Zero-shot Event Extraction via Transfer Learning:
Challenges and Insights

Qing Lyu, Hongming Zhang, Elior Sulem, Dan Roth
Department of Computer & Information Science
University of Pennsylvania
Task Definition

- An event is represented as a trigger + several arguments.
- Example from ACE-2005:

Event type: TRANSFER-OWNERSHIP

\[
\text{China has purchased two nuclear submarines from Russia last month.}
\]

<table>
<thead>
<tr>
<th>Buyer-Arg</th>
<th>Trigger</th>
<th>Artifact-Arg</th>
<th>Seller-Arg</th>
<th>Time-Arg</th>
</tr>
</thead>
</table>

- Event Extraction (EE) = Trigger Identification (TI) + Trigger Classification (TC) + Argument Identification (AI) + Argument Classification (AC)
Predominant approaches: **supervised**, both expensive & inflexible.

Recent efforts explored zero-shot event extraction, usually requiring some event types to be **seen** (Huang et al., 2018) / only dealing with triggers or arguments **alone** (Peng et al., 2016; Liu et al., 2020).

Their performance is still far from supervised methods, but little is known about **why**.

Our work:

- Proposes a zero-shot event extraction system that tackles **both** triggers and arguments without any **event training data**, via transfer learning from Question Answering (QA) / Textual Entailment (TE).
- Provides insights into the remaining challenges behind the performance gap.
Approach: Trigger Extraction

“The text piece”'s (predicate + core SRL arguments) = premise

“China purchased two nuclear submarines from Russia”

“China purchased two nuclear submarines from Russia”

“This text is about {event type}” = hypothesis

“This text is about a transfer of ownership” (hypothesis for TRANSFER-OWNERSHIP)

Entailment confidence >=? threshold

0.995 > 0.99

“purchased” is the trigger of a TRANSFER-OWNERSHIP event

---

1An alternative uses Yes/No QA instead of TE, which is similar and thus not illustrated.
Approach: Argument Extraction

For each extracted trigger (span + event type),
e.g. “purchased” + TRANSFER-OWNERSHIP

For each extracted trigger (span + event type),
e.g. “purchased” + TRANSFER-OWNERSHIP

- corresponding “text piece”
  (predicate + all SRL arguments)
  “China purchased two nuclear submarines from Russia last month”

- predefined question for each argument type
  “What is bought?”
  (question for Artifact-Arg)

- answer span
  “two nuclear submarines”
  confidence ≥ 0.95?

- head span
  “submarines”
  is the Artifact-Arg

The questions are written based on the definition of each event type.
Experimental Setup

- **Dataset**: ACE-2005 (LDC2006T06), ERE (LDC2015E29)

- **Settings**:
  - *scratch*: the system performs all subtasks without any gold annotation
  - *gold TI*: gold trigger spans are given
  - *gold TI+TC*: gold trigger spans and types are given

- **Pretrained models**\(^3\):
  - Architecture: BERT/RoBERTa/BART - base/large
  - Pretraining data:

<table>
<thead>
<tr>
<th>Target EE Subtask</th>
<th>Pretraining Dataset</th>
<th>Pretraining Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trigger Extraction</td>
<td><strong>MNLI</strong> (Williams et al., 2018)</td>
<td>TE</td>
</tr>
<tr>
<td></td>
<td><strong>BoolQ</strong> (Clark et al., 2019)</td>
<td>Yes/No QA</td>
</tr>
<tr>
<td>Argument Extraction</td>
<td><strong>QAMR</strong> (Michael et al., 2018)</td>
<td>Extractive QA</td>
</tr>
<tr>
<td></td>
<td><strong>SQuAD2.0</strong> (Rajpurkar et al., 2018)</td>
<td>Extractive QA</td>
</tr>
</tbody>
</table>

Table 1: Datasets used to pretrain the TE/QA models.

\(^3\)Optimal configuration highlighted in green.
### Results: ACE-2005

<table>
<thead>
<tr>
<th>Setting</th>
<th>System</th>
<th>TI</th>
<th>TI+TC</th>
<th>AI</th>
<th>AI+AC</th>
</tr>
</thead>
<tbody>
<tr>
<td>scratch</td>
<td>Lin et al. 20</td>
<td>78.2</td>
<td>74.7</td>
<td>59.2</td>
<td>56.8</td>
</tr>
<tr>
<td>(supervised)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>scratch</td>
<td>Huang et al. 18</td>
<td>55.6</td>
<td>49.1</td>
<td>27.8</td>
<td>15.8</td>
</tr>
<tr>
<td>(zero-shot)</td>
<td>Zhang et al. 20</td>
<td>58.3</td>
<td>53.5</td>
<td>16.3</td>
<td>6.3</td>
</tr>
<tr>
<td>Ours</td>
<td>45.5</td>
<td>41.7</td>
<td>27.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gold TI</td>
<td>Huang et al. 18</td>
<td>-</td>
<td>33.5</td>
<td>-</td>
<td>14.7</td>
</tr>
<tr>
<td>(zero-shot)</td>
<td>Zhang et al. 20</td>
<td>-</td>
<td>82.9</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>-</td>
<td>83.7</td>
<td>38.9</td>
<td>24.2</td>
<td></td>
</tr>
<tr>
<td>gold TI+TC</td>
<td>Liu et al. 20</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25.8</td>
</tr>
<tr>
<td>(zero-shot)</td>
<td>Ours</td>
<td>-</td>
<td>-</td>
<td>44.3</td>
<td>27.4</td>
</tr>
</tbody>
</table>

Table 2: The F1 score on ACE-2005. SOTA results among zero-shot methods are in boldface.
Table 3: The F1 score on the ERE. The optimal model is chosen on ACE dev and directly evaluated on ERE.
Analysis

- **Remaining challenges**: Manually annotated in 100 wrong predictions

- **Error attribution**: 
  - *Model-Error*: the intrinsic fragility of pretrained TE/QA models
  - *Usage-Error*: our usage of the models
  - *Task-Error*: the task itself

- **Ablation study**: 
  To isolate their individual impact, we alter *certain conditions* that have caused the target error⁴, and see *how many errors are corrected after when predicting again*.

⁴See Section 5 of our paper for details.
Figure 1: Error types in trigger and argument extraction in 100 wrong predictions. The count sum exceeds 100 since a prediction can contain multiple types of error.
Figure 1: Error types in trigger and argument extraction in 100 wrong predictions. The count sum exceeds 100 since a prediction can contain multiple types of error.
Analysis: Example

- **Distracting Context** (18%): Usage-Error
  - e.g. “The woman’s parents ... found the decomposing body.”
    Gold type: Not a trigger       Predicted type: DIE

- **Insufficient Context** (19%): Usage-Error
  - e.g. “(Turkey sent 1,000 troops ... and said) it would send more”
    Gold type: TRANSPORT       Predicted type: TRANSFER-MONEY

- Ablation study: 18% **Distracting Context** errors and 59% **Insufficient Context** errors are corrected when predicting again.
Figure 1: Error types in trigger and argument extraction in 100 wrong predictions. The count sum exceeds 100 since a prediction can contain multiple types of error.
“Competitive” Entity (24%): Model-Error
- e.g. “A unit meets in confidential sessions to review terrorist activities in Europe.”

Question for Place-Arg: “Where is the meeting?”
Gold answer: No Answer  
Predicted answer: “Europe”

Non-competitive NA Questions (19%): Model-Error
- e.g. “Iraqi forces responded with artillery fire.”

Question for Time-Arg: “When is the fire?”
Gold answer: No Answer  
Predicted answer: “artillery”

Ablation study: Adding training data on NA questions (SQuAD2.0) even hurts the performance.\(^5\)

\(^5\)We propose and test three hypotheses behind this. See Section 5.2.2 of our paper for details.
Conclusions

- We propose the first complete zero-shot event extraction system via transfer learning from TE and QA.
- While QA/TE models perform exceptionally well on standard benchmarks (SQuAD, QAMR, MNLI), they do not generalize as expected when being used on event extraction datasets.
- We analyze the limited success and several main challenges of this promising approach, and point out future research directions.
Thank you for listening!