

TEST POSITIVE at W-NUT 2020 Shared Task-3: Joint Event Multi-task Learning for Slot Filling in Noisy Text

Chacha Chen¹

Penn State
chachachen@psu.edu

Chieh-Yang Huang

Penn State
chiehyang@psu.edu

Yaqi Hou

UNC at Chapel Hill
yaqi.hou@unc.edu

Yang Shi

NC State
yshi26@ncsu.edu

Enyan Dai

Penn State
emd5759@psu.edu

Jiaqi Wang^{*}

Penn State
jqwang@psu.edu

Abstract

The competition of extracting COVID-19 events from Twitter is to develop systems that can automatically extract related events from tweets. The built system should identify different pre-defined slots for each event, in order to answer important questions (e.g., *Who is tested positive? What is the age of the person? Where is he/she?*). To tackle these challenges, we propose the Joint Event Multi-task Learning (JOELIN) model. Through a unified global learning framework, we make use of all the training data across different events to learn and fine-tune the language model. Moreover, we implement a type-aware post-processing procedure using named entity recognition (NER) to further filter the predictions. JOELIN outperforms the BERT baseline by 17.2% in micro F1.¹

1 Introduction

In this work, we report the system architecture and results of the team TEST_POSITIVE in the competition of W-NUT 2020 shared Task-3: extracting COVID-19 event from Twitter.

Since February 2020, the pandemic COVID-19 has been spreading all over the world, posing a significant threat to mankind in every aspect. The information sharing about a pandemic has been critical in stopping virus spreading. With the recent advance of social networks and machine learning, we are able to automatically detect potential events of COVID cases, and identify key information to prepare ahead.

We are interested in COVID-19 related event extraction from tweets. With the prevalence of coronavirus, Twitter has been a valuable source of news and information. Twitter users share COVID-19 related topics about personal narratives and news on social media (Müller et al., 2020). The information

could be helpful for doctors, epidemiologists, and policymakers in controlling the pandemic. However, manual extracting useful information from tremendous amount of tweets is impossible. Hence, we aim to develop a system to automatically extract structured knowledge from Twitter.

Extracting COVID-19 related events from Twitter is non-trivial due to the following challenges:

(1) **How to deal with limited annotations in heterogeneous events and subtasks?** The creation of the annotated data relies completely on human labors, and thus only a limited amount of data can be obtained in each event categories. There are a variety types of events and subtasks. Many existing works solve these low resource problem by different approaches, including crowdsourcing (Müller et al., 2020; Finin et al., 2010; Potthast et al., 2018), unsupervised training (Xie et al., 2019; Hsu et al., 2017), or multi-task learning (Zhang and Yang, 2017; Pentyala et al., 2019). Here we adopt multi-task training paradigm to benefit from the inter-event and intra-event (subtasks) information sharing. In this way, JOELIN learns a shared embedding network globally from all events data. In this way, we implicitly augment the dataset by global training and fine-tuning the language model.

(2) **How to make type-aware predictions?** Existing work (Zong et al., 2020) did not encode the information of different subtask types into the model, while it could be useful in suggesting the candidate slot entity type. In order to make type-aware predictions, we propose a NER-based post-processing procedure in the end of JOELIN pipeline. We use NER to automatically tag the candidate slots and remove the candidate whose entity type does not match the corresponding subtask type. For example, as shown in Figure 1, in subtask “Who”, “my wife’s grandmother” is a valid candidate slot, while “old persons home”, tagged as location entity, would be replaced with “Not Specified” during the

¹<https://github.com/Chacha-Chen/JOELIN>

post-processing.

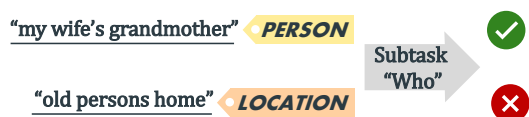


Figure 1: Illustration of NER-based post-processing.

In summary, JOELIN is enabled by the following technical contributions:

- **A joint event multi-task learning framework for different events and subtasks.** With the unified global training framework, we train and fine-tune the language model across all events and make predictions based on multi-task learning to learn from limited data.
- **A NER-based type-aware post-processing approach.** We leverage NER tagging on the model predictions and filter out wrong predictions based on subtask types. In this way, JOELIN benefits from subtask type prior knowledge and further boosts the performance.

2 Related Work

Event Extraction from Twitter Impressive efforts have been made to detect events from Twitter. Existing works include domain specific event extraction and open domain event extraction. For domain specific extraction, approaches mainly focus on extracting a particular type of events, including natural disasters (Sakaki et al., 2010), traffic events (Dabiri and Heaslip, 2019), user mobility behaviors (Yuan et al., 2013), and etc. The open domain scenario is more challenging and usually relies on unsupervised approaches. Existing works usually create clusters with event-related keywords (Parikh and Karlapalem, 2013), or named entities (McMinn and Jose, 2015; Edouard et al., 2017). Additionally, Ritter et al. (2012) and Zhou et al. (2015) design general pipelines to extract and categorize events in supervised and unsupervised manner respectively.

Different from previous works, we deal with COVID-19 related event extraction in particular. Zong et al. (2020) provide a BERT baseline for the same task. But we create a unified framework to learn simultaneously for different categories of events and subtasks.

Type-aware Slot Filling Yang et al. (2016) formulate entity type constraints and use integer linear programming to combine them with relation classification. Adel and Schütze (2019) propose to in-

tegrate entity and relation classes in convolutional neural networks and learn the correlation from data. We propose a NER-based post-processing technique for type-aware slot filling. By filtering out entity mis-matched predictions, JOELIN can efficiently boost the performance with minimum hand-crafted rules.

COVID-19 Twitter Analysis With the quarantine situation, people can share thoughts and make comments about COVID-19 on Twitter. It has become a research source for researchers to explore and study. Singh et al. (2020) show that Twitter conversations indicate a spatio-temporal relationship between information flow and new cases of COVID-19. There is some work about COVID-19 datasets. Banda et al. (2020) provide a large-scale curated dataset of over 152 million tweets. Chen et al. (2020) collect tweets and forms a multilingual COVID-19 Twitter dataset. Based on the collected data, Jahanbin and Rahmanian (2020) propose a model to predict COVID-19 breakout by monitoring and tracking information on Twitter. Though there are some works about COVID-19 tweets analysis (Müller et al., 2020; Jimenez-Sotomayor et al., 2020; Lopez et al., 2020), the work about automatically extracting structured knowledge of COVID-19 events from tweets is still limited.

3 Method

In this section, we introduce our approach JOELIN and its data pre-processing and post-processing steps in detail. First, we pre-process the noisy Twitter data following the data cleaning procedures in Müller et al. (2020). Second, we train JOELIN and fine-tune the pre-trained language model end-to-end. Specifically, we design the JOELIN classifier in a joint event multi-task learning framework. Moreover, we provide four options of embedding types and ensemble the outputs with the highest validation score. Finally, we further utilize NER techniques to post-process our results with minimum hand-crafted rules.

3.1 Data Pre-processing

Prior to training, the original tweets are cleaned following Müller et al. (2020). The punctuations are standardized and unicode emoticons are expanded into textual ASCII representations². All Twitter usernames are replaced with a special token <USER> for pseudonymisation, URLs with

²[https://pypi.org/project/emoj/](https://pypi.org/project/emoji/)

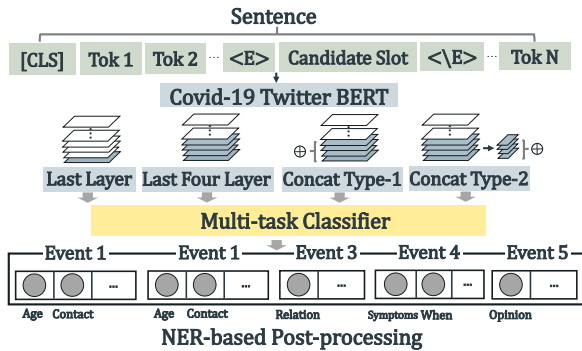


Figure 2: Our approach comprises of 2 main components: (1) global language model across events and sub-tasks; (2) multi-task learning classifier.

<URL>, and COVID-19 related tags, such as #COVID19, #coronavirus, #COVID etc., with <COVID_TAG>. Note that the data cleaning step is designed as a hyper-parameter and can be on or off during the experiments.

We construct the training instance as follows. The annotated data is a collection of tweets. Each tweet is accompanied by hand-labeled candidate chunks. Each candidate chunk is extracted and sandwiched by a pair of tokens <E> and </E>. The masked text, together with the annotated label, will then serve as one instance of the input.

3.2 The JOELIN Model

JOELIN consists of four modules as shown in Figure 2: the pre-trained COVID Twitter BERT (CT-BERT) (Müller et al., 2020), four different embedding layers, joint event multi-task learning framework with global parameter sharing, and the output ensemble module.

COVID Twitter BERT It has been a common practice that pre-trained language models, e.g., BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), are used for a supervised fine-tuning for specific downstream tasks. In this work, we use CT-BERT as JOELIN pre-trained language model. The CT-BERT is trained on a corpus of 160M tweets related to COVID-19. CT-BERT shows great improvement compared to BERT-LARGE and RoBERTa. We further fine-tune CT-BERT with the provided dataset.

Feature Extraction With the hidden representation of token <E> given by CT-BERT, we further apply various choices of different feature extraction methods to choose the more useful features. Inspired by Devlin et al. (2018), we implemented

the following four feature extraction methods:

1. *Last hidden layer*: we directly use the last hidden layer of CT-BERT as our classifier input.
2. *Summation of last four*: we sum the last four hidden layer outputs as the classifier input.
3. *Concatenation of last four (type-1)*: we directly concatenate the last four layers, and flatten the vector before feeding it to the classifier.
4. *Concatenation of last four (type-2)*: Each of last four layers is passed through a fully-connected layer and reduced to a quarter of its original hidden size. We flatten the vectors before passing through the classifier.

Joint Event Multi-task Learning To tackle the challenge of limited annotated data, we apply a global parameter sharing model across all events. Specifically, we jointly learn and fine-tune the language embedding across different events and apply a multi-task classifier for prediction. As shown in Figure 2, the language embedding as well as the feature extraction mechanism are jointly learned and fine-tuned globally. We then apply a fully-connected layer as our classifier for all the sub-tasks in different categories of events. In this way, JOELIN benefits from using data of all the events and their subtasks. Compared with training separate models for each event, joint training across different tasks significantly boosts the performance.

Model Ensemble It has long been observed that ensembles of models boost overall performance. Hence, in this work, we train multiple models with different feature extraction approaches, and we select the top 5 models with best performance and ensemble them by majority voting. Note that we choose the best model for each subtask in validation. In testing, we use the average of the top 5.

3.3 NER-based Post-processing

We further filter our prediction based on NER for post-processing. Specifically, we use spaCy’s NER model³ to tag the predicted candidate slots. We collect all the tags of each word in candidate slots. Then we compare the entity tags with the subtask. If one of the candidate tags does not match the subtask type, we invalidate the prediction by replacing it with “NOT SPECIFIED”. For example, if the subtask is “who”, we nullify those candidate slots whose tags are not related to persons, as shown in Figure 1.

³<https://spacy.io/>

4 Experiments and Analysis

4.1 Dataset

The dataset⁴ is composed of annotated tweets sampled from January 15, 2020 to April 26, 2020. It contains 7,500 tweets for the following 5 events: (1) tested positive, (2) tested negative, (3) can not test, (4) death, and (5) cure and prevention. Each event contains several slot subtasks.

4.2 Implementation Details

We randomly split the dataset into training and validation in a 80:20 ratio. The model is trained with the AdamW optimizer (Loshchilov and Hutter, 2017) toward minimizing the binary cross entropy loss with batch size of 32 and learning rate of $2e-5$. To deal with the class imbalance issue, we apply class weighting on the loss function. With grid-search, the best weight is 10 and 1 for positive and negative samples respectively.

4.3 Results and Discussion

We evaluate JOELIN with BERT and CT-BERT baselines. We measure the performance of different models with F1 score and micro F1 score, in consideration of imbalanced sample sizes. The overall results are shown in Table 1. Compared with the performance of BERT (Zong et al., 2020) and CT-BERT (Müller et al., 2020), JOELIN significantly outperforms the best baseline CT-BERT by 7.6% in micro F1. In terms of performance on subtasks, JOELIN outperforms the best baseline CT-BERT by up to 44.9% in recent travel of event TESTED POSITIVE. The performance gains of JOELIN are attributed to the well-designed joint event multi-task learning framework and the type-aware NER-based post-processing.

4.4 Ablation Study

We conduct an ablation study to understand the contribution of type-aware post-processing in JOELIN. We remove the post-processing step as a reduced model (JOELIN-P) and compare the micro F1 scores. As shown in Table 2, JOELIN has better micro F1 score in comparison with the reduced model JOELIN-P. It supports the claim that our proposed type-aware post processing with NER can significantly boost the performance.

5 Conclusion

In this work, we build JOELIN upon a joint event multi-task learning framework. We use NER-based

⁴https://github.com/viczong/extract_COVID19_events_from_Twitter

Sub-task	BERT	CT-BERT	JOELIN
TESTED POSITIVE			
age	0.519	0.571	0.769
close_contact	0.262	0.333	0.420
employer	0.394	0.391	0.453
gender_male	0.664	0.669	0.711
gender_female	0.635	0.698	0.779
name	0.740	0.774	0.807
recent_travel	0.227	0.391	0.567
relation	0.476	0.621	0.769
when	0.571	0.571	0.741
where	0.560	0.631	0.660
TESTED NEGATIVE			
age	0.000	0.750	0.750
close_contact	0.000	0.133	0.133
gender_male	0.479	0.660	0.706
gender_female	0.214	0.649	0.766
how_long	0.000	0.400	0.800
name	0.519	0.646	0.675
relation	0.449	0.720	0.784
when	0.000	0.471	0.471
where	0.372	0.578	0.651
CAN NOT TEST			
relation	0.516	0.608	0.771
symptoms	0.517	0.704	0.757
name	0.382	0.545	0.550
when	0.000	0.000	0.000
where	0.509	0.500	0.638
DEATH			
age	0.727	0.722	0.789
name	0.642	0.715	0.774
relation	0.378	0.646	0.680
symptoms	0.000	0.000	0.444
when	0.633	0.605	0.690
where	0.483	0.613	0.628
CURE AND PREVENTION			
opinion	0.520	0.573	0.627
what_cure	0.583	0.671	0.671
who_cure	0.389	0.515	0.545
micro avg. F1	0.576	0.647	0.696

Table 1: Overall performance of JOELIN compared with BERT and CT-BERT on validation data. The results are reported with F1 score. For JOELIN, we report the best of all subtask results.

Model	Micro F1
JOELIN-P	0.488
JOELIN	0.511

Table 2: Ablation model comparison on test data.

post-processing to generate type-aware predictions. The results show JOELIN significantly boosts the performance of extracting COVID-19 events from noisy tweets over BERT and CT-BERT baselines. In the future, we would like to extend JOELIN to open domain event extraction tasks, which is more challenging and requires a more general pipeline.

References

- Heike Adel and Hinrich Schütze. 2019. Type-aware convolutional neural networks for slot filling. *Journal of Artificial Intelligence Research*, 66:297–339.
- Juan M Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, and Gerardo Chowell. 2020. A large-scale covid-19 twitter chatter dataset for open scientific research—an international collaboration. *arXiv preprint arXiv:2004.03688*.
- Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Covid-19: The first public coronavirus twitter dataset. *arXiv preprint arXiv:2003.07372*.
- Sina Dabiri and Kevin Heaslip. 2019. Developing a twitter-based traffic event detection model using deep learning architectures. *Expert systems with applications*, 118:425–439.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Amosse Edouard, Elena Cabrio, Sara Tonelli, and Nhan Le Thanh. 2017. Graph-based event extraction from twitter.
- Tim Finin, William Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, pages 80–88.
- Wei-Ning Hsu, Yu Zhang, and James Glass. 2017. Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 16–23. IEEE.
- Kia Jahanbin and Vahid Rahmanian. 2020. Using twitter and web news mining to predict covid-19 outbreak. *Asian Pacific Journal of Tropical Medicine*, 13.
- Maria Renee Jimenez-Sotomayor, Carolina Gomez-Moreno, and Enrique Soto-Perez-de Celis. 2020. Coronavirus, ageism, and twitter: An evaluation of tweets about older adults and covid-19. *Journal of the American Geriatrics Society*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Christian E Lopez, Malolan Vasu, and Caleb Gallemore. 2020. Understanding the perception of covid-19 policies by mining a multilanguage twitter dataset. *arXiv preprint arXiv:2003.10359*.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Andrew J McMinn and Joemon M Jose. 2015. Real-time entity-based event detection for twitter. In *International conference of the cross-language evaluation forum for european languages*, pages 65–77. Springer.
- Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.
- Ruchi Parikh and Kamalakar Karlapalem. 2013. Et: events from tweets. In *Proceedings of the 22nd international conference on world wide web*, pages 613–620.
- Shiva Pentylala, Mengwen Liu, and Markus Dreyer. 2019. Multi-task networks with universe, group, and task feature learning. *arXiv preprint arXiv:1907.01791*.
- Martin Potthast, Tim Gollub, Kristof Komlossy, Sebastian Schuster, Matti Wiegmann, Erika Patricia Garces Fernandez, Matthias Hagen, and Benno Stein. 2018. Crowdsourcing a large corpus of clickbait on twitter. In *Proceedings of the 27th international conference on computational linguistics*, pages 1498–1507.
- Alan Ritter, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112.
- Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860.
- Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga, and Yanchen Wang. 2020. A first look at covid-19 information and misinformation sharing on twitter. *arXiv preprint arXiv:2003.13907*.
- Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*.
- Bishan Yang, Ndapandula Nakashole, Bryan Kisiel, Emmanouil A Platanios, Abulhair Saparov, Shashank Srivastava, Derry Wijaya, and Tom M Mitchell. 2016. Cmuml micro-reader system for kbp 2016 cold start slot filling, event nugget detection, and event argument linking. In *TAC*.

Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat Thalmann. 2013. Who, where, when and what: discover spatio-temporal topics for twitter users. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 605–613.

Yu Zhang and Qiang Yang. 2017. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*.

Deyu Zhou, Liangyu Chen, and Yulan He. 2015. An unsupervised framework of exploring events on twitter: Filtering, extraction and categorization. In *Twenty-ninth aaii conference on artificial intelligence*.

Shi Zong, Ashutosh Baheti, Wei Xu, and Alan Ritter. 2020. [Extracting covid-19 events from twitter](#).