Cross-lingual Representation Learning for Natural Language Processing

Ph.D. Dissertation Defense

by

Wasi Uddin Ahmad

Department of Computer Science

University of California, Los Angeles

09.07.2021

Doctoral Committee

Dr. Kai-Wei Chang (Chair)

Dr. Junghoo Cho

Dr. Yizhou Sun

Dr. Guy Van Den Broeck

Natural Language Processing (NLP)

What does an NLP system need to know?

- Languages consist of many levels of structure
  - Morphology, syntax, semantics, pragmatics
- Humans fluently integrate all of these in understanding languages
- Ideally, so would an NLP system!

Courtesy: http://www.cs.cmu.edu/~ytsvetko/jsalt-part1.pdf

Multilingual NLP

- NLP systems capable of understanding many languages
- Why do we need multilingual NLP systems?
    1. Commercial value
    2. Social well-being
    3. Information dissemination

[1] https://www.ethnologue.com/browse/families

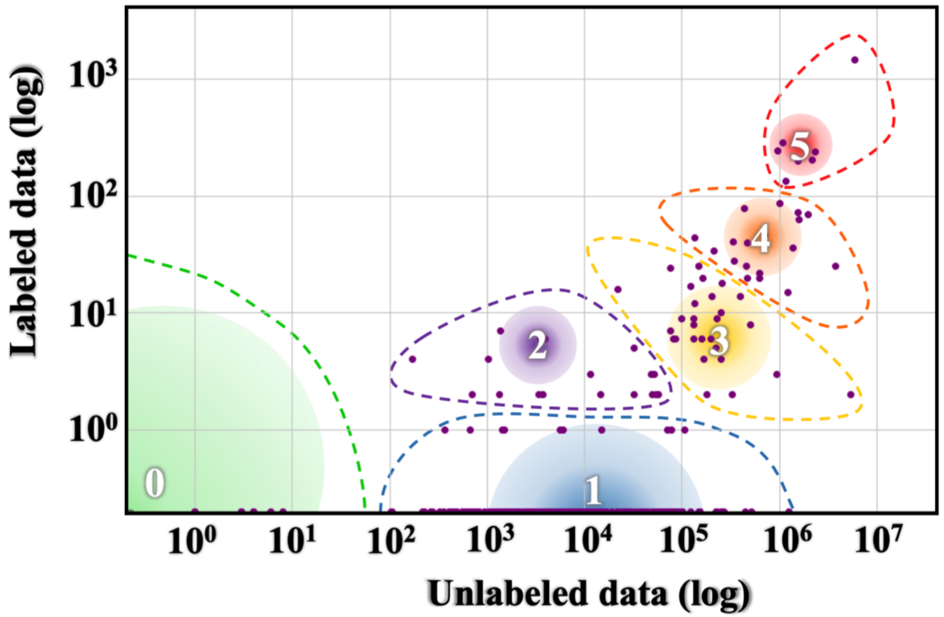
Multilingual NLP

- NLP systems capable of understanding many languages
- Challenges
  1. Linguistic diversity
     - 7000+ word languages, 14+ language families[1]
     - Languages diverge across all levels of language structure
  2. Inequality in available language resources
     - Labeled and unlabeled resources vary across languages

[1] https://www.ethnologue.com/browse/families

Inequality In Language Resources

Development in NLP technology mostly benefited the resource-rich languages (class 5)



Class	#Langs	#Speakers	% of Total Langs
0	2191	1.2B	88.38%
1	222	30M	5.49%
2	19	5.7M	0.36%
3	28	1.8B	4.42%
4	18	2.2B	1.07%
5	7	2.5B	0.28%

Image reference: The State and Fate of Linguistic Diversity and Inclusion in the NLP World, ACL 2020.

High-resource Languages

Languages having large collection of labeled or unlabeled corpora or manually crafted linguistic resources sufficient for building statistical NLP solutions.

Examples: English, Chinese, etc.

Low-resource Languages

Languages lacking large collection of labeled or unlabeled corpora or manually crafted linguistic resources sufficient for building statistical NLP solutions.

Examples: Swahili, Nepali, etc.

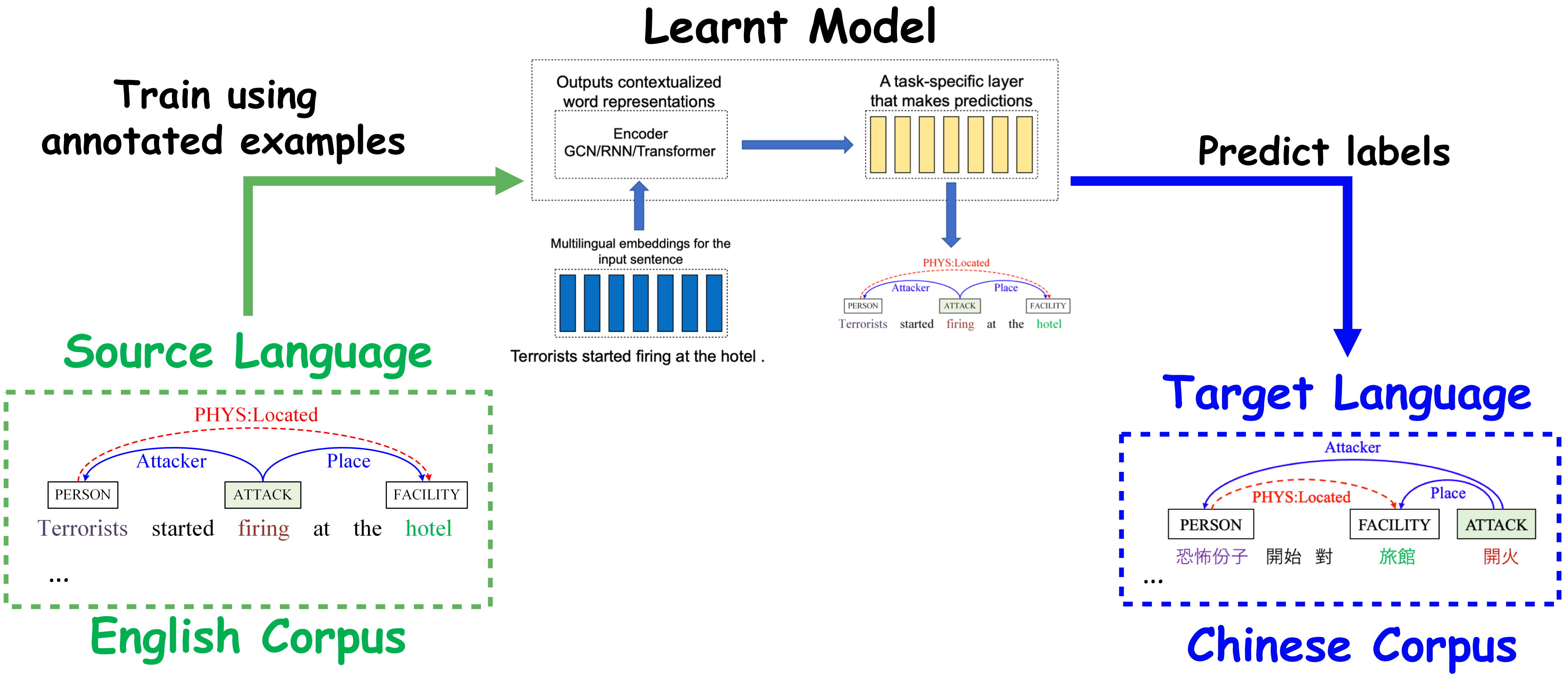
Cross-lingual Transfer

Learn/finetune language representations
in high-resource language(s)

⬇

Use/adapt the learnt representations
in low-resource language(s)

Cross-lingual Transfer



Challenges: Cross-lingual Transfer

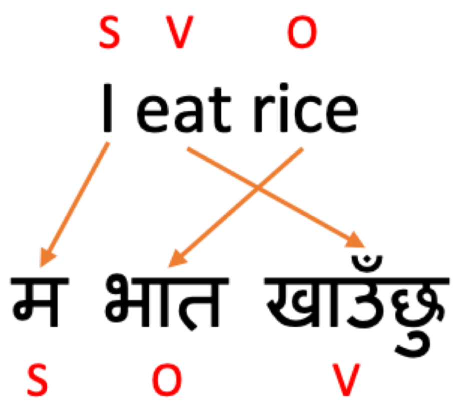
1. Languages differ at levels of morphology, syntax, and semantics

Syntactic Differences

Syntactic differences in terms of word order, word grammar

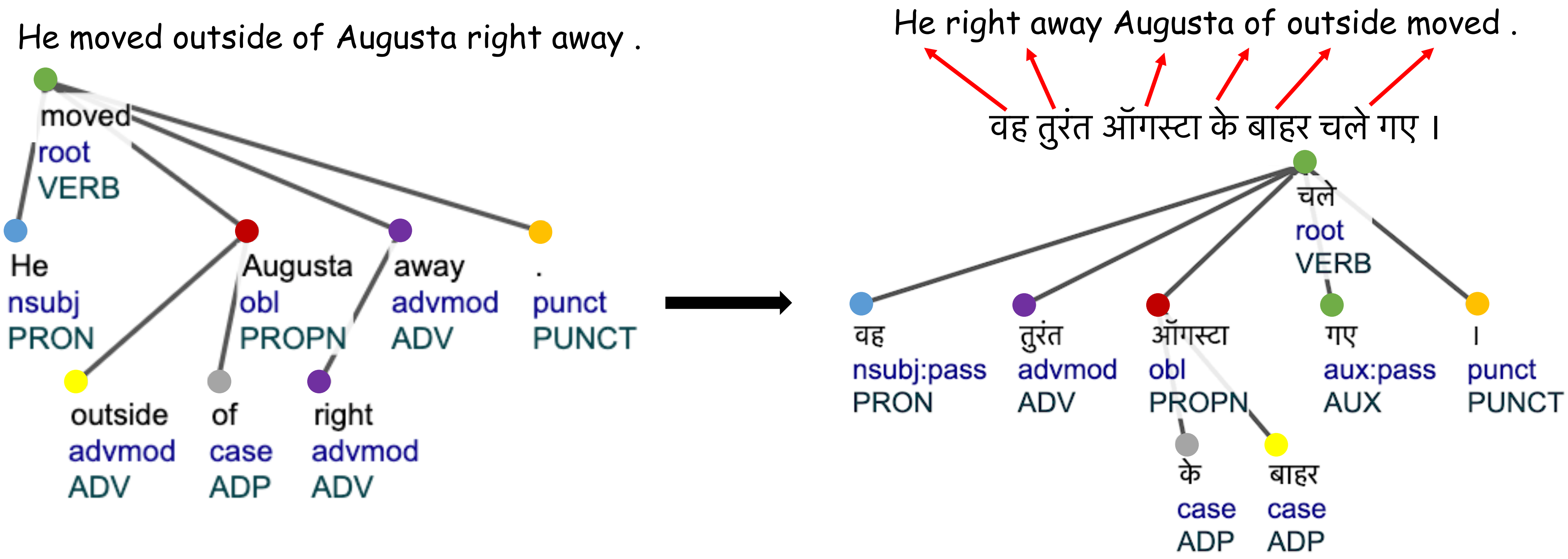
NLP systems typically process a natural language text as a sequence of words, thus word order matters!

English: Subject-Verb-Object (**SVO**)
Nepali: Subject-Object-Verb (**SOV**)



Syntactic Differences

Does utilizing universal language syntax can bridge the syntactic differences across languages?



Thesis Goals (1)

Encoding universal language syntax to bridge typological differences across languages

Challenges: Cross-lingual Transfer

1. Languages differ at levels of morphology, syntax, and semantics

2. Cross-lingual representation learning models often carry language specific information
   - Case1: When models are fine-tuned on high-resource languages
   - Case2: When models are jointly pre-trained on many languages with different scale of pre-training data

Thesis Goals (2)

Using unlabeled resources to facilitate cross-lingual representation learning

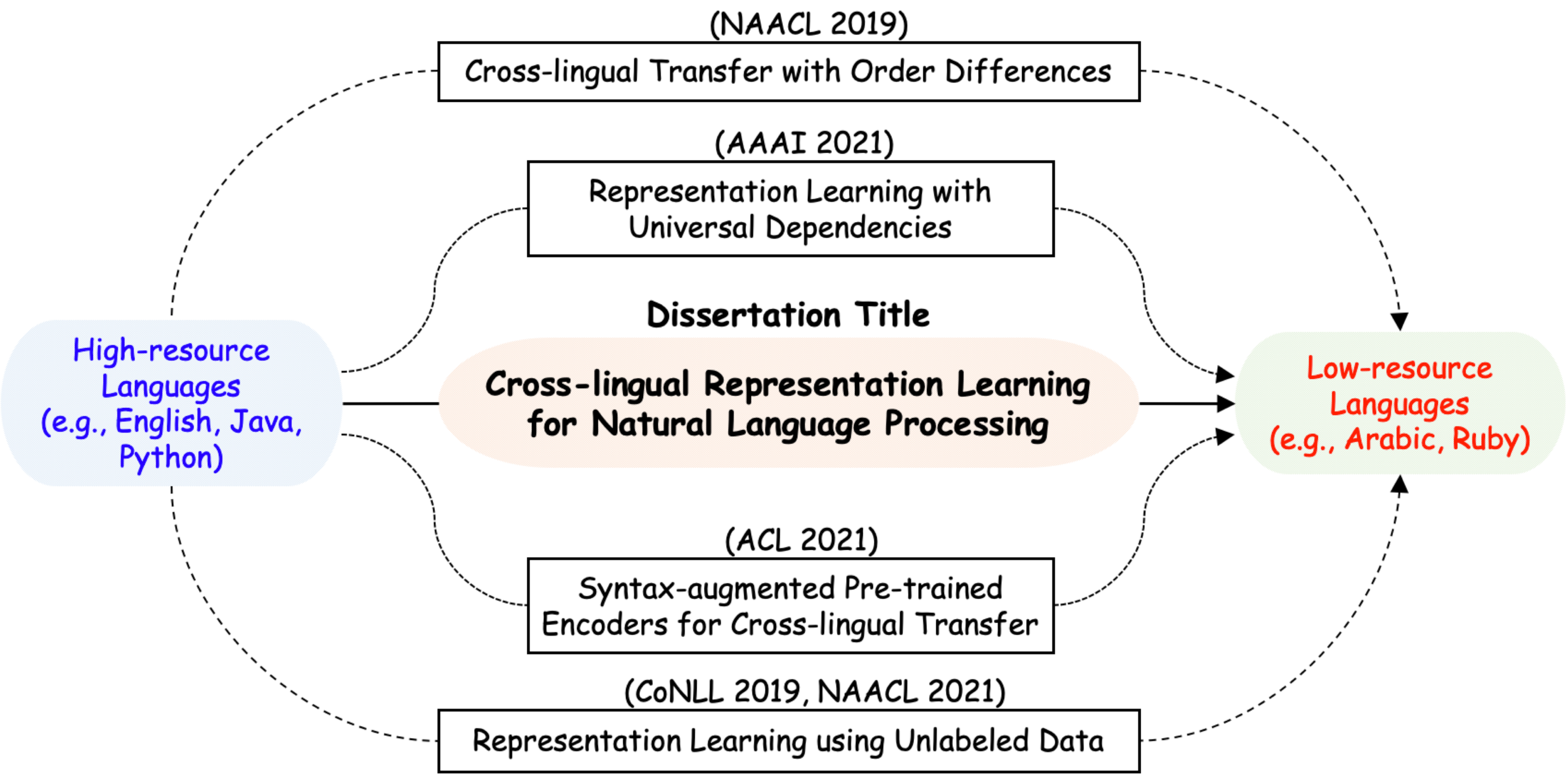
Thesis Statement

Encoding universal language syntax to bridge typological differences across languages

and

utilize unlabeled resources to facilitate cross-lingual representation learning

Overview of Our Works



Contributions

[1] What type of neural architectures are suitable to learn transferable representations? What is the impact of the distance between the source and target languages?

[2] How to improve cross-lingual representations to develop cross-lingual information extraction system?

[3] Does incorporating universal language syntax into multilingual encoders improve cross-lingual transfer?

[4] How to use unlabeled resources to learn robust and generalizable cross-lingual representations?

Outline

[1] Order-free neural architectures improve cross-lingual transfer and more effective when transferred to distant languages [at NAACL'19]

[2] Syntactic distance encoding in representation learning for cross-lingual information extraction [at AAAI'21]

[3] Incorporating universal language syntax into multilingual encoders for cross-lingual transfer [at ACL'21]

[4] Adversarial learning using unlabeled language resources to learn language-agnostic representation [at CoNLL'19]

[4] Unsupervised cross-lingual representation learning for natural and programming languages [at NAACLl'21]

Background

Considering an input text sequence with 8 words.

Input	He	moved	outside	of	Augusta	right	away	.
Word id	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8

- Words are embedded into vectors: $w_i' = w_i W_e$
- The embeddings matrix: $H^0 = [x_1^0, x_2^0, \dots, x_n^0]$ where $x_i^0 = w_i'$

Background

Recurrent neural networks implicitly capture word order

Embeddings matrix: $H^0 = [x_1^0, x_2^0, \dots, x_n^0]$

$$\downarrow$$

$$\overrightarrow{h}_t^l = \overrightarrow{LSTM}_t^l(\overrightarrow{h}_{t-1}^l, x_t^l)$$

$$\overleftarrow{h}_t^l = \overleftarrow{LSTM}_t^l(\overleftarrow{h}_{t+1}^l, x_t^l)$$

$$x_t^l = [\overrightarrow{h}_t^l, \overleftarrow{h}_t^l];\ H^l = [x_1^l, x_2^l, \dots, x_n^l]$$

Background

Self-attention* does not model word order

Embeddings matrix: $H^0 = [x_1^0, x_2^0, \dots, x_n^0]$

$\Downarrow$

$$\mathcal{Q} = H^{l-1}W_l^Q, \mathcal{K} = H^{l-1}W_l^K, \mathcal{V} = H^{l-1}W_l^V$$

$$\mathcal{O} = Attention\,(\mathcal{Q}, \mathcal{K}, \mathcal{V}, \mathcal{M}, d_k) = softmax\left(\frac{\mathcal{Q}\mathcal{K}^T + \mathcal{M}}{\sqrt{d_k}}\right)\mathcal{V}$$

$$H^l = FFNN(\mathcal{O}) = [x_1^l, x_2^l, \dots, x_n^l]$$

* Vaswani et al., 2017

Self-Attention Mechanism

Input "Natural Language" is treated as a bag of words

Input	Word Embedding	Queries	Keys	Values	Attention Scores	Divided by $\sqrt{d_k}$	Softmax Scores	Softmax × Values	Weighted Sum
Natural	x_1	q_1	k_1	v_1	$q_1 \cdot k_1 = 112$ $q_1 \cdot k_2 = 96$	14 12	0.88 0.12	v_1, v_2	o_1
Language	x_2	q_2	k_2	v_2	$q_2 \cdot k_1 = 108$ $q_2 \cdot k_2 = 144$	13 18	0.08 0.92	v_1, v_2	o_2

Image idea courtesy: https://jalammar.github.io/illustrated-transformer

Background

Self-attention* requires to be provided position information

Input	He	moved	outside	of	Augusta	right	away	.
Word id	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
Position id	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8

- Words are embedded into vectors: $w_i' = w_i W_e$
- Positions are embedded into vectors: $p_i' = p_i W_p$
- The embeddings matrix: $H^0 = [x_1^0, x_2^0, \dots, x_n^0]$ where $x_i^0 = w_i' + p_i'$

* Vaswani et al., 2017

Background

Self-attention* requires to be provided order information

$$\mathcal{O} = Attention\,(\mathcal{Q}, \mathcal{K}, \mathcal{V}, \mathcal{M}, d_k) = softmax\left(\frac{\mathcal{Q}\mathcal{K}^T + \mathcal{M}}{\sqrt{d_k}}\right)\mathcal{V}$$

Re-writing the attention weight between token at position i and j

$$\alpha_{ij} = \frac{1}{\sqrt{d_k}}(x_i^l W_l^Q)(x_j^l W_l^K)$$

For layer 1 → $\alpha_{ij} = \frac{1}{\sqrt{d_k}}((w_i' + p_i')W_1^Q)((w_j' + p_j')W_1^K)$

Indicates absolute position of tokens

Recap Thesis Proposal

Order-free model: self-attention with relative positions

$$\mathcal{O} = Attention\,(\mathcal{Q}, \mathcal{K}, \mathcal{V}, \mathcal{M}, d_k) = softmax\left(\frac{\mathcal{Q}\mathcal{K}^T + \mathcal{M}}{\sqrt{d_k}}\right)\mathcal{V}$$

Re-writing the attention weight between token at position i and j

$$\alpha_{ij} = \frac{1}{\sqrt{d_k}}(x_i^l W_l^Q)(x_j^l W_l^K)$$

For layer 1 → $\alpha_{ij} = \frac{1}{\sqrt{d_k}}(w_i' W_1^Q)(w_j' W_1^K + r_{|i-j|}^1)$

* Published at NAACL 2019

Absolute distance between tokens

Recap Thesis Proposal

Representations learnt by order-free models transfer better

XLT-performance (OF-models) > XLT-performance (OS-models)

* Published at NAACL 2019

Recap Thesis Proposal

Given

- Labeled resources in source languages (X^a)
- Unlabeled resources in auxiliary languages (X^b)

Objective

- Adversarially train a model M and a discriminator D such that M does not carry language-specific information

Summary

- We showed that representations learnt by M transfer better

* Published at CoNLL 2019

Outline

[1] Order-free neural architectures improve cross-lingual transfer and more effective when transferred to distant languages [at NAACL'19]

[2] Syntactic distance encoding in representation learning for cross-lingual information extraction [at AAAI'21]

[3] Incorporating universal language syntax into multilingual encoders for cross-lingual transfer [at ACL'21]

[4] Adversarial learning using unlabeled language resources to learn language-agnostic representation [at CoNLL'19]

[4] Unsupervised cross-lingual representation learning for natural and programming languages [at NAACLl'21]

Information Extraction (IE)

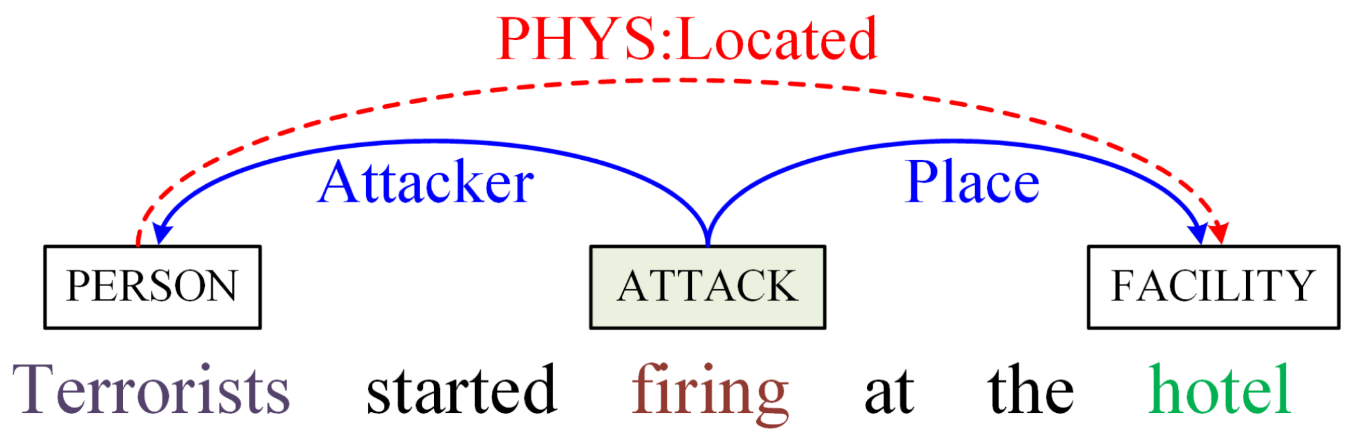


Figure: A relation (red dashed) between two entities and an event of type Attack (triggered by "firing") including two arguments and their role labels (blue) are highlighted.

Challenges

Representations should capture long-range dependencies

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

Distance (fire, hospitalized)

- Sequential = 15

Challenges

Representations should not be sensitive to word order

English follows Subject-Verb-Object (SVO)

A Pakistani court in central Punjab province has sentenced a Christian man to life imprisonment.

Bengali follows Subject-Object-Verb (SOV)

মধ্য পাঞ্জাব প্রদেশের একটি পাকিস্তানি আদালত একজন খ্রিস্টান ব্যক্তিকে যাবজ্জীবন কারাদণ্ড দিয়েছে।

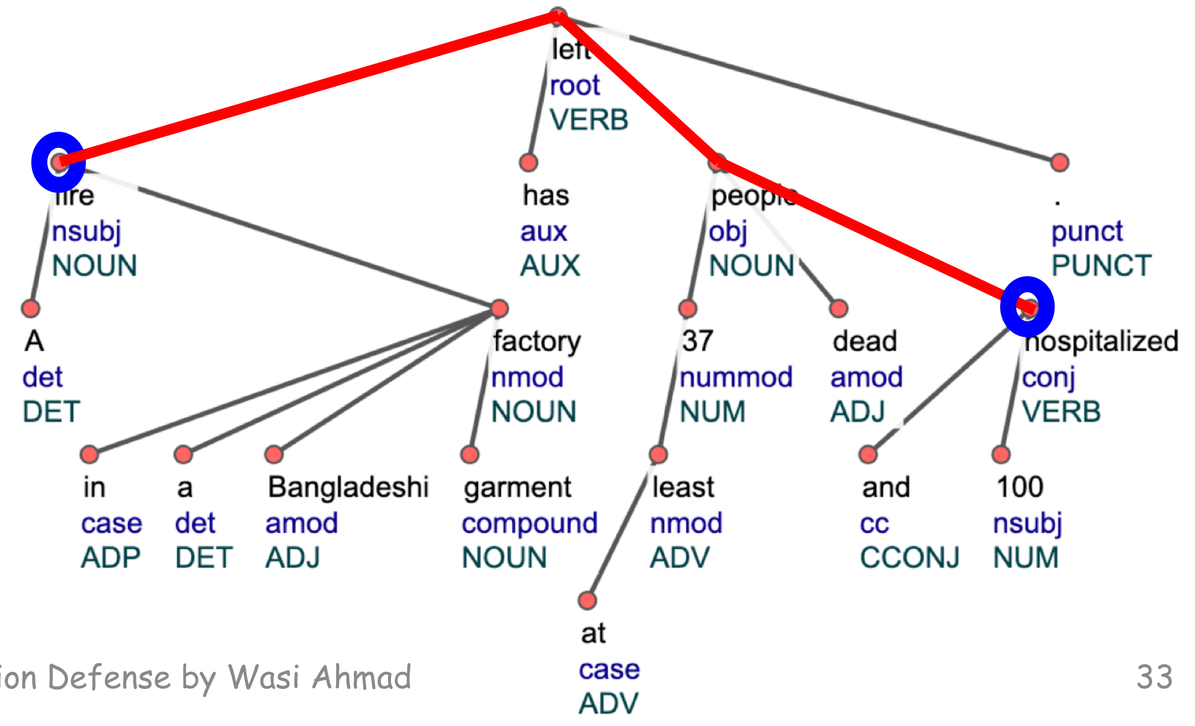
Motivation: Encoding Syntax Structure

Encoding syntax to mitigate long-range dependencies issues

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

Distance (fire, hospitalized)

- Sequential = 15
- Syntactic = 3



Motivation: Encoding Syntax Structure

Encoding syntax to mitigate long-range dependencies issues

According to the popular IE dataset, ACE05

Language	Sequential Distance			Structural Distance		
	English	Chinese	Arabic	English	Chinese	Arabic
Relation mentions	4.8	3.9	25.8	2.2	2.6	5.1
Event mentions and arguments	9.8	21.7	58.1	3.1	4.6	12.3

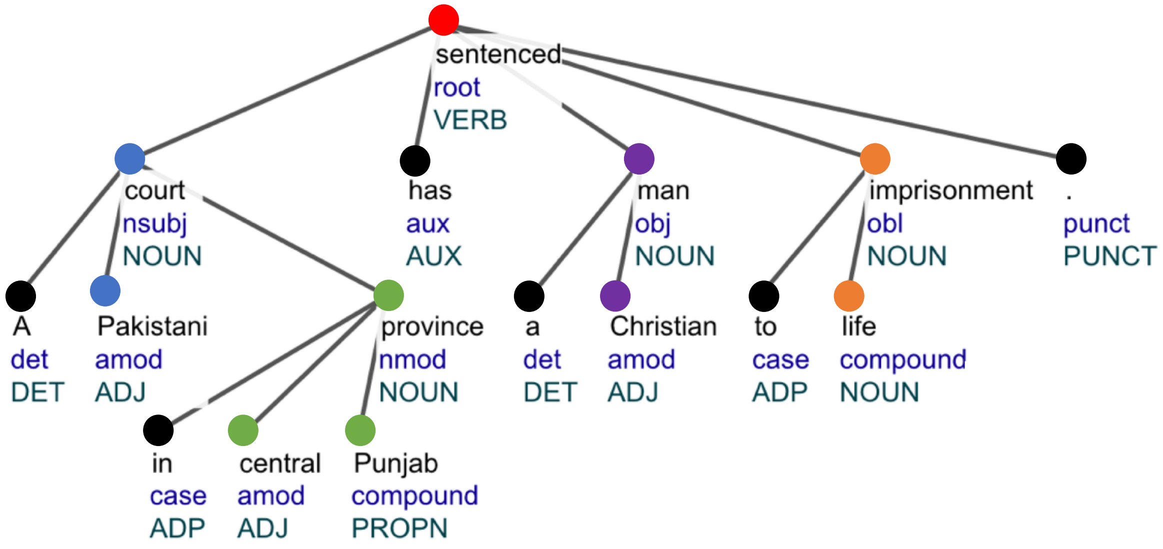
Table: Average sequential and structural (shortest path) distance between relation mentions and event mentions and their candidate arguments. Distances are computed by ignoring the order of mentions.

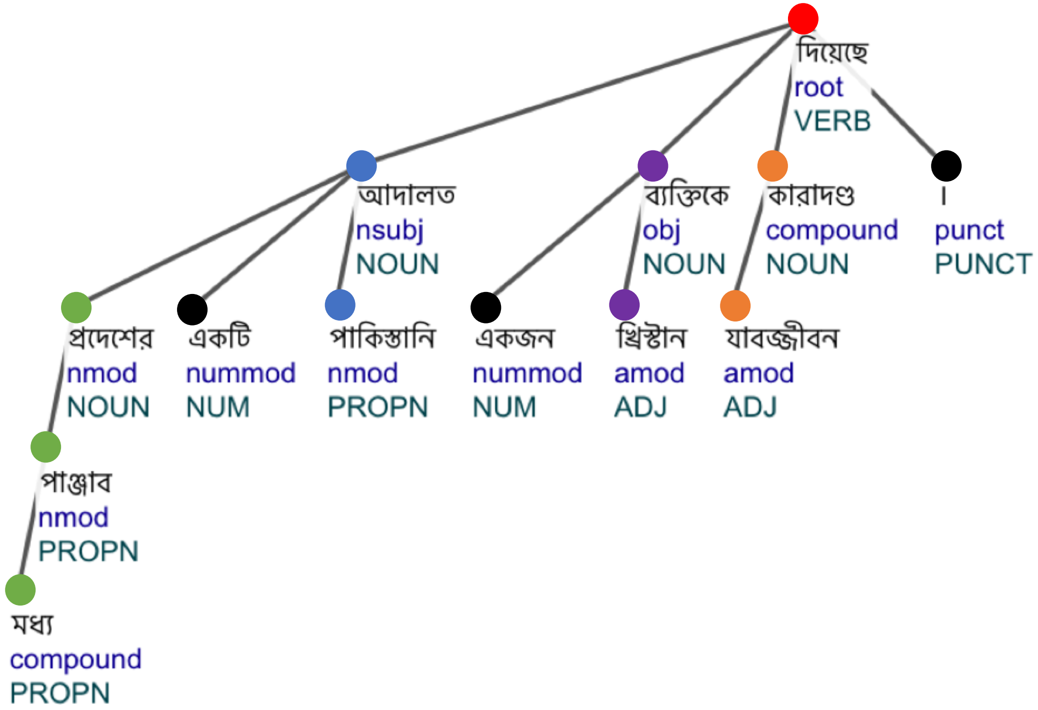
Motivation: Encoding Syntax Structure

Encoding syntax to mitigate word order differences issue

A Pakistani court in central Punjab province has sentenced a Christian man to life imprisonment.

মধ্য পাঞ্জাব প্রদেশের একটি পাকিস্তানি আদালত একজন খ্রিস্টান ব্যক্তিকে যাবজ্জীবন কারাদণ্ড দিয়েছে।



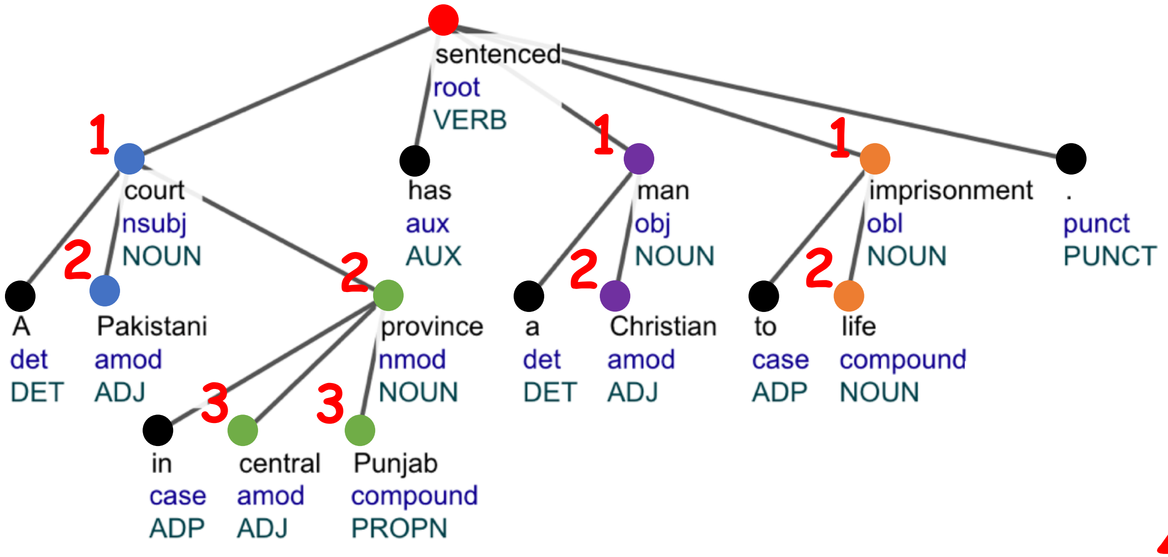


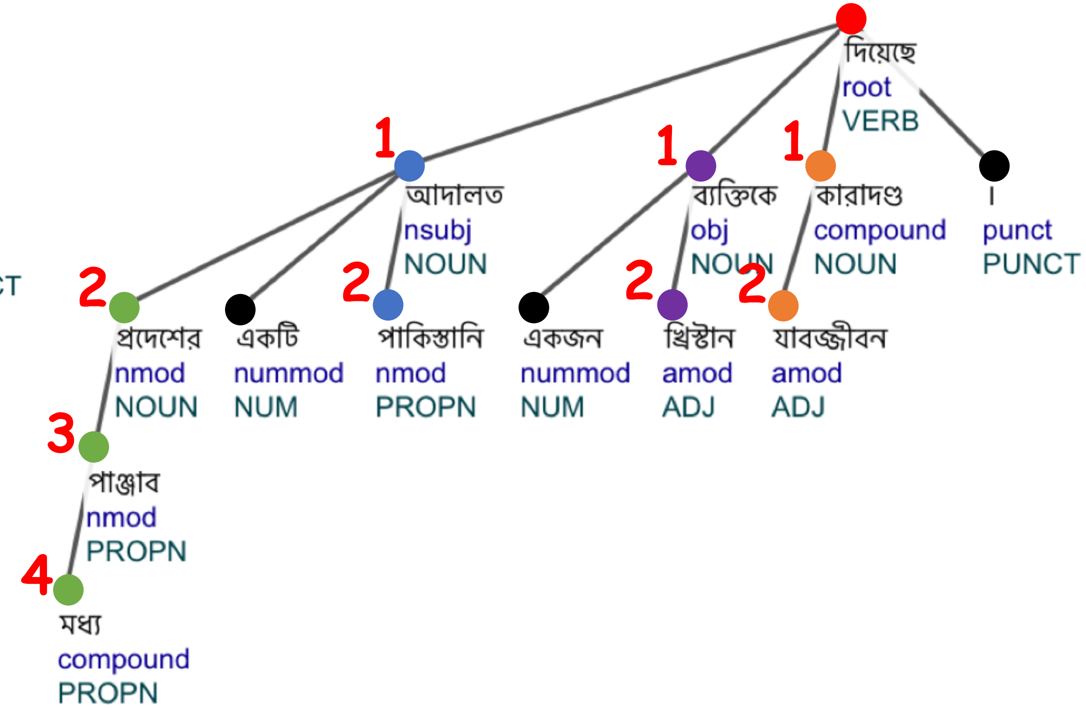
Proposal

Adjust attention between tokens based on syntactic distance

A Pakistani court in central Punjab province has <u>sentenced</u> a Christian man to life imprisonment.

মধ্য পাঞ্জাব প্রদেশের একটি পাকিস্তানি আদালত একজন খ্রিস্টান ব্যক্তিকে যাবজ্জীবন কারাদণ্ড <u>দিয়েছে</u>।





* Published at AAAI 2021

Proposal

Adjust attention between tokens based on syntactic distance

- Pay more attention to tokens that are closer and less attention to tokens that are far away

$$\mathcal{O} = Attention\,(\mathcal{Q}, \mathcal{K}, \mathcal{V}, \mathcal{M}, d_k) = F\left(softmax\left(\frac{\mathcal{Q}\mathcal{K}^T + \mathcal{M}}{\sqrt{d_k}} \right) \right) \mathcal{V}$$

Where, $F(P)_{ij} = \frac{P_{ij}}{Z_i D_{ij}}$

and D_{ij} is syntactic distance between token at position i and j

and Z_i is the normalization factor.

* Published at AAAI 2021

Proposal

Adjust attention between tokens based on syntactic distance

In multi-head attention,

- At each head, restrict tokens to attend tokens that are within a certain distance K

$$\mathcal{O} = Attention\,(\mathcal{Q}, \mathcal{K}, \mathcal{V}, \mathcal{M}, d_k) = F\left(softmax\left(\frac{\mathcal{QK}^T + \mathcal{M}}{\sqrt{d_k}}\right)\right)\mathcal{V}$$

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

* Published at AAAI 2021

Proposal Summary

Syntactic distance-aware self-attention

$$\mathcal{O} = Attention\,(\mathcal{Q},\mathcal{K},\mathcal{V},\mathcal{M},d_k) = F\left(softmax\left(\frac{\mathcal{Q}\mathcal{K}^T + \mathcal{M}}{\sqrt{d_k}}\right)\right)\mathcal{V}$$

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

(1) Allow tokens to attend tokens that are within distance K.

(2) Pay more attention to tokens that are closer and less attention to tokens that are faraway in the syntax tree.

* Published at AAAI 2021

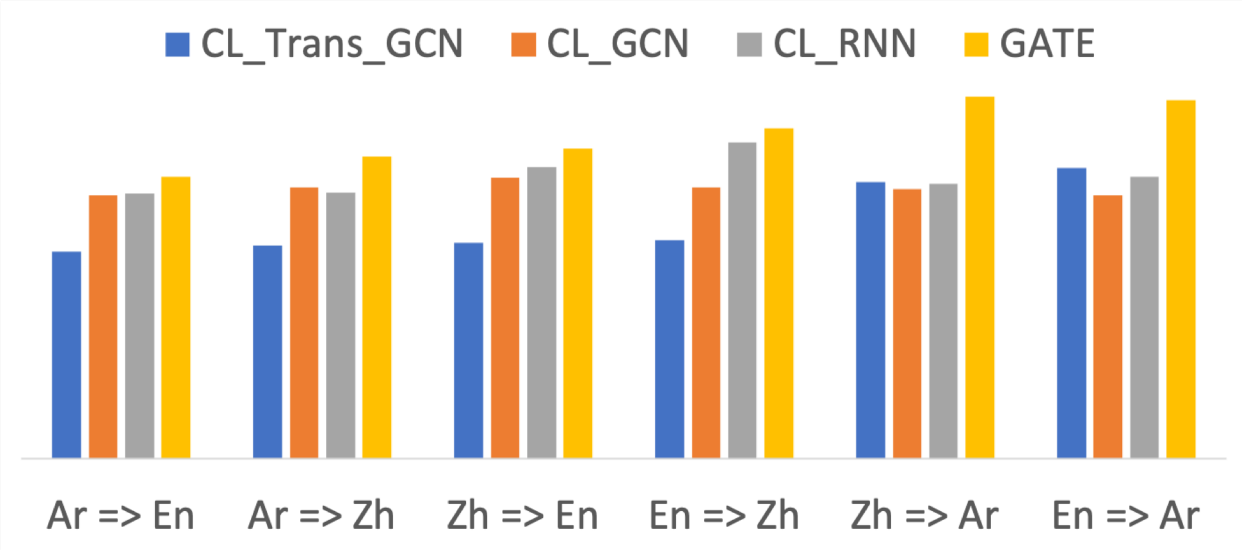
Experiment Results

Event Argument Role Labeling

- Dataset: ACE05 (En, Zh, Ar)
- Performance metric: F-score

Baseline models: CL_Trans_GCN, CL_GCN, CL_RNN

Our proposed model: GATE



* Published at AAAI 2021

Limitations

- Hard restrictions on attention
  - Blocking tokens to attend tokens that are beyond distance K.
- Tree structure cannot be decoded from syntactic distances
  - Parent-child relationship can be decoded given depth of tokens
- Part-of-speech (POS) tags are used as input features
  - POS tags could play a role in determining attention weights
- Cannot be applied to pre-trained language encoders
  - Because of the encoders' own custom vocabulary

Outline

[1] Order-free neural architectures improve cross-lingual transfer and more effective when transferred to distant languages [at NAACL'19]

[2] Syntactic distance encoding in representation learning for cross-lingual information extraction [at AAAI'21]

[3] Incorporating universal language syntax into multilingual encoders for cross-lingual transfer [at ACL'21]

[4] Adversarial learning using unlabeled language resources to learn language-agnostic representation [at CoNLL'19]

[4] Unsupervised cross-lingual representation learning for natural and programming languages [at NAACLl'21]

Proposal

Bias self-attention to provide syntactic clues

$$\mathcal{O} = \ Attention\left(\mathcal{Q} + \mathcal{G}G_l^Q, \mathcal{K} + \mathcal{G}G_l^K, \mathcal{V}, \mathcal{M}, d_k\right)$$

- Where $\mathcal{G}$ is syntax representations learned by a graph attention network (GAT).
- We call the addition terms $(\mathcal{G}G_l^Q, \mathcal{G}G_l^k)$ syntax-bias.
- Intuition - attend tokens with a specific part-of-speech tag sequence or dependencies.

* Published at ACL 2021

Graph Attention Network (GAT)

GAT also uses multi-head attention*

Embeddings matrix: $\mathcal{G}^0 = [g_1^0, g_2^0, \ldots, g_n^0]$

$$\downarrow$$

$$\mathcal{G}^l = \ Attention\ (\mathcal{T}, \mathcal{T}, \mathcal{V}, \mathcal{M}, d_k)$$

- GAT does not employ position representations
  - Only uses word and part-of-speech embeddings, i.e., $g_i^0 = w_i W_e + pos_i W_{pos}$

* Vaswani et al., 2017

Graph Attention Network (GAT)

GAT also uses multi-head attention*

Embeddings matrix: $\mathcal{G}^0 = [g_1^0, g_2^0, \ldots, g_n^0]$

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

$$\mathcal{G}^l = \ Attention\ (\mathcal{T}, \mathcal{T}, \mathcal{V}, \mathcal{M}, d_k)$$

- In GAT, typically $K = 1$
  - Allowing attention between adjacent words only
  - In our work, we find $K = [2, 4]$ is helpful for downstream tasks

* Vaswani et al., 2017

Graph Attention Network (GAT)

GAT also uses multi-head attention*

Embeddings matrix: $\mathcal{G}^0 = [g_1^0, g_2^0, \ldots, g_n^0]$

$$\Downarrow$$

$$\mathcal{G}^l = Attention\,(\mathcal{T}, \mathcal{T}, \mathcal{V}, \mathcal{M}, d_k)$$

- GAT does not use feed-forward sublayer
  - As a result, GAT is light-weight
  - Representations learnt at head h_n at layer l goes to head h_n at layer $l+1$

* Vaswani et al., 2017

Multi-task Fine-tuning

- Both pre-trained encoder and GAT are fine-tuned on the downstream tasks
- GAT is additionally fine-tuned to predict the tree structure
  - Use GAT's output representation to predict the tree distance between all pairs of tokens and the tree depth of tokens
  - Ensures GAT encodes the tree structure accurately

* Published at ACL 2021

Multi-task Fine-tuning

Fine-tune multilingual encoder and GAT on downstream tasks in the source language

$$\mathcal{L} = \mathcal{L}_{task} + \alpha(\mathcal{L}_{dist} + \mathcal{L}_{depth})$$

Task loss in source language

Least square loss for predicting distance between all pairs of tokens

$$d_{\theta_1}(g_i, g_j)^2 = \left(\theta_1(g_i - g_j)\right)^T \left(\theta_1(g_i - g_j)\right)$$

Least square loss for predicting tree depth of tokens

$$d_{\theta_2}(g_i) = (\theta_2 g_i)^T (\theta_2 g_i)$$

* Published at ACL 2021

Experiments

Dataset

- Text classification: XNLI, PAWS-X
- Named entity recognition: Wikiann, CoNLL
- Question answering: MLQA, XQuAD
- Semantic parsing: mTOP, mATIS++

Languages

Source: en; Target: ar, bg, de, el, es, fr, hi, ru, tr, ur, vi, zh, ko, ja, nl, pt

Models

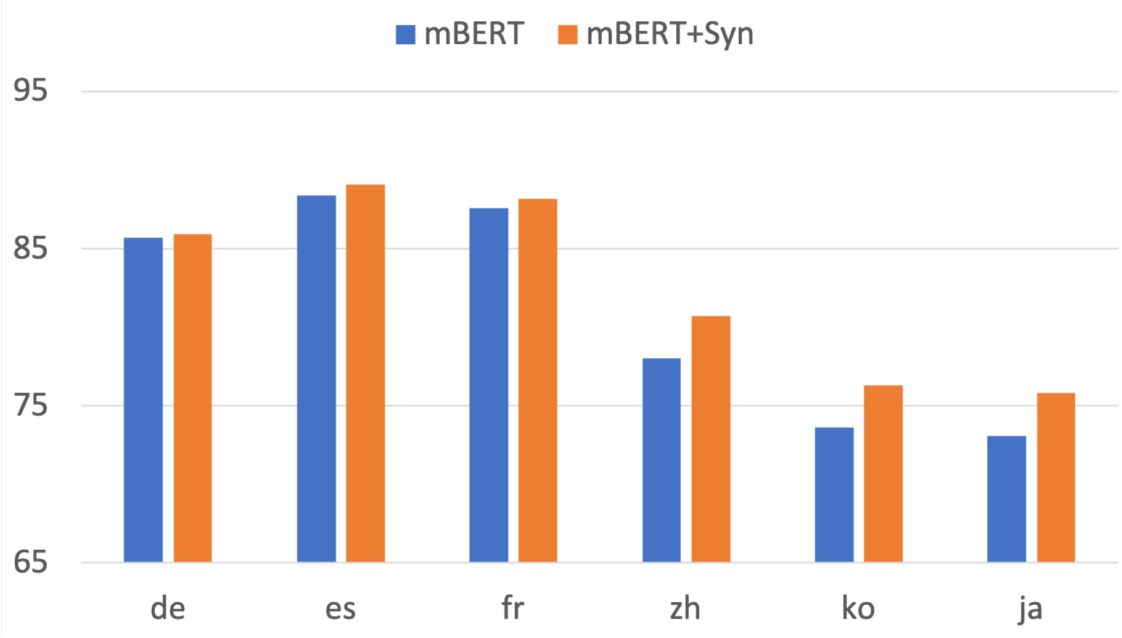
- mBERT: fine-tuned on the pre-processed datasets
- mBERT+Syn: proposed approach

* Published at ACL 2021

Results: Text Classification

Zero-shot transfer results on PAWS-X

- Given a pair of sentences, predict if they are paraphrase



* Published at ACL 2021

Results: Text Classification

Zero-shot generalized transfer results on PAWS-X

- Given a pair of sentences from two different languages, predict if they are paraphrase

s_1/s_2	en	de	es	fr	ja	ko	zh
en	-	0.7	1.6	1.4	4.7	2.5	5.4
de	0.5	-	2.0	2.1	5.1	3.5	5.9
es	1.0	2.1	-	1.7	4.6	3.0	6.6
fr	0.9	1.7	1.9	-	5.0	2.7	5.4
ja	5.2	5.3	5.6	5.1	-	5.9	5.1
ko	3.1	2.8	4.3	3.9	6.4	-	5.1
zh	5.8	5.5	6.3	6.0	6.1	4.5	-

* Published at ACL 2021

Limitations

- *Assumption*: we have access to an off-the-shelf universal parser to collect POS tags and dependency parse structure of the input sequences
- Parsers often normalize the input that lead to inconsistent characters between input text and the output tokenized text (e.g., happens for languages, such as Arabic)

Related Works

Revising positional encoding to mitigate word order issues

- Sinusoidal encoding [Stickland et al., 2020]
- Freezing positional encoding [Liu et al., 2020]
- Applying CNN to encode local n-gram features [Liu et al., 2020]
- Structure-aware position representation [Ding et al., 2020, Wang et al., 2019]

Syntax-aware self-attention

- Dependency-aware self-attention [Deguchi et al., 2019, Bugliarello et al., 2020]
- Syntax-aware Local Attention [Li et al., 2021]
- Syntax-augmented BERT [Sachan et al., 2021]
- Distance-aware Transformer [Wu et al., 2021]

Outline

[1] Order-free neural architectures improve cross-lingual transfer and more effective when transferred to distant languages [at NAACL'19]

[2] Syntactic distance encoding in representation learning for cross-lingual information extraction [at AAAI'21]

[3] Incorporating universal language syntax into multilingual encoders for cross-lingual transfer [at ACL'21]

[4] Adversarial learning using unlabeled language resources to learn language-agnostic representation [at CoNLL'19]

[4] Unsupervised cross-lingual representation learning for natural and programming languages [at NAACLl'21]

Representation Learning for PL & NL

- Developers use programming languages (PL) to develop software and natural language (NL) to document them
- Learning representations for PL & NL can benefit many downstream tasks
  - Code summarization
  - Code generation
  - Code translation
- Can we apply NLP technology to jointly learn representations for PL & NL?

NL vs. PL

NL and PL have similarities

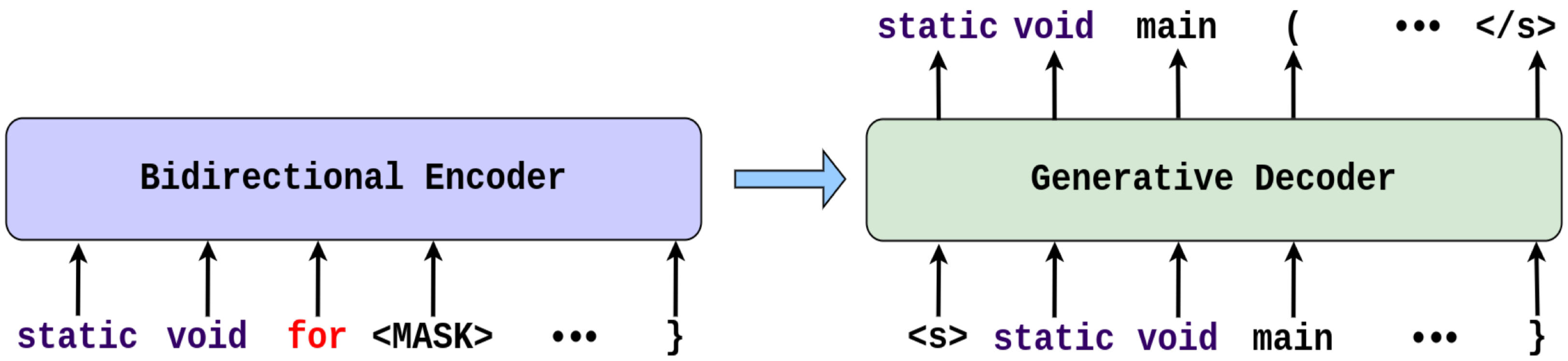
Natural Language	**Programming Language**
Word meaning	Tokens' meaning
Dependency structure	Abstract syntax tree structure
Coreference, events reasoning	Data flow structure
Discourse analysis	Program structure

Unsupervised Representation Learning

- Useful when there is abundant unlabeled data
  - Ex., Java/Python functions from Github
  - Ex., questions/answers from StackOverflow
- Benefits low-resource learning
  - Learning program translation using a few thousands of examples
- How to use unlabeled data for representation learning?

Our Proposal: PLBART

Pre-training Transformer via denoising autoencoding in program and natural languages jointly



* Published at NAACL 2021

Denoising Autoencoding

Three noise functions

- Token masking, token deletion, token infilling [Lewis et al., 2020]

PLBART Encoder Input	PLBART Decoder Output
Is 0 the **[MASK]** Fibonacci **[MASK]** ? <En>	<En> Is 0 the **first** Fibonacci **number** ?
public static main (String args [ ]) { date = Date ( ) ; System . out . (String . format (" Current Date : % tc " ,)) ; } <java>	<java> public static **void** main (String args [ ]) { **Date** date = **new** Date ( ) ; System . out . **printf** (String . format (" Current Date : % tc " , **date**)) ; }
def addThreeNumbers (x , y , z) : NEW_LINE INDENT return **[MASK]** <python>	<python> def addThreeNumbers (x , y , z) : NEW_LINE INDENT return **x + y + z**

* Published at NAACL 2021

Pre-training Corpus

- Java/Python functions from Github
- Questions/answers from StackOverflow
- Up/down sample to balance the corpora

	Java	Python	NL
All Size	352 GB	224 GB	79 GB
All - Nb of tokens	36.4 B	28 B	6.7 B
All - Nb of documents	470 M	210 M	47 M

* Published at NAACL 2021

PLBART on Code Translation

Train/Valid/Test: 10,300/500/1000 [Lu et al. 2021]

Methods	Java to C#			C# to Java		
	BLEU	EM	CodeBLEU	BLEU	EM	CodeBLEU
Transformer	55.84	33.00	63.74	50.47	37.90	61.59
RoBERTa (code)	77.46	56.10	83.07	71.99	57.90	80.18
CodeBERT	79.92	59.00	85.10	72.14	58.80	79.41
GraphCodeBERT	80.58	59.40	-	72.64	58.80	-
PLBART	**83.02**	**64.60**	**87.92**	**78.35**	**65.00**	**85.27**

* Published at NAACL 2021

Summary

- Study shows that PLBART learns program syntax, style, and logical flow
  - e.g., identifier naming convention, “if” block inside an “else” block is equivalent to “else if” block
- PLBART achieves state-of-the-art performance in a wide range of downstream tasks
  - Code summarization, code generation, code translation, program repair, clone detection, and vulnerability prediction

This Thesis

[ACL, 2021] Syntax-augmented Multilingual BERT for Cross-lingual Transfer. **Wasi Ahmad**, Haoran Li, Kai-Wei Chang, and Yashar Mehdad.

[NAACL, 2021] Unified Pre-training for Program Understanding and Generation. **Wasi Ahmad***, Saikat Chakraborty*, Baishakhi Ray, and Kai-Wei Chang.

[AAAI, 2021] GATE: *Graph Attention Transformer Encoder for Cross-lingual Relation and Event Extraction*. **Wasi Ahmad**, Nanyun Peng, and Kai-Wei Chang.

[CoNLL, 2019] Cross-lingual Dependency Parsing with Unlabeled Auxiliary Languages. **Wasi Ahmad**, Zhisong Zhang, Xuezhe Ma, Kai-Wei Chang, and Nanyun Peng.

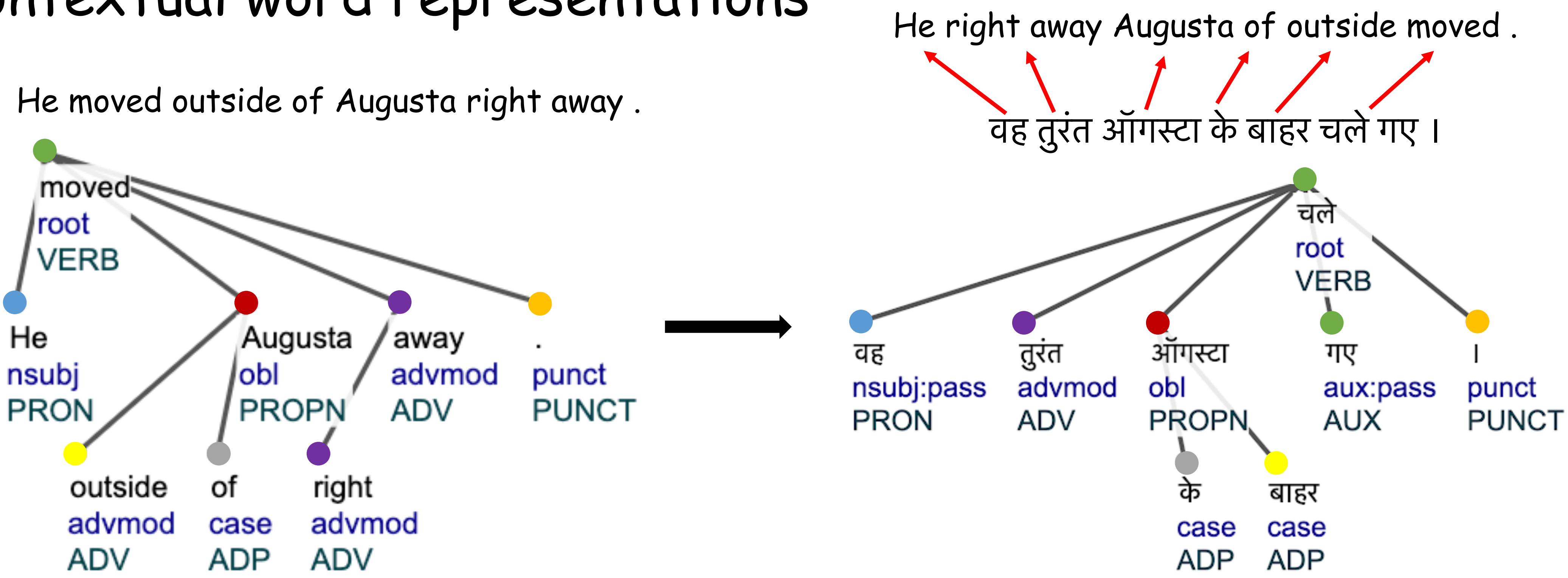
[NAACL, 2019] On Difficulties of Cross-Lingual Transfer with Order Differences: A Case Study on Dependency Parsing. **Wasi Ahmad***, Zhisong Zhang*, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng.

Summary of Findings

- Models carrying less word order information transfers better, specially to distant languages
- Incorporating universal language syntax into multilingual representations improve cross-lingual transfer
- Unlabeled data can be leveraged to learn representations to benefit cross-lingual transfer

Future Works

Role of language syntax in improving alignment of multilingual contextual word representations



Future Works

Cross-lingual representation learning across domains

- In social media, users often use code-mixed language
- Develop ways for feature representations that smoothens the differences in the two languages

References

[Wu et al., 2021] DA-Transformer: Distance-aware Transformer

[Li et al., 2021] Improving BERT with Syntax-aware Local Attention

[Sachan et al., 2021] Do Syntax Trees Help Pre-trained Transformers Extract Information?

[Liu et al., 2021] On the Importance of Word Order Information in Cross-lingual Sequence Labeling

[Lu et al., 2021] Codexglue: A machine learning benchmark dataset for code understanding and generation

[Lewis et al., 2020] BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

[Liu et al., 2020] Improving Zero-Shot Translation by Disentangling Positional Information

References

[Stickland et al., 2020] Recipes for Adapting Pre-trained Monolingual and Multilingual Models to Machine Translation

[Bugliarello et al., 2020] Enhancing Machine Translation with Dependency-Aware Self-Attention

[Ding et al., 2020] Self-Attention with Cross-Lingual Position Representation

[Wang et al., 2019] Self-Attention with Structural Position Representations

[Deguchi et al., 2019] Dependency-Based Self-Attention for Transformer NMT

[Vaswani et al., 2017] Attention Is All You Need

Other Publications (2017-21)

[EMNLP, 2021] Improving Zero-Shot Cross-Lingual Transfer Learning via Robust Training. Kuan-Hao Huang, **Wasi Ahmad**, Nanyun Peng, and Kai-Wei Chang.

[EMNLP-findings, 2021] Retrieval Augmented Code Generation and Summarization. Md Rizwan Parvez, **Wasi Ahmad**, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang.

[ACL, 2021] Select, Extract and Generate: Neural Keyphrase Generation with Layer-wise Coverage Attention. **Wasi Ahmad**, Xiao Bai, Soomin Lee, and Kai-Wei Chang.

[ACL, 2021] Intent Classification and Slot Filling for Privacy Policies. **Wasi Ahmad***, Jianfeng Chi*, Tu Le, Thomas Norton, Yuan Tian, and Kai-Wei Chang.

[EMNLP-findings, 2020] PolicyQA: A Reading Comprehension Dataset for Privacy Policies. **Wasi Ahmad***, Jianfeng Chi*, Yuan Tian, and Kai-Wei Chang.

[ACL, 2020] A Transformer-based Approach for Source Code Summarization. **Wasi Ahmad**, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang.

Other Publications (2017-21)

[SIGIR, 2019] Context Attentive Document Ranking and Query Suggestion. **Wasi Ahmad**, Kai-Wei Chang, and Hongning Wang.

[Journal of Computational Biology, 2019] Word and sentence embedding tools to measure semantic similarity of Gene Ontology terms by their definitions. Dat Duong, **Wasi Ahmad**, Eleazar Eskin, Kai-Wei Chang, and Jingyi Jessica Li.

[ICLR, 2018] Multi-Task Learning for Document Ranking and Query Suggestion. **Wasi Ahmad**, Kai-Wei Chang, and Hongning Wang.

[SIGIR, 2018] Intent-aware Query Obfuscation for Privacy Protection in Personalized Web Search. **Wasi Ahmad**, Kai-Wei Chang, and Hongning Wang.

[LREC, 2018] A Corpus to Learn Refer-to-as Relations for Nominals. **Wasi Ahmad** and Kai-Wei Chang.

Thanks to Collaborators

- Kai-Wei Chang
- Nanyun Peng
- Hongning Wang
- Baishakhi Ray
- Yuan Tian
- Eduary Hovy
- Thomas Norton
- Eleazar Eskin
- Jingyi Jessica Li
- Yashar Mehdad
- Xiao Bai
- Soomin Lee
- Haoran Li
- Xuezhe Ma
- Dat Duong
- Zhisong Zhang
- Jianfeng Chi
- Tu Le
- Kuan-Hao Huang
- Saikat Chakraborty
- Md Rizwan Parvez
- Xueing Bai
- Chao Jiang
- Zhechao Huang
- Md Masudur Rahman

Thank You!

Questions?