Unregulated Chinese-to-English Data Expansion Does NOT Work for Neural Event Detection

Zhongqiu Li, Yu Hong*, Jie Wang, Shiming He, Jianmin Yao, Guodong Zhou
School of Computer Science and Technology, Soochow University, China
{leosanqiu, tianxianer, jawn233, smhelpai}@gmail.com
{jyao, gdzhou}@suda.edu.cn

Abstract

We leverage cross-language data expansion and retraining to enhance neural Event Detection (abbr., ED) on English ACE corpus. Machine translation is utilized for expanding English training set of ED from that of Chinese. However, experimental results illustrate that such strategy actually results in performance degradation. The survey of translations suggests that the mistakenly-aligned triggers in the expanded data negatively influences the retraining process. We refer this phenomenon to "trigger falsification". To overcome the issue, we apply heuristic rules for regulating the expanded data, fixing the distracting samples that contain the falsified triggers. The supplementary experiments show that the rule-based regulation is beneficial, yielding the improvement of about 1.6% F1-score for ED. We additionally prove that, instead of transfer learning from the translated ED data, the straight data combination by random pouring surprisingly performs better.

1 Introduction

We tackle ED, a task of recognizing trigger words (triggers for short) that signal different types of events (Ahn, 2006). For example, the trigger "*meeting*" in (1) signals the CONTACT-MEET event.

(1) "*We have the transcript of the **meeting**.*"
Trigger: *meeting*
Type: CONTACT-MEET

The current study of ED, in general, utilizes neural classification models for determining the event types in the word-by-word manner, including the predefined ACE[1] event types and Non-trigger type. Correspondingly, supervised learning is applied to pursue the semantic-level distributed representations of words, so as to provide perceptible evidence for decoding event types.

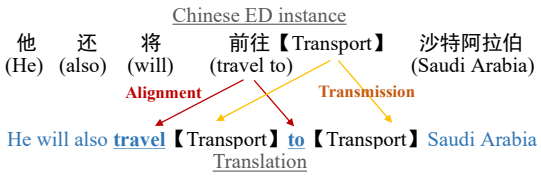

Figure 1: A case study of trigger designation and class-label transmission which is bridged by word alignment.

Expanding the training data is able to enhance the ED-oriented neural classification models. It is because there are a larger amount of knowledge and diverse pragmatic phenomena can be introduced into the supervised learning process, as claimed in different tasks of natural language processing. We childishly embrace this method, conducting expansion by translating Chinese ED corpus and pouring translations into the English training set, where word alignment (Sabet et al., 2020) is used for designating triggers in the translations.

What is beyond our expectation is that, however, cross-language data expansion actually results in performance degradation. We survey the translations which were used for expansion, and observe that a large number of low-quality instances were involved. The survey also suggests that the falsified triggers in such instances probably misled neural models during supervised learning. We provide an example in Figure 1 where Chinese event mention is taken from the publicly-shared ACE corpus of Chinese ED, and word alignment is carried out for trigger designation in English translation, as well as transmission of class label information. It can be found that the trigger "*to*" is mistakenly designated and labeled for event class (viz., TRANSPORT) due to inexact alignment, and undoubtedly it will cause severe misleading (in terms of the high occurrence frequency of "*to*" in English).

To overcome the issue, we explore a variety of easily-accessible rules to purify the translated ED instances (Section 2). Using the purified instances

*Corresponding author.
[1]https://catalog.ldc.upenn.edu/LDC2006T06

Proceedings of the 29th International Conference on Computational Linguistics, pages 2633–2638
October 12–17, 2022.

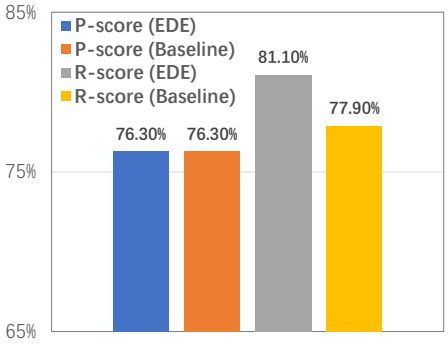

Figure 2: P and R-scores obtained by the baseline and ED𝔼♣, where RoBERTa-**base** is used.

as external data, we conduct retraining and transfer learning to strengthen the baseline ED models, including the ones which are grounded on basic and large RoBERTa (Liu et al., 2019), respectively (Section 3). Experimental results (Section 4) show that 1) the rule-based regulation helps to avoid performance degradation and yields substantial improvements, and 2) conventional expansion by combining datasets is beneficial while, on the contrary, transfer learning is less useful. We overview the related work in Section 5 before concluding this paper (Section 6).

2 Rule-based Purification Against Trigger Falsification

We apply Google translation toolkit[2] for translating Chinese event mentions into English, and use *SimAlign* (Sabet et al., 2020) to pursue the alignment between triggers in Chinese mentions and words in the corresponding translations. The aligned words are designated as triggers of translations and assigned with the manually-labeled event types in Chinese corpus.

Word alignment unavoidably falsify triggers in the translations. Therefore, we explore five heuristic rules to regulate the falsified triggers.

Unbinding prepositions It has been exhibited in Figure 1 that some prepositions (e.g., "*to*") are mistakenly designated as triggers due to inexact alignment, i.e., a Chinese trigger is aligned to the constituent that contains both verb and preposition. The number of prepositions that serve as triggers in translations is up to 326, occupying 8% of all the designated triggers. In the cases, we unbind verbs from prepositions, and designate the latter as `Non-trigger` words.

[2]https://translate.google.com

Unbinding participles In some cases, a single Chinese trigger is aligned to the present or past-participle phrase, where the participle that stands for an attributive is redundant for signaling a certain event type and, more seriously, it is common and generally leads a variety of word senses. For example, the past-participle "*opened*" in (2) is redundant. There are 38 participles found to be mistakenly designated as triggers, occupying about 1% of all the designated triggers. We repeal the designation.

(2) 坦克向两辆正常行驶的民用车辆开火
Translation: *Tanks opened fire on two normal civilian vehicles*
Chinese trigger: 开火; **Type**: `ATTACK`
Alignment: 开火="***opened*** *fire*"

Binary-choice exclusion Occasionally, a single English word is aligned with a pair of Chinese words, including not only a `Non-trigger` word but trigger. For example, both the `Non-trigger` word "提出" (i.e., "*bring*") and trigger "上诉" ("*lawsuit*") are aligned to the English word "*appealing*" in (3). In the cases, we exclude the `Non-trigger` type, but instead merely assign the concrete event type (such as `SUE` in (3)) to the aligned English word. There are 58 binary-choice cases occurred in the translations, occupying 1.4% of all the designated triggers.

(3) 我们正(提出)(上诉)
Translation: *We are appealing*
Chinese trigger: 上诉; **Type**: `SUE`
Alignment: (提出)(上诉)="*appealing*"

Correcting far-fetched triggers Before alignment, some Chinese triggers are segmented into formal characters or the ones holding less senses. As a result, the Chinese triggers are easily aligned with function words (prepositions and conjunctions) instead of content words in English. Grounded on the alignment results, the trigger designation method produces a series of far-fetched triggers. For example, the Chinese trigger "身中" (i.e., "*injured*") in (4) is mistakenly segmented into the characters "身" (i.e., *body*) and "中" ("*in*"), and the aligned preposition "*in*" is designated as the `INJURY` trigger. The number of English prepositions and conjunctions that were designated as triggers is up to 226, occupying 5.5%. We correct the errors by designating them as `Non-trigger` words.

(4) 发射了80发胡椒弹并(身中)约57发
Segmentation: (发射)(了)(80)(发)(胡椒)—

(弹)(并)(身)(中)(约)(57)(发)
Translation: *Fired 80 pepper bombs at him, with about 57 (in) his body*
Chinese trigger: 身中; **Type**: `INJURY`
Alignment: (身中)="*in*"

Skipping the omissions A large number of Chinese triggers fail to be aligned with any English word. For example, although the Chinese trigger "启用" is semantically equivalent to the English word "*opened*" in (5), the alignment is neglected. This results in the omission of triggers in translations. More seriously, the omitted triggers will be designated as `Non-trigger` word, and thus mislead classification models during training. Therefore, we skip the mentions in which trigger omission occurs. There are 426 cases of trigger omission found in the designation process, occupying 10.4% of all the Chinese triggers.

(5) 重新改建的勤务中心是在上午落成(启用)
Translation: *The remodeled service center was completed and (opened) in the morning*
Chinese trigger: 启用; **Type**: `Start-Org`
Alignment: (启用)="*None*"

3 Enhancing Classification Models

We use pretrained language models for ED, including RoBERTa-base and RoBERTa-large (Liu et al., 2019). RoBERTa-base is constructed by 12 transformer layers (Vaswani et al., 2017), each of which contains a 12-head attention network and 768 hidden states. RoBERTa-large is constructed by 24 transformer layers, each of which contains a 16-head attention network and 1,024 hidden states. The input of both RoBERTa models is a sentence no matter whether it appears as an event mention containing triggers. The maximum input length is set to 256 tokens, and padding is used if the input sentence fails to reach the length (Section 4.2 presents other hyperparameters). The initial word embeddings are obtained using look-up tables, and they are slightly strengthened by element-wise fusion with position embeddings. Besides, both RoBERTa models are connected with a linear fully-connected layer and Softmax layer (Bridle, 1990). For each word in the input sentence, the RoBERTa models conduct 34-class classification, towards not only the predefined 33 ACE event types but `Non-trigger` type.

We intend to enhance the classification models by transfer learning (Bengio, 2012) and data expansion (Journal and Alabert, 1989), using the translated Chinese ED corpus (**MT-ED** for short) as the external data. The aforementioned rule-based purification is utilized for regulating MT-ED. The considered models in experiments are as below:

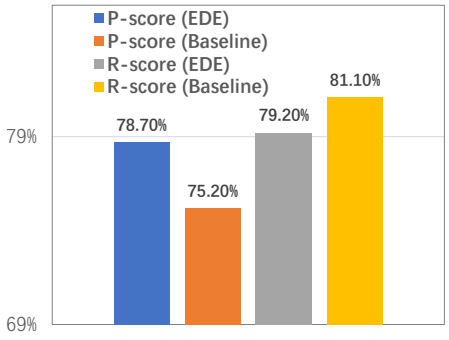

Figure 3: P and R-scores obtained by the baseline and ED$\mathbb{E}^{\clubsuit}$, where RoBERTa-**large** is used.

Baselines The baselines denotes the RoBERTa-based classifiers which are merely trained on the original training set. Such a training set contains ED instances that were split from the English corpus of the publicly shared ACE-2005 tasks.

ED$\mathbb{T}$ Transfer learning is used to enhance the RoBERTa-based classifiers. We first train the classifiers on MT-ED, and then train them on the original training set. Within the double-stage training process, the parameters obtained in the first stage (on MT-ED) are transferred to the second stage (on the original set). We refer the classifiers to ED$\mathbb{T}$s.

ED$\mathbb{E}$ We use MT-ED to expand the original training set by straight pouring, without any additional handling. Using the expanded data set, we train the RoBERTa-based classifiers from scratch. We refer the obtained classifiers to ED$\mathbb{E}$s.

4 Experimentation

4.1 Corpus and Evaluation Measure

We carry out experiments on the ACE-2005 benchmark dataset of English ED task, which comprises 599 documents. The documents contain about 5.2K manually-labeled triggers for 33 predefined event classes, and 280K `Non-trigger` words. We follow the common practice to set up the training, validation and test sets, which hold 529, 30, and 40 documents, respectively.

Besides, we use a set of Chinese ED instances which are taken from the ACE-2005 multilingual training corpus. Such data set comprises 633 documents scripted in Chinese, and involves about

3.3K triggers of 33 ACE event classes as well as 170K `Non-trigger` words. After purification, we collect 2.6K translated mentions, 2.6K triggers and about 218K `Non-trigger` words for building MT-ED. It is taken into consideration during transfer learning and data expansion (Section 3).

We evaluate all the considered classification models using the measure of Precision (P), Recall (R) and F1-score.

4.2 Hyperparameter Settings

The hyperparameters of both RoBERTa-base and RoBERTa-large are set as follows. The learning rate is set to 1e-5. We set epoch to 16 and batch size to 8. AdamW (Loshchilov and Hutter, 2017) optimizer is used where ε is set to 10e-8.

4.3 Results and Analysis

First, we examine the feasibility of cross-language data expansion for enhancing ED. The performance is indicated by ED$\mathbb{E}^*$ in Table 1, where the mark "*" denotes that ED$\mathbb{E}$ is trained on the unpurified MT-ED. It can be observed that, compared to the baseline, ED$\mathbb{E}^*$ obtains worse performance. By contrast, training ED$\mathbb{E}$ using the purified MT-ED produces substantial performance gains, as indicated by ED$\mathbb{E}^{\clubsuit}$ in Table 1. The test results reveal the necessity of data purification when MT-ED is combined with the original training set.

We compare RoBERTa-base to RoBERTa-large when different training sets are used, including the original training set, as well as the expanded version with the purified MT-ED. Table 1 shows P and R-scores they achieved, which are opposite to each other. Specifically, as indicated by *baseline* and ED$\mathbb{E}^{\clubsuit}$, RoBERTa-base achieves much higher R-score and slightly lower P-score when data expansion is used, but on the contrary, data expansion has exactly the opposite effect for RoBERTa-large. We also evaluate the performance of binary classification for triggers and `Non-trigger` words. Figure 2 shows the P and R-scores obtained by the baseline and ED$\mathbb{E}^{\clubsuit}$ when RoBERTa-base is used, while Figure 3 shows that of RoBERTa-large. It can be observed that ED$\mathbb{E}^{\clubsuit}$ achieves much higher R-score than baseline when RoBERTa-base is considered, but both of them achieve the same P-scores. On the contrary, the P and R-scores obtained when RoBERTa-large is considered change to be opposite states. The phenomena imply that the deeper neural networks like RoBERTa-large most probably overfit the common or homogeneous event instances in the original training set and MT-ED, though a small amount of novel knowledge within MT-ED is impervious to them.

RoBERTa-**base**	**P** (%)	**R** (%)	**F1** (%)
Baseline	**72.7**	74.2	73.4
ED$\mathbb{E}^*$	70.3	76.0	73.0
ED$\mathbb{E}^{\clubsuit}$	72.5	**76.9**	**74.6**$^{\triangle}$
RoBERTa-**large**	**P** (%)	**R** (%)	**F1** (%)
Baseline	72.9	**78.5**	75.6
ED$\mathbb{E}^*$	75.9	74.9	75.4
ED$\mathbb{E}^{\clubsuit}$	**76.2**	76.7	**76.5**$^{\triangle}$

Table 1: Performance of 34-class classification for ED when **data expansion** is used. The mark "*" denotes the use of unpurified MT-ED data for expansion, "$\clubsuit$" is that of purified, and "$\triangle$" indicates the significance level that p-value (Dror et al., 2018) is smaller than 0.05.

RoBERTa-**base**	**P** (%)	**R** (%)	**F1** (%)
ED$\mathbb{T}^{\clubsuit}$	69.7	76.7	73.0
ED$\mathbb{E}^{\clubsuit}$	**72.5**	**76.9**	**74.6**
RoBERTa-**large**	**P** (%)	**R** (%)	**F1** (%)
ED$\mathbb{T}^{\clubsuit}$	**77.6**	75.1	76.3
ED$\mathbb{E}^{\clubsuit}$	76.2	**76.7**	**76.5**

Table 2: Comparison between ED$\mathbb{E}^{\clubsuit}$ and ED$\mathbb{T}^{\clubsuit}$.

In a separate experiment, we compare the effect of data expansion to that of transfer learning, where ED$\mathbb{T}^{\clubsuit}$ and ED$\mathbb{E}^{\clubsuit}$ are considered. Table 2 shows the comparison results. It can be observed that ED$\mathbb{E}^{\clubsuit}$ outperforms ED$\mathbb{T}^{\clubsuit}$ for F1-score no matter what kind of RoBERTa (base or large) is used. Note that the scale of external data they take from MT-ED is the same. The comparison results suggest that asynchronous learning from exotic event knowledge to local contributes less to ED, compared to synchronous learning on the shuffled data.

5 Related Work

Conventional ED models rely heavily on elaborate feature engineering, such as that of context-independent features (Ji and Grishman, 2008), as well as cross-event (Liao and Grishman, 2010) and cross-entity (Hong et al., 2011) statistical features. In order to pursue the perception of deep event semantics, the current study concentrates on the utilization of neural networks, designing and developing a series of reliable neural ED models, including those which are grounded on CNN (Nguyen and Grishman, 2015), DMCNN (Chen et al., 2015), RNN (Nguyen et al., 2016), GAN (Hong et al.,

2018), GCN (Li et al., 2020) and VAE (Huang and Ji, 2020). Recently, the pretrained language models like BERT (Yang et al., 2019), RoBERTa (Wang et al., 2021) and AD-DMBERT (Wang et al., 2019) are used, yielding substantial improvements.

Data-driven enhancement strategies have been explored for ED, most of which are implemented by data augmentation. Yang et al. (2019) produce new ED instances by entity replacement. It is potentially effective to enhance entity-aware neural encoders for detecting events that hold entities. Tong et al. (2020) leverage knowledge distillation, which is beneficial for bringing open-domain knowledge into the understanding of local events. Veyseh et al. (2021) use GPT-2 to generate new training data. Teacher-student learning is applied for attenuating the effect of the generated noises.

6 Conclusion

We use cross-language data expansion to enhance neural ED models. Experimental results demonstrate that unregulated data expansion yields less improvement or even causes performance degradation. By contrast, data purification by simple heuristic rules produces substantial performance gains. In addition, it is proven that data expansion contributes more to ED then transfer learning.

Conducting multilingual data expansion potentially contributes to the enhancement of ED models. It is because diverse pragmatics in different languages and exotic event knowledge are informative for versatile encoding. However, it is challenging due to the lack of shareable purification rules among different languages for trigger alignment. Therefore, we will develop an automatic purification model that generalize well in different languages, where the encoding of syntactic information and reinforcement learning will be used.

Acknowledgements

The research is supported by National Key R&D Program of China (2020YFB1313601), National Science Foundation of China (62076174, 62076175).

References

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8.

Yoshua Bengio. 2012. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36. JMLR Workshop and Conference Proceedings.

John S Bridle. 1990. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In *Advances in neural information processing systems*, pages 211–217.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392.

Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou, and Qiaoming Zhu. 2011. Using cross-entity inference to improve event extraction. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 1127–1136.

Yu Hong, Wenxuan Zhou, Jingli Zhang, Guodong Zhou, and Qiaoming Zhu. 2018. Self-regulation: Employing a generative adversarial network to improve event detection. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 515–526.

Lifu Huang and Heng Ji. 2020. Semi-supervised new event type induction and event detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 718–724.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Proceedings of ACL-08: Hlt*, pages 254–262.

AG Journal and F Alabert. 1989. Non-gaussian data expansion in the earth sciences. *Terra Nova*, 1(2):123–134.

Manling Li, Qi Zeng, Ying Lin, Kyunghyun Cho, Heng Ji, Jonathan May, Nathanael Chambers, and Clare Voss. 2020. Connecting the dots: Event graph schema induction with path language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 684–695.

Shasha Liao and Ralph Grishman. 2010. Using document level cross-event inference to improve event

extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 789–797.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309.

Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 365–371.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. *arXiv preprint arXiv:2004.08728*.

Meihan Tong, Bin Xu, Shuai Wang, Yixin Cao, Lei Hou, Juanzi Li, and Jun Xie. 2020. Improving event detection via open-domain trigger knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5887–5897.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Amir Pouran Ben Veyseh, Viet Lai, Franck Dernoncourt, and Thien Huu Nguyen. 2021. Unleash gpt-2 power for event detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6271–6282.

Xiaozhi Wang, Xu Han, Zhiyuan Liu, Maosong Sun, and Peng Li. 2019. Adversarial training for weakly supervised event detection. 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 998–1008.

Ziqi Wang, Xiaozhi Wang, Xu Han, Yankai Lin, Lei Hou, Zhiyuan Liu, Peng Li, Juanzi Li, and Jie Zhou. 2021. Cleve: Contrastive pre-training for event extraction. *arXiv preprint arXiv:2105.14485*.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5284–5294.