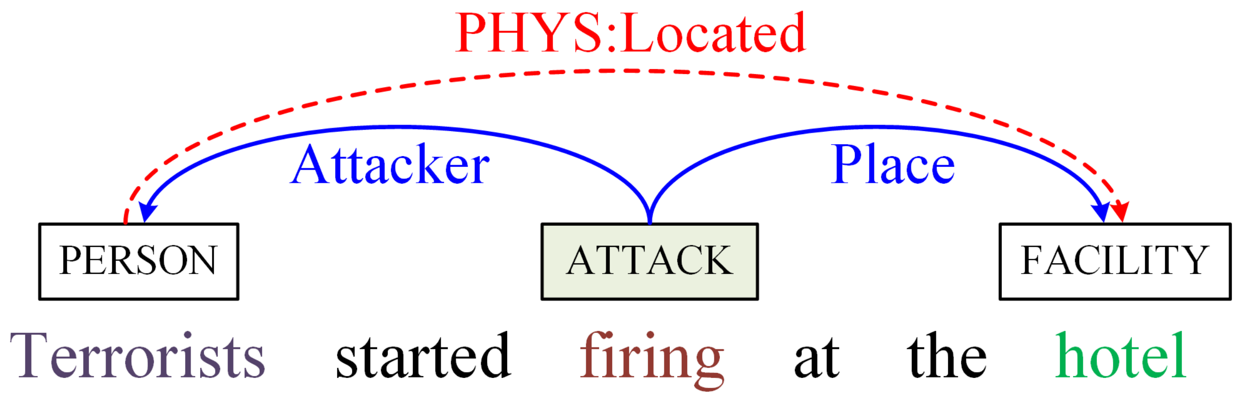
GATE: Graph Attention Transformer Encoder for Cross-Lingual Relation and Event Extraction

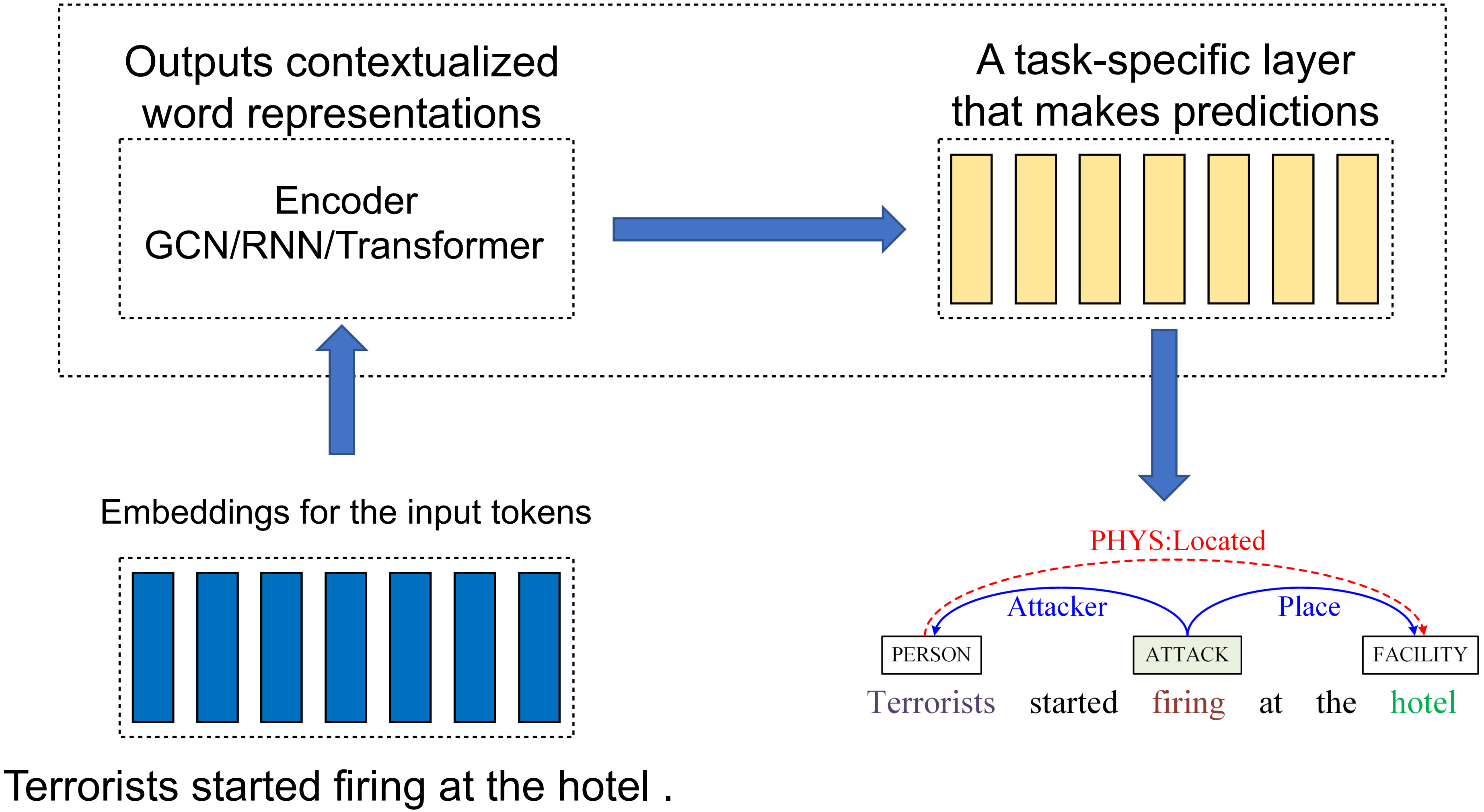
Wasi Ahmad, Nanyun Peng, and Kai-Wei Chang.
University of California, Los Angeles
AAAI 2021

Relation and Event Extraction

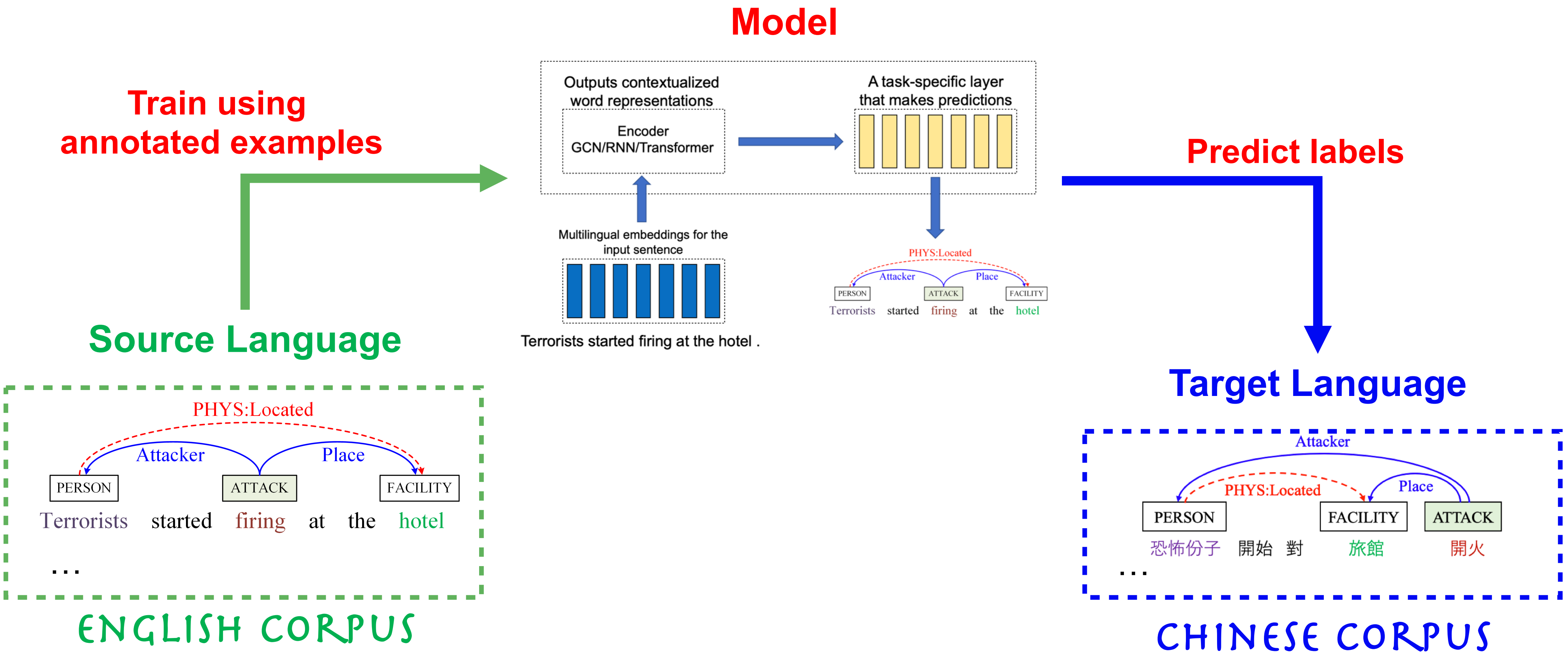


A relation (red dashed) between two entities and an event of type Attack (triggered by “firing”) including two arguments and their role labels (blue) are highlighted.

Relation and Event Extraction



Cross-lingual Relation and Event Extraction



Cross-lingual Transfer

<u>Challenge</u>

Different languages have different properties

(e.g., word order)

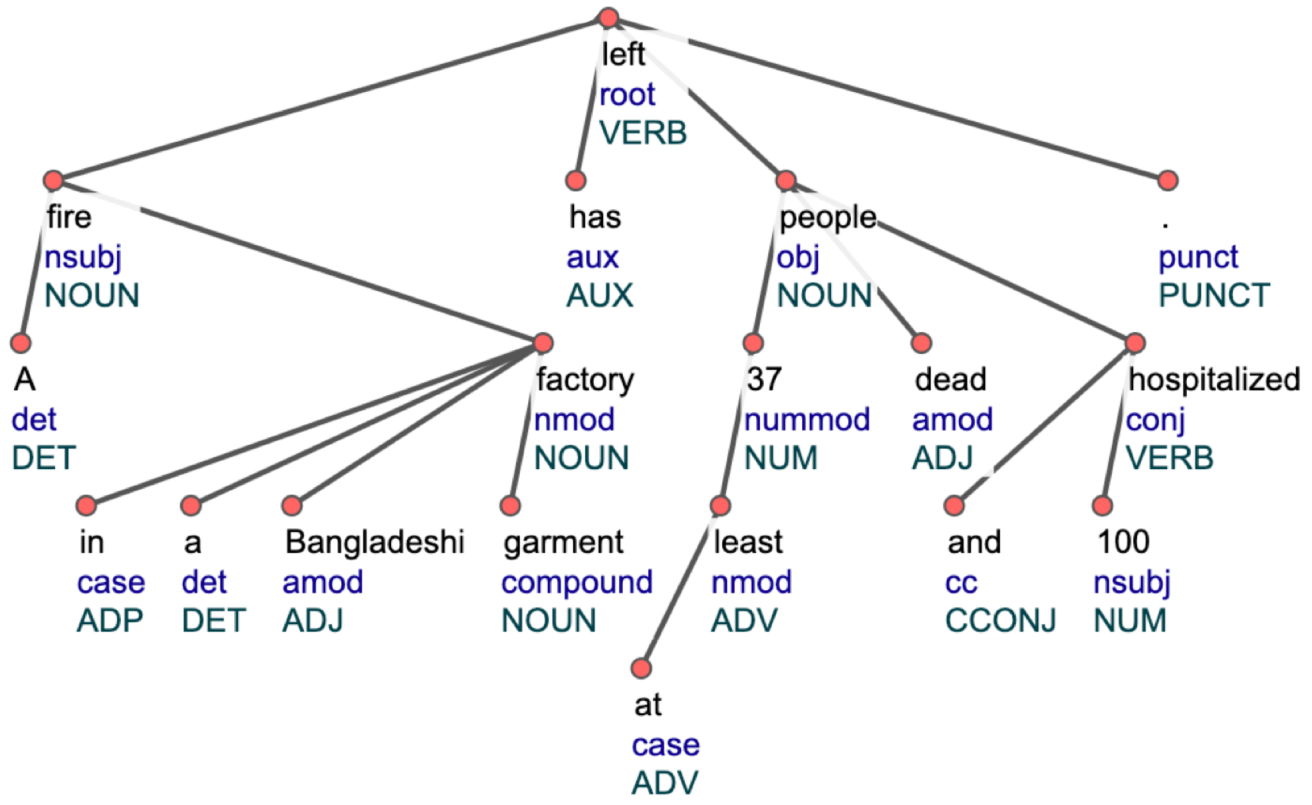
<u>Countermeasure</u>

Learning language-agnostic representation

(to improve cross-lingual transfer)

Language-agnostic Representation

A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized .



Use of Dependency Structure

A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized .

Distance (fire, hospitalized) = 15

- Capturing long-range dependency is crucial

Use of Dependency Structure

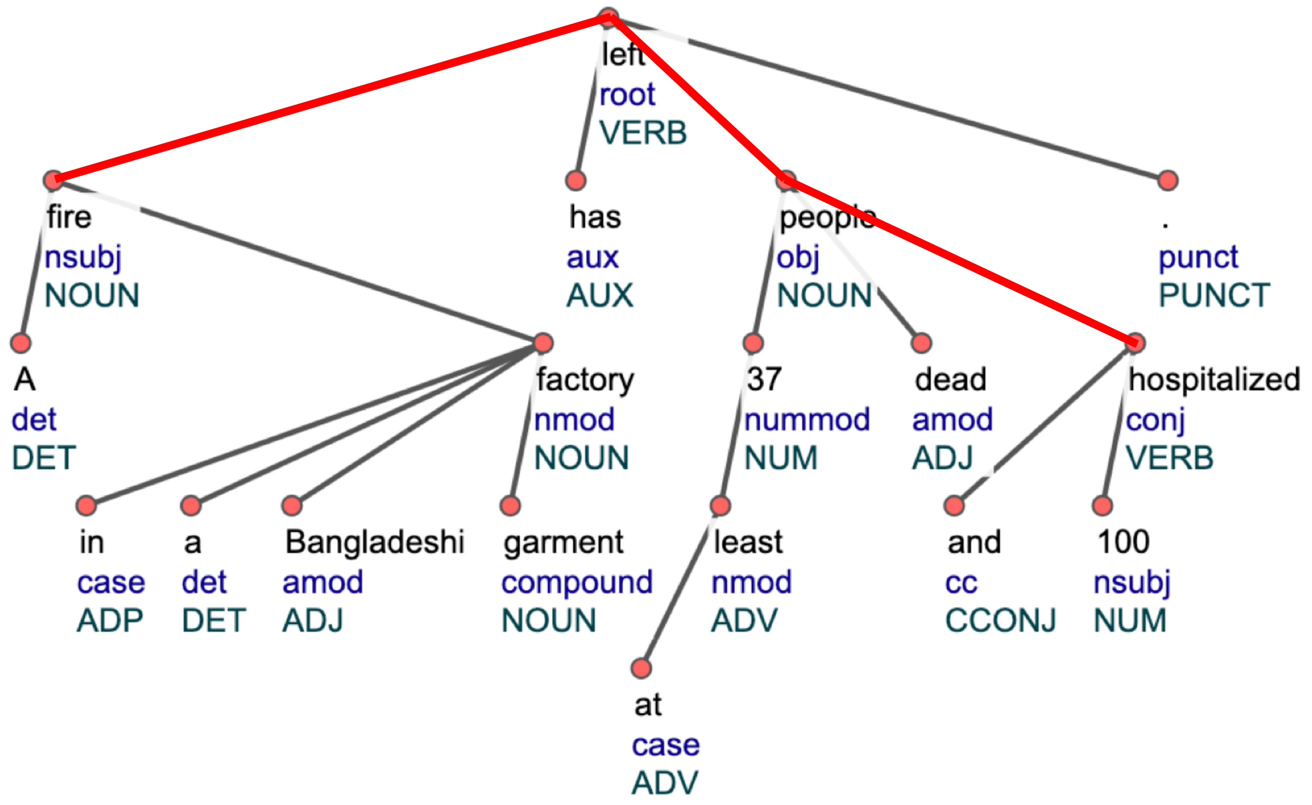
A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized .

Distance (fire, hospitalized) = 15

- Capturing long-range dependency is crucial
- Syntactic distance between two words in a sentence is typically smaller than the sequential distance
  - Avg. sequential distance: [English] 9.8; [Arabic] 58.1
  - Avg. syntactic distance: [English] 3.1; [Arabic] 12.3

Use of Dependency Structure

A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized .



Distance
Sequential = 15
Syntactic = 4

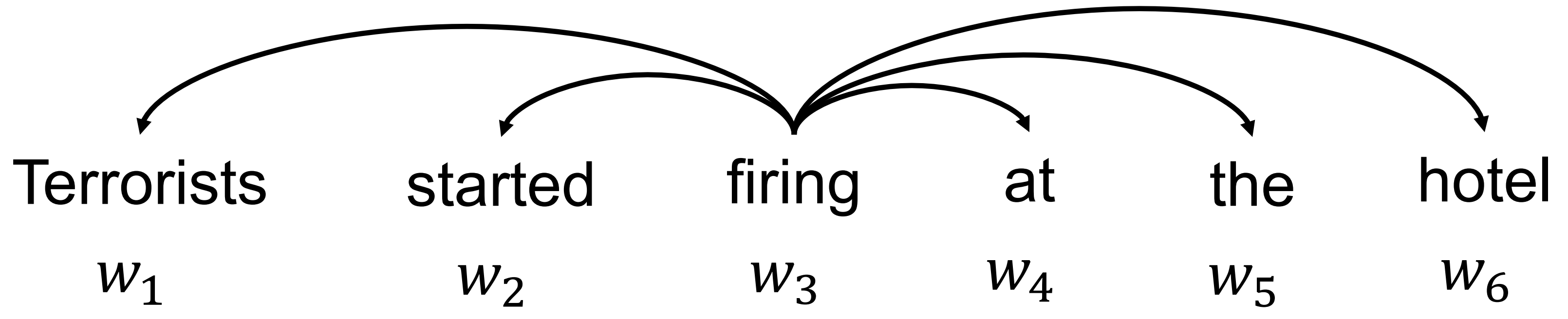
Dependency-guided Self-Attention

Self-attention

Attention mask

$$e_{ij} = \frac{1}{\sqrt{d_k}}\left[(x_i W_l^Q)(x_j W_l^K)^T + M\right]$$

$$M_{ij} = \begin{cases} 0, & \text{allow to attend} \\ -\infty, & \text{prevent from attending} \end{cases}$$



Dependency-guided Self-Attention

Restricting attention to adjacent tokens

$$e_{ij} = \frac{1}{\sqrt{d_k}} \left[(x_i W_l^Q)(x_j W_l^K)^T + M \right]$$

$$M_{ij} = \begin{cases} 0, & D_{ij} = 1 \\ -\infty, & \text{otherwise} \end{cases}$$

Syntactic Distance Matrix, D

	w_1	w_2	w_3	w_4	w_5	w_6
w_1	1	1	2	4	4	3
w_2	1	1	1	3	3	2
w_3	2	1	1	2	2	1
w_4	4	3	2	1	2	1
w_5	4	3	2	2	1	1
w_6	3	2	1	1	1	1

Terrorists	started	firing	at	the	hotel
w_1	w_2	w_3	w_4	w_5	w_6

Dependency-guided Self-Attention

Attend tokens that are at most δ-hop away

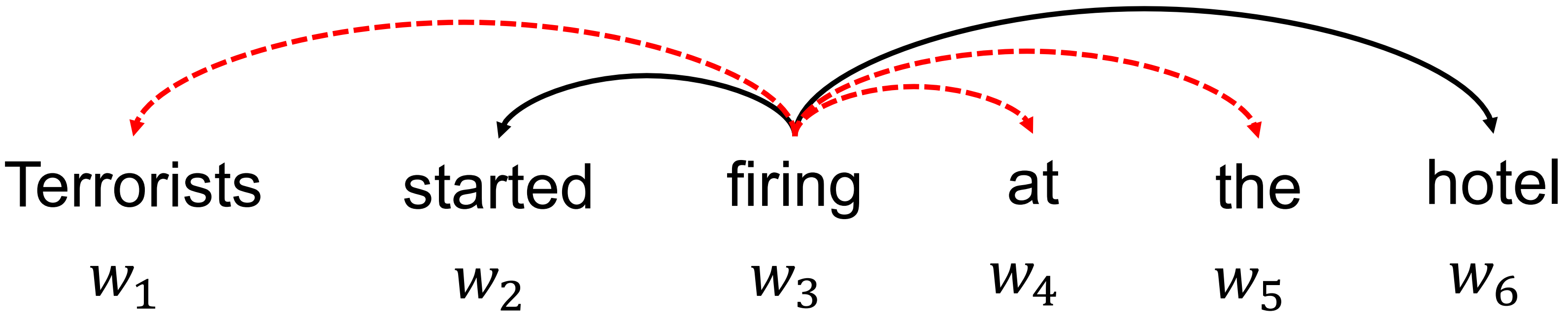
$$e_{ij} = \frac{1}{\sqrt{d_k}}\left[(x_i W_l^Q)(x_j W_l^K)^T + M\right]$$

$$M_{ij} = \begin{cases} 0, & D_{ij} \le \delta \\ -\infty, & \text{otherwise} \end{cases}$$

$\delta = 2$

Syntactic Distance Matrix, D

	w_1	w_2	w_3	w_4	w_5	w_6
w_1	1	1	2	4	4	3
w_2	1	1	1	3	3	2
w_3	2	1	1	2	2	1
w_4	4	3	2	1	2	1
w_5	4	3	2	2	1	1
w_6	3	2	1	1	1	1



Dependency-guided Self-Attention

Our proposal

$$A_l = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right) V_l,$$

$$M_{ij} = \begin{cases} 0, & D_{ij} \leq \delta \\ -\infty, & \text{otherwise} \end{cases}$$

Allow tokens to attend tokens that are within distance δ

Dependency-guided Self-Attention

Our proposal

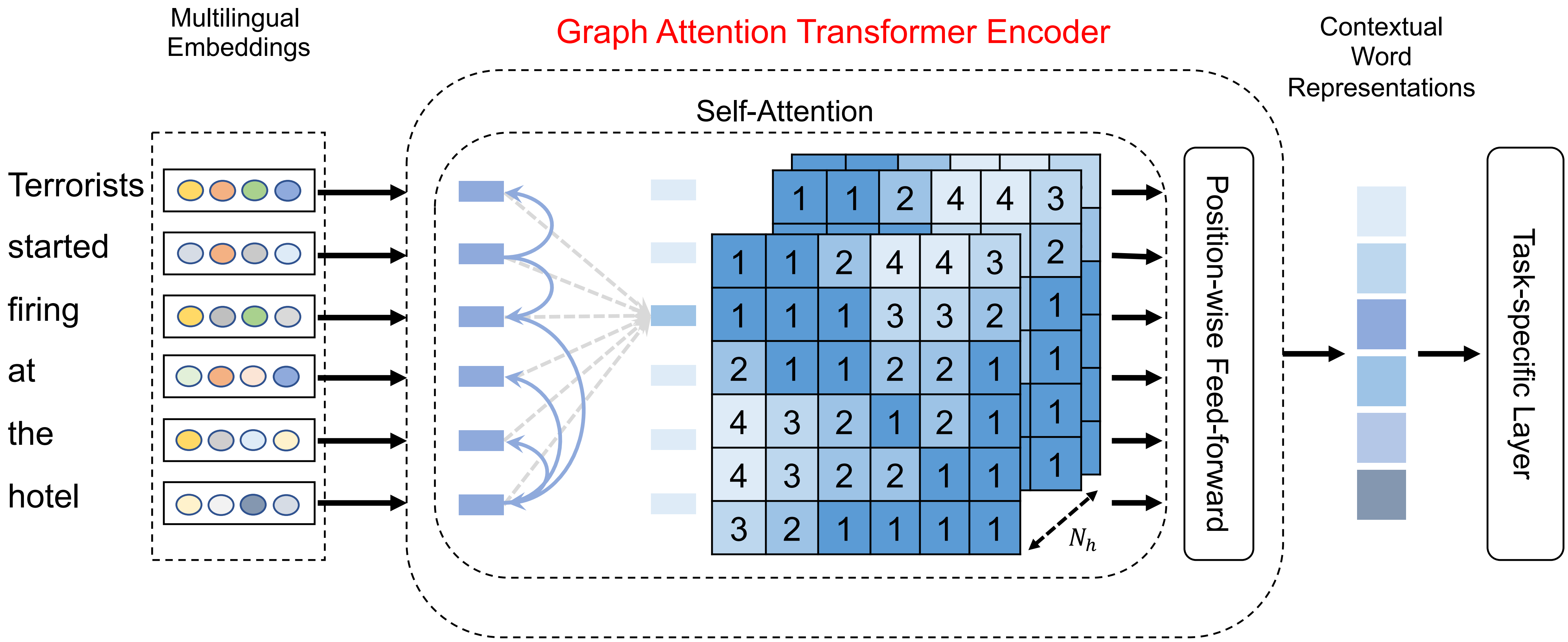
$$A_l = F\left(\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right)\right)V_l.$$

$$F(P)_{ij} = \frac{P_{ij}}{Z_i D_{ij}}, M_{ij} = \begin{cases} 0, & D_{ij} \leq \delta \\ -\infty, & \text{otherwise} \end{cases}$$

Allow tokens to attend tokens that are within distance δ

Pay more attention to tokens that are closer and less attention to tokens that are faraway in the dependency tree

Our Proposal



Evaluation Setup

Dataset: ACE 2005

- Languages – English (En), Chinese (Zh), Arabic (Ar)
- Single-source transfer (En->Ar, Ar-> Zh, etc.)
- Multi-source transfer (En+Ar->Zh, Zh+Ar-> En, etc.)

Data Statistics

	English	Chinese	Arabic
Relations Mentions	8,738	9,317	4,731
Event Mentions	5,349	3,333	2,270
Event Arguments	9,793	8,032	4,975

Distance Statistics

	Sequential			Syntactic		
	En	Zh	Ar	En	Zh	Ar
Relation Mention	4.8	3.9	25.8	2.2	2.6	5.1
Event Mention & Arg.	9.8	21.7	58.1	3.1	4.6	12.3

Evaluation Setup

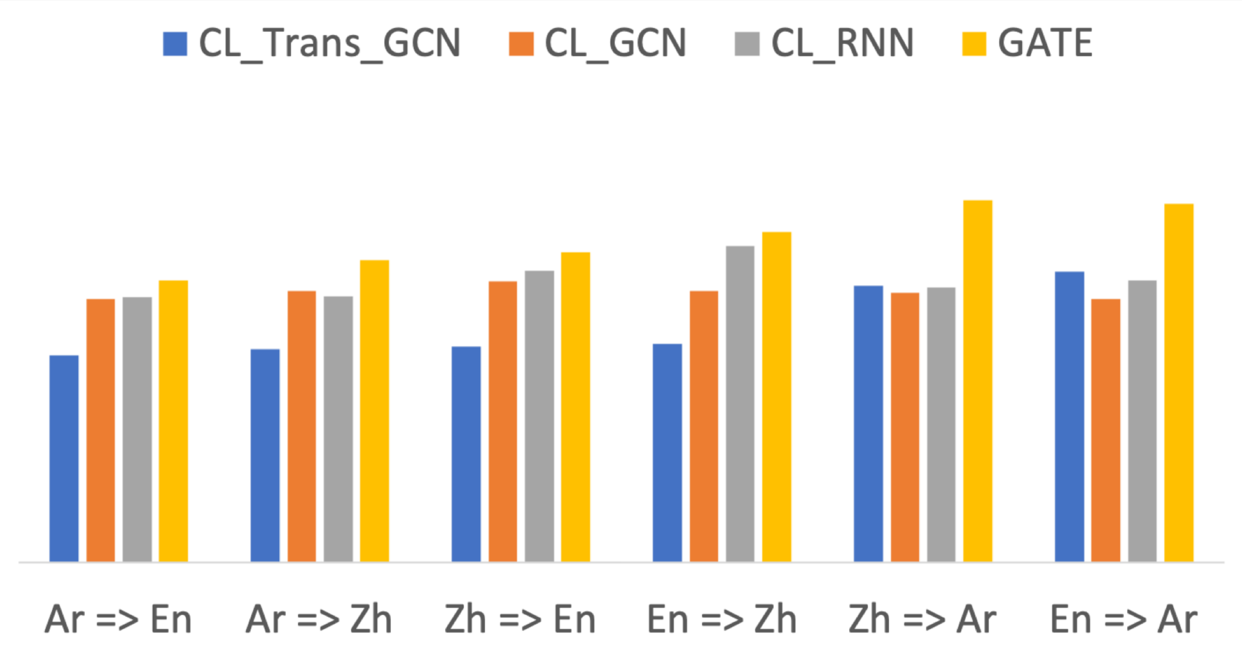
Baseline Methods

- CL_Trans_GCN [Liu et al. 2019]
- CL_GCN [Subburathinam et al. 2019]
- CL_RNN [Ni and Florian 2019]
- Transformer [Vaswani et al. 2017]
- Transformer_RPR [Shaw et al. 2018]

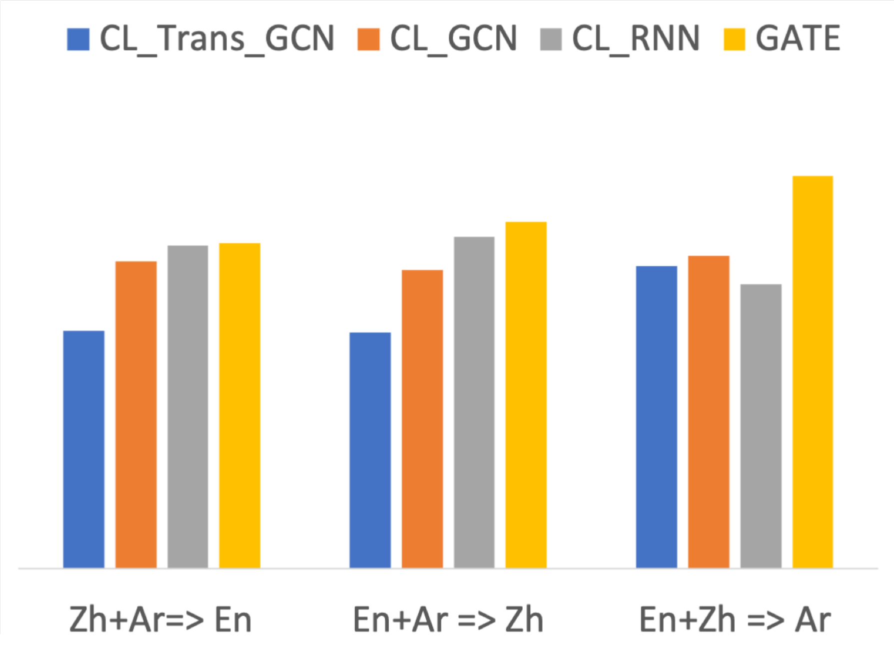
Source Code is Publicly Available
https://github.com/wasiahmad/GATE

EARL Results

Single-source Transfer

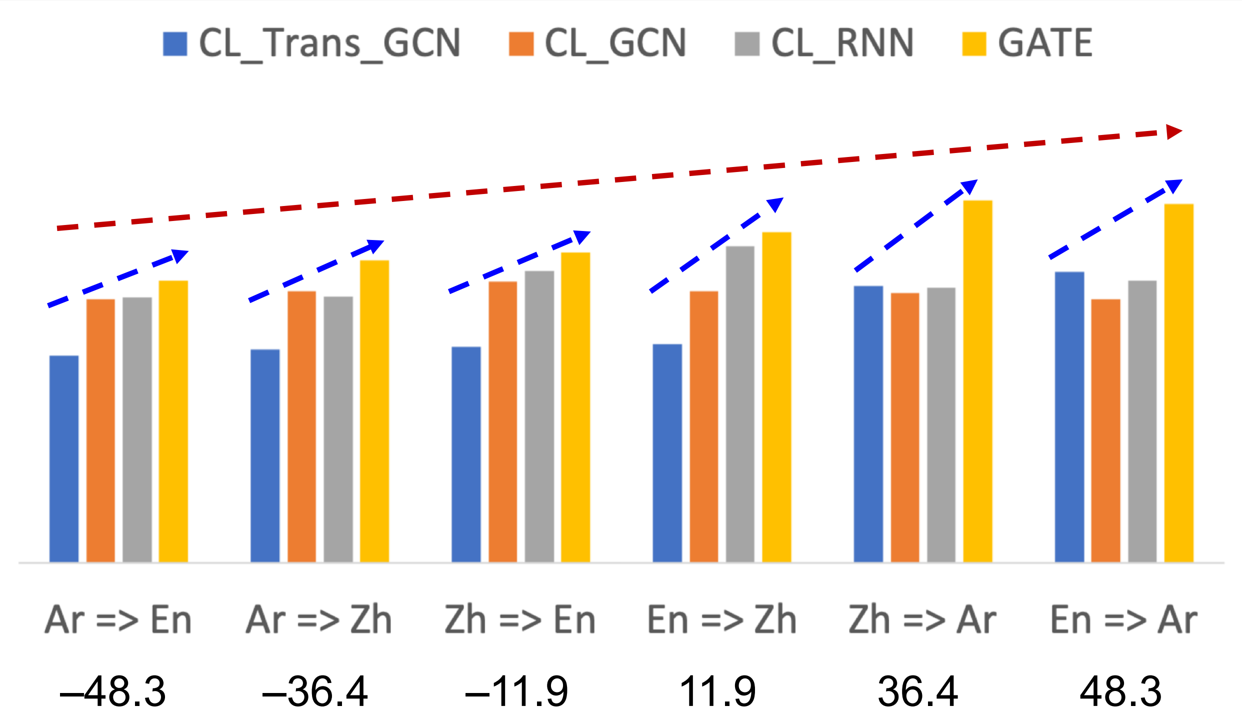


Multi-source Transfer

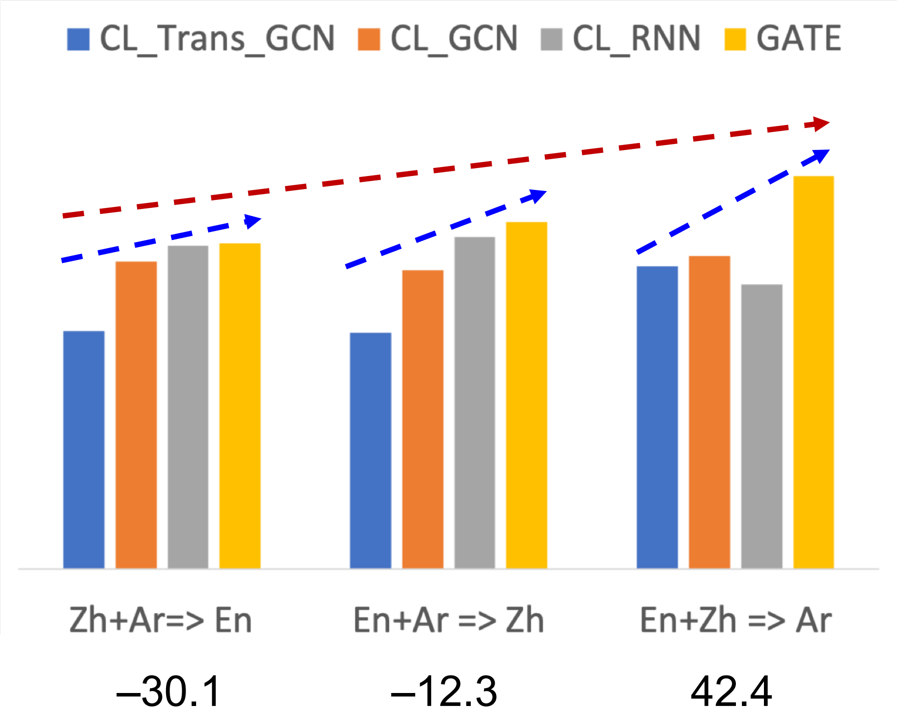


EARL Results

Single-source Transfer



Multi-source Transfer



Avg. sequential distance between event triggers and their arguments: (Arabic) 58.1 – (English) 9.8 = 48.3

Sensitivity Towards Language

A model would transfer well on target languages if the model is less sensitive towards the source language characteristics

(e.g., word order, grammar structure)

Sensitivity Towards Language

Evaluate the model on the target language sentences and their translation in source language.

<u>Hypothesis</u>

Lower cross-lingual gap indicates the model is less sensitive.

Sensitivity Towards Language

Collecting Translations

- Used Google Translation
- Translated English (test set) sentences into Chinese and Arabic

```
English sentence: her <i> stockbroker </i> was also <b> charged </b> .
Chinese translation: 她的<i>股票经纪人</i>也<b>被起诉</b> 。
Arabic translation: كما <b>اتهم</b> <i>سمسار الأوراق المالية</i> لها.
```

Sensitivity Towards Language

- Source Language: Chinese; Target Language: English

Model	EARL		RE	
	English	Chinese*	English	Chinese*
CL_GCN	51.5	56.3	46.9	50.7
CL_RNN	55.6	59.3	56.8	**62.0**
GATE	**63.8**	**64.2**	**58.8**	57.0

Cross-lingual GAP

Model	EARL	RE
CL_GCN	+4.8	+3.8
CL_RNN	+3.3	+5.2
GATE	+0.4	-1.8

Sensitivity Towards Language

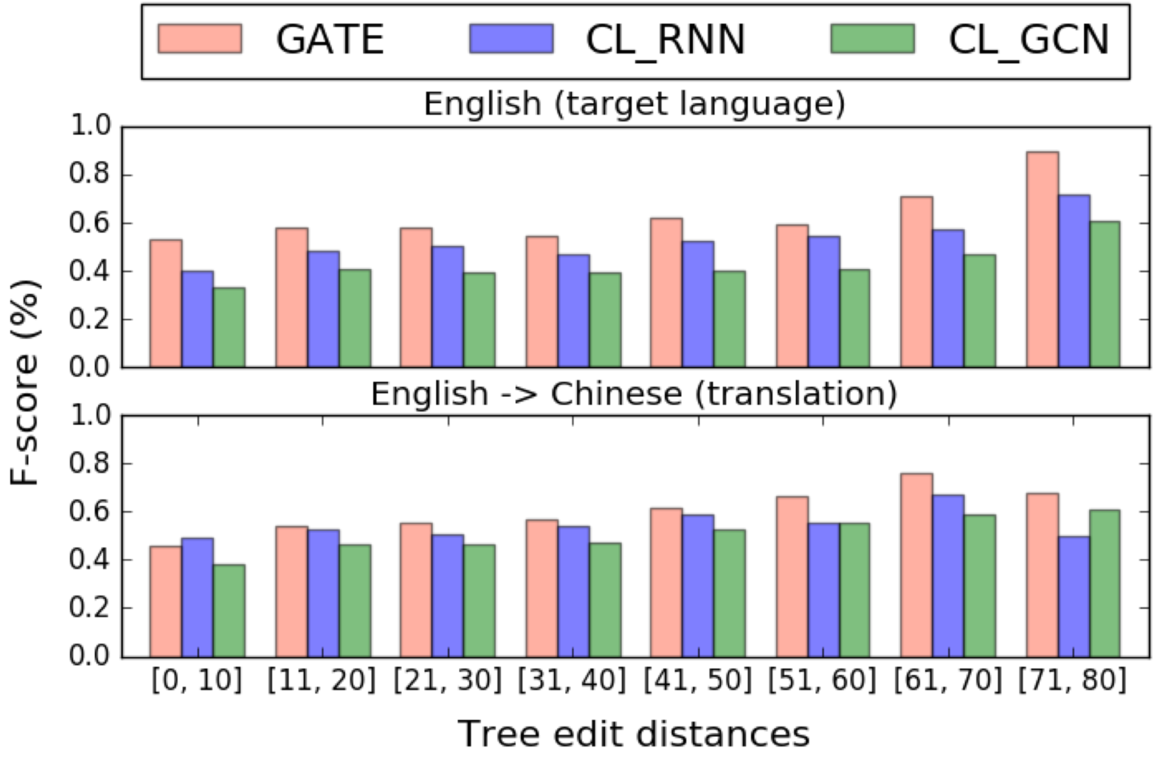
We quantify the difference between languages using dependency structure

- Tree edit distance using the APTED algorithm

Sensitivity Towards Language

We quantify the difference between languages using dependency structure

- Tree edit distance using the APTED algorithm

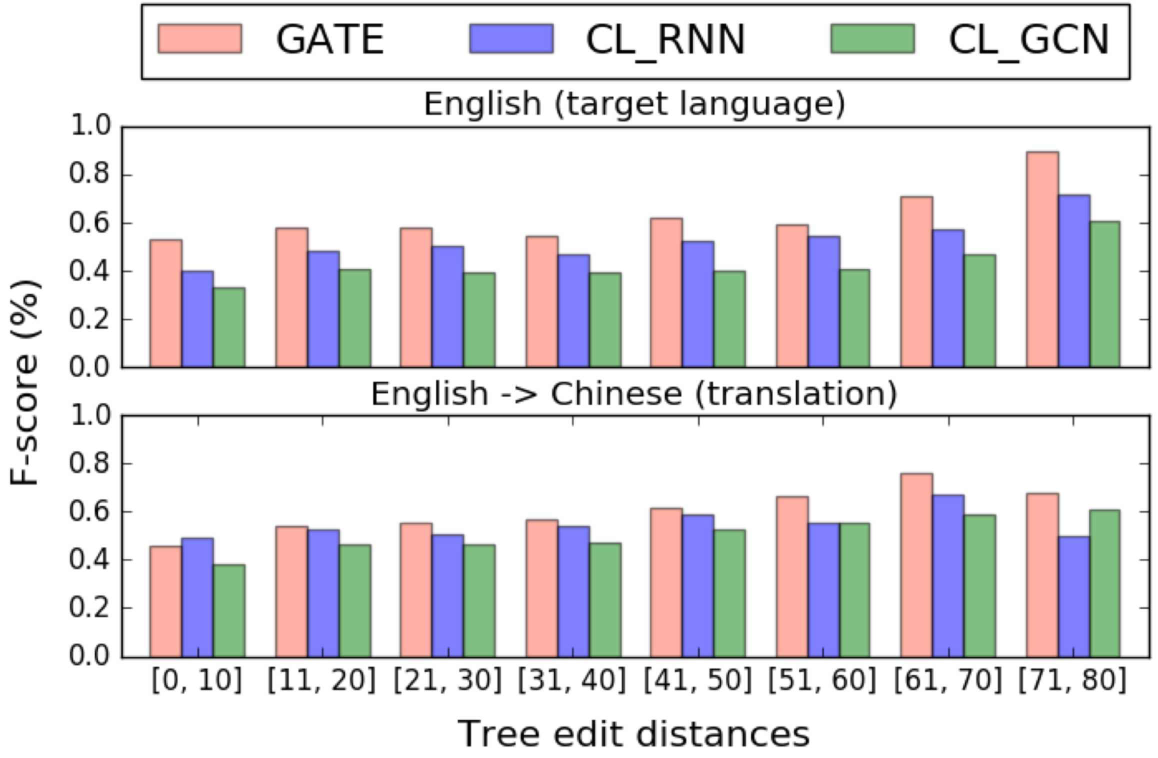


Larger edit distance indicates larger difference in the sentence structure

Sensitivity Towards Language

We quantify the difference between languages using dependency structure

- Tree edit distance using the APTED algorithm



When tree edit distance gets large, GATE performs significantly better

Larger edit distance indicates larger difference in the sentence structure

Conclusion

Source Code is Publicly Available
https://github.com/wasiahmad/GATE

- Proposed a dependency-guided self-attention mechanism
  - To embed structure in contextual representations
- Comprehensive empirical study
  - Both single- and multi-source transfer
- Future work
  - Other ways of encoding dependency structure

References

[Liu et al. 2019] Liu, J.; Chen, Y.; Liu, K.; and Zhao, J. 2019. Neural Cross-Lingual Event Detection with Minimal Parallel Resources. In Proceedings of EMNLP-IJCNLP, 738–748.

[Subburathinam et al. 2019] Subburathinam, A.; Lu, D.; Ji, H.; May, J.; Chang, S.-F.; Sil, A.; and Voss, C. 2019. Cross-lingual Structure Transfer for Relation and Event Extraction. In Proceedings of EMNLPIJCNLP, 313–325.

[Ni and Florian 2019] Ni, J.; and Florian, R. 2019. Neural Cross-Lingual Relation Extraction Based on Bilingual Word Embedding Mapping. In Proceedings of EMNLP-IJCNLP, 399–409.

[Vaswani et al. 2017] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In NeurIPS, 5998–6008.

[Shaw et al. 2018] Shaw, P.; Uszkoreit, J.; and Vaswani, A. 2018. Self-Attention with Relative Position Representations. In Proceedings of NAACL, 464–468.

[Wang et al. 2019] Wang, X.; Tu, Z.;Wang, L.; and Shi, S. 2019. Self-Attention with Structural Position Representations. In Proceedings of EMNLP-IJCNLP, 1403–1409.

Thank You