Zero-shot Event Detection using a Textual Entailment Model as an Enhanced Annotator

Ziqian Zeng, Runyu Wu, Yuxiang Xiao, Xiaoda Zhong, Hanlin Wang,
Zhengdong Lu and Huiping Zhuang

South China University of Technology

Topic: Event Detection (ED)

Event Extraction

aims at extracting such event information from **unstructured plain texts** into a **structured form**, which mostly describes “who, when, where, what, why” and “how” of real-world events that happened.

Typically, an event in a text is expressed by the following components:

- Event type
- Event trigger
- Event argument
- Argument role: the relationship between an argument and the event in which it participates.

Buyer, Seller, Time, and Artifact are roles of arguments that are specific for the transfer ownership event type.

The goal of **Event Detection (ED)** is to **detect the occurrences of events and categorize them**.

Topic: Textual Entailment

Textual Entailment Task:

to identify the **directional relation** between text pairs.

	ID	sentence	label
Premise		A dog jumping for a Frisbee in the snow.	
Hypothesis	Example 1	An animal is outside in the cold weather, playing with a plastic toy.	*entailment*
	Example 2	A cat washed his face and whiskers with his front paw.	*contradiction*
	Example 3	A pet is enjoying a game of fetch with his owner.	*neutral*

Problems & Our Work

- Problems of previous works:
  - ED methods are mostly accomplished in a supervised manner which requires a large number of annotated data.
  - The aforementioned methods treat the TE model as a frozen annotator which is used solely for inference on the test set.

Problems & Our Work

- Our Work:
  - We turn the TE model into an enhanced annotator by utilizing it to annotate massive amounts of unlabeled data and subsequently finetune it.
  - To improve the efficiency, we propose to use keywords to filter out sentences with a low probability of expressing events.
  - To improve the coverage of keywords, we expand the limited number of seed keywords using WordNet.
  - The experimental results show that our method can outperform other baselines by 15% on the ACE05 dataset.

Problems & Our Work

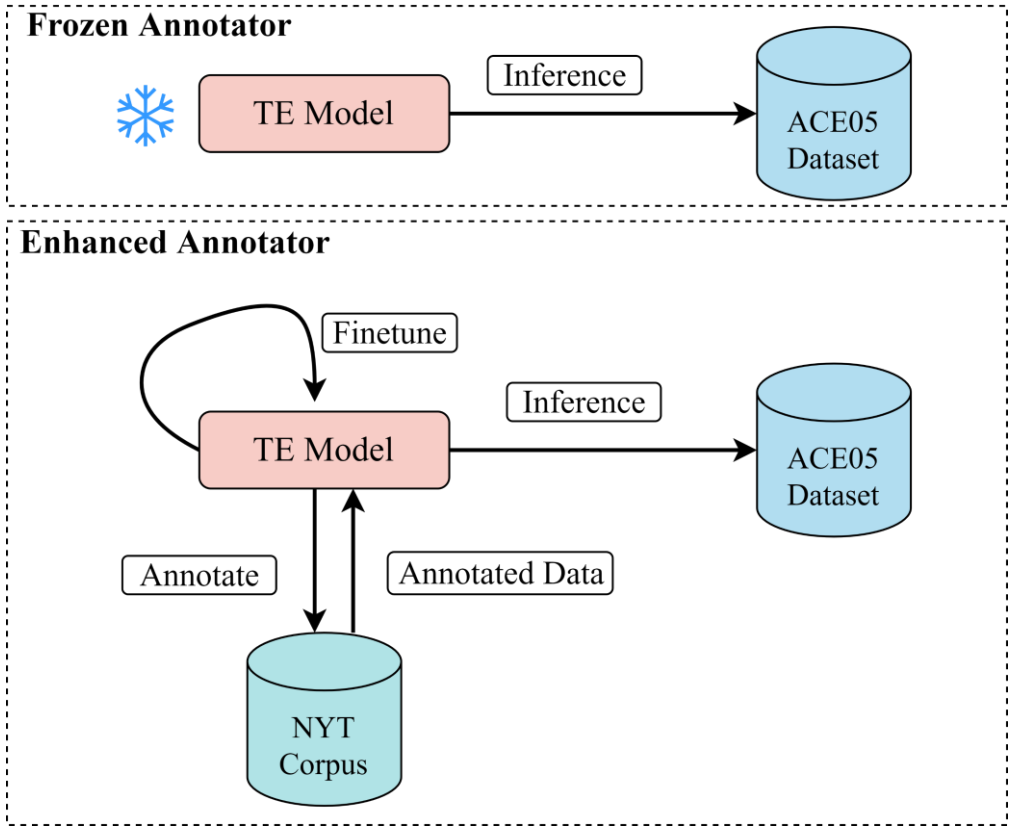


Figure 1. The illustration of the difference between a textual entailment model as a frozen annotator and an enhanced annotator

Data Annotation

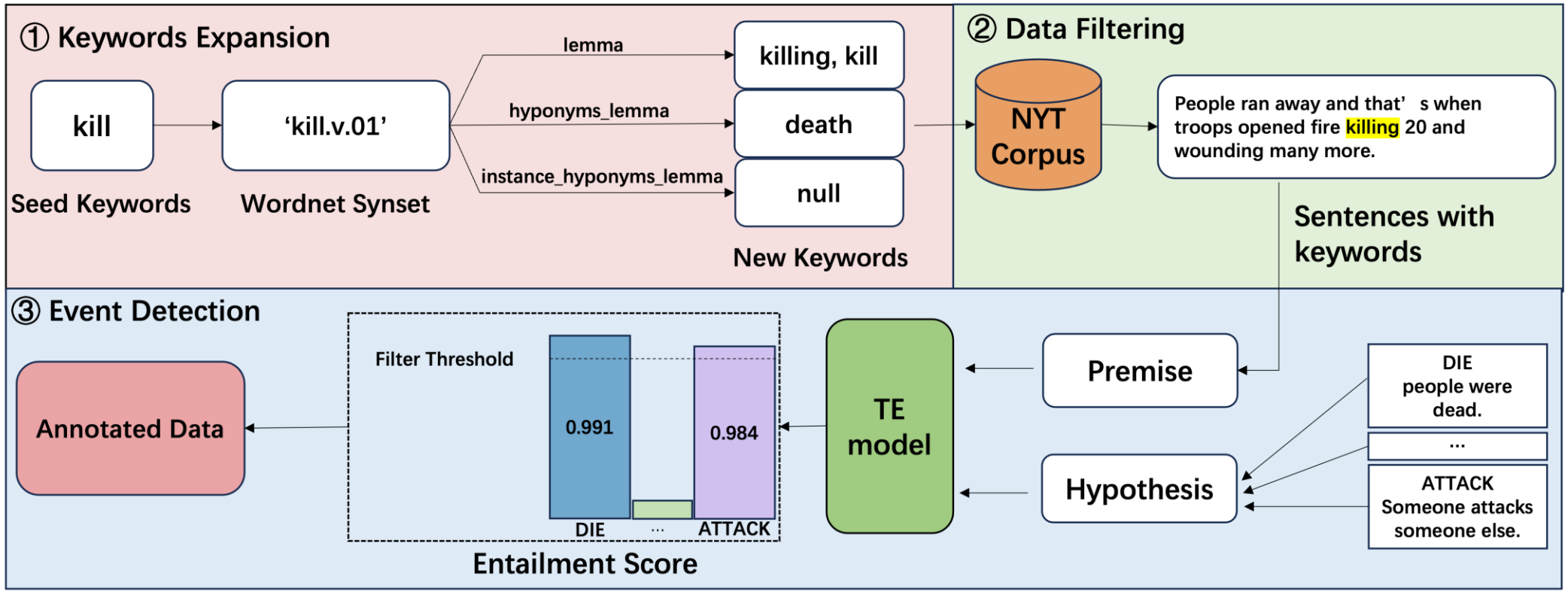
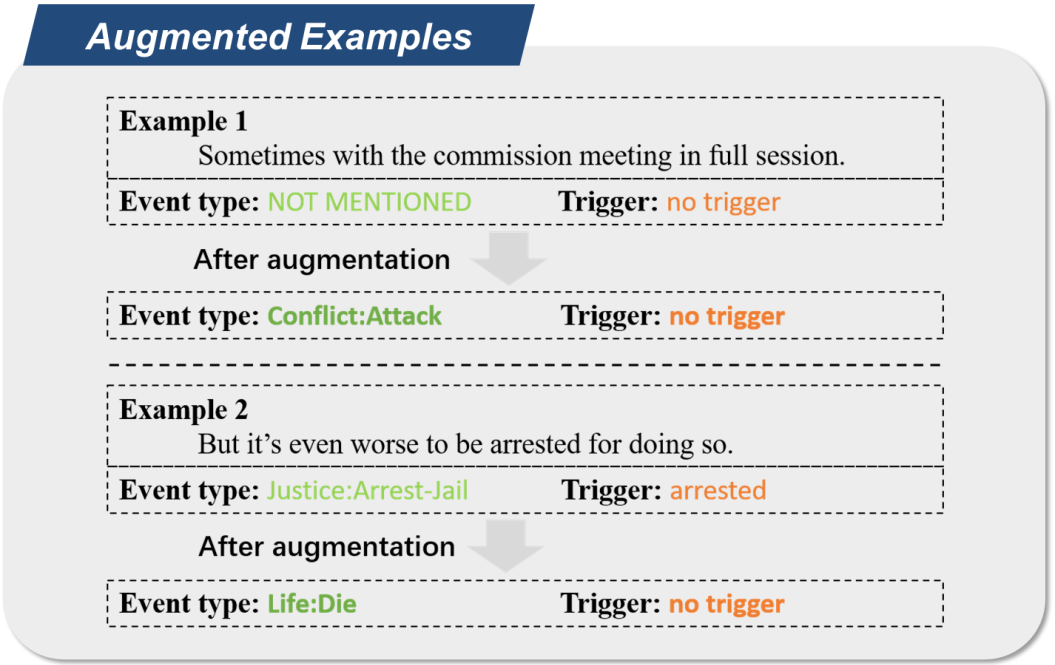


Figure 2. The general workflow of using a pre-trained TE model and keyword expansion to annotate unlabeled data.

TE Model Finetuning

- For the event detection task, we use the annotated NYT data to finetune the TE model.
- In case triggers are needed in downstream tasks, we also propose a method to identify triggers given detected event types as inputs. We finetune the BERT (Devlin et al., 2018) model using the annotated NYT data via prompt tuning.
- If a sentence does not express any event, we let the trigger classification model to predict "no trigger." We propose two data augmentation methods to generate "no trigger" data.



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.

Experiments

- Experimental Settings
  - Datasets
  - Compared Methods
- Experiment Results
  - Event Detection Results
  - Trigger Classification Results
  - Low-resource Analysis
  - Hyperparameter Analysis

Datasets

1. **ACE05-E+** (Lin et al., 2020) dataset is a widely used dataset for the event extraction task, which predefines 8 event types and 33 subtypes.

Splits	Train	Dev	Test
Sentences	19,240	902	676
Events	4,419	468	424

Table 1: Statistics of ACE05-E+ Dataset.

2. **Annotated NYT Data** We extract sentences that contain keywords in the New York Times (NYT) corpus (Sandhaus, 2008). Finally, we collected **322,570** data, including **268,406** single-event data and **54,164** multi-event data. The single-event (multi-event) data express one (more than one) event within a sentence

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In Proceedings of ACL, pages 7999–8009.
Evan Sandhaus. 2008. The new york times annotated corpus. Linguistic Data Consortium,Philadelphia, 6(12):e26752.

Compared Methods

Zero-shot event detection baseline methods:

- Liberal_EE (Huang et al., 2016)
- ZS4IE (Sainz et al.,2022),
- ZS_Transfer (Lyu et al., 2021)
- ZS_CLEVE (Wang et al., 2021)
- Label_Aware (Zhang et al., 2021)
- Chat4ED (Li et al., 2023)

Upper-bound supervised methods:

- CLEVE
- OneIE (Lin et al., 2020)
- TBNNAM (Liu et al., 2019)

Experiments

- Experimental Settings
  - Datasets
  - Compared Methods
- Experiment Results
  - Event Detection Results
  - Trigger Classification Results
  - Low-resource Analysis
  - Hyperparameter Analysis

Event Detection Results

Our method outperforms the baseline ZS_CLEVE by 15%. Our method can achieve **86%** performance of the upper-bound supervised CLEVE. Without using expanded keywords, our method drops **3%**, which shows the **effectiveness of the keyword expansion** strategy.

Methods	P	R	F1
CLEVE (Wang et al., 2021)	78.1	81.5	79.8
OneIE (Lin et al., 2020)	74.3	70.3	72.2
TBNNAM (Liu et al., 2019)	76.2	64.5	69.9
Liberal_ EE (Huang et al., 2016)	55.7	45.1	49.8
ZS4IE (Sainz et al., 2022)	32.0	52.9	39.9
ZS_Transfer (Lyu et al., 2021)	31.7	60.6	41.7
ZS_CLEVE (Wang et al., 2021)	62.0	47.3	53.7
Label_Aware (Zhang et al., 2021)	54.1	53.1	53.6
Chat4ED (Li et al., 2023)	9.4	44.3	15.5
ZS_TE (our method)	**65.6**	72.3	**68.8**±0.003
w/o keyword expansion	54.0	**83.6**	65.6±0.006

Table 2: Precision, recall, and F1 scores (%) in the event detection task.

Furthermore, the combination of single-event and multi-event data yields the best F1 score.

Data Combinations	P	R	F1
Single	58.0	74.9	65.3±0.018
Multi	37.3	**94.5**	53.5±0.012
Single + Multi	**65.6**	72.3	**68.8**±0.003

Table 3: Precision, recall, and F1 scores (%) of our methods in the event detection task using different data combinations.

Trigger Classification Results

The trigger classification result drops 9%. The possible reason is that BERT model may not be proficient in identifying and classifying words.

ZS_TE (our method)	P	R	F1
Event Detection	65.6	72.3	68.8±0.003
Trigger Classification	66.9	54.1	59.8±0.002

Table 4: Precision, recall, and F1 scores (%) in the event detection and trigger classification task.

Low-resource Analysis

- We evaluate our method and two supervised methods on a low-resource setting in which we use 10%~50% ACE data for training.
- Our method consistently outperforms TBNNAM(Liu et al., 2019) by a large margin indifferent proportions.

Note that OneIE used trigger-level an notations while our method and TBNNAM do not use them. Direct comparison between OneIE and trigger-free methods is not fair. OneIE here serves as a reference rather than a baseline.

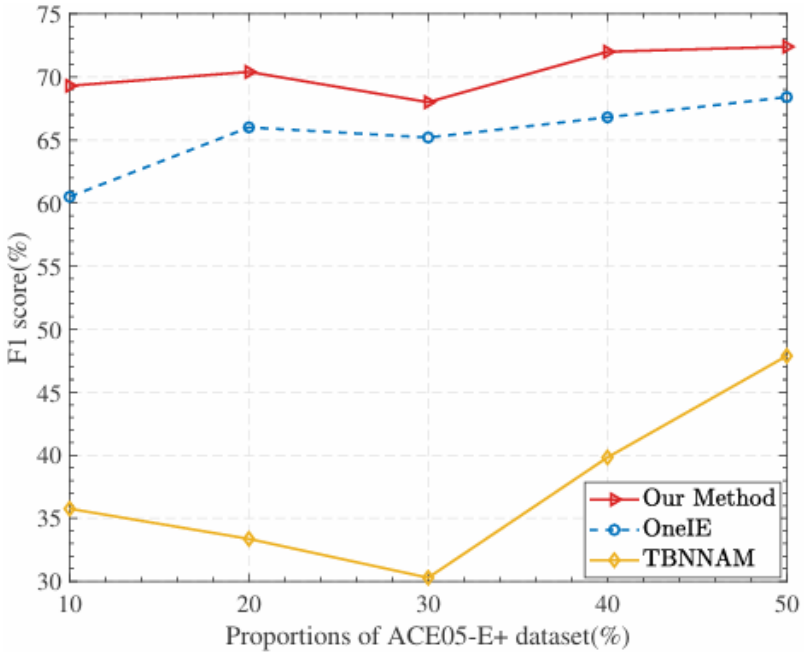


Figure 4: F1 scores (%) of our method and OneIE in the event detection task in different low-resource settings.

Hyperparameter Analysis

- The search range of confidence threshold γ is $\{0.5,\cdots,0.9\}$. As shown in Figure, 0.9 yields the best performance and stability among all threshold values.
- When the confidence threshold γ is larger, the performance is better because a high confidence threshold γ can rule out more wrong event types.

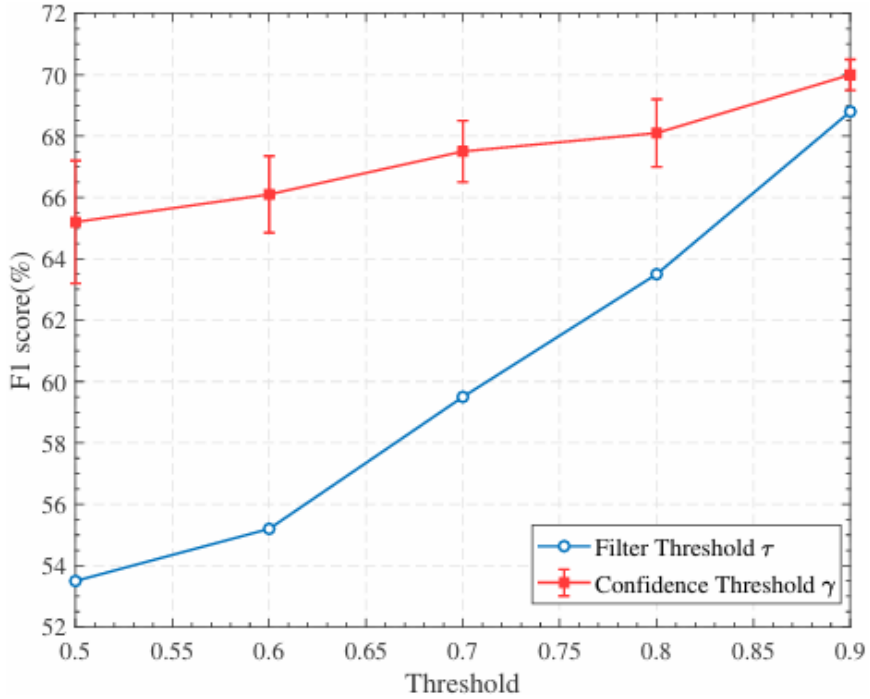


Figure 5: F1 scores (%) in the event detection task under different filter threshold τ and confidence threshold γ.

THANKS

Speaker: Yuxiang Xiao

South China University of Technology