Event Extraction Model Given Corrupted Label Information with Text Masking

Efe Kalyoncu ek2608@nyu.edu
New York University

Ragnor Wu ww2342@nyu.edu
New York University

Emily Hao yh4247@nyu.edu
New York University

Abstract

In the realm of Natural Language Processing, event extraction is a crucial task that enables the conversion of unstructured text into structured data, reflecting occurrences and their relevant details as actionable insights. The conventional models rely heavily on accurate, clean, and well-annotated training datasets, a rarely met condition in real-world scenarios. In many cases, label information is incomplete, noisy, or corrupted due to human error, ambiguous contexts, or inconsistencies in annotation guidelines. This poses a significant challenge for traditional event extraction algorithms. In this paper we aim to investigate and develop a novel algorithmic solution that can cope with these challenges while maintaining sufficient performance in the event extraction task.

1. Introduction

Natural Language Processing (NLP) has revolutionized our ability to extract structured information from unstructured text, enabling us to distill meaningful insights from a vast sea of textual data. At the heart of this transformative field lies the task of event extraction—a process that seeks to uncover and represent occurrences and their pertinent details in a structured format. This structured data, in turn, serves as a foundation for actionable insights and further downstream applications.

Conventionally, the efficacy of event extraction models has hinged upon the availability of accurate, clean, and meticulously annotated training datasets. However, the reality of NLP applications in the real world rarely aligns with these ideal conditions. Label information, crucial for training these models, often is not as available as we need it to be to train large language models to a good convergence point. Furthermore the classical approach of using a classification layer on a model that was trained on token replacement is not as intuitive as using the model with the intended training task. While empirically it is observed that with different objectives these models can still perform well, the weights of the models will always be optimized for the initial training task.

However for token classification purposes, masked language training is usually not used.

The reason for this is because it is hard to create an objective function, and a coupling loss function to train with. One of the reasons for this is because for token classification, the information that needs to be generated is of the same size as the input. However when doing masked language training, only 10% of the data is masked, because of how dense language is in terms of information content. However masked prediction algorithms that mask a larger percentage of the information do exists, in particular in computer vision field, training done with masks obstruct up to 90% of the data before reconstruction. This is due to redundant nature of computer vision where a single pixel does not affect the meaning of an image as much as a single word impacts the meaning of a sentence.

Our idea is rooted from the fact that computer vision models masks more of their information. If we consider the contents of a simple word classification task in every day English, it is easy to notice that the data is highly reflexive. If one is asked what is the event of "We'll check out some of the demonstrations", the answer, "The event is demonstrations" already exists within the first sentence. So a lot of the information in this case is redundant as well. So it is possible to look for a more aggressive masking rate, one high enough to classify tokens with the masked information.

Our endeavor is two-fold:

1. **Generation of sentences relating to the events:** We aim to have a rule based system that generates sentences that includes the tokens that are classified as the event that is coherent enough that it is useful to the model.

2. **Regenerating these sentences given the initial sentences we extracted the event from:** Once we have sentences that are generated based on the event of another sentence, we need to be able to regenerate the sentence through our model. If for a test sample, we can generate the output of the rule based system that takes in the event as an input with our model, without giving any event information, then it is possible to extract what the event was through the use of model alone.

2. Methodology

In this section, we present the methodology employed in our project, outlining the strategies and techniques used to address the challenges posed by corrupted label information in event extraction.

2.1 Dataset

The foundation of our work relies on a carefully chosen dataset, which serves as the basis for training and evaluation. We will be using ACE 2005 dataset and try to extract the tagged events from the sentences they are in. However since not every event in the dataset could be used to generate a coherent English sentence with the system we had, we used a subsection of the data.

2.2 Data Preprocessing

Before training our event extraction model, we perform data preprocessing to prepare the dataset for ingestion. As part of the degradation process new "events" are constructed using NLP techniques to replicate the intent of the original event using different words while preserving grammatical fidelity. Using experimental approaches ranging from straightforward to complex the following processing was implemented:

1. **Single Word Events**: A large percentage of events were single words. First we used stopwords to remove distorted events such "this", "there". For the remaining, if the event was singular, we added "There was a" in front to create a full sentence. For plural events, we added "There were".

    - nomination $\longrightarrow$ ***There was a*** *nomination.*
    - lawsuits $\longrightarrow$ ***There were*** *lawsuits.*

2. **Events with Years**: Events that start with a year such as "1992" were given "From" or "In" the year in front to form a complete sentence.

    - 2004, she serves 6 months, then the day before she is to be released to do 5 yrs probation $\longrightarrow$ ***From the year 2004****, she serves 6 months, then the day before she is to be released to do 5 yrs probation.*

3. **POS Tagging with Synonym Replacement**: Events were tagged with parts-of-speech so that nouns and verbs could be replaced with thoughtful synonyms. This extended to bigrams as well.

    - prepare you for combat $\longrightarrow$ *Prepare you for* ***fighting****.*
    - we got married $\longrightarrow$ *we* ***were wed***
    - miller was shot $\longrightarrow$ *Miller was* ***hit with bullet****.*

4. **Proper Noun Substitution**: Proper nouns such as country abbreviations, capital cities were replaced to modify the sentence without changing the context.

    - U.S. forces moving ever closer to Baghdad $\longrightarrow$ ***united states*** *armed forces moving ever closer to the* ***iraqi capital of baghdad.***
    - a russian soyuz capsule dropped them off this morning after docking perfectly at the orbiting outpost $\longrightarrow$ *a* ***soviet*** *soyuz capsule dropped them off this morning after docking perfectly at the orbiting outpost.*

5. **Relative Time**: Substitutions were made for relative times of day.

    - the Minister of Defense, Reggie White, died of heart attack this morning in his home $\longrightarrow$ *the minister of defense, reggie white, passed away of heart attack* ***earlier today*** *in his home.*

2.3 Loading the Data

We need to load the data in a specific way depending on the approach. For the classification approach, we simply load the data, alongside a mask that indicates the locations of the tokens that are tagged as event tokens. So "There was a war" is accompanied by "[0, 0, 0, 1]"

For the generative approach where the event length is known the structure is that two sentences are given, one of them being masked as follows:

> **Original:** Earlier, from 1979 to 1983, he headed Iraq's Mukhabarat, or intelligence service, a period when the organization arranged executions of regime opponents in Iraq and overseas, the official said. In the year 1979 to 1983, he headed Iraq's Mukhabarat, or intelligence service, a period when the organization arranged executions of regime opponents in Iraq and overseas.
>
> **Given:** Earlier, from 1979 to 1983, he headed Iraq's Mukhabarat, or intelligence service, a period when the organization arranged executions of regime opponents in Iraq and overseas, the official said. In the year mask mask mask mask mask mask mask mask mask mask mask mask mask mask mask mask mask mask mask mask mask mask mask mask mask mask mask.

For the last generative approach, alongside masking the text, we also add random padding to the the given text to make it so the model does not make conclusions from the masks size. For example:

> **Original:** Ask them about going and not just a war, Bob, but an invasion and occupying for up to 10 years a sovereign Arab nation in the midst of one of the most distable and volatile regions in the world. There was a war.
>
> **Given:** Ask them about going and not just a war, Bob, but an invasion and occupying for up to 10 years a sovereign Arab nation in the midst of one of the most distable and volatile regions in the world. There was a mask mask mask mask mask. In this methodology the size of the padding is randomly decided so that no conculsion about size of the true event can be made.

2.4 Transformer-based Model

Our methodology heavily relies on Transformer-based models, specifically BERT and RoBERTa, which have demonstrated exceptional performance in various NLP tasks. We use these models as the backbone for our event extraction model.

2.5 Model Architecture

Our event extraction model consists of the following components:

1. Encoder: We utilize the pre-trained Transformer model as an encoder to extract contextual information from the input text.

2. Label Integration: A key innovation in our approach is the strategic integration of event labels into the training process. We design a mechanism to progressively degrade and obscure this label information within sentences, challenging the model to effectively restore it.

3. Training Objective: Our model is trained using a supervised learning objective, aiming to minimize the loss between predicted event labels and ground truth labels.

2.6 Data Partitioning

In this paper we have split the data with a 90-10 split, using 90% of the available data for our training and test our model on the remaining 10% of the data. The data was shuffled prior to training and partitioning so we had no control over how the training and test sets looked.

2.7 Evaluation Plan

To gauge the efficacy of our approach, we will leverage ACE 2005 data set, an established benchmark data set. Our primary yardstick of success will be the F1 score—an indicator of the precision and recall of events extracted. In our report, we will compare our model's performance against other models found in prior research papers, providing a baseline for our contributions.

3. Results

3.1 Training with Classification Head

To have a baseline approach we initially tried to train using a classification head on our model. So given a sentence, each word was labelled 0 for not a part of the event, and 1 for a part of the event. Using this methodology, we managed to get 81.52% with Bert based fine tuning, 82.57% with RoBERTa based fine tuning and 73.58% with Elektra based fine tuning.

1. BERT 75.07% Recall, 81.52% Precision, F1: 78.16%

2. RoBERTa: 76.72% Recall, 82.57% Precision, F1: 79.53%

3. Elektra: 70.04% Recall, 73.58% Precision, F1: 71.77%

3.2 Predict the Events Given Event Length

The second thing we tried was to predict the specific event tokens, given the event length. This task is easier compared to the simple classification task, as the model is given prior information about length of the event, so it only needs to find the initial index of the event. Using this approach, our findings for precision have improved as we expected, to 99.8% for Bert, 92.68% for RoBERTa and 97.81% for Elektra. We do not have recall for this part, since the number of tokens being given, meant that even if it did not find any matches, it had to fill it with random words anyways.

3.3 Predict the Events by Predicting a Sentence with no Information

Our last approach is the one where we try to predict the contents of an arbitrary second sentence given our initial sentence, where we can deterministically extract the event from the second sentence. This approach is harder than both approaches, as the model can simply predict unrelated words, as well as predicting which words in the initial sentence was incorrect. Despite this, we have seen an improvement in performance from the first method with every model as follows:

1. BERT 94.44% Recall, 92.11% Precision, F1: 93.22%
2. RoBERTa: 88% Recall, 100% Precision, F1: 93.62%
3. Elektra: 99.53% Recall, 86.06% Precision, F1: 92.31%

4. Discussion

The results show that generative approaches that utilizes the native training objective of large language models can outperform adding different heads to the model, if the generation task is well defined. In all three models we have seen substantial improvement compared to using a standard classification head.

We see an interesting mix of precision and recall on different models that were used. Elektra in particular was over generating words, which caused it to have the worst F1 score, despite being the model with the highest recall. RoBERTa on the other hand was very precise, but more willing to skip words, with BERT having a mix of both aspects. As the data size was not substantial, it is important to note that the differences in performance in different categories could have been due to random initialization. The models would need to be tested further to see if this pattern persists.

It is also important to note that while in this paper we show that the generative approach to token classification is feasible, it requires the initial sentences to be well behaving enough that they can be processed into the secondary sentences. This means that our method as is would not be able to handle sentences where the events are complex enough that our automatic sentence generation could not make coherent English sentences with them. However more complex rule based systems could be used to handle such cases. However it is important to realize that the more complex the rule based sentence generation system is, harder it would be to invert the generated sentence.

5. Conclusion

In the end, we observed that using rule based NLP systems for label generation and using model native objective functions can yield better results than using alternative objective functions and model heads when doing large language model training. While using different objective functions have the advantage that they are easier to implement, the implementation of even simple rule based systems can improve the performance of the system depending on the structure of the data.

6. Future Work

While our current study has made significant strides in addressing the challenges of event extraction with corrupted label information, there are also some aspects in terms of which improvements could be done:

6.1 Utilizing Larger Datasets

One direction for future work involves utilizing larger and more diverse datasets for event extraction. A larger dataset can provide a broader spectrum of linguistic variations, contexts, and event types. This allows our model to generalize more effectively. Exploring multiple domains and languages could also lead to a more generalized and robust event extraction system. Moreover, incorporating data from multiple sources, such as news articles, social media, and academic publications, can also enrich the training data and improve the model's adaptability to different textual genres.

6.2 Enhancing Sentence Generation Robustness

To further improve the quality of training data, future research can focus on enhancing the robustness of the sentence generation system. In our project, we developed several ways to generate sentences given the event types and the original sentences. However, these methods are quite primitive and simple. With a better sentence generation system, the raw data could be utilized more effectively. This might involve developing advanced techniques for generating realistic and contextually appropriate sentences. Exploring state-of-the-art language models, such as GPT-4 or successor models, may provide better sentence generation capabilities, leading to more reliable training data.

Acknowledgments

This project was carried out as a part of the curriculum for the Natural Language Processing course at NYU Courant Institute of Mathematical Sciences. We would like to acknowledge the support of our course instructor, Prof. Adam Meyers and our teaching assistant Tarun Sharma for their valuable feedback and guidance.

References

[1] H. Ji and R. Grishman, "Refining event extraction through cross-document inference," in *Proceedings of ACL-08: HLT* (J. D. Moore, S. Teufel, J. Allan, and S. Furui, eds.), (Columbus, Ohio), pp. 254–262, Association for Computational Linguistics, June 2008.

[2] J. Liu, Y. Chen, K. Liu, W. Bi, and X. Liu, "Event extraction as machine reading comprehension," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (B. Webber, T. Cohn, Y. He, and Y. Liu, eds.), (Online), pp. 1641–1651, Association for Computational Linguistics, Nov. 2020.

[3] Q. Li, H. Ji, and L. Huang, "Joint event extraction via structured prediction with global features," in *Proceedings of the 51st Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers) (H. Schuetze, P. Fung, and M. Poesio, eds.), (Sofia, Bulgaria), pp. 73–82, Association for Computational Linguistics, Aug. 2013.

[4] S. Yang, D. Feng, L. Qiao, Z. Kan, and D. Li, "Exploring pre-trained language models for event extraction and generation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (A. Korhonen, D. Traum, and L. Màrquez, eds.), (Florence, Italy), pp. 5284–5294, Association for Computational Linguistics, July 2019.

[5] W. Li, D. Cheng, L. He, Y. Wang, and X. Jin, "Joint event extraction based on hierarchical event schemas from framenet," *IEEE Access*, vol. 7, pp. 25001–25015, 2019.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[8] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," 2020.

[9] W. Xiang and B. Wang, "A survey of event extraction from text," *IEEE Access*, vol. 7, pp. 173111–173137, 2019.

[10] S. Liu, Y. Chen, S. He, K. Liu, and J. Zhao, "Leveraging FrameNet to improve automatic event detection," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (K. Erk and N. A. Smith, eds.), (Berlin, Germany), pp. 2134–2143, Association for Computational Linguistics, Aug. 2016.