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

China has purchased two nuclear submarines from Russia last month.

Buyer-Arg Trigger Artifact-Arg Seller-Arg Time-Arg

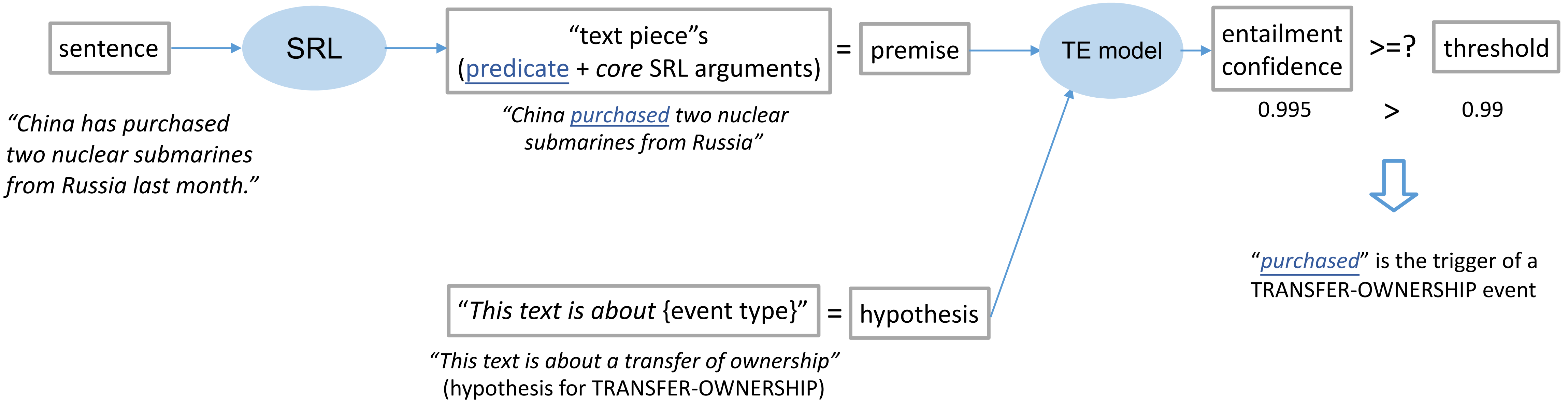
- **Event Extraction** (**EE**) = Trigger Identification (**TI**) + Trigger Classification (**TC**) + Argument Identification (**AI**) + Argument Classification (**AC**)

Context

- 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

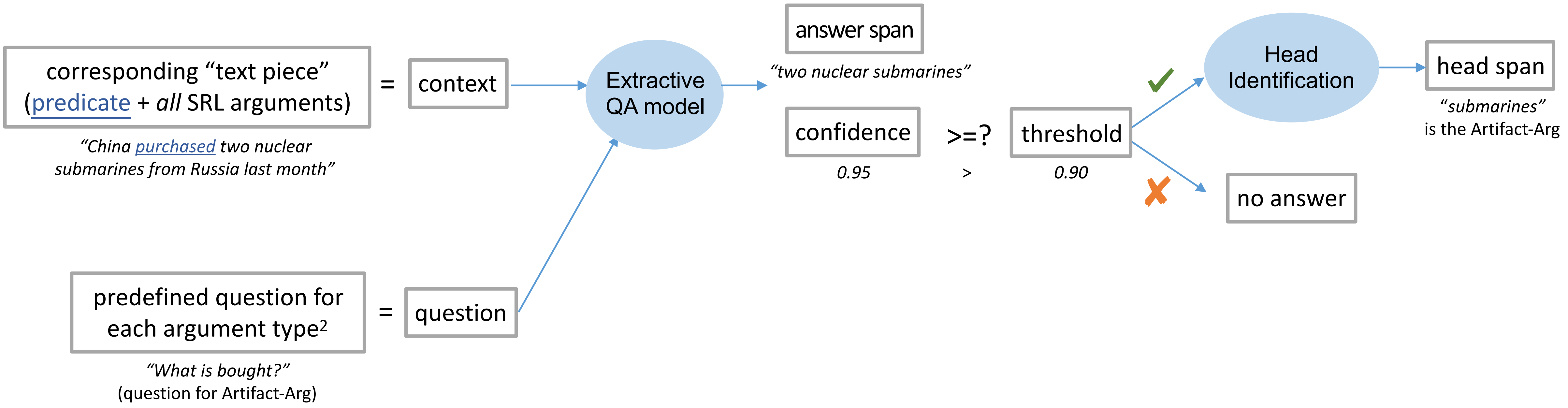


[1]An alternative uses Yes/No QA instead of TE, which is similar and thus not illustrated.

Approach: Argument Extraction

For each extracted trigger (span + event type),

e.g. "*purchased*" + TRANSFER-OWNERSHIP



[2]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:

Target EE Subtask	Pretraining Dataset	Pretraining Task
Trigger Extraction	MNLI (Williams et al., 2018)	TE
	BoolQ (Clark et al., 2019)	Yes/No QA
Argument Extraction	QAMR (Michael et al., 2018)	Extractive QA
	SQuAD2.0 (Rajpurkar et al., 2018)	Extractive QA

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

[3]Optimal configuration highlighted in green.

Results: ACE-2005

Setting	System	TI	TI+TC	AI	AI+AC
scratch **(supervised)**	Lin et al. 20	78.2	74.7	59.2	56.8
scratch **(zero-shot)**	Huang et al. 18	55.6	49.1	**27.8**	15.8
	Zhang et al. 20	**58.3**	**53.5**	16.3	6.3
	Ours	45.5	41.7	27.0	**16.8**
gold TI **(zero-shot)**	Huang et al. 18	-	33.5	-	14.7
	Zhang et al. 20	-	82.9	-	-
	Ours	-	**83.7**	**38.9**	**24.2**
gold TI+TC **(zero-shot)**	Liu et al. 20	-	-	-	25.8
	Ours	-	-	**44.3**	**27.4**

Table 2: The F1 score on ACE-2005. SOTA results among zero-shot methods are in boldface.

Results: ERE

Setting	System	TI	TI+TC	AI	AI+AC
scratch **(supervised)**	Lin et al. 20	68.4	57.0	50.1	46.5
scratch	Ours	39.8	31.8	23.0	15.0
gold TI		-	58.4	30.8	18.8
gold TI+TC		-	-	47.9	27.5
(zero-shot)					

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 error[4], and see *how many errors are corrected after when predicting again*.

[4]See Section 5 of our paper for details.

Analysis

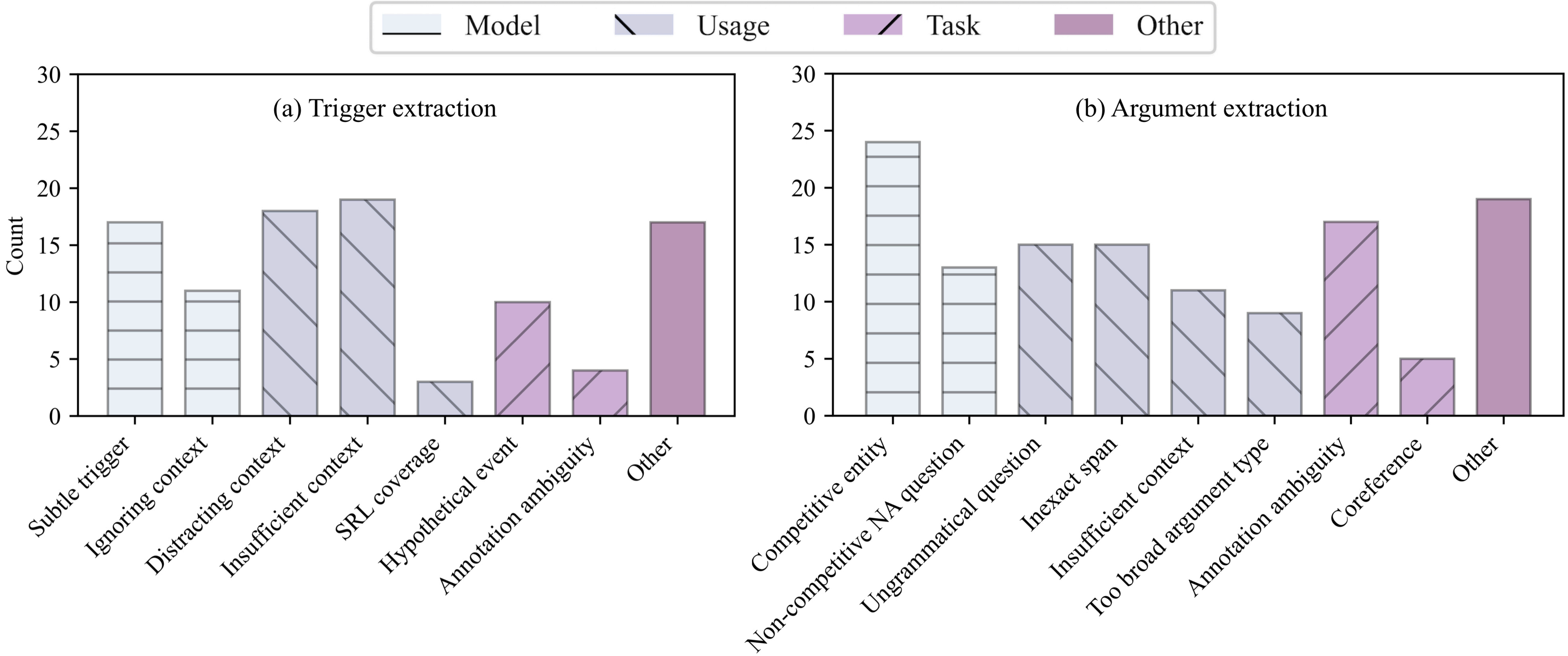


Figure 1: Error types in trigger and argument extraction in 100 wrong predictions. The count sum exceeds 100 since a prediction can contain multiple types of error.

Analysis

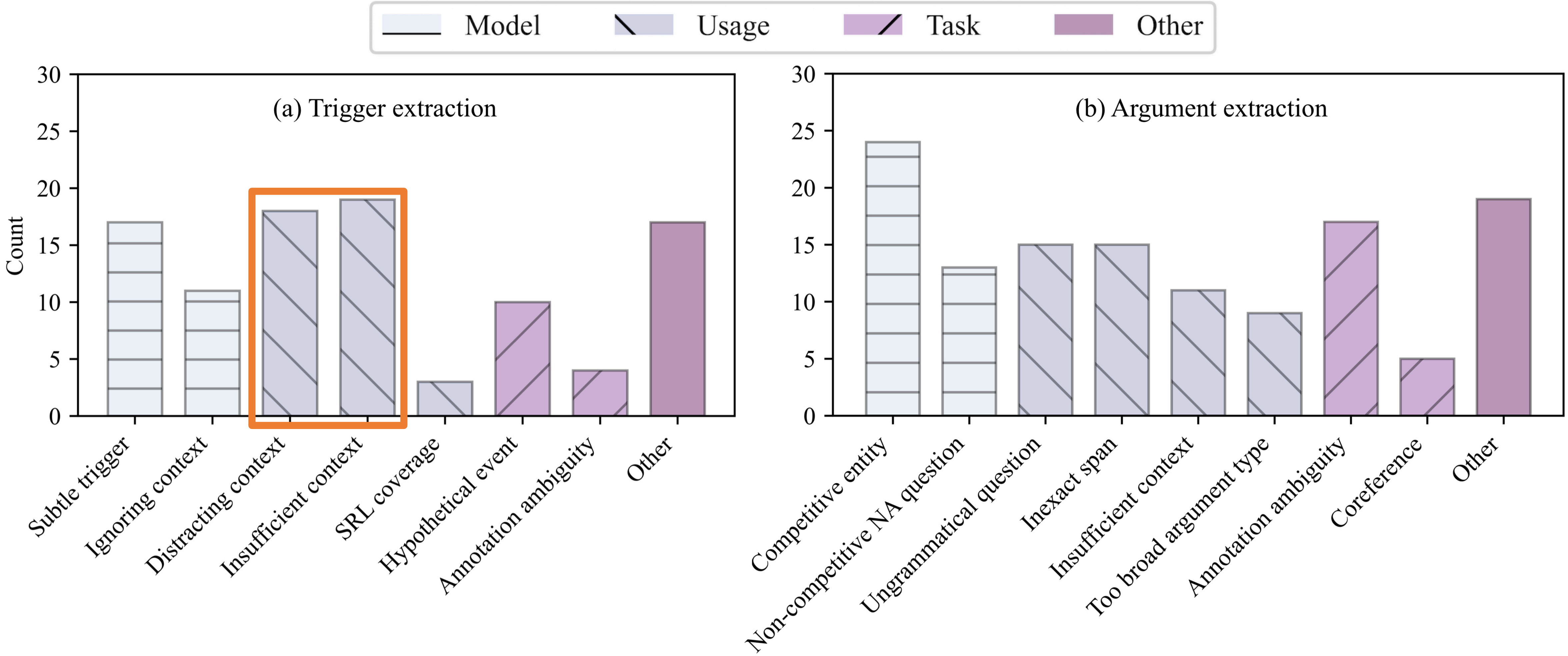


Figure 1: Error types in trigger and argument extraction in 100 wrong predictions. The count sum exceeds 100 since a prediction can contain multiple types of error.

Analysis: Example

- **Distracting Context** (18%): Usage-Error
  - e.g. “The woman’s parents ... found the <u>decomposing</u> 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 <u>send</u> more”
    Gold type: TRANSPORT Predicted type: TRANSFER-MONEY
- Ablation study: 18% **Distracting Context** errors and 59% **Insufficient Context** errors are corrected when predicting again.

Analysis

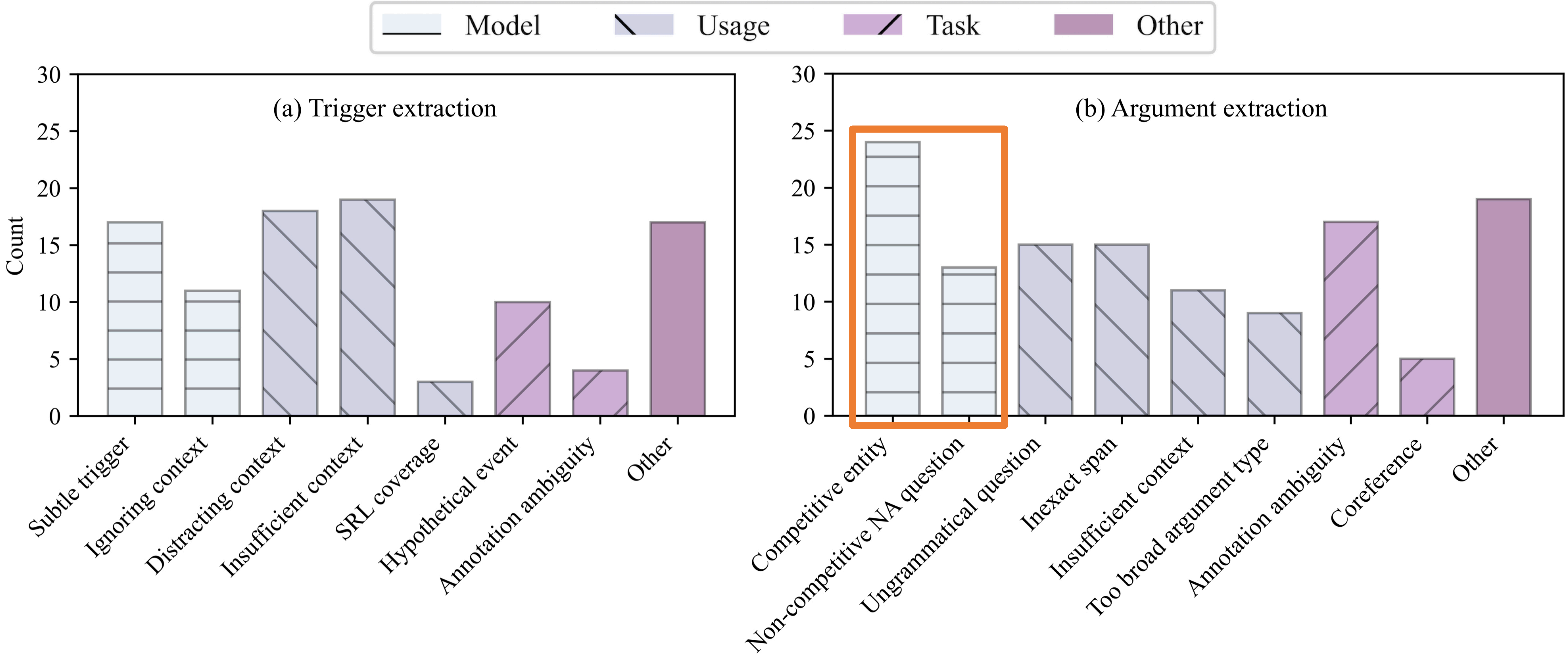


Figure 1: Error types in trigger and argument extraction in 100 wrong predictions. The count sum exceeds 100 since a prediction can contain multiple types of error.

Analysis: Example

- “*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!