Improving Event Duration Question Answering by Leveraging Existing Temporal Information Extraction Data

LREC 2022 Marseille

Felix Giovanni Virgo, Fei Cheng, Sadao Kurohashi
Graduate School of Informatics, Kyoto University

京都大学 KYOTO UNIVERSITY

Introduction

Duration Question Answering: McTACO (Zhou et.al, 2019)

If you have ever heard, "Eat a good breakfast", thats why.
How long does it take to ***eat breakfast****?*

— 15 minutes ✔ *plausible*
— several days ✘ *not plausible*
— 20 minutes ✔ *plausible*

The performance of modern pre-trained NLP models for this task is still far behind humans due to limited training data.

There are plenty of auxiliary resources containing duration information, e.g. UDS-T dataset (Vashishtha et.al., 2019), that can be used to improve McTACO.

However, a straightforward two-stage fine-tuning is less likely to succeed since there are discrepancy between the two tasks:
- UDS-T : Duration Unit Classification
- McTACO : Duration Question Answering

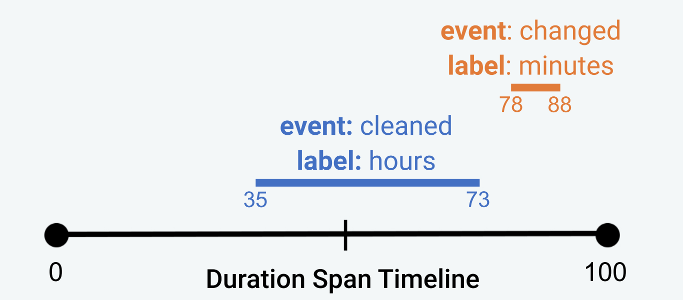
We need to bridge the discrepancy between the two tasks.

Duration Task Recasting

We bridge the discrepancy by recasting Duration Information Extraction dataset into Question Answering.

UDS-T Duration Classification

Their worker even cleaned 3 of my windows and changed a lightbulb for me.



Recasting Steps:
1. Irrelevant Contexts Removal
2. Question Generation
3. Candidate Answer Generation
   - Positive answer generation
   - Negative answer generation

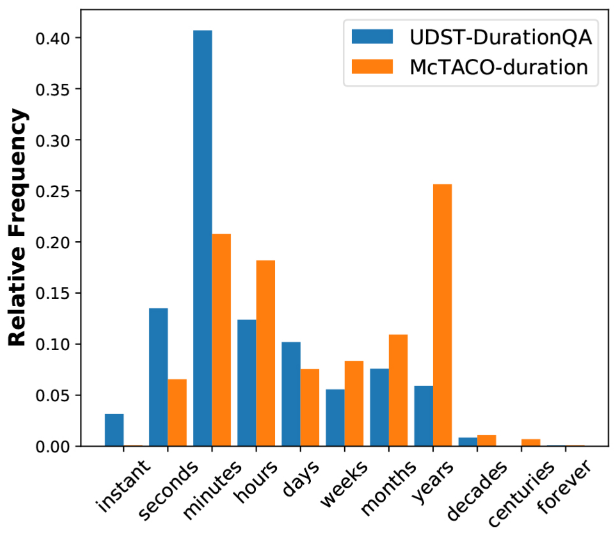
UDS-T Duration QA

Their worker even cleaned 3 of my windows and changed a lightbulb for me.

How long does it take for their worker to clean 3 of my windows?

- 2 hours ✔ *plausible*
- a few hours ✔ *plausible*
- several years ✘ *not plausible*
- 4 months ✘ *not plausible*

Statistics



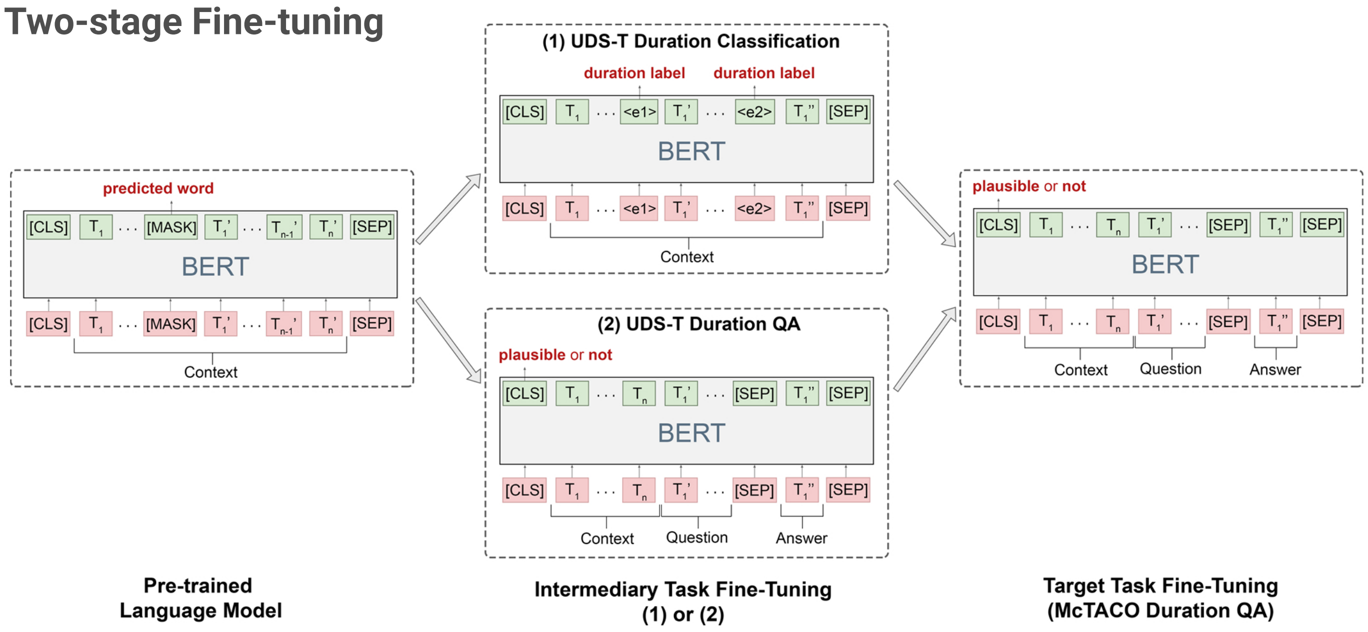
The duration distribution of our recast data is relatively similar to McTACO, except for *minutes* and *years*.

Number of QA pairs

	train + dev	test
McTACO-duration :	1.1k	3.0k
UDST-DurationQA :	39.9k + 4.9k	4.8k

Experiments

Two-stage Fine-tuning



Results

Model	EM	F1
RoBERTa-large → McTACO-duration	40.45	67.42
RoBERTa-large → UDS-T (duration cls.) → McTACO-duration	39.49	64.95
RoBERTa-large → UDST-DurationQA (unit only) → McTACO-duration	42.78	66.97
RoBERTa-large → UDST-DurationQA → McTACO-duration	**45.86**	**70.52**

Additional Experiments

Are the setups that leverage two-stage fine-tuning more effective than multi-task learning?

Model	EM	F1
Two-stage Fine-tuning	**45.86**	**70.52**
Multi-task Learning	41.72	66.93

How does our proposed method compare to a SOTA pretrained temporal common sense language model?

Model	EM	F1
TACOLM (Zhou et. al., 2020) → McTACO	34.60	-
BERT-base → McTACO	33.76	60.98
BERT-base → UDST-DurationQA → McTACO	**36.52**	**63.22**