UCD-CS at W-NUT 2020 Shared Task-3: A Text to Text Approach for COVID-19 Event Extraction on Social Media

Congcong Wang, David Lillis

School of Computer Science, University College Dublin, Ireland.

ABSTRACT

This poster presents our approach in the shared task: *COVID-19 event extraction from Twitter (W-NUT Task-3)*. **Our approach treats the event extraction task as a question answering task by leveraging the transformer-based T5 text-to-text model.** According to the official evaluation scores returned, namely F1, our submitted run achieves competitive performance. However, **we argue that this evaluation may underestimate the actual performance of runs based on text-generation**. Although some such runs may answer the slot questions well, they may not be an exact string match for the gold standard answers. To measure the extent of this underestimation, **we adopt a simple exact-answer transformation method aiming at converting the well-answered predictions to exactly-matched predictions. The results show that after this transformation our run overall reaches the same level of performance as the best participating run and state-of-the-art F1 scores in three of five COVID-related events.**

OBJECTIVES

The major objective of W-NUT Task-3 is to seek computational linguistic techniques for extracting text spans from a corpus of raw tweets to answer a set of predefined slot questions.

- The corpus used in the task can be described in two parts. **Part 1, the corpus consists of tweets categorised into five broad event types: (1) TESTED POSITIVE, (2) TESTED NEGATIVE, (3) CAN NOT TEST, (4) DEATH and (5) CURE AND PREVENTION. Part 2, for the tweets in each event, a set of questions or slot-filling types are defined to help gather more fine-grained information about the tweets. The human annotations of the corpus are simply the answers to the predefined questions** (see Figure 1).

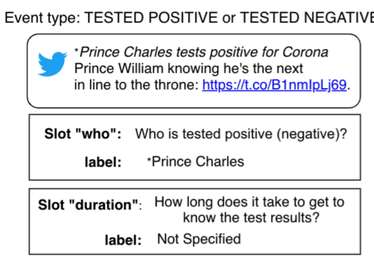

Figure 1: An example of the event extraction task: this shows a tweet represents a `TESTED POSITIVE` or `TESTED NEGATIVE` event. The objective is to extract answers to the slot questions concerning the event.

- **The annotation process in part 2 is conducted by annotators who select answers from a drop-down list of candidate choices.** The choices are automatically-extracted text spans obtained through a Twitter tagging tool or predefined choices such as ``not specified'', ``yes'', ``no'', etc. This explains why the label for slot ``who'' in Figure 1 has the symbol * at the beginning.

METHODS

Figure 2 presents the system architecture of our approach. The core component of the system is the sequence construction, which converts each tweet in the dataset into a source sequence and target sequence that well fits to train the transformer encoder-decoder model in a text-to-text fashion. **As introduced, each tweet in the dataset is annotated in two parts, 1), labels indicating the event types, 2), answers to the slot questions. Our approach uses both parts of the annotations to construct the source and target sequences** as follows:

- **Source:** This sequence is constructed from the raw tweets through mapping their major four attributes, i.e., as outlined in Figure 2, the tweet text is mapped to "context", event type question (part 1) or slot question (part 2) 3 to "question", and the candidate choices to "choices". The leads to the final source sequence concatenating all these fields in the form: **"context:{tweet text} question: {slot/event question} choices:{candidates}".**
- **Target:** The target sequence needs both part 1 and part 2 annotations in training. They are simply mapped from the event labels for part 1 ("yes" or "no" indicating if the tweet falls into the event as specified in the question field of the source sequence) and annotation answers for part 2. Where the annotations are not available in inference, the source sequence and the target sequence with the start decoder token *<pad>* are fed as input to the model for generating the answers directly.

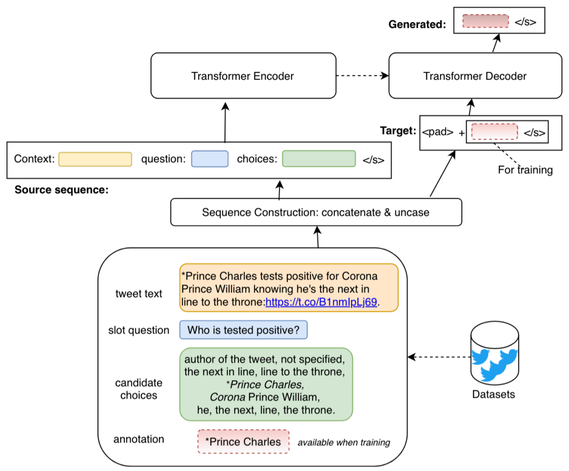

Figure 2: The system architecture of our approach

In our experiments, we choose **T5** (Raffel et al.,2020) as the encoder-decoder model's architecture and train it on this task via fine-tuning its pre-trained both small, base and large weights with 12 epochs. We evaluated it on the test set at epoch 12 (run-1), 11 (run-2), and 10 (run-3) respectively.

RESULTS

Takeaways

- **Larger model gives better performance but the marginal gain decreases** (see Table 2a).
- **Our approach-based officially-submitted run-2 exhibits a significant advantage in recall over other runs** (0.7118 versus the second-highest of 0.6039, see Table 2b). However, **this advantage is combined with a trade-off in terms of precision**. This reveals that **our run is somewhat "active" at finding the answers to the slot questions.**
- **Evaluating at per event level, our run achieves the best recall in every event type but quite behind in precision, especially for *cannot test*, *death* and *cure*.** To further analyse the cause of low precision, we rethink the metrics used to evaluate runs with a similar approach to ours via post-processing.

	F1	P	R	Params
small-2	0.5308	0.4308	0.6913	1.0x
base-2	0.6225	0.5449	0.7258	3.7x
large-2	0.6392	0.5800	0.7118	12.8x

(a) Evaluation results of different sizes of models where **large-2** is the officially submitted **run-2** and Params refers to the model's parameters relative to `t5-small` that has around 60M parameters.

team name	F1	P	R
winners	0.6598⋆	0.7272	0.6039
HLTRI	0.6476	0.7532⋆	0.5679
Our (run-2)	0.6392	0.5800	0.7118⋆
VUB	0.6160	0.6875	0.5580
UPennHLP	0.5237	0.6754	0.4277
Test_Positive	0.5114	0.5377	0.4875

(b) Evaluation results of submitted runs ranked by F1. **Our (run-2)** is named **UCD_CS** officially.

Ours	F1	P	R
run-1	0.6429⋆	0.5815	0.7188 ⋆
run-2	0.6392	0.5800	0.7118
run-3	0.6367	0.5920⋆	0.6887

(c) Evaluation results of our `t5-large` runs at different epochs.

Ours	F1	P	R
post-run-1	0.6571⋆	0.5956	0.7327⋆
post-run-2	0.6517	0.5921	0.7247
post-run-3	0.6495	0.6050⋆	0.7012

(d) Evaluation results of our `t5-large` runs after post-processing.

Table 2: F1, precision (P) and recall (R) scores averaged over the five COVID-19 events where ⋆ refers to the highest in each column.

	F1			P			R		
	best	run-2	post-run-2	best	run-2	post-run-2	best	run-2	post-run-2
positive	0.6973	0.6778	**0.6989**	0.8569	0.7380	0.7620	0.6267	0.6267	**0.6454**
negative	0.7030	0.7030	**0.7047**	0.7107	0.6873	0.6890	0.7194	0.7194	**0.7212**
can_not_test	0.6523	0.5660	0.5667	0.6863	0.4646	0.4656	0.7240	0.7240	0.7240
death	0.6942	0.6048	0.6191	0.7240	0.4917	0.5041	0.7855	0.7855	**0.8020**
cure	0.6205	0.6078	**0.6236**	0.8405	0.4961	0.6236	0.7843	0.7843	**0.8028**

Table 4: The evaluation results of our run-2 and post-run-2 at event type level where best represents the highest score across all participating runs. These in bold stand for new state-of-the-art scores in W-NUT task-3 after applying **TransM**.

POST-PROCESSING

Having observed some of the mismatches in our experiment (see Table 3), we subsequently transformed these to exact-matched answers using a simple approach based on Levenshtein string edit distance, which we name **TransM**. **After this transformation, our best run reaches the same level of F1 performance as the best participating run (see Table 2d), leading to state-of-the-art F1 scores in *positive, negative,* and *cure* events** (see Table 4).

Example 4	Tweet text	@joshua4congress my sister is a vet with an active-duty husband and a 6wk old baby. she has a fever and symptoms but can't get tested bc she lives in a military base in texas. they require exposure to a confirmed positive. she had to give birth without family the day after our grandma died.
	Slot question (where)	**where is the can't-be-tested situation reported?**
	Our raw prediction	a military base in texas
	Post-processed prediction	a military base
	Ground truth	texas
Example 5	Tweet text	so in a few weeks time 4 year olds will be expected to be back in school but can't get tested. doesn't make sense. #covid19 #dailybriefings
	Slot question (who)	**who can not get a test?**
	Our raw prediction	4 year olds
	Post-processed prediction	year olds
	Ground truth	a few weeks time 4 year olds

Table 3: Examples of mismatches (uncased), i.e., predictions that are self-evidently correct answers, but do not exactly match the ground truths in violet text. The orange text refers to the original submitted predictions generated by our approach and the blue text refers to the transformed predictions from the raw predictions based on their edit distances to candidate answers.

CONCLUSION

We presents our text-to-text based approach at W-NUT 2020 shared task 3. We show that the principal idea behind the approach is adaptability to other domain-similar tasks such as informativeness classification of COVID-19 tweets. **We expect to conduct more work on this adaptability in the future. It is even more interesting to test the idea in zero-shot learning.** For example, how well it performs if transferring the model that is trained on the event extraction corpus to do inference in the informativeness task directly without further training. In addition, we empirically present that our system is effective, achieving competitive performance and arguably the state-of-the-art F1 scores in three of five COVID-events in the shared task. Despite the effectiveness, one concern of our approach is the model size. Our best performed model is fine-tuned using the large version of T5 with around 770M parameters. **This makes it important to compress the model efficiently in the future**.

ACKNOWLEDGEMENTS

We would like to thank Google's TensorFlow Re-search Cloud (TFRC) team who provided TPUscredits to support this research.

REFERENCES

- Shi Zong, Ashutosh Baheti, Wei Xu, and Alan Ritter.2020. Extracting COVID-19 Events from Twitter.
- Colin Raffel, Noam Shazeer, Adam Roberts, KatherineLee, Sharan Narang, Michael Matena, Yanqi Zhou,Wei Li, and Peter J Liu. 2020. Exploring the limitsof transfer learning with a unified text-to-text transformer.Journal of Machine Learning Research,21(140):1–67.
- Vladimir I Levenshtein. 1966. Binary codes capableof correcting deletions, insertions, and reversals. InSoviet physics doklady, volume 10, pages 707–710.

RESEARCH POSTER PRESENTATION DESIGN © 2015
www.PosterPresentations.com