METS-CoV: A Dataset of Medical Entity and Targeted Sentiment on COVID-19 Related Tweets

Peilin Zhou[1,2] **Zeqiang Wang**[1,2] **Dading Chong**[3] **Zhijiang Guo**[4]
Yining Hua[5] **Zichang Su**[2,6] **Zhiyang Teng**[7] **Jiageng Wu**[1,2] **Jie Yang**[1,2*]
[1]School of Public Health and the Second Affiliated Hospital, Zhejiang University, China
[2]The Key Laboratory of Intelligent Preventive Medicine of Zhejiang Province, China
[3]School of Electronic and Computer Engineering, Peking University, China
[4]Department of Computer Science and Technology, University of Cambridge, UK
[5]Department of Biomedical Informatics, Harvard Medical School, USA
[6]Chu Kochen Honors College, Zhejiang University, China
[7]School of Engineering, Westlake University, China
{zhoupalin, jieynlp}@gmail.com, {wzq99, suzc,jiagengwu}@zju.edu.cn
1601213984@pku.edu.cn, tengzhiyang@westlake.edu.cn
zg283@cam.ac.uk, yining_hua@hms.harvard.edu

A Appendix

A.1 Detailed Statistics of METS-CoV-NER Dataset

Supplement Table 1 shows the detailed statistics of METS-CoV-NER dataset, including the number of tweets, the number of unique entity mentions and the number of unseen entity mentions for each entity type. Note that the entity mentions in METS-CoV-NER dataset are all lower-cased when calculating data statistics.

Supplement Table 1: Detailed Statistics of METS-CoV-NER.

Entity Type	Number of Tweets			Number of Unique Entity Mentions			Number of Unseen Entity Mentions	
	Train	Dev	Test	Train	Dev	Test	Dev	Test
Person	1,537	343	330	1,350	330	328	257	257
Location	932	198	196	546	180	166	95	85
Organization	1,458	295	293	1,025	259	275	171	190
Disease	969	200	182	358	102	109	43	46
Drug	772	175	170	374	128	110	69	56
Symptom	2,446	508	531	1,098	292	338	136	178
Vaccine	1,018	189	209	273	77	93	35	41

A.2 Training Details

For NER benchmarking, all the experiments of NER models are conducted using NCRF++ (Yang and Zhang, 2018). The tag scheme for NER is BIOES. All PLM-based NER methods are trained using an AdamW optimizer with the initial learning rate of 0.00003 for 100 epochs. For word LSTM models, GloVe 100-dimension (Pennington et al., 2014) is used to initialize word embeddings and character embeddings are randomly initialized. We use a mini-batch stochastic gradient descent (SGD) optimizer with a decayed learning rate to update parameters. The main hyperparameters of

*Corresponding Author.

Submitted to the 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks. Do not distribute.

Supplement Table 2: Hyperparameters of NER models. (* means uncased model)

Model	epoch	batch_size	optimizer	learning_rate
WLSTM	100	16	SGD	1
WLSTM + CCNN	100	16	SGD	1
WLSTM + CLSTM	100	16	SGD	1
WLSTM + CRF	100	16	SGD	0.015
WLSTM + CCNN + CRF	100	16	Adadelta	0.1
WLSTM + CLSTM + CRF	100	16	SGD	0.015
BERT-base*	100	32	AdamW	0.00003
BERT-base	100	32	AdamW	0.00003
BERT-large*	100	32	AdamW	0.00003
BERT-large	100	32	AdamW	0.00003
RoBERTa-base	100	32	AdamW	0.00003
RoBERTa-large	100	32	AdamW	0.00003
BART-base	100	32	AdamW	0.00003
BART-large	100	32	AdamW	0.00003
BERTweet-covid19-base*	100	32	AdamW	0.00003
BERTweet-covid19-base	100	32	AdamW	0.00003
COVID-TWITTER-BERT*	100	16	AdamW	0.00003

Supplement Table 3: Hyperparameters of TSA models.

Model	epoch	batch_size	optimizer	learning_rate
LSTM	40	16	Adam	0.001
TD-LSTM	40	16	Adam	0.001
MemNet	40	16	Adam	0.001
IAN	40	16	Adam	0.001
MGAN	40	16	Adam	0.001
TNet-LF	40	16	Adam	0.001
ASGCN	40	16	Adam	0.001
AEN	40	16	Adam	0.00002
LCF	40	16	Adam	0.00002
BERT-SPC	10	16	Adam	0.00002
depGCN	10	16	Adam	0.00002
kumaGCN	10	16	Adam	0.00002
dotGCN	10	16	Adam	0.00002
BERT-SPC(COVID-TWIITER-BERT)	10	16	Adam	0.00002
depGCNC(COVID-TWIITER-BERT)	10	16	Adam	0.00002
kumaGCNC(COVID-TWIITER-BERT)	10	16	Adam	0.00002
dotGCNC(COVID-TWIITER-BERT)	10	16	Adam	0.00002

NER models are shown in Supplement Table 2 Other hyperparamteres, such as hidden dimension, dropout rate and etc, are consistent with Yang and Zhang (2018).

For TSA benchmarking, we conducted all the experiments using the codes released by the authors. And the hyperparamters of TSA models are summarized in Supplement Table 3.

We reported the results based on experiments on 5 different random seeds: 22, 32, 42, 52 and 62. Mean ± std were reported in this paper. More detailed training information is available in `https://github.com/YLab-Open/METS-CoV/`.

A.3 Dataset documentation and intended uses

The following questions are copied from "Datasheets for Datasets" (Gebru et al., 2021).

A.3.1 Motivation

- **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

  In order to explore the impact of pandemics on people's lives, it is crucial to understand the public's concerns and attitudes towards pandemic-related entities (e.g., drugs, vaccines) on social media. However, models trained on existing named entity recognition (NER) or targeted sentiment analysis (TSA) datasets have limited ability to understand COVID-19-related social media texts because these datasets are not designed or annotated from a medical perspective. In this work, we release **METS-CoV** (**M**edical **E**ntities and **T**argeted **S**entiments on **CoV**id-19-related tweets), a dataset that contains 10,000 tweets annotated with 7 types of entities, including 4 medical entity types (*Disease*, *Drug*, *Symptom*, and *Vaccine*) and 3 general entity types (*Person*, *Location*, and *Organization*). In addition, 4 types of entities (*Person*, *Organization*, *Drug*, and *Vaccine*) are selected and annotated with user sentiments to explore the attitudes of tweet users toward specific entities. Unlike other general NER and TSA datasets, METS-CoV is built from a public health research perspective and can contribute to developing NLP tools in the medical domain. Based on METS-CoV, we also benchmark the performance of classical machine learning models and state-of-the-art deep learning models (including pretraining language models) on NER and TSA tasks.

- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

  YLab, an interdisciplinary research group at Zhejiang University(`https://ylab.top`).

- **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

  No fund and out of research interest.

A.3.2 Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

  COVID-19 related tweets as well as their annotations including medical entities and targeted sentiment. To avoid the privacy issue, the user profiles were removed before annotation.

- **How many instances are there in total (of each type, if appropriate)?**

  This dataset contains 10,000 COVID-19 related English tweets.

- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

  This dataset is a sample from a larger set. Details could be checked in the Sec 3.1 and Sec 3.2 of our submitted paper.

- **What data does each instance consist of?** "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

  Each instance has pre-processed text of tweets and annotations including entities and targeted sentiment.

- **Is there a label or target associated with each instance?** If so, please provide a description.

  This dataset is annotated with 7 types of entities, including 4 medical entity types (*Disease*, *Drug*, *Symptom*, and *Vaccine*) and 3 general entity types (*Person*, *Location*, and *Organization*). In addition, 4 types of entities (*Person*, *Organization*, *Drug*, and *Vaccine*) are selected

and annotated with user sentiments to explore the attitudes of tweet users toward specific entities.

- **Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

  No.

- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.

  No.

- **Are there recommended data splits (e.g., training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.

  Yes. We perform a train-dev-test splitting of our dataset with a ratio of 70:15:15.

- **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.

  There might be some annotation errors in the dataset, as it is inevitable that the targeted sentiment annotations contain the subjectivity of the annotators. To reduce subjectivity, we make strict annotation guidelines, conduct multi-rounds of pre-annotation training, and have annotators work in pairs with a third-party validation step.

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

  Following Twitter's automation rules and data security policy, we can not directly provide the original tweets to dataset consumer. Therefore, We will release the Tweet IDs and corresponding annotations. Based on the Tweet IDs provided by us, the dataset consumers could download the original tweets freely via official Twitter API by themselves.

- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?** If so, please provide a description.

  No, all these tweets are public available.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.

  Directly removing offensive words might have negative impacts on the performance of TSA models and even misguide the medical research. Because these words are usually helpful to judge the real sentiment of users towards targeted entities. Therefore, all the tweets maintain their original setting in our dataset.

If the dataset does not relate to people, you may skip the remaining questions in this section.

- **Does the dataset identify any subpopulations (e.g., by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

  No.

- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** If so, please describe how.

  No.

- **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** If so, please provide a description.

  No.

A.3.3 Collection Process

- **How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

  Please check Sec 3.1 in the main paper.

- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** How were these mechanisms or procedures validated?

  All the tweets were collected via Twitter's official API.

- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

  Please check Sec 3.1 and 3.2 in the main paper.

- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

  This dataset was voluntarily annotated by the authors and members of Prof. Jie Yang's group.

- **Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

  We collect COVID-19 related tweets ranging from February 1, 2020, to September 30, 2021.

- **Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

  Yes. We have conducted an internal ethical review process by the Zhejiang University ethical team.

If the dataset does not relate to people, you may skip the remaining questions in this section.

- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

  All the tweets were collected via Twitter's official API.

- **Were the individuals in question notified about the data collection?** If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

  No. All these tweets are public available.

- **Did the individuals in question consent to the collection and use of their data?** If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

  No. All these tweets are public available and thus users' consent could be waived.

- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

  N/A.

- **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

  No.

A.3.4 Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** If so, please provide a description. If not, you may skip the remaining questions in this section.

  We removed the HTML tags in the tweets because they are meaningless for medical research. Following (Müller et al., 2020), we replaced all unicode emoticons with textual ASCII representations using the Python emoji library.

- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** If so, please provide a link or other access point to the "raw" data.

  N/A.

- **Is the software that was used to preprocess/clean/label the data available?** If so, please provide a link or other access point.

  Yes. All the annotation work is done using the YEDDA annotation platform by Yang et al. (2018). The link of YEDDA is as follows: `https://github.com/jiesutd/YEDDA`. Screenshots are shown in Figure 1 and Figure 2.

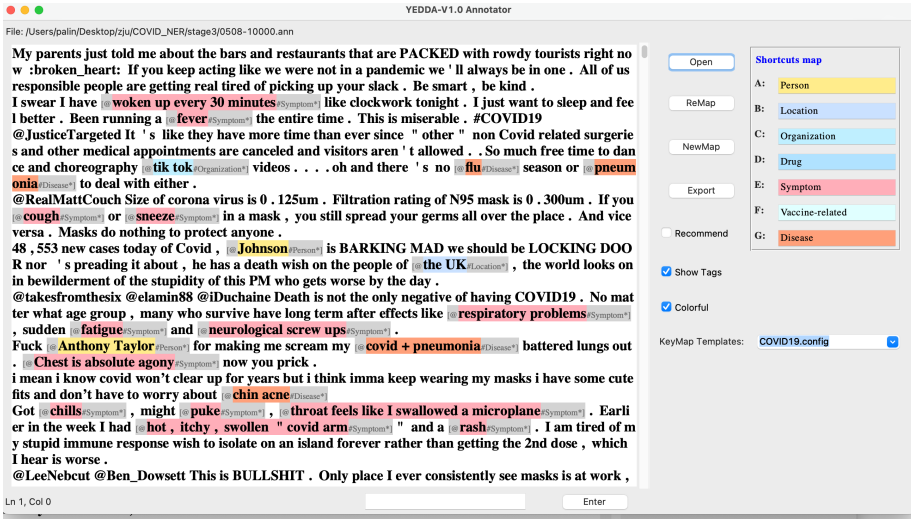

Figure 1: The screenshot of NER annotation using YEDDA platform.

A.3.5 Uses

- **Has the dataset been used for any tasks already?** If so, please provide a description.

  The METS-CoV is a new dataset to collect medical entities and corresponding sentiments of COVID-19 related tweets. We benchmark the performance of classical machine learning

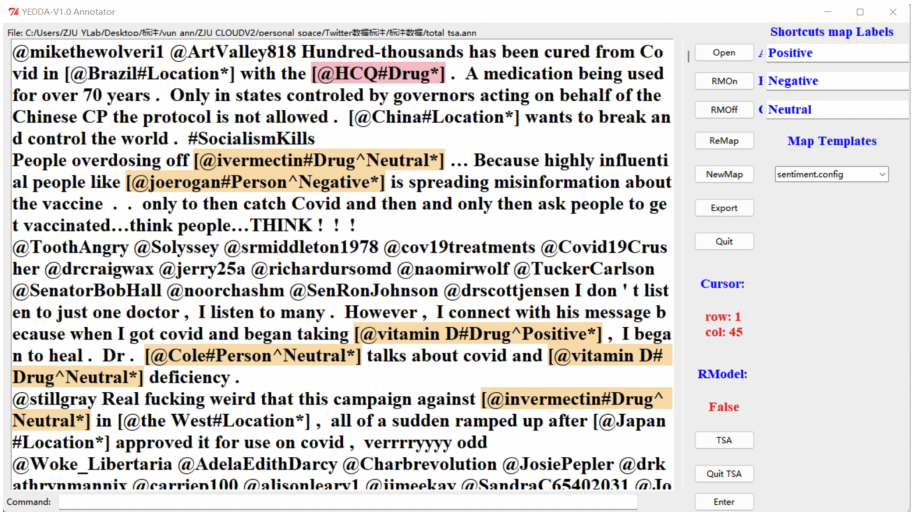

Figure 2: The screenshot of TSA annotation using YEDDA platform.

models and state-of-the-art deep learning models on NER and TSA tasks with extensive experiments.

- **Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.

  Our data, annotation guidelines, benchmark models, and source code are publicly available (`https://github.com/YLab-Open/METS-CoV`) to ensure reproducibility.

- **What (other) tasks could the dataset be used for?**

  METS-CoV cound be used for the joint learning of NER and TSA tasks, and supporting other medical analysis on social media.

- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

  No.

- **Are there tasks for which the dataset should not be used?** If so, please provide a description.

  Could not be used to predict individual health risks.

A.3.6 Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.

  No.

- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?

  It is released on Github at `https://github.com/YLab-Open/METS-CoV`. No DOI.

- **When will the dataset be distributed?**

  Before the conference.

- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license

and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Apache License 2.0. (`https://github.com/YLab-Open/METS-CoV/blob/main/LICENSE`)

- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

  No.

- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

  No.

A.3.7 Maintenance

- **Who will be supporting/hosting/maintaining the dataset?**

  YLab, an interdisciplinary research group at Zhejiang University(`https://ylab.top`).

- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

  Peilin Zhou's email address: zhoupalin@gmail.com; Professor Jie Yang's email address: jieynlp@gmail.com.

- **Is there an erratum?** If so, please provide a link or other access point.

  No.

- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

  No. If we plan to update the dataset in the future, we will elaborate the reason on our GitHub repository.

- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.

  Yes. The limits do exist because Twitter wanted to guarantee the users the "right to be forgotten." When the user decides to hide or delete a tweet, we should not be able to further access it. Previous Twitter datasets generally set up a provided-upon-request convention: if the tweet ID can no longer be hydrated, users can send emails to the dataset development team requesting the desensitized version of the tweets. We also follow this convention.

- **Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

  Yes. If we plan to update the data, we will maintain the old version and then release the follow-up version, for example, METS-CoV-2.0

- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

  Yes. For data annotation, researchers could carefully check our annotation guidelines to ensure the consistency. And if others want to contribute to the dataset, they could submit a pull request or contact us via email.

A.4 Accessibility

1. Links to access the dataset and its metadata. (`https://github.com/YLab-Open/METS-CoV`)
2. The data is saved in a CSV format, where an example is shown in the README.md file.
3. YLab research group will maintain this dataset on the official Github account.
4. Apache License 2.0. (`https://github.com/YLab-Open/METS-CoV/blob/main/LICENSE`)

A.5 Data Usage

The authors bear all responsibility in case of violation of rights.

References

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

Martin Müller, Marcel Salathé, and Per Egil Kummervold. 2020. COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. *CoRR* abs/2005.07503 (2020). arXiv:2005.07503 `https://arxiv.org/abs/2005.07503`

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. `http://www.aclweb.org/anthology/D14-1162`

Jie Yang and Yue Zhang. 2018. NCRF++: An Open-source Neural Sequence Labeling Toolkit. In *Proceedings of ACL 2018, System Demonstrations*. Association for Computational Linguistics, Melbourne, Australia, 74–79. `https://doi.org/10.18653/v1/P18-4013`

Jie Yang, Yue Zhang, Linwei Li, and Xingxuan Li. 2018. YEDDA: A Lightweight Collaborative Text Span Annotation Tool. In *Proceedings of ACL 2018, System Demonstrations*. Association for Computational Linguistics, Melbourne, Australia, 31–36. `https://doi.org/10.18653/v1/P18-4006`