Political Event Coding as Text-to-Text Sequence Generation

Yaoyao Dai and **Benjamin J. Radford**
University of North Carolina at Charlotte
benjamin.radford@uncc.edu

Andrew Halterman
Michigan State University
halterm3@msu.edu

Abstract

We report on the current status of an effort to produce political event data from unstructured text via a Transformer language model. Compelled by the current lack of publicly available and up-to-date event coding software, we seek to train a model that can produce structured political event records at the sentence level. Our approach differs from previous efforts in that we conceptualize this task as one of text-to-text sequence generation. We motivate this choice by outlining desirable properties of text generation models for the needs of event coding. To overcome the lack of sufficient training data, we also describe a method for generating synthetic text and event record pairs that we use to fit our model.

1 Introduction

Political event records that are automatically derived from text are an important source of structured data for researchers in social science. Existing approaches to generating event data often rely on dictionary methods, which are brittle and go obsolete, or classifiers trained on hand-labeled text that required large amounts of expensive data. This paper introduces a new proof of concept model for generating structured event data from news text that does not require dictionaries or hand-labeled document. We generate synthetic news stories using a novel combination of rules based generation and a paraphrasing model and train a text-to-text Transformer to produce event records from text. When evaluated on synthetic test data, the model correctly identifies high-level event types 83% of the time and reaches accuracies of 63% and 56.9% on the source and target actors, respectively. The article concludes with a brief real-world evaluation and a discussion of the model's limitations.

1.1 Political Event Data

Political event data describe who did what to whom and, usually, where and when that action occurred. While the actors and actions themselves are derived directly from source texts, locations and times are often determined via the textual metadata. Therefore, the core component of political event data is the source actor, target actor, action three-tuple.

Precisely what actors and actions are included in an event dataset varies; some tend to be specific to certain classes of events while others seek to capture the full range of politically-relevant interactions. Examples of the latter include the Global Database of Events, Language, and Tone (GDELT)[1], the Integrated Crisis Early Warning System (ICEWS)[2], and the various Phoenix datasets.[3]

1.2 Generating Event Data

Historically political event data have been made by hand (Azar, 1980; McClelland, 2006), by rules-based software tools (Schrodt, 1998, 2001, 2011; Schrodt et al., 2014; Norris et al., 2017), and via machine learning. Rules based software typically relies on large hand-curated dictionaries to perform pattern matching. These dictionaries will conform to a predetermined event ontology like that defined by CAMEO, the Conflict and Mediation Event Observations (Schrodt, 2012). Neural networks have been used in conjunction with PETRARCH, one rules based coding software, to generate event data by (Radford, 2021a) and to classify events into quad-class categories by (Beieler, 2016). Recently, workshops like ProtestNews at CLEF 2019, AESPEN at LREC 2020, and CASE have prompted even more work on machine learning approaches to coding event data from text (Hürriyetoğlu et al., 2020; Hürriyetoğlu et al., 2020;

[1] https://gdeltproject.org
[2] https://dataverse.harvard.edu/dataverse/icews
[3] https://clinecenter.illinois.edu/project/machine-generated-event-data-projects/phoenix-data

Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), pages 117 - 123
December 7-8, 2022

Hürriyetoğlu, 2021). Transformer models were used for zero-shot classification of previously unseen event types and for cross-context and multilingual protest detection (Haneczok et al., 2021; Barker et al., 2021; Kent and Krumbiegel, 2021; Radford, 2021b). These zero-shot methods rely primarily on textual entailment formulations of the event data coding task (Yin et al., 2019).

2 Methodology

2.1 Training Data

The ideal training data for our model would be a dataset of source texts and their associated coded events, produced by an existing event coder. However, due to copyright restrictions, there are no publicly available large scale event datasets that include event's associated source texts. We therefore propose generating synthetic news stories from coded events. The use of synthetic text is growing in political science (Halterman, 2022), but we introduce a novel technique using a combination of rule-based generation and a paraphrase model to generate synthetic text that contains the content we require. To generate positive examples, we generate synthetic stories through a rule-based process. We parse the CAMEO and Petrarch dictionaries available from the Open Event Data Alliance and piece together pseudo-sentences by substituting random actors, agents, and word synonyms in the placeholders denoted within the CAMEO verbs dictionary.

To ensure that our model learns to refrain from returning coded events when no event is reported, we also include negative samples, drawn randomly from sentences published in the *New York Times* (NYT) between 1970 and 2022 and assumed to have no event present.[4] From these two sources, we draw 4.08 million samples (4,000,000 training, 40,000 validation, and 40,000 test set) with a ratio of 39 positive to 1 negative. One sample represents approximately a single sentence.

Because the heuristic approach to generating positive examples often results in bizarre, poorly-formed, and repetitive sentences, we post-process 50% of all samples (positive and negative) by running them through a Transformer model for paraphrase generation.[5] This model attempts to output a sentence that is not identical to, but semantically similar to, an input sentence. The paraphraser is set by default to produce a sentence of no more than 30 tokens in length.[6] Unfortunately, this induces a bias in our model towards coding shorter and simpler sentences than are typical for new text and we intend to adjust the paraphraser parameters in future iterations.[7] Our target values are comma-delimited three-tuples of action category, source actor, and target actor. An example of a raw synthetic story, a paraphrased story, and the associated event code is given below.

Raw synthetic story: "Ministries For Public Health And Social Welfare Rossija said could beat Jibouti within Unmanned Aircraft."

Paraphrased text: "Rossija said that he could beat Jibouti with Unmanned Aircraft".[8]

Event record: 138 (Threaten with military force), RUSGOVHLH (Russian government healthcare), DJIMIL (Djibouti military).

In the raw synthetic story, the randomly-drawn actors are "Ministries for Public Health and Social Welfare Rossija," and "Jibouti," the verb phrase is "[SOURCE ACTOR] said could [VERB] [TARGET ACTOR]," the verb is "beat," and "Unmanned Aircraft" is a synonym for "aircraft." "Within" is one of a set of available random prepositions.

To minimize the leakage of actors or phrases from the training set into the (out-of-sample) validation or test sets, we partition the dictionaries prior to generating synthetic samples. Specifically, we partition the NYT sentences, verb phrases, agents, countries, and actors into their own training, validation, and test sets prior to constructing our three respective data partitions. We then construct synthetic samples for each of the training, validation, and test sets by sampling only from those words and phrases found in the corresponding partitions

[4]It is likely that these sentences from the NYT contain a small number of relevant socio-political events. We have not attempted to remove these false negatives from the corpus and therefore admit that the negative examples in our training data likely contain a small proportion of errors.

[5]`https://huggingface.co/ramsrigouthamg/t5_sentence_paraphraser`

[6]This is likely too short and we recommend longer maximum sequence lengths be used in future work. However, producing longer paraphrased sentences requires greater computational resources or computation time than were available for this study. We therefore leave the paraphraser set to the default 30 tokens maximum output.

[7]An open-ended text generation model like GPT-2, applied after the paraphraser, could expand the paraphrase in such a way that results better simulate the target distribution of news texts (Radford et al., 2019).

[8]It is possible that the paraphraser model changes the content of some texts such that they no longer correspond to the codes associated with their original associated synthetic event records. Nonetheless, paraphrase-based data augmentation is becoming common in NLP applications (Kumar et al., 2019; Corbeil and Abdi Ghavidel, 2020; Beddiar et al., 2021).

of dictionaries. Synsets, words that are effectively synonyms of one another, are not partitioned in such a way.

2.2 Model

Text-to-Text Transfer Transformer (T5) is a language model tailored for text generation (Raffel et al., 2020). It comes in a variety of sizes, of which we select T5-Base version 1.1 with 250m parameters. T5 was trained on a variety of natural language tasks, distinguished by prepending a keyword describing the task to the input ("context") to the model. We fine tune T5 on our synthetic dataset for a single epoch with a learning rate of 5.6×10^{-6} and all other hyperparameters held at their default values. We decode our output using the default configuration for T5 in HuggingFace's pipeline (greedy search).[9] Alternative configurations may lead to different output values.

Continuing with the example from Section 2.1, the input to our model would be the Sentencepiece-tokenized version of either the raw synthetic story or the paraphrased story, drawn with equal probability (50% each) (Kudo and Richardson, 2018). The desired output of the model is the comma-delimited, semicolon-terminated event record "138, RUSGOVHLH, DJIMIL;".

3 Results

This section provides descriptions of our results in two experimental settings: an out-of-sample test set evaluation using data generated via the same process as the training data and an out-of-distribution case study using a small sample of real world news text.

3.1 Within Distribution Performance

We evaluate the test set accuracy of the model on the event category, actor coding, and exact match accuracy on the full event triple. At the coarser level of 20 event types, the model achieves 83.4% accuracy and reaches 77.8% accuracy for the full set of 295 fine-grained action codes. Source and target actor exact match accuracy are 63% and 56.9%, respectively. Because actors are represented by sets of three-character sub codes, we can compute the precision, recall, and F1-score of these sub codes. We find values of 0.73, 0.78, and 0.75, respectively. Evaluating against the complete event record, our model achieves 30.7% exact match accuracy in the test set.

The model only fails to code events for 15 input samples that contain events and erroneously codes events for 53 samples that should not contain events, corresponding to an F1-score of 0.999. These scores likely reflect the differences in samples generated by our synthetic process versus those drawn from the NYT more so than they do strong model performance. Overall, we find these results promising but acknowledge that synthetic data often fail to sufficiently mimic their real world targets. For this reason, we turn now to a small case study with real world data that are representative of data that would typically be used in event coding applications.

3.2 Real World Performance

While we reserve a full real-world evaluation of our neural event coder for a follow-up paper, here we demonstrate its use in a very short case study: the top ten articles on the Associated Press's World page as retrieved on September 6, 2022. We first attempt to code the introductory sentence from each of the ten articles to no success: not a single sentence produced an event. However, as we noted before, these sentences are far longer and more complex than those generated by our heuristic process. If we first use the paraphraser model to paraphrase these sentences such that they better resemble the distribution of the training data, we find three events.[10] Furthermore, if we code the headlines rather than the introductory sentences, six out of the ten produce event data records. See Table 1 for the headlines and coded events. Most of the event records produced from headlines are at least partially correct. The verb codes correspond to "make pessimistic comment," "threaten with military force," "express intent to cooperate militarily," "praise or endorse," "investigate," and "reduce relations," in order. Example 2 ("UN agency calls for safety zone around Ukraine nuclear plant") was correctly coded as 0256 "appeal for de-escalation of military engagement" when using the paraphrased introductory sentence but incorrectly coded as a threat when using the headline. While the actor countries tend to match those described in the headlines, the model is ambitious about inferring unstated actor affiliations. For instance, in example 9, the target actor ("cabinet") is incorrectly assumed

[9] https://huggingface.co/docs/transformers/main_classes/pipelines

[10] Sentences, paraphrases, and events given in the Appendix.

Headline	Headline Event
1. New UK leader vows to tackle energy crisis, ailing economy	012 (Statement), GBR, NGOENV
2. UN agency calls for safety zone around Ukraine nuclear plant	138 (Threaten), IGOUNODEV, UKRUNR
3. EXPLAINER: Why Truss went to Scotland to become UK leader	–
4. US: Russia to buy rockets, artillery shells from North Korea	0312:0312 (Intent to cooperate), RUS, PRK
5. Rallies show Pakistan's ex-PM Khan remains political force	051 (Diplomatic cooperation), NGOHRI, PAKOPP
6. 'This is it, folks': Boris Johnson bids an ambiguous goodbye	–
7. Fears high as Canadian police search for stabbing suspect	090 (Investigate), CAN, CAN
8. UN: Tribal clashes in Sudan kill 380 in Jan.-Aug. period	–
9. Chile's Boric shakes up cabinet after constitution loss	160 (Reduce relations), CHL, HRVGOV
10. Tension rises as Turkey, Greece voice festering grievances	–

Table 1: World section headlines from the AP on September 6, 2022 and associated predicted event records. Top-level CAMEO action categories are given in parentheses; specific action codes can be found in the CAMEO codebook (Schrodt, 2012).

by the model to be the Herzegovina government. However, sometimes these assumptions are warranted: the model correctly identifies "Pakistan's ex-PM Khan" as a Pakistani opposition figure. In fact, inspection of the CAMEO actors dictionary reveals that Imran Khan is not ever coded as PAKOPP in the dictionary and therefore cannot be coded as such in the training data–this label is inferred by the model entirely out-of-sample.

4 Conclusion

Text-to-text is a flexible modeling task that is amenable to complicated output data types. Using multiple classification heads is an alternative method for event coding text via large language model, but it offers less flexibility for future improvements. For example, a text-to-text model can be trained to generate an arbitrary number of events from a single input text.[11] A more traditional classification-based approach is suitable for coding only up to a predefined number of events. We also appreciate that the open-ended nature of text output means that we do not need to generate all possible actor combinations prior to training as we would in a multiclass classification setup. The text-to-text model can simply append actor codes to one another as necessary, even if it has not previously seen a sample with the particular given combination.

We leave a more formal evaluation of our methodology and model to a follow-up paper in which we plan to employ expert human annotators to generate comparable event data against which we can benchmark our model. Nonetheless, we believe the results presented here are promising for future development of text-to-text models for political event data coding.

4.1 Limitations

We regret that we cannot distribute the entirety of our datasets due to copyright issues. A portion of our samples are drawn from a corpus derived from the *New York Times* and we therefore lack the ability to redistribute them. We do make available the samples generated via our heuristic and paraphrase approach, though. In future iterations of this work, we plan to replace the NYT derived data with samples drawn from open sources.[12]

Our actor resolution step is also limited by our reliance on the existing CAMEO dictionaries and the world knowledge built into T5. Without access to an external data set such as Wikipedia, our accuracy for obscure political entities or people whose roles change frequently will be limited.

Our model exhibits a strange sensitivity to punctuation, especially periods. The model appears to more readily code events when the sentence in question does not end with a period. We have been unable to identify a source of this bias in our training data.

We append a semicolon to the end of every event record. In our next version of this model, we will train on paragraph-length texts and allow the model to output an arbitrary number of semicolon-delimited event records per input example.

We hope to compare our model's performance

[11] We have preliminary work in which we generate up to four events per input paragraph.

[12] For example, we are considering *Voice of America* and Common Crawl as substitute text sources.

directly with that of Petrarch or TABARI. Unfortunately, this will require a functional instance of the software in question which we do not currently possess.

4.2 Broader Impacts

Political event coding software has been publicly available for decades now, as have been the dictionaries of actors and verb phrases that they require. As such, we do not believe that our work poses any additional risk for misuse. Furthermore, we rely on a synthetic data generation technique that allows us to train our model with limited access to real-world text data that may contain sensitive information or reflect undesirable societal biases. As always, we implore others interested in our work to not use it for evil.

References

Edward E. Azar. 1980. The Conflict and Peace Data Bank (COPDAB) Project. *The Journal of Conflict Resolution*, 24(1):143–152.

Ken Barker, Parul Awasthy, Jian Ni, and Radu Florian. 2021. IBM MNLP IE at CASE 2021 task 2: NLI reranking for zero-shot text classification. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 193–202, Online. Association for Computational Linguistics.

Djamila Romaissa Beddiar, Md Saroar Jahan, and Mourad Oussalah. 2021. Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media*, 24:100153.

John Beieler. 2016. Generating politically-relevant event data. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 37–42, Austin, Texas. Association for Computational Linguistics.

Jean-Philippe Corbeil and Hadi Abdi Ghavidel. 2020. Bet: A backtranslation approach for easy data augmentation in transformer-based paraphrase identification context. *arXiv preprint*.

Andrew Halterman. 2022. Synthetically generated text for supervised text analysis. *PolMeth conference paper*.

Jacek Haneczok, Guillaume Jacquet, Jakub Piskorski, and Nicolas Stefanovitch. 2021. Fine-grained event classification in news-like text snippets - shared task 2, CASE 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 179–192, Online. Association for Computational Linguistics.

Ali Hürriyetoğlu, editor. 2021. *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*. Association for Computational Linguistics, Online.

Ali Hürriyetoğlu, Vanni Zavarella, Hristo Tanev, Erdem Yörük, Ali Safaya, and Osman Mutlu. 2020. Automated extraction of socio-political events from news (AESPEN): Workshop and shared task report. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 1–6, Marseille, France. European Language Resources Association (ELRA).

Ali Hürriyetoğlu, Erdem Yörük, Deniz Yüret, Çağrı Yoltar, Burak Gürel, Fırat Duruşan, Osman Mutlu, and Arda Akdemir. 2020. Overview of clef 2019 lab protestnews: Extracting protests from news in a cross-context setting.

Samantha Kent and Theresa Krumbiegel. 2021. CASE 2021 task 2 socio-political fine-grained event classification using fine-tuned RoBERTa document embeddings. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 208–217, Online. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Ashutosh Kumar, Satwik Bhattamishra, Manik Bhandari, and Partha Talukdar. 2019. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3609–3619, Minneapolis, Minnesota. Association for Computational Linguistics.

Charles McClelland. 2006. World Event/Interaction Survey (WEIS) Project, 1966-1978. *Inter-university Consortium for Political and Social Research [Distributor]*.

Clayton Norris, Philip Schrodt, and John Beieler. 2017. Petrarch2: Another event coding program. *Journal of Open Source Software*, 2(9):133.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Benjamin J. Radford. 2021a. Automated dictionary generation for political event coding. *Political Science Research and Methods*, 9(1):157–171.

Benjamin J. Radford. 2021b. CASE 2021 task 2: Zero-shot classification of fine-grained sociopolitical events with transformer models. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 203–207, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Philip A. Schrodt. 1998. KEDS: Kansas Event Data System V 1.0.

Philip A. Schrodt. 2001. Automated Coding of International Event Data Using Sparse Parsing Techniques.

Philip A. Schrodt. 2011. TABARI: Textual Analysis by Augmented Replacement Instructions V. 0.7.6.

Philip A. Schrodt. 2012. CAMEO Conflict and Mediation Event Observations Event and Actor Codebook.

Philip A. Schrodt, John Beieler, and Muhammed Idris. 2014. Three's a Charm?: Open Event Data Coding with EL:DIABLO, PETRARCH, and the Open Event Data Alliance.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3914–3923.

A Appendix

Below are the headlines, introductory paragraphs, and automatically paraphrased paragraphs drawn from the Associated Press World section.

Example 1

Headline: New UK leader vows to tackle energy crisis, ailing economy

First paragraph: "Liz Truss became U.K. prime minister on Tuesday and immediately faced up to the enormous tasks ahead of her: curbing soaring prices, boosting the economy, easing labor unrest and fixing a national health care system burdened by long waiting lists and staff shortages."

Paraphrase: "Liz Truss became the Prime Minister of the United Kingdom on Tuesday and immediately faced the"

Example 2

Headline: UN agency calls for safety zone around Ukraine nuclear plant

First paragraph: "The U.N. atomic watchdog agency urged Russia and Ukraine on Tuesday to establish a "nuclear safety and security protection zone" around the Zaporizhzhia power plant amid mounting fears the fighting could trigger a catastrophe in a country still scarred by the Chernobyl disaster."

Paraphrase: "the United Nations nuclear watchdog group urged Russia and Ukraine to establish a "n"

Paraphrase code: `(0256, IGOUNOKID, RUS)`

Example 3

Headline: EXPLAINER: Why Truss went to Scotland to become UK leader

First paragraph: "Liz Truss, a onetime accountant who has served in Parliament for the past 12 years, became Britain's prime minister on Tuesday after Queen Elizabeth II formally asked her to form a government."

Paraphrase: "Liz Truss, a one-time accountant who has served in Parliament for the"

Example 4

Headline: US: Russia to buy rockets, artillery shells from North Korea

First paragraph: "The Russian Ministry of Defense is in the process of purchasing millions of rockets and artillery shells from North Korea for its ongoing fight in Ukraine, according to a newly downgraded U.S. intelligence finding."

Paraphrase: "according to a recently downgraded US intelligence report, the Russian Ministry of Defense is in"

Example 5

Headline: Rallies show Pakistan's ex-PM Khan remains political force

First paragraph: "Since he was toppled by parliament five months ago, former Prime Minister Imran Khan has demonstrated his popularity with rallies that have drawn huge crowds and signaled to his rivals that he remains a considerable political force."

Paraphrase: "former Prime Minister Imran Khan has resurrected his popularity since being deposed"

Paraphrase code: `(051, ELI, PAKGOV)`

Example 6

Headline: 'This is it, folks': Boris Johnson bids an ambiguous goodbye

First paragraph: "Boris Johnson's term as British leader was a mix of high drama and low disgrace. But he left office Tuesday with a casual shrug of a farewell: “Well, this is it, folks.””

Paraphrase: “Boris Johnson’s term as British leader was a mix of high drama and low”

Example 7

Headline: Fears high as Canadian police search for stabbing suspect

First paragraph: “Fears ran high Tuesday on an Indigenous reserve in the Canadian province of Saskatchewan after police warned that the suspect in a deadly stabbing rampage over the weekend might be nearby and officers surrounded a house with guns drawn.”

Paraphrase: “fear erupted on an Indigenous reserve in the Canadian province of Saskatchewan on Tuesday,”

Paraphrase code: `(012, CAN, CVL)`

Example 8

Headline: UN: Tribal clashes in Sudan kill 380 in Jan.-Aug. period

First paragraph: “Around 380 people were killed in tribal clashes in Sudan between January and August, most of them in the conflict-wracked Darfur region, the U.N. said Tuesday.”

Paraphrase: “380 people were killed in tribal clashes in Sudan between January and August, the bulk”

Example 9

Headline: Chile’s Boric shakes up cabinet after constitution loss

First paragraph: “Chile’s President Gabriel Boric shook up his cabinet Tuesday in an effort to relaunch his government less than 48 hours after he was dealt a resounding blow when citizens overwhelmingly rejected a new progressive constitution he had championed.”

Paraphrase: “Chile’s President Gabriel Boric shook up his cabinet Tuesday in an attempt to”

Example 10

Headline: Tension rises as Turkey, Greece voice festering grievances

First paragraph: “Troubled relations between regional rivals Turkey and Greece worsened Tuesday, with Turkey’s president doubling down on a thinly veiled invasion threat and Athens responding that it’s ready to defend its sovereignty."

Paraphrase: “tensions between regional rivals Turkey and Greece worsened on Tuesday, with Turkey”