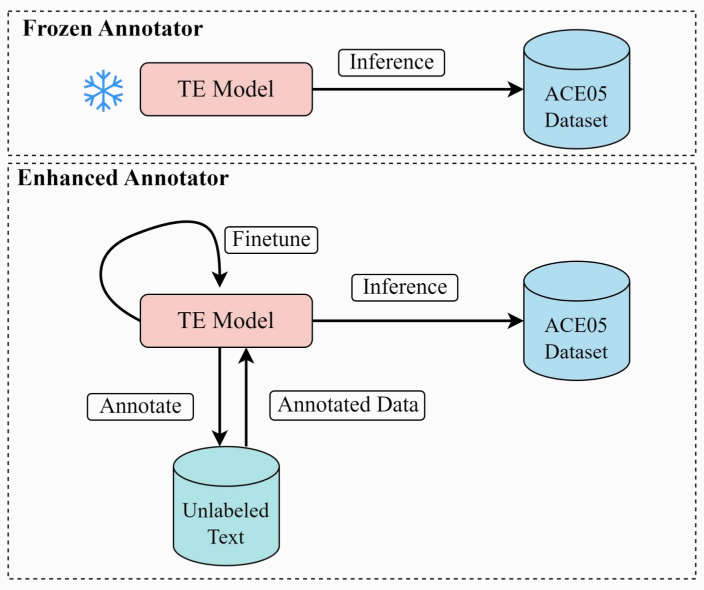
Zero-shot Event Detection using a Textual Entailment Model as an Enhanced Annotator

Ziqian Zeng, Runyu Wu, Yuxiang Xiao, Xiaoda Zhong, Hanlin Wang, Zhengdong Lu, Huiping Zhuang
South China University of Technology

Introduction

- Zero-shot event detection is a challenging task. Recent research work proposed to use a pre-trained textual entailment (TE) model to solve this task. However, those methods treated the TE model as a frozen annotator. We treat the TE model as an annotator that can be enhanced.
- We propose to use a TE model to annotate large-scale unlabeled text and use annotated data to finetune the TE model, yielding an improved TE model.
- To improve the efficiency, we propose to use keywords to filter out sentences with a low probability of expressing event(s).
- To improve the coverage of keywords, we expand limited number of seed keywords using WordNet, so that we can use the TE model to annotate unlabeled text efficiently.

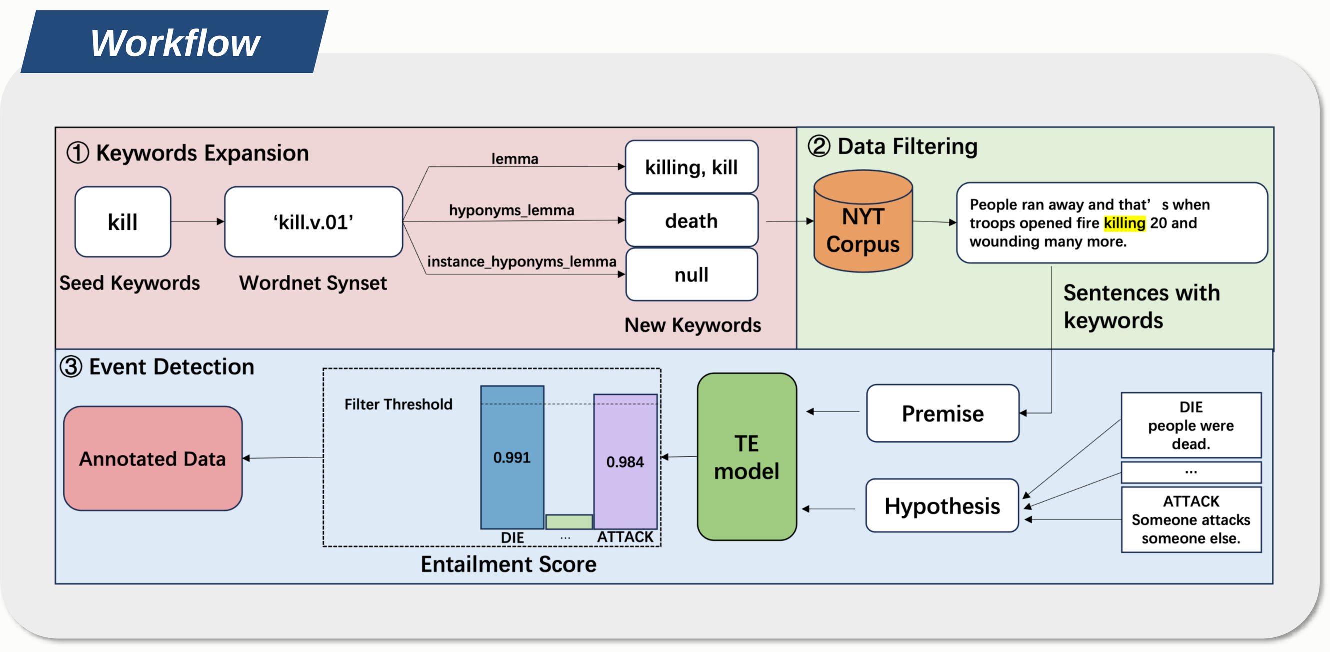


Methodology

1. Data Annotation

First, we expand keywords using Word Net (Miller, 1995). Secondly, we extract sentences that contain keywords from the New York Times (NYT) corpus (Sandhaus, 2008) and then use a pre-trained TE model to annotate them.

Workflow



2. 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 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.

Augmented Examples

Example 1
Sometimes with the commission meeting in full session.
Event type: NOT MENTIONED Trigger: no trigger
After augmentation
Event type: Conflict:Attack Trigger: no trigger

Example 2
But it's even worse to be arrested for doing so.
Event type: Justice:Arrest-Jail Trigger: arrested
After augmentation
Event type: Life:Die Trigger: no trigger

Experiments Settings

Datasets

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

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

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.

Zero-shot event detection baseline methods & Supervised upper-bound methods

Results

1. Event Detection

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**$\pm$0.003
w/o keyword expansion	54.0	**83.6**	65.6$\pm$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$\pm$0.018
Multi	37.3	**94.5**	53.5$\pm$0.012
Single + Multi	**65.6**	72.3	**68.8**$\pm$0.003

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

2. Trigger Classification

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$\pm$0.003
Trigger Classification	66.9	54.1	59.8$\pm$0.002

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

3. Low-resource Settings

We evaluate our method and two supervised methods on a low-resource setting in which we use 10%~50% ACE data for training.

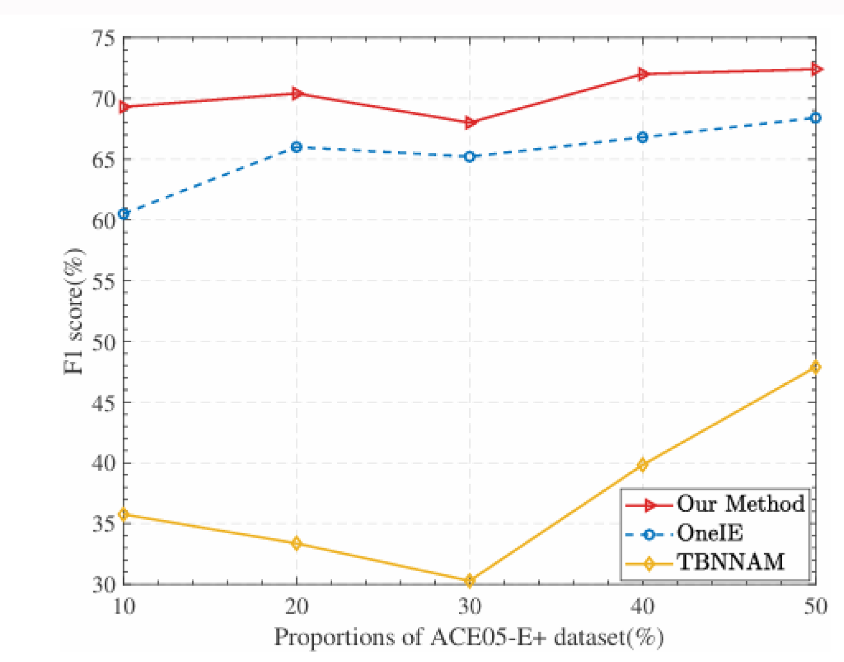


Figure 4: F1 scores (%) of our method and OneIE in the event detection task in different low-resource settings.

4. 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.

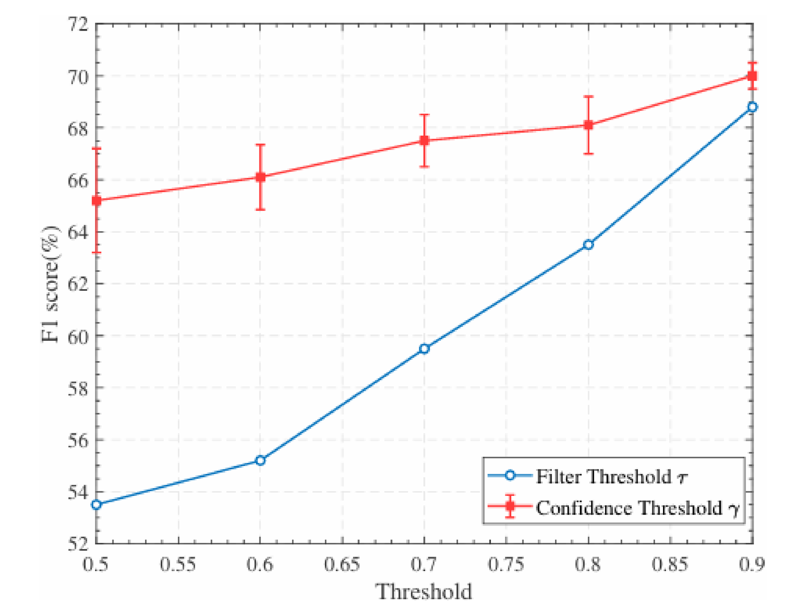


Figure 5: F1 scores (%) in the event detection task under different filter threshold τ and confidence threshold γ.

References

George A Miller. 1995. Wordnet: a lexical database for english. Communications of the ACM, 38(11):39–41.
Evan Sandhaus. 2008. The new york times annotated corpus. Linguistic Data Consortium, Philadelphia, 6(12):e26752.
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.