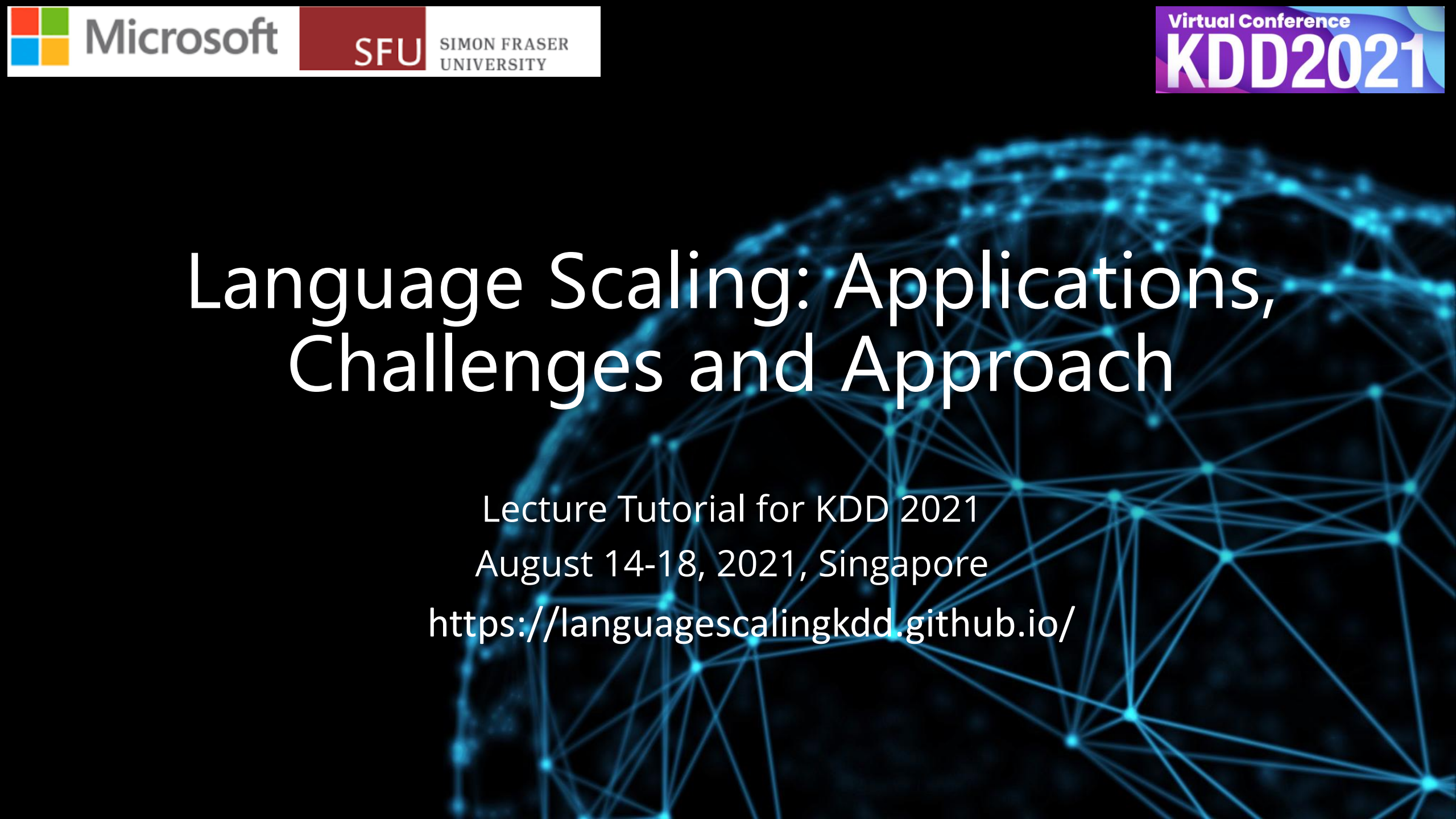


Language Scaling: Applications, Challenges and Approach

Lecture Tutorial for KDD 2021

August 14-18, 2021, Singapore

https://languagescalingkdd.github.io/

Outline

- Introduction [Daxin Jiang]
  - Motivating examples in Microsoft products
  - Problem description
  - Categorization of applications
  - Challenges
- Methodology [Daxin Jiang]
  - Model Transfer
  - Data Transfer
- Applications*
  - Dependency Parsing [Xiubo Geng]
  - Machine Reading Comprehension [Ming Gong]
  - Grammar Error Correction [Linjun Shou]
- Summary & Future directions [Jian Pei]

*For more applications, please refer to our tutorial at The Web Conference 2021

Daxin Jiang
Software Technology Center at Asia (STCA) of Microsoft

Linjun Shou Xiubo Geng Ming Gong
Software Technology Center at Asia (STCA) of Microsoft

Jian Pei
Simon Fraser University

Why language matters (society)

[Why Languages Matter | SIL International]

Eradicate extreme poverty and hunger
Higher literacy rates often result in higher per capita incomes.

Achieve universal primary education
Primary education programs that begin in the mother tongue help students gain literacy and numeracy skills more quickly.

Promote gender equality and empower women
Nearly two-thirds of the world's 875 million illiterate people are women.

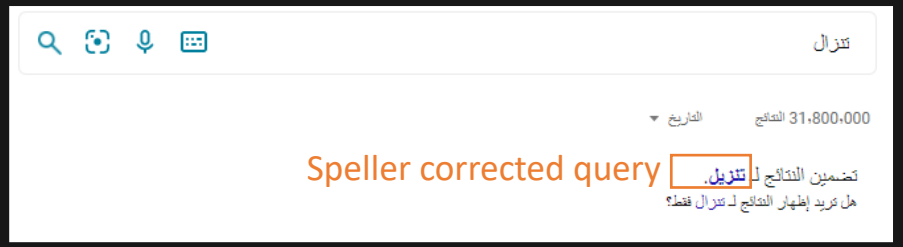
Reduce child mortality
The mortality rate for children under five years of age is reduced when vital health information about disease prevention and treatment is available in local languages.

Why Language Matters (Industry)

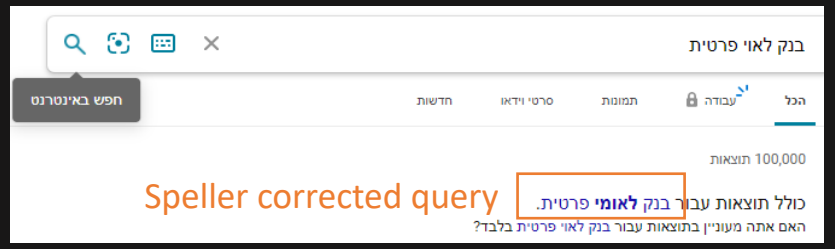
- Microsoft's mission statement: to empower ***every person and every organization on the planet*** to achieve more.
- NLP is widely adopted in Microsoft products and services

Microsoft

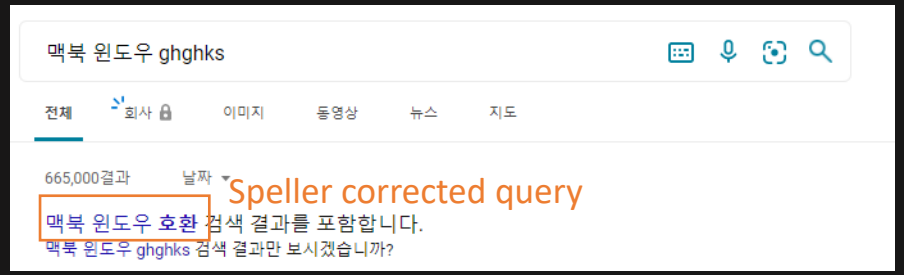
ar-EG
{تنزال}
English Translation:
download



he-IL
{בנק לאוי פרטית}
English Translation:
Private National Bank

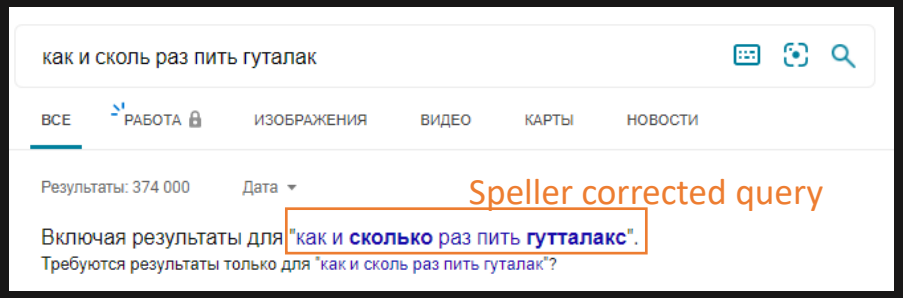


ko-KR
{맥북 윈도우 ghghks}
English Translation:
MacBook Windows
Compatible

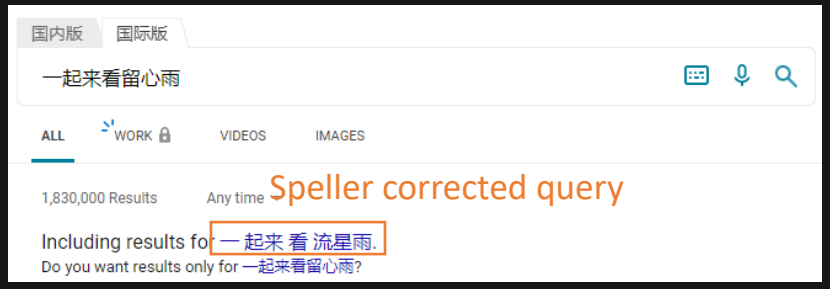


ja-JP
{田中みどらら}
English Translation:
Ms. Tanaka

ru-RU
{как и сколь раз пить гуталак}
English Translation:
how many times to
drink guttalax



zh-CN
{一起来看流心雨}
English Translation:
Watch meteor
together (TV series)



Scalability vs. Specificity

Scalability: scaling out to many languages
Specificity: language specific features

Japanese Speller

- Four scripts (Kanji, Hiragana, Katakana and Romaji)
- Two IME (Romaji on PC, and Kana on mobile

Example: みやさき vs. 宮崎

	typo	correction
Kana	みやさき	みやざき
Romaji	miyasaki	miyazaki

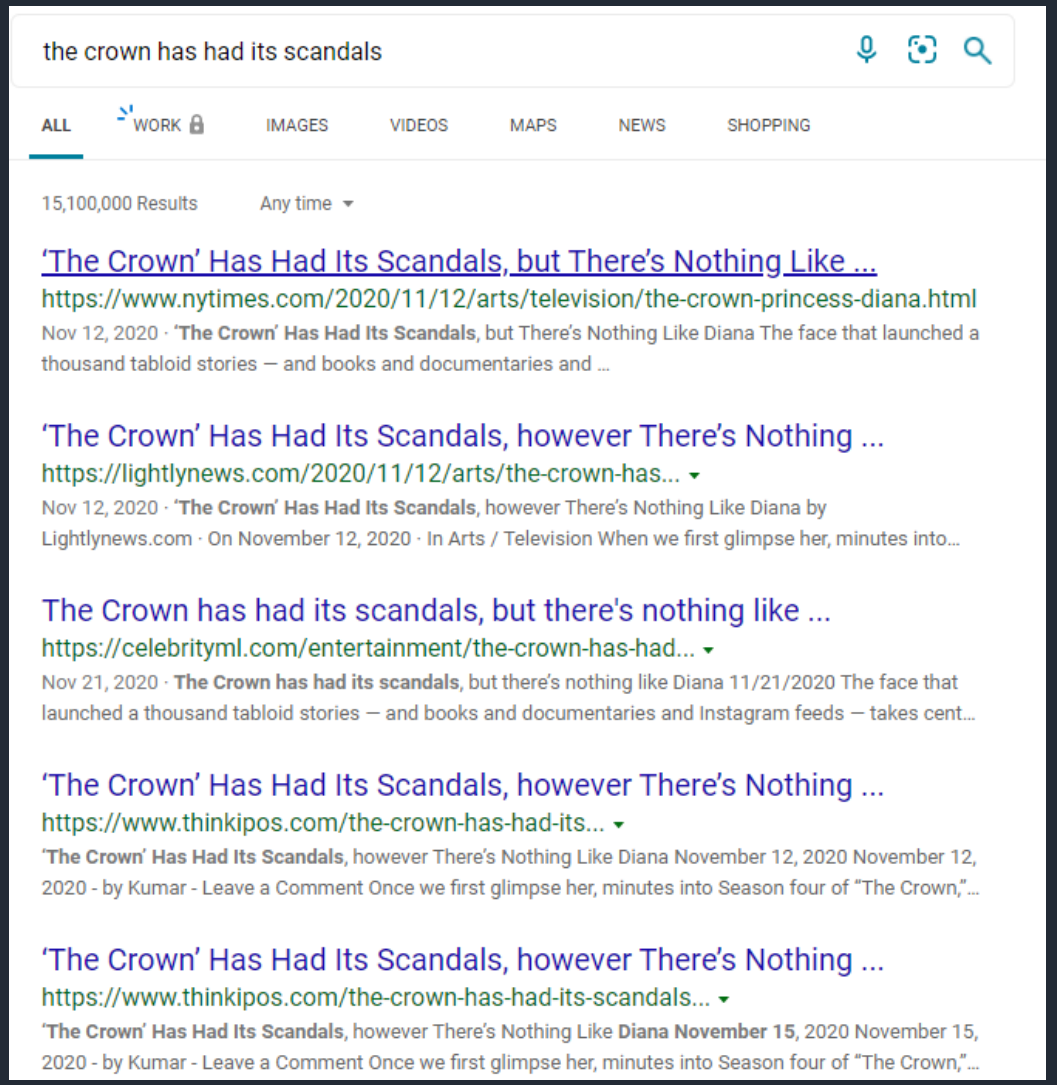
1 keystroke difference

- Need language-specific training data
- Trade off between scalability and specificity to meet the highest ROI (Return Over Invest)

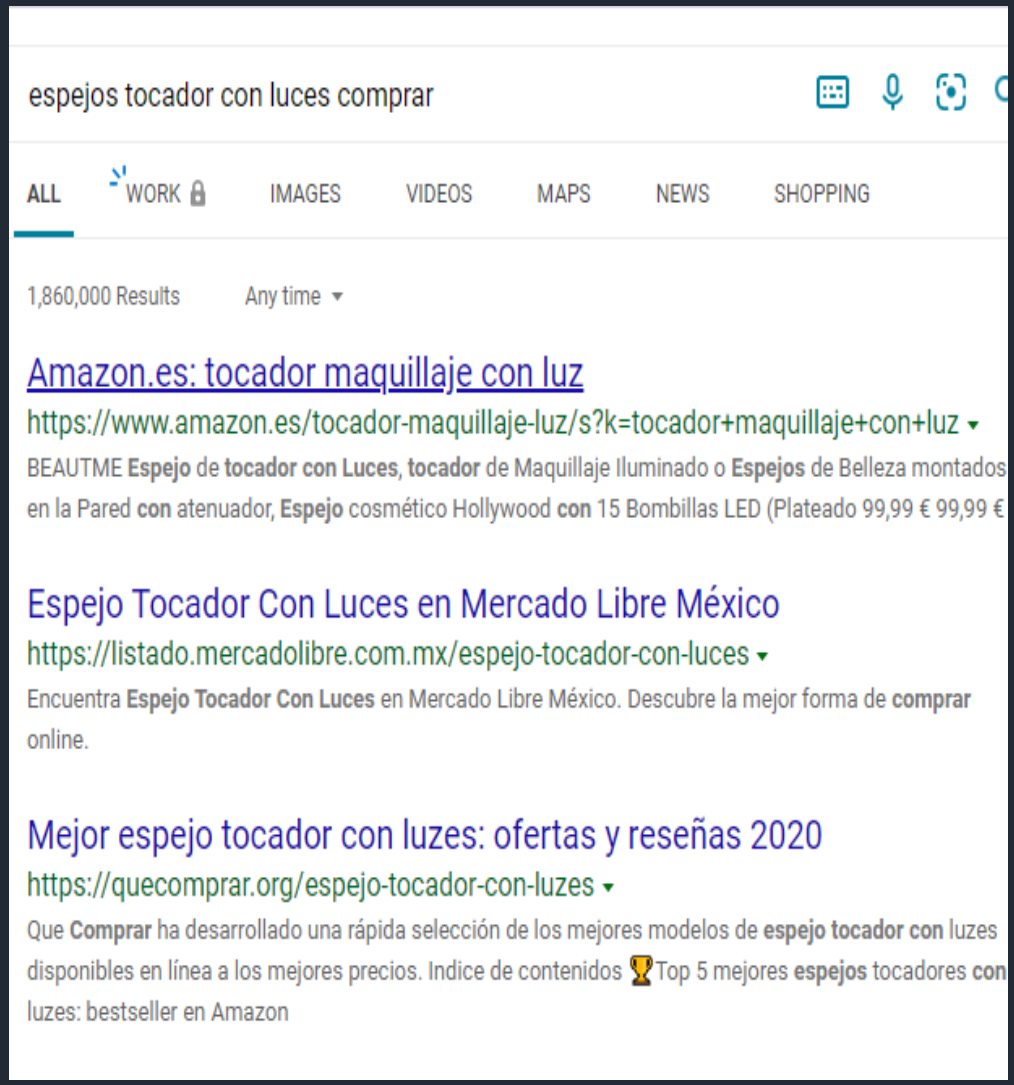
Speller | **Web Relevance** | Suggested Replies | Natural Language Understanding

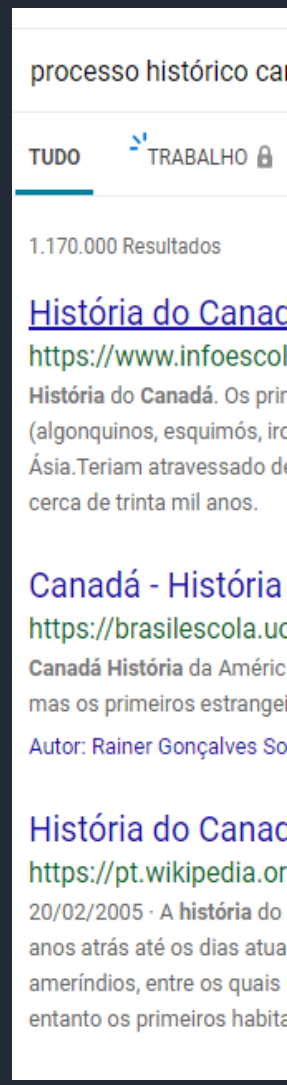
Microsoft

English

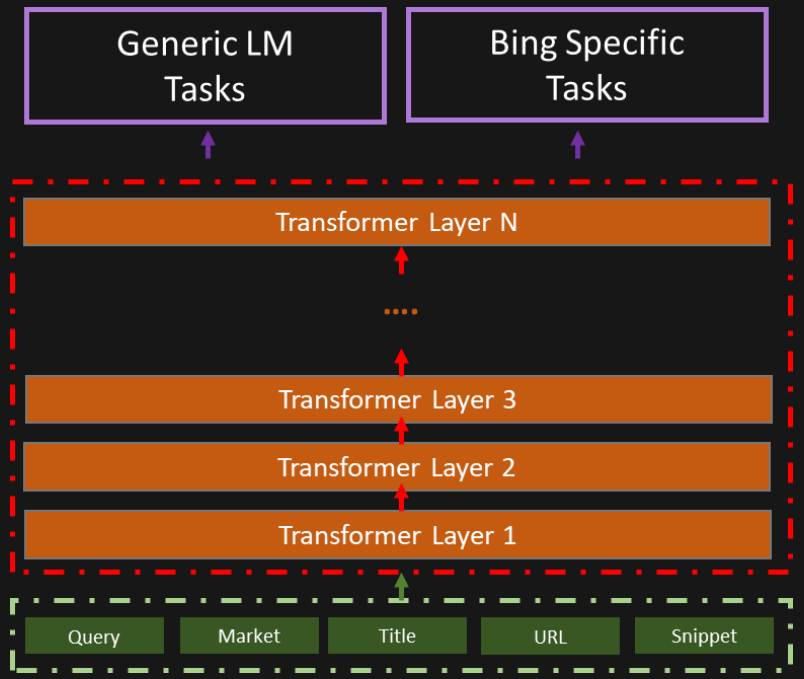


Spanish

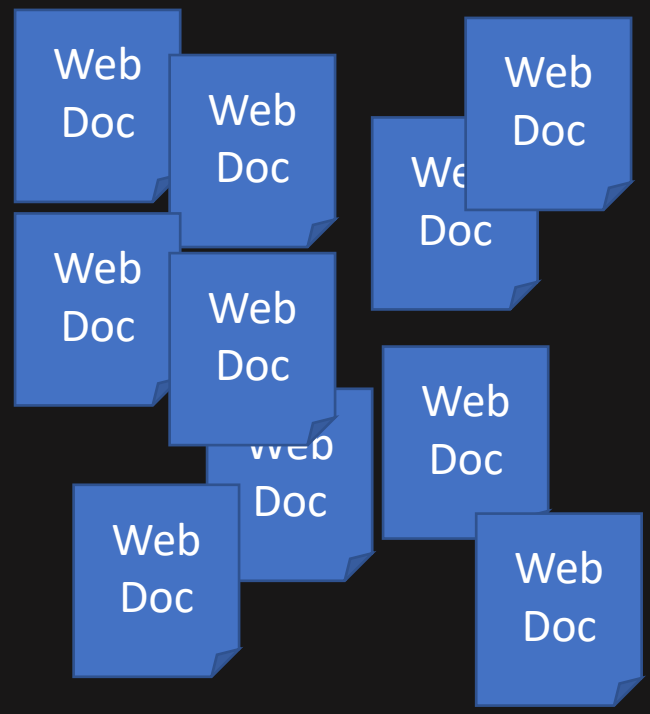




Efficiency



Huge model

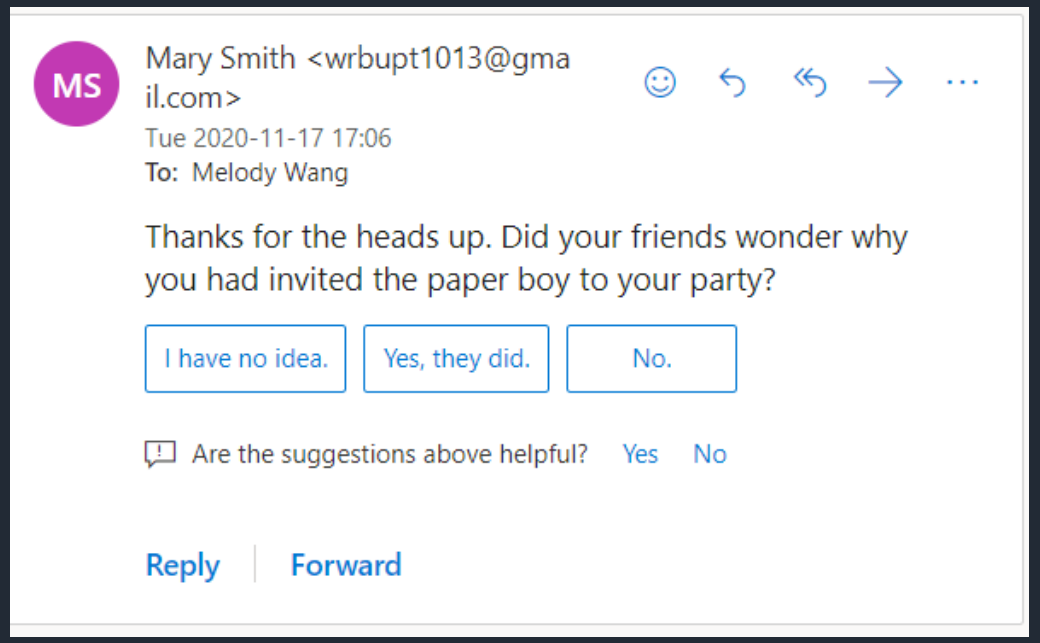


Huge index

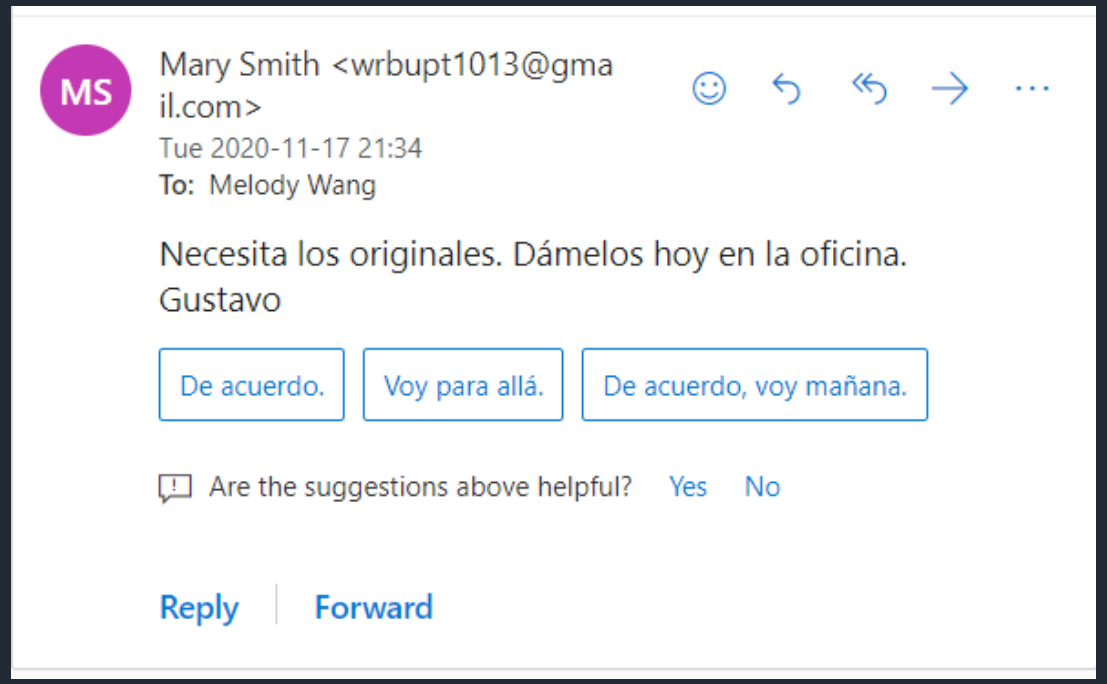
For every query, rank the top O(10) most relevant documents within several hundred milliseconds

Microsoft

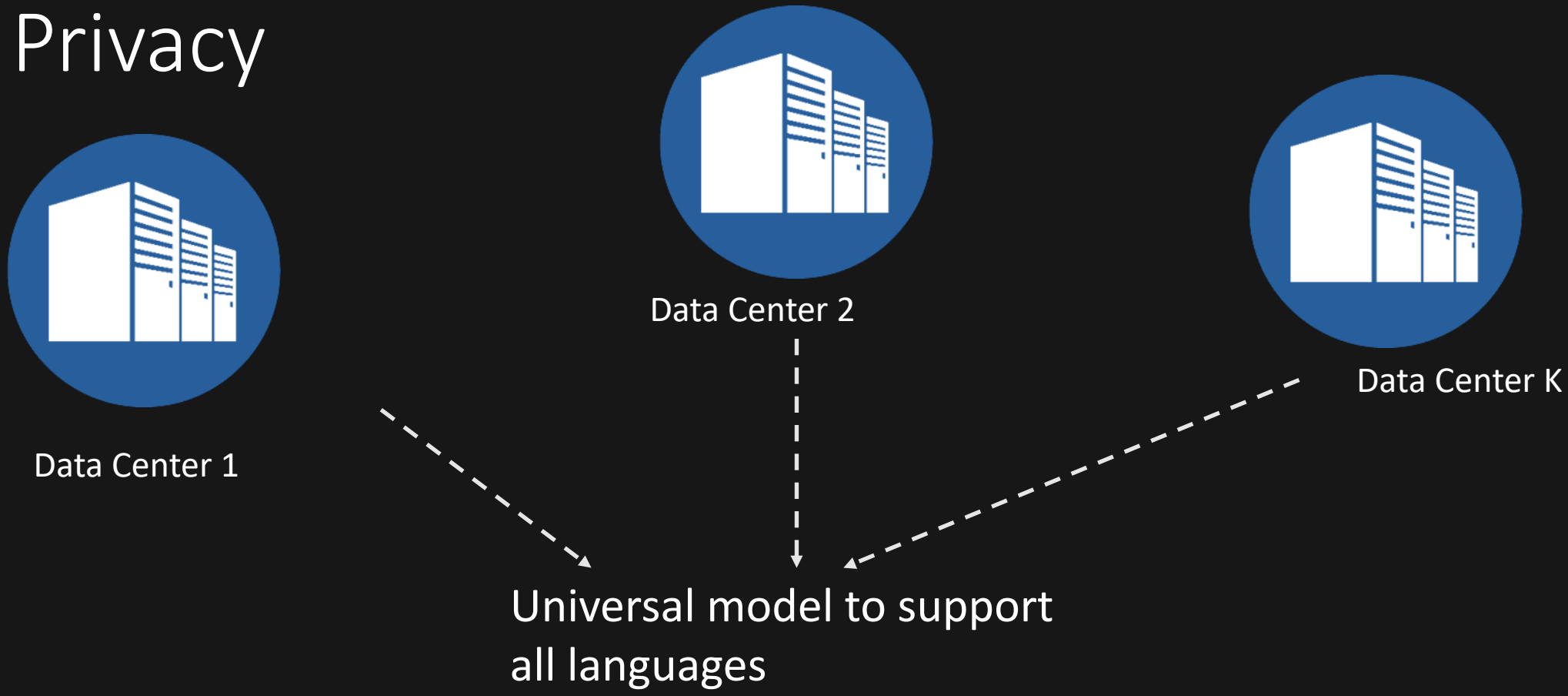
English



Spanish



Privacy



According to compliance restrictions, data cannot be moved out of the data center
How to use the data in all data centers to train a universal model?

Speller | Web Relevance | Suggested Replies | **Natural Language Understanding**

Microsoft

Teams Voice Skill:

Join my next meeting

Domain: Calendar

Intent: Connect_to_meeting

Slot: <start_time>next</start_time>

Start the presentation

Quando è il mio incontro con Tom

Message à Carlos que je serai en retard

打开未读消息

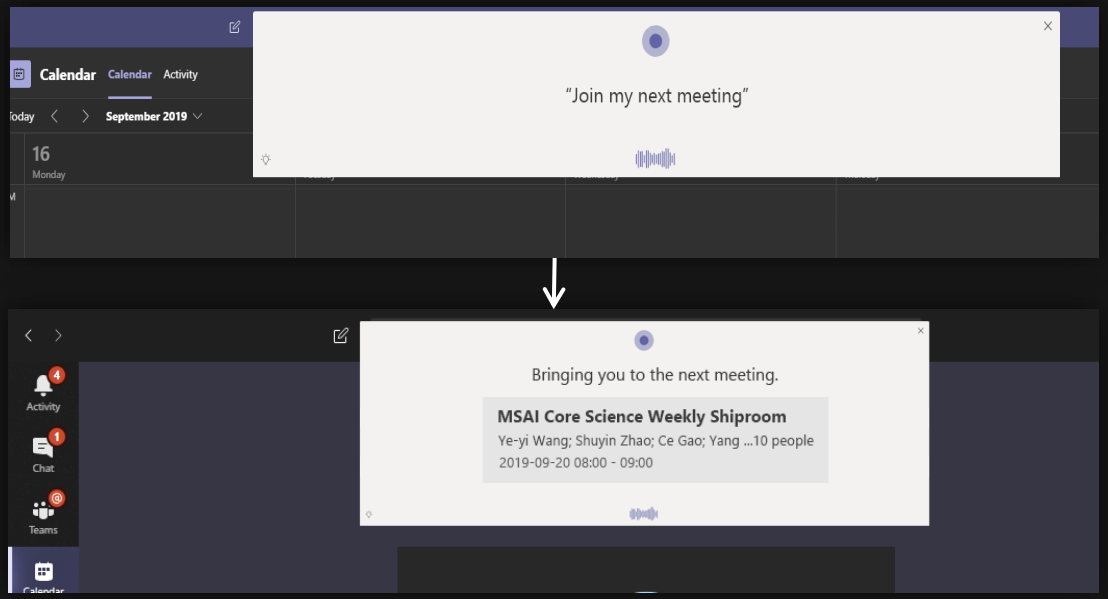
Compartilhe o arquivo com Bob

Agrega Alice a este chat

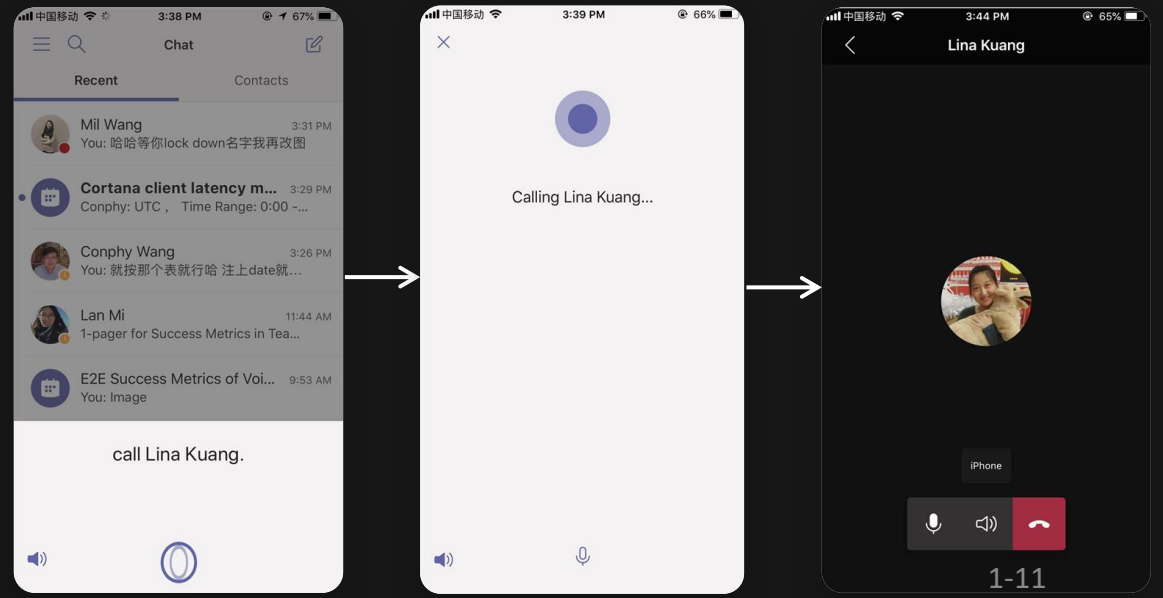
Füge Alice zum Cortana - Kanal hinzu

次の会議に参加する

Desktop Experience



Mobile Experience

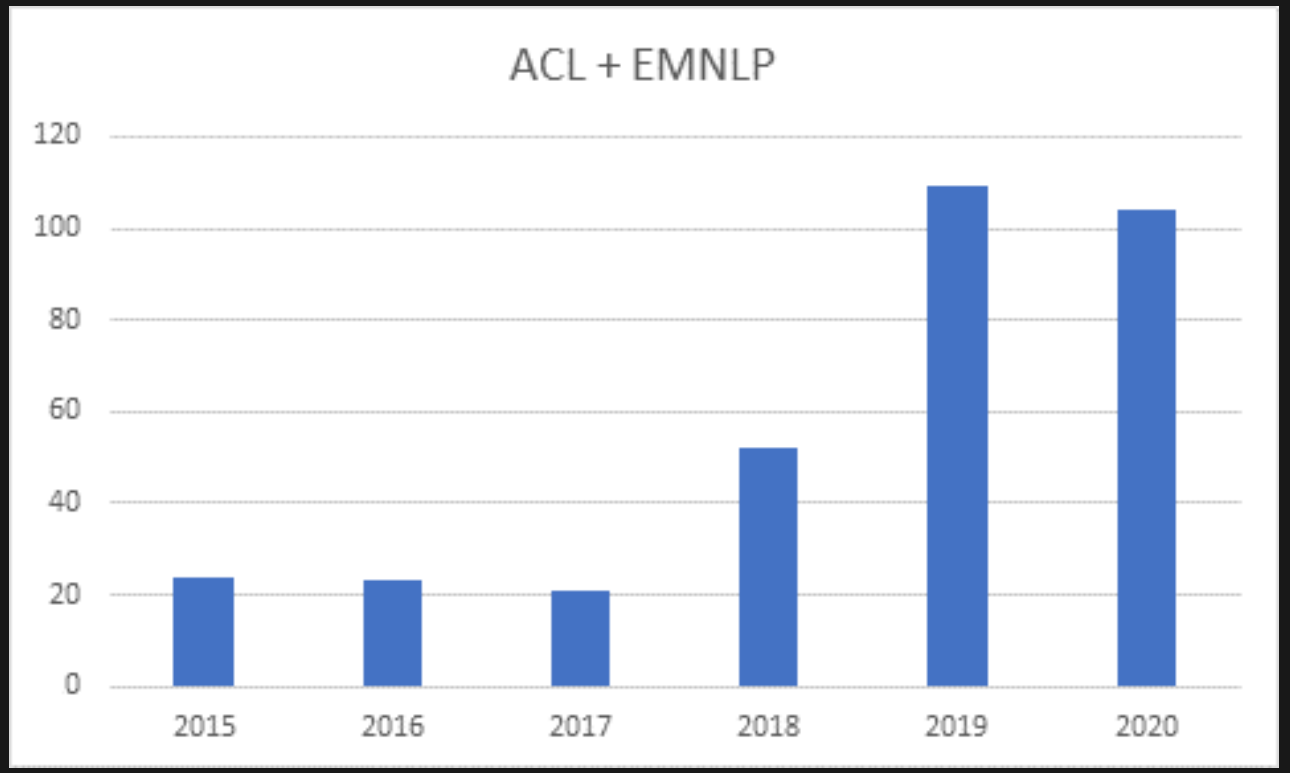


M Domains × N Languages

- Need large amounts of training data
- Need to handle many models
  - Domain classification model, intent detection model, slot tagging model
  - Training, deployment, maintenance

Why Language Matters (Research)

Searching for “multilingual”, “cross-lingual” and “bilingual” in the ACL anthology (ACL+EMNLP)



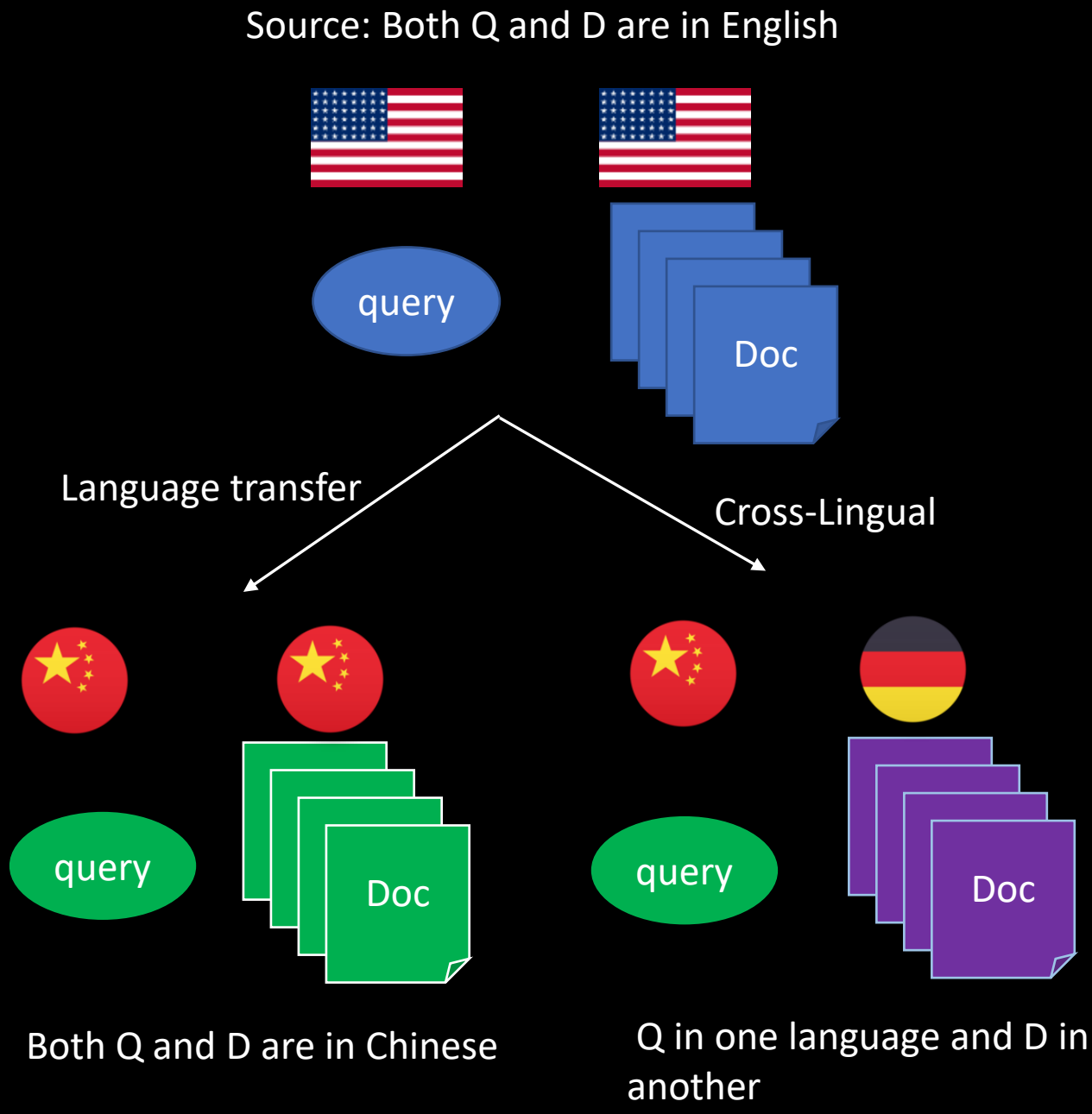
Problem Description for Language Scaling

Given an NLP application (such as spelling correction, Web search, suggested reply or NLU)

1. Usually have relatively rich English training data, and well-trained English model
2. Scale out the model to **100+ languages** ***effectively*** and ***efficiently***

A Related Problem

- **Language transfer**
  - Given English training data, transfer to other languages
- **Cross-lingual**
  - e.g., in Web search, query is in one language and document is in another language
  - Special case for languages transfer
    - Similar multi-lingual representation and cross-lingual alignment



NLP Tasks

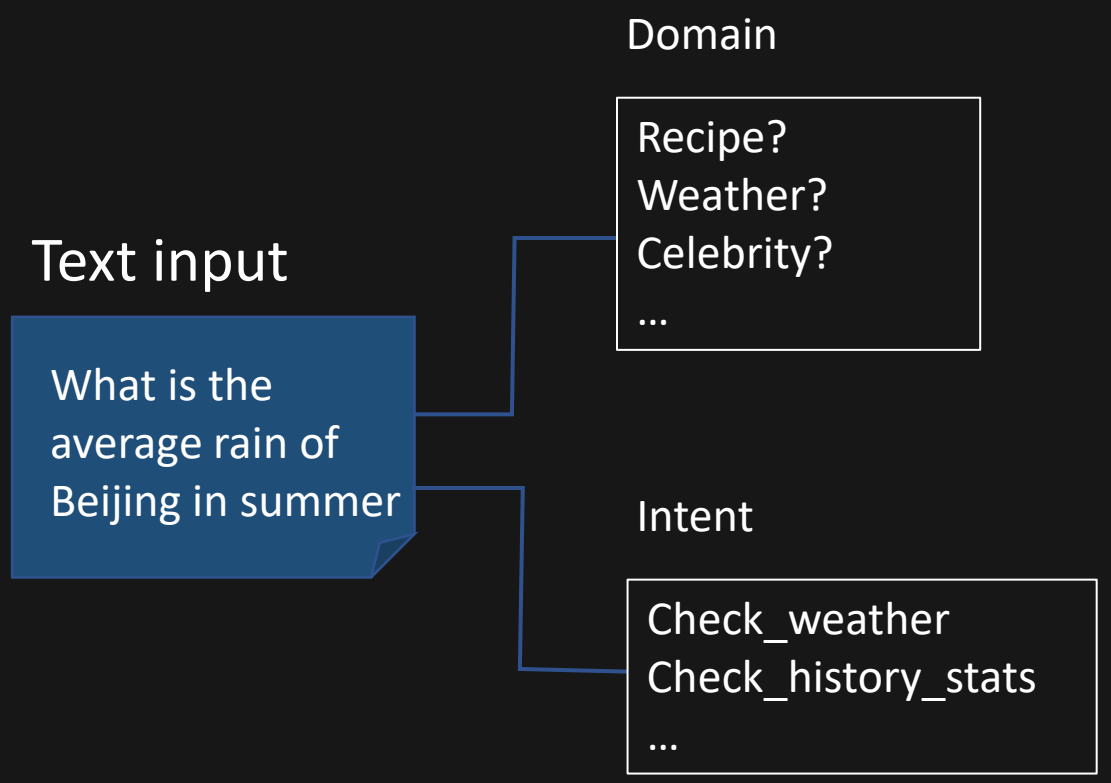
Type	Category	Sub Category	Example
NLU	Text Classification	Single text 	Domain identification, Intent detection, Sentiment classification
		Text pair 	Information retrieval, Natural language inference
	Sequence Labeling	Single text	Named entity recognition, Slot tagging
		Text pair 	Machine reading comprehension
	Structure Prediction	Single text 	Dependency parsing, constituency parsing, semantic role labeling
NLG	Text Generation	Token level 	Spelling correction, Sentence auto completion
		Sentence level	Machine translation, Conversation, Question generation

 **Will be presented in this tutorial**

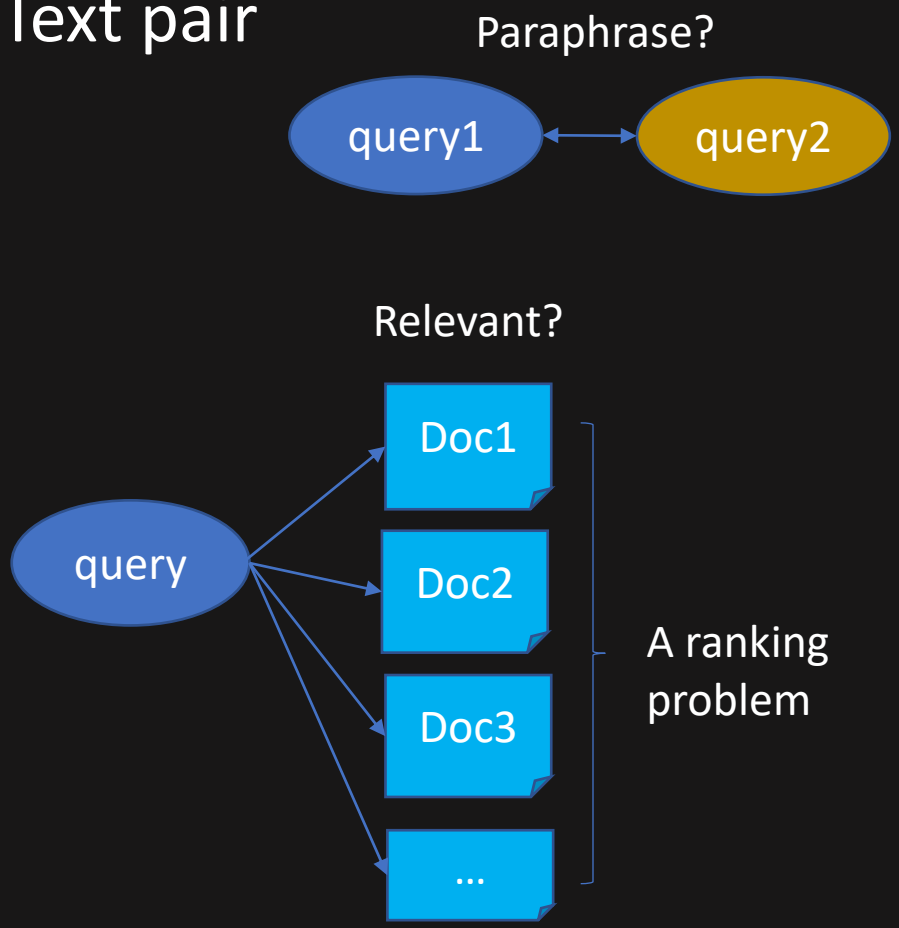
 **Has been presented in a related tutorial in The Web Conference 2021**

Text Classification

- Single text

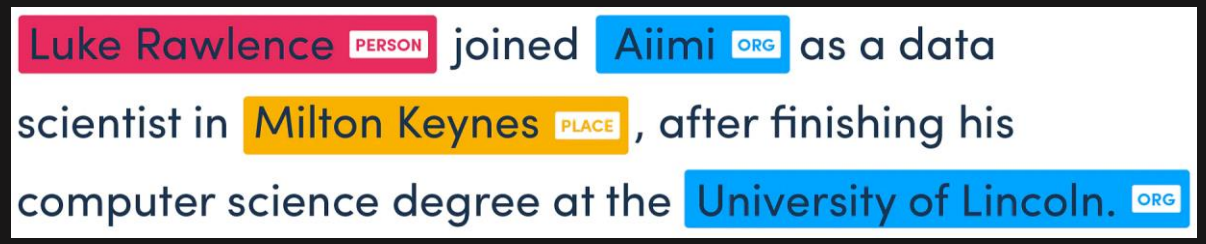


- Text pair

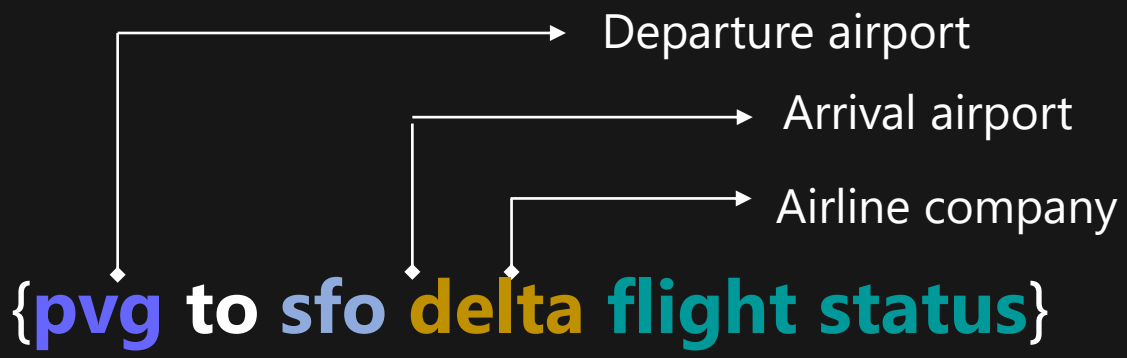


Sequence Labeling

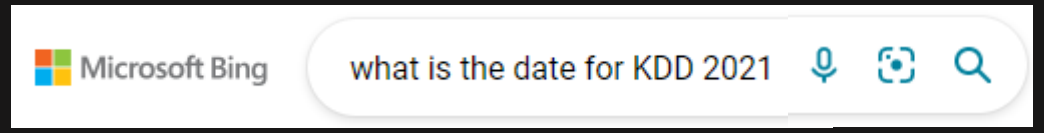
- Single text

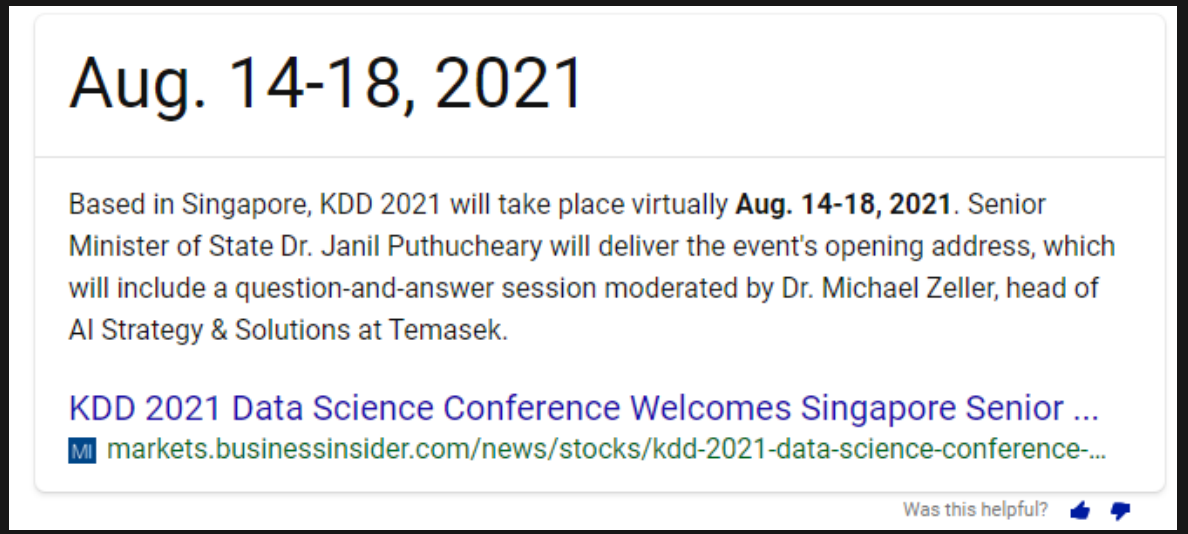


Example from https://www.aiimi.com/



- Text pair

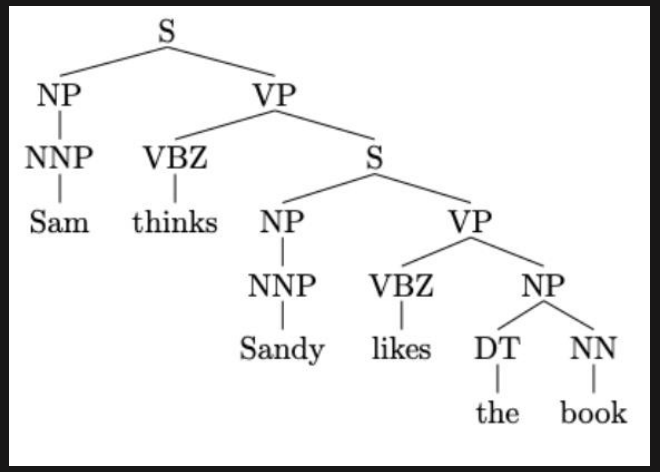




Structure Prediction

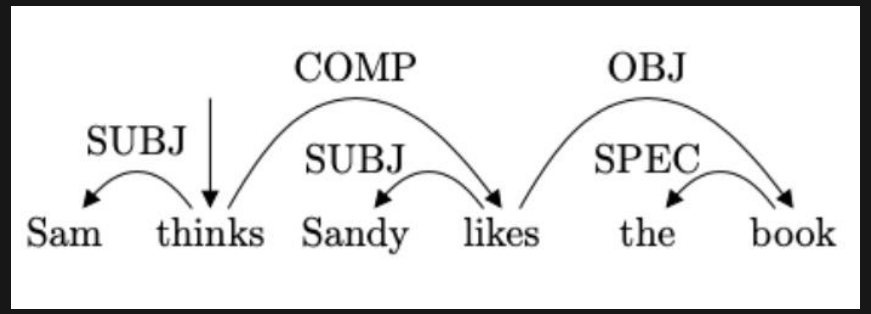
- Given a sentence, parse the syntactic structure of the sentence
  - e.g., constituency parsing, dependency parsing

Constituency parsing: how words group into nested constituents



The structure of a sentence is represented by a tree, which is based on the phrase structure grammar.

Dependency parsing: how words depend on each other



Each word has a directed edge towards its head, and there is a label to describe the dependency type.

Examples are from: https://www.cs.princeton.edu/courses/archive/fall19/cos484/lectures/lec10.pdf

Text Generation

Token level

Spelling Correction

Britnay spears vidios

↓

Britney spears videos

Sentence Auto Completion

The web conference

The web conference is a yearly international

The Web Conference is a yearly international academic conference on the topic of the future development

Sentence level

Machine Translation

The Web Conference is a yearly international academic conference on the topic of the future direction of the World Wide Web.

↓

La Conferencia Web es una conferencia académica internacional anual sobre el tema de la futura dirección de la World Wide Web.

Conversation

A: I've been hearing some strange noises around the house at night.

B: oh no! That's scary! What do you think it is?

A: I don't know, that's what's making me anxious.

Response: ???

*EMPATHETICDIALOGUES

* Rashkin, H. et al. Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset. ACL'19.

Challenges for Language Scaling

Lack of Training Data

- Examples of labeling items
  - Passage-QA: **millions** of QA pairs labeled for English
  - Web relevance: **millions** of query-document pairs labeled for English
- Unrealistic to label so much data for each language

- Language-specific features, data privacy issues, cold start problem are all related to this challenge
- Major focus on this tutorial

Too Many Models

- Suppose we have M applications and N languages, we need *O*(M*N) models
- In reality, M=*O*(**100**), N=*O*(**100**)
- The cost for model building, serving, and maintenance is huge

- Multi-task learning
- Adapters

Model Size and Latency

- Recent deep learning and pre-trained models are getting larger and larger
- TULRv2: 270M parameters
- GPT-3: 175B parameters
- Unrealistic to serve such models for online service

- Hardware acceleration
- Computation library optimization
- Model compression

Outline

Daxin Jiang

Software Technology Center at Asia (STCA) of Microsoft

- Introduction [Daxin Jiang]
  - Motivating examples in Microsoft products
  - Problem description
  - Categorization of applications
  - Challenges
- Methodology [Daxin Jiang]
  - Model Transfer
  - Data Transfer
- Applications*
  - Dependency Parsing [Xiubo Geng]
  - Machine Reading Comprehension [Ming Gong]
  - Grammar Error Correction [Linjun Shou]
- Summary & Future directions [Jian Pei]

Linjun Shou Xiubo Geng Ming Gong

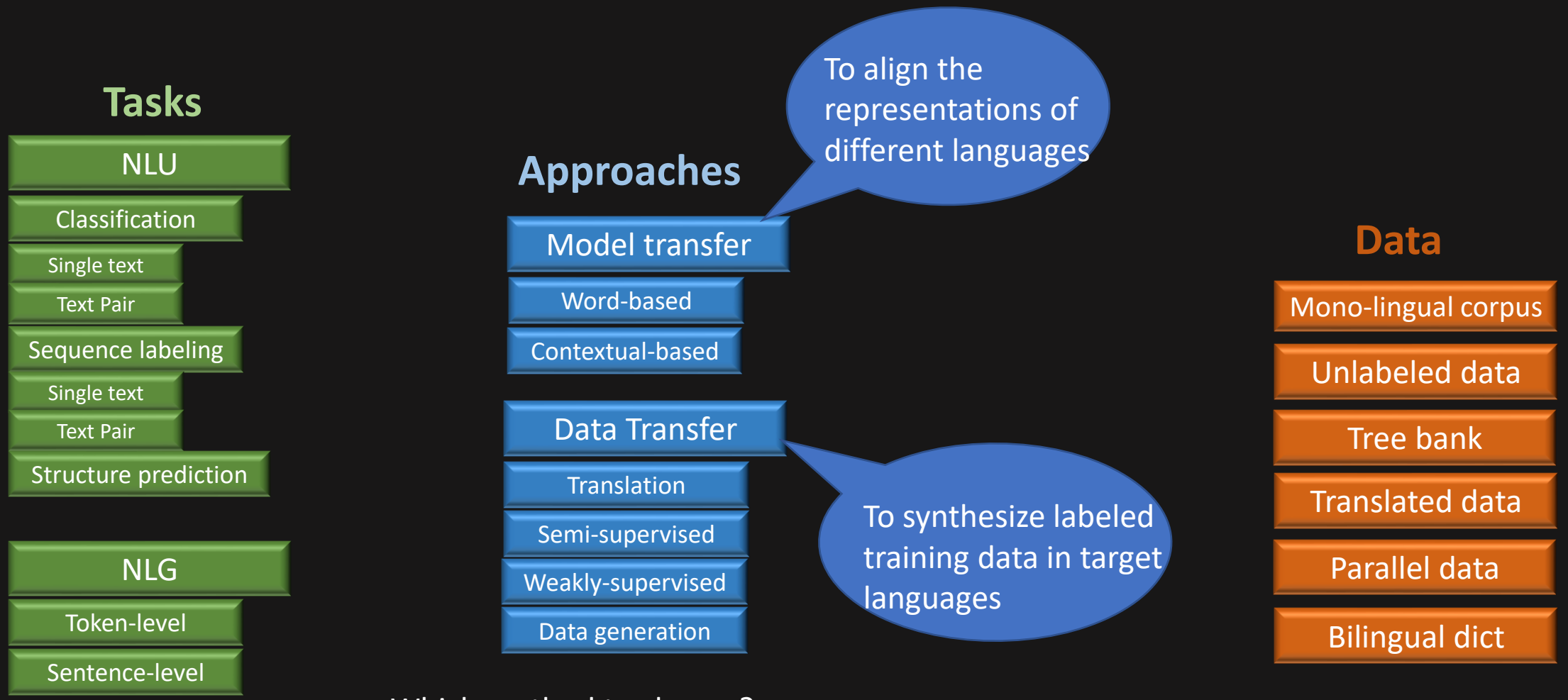
Software Technology Center at Asia (STCA) of Microsoft

Jian Pei

Simon Fraser University

*For more applications, please refer to our tutorial at The Web Conference 2021

Overview of Approaches



Which method to choose?

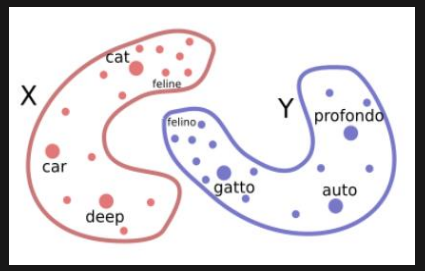
- In general, there is no single “winning” method
- Similar tasks, similar methods
- Method selection is often constrained by data availability

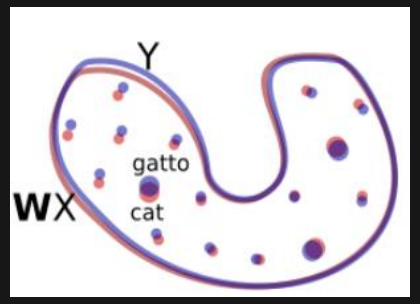
Approaches: Model Transfer and Data Transfer

(1) Model Transfer

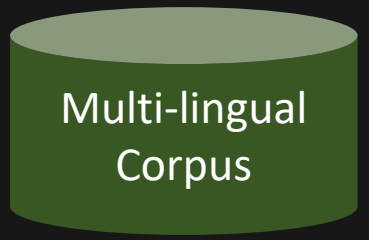
Zero-shot approach

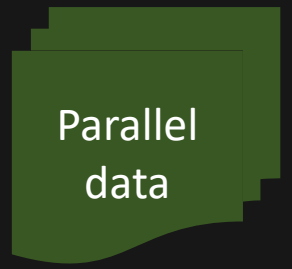
Cross-lingual model



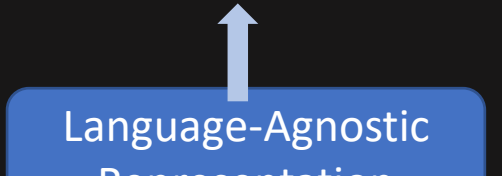


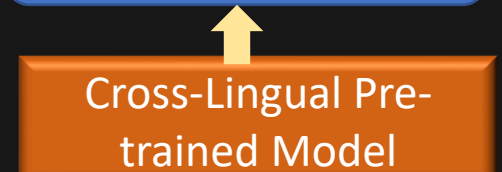
From MUSE





English training data





Test data in target languages

(1) Pre-train a cross-lingual model to align different languages.

(2) Use English training data to fine-tune the model

(3) Apply cross-lingual model and feed into fined-tuned model

Cross-Lingual Models

- Goal: **represent** different languages in a shared vector space, such that the texts with similar meaning are **aligned** close to each other, no matter in which languages they are expressed
- Cross-lingual *word* embedding
- Cross-lingual *contextual* embedding

Cross-Lingual Word Embedding

$$X^t \approx W X^s$$

- Mapping-based methods

$$J = \underbrace{\mathcal{L}(X^s) + \mathcal{L}(X^t)}_{1} + \underbrace{\Omega(X^s, X^t, W)}_{2}$$

- **Joint-learning methods**

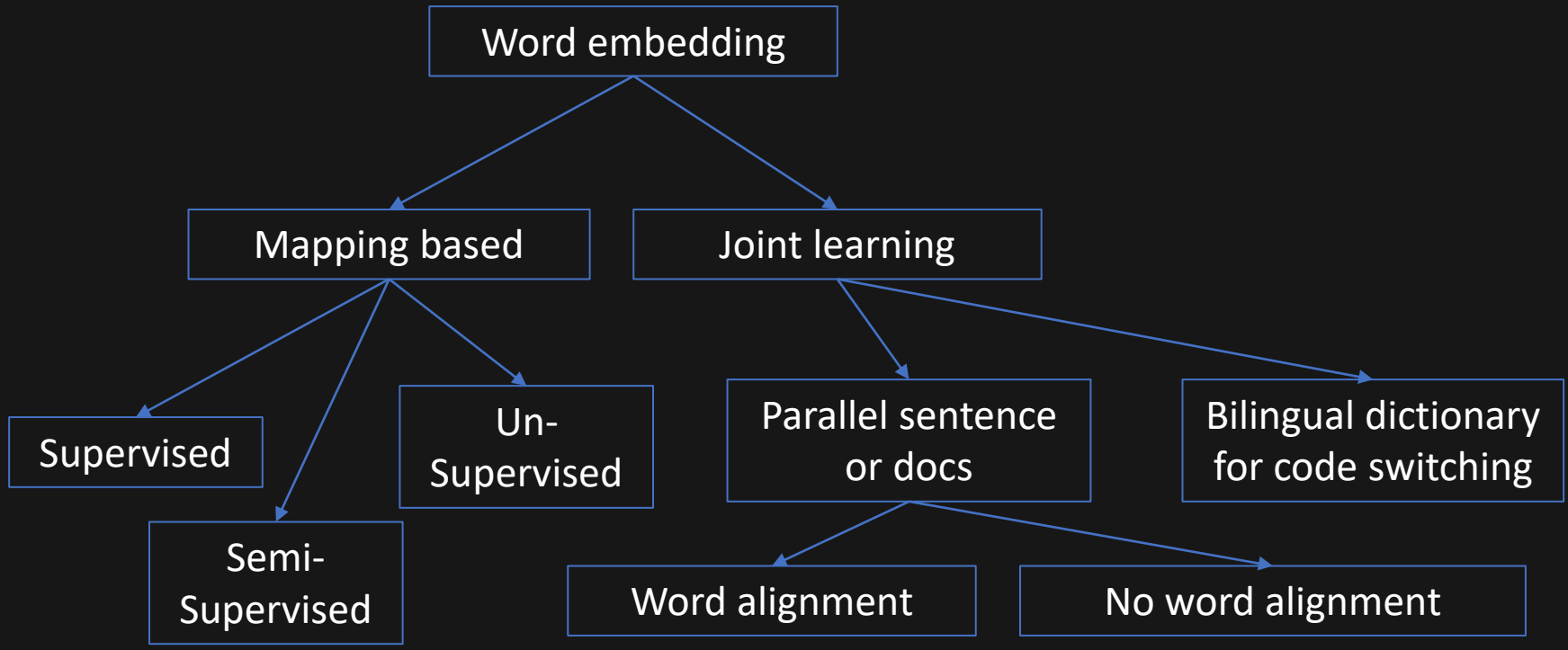
$$J = \mathcal{L}(X^s) + \mathcal{L}(X^t) + \Omega(X^s, X^t)$$

$\boldsymbol{X^s, X^t}$: word embedding matrix for source and target languages, respectively

$\boldsymbol{W}$: mapping matrix from X^s to X^t

$\boldsymbol{\mathcal{L}, \Omega, J}$: Loss functions for individual language embedding, cross-lingual mapping, and total loss, respectively

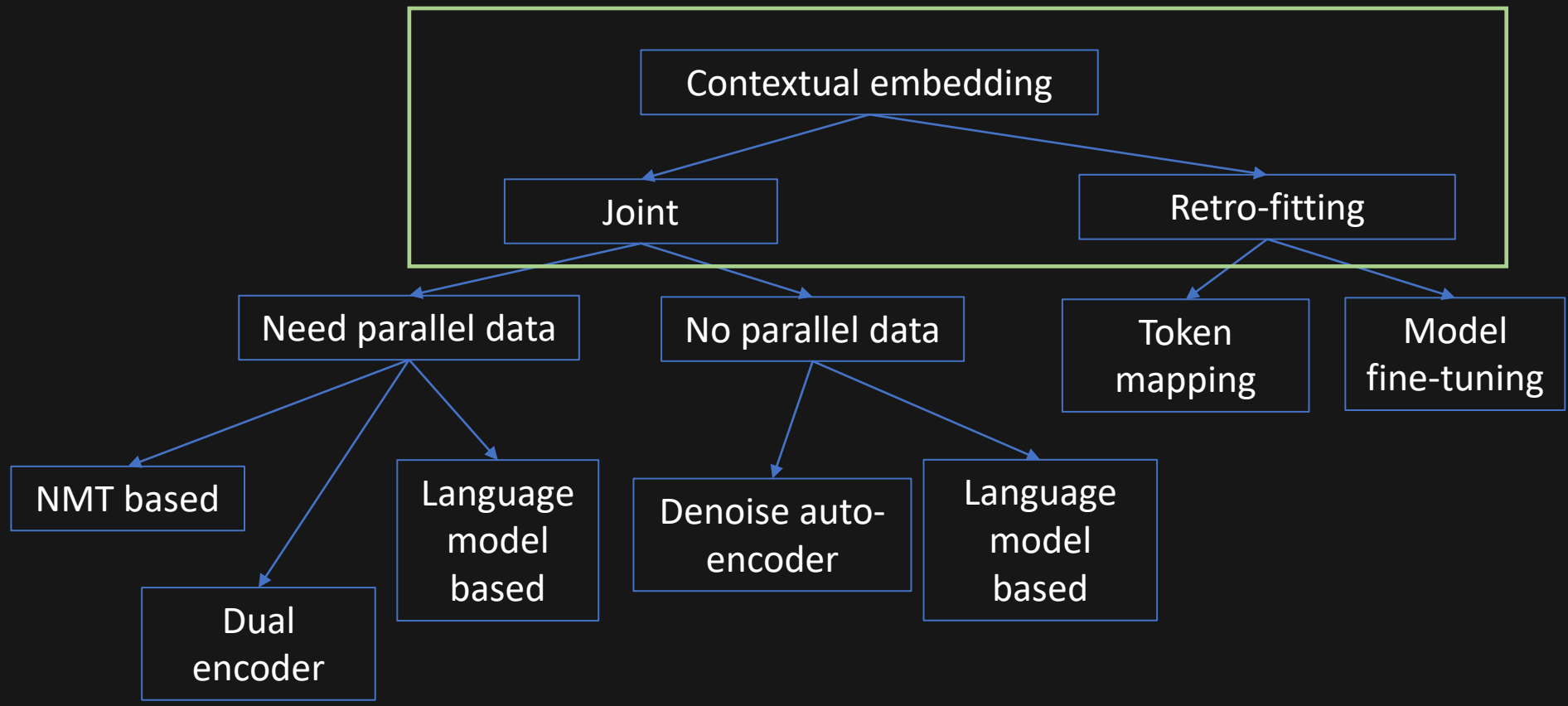
Cross-Lingual Word Embedding



Sebastian Ruder, Ivan Vulić, Anders SØgaard

- A survey of cross-lingual word embedding models,. Journal of Artificial Intelligence Research, May 2019.
- 2019 ACL Tutorial Unsupervised Cross-lingual Representation Learning (ruder.io)

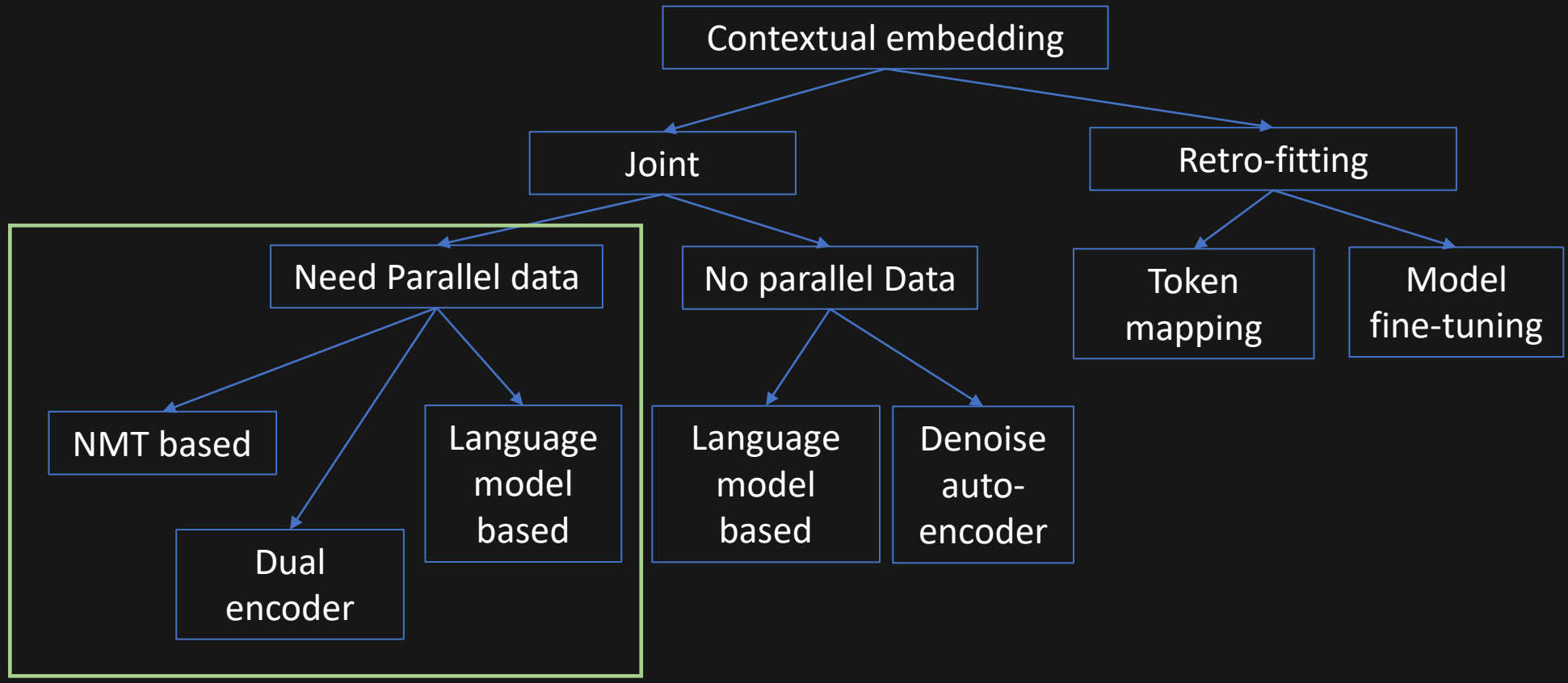
Cross-Lingual Contextual Embedding



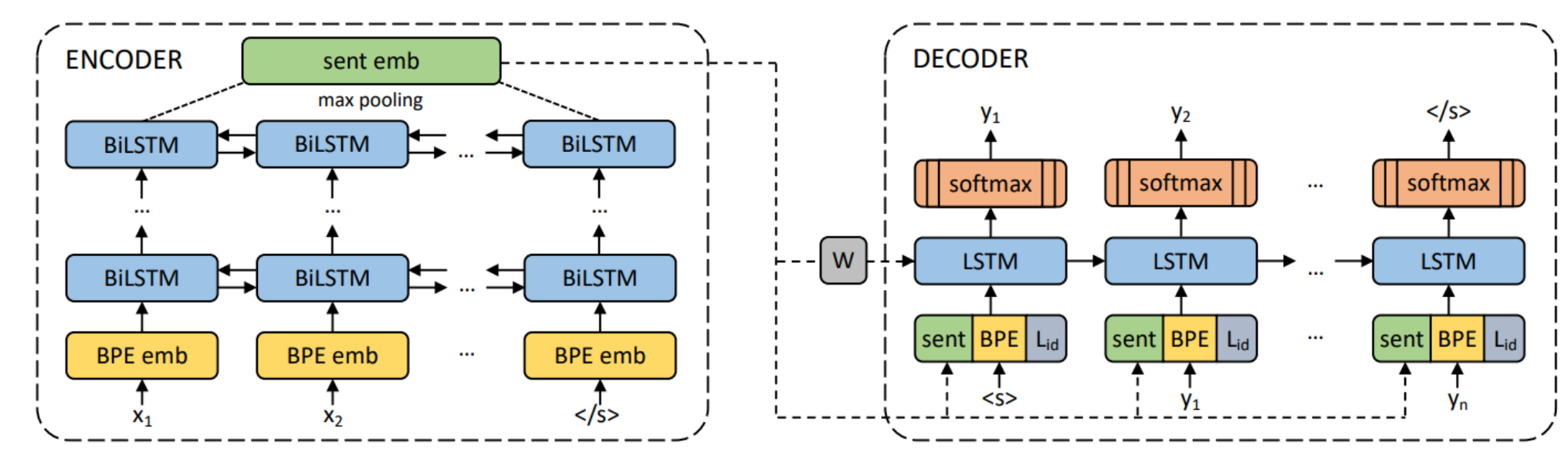
Joint or Retro-fitting?

- *Joint*: the alignment between languages happens simultaneously with the learning of contextual embedding
- *Retro-fitting*: the alignment happens after the contextual embedding for individual languages

Joint Learning with Parallel Data



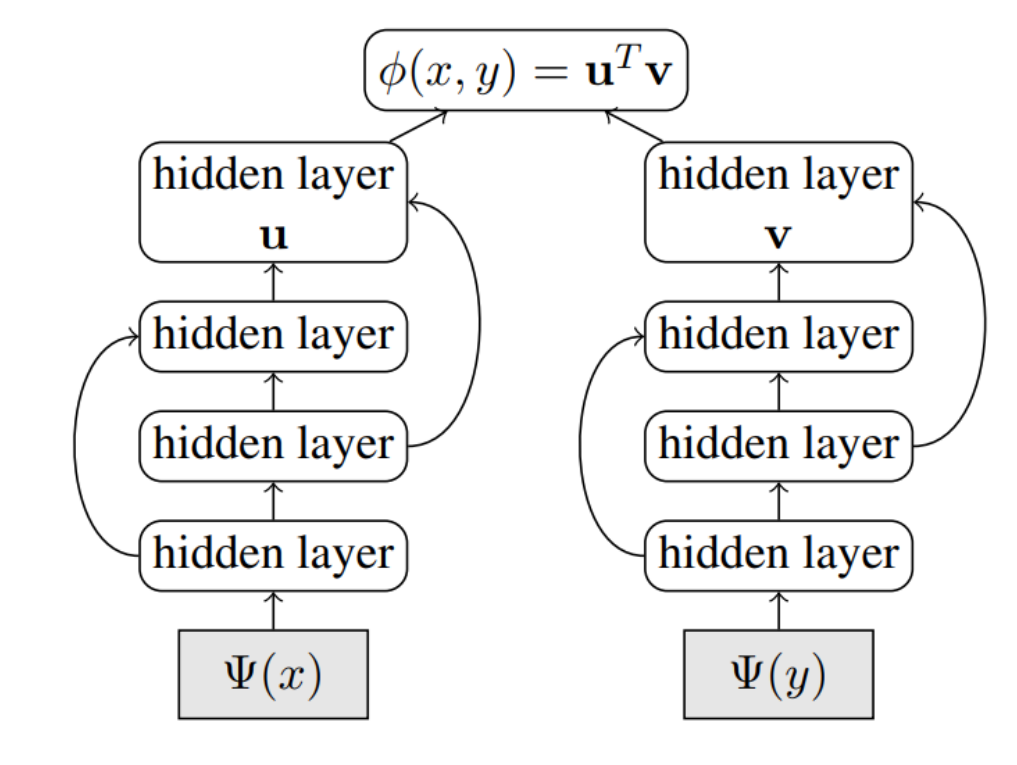
Neural Machine Translation (NMT) based Method



Mikel Artetxe, Holger Schwenk: Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. Trans. Assoc. Comput. Linguistics 7: 597-610 (2019)

- Encoder-decoder architecture for machine translation
- M-to-N translation: no indicator to distinguish the input languages => implicitly enforces different languages to have a uniform encoding

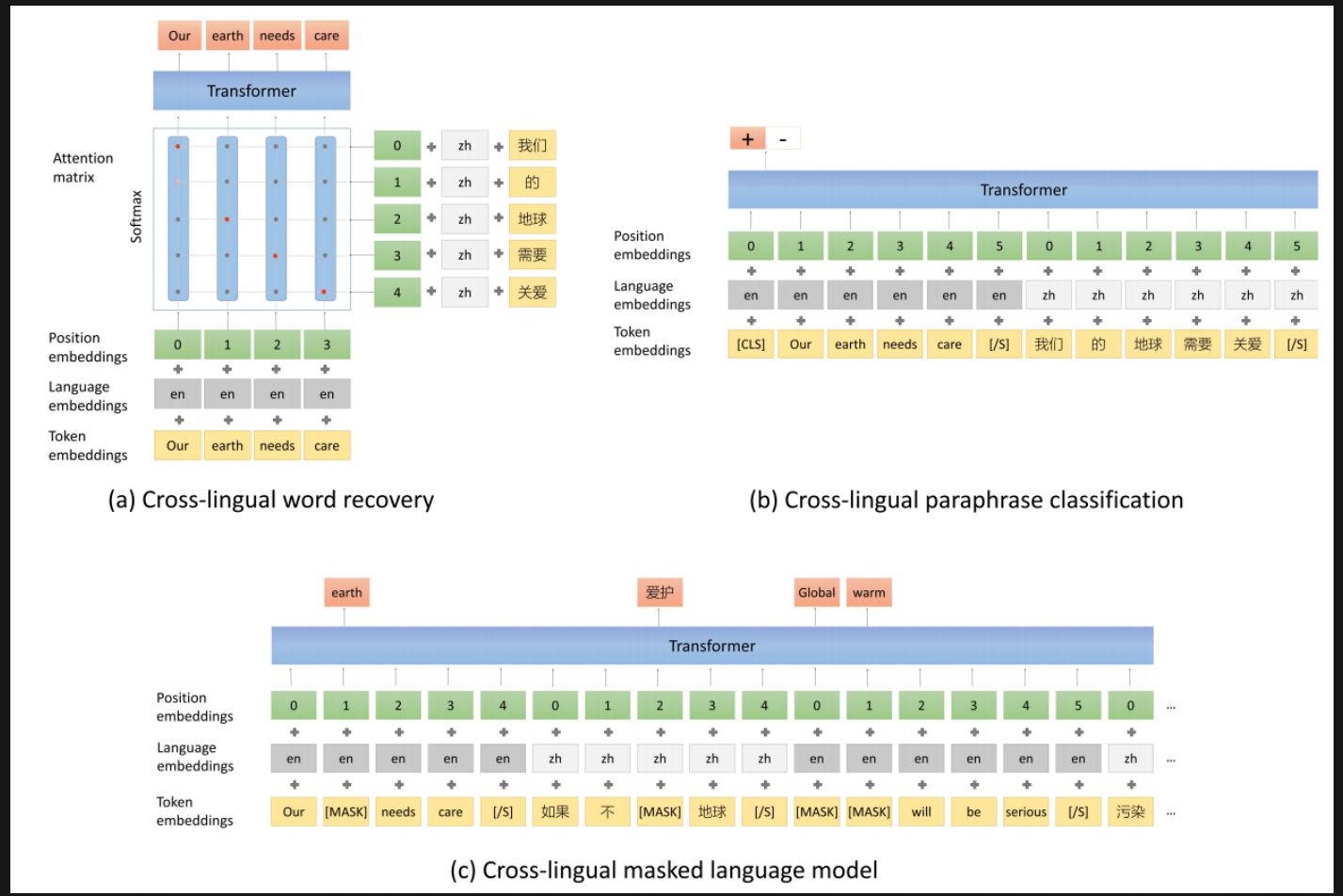
Dual Encoder



- *x* is in one language, and *y* is in another language
- Using parallel data to align the encoding *u* and *v*
- Dual encoder is explicit alignment, while NMT based approach is implicit alignment

Mandy Guo, Qinlan Shen, Yinfei Yang, et al.
Effective Parallel Corpus Mining using Bilingual Sentence Embeddings. WMT 2018: 165-176

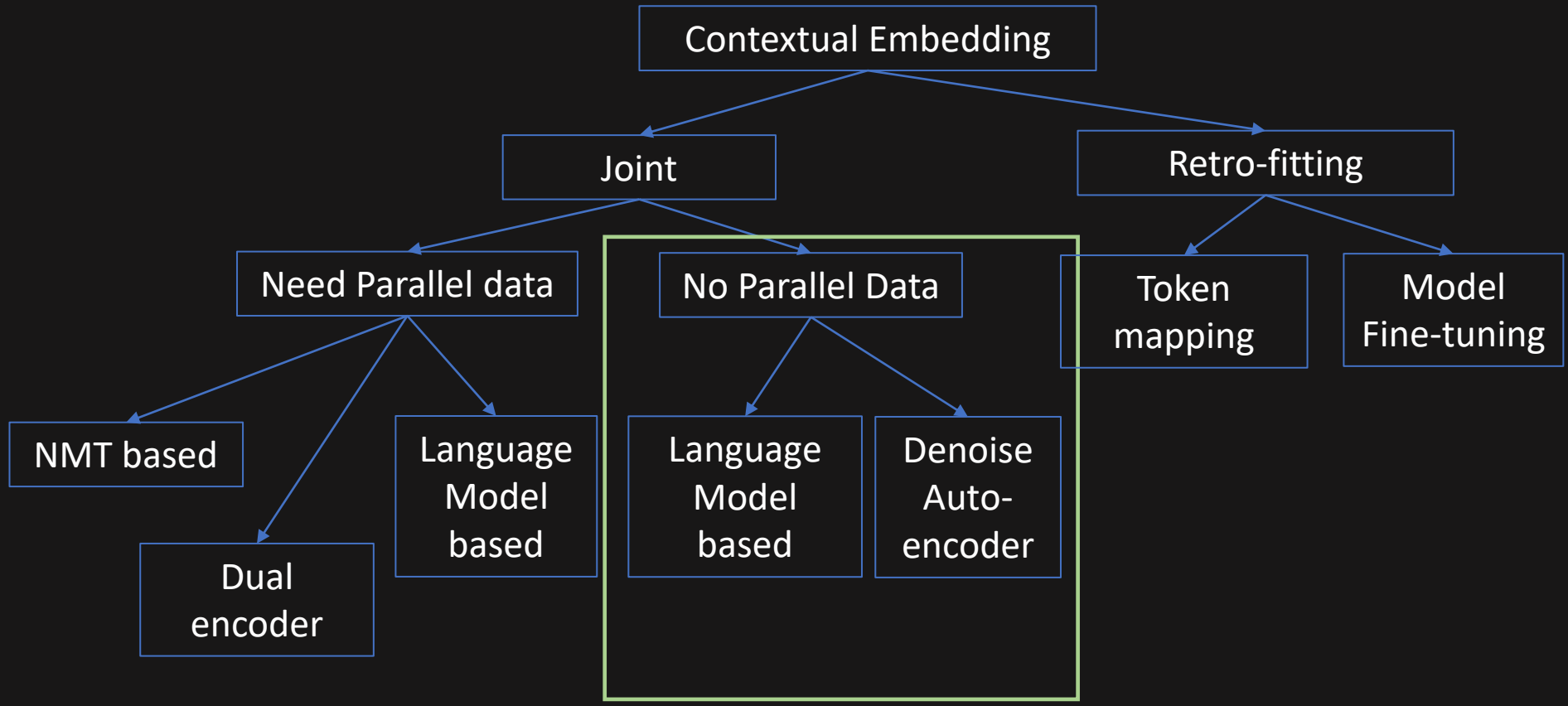
Language Model based Methods



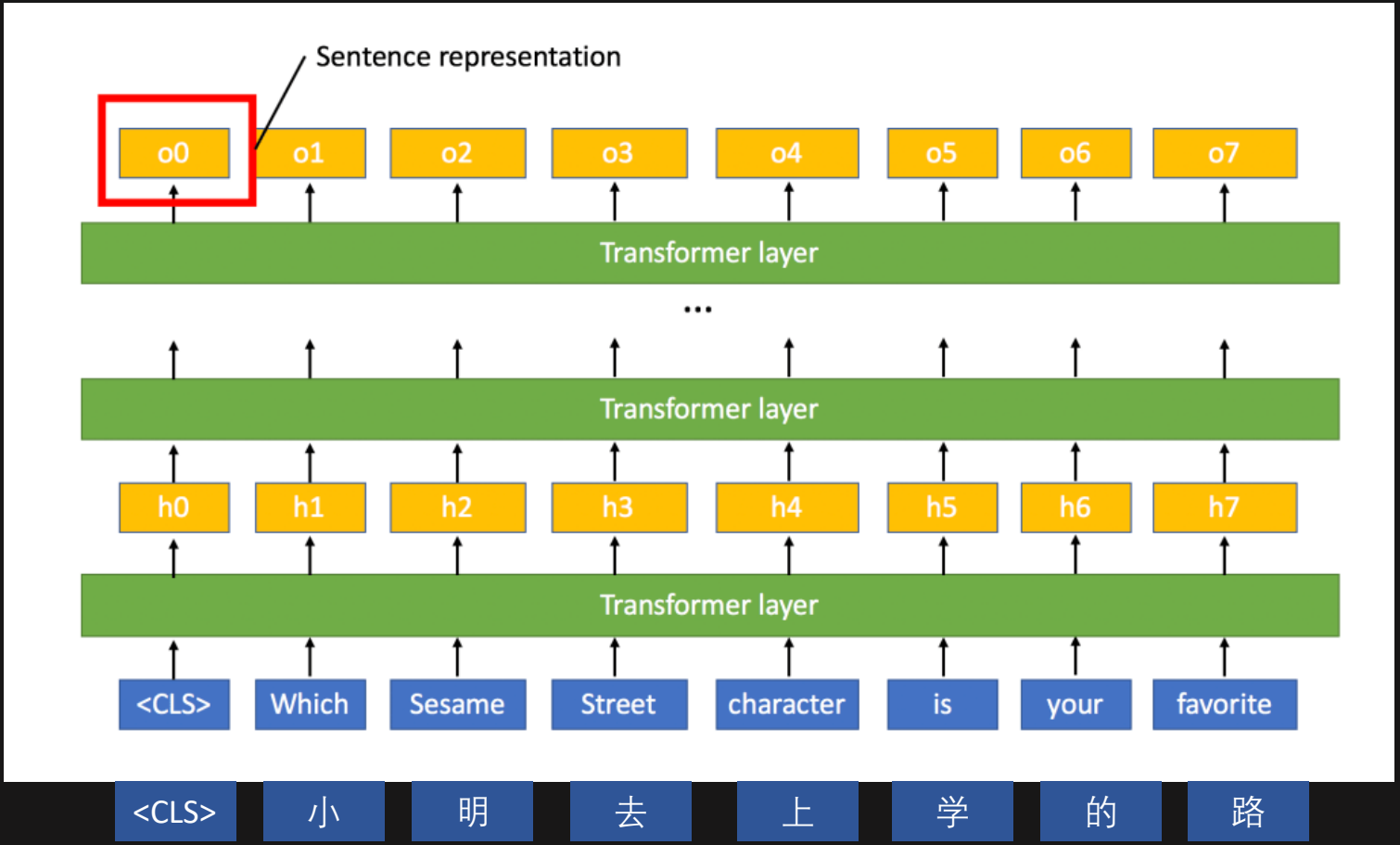
- Transformer based language model
- Cross-lingual representation at **both sentence level and token level**

Haoyang Huang, Yaobo Liang, Nan Duan et al. Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks. ***EMNLP 2019.***

Joint Learning without Parallel Data



Multi-Lingual BERT



- Share the same architecture and training scheme with monolingual BERT
- 104 languages Wikipedia pages
- 110k shared WordPiece vocabulary

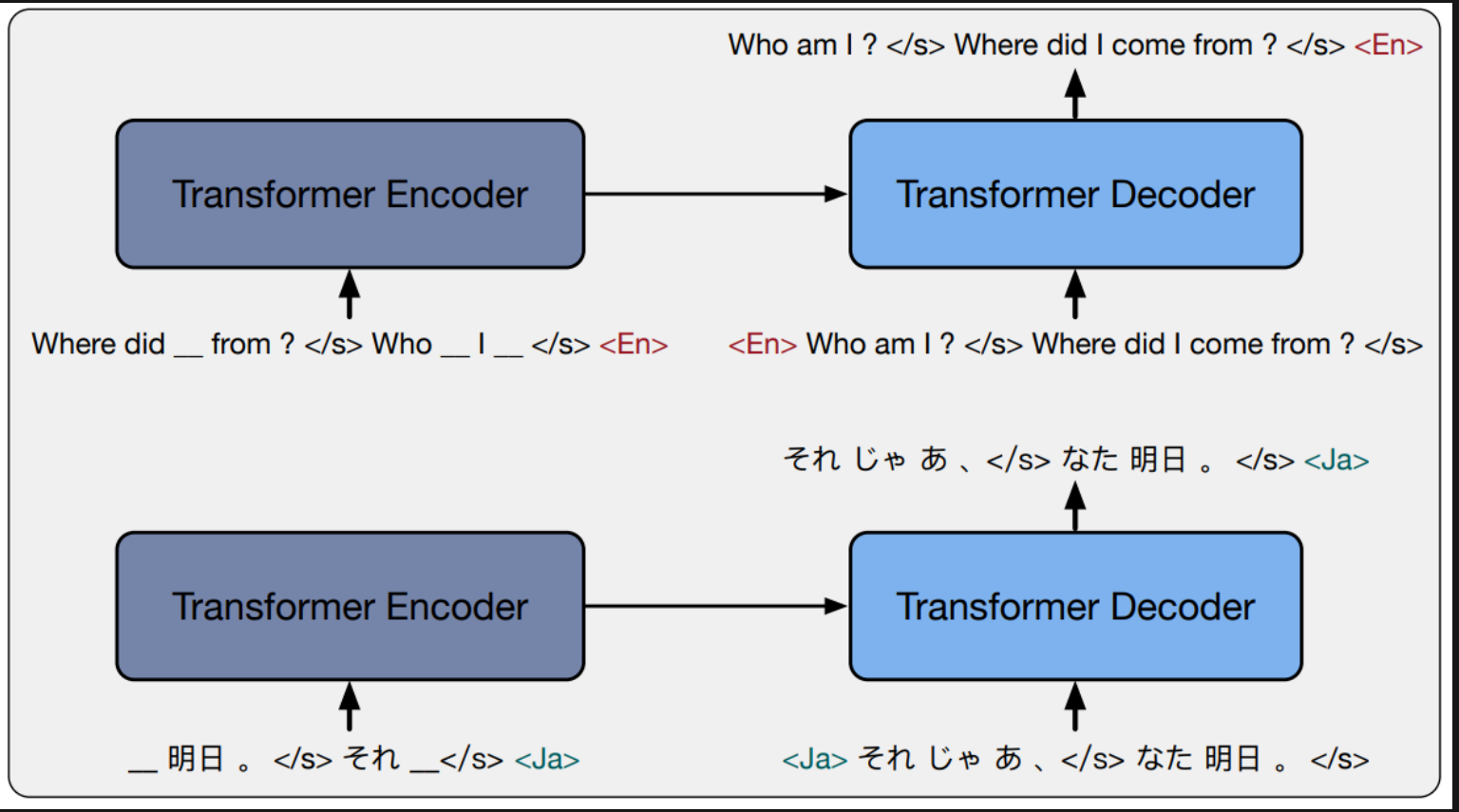
bert/multilingual.md at master · google-research/bert · GitHub

Why mBERT has Cross-Lingual Effectiveness

- Vocabulary overlap in different languages
  - Cognate words
  - Foreign words
  - However, it has been shown that mBERT still has cross-lingual effectiveness when there is zero vocabulary overlap
- Common structure in different languages
  - Word order
  - Most frequent words

- Pires T. et al. How multilingual is Multilingual BERT? ACL'19.
- Wu, S. and Dredze M. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT, EMNLP'19.
- Artetxe M. et al. On the Cross-lingual Transferability of Monolingual Representations, ACL'20.
- Conneau, A. et al. Emerging Cross-lingual Structure in Pretrained Language Models, ACL'20.
- Karthikeyan, K. et al. Cross-Lingual Ability of Multilingual BERT: An Empirical Study, ICLR'20.
- Libovický, J. et al. On the Language Neutrality of Pre-trained Multilingual Representations, EMNLP Findings'20.

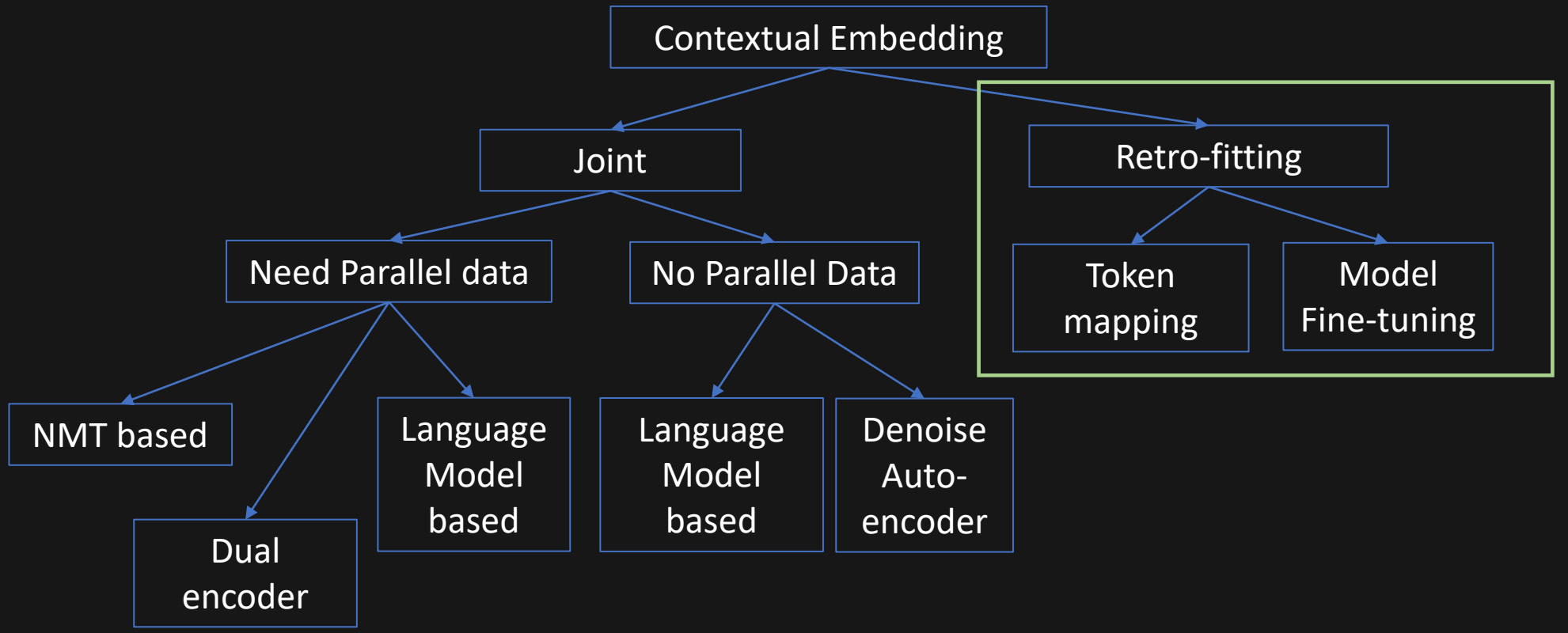
Multi-Lingual Bidirectional and Auto-Regressive Transformers (mBART)



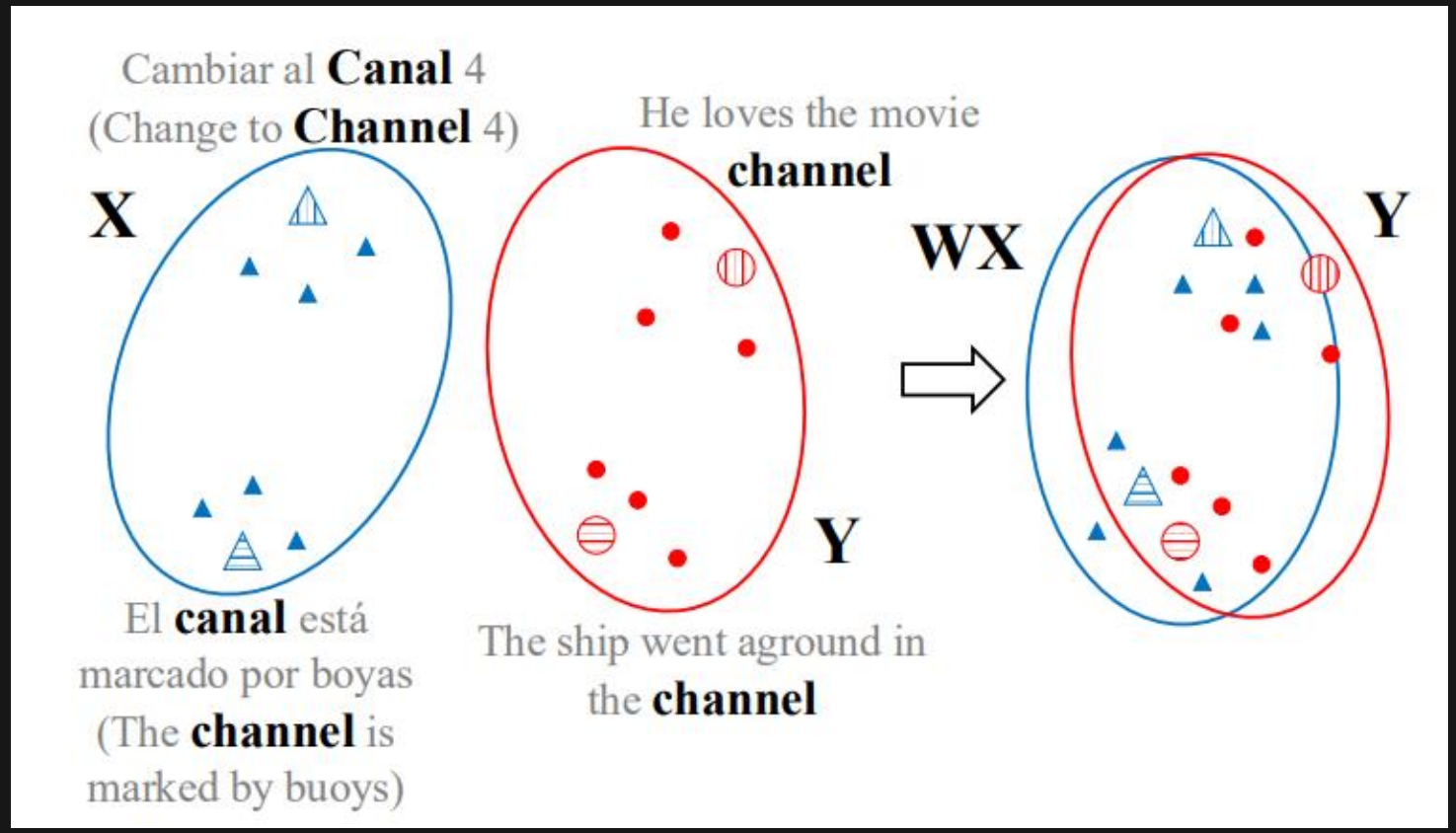
- Same architecture and training scheme with monolingual BART
- 25 languages crawled web pages
- 250K shared Sentence Piece vocabulary

Yinhan Liu, Jiatao Gu, Naman Goyal, et al.: Multilingual Denoising Pre-training for Neural Machine Translation. Trans. Assoc. Comput. Linguistics 8: 726-742 (2020)

Joint learning without Parallel Data



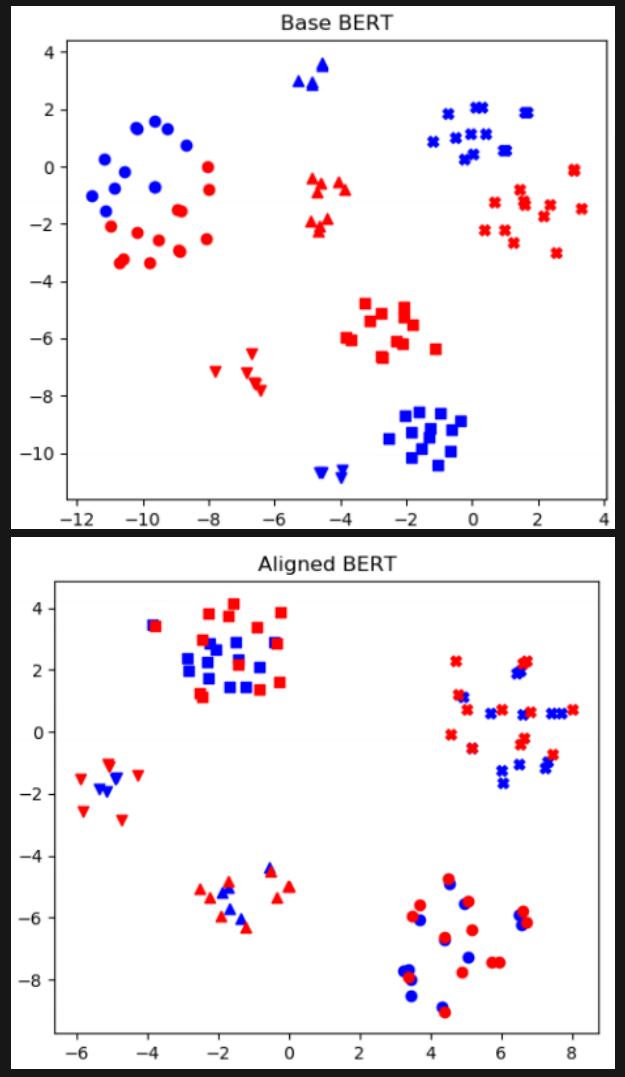
Contextual Token Mapping



- Learn the mapping matrix W
  $\min_W \sum_{i=1}^{n} \|Wx_i - y_i\|^2$
  where $W^\intercal W = I$
- Instead of using static words in dictionary, use parallel sentences to align the word occurrences in sentences

Yuxuan Wang, Wanxiang Che, Jiang Guo, et al. Cross-Lingual BERT Transformation for Zero-Shot Dependency Parsing, EMNLP'19.

Model Fine-Tuning



- Upper figure: mBERT roughly aligns the words
- Lower figure: effect of fine-tuning
- After mBERT training, fine-tuning by

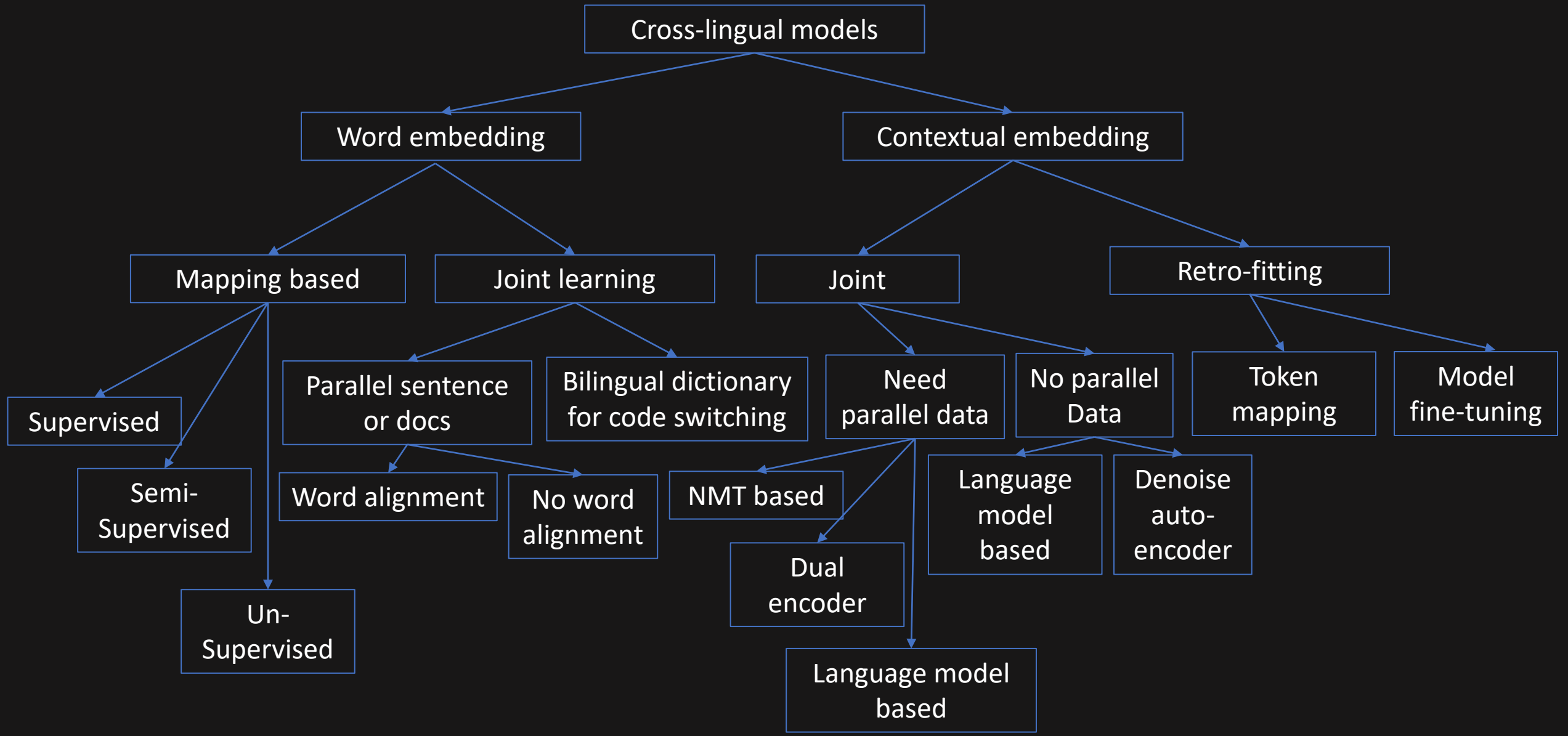
$$\mathcal{L}(f, C) = -\sum_{(s,t)\in C}\sum_{(i,j)\in a(s,t)} \|f(i,s) - f(j,t)\|^2$$

(s,t) are parallel sentences, (i,j) are aligned words, f() is the encoding function

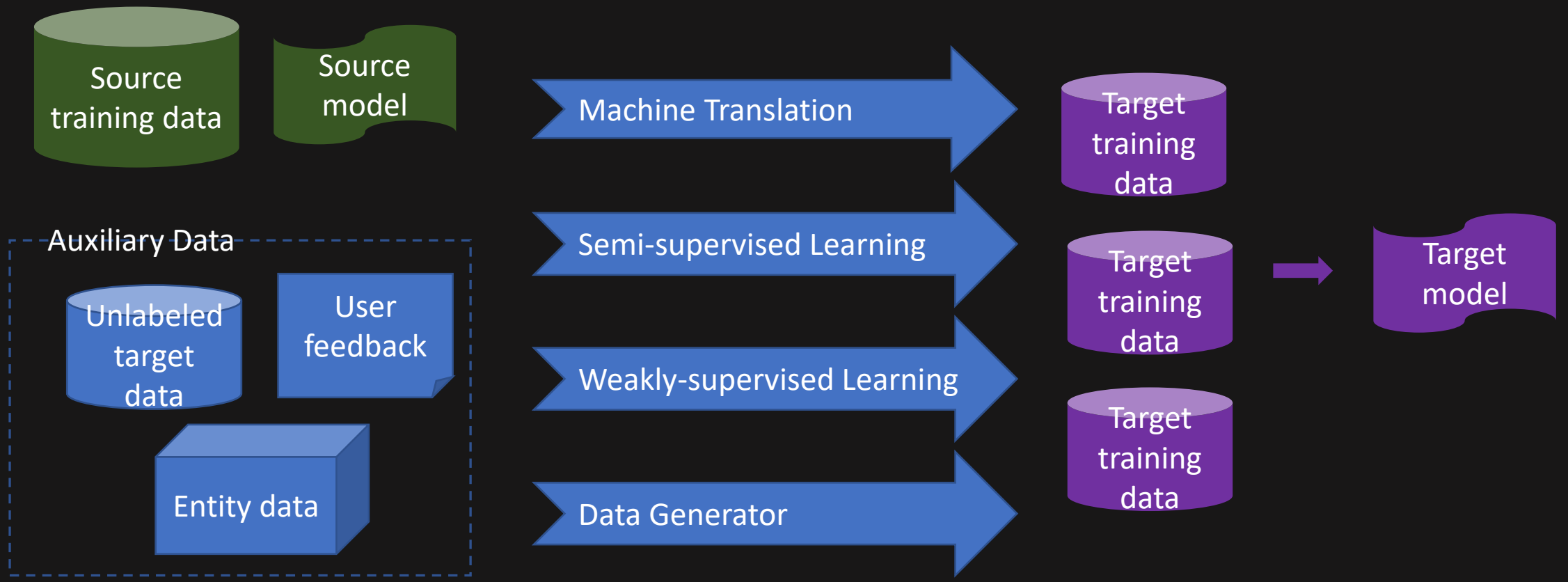
$$R(f, C) = \sum_{t\in C}\sum_{i=1}^{\text{len}(t)} \|f(i,t) - f_0(i,t)\|^2$$

To avoid trivial alignment (e.g., f is constant), add regularization

Steven Cao, Nikita Kitaev, Dan Klein, Multilingual Alignment of Contextual Word Representations, ICLR'20.

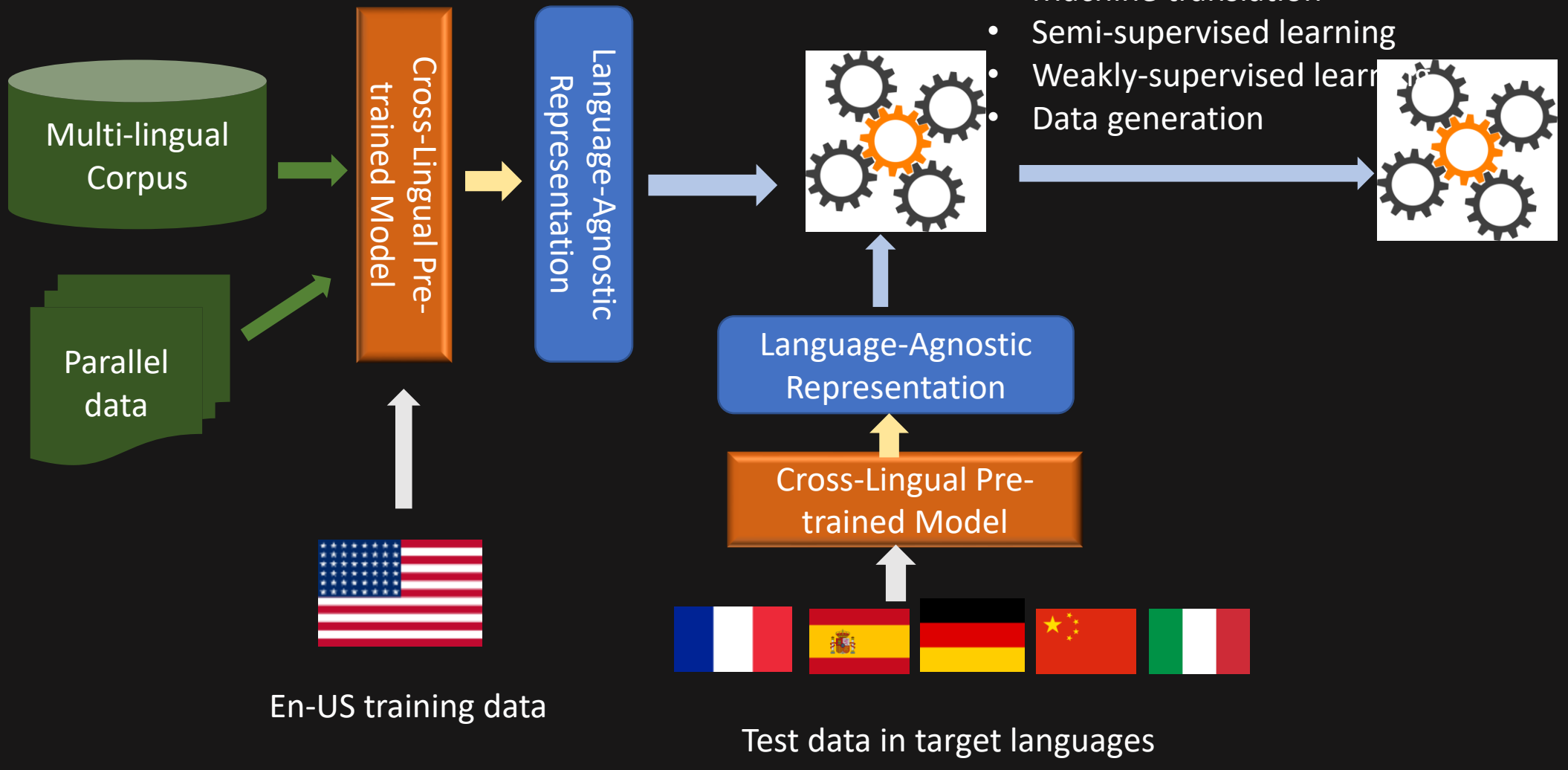

Cross-lingual models
Word embedding
Contextual embedding
Mapping based
Joint learning
Joint
Retro-fitting
Supervised
Semi-Supervised
Un-Supervised
Parallel sentence or docs
Bilingual dictionary for code switching
Word alignment
No word alignment
Need parallel data
No parallel Data
Token mapping
Model fine-tuning
NMT based
Dual encoder
Language model based
Language model based
Denoise auto-encoder

Approaches: Model Transfer and Data Transfer (2) Data Transfer



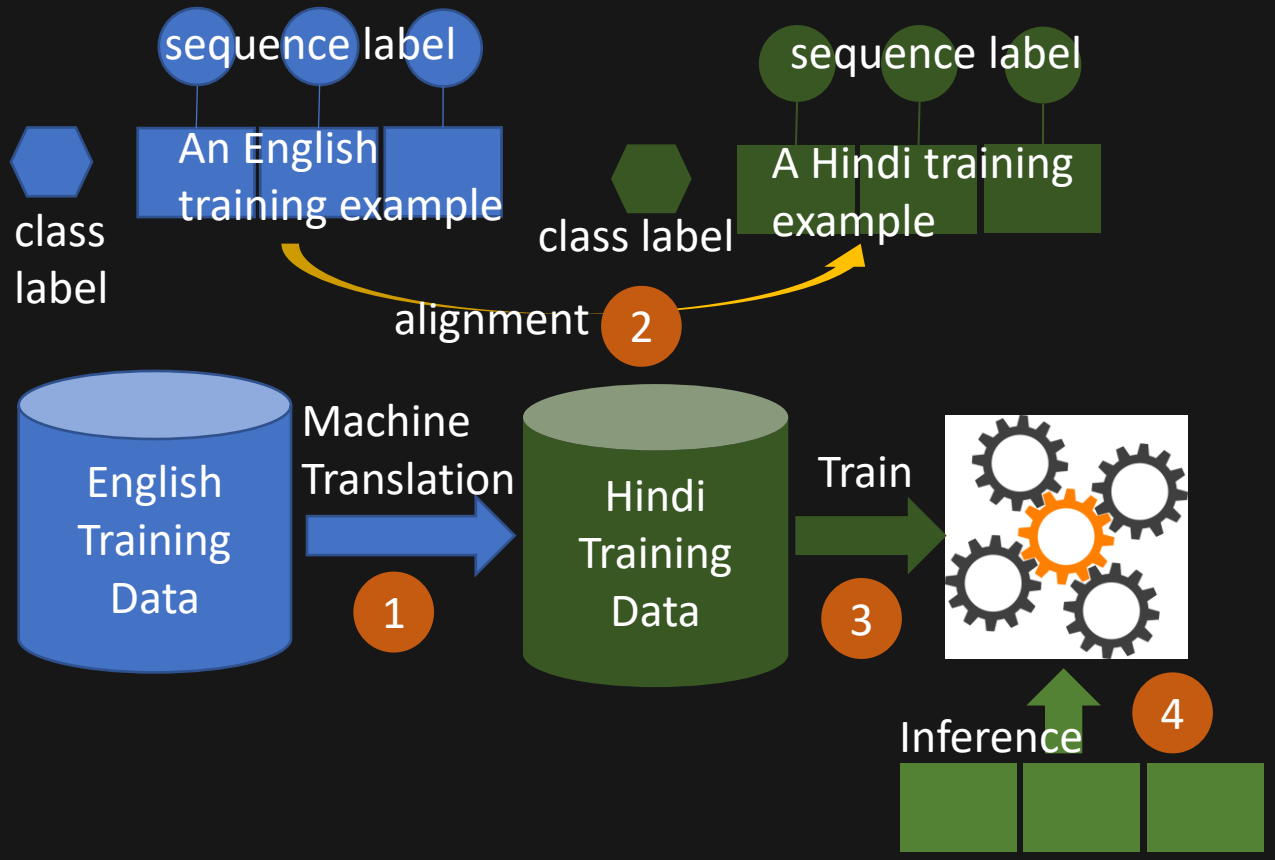
Approaches: Model Transfer and Data Transfer

(1) Model Transfer + (2) Data Transfer

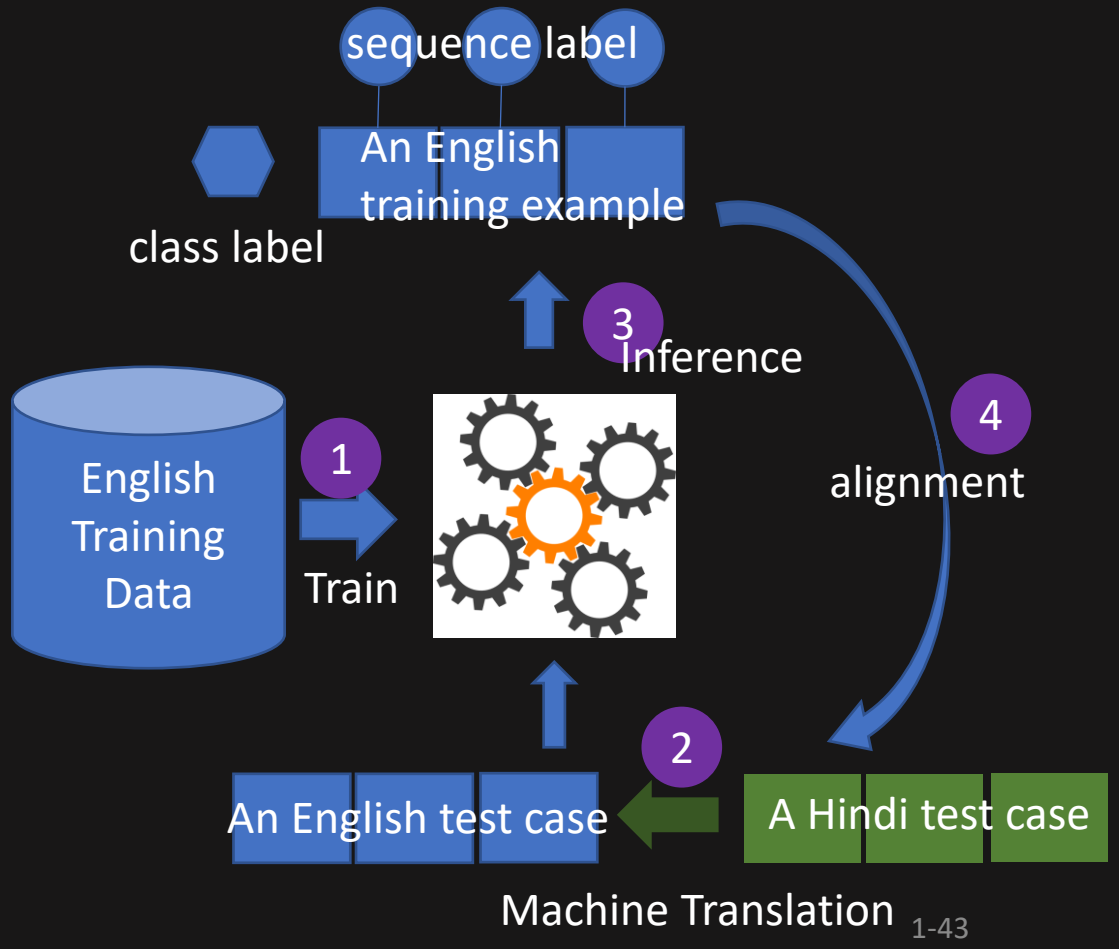


Machine Translation

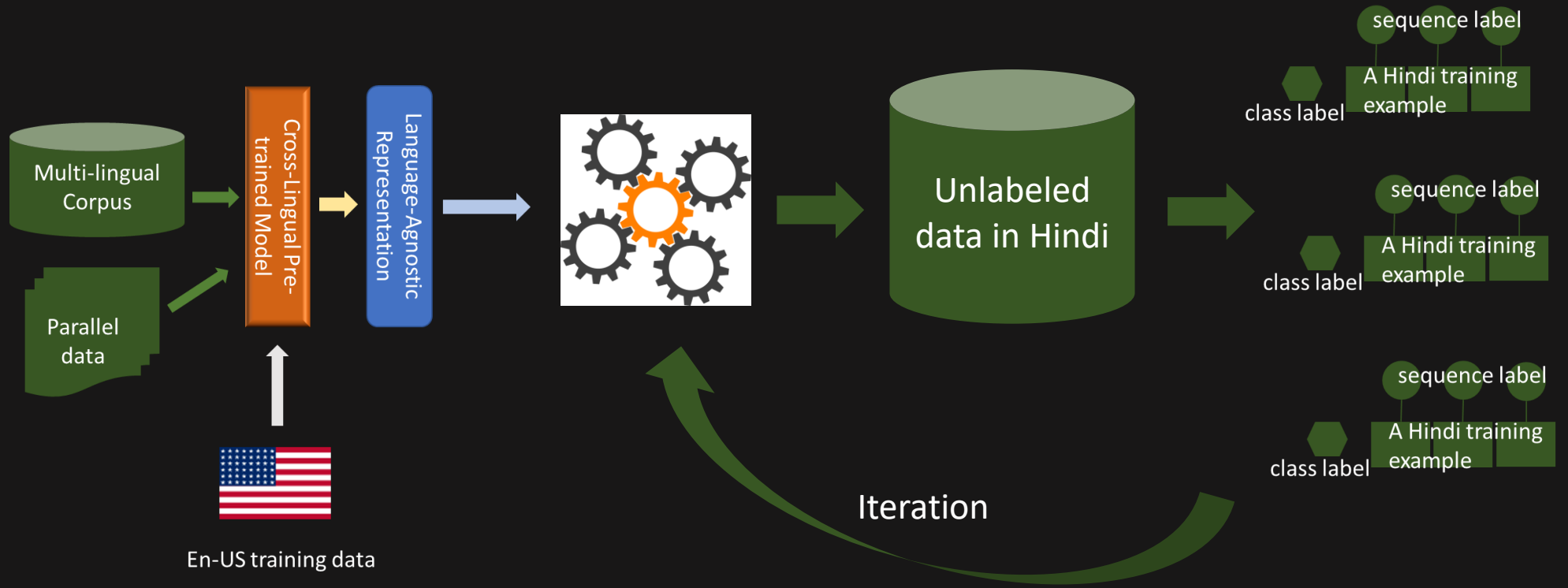
- Translate-on-train



- Translate-on-test



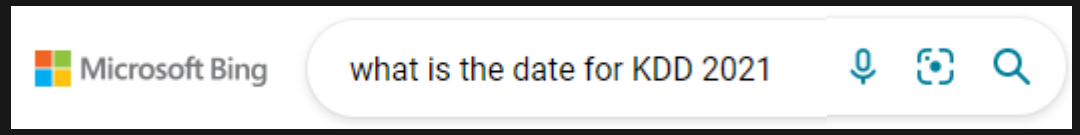
Semi-Supervised Learning

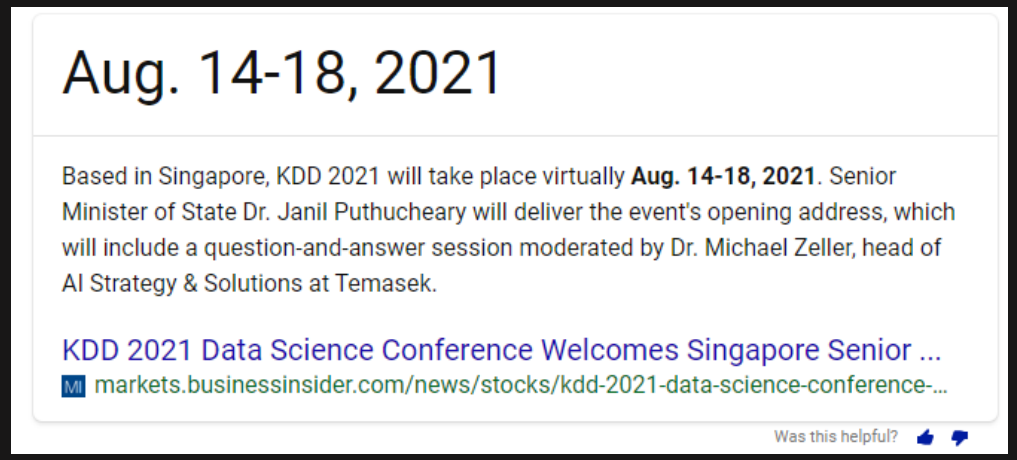


1. Build an initial model by model transfer approach
2. Apply the initial model on unlabeled data to get labels
3. Use the labeled data to improve the initial model
4. Iterate the above steps 2-3 for several rounds (self learning, model ensemble, reinforcement learning)

Weakly-Supervised Learning

- Collect auxiliary data
  - E.g., user feedback, knowledge data





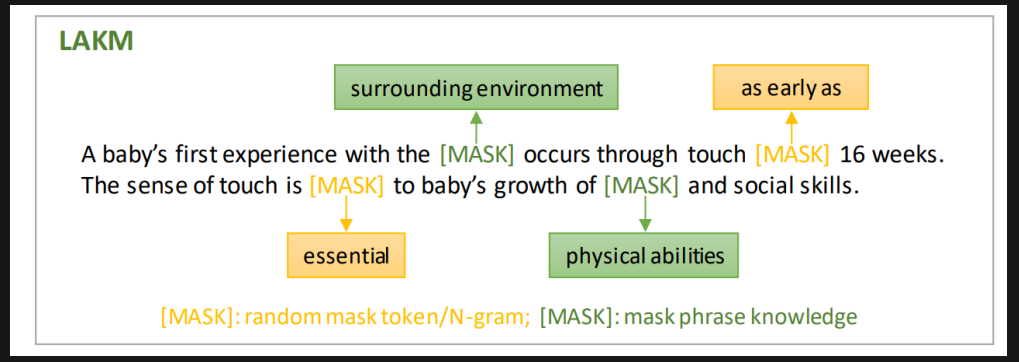
- Rich user behaviors in search logs
- Any single behavior is a weak signal, but more powerful when combined in a ML model
- Generally applicable to all languages

Shou, L. et al. Mining Implicit Relevance Feedback from User Behavior for Web Question Answering. KDD'20.

[Question]: who were the kings of the southern kingdom
[Passage]: In the southern kingdom there was only one dynasty, that of king David, except usurper Athaliah from the northern kingdom, who by marriage, []
[Answer - ground truth]: king David
[Answer - model predication:] David, except usurper Athaliah

[Question]: What is the suggested initial does dosage of chlordiazepoxide
[Passage]: If the drug is administered orally, the suggested initial dose is 50 to 100 mg, to be followed by repeated doses as needed until agitation is controlled up to 300 mg per day. []
[Answer - ground truth]: 50 to 100 mg
[Answer - model predication:] 100 mg

Examples of boundary detection error



Mask entities and phrases in sentences and let the model to **recover** them as a pre-training task

Yuan F. et al. **Enhancing Answer Boundary Detection for Multilingual Machine Reading Comprehension**. ACL 2020.

Data Generation

Natural Language Understanding (NLU)

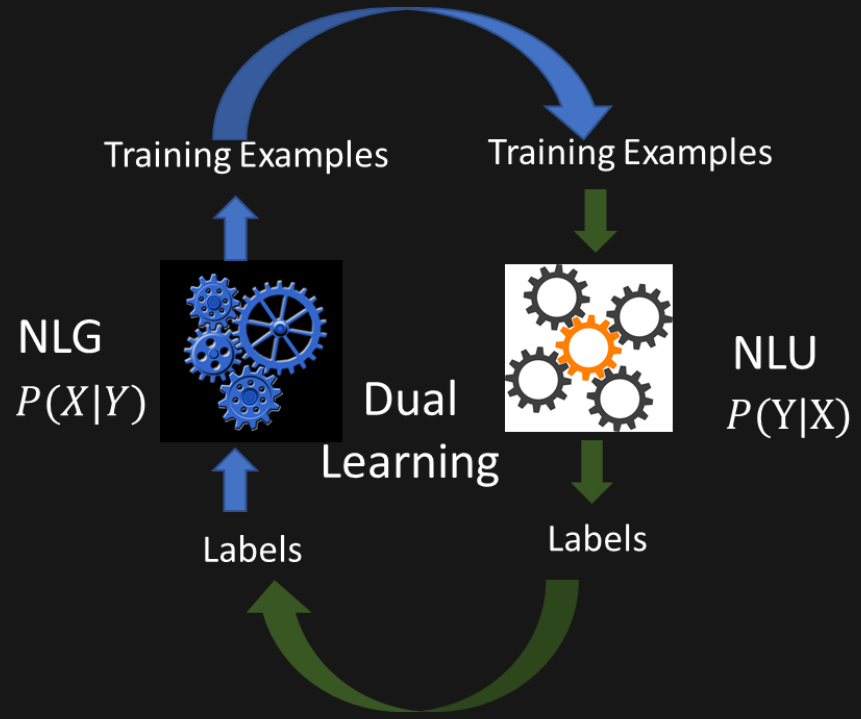
Intent: check_status

Departure airport

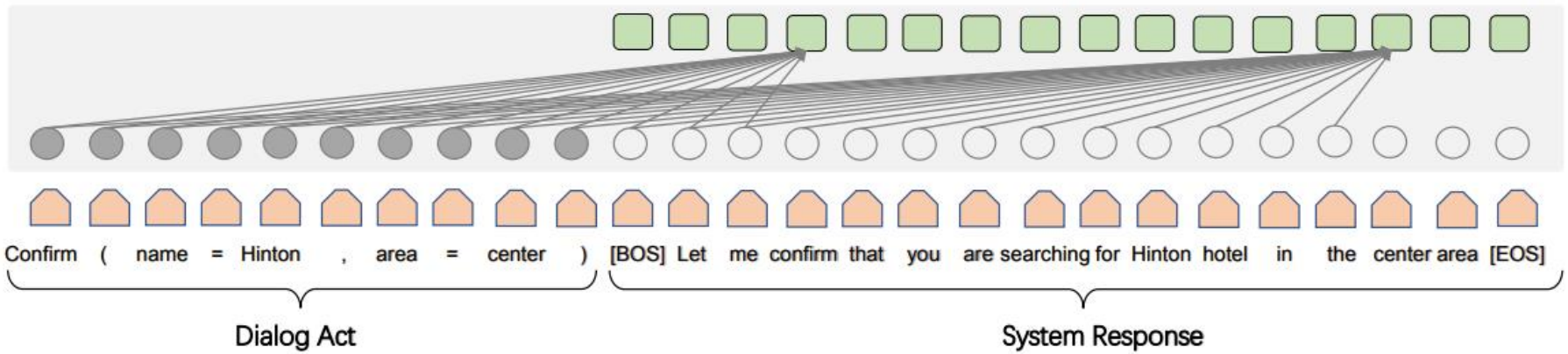
Arrival airport

Airline company

{pvg to sfo delta flight status}



Natural Language Generation (NLG)



Peng, B. et a. Few-shot Natural Language Generation for Task-Oriented Dialog. EMNLP Finding'20.

Zhu, S. et al. Dual Learning for Semi-Supervised Natural Language Understanding. IEEE/ACM Trans. On Audio, Speech, And Language Processing. 2020

Summary: Data Transfer Methods

The central idea of data transfer is to produce pseudo training data (X,Y) in the target language, where X is the training instance, and Y is the label

Methods	Data instances (X)	How to derive labels (Y)
Machine translation	Translated	Alignment
Semi-supervised learning	Real data	Self-learning with iteration
Weakly-supervised learning	Real data	Implicitly/partially derived from auxiliary data
Data generation	Synthesized	Generated

Outline

- Introduction [Daxin Jiang]
  - Motivating examples in Microsoft products
  - Problem description
  - Categorization of applications
  - Challenges
- Methodology [Daxin Jiang]
  - Model Transfer
  - Data Transfer
- Applications*
  - Dependency Parsing [Xiubo Geng]
  - Machine Reading Comprehension [Ming Gong]
  - Grammar Error Correction [Linjun Shou]
- Summary & Future directions [Jian Pei]

*For more applications, please refer to our tutorial at The Web Conference 2021

Daxin Jiang
Software Technology Center at Asia (STCA) of Microsoft

Linjun Shou Xiubo Geng Ming Gong
Software Technology Center at Asia (STCA) of Microsoft

Jian Pei
Simon Fraser University

Microsoft

Cross Lingual Dependency Parsing

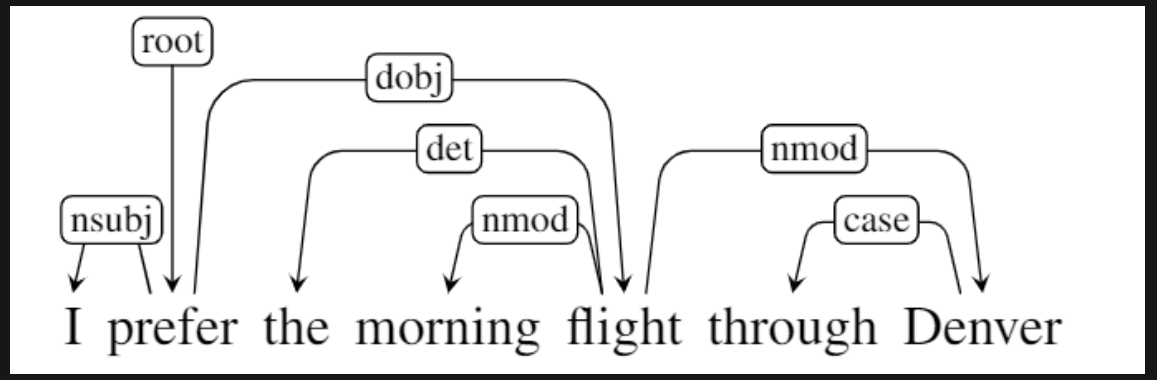
Xiubo Geng
xigeng@microsoft.com

Cross-lingual Dependency Parsing

- Background
- Model Transfer
  - Delexicalized model and improvements
  - Multilingual pre-trained models
- Data Transfer
  - Translation-Based Approach
    - Treebank Translation (source training data + translator)
    - Annotation Projection (parallel data + source model)
  - Unsupervised Approach
    - Source reordering
    - Direct unsupervised adaption

Dependency Parsing

- Given a sentence, parse it into a dependency tree
- A dependency tree is a directed graph G = (V, A), where
  - The nodes in V are words $w_1, \ldots, w_n$ of a sentence
  - The arcs in A have the form (w_i, l, w_j), where l is label



Dependency Parsing

Type	Category	Sub Category	Example
NLU	Text Classification	Single text	Domain identification, Intent detection, Sentiment classification
		Text pair	Information retrieval, Natural language inference
	Sequence Labeling	Single text	Named entity recognition, Slot tagging
		Text pair	Machine reading comprehension
	Structure Prediction	Single text	Dependency parsing, constituency parsing, semantic role labeling
NLG	Text Generation	Token level	Spelling correction, Sentence auto completion
		Sentence level	Machine translation, Conversation, Question generation

Cross-lingual Dependency Parsing

- Given treebanks of rich-resource languages (e.g. English), derive dependency parsing models for low-resource languages
- Benchmarks – Universal Dependencies (UD)
  - A framework for consistence annotation of grammar (POS, syntactic dependencies, morphological features)
  - Nearly 200 treebanks in over 100 languages

https://universaldependencies.org/introduction.html

Dependency Parsing

- Evaluation Metrics
  - Labeled attachment score (LAS) =
    percentage of words that get the correct head and label
  - Unlabeled attachment score (UAS) =
    percentage of words that get the correct head

https://universaldependencies.org/introduction.html

Monolingual Dependency Parsing

- Graph-based Parser
  - Define a space of candidate dependency trees for a sentence
  - Learn to score dependency trees, factored into subgraphs
  - Parse by finding the highest-scoring dependency tree
- Arc-factored Model
  - Every arc is a separate subgraph

$$s(G = (V, A)) = \sum_{(w_i, l, w_j) \in A} s(w_i, l, w_j)$$

score of the labeled arc

  - First parse an unlabeled tree, then label each arc
  - So we need two scoring functions
    - $s^{(arc)}(w_i, w_j)$ to construct the tree
    - $s^{(label)}(l|w_i, w_j)$ to label each predicted arc

Timothy Dozat and Christopher D. Manning. Deep biaffine attention for neural dependency parsing. ICLR 2017.

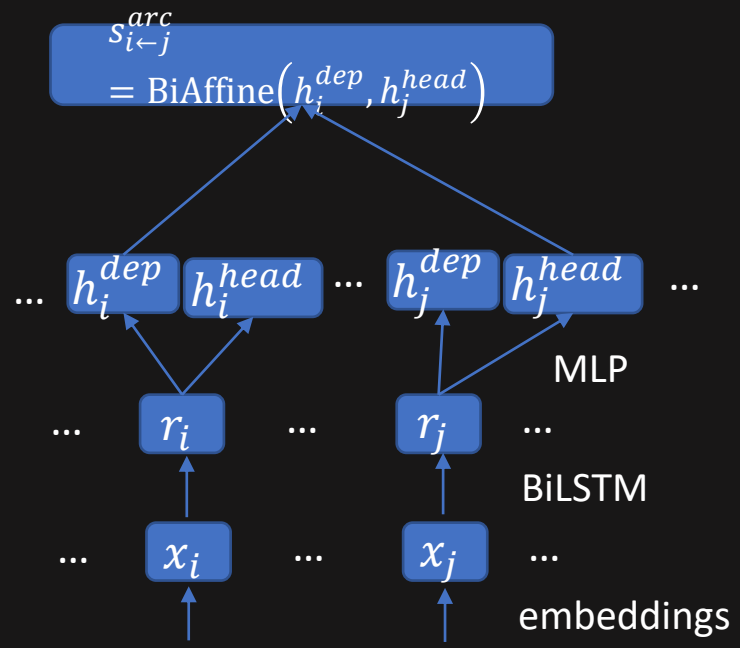
Monolingual Model

- Biaffine attention scoring for arc prediction

$$h_i^{arc-dep} = MLP^{arc-dep}(r_i)$$
$$h_j^{arc-head} = MLP^{arc-head}(r_j)$$
$$s_{i\leftarrow j}^{arc} = \text{BiAffine}\left(h_i^{arc-dep}, h_j^{arc-head}\right)$$
$$\text{BiAffine}(h_i, h_j) = h_j^T U^1 h_i + (h_j, h_i)^T U^2 + b$$

- Biaffine classifier to predict dependency labels

$$h_i^{label-dep} = MLP^{label-dep}(r_i)$$
$$h_j^{label-head} = MLP^{label-head}(r_j)$$
$$s_{i\leftarrow j}^{label} = \text{BiAffine}(h_i^{label-dep}, h_j^{label-head})$$



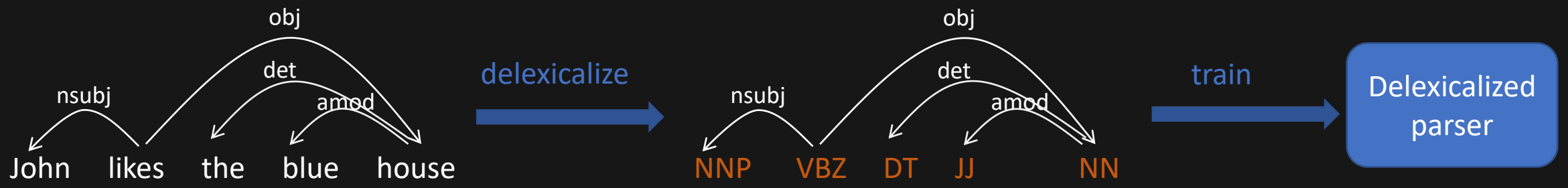
Timothy Dozat and Christopher D. Manning. Deep biaffine attention for neural dependency parsing. ICLR 2017.

Cross-lingual Dependency Parsing

- Background
- Model Transfer
  - Delexicalized model and improvements
  - Multilingual pre-trained models
- Data Transfer
  - Translation-Based Approach
    - Treebank Translation (source training data + translator)
    - Annotation Projection (parallel data + source model)
  - Unsupervised Approach
    - Source reordering
    - Direct unsupervised adaption

Traditional Approaches

- Delexicalized model
  - Using shared POS features across languages
  - Success for closely related languages



McDonald, Ryan, et al. "Universal dependency annotation for multilingual parsing." *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2013.

Traditional Approaches

- Fine-Grained Features

Feature	Description
81A	Order of Subject, Object and Verb
85A	Order of Adposition and Noun
86A	Order of Genitive and Noun
87A	Order of Adjective and Noun
88A	Order of Demonstrative and Noun
89A	Order of Numeral and Noun

Typological Features

Cluster	Lang.	Sample words
60	EN	was, wasn't, was'nt, wasn"t, hasnít, doesn't, ...
60	ES	estaba, estaráen, estubo, fúe, quedaba, ...
101	EN	very, mildly, wholly, terribly, gloriously, ...
101	ES	muchomás, fuerte, fuertes, duro, duros, poco, ...
153	EN	chicken, bird, ostriches, beef, pork, burger, steak, ...
153	ES	pollo, achote, manzana, tortugas, marsupiales, ...
195	EN	The, ...
195	ES	El, La, Los, Las, LoS, ...
236	EN	dry, wet, moist, lifeless, dullish, squarish, limpid, ...
236	ES	seco, secos, semiseco, semisecos, mojado, humedo, ...

Cross-lingual word clusters

Tahira Naseem, Regina Barzilay, and Amir Globerson. Selective sharing for multilingual dependency parsingAmmar. ACL 2012
Waleed, et al. "Many languages, one parser." *TACL 2016.*
Täckström, Oscar, Ryan McDonald, and Jakob Uszkoreit. "Cross-lingual word clusters for direct transfer of linguistic structure." NAACL-HLT 2012

Modern Approach

- Pre-trained multilingual embeddings (mBERT) performs relatively well for zero-shot model transfer
- There is still a large gap between zero-shot transfer and supervised learning, especially for languages with large distance to English.

	Dist	mBERT(S)	Baseline(Z)	mBERT(Z)	mBERT(Z+POS)
en	0.00	91.5/81.3	90.4/**88.4**	91.5/81.3	**91.8**/82.2
no	0.06	93.6/85.9	80.8/**72.8**	80.6/68.9	**82.7**/72.1
sv	0.07	91.2/83.1	81.0/73.2	82.5/71.2	**84.3/73.7**
fr	0.09	91.7/85.4	77.9/72.8	82.7/72.7	**83.8/76.2**
pt	0.09	93.2/87.2	76.6/**67.8**	77.1/64.0	**78.3**/66.9
da	0.10	89.5/81.9	76.6/67.9	77.4/64.7	**79.3/68.1**
es	0.12	92.3/86.5	74.5/66.4	78.1/64.9	**79.0/68.9**
it	0.12	94.8/88.7	80.8/75.8	84.6/74.4	**86.0/77.8**
ca	0.13	94.3/89.5	73.8/65.1	78.1/64.6	**79.0/67.9**
hr	0.13	92.4/83.8	61.9/52.9	**80.7**/65.8	80.4/**68.2**
pl	0.13	94.7/79.9	74.6/62.2	82.8/59.4	**85.7/65.4**
sl	0.13	88.0/77.8	68.2/56.5	72.6/51.4	**75.9/59.2**
uk	0.13	90.6/83.4	60.1/52.3	**76.7**/60.0	76.5/**65.5**
bg	0.14	95.2/85.5	79.4/**68.2**	83.3/62.3	**84.4**/68.1
cs	0.14	94.2/86.6	63.1/53.8	76.6/58.7	**77.4/63.6**
de	0.14	86.1/76.5	71.3/61.6	80.4/66.3	**83.5/71.2**
he	0.14	91.9/83.6	55.3/48.0	**67.5**/48.4	67.0/**54.3**
nl	0.14	94.0/85.0	68.6/60.3	78.0/64.8	**79.9/67.1**
ru	0.14	94.7/88.0	60.6/51.6	**73.6**/58.5	73.2/**61.5**
ro	0.15	92.2/83.2	65.1/54.1	**77.0**/58.5	76.9/**62.6**
id	0.17	86.3/75.4	49.2/43.5	**62.6**/45.6	59.8/**48.6**
sk	0.17	93.8/83.3	66.7/58.2	82.7/63.9	**82.9/67.8**
lv	0.18	87.3/75.3	**70.8/49.3**	66.0/41.4	70.4/48.5
et	0.20	88.8/79.7	65.7/44.9	66.9/44.3	**70.8/50.7**
fi	0.20	91.3/81.8	66.3/48.7	68.4/47.5	**71.4/52.5**
zh*	0.23	88.3/81.2	42.5/25.1	**53.8**/26.8	53.4/**29.0**
ar	0.26	87.6/80.6	38.1/28.0	43.9/28.3	**44.7/32.9**
la	0.28	85.2/73.1	48.0/**35.2**	47.9/26.1	**50.9**/32.2
ko	0.33	86.0/74.8	34.5/16.4	**52.7**/27.5	52.3/**29.4**
hi	0.40	94.8/86.7	35.5/26.5	49.8/33.2	**58.9/44.0**
ja*	0.49	94.2/87.4	28.2/20.9	36.6/15.7	**41.3/30.9**
AVER	0.17	91.3/82.6	64.1/53.8	71.4/54.2	**73.0/58.9**

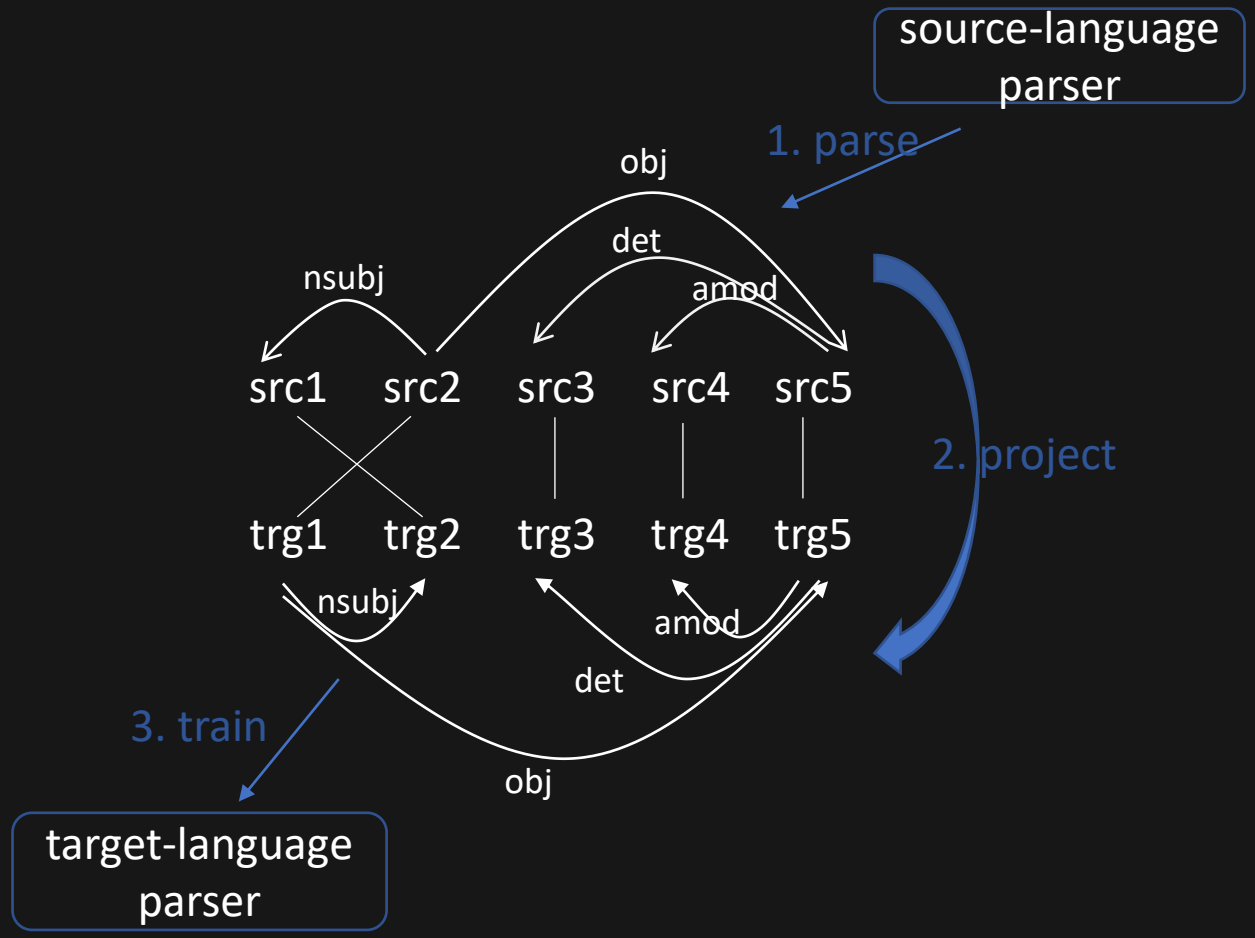
mBERT(S) : supervised learning
Baseline(Z): dictionary supervised cross-lingual word embeddings
mBERT(Z) : zero-shot transfer from English training data
mBERT(Z+POS) : gold POS is used as input
Languages are sorted according to word order distance to English

Wu, Shijie, and Mark Dredze. "Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT." EMNLP 2019

Cross-lingual Dependency Parsing

- Background
- Model Transfer
  - Delexicalized model and improvements
  - Multilingual pre-trained models
- Data Transfer
  - Translation-Based Approach
    - Treebank Translation
    - Annotation Projection
  - Unsupervised Approach
    - Source reordering
    - Direct unsupervised adaption

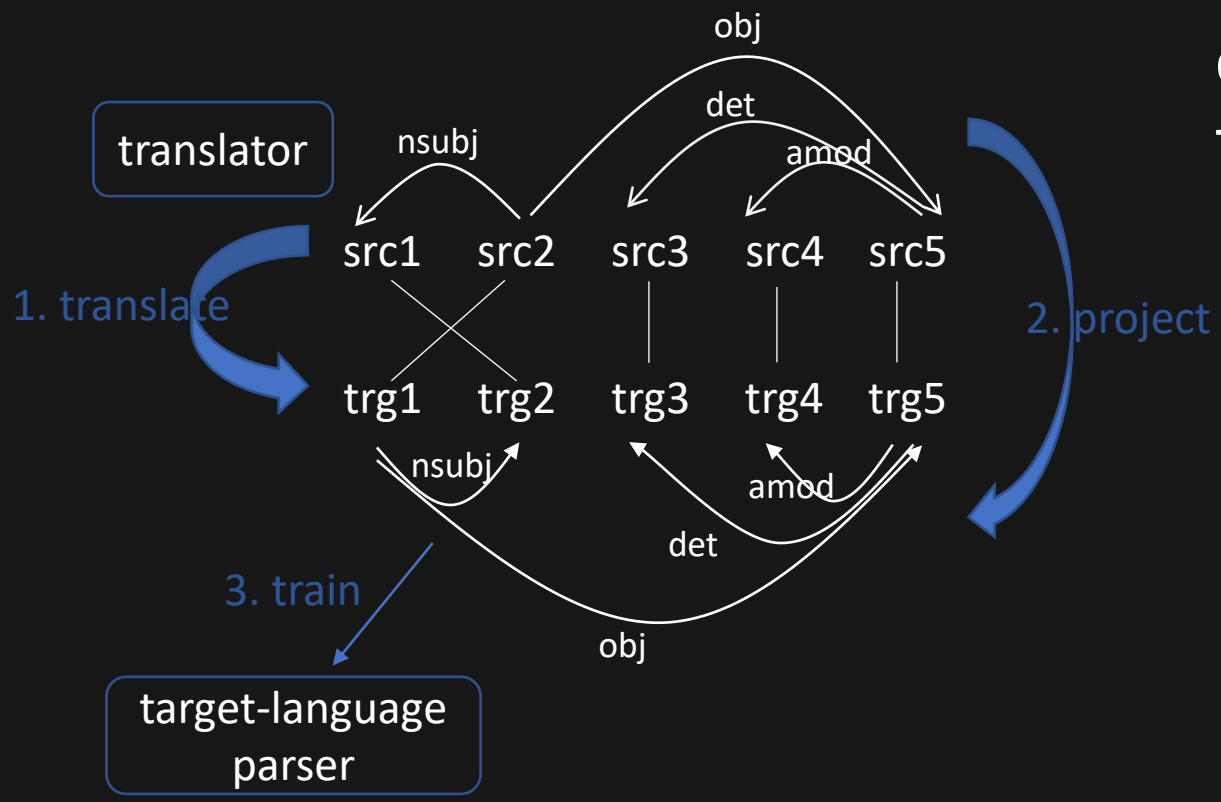
Annotation Projection



- Given a word-aligned parallel corpus and a source-language parser
  - Annotate the source sentences with the parser
  - Use word alignment to map the source annotation to target sentences
  - Train a target-language parser with the projected annotation

Yarowsky, D., Ngai, G., & Wicentowski, R. (2001). Inducing Multilingual Text Analysis Tools via Robust Projection Across Aligned Corpora. In Proceedings of HLT.
Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., & Kolak, O. (2005). Bootstrapping Parsers via Syntactic Projection across Parallel Texts. Natural Language Engineering.

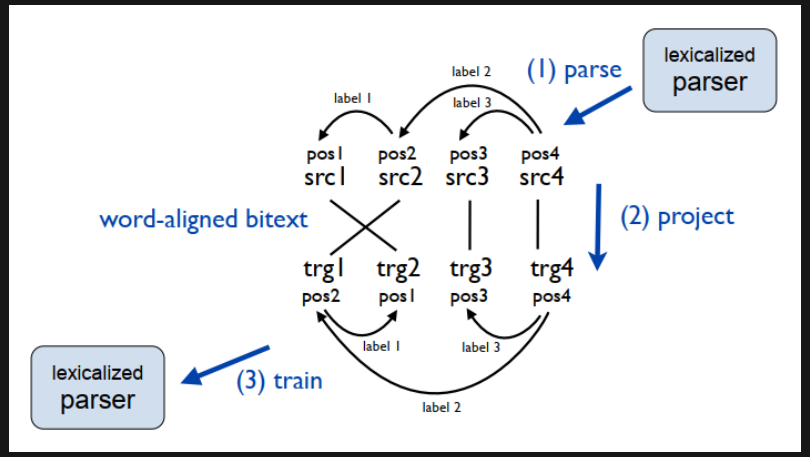
Treebank Translation

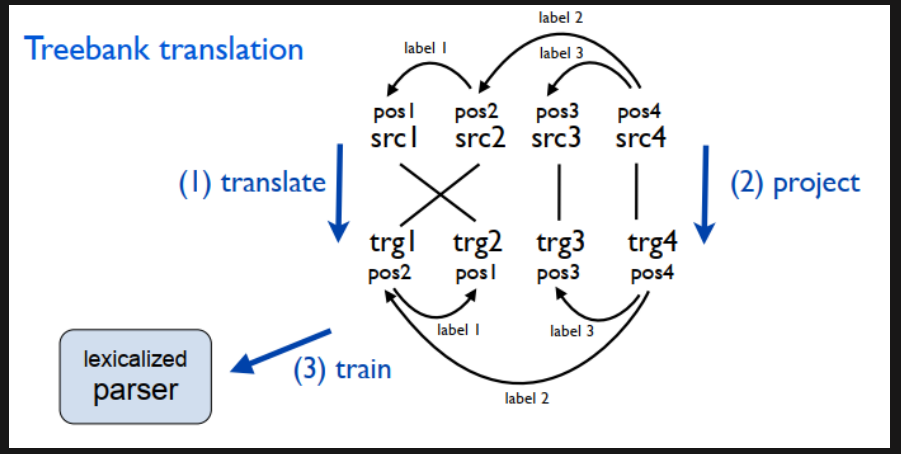


- Given a source-language treebank corpus and a source-to-target translator
  - Translate source sentences into target sentences with word-to-word or phrase-to-phrase alignment
  - Map source annotation to target sentences
  - Train a target-language parser with the projected annotation

Tiedemann, Jörg, and Zeljko Agić. "Synthetic treebanking for cross-lingual dependency parsing." *Journal of Artificial Intelligence Research* 55 (2016)

Discussion

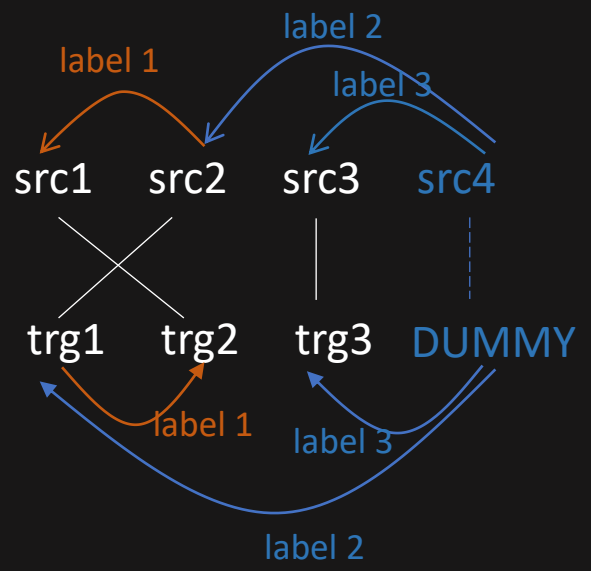




	Annotation Projection	Treebank Translation
Data Required	Source-Target sentence pairs	Annotated source treebank
Model Required	Source annotator	Translator
Source-language annotation	By an annotation model	By human annotators
Target-language sentence	By human	By a translation model
Target-language annotation	Projected from source annotation	Projected from source annotation

Projecting Strategy

- One-to-one
  - Directly copy dependency relations
  - Unaligned source language words are covered by additional DUMMY nodes
  - Unaligned target language words are deleted

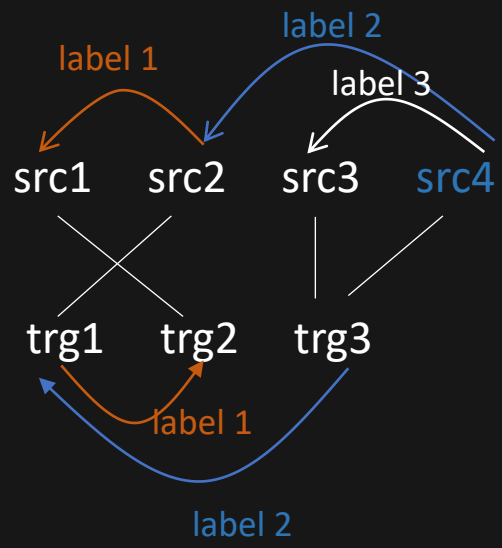


one-to-one

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping Parsers via Syntactic Projection across Parallel Texts. Natural Language Engineering, 11(3):311–325
Tiedemann, Jörg. "Rediscovering annotation projection for cross-lingual parser induction." *COLING 2014.*

Projecting Strategy

- One-to-one
- Many-to-one
  - Only keep the link to the head of the aligned source language words
  - Delete all other links

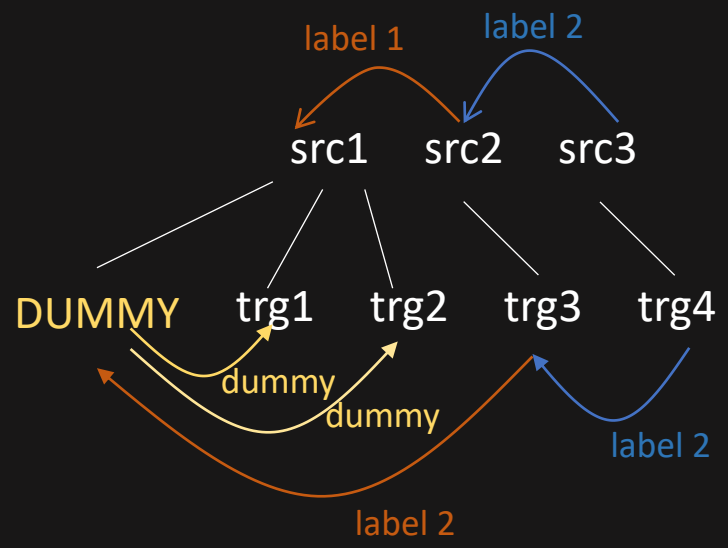


many to one

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping Parsers via Syntactic Projection across Parallel Texts. Natural Language Engineering, 11(3):311–325
Tiedemann, Jörg. "Rediscovering annotation projection for cross-lingual parser induction." *COLING 2014.*

Projecting Strategy

- One-to-one
- Many-to-one
- One-to-many
  - Introduce an additional DUMMY node
  - The DUMMY node acts as the immediate parent in the target language
  - The DUMMNY node captures the dependency relation of the source side annotation



one-to-many

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping Parsers via Syntactic Projection across Parallel Texts. Natural Language Engineering, 11(3):311–325
Tiedemann, Jörg. "Rediscovering annotation projection for cross-lingual parser induction." *COLING 2014.*

Projecting Strategy

- One-to-one
- Many-to-one
- One-to-many
- Many-to-many
  - First apply one-to-many, then apply many-to-one

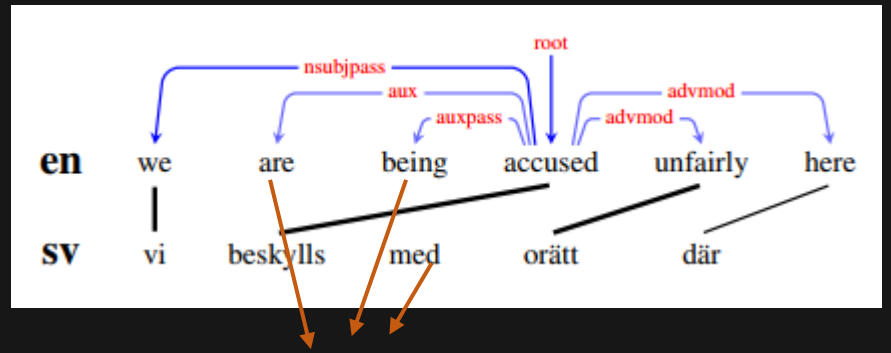
Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping Parsers via Syntactic Projection across Parallel Texts. Natural Language Engineering, 11(3):311–325
Tiedemann, Jörg. "Rediscovering annotation projection for cross-lingual parser induction." *COLING 2014.*

Discussion

- Advantages
  - Synthesize target-language data for model training
  - Competitive and complementary to cross-lingual models

Discussion

- Advantages
  - Synthesize target-language data for model training
  - Competitive and complementary to cross-lingual models
- Limitations
  - Non-isomorphic alignment between source and target language



The English words “are” and “being”, and the Swedish word “med”, do not have corresponding word-level translation

Discussion

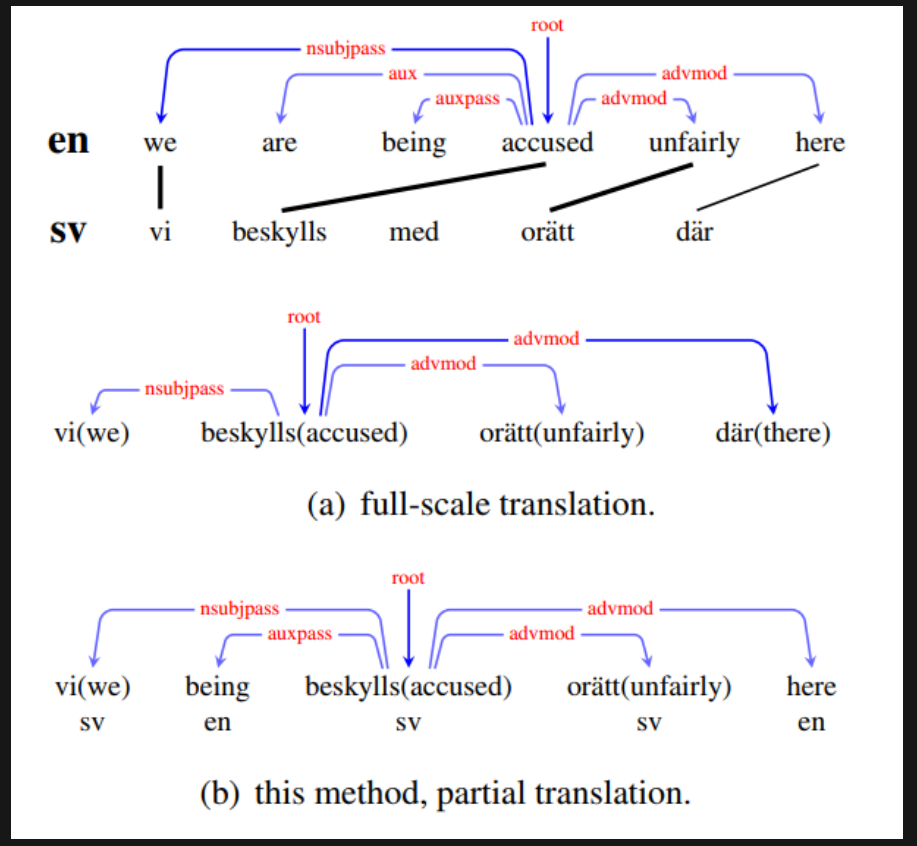
- Advantages
  - Synthesize target-language data for model training
  - Competitive and complementary to cross-lingual models
- Limitations
  - Non-isomorphic alignment between source and target language
  - Translation error or parsing error

Discussion

- Advantages
  - Synthesize target-language data for model training
  - Competitive and complementary to cross-lingual models
- Limitations
  - Non-isomorphic alignment between source and target language
  - Translation error or parsing error
  - Conflicting annotation

Code-Mixed Treebank

- Translate a source treebank into a code-mixed treebank
  - Source words with highly confident target alignments are translated into target words
  - Continuous spans of target words are reordered according to translation
- Make the best use of source syntax
  - With the minimum noise introduced
  - Capture the structure of target-language sentences as much as possible



Meishan Zhang, Yue Zhang, and Guohong Fu. Cross-lingual dependency parsing using code-mixed treebank. EMNLP 2019

Code-Mixed Treebank

Lang.	Delex		PartProj		Src		Tgt		Src + Tgt		Mix		**Src + Mix**	
	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS
DE	64.10	53.77	69.90	61.28	66.87	57.46	70.84	62.30	72.41	63.74	71.41	63.46	**72.78**	**64.38**
ES	71.53	63.33	75.81	66.83	75.63	65.85	76.49	67.39	77.00	67.95	81.18	71.80	**81.44**	**71.66**
FR	75.13	67.26	75.54	67.63	78.13	70.63	76.91	69.39	78.75	71.17	83.20	76.32	**83.77**	**76.48**
IT	77.71	69.27	77.71	69.27	81.11	72.83	79.30	71.65	81.56	74.09	85.30	77.43	**86.13**	**78.38**
PT	74.03	67.70	79.44	71.30	77.37	69.36	78.32	70.67	79.73	71.84	83.54	75.34	**84.05**	**75.89**
AVG	72.50	64.27	75.68	67.26	75.82	67.23	76.37	68.28	77.89	69.76	80.93	72.87	**81.63**	**73.36**

- All lexicalized model outperform Delex, showing effectiveness of lexicalized features
- The proposed code-mixed approach (Mix) increases the accuracy

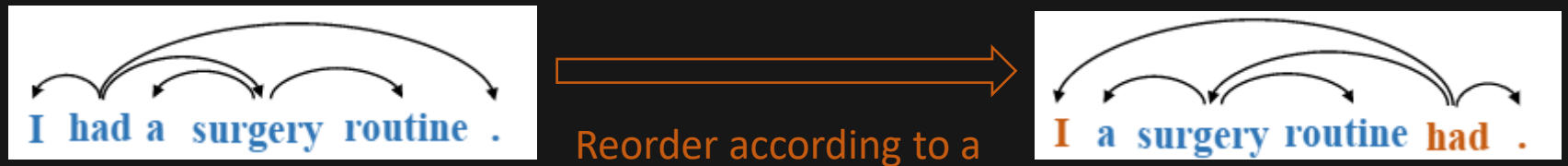
Meishan Zhang, Yue Zhang, and Guohong Fu. Cross-lingual dependency parsing using code-mixed treebank. EMNLP 2019

Cross-lingual Dependency Parsing

- Background
- Model Transfer
  - Delexicalized model and improvements
  - Multilingual pre-trained models
- Data Transfer
  - Translation-Based Approach
    - Treebank Translation (source training data + translator)
    - Annotation Projection (parallel data + source model)
  - Unsupervised Approach
    - Source reordering
    - Direct unsupervised adaption

Source Reordering

- Given a source-language treebank and unannotated target-language corpus
- Rearrange the order of source sentences to meet the word order in target language
- Under the guidance of a POS language model



Source-language tree

Reordered source-language tree

Liu, Lu, et al. "Cross-Lingual Dependency Parsing by POS-Guided Word Reordering." EMNLP Findings. 2020.

Source Reordering

- Given a source-language treebank and unannotated target-language corpus
- Rearrange the order of source sentences to meet the word order in target language
- Under the guidance of a POS language model

$$x^* = \arg \max_{x' \in R(x)} p_{\mathcal{T}}(x')$$

Set of all possible permutations of the words in x

The POS language model learned from the target-language corpus

Liu, Lu, et al. "Cross-Lingual Dependency Parsing by POS-Guided Word Reordering." EMNLP Findings. 2020.

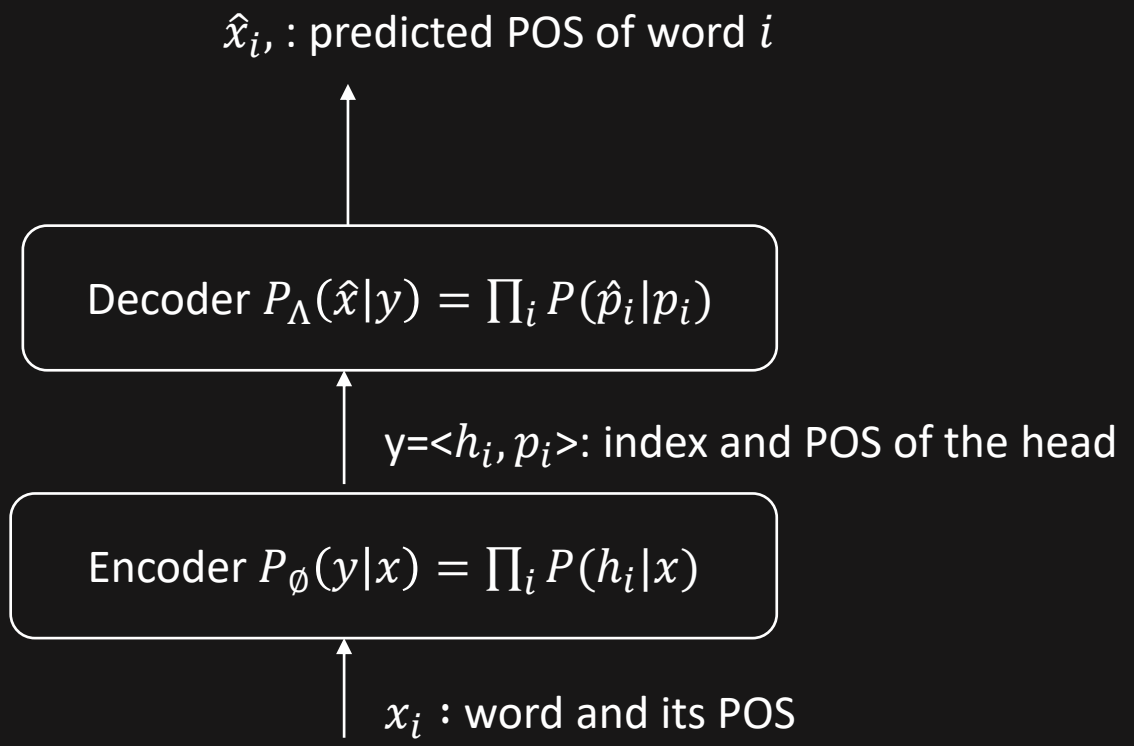
Unsupervised Adaptation

- Leverage the unsupervised parser based on CRF autoencoder framework
  - Applicable for both supervised and unsupervised learning
- Approach
  - Train a source parser
  - Use it to initialize and regularize a target parser
  - Unsupervised training on unannotated target data

Li, Zhao, and Kewei Tu. "Unsupervised cross-lingual adaptation of dependency parsers using CRF autoencoders.“ EMNLP Findings. 2020.

Unsupervised Adaptation

- Supervised Learning
$$\mathcal{L} = -\sum \log P(y^*, \hat{x}|x)$$
- Unsupervised Learning
$$\mathcal{L} = -\sum \max_{y} \log P(y, \hat{x}|x)$$
- Cross-lingual adaptation
  - Train a CRF autoencoder in a supervised way with multilingual BERT
  - Use the source model to initialized a second CRF autoencoder
  - Train it in an unsupervised way on unannotated target data with regularization by source model



Li, Zhao, and Kewei Tu. "Unsupervised cross-lingual adaptation of dependency parsers using CRF autoencoders." EMNLP Findings. 2020.

Data Transfer: Summary

	Required resource	Input sentence(X)	How to derive labels (Y)	Training data languages
Treebank translation	Source treebank + translator	translated	Alignment according to source treebank	Target language
Annotation projection	Parallel data + source annotator	from parallel data	Alignment according to source treebank annotated by source model	Target language
Source reordering	Source treebank +unannotated target sentences	Reordered from source sentence	Direct copy from source treebank	Source language
Direct unsupervised adaptation	Source treebank + unannotated target sentences	Real data	No label	Target language

Key Takeaway

- Dependency parsing parses a sentence into a dependency tree
  - Predicts the head and arc label for each token
- Model Transfer
  - Traditional approach leverages delexicalized features
  - Model approach relies on multilingual embeddings
- Data Transfer
  - Translation-based approach aligns annotation from source language to target language
  - Unsupervised learning approach directly uses unannotated target-language corpus

Outline

Daxin Jiang

Software Technology Center at Asia (STCA) of Microsoft

- Introduction [Daxin Jiang]
  - Motivating examples in Microsoft products
  - Problem description
  - Categorization of applications
  - Challenges
- Methodology [Daxin Jiang]
  - Model Transfer
  - Data Transfer
- Applications*
  - Dependency Parsing [Xiubo Geng]
  - Machine Reading Comprehension [Ming Gong]
  - Grammar Error Correction [Linjun Shou]
- Summary & Future directions [Jian Pei]

Linjun Shou Xiubo Geng Ming Gong

Software Technology Center at Asia (STCA) of Microsoft

Jian Pei

Simon Fraser University

*For more applications, please refer to our tutorial at The Web Conference 2021

Cross Lingual Sequence Labelling

- Machine Reading Comprehension (MRC) as Example

Ming Gong
migon@microsoft.com

Microsoft

Outline

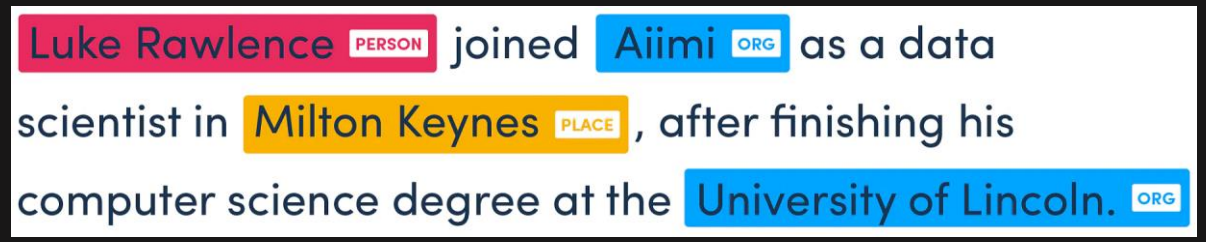
- Sequence Labeling Application Overview
- CLMRC Definition & Benchmark Datasets
- CLMRC Baseline Approach & Challenges
- Advanced Approaches for CLMRC
- Key Take-aways

Cross-lingual Sequence Labelling

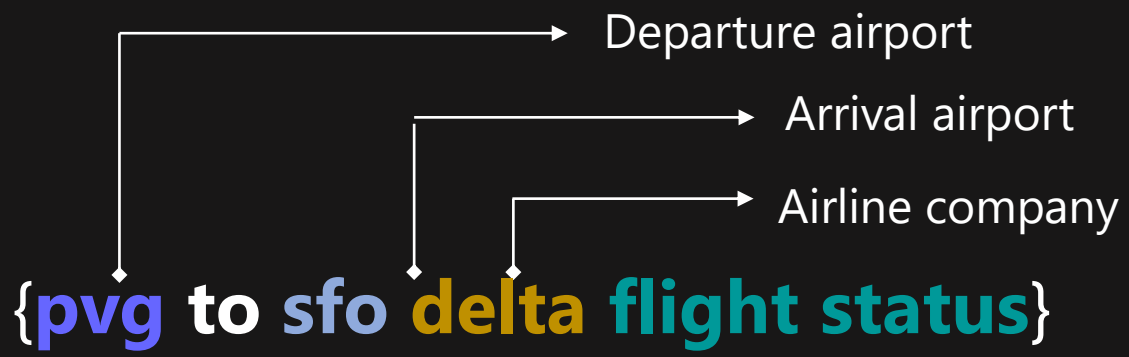
Type	Category	Sub Category	Example
NLU	Text Classification	Single text	Domain identification, Intent detection, Sentiment classification
		Text pair	Information retrieval, Natural language inference
	Sequence Labeling	Single text	Named entity recognition, Slot tagging
		Text pair	Extractive machine reading comprehension (MRC)
	Structure Prediction	Single text	Dependency parsing, semantic role labeling
NLG	Text Generation	Token level	Spelling correction, Sentence auto completion, Grammar error correction
		Sentence level	Machine translation, Conversation, Question generation

Sequence Labeling Example Tasks

- Single text
  - Given a text and a set of labels, predict the label for each token in the text

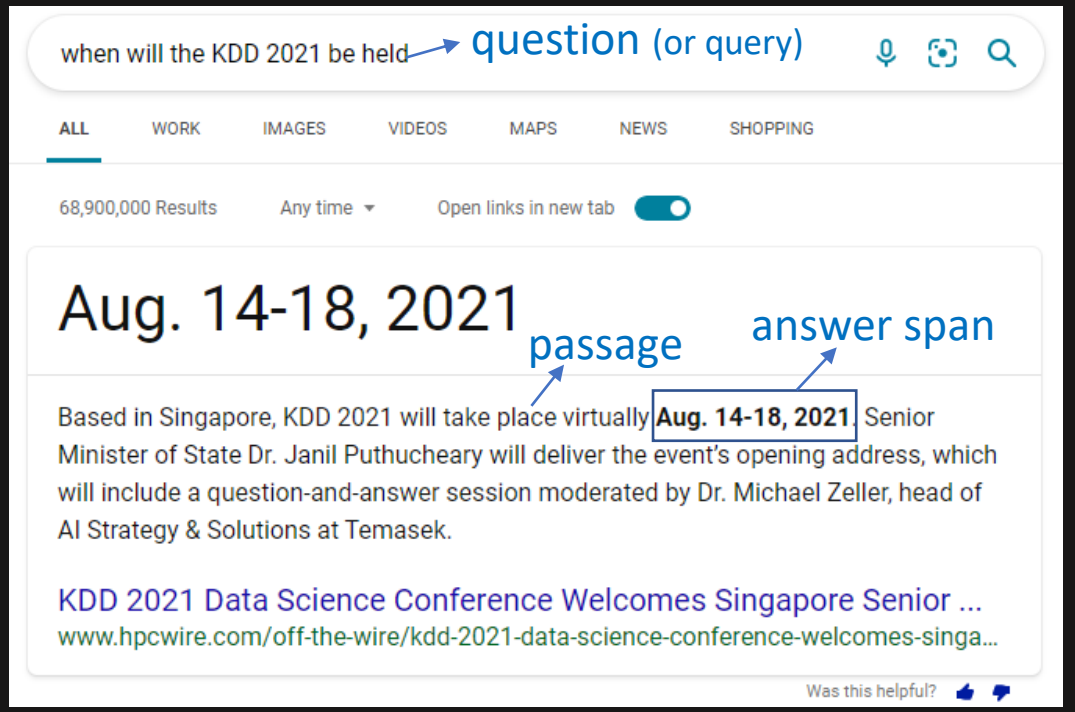


Application1: Named Entity Recognition (NER)



Application2: Slot Tagging

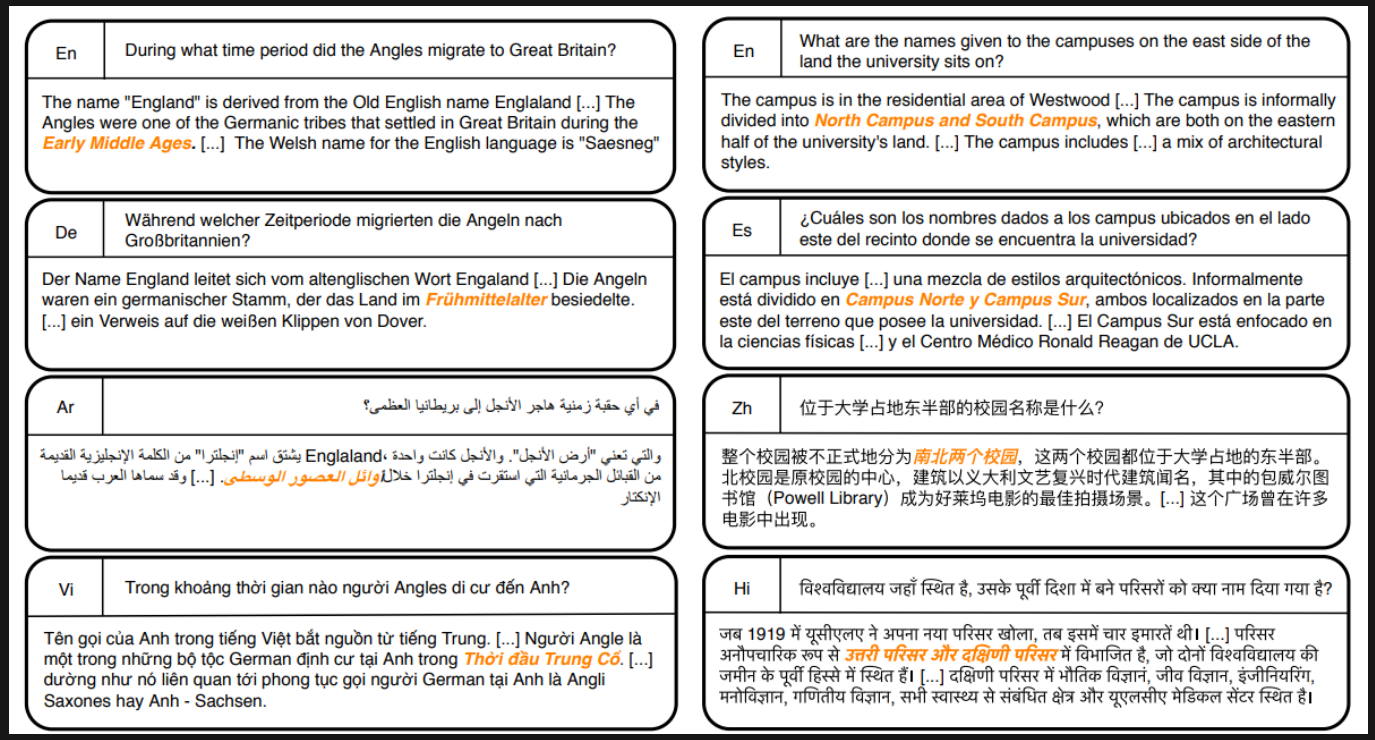
- Text pair
  - Given two texts, label the tokens in one text w.r.t. the other text.



Application3: Machine reading comprehension (MRC)

Cross-lingual MRC (CLMRC): Definition & Dataset

- Benchmark Datasets
  - *MLQA*: Patrack et al. MLQA: Evaluating Cross-lingual Extractive Question Answering. ACL 2020.
  - *TydiQA*: Clark, J. et al. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. TACL 2020.
  - *XQuAD*: Artetxe, M. et al. On the Cross-lingual Transferability of Monolingual Representations. ACL 2020.



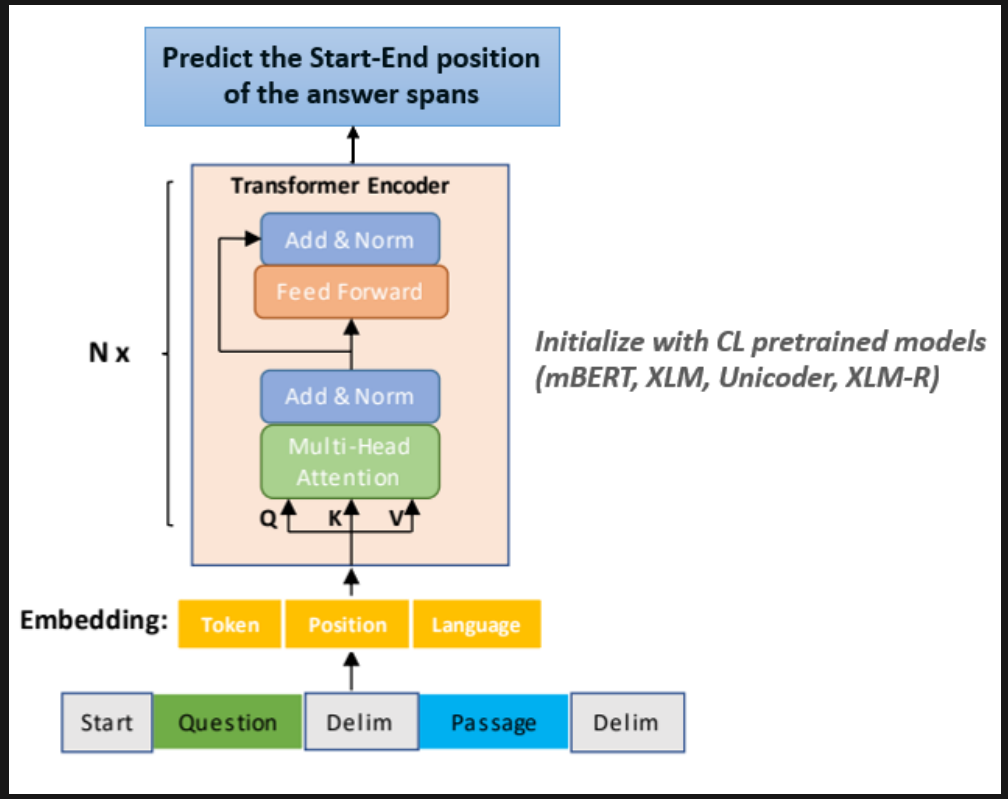
Examples from MLQA

Typical Experiment Setting

- There are only English labeled data for training.
- For non-EN languages, there are only test sets in specific languages.

Baseline Approach for CLMRC

- Zero-Shot Train
  - CL Pretrained Model + *English Labeled Data*
- Translation Train
  - CL Pretrained Model + *Translated Data in Target Languages*
  - CL Pretrained Model + *English Data + Translated Data in Target Languages*



Challenges of Baseline Approach for CLMRC

A CL pretrain models don't have sufficient transfer capability for phrase boundary tasks

B Machine Translation Data Quality

1. MT noises impact more for pair-wise tasks
2. Answer span alignment is challenging after MT

C Machine translation data is not available in some cases or quality is bad for tail languages

D Answer span boundary detection leads to the major errors of CLMRC model prediction

Challenges of Baseline Approach for CLMRC (A)

A CL pretrain models don't have sufficient transfer capability for phrase boundary tasks

B Machine Translation Data Quality

- MT noises impact more for pair-wise tasks
- Answer span alignment is challenging after MT

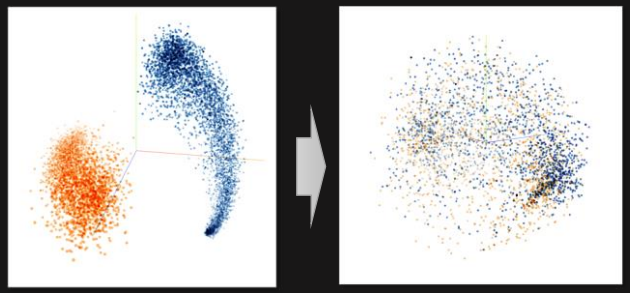
C Machine translation data is not available in some cases or quality is bad for tail languages

D Answer span boundary detection leads to the major errors of CLMRC model prediction

Classification *vs* Sequence Labeling using CL Pretrained models

Language	MRC EM	MRC Gap to en	NLI ACC	NLI Gap to en
en	62.4	–	85.0	–
es	49.8	-12.6	78.9	-6.1
de	47.6	-14.8	77.8	-7.2
ar	36.3	-26.1	73.1	-11.9
hi	27.3	-35.1	69.6	-15.4
vi	41.8	-20.6	76.1	-8.9
zh	39.6	-22.8	76.5	-8.5

Table 1: The gap between target languages and English on Machine Reading Comprehension (MRC) (Lewis et al., 2019) is significantly larger than sentence level classification task like Natural Language Inference (NLI) (Conneau et al., 2018). In this experiment, we fine-tune XLM (Conneau and Lample, 2019) on English and directly test on other languages.



For phrase boundary task, CL pretraining models *don't have sufficient transfer capability*

→ *Require better semantic alignment*

Fei Yuan, Linjun Shou, Xuanyu Bai, Ming Gong, Yaobo Liang, Nan Duan, Yan Fu, Daxin Jiang. Enhancing Answer Boundary Detection for Multilingual Machine Reading Comprehension. ACL, 2020.

Challenges of Baseline Approach for CLMRC (B-1)

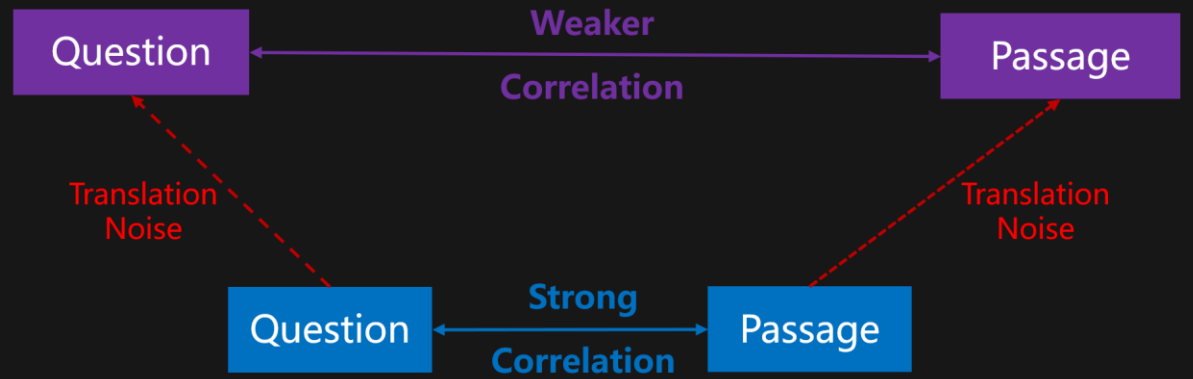
A CL pretrain models don't have sufficient transfer capability for phrase boundary tasks

B Machine Translation Data Quality

1. **MT noises impact more for pair-wise tasks**
2. Answer span alignment is challenging after MT

C Machine translation data is not available in some cases or quality is bad for tail languages

D Answer span boundary detection leads to the major errors of CLMRC model prediction



Challenges of Baseline Approach for CLMRC (B-2)

A CL pretrain models don't have sufficient transfer capability for phrase boundary tasks

B Machine Translation Data Quality

1. MT noises impact more for pair-wise tasks
2. **Answer span alignment is challenging after MT**

C Machine translation data is not available in some cases or quality is bad for tail languages

D Answer span boundary detection leads to the major errors of CLMRC model prediction

Question (EN): {Where is the Earth during the full moon?}

Passage (EN)

The full moon is the lunar phase when the Moon appears fully illuminated from Earth's perspective. This occurs when Earth is **located between the Sun and the Moon**.

Passage (Translate to DE)

Der Vollmond ist die Mondphase, in der der Mond vollständig aus der Perspektive der Erde erleuchtet erscheint. Dies geschieht, wenn sich die Erde **zwischen Sonne und Mond befindet**.

Answer span translate to DE: befindet sich zwischen der Sonne und dem Mond

Miss alignment with the correct span in DE passage

Challenges of Baseline Approach for CLMRC (C)

A CL pretrain models don't have sufficient transfer capability for phrase boundary tasks

B Machine Translation Data Quality

- MT noises impact more for pair-wise tasks
- Answer span alignment is challenging after MT

C Machine translation data is not available in some cases or quality is bad for tail languages

D Answer span boundary detection leads to the major errors of CLMRC model prediction

Challenges of Baseline Approach for CLMRC (D)

A CL pretrain models don't have sufficient transfer capability for phrase boundary tasks

B Machine Translation Data Quality

- MT noises impact more for pair-wise tasks
- Answer span alignment is challenging after MT

C Machine translation data is not available in some cases or quality is bad for tail languages

D Answer span boundary detection leads to the major errors of CLMRC model prediction

Error analysis on MLQA for zero-shot CLMRC model using XLM (Lewis et al., 2019) showed that major errors come from answer spans *partially overlap with golden span*.

[Question]: who were the kings of the southern kingdom
[Passage]: In the southern kingdom there was only one dynasty, that of king David, except usurper Athaliah from the northern kingdom, who by marriage, []
[Answer - ground truth]: king David
[Answer - model predication:] David, except usurper Athaliah

[Question]: What is the suggested initial does dosage of chlordiazepoxide
[Passage]: If the drug is administered orally, the suggested initial dose is 50 to 100 mg, to be followed by repeated doses as needed until agitation is controlled up to 300 mg per day. []
[Answer - ground truth]: 50 to 100 mg
[Answer - model predication:] 100 mg

Table 2: Bad answer boundary detection cases of multilingual MRC model.

	#Test	#Error	Boundary error
es	4, 054	955	*66.4%*
de	5,390	1,648	*75.2%*

Fei Yuan, Linjun Shou, Xuanyu Bai, Ming Gong, Yaobo Liang, Nan Duan, Yan Fu, Daxin Jiang. Enhancing Answer Boundary Detection for Multilingual Machine Reading Comprehension. ACL, 2020.

Advanced Approaches for CLMRC

Microsoft

Challenges of the Baseline Approach

A CL pretrain models don't have sufficient transfer capability for phrase boundary tasks

B Machine Translation Data Quality

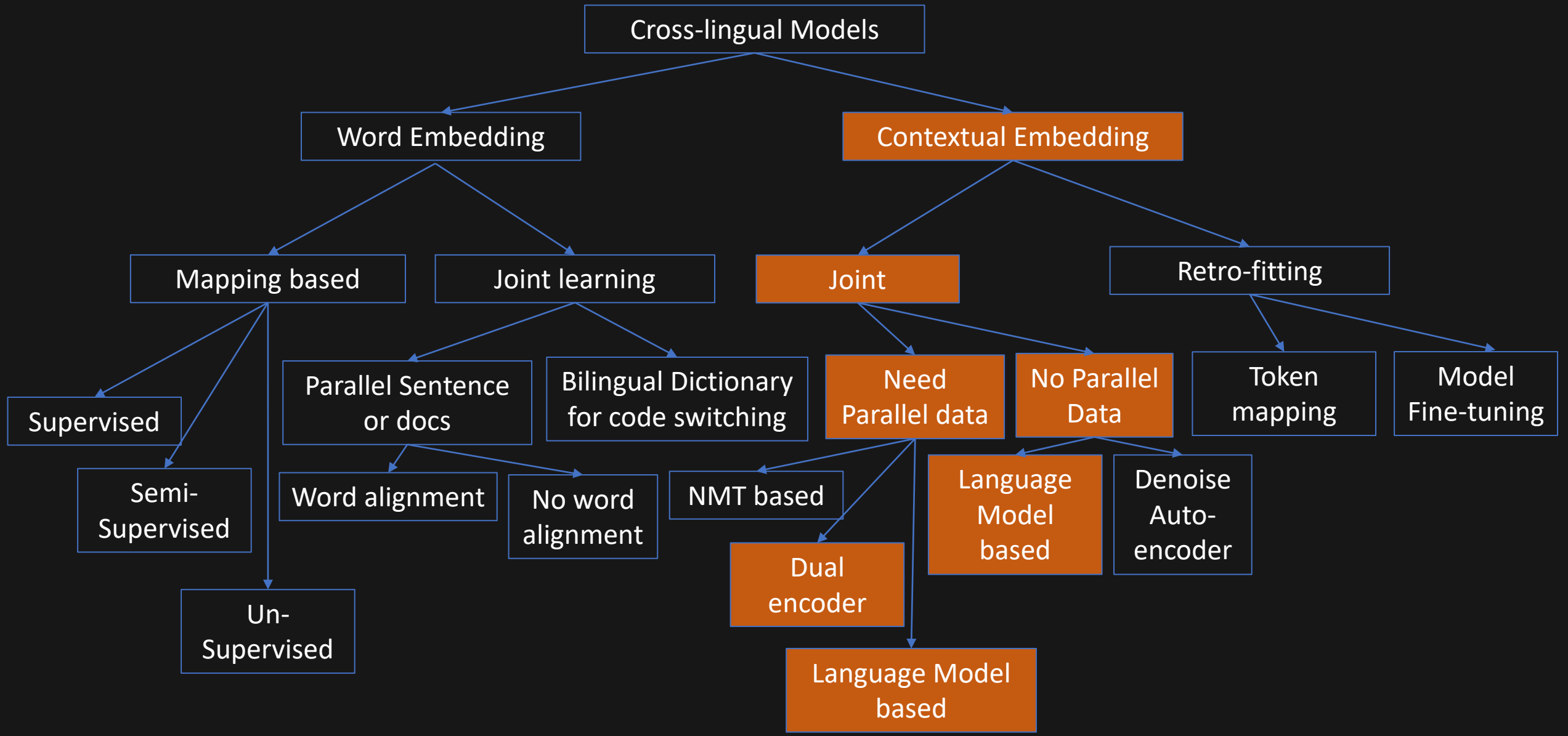
1. MT noises impact more for pair-wise tasks
2. Answer span alignment is challenging after MT

C Machine translation data is not available in some cases or quality is bad for tail languages

D Answer span boundary detection leads to the major errors of CLMRC model prediction

Advanced Approaches

Model transfer	Data transfer
• New Mix-MRC task • Dual encoders • Adversarial training	+ MT data
• Multiple teacher models + Language branch MT data • Instance level best teacher selection through reinforcement learning • Back-translation & Verification to improve answer span alignment	
• Unlabeled language specific data • Generated language specific data	
• New LAKM task + Language specific phrase knowledge data • Calibration network for primary boundary correction	


Cross-lingual Models
Word Embedding
Contextual Embedding
Mapping based
Joint learning
Joint
Retro-fitting
Supervised
Semi-Supervised
Un-Supervised
Parallel Sentence or docs
Bilingual Dictionary for code switching
Word alignment
No word alignment
Need Parallel data
No Parallel Data
NMT based
Dual encoder
Language Model based
Language Model based
Denoise Auto-encoder
Token mapping
Model Fine-tuning

Advanced Approaches for CLMRC

Microsoft

Challenges of the Baseline Approach

A CL pretrain models don't have sufficient transfer capability for phrase boundary tasks

B Machine Translation Data Quality

- MT noises impact more for pair-wise tasks
- Answer span alignment is challenging after MT

C Machine translation data is not available in some cases or quality is bad for tail languages

D Answer span boundary detection leads to the major errors of CLMRC model prediction

Advanced Approaches

Model transfer		Data transfer
• New Mix-MRC task • Dual encoders • Adversarial training	+	MT data

Mixed MRC Task with MT Data for Semantic Transfer

- Motivation: if human truly understand two languages, they could well perform MRC task in mixed languages. Could we enhance model semantic transfer capability using the MixMRC task?

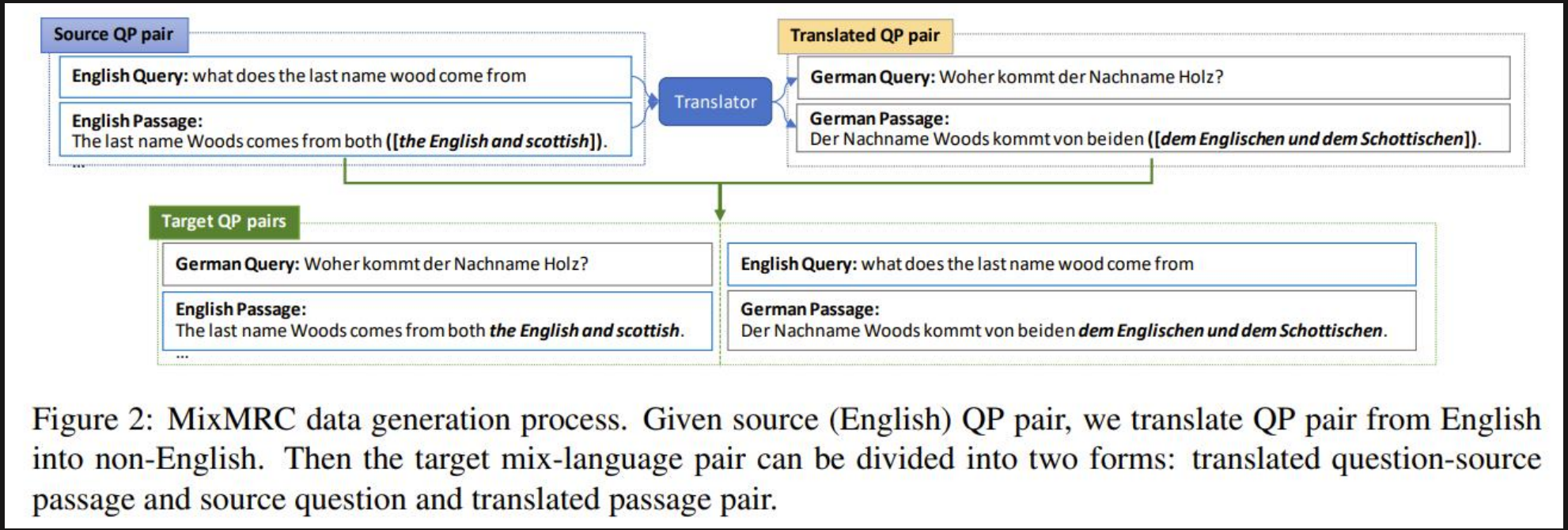


Figure 2: MixMRC data generation process. Given source (English) QP pair, we translate QP pair from English into non-English. Then the target mix-language pair can be divided into two forms: translated question-source passage and source question and translated passage pair.

Fei Yuan, Linjun Shou, Xuanyu Bai, Ming Gong, Yaobo Liang, Nan Duan, Yan Fu, Daxin Jiang. Enhancing Answer Boundary Detection for Multilingual Machine Reading Comprehension. ACL, 2020.

Mixed MRC Task with MT Data for Semantic Transfer

- Translation Train Results

Model	Methods	MLQA (EM / F1) en	MLQA (EM / F1) es	MLQA (EM / F1) de	MTQA (EM / F1) en	MTQA (EM / F1) fr	MTQA (EM / F1) de
M-BERT	Lewis et al. (2019)	65.2 / 77.7	37.4 / 53.9	47.5 / 62.0	-	-	-
	Baseline	65.4 / 79.0	50.4 / 68.5	46.2 / 60.6	67.0 / 86.9	52.9 / 78.2	59.8 / 81.4
	LAKM	**66.9** / 80.1	51.5 / 69.5	49.9 / 64.4	**68.8** / 87.6	56.8 / 78.8	62.4 / 81.9
	mixMRC	65.4 / 79.4	50.5 / 69.1	49.1 / 64.0	67.9 / 86.8	56.4 / 77.8	62.4 / 81.0
	mixMRC + LAKM	64.7 / 79.2	**52.1** / 70.4	**50.9** / 65.6	68.6 / 87.0	**57.5** / 78.5	**62.9** / 81.3
XLM	Lewis et al. (2019)	62.4 / 74.9	47.8 / 65.2	46.7 / 61.4	-	-	-
	Baseline	64.1 / 77.6	50.4 / 68.4	47.4 / 62.0	67.1 / 86.8	51.5 / 75.8	61.6 / 81.3
	LAKM	**64.6** / 79.0	52.2 / 70.2	50.6 / 65.4	**68.3** / 87.3	52.5 / 75.9	61.9 / 81.2
	mixMRC	63.8 / 78.0	52.1 / 69.9	49.8 / 64.8	66.5 / 85.9	52.9 / 75.0	62.1 / 80.5
	mixMRC + LAKM	64.4 / 79.1	**52.2** / 70.3	**51.2** / 66.0	68.2 / 86.8	**53.6** / 75.9	**62.5** / 80.9

Table 7: Experimental results on MLQA and MTQA dataset under translation condition (%).

Fei Yuan, Linjun Shou, Xuanyu Bai, Ming Gong, Yaobo Liang, Nan Duan, Yan Fu, Daxin Jiang. Enhancing Answer Boundary Detection for Multilingual Machine Reading Comprehension. ACL, 2020.

Dual Language Encoders for Semantic Transfer - DualBERT

- Simultaneously model the training data in both source and target language to better exploit the relations between two languages

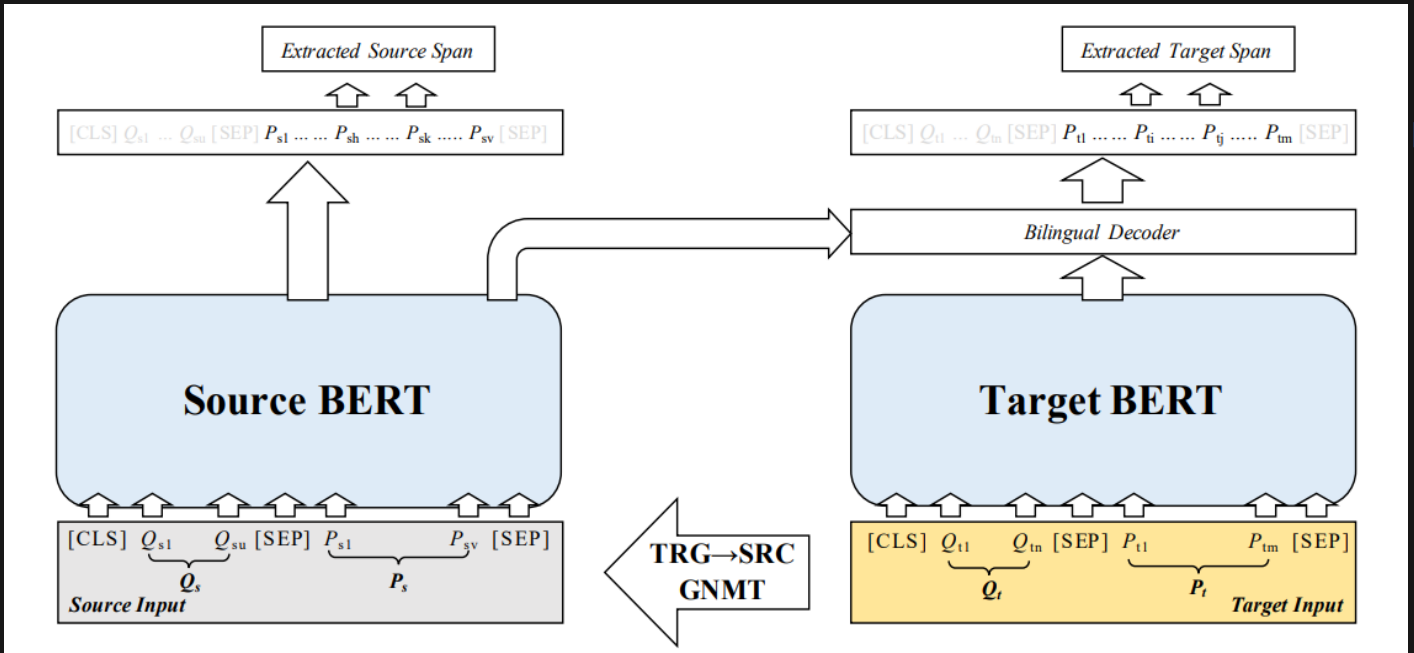


Figure 2: System overview of the Dual BERT model for cross-lingual machine reading comprehension task.

Experiment Results

#	System	CMRC 2018 Dev EM	CMRC 2018 Dev F1	CMRC 2018 Test EM	CMRC 2018 Test F1	CMRC 2018 Challenge EM	CMRC 2018 Challenge F1	DRCD Dev EM	DRCD Dev F1	DRCD Test EM	DRCD Test F1
6	$\text{BERT}_{B_{cn}}$	63.6	83.9	67.8	86.0	18.4	42.1	83.4	90.1	81.9	89.0
7	$\text{BERT}_{B_{mul}}$	64.1	84.4	68.6	86.8	18.6	43.8	83.2	89.9	82.4	89.5
8	**Dual BERT**	65.8	86.3	70.4	88.1	23.8	47.9	84.5	90.8	83.7	90.3

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, Guoping Hu. **Cross-Lingual Machine Reading Comprehension.** EMNLP 2019.

Dual Language Encoders for Semantic Transfer - FILTER

- Simultaneously model the training data in both source and target language to better exploit the relations between two languages

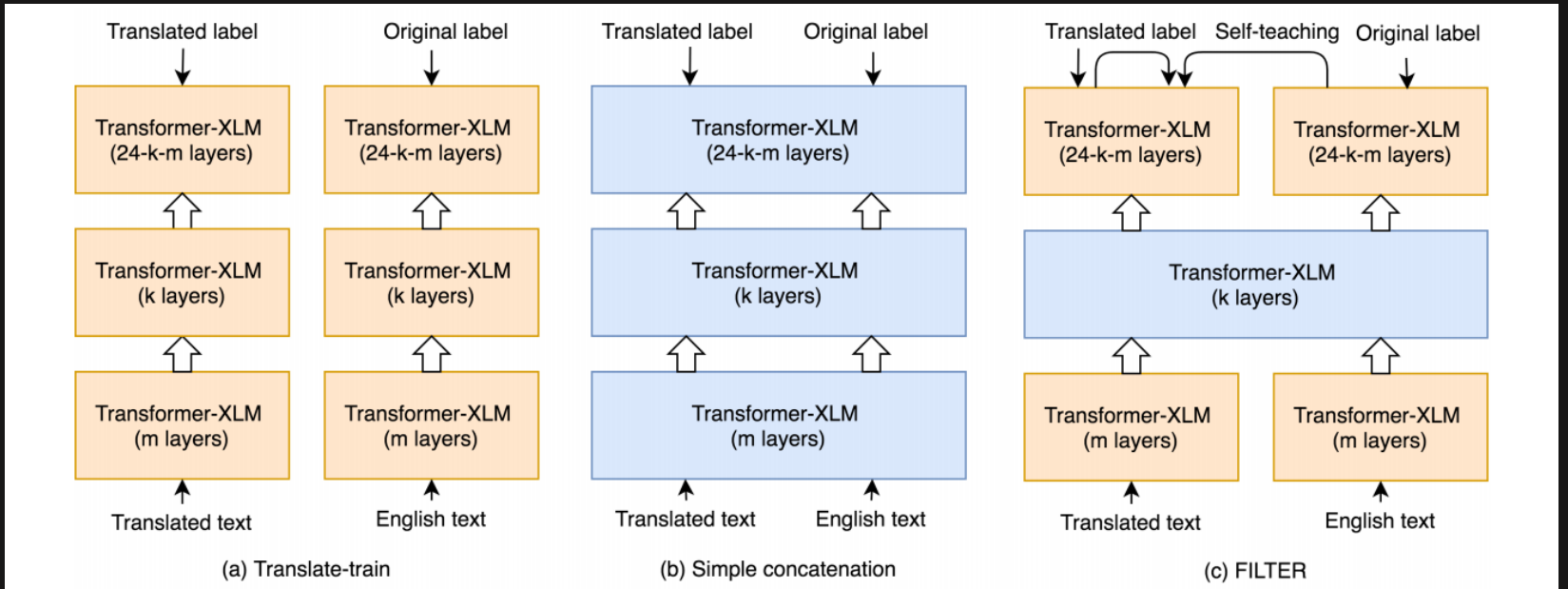


Figure 2: Comparison between different methods for finetuning XLM-R model for the XTREME benchmark. (a) Translate-train baseline. (b) Another baseline via simple concatenation of translated text. (c) Proposed FILTER approach. (a) and (b) can be considered as special instantiations of FILTER by setting $m = 24, k = 0$ and $m = 0, k = 24$, respectively.

Yuwei Fang, Shuohang Wang, Zhe Gan, Siqi Sun, Jingjing Liu. **FILTER: An Enhanced Fusion Method for Cross-lingual Language Understanding.** AAAI 2021.

Dual Language Encoders for Semantic Transfer - FILTER

Model	Pair sentence		Structured prediction		Question answering		
	XNLI	PAWS-X	POS	NER	XQuAD	MLQA	TyDiQA-GoldP
Metrics	Acc.	Acc.	F1	F1	F1 / EM	F1 / EM	F1 / EM
Cross-lingual zero-shot transfer (models are trained on English data)							
mBERT	65.4	81.9	70.3	62.2	64.5 / 49.4	61.4 / 44.2	59.7 / 43.9
XLM	69.1	80.9	70.1	61.2	59.8 / 44.3	48.5 / 32.6	43.6 / 29.1
XLM-R	79.2	86.4	72.6	65.4	76.6 / 60.8	71.6 / 53.2	65.1 / 45.0
InfoXLM	81.4	-	-	-	- / -	73.6 / 55.2	- / -
X-STILTs	80.0	87.9	74.4	64.0	78.7 / 63.3	72.4 / 53.7	**76.0 / 59.5**
Translate-train (models are trained on English training data and its translated data on the target language)							
mBERT	74.0	86.3	-	-	70.0 / 56.0	65.6 / 48.0	55.1 / 42.1
mBERT, multi-task	75.1	88.9	-	-	72.4 / 58.3	67.6 / 49.8	64.2 / 49.3
XLM-R, multi-task (Ours)	82.6	90.4	-	-	80.2 / 65.9	72.8 / 54.3	66.5 / 47.7
FILTER (Ours)	83.6	91.2	75.5	66.7	82.3 / 67.8	75.8 / 57.2	68.1 / 49.7
FILTER + Self-Teaching (Ours)	**83.9**	**91.4**	**76.2**	**67.7**	**82.4 / 68.0**	**76.2 / 57.7**	68.3 / 50.9

Yuwei Fang, Shuohang Wang, Zhe Gan, Siqi Sun, Jingjing Liu. **FILTER: An Enhanced Fusion Method for Cross-lingual Language Understanding.** AAAI 2021.

Multi-lingual Adversarial Training for Semantic Transfer

- Adversarial Training to bring the embeddings of different languages closer to each other to achieve effective cross-lingual transfer

Adversarial Loss

$$\mathcal{L}_{QA} + \mathcal{L}_{adv} = \mathcal{L}_{QA} - \sum_{l=1}^{L} KL[U(\boldsymbol{g}_l) || log\boldsymbol{p}_l]$$

$$\mathcal{L}_D = - \sum_{l=1}^{L} \mathbb{1}(\boldsymbol{g}_l) log\boldsymbol{p}_l \ \text{(Discriminator)}$$

Mihaela Bornea, Lin Pan, Sara Rosenthal, Radu Florian, Avirup Sil. **Multilingual Transfer Learning for QA Using Translation as Data Augmentation.** AAAI 2021.

Multi-lingual Adversarial Training for Semantic Transfer

- Bringing the embeddings of different languages closer to each other to achieve effective cross-lingual transfer

Model	Method	MLQA Languages (G-XLT) ar	de	en	es	hi	vi	zh	G-XLT	XLT
MBERT$_{QA}$	ZS	46.9	51.4	60.2	55.0	47.0	52.0	49.7	51.7 ($\pm$0.4)	61.7 ($\pm$0.3)
Trans	T(Q)	**53.8**	60.8	73.5	65.4	**53.2**	63.2	56.7	60.9 ($\pm$0.2)	**64.9** ($\pm$0.2)
	T(C)	44.8	51.7	62.0	57.6	42.7	55.8	50.8	52.2 ($\pm$1.0)	58.5 ($\pm$0.9)
	T(Q+C)	48.9	58.4	70.4	63.6	46.8	61.4	54.4	57.7 ($\pm$0.1)	64.3 ($\pm$0.0)
	T(ALL)	52.6	**61.1**	**73.8**	**66.3**	50.6	**64.8**	**58.2**	**61.1** ($\pm$ 0.1)	64.2 ($\pm$0.2)
AT	(en-zh)	50.5	56.7	68.0	60.8	50.7	57.4	51.5	56.5 ($\pm$ 0.1)	62.8 ($\pm$ 0.1)
	(en-all)	**54.1**	**61.1**	**73.6**	**65.5**	**54.2**	**63.4**	**56.8**	**61.2** ($\pm$ 0.1)	**65.2** ($\pm$ 0.1)

Mihaela Bornea, Lin Pan, Sara Rosenthal, Radu Florian, Avirup Sil. **Multilingual Transfer Learning for QA Using Translation as Data Augmentation.** AAAI 2021.

Advanced Approaches for CLMRC

Microsoft

Challenges of the Baseline Approach

A CL pretrain models don't have sufficient transfer capability for phrase boundary tasks

B Machine Translation Data Quality

1. MT noises impact more for pair-wise tasks
2. Answer span alignment is challenging after MT

C Machine translation data is not available in some cases or quality is bad for tail languages

D Answer span boundary detection leads to the major errors of CLMRC model prediction

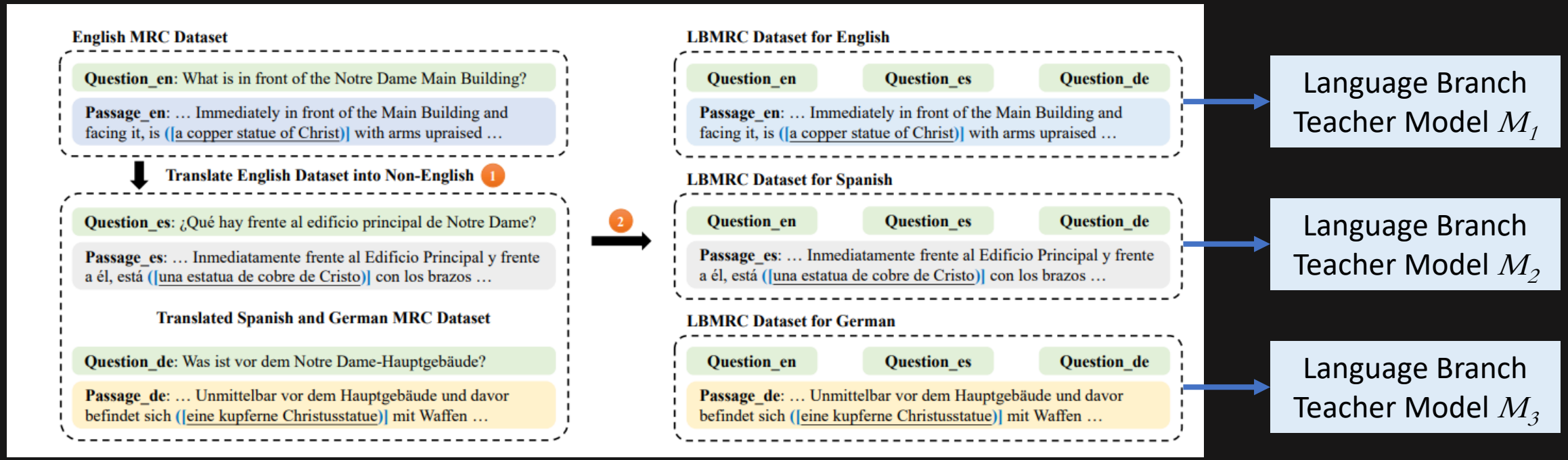
Model transfer	Data transfer
• Multiple teacher models + Language branch MT data • Instance level best teacher selection through reinforcement learning • Back-translation & Verification to improve answer span alignment	

B

Microsoft

LBMRC: Language Branch Knowledge Distillation for Reducing MT Noise

- **Motivation:** How to reduce the negative impact of the noisy MT data?
  → Leverage multiple teacher models to overcome the MT noise

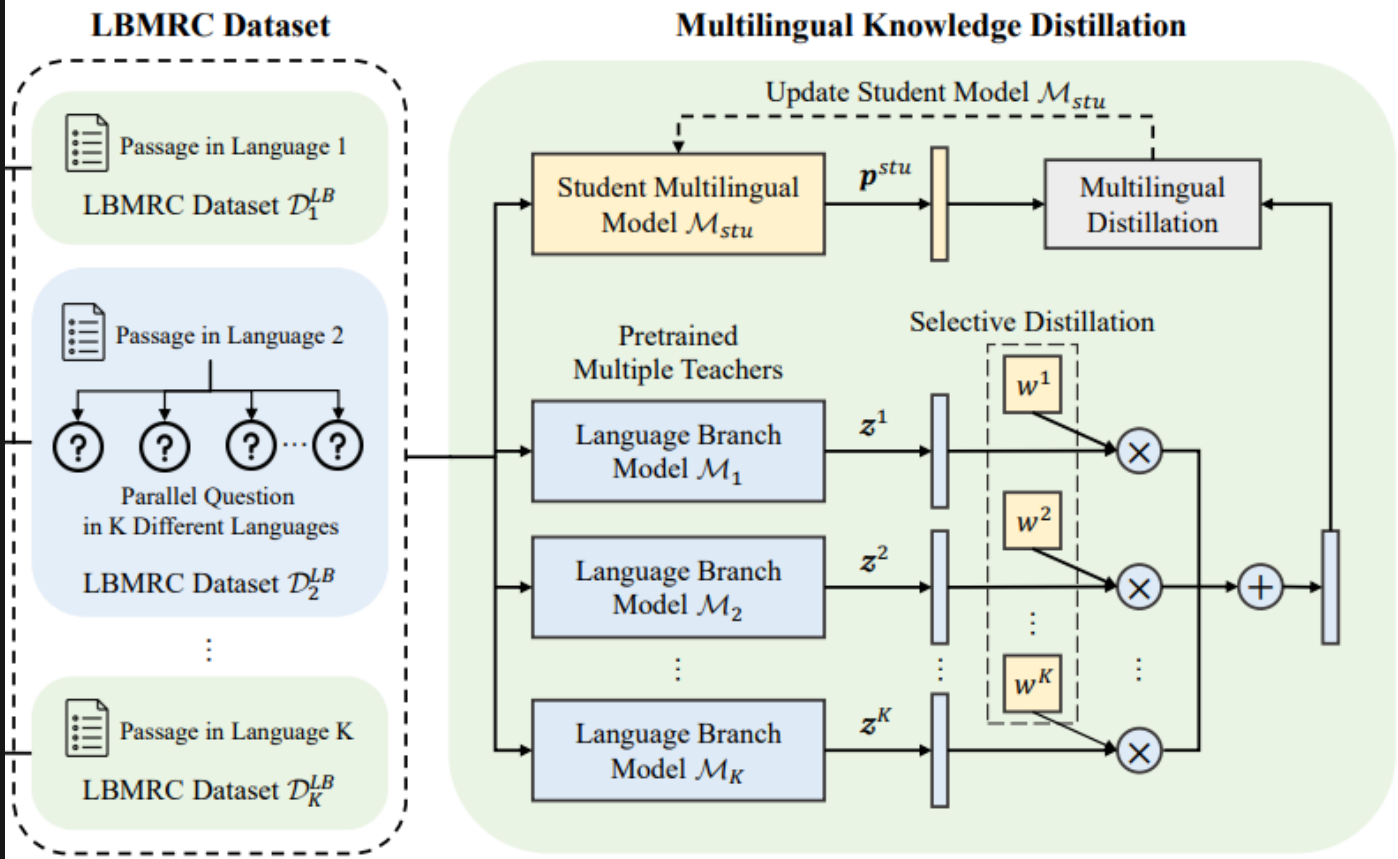


Junhao Liu, Linjun Shou, Jian Pei, Ming Gong, Min Yang, and Daxin Jiang. **Cross-lingual Machine Reading Comprehension with Language Branch Knowledge Distillation.** COLING. 2020

B

Microsoft

LBMRC: Language Branch Knowledge Distillation for Reducing MT Noise



Methods	MLQA (EM / F1)					
	en	es	de	ar	hi	vi
Lewis¶	62.4 / 74.9	47.8 / 65.2	46.7 / 61.4	34.4 / 54.0	33.4 / 50.7	39.4 / 59.3
Baseline	63.4 / 77.3	49.7 / 68.2	48.5 / 63.7	37.5 / 56.9	37.0 / 54.3	42.4 / 63.5
LAKM	64.6 / 79.0	52.2 / 70.2	50.6 / 65.4	-	-	-
mixMRC	63.8 / 78.0	52.1 / 69.9	49.8 / 64.8	38.5 / 58.4	40.1 / 57.1	45.2 / 66.2
mixMRC + LAKM	64.4 / 79.1	52.2 / 70.3	51.2 / 66.0	-	-	-
Ours-hyper	**64.8 / 79.3**	53.9 / 71.8	52.1 / 66.8	**40.4 / 60.0**	42.8 / 59.8	46.1 / 67.2
Ours-imp	64.7 / 79.2	**54.3 / 72.0**	**52.4 / 66.9**	40.1 / 59.9	**42.9 / 59.9**	**46.5 / 67.5**

Table 1: EM and F1 score of 6 languages on the MLQA dataset. The left 3 languages (en, es, de) are under translation condition while the right part (ar, hi, vi) results are under the zero-shot transfer method. The results with ¶ are adopted from Lewis et al. (2019).

Methods	XQuAD (EM / F1)					
	ar	hi	vi	el	ru	tr
Baseline	43.2 / 62.6	46.0 / 63.1	48.7 / 70.4	49.4 / 68.5	55.2 / 72.3	44.1 / 63.1
mixMRC	42.4 / 63.6	50.0 / 66.2	52.7 / 72.6	51.1 / 72.1	58.7 / 75.9	47.8 / 65.8
Ours-hyper	**44.5 / 65.0**	52.0 / 67.4	55.5 / 74.6	52.2 / 73.1	59.3 / 76.6	**50.8 / 68.3**
Ours-imp	44.0 / 64.6	**52.5 / 67.9**	**55.6 / 74.9**	**52.4 / 73.3**	**59.6 / 76.6**	50.2 / 67.7

Table 3: EM and F1 score of 6 languages on the XQuAD dataset under the zero-shot transfer setting.

Junhao Liu, Linjun Shou, Jian Pei, Ming Gong, Min Yang, and Daxin Jiang. **Cross-lingual Machine Reading Comprehension with Language Branch Knowledge Distillation.** COLING. 2020

B

Microsoft

Reinforced Multi-Teacher Selection for Knowledge Distillation (RL-KD)

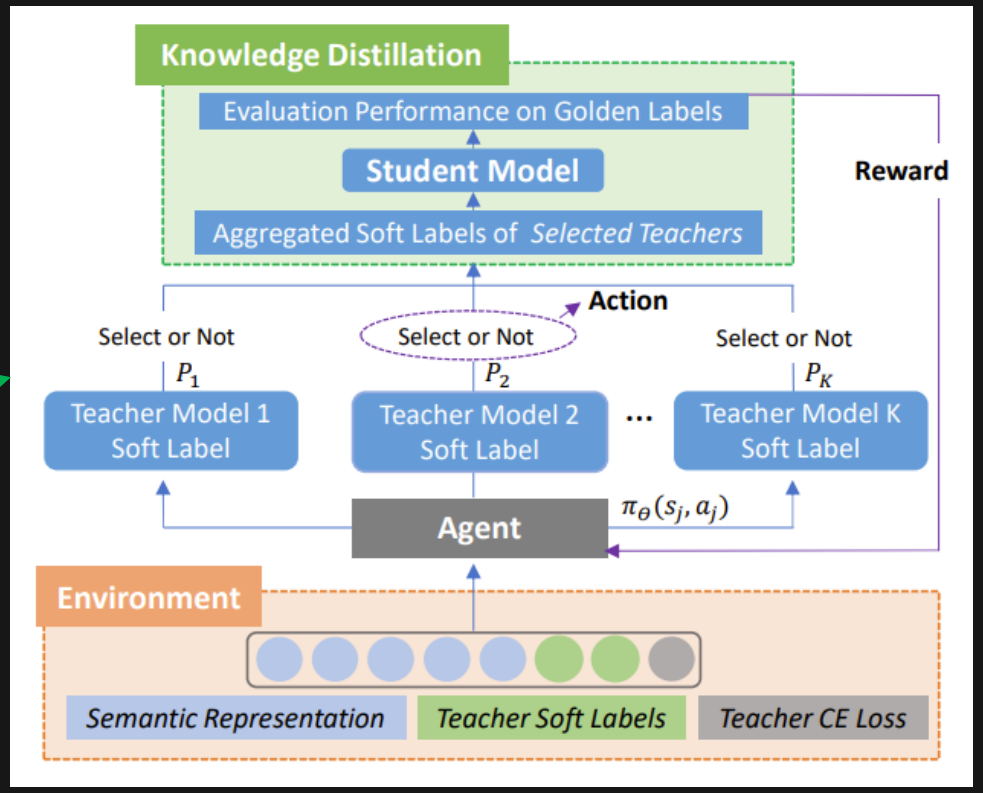
- **Motivation: we have multiple teachers, how to select best teacher model for each case to generate soft label for better KD?**

Baseline Methods:

- Vanilla KD from teacher ensemble with static weights

New Approach:

- RL method to select from multiple teacher models at instance level



Fei Yuan, Linjun Shou, Jian Pei, Wutao Lin, Ming Gong, and Daxin Jiang. **Reinforced Multi-Teacher Selection for Knowledge Distillation**. AAAI 2021.

B

Microsoft

Reinforced Multi-Teacher Selection for Knowledge Distillation (RL-KD)

- **Motivation: we have multiple teachers, how to select best teacher model for each case to generate soft label for better KD?**

Baseline Methods:

- Vanilla KD from teacher ensemble with static weights

New Approach:

- RL method to select from multiple teacher models at instance level

Teacher	Student	Strategy	QQP	MRPC	MNLI-(mm/m)	RTE	QNLI	SST-2	AVG.
-	$BERT_3$	FT	88.3	72.1	74.6 / 74.8	60.7	83.3	85.9	77.1
$BERT_{12}$	$BERT_3$	V-KD	88.4	74.0	75.3 / 75.6	56.7	83.7	87.5	77.3
$Robert_{12}$	$BERT_3$	V-KD	85.3	73.0	71.9 / 71.7	55.2	81.9	86.0	75.0
$XLNet_{12}$	$BERT_3$	V-KD	85.4	74.3	72.0 / 71.6	55.2	81.4	87.4	75.3
$ALBERT_{12}$	$BERT_3$	V-KD	88.4	74.0	76.4 / 75.6	56.7	83.7	87.5	77.5
Rand-Single-Ensemble	$BERT_3$	V-KD	87.3	70.6	75.0 / 74.5	51.6	83.7	86.0	75.5
W-Ensemble	$BERT_3$	V-KD	85.3	74.8	72.2 / 72.0	56.7	82.4	87.5	75.8
LR-Dev-Ensemble	$BERT_3$	V-KD	88.6	71.8	75.4 / 75.4	54.9	84.2	86.8	76.7
Best-Single-Ensemble	$BERT_3$	V-KD	88.6	**77.2**	75.1 / 74.7	56.0	84.1	87.0	77.5
Our Method ($reward_1$)	$BERT_3$	RL-KD	**89.1**	76.0	76.9 / **76.8**	61.4	**85.4**	**89.1**	79.2
Our Method ($reward_2$)	$BERT_3$	RL-KD	**89.1**	76.2	**77.4** / 76.3	63.5	84.8	88.5	79.4
Our Method ($reward_3$)	$BERT_3$	RL-KD	89.0	76.7	76.7 / 75.7	**64.6**	85.3	**89.1**	**79.6**

(1)reward1: use the minus of ground-truth loss (CE) of student model as the reward function for teacher model selection.
(2)reward2: besides (1), also introduce the minus of knowledge distillation loss (DL) into the reward function.
(3)reward3: besides (2), also take the Accuracy metric of student model on dev set into account for better generalization

Fei Yuan, Linjun Shou, Jian Pei, Wutao Lin, Ming Gong, and Daxin Jiang. **Reinforced Multi-Teacher Selection for Knowledge Distillation**. AAAI 2021.

B

Microsoft

Improve Span Alignment for MT: Back-translation + Verification

- **Motivation:** the answer spans after translation may not exist in translated passage or have boundary errors.

- Translate source span to target span and further improve boundary by:
  1. *Simple Match*
  2. *Answer Aligner*
  3. *Answer Verifier*

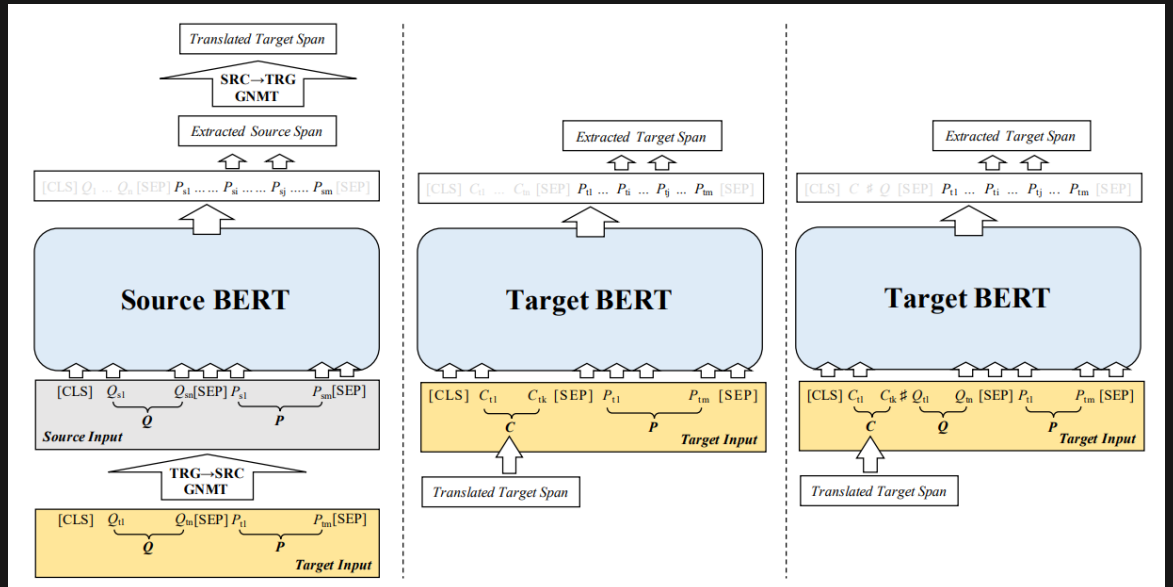


Figure 1: Back-translation approaches for cross-lingual machine reading comprehension (Left: GNMT, Middle: Answer Aligner, Right: Answer Verifier)

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, Guoping Hu. **Cross-Lingual Machine Reading Comprehension.** EMNLP 2019.

Improve Span Alignment for MT: Back-translation + Verification

Experiment Results using different answer span alignment approaches

Without any alignment approaches

#	System	CMRC 2018 Dev EM	CMRC 2018 Dev F1	CMRC 2018 Test EM	CMRC 2018 Test F1	CMRC 2018 Challenge EM	CMRC 2018 Challenge F1	DRCD Dev EM	DRCD Dev F1	DRCD Test EM	DRCD Test F1
1	GNMT+BERT$_{SQ-B_{en}}$ ♠	15.9	40.3	20.8	45.4	4.2	20.2	28.1	50.0	26.6	48.9
2	GNMT+BERT$_{SQ-L_{en}}$ ♠	16.8	42.1	21.7	47.3	5.2	22.0	28.9	52.0	28.7	52.1
3	GNMT+BERT$_{SQ-L_{en}}$+SimpleMatch ♠	26.7	56.9	31.3	61.6	9.1	35.5	36.9	60.6	37.0	61.2
4	GNMT+BERT$_{SQ-L_{en}}$+Aligner	46.1	66.4	49.8	69.3	16.5	40.9	60.1	70.5	59.5	70.7
5	GNMT+BERT$_{SQ-L_{en}}$+Verifier	64.7	84.7	68.9	86.8	20.0	45.6	83.5	90.1	82.6	89.6

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, Guoping Hu. **Cross-Lingual Machine Reading Comprehension.** EMNLP 2019.

Advanced Approaches for CLMRC

Microsoft

Challenges of the Baseline Approach

A CL pretrain models don't have sufficient transfer capability for phrase boundary tasks

B Machine Translation Data Quality

- MT noises impact more for pair-wise tasks
- Answer span alignment is challenging after MT

C Machine translation data is not available in some cases or quality is bad for tail languages

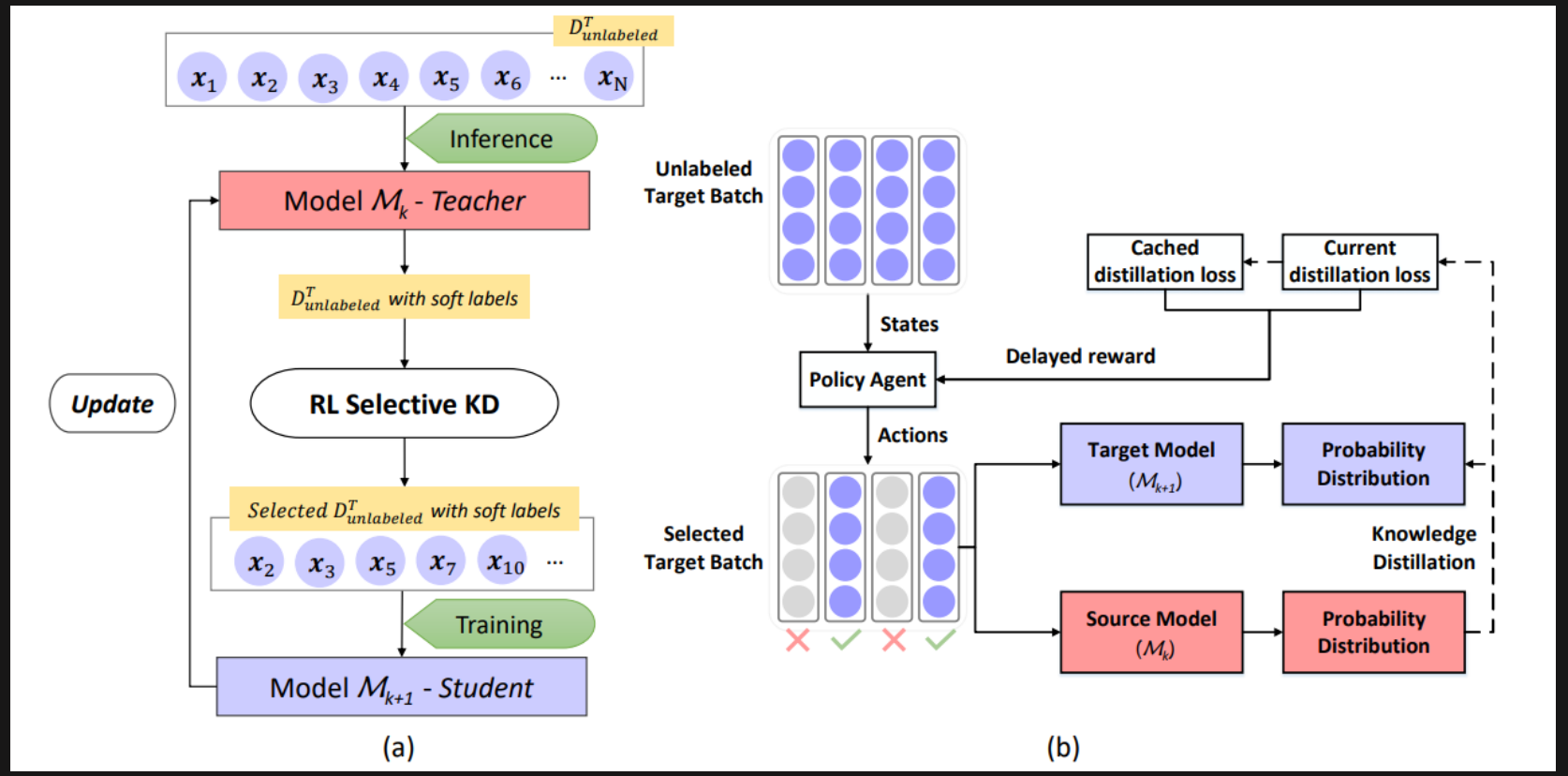
→ **Data transfer**

- Unlabeled language specific data
- Generated language specific data

Answer span boundary detection leads to the major errors of CLMRC model prediction

Unlabeled Language Specific Data + Iterative Knowledge Distillation

- **Motivation:** real language specific data (e.g. <question, passage> pairs) without label is relatively easy to get in many application scenarios. Could we generate good labels (e.g. answer spans) for them?



Shining Liang, Ming Gong, Jian Pei, Linjun Shou, Wanli Zuo, Xianglin Zuo, Daxin Jiang. Reinforced Iterative Knowledge Distillation for Cross-Lingual Named Entity Recognition. KDD 2021.

Unlabeled Language Specific Data + Iterative Knowledge Distillation

- Experiment Results

(a) Results on CoNLL.

Method	es	nl	de	Average
Täckström [41]	59.30	58.40	40.40	52.70
Tsai et al. [6]	60.55	61.56	48.12	56.74
Ni et al. [14]	65.10	65.40	58.50	63.00
Mayhew et al. [25]	65.95	66.50	59.11	63.85
Xie et al. [16]	72.37	71.25	57.76	67.13
Wu and Dredze [43]†	74.50	79.50	71.10	75.03
Moon [36]†	75.67	80.38	71.42	75.82
Wu et al. [21]	76.75	80.44	73.16	76.78
Wu et al. [22] (mBERT)	76.94	80.89	73.22	77.02
RIKD (mBERT)	**77.84**¶	**82.46**¶	**75.48**¶	**78.59**¶
Wu et al. [22] (XLM-R)	78.77	80.99	74.67	78.14
RIKD (XLM-R)	**79.46**¶	**81.40**¶	**78.40**¶	**79.75**¶
Täckström [41] ‡	61.90	59.90	36.40	52.73
Moon [36] ‡	76.53	83.35	72.44	77.44
Wu et al. [22] ‡	78.00	81.33	75.33	78.22

(b) Results on WikiAnn.

Method	ar	hi	zh	Average
Wu and Dredze [44]	42.30	67.60	52.90	54.27
Wu et al. [22] (mBERT)	43.12	69.54	48.12	53.59
RIKD (mBERT)	**45.96**¶	**70.28**¶	**50.40**¶	**55.55**¶
Pfeiffer et al. [15]	41.80	-	20.50	-
Wu and Dredze [44]	45.50	66.60	43.90	52.00
Wu et al. [22] (XLM-R)	50.91	72.48	31.14	51.51
RIKD (XLM-R)	**54.46**¶	**74.42**¶	**37.48**¶	**55.45**¶

Shining Liang, Ming Gong, Jian Pei, Linjun Shou, Wanli Zuo, Xianglin Zuo, Daxin Jiang. Reinforced Iterative Knowledge Distillation for Cross-Lingual Named Entity Recognition. KDD 2021.

Generated Language Specific Data for Model Enhancement

- **Motivation:** Language specific plain text data (i.e. passages) is easier to get. Could we generate both questions and labels (i.e. answer spans) for them?

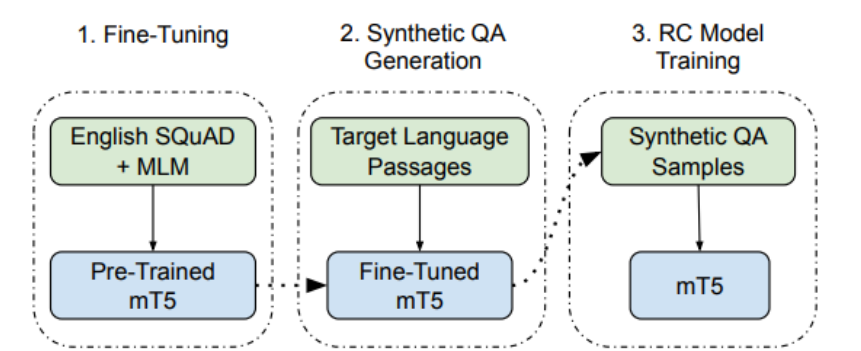


Figure 1: End-to-End pipeline: 1) Fine-tuning the generative model using SQuAD English samples and multilingual MLM. 2) Generating synthetic samples from Wikipedia passages of the target language using the fine-tuned generative model. 3) Training the downstream reading comprehension model using synthetic samples.

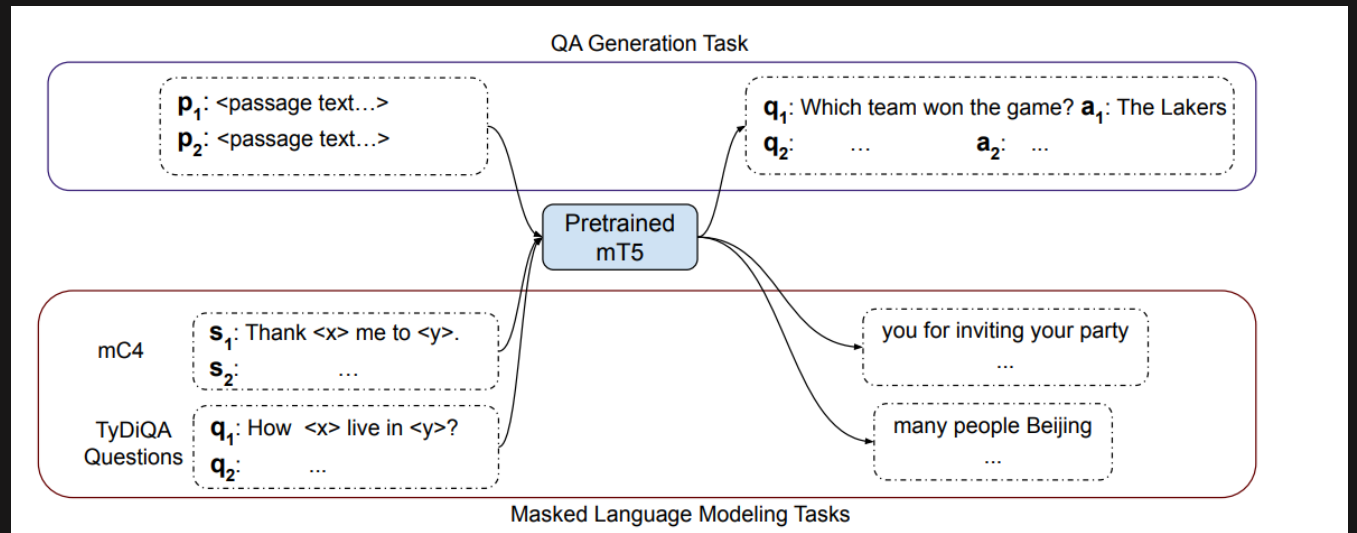


Figure 2: Multi-task fine-tuning of the multilingual pre-trained mT5 model: 1) QA generation task, which uses SQuAD English samples, 2) MLM task on a subset of mC4, 3) MLM on the questions from the TyDiQA Gold Passage Task. The MLM variant used does not include sentinel tokens in the decoder output.

- Siamak Shakeri, Noah Constant, Mihir Sanjay Kale, Linting Xue. Towards Zero-Shot Multilingual Synthetic Question and Answer Generation for Cross-Lingual Reading Comprehension. Arxiv 2021.
- Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, Jacopo Staiano. Synthetic Data Augmentation for Zero-Shot Cross-Lingual Question Answering, arxiv 2020

Generated Language Specific Data for Model Enhancement

- Experiment Results

Dataset	en	ar	de	el	es	hi	ru	th	tr	vi	zh	avg
SQuAD en (paper)	88.4	75.2	80.0	77.5	81.8	73.4	74.7	73.4	76.5	79.4	75.9	77.8
SQuAD en (ours)	88.6	75.0	80.4	76.5	81.6	73.9	74.1	73.8	**76.2**	**80.1**	76.4	77.4
SQuAD en + ru	88.2	76.6	81.2	79.1	82.6	76.1	77.6	72.1	75.1	78.4	77.4	**78.6**
SQuAD en + hi	**88.9**	76.7	81.5	79.4	**82.9**	73.4	78.7	**74.1**	75.0	79.1	78.0	**78.6**
SQuAD en + de	88.0	72.7	79.7	73.0	82.0	73.6	76.4	71.6	74.8	78.7	76.2	76.6
SQuAD en + ar	88.0	73.3	81.2	78.8	82.4	75.1	78.5	71.4	75.6	77.3	**78.2**	77.8
SQuAD en + es	88.2	**77.2**	**81.8**	**79.9**	81.3	**76.4**	**79.2**	72.3	75.8	79.5	77.7	**78.7**
Supervised	87.3	79.4	82.7	81.8	83.8	78.0	81.9	74.7	80.2	80.4	83.2	81.2

Table 5: Performance of fine-tuned mT5 Large models on XQuAD. *Supervised* refers to training on SQuAD en + `translate-train` dataset of the target language.

Towards Zero-Shot Multilingual Synthetic Question and Answer Generation for Cross-Lingual Reading Comprehension. Arxiv 2021.

Advanced Approaches for CLMRC

Microsoft

Challenges of the Baseline Approach

A CL pretrain models don't have sufficient transfer capability for phrase boundary tasks

B Machine Translation Data Quality

- MT noises impact more for pair-wise tasks
- Answer span alignment is challenging after MT

C Machine translation data is not available in some cases or quality is bad for tail languages

D Answer span boundary detection leads to the major errors of CLMRC model prediction

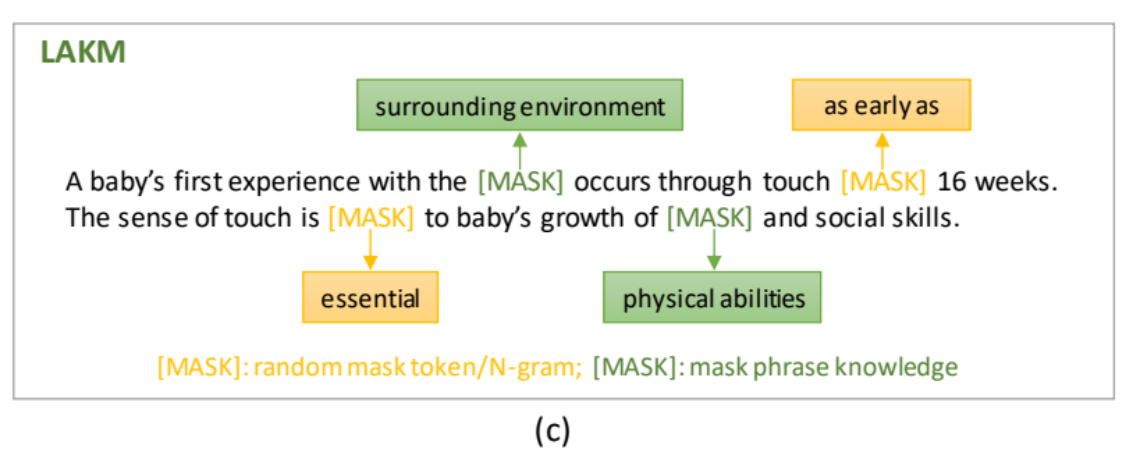
➡

Model transfer **Data transfer**

- New LAKM task + Language specific phrase knowledge data
- Calibration network for primary boundary correction

Enhance boundary detection: LAKM task + Language specific phrase data

- **Language Agnostic Knowledge Masking** task (LAKM) targets introducing language specific boundary prior knowledge into the model.



(c)

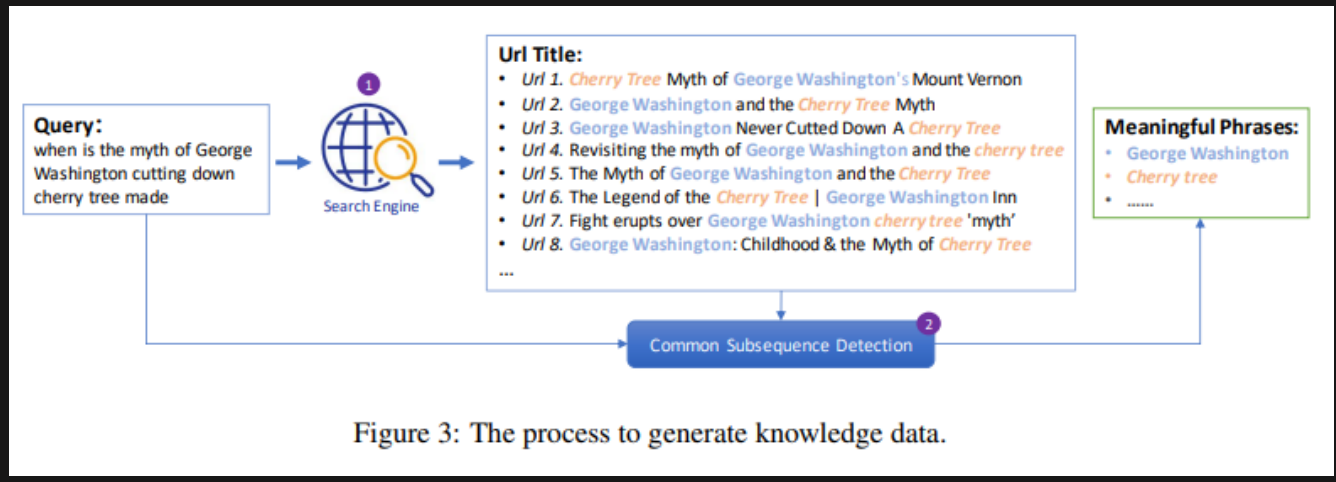


Figure 3: The process to generate knowledge data.

- As an auxiliary task of the main MRC task in the fine-tuning stage (i.e. multi-task training)

- Phrase Candidates Generation (for recall)
- Phrase Filtering (for precision)

Fei Yuan, Linjun Shou, Xuanyu Bai, Ming Gong, Yaobo Liang, Nan Duan, Yan Fu, Daxin Jiang. Enhancing Answer Boundary Detection for Multilingual Machine Reading Comprehension. ACL, 2020.

Enhance boundary detection: LAKM task + Language specific phrase data

- Experiment Results

	MLQA (EM / F1)		
	en	es	de
Baseline	65.2 / 77.7	46.6 / 64.3	44.3 / 57.9
LAKM	**66.8 / 80.0**	**48.0 / 65.9**	**45.5 / 60.5**

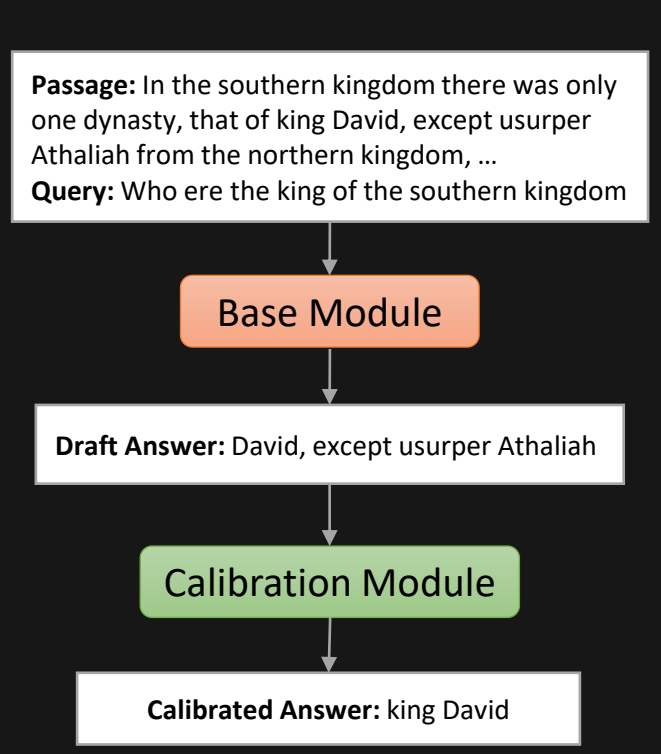
	MTQA (EM / F1)		
	en	fr	de
Baseline	65.8 / 86.6	41.3 / 70.9	50.7 / 76.2
LAKM	**67.8 / 87.2**	**44.6 / 72.1**	**54.5 / 77.8**

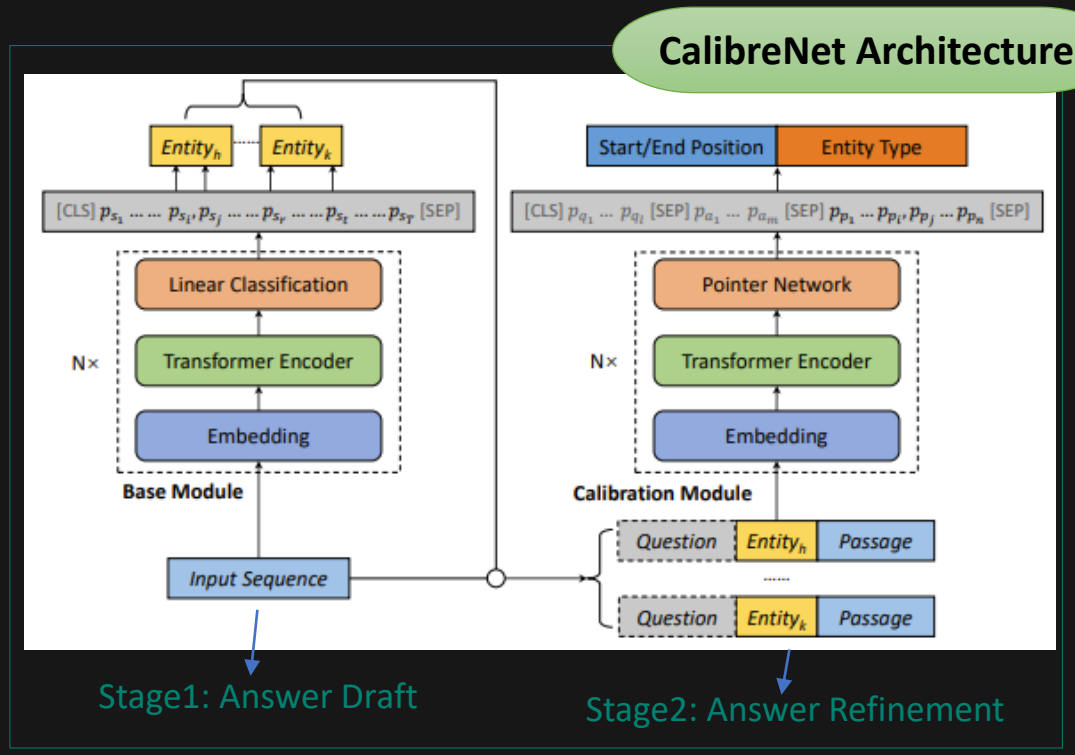
Table 9: Zero Shot experimental results on MLQA and MTQA datasets (%). We only use English MRC training data and don't use translation data.

Fei Yuan, Linjun Shou, Xuanyu Bai, Ming Gong, Yaobo Liang, Nan Duan, Yan Fu, Daxin Jiang. Enhancing Answer Boundary Detection for Multilingual Machine Reading Comprehension. ACL, 2020.

Post-processing for Boundary Refinement: CalibreNet

- Propose a Two-stage Approach inspired by how human write articles
  - Answer draft (*base module*) → Answer refinement (*calibration module*)



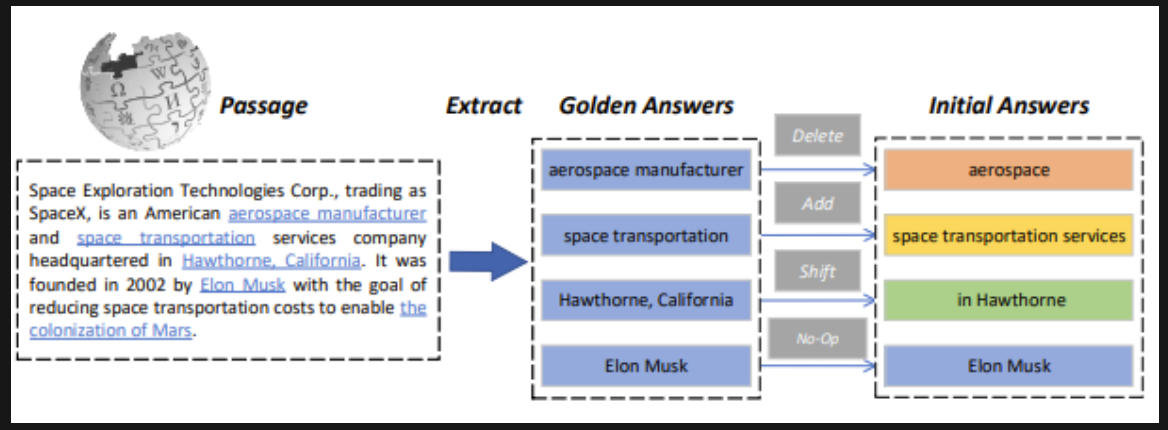


Challenge: how to collect training data for Calibration module, especially non-EN languages?

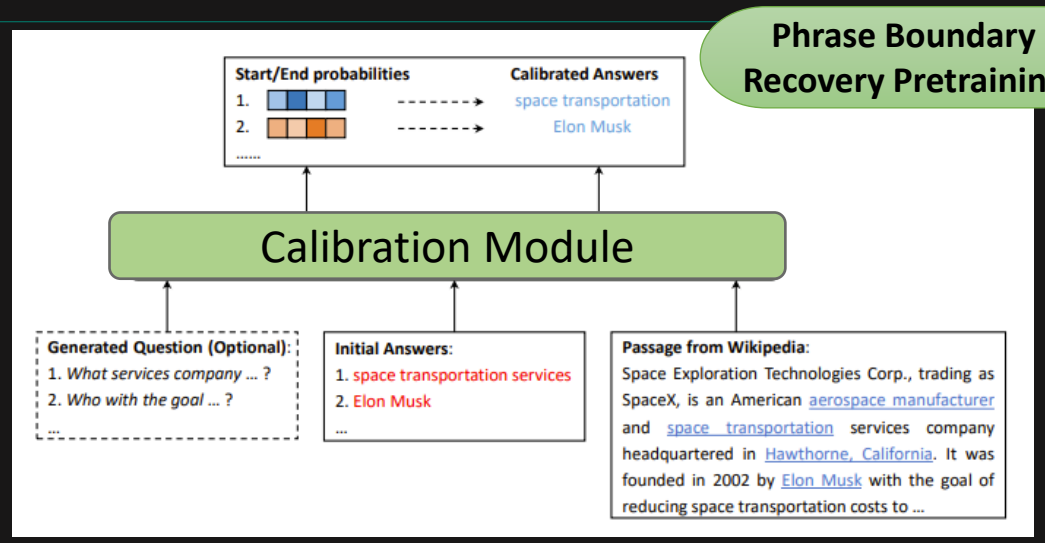
Shining Liang, Linjun Shou, Jian Pei, Ming Gong, Wanli Zuo, and Daxin Jiang. CalibreNet: CalibrationNetworks for Multilingual Sequence Labeling. WSDM. 2021.

Post-processing for Boundary Refinement: CalibreNet

- Design a novel **Phrase Boundary Recovery task** (PBR) to pretrain the calibration module on *large-scale data synthesized from Wikipedia docs* in multiple languages



Synthesize a large number of noisy entity spans by garbling the true entity phrases from Wikipedia docs



Wikipedia hyperlinks as the ground truth for **denoising pretraining**

Results on MLQA (F1/EM)

Model	Methods	en	es	de	ar	hi	vi	Avg.
XLM R_{base}	$MLQA_{org}$	77.80/65.30	67.20/49.70	60.80/47.10	53.00/34.70	57.90/41.70	63.10/43.10	63.30/46.90
	$MLQA_{re\text{-}imp}$	79.21/65.96	67.73/49.94	60.92/46.31	55.76/36.89	59.58/42.92	66.30/45.64	64.92/47.94
	CalibreNet ($MLQA_{re\text{-}imp}$)	79.68/66.51	68.04/50.77	61.66/47.55	56.14/37.83	59.97/43.84	66.92/46.59	**65.40/48.84**¶

Shining Liang, Linjun Shou, Jian Pei, Ming Gong, Wanli Zuo, and Daxin Jiang. CalibreNet: CalibrationNetworks for Multilingual Sequence Labeling. WSDM. 2021.

Key Take-aways

Microsoft

- MRC is more challenging in language scaling and the baseline approaches (CL pretrained model + MT data) don't work well due to

A CL pretrain models don't have sufficient transfer capability for phrase boundary tasks

B Machine Translation Data Quality

1. MT noises impact more for pair-wise tasks
2. Answer span alignment is challenging after MT

C Machine translation data is not available in some cases or quality is bad for tail languages

D Answer span boundary detection leads to the major errors of CLMRC model prediction

Advanced Approaches

	Model transfer	Data transfer
→	• New Mix-MRC task • Dual encoders • Adversarial training	+ MT data
→	• Multiple teacher models + Language branch MT data • Instance level best teacher selection through reinforcement learning • Back-translation & Verification to improve answer span alignment	
→	• Unlabeled language specific data • Generated language specific data	
→	• New LAKM task + Language specific phrase knowledge data • Calibration network for primary boundary correction	

Outline

Daxin Jiang
Software Technology Center at Asia (STCA) of Microsoft

- Introduction [Daxin Jiang]
  - Motivating examples in Microsoft products
  - Problem description
  - Categorization of applications
  - Challenges
- Methodology [Daxin Jiang]
  - Model Transfer
  - Data Transfer
- Applications*
  - Dependency Parsing [Xiubo Geng]
  - Machine Reading Comprehension [Ming Gong]
  - Grammar Error Correction [Linjun Shou]
- Summary & Future directions [Jian Pei]

Linjun Shou Xiubo Geng Ming Gong
Software Technology Center at Asia (STCA) of Microsoft

Jian Pei
Simon Fraser University

*For more applications, please refer to our tutorial at The Web Conference 2021

Microsoft

Grammar Error Correction & Question Generation

Linjun Shou
lisho@microsoft.com

Cross-lingual GEC & QG

Type	Category	Sub Category	Example
NLU	Text Classification	Single text	Domain identification, Intent detection, Sentiment classification
		Text pair	Information retrieval, Natural language inference
	Sequence Labeling	Single text	Named entity recognition, Slot tagging
		Text pair	Machine reading comprehension
	Structure Prediction	Single text	Dependency parsing, semantic role labeling
NLG	Text Generation	Token level	Spelling correction, Sentence auto completion, **Grammar error correction**
		Sentence level	Machine translation, Conversation, **Question generation**

Grammar Error Correction

Grammatical Error Correction (GEC) is the task of correcting different kinds of errors in text such as spelling, punctuation, grammatical, and word choice errors.

Input (Erroneous)	**Output (Corrected)**
She see Tom is catched by policeman in park at last night.	She saw Tom caught by a policeman in the park last night.

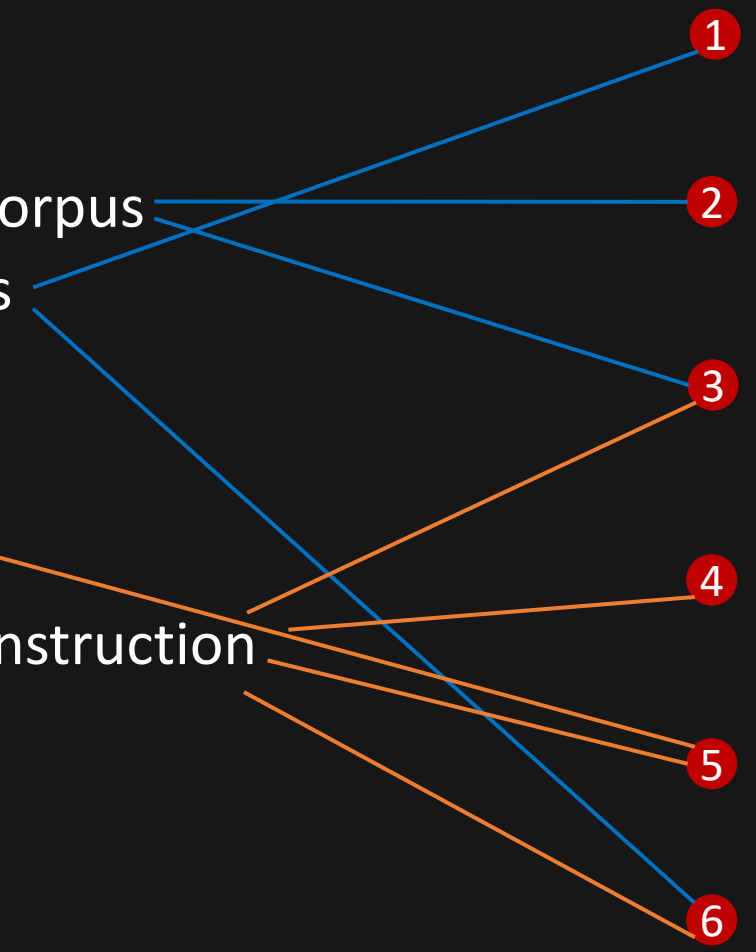
Question Generation

Question Generation (QG) is the task of generating a question based on passage context and an input answer.

Input Passage	… In case of European Union law which should have been transposed into the laws of member stages, such Directives, the European Commission can take proceedings against the member stage under the *Treaty on the Functioning of the European Union*, …
Input Answer	Treaty on the Functioning of the European Union
Reference	Under what treaty can the European Commission take actions against member states?

Approaches for Cross-lingual Transfer

- **Model Transfer**
  - w/ monolingual corpus
  - w/ parallel corpus
- **Data Transfer**
  - Back translation
  - Synthetic data construction



Representative Works

Cross-Lingual Natural Language Generation via Pre-Training

Zewen Chi[†*], Li Dong[‡], Furu Wei[‡], Wenhui Wang[‡], Xian-Ling Mao[†], Heyan Huang[†]
[†]Beijing Institute of Technology
[‡]Microsoft Research
{czw,maoxl,hhy63}@bit.edu.cn
{lidong1,fuwei,Wenhui.Wang}@microsoft.com

Multilingual Denoising Pre-training for Neural Machine Translation

Yinhan Liu[‡*], Jiatao Gu[†*], Naman Goyal[†*], Xian Li[†], Sergey Edunov[†],
Marjan Ghazvininejad[†], Mike Lewis[†], and Luke Zettlemoyer[†]

mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer

Linting Xue* Noah Constant* Adam Roberts*
Mihir Kale Rami Al-Rfou Aditya Siddhant Aditya Barua Colin Raffel
Google Research

A Simple Recipe for Multilingual Grammatical Error Correction

Sascha Rothe
Google
rothe@google.com
Jonathan Mallinson
Google
jonmall@google.com
Eric Malmi
Google
emalmi@google.com
Sebastian Krause
Google
bastik@google.com
Aliaksei Severyn
Google
severyn@google.com

An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction

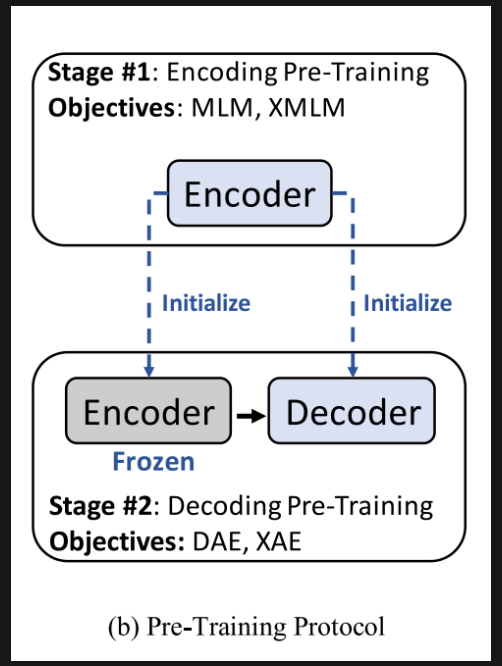
Shun Kiyono[1,2] Jun Suzuki[2,1] Masato Mita[1,2] Tomoya Mizumoto[1,2*] Kentaro Inui[2,1]
[1] RIKEN Center for Advanced Intelligence Project [2] Tohoku University
{shun.kiyono, masato.mita, tomoya.mizumoto}@riken.jp;
{jun.suzuki,inui}@ecei.tohoku.ac.jp

Cross-lingual Transfer Learning for Grammatical Error Correction

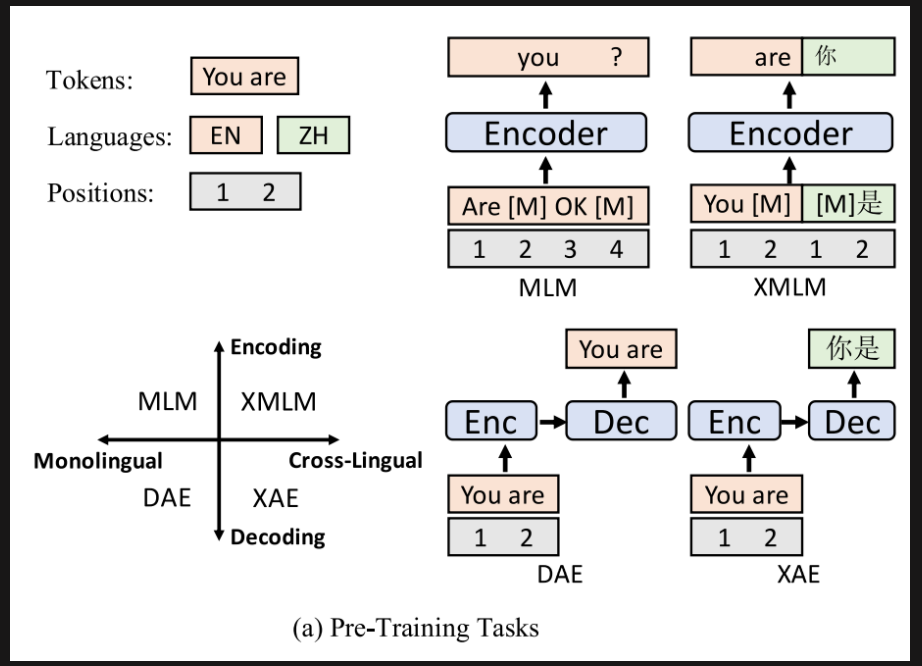
Ikumi Yamashita Satoru Katsumata* Masahiro Kaneko
Aizhan Imankulova Mamoru Komachi
Tokyo Metropolitan University, Japan
yamashita-ikumi@ed.tmu.ac.jp
satoru.katsumata@retrieva.jp
{kaneko-masahiro, imankulova-aizhan}@ed.tmu.ac.jp
komachi@tmu.ac.jp

Model Transfer - XNLG

- **Model Transfer**
  - w/ monolingual corpus
  - w/ parallel corpus



(b) Pre-Training Protocol

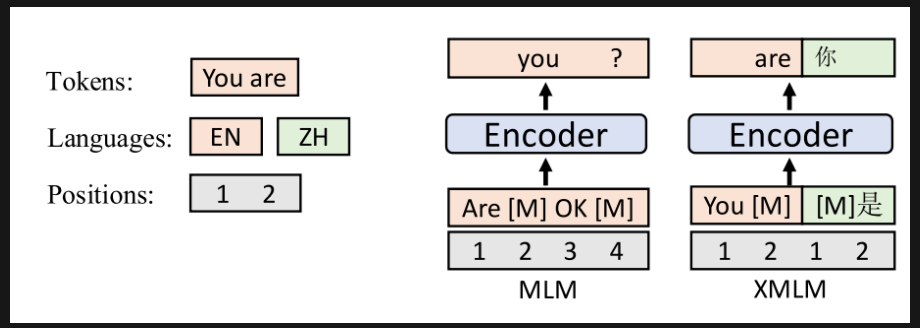


(a) Pre-Training Tasks

Chi, Z., Dong, L., Wei, F., Wang, W., Mao, X., & Huang, H. (2020). Cross-Lingual Natural Language Generation via Pre-Training. *ArXiv, abs/1909.10481*.

Model Transfer - XNLG

- **Encoder Pretraining**
  - Monolingual data
  - Parallel data



$$\mathcal{L}_{\text{MLM}}^{(x)} = -\sum_{i \in M_x} \log p(x_i | x_{\setminus M_x})$$

$$\mathcal{L}_{\text{XMLM}}^{(x,y)} = -\sum_{i \in M_x} \log p(x_i | x_{\setminus M_x}, y_{\setminus M_y}) \quad (2)$$

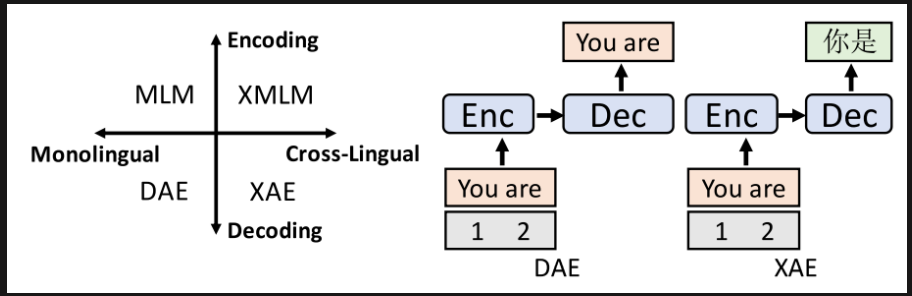
$$-\sum_{i \in M_y} \log p(y_i | x_{\setminus M_x}, y_{\setminus M_y}) \quad (3)$$

where M_x, M_y represent the masked positions of x and y, respectively.

Chi, Z., Dong, L., Wei, F., Wang, W., Mao, X., & Huang, H. (2020). Cross-Lingual Natural Language Generation via Pre-Training. *ArXiv, abs/1909.10481*.

Model Transfer - XNLG

- **Decoder Pretraining (w/ Encoder fixed)**
  - Monolingual data
  - Parallel data



$$\mathcal{L}_2 = \sum_{(x,y)\in\mathcal{D}_\mathrm{p}} \mathcal{L}_\mathrm{XAE}^{(x,y)} + \sum_{x\in\mathcal{D}_\mathrm{m}} \mathcal{L}_\mathrm{DAE}^{(x)}$$

$$\mathcal{L}_\mathrm{XAE}^{(x,y)} = -\log p(y|x) - \log p(x|y)$$

- Word order is locally shuffled
- Randomly drop tokens with probability = 0.1
- substitute tokens with the special padding token [P] with a probability of 0.1.

Chi, Z., Dong, L., Wei, F., Wang, W., Mao, X., & Huang, H. (2020). Cross-Lingual Natural Language Generation via Pre-Training. *ArXiv, abs/1909.10481*.

Model Transfer - XNLG

- **Results on Zero-shot QG**

Models	BL-4	MTR	RG-L
XLM	0.25	0.62	2.56
PIPELINE (XLM)	4.42	9.59	21.22
w/ Google Translator	9.95	14.92	29.37
XNLG	**16.37**	**18.74**	**34.93**

Table 2: Evaluation results of zero-shot Chinese-Chinese question generation. Same shorthands apply as in Table 1.

Models	Rel	Flu	Corr
XLM	0	0	0
PIPELINE (XLM)	0.50	0.80	0.03
w/ Google Translator	1.31	**1.43***	0.69
XNLG	**1.68***	1.29	**0.89***

Table 3: Human evaluation results of zero-shot Chinese-Chinese question generation. Rel is short for relatedness, Flu for fluency, and Corr for correctness. "*" indicates the improvements are significant at $p < 0.05$.

Chi, Z., Dong, L., Wei, F., Wang, W., Mao, X., & Huang, H. (2020). Cross-Lingual Natural Language Generation via Pre-Training. *ArXiv, abs/1909.10481*.

Model Transfer – mBART

- **Model Transfer**
  - w/ monolingual corpus

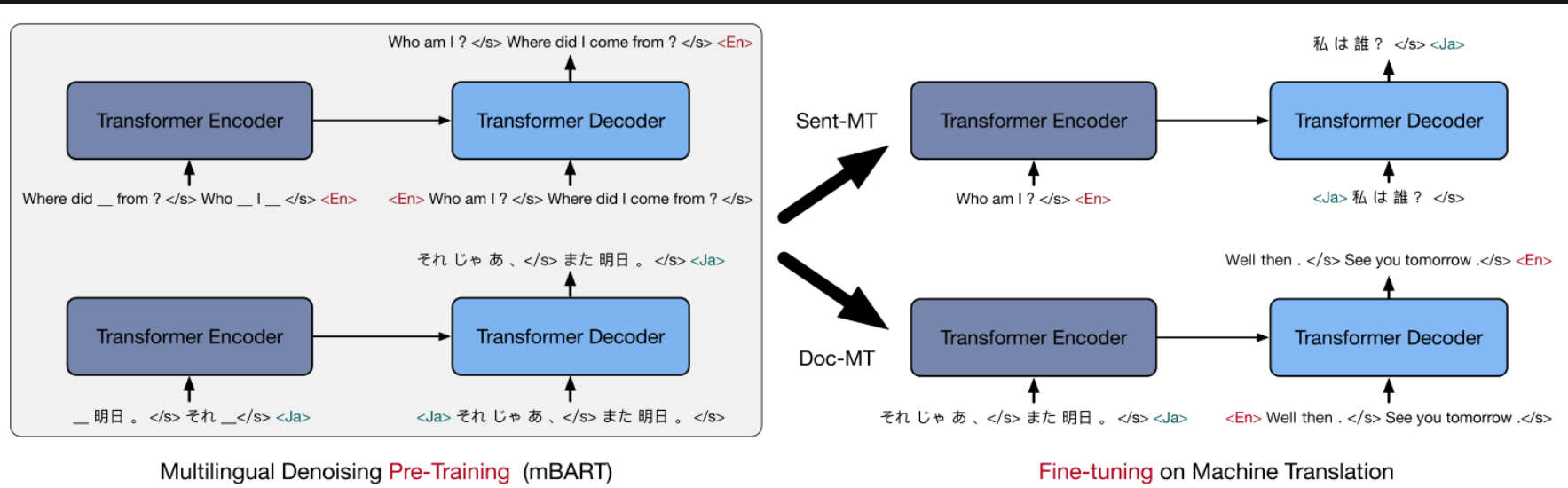


Figure 2: Framework for our multilingual denoising pre-training (left) and fine-tuning on downstream MT tasks (right), where we use (1) sentence permutation and (2) word-span masking as the injected noise. A special language id token is added at both the encoder and decoder. One multilingual pre-trained model is used for all tasks.

Liu, Yinhan et al. "Multilingual Denoising Pre-training for Neural Machine Translation." Transactions of the Association for Computational Linguistics 8 (2020): 726-742.

Model Transfer – mBART

- **Model Transfer**
  - w/ monolingual corpus

$$\mathcal{L}_\theta = \sum_{\mathcal{D}_i \in \mathcal{D}} \sum_{X \in \mathcal{D}_i} \log P(X|g(X);\theta)$$

Noise function *g(X)*:

- mask 35% of the words in each instance by random sampling a span length according to a Poisson distribution (λ = 3.5)
- Permute the order of sentences within each instance.

Code	Language	Tokens/M	Size/GB
En	English	55608	300.8
Ru	Russian	23408	278.0
Vi	Vietnamese	24757	137.3
Ja	Japanese	530 (*)	69.3
De	German	10297	66.6
Ro	Romanian	10354	61.4
Fr	French	9780	56.8
Fi	Finnish	6730	54.3
Ko	Korean	5644	54.2
Es	Spanish	9374	53.3
Zh	Chinese (Sim)	259 (*)	46.9
It	Italian	4983	30.2
Nl	Dutch	5025	29.3
Ar	Arabic	2869	28.0
Tr	Turkish	2736	20.9
Hi	Hindi	1715	20.2
Cs	Czech	2498	16.3
Lt	Lithuanian	1835	13.7
Lv	Latvian	1198	8.8
Kk	Kazakh	476	6.4
Et	Estonian	843	6.1
Ne	Nepali	237	3.8
Si	Sinhala	243	3.6
Gu	Gujarati	140	1.9
My	Burmese	56	1.6

Table 1: **Languages and Statistics of the CC25 Corpus.** A list of 25 languages ranked with monolingual corpus size. Throughout this paper, we replace the language names with their ISO codes for simplicity. (*) Chinese and Japanese corpus are not segmented, so the tokens counts here are sentences counts

Liu, Yinhan et al. “Multilingual Denoising Pre-training for Neural Machine Translation.” Transactions of the Association for Computational Linguistics 8 (2020): 726-742.

Model Transfer – mT5

- **Model Transfer**
  - w/ monolingual corpus

Model	Architecture	Parameters	# languages	Data source
mBERT (Devlin, 2018)	Encoder-only	180M	104	Wikipedia
XLM (Conneau and Lample, 2019)	Encoder-only	570M	100	Wikipedia
XLM-R (Conneau et al., 2020)	Encoder-only	270M – 550M	100	Common Crawl (CCNet)
mBART (Lewis et al., 2020b)	Encoder-decoder	680M	25	Common Crawl (CC25)
MARGE (Lewis et al., 2020a)	Encoder-decoder	960M	26	Wikipedia or CC-News
mT5 (ours)	Encoder-decoder	300M – 13B	101	Common Crawl (mC4)

Table 1: Comparison of mT5 to existing massively multilingual pre-trained language models. Multiple versions of XLM and mBERT exist; we refer here to the ones that cover the most languages. Note that XLM-R counts five Romanized variants as separate languages, while we ignore six Romanized variants in the mT5 language count.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. NAACL.

Model Transfer – mT5

- Multi-lingual version of T5
- mC4 in 101 languages

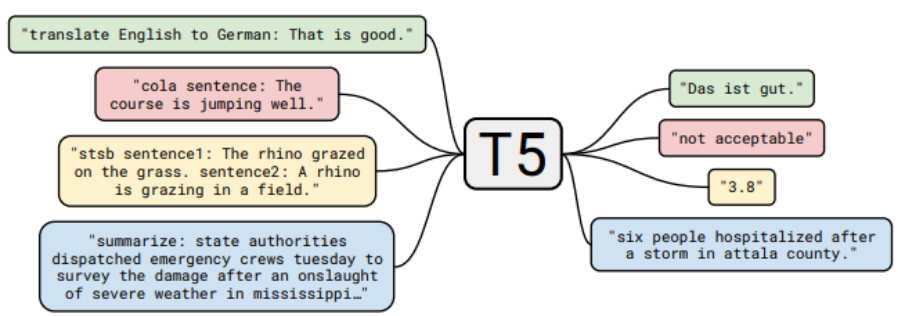


Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks. It also provides a standard testbed for the methods included in our empirical survey. "T5" refers to our model, which we dub the "**T**ext-**t**o-**T**ext **T**ransfer **T**ransformer".

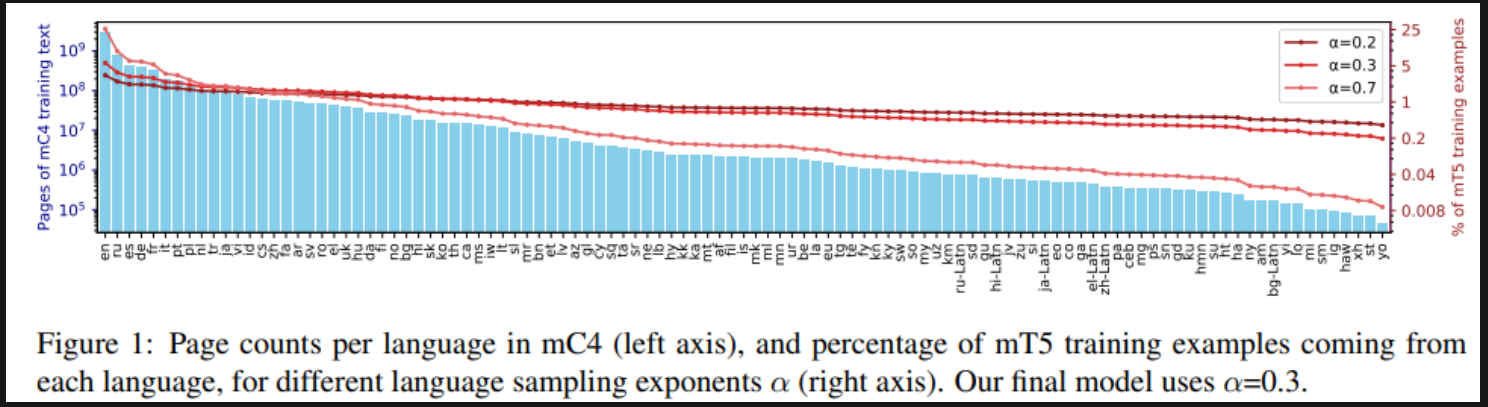


Figure 1: Page counts per language in mC4 (left axis), and percentage of mT5 training examples coming from each language, for different language sampling exponents α (right axis). Our final model uses α=0.3.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. NAACL.

Model Transfer – mT5

Pretraining Tasks:

Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style Devlin et al. (2018)	Thank you <M> <M> me to your party apple week .	*(original text)*
Deshuffling	party me for your to . last fun you inviting week Thank	*(original text)*
MASS-style Song et al. (2019)	Thank you <M> <M> me to your party <M> week .	*(original text)*
I.i.d. noise, replace spans	Thank you <X> me to your party <Y> week .	<X> for inviting <Y> last <Z>
I.i.d. noise, drop tokens	Thank you me to your party week .	for inviting last
Random spans	Thank you <X> to <Y> week .	<X> for inviting me <Y> your part

Objective	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
Prefix language modeling	80.69	18.94	77.99	65.27	**26.86**	39.73	**27.49**
BERT-style (Devlin et al., 2018)	**82.96**	**19.17**	**80.65**	**69.85**	**26.78**	**40.03**	**27.41**
Deshuffling	73.17	18.59	67.61	58.47	26.11	39.30	25.62

Table 4: Performance of the three disparate pre-training objectives described in Section 3.3.1.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. NAACL.

Model Transfer – mT5

- **Results**

Model	Sentence pair		Structured	Question answering		
	XNLI	PAWS-X	WikiAnn NER	XQuAD	MLQA	TyDiQA-GoldP
Metrics	Acc.	Acc.	F1	F1 / EM	F1 / EM	F1 / EM
Cross-lingual zero-shot transfer (models fine-tuned on English data only)						
mBERT	65.4	81.9	62.2	64.5 / 49.4	61.4 / 44.2	59.7 / 43.9
XLM	69.1	80.9	61.2	59.8 / 44.3	48.5 / 32.6	43.6 / 29.1
InfoXLM	81.4	-	-	- / -	73.6 / 55.2	- / -
X-STILTs	80.4	87.7	64.7	77.2 / 61.3	72.3 / 53.5	76.0 / 59.5
XLM-R	79.2	86.4	65.4	76.6 / 60.8	71.6 / 53.2	65.1 / 45.0
VECO	79.9	88.7	65.7	77.3 / 61.8	71.7 / 53.2	67.6 / 49.1
RemBERT	80.8	87.5	**70.1**	79.6 / 64.0	73.1 / 55.0	77.0 / 63.0
mT5-Small	67.5	82.4	50.5	58.1 / 42.5	54.6 / 37.1	35.2 / 23.2
mT5-Base	75.4	86.4	55.7	67.0 / 49.0	64.6 / 45.0	57.2 / 41.2
mT5-Large	81.1	88.9	58.5	77.8 / 61.5	71.2 / 51.7	69.9 / 52.2
mT5-XL	82.9	89.6	65.5	79.5 / 63.6	73.5 / 54.5	75.9 / 59.4
mT5-XXL	**85.0**	**90.0**	69.2	**82.5 / 66.8**	**76.0 / 57.4**	**80.8 / 65.9**
Translate-train (models fine-tuned on English data plus translations in all target languages)						
XLM-R	82.6	90.4	-	80.2 / 65.9	72.8 / 54.3	66.5 / 47.7
FILTER + Self-Teaching	83.9	91.4	-	82.4 / 68.0	76.2 / 57.7	68.3 / 50.9
VECO	83.0	91.1	-	79.9 / 66.3	73.1 / 54.9	75.0 / 58.9
mT5-Small	64.7	79.9	-	64.3 / 49.5	56.6 / 38.8	48.2 / 34.0
mT5-Base	75.9	89.3	-	75.3 / 59.7	67.6 / 48.5	64.0 / 47.7
mT5-Large	81.8	91.2	-	81.2 / 65.9	73.9 / 55.2	71.1 / 54.9
mT5-XL	84.8	91.0	-	82.7 / 68.1	75.1 / 56.6	79.9 / 65.3
mT5-XXL	**87.8**	**91.5**	-	**85.2 / 71.3**	**76.9 / 58.3**	**82.8 / 68.8**
In-language multitask (models fine-tuned on gold data in all target languages)						
mBERT	-	-	89.1	-	-	77.6 / 68.0
mT5-Small	-	-	83.4	-	-	73.0 / 62.0
mT5-Base	-	-	85.4	-	-	80.8 / 70.0
mT5-Large	-	-	88.4	-	-	85.5 / 75.3
mT5-XL	-	-	90.9	-	-	87.5 / 78.1
mT5-XXL	-	-	**91.2**	-	-	**88.5 / 79.1**

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. NAACL.

Model Transfer – mT5

Zero shot generation Challenges – Illegal predictions

- Normalization
- Grammatical adjustment
- Accidental translation

Target	Prediction	Explanation
จำนวนเฉพาะ	จำนวนเฉพาะ	Decomposed Thai ำ into ◌ั + า
लोयर डे माइज़ियर	लोयर डे माइज़ियर	Decomposed Hindi ज़ into ज + ◌़
27 - 30 %	27 - 30 %	Replaced full-width percent sign
12 . ª	12 . a	Removed superscript
للبكتريا اللاهوائية	البكتريا اللاهوائية	Arabic "for anaerobic bacteria" ⇒ "anaerobic bacteria"
строками битов	строки битов	Russian "bit strings (instrumental)" ⇒ "bit strings (nominative)"
seis años	six years	Translated from Spanish
Zweiten Weltkrieg	the Second World War	Translated from German
新英格兰爱国者队	New英格兰爱国者队	Partially translated Chinese "New England Patriots"
хлоропласт	chloroпласт	Partially translated Russian "chloroplast"

Table 5: Illegal mT5-XXL predictions on XQuAD zero-shot, illustrating normalization (top), grammatical adjustment (middle) and translation (bottom).

During fine-tuning stage, combine with pretraining task for multi-task learning

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. NAACL.

Data Transfer –Domain Adaptation

mT5 Pretraining

Input:	A Simple [x] Multilingual Grammatical Error [y]
Target:	[x] Recipe for [y] Correction

- Span-prediction task for Seq2Seq pretraining
- mC4 data -> 50 billion documents in 101 languages

GEC Pretraining

Input:	Simple recipe for Multingual Grammatical Correction Error
Target:	A Simple Recipe for Multilingual Grammatical Error Correction

- Adapt to GEC task
- Denoising pretraining
  - *Drop spans of tokens*
  - *Swap tokens*
  - *Drop spans of characters*
  - *Swap characters*
  - *Insert*
  - *Lower/Upper case of words*

GEC Fine-tuning

Rothe, Sascha et al. “A Simple Recipe for Multilingual Grammatical Error Correction.” ArXiv abs/2106.03830 (2021): n. pag.

Data Transfer - Domain Adaptation

Models	CoNLL-14	BEA test	Czech	German	Russian
Omelianchuk et al.*	66.5	**73.6**	-	-	-
Lichtarge et al.*	**66.8**	73.0	-	-	-
Náplava and Straka	63.40	69.00	80.17	73.71	50.20
Katsumata and Komachi*	63.00	66.10	73.52	68.86	44.36
gT5 base	54.10	60.2	71.88	69.21	26.24
gT5 xxl	65.65	69.83	**83.15**	**75.96**	**51.62**

Table 2: $F_{0.5}$ Scores. Models denoted with * are ensemble models. We used the M^2 scorer for CoNLL-14, Russian, Czech and German, and the ERRANT scorer (Bryant et al., 2019b) for BEA test.

Rothe, Sascha et al. "A Simple Recipe for Multilingual Grammatical Error Correction." ArXiv abs/2106.03830 (2021): n. pag.

Data Transfer - An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction

Q1: How to generate better pseudo data (X)

Q2: Seed Corpus Selection (Y)

- Clean data
- Grammatical complexity

Q3: Joint Training or Pretraining

$$\mathcal{L}(\mathcal{D}, \boldsymbol{\Theta}) = -\frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{X},\boldsymbol{Y})\in\mathcal{D}} \log(p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\Theta})), \quad (1)$$

where $p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\Theta})$ denotes the conditional probability of $\boldsymbol{Y}$ given $\boldsymbol{X}$.

Kiyono, Shun et al. “An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction.” EMNLP/IJCNLP (2019).

Data Transfer - An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction

Q1: How to Generate Better Pseudo Data:

1. Direct Noise
   - Inject noise directly
2. Back translation (Noisy)
   - Train a Seq2Seq model to generate noisy sentences, given grammatically correct sentence
   - Add noise during decoding stage
3. Back translation (sample)
   - Seq2Seq
   - Random Sample during decoding (beam search)

Kiyono, Shun et al. “An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction.” EMNLP/IJCNLP (2019).

Data Transfer - An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction

Q1: How to Generate Better Pseudo Data:

1. Direct Noise
2. Back translation (Noisy)
3. Back translation (sample)

Method	Prec.	Rec.	$F_{0.5}$
Baseline	46.6	23.1	38.8
BACKTRANS (SAMPLE)	44.6	27.4	39.6
BACKTRANS (NOISY)	42.5	**31.3**	39.7
DIRECTNOISE	**48.9**	25.7	**41.4**

Table 2: Performance of models on BEA-valid: a value in **bold** indicates the best result within the column. The seed corpus $\mathcal{T}$ is SimpleWiki.

Kiyono, Shun et al. “An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction.” EMNLP/IJCNLP (2019).

Data Transfer - An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction

Q2: Seed Corpus Selection:

Corpus	#sent (pairs)	Notes
SimpleWiki	1,369,460	Simple grammar
Wikipedia	145,883,941	Complex grammar
Gigaword	131,864,979	Clean text

Kiyono, Shun et al. “An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction.” EMNLP/IJCNLP (2019).

Data Transfer - An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction

Q2: Seed Corpus Selection:

Method	Seed Corpus $\mathcal{T}$	Prec.	Rec.	$F_{0.5}$
Baseline	N/A	46.6	23.1	38.8
BACKTRANS (NOISY)	Wikipedia	43.8	30.8	40.4
BACKTRANS (NOISY)	SimpleWiki	42.5	31.3	39.7
BACKTRANS (NOISY)	Gigaword	43.1	33.1	40.6
DIRECTNOISE	Wikipedia	48.3	25.5	41.0
DIRECTNOISE	SimpleWiki	48.9	25.7	41.4
DIRECTNOISE	Gigaword	48.3	26.9	41.7

Table 3: Performance on BEA-valid when changing the seed corpus $\mathcal{T}$ used for generating pseudo data ($|\mathcal{D}_p| = 1.4$M).

Kiyono, Shun et al. “An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction.” EMNLP/IJCNLP (2019).

Data Transfer - An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction

Q3: Joint Training vs Pretraining:

Optimization	Method	$\lvert\mathcal{D}_p\rvert$	Prec.	Rec.	$F_{0.5}$
N/A	Baseline	0	46.6	23.1	38.8
PRETRAIN	BACKTRANS (NOISY)	1.4M	49.6	24.3	41.1
PRETRAIN	DIRECTNOISE	1.4M	48.4	21.2	38.5
JOINT	BACKTRANS (NOISY)	1.4M	43.8	30.8	40.4
JOINT	DIRECTNOISE	1.4M	48.3	25.5	41.0
PRETRAIN	BACKTRANS (NOISY)	14M	50.6	30.1	44.5
PRETRAIN	DIRECTNOISE	14M	49.8	25.8	42.0
JOINT	BACKTRANS (NOISY)	14M	43.0	32.3	40.3
JOINT	DIRECTNOISE	14M	48.7	23.5	40.1

Table 4: Performance of the model with different optimization settings on BEA-valid. The seed corpus $\mathcal{T}$ is Wikipedia.

PRETRAIN is crucial for utilizing extensive pseudo data.

Kiyono, Shun et al. “An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction.” EMNLP/IJCNLP (2019).

Data Transfer - An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction

Q3: Joint Training vs Pretraining:

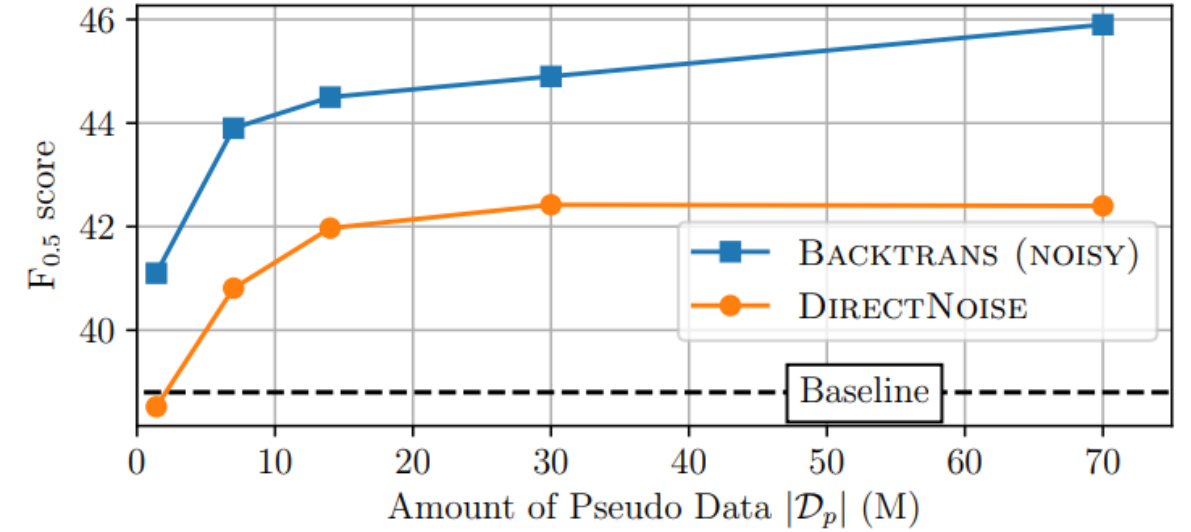


Figure 1: Performance on BEA-valid for different amounts of pseudo data ($|\mathcal{D}_p|$). The seed corpus $\mathcal{T}$ is Wikipedia.

Kiyono, Shun et al. "An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction." EMNLP/IJCNLP (2019).

Data Transfer - Cross-lingual Transfer Learning for Grammatical Error Correction

- Effect of Language Similarity
- Effect of parallel corpus

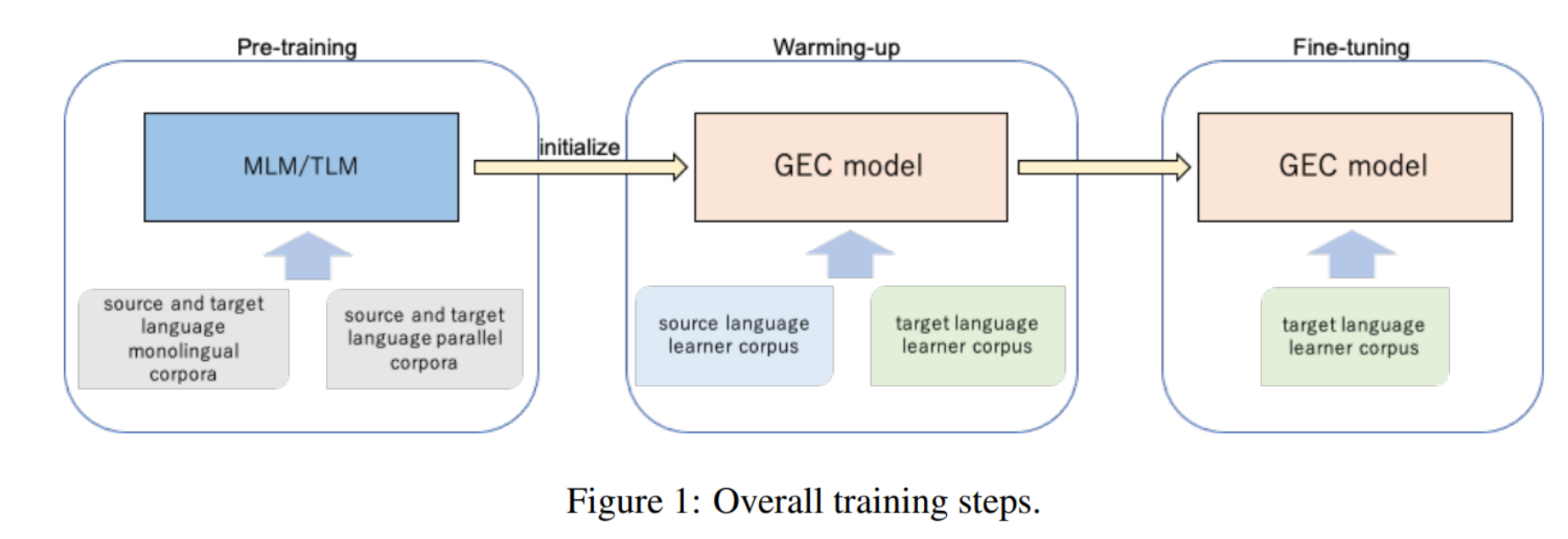


Figure 1: Overall training steps.

Rothe, Sascha et al. "A Simple Recipe for Multilingual Grammatical Error Correction." ArXiv abs/2106.03830 (2021): n. pag.

Data Transfer - Cross-lingual Transfer Learning for Grammatical Error Correction

target	source similarity		
	high	moderate	low
Russian	Czech	English	Japanese
Czech	Russian	English	Japanese
English	German	Russian	Japanese

Table 2: Languages used in the experiment.

- Better transfer for similar languages
- Parallel data helps better transfer

	Model	P	R	$F_{0.5}$
PLAIN	Ru-only	19.29	14.08	17.96
	Cs→Ru	19.05	12.78	17.35
	En→Ru	23.76	13.65	20.70
	Ja→Ru	20.70	13.57	18.73
MLM	Ru-only	19.95	**23.15**	20.52
	Cs→Ru	26.36	19.02	24.47
	En→Ru	26.02	19.74	24.47
	Ja→Ru	27.23	16.13	23.93
TLM	Cs→Ru	**28.51**	22.47	**27.06**
	En→Ru	27.60	22.18	26.31
	Ja→Ru	26.11	19.99	24.61

Table 4: Russian GEC results.

	Model	P	R	$F_{0.5}$
PLAIN	Cs-only	52.05	39.29	48.88
	Ru→Cs	59.93	38.73	54.01
	En→Cs	61.35	39.01	55.05
	Ja→Cs	56.22	38.53	51.49
MLM	Cs-only	57.46	47.40	55.12
	Ru→Cs	63.58	47.15	59.43
	En→Cs	63.54	48.63	59.87
	Ja→Cs	62.15	47.35	58.50
TLM	Ru→Cs	**65.09**	**50.82**	**61.63**
	En→Cs	63.20	48.47	59.58
	Ja→Cs	63.84	45.70	59.15

Table 5: Czech GEC results.

Rothe, Sascha et al. “A Simple Recipe for Multilingual Grammatical Error Correction.” ArXiv abs/2106.03830 (2021): n. pag.

Key Takeaways

Model Transfer

- Multi-lingual Denoising pretraining is effective for cross-lingual Seq2seq pretraining
- Parallel corpus helps language transfer
- Bigger model / Larger data lead to better performance

Data transfer

- GEC pretraining could boost performance on top of general pretrained models

Outline

Daxin Jiang

Software Technology Center at Asia (STCA) of Microsoft

- Introduction [Daxin Jiang]
  - Motivating examples in Microsoft products
  - Problem description
  - Categorization of applications
  - Challenges
- Methodology [Daxin Jiang]
  - Model Transfer
  - Data Transfer
- Applications*
  - Dependency Parsing [Xiubo Geng]
  - Machine Reading Comprehension [Ming Gong]
  - Grammar Error Correction [Linjun Shou]
- Summary & Future directions [Jian Pei]

Linjun Shou Xiubo Geng Ming Gong

Software Technology Center at Asia (STCA) of Microsoft

Jian Pei

Simon Fraser University

*For more applications, please refer to our tutorial at The Web Conference 2021

Summary

- Language scaling is critical in business and many applications
  - Key technical challenge: lack of cross-lingual data

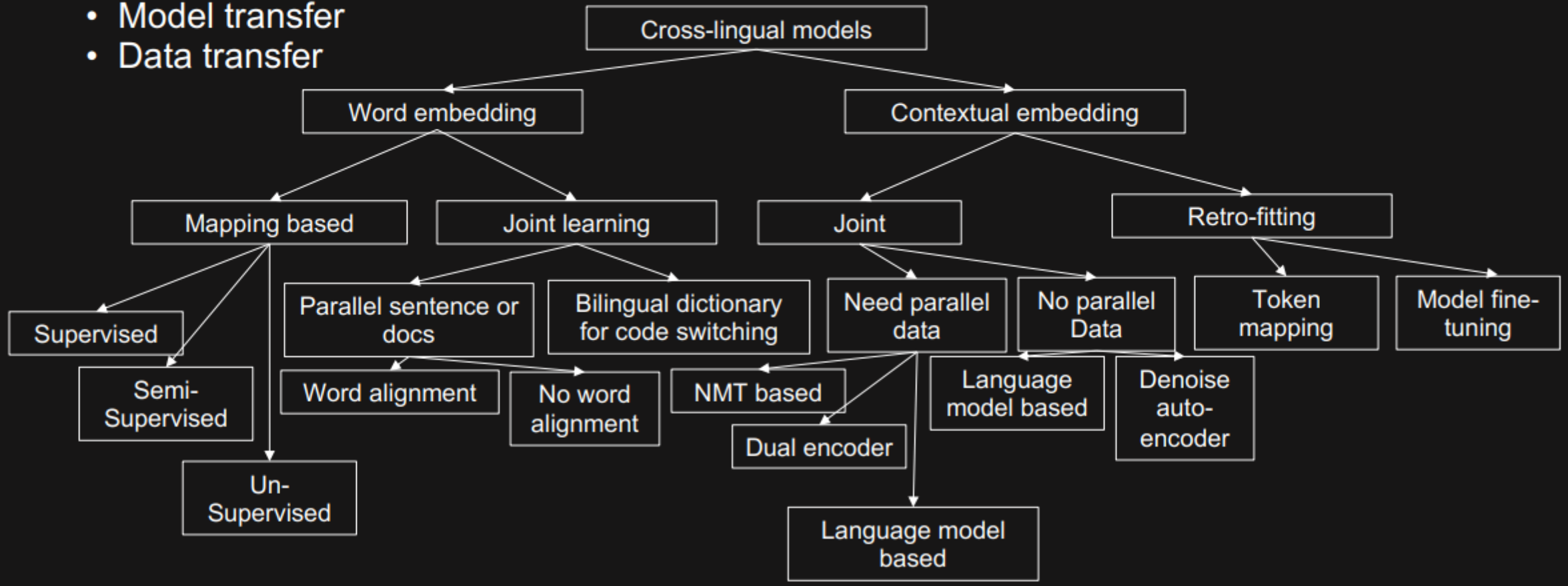
Type	Category	Sub Category	Example
NLU	Text Classification	Single text	Domain identification, Intent detection, Sentiment classification
		Text pair	Information retrieval, Natural language inference
	Sequence Labeling	Single text	Named entity recognition, Slot tagging
		Text pair	Machine reading comprehension
	Structure Prediction	Single text	Dependency parsing, constituency parsing, semantic role labeling
NLG	Text Generation	Token level	Spelling correction, Sentence auto completion
		Sentence level	Machine translation, Conversation, Question generation

Summary

- Methodologies
  - Model transfer
  - Data transfer

Application examples:

- Dependency parsing
- Machine reading comprehension
- Grammar error correction



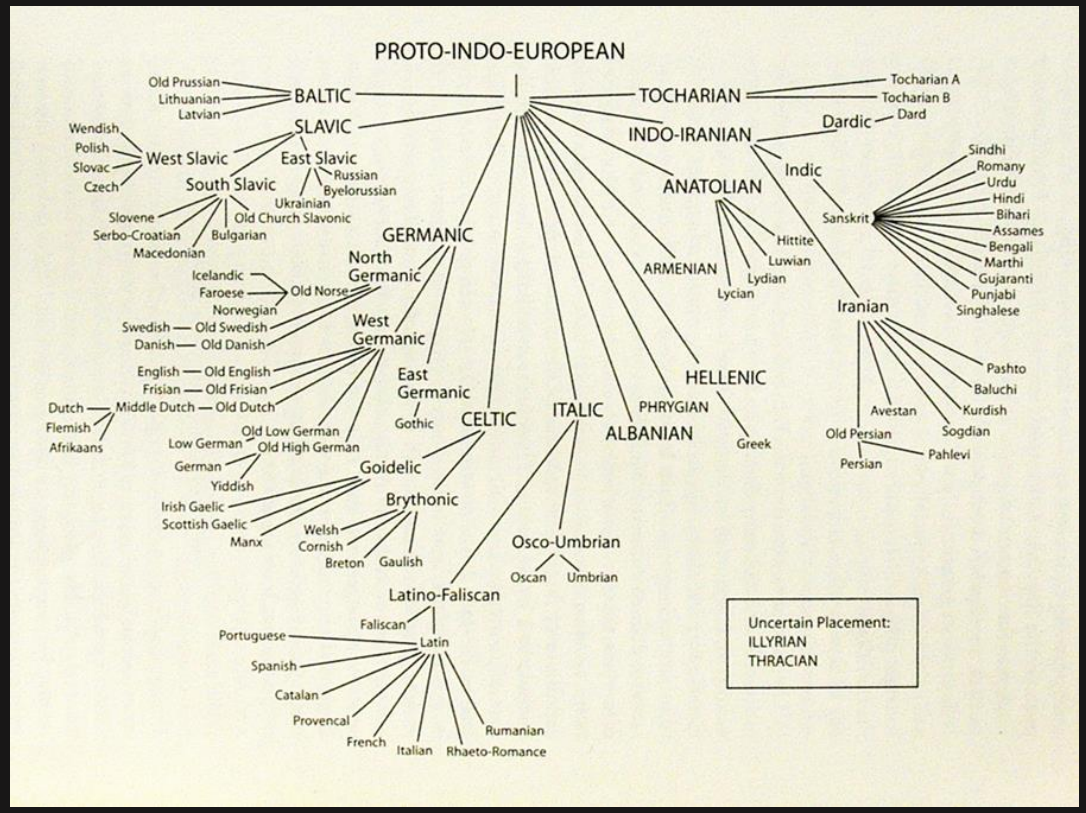
Future Directions (1)

- Linguistic knowledge
  - Universal POS tagging and Dependency Parsing (Universal Dependencies)
  - Typological databases such as the WALS (Dryer and Haspelmath, 2013), PHOIBLE (Moran et al., 2014), Ethnologue (Lewis et al., 2015), and Glottolog (Hammarstrom et al., 2015)
- Early works use topological databases to measure the distances between languages or select the intermediate transfer languages
- Look forward to systematic methods that integrate linguistic knowledge into transfer approaches

WALS: Matthew S. Dryer and Martin Haspelmath. 2013. The World Atlas of Language Structures Online. Max Planck Institute for Evolutionary Anthropology
PHOIBLE: Steven Moran, Daniel McCloy, and Richard Wright. 2014. PHOIBLE Online. Max Planck Institute for Evolutionary Anthropology, Leipzig
Ethnologue: M. Paul Lewis, Gary F. Simons, and Charles D. Fennig. 2015. Ethnologue: Languages of the World, Eighteenth edition. SIL International, Dallas, Texas
Glottolog: Harald Hammarstrom, Robert Forkel, Martin Haspel-math, and Sebastian Bank. 2015. Glottolog 2.6. Max Planck Institute for the Science of Human History, Jena

Future Directions (2)

- Best practice to train multi-lingual models
  - how to mitigate “catastrophic forgetting”?
  - do we align models with language families?
  - how to measure the “capacity” of models and the “size” of languages to determine the best combination of languages



Picture from The English Cowpath: The Proto-Indo-European Homeland Puzzle


thank you
danke
謝謝
merci
gracias
obrigado
dziękuję
bedankt
go raibh maith agat
teşekkür ederim
ngiyabonga
спасибо
рахмат
Баярлалаа
dank je
tapadh leat
mochchakkeram
sukriya
kop khun krap
terima kasih
grazie
arigatô
dakujem
takk
taiku
sagolun
hvala
mauruuru
kiitos
dankie
faafetai lava
vinaka
blagodaram
misaotra
matondo
paldies
grazzi
mahalo
хвала
asante
manana
obrigada
tenki
murakoze
chokrane
mamnun
дякую
djiere dieuf
tau
akun
dankon
aciu
köszönöm
gràcie
bayarlalaa
enkosi
nanni
nandri
dhanyavad
sobodi
dēkuji
mèsi
didi madloba
kam sah hamnida
তোমাকে ধন্যবাদ
감사합니다
ありがとう
najis tuke
rahmat
tanemirt
rahmet
xièxie
ευχαριστώ
diolch
dhanyavadagalu
shukriya
merce
мерси
trugarez
chnorakaloutioun
gratias ago
gràcies
sulpáy
welalin
tack
barka
mersi
kia ora
spas
спасибі