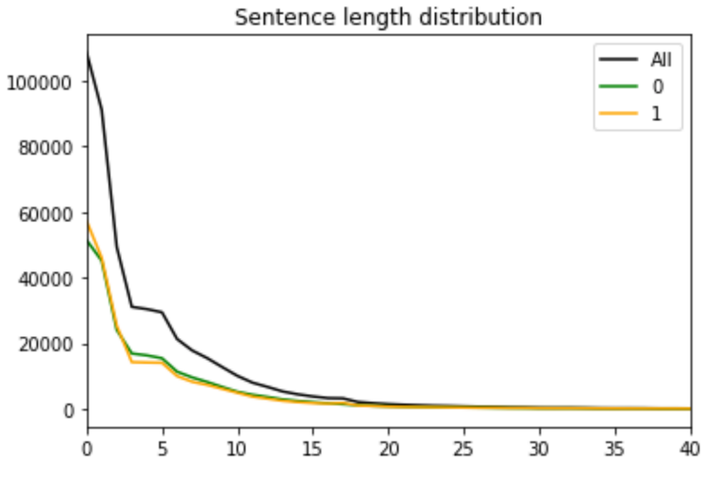
E2E Challenge Dataset
Conversations from three environments: restaurants, theatres and interactions between taxi drivers and their customers

48,721 utterances



Taskmaster
Goal-oriented conversational data, multiple categories: flights, food ordering, hotels, movies, restaurant search and sports

474,624 utterances



~Half of the sentences in the datasets cut short to simulate not-finished utterance

Networks from ML Workshop Prague 2018

- Workshop focused on predicting the sentiment of IMDB movie reviews
- Use of the pre-trained GloVe word vector representations
- Applying the same network to classify finished/not-finished utterances

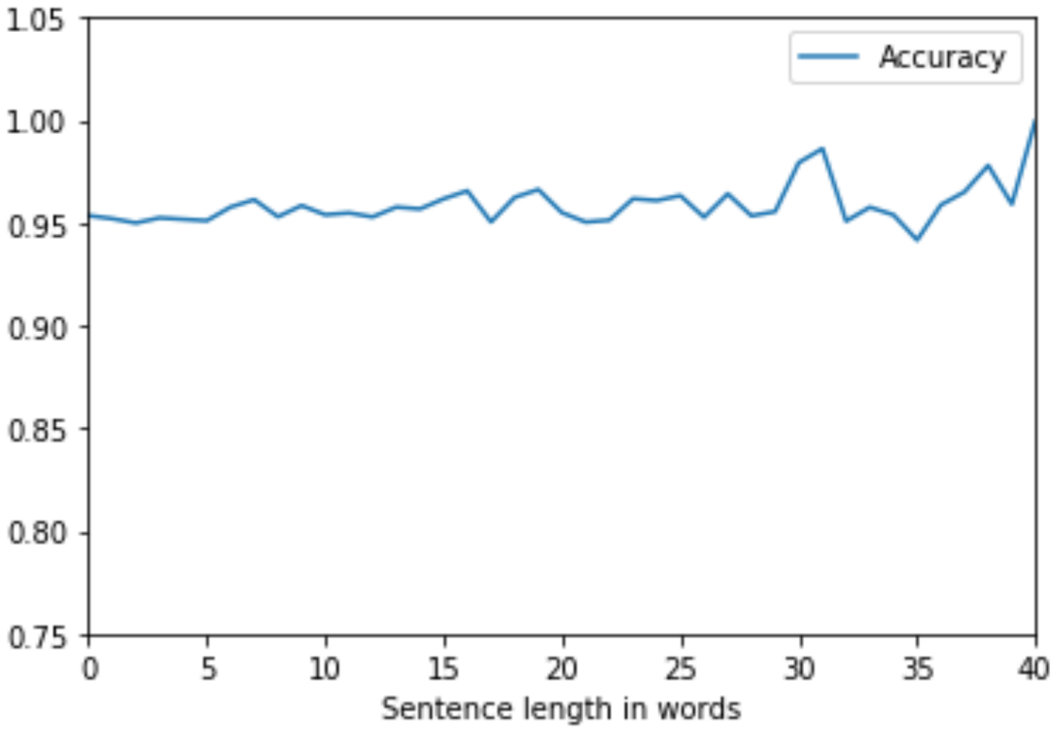
Networks from ML Workshop Prague 2018

Model accuracies based on the training datasets
(accuracy decreases with increasing size of the dataset)

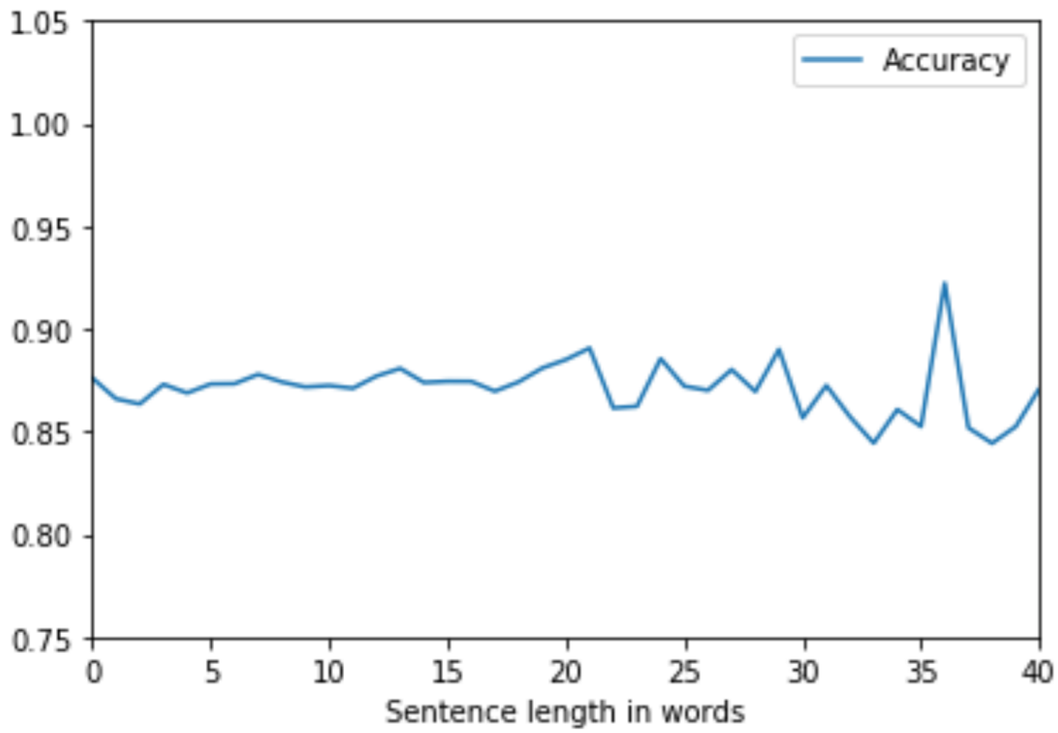
	Epoch 1	Epoch 5	Epoch 10
MultiWOZ	0.8508	0.9055	0.9202
E2E	0.8931	0.9478	0.9584
Taskmaster	0.8360	0.8636	0.8666

Networks from ML Workshop Prague 2018 - analysis of accuracy based on sentence length

E2E dataset



Taskmaster dataset



Classification examples

0 - utterance is not-finished, 1 - utterance is finished

```
0 [0.00060312] could you
1 [0.9447847] is there one in the centre with free parking
0 [0.00611118] yes please there
1 [0.9947095] great that was all i needed today thank you for your help
1 [0.03836118] no i am sorry
0 [0.00016204] train tr leaves
0 [0.7648024] the pick up location will
1 [0.00033437] it has stars
0 [3.3340782e-06] i also
1 [5.439254e-05] got it when are you planning to leave
1 [0.0008164] three hamburgers
0 [0.00010264] i found
1 [0.9996666] what date and time would you like to go
1 [8.300042e-06] round trip
1 [0.9671395] which one do you like
```

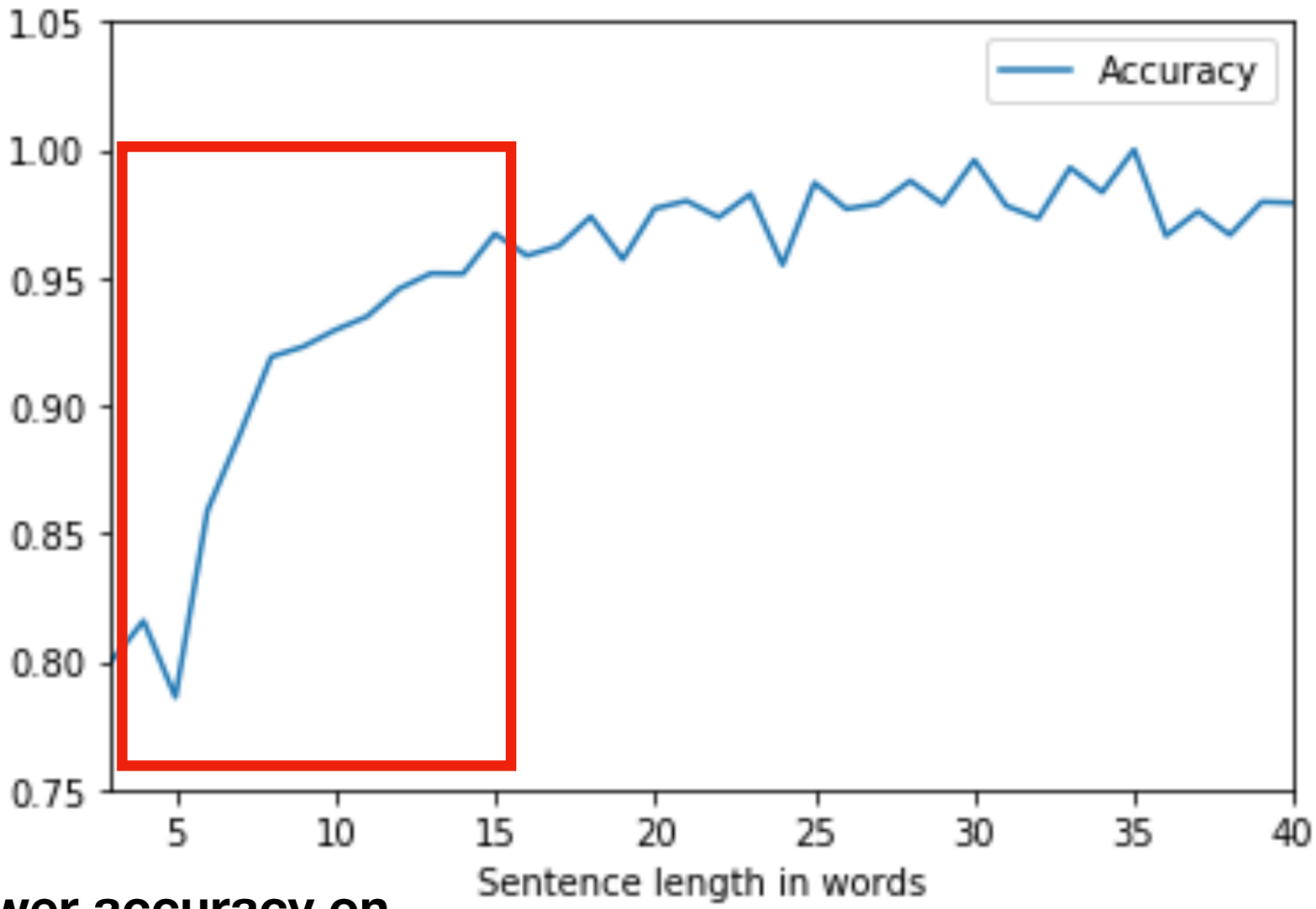
BERT models

- Applied only on the Taskmaster dataset
  (lowest accuracy while using the previous model)
- **BERT tokeniser**
  - Transforming the sentences into an input accepted by the BERT model
  - Using separate simple neural network based on six Conv1D layers
- **Full BERT model**

Accuracy of the model based on BERT tokeniser (Taskmaster dataset)

	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5
Accuracy	0.9048	0.9313	0.9453	0.9544	0.9598

Model based on the BERT tokeniser - analysis of accuracy based on sentence length



Lower accuracy on short sentences (3-15 words)

Model based on the BERT tokeniser - analysis of accuracy based on sentence length

Only one naturally ended sentence was classified as non-ended:

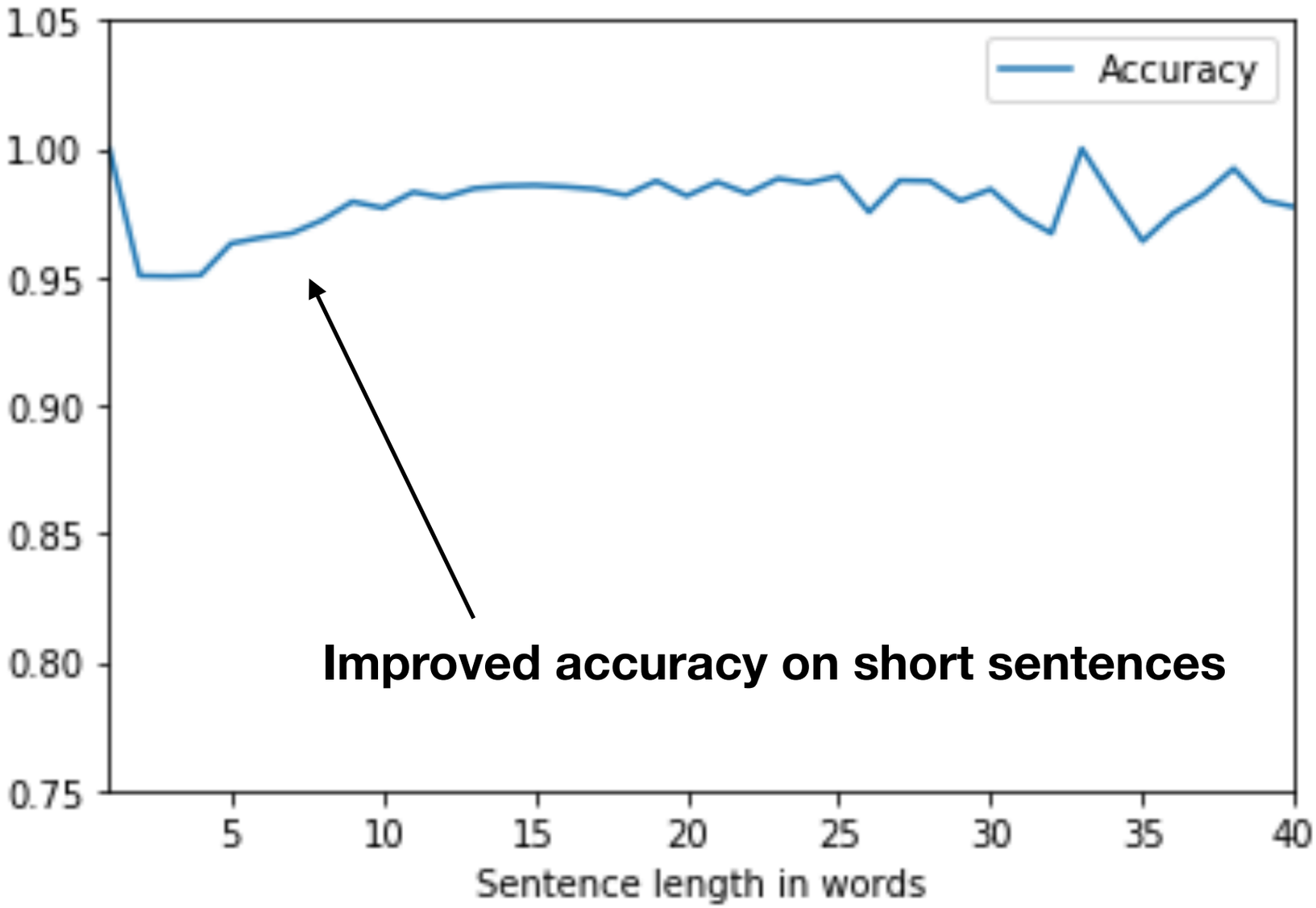
`'find one that has some pretty good customer reviews'`

These sentences are non-ended, however they were classified as ended (the ones that could be taken as ended (= wrong split) are marked with an exclamation mark):

`'no that's it thank you'` ❗
`'i'm feeling italian i think'` ❗
`'do you have any requirements'` ❗
`'expensive alright that sounds good'` ❗
`'is there anything else i'`
`'and if any of them'`
`'could you tell me the synopsis'` ❗
`'is there anything else that you'`
`'and citizen hotel it will cost you'`
`'yes i did are any of your preferences'`
`'yes they come with organic soy sauce is'`
`'okay let me look into that for you'` ❗
`'do you have any requirements for this restaurant and'`
`'make sure it has good reviews i don't want'`
`'they are at truck so loud and the information will be'`
`'non stop if possible i want to search if i can do it'` ❗

Testing on small dataset of 550 sentences
(50 sentences for each length in interval [5,15] words)

Full BERT model - analysis of accuracy based on sentence length



Full BERT model - analysis of accuracy based on sentence length

Falsely classified as non-ended:

`'it is called one room'`

`'okay perfect the price for that is'`

`'okay i have booked this hotel for you enjoy your stay your welcome bye'`

`'okay i can help with that do you want to stay in san francisco'`

Falsely classified as naturally ended (in the dataset they are labeled as non-ended):

`'okay let's pick from one'`

`'and can i spell it'` ❗

`'the nd rating would be'`

`'thank you for your help'` ❗

`'thank you for your help today'` ❗

`'yes the showing is on thursday'` ❗

`'or would you prefer something more'`

`'spiderman is very popular right now'` ❗

`'great could i see the trailer'` ❗

`'can i get a hangover with cheddar cheese fried egg'` ❗

`'and i want it to be downtown but kind of where'`

`'well i'd like an action movie and something in d close by'`

`'hello i need to find a movie on google play that's a comedy'` ❗

`'and then how far is that from the audubon butterfly place that i wanted'` ❗

`'yeah i was wondering if you could help me find a movie to watch'` ❗

`'oh yeah i heard that's a really good one i'd like to see saving private'` ❗

`'i have a cozy menu it should be near the beach it should have star'`

Testing on small dataset of 550 sentences
(50 sentences for each length in interval [5,15] words)

The majority of wrongly classified sentences comes from a wrong split (even though it is labeled as non-finished, it could be taken as finished)

= potentially higher accuracy

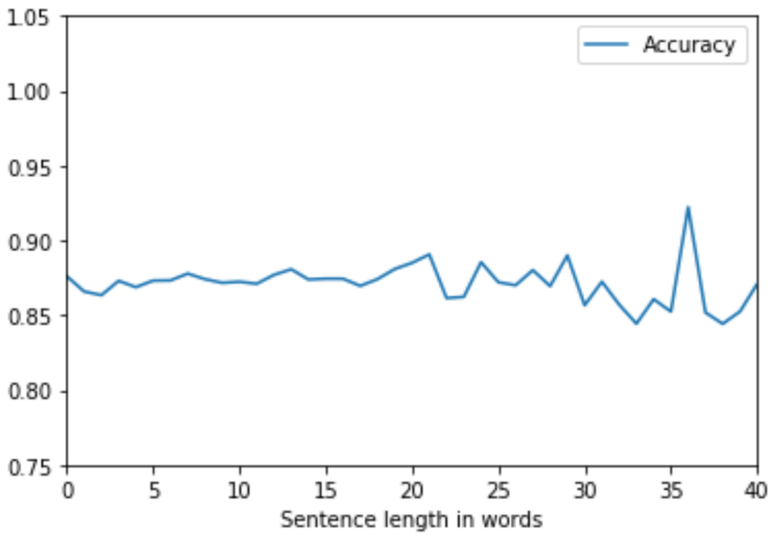
Computational demands

- Times based on the AlquistAI cloud instance (run on CPU), all models are trained on the Taskmaster dataset

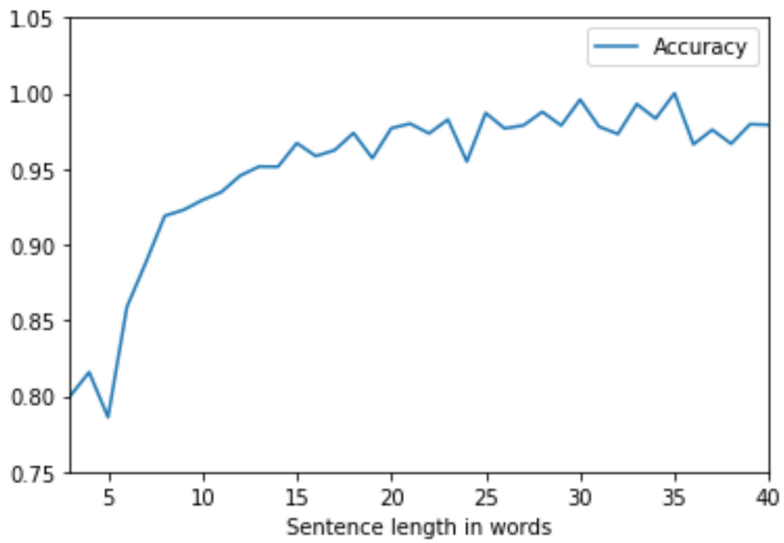
	Epoch 1 Accuracy	Training time	1 sentence prediction time in seconds	100 sentences prediction time in seconds
“Workshop” network	0.8632	~1 minute	0.0258	0.0343
BERT Tokeniser	0.9598	~10 minutes	0.0279	0.0370
Full BERT model	0.9626	~4 hours	0.2402	9.3449

Accuracies comparison based on the Taskmaster dataset

“Workshop” network



BERT tokeniser model



Full BERT model