Scaling out NLP Applications to 100+ Languages

Lecture Tutorial for The Web Conference 2021

April 19-23, 2021, Ljubljana

https://languagescaling.github.io/

Outline

- Introduction [Daxin Jiang]
  - Motivating examples in Microsoft products
  - Problem description
  - Categorization of applications
  - Challenges and major approaches
- Applications
  - Natural Language Inference [Linjun Shou]
  - Information retrieval [Xiubo Geng]
  - Machine Reading Comprehension [Ming Gong]
- Future directions [Daxin Jiang]

NLP Group
Software Technology Center at Asia (STCA) of Microsoft

Linjun Shou

Xiubo Geng

Ming Gong

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.

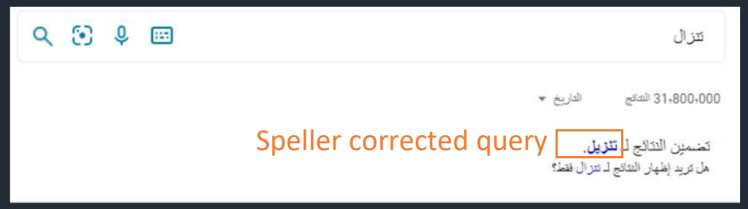
Why language matters (industry)

- 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

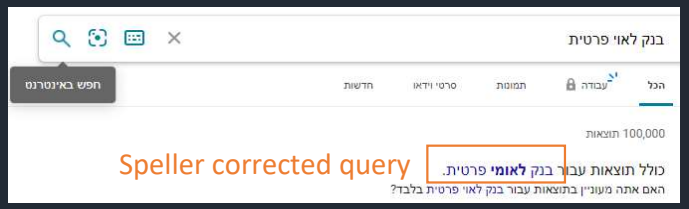
Speller | Web Relevance | Suggested Replies | Natural Language Understanding

Microsoft

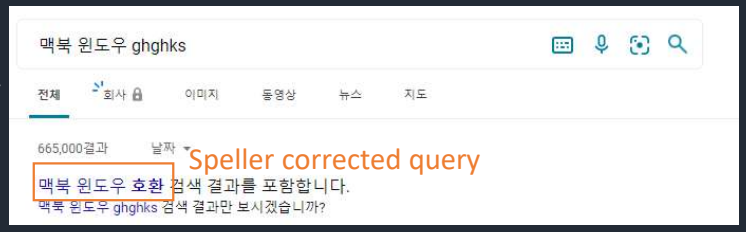
ar-EG
{تنزال}
English Translation:
download



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

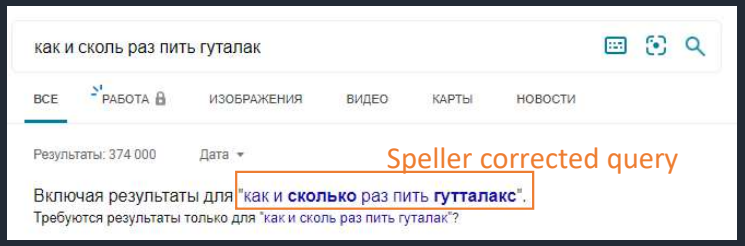


ko-KR
{맥북 윈도우 ghghks}
English Translation:
MacBook Windows
Compatible

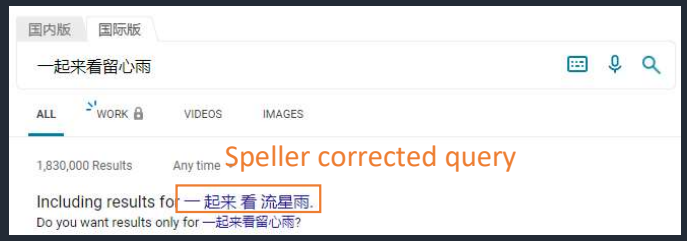


ja-JP
{田中みどらら}
English Translation:
Ms. Tanaka

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

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

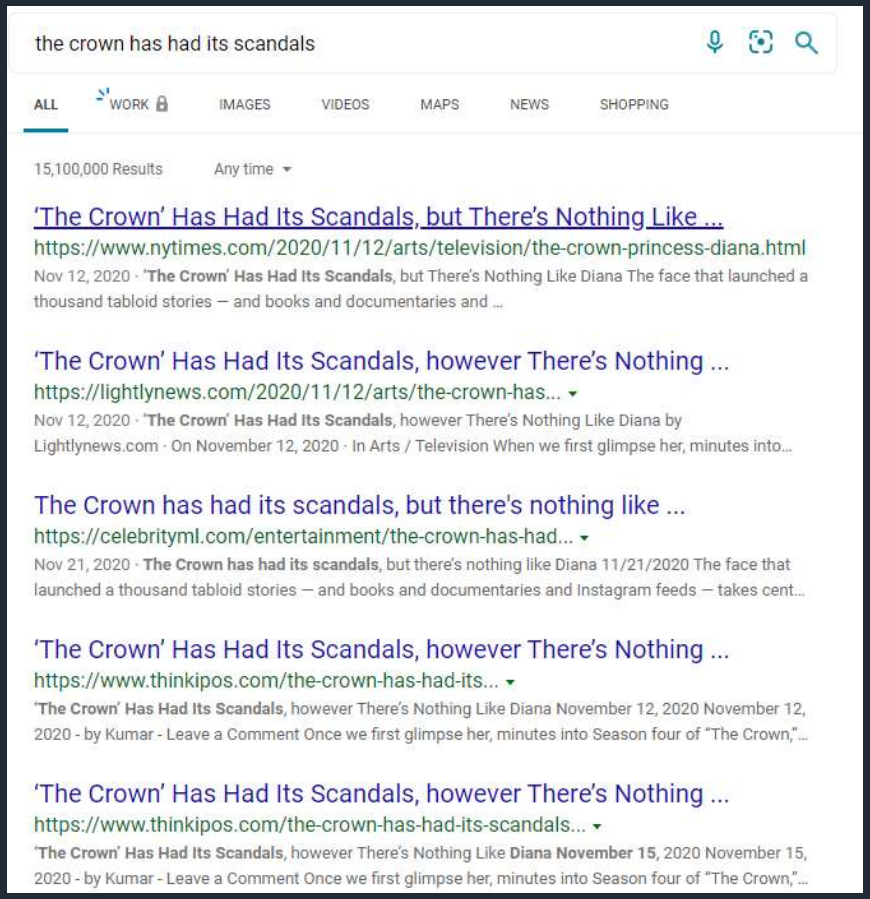
1 keystroke difference

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

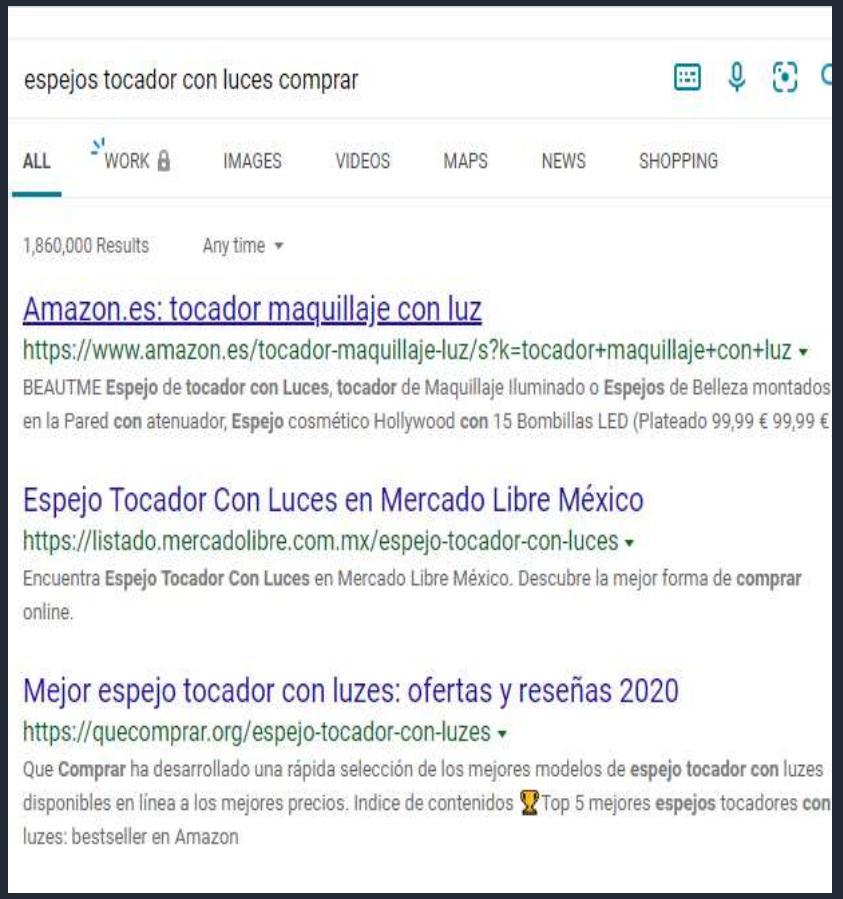
Speller | **Web Relevance** | Suggested Replies | Natural Language Understanding

Microsoft

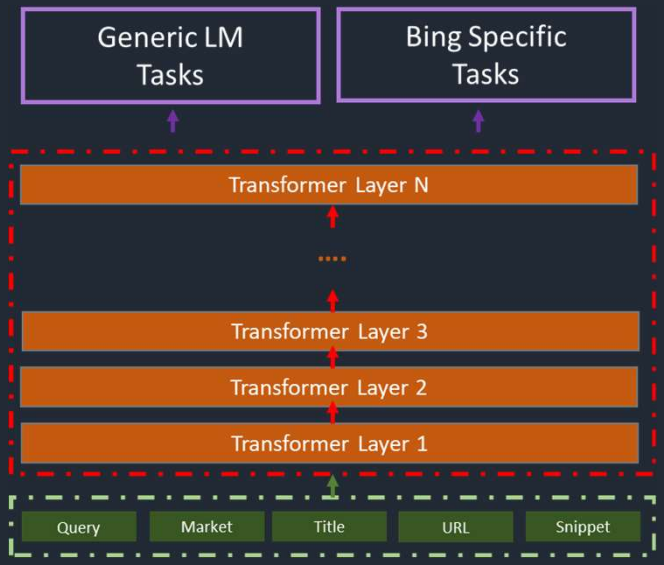
English



Spanish



Efficiency



Huge model

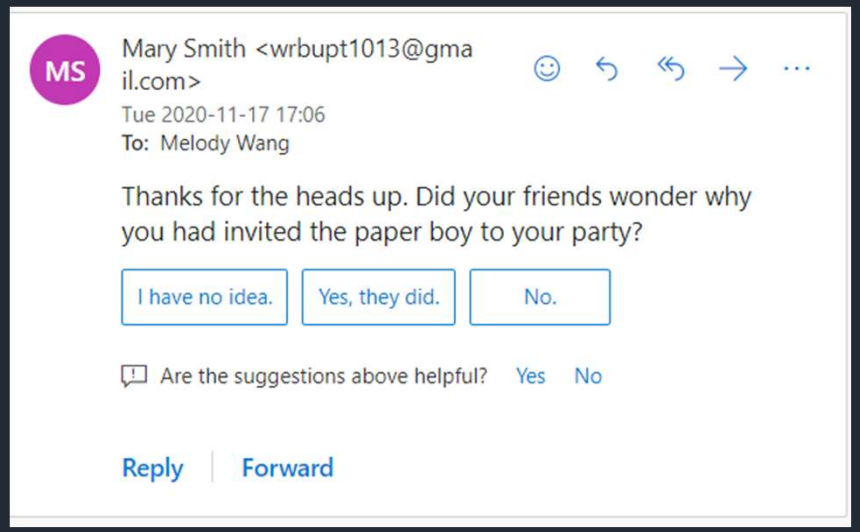
Huge index

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

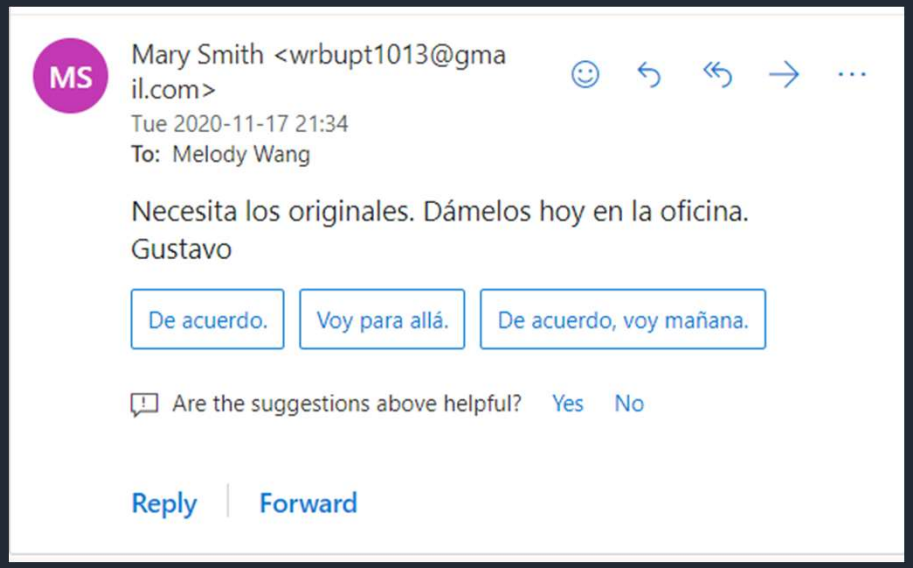
Speller | Web Relevance | **Suggested Replies** | Natural Language Understanding

Microsoft

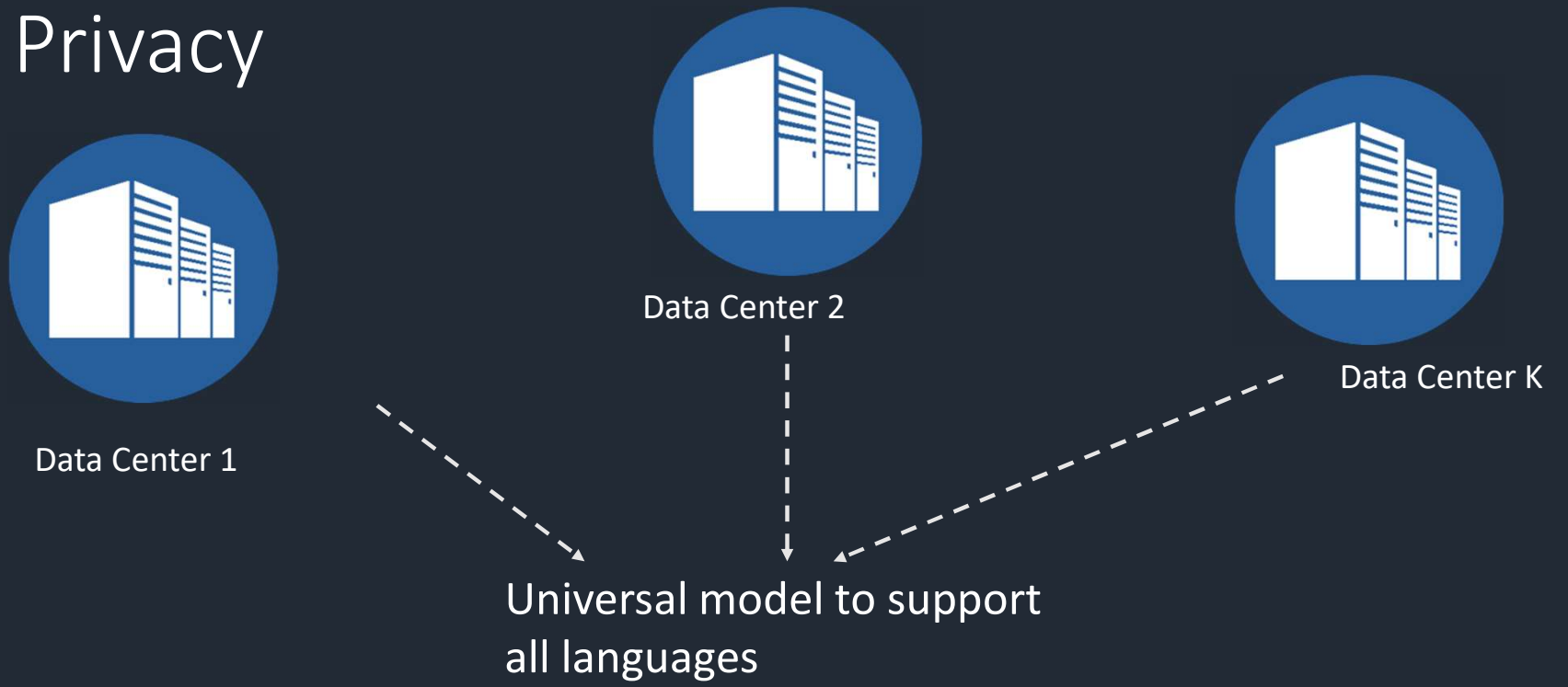
English



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

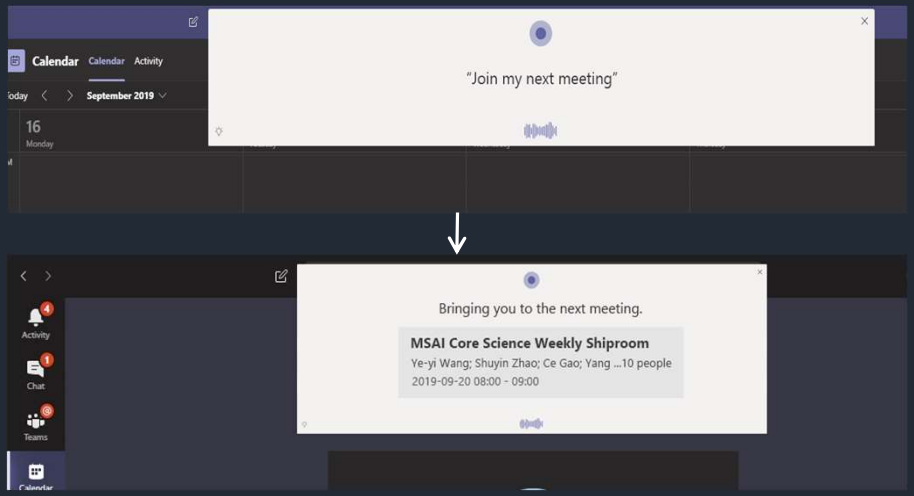
Compartilhe o arquivo com Bob

Agrega Alice a este chat

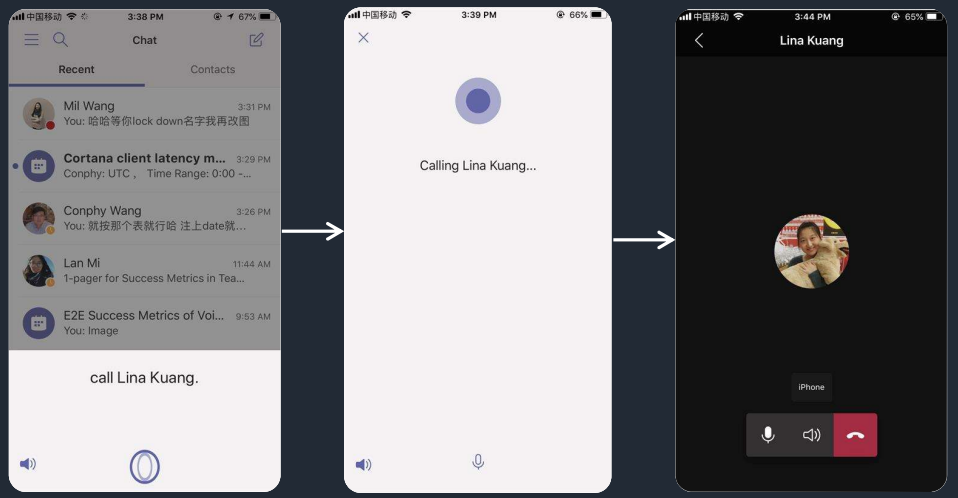
Füge Alice zum Cortana - Kanal hinzu

次の会議に参加する

Desktop Experience



Mobile Experience

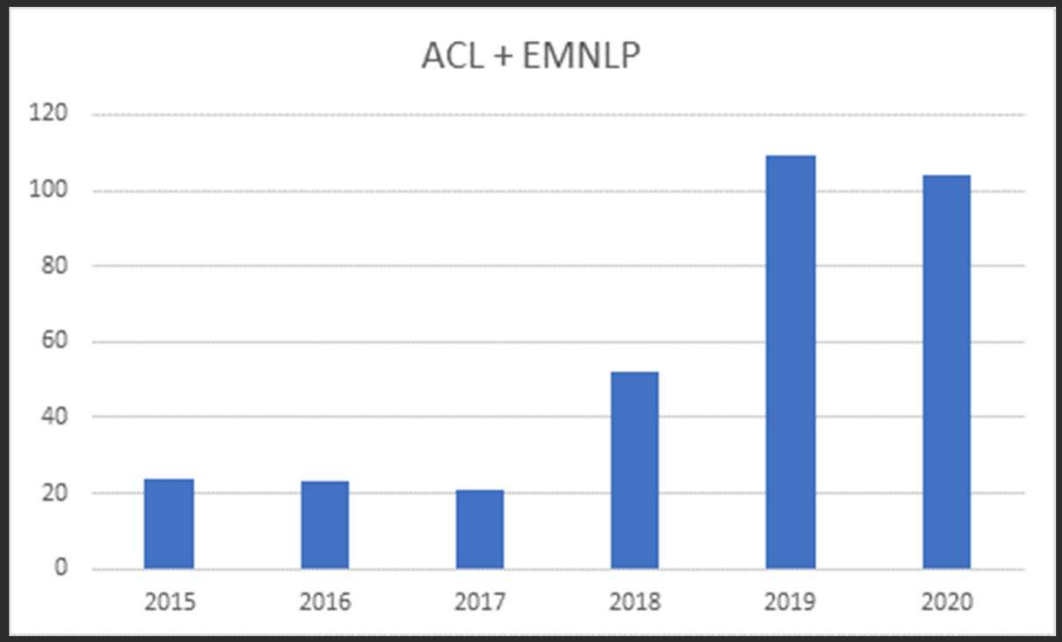


M Domains × N Languages

- Need large amount of training data
- Need to handle large number of 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)



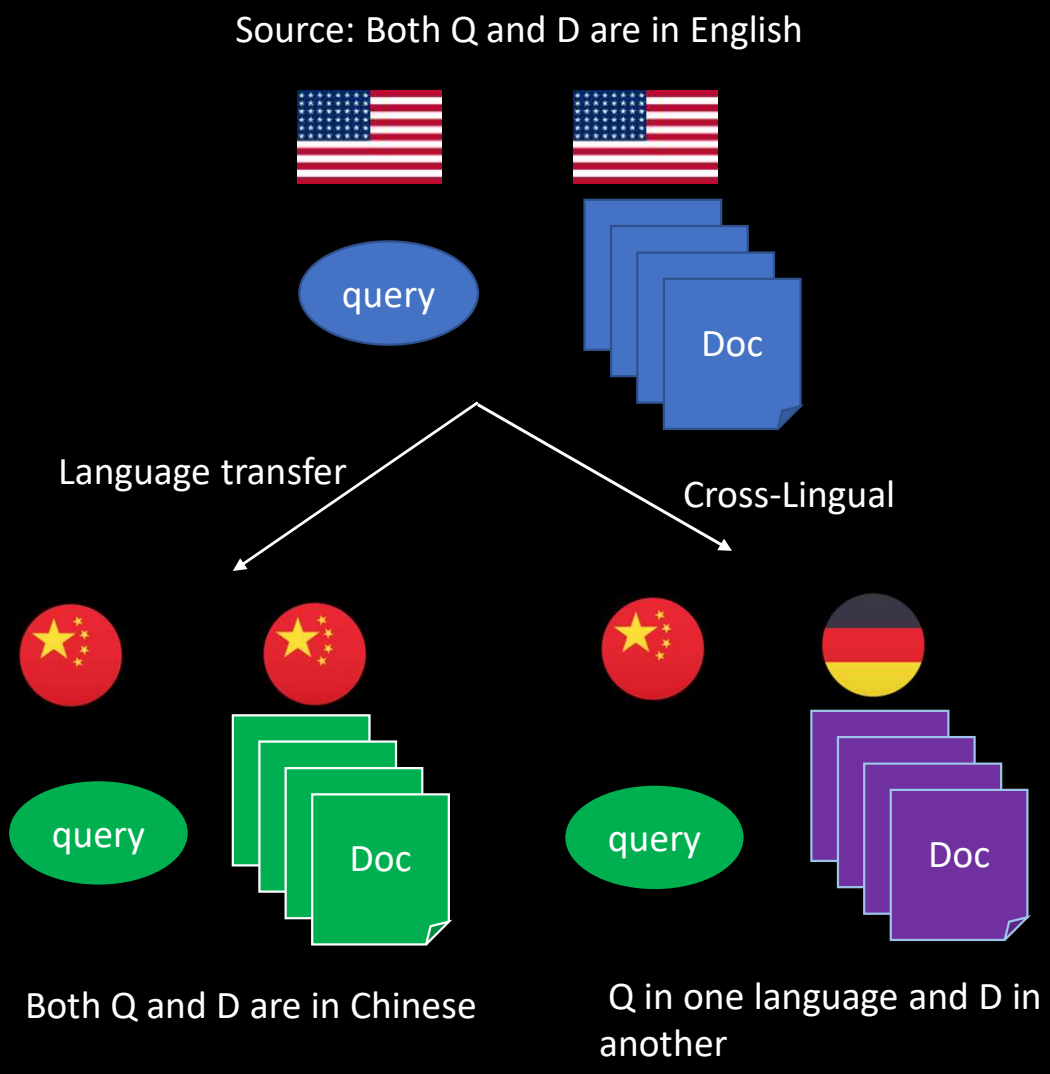
Problem Description for Language Scaling

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

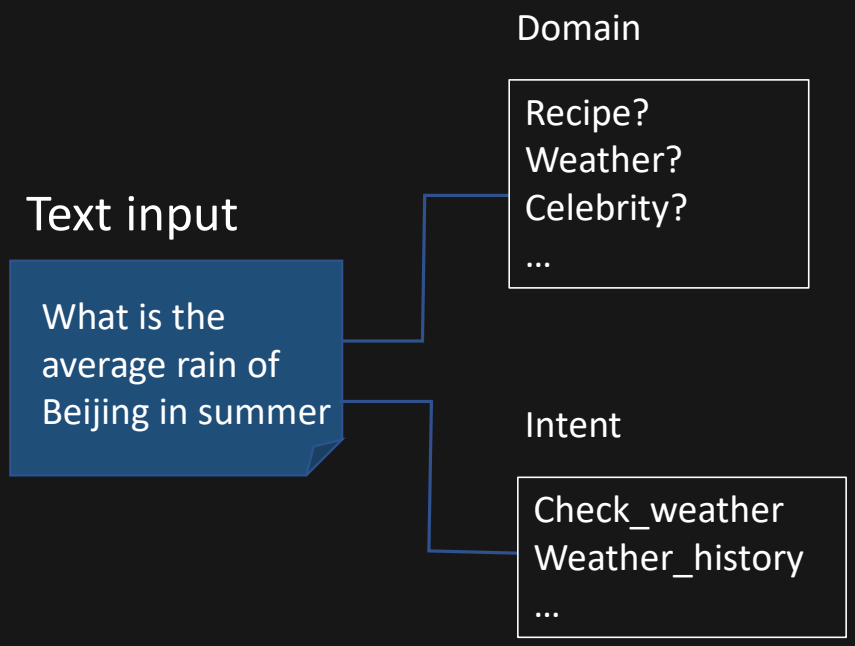


NLP Tasks

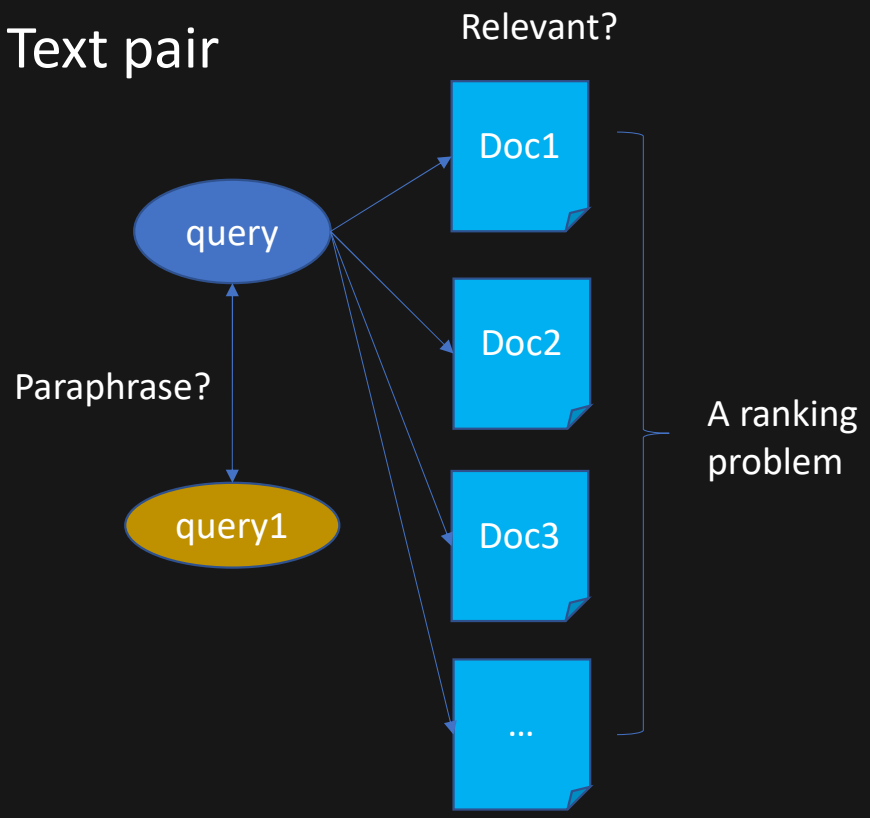
Type	Category	Sub Category	Example
NLU	Text Classification	Single text	Domain identification, Intent detection, Sentiment classification
		Text pair	Information retrieval, Natural language inference
	Sequence Labeling	Single text	Named entity recognition, Slot tagging
		Text pair	Machine reading comprehension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

Text Classification

- Single text



- Text pair

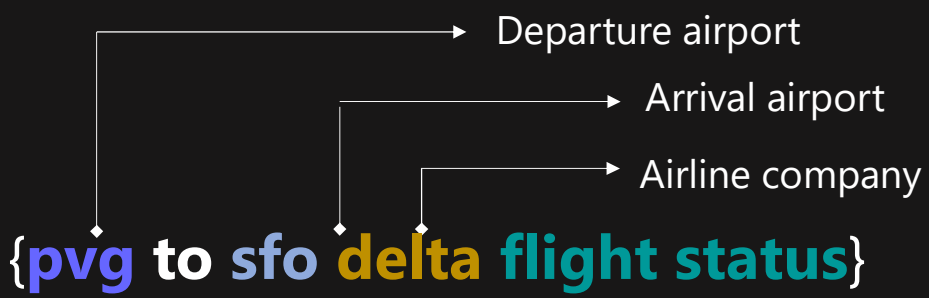


Sequence labeling

- Single text

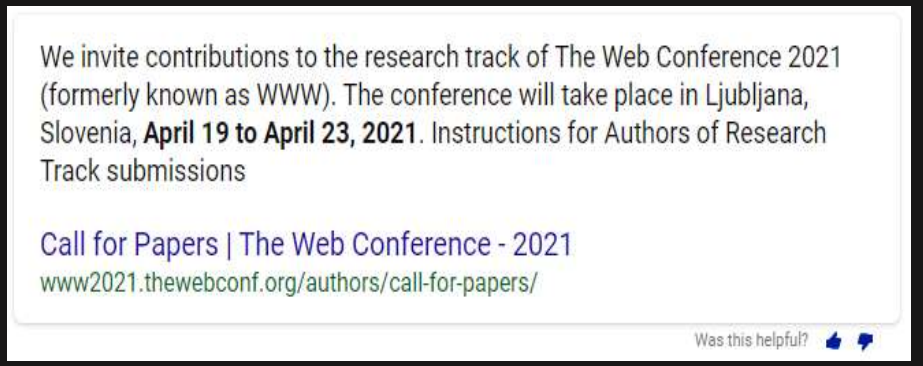
Luke Rawlence PERSON joined Aiimi ORG as a data scientist in Milton Keynes PLACE, after finishing his computer science degree at the University of Lincoln. ORG

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



- Text pair

what is the date for the web conference 2021



Text Generation

Sentence level

Token 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

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

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

Challenges for Language Scaling

Size 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, cold start issues are all related to this challenge
- Major focus on this tutorial

Number of 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

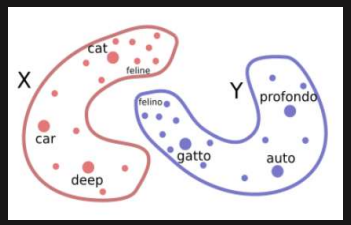
- Hardware acceleration
- Computation library optimization
- Model compression

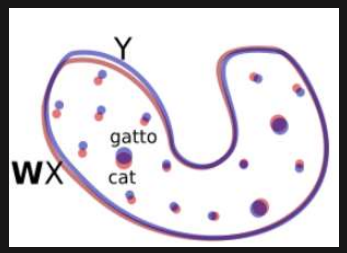
Approaches: Model Transfer and Data Transfer

(1) Model Transfer

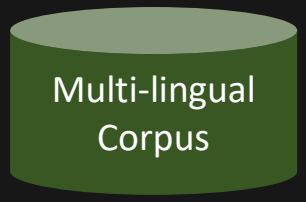
Zero-shot approach

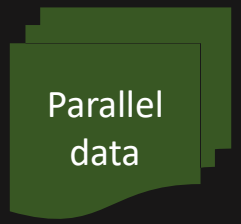
Cross-lingual model



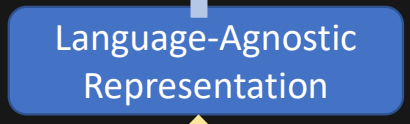


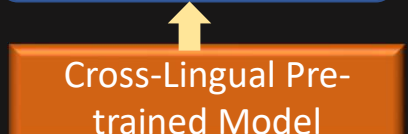
From MUSE





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

- Mapping-based methods

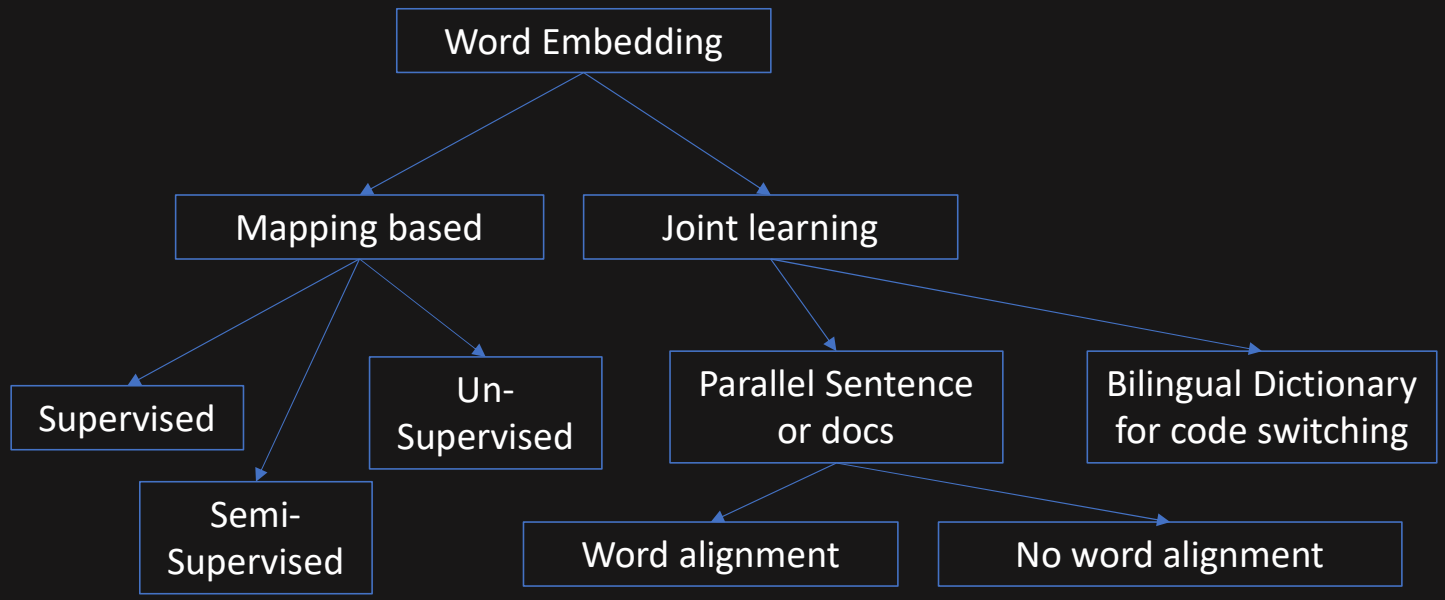
$$X^t \approx WX^s$$

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

- **Joint-learning methods**

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

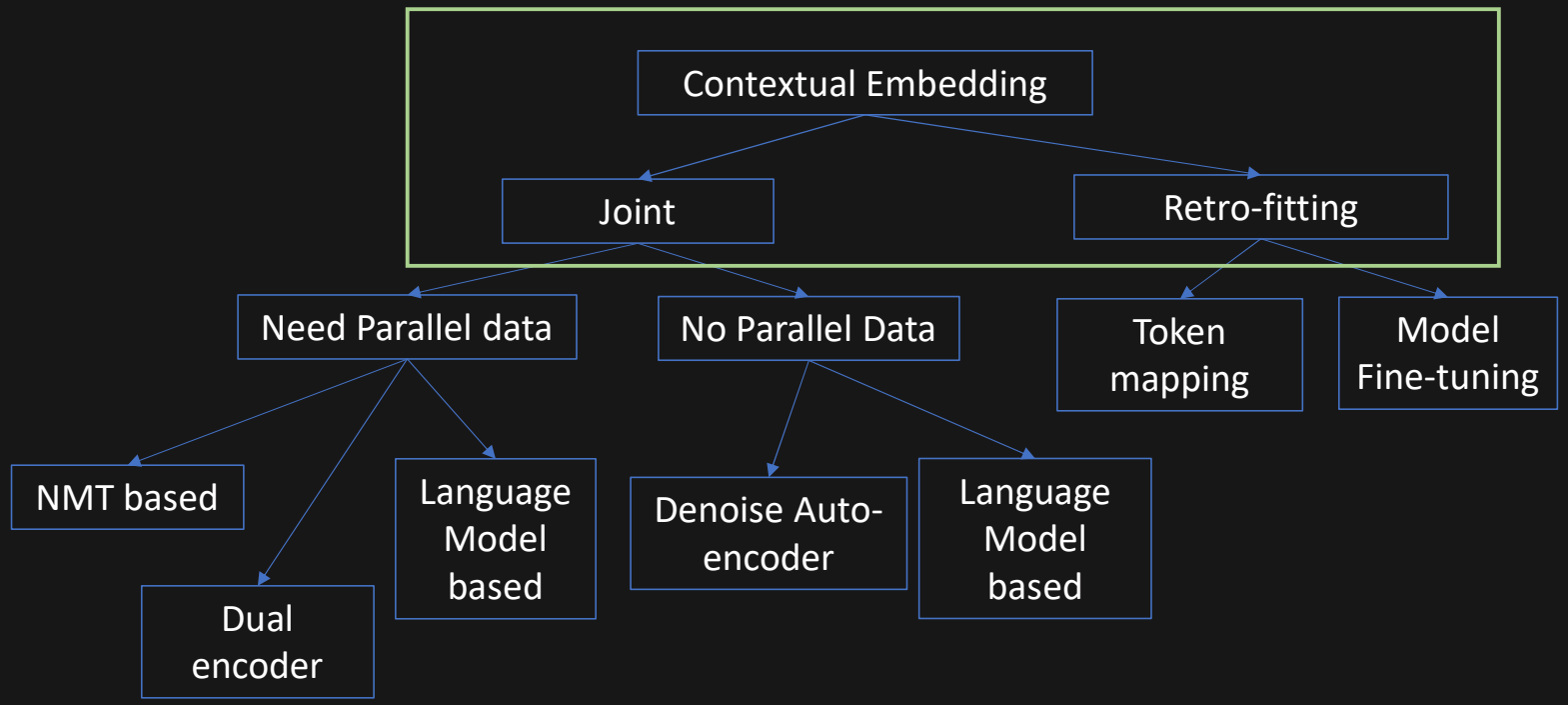
Cross-Lingual Word Embedding



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)]

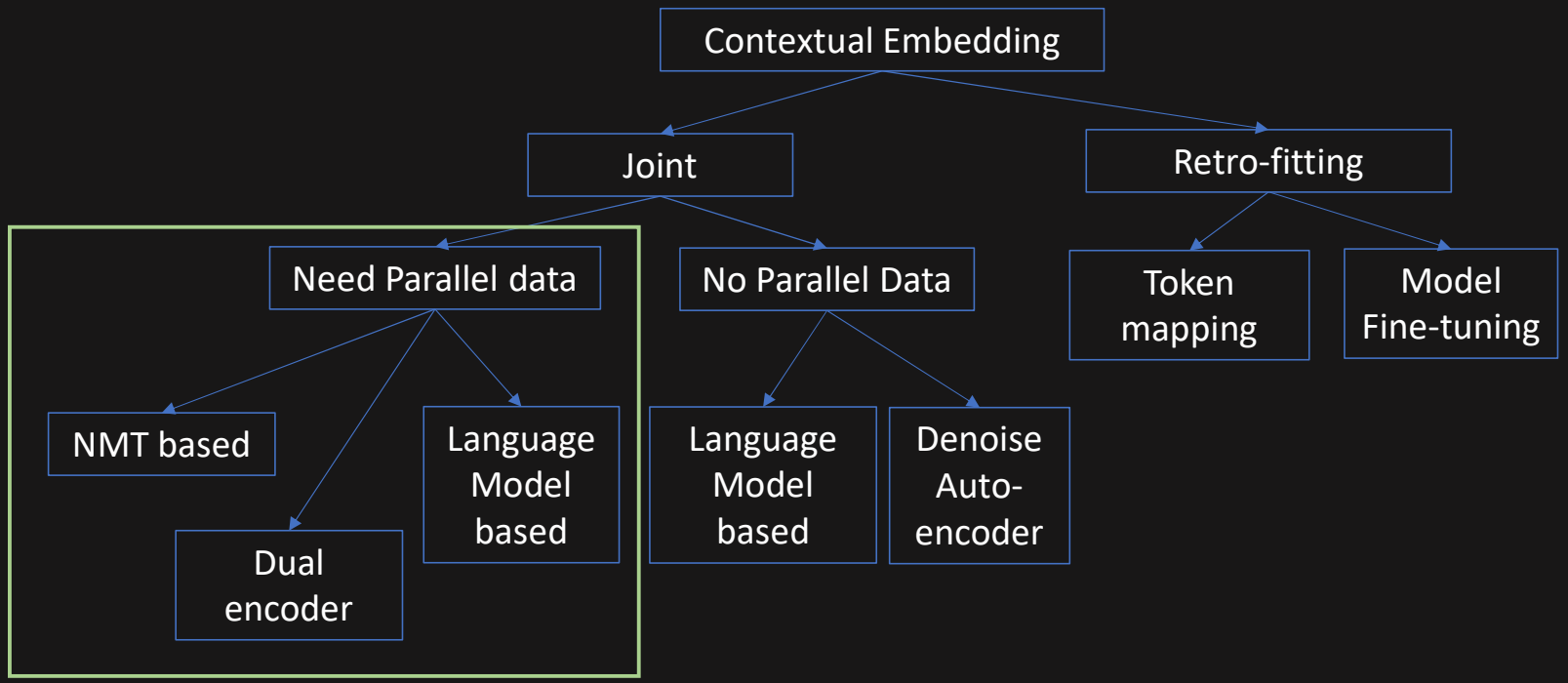
Cross-lingual Contextual embedding



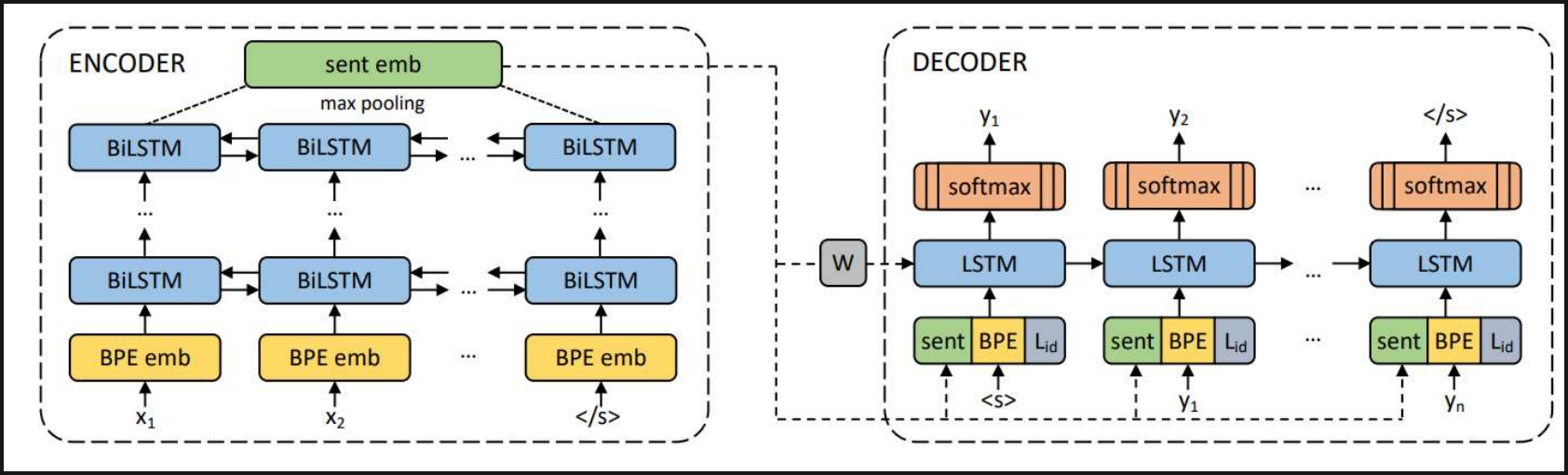
Joint or Retro-fitting?

- *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



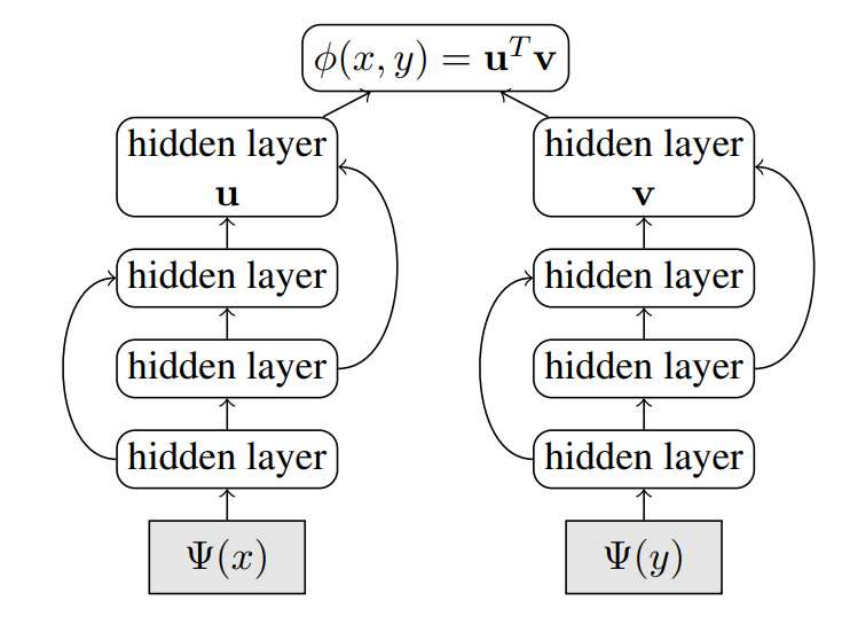
Neural machine translation (NMT) based method



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

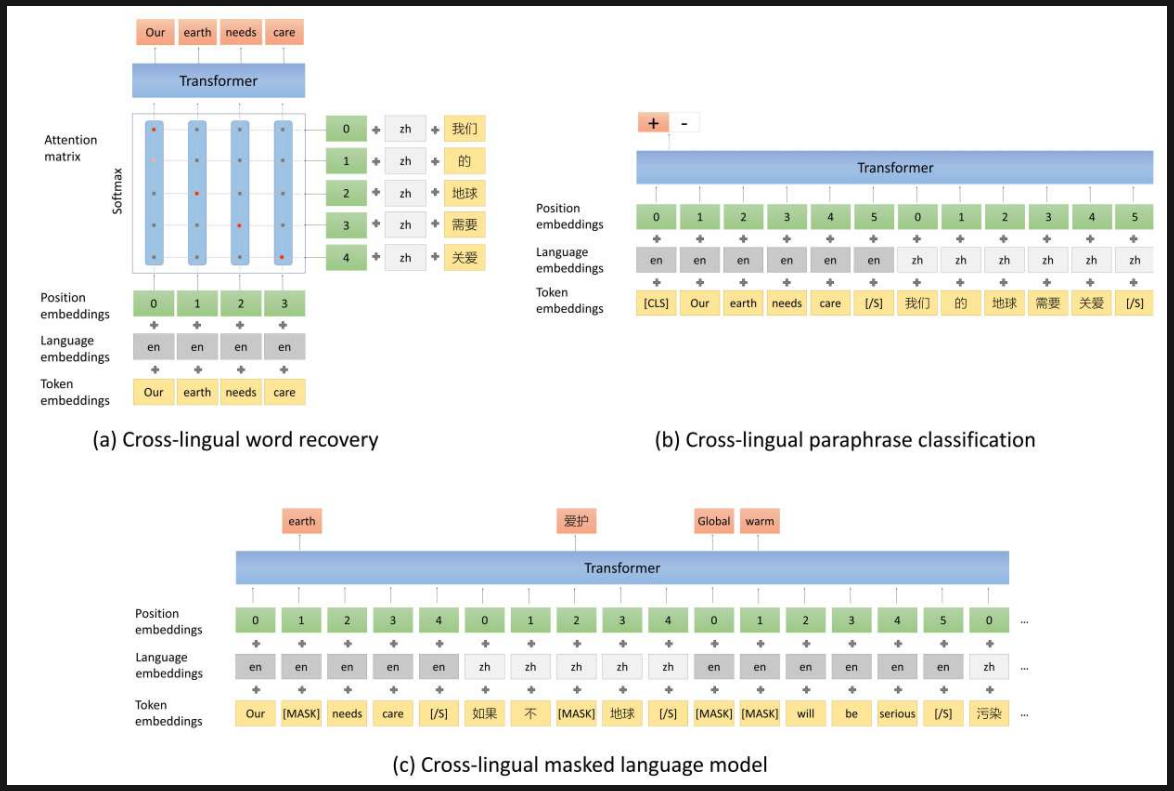
Dual encoder



- *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

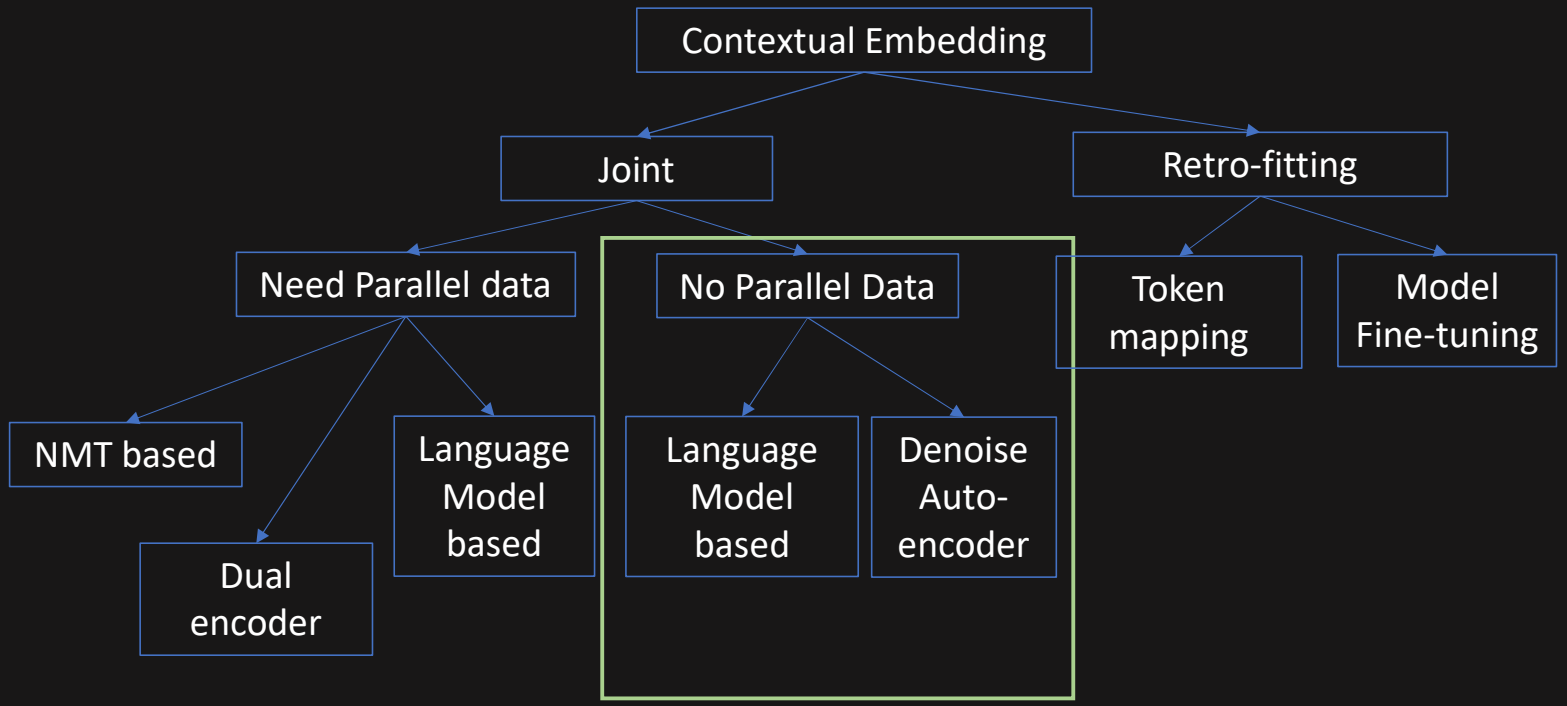
Language model based methods



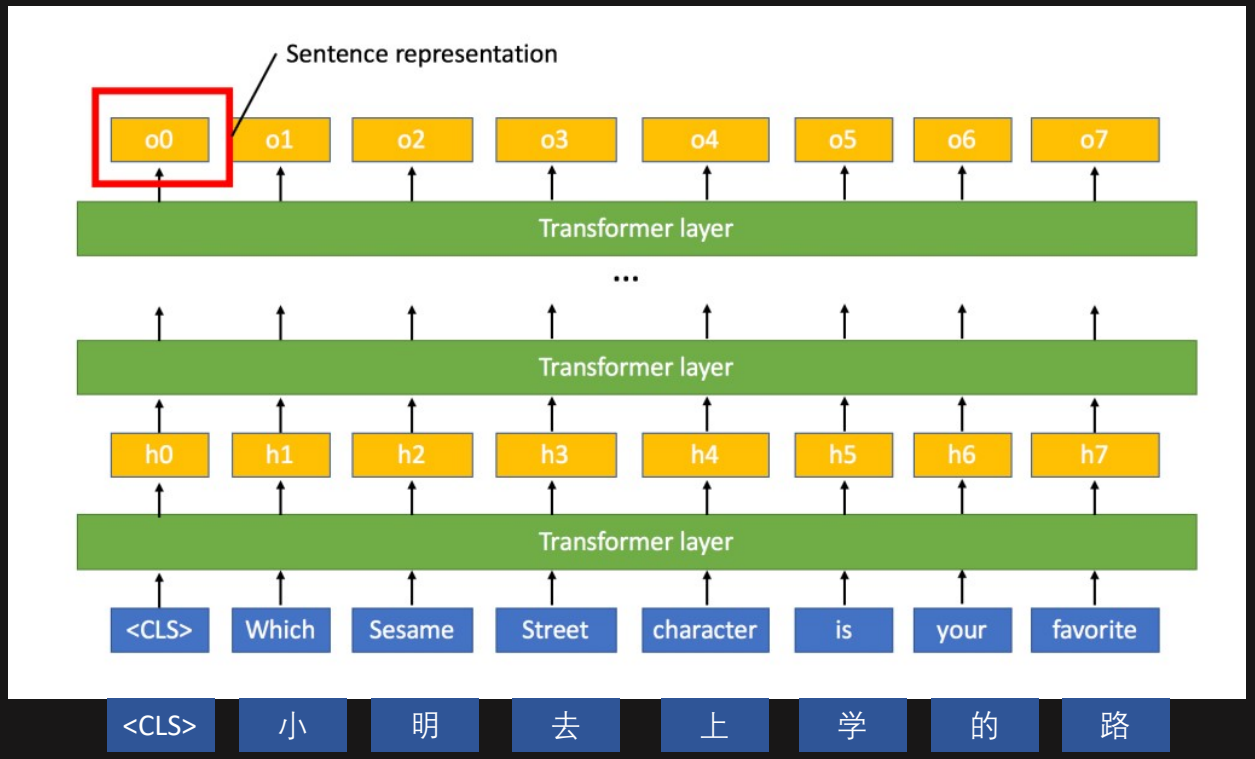
- Transformer based language model
- Cross-lingual representation at both sentence level and token level

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



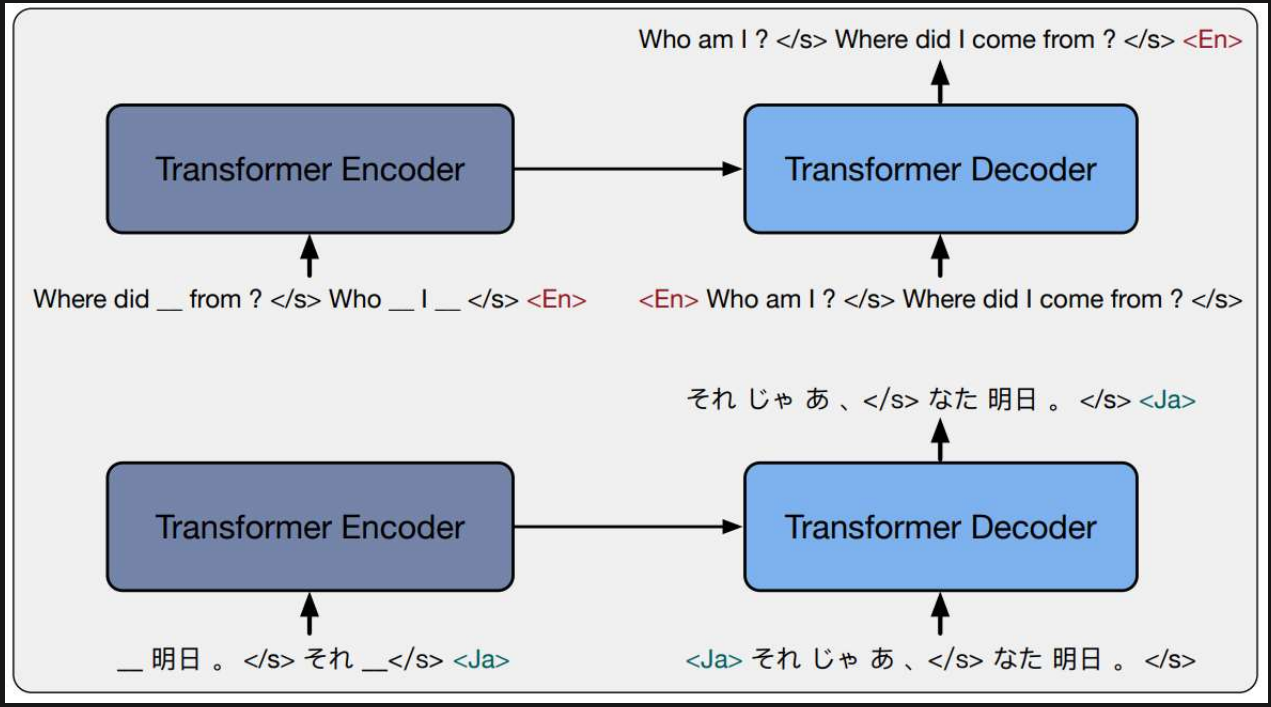
- Share the same architecture and training scheme with monolingual BERT
- 104 languages Wikipedia pages
- 110k shared WordPiece vocabulary

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

Why mBERT has cross-lingual effectiveness

- Overlap vocabulary 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
- Shijie Wu, Mark Dredze. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT, EMNLP'19
- …

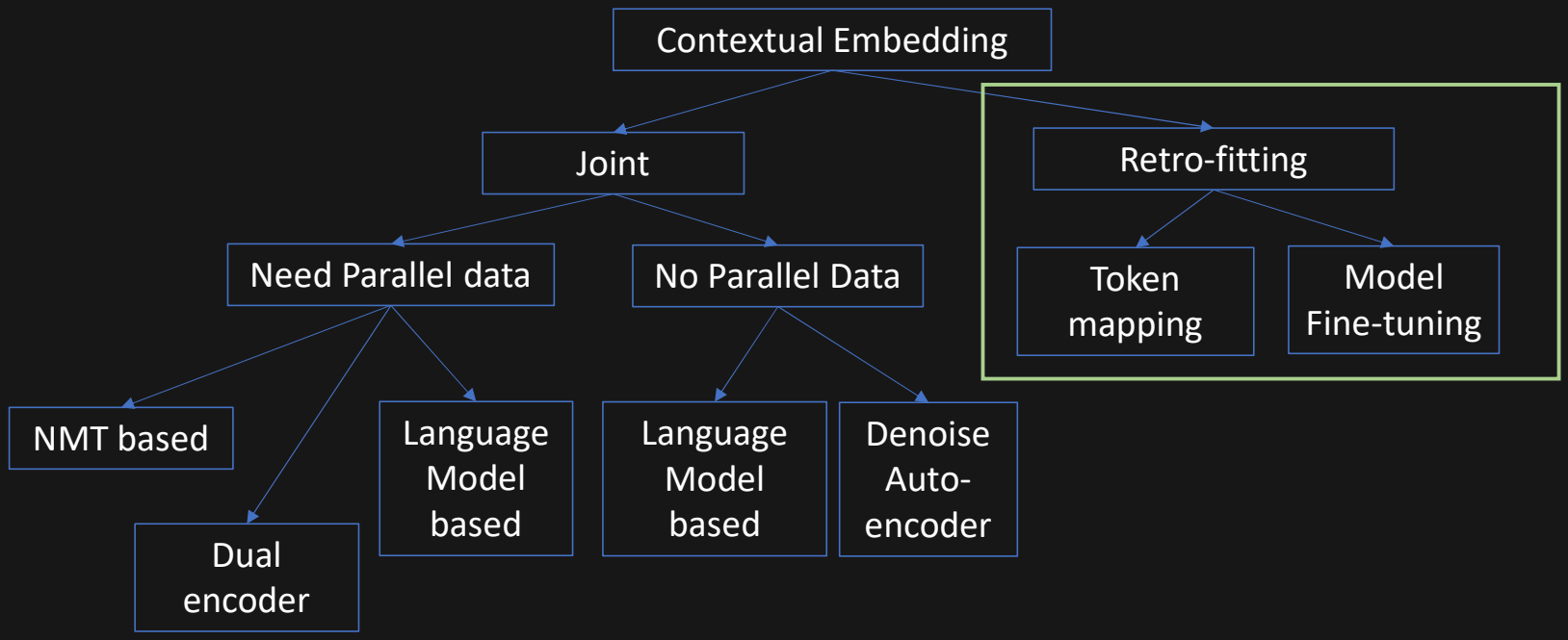
Multi-Lingual Bidirectional and Auto-Regressive Transformers (mBART)



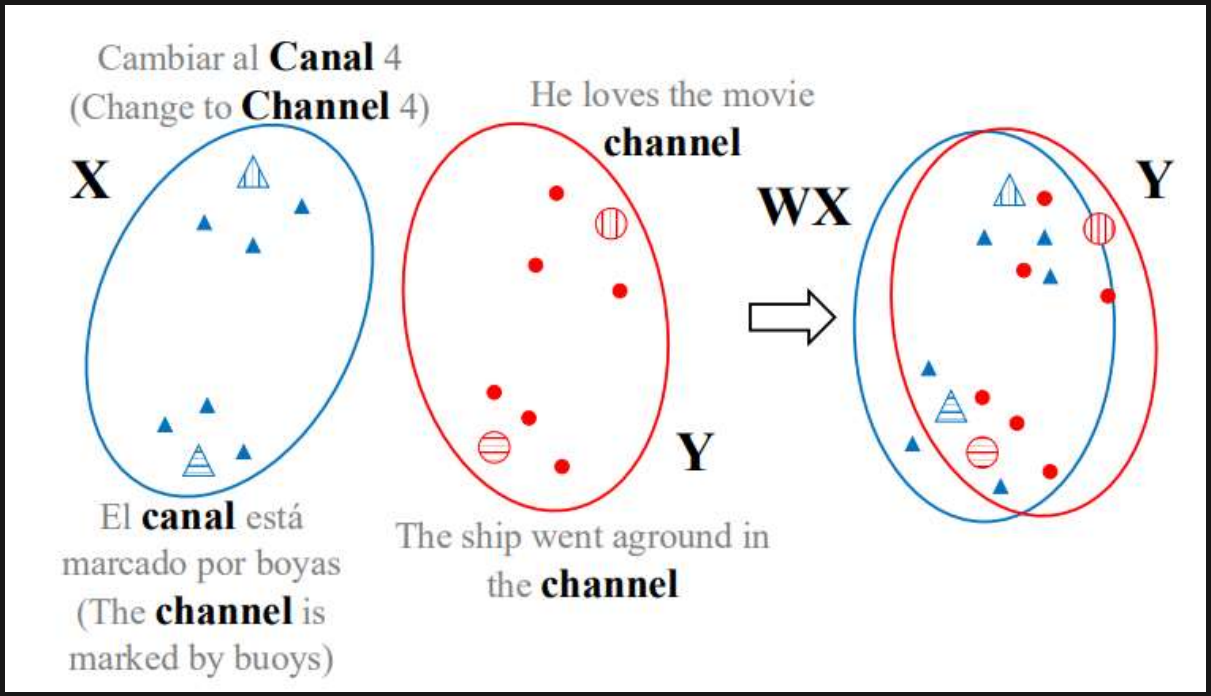
- Same architecture and training scheme with monolingual BART
- 25 languages crawled web pages
- 250K shared Sentence Piece vocabulary

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



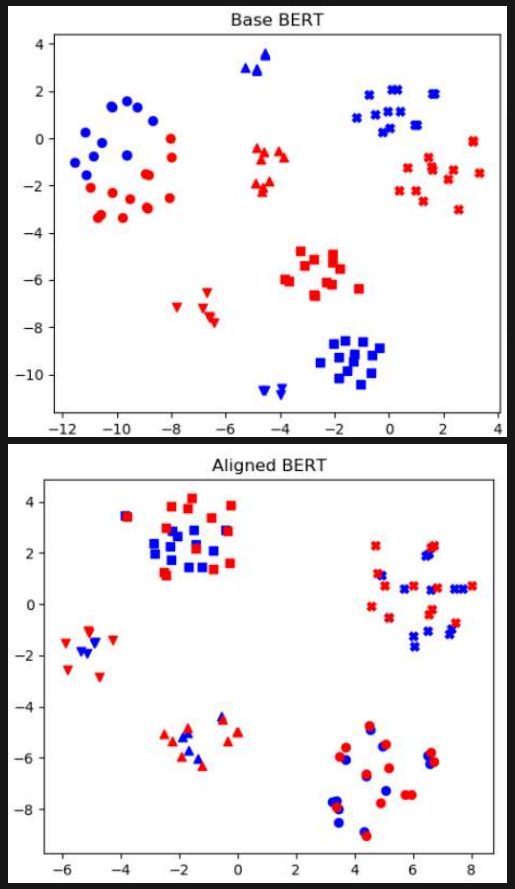
Contextual token mapping



- 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

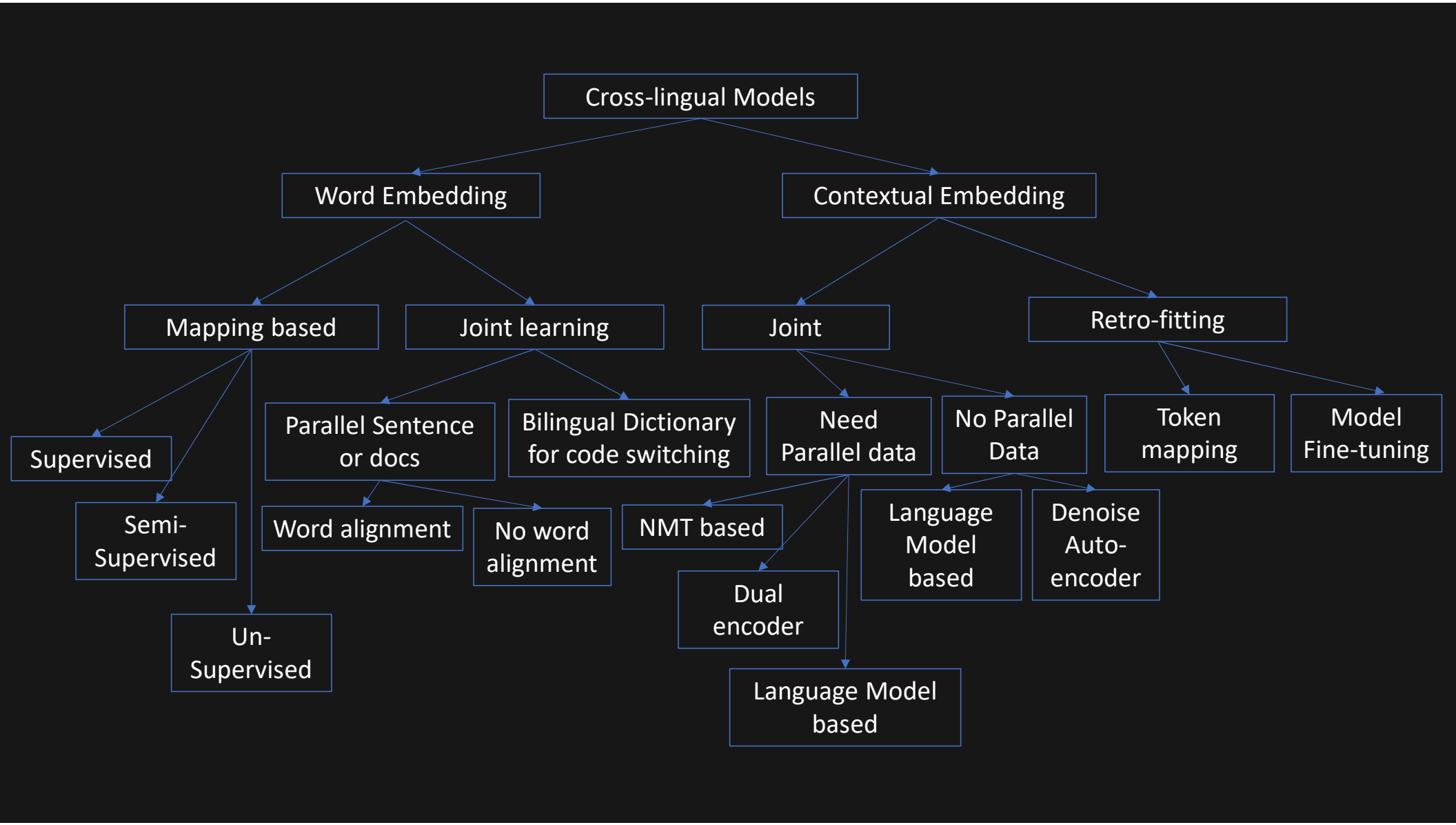
$$\mathcal{L}(f, C) = -\sum_{(s,t)\in C}\sum_{(i,j)\in a(s,t)} \|f(i,s) - f(j,t)\|^2$$

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

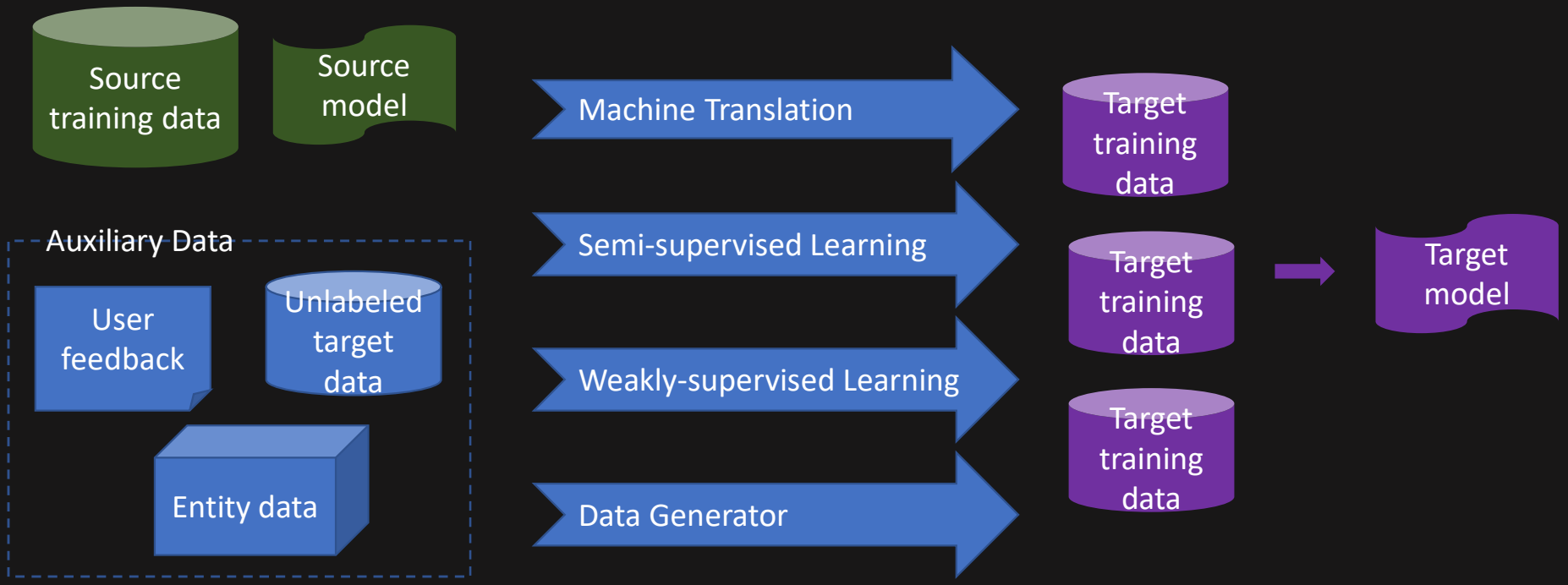
$$R(f, C) = \sum_{t\in C}\sum_{i=1}^{\text{len}(t)} \|f(i,t) - f_0(i,t)\|^2$$

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.

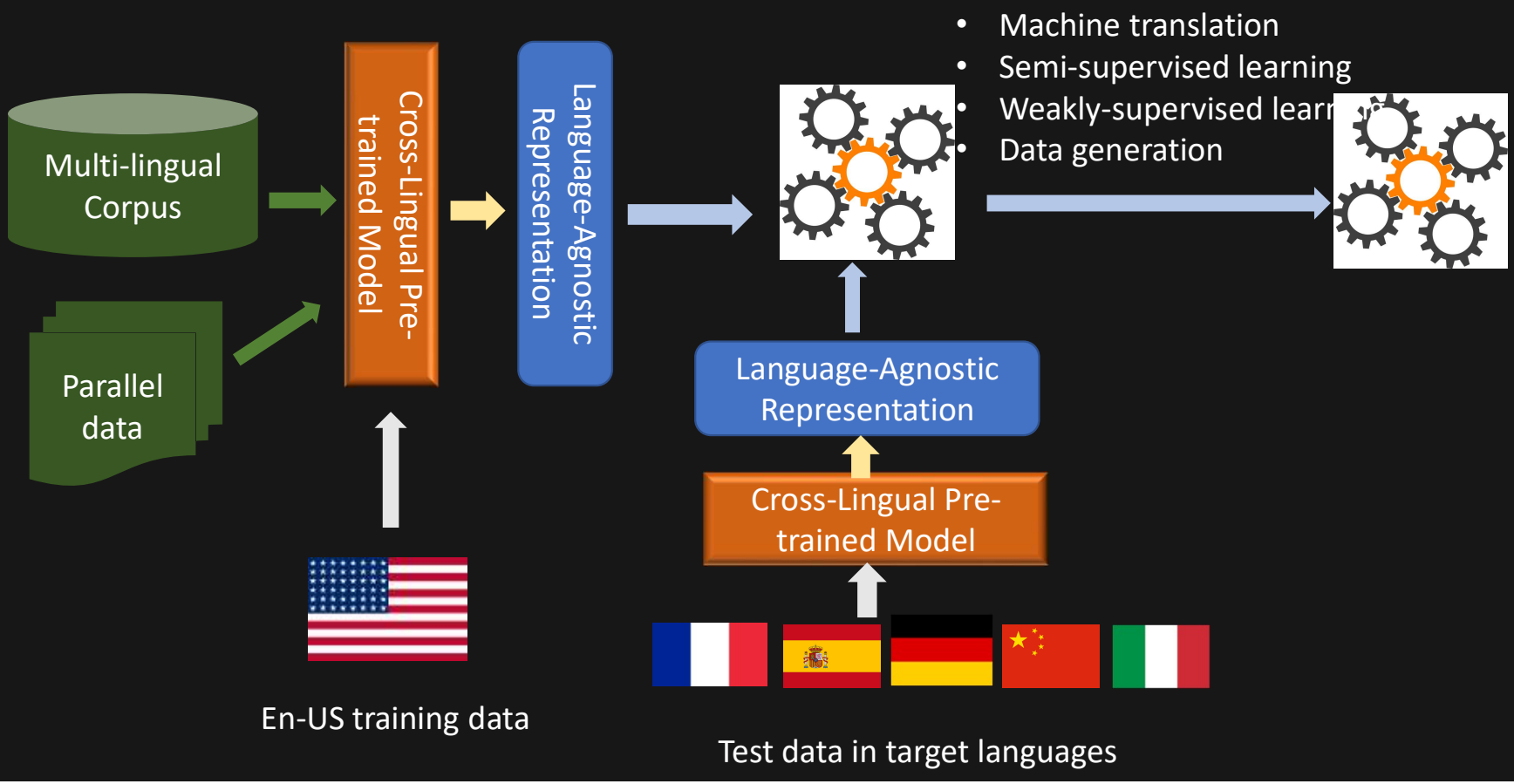

Cross-lingual Models
Word Embedding
Contextual Embedding
Mapping based
Joint learning
Joint
Retro-fitting
Supervised
Semi-Supervised
Un-Supervised
Parallel Sentence or docs
Bilingual Dictionary for code switching
Word alignment
No word alignment
Need Parallel data
No Parallel Data
Token mapping
Model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

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



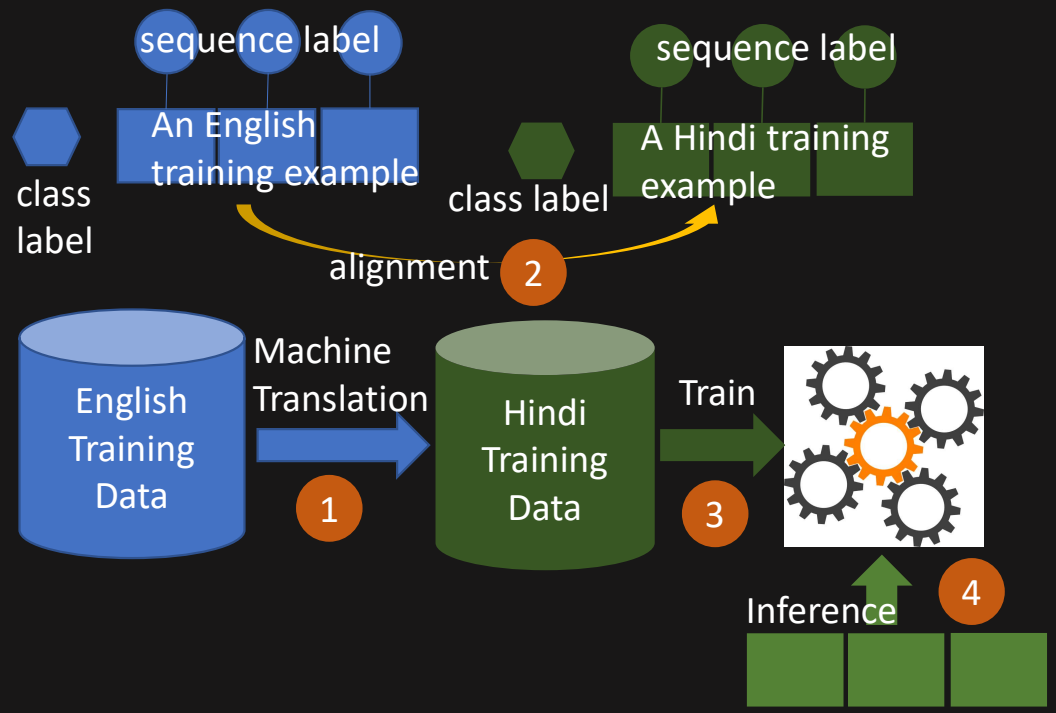
Approaches: Model Transfer and Data Transfer

(1) Model Transfer + (2) Data Transfer

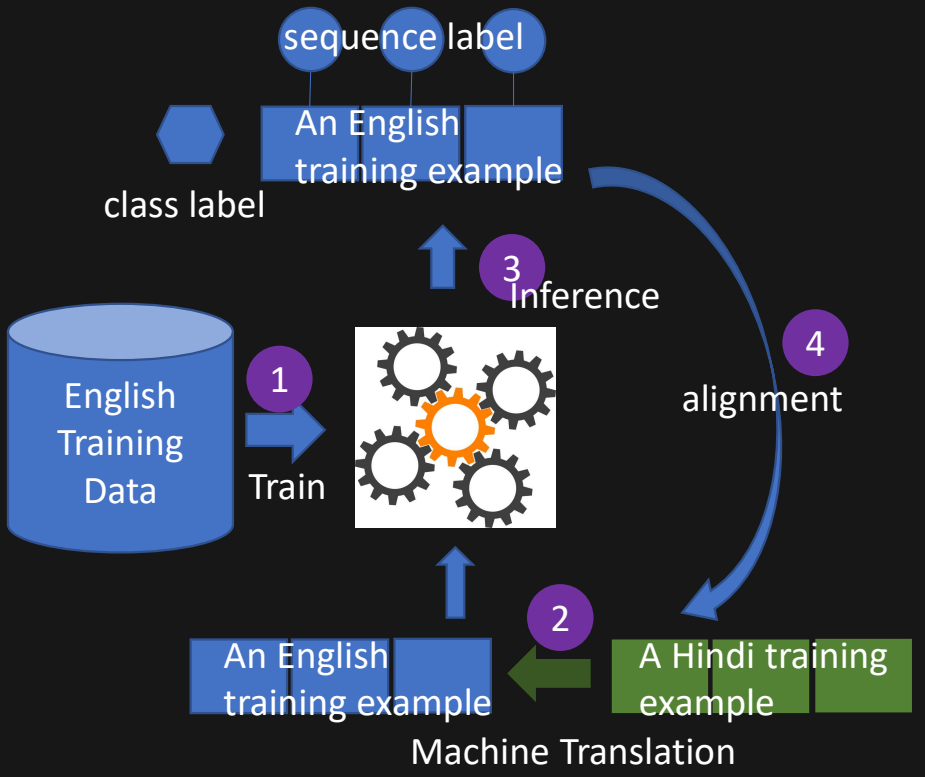


Machine translation

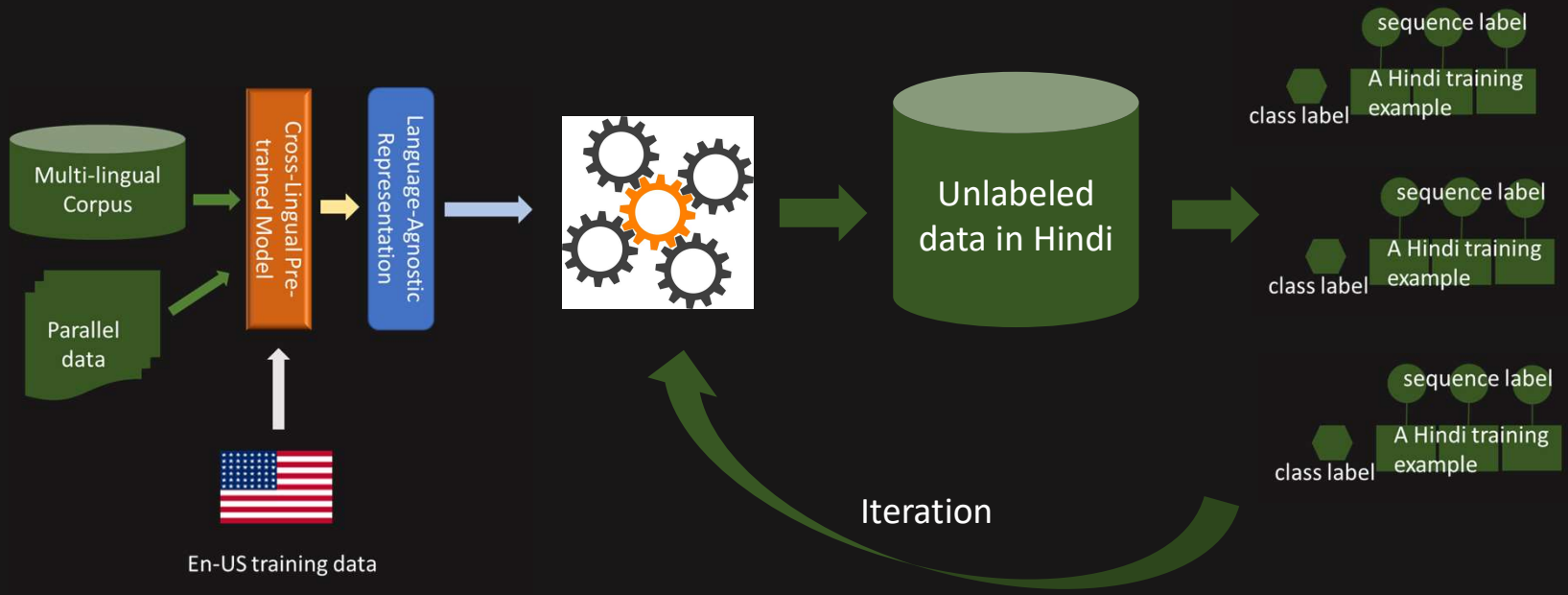
- Translate-on-train



- Translate-on-test



Semi-supervised learning

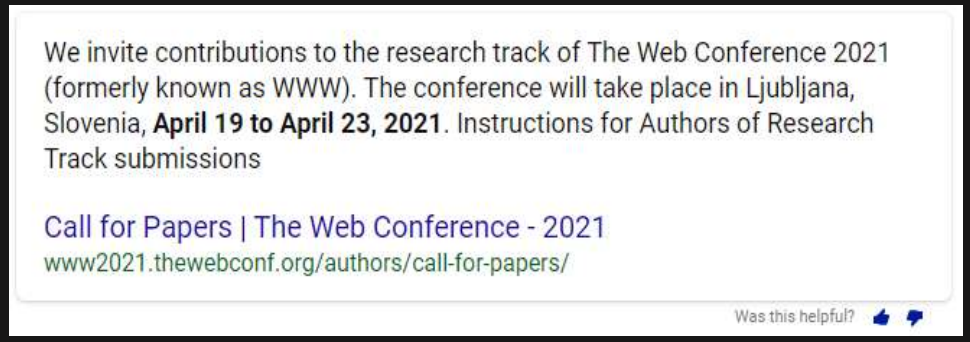


1. Build an initial model by model transfer approach
2. Apply the initial model on unlabeled data to get labels
3. Use the labeled data to improve the initial model
4. Iterate the above steps 2-3 for several rounds (self learning, model ensemble, reinforcement learning)

Weakly-supervised learning

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

Q: {*what is the date for the web conference 2021*}



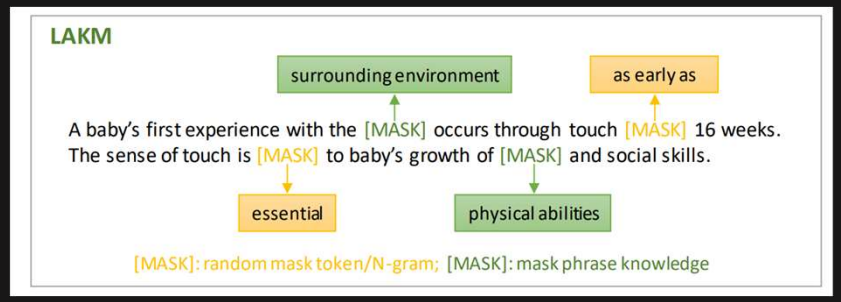
- Rich user behaviors in search logs
- Any single behavior is a weak signal, but more powerful when combined in a ML model
- Generally applicable to all languages

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)

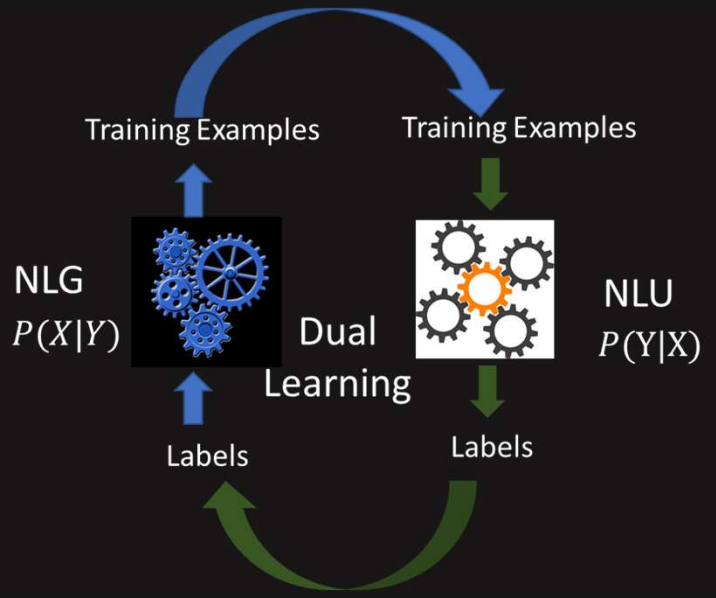
Intent: check_status

Departure airport

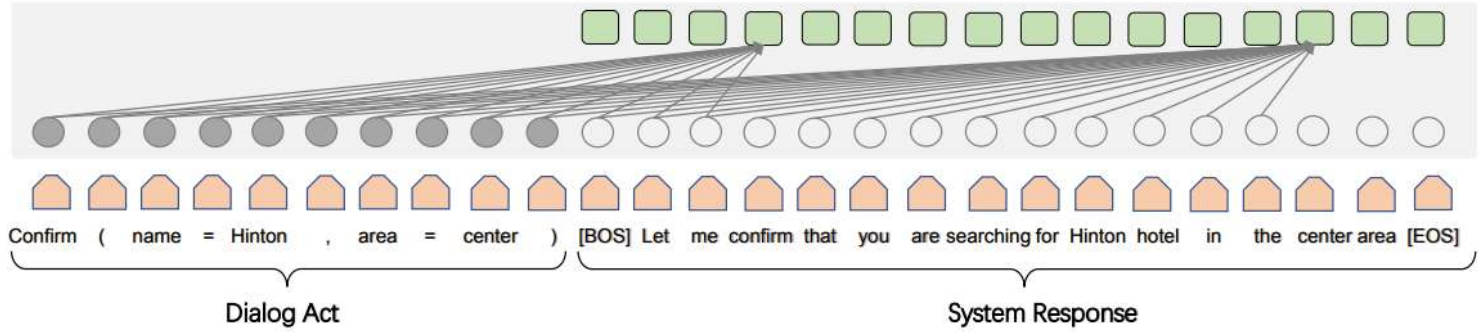
Arrival airport

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.

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

NLP Group
Software Technology Center at Asia (STCA) of Microsoft

- Introduction [Daxin Jiang]
  - Motivating examples in Microsoft products
  - Problem description
  - Categorization of applications
  - Challenges and major approaches
- Applications
  - **Natural Language Inference [Linjun Shou]**
  - Information retrieval [Xiubo Geng]
  - Machine Reading Comprehension [Ming Gong]
- Future directions [Daxin Jiang]

Linjun Shou

Xiubo Geng

Ming Gong

Cross Lingual Natural Language Inference

Linjun Shou
lisho@microsoft.com

Cross-lingual Natural Language Inference

Microsoft

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
NLG	Text Generation	Token level	Spelling correction, Sentence auto completion
		Sentence level	Machine translation, Conversation, Question generation

Cross-lingual Natural Language Inference

- Task Definition:
  - Given *Premise and Hypothesis*, predict *Entailment/Neural/Contradiction*.
- Benchmark Datasets
  - *XNLI*: Conneau, Alexis et al. "XNLI: Evaluating Cross-lingual Sentence Representations." EMNLP (2018).

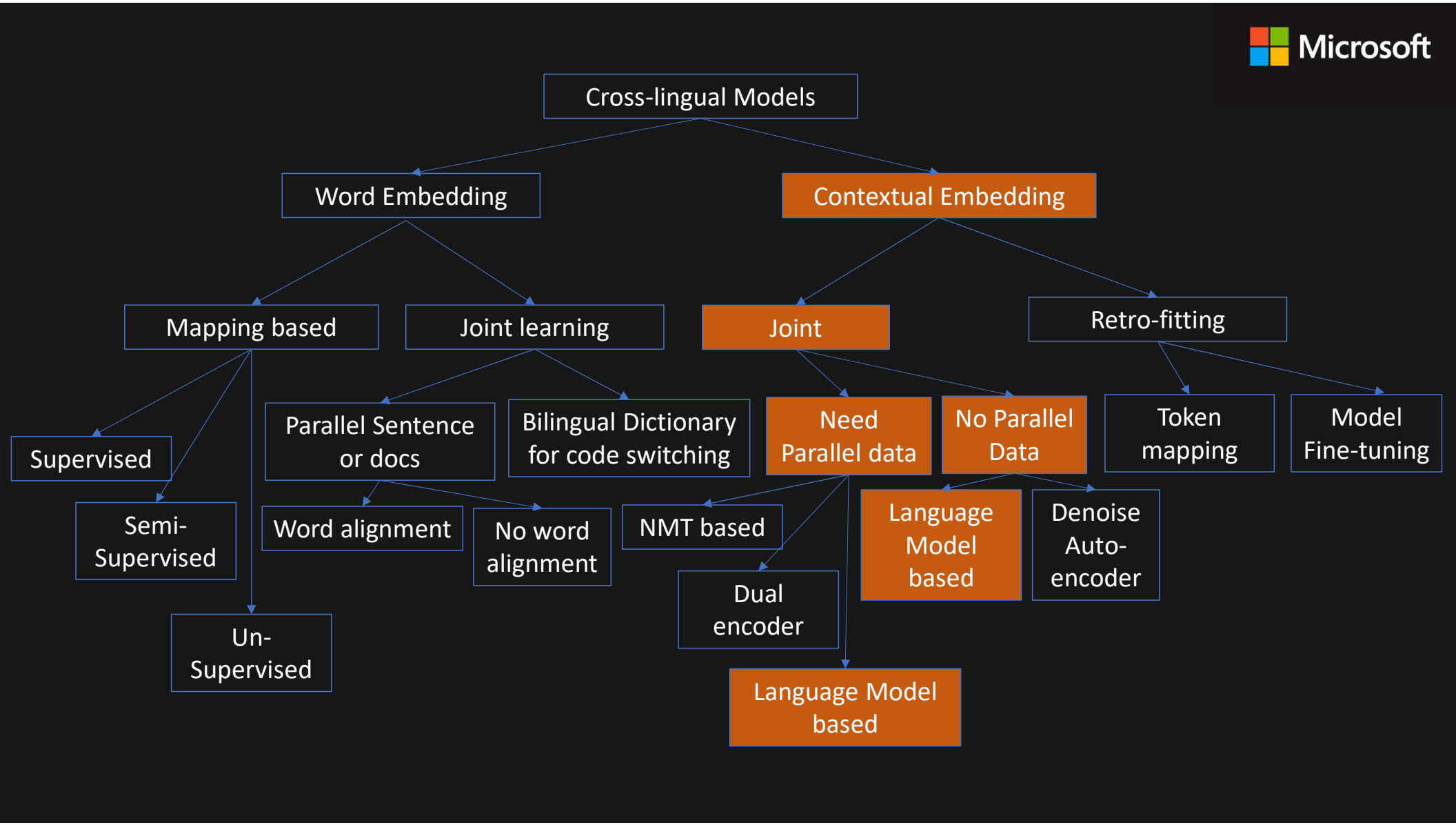
Language	Premise / Hypothesis	Genre	Label
English	You don't have to stay there. You can leave.	Face-To-Face	Entailment
French	La figure 4 montre la courbe d'offre des services de partage de travaux. Les services de partage de travaux ont une offre variable.	Government	Entailment
Spanish	Y se estremeció con el recuerdo. El pensamiento sobre el acontecimiento hizo su estremecimiento.	Fiction	Entailment
German	Während der Depression war es die ärmste Gegend, kurz vor dem Hungertod. Die Weltwirtschaftskrise dauerte mehr als zehn Jahre an.	Travel	Neutral
Swahili	Ni silaha ya plastiki ya moja kwa moja inayopiga risasi. Inadumu zaidi kuliko silaha ya chuma.	Telephone	Neutral
Russian	И мы занимаемся этим уже на протяжении 85 лет. Мы только начали этим заниматься.	Letters	Contradiction
Chinese	让我告诉你，美国人最终如何看待你作为独立顾问的表现。 美国人完全不知道您是独立律师。	Slate	Contradiction
Arabic	تحتاج الوكالات لأن تكون قادرة على قياس مستويات النجاح. لا يمكنالوكالات أ اتعرف ما إذا كانت ناجحة أم لا	Nine-Eleven	Contradiction

Table 1: Examples (premise and hypothesis) from various languages and genres from the XNLI corpus.

Approaches for Cross-lingual Transfer

- **Model Transfer**
  - w/ monolingual corpus
  - w/ parallel corpus
- **Data Transfer**
  - Using machine translation
  - Code Switch data construction

Representative Works


Microsoft
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

Multi-lingual Sentence Encoders

- Use parallel corpus for sentence embedding alignment

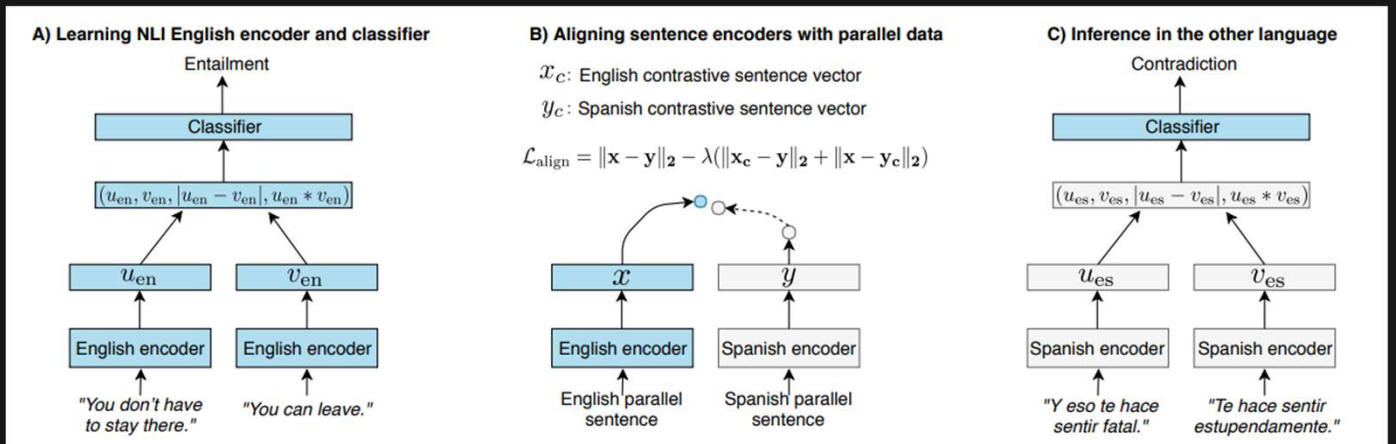


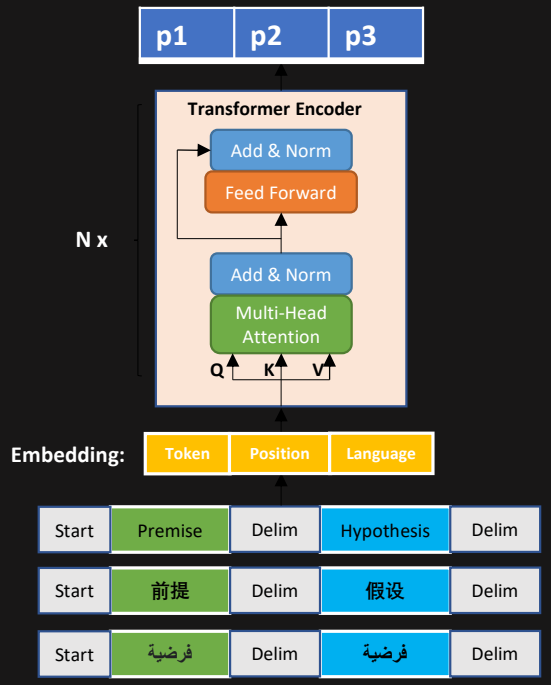
Figure 1: **Illustration of language adaptation by sentence embedding alignment.** A) The English encoder and classifier in blue are learned on English (*in-domain*) NLI data. The encoder can also be pretrained (*transfer learning*). B) The Spanish encoder in gray is trained to mimic the English encoder using parallel data. C) After alignment of the encoders, the classifier can make predictions for Spanish.

Conneau, Alexis et al. "XNLI: Evaluating Cross-lingual Sentence Representations." EMNLP (2018).

Multi-lingual BERT – Multi-lingual Pretrained Model

Transfer Text Pair into Single Text Classification

- Same architecture and training scheme with monolingual BERT
- MLM pretraining on 104 languages Wikipedia pages
- 110k shared Word-Piece vocabulary



Conneau, Alexis et al. “XNLI: Evaluating Cross-lingual Sentence Representations.” EMNLP (2018).

XLM - Cross-lingual Language Model Pretraining

- Translation Language Model: Leverage Parallel data for alignment pretraining

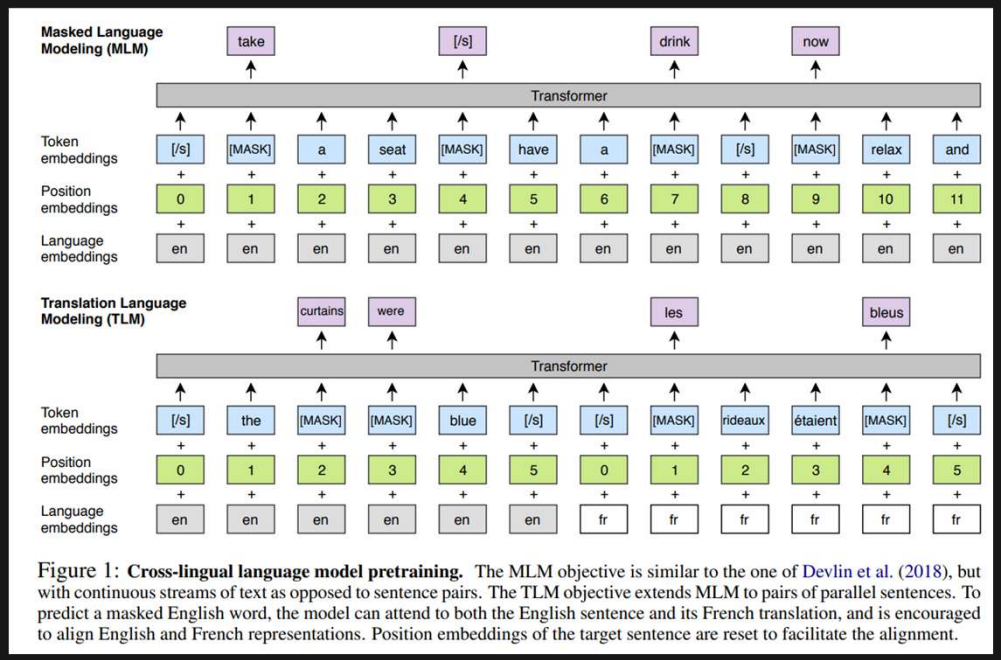


Figure 1: **Cross-lingual language model pretraining.** The MLM objective is similar to the one of Devlin et al. (2018), but with continuous streams of text as opposed to sentence pairs. The TLM objective extends MLM to pairs of parallel sentences. To predict a masked English word, the model can attend to both the English sentence and its French translation, and is encouraged to align English and French representations. Position embeddings of the target sentence are reset to facilitate the alignment.

Lample, Guillaume and Alexis Conneau. "Cross-lingual Language Model Pretraining." NeurIPS (2019).

XLM - Cross-lingual Language Model Pretraining

- Translation Language Model: Leverage Parallel data for alignment pretraining

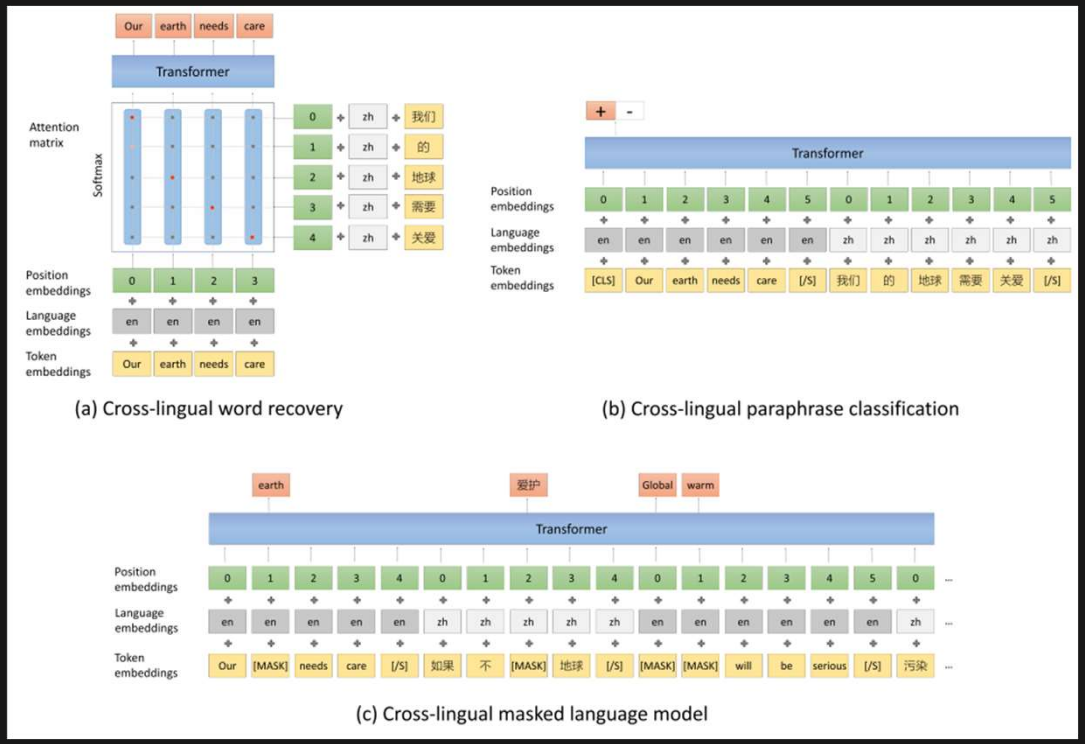
	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Δ
Machine translation baselines (TRANSLATE-TRAIN)																
Devlin et al. (2018)	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
XLM (MLM+TLM)	85.0	80.2	80.8	80.3	78.1	79.3	78.1	74.7	76.5	76.6	75.5	78.6	72.3	70.9	63.2	76.7
Machine translation baselines (TRANSLATE-TEST)																
Devlin et al. (2018)	81.4	-	74.9	74.4	-	-	-	-	70.4	-	-	70.1	-	-	62.1	-
XLM (MLM+TLM)	85.0	79.0	79.5	78.1	77.8	77.6	75.5	73.7	73.7	70.8	70.4	73.6	69.0	64.7	65.1	74.2
Evaluation of cross-lingual sentence encoders																
Conneau et al. (2018b)	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4	65.6
Devlin et al. (2018)	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
Artetxe and Schwenk (2018)	73.9	71.9	72.9	72.6	73.1	74.2	71.5	69.7	71.4	72.0	69.2	71.4	65.5	62.2	61.0	70.2
XLM (MLM)	83.2	76.5	76.3	74.2	73.1	74.0	73.1	67.8	68.5	71.2	69.2	71.9	65.7	64.6	63.4	71.5
XLM (MLM+TLM)	**85.0**	**78.7**	**78.9**	**77.8**	**76.6**	**77.4**	**75.3**	**72.5**	**73.1**	**76.1**	**73.2**	**76.5**	**69.6**	**68.4**	**67.3**	**75.1**

Table 1: **Results on cross-lingual classification accuracy.** Test accuracy on the 15 XNLI languages. We report results for machine translation baselines and zero-shot classification approaches based on cross-lingual sentence encoders. XLM (MLM) corresponds to our unsupervised approach trained only on monolingual corpora, and XLM (MLM+TLM) corresponds to our supervised method that leverages both monolingual and parallel data through the TLM objective. Δ corresponds to the average accuracy.

Lample, Guillaume and Alexis Conneau. “Cross-lingual Language Model Pretraining.” NeurIPS (2019).

Unicoder

- Leverage Parallel data for alignment pretraining
  - MMLM
  - TLM
  - Cross lingual word recovery
  - Cross-lingual paraphrase classification



Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Ming Zhou. Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks. EMNLP, 2019.

Unicoder

- Leverage Parallel data for alignment pretraining
  - MMLM
  - TLM
  - Cross lingual word recovery
  - Cross-lingual paraphrase classification

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	average
Machine translate at training (TRANSLATE-TRAIN)																
Conneau et al. (2018)	73.7	68.3	68.8	66.5	66.4	67.4	66.5	64.5	65.8	66.0	62.8	67.0	62.1	58.2	56.6	65.4
Multilingual BERT (Devlin et al., 2018)	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
Multilingual BERT from Wu and Dredze 2019	82.1	76.9	78.5	74.8	72.1	75.4	74.3	70.6	70.8	67.8	63.2	76.2	65.3	65.3	60.6	71.6
XLM (Lample and Conneau, 2019)	85.0	80.2	80.8	80.3	78.1	79.3	78.1	74.7	76.5	76.6	75.5	78.6	72.3	70.9	63.2	76.7
Unicoder	85.1	80.0	81.1	79.9	77.7	80.2	77.9	75.3	76.7	76.4	75.2	79.4	71.8	71.8	64.5	76.9
Machine translate at test (TRANSLATE-TEST)																
Conneau et al. (2018)	73.7	70.4	70.7	68.7	69.1	70.4	67.8	66.3	66.8	66.5	64.4	68.3	64.2	61.8	59.3	67.2
Multilingual BERT (Devlin et al., 2018)	81.4	-	74.9	74.4	-	-	-	-	70.4	-	-	70.1	-	-	62.1	-
XLM (Lample and Conneau, 2019)	85.0	79.0	79.5	78.1	77.8	77.6	75.5	73.7	73.7	70.8	70.4	73.6	69.0	64.7	65.1	74.2
Unicoder	85.1	80.1	80.3	78.2	77.5	78.0	76.2	73.3	73.9	72.8	71.6	74.1	70.3	65.2	66.3	74.9
Evaluation of cross-lingual sentence encoders (Cross-lingual TEST)																
Conneau et al. (2018)	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4	65.6
Multilingual BERT (Devlin et al., 2018)	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
Multilingual BERT from Wu and Dredze 2019	82.1	73.8	74.3	71.1	66.4	68.9	69	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
Artetxe and Schwenk (2018)	73.9	71.9	72.9	72.6	73.1	74.2	71.5	69.7	71.4	72.0	69.2	71.4	65.5	62.2	61.0	70.2
XLM (Lample and Conneau, 2019)	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
Unicoder	85.1	79.0	79.4	77.8	77.2	77.2	76.3	72.8	73.5	76.4	73.6	76.2	69.4	69.7	66.7	75.4
Multi-language Fine-tuning																
XLM (Lample and Conneau, 2019)	85.0	80.8	81.3	80.3	79.1	80.9	78.3	75.6	77.6	78.5	76.0	79.5	72.9	72.8	68.5	77.8
Unicoder w/o Word Recovery	85.2	80.5	81.8	80.9	79.7	81.1	79.3	76.2	78.2	78.5	76.4	79.7	73.4	73.6	68.8	78.2
Unicoder w/o Paraphrase Classification	85.5	81.1	82.0	81.1	80.0	81.3	79.6	76.6	78.2	78.2	75.9	79.9	73.7	74.2	69.3	78.4
Unicoder w/o Cross-lingual Language Model	85.5	81.9	81.8	80.5	80.5	81.0	79.3	76.4	78.1	78.3	76.3	79.6	72.9	73.0	68.7	78.3
Unicoder	**85.6**	**81.1**	**82.3**	**80.9**	**79.5**	**81.4**	**79.7**	**76.8**	**78.2**	**77.9**	**77.1**	**80.5**	**73.4**	**73.8**	**69.6**	**78.5**

Table 2: Test accuracy on the 15 XNLI languages. This table is organized by fine-tuning and test approaches. TRANSLATE-TRAIN is to machine translate English training data to target language and fine-tune with this translated data; TRANSLATE-TEST is machine translate target language test data to English, the fine-tuning is conducted on English; Cross-lingual TEST is to fine-tune on English and directly test on target language; Multi-language Fine-tune is to fine-tune on machine translated training data on all languages.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Ming Zhou. Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks. EMNLP, 2019.

XLM-RoBERTa

- MLM
- Bigger data covering 100 languages
- Sentence piece for tokenization

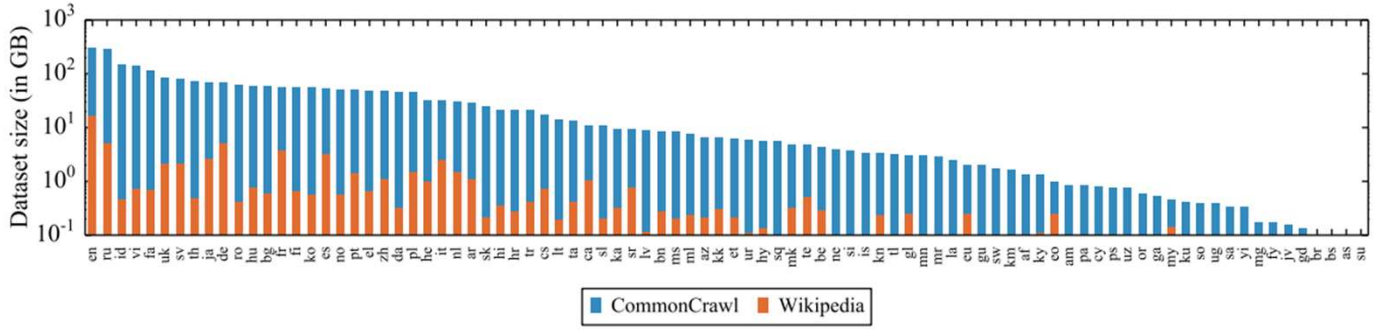


Figure 1: Amount of data in GiB (log-scale) for the 88 languages that appear in both the Wiki-100 corpus used for mBERT and XLM-100, and the CC-100 used for XLM-R. CC-100 increases the amount of data by several orders of magnitude, in particular for low-resource languages.

Conneau, Alexis et al. “Unsupervised Cross-lingual Representation Learning at Scale.” ACL (2020).

XLM-RoBERTa

Model	D	#M	#lg	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Fine-tune multilingual model on English training set (Cross-lingual Transfer)																			
Lample and Conneau (2019)	Wiki+MT	N	15	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
Huang et al. (2019)	Wiki+MT	N	15	85.1	79.0	79.4	77.8	77.2	77.2	76.3	72.8	73.5	76.4	73.6	76.2	69.4	69.7	66.7	75.4
Devlin et al. (2018)	Wiki	N	102	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
Lample and Conneau (2019)	Wiki	N	100	83.7	76.2	76.6	73.7	72.4	73.0	72.1	68.1	68.4	72.0	68.2	71.5	64.5	58.0	62.4	71.3
Lample and Conneau (2019)	Wiki	1	100	83.2	76.7	77.7	74.0	72.7	74.1	72.7	68.7	68.6	72.9	68.9	72.5	65.6	58.2	62.4	70.7
XLM-R$_{\text{Base}}$	CC	1	100	85.8	79.7	80.7	78.7	77.5	79.6	78.1	74.2	73.8	76.5	74.6	76.7	72.4	66.5	68.3	76.2
XLM-R	CC	1	100	**89.1**	**84.1**	**85.1**	**83.9**	**82.9**	**84.0**	**81.2**	**79.6**	**79.8**	**80.8**	**78.1**	**80.2**	**76.9**	**73.9**	**73.8**	**80.9**
Translate everything to English and use English-only model (TRANSLATE-TEST)																			
BERT-en	Wiki	1	1	88.8	81.4	82.3	80.1	80.3	80.9	76.2	76.0	75.4	72.0	71.9	75.6	70.0	65.8	65.8	76.2
RoBERTa	Wiki+CC	1	1	**91.3**	82.9	84.3	81.2	81.7	83.1	78.3	76.8	76.6	74.2	74.1	77.5	70.9	66.7	66.8	77.8
Fine-tune multilingual model on each training set (TRANSLATE-TRAIN)																			
Lample and Conneau (2019)	Wiki	N	100	82.9	77.6	77.9	77.9	77.1	75.7	75.5	72.6	71.2	75.8	73.1	76.2	70.4	66.5	62.4	74.2
Fine-tune multilingual model on all training sets (TRANSLATE-TRAIN-ALL)																			
Lample and Conneau (2019)†	Wiki+MT	1	15	85.0	80.8	81.3	80.3	79.1	80.9	78.3	75.6	77.6	78.5	76.0	79.5	72.9	72.8	68.5	77.8
Huang et al. (2019)	Wiki+MT	1	15	85.6	81.1	82.3	80.9	79.5	81.4	79.7	76.8	78.2	77.9	77.1	80.5	73.4	73.8	69.6	78.5
Lample and Conneau (2019)	Wiki	1	100	84.5	80.1	81.3	79.3	78.6	79.4	77.5	75.2	75.6	78.3	75.7	78.3	72.1	69.2	67.7	76.9
XLM-R$_{\text{Base}}$	CC	1	100	85.4	81.4	82.2	80.3	80.4	81.3	79.7	78.6	77.3	79.7	77.9	80.2	76.1	73.1	73.0	79.1
XLM-R	CC	1	100	**89.1**	**85.1**	**86.6**	**85.7**	**85.3**	**85.9**	**83.5**	**83.2**	**83.1**	**83.7**	**81.5**	**83.7**	**81.6**	**78.0**	**78.1**	**83.6**

Table 1: **Results on cross-lingual classification.** We report the accuracy on each of the 15 XNLI languages and the average accuracy. We specify the dataset D used for pretraining, the number of models #M the approach requires and the number of languages #lg the model handles. Our *XLM-R* results are averaged over five different seeds. We show that using the translate-train-all approach which leverages training sets from multiple languages, *XLM-R* obtains a new state of the art on XNLI of 83.6% average accuracy. Results with † are from Huang et al. (2019).

Conneau, Alexis et al. “Unsupervised Cross-lingual Representation Learning at Scale.” ACL (2020).

InfoXLM

- Leverage Parallel data for alignment pretraining
  - MMLM – Multilingual MLLM
  - TLM – Translation LM
  - XLCO – Cross lingual contrastive learning

$$\mathcal{L} = \mathcal{L}_{\text{MMLM}} + \mathcal{L}_{\text{TL}} \quad + \mathcal{L}_{\text{XLCO}}$$

$$\mathcal{L}_{\text{XLCO}} = -\log \frac{\exp(g\boldsymbol{\theta}_Q(c_1)^T g\boldsymbol{\theta}_K(c_2))}{\sum_{c' \in \mathcal{N}} \exp(g\boldsymbol{\theta}_Q(c_1)^T g\boldsymbol{\theta}_K(c'))}$$

Chi, Zewen et al. “InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training.” ArXiv abs/2007.07834 (2020): n. pag.

InfoXLM

- Leverage Parallel data for alignment pretraining

Models	#M	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Fine-tune multilingual model on English training set (Cross-lingual Transfer)																	
MBERT*	N	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
XLM (w/o TLM)*	N	83.7	76.2	76.6	73.7	72.4	73.0	72.1	68.1	68.4	72.0	68.2	71.5	64.5	58.0	62.4	71.3
XLM*	N	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
XLM (w/o TLM)*	1	83.2	76.7	77.7	74.0	72.7	74.1	72.7	68.7	68.6	72.9	68.9	72.5	65.6	58.2	62.4	70.7
UNICODER	1	85.4	79.2	79.8	78.2	77.3	78.5	76.7	73.8	73.9	75.9	71.8	74.7	70.1	67.4	66.3	75.3
XLM-R*	1	85.8	79.7	80.7	78.7	77.5	79.6	78.1	74.2	73.8	76.5	74.6	76.7	72.4	66.5	68.3	**76.2**
XLM-R (reimpl)	1	84.7	79.1	79.4	77.4	76.6	78.4	76.0	73.5	72.6	75.5	73.0	74.5	71.0	65.7	67.6	75.0
INFOXLM	1	86.4	80.6	80.8	78.9	77.8	78.9	77.6	75.6	74.0	77.0	73.7	76.7	72.0	66.4	67.1	**76.2**
XLM-R_{LARGE}*	1	89.1	84.1	85.1	83.9	82.9	84.0	81.2	79.6	79.8	80.8	78.1	80.2	76.9	73.9	73.8	80.9
XLM-R_{LARGE} (reimpl)	1	88.9	83.6	84.8	83.1	82.4	83.7	80.7	79.2	79.0	80.4	77.8	79.8	76.8	72.7	73.3	80.4
INFOXLM$_{LARGE}$	1	89.7	84.5	85.5	84.1	83.4	84.2	81.3	80.9	80.4	80.8	78.9	80.9	77.9	74.8	73.7	**81.4**
Fine-tune multilingual model on all training sets (Translate-Train-All)																	
XLM (w/o TLM)*	1	84.5	80.1	81.3	79.3	78.6	79.4	77.5	75.2	75.6	78.3	75.7	78.3	72.1	69.2	67.7	76.9
XLM*	1	85.0	80.8	81.3	80.3	79.1	80.9	78.3	75.6	77.6	78.5	76.0	79.5	72.9	72.8	68.5	77.8
XLM-R*	1	85.4	81.4	82.2	80.3	80.4	81.3	79.7	78.6	77.3	79.7	77.9	80.2	76.1	73.1	73.0	79.1
XLM-R (reimpl)	1	85.0	81.0	81.9	80.6	79.7	81.4	79.5	77.7	77.3	79.5	77.5	79.1	75.3	72.2	70.9	78.6
INFOXLM	1	86.1	82.0	82.8	81.8	80.9	82.0	80.2	79.0	78.8	80.5	78.3	80.5	77.4	73.0	71.6	**79.7**

Chi, Zewen et al. "InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training." ArXiv abs/2007.07834 (2020): n. pag.

Code-Switch Data Augmentation

- Construct multi-lingual Code-switch data for augmentation

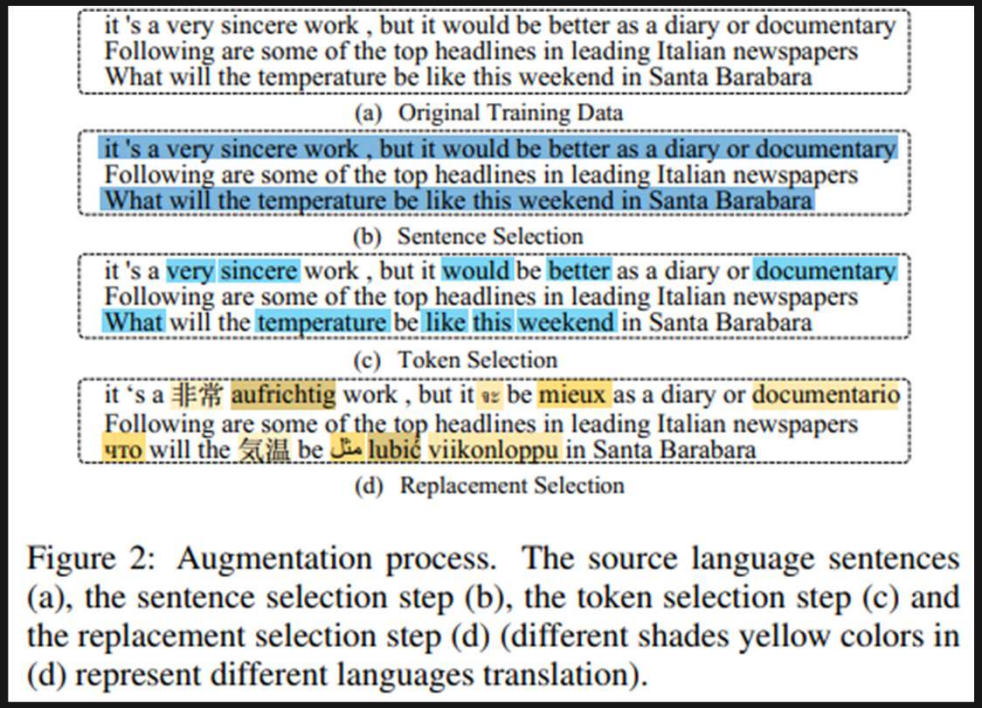


Figure 2: Augmentation process. The source language sentences (a), the sentence selection step (b), the token selection step (c) and the replacement selection step (d) (different shades yellow colors in (d) represent different languages translation).

Qin, L. et al. “CoSDA-ML: Multi-Lingual Code-Switching Data Augmentation for Zero-Shot Cross-Lingual NLP.” IJCAI (2020).

Code-Switch Data Augmentation

- Construct multi-lingual Code-switch data for augmentation

Model	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Average
Artetxe and Schwenk [2018]	73.9	71.9	72.9	72.6	73.1	74.2	71.5	69.7	71.4	72.0	69.2	71.4	65.5	62.2	61.0	70.2
XLM [Conneau and Lample, 2019]	84.1	77.1	78.0	75.0	74.1	75.1	72.4	70.0	70.6	71.5	68.3	73.2	66.7	67.5	62.2	72.4
+CoSDA-ML	**84.4***	**79.0***	**79.2***	**77.9***	**76.8***	**77.6***	**75.7***	**72.6***	**73.4***	**75.3***	**72.6***	**75.1***	**71.2***	**70.0***	**68.3***	**75.3***
mBERT from Wu and Dredze [2019]	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
+CoSDA-ML	82.9	76.7	76.9	74.1	70.9	72.7	73.2	63.9	68.0	73.6	59.8	73.8	65.5	51.0	62.3	69.7

Table 1: Natural Language Inference experiments.

Qin, L. et al. “CoSDA-ML: Multi-Lingual Code-Switching Data Augmentation for Zero-Shot Cross-Lingual NLP.” IJCAI (2020).

Key Takeaway

- Parallel corpus helps language transfer
- Bigger model leads to better performance
- Larger data leads to better performance

Outline

NLP Group
Software Technology Center at Asia (STCA) of Microsoft

- Introduction [Daxin Jiang]
  - Motivating examples in Microsoft products
  - Problem description
  - Categorization of applications
  - Challenges and major approaches
- Applications
  - Natural Language Inference [Linjun Shou]
  - **Information retrieval [Xiubo Geng]**
  - Machine Reading Comprehension [Ming Gong]
- Future directions [Daxin Jiang]

Linjun Shou

Xiubo Geng

Ming Gong

Cross Lingual Semantic Retrieval

Xiubo Geng
xigeng@microsoft.com

Cross-lingual Semantic Retrieval

- Introduction
- Cross-lingual Models
  - Base architecture
    - Dual encoder
    - NMT based model
  - Advanced approaches
    - Better loss function for training
    - Better scoring function for inference
- Data Augmentation
  - Using both monolingual and parallel data
  - Using only monolingual data

Microsoft

Introduction

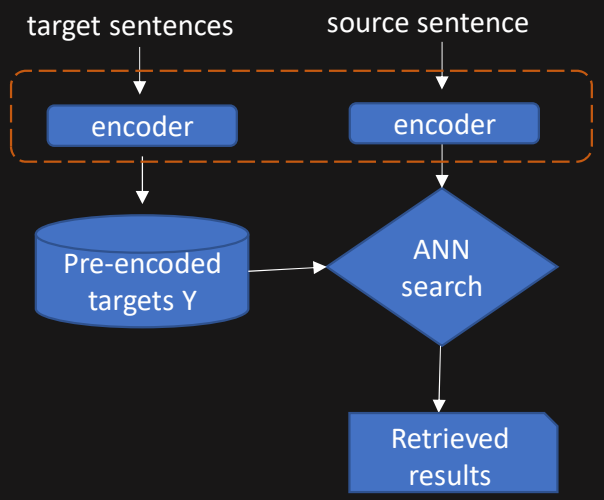
- Cross-lingual Semantic Retrieval
  - Given a sentence, retrieve relevant sentences from a large candidate pool in other languages
- Application Scenarios
  - Parallel corpus mining
    - Given a sentence in a language, find its translations in another language from a large corpus
  - Cross-lingual question retrieval
    - Given a question in a language, find questions with similar meaning from a candidate pool in another language

Cross-lingual Semantic Retrieval

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
NLG	Text Generation	Token level	Spelling correction, Sentence auto completion
		Sentence level	Machine translation, Conversation, Question generation

Introduction

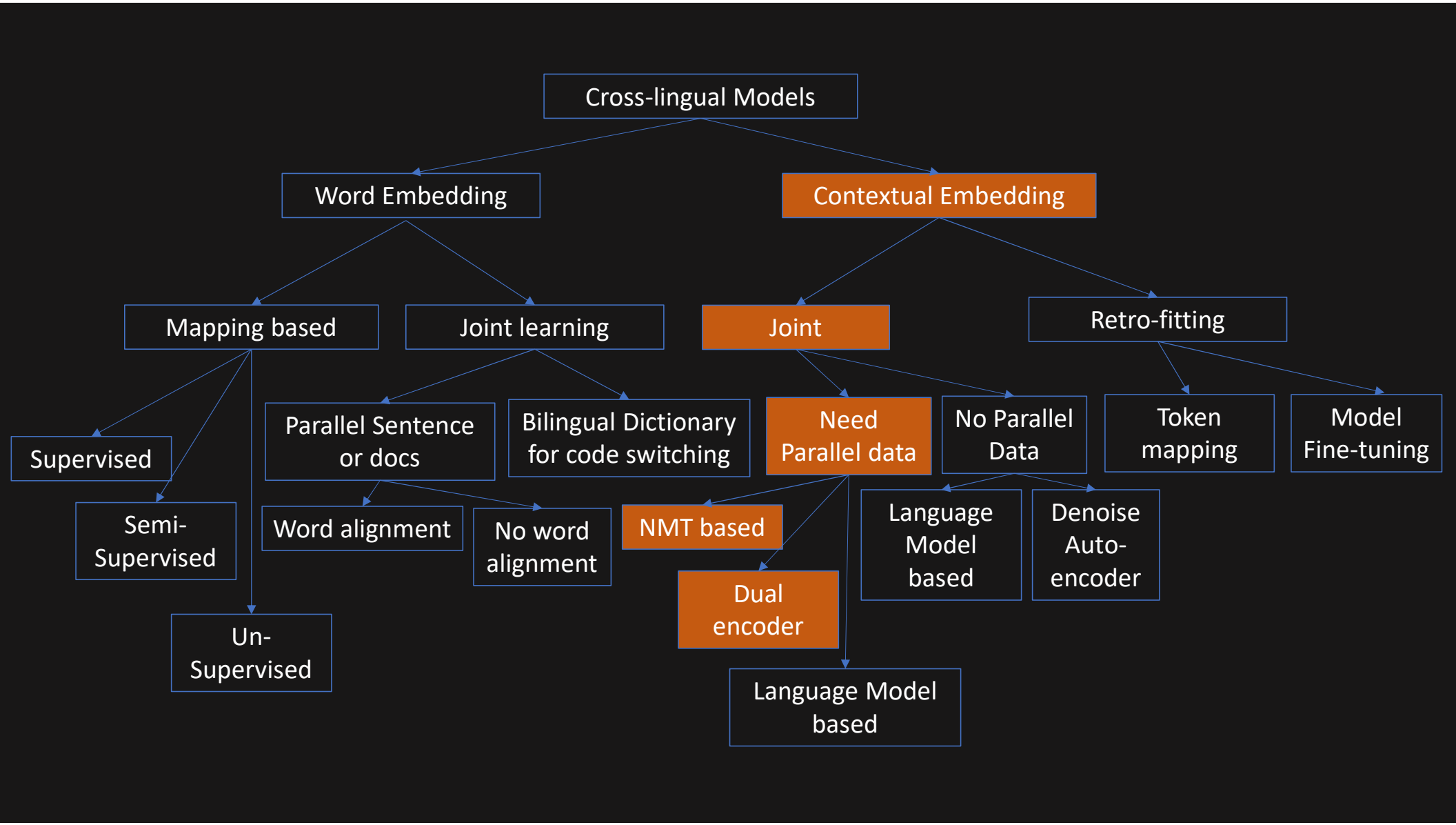
- Language Alignment
  - The input sentence pairs are in different languages
  - Align embeddings of different languages into one common space
- Efficient Retrieval
  - Efficiency is critical since the candidate pool is usually very large
  - Not feasible to conduct full interaction between input sentence pair
  - Pre-encode all candidates, and conduct approximate nearest neighbor (ANN) search



Microsoft

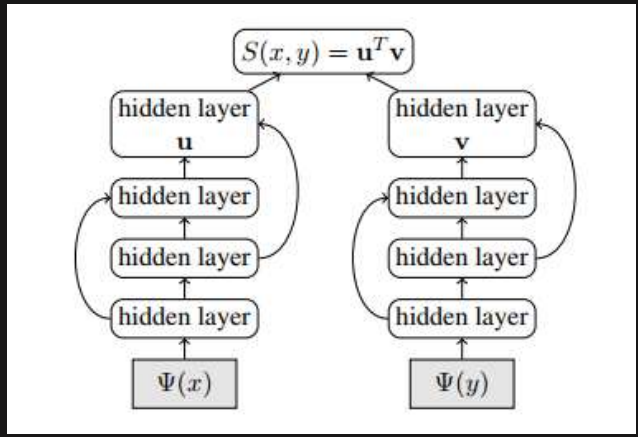
Cross-lingual Semantic Retrieval

- Introduction
- Cross-lingual Models
  - Base architecture
    - Dual encoder
    - NMT based model
  - Advanced approaches
    - Better loss function for training
    - Better scoring function for inference
- Data Augmentation
  - Using both monolingual and parallel data
  - Using only monolingual data


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
Need Parallel data
No Parallel Data
Token mapping
Model Fine-tuning
Word alignment
No word alignment
NMT based
Dual encoder
Language Model based
Language Model based
Denoise Auto-encoder

Dual Encoder

- Dual Encoder Model
  - Source sentence and target sentence are encoded separately by a dual encoder model
  - Similarity between two embeddings
- Encoder
  - Deep Averaging Networks (DAN)
  - Transformers
  - …



Guo, Mandy, et al. Effective parallel corpus mining using bilingual sentence embeddings. WMT 2018.

Dual Encoder

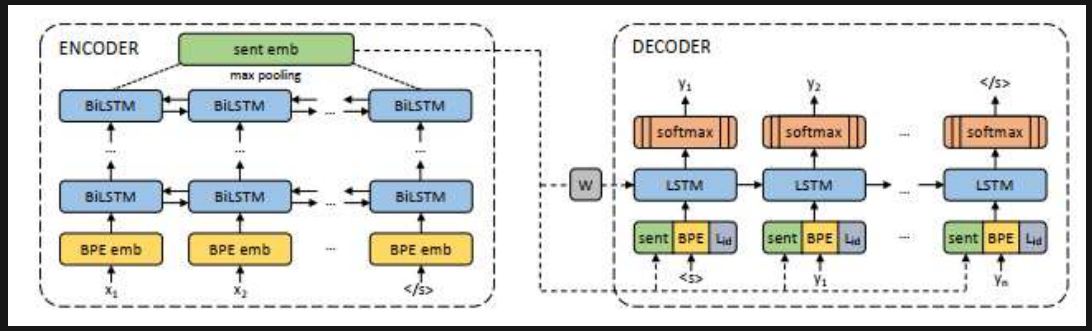
- Model Training
  - Parallel sentences as training data
  - Approximated ranking probability

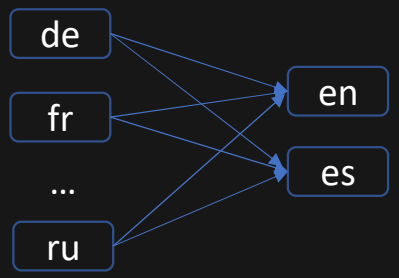
$$P(y|x) = \frac{e^{\emptyset(x,y)}}{\sum_{\bar{y}} e^{\emptyset(x,\bar{y})}} \quad \Longrightarrow \quad P_{approx}(y|x) = \frac{e^{\emptyset(x,y)}}{\sum_{k=1}^{K} e^{\emptyset(x,y_k)}}$$

  - Hard negatives
    - Semantically similar translations but not quite identical to the correct translation
    - First train a baseline model with randomly sampled negatives
    - Then select hard negatives which are inaccurate sentences with high relevance score according to the baseline model

Guo, Mandy, et al. Effective parallel corpus mining using bilingual sentence embeddings. WMT 2018.

NMT (Encoder-Decoder) Model





- Encoder
  - Max-pooling over BiLSTM encoder without language signals
- Decoder
  - LSTM with language ID as input
- A single encoder and decoder for all languages
  - Shared BPE vocabulary learned on all training data
  - Sentences of different languages are encoded into the same space when they are translated to the same language

Artetxe, Mikel, and Holger Schwenk. "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond." *TACL 2019.*

NMT (Encoder-Decoder) Model

- Training Data
  - Difficult to obtain training data if each input sentence is jointly translated into all other languages
  - Using only two target languages (English and Spanish) gets similar results
  - Scaling to almost 100 languages
- Inference
  - Get language-independent representation for each sentence with encoder

	TRAIN				TEST			
	de-en	fr-en	ru-en	zh-en	de-en	fr-en	ru-en	zh-en
Azpeitia et al. (2017)	83.33	78.83	-	-	83.74	79.46	-	-
Grégoire and Langlais (2017)	-	20.67	-	-	-	20	-	-
Zhang and Zweigenbaum (2017)	-	-	-	43.48	-	-	-	45.13
Azpeitia et al. (2018)	84.27	80.63	80.89	76.45	85.52	81.47	81.30	77.45
Bouamor and Sajjad (2018)	-	75.2	-	-	-	76.0	-	-
Chongman Leong and Chao (2018)	-	-	-	58.54	-	-	-	56
Schwenk (2018b)	76.1	74.9	73.3	71.6	76.9	75.8	73.8	71.6
Artetxe and Schwenk (2018)	94.84	91.85	90.92	91.04	95.58	92.89	92.03	**92.57**
Proposed method	**95.43**	**92.40**	**92.29**	**91.20**	**96.19**	**93.91**	**93.30**	92.27

Artetxe, Mikel, and Holger Schwenk. "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond." *TACL 2019.*

Discussion

- Dual Encoder
  - Parallel corpus for training data
  - Explicit language alignment between two languages
- NMT Model
  - Parallel corpus for training data
  - Implicit language alignment among multiple languages

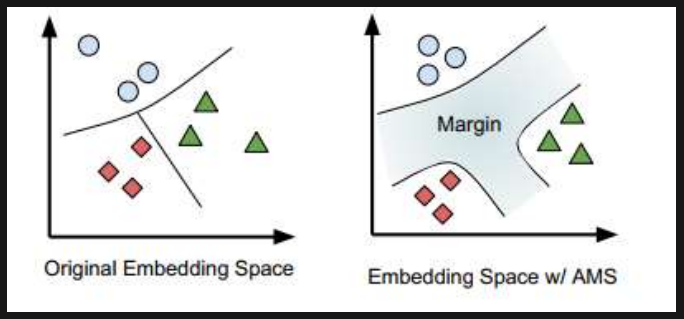
Cross-lingual Semantic Retrieval

- Introduction
- Cross-lingual Models
  - Base architecture
    - Dual encoder
    - NMT based model
  - Advanced approaches
    - Better loss function for training
    - Better scoring function for inference
- Data Augmentation
  - Using both monolingual and parallel data
  - Using only monolingual data

Better Loss Function for Training

- Dual Encoder Model
  - Originally, maximizing the softmax probability of positive translation pairs over negative ones
  - Additive margin softmax: maximizing the margin between matched sentence pairs compared to similar but inaccurate pairs

The margin tends to structure the embedding space to be more compact, and achieves state-of-the-art results



Yang Y. et al.. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. IJCAI'19

Better Loss Function for Training

- Bidirectional dual encoder
  - Parallel sentences are true translations of each other

$$L_s = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\phi(x_i,y_i)}}{e^{\phi(x_i,y_i)}+\sum_{n=1,n\neq i}^{N}e^{\phi(x_i,y_n)}} \qquad L'_s = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\phi(y_i,x_i)}}{e^{\phi(y_i,x_i)}+\sum_{n=1,n\neq i}^{N}e^{\phi(y_i,x_n)}}$$

- Additive Margin Softmax

additive margin for positive pairs

$$L_{ams} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\phi(x_i,y_i)-m}}{e^{\phi(x_i,y_i)-m}+\sum_{n=1,n\neq i}^{N}e^{\phi(x_i,y_n)}}$$

Yang Y. et al.. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. IJCAI'19

Better Loss Function for Training

Models	en-fr			en-es			en-ru			en-zh		
	P@1	P@3	P@10	P@1	P@3	P@10	P@1	P@3	P@10	P@1	P@3	P@10
[Guo *et al.*, 2018]	48.9	62.3	73.0	54.9	67.8	78.1	-	-	-	-	-	-
[Artetxe and Schwenk, 2018]	83.3	-	-	85.8	-	-	-	-	-	-	-	-
DE	80.7	87.9	91.2	85.6	92.7	95.1	83.9	89.7	92.0	82.2	91.1	94.1
BiDE	82.3	90.7	94.2	86.3	93.0	95.6	85.7	92.3	95.1	83.1	91.3	94.6
BiDE+AM	**86.1**	**93.5**	**96.1**	**89.0**	**95.2**	**97.2**	**89.2**	**94.8**	**96.9**	**87.9**	**94.7**	**97.1**

Comparison on the UN corpus

- Results
  - Bidirectional dual model (BiDE) outperforms previous dual model (DE)
  - Additive margin (AM) improves the accuracy further compared to BiDE

Yang Y. et al.. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. IJCAI'19

Better Scoring Function for Inference

- Parallel sentence pair mining by nearest neighbor retrieval by cosine similarity with a hard threshold
- However, scale of cosine similarity is not globally consistent

irrelevant

correct translation →

(A)	*Les produits agricoles sont constitués de thé, de riz, de sucre, de tabac, de camphre, de fruits et de soie.*
0.818	Main crops include wheat, sugar beets, potatoes, cotton, tobacco, vegetables, and fruit.
0.817	The fertile soil supports wheat, corn, barley, tobacco, sugar beet, and soybeans.
0.814	Main agricultural products include grains, cotton, oil, pigs, poultry, fruits, vegetables, and edible fungus.
0.808	The important crops grown are cotton, jowar, groundnut, rice, sunflower and cereals.

(B)	*Mais dans le contexte actuel, nous pourrons les ignorer sans risque.*
0.737	But, in view of the current situation, we can safely ignore these.
0.499	But without the living language, it risks becoming an empty shell.
0.498	While the risk to those working in ceramics is now much reduced, it can still not be ignored.
0.488	But now they have discovered they are not free to speak their minds.

Mikel Artetxe and Holger Schwenk. Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. ACL 2019

Better Scoring Function for Inference

- Margin-based Scoring

similarity between x and y

$$score(x, y) = margin(\cos(x, y), \sum_{z \in NN_k(x)} \frac{\cos(x, y)}{2k} + \sum_{z \in NN_k(y)} \frac{\cos(y, z)}{2k})$$

similarity between x and x's nearest neighbors

similarity between y and y's nearest neighbors

	en-de	en-fr	en-ru	en-zh
Azpeitia et al. (2017)	83.7	79.5	-	-
Azpeitia et al. (2018)	85.5	81.5	81.3	77.5
Bouamor and Sajjad (2018)	-	76.0	-	-
Schwenk (2018)	76.9	75.8	73.8	71.6
Proposed method (Europarl)	**95.6**	**92.9**	-	-
Proposed method (UN)	-	-	**92.0**	**92.6**

	en-fr	en-es
Guo et al. (2018)	48.90	54.94
Proposed method	**83.27**	**85.78**

- significantly improves baselines on BUCC and UN corpus
- Can be used for both dual encoder and NMT-based models

Mikel Artetxe and Holger Schwenk. Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. ACL 2019

Microsoft

Cross-lingual Semantic Retrieval

- Introduction
- Cross-lingual Models
  - Basic models
  - Advanced approaches
- Data Augmentation
  - Using both monolingual and parallel data
  - Using only monolingual data

Microsoft

Motivation

- Parallel data is not easy to get, especially for low-resource languages
- Large volume of monolingual data
- The goal is to generate multilingual sentence embeddings with limited, even no parallel data

Microsoft

Using Both Monolingual and Parallel Data

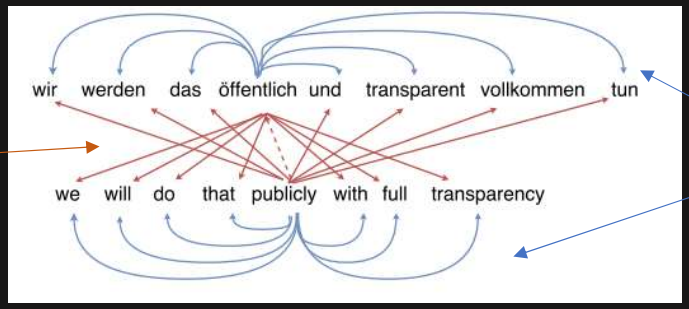
- Learn high quality multilingual sentence embeddings with limited parallel corpus
- A multi-task Approach
  - Multilingual skip-gram for word embedding
  - Cross-lingual sentence similarity
  - Shared word embedding layer for joint learning

Singla, Karan, Doğan Can, and Shrikanth Narayanan. "A multi-task approach to learning multilingual representations." *ACL* 2018.

Using Both Monolingual and Parallel Data

- Multilingual Skip-gram
  - Monolingual (context, word) pairs L1->L1, L2->L2, sampled from monolingual and parallel corpora
  - Cross-lingual (context, word) pairs L1->L2, L2->L1, sampled from parallel corpora

cross-lingual pair of (context, word)



cross-lingual pair of (context, word)

Embeddings of words not exist in parallel data can be learned from monolingual pairs
Cross-lingual embeddings are aligned according to cross-lingual pairs

Singla, Karan, Doğan Can, and Shrikanth Narayanan. "A multi-task approach to learning multilingual representations." *ACL* 2018.

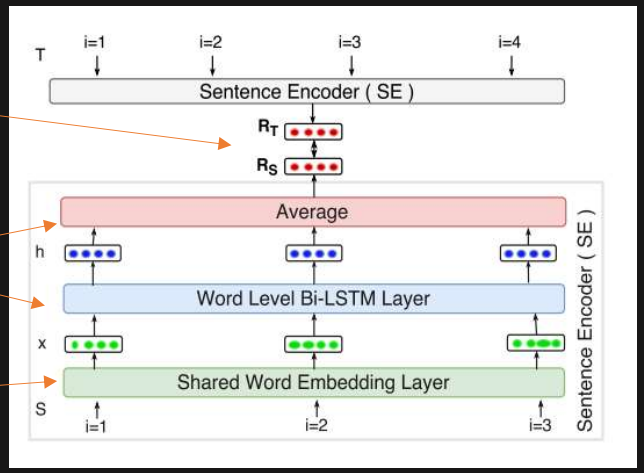
Using Both Monolingual and Parallel Data

- Dual Encoder for Sentence Similarity

Similarity between sentence pairs for model training

BiLSTM and average pooling to get sentence embedding based on word embedding

Shared word embeddings with skip-gram



- Multi-lingual skip-gram and cross-lingual sentence similarity models are trained jointly
- Inform each other through the shared word embeddings

Singla, Karan, Doğan Can, and Shrikanth Narayanan. "A multi-task approach to learning multilingual representations." *ACL* 2018.

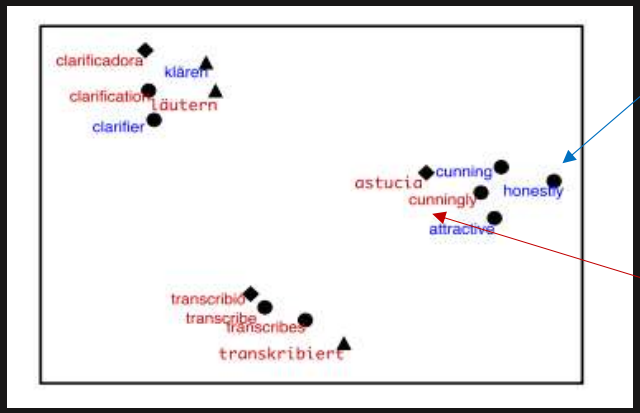
Using Both Monolingual and Parallel Data

size of parallel data

monolingual data size

Mono \ Parallel	20K	50K	100K	500K
no-mono	60.3	68.3	82.1	89.5
20K	57.4	68.7	80.2	89.5
50K	**62.7**	69.0	83.5	89.5
100K	61.5	71.9	85.1	89.6
200K	58.1	**72.1**	85.5	90.0
500K	52.6	64.8	**87.4**	**90.4**

Multi-task learning produces consistently better embeddings when the amount of monolingual data is neither too large nor too small



words exist in parallel corpus

words exist only in monolingual corpus

nearest neighbors of English words (clarification, transcribe, cunningly)

Multi-task learning aligns cross-lingual words which only exists monolingual corpus

Singla, Karan, Doğan Can, and Shrikanth Narayanan. "A multi-task approach to learning multilingual representations." *ACL* 2018.

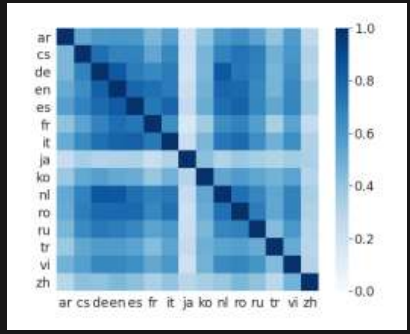
Microsoft

Cross-lingual Semantic Retrieval

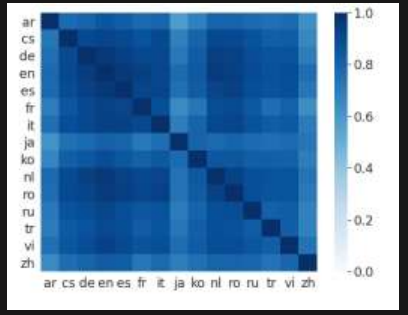
- Introduction
- Cross-lingual Models
  - Basic models
  - Advanced approaches
- Data Augmentation
  - Using both monolingual and parallel data
  - Using only monolingual data

Using Only Monolingual Data

- A multilingual model mBART trained from monolingual data
- Cross-lingual retrieval with mBART
  - Good retrieval accuracy without any fine-tuning
  - Big improvement (27%) on all languages when just fine-tuned on Japanese-English parallel data
- Thinking
  - mBART is a good starting point for cross-lingual retrieval
  - It can be self-improved by parallel data mined from the model itself



Sentence retrieval using mBART encoder



mBART fine-tuned on Japanese-English parallel data

Tran et al. Cross-lingual retrieval for iterative self-supervised training. NeurIPS 2020.

Using Only Monolingual Data

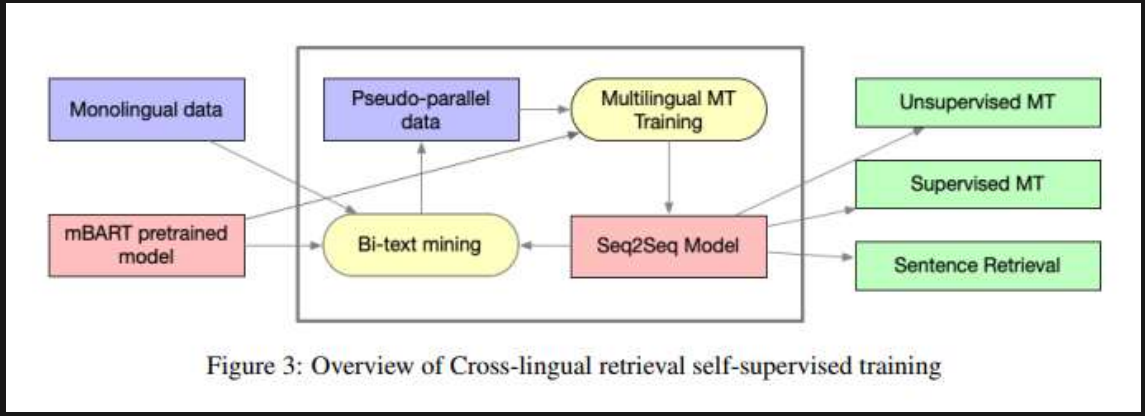


Figure 3: Overview of Cross-lingual retrieval self-supervised training

- Iterative approach
  - Pretrained mBART as an initial model to retrieve parallel sentences
  - A sequence-to-sequence model is learned from the parallel sentences
  - Iteratively retrieve and update the seq2seq model

Tran et al. Cross-lingual retrieval for iterative self-supervised training. NeurIPS 2020.

Using Only Monolingual Data

- After 2-3 iterations, the retrieval accuracy is significantly increased
- However, there is still a gap between unsupervised and supervised approach

Language	ar	de	es	et	fi	fr	hi	it
XLMR [14]	47.5	88.8	75.7	52.2	71.6	73.7	72.2	68.3
mBART [34]	39	86.8	70.4	52.7	63.5	70.4	44	68.6
CRISS Iter 1	72	97.5	92.9	85.6	88.9	89.1	86.8	88.7
CRISS Iter 2	76.4	**98.4**	95.4	**90**	92.2	91.8	91.3	91.9
CRISS Iter 3	**78.0**	98.0	**96.3**	89.7	**92.6**	**92.7**	**92.2**	**92.5**
LASER (supervised) [6]	92.2	99	97.9	96.6	96.3	95.7	95.2	95.2

Language	ja	kk	ko	nl	ru	tr	vi	zh
XLMR [14]	60.6	48.5	61.4	80.8	74.1	65.7	74.7	68.3 (71.6)
mBART [34]	24.9	35.1	42.1	80	68.4	51.2	63.9	14.8
CRISS Iter 1	76.8	67.7	77.4	91.5	89.9	86.9	89.9	69
CRISS Iter 2	**84.8**	74.6	**81.6**	92.8	**90.9**	92	92.5	81
CRISS Iter 3	84.6	**77.9**	81.0	**93.4**	90.3	**92.9**	**92.8**	**85.6**
LASER (supervised) [6]	94.6	17.39	88.5	95.7	94.1	97.4	97	95

Tran et al. Cross-lingual retrieval for iterative self-supervised training. NeurIPS 2020.

Microsoft

Key Takeaways

- Two Model Architectures for Cross-lingual Semantic Retrieval
  - Dual encoder, explicit language alignment
  - NMT model, implicit language alignment
- Advanced Approaches
  - Improve model training by additive margin
  - Improve inference by margin-based scoring
  - Enhance low-resource language performance by leveraging monolingual data

Outline

NLP Group
Software Technology Center at Asia (STCA) of Microsoft

- Introduction [Daxin Jiang]
  - Motivating examples in Microsoft products
  - Problem description
  - Categorization of applications
  - Challenges and major approaches
- Applications
  - Natural Language Inference [Linjun Shou]
  - Information retrieval [Xiubo Geng]
  - **Machine Reading Comprehension [Ming Gong]**
- Future directions [Daxin Jiang]

Linjun Shou

Xiubo Geng

Ming Gong

Cross Lingual Sequence Labelling

- Machine Reading Comprehension as Example

Ming Gong
migon@microsoft.com

Microsoft

Outline

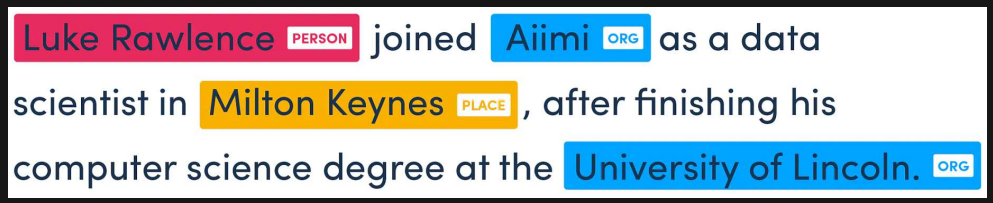
- Sequence Labeling Application Overview
- CLMRC Benchmark Datasets
- CLMRC Baseline Approach & Challenges
- Advanced Approaches for CLMRC

Cross-lingual Sequence Labelling

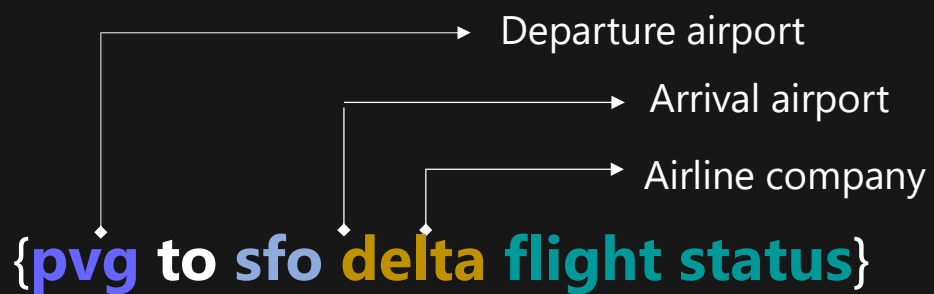
Type	Category	Sub Category	Example
NLU	Text Classification	Single text	Domain identification, Intent detection, Sentiment classification
		Text pair	Information retrieval, Natural language inference
	Sequence Labeling	Single text	Named entity recognition, Slot tagging
		Text pair	Extractive Machine reading comprehension
NLG	Text Generation	Token level	Spelling correction, Sentence auto completion
		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)



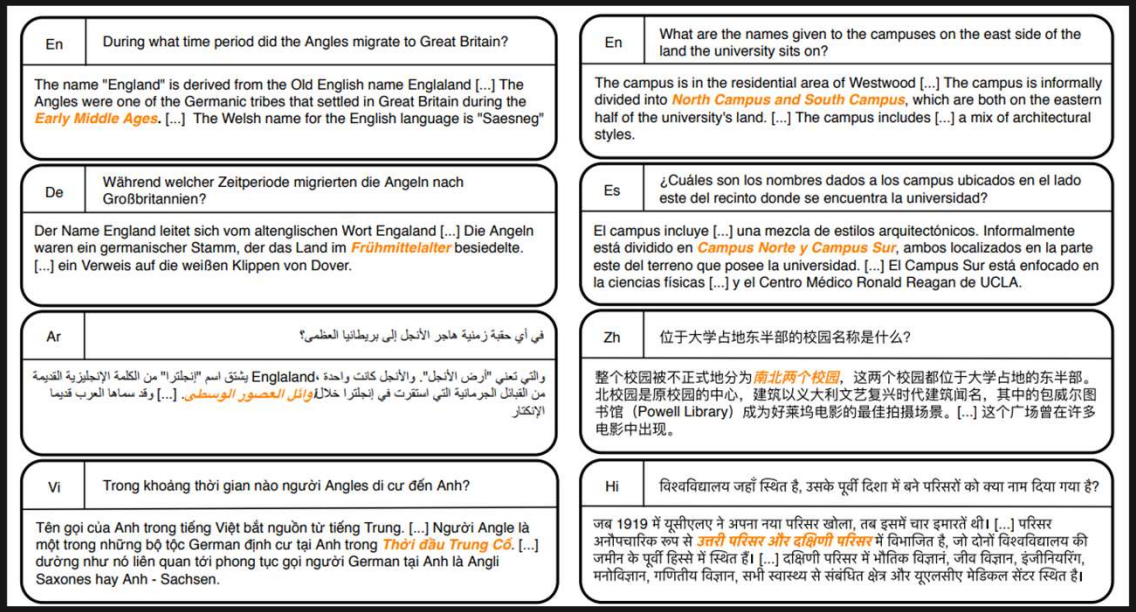
Application2: Slot Tagging

- 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.



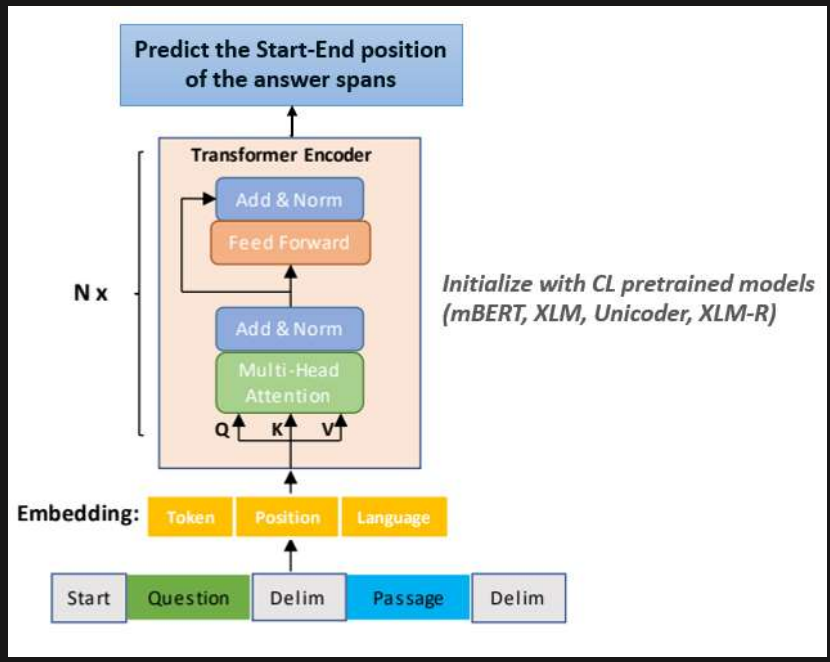
Examples from MLQA

Typical Setting

- There are only English labeled data for training.
- for non-EN languages, there are only small sets of labeled data as test set.

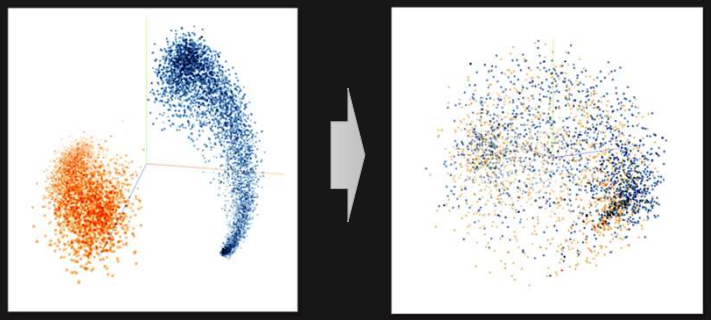
Baseline Approach for CLMRC

- 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 for CLMRC (1/4)

- Classification *vs* Sequence Labeling using CL Pretrained models
  - CL pretrained models (*mBERT, XLM, Unicoder, XLM-R*) work decently well in **sentence level tasks**, like sentence classifier, sentence pair matching etc.
  - However, performance on phrase boundary related tasks are limited.



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

→ *better semantic alignment is required*

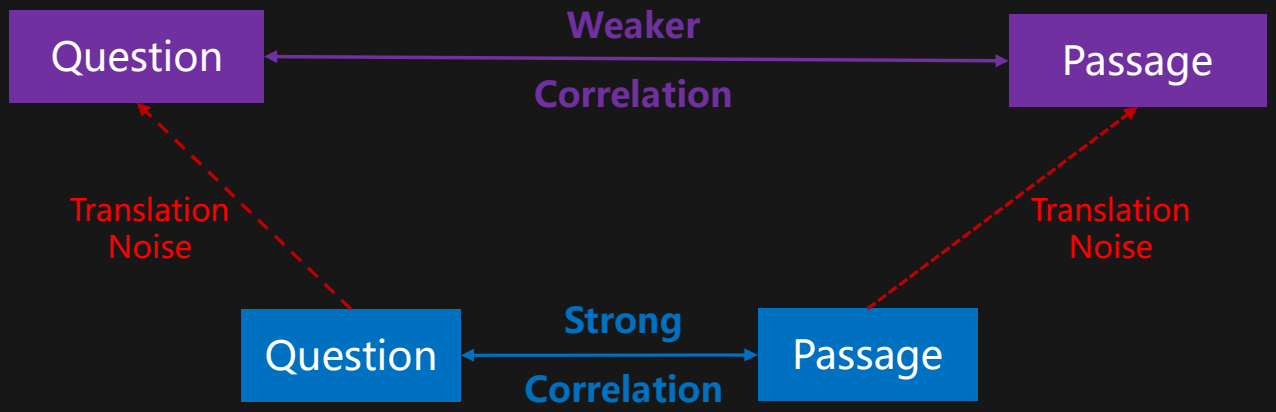
Language	MRC EM	MRC Gap to en	NLI ACC	NLI Gap to en
en	62.4	–	85.0	–
es	49.8	-12.6	78.9	-6.1
de	47.6	-14.8	77.8	-7.2
ar	36.3	-26.1	73.1	-11.9
hi	27.3	-35.1	69.6	-15.4
vi	41.8	-20.6	76.1	-8.9
zh	39.6	-22.8	76.5	-8.5

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.

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 for CLMRC (2/4)

- Machine translation (MT) data is noisy which impact more for pair-wise sequence labeling task



Challenges for CLMRC (3/4)

- Besides of query and passage, answer span alignment in passage is challenging after machine translation (MT).

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:
zwischen Sonne und Mond

Only partial of the correct span in DE
(*missed "befindet"*)

Challenges for CLMRC (4/4)

- Precise answer boundary detection is critical for MRC which result in the major errors of CLMRC model inference.

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*.

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

[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.

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

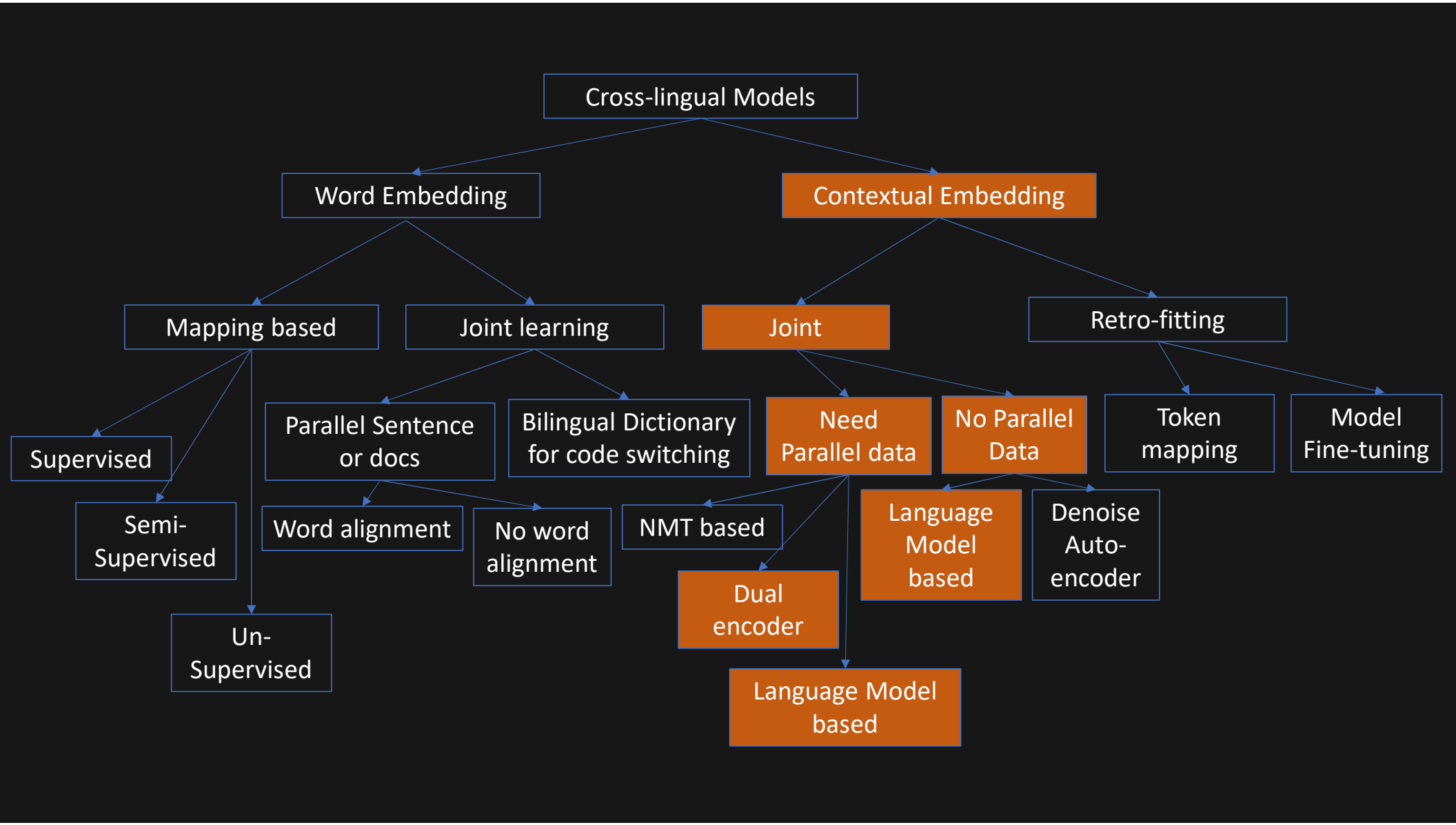
B Machine Translation Data Quality

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

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

Representative Works

1. **Cross-Lingual Machine Reading Comprehension**
   Yiming Cui[†‡], Wanxiang Che[†], Ting Liu[†], Bing Qin[†], Shijin Wang[‡§], Guoping Hu[‡]
2. **Enhancing Answer Boundary Detection for Multilingual Machine Reading Comprehension**
   Fei Yuan[1*] Linjun Shou[2†] Xuanyu Bai[2] Ming Gong[2] Yaobo Liang[3] Nan Duan[3] Yan Fu[1] Daxin Jiang[2]
3. **CalibreNet: Calibration Networks for Multilingual Sequence Labeling**
   Shining Liang[1,2,*], Linjun Shou[3], Jian Pei[4], Ming Gong[3], Wanli Zuo[1,2†], Daxin Jiang[3†]
4. **Cross-lingual Machine Reading Comprehension with Language Branch Knowledge Distillation**
   Junhao Liu[1,2*] Linjun Shou[3*] Jian Pei[4] Ming Gong[3] Min Yang[1†] Daxin Jiang[3†]
5. **Reinforced Multi-Teacher Selection for Knowledge Distillation**
   Fei Yuan[1*] Linjun Shou[2] Jian Pei[3] Wutao Lin[2] Ming Gong[2] Yan Fu[1] Daxin Jiang[2†]
6. **FILTER: An Enhanced Fusion Method for Cross-lingual Language Understanding**
   Yuwei Fang*, Shuohang Wang*, Zhe Gan, Siqi Sun, Jingjing Liu
7. **Multilingual Transfer Learning for QA Using Translation as Data Augmentation**
   Mihaela Bornea, Lin Pan, Sara Rosenthal, Radu Florian, Avirup Sil


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

A New task to enhance semantic transfer

Microsoft

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?

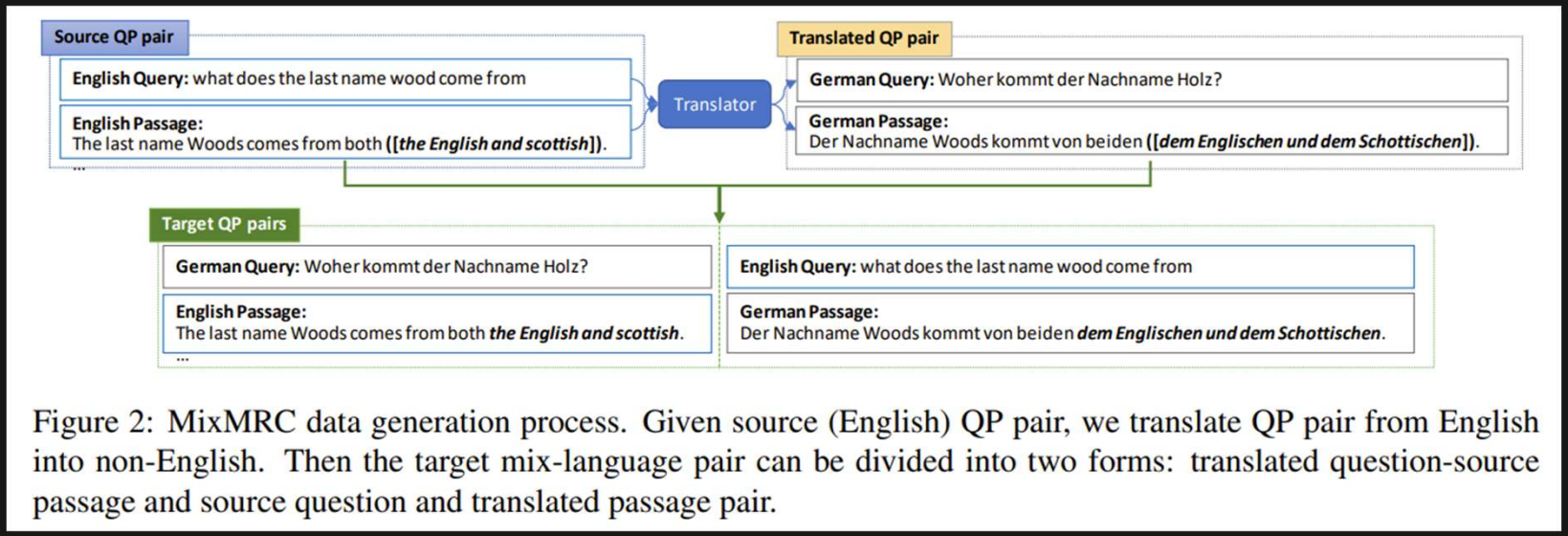


Figure 2: MixMRC data generation process. Given source (English) QP pair, we translate QP pair from English into non-English. Then the target mix-language pair can be divided into two forms: translated question-source passage and source question and translated passage pair.

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.

A New task to enhance semantic transfer

Microsoft

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.

A Dual Encoder to enhance semantic transfer

Microsoft

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

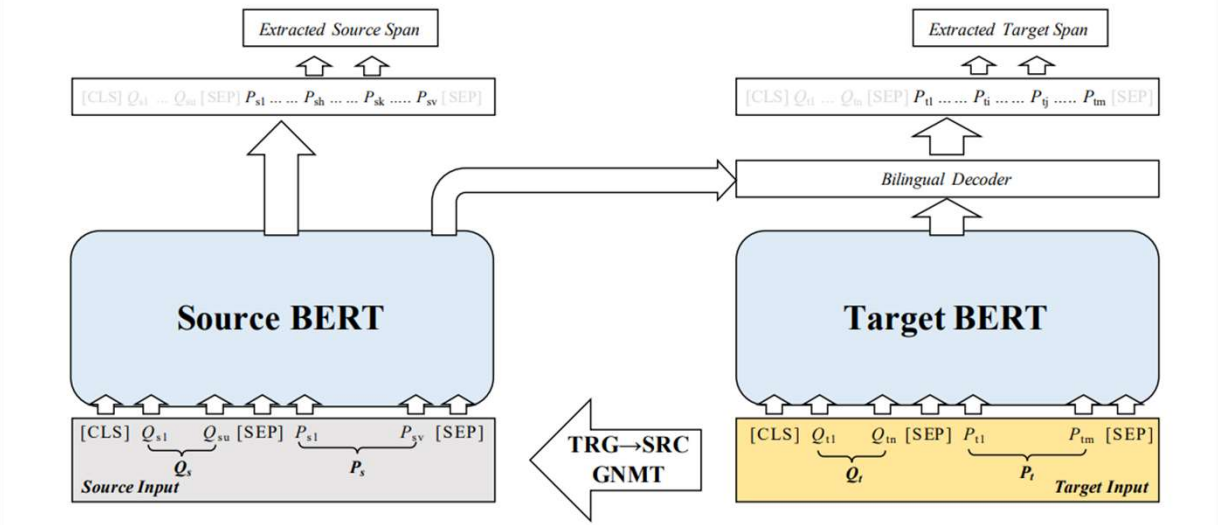


Figure 2: System overview of the Dual BERT model for cross-lingual machine reading comprehension task.

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

Microsoft

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

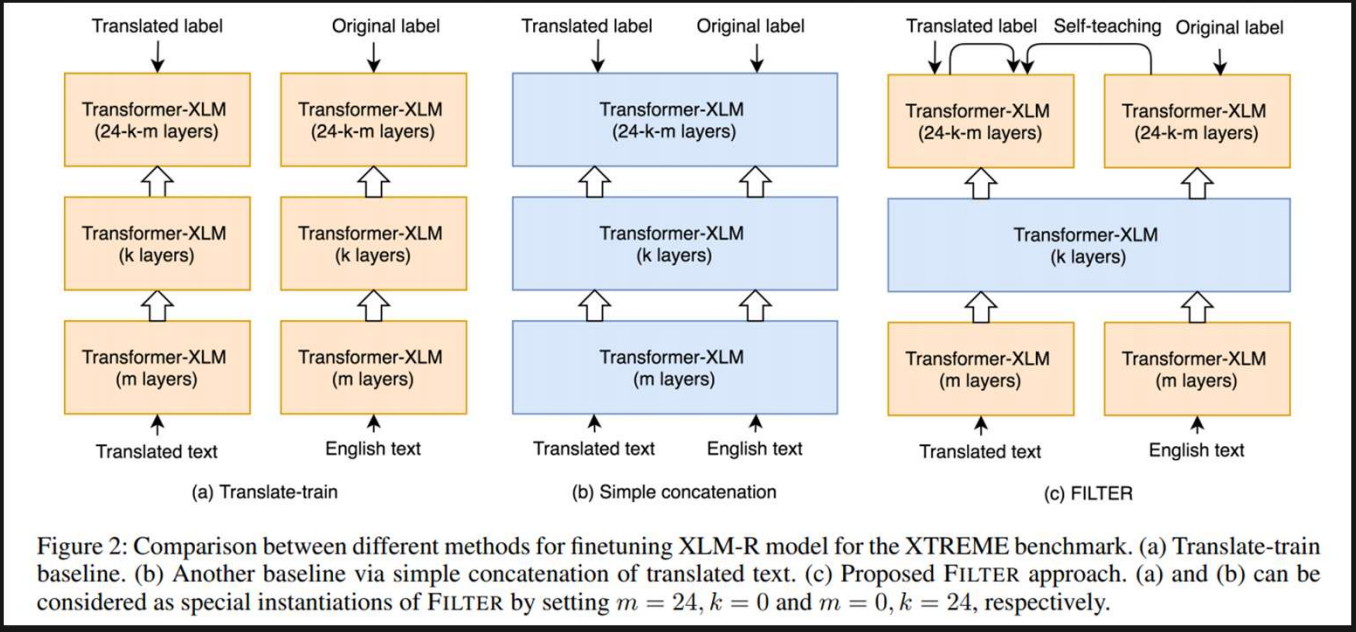


Figure 2: Comparison between different methods for finetuning XLM-R model for the XTREME benchmark. (a) Translate-train baseline. (b) Another baseline via simple concatenation of translated text. (c) Proposed FILTER approach. (a) and (b) can be considered as special instantiations of FILTER by setting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

B Better Answer span alignment for MT

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.

1. Translate from target language back to source language
2. Use source MRC model to predict answer span
3. Translate source span to target span and further improve boundary by:
   - *Simple Match*
   - *Answer Aligner*
   - *Answer Verifier*

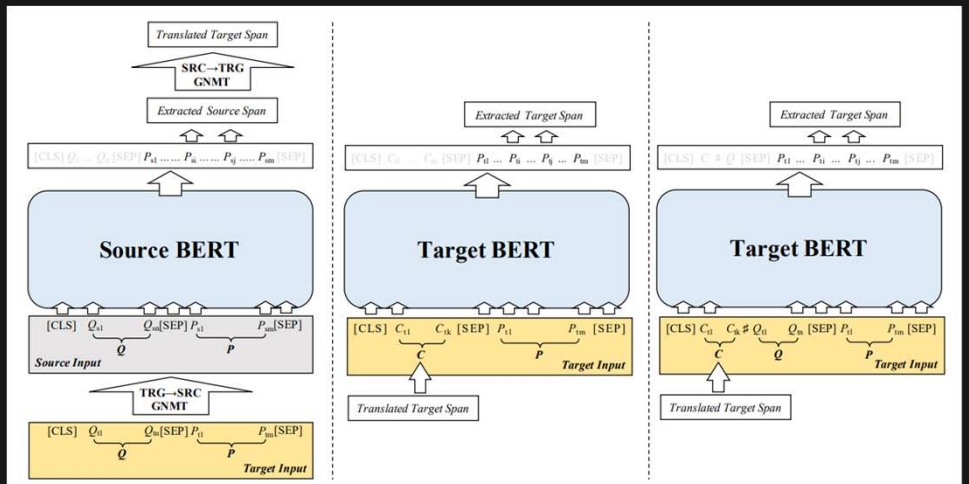


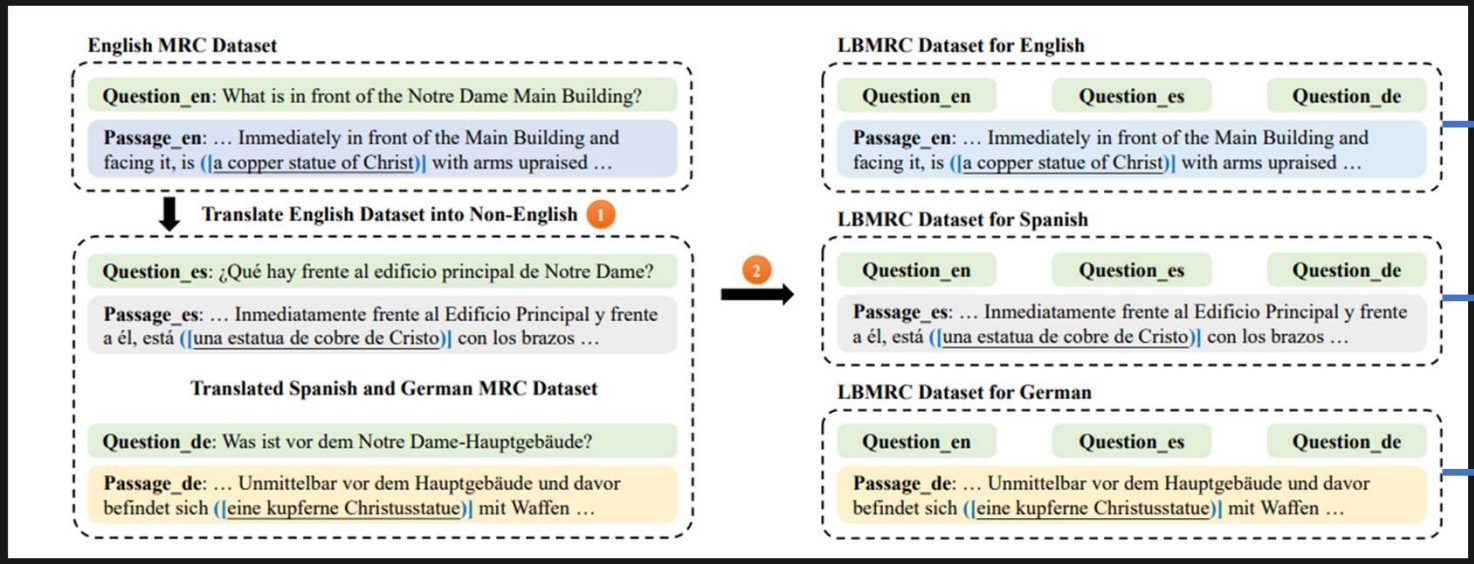
Figure 1: Back-translation approaches for cross-lingual machine reading comprehension (Left: GNMT, Middle: Answer Aligner, Right: Answer Verifier)

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

Microsoft

LBMRC: Language Branch Knowledge Distillation for Answer Span Correction in MT Data

- **Motivation:** How to correct the noisy answer span in MT data? → Leverage multiple teacher models for de-noising labeling

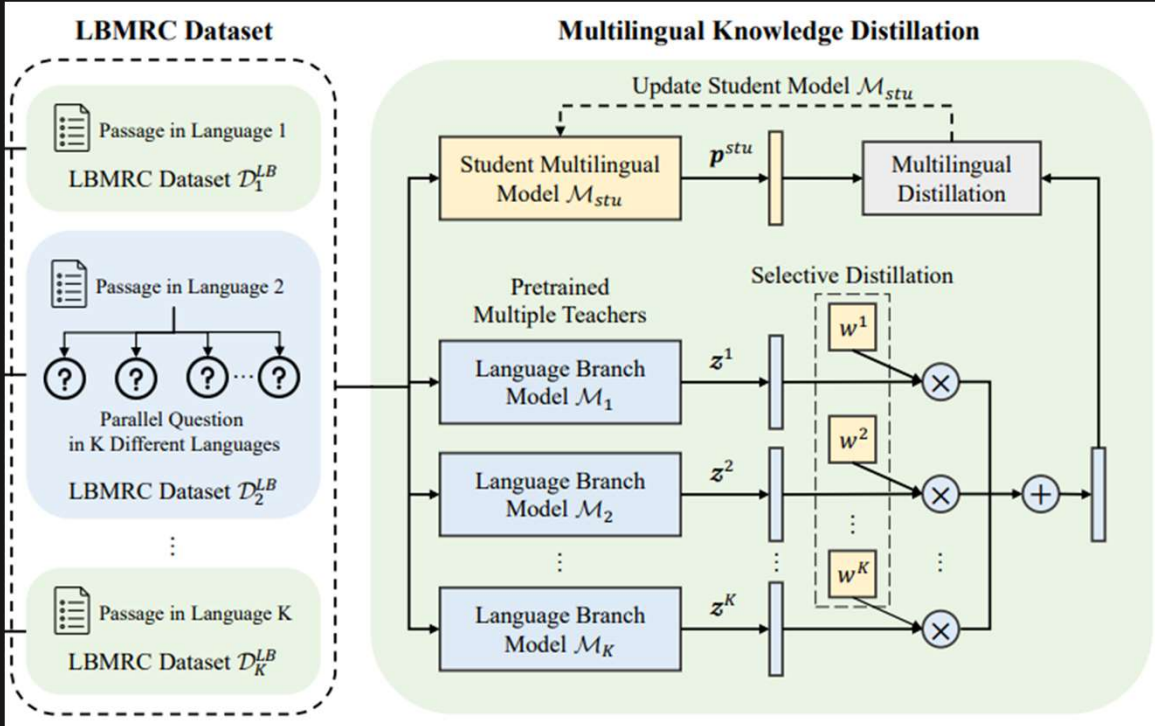


Junhao Liu, Linjun Shou, Jian Pei, Ming Gong, Min Yang, and Daxin Jiang. **Cross-lingual Machine Reading Comprehension with Language Branch Knowledge Distillation.** COLING. 2020

B Better leverage MT noisy data

Microsoft

LBMRC: Language Branch Knowledge Distillation for De-noise Training with MT Data



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

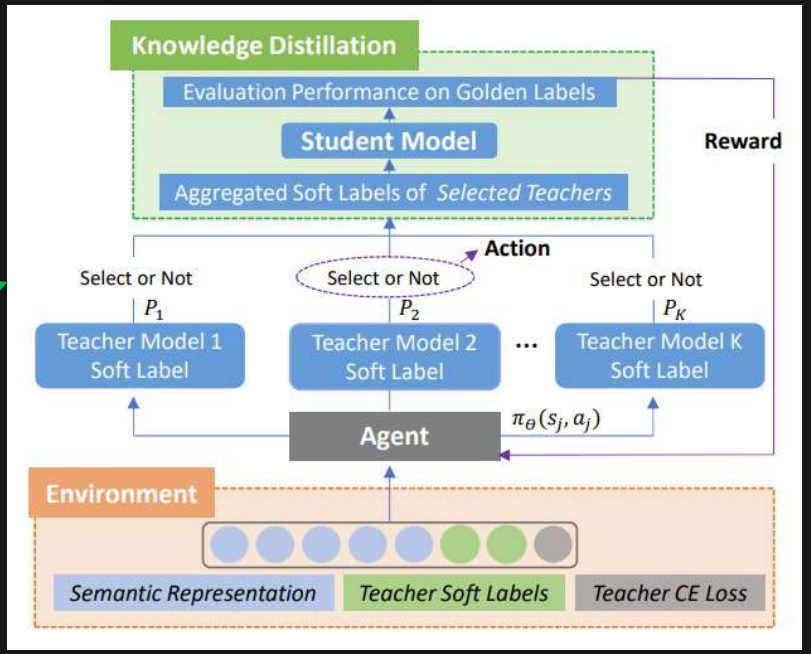
Microsoft

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 single teacher
- Vanilla KD from teacher ensemble

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 Better leverage MT noisy data

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 single teacher
- Vanilla KD from teacher ensemble

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

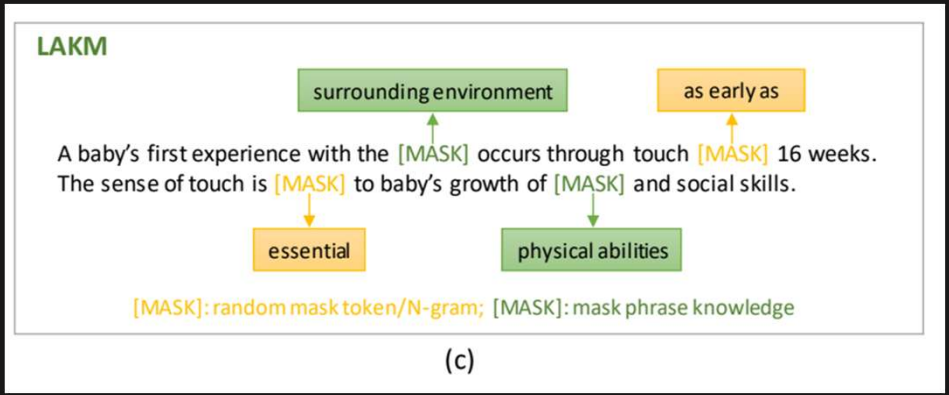
Fei Yuan, Linjun Shou, Jian Pei, Wutao Lin, Ming Gong, and Daxin Jiang. **Reinforced Multi-Teacher Selection for Knowledge Distillation**. AAAI 2021.

C Enhance answer boundary detection

Microsoft

New Tasks for CL Pretrained Models Fine-tuning

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



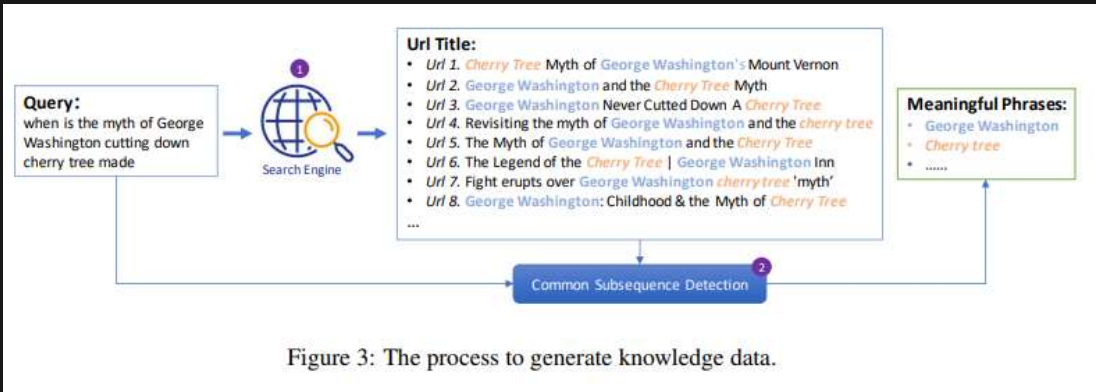


Figure 3: The process to generate knowledge data.

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

- Phrase Candidates Generation: recall
- Phrase Filtering: 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.

C Enhance answer boundary detection

Microsoft

New Tasks for CL Pretrained Models Fine-tuning

- Zero-shot 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.

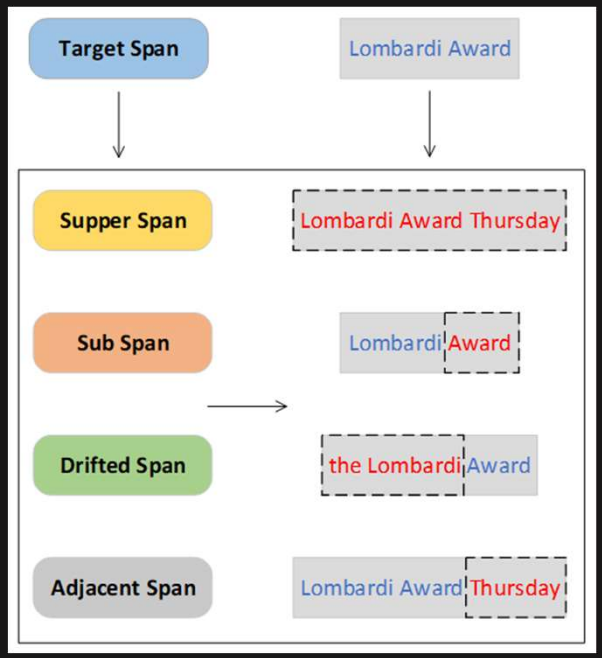
Fei Yuan, Linjun Shou, Xuanyu Bai, Ming Gong, Yaobo Liang, Nan Duan, Yan Fu, Daxin Jiang. Enhancing Answer Boundary Detection for Multilingual Machine Reading Comprehension. ACL, 2020.

Microsoft

Post-processing for Boundary Refinement: CalibreNet

- 4 types of answer mismatch by empirical study:
  - Supper span
  - Sub span
  - Drifted span
  - Adjacent span

- ***Thinking:*** *is it possible to add another module for answer span refinement?*



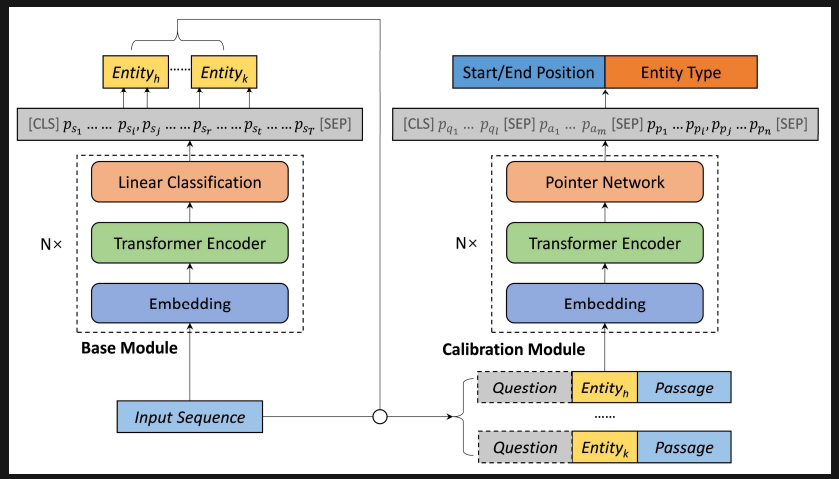
Shining Liang, Linjun Shou, Jian Pei, Ming Gong, Wanli Zuo, and Daxin Jiang. CalibreNet: CalibrationNetworks for Multilingual Sequence Labeling. WSDM. 2021.

C Enhance answer boundary detection

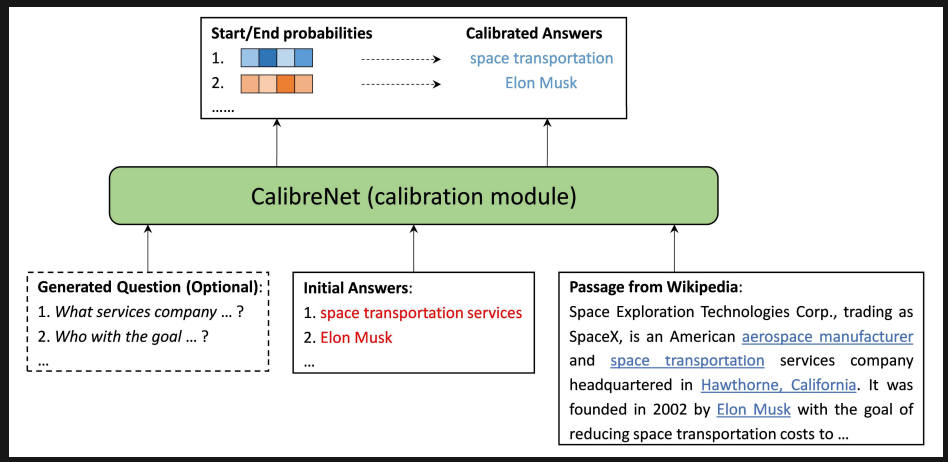
Microsoft

Post-processing for Boundary Refinement: CalibreNet

- **Phrase Boundary Recovery** pre-training task (PBR task) to pre-train CalibreNet on large-scale multilingual synthesized datasets.



(a) CalibreNet



(b) PBR task

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 don't work well due to

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

→
- New Mix-MRC task using MT data
- Dual encoders using MT data

B Machine Translation Data Quality

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

→
- Back-translation + Verification to improve answer span alignment
- Leverage language branch MT data to train multiple teacher models for better labeling
- Leverage reinforcement learning to select best teacher per instance for better labeling

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

→
- New LAKM task to infuse language specific boundary prior knowledge
- CalibraNet to add a new module to refine primary answer span

Outline

NLP Group
Software Technology Center at Asia (STCA) of Microsoft

- Introduction [Daxin Jiang]
  - Motivating examples in Microsoft products
  - Problem description
  - Categorization of applications
  - Challenges and major approaches
- Applications
  - Natural Language Inference [Linjun Shou]
  - Information retrieval [Xiubo Geng]
  - Machine Reading Comprehension [Ming Gong]
- **Future directions [Daxin Jiang]**

Linjun Shou

Xiubo Geng

Ming Gong

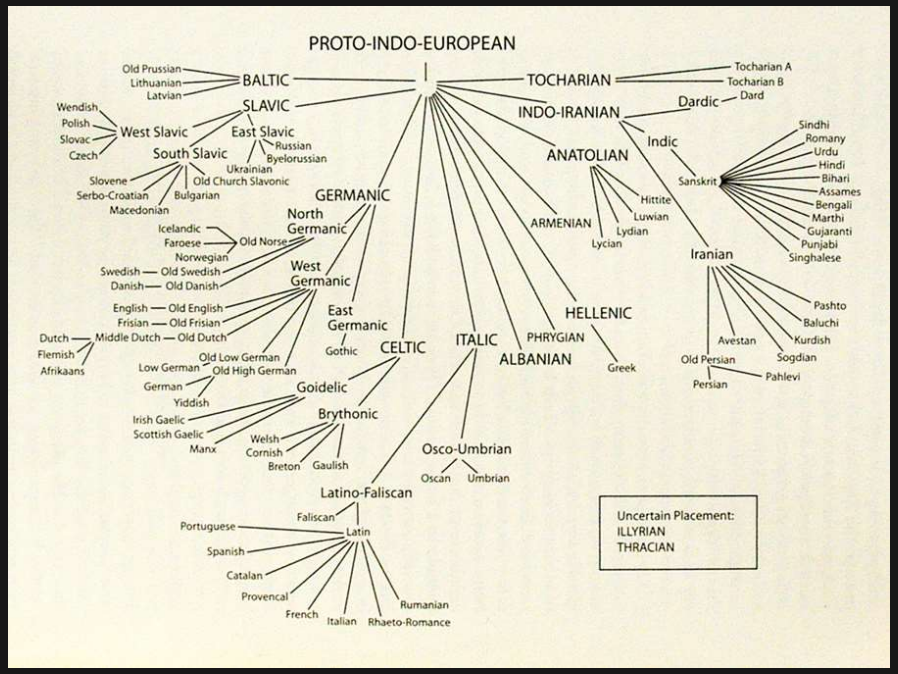
Future directions (1)

- 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)

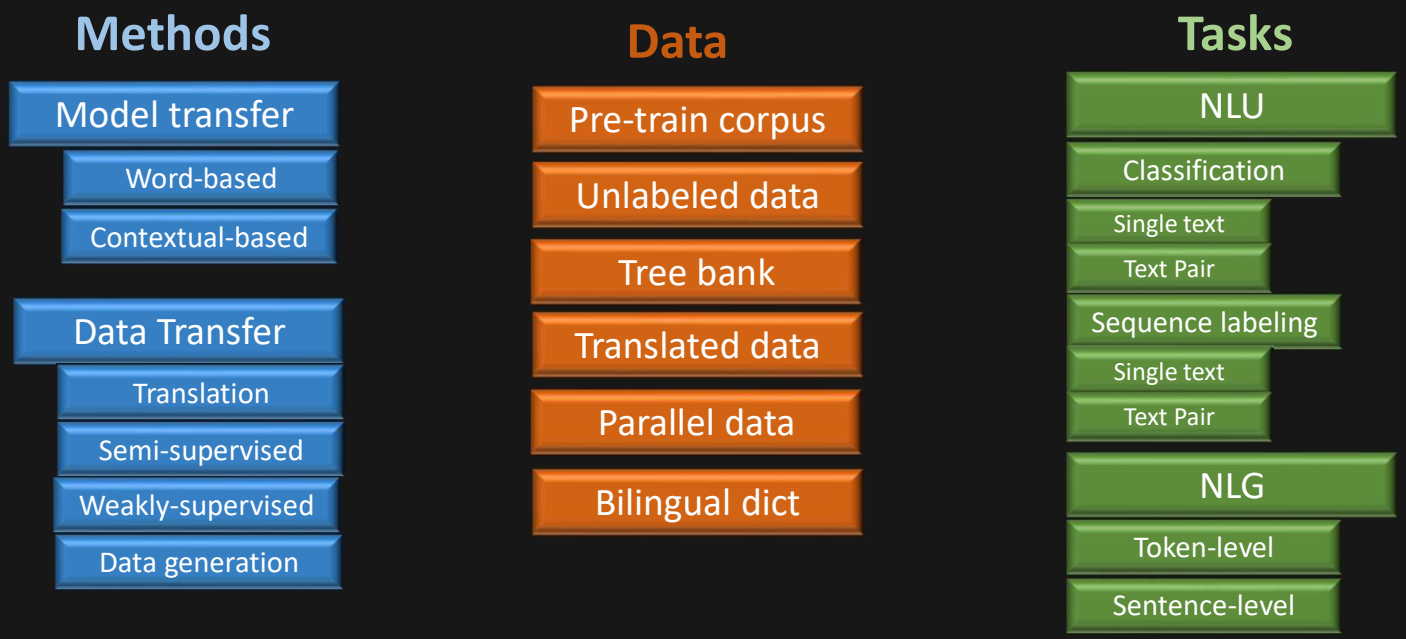
- Best practice to train multi-lingual models
  - how to mitigate “catastrophic forgetting”?
  - do we align models with language families?
  - how to measure the “capacity” of models and the “size” of languages to determine the best combination of languages



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

Future directions (3)

- Best practice to combine data and methods to reach best ROI



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