Transfer Learning for IE

New Frontiers of Information Extraction (Part IV)

Lifu Huang
Computer Science Department
Virginia Tech

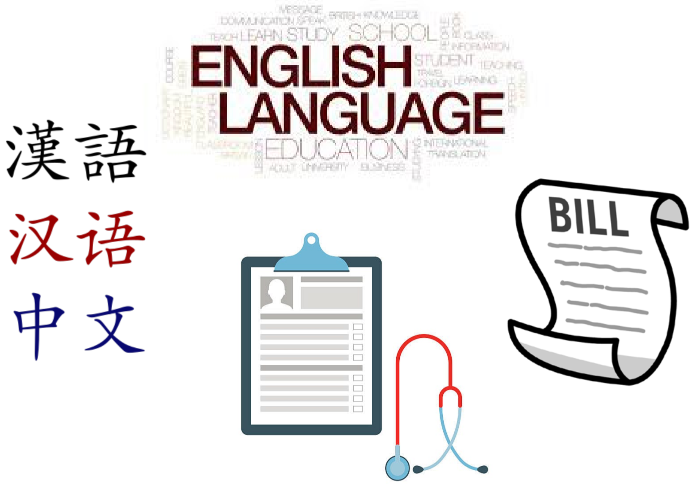
July 2022

NAACL Tutorials

New Frontiers of Information Extraction

Why Transferability is Important

- Current status of information extraction
  - Domains: news, biomedical, clinical, legal, agriculture
  - Languages: English, Chinese, Spanish, Arabic
  - Number of Target Types: 3-100+ for entity recognition, ~100 for relation extraction, 33/38 for event extraction
- However, for other languages and domains, learning resources are insufficient.



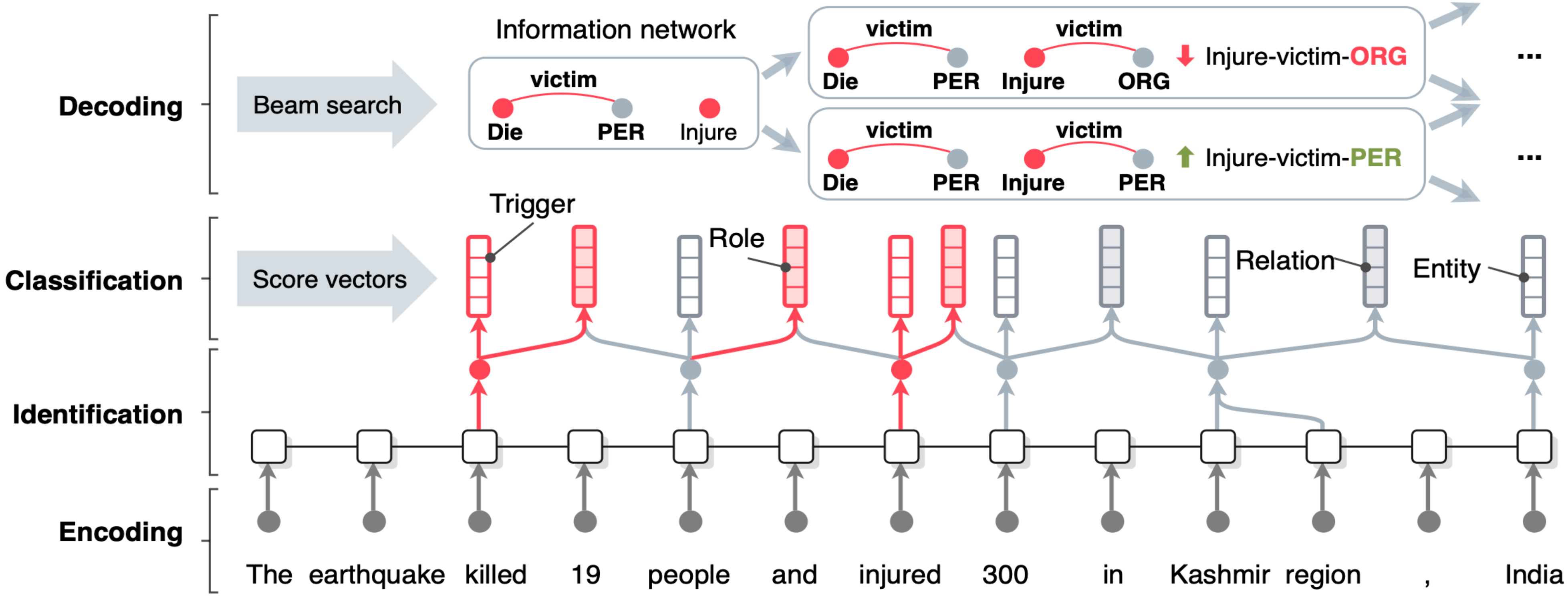
High-resource domains

Transfer

Low-resource domains

A "Typical" Neural Model for IE

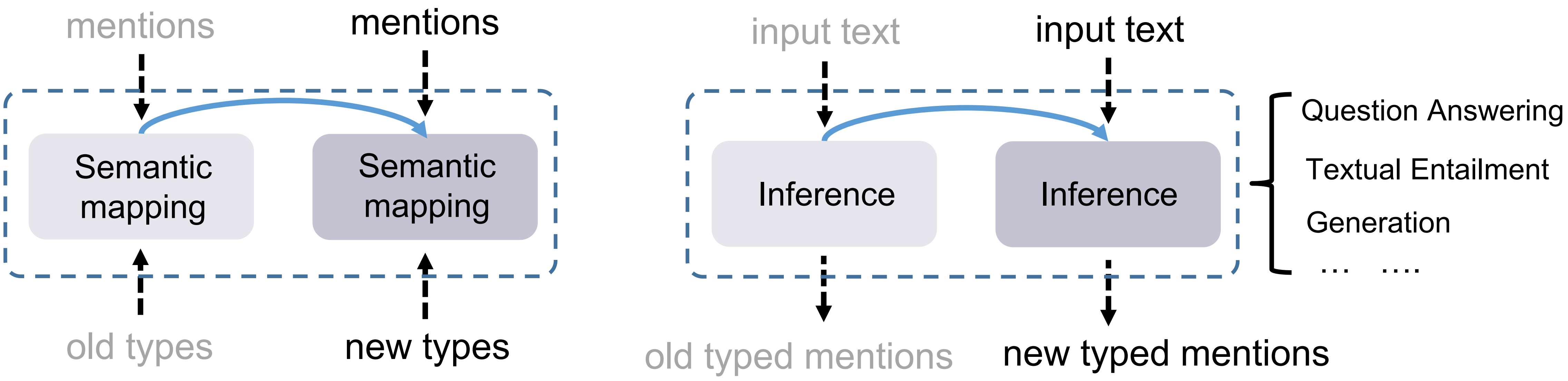
- Top-down classification: given a text, the model aims to classify each token or each pair of tokens into one of the target types
  - Pros: can extract mentions with high quality
  - Cons: require a large amount of annotations; cannot transfer to new domains or languages



OneIE
(Lin et al., 2020)

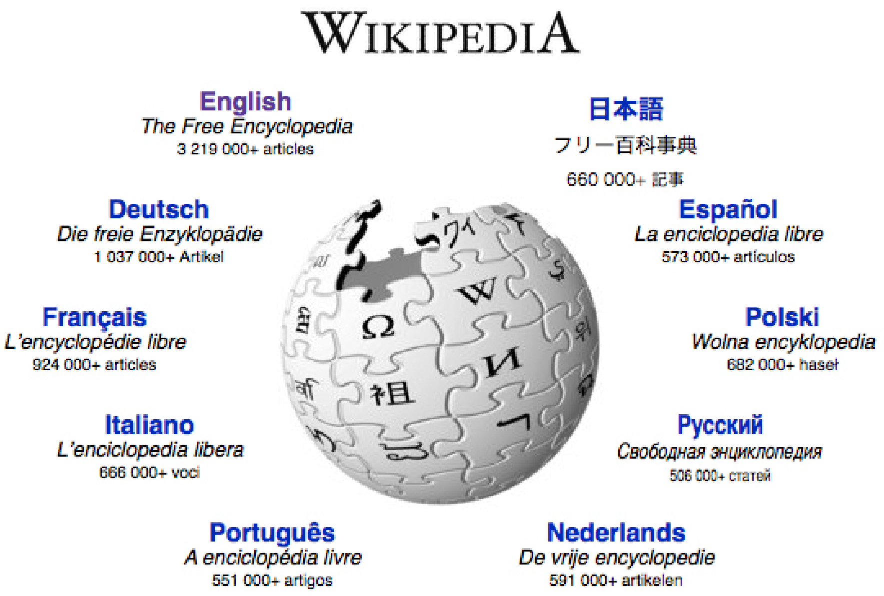
Challenge 1: Cross-type Transfer

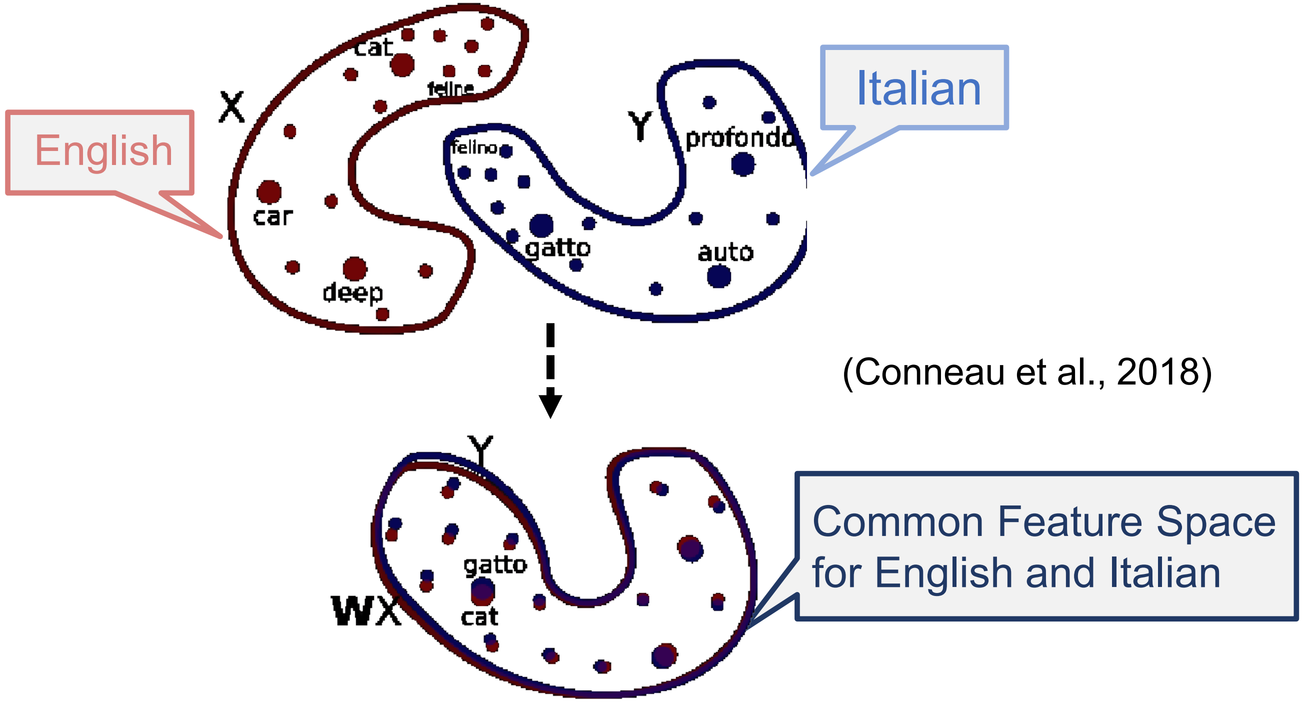
- How to transfer the knowledge and resources from old types to new types with little to no annotations?
  - Type-agnostic semantic mapping between mentions and types – (common semantic space for both mentions and types)
  - Type-agnostic inference from unstructured text to (structured) mentions



Challenge 2: Cross-lingual Transfer

- How to transfer the knowledge and resources across languages, especially from high-resource languages to low-resource languages?
  - Language universal resources, e.g., Wikipedia markups, linguistic knowledge bases, data annotation projection
  - Common semantic or feature space across languages

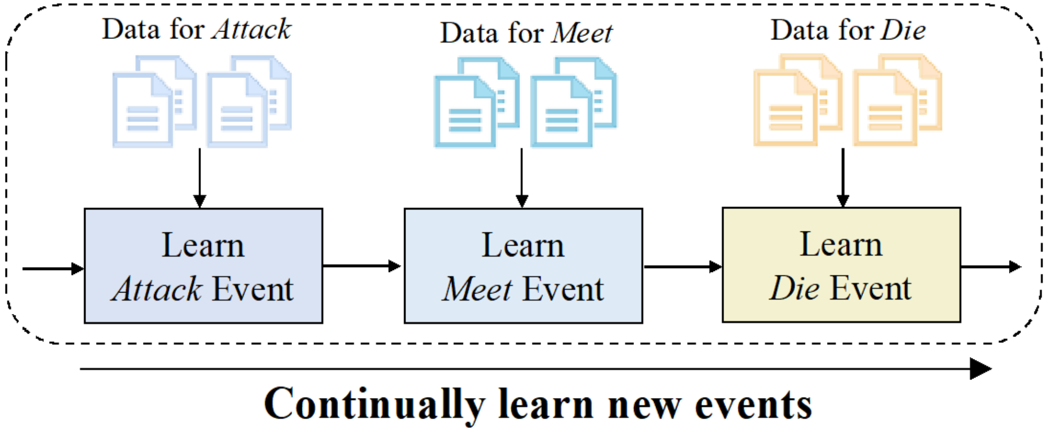




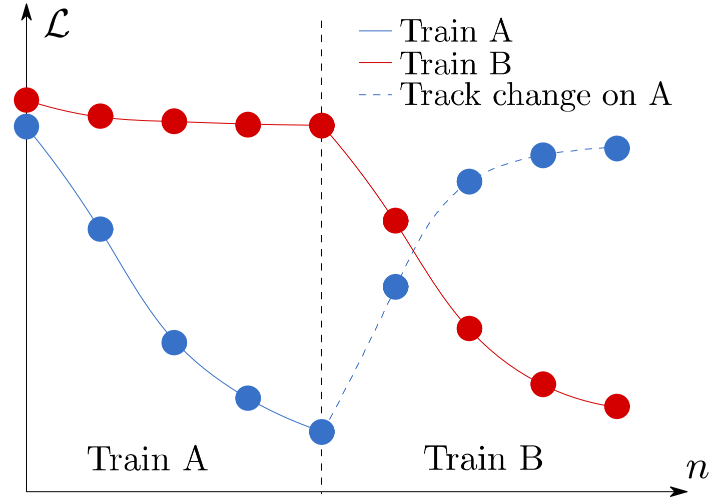
(Conneau et al., 2018)

Challenge 3: Continual Learning

- How to continually update the model on new annotations or tasks while retaining the capability learned from old tasks?
  - Catastrophic Forgetting: the model's performance on previously learned tasks significantly drops after it is trained on new data
    - Solutions: experience replay, knowledge distillation, regularization, task-specific adapter
  - Knowledge Transfer: transfer the knowledge from old tasks to new tasks



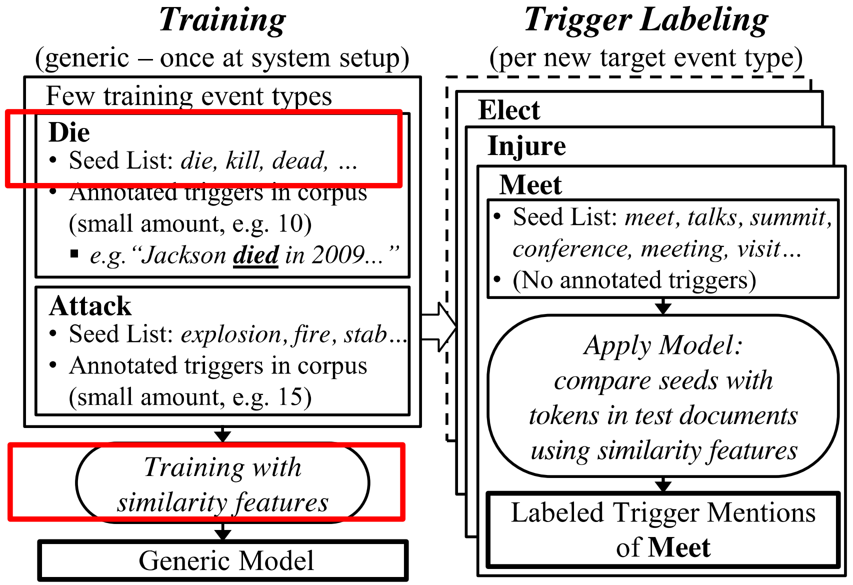
(Cao et al., 2020)



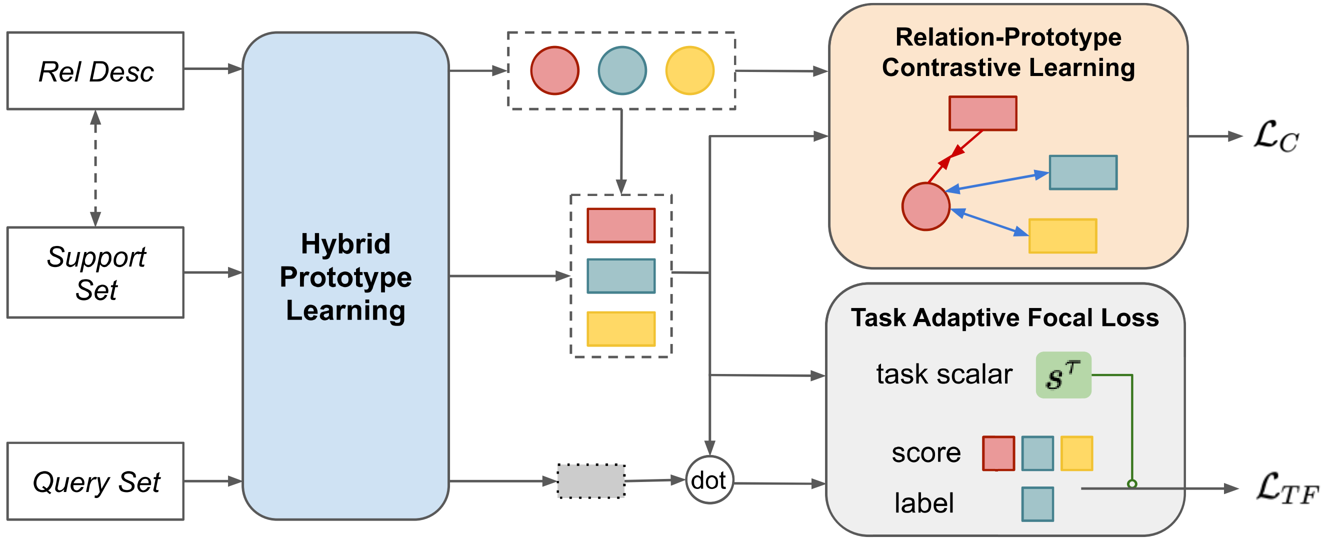
(Wiewel et al., 2019)

Cross-type Transfer: Type-agnostic Semantic Mapping

- Learning label representations based on a few seed examples, e.g., *triggers* for event extraction, *entity-relation instances* for relation extraction



(Bronstein et al., 2015)

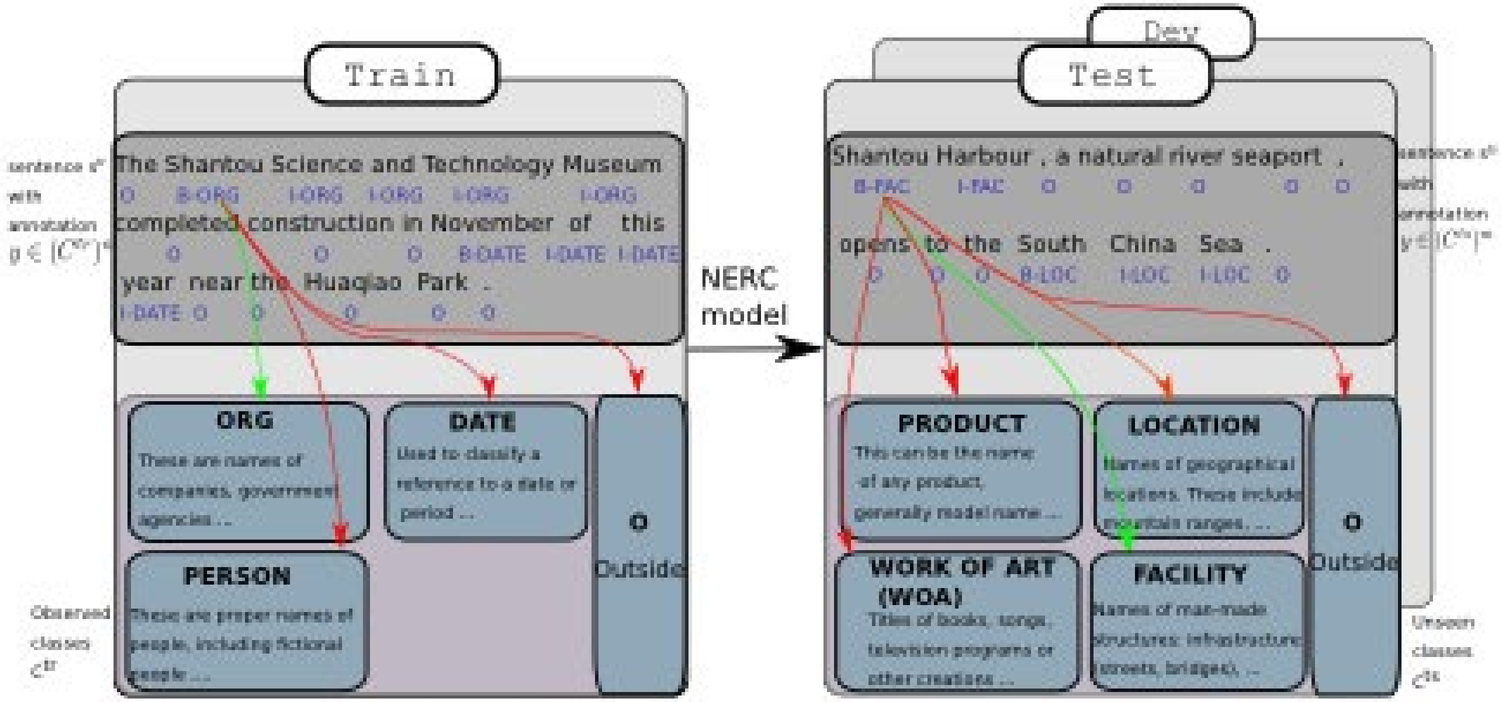


(Han et al., 2021)

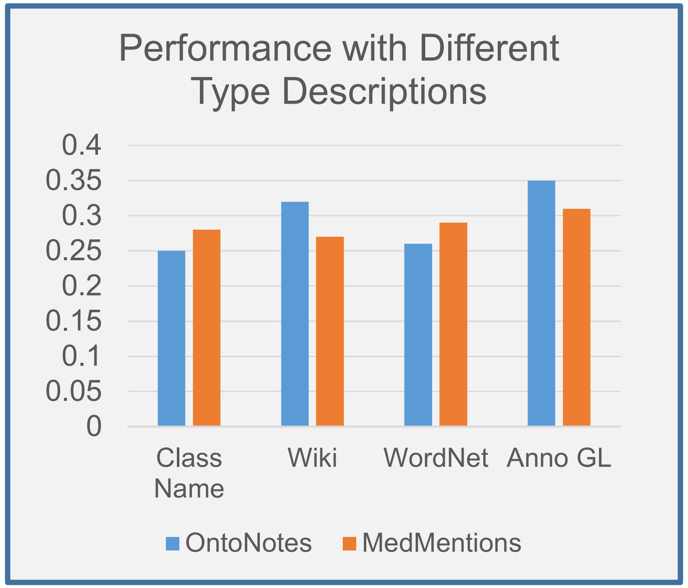
How to select the best seed examples?
- semantic popularity
- TF-IDF
- …

Cross-type Transfer: Type-agnostic Semantic Mapping

- Learning label representations based on type descriptions
  - Cross-attention Encoding: for each token in an input sentence, learn a type-specific representation by concatenating the sentence with the type description
  - Modeling the negative class (*other*): for each token, learn a negative class specific representation based on the max-pooling of all type-specific representations



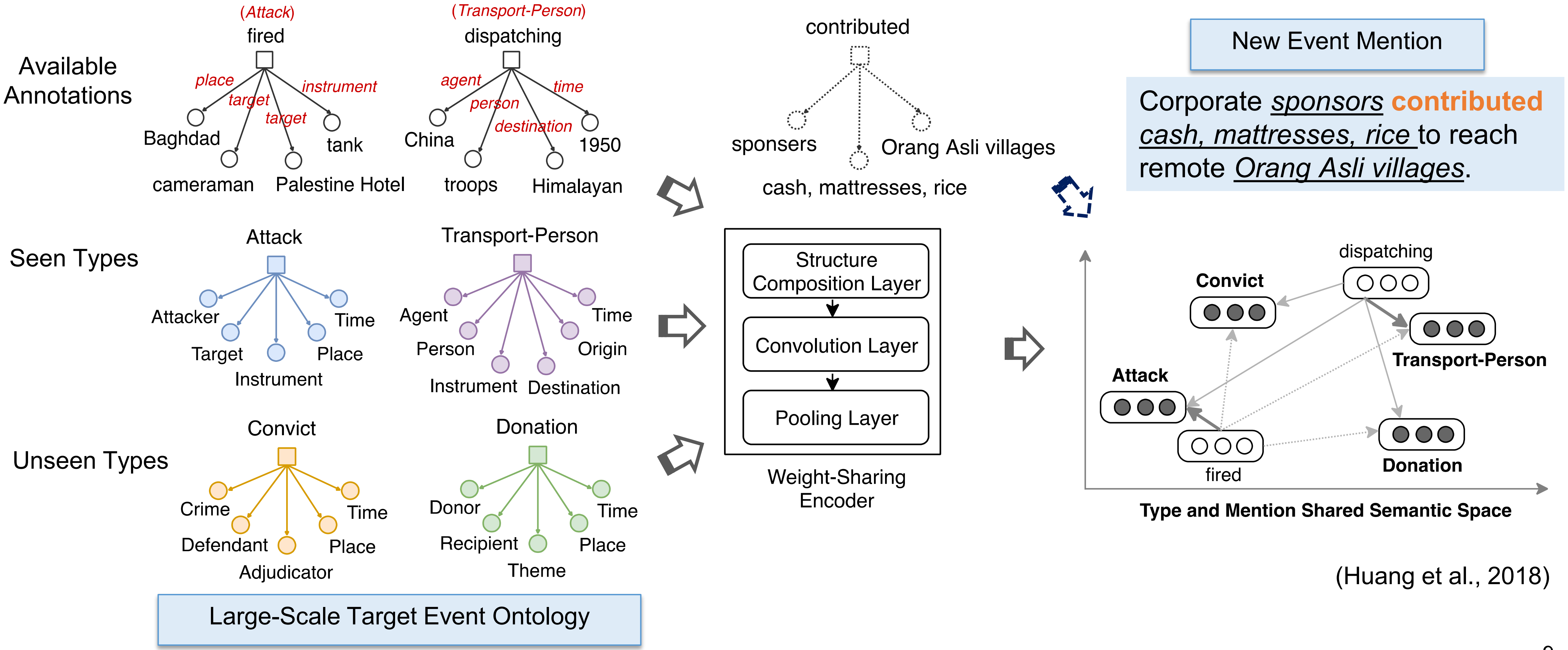
(Aly et al., 2021)



How to best describe the types?

Cross-type Transfer: Type-agnostic Semantic Mapping

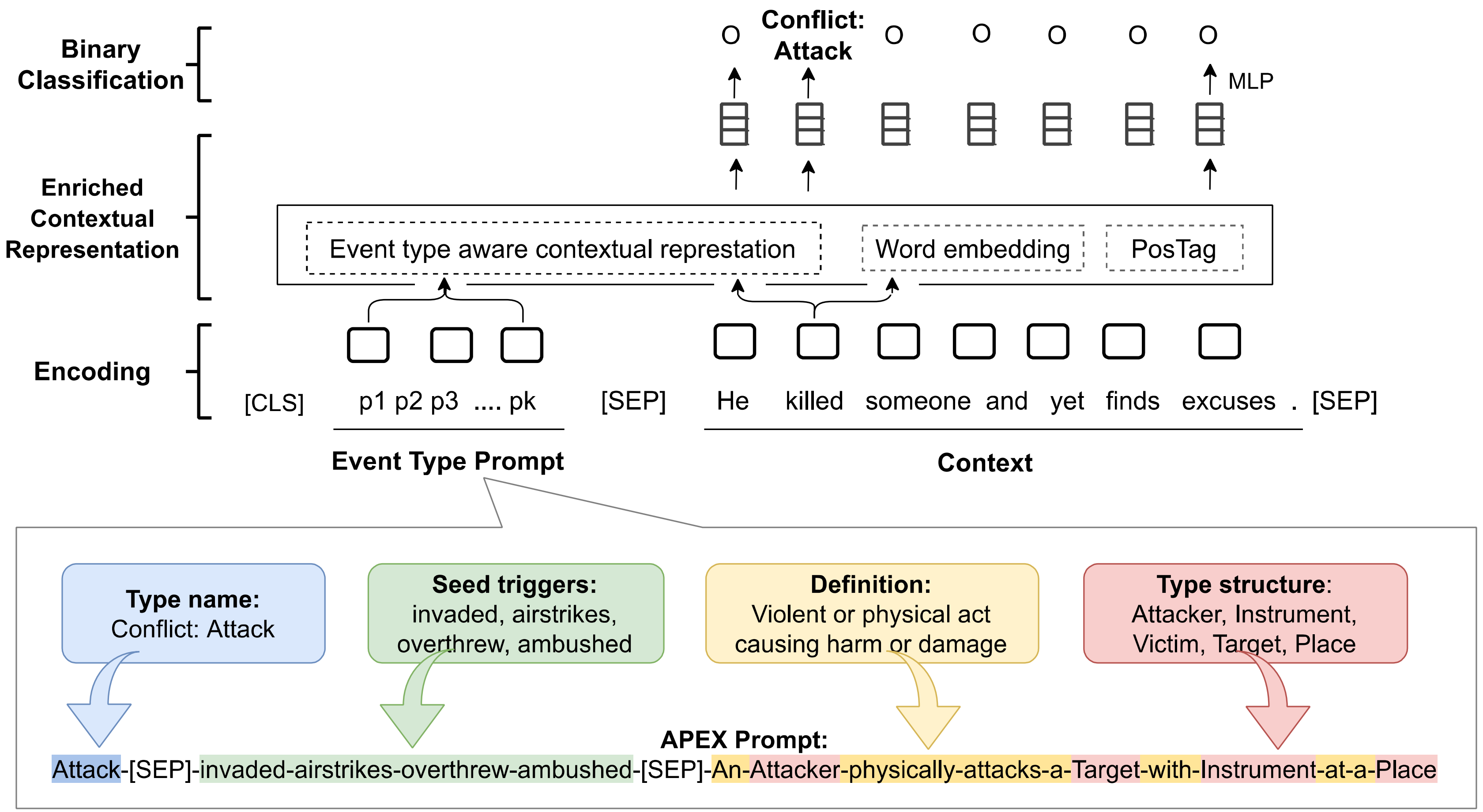
- Learning label and mention representations based on structures



(Huang et al., 2018)

Cross-type Transfer: Type-agnostic Semantic Mapping

- Which form provides the best label representations?
  - Detect mentions for each type by taking a type specific representation as a prompt (Wang et al., 2022)

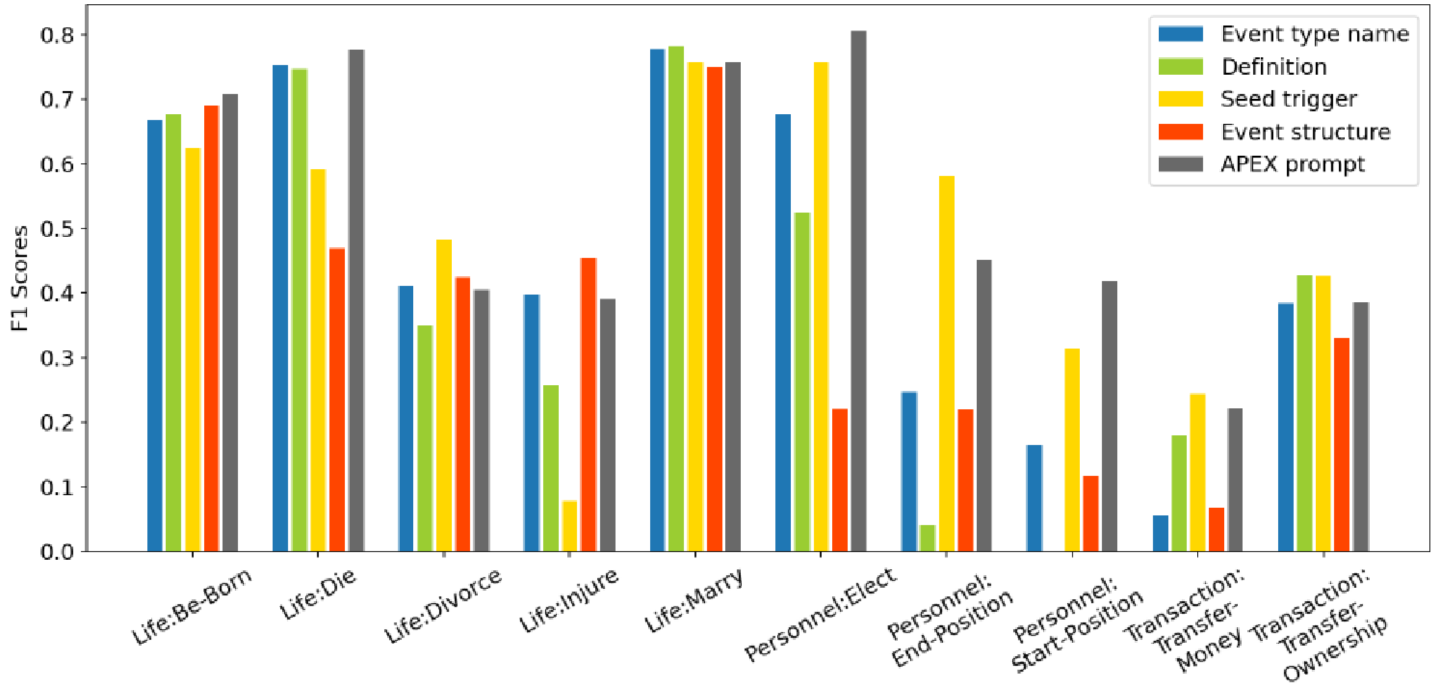


- (a) Type Name
- (b) Seed Examples
- (c) Definition
- (d) Type Structure
- (e) Soft Prompt
- (f) APEX Prompt (combination of a-d)

e.g., "*Attack*"

Cross-type Transfer: Type-agnostic Semantic Mapping

- Which form provides the best label representations?



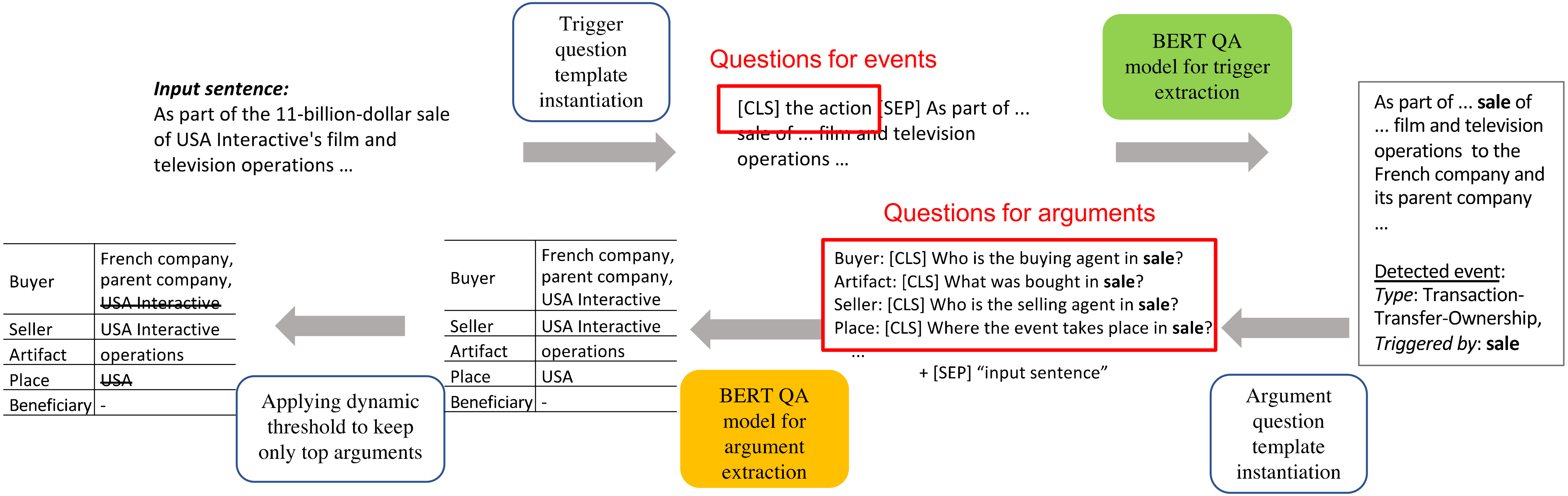
Performance on all novel event types of ACE under Zero-shot transfer

(Wang et al., 2022)

- **Seeds Triggers** are not always selected as the best
  - e.g., *extermination* for Life:Die
  - e.g., *paralyzed*, *dismember* for Life:Injure
- It's hard to determine if the **definition** is appropriate
  - e.g., "*a person entity begins working or change office*" for Personnel:Start-Position.
- **APEX Prompt** generally performs well

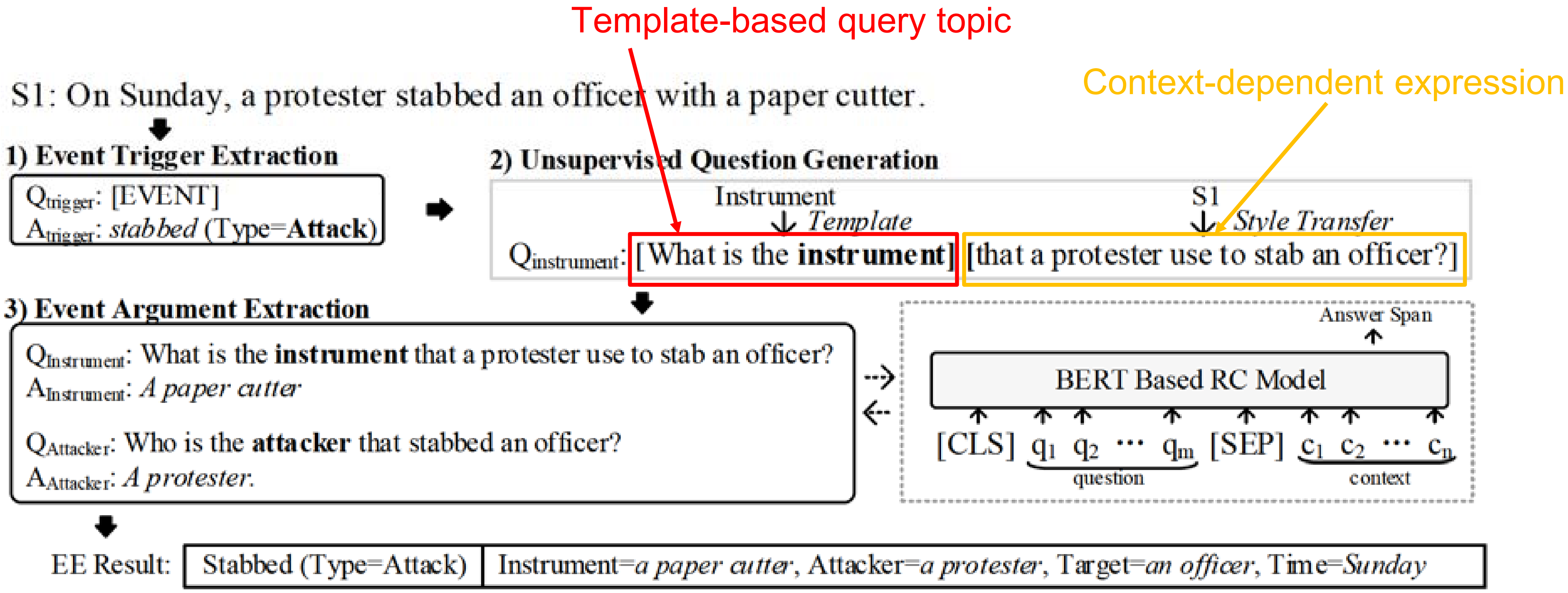
Cross-type Transfer: QA-based Event Extraction

- Questions are constructed based on templates for each role and the predicted answer serves as the extracted argument (Du and Cardie, 2020)
  - The input sequences for the two QA models share a standard BERT-style format: **[CLS] <question> [SEP] <sentence> [SEP]**



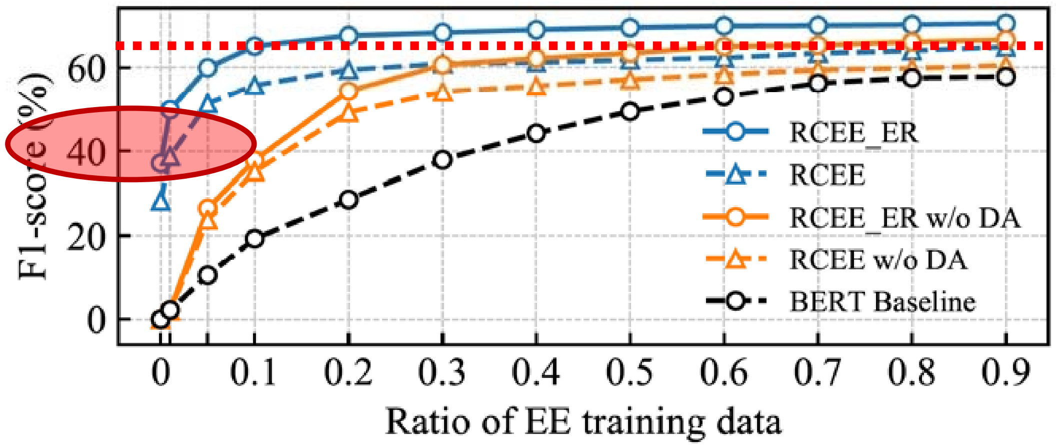
Cross-type Transfer: QA-based Event Extraction

- Questions can also be automatically generated in an unsupervised way (Liu et al., 2020)



Cross-type Transfer: QA-based Event Extraction

- Impact of pretraining on MRC datasets
  - Using 10% of EE training data, the approach achieves comparable performance as the baseline without MRC-based pre-training that is trained on 70% of the training set.
  - Without using any event annotations, the approach still achieves 37% F-score under zero-shot transfer



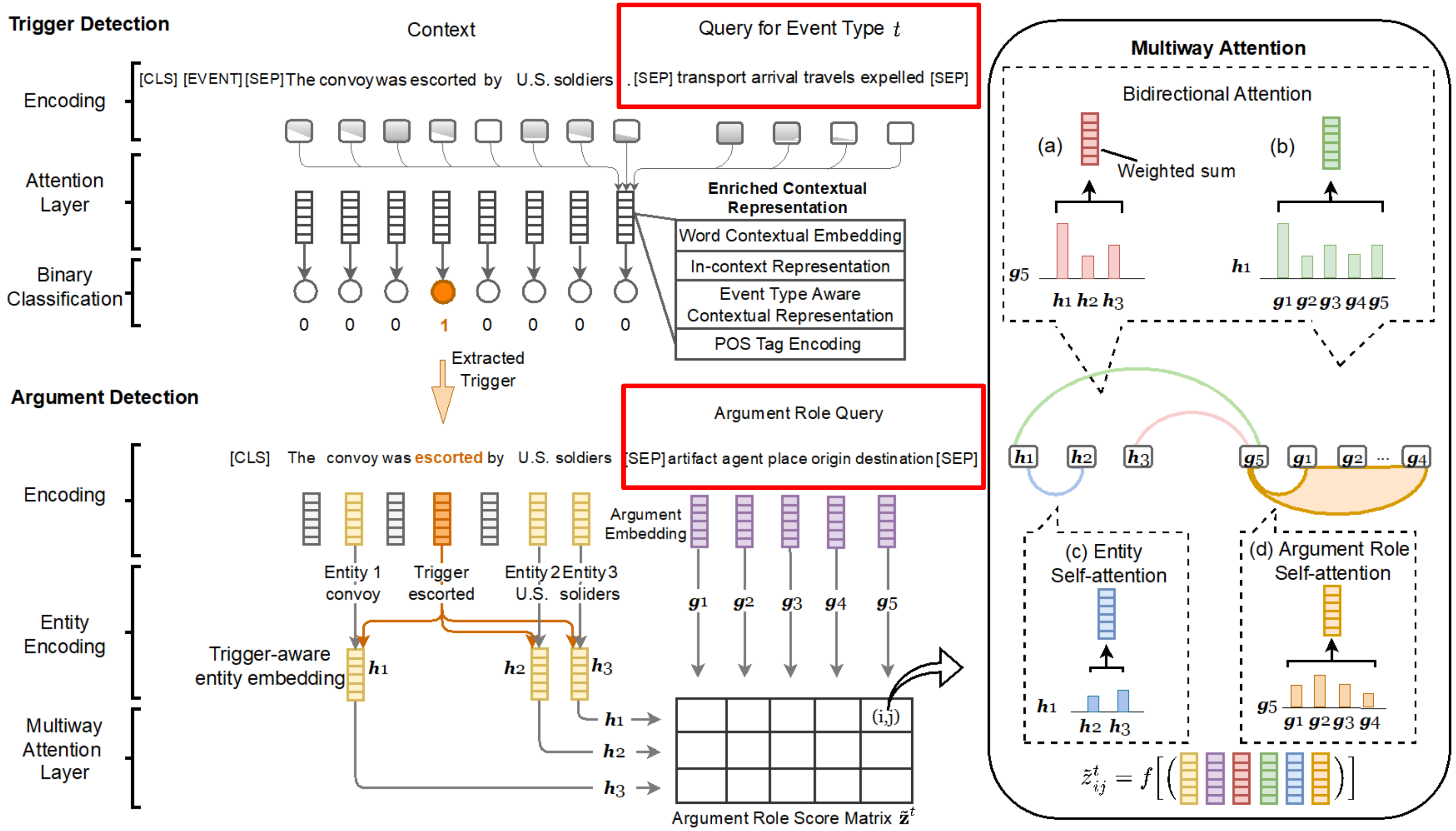
(Liu et al., 2020)

Performance on different ratios of EE training data

DA: pre-train the question answering model on MRC datasets

Cross-type Transfer: QA-based Event Extraction

- Query and Extract: directly take event type and argument roles as query to extract event triggers and arguments (Wang et al., 2022).



1. Encode a sentence and a type-specific query together
2. Learn a type-specific contextual rep. for each token based on attention mechanisms
3. Predict a binary label for each token

Cross-type Transfer: QA-based Event Extraction

- Query-and-Extract: rely more on semantic mapping between mentions and types rather than machine reading comprehension (Wang et al., 2022)
- Pros
  - Does not require any questions created for event types or argument roles
  - Can extract arguments for all possible argument roles at one time
- Cons
  - Cannot leverage available annotations for question answering

Model	Trigger Extraction	Argument Extraction
BERT_QA (Du and Cardie, 2020)	31.6	17.0
Query_and_Extract (Wang et al., 2022)	47.8	43.0

Performance on all novel event types of ACE under Zero-shot transfer (Wang et al., 2022)

Cross-type Transfer: Generation-based Event Extraction

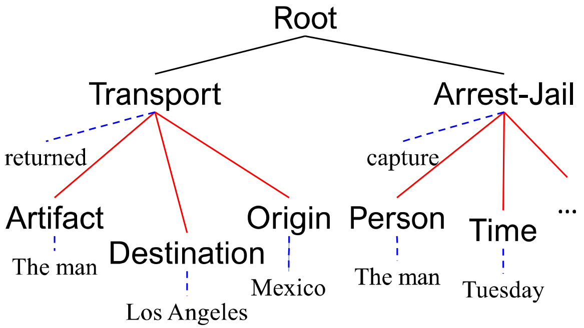
- **Text2Event**: translating *natural language text* to *event structures* with controllable sequence-to-structure generation (Lu et al., 2021)

The man returned to Los Angeles from Mexico following his capture Tuesday by bounty hunters.

Event Type	Transport
Trigger	returned
Artifact	The man
Destination	Los Angeles
Origin	Mexico

Event Type	Arrest-Jail
Trigger	capture
Person	The man
Time	Tuesday
Agent	bounty hunters

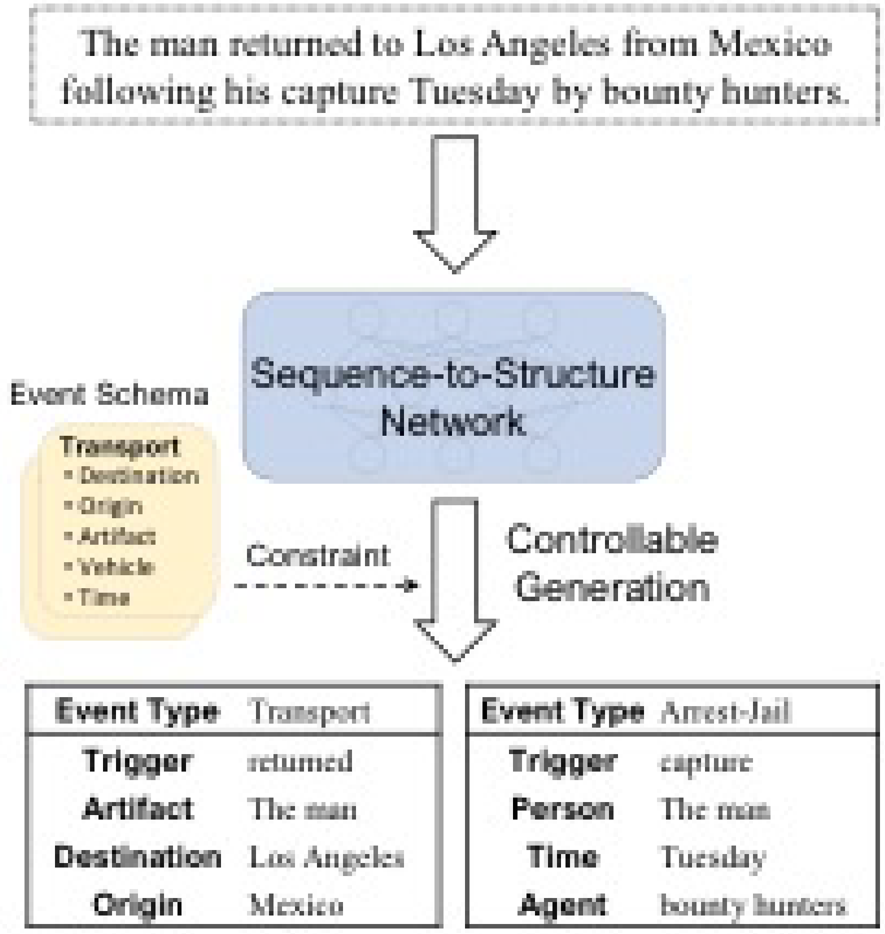
record to labeled tree



linearize

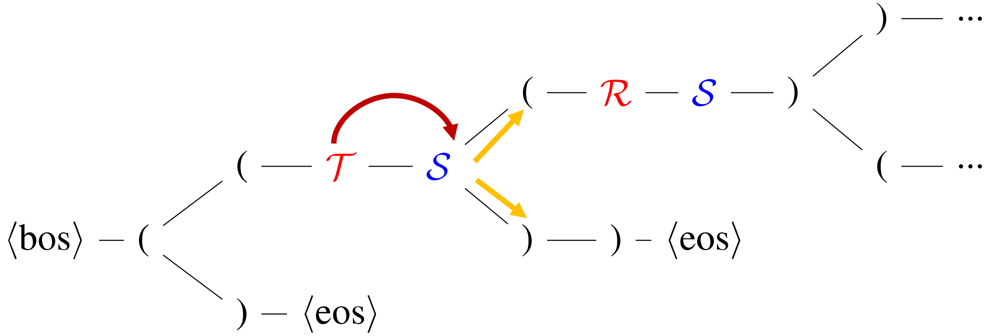
DFS

```
((Transport returned
  (Artifact The man)
  (Destination Los Angeles)
  (Origin Mexico))
 (Arrest-Jail capture
  (Person The man)
  (Time Tuesday)
  (Agent bounty hunters))
```

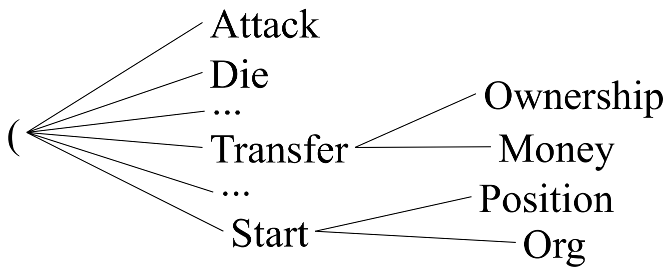


- **Text2Event**: translating *natural language text* to *event structures* with controllable sequence-to-structure generation (Lu et al., 2021)

Trie-based Constrained Decoding

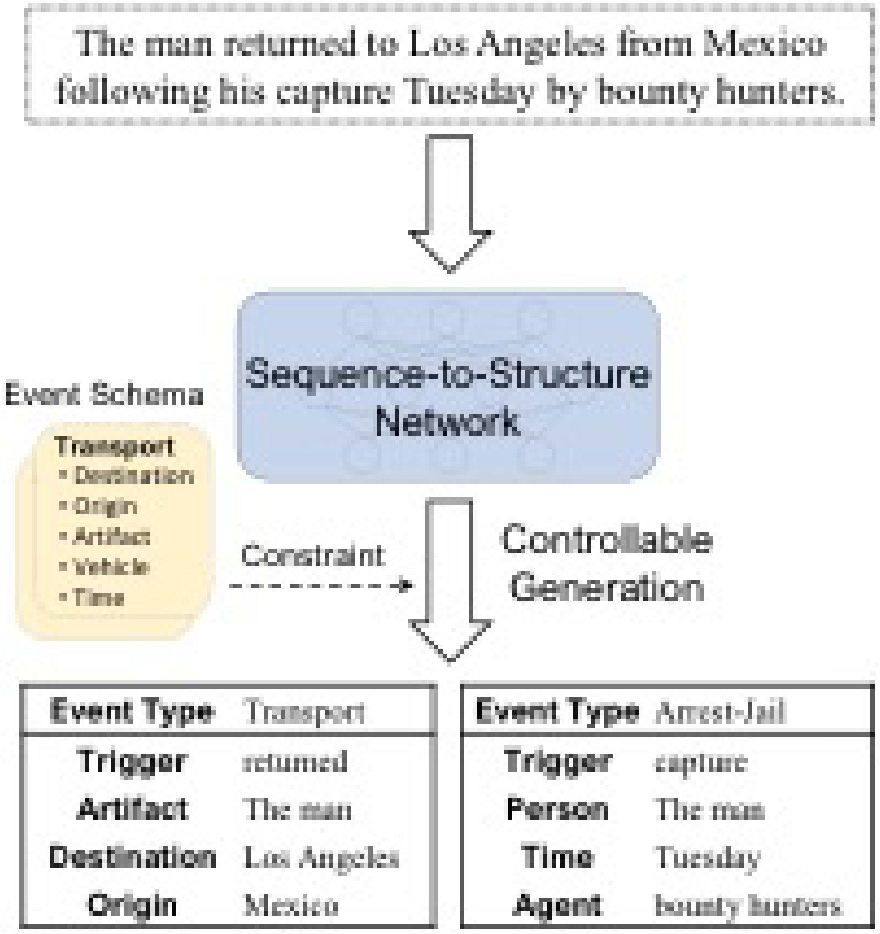


(a) The trie of event structure.



(b) The trie of event type $\mathcal{T}$.

$\mathcal{T}$ Event Types
$\mathcal{R}$ Argument Roles
$\mathcal{S}$ Mention Strings



Cross-type Transfer: Generation-based Event Extraction

- **Text2Event**: translating *natural language text* to *event structures* with controllable sequence-to-structure generation (Lu et al., 2021)

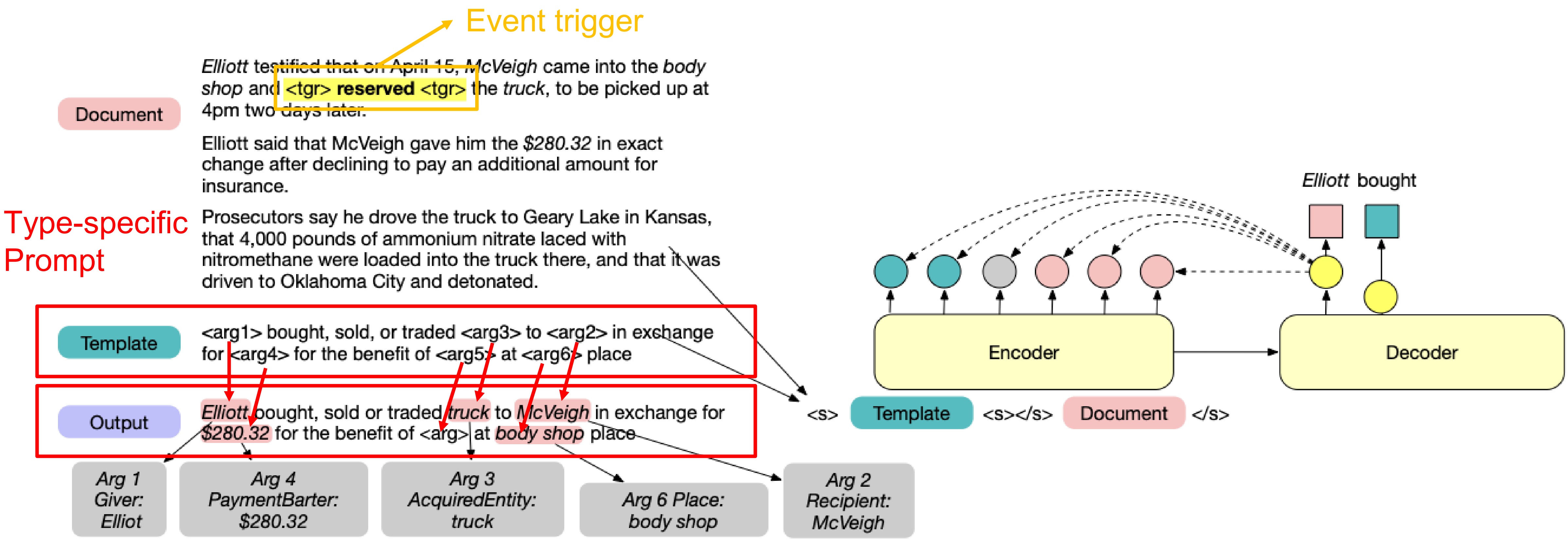
Settings	Trig-C P	Trig-C R	Trig-C F1	Arg-C P	Arg-C R	Arg-C F1
OneIE (Token + Entity Annotation)						
Non-transfer	78.1	62.3	69.3	50.9	37.9	43.5
Transfer	78.9	61.7	69.2	57.1	40.0	47.0
Gain			-0.1			+3.5
EEQA (Token Annotation)						
Non-transfer	69.9	67.3	68.6	36.5	37.4	36.9
Transfer	79.5	61.7	69.5	33.9	41.2	37.2
Gain			+0.9			+0.3
TEXT2EVENT (Parallel Text-Record Annotation)						
Non-transfer	79.4	61.1	69.0	58.4	40.9	48.0
Transfer	82.1	65.3	72.7	58.8	45.4	51.2
Gain			+3.7			+3.2

Transfer: first pre-train the model on source types, and then fine-tune on the annotations of target types.

- More data efficient and can make better use of supervision signals
- Effectively transfer knowledge across different types

Cross-type Transfer: Generation-based Argument Extraction

- Event type specific prompts, e.g., a template-based event type description, can better guide the model to generate events/arguments (Li et al., 2021)
  - All arguments for one event can be extracted in a single pass.



Summary – Cross-type Transfer

	Pros	Cons
Semantic Mapping	- Easy to setup; - Require minimal resource;	- Difficult to find the globally optimal form to represent the target types;
Question Answering		
Generation		

Summary – Cross-type Transfer

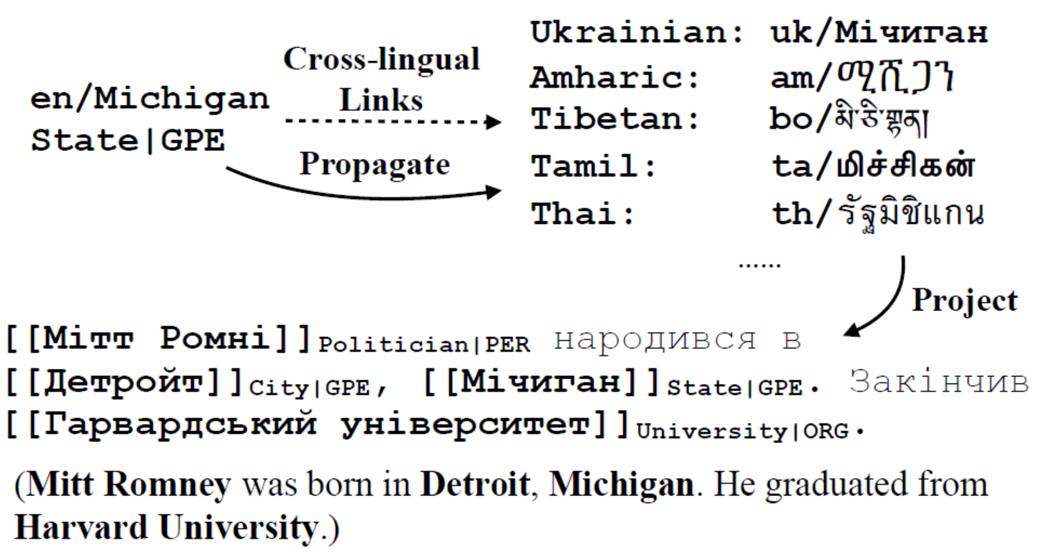
	Pros	Cons
Semantic Mapping	- Easy to setup; - Require minimal resource;	- Difficult to find the globally optimal form to represent the target types;
Question Answering	- Can leverage large-scale QA datasets; - Leverage the inference capability of pre-trained language models; - Does not require entity extraction for event extraction task;	- Require template or auto-generated questions as input, however it's hard to determine the optimal questions; - High computational cost as it can only extract for one event type or argument role at each time;
Generation		

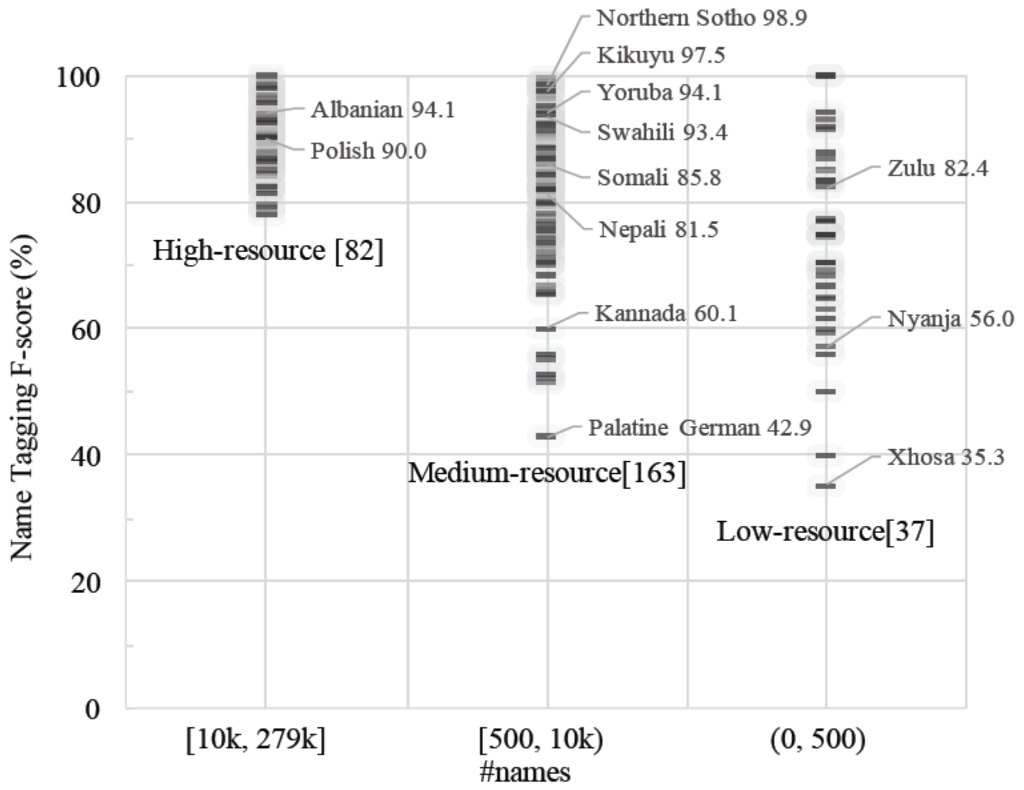
Summary – Cross-type Transfer

	Pros	Cons
Semantic Mapping	- Easy to setup; - Require minimal resource;	- Difficult to find the globally optimal form to represent the target types;
Question Answering	- Can leverage large-scale QA datasets; - Leverage the inference capability of pre-trained language models; - Does not require entity extraction for event extraction task;	- Require template or auto-generated questions as input, however it's hard to determine the optimal questions; - High computational cost as it can only extract for one event type or argument role at each time;
Generation	- Leverage the generation capability of pre-trained language models; - Computationally efficient: extract trigger and all arguments in a single pass;	- Hard to control; - Each type requires a carefully defined template which is hard to tell whether it's optimal or not;

Cross-lingual Transfer: Language Universal Resources

- Leveraging “silver standard” multilingual annotations from Wikipedia markups (Pan et al., 2017)

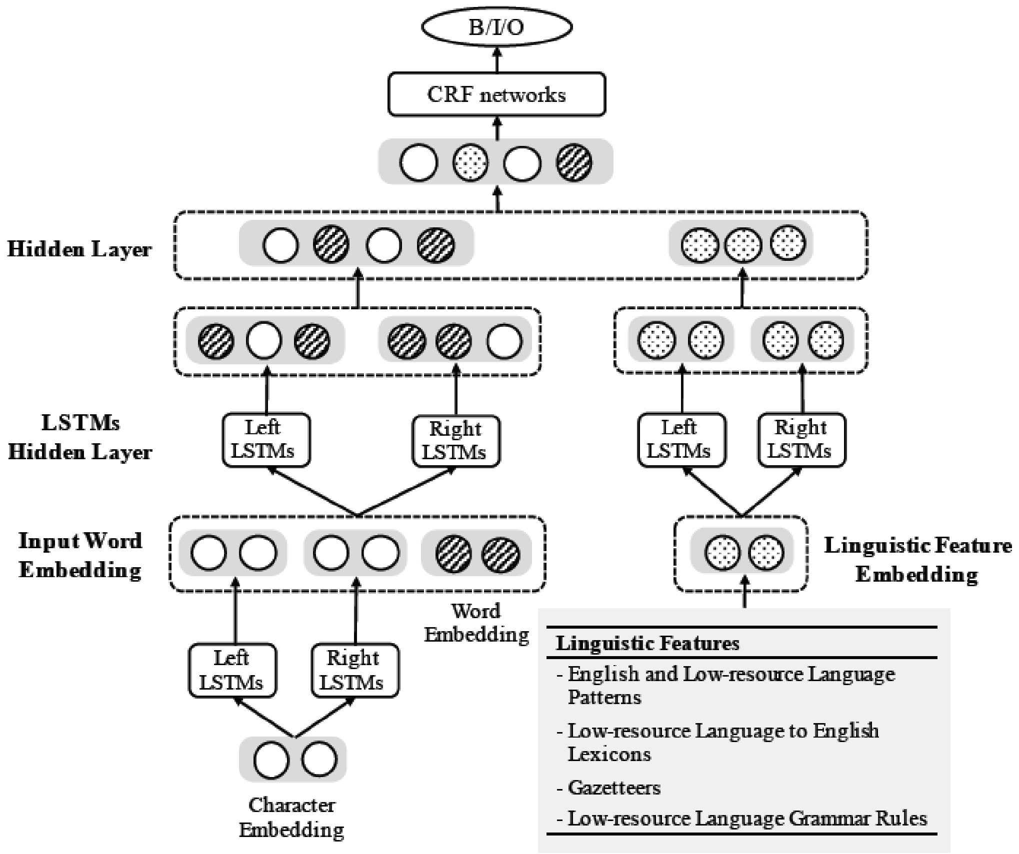


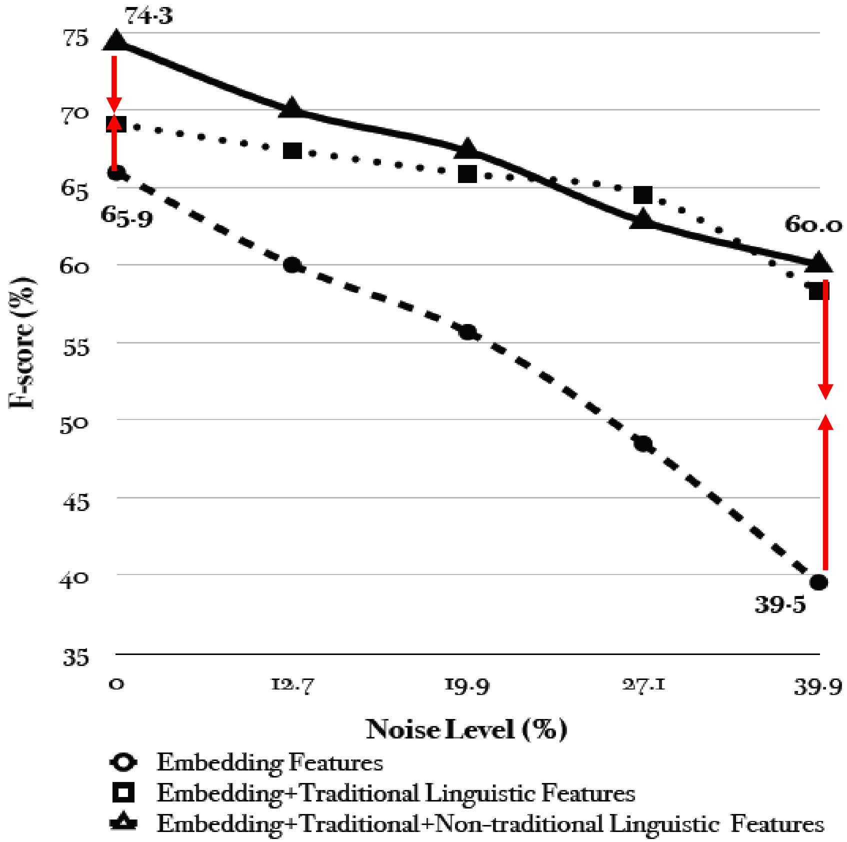


- Self-training to propagate labels
- However, such training data is usually noisy

Cross-lingual Transfer: Language Universal Resources

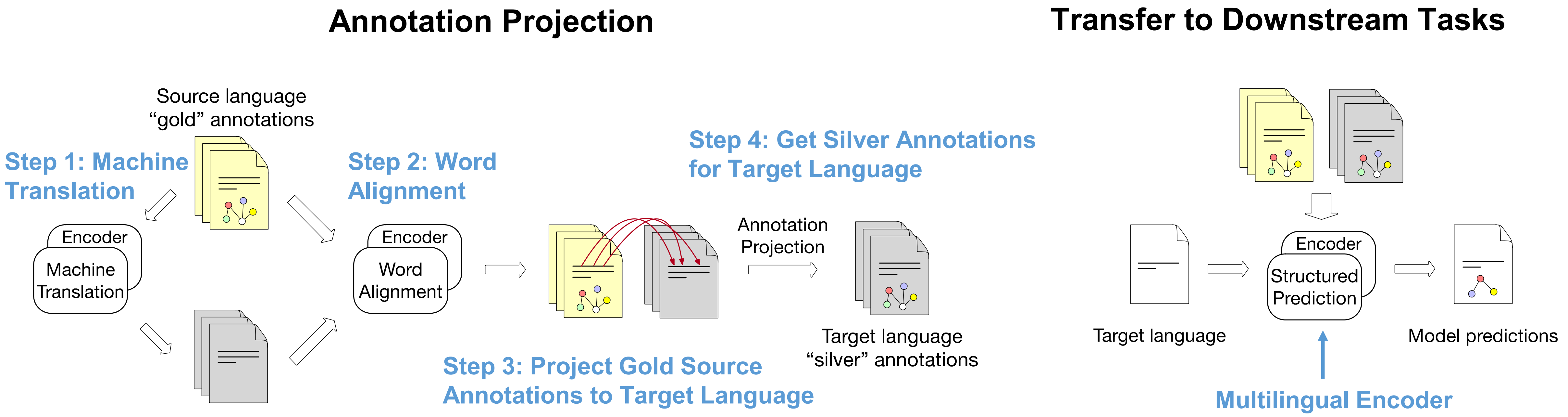
- Making DNNs more robust to the data noise by integrating language-universal linguistic features (Zhang et al., 2017)





Cross-lingual Transfer: Language Universal Resources

- Cross-lingual Annotation Projection through machine translation, statistical and neural word aligners, dictionaries, multilingual pretrained language models, etc.



(Yarmohammadi et al., 2021)

Cross-lingual Transfer: Language Universal Resources

- Performance of Zero-shot Cross-lingual Transfer w/ and w/o Data Projection

	MT	Align	Entity	Relation	Trig-I	Trig-C	Arg-I	Arg-C	AVG
mBERT (base, multilingual)									
(Z)	-	-	59.3	25.7	23.8	22.2	17.2	13.8	27.0
(A)	public	FA	-2.2	-13.9	+6.5	+2.5	+10.7	+11.5	+2.5
(B)	public	mBERT	-6.2	-5.1	+16.0	+10.6	+11.5	+12.1	+6.5
(B)	public	XLM-R	-12.7	-17.9	+11.1	+8.0	+8.5	+8.1	+0.9
(C)	public	mBERT_{ft}	-1.1	+0.9	+12.8	+9.8	+10.9	+13.6	+7.8
(C)	public	XLM-R_{ft}	-0.1	-4.2	+16.0	+11.9	+11.2	+11.3	+7.7
(C)	public	$\text{XLM-R}_{ft.s}$	-0.2	-1.6	+13.4	+11.5	+9.0	+11.7	+7.3
(D)	public	GBv4_{ft}	-1.9	+2.8	+14.3	+9.9	+12.7	+13.3	+8.5
(D)	public	L128K_{ft}	-1.7	+0.6	+11.6	+8.3	+10.7	+9.0	+6.4
(D)	public	$\text{L128K}_{ft.s}$	-1.3	+3.6	+12.7	+8.4	+8.3	+10.3	+7.0
(E)	GBv4	mBERT_{ft}	+1.0	+4.7	+13.6	+10.3	+9.3	+11.3	+8.4
(E)	GBv4	XLM-R_{ft}	-0.5	+5.5	+12.6	+10.8	+15.1	+14.4	+9.6
(E)	L128K	mBERT_{ft}	+2.6	+5.2	+12.9	+13.4	+18.8	+19.6	+12.1
(E)	L128K	XLM-R_{ft}	+2.5	+6.3	+11.2	+5.1	+17.1	+19.2	+10.2

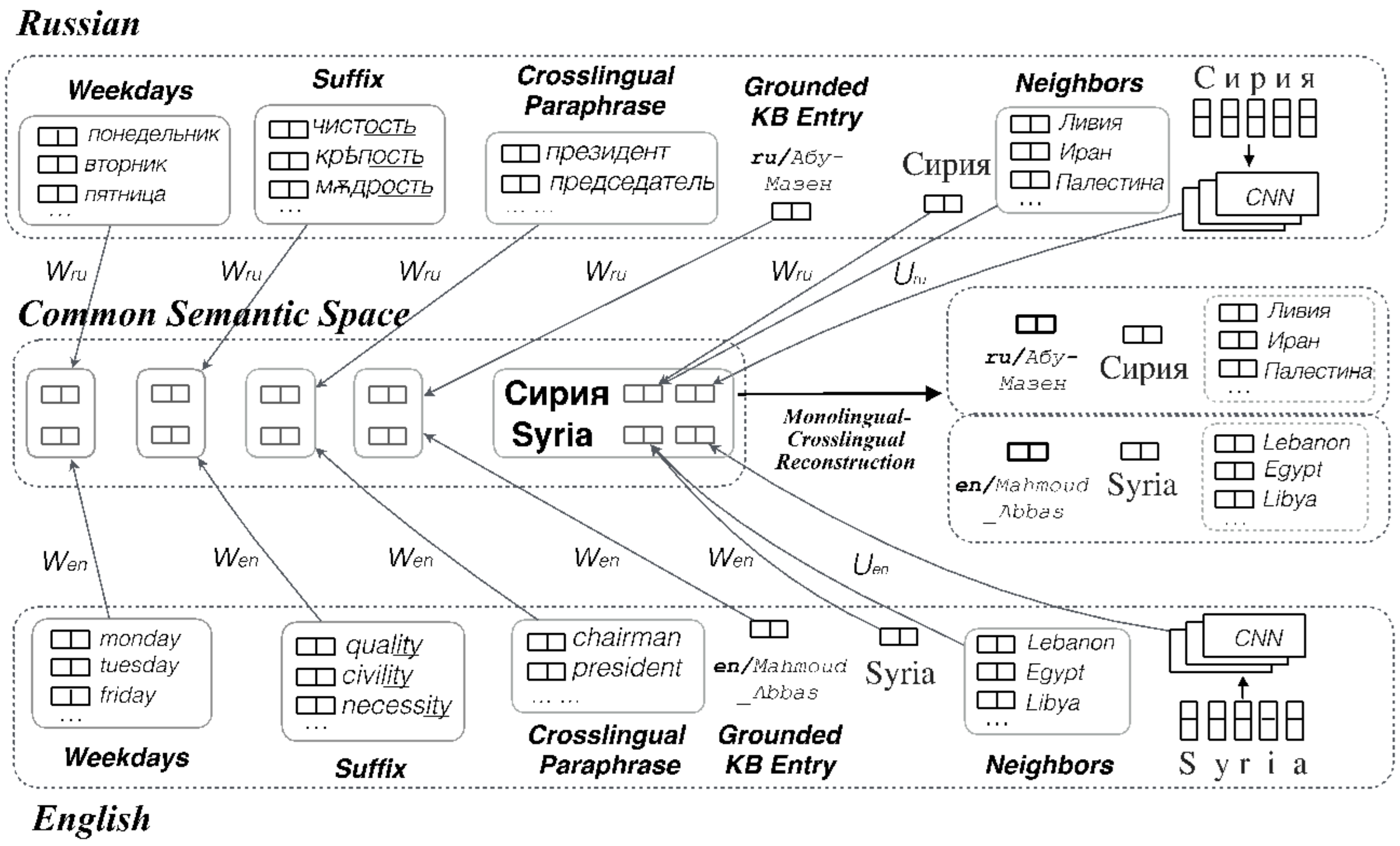
Zero-shot cross-lingual transfer w/o data projection

Data projection generally helps, no matter which machine translation or word aligners are used

Performance on Arabic Information Extraction Tasks with Cross-lingual Transfer (English→Arabic)

(Yarmohammadi et al., 2021)

- Learning language-agnostic semantic features – Multilingual Common Semantic Space

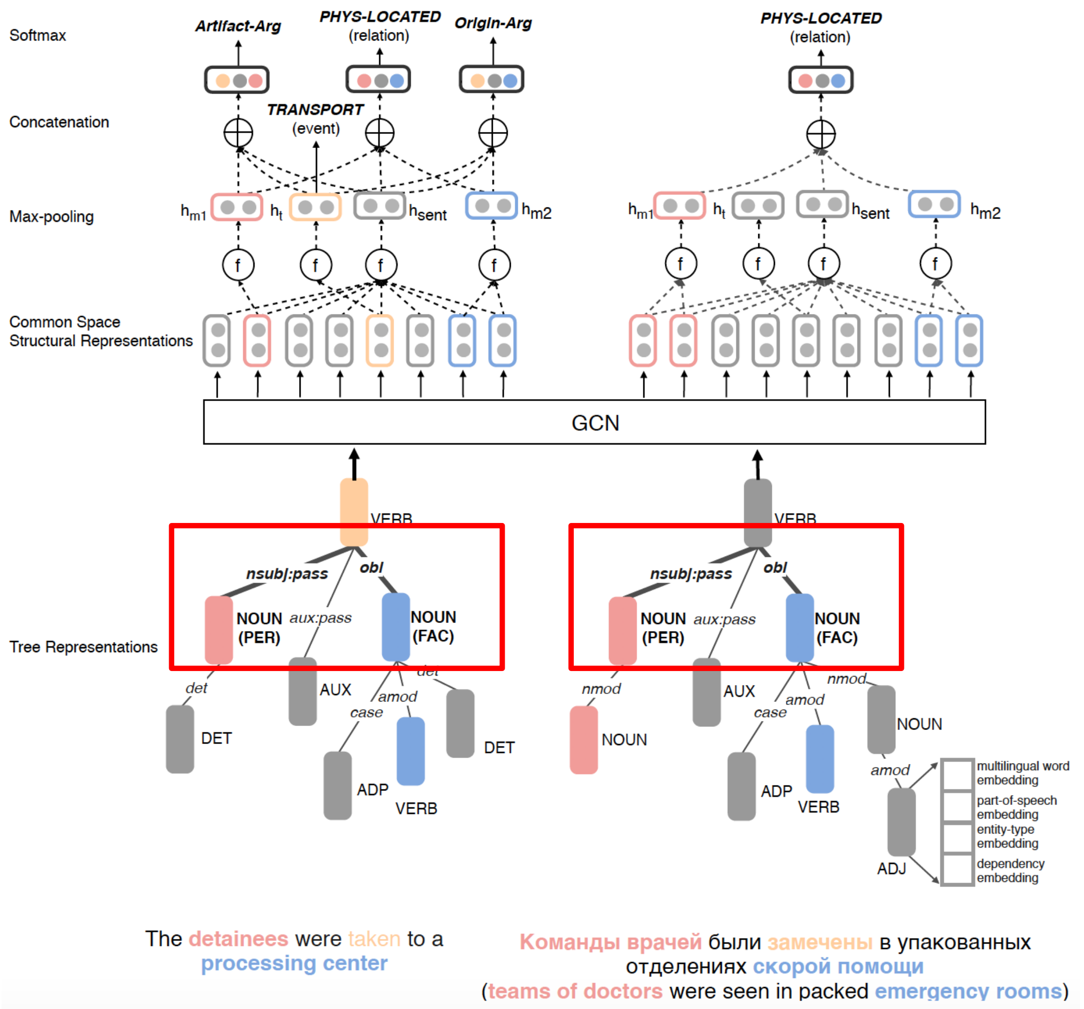


(Huang et al., 2018)

- Hypothesis: Cluster distribution tends to be consistent across languages
- Linguistic-driven cluster consistency across languages is more beneficial to information extraction

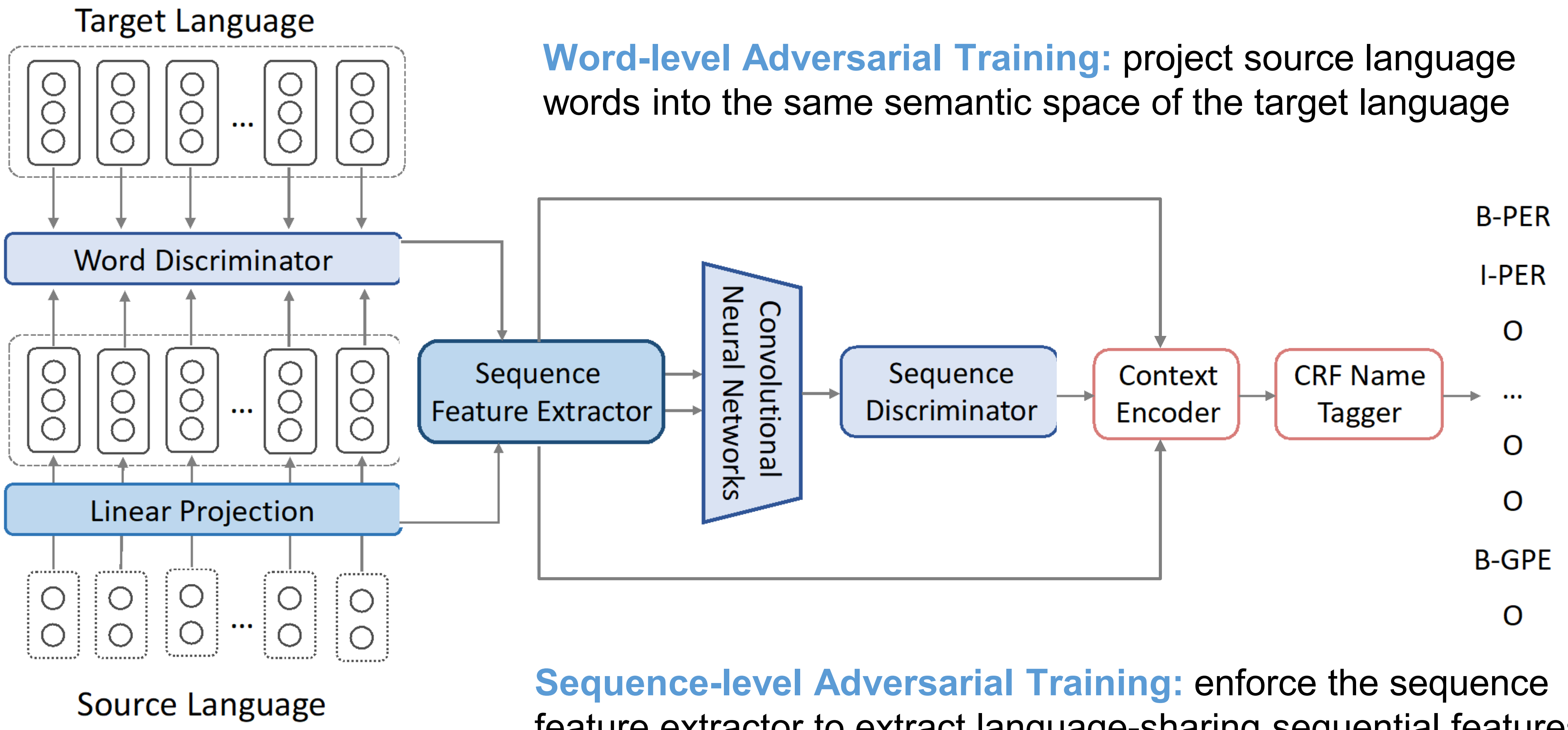
Cross-lingual Transfer: Language-agnostic Feature Representations

- Leveraging language-universal structural feature representations, e.g., dependency structures
  - Dependency substructures covering trigger and arguments are similar across languages (Subburathinam et al., 2019)
  - Pros
    - Agnostic to language word order
    - Capturing long-distance arguments
  - Cons: GCNs struggle to model words with long-range dependencies or are not connected in the dependency tree



Cross-lingual Transfer: Language-agnostic Feature Representations

- Learning language-agnostic feature representations with adversarial training



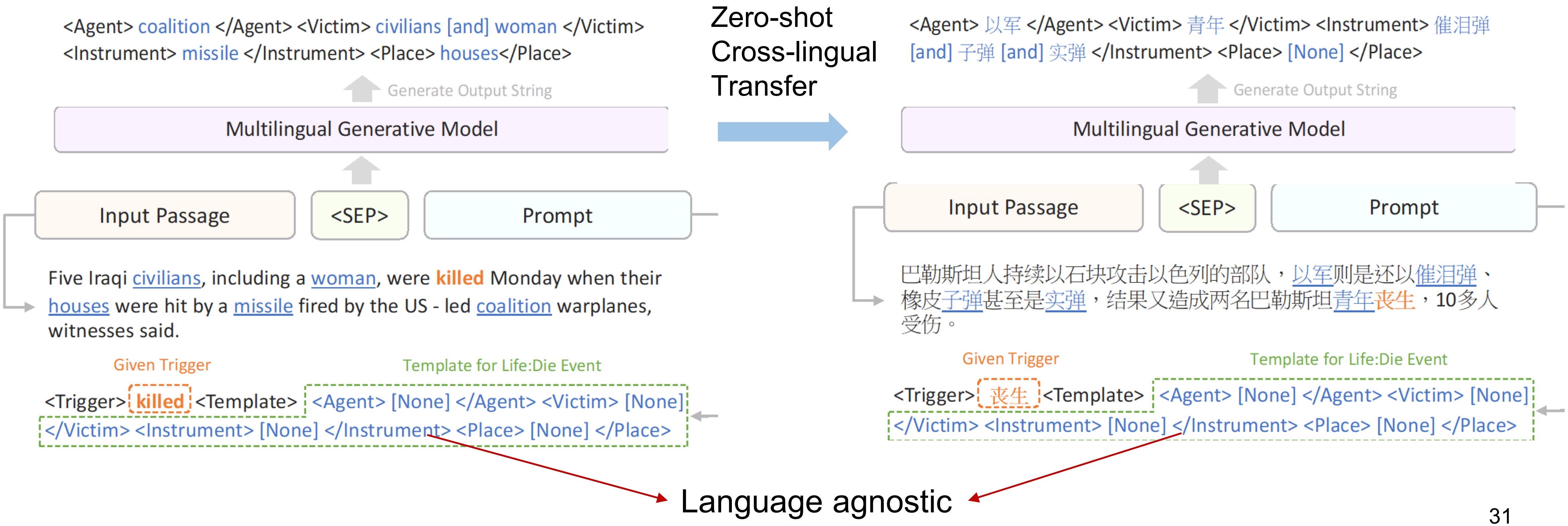
(Huang et al., 2019)

Word-level Adversarial Training: project source language words into the same semantic space of the target language

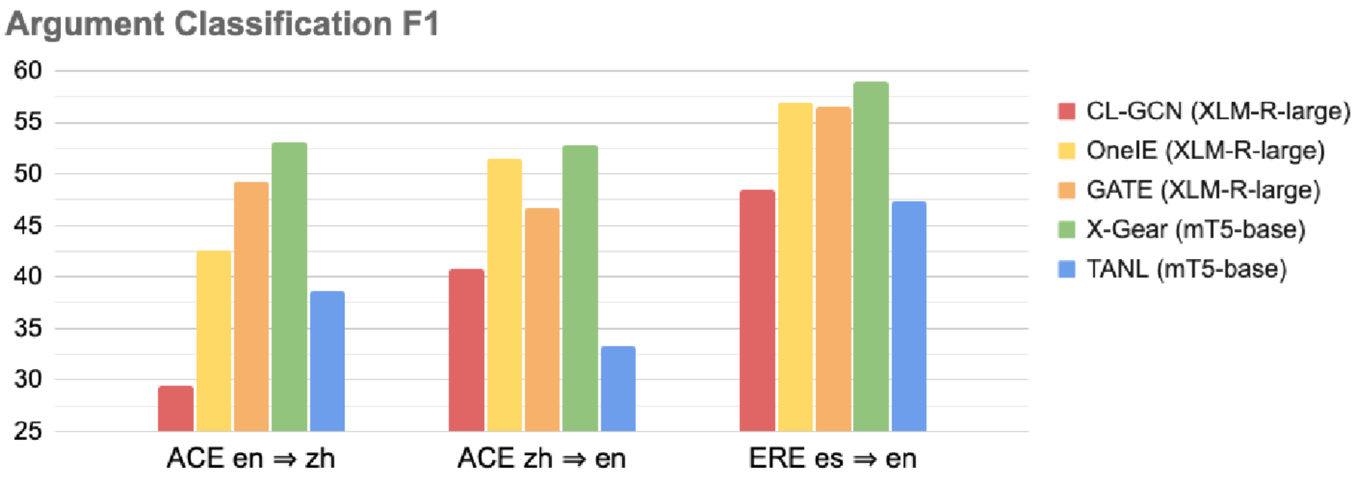
Sequence-level Adversarial Training: enforce the sequence feature extractor to extract language-sharing sequential features

Cross-lingual Transfer: Language-agnostic Feature Representations

- Leveraging language-agnostic feature representations from multilingual encoders / language models
  - X-Gear (Huang et al., 2022) : Leverage a multilingual pre-trained generative language model to generate events based on language-agnostic templates

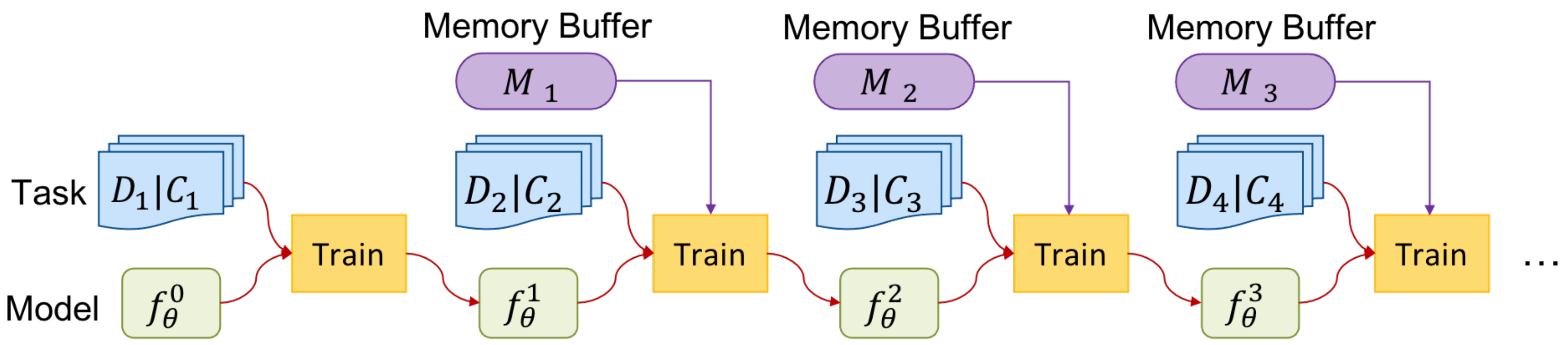


- X-Gear: Cross-lingual Zero-shot Transfer for Argument Extraction (Huang et al., 2022)
  - X-Gear consistently outperforms other approaches
    - CL-GCN: based on universal dependency structures,
    - OneIE/GATE: based on multilingual embeddings learned from pretrained multilingual language models



Continual Learning for IE

- How to mitigate the catastrophic forgetting?
  - Experience Replay: store K exemplars from old tasks into a memory and replay them periodically to prevent model forgetting previous knowledge when it's being trained on a new task
  - Knowledge Distillation: if a model extracts similar features or makes similar predictions for the same input as the old model, we can assume it preserves the knowledge
  - Task-specific Adapter: incrementally adding task-specific tunable parameters for each new task while fixing other parameters

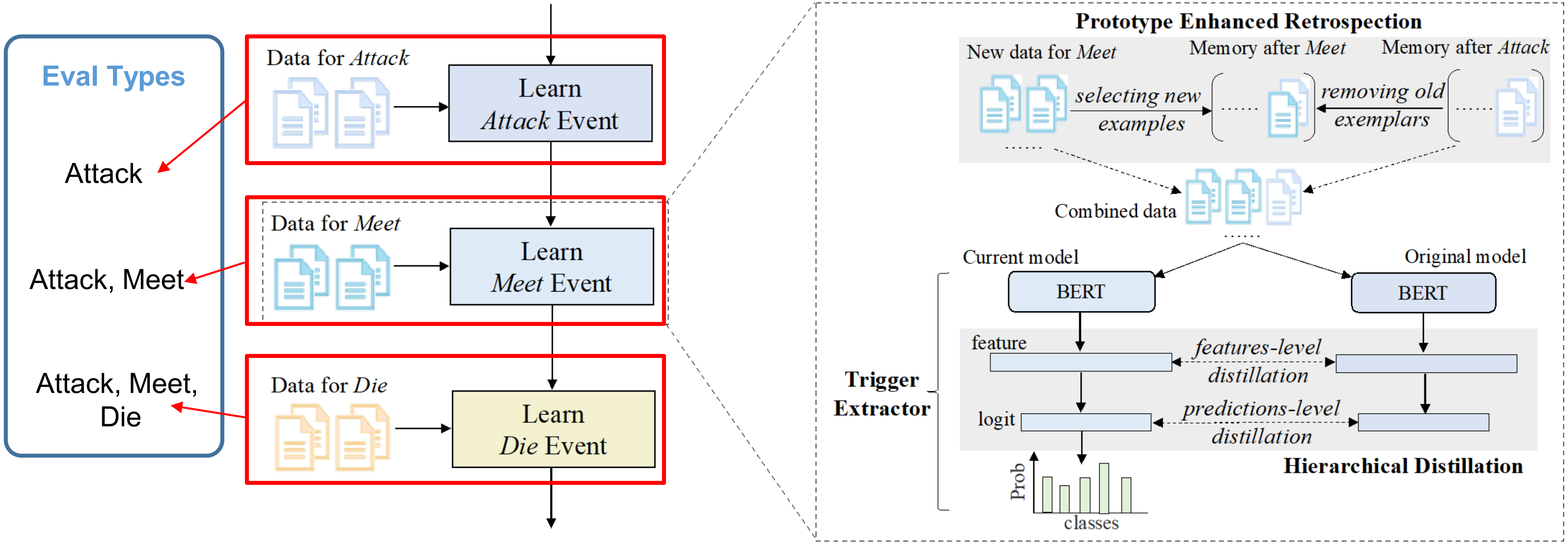


Continual Learning for IE

How to select examples?

- Selecting the examples that are closer to the prototype of each old type
- Removing the examples that are far from the prototype of each old type

$$\boldsymbol{\mu}_c = \frac{1}{N_c}\sum_{i=1}^{N_c} \boldsymbol{z}_i$$



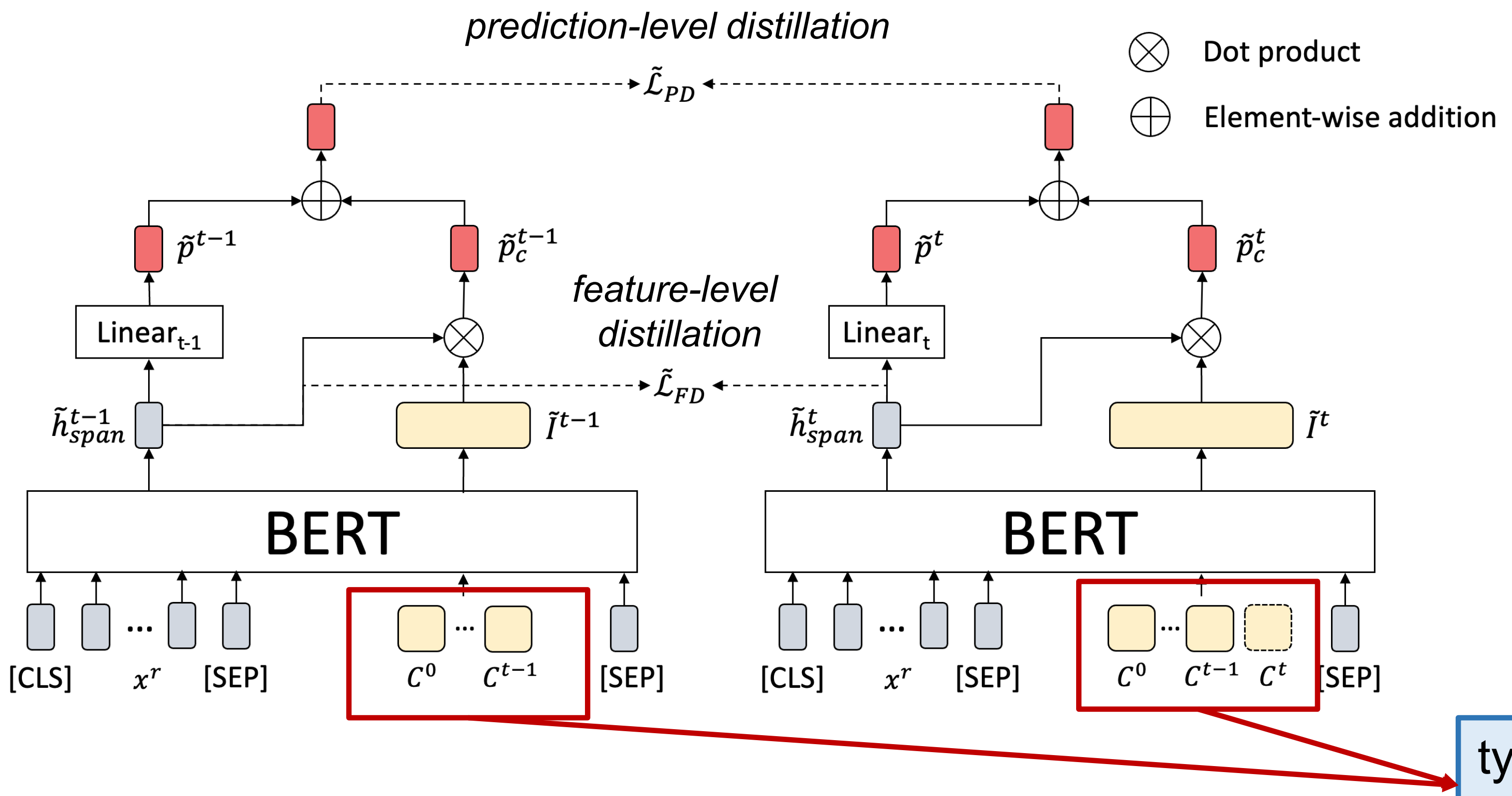
(Cao et al., 2020)

Knowledge distillation

- Feature-level Distillation: encourage the new model to extract similar features for the same input as the original model
- Prediction-level Distillation: encourage the new model to make similar predictions for the same input as the original model

Continual Learning for IE

- Episodic Memory Prompting (EMP): incrementally integrating the representations of new labels for each new task



$$\tilde{p}^t = linear(\tilde{\mathbf{h}}^t_{span}).$$

$$\tilde{p}^t_c = \mathrm{MLP}(\tilde{\mathbf{I}}^t) \cdot \tilde{\mathbf{h}}^t_{span}$$

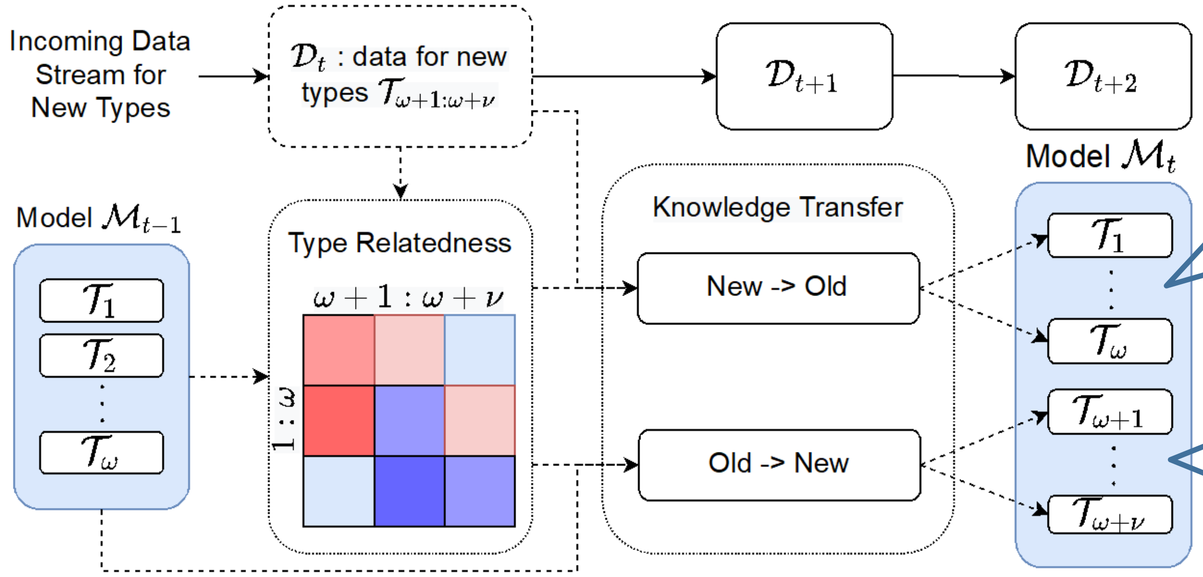
$$\tilde{\mathcal{L}}_C = -\sum_{(\tilde{x}^t, y^t) \in \mathcal{D}_t} \log\left(\tilde{p}^t + \tilde{p}^t_c\right).$$

type representations, which are initialized by the event type names

(Liu et al., 2022)

Continuous Learning for IE

- Knowledge Transfer (Yu et al., 2021)
  - Event detection: inner product between a token embedding and type embeddings
  - New → Old: Use new data to update the knowledge of old model by self-training
  - Old → New: Transfer old knowledge to new types by initializing the type embeddings for new types based on learned types



Self-training: encouraging the probability of each instance from the new task over old types to be consistent between the old and new models

Pseudo label dist. from old model

Label dist. from new model

$$\mathcal{L}_S = -\sum_{\substack{(x,y)\in\mathcal{D}_t \\ c\in\mathcal{O}_{t-1}\cup\{c^\phi\}}} q^{t-1}(c|x)\log p(c|x)$$

Inst. from new task

Old label embeddings

1. learn new type embeddings based on old types

$$\boldsymbol{\omega} = \frac{1}{h}\sum_{i=1}^{h}\sum_{c\in\mathcal{O}_{t-1}} p(c|x_i)\boldsymbol{c}.$$

x_i: an inst. from new type
$\boldsymbol{c}$: the embedding of an old type

2. learn new type embeddings from new instances

$$\boldsymbol{\nu} = \frac{1}{h}\sum_{i=1}^{h} p(c^\phi|x_i)\frac{d\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|_2},$$

$\boldsymbol{x}_i$: embedding of x_i from BERT
$p(c^\emptyset|x_i)$: how much x_i is different from old types

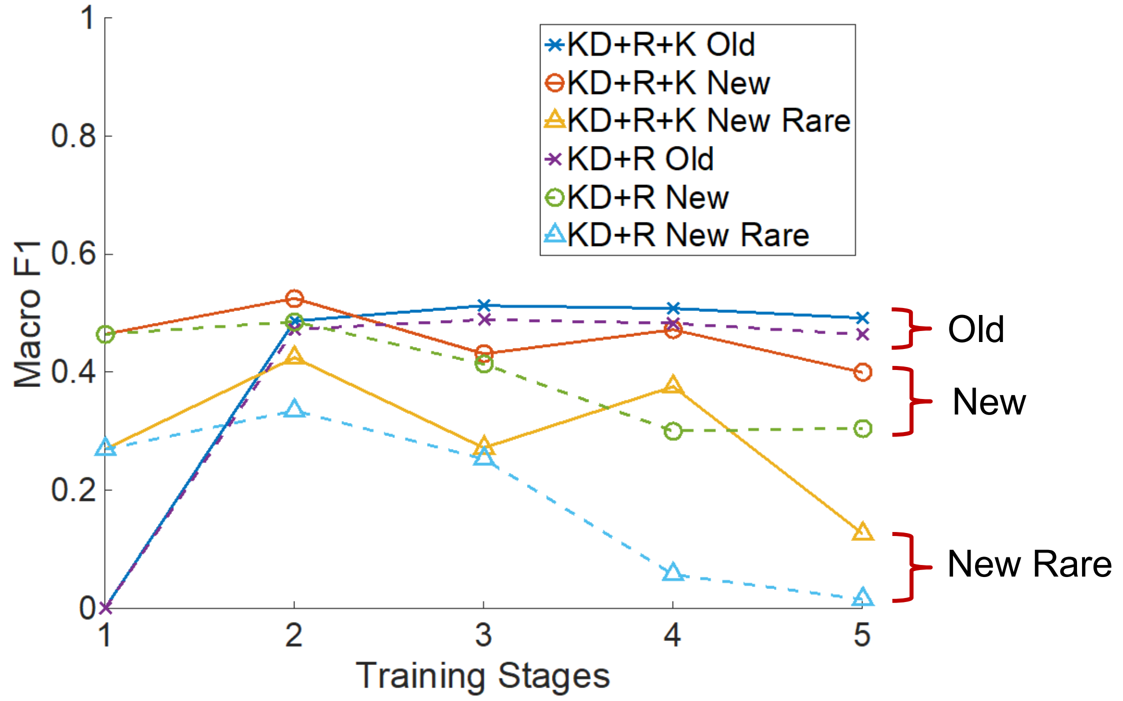
Downside: all these approaches require to store exemplars from old tasks, which is not realistic

Continuous Learning for IE

- Knowledge Transfer Improves Learning on Old and New Types (Yu et al., 2021)

  - Comparing with baseline (KD+R), Knowledge transfer improve performance on both new and old types

  - More improvements on rare new types, showing that sharing knowledge can help learning long-tail events



Old: old types learned in previous stages
New: new types learned in this stage
New Rare: new types with fewer than 120 training mentions

References

Lin et al., 2020, OneIE: A Joint Neural Model for Information Extraction with Global Features. ACL'2020

Conneau et al., 2018. Word Translation without Parallel Data. ICLR'2018

Cao et al., 2020, Incremental Event Detection via Knowledge Consolidation Networks. EMNLP'2020

Wiewel et al., 2019, Localizing Catastrophic Forgetting in Neural Networks.

Bronstein et al., 2015, Seed-Based Event Trigger Labeling: How far can event descriptions get us? IJCNLP'2015

Han et al., 2021, Exploring Task Difficulty for Few-Shot Relation Extraction. EMNLP'2021

Aly et al., 2021, Leveraging Type Descriptions for Zero-shot Named Entity Recognition and Classification. ACL'2021

Huang et al., 2018, Zero-Shot Transfer Learni

Wang et al., 2022, The Art of Prompting- Ever ... pts. in arXiv. 2022

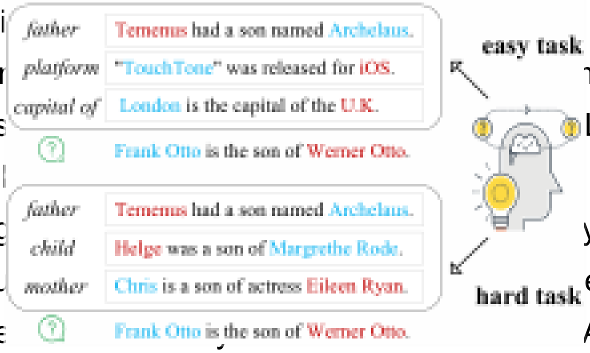
Du and Cardie, 2020, Event Extraction by Ans ... LP'2020

Liu et al., 2020, Event Extraction as Machine

Wang et al., 2022, Query and Extract Refining ... y Decoding. ACL'2022

Lu et al., 2021, Text2Event- Controllable Sequ ... end Event Extraction. ACL'2021

Li et al., 2021, Document-Level Event Argume ... ACL'2021



Pan et al., 2017, Cross-lingual Name Tagging and Linking for 282 Languages. ACL'2017

Zhang et al., 2017, Embracing non-traditional linguistic resources for low-resource language name tagging. IJCNLP'2017

Yarmohammadi et al., 2021, Everything Is All It Takes- A Multipronged Strategy for Zero-Shot Cross-Lingual Information Extraction. EMNLP'2021

Huang et al., 2018, Multi-lingual Common Semantic Space Construction via Cluster-consistent Word Embedding. EMNLP'2018

Subburathinam et al., 2019, Cross-lingual Structure Transfer for Relation and Event Extraction. EMNLP'2019

Huang et al., 2019, Cross-lingual Multi-Level Adversarial Transfer to Enhance Low-Resource Name Tagging. NAACL'2019

Huang et al., 2022, Multilingual Generative Language Models for Zero-Shot Cross-Lingual Event Argument Extraction.

Liu et al., 2022 Incremental Prompting: Episodic Memory Prompt for Lifelong Event Detection. In arXiv. 2022

Yu et al., 2021, Lifelong Event Detection with Knowledge Transfer. EMNLP'2021

Thank You