

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 [Dixin 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 [Dixin 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.



Photo: Rodney Ballard

Achieve universal primary education

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



Photo: Rodney Ballard

Promote gender equality and empower women

Nearly two-thirds of the world's 875 million illiterate people are women.



Photo: Marc Ewell

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



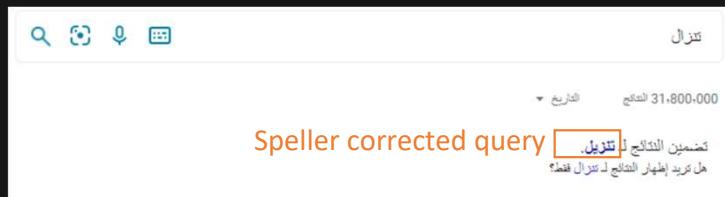
Speller

Web Relevance | Suggested Replies | Natural Language Understanding

ar-EG

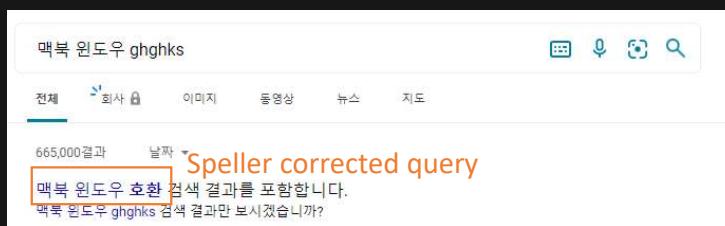
{تنزال}

English Translation:
download



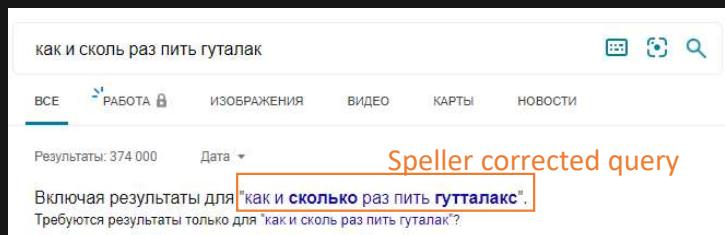
ko-KR

{맥북 윈도우 ghghks}
English Translation:
MacBook Windows
Compatible



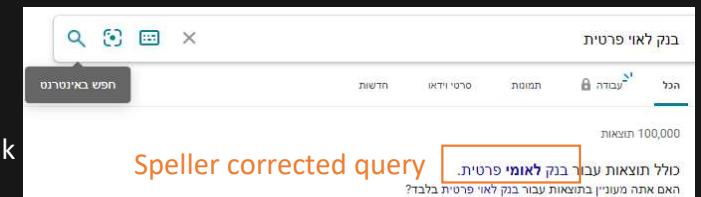
ru-RU

{как и сколь раз пить
гутталак}
English Translation:
how many times to
drink guttalax



he-IL

בנק לאומי פרטיט
English Translation:
Private National Bank



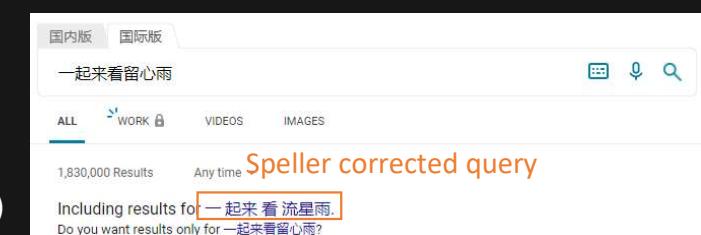
ja-JP

{田中みどらる}
English Translation:
Ms. Tanaka



zh-CN

{一起来看流星雨}
English Translation:
Watch meteor
together (TV series)



Scalability vs. Specificity

Scalability: scaling out to many languages

Specificity: language specific features

Japanese Speller

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

Example: みやさき vs. 宮崎

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

1 keystroke difference



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

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

▼ English

the crown has had its scandals

ALL WORK IMAGES VIDEOS MAPS NEWS SHOPPING

15,100,000 Results Any time ▾

['The Crown' Has Had Its Scandals, but There's Nothing Like ...](https://www.nytimes.com/2020/11/12/arts/television/the-crown-princess-diana.html)
 https://www.nytimes.com/2020/11/12/arts/television/the-crown-princess-diana.html
 Nov 12, 2020 · 'The Crown' Has Had Its Scandals, but There's Nothing Like Diana The face that launched a thousand tabloid stories – and books and documentaries and ...

['The Crown' Has Had Its Scandals, however There's Nothing ...](https://lightlynews.com/2020/11/12/arts/the-crown-has...)
 https://lightlynews.com/2020/11/12/arts/the-crown-has... ▾
 Nov 12, 2020 · 'The Crown' Has Had Its Scandals, however There's Nothing Like Diana by Lightlynews.com - On November 12, 2020 - In Arts / Television When we first glimpse her, minutes into...

[The Crown has had its scandals, but there's nothing like ...](https://celebrityml.com/entertainment/the-crown-has-had...)
 https://celebrityml.com/entertainment/the-crown-has-had... ▾
 Nov 21, 2020 · The Crown has had its scandals, but there's nothing like Diana 11/21/2020 The face that launched a thousand tabloid stories – and books and documentaries and Instagram feeds – takes cent...

['The Crown' Has Had Its Scandals, however There's Nothing ...](https://www.thinkipos.com/the-crown-has-had-its...)
 https://www.thinkipos.com/the-crown-has-had-its... ▾
 'The Crown' Has Had Its Scandals, however There's Nothing Like Diana November 12, 2020 November 12, 2020 - by Kumar - Leave a Comment Once we first glimpse her, minutes into Season four of "The Crown," ...

['The Crown' Has Had Its Scandals, however There's Nothing ...](https://www.thinkipos.com/the-crown-has-had-its-scandals...)
 https://www.thinkipos.com/the-crown-has-had-its-scandals... ▾
 'The Crown' Has Had Its Scandals, however There's Nothing Like Diana November 15, 2020 November 15, 2020 - by Kumar - Leave a Comment Once we first glimpse her, minutes into Season four of "The Crown," ...

espejos tocador con luces comprar

ALL WORK IMAGES VIDEOS MAPS NEWS SHOPPING

1,860,000 Results Any time ▾

[Amazon.es: tocador maquillaje con luz](https://www.amazon.es/tocador-maquillaje-luz/s?k=tocador+maquillaje+con+luz)
 https://www.amazon.es/tocador-maquillaje-luz/s?k=tocador+maquillaje+con+luz ▾
 BEAUTME Espejo de tocador con Luces, tocador de Maquillaje Iluminado o Espejos de Belleza montados en la Pared con atenuador, Espejo cosmético Hollywood con 15 Bombillas LED (Plateado 99,99 € 99,99 €)

[Espejo Tocador Con Luces en Mercado Libre México](https://listado.mercadolibre.com.mx/espejo-tocador-con-luces)
 https://listado.mercadolibre.com.mx/espejo-tocador-con-luces ▾
 Encuentra Espejo Tocador Con Luces en Mercado Libre México. Descubre la mejor forma de comprar online.

[Mejor espejo tocador con luces: ofertas y reseñas 2020](https://quecomprar.org/espejo-tocador-con-luzes)
 https://quecomprar.org/espejo-tocador-con-luzes ▾
 Que Comprar ha desarrollado una rápida selección de los mejores modelos de espejo tocador con luces disponibles en línea a los mejores precios. Índice de contenidos 🏆 Top 5 mejores espejos tocadores con luces: bestseller en Amazon

processo histórico car

TUDO TRABALHO

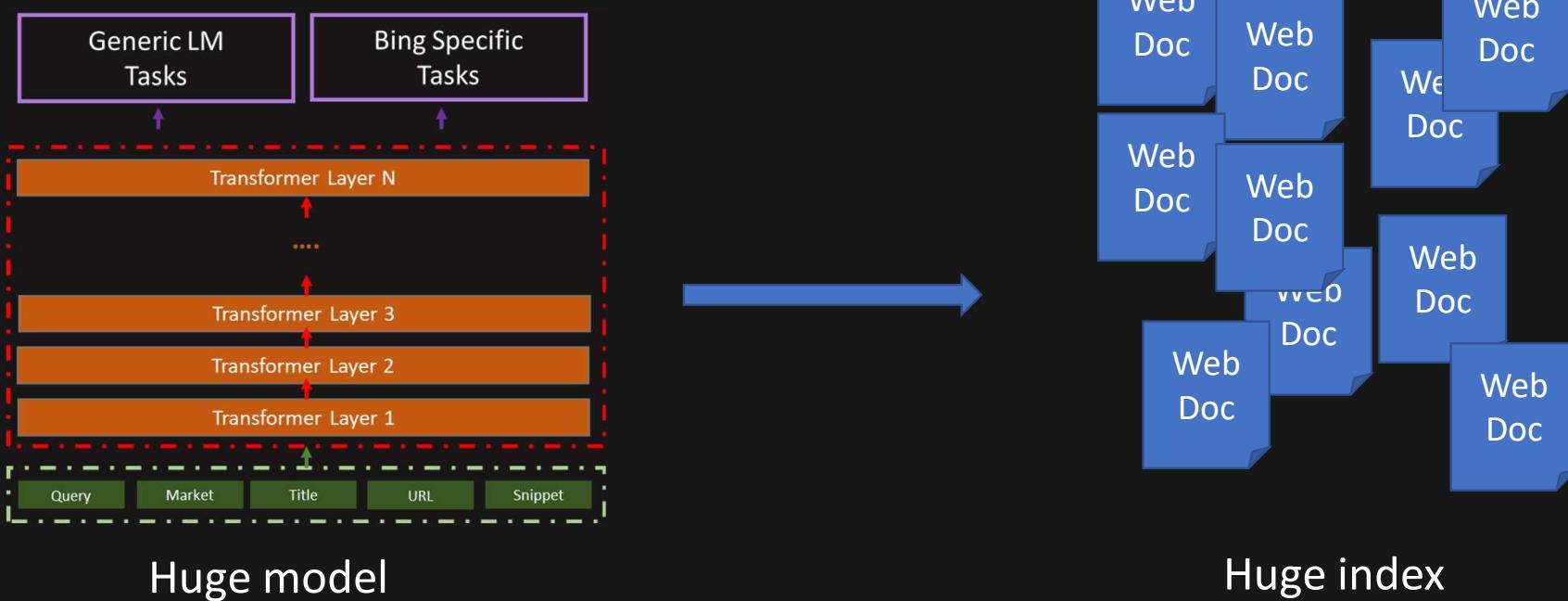
1.170.000 Resultados

[História do Canadá](https://www.infoescola.br/historia-do-canada)
 https://www.infoescola.br/historia-do-canada
 História do Canadá. Os primeiros habitantes (algonquinos, esquimós, iroqueses) e sua cultura. A colonização europeia, a independência e o processo de unificação do Canadá. Os principais eventos da história canadense.

[Canadá - História](https://brasilescola.uol.com.br/historia/canada.htm)
 https://brasilescola.uol.com.br/historia/canada.htm
 Canadá História da América do Norte. A origem dos povos canadenses, suas tradições, costumes, cultura, economia, política, geografia, ecologia, turismo, gastronomia, etc.

[História do Canadá](https://pt.wikipedia.org/wiki/Hist%C3%B3ria_do_Canad%C3%A1)
 https://pt.wikipedia.org/wiki/Hist%C3%B3ria_do_Canad%C3%A1
 20/02/2005 · A história do Canadá abrange os últimos 10 mil anos, desde os primeiros habitantes da América do Norte até os dias atuais.

Efficiency



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

English

 Mary Smith <wrbutpt1013@gmail.com>    

Tue 2020-11-17 17:06
To: Melody Wang

Thanks for the heads up. Did your friends wonder why you had invited the paper boy to your party?

Are the suggestions above helpful? Yes No

[Reply](#) | [Forward](#)

Spanish

 Mary Smith <wrbutpt1013@gmail.com>    

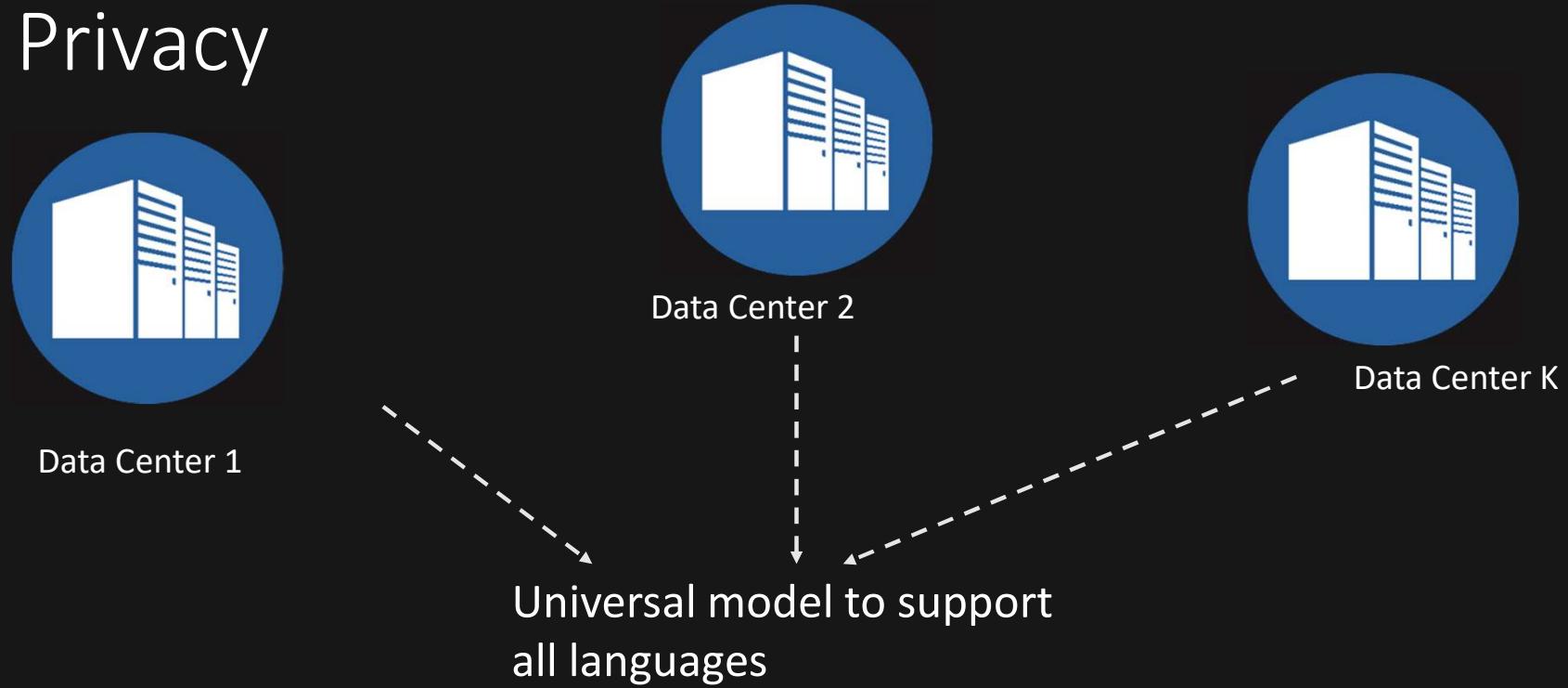
Tue 2020-11-17 21:34
To: Melody Wang

Necesita los originales. Dámelos hoy en la oficina.
Gustavo

Are the suggestions above helpful? Yes No

[Reply](#) | [Forward](#)

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**

Teams Voice Skill:

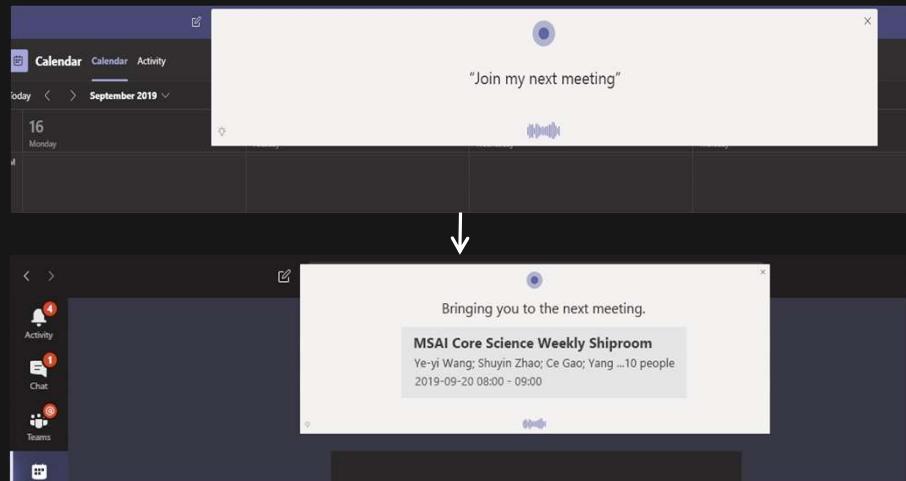
Join my next meeting

Domain: Calendar

Intent: Connect_to_meeting

Slot: <start_time>next</start_time>

Desktop Experience



Start the presentation

Quando è il mio incontro con Tom

打开未读消息

Message à Carlos que je serai en retard

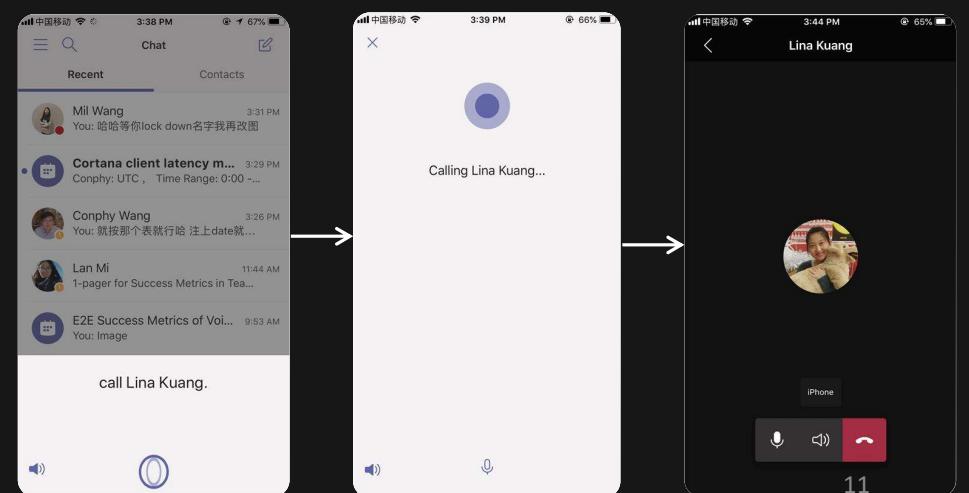
Agrega Alice a este chat

Compartilhe o arquivo com Bob

次の会議に参加する

Füge Alice zum Cortana - Kanal hinzu

Mobile Experience

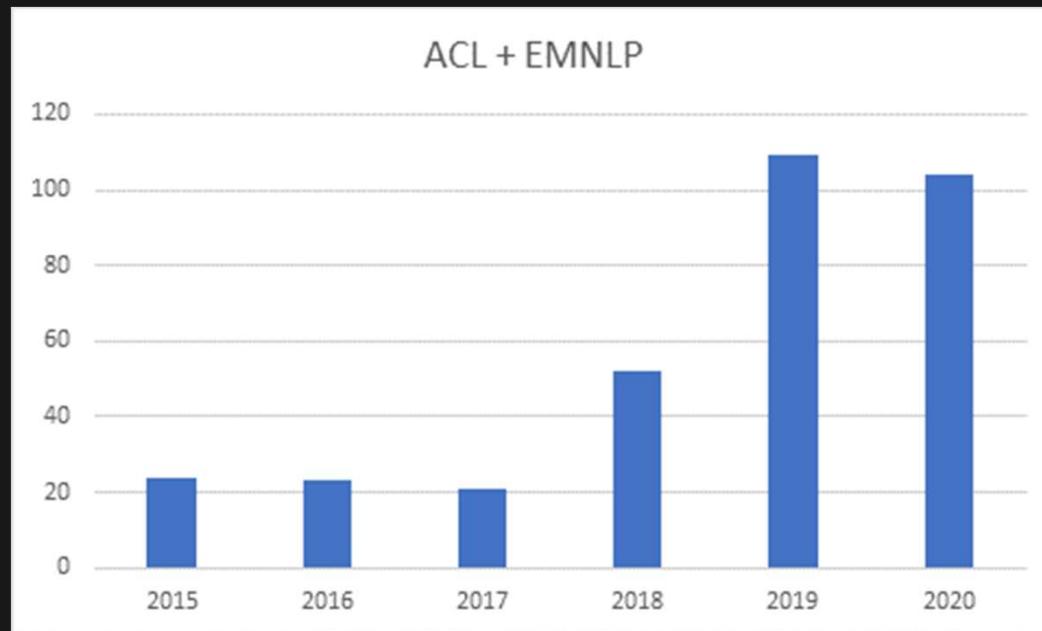


M Domains × N Languages

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

Why Language Matters (Research)

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



Problem Description for Language Scaling



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

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

Source: Both Q and D are in English

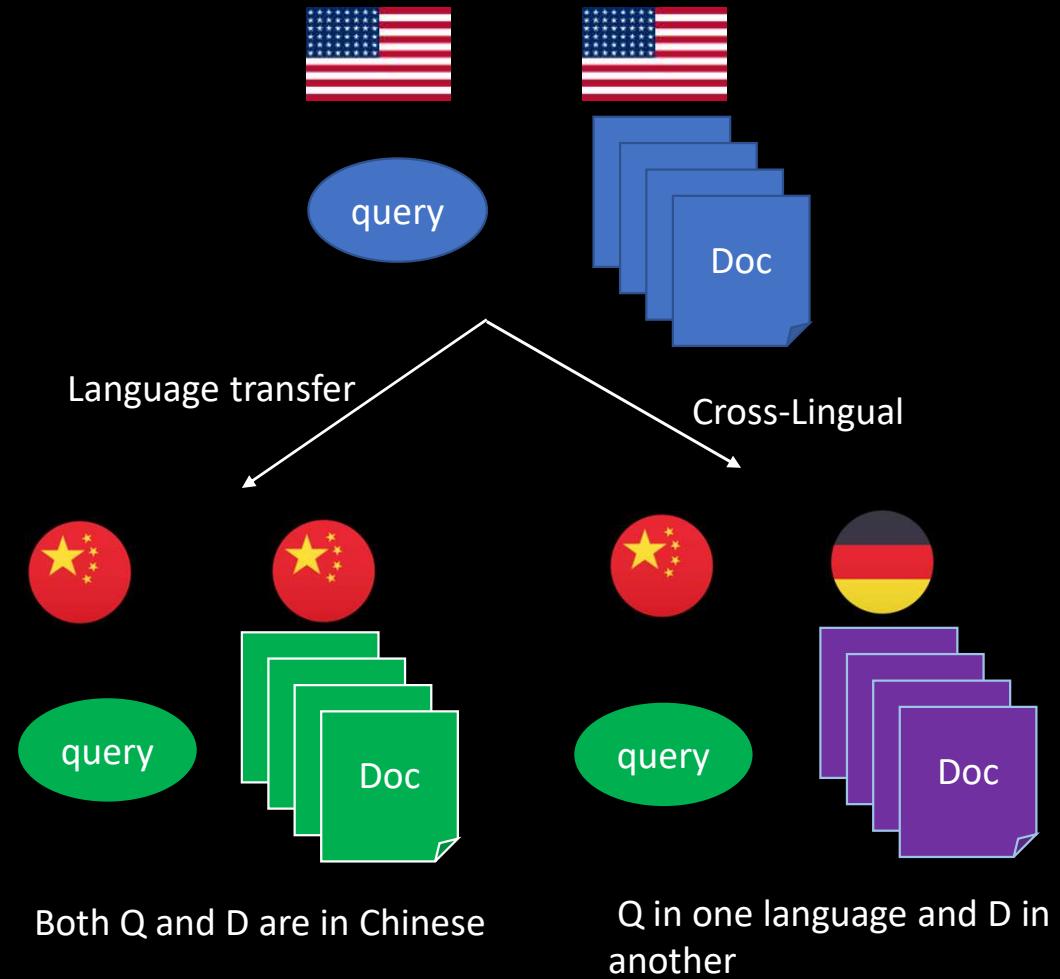
A Related Problem

- **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

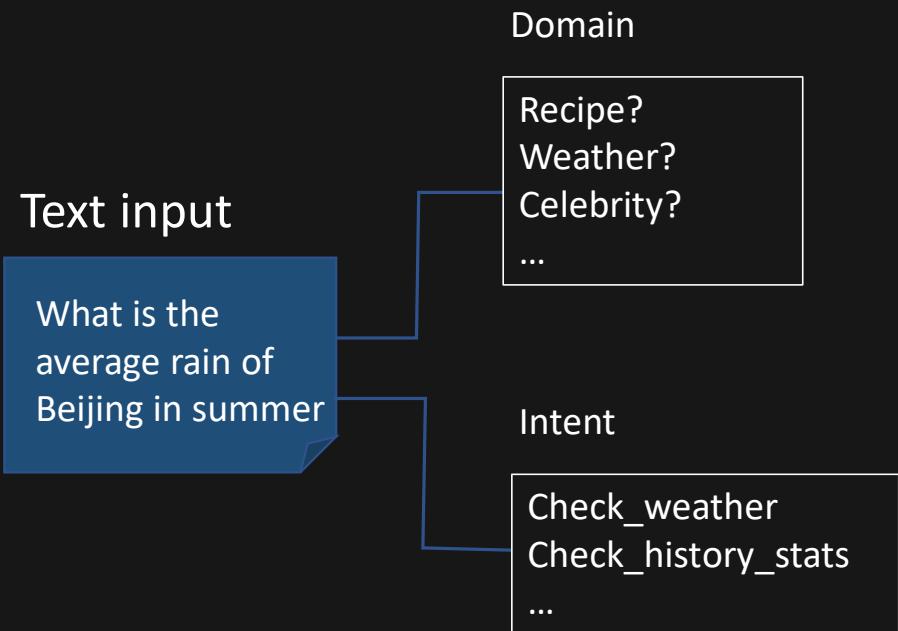


NLP Tasks

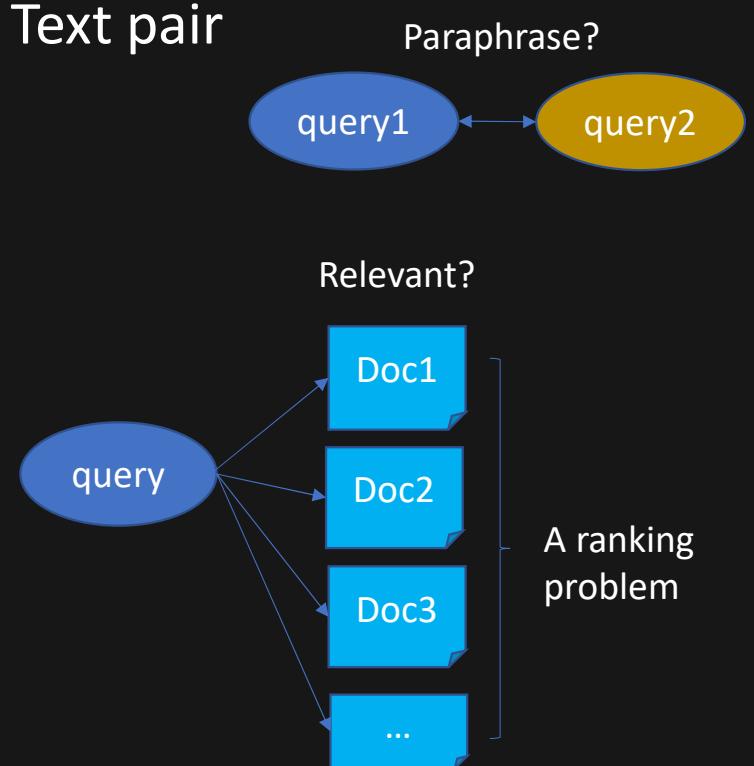
Type	Category	Sub Category	Example
NLU	Text Classification	Single text	Domain identification, Intent detection, Sentiment classification
		Text pair	Information retrieval, Natural language inference
	Sequence Labeling	Single text	Named entity recognition, Slot tagging
		Text pair	Machine reading comprehension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

Text Classification

- Single text



- Text pair

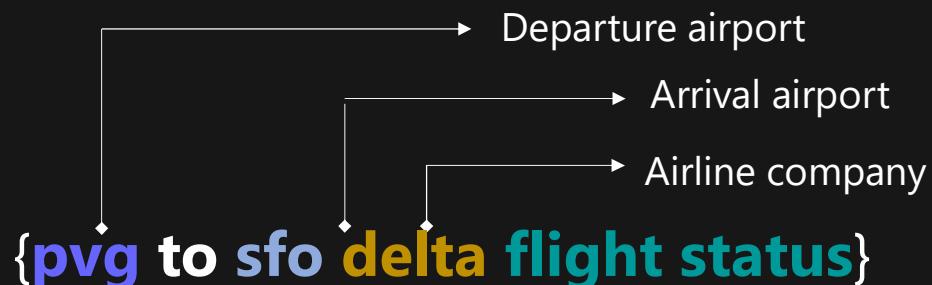


Sequence Labeling

- 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

We invite contributions to the research track of The Web Conference 2021 (formerly known as WWW). The conference will take place in Ljubljana, Slovenia, **April 19 to April 23, 2021**. Instructions for Authors of Research Track submissions

[Call for Papers | The Web Conference - 2021
www2021.thewebconf.org/authors/call-for-papers/](http://www2021.thewebconf.org/authors/call-for-papers/)

Was this helpful?

Text Generation

Spelling Correction

Britnay spears vidios

Britney speaks videos

Sentence Auto Completion

The web conference

The web conference is a yearly international

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

Sentence level

Machine Translation

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

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

Conversation

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

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

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

Response: ???

*EMPATHETICDIALOGUES

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

Challenges for Language Scaling

Lack of Training Data

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

Too Many Models

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

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

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

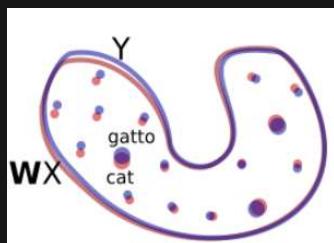
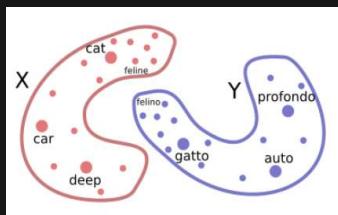
- Multi-task learning
- Adapters

- Hardware acceleration
- Computation library optimization
- Model compression

Approaches: Model Transfer and Data Transfer

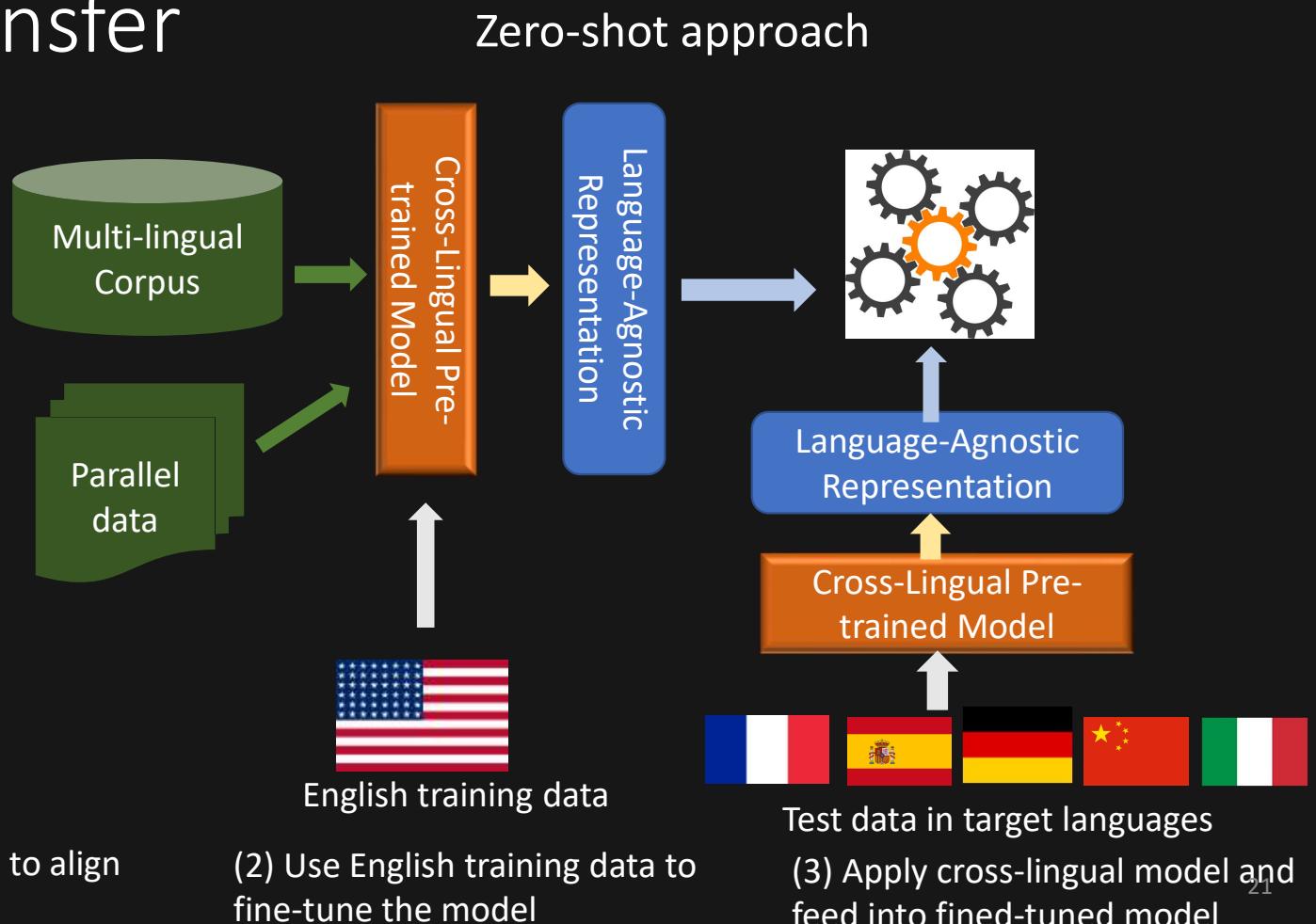
(1) Model Transfer

Cross-lingual model



From MUSE

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



English training data

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

Test data in target languages

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

Zero-shot approach

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

$$J = \underbrace{\mathcal{L}(X^s) + \mathcal{L}(X^t)}_1 + \underbrace{\Omega(\underline{X^s}, \underline{X^t}, W)}_2$$

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



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

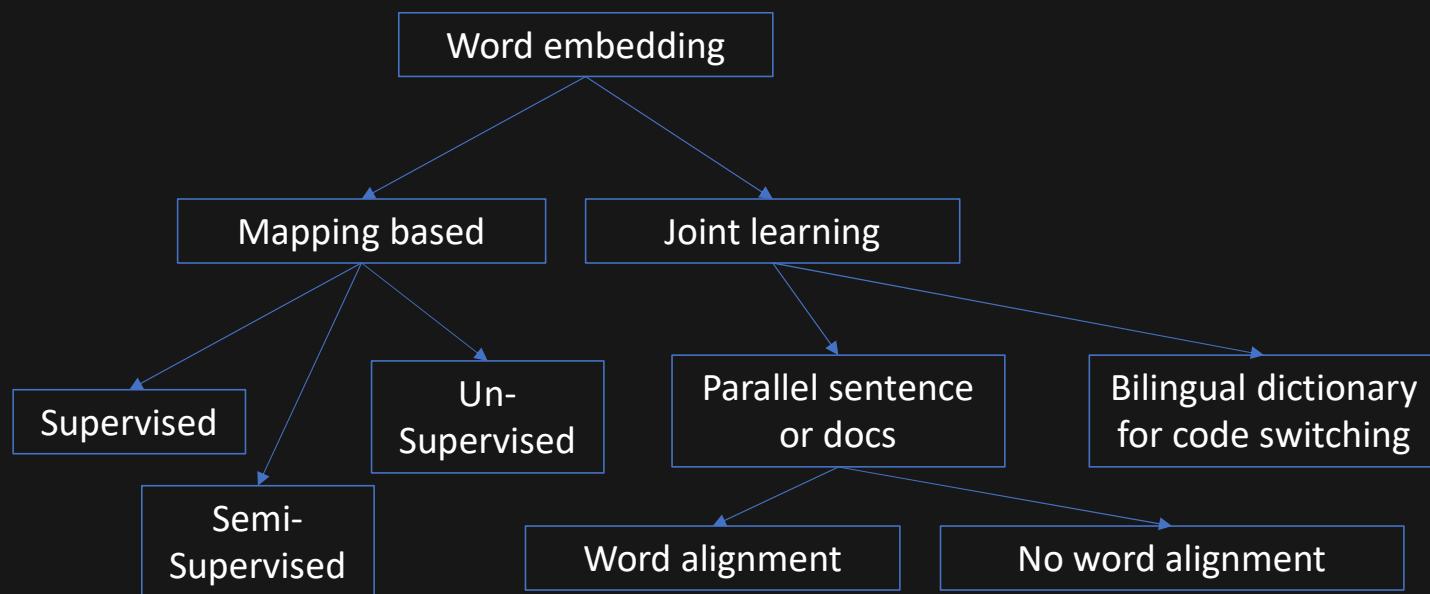
\mathbf{W} : mapping matrix from X^s to X^t

- Joint-learning methods

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

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

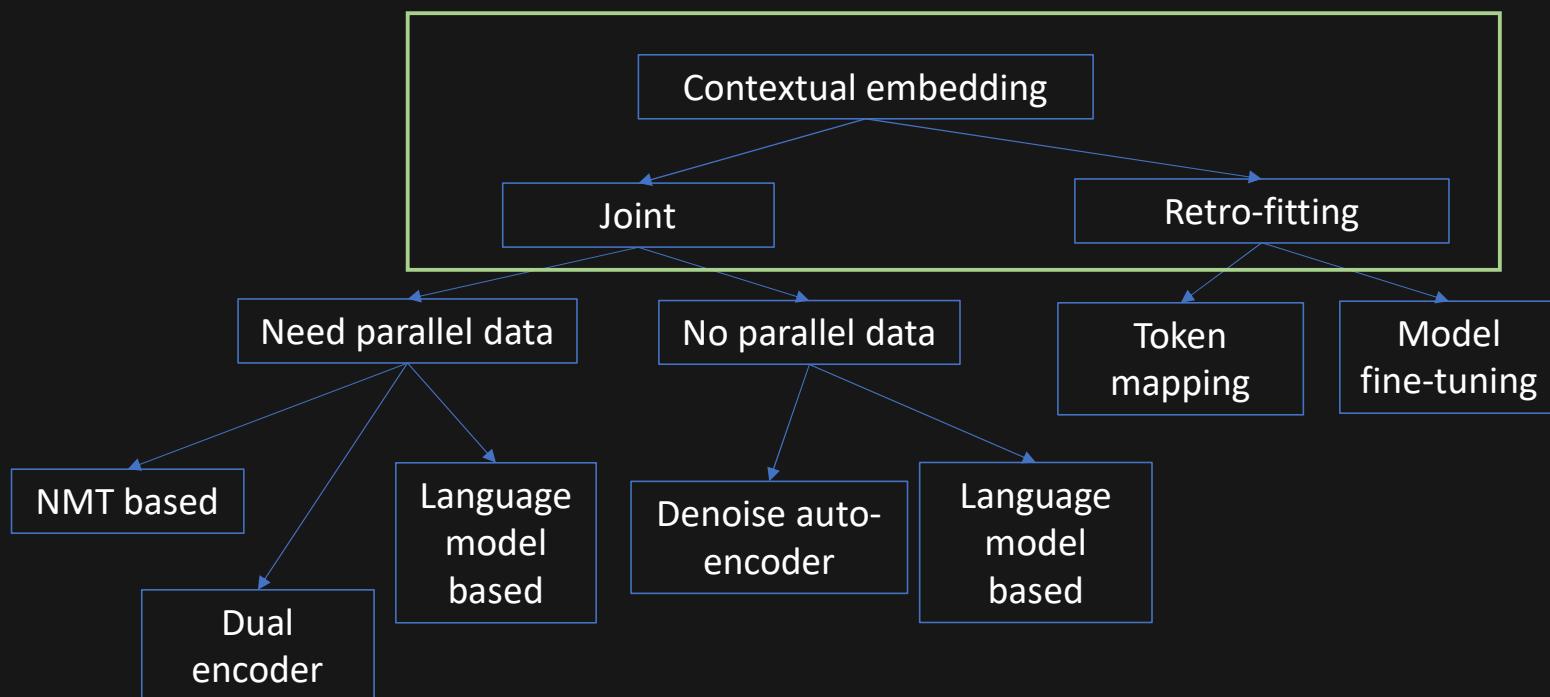
Cross-Lingual Word Embedding



Sebastian Ruder, Ivan Vulić, Anders SØgaard

- A survey of cross-lingual word embedding models,. Journal of Artificial Intelligence Research, May 2019.
- 2019 ACL Tutorial [Unsupervised Cross-lingual Representation Learning \(ruder.io\)](https://ruder.io/Unsupervised-Cross-lingual-Representation-Learning.html)

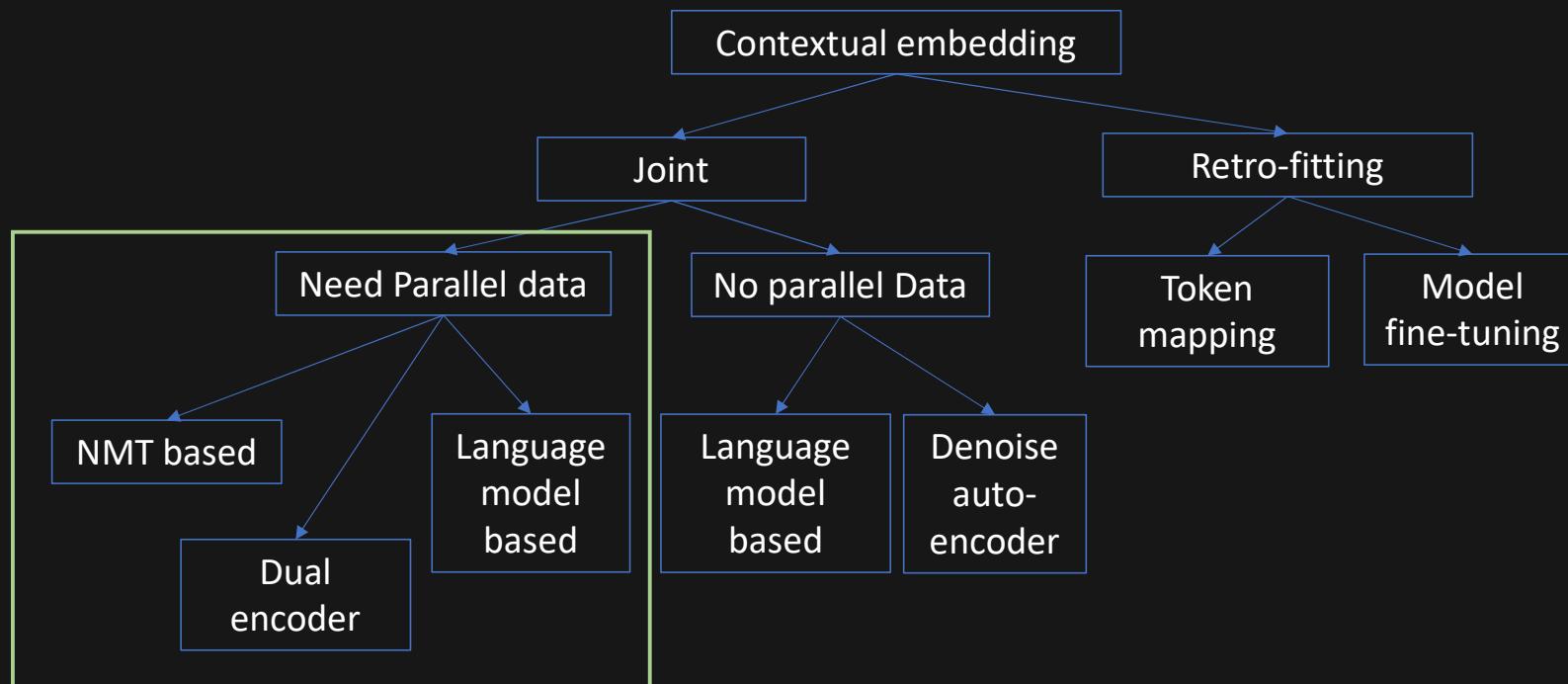
Cross-Lingual Contextual Embedding



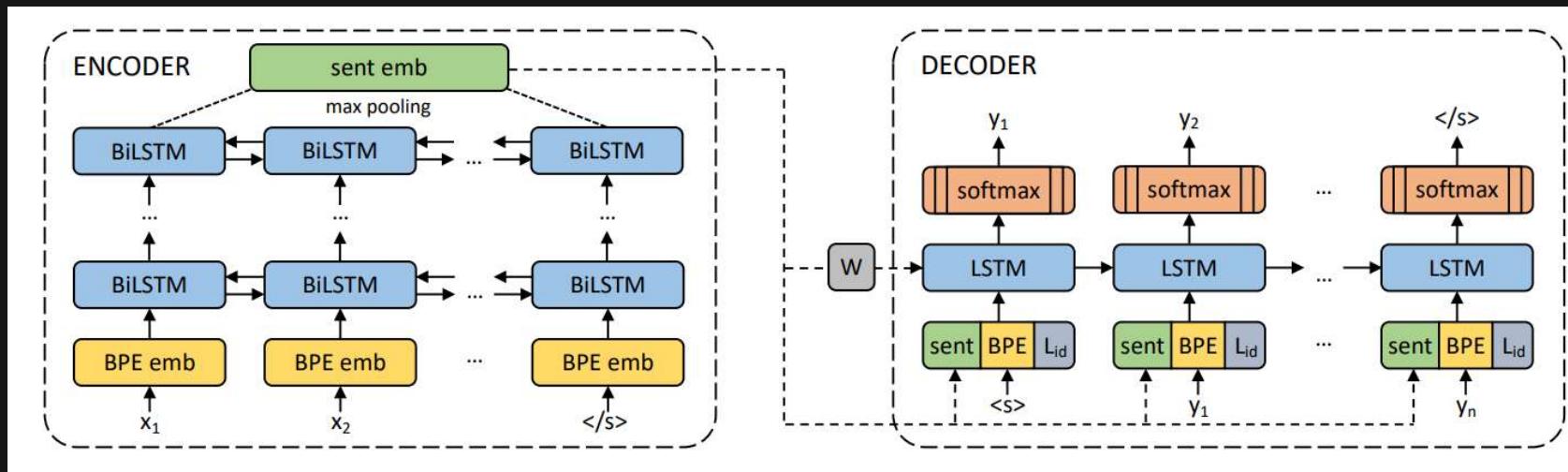
Joint or Retro-fitting?

- *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



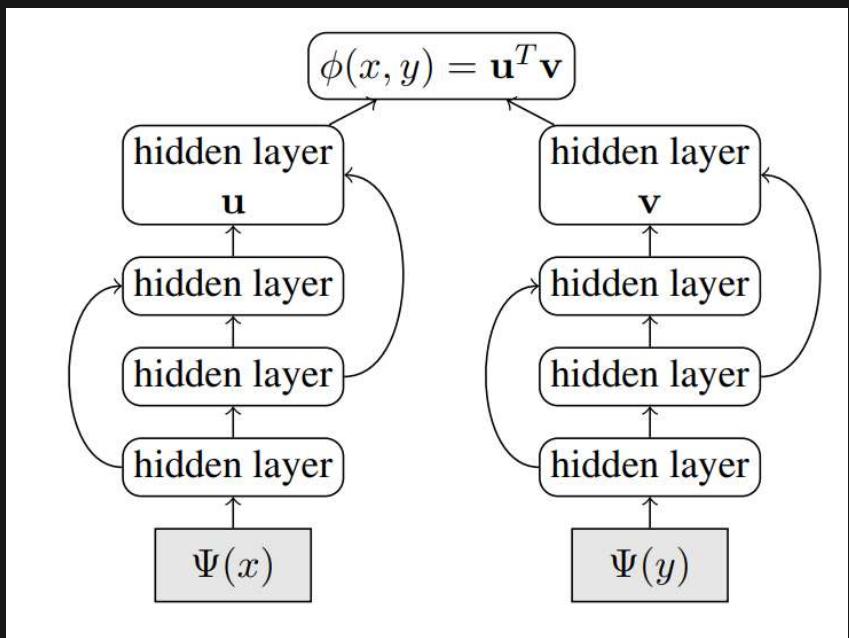
Neural Machine Translation (NMT) based Method



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

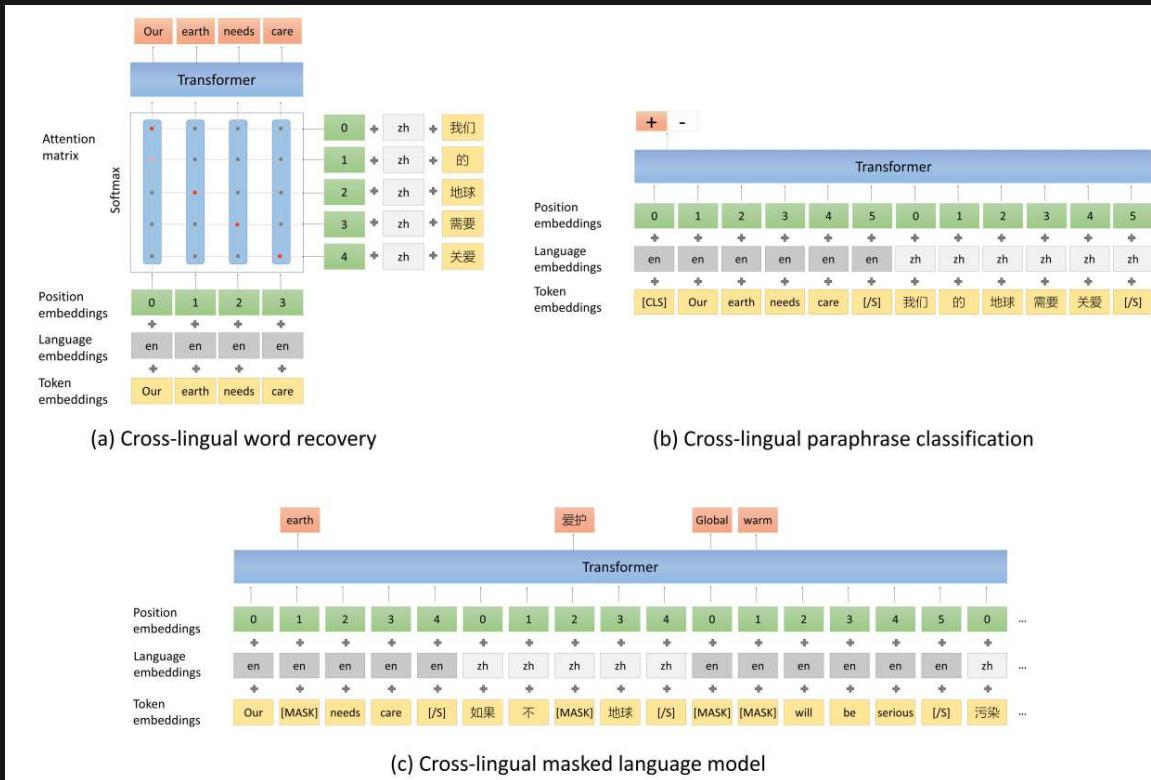


- x is in one language, and y is in another language
- Using parallel data to align the encoding u and v
- Dual encoder is explicit alignment, while NMT based approach is implicit alignment

Mandy Guo, Qinlan Shen, Yinfai Yang, et al.

Effective Parallel Corpus Mining using Bilingual Sentence Embeddings. WMT 2018: 165-176

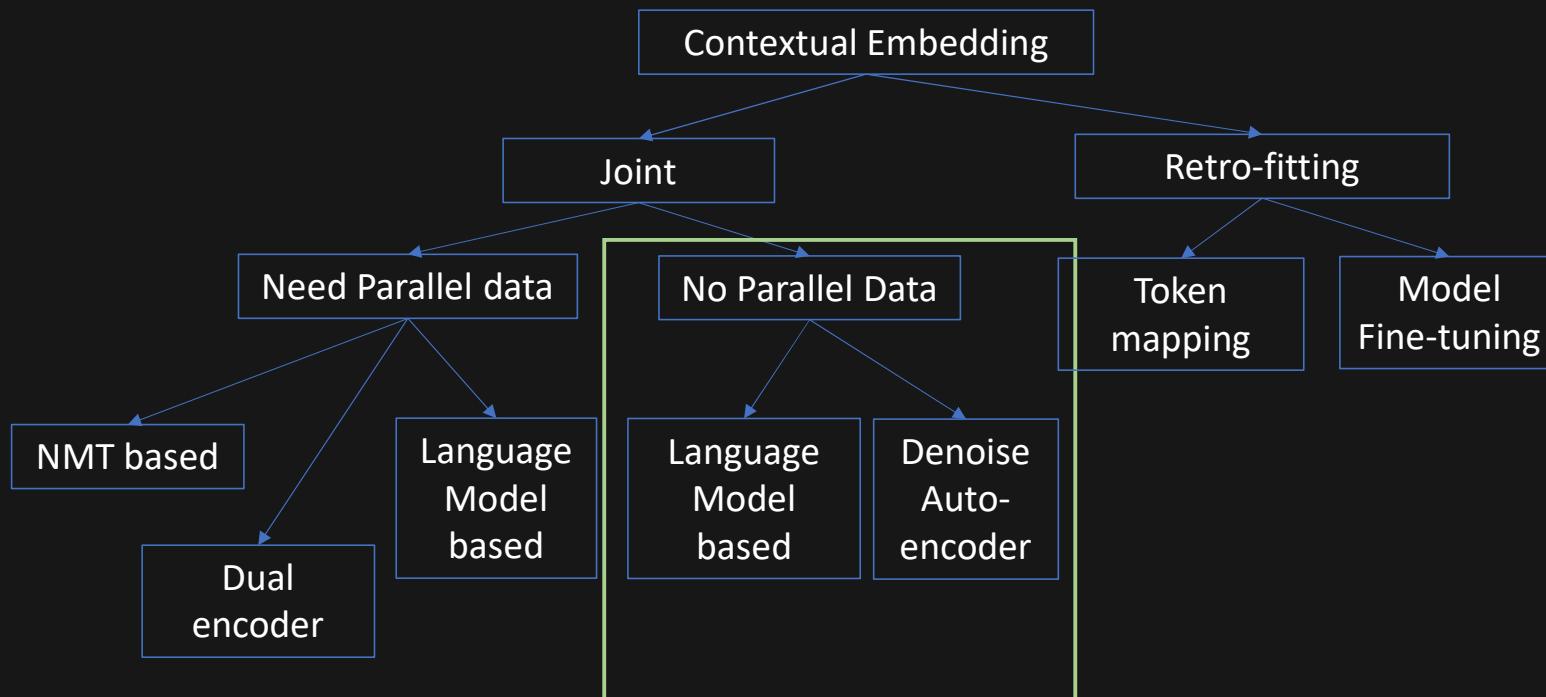
Language Model based Methods



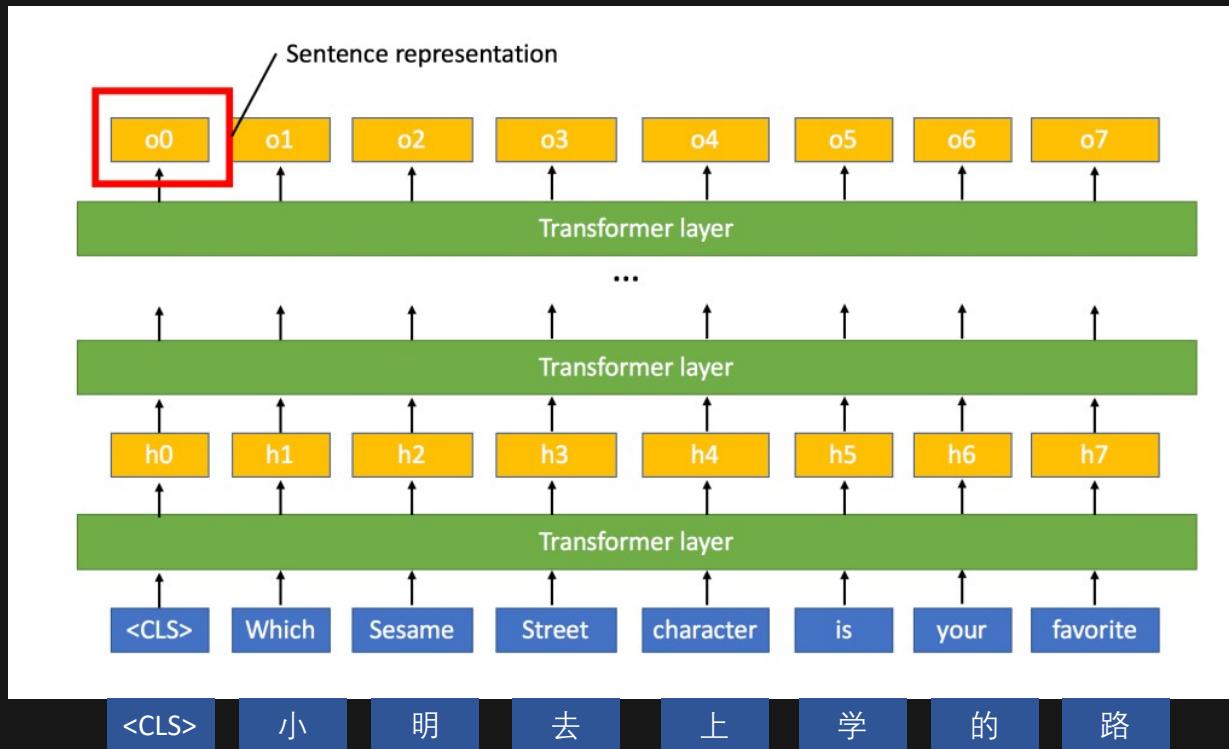
- Transformer based language model
- Cross-lingual representation at **both sentence level and token level**

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



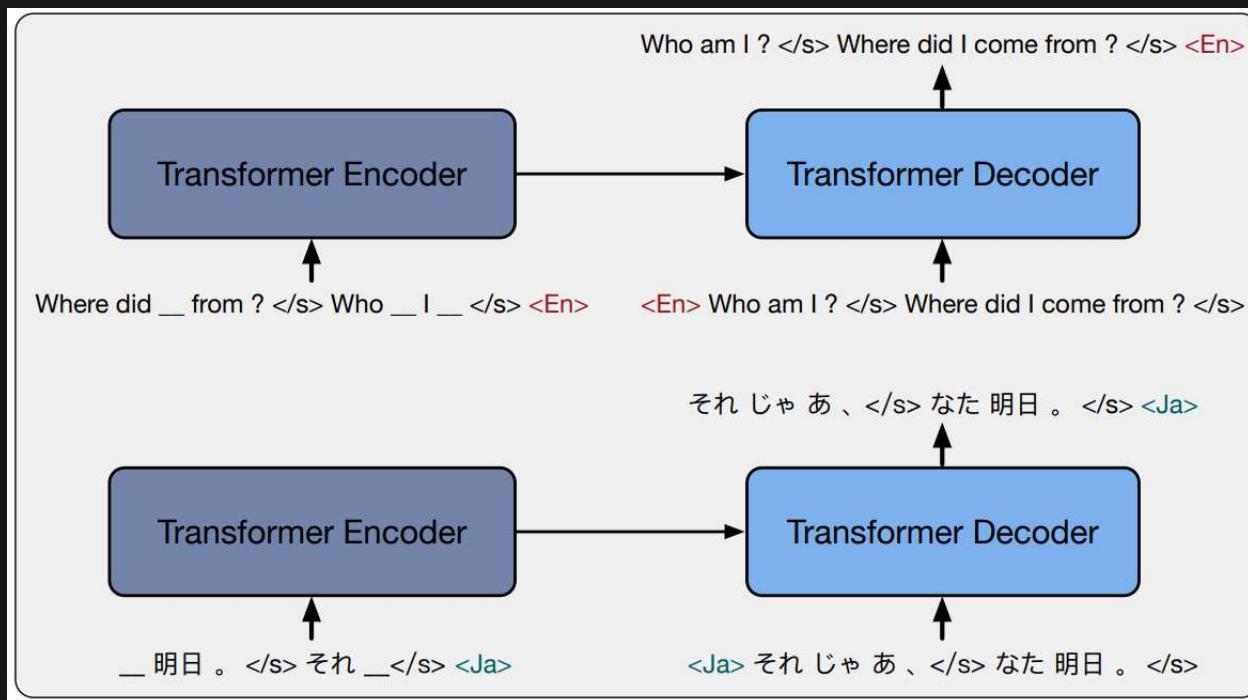
- Share the same architecture and training scheme with monolingual BERT
- 104 languages Wikipedia pages
- 110k shared WordPiece vocabulary

[bert/multilingual.md at master · google-research/bert · GitHub](https://github.com/google-research/bert/blob/master/multilingual.md)

Why mBERT has Cross-Lingual Effectiveness

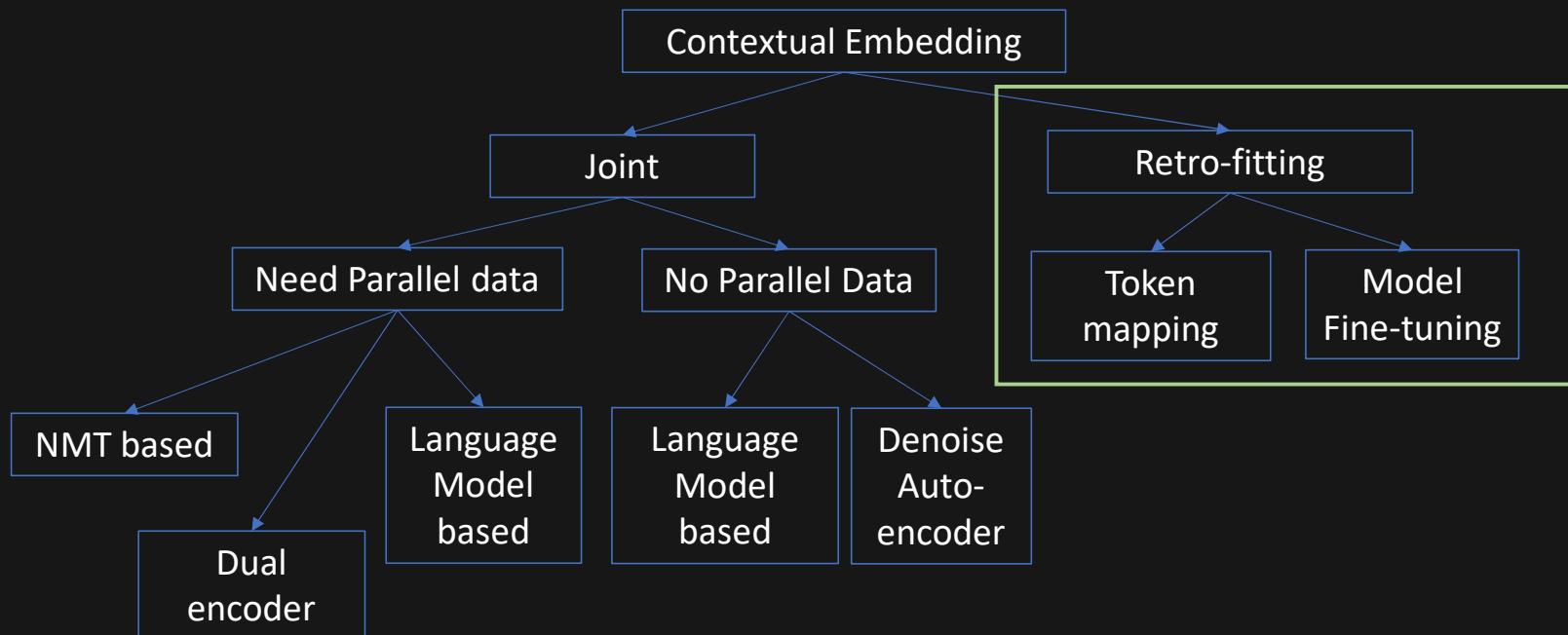
- Vocabulary overlap in different languages
 - Cognate words
 - Foreign words
 - However, it has been shown that mBERT still has cross-lingual effectiveness when there is zero vocabulary overlap
- Common structure in different languages
 - Word order
 - Most frequent words
- Pires T. et al. How multilingual is Multilingual BERT? ACL'19.
- Wu, S. and Dredze M. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT, EMNLP'19.
- Artetxe M. et al. On the Cross-lingual Transferability of Monolingual Representations, ACL'20.
- Conneau, A. et al. Emerging Cross-lingual Structure in Pretrained Language Models, ACL'20.
- Karthikeyan, K. et al. Cross-Lingual Ability of Multilingual BERT: An Empirical Study, ICLR'20.
- Libovický, J. et al. On the Language Neutrality of Pre-trained Multilingual Representations, EMNLP Findings'20.

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

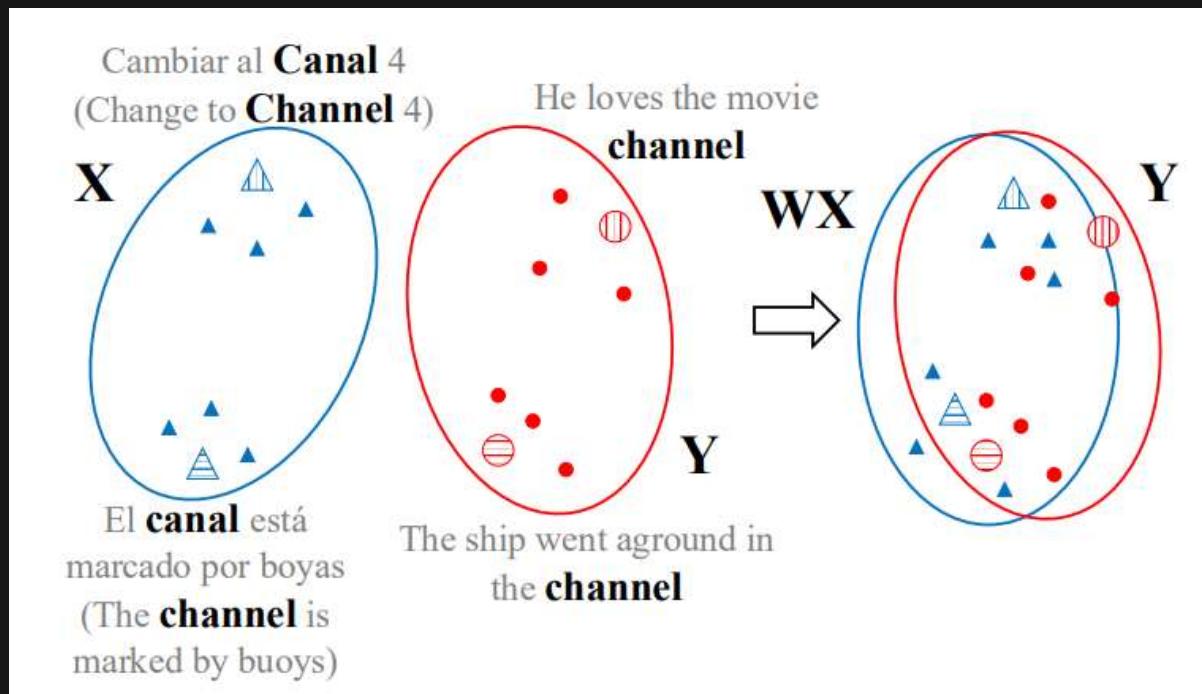


- Same architecture and training scheme with monolingual BART
- 25 languages crawled web pages
- 250K shared Sentence Piece vocabulary

Joint learning without Parallel Data

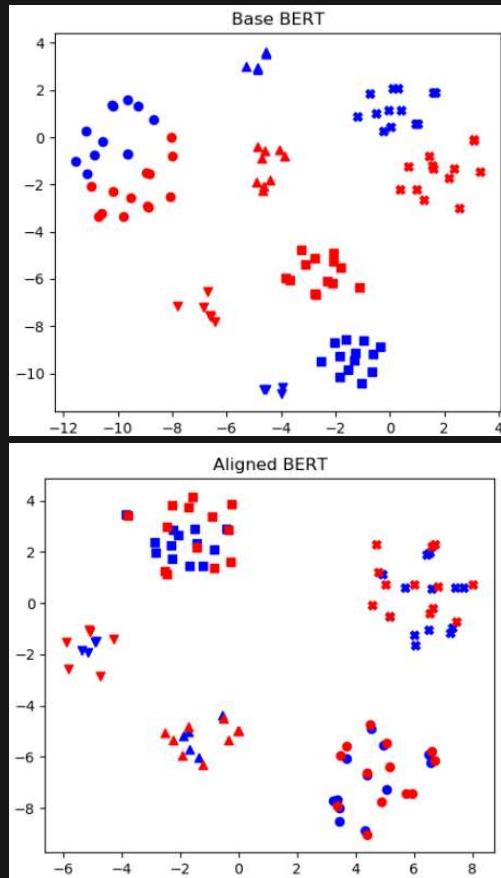


Contextual Token Mapping



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

Model Fine-Tuning



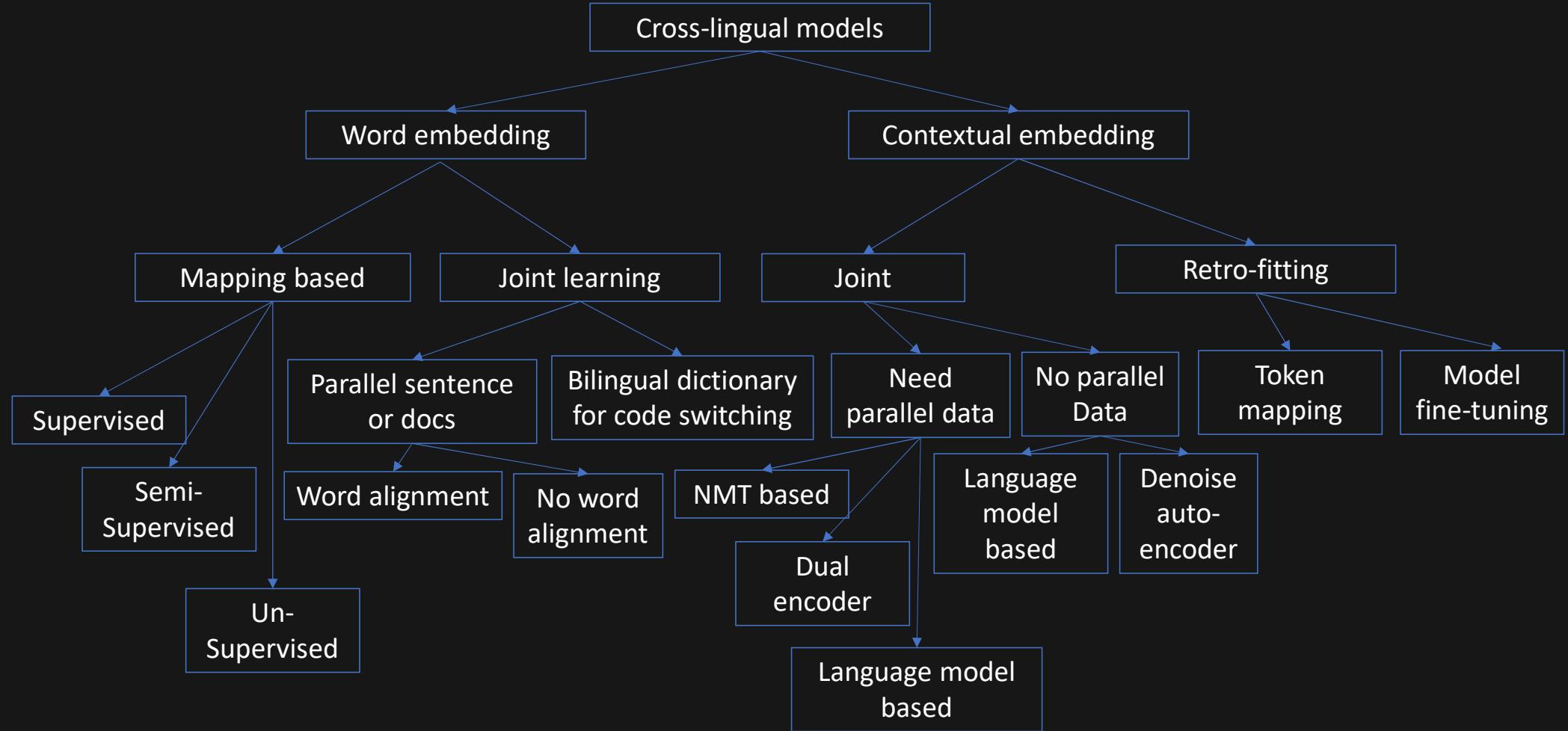
- Upper figure: mBERT roughly aligns the words
- Lower figure: effect of fine-tuning
- After mBERT training, fine-tuning by

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

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

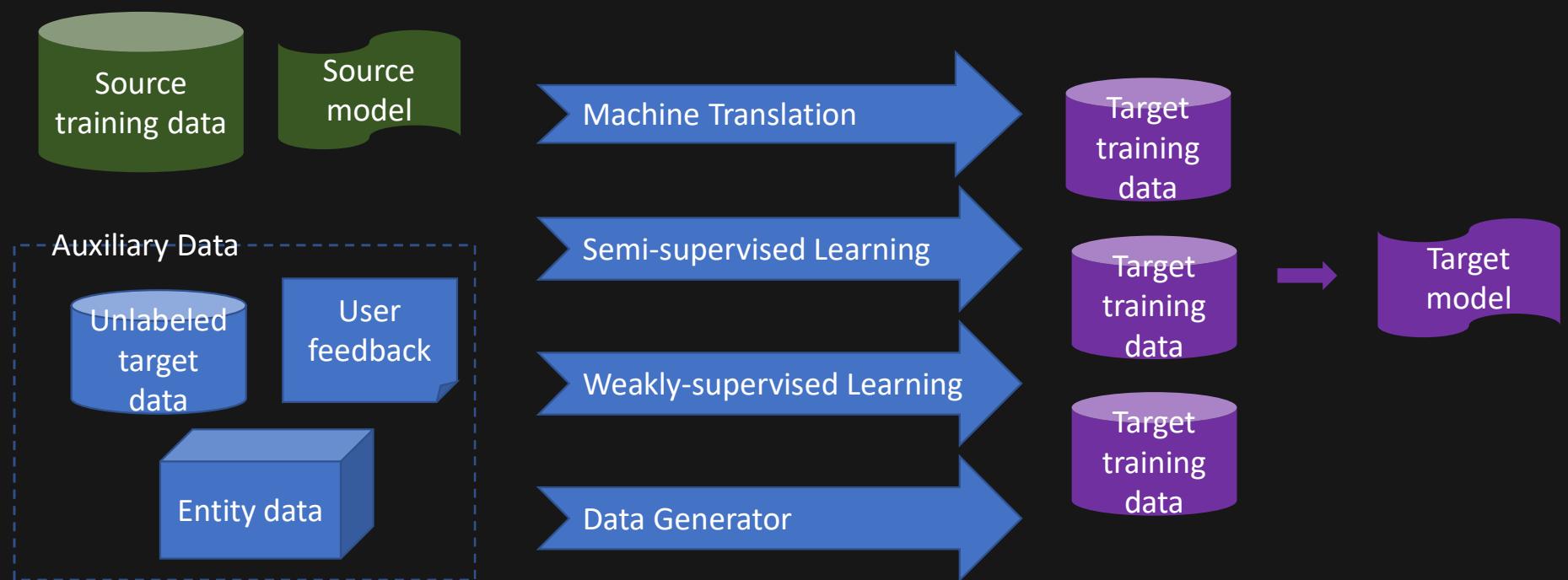
$$R(f, C) = \sum_{t \in C} \sum_{i=1}^{\text{len}(t)} \|f(i, t) - f_0(i, t)\|^2$$

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



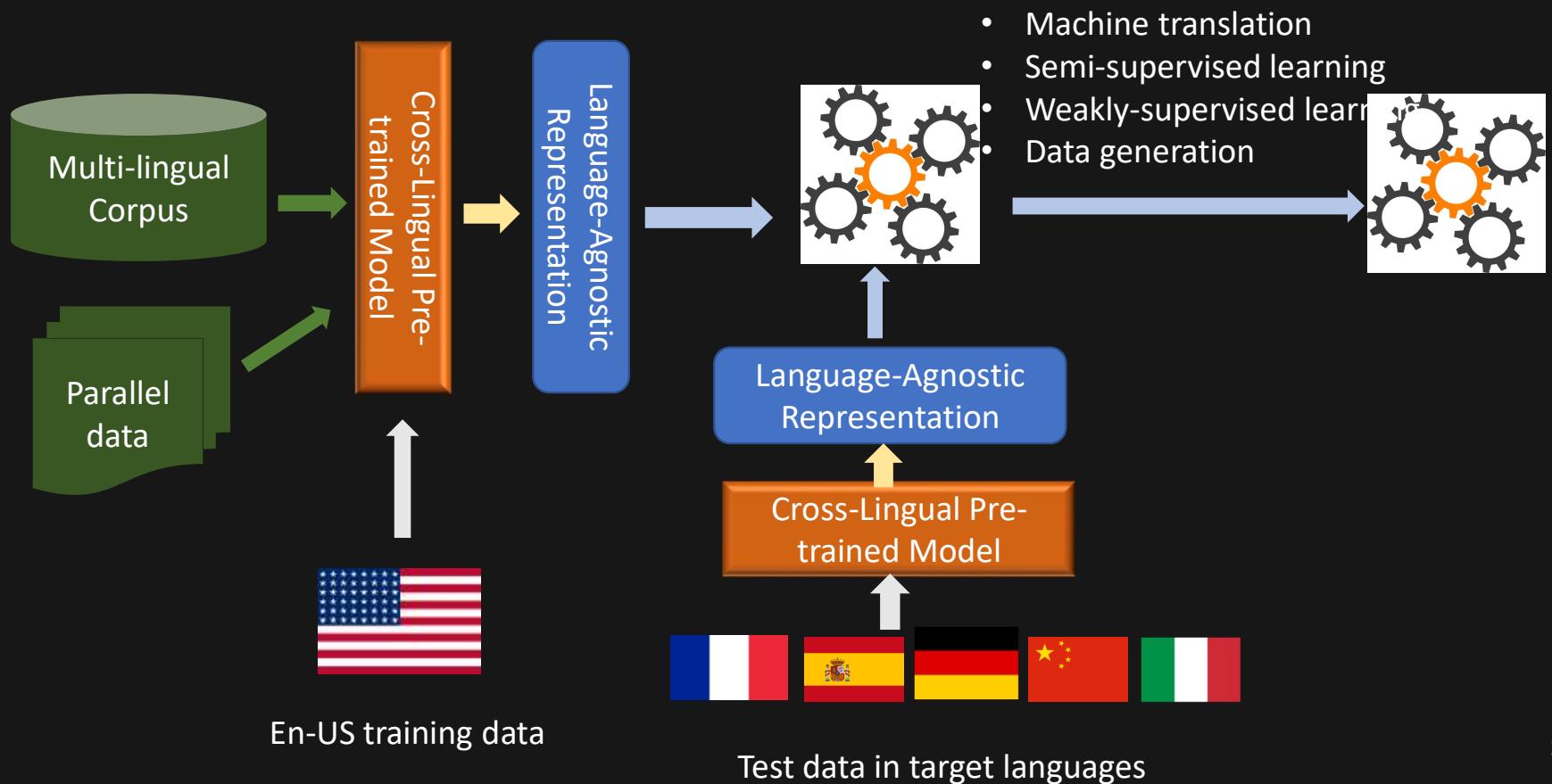
Approaches: Model Transfer and Data Transfer

(2) Data Transfer



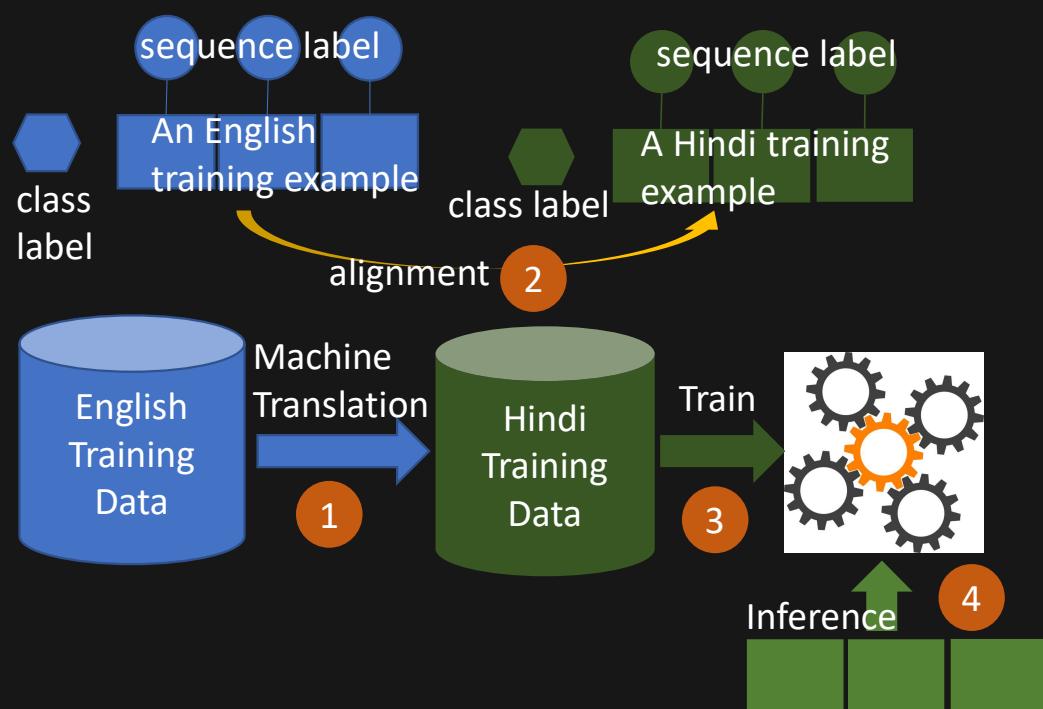
Approaches: Model Transfer and Data Transfer

(1) Model Transfer + (2) Data Transfer

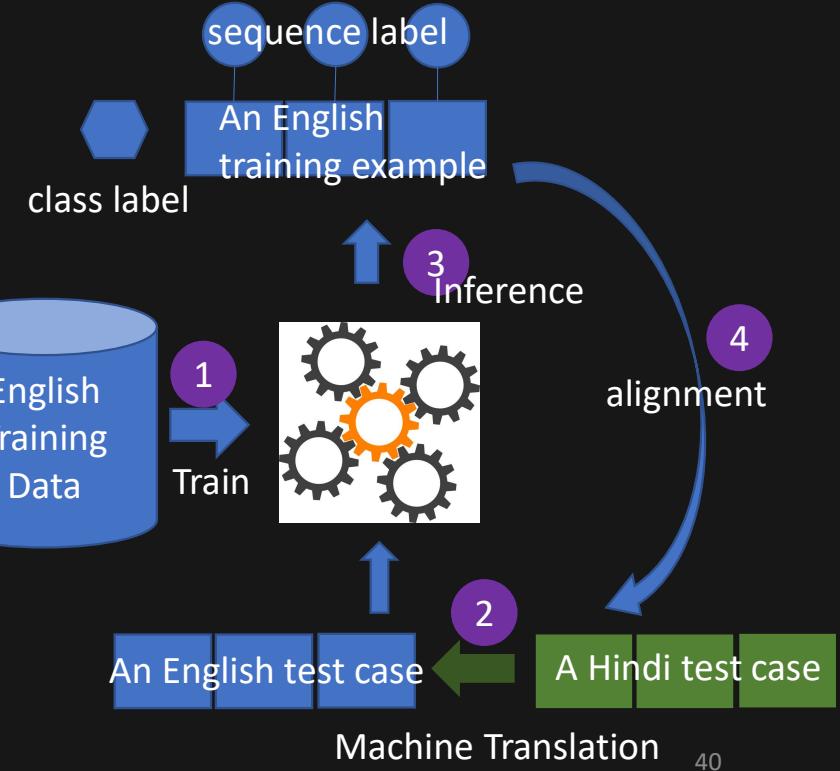


Machine Translation

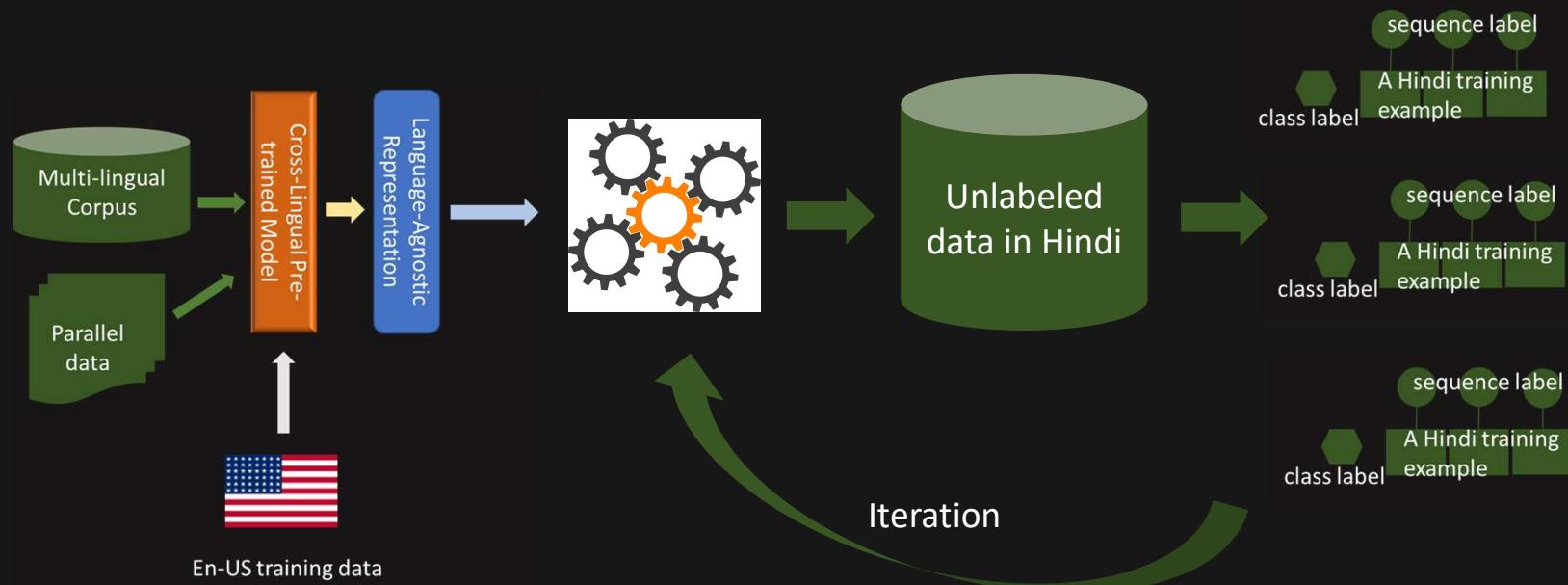
- Translate-on-train



- 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*}

We invite contributions to the research track of The Web Conference 2021 (formerly known as WWW). The conference will take place in Ljubljana, Slovenia, **April 19 to April 23, 2021**. Instructions for Authors of Research Track submissions

Call for Papers | The Web Conference - 2021
www2021.thewebconf.org/authors/call-for-papers/

Was this helpful?  

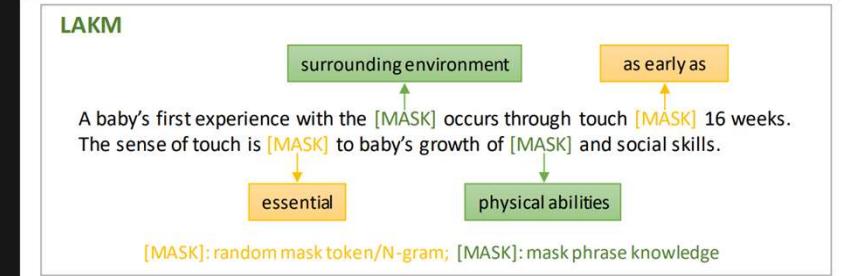
- Rich user behaviors in search logs
- Any single behavior is a weak signal, but more powerful when combined in a ML model
- Generally applicable to all languages

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

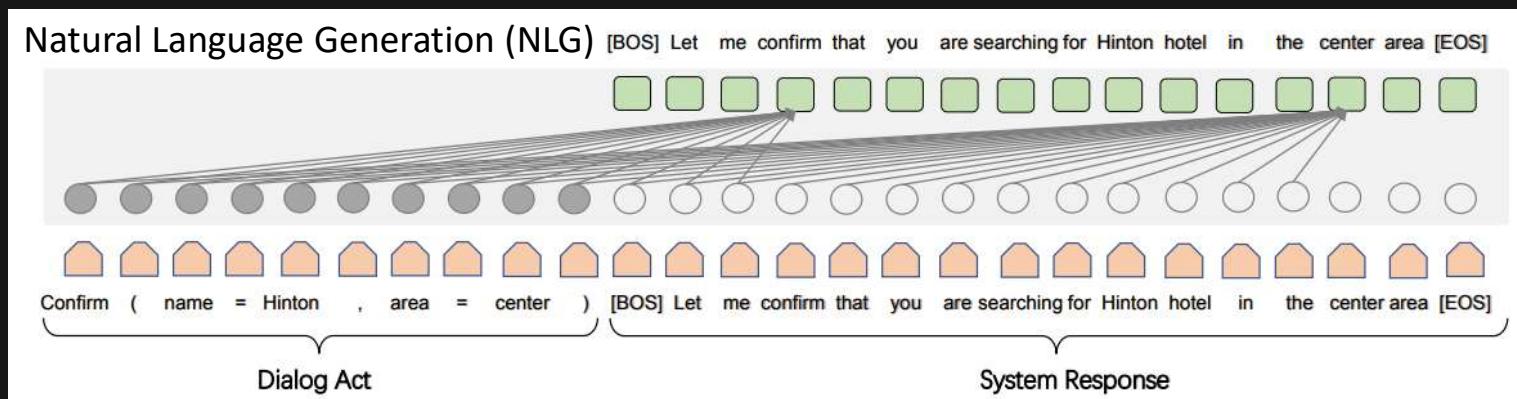
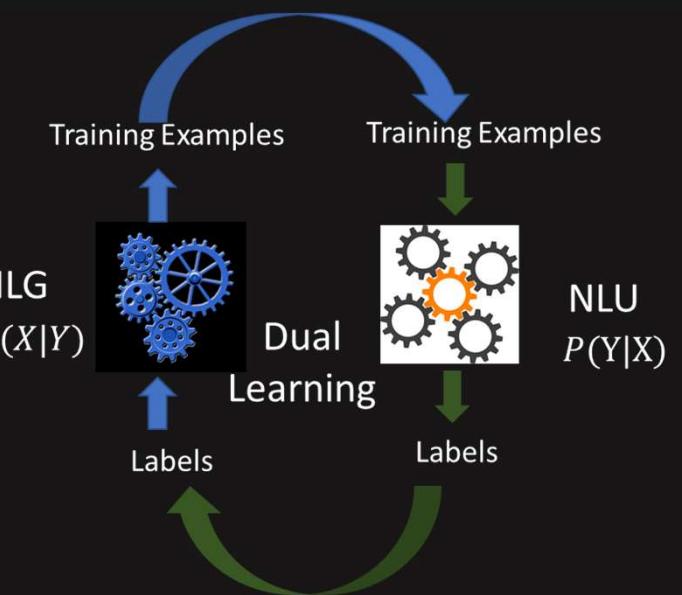
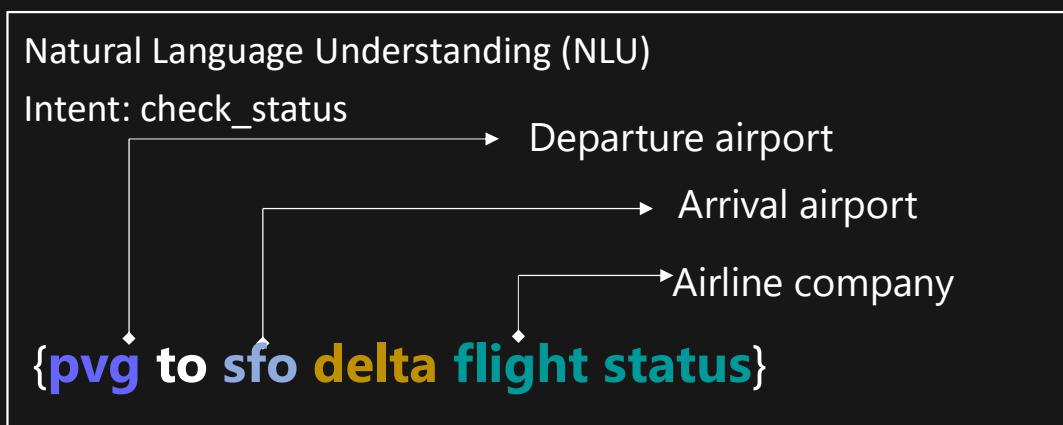
Examples of boundary detection error



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



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

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

Summary: Data Transfer Methods

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

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

Outline



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

- Introduction [Dixin 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 [Dixin Jiang]



Linjun Shou



Xiubo Geng



Ming Gong



Cross Lingual Natural Language Inference

Linjun Shou
lisho@microsoft.com

Cross-lingual Natural Language Inference

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/Neutral/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 يمكن الوكالات أن تعرف 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

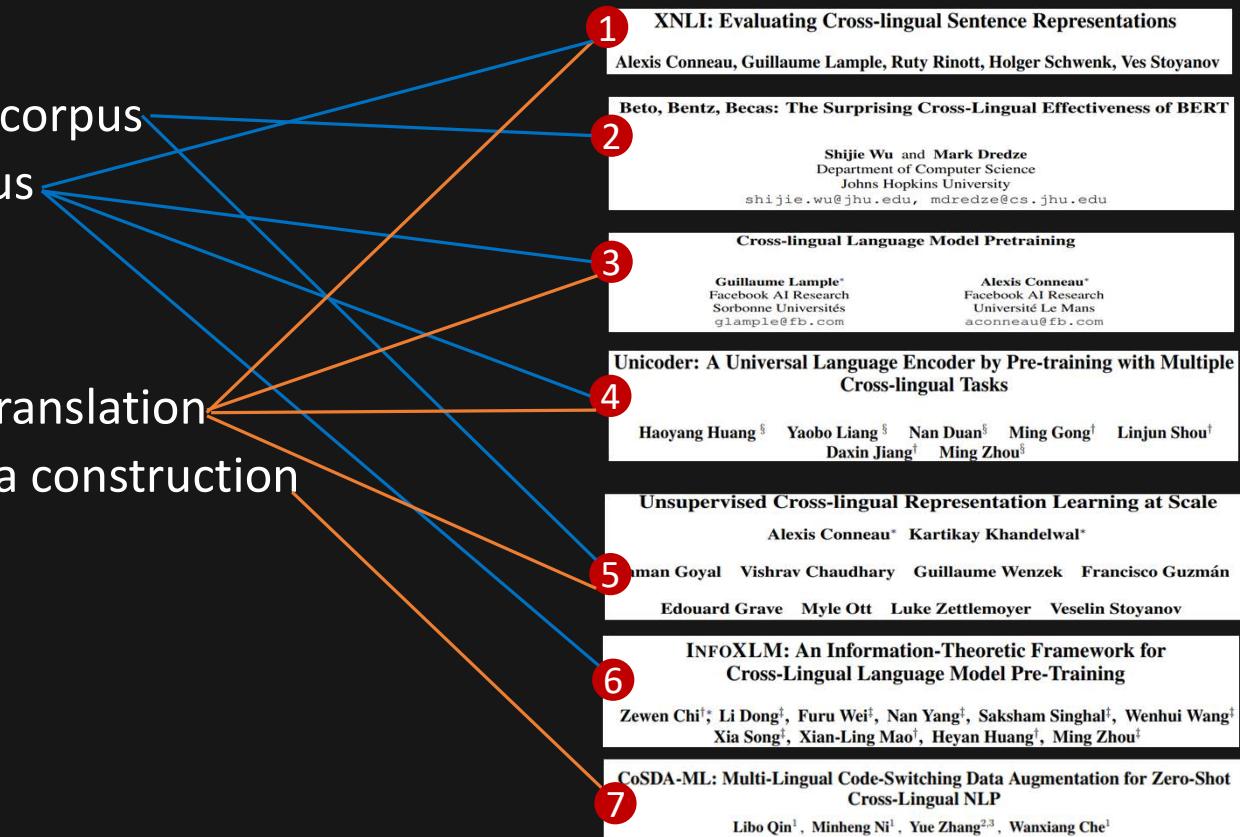
Representative Works

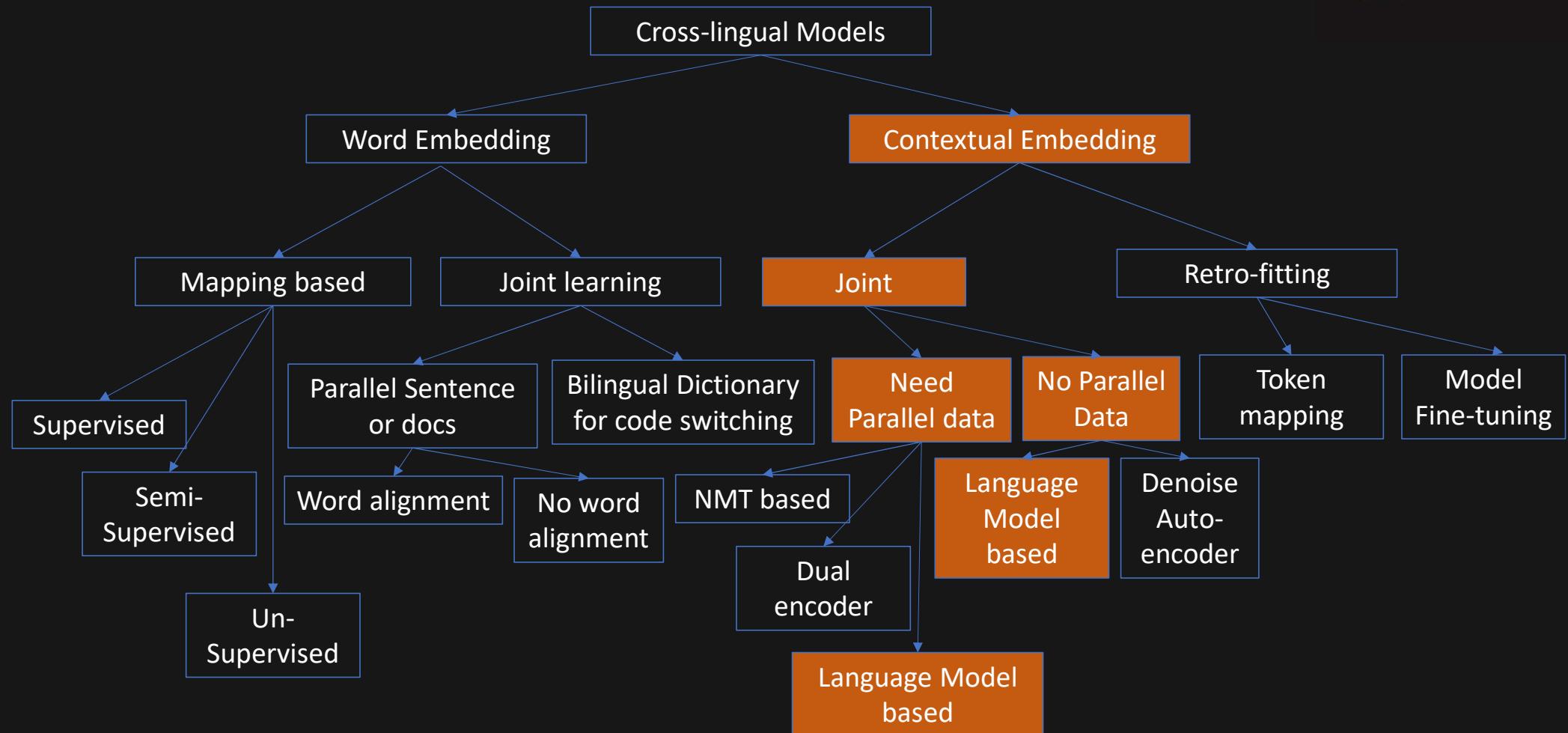
• Model Transfer

- w/ monolingual corpus
- w/ parallel corpus

• Data Transfer

- Using machine translation
- Code Switch data construction





Multi-lingual Sentence Encoders

- Use parallel corpus for sentence embedding alignment

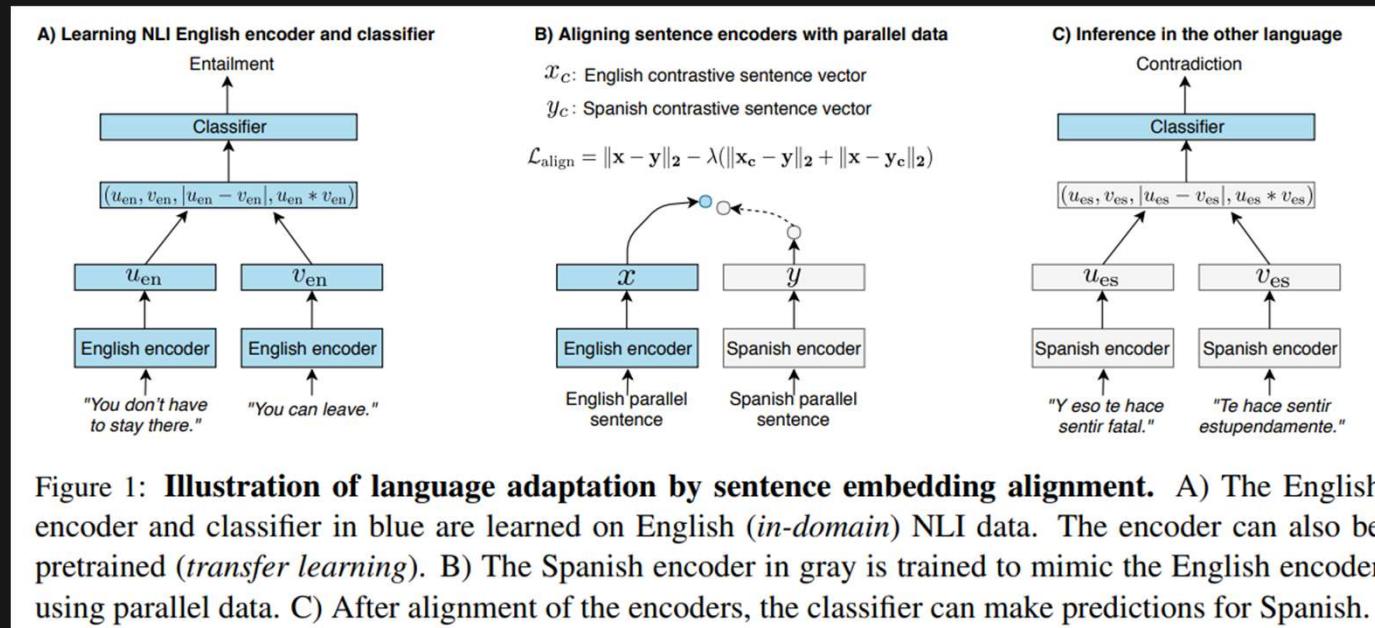


Figure 1: **Illustration of language adaptation by sentence embedding alignment.** A) The English encoder and classifier in blue are learned on English (*in-domain*) NLI data. The encoder can also be pretrained (*transfer learning*). B) The Spanish encoder in gray is trained to mimic the English encoder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

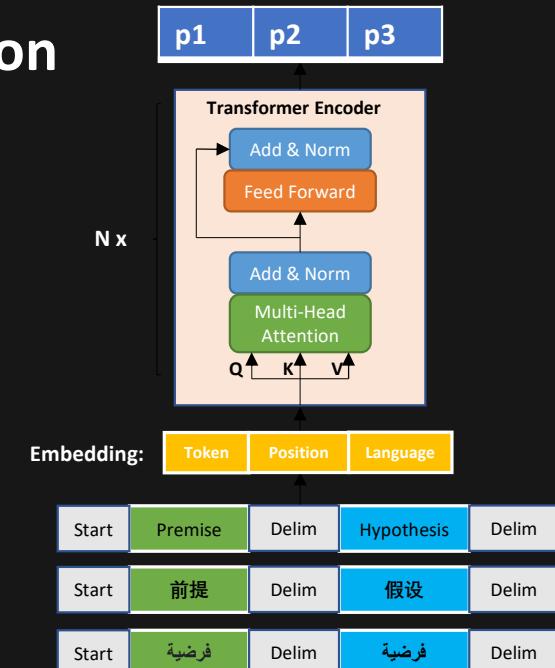
Conneau, Alexis et al. “XNLI: Evaluating Cross-lingual Sentence Representations.” EMNLP (2018).

Multi-lingual BERT – Multi-lingual Pretrained Model

Transform 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

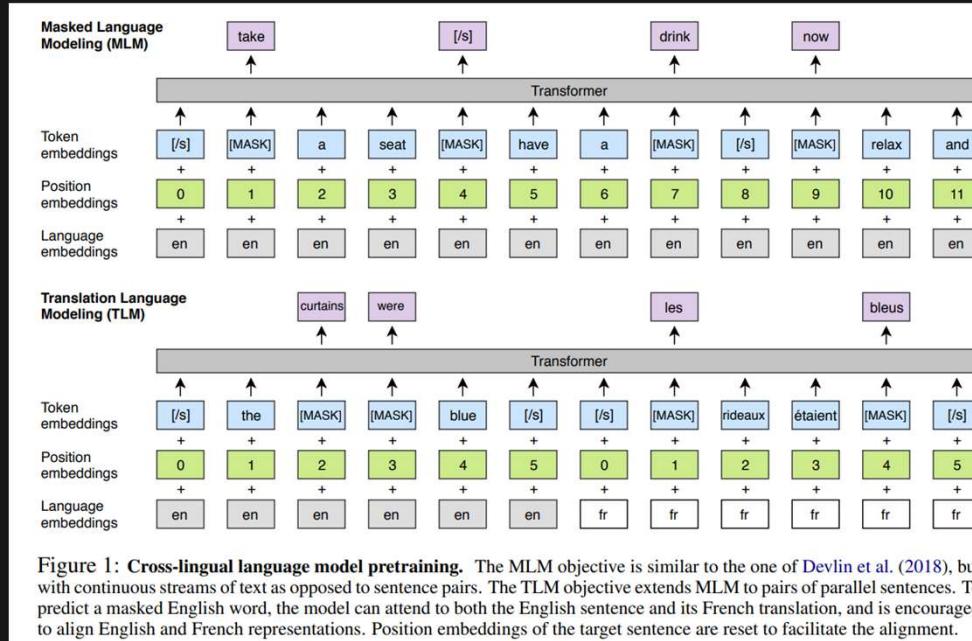
Model (zero-shot)	XNLI Avg Acc
X-LSTM	65.6
mBERT	66.3



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

XLM - Cross-lingual Language Model Pretraining

- Translation Language Model: Leverage Parallel data for alignment pretraining



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

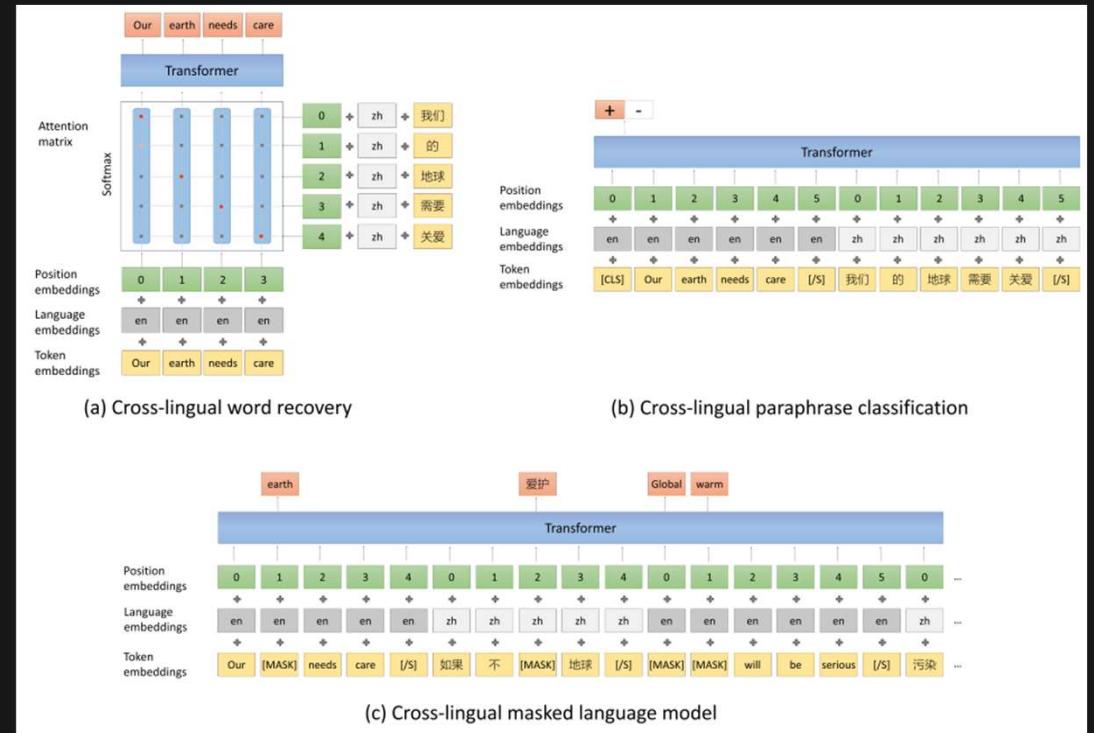
	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Δ
<i>Machine translation baselines (TRANSLATE-TRAIN)</i>																
Devlin et al. (2018)	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
XLM (MLM+TLM)	85.0	80.2	80.8	80.3	78.1	79.3	78.1	74.7	76.5	76.6	75.5	78.6	72.3	70.9	63.2	76.7
<i>Machine translation baselines (TRANSLATE-TEST)</i>																
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
<i>Evaluation of cross-lingual sentence encoders</i>																
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, Dixin 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

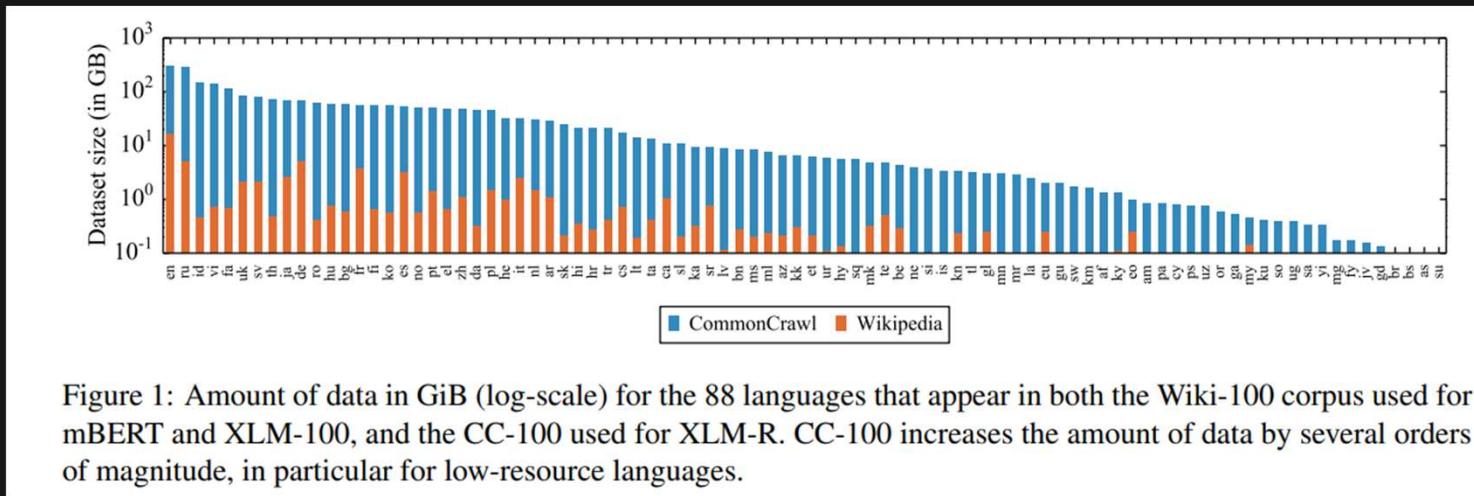
	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	average	
<i>Machine translate at training (TRANSLATE-TRAIN)</i>																	
Conneau et al. (2018)	73.7	68.3	68.8	66.5	66.4	67.4	66.5	64.5	65.8	66.0	62.8	67.0	62.1	58.2	56.6	65.4	
Multilingual BERT (Devlin et al., 2018)	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	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	
<i>Machine translate at test (TRANSLATE-TEST)</i>																67.2	
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	
<i>Evaluation of cross-lingual sentence encoders (Cross-lingual TEST)</i>																65.6	
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	
<i>Multi-language Fine-tuning</i>																77.8	
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	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, Dixin 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



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
<i>Fine-tune multilingual model on English training set (Cross-lingual Transfer)</i>																			
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_{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
<i>Translate everything to English and use English-only model (TRANSLATE-TEST)</i>																			
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
<i>Fine-tune multilingual model on each training set (TRANSLATE-TRAIN)</i>																			
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
<i>Fine-tune multilingual model on all training sets (TRANSLATE-TRAIN-ALL)</i>																			
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_{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

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

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

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
<i>Fine-tune multilingual model on English training set (Cross-lingual Transfer)</i>																	
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
<i>Fine-tune multilingual model on all training sets (Translate-Train-All)</i>																	
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

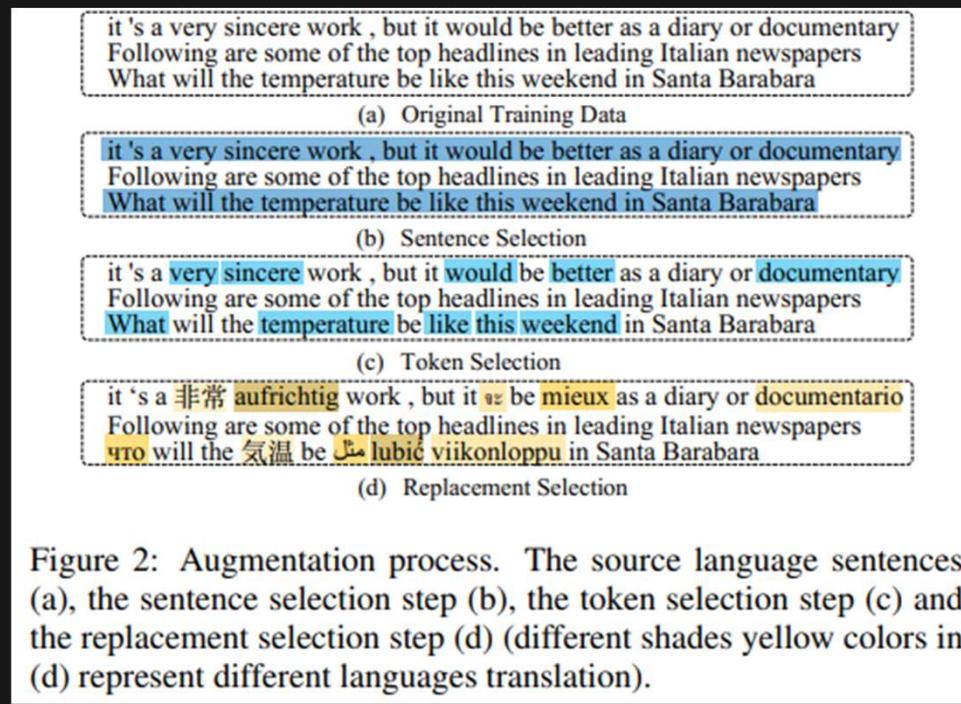


Figure 2: Augmentation process. The source language sentences (a), the sentence selection step (b), the token selection step (c) and the replacement selection step (d) (different shades yellow colors in (d) represent different languages translation).

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 (Unicoder/XLM/InfoXLM)
- Bigger model leads to better performance (XLM-R)
- Larger data leads to better performance (XLM-R)

Outline

- Introduction [Dixin 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 [Dixin Jiang]



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



Linjun Shou

Xiubo Geng

Ming Gong



Cross Lingual Semantic Retrieval

Xiubo Geng
xigeng@microsoft.com

Cross-lingual Semantic Retrieval

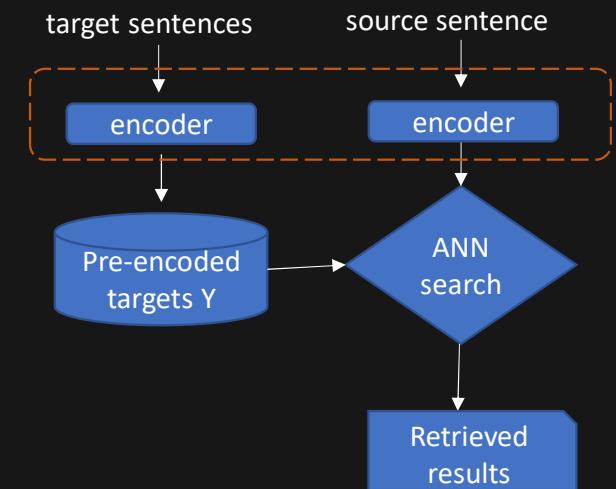
- Introduction
- Cross-lingual Models
 - Basic architecture
 - Dual encoder
 - NMT based model
 - Advanced approaches
 - Better loss function for training
 - Better scoring function for inference
- Data Augmentation
 - Using both monolingual and parallel data
 - Using only monolingual data

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

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

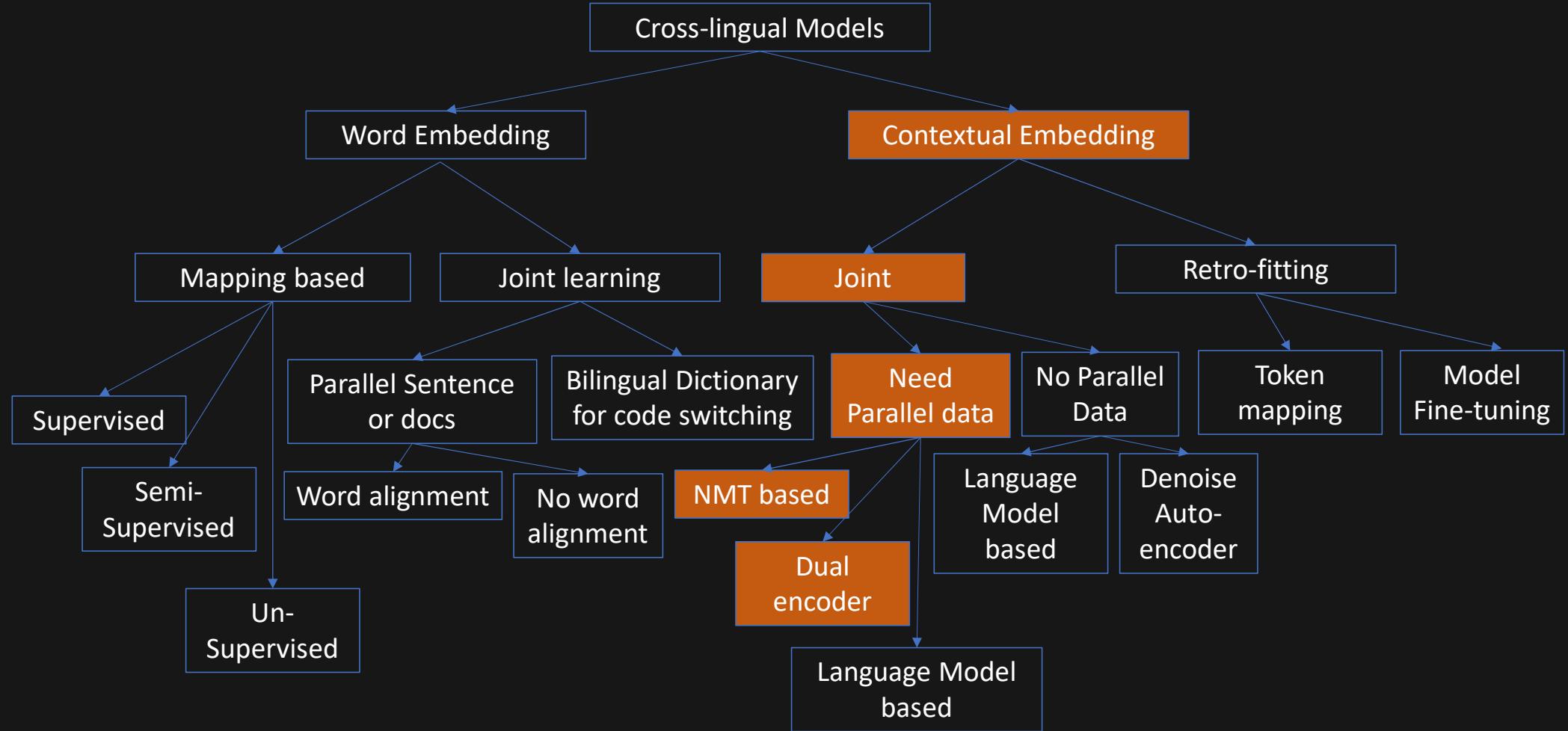


Cross-lingual Semantic Retrieval

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

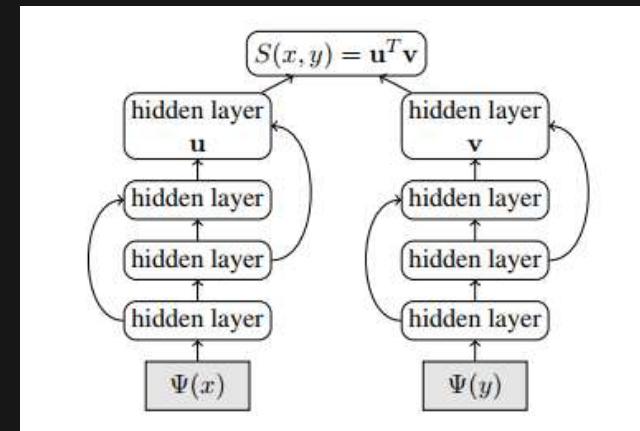
Cross-lingual Semantic Retrieval

- Introduction
- Cross-lingual Models
 - Basic 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



Dual Encoder

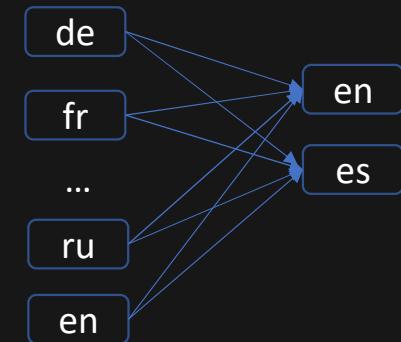
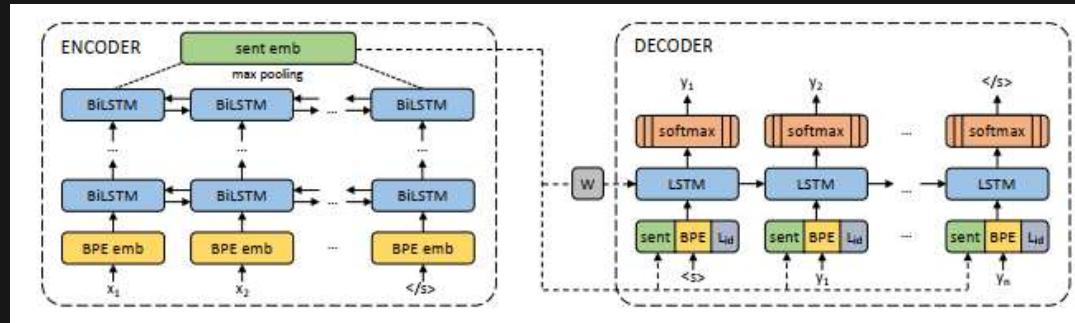
- Dual Encoder Model
 - Source sentence and target sentence are encoded separately by a dual encoder model
 - Similarity between two embeddings
- Encoder
 - Deep Averaging Networks (DAN)
 - Transformers
 - ...



Dual Encoder

- Model Training
 - Parallel sentences as training data
 - Approximated ranking probability
$$P(y|x) = \frac{e^{\phi(x,y)}}{\sum_{\bar{y}} e^{\phi(x,\bar{y})}}$$
 \longrightarrow $P_{approx}(y|x) = \frac{e^{\phi(x,y)}}{\sum_{k=1}^K e^{\phi(x,y_k)}}$
- 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

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

NMT (Encoder-Decoder) Model

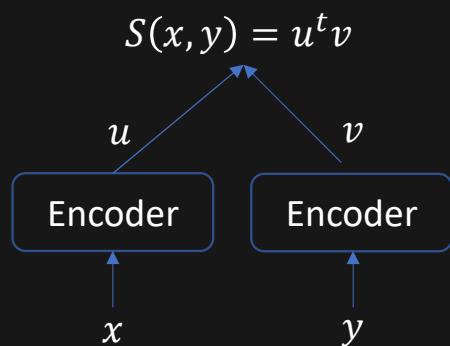
- Training Data
 - Difficult to obtain training data if each input sentence is jointly translated into all other languages
 - Using only two target languages (English and Spanish) gets similar results
 - Scaling to almost 100 languages
- Inference
 - Get language-independent representation for each sentence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

NMT (Encoder-Decoder) Model

- List of 93 source languages
- Evaluation results on cross-lingual similarity search
 - 37 languages with below 5% error rate
 - 48 languages with below 10% error rate
 - 55 languages with below 20% error rate
 - Only 15 languages with error rates above 50%

	af	am	ar	ay	az	be	ber	bg	bn	br	bs	ca	cbk	cs	da	de	
train sent.	67k	88k	8.2M	14k	254k	5k	62k	4.9M	913k	29k	4.2M	813k	1k	5.5M	7.9M	8.7M	
en→xx err.	11.20	60.71	8.30	n/a	44.10	31.20	29.80	4.50	10.80	83.50	3.95	4.00	24.20	3.10	3.90	0.90	
xx→en err.	9.90	55.36	7.80	n/a	23.90	36.50	33.70	5.40	10.00	84.90	3.11	4.20	21.70	3.80	4.00	1.00	
test sent.	1000	168	1000	—	1000	1000	1000	1000	1000	354	1000	1000	1000	1000	1000	1000	
	dtp	dv	el	en	eo	es	et	eu	fi	fr	ga	gl	ha	he	hi	hr	
train sent.	1k	90k	6.5M	2.6M	397k	4.8M	5.3M	1.2M	7.9M	8.8M	732	349k	127k	4.1M	288k	4.0M	
en→xx err.	92.10	n/a	5.30	n/a	2.70	1.90	3.20	5.70	3.70	4.40	93.80	4.60	n/a	8.10	5.80	2.80	
xx→en err.	93.50	n/a	4.80	n/a	2.80	2.10	3.40	5.00	3.70	4.30	95.80	4.40	n/a	7.60	4.80	2.70	
test sent.	1000	—	1000	—	1000	1000	1000	1000	1000	1000	1000	1000	—	1000	1000	1000	
	hu	hy	ia	id	ie	io	is	it	ja	ka	kab	kk	km	ko	ku	kw	
train sent.	5.3M	6k	9k	4.3M	3k	3k	2.0M	8.3M	3.2M	296k	15k	4k	625	1.4M	50k	2k	
en→xx err.	3.90	59.97	5.40	5.20	14.70	17.40	4.40	4.60	3.90	60.32	39.10	80.17	77.01	10.60	80.24	91.90	
xx→en err.	4.00	67.79	4.10	5.80	12.80	15.20	4.40	4.80	5.40	67.83	44.70	82.61	81.72	11.50	85.37	93.20	
test sent.	1000	742	1000	1000	1000	1000	1000	1000	1000	746	447	1000	575	722	1000	410	1000
	kzj	la	lfn	lt	lv	mg	mhr	mk	ml	mr	ms	my	nb	nds	nl	oc	
train sent.	560	19k	2k	3.2M	2.0M	355k	1k	4.2M	373k	31k	2.9M	2k	4.1M	12k	8.4M	3k	
en→xx err.	91.60	41.60	35.90	4.10	4.50	n/a	87.70	5.20	3.35	9.00	3.40	n/a	1.30	18.60	3.10	39.20	
xx→en err.	94.10	41.50	35.10	3.40	4.70	n/a	91.50	5.40	2.91	8.00	3.80	n/a	1.10	15.60	4.30	38.40	
test sent.	1000	1000	1000	1000	1000	—	1000	1000	687	1000	1000	—	1000	1000	1000	1000	
	pl	ps	pt	ro	ru	sd	si	sk	sl	so	sq	sr	sv	sw	ta	te	
train sent.	5.5M	4.9M	8.3M	4.9M	9.3M	91k	796k	5.2M	5.2M	85k	3.2M	4.0M	7.8M	173k	42k	33k	
en→xx err.	2.00	7.20	4.70	2.50	4.90	n/a	n/a	3.10	4.50	n/a	1.80	4.30	3.60	45.64	31.60	18.38	
xx→en err.	2.40	6.00	4.90	2.70	5.90	n/a	n/a	3.70	3.77	n/a	2.30	5.00	3.20	39.23	29.64	22.22	
test sent.	1000	1000	1000	1000	1000	—	1000	823	—	1000	1000	1000	390	307	234	—	
	tg	th	tl	tr	tt	ug	uk	ur	uz	vi	wuu	yue	zh				
train sent.	124k	4.1M	36k	5.7M	119k	88k	1.4M	746k	118k	4.0M	2k	4k	8.3M				
en→xx err.	n/a	4.93	47.40	2.30	72.00	59.90	5.80	20.00	82.24	3.40	25.80	37.00	4.10				
xx→en err.	n/a	4.20	51.50	2.60	65.70	49.60	5.10	16.20	80.37	3.00	25.20	38.90	5.00				
test sent.	—	548	1000	1000	1000	1000	1000	1000	428	1000	1000	1000	1000	1000	1000	1000	

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

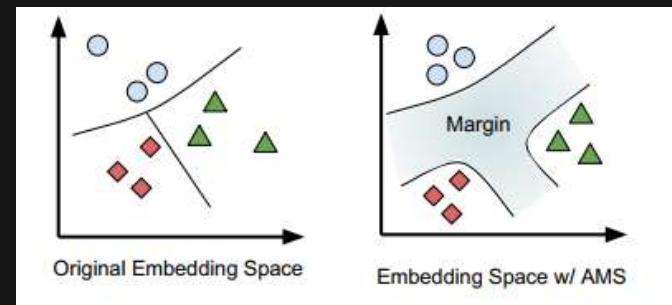
Cross-lingual Semantic Retrieval

- Introduction
- Cross-lingual Models
 - Basic 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



Better Loss Function for Training

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

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

- Additive Margin Softmax

additive margin for positive pairs

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

Better Loss Function for Training

Models	en-fr			en-es			en-ru			en-zh		
	P@1	P@3	P@10									
[Guo <i>et al.</i> , 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

Better Scoring Function for Inference

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

irrelevant { correct translation →	<p>(A) <i>Les produits agricoles sont constitués de thé, de riz, de sucre, de tabac, de camphre, de fruits et de soie.</i></p> <hr/> <p>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.</p> <hr/> <p>(B) <i>Mais dans le contexte actuel, nous pourrons les ignorer sans risque.</i></p> <hr/> <p>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.</p>
---	--

Better Scoring Function for Inference

- Margin-based Scoring

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

similarity between x and y

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

Better Scoring Function for Inference

- Variants of Margin Functions
 - Absolute $\text{margin}(a, b) = a$
 - Distance $\text{margin}(a, b) = a - b$
 - Ratio $\text{margin}(a, b) = a/b$
- Both distance and ratio functions outperforms cosine similarity by more than 10%
- Ratio function is slightly better than distance function

Func.	Retrieval	EN-DE			EN-FR		
		P	R	F1	P	R	F1
Abs.	Forward	78.9	75.1	77.0	82.1	74.2	77.9
	Backward	79.0	73.1	75.9	77.2	72.2	74.7
	(cos)	84.9	80.8	82.8	83.6	78.3	80.9
	Max. score	83.1	77.2	80.1	80.9	77.5	79.2
Dist.	Forward	94.8	94.1	94.4	91.1	91.8	91.4
	Backward	94.8	94.1	94.4	91.5	91.4	91.4
	Intersection	94.9	94.1	94.5	91.2	91.8	91.5
	Max. score	94.9	94.1	94.5	91.2	91.8	91.5
Ratio	Forward	95.2	94.4	94.8	92.4	91.3	91.8
	Backward	95.2	94.4	94.8	92.3	91.3	91.8
	Intersection	95.3	94.4	94.8	92.4	91.3	91.9
	Max. score	95.3	94.4	94.8	92.4	91.3	91.9

Comparison on BUCC dataset

Cross-lingual Semantic Retrieval

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

Motivation

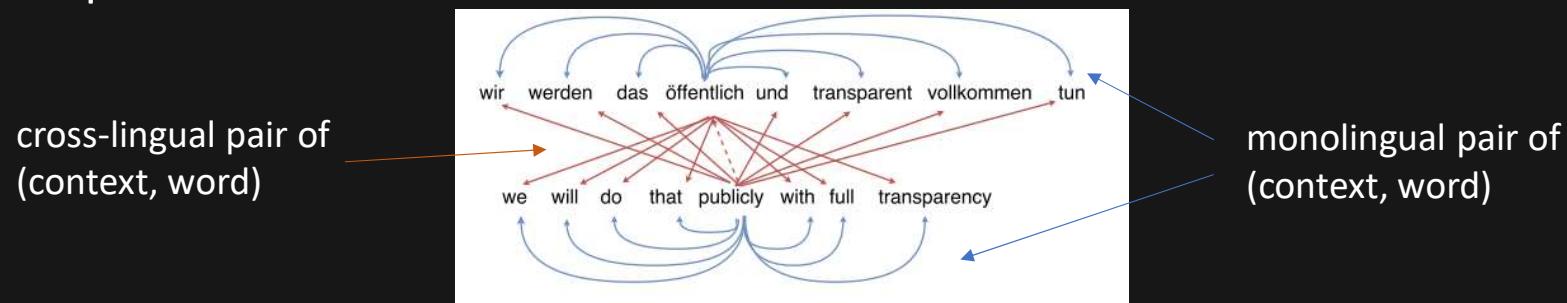
- Parallel data is not easy to get, especially for low-resource languages
- Large volume of monolingual data
- The goal is to generate multilingual sentence embeddings with limited, even no parallel data

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

Using Both Monolingual and Parallel Data

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



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

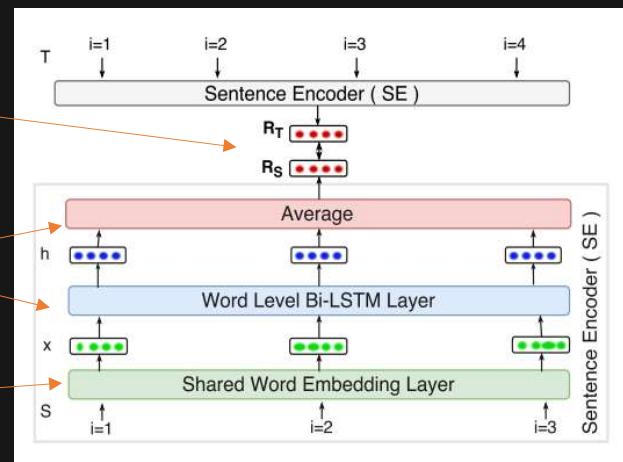
Using Both Monolingual and Parallel Data

- Dual Encoder for Sentence Similarity

Similarity between sentence pairs for model training

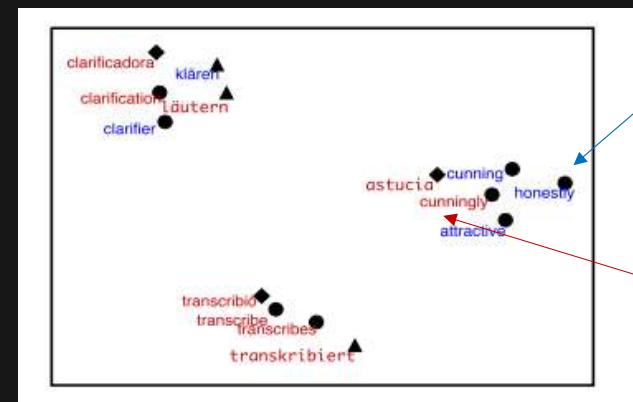
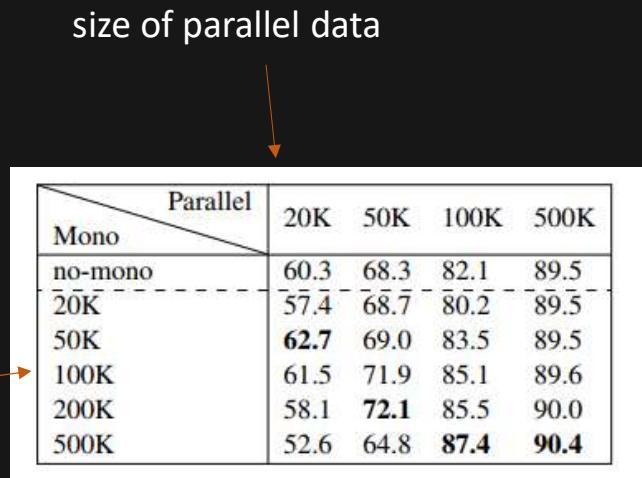
BiLSTM and average pooling to get sentence embedding based on word embedding

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

Using Both Monolingual and Parallel Data



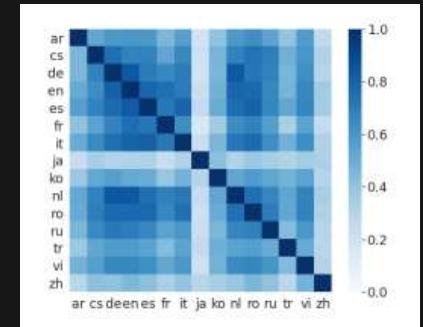
Multi-task learning produces consistently better embeddings when the amount of monolingual data is neither too large nor too small

Cross-lingual Semantic Retrieval

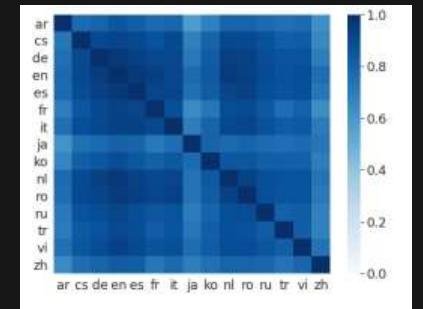
- Introduction
- Cross-lingual Models
 - Basic models
 - Advanced approaches
- Data Augmentation
 - Using both monolingual and parallel data
 - Using only monolingual data

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



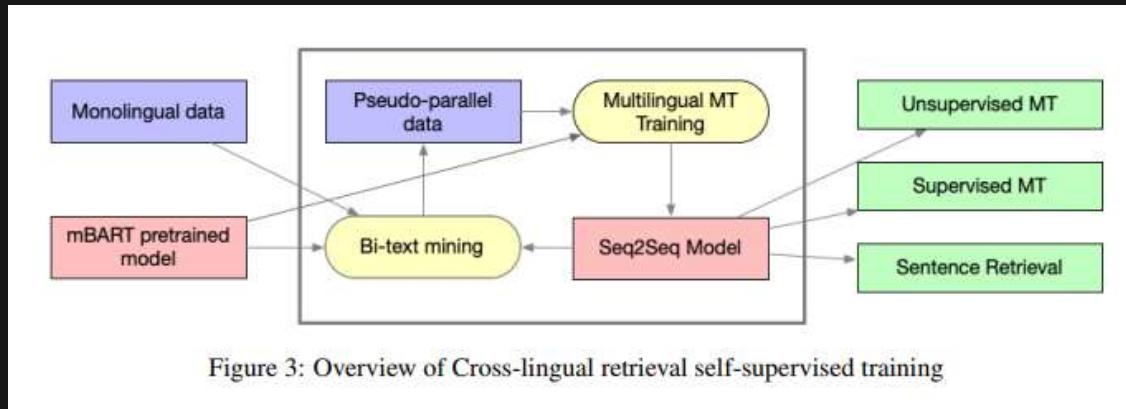
Sentence retrieval using mBART encoder



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



- 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

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

- Introduction [Dixin 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 [Dixin Jiang]



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



Linjun Shou



Xiubo Geng



Ming Gong

Cross Lingual Sequence Labelling

- Machine Reading Comprehension (MRC) as Example

Ming Gong

migon@microsoft.com

Outline

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

Cross-lingual Sequence Labelling

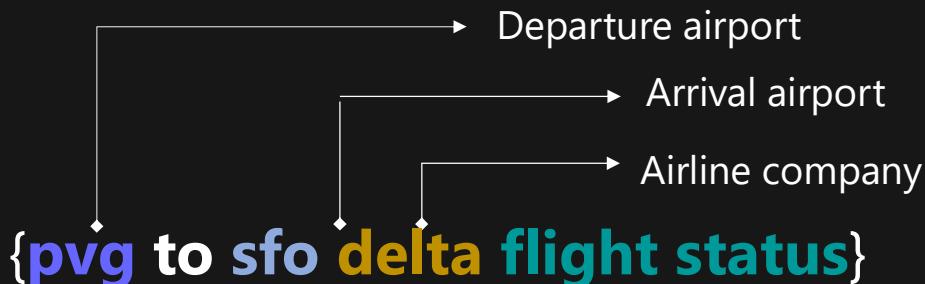
Type	Category	Sub Category	Example
NLU	Text Classification	Single text	Domain identification, Intent detection, Sentiment classification
		Text pair	Information retrieval, Natural language inference
	Sequence Labeling	Single text	Named entity recognition, Slot tagging
		Text pair	Extractive machine reading 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

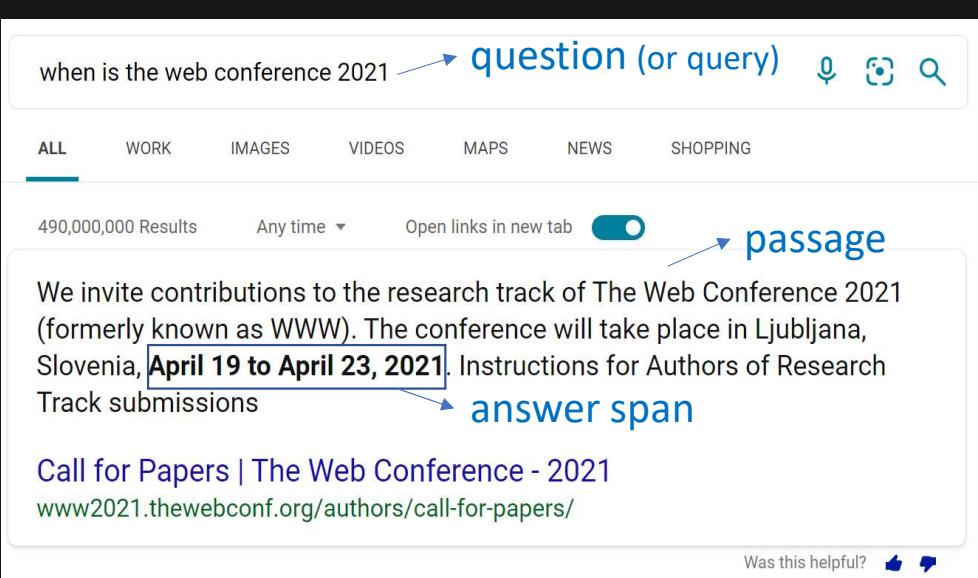
Luke Rawlence PERSON joined Aiimi ORG as a data scientist in Milton Keynes PLACE, after finishing his computer science degree at the University of Lincoln. ORG

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.



when is the web conference 2021

question (or query) 🔍 🔎 🔍

ALL WORK IMAGES VIDEOS MAPS NEWS SHOPPING

490,000,000 Results Any time Open links in new tab

We invite contributions to the research track of The Web Conference 2021 (formerly known as WWW). The conference will take place in Ljubljana, Slovenia, April 19 to April 23, 2021. Instructions for Authors of Research Track submissions

answer span

Call for Papers | The Web Conference - 2021
www2021.thewebconf.org/authors/call-for-papers/

Was this helpful? ⤵ ⤵

Application3: Machine reading comprehension (MRC)

Cross-lingual MRC (CLMRC): Definition & Dataset

- Benchmark Datasets

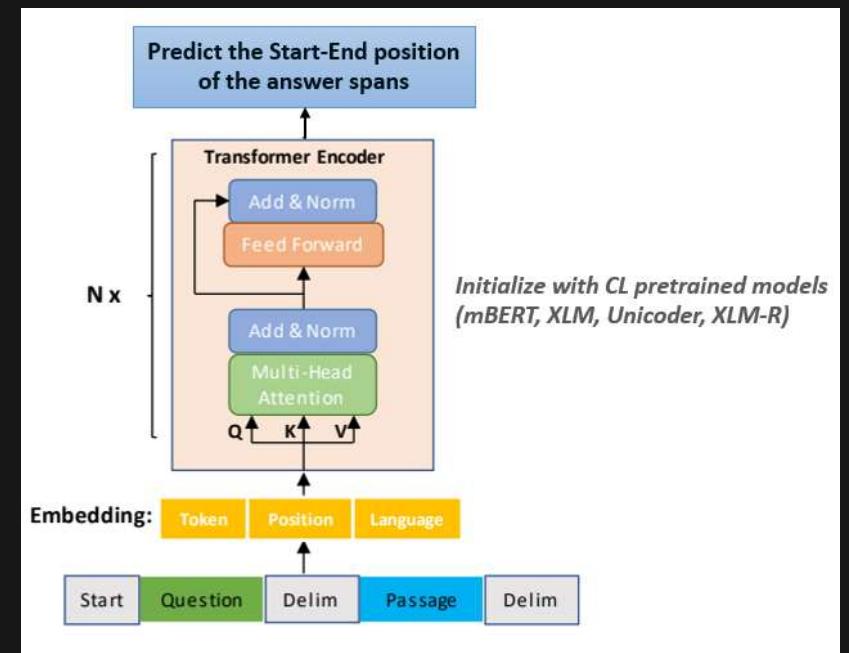
- MLQA*: Patrak et al. MLQA: Evaluating Cross-lingual Extractive Question Answering. ACL 2020.
- TydiQA*: Clark, J. et al. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. TACL 2020.
- XQuAD*: Artetxe, M. et al. On the Cross-lingual Transferability of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

En	During what time period did the Angles migrate to Great Britain?	En	What are the names given to the campuses on the east side of the land the university sits on?
	The name "England" is derived from the Old English name Englaland [...] The Angles were one of the Germanic tribes that settled in Great Britain during the Early Middle Ages . [...] The Welsh name for the English language is "Saesneg"		The campus is in the residential area of Westwood [...] The campus is informally divided into North Campus and South Campus , which are both on the eastern half of the university's land. [...] The campus includes [...] a mix of architectural styles.
De	Während welcher Zeitperiode migrierten die Angeln nach Großbritannien?	Es	Cuáles son los nombres dados a los campus ubicados en el lado este del recinto donde se encuentra la universidad?
	Der Name England leitet sich vom altenglischen Wort Englaland [...] Die Angeln waren ein germanischer Stamm, der das Land im Frühmittelalter besiedelte. [...] ein Verweis auf die weißen Klippen von Dover.		El campus incluye [...] una mezcla de estilos arquitectónicos. Informalmente está dividido en Campus Norte y Campus Sur , ambos localizados en la parte este del terreno que posee la universidad. [...] El Campus Sur está enfocado en la ciencias físicas [...] y el Centro Médico Ronald Reagan de UCLA.
Ar	في أي حقبة زمنية هاجر الأنجل إلى بريطانيا العظمى؟	Zh	位于大学占地东半部的校园名称是什么？
	والتي تعني "رض الأجل". والأجل كانت واحدة، England يشتق اسم "الجبل" من الكلمة الإنجليزية القديمة من القبائل герمانية التي استقرت في إنجلترا خلال غزو العصور الوسطى . [...] وقد سماها العرب قديماً الإنكلز.		整个校园被不正式地分为 南北两个校园 ，这两个校园都位于大学占地的东半部。北校园是原校园的中心，建筑以意大利文艺复兴时代建筑闻名，其中的包威尔图书馆（Powell Library）成为好莱坞电影的最佳拍摄场景。 [...] 这个广场曾在许多电影中出现。
Vi	Trong khoảng thời gian nào người Angles di cư đến Anh?	Hi	विश्वविद्यालय जहाँ स्थित है, उसके पूर्वी दिशा में बने परिसरों को क्या नाम दिया गया है?
	Tên gọi của Anh trong tiếng Việt bắt nguồn từ tiếng Trung. [...] Người Angle là một trong những bộ tộc German định cư tại Anh trong Thời đại Trung Cổ . [...] dường như nó liên quan tới phong tục gọi người German tại Anh là Angli Saxones hay Anh - Sachsen.		जब 1919 में यूसीएलए ने अपना नया परिसर खोला, तब इसमें घार इमारतें थी। [...] परिसर अनोन्यारिक रूप से उत्तरी परिसर और दक्षिणी परिसर में विभाजित है, जो दोनों विश्वविद्यालय की जमीन के पूर्वी हिस्से में स्थित है। [...] उद्दिष्टी परिसर में भौतिक विज्ञान, जीव विज्ञान, इंजीनियरिंग, मनोविज्ञान, गणितीय विज्ञान, सांस्कृतिक स्वास्थ्य से संबंधित क्षेत्र और यूरोपीय मैडिकल सेंटर स्थित हैं।

Examples from MLQA

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*

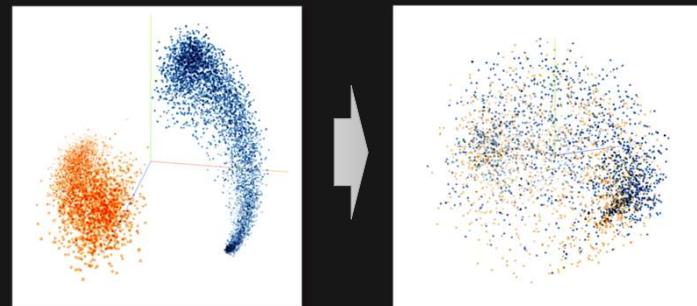


Challenges for CLMRC (1/4)

- 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.

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

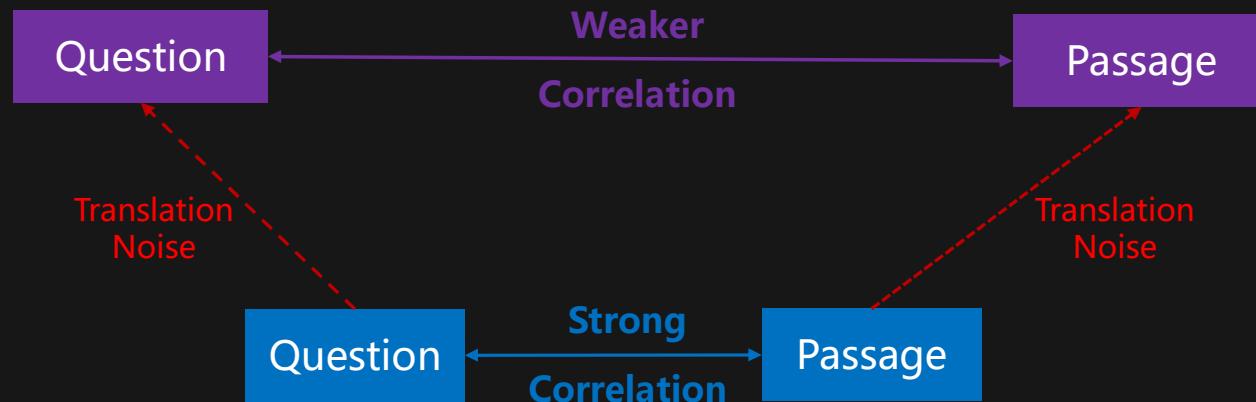
Table 1: The gap between target languages and English on Machine Reading Comprehension (MRC) (Lewis et al., 2019) is significantly larger than sentence level classification task like Natural Language Inference (NLI) (Conneau et al., 2018). In this experiment, we fine-tune XLM (Conneau and Lample, 2019)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



*For phrase boundary task, CL pretraining models **don't have sufficient transfer capability**
→ better semantic alignment is required*

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:
**befindet sich zwischen der
Sonne und dem Mond**

*Miss alignment with the
correct span in DE passage*

Challenges for CLMRC (4/4)

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

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

	#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.

Advanced Approaches for CLMRC

Challenges of the Baseline Approach

A

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

B

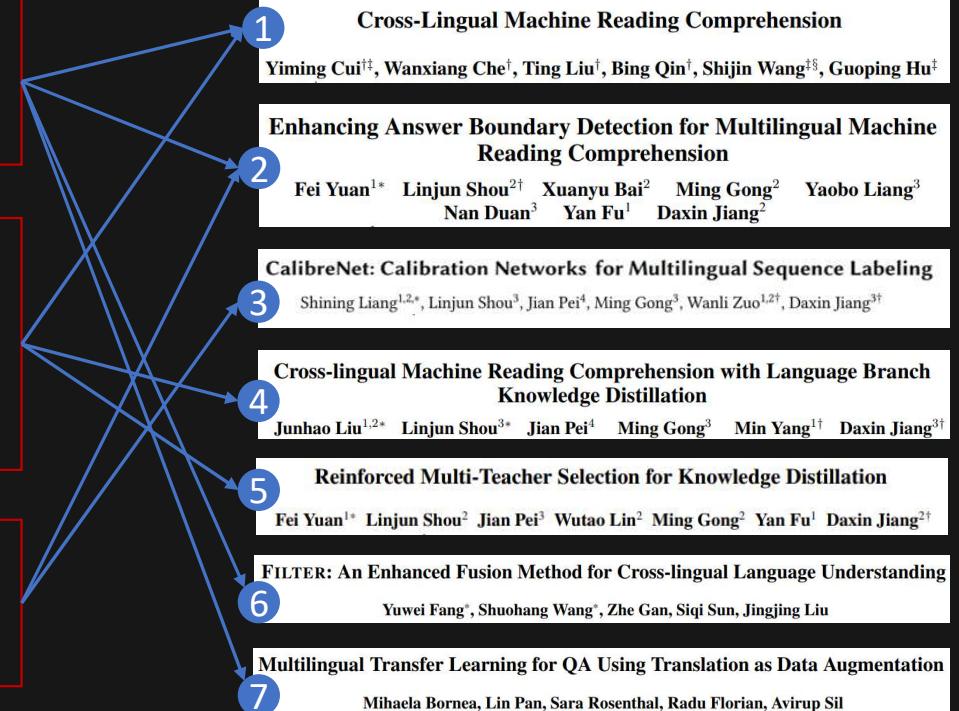
Machine Translation Data Quality

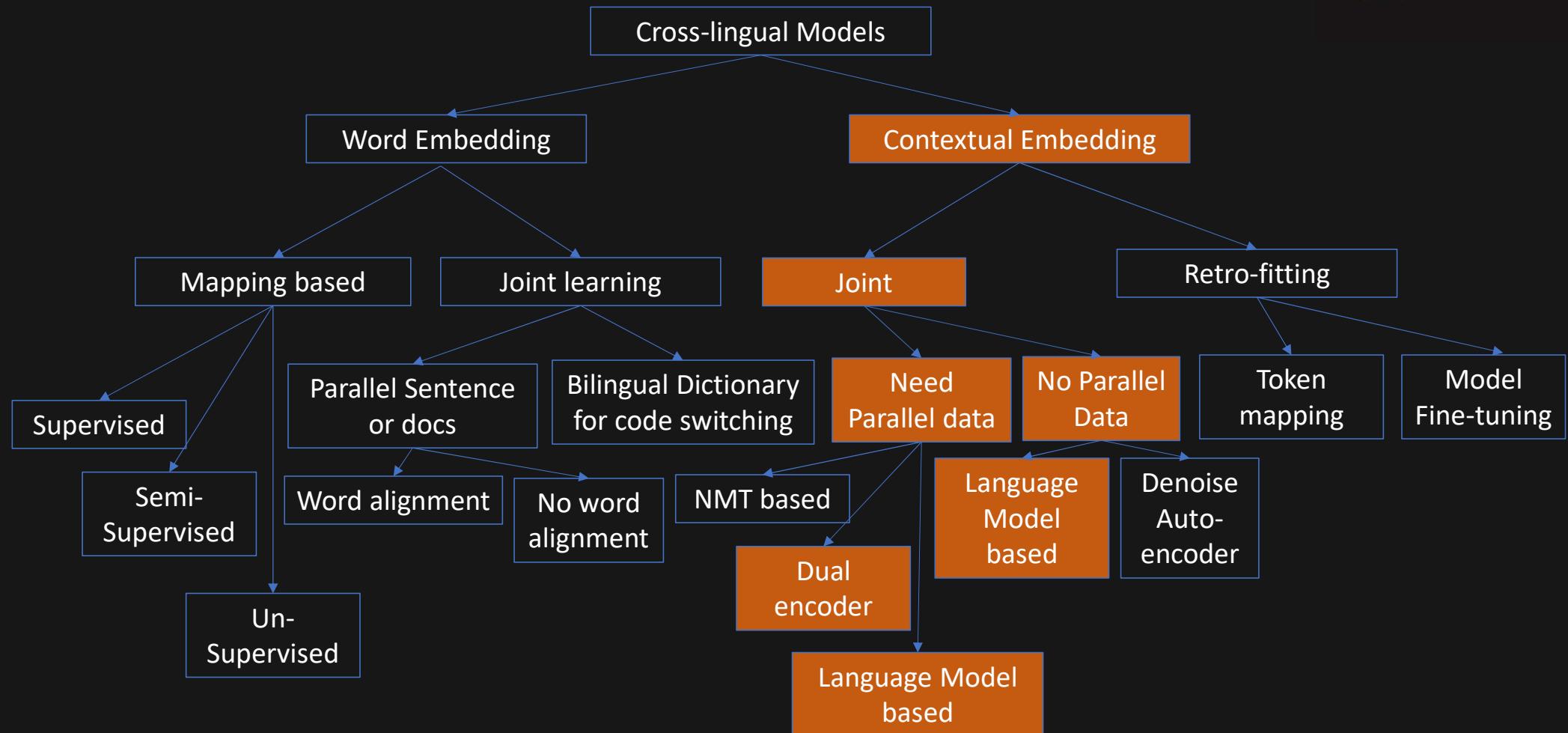
- MT noises impact more for pair-wise tasks
- Answer span alignment is challenging after MT

C

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

Representative Works





Advanced Approaches for CLMRC

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 prediction

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 Bai² Ming Gong² Yaobo Liang³
Nan Duan³ Yan Fu¹ Dixin Jiang²

3

CalibreNet: Calibration Networks for Multilingual Sequence Labeling

Shining Liang^{1,2,*}, Linjun Shou³, Jian Pei⁴, Ming Gong³, Wanli Zuo^{1,2†}, Dixin Jiang^{3†}

4

Cross-lingual Machine Reading Comprehension with Language Branch Knowledge Distillation

Junhao Liu^{1,2*} Linjun Shou^{3*} Jian Pei⁴ Ming Gong³ Min Yang^{1†} Dixin Jiang^{3†}

5

Reinforced Multi-Teacher Selection for Knowledge Distillation

Fei Yuan^{1*} Linjun Shou² Jian Pei³ Wutao Lin² Ming Gong² Yan Fu¹ Dixin 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

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?

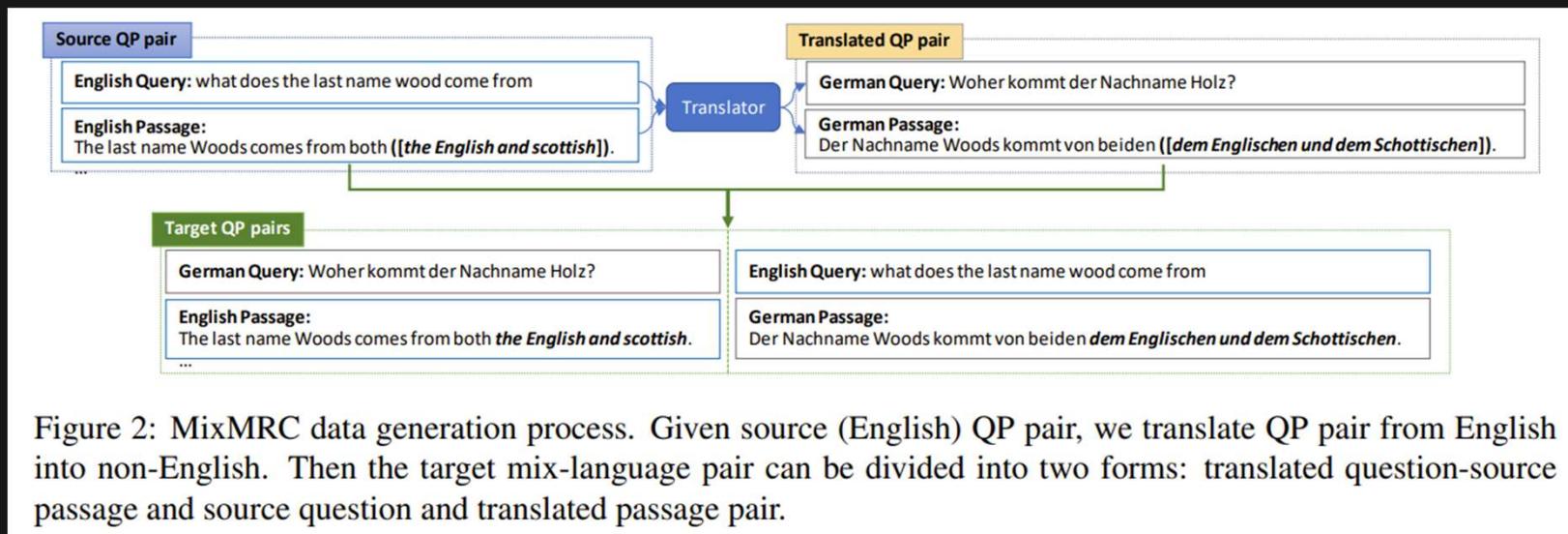


Figure 2: MixMRC data generation process. Given source (English) QP pair, we translate QP pair from English into non-English. Then the target mix-language pair can be divided into two forms: translated question-source passage and source question and translated passage pair.

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

Mixed MRC Task with MT Data for Semantic Transfer

- Translation Train Results

Model	Methods	MLQA (EM / F1)			MTQA (EM / F1)		
		en	es	de	en	fr	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, Dixin Jiang. Enhancing Answer Boundary Detection for Multilingual Machine Reading Comprehension. ACL, 2020.

Dual Language Encoders for Semantic Transfer - DualBERT

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

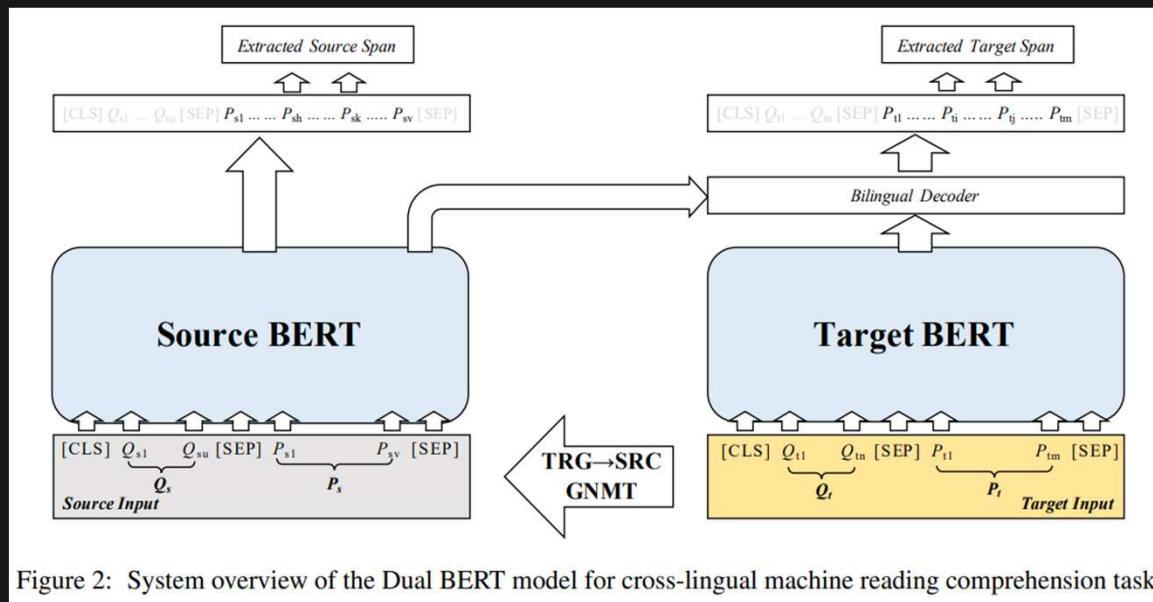


Figure 2: System overview of the Dual BERT model for cross-lingual machine reading comprehension task.

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

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

Dual Language Encoders for Semantic Transfer - FILTER

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

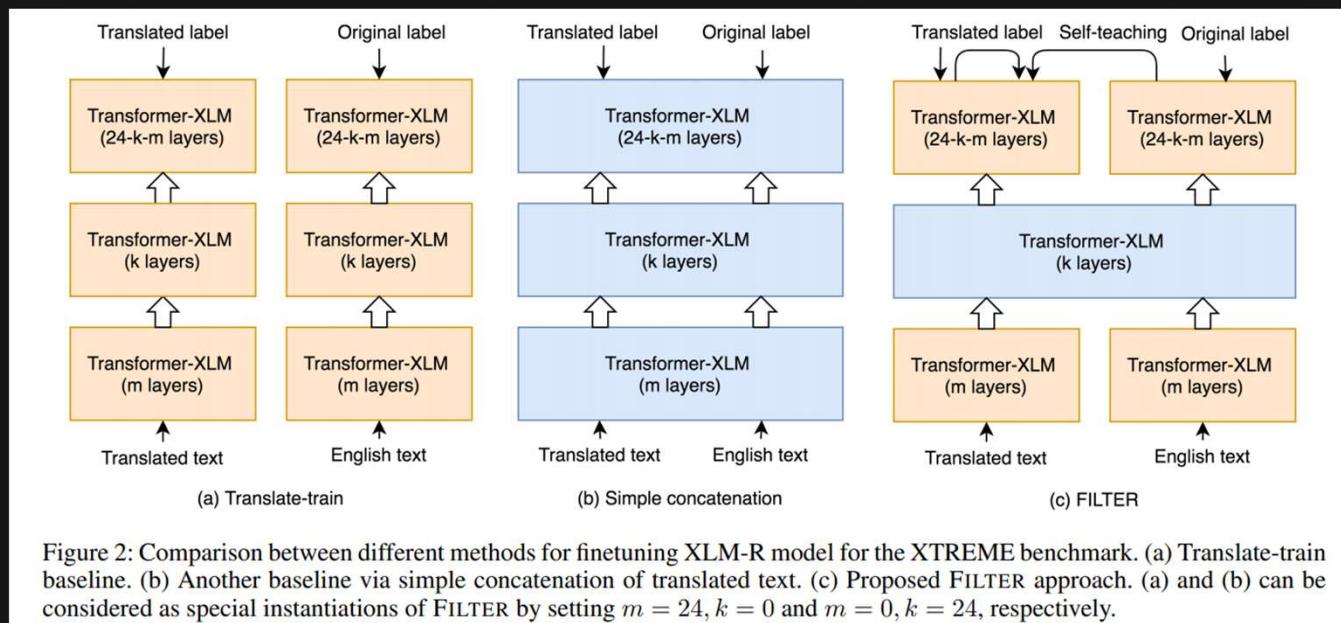


Figure 2: Comparison between different methods for finetuning XLM-R model for the XTERME benchmark. (a) Translate-train baseline. (b) Another baseline via simple concatenation of translated text. (c) Proposed FILTER approach. (a) and (b) can be considered as special instantiations of FILTER by setting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

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

Dual Language Encoders for Semantic Transfer - FILTER

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

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

Multi-lingual Adversarial Training

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

$$\begin{aligned}\mathcal{L}_{QA} + \mathcal{L}_{adv} &= \mathcal{L}_{QA} - \sum_{l=1}^L KL[U(\mathbf{g}_l) || \boxed{\log \mathbf{p}_l}] \\ \mathcal{L}_D &= - \sum_{l=1}^L \mathbb{1}(\mathbf{g}_l) \underline{\log \mathbf{p}_l} \quad (\text{Discriminator})\end{aligned}$$

Adversarial Loss

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

Multi-lingual Adversarial Training

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

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

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

Advanced Approaches for CLMRC

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 prediction

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 Bai² Ming Gong² Yaobo Liang³
Nan Duan³ Yan Fu¹ Dixin Jiang²

3 CalibreNet: Calibration Networks for Multilingual Sequence Labeling

Shining Liang^{1,2,*}, Linjun Shou³, Jian Pei⁴, Ming Gong³, Wanli Zuo^{1,2†}, Dixin Jiang^{3†}

4 Cross-lingual Machine Reading Comprehension with Language Branch Knowledge Distillation

Junhao Liu^{1,2*} Linjun Shou^{3*} Jian Pei⁴ Ming Gong³ Min Yang^{1†} Dixin Jiang^{3†}

5 Reinforced Multi-Teacher Selection for Knowledge Distillation

Fei Yuan^{1*} Linjun Shou² Jian Pei³ Wutao Lin² Ming Gong² Yan Fu¹ Dixin 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

Improve Span Alignment for MT: Back-translation + Verification

- **Motivation:** the answer spans after translation may not exist in translated passage or have boundary errors.
- Translate source span to target span and further improve boundary by:
 1. *Simple Match*
 2. *Answer Aligner*
 3. *Answer Verifier*

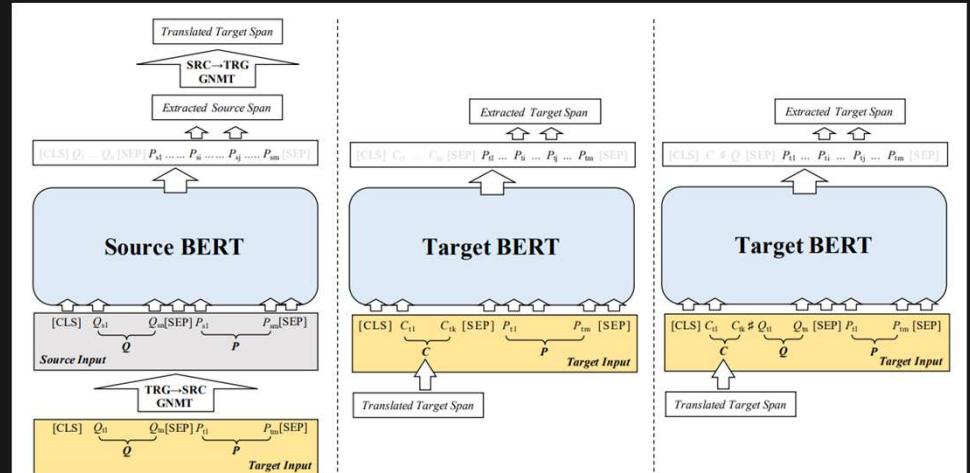


Figure 1: Back-translation approaches for cross-lingual machine reading comprehension (Left: GNMT, Middle: Answer Aligner, Right: Answer Verifier)

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

Improve Span Alignment for MT: Back-translation + Verification

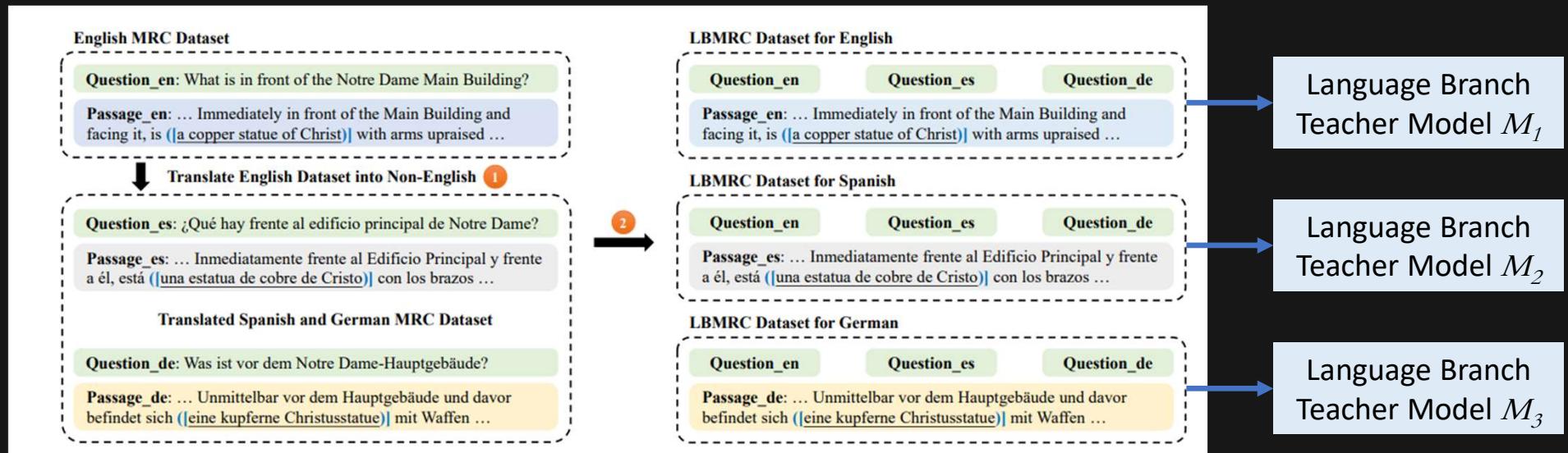
Experiment Results using different answer span alignment approaches

#	System	CMRC 2018						DRCD			
		Dev		Test		Challenge		Dev		Test	
		EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Without any alignment approaches	1 GNMT+BERT _{SQ-B_{en}}	15.9	40.3	20.8	45.4	4.2	20.2	28.1	50.0	26.6	48.9
	2 GNMT+BERT _{SQ-L_{en}}	16.8	42.1	21.7	47.3	5.2	22.0	28.9	52.0	28.7	52.1
	3 GNMT+BERT _{SQ-L_{en}} +SimpleMatch	26.7	56.9	31.3	61.6	9.1	35.5	36.9	60.6	37.0	61.2
	4 GNMT+BERT _{SQ-L_{en}} +Aligner	46.1	66.4	49.8	69.3	16.5	40.9	60.1	70.5	59.5	70.7
	5 GNMT+BERT _{SQ-L_{en}} +Verifier	64.7	84.7	68.9	86.8	20.0	45.6	83.5	90.1	82.6	89.6

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

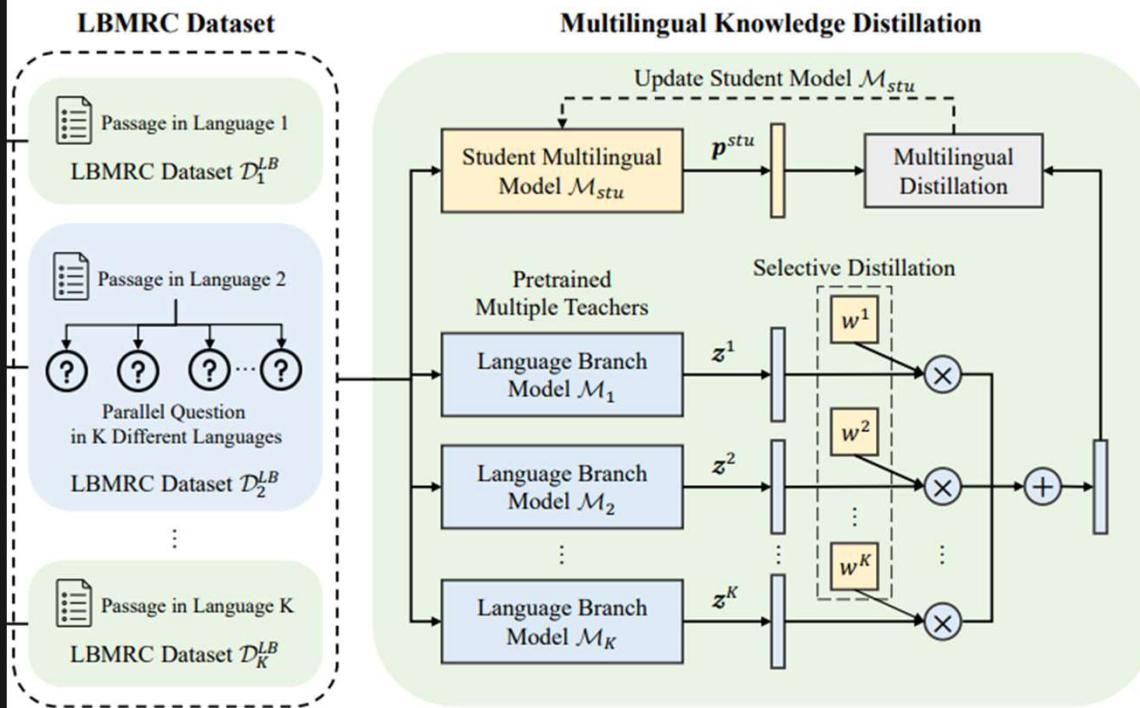
LBMRC: Language Branch Knowledge Distillation for Reducing MT Noise

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



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

LBMRC: Language Branch Knowledge Distillation for Reducing MT Noise



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

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

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

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

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

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

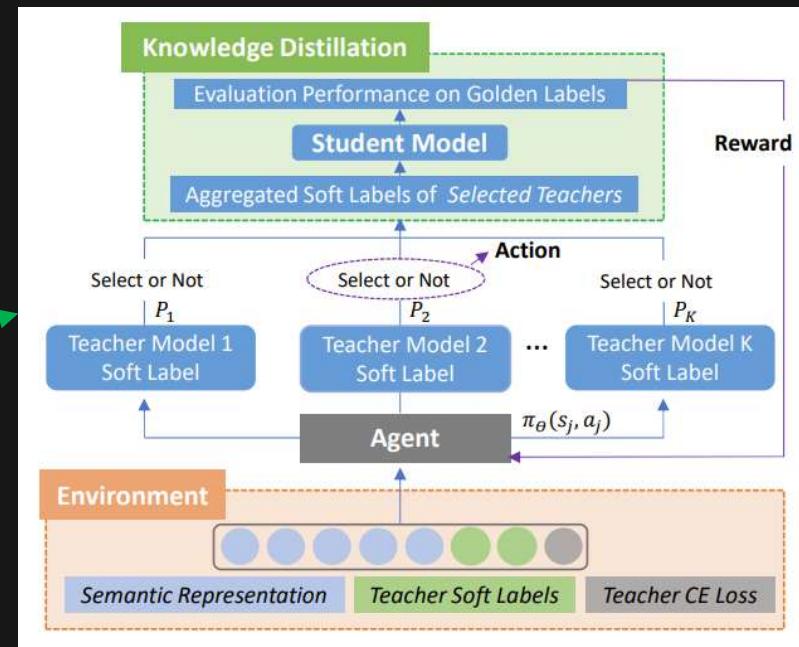
- **Motivation:** we have multiple teachers, how to select best teacher model for each case to generate soft label for better KD?

Baseline Methods:

- Vanilla KD from teacher ensemble with static weights

New Approach:

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



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

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

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

Baseline Methods:

- Vanilla KD from teacher ensemble with static weights

New Approach:

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

Teacher	Student	Strategy	QQP	MRPC	MNLI-(mm/m)	RTE	QNLI	SST-2	AVG.
-	BERT ₃	FT	88.3	72.1	74.6 / 74.8	60.7	83.3	85.9	77.1
BERT ₁₂	BERT ₃	V-KD	88.4	74.0	75.3 / 75.6	56.7	83.7	87.5	77.3
Robert ₁₂	BERT ₃	V-KD	85.3	73.0	71.9 / 71.7	55.2	81.9	86.0	75.0
XLNet ₁₂	BERT ₃	V-KD	85.4	74.3	72.0 / 71.6	55.2	81.4	87.4	75.3
ALBERT ₁₂	BERT ₃	V-KD	88.4	74.0	76.4 / 75.6	56.7	83.7	87.5	77.5
Rand-Single-Ensemble	BERT ₃	V-KD	87.3	70.6	75.0 / 74.5	51.6	83.7	86.0	75.5
W-Ensemble	BERT ₃	V-KD	85.3	74.8	72.2 / 72.0	56.7	82.4	87.5	75.8
LR-Dev-Ensemble	BERT ₃	V-KD	88.6	71.8	75.4 / 75.4	54.9	84.2	86.8	76.7
Best-Single-Ensemble	BERT ₃	V-KD	88.6	77.2	75.1 / 74.7	56.0	84.1	87.0	77.5
Our Method (<i>reward</i> ₁)	BERT ₃	RL-KD	89.1	76.0	76.9 / 76.8	61.4	85.4	89.1	79.2
Our Method (<i>reward</i> ₂)	BERT ₃	RL-KD	89.1	76.2	77.4 / 76.3	63.5	84.8	88.5	79.4
Our Method (<i>reward</i> ₃)	BERT ₃	RL-KD	89.0	76.7	76.7 / 75.7	64.6	85.3	89.1	79.6

(1)*reward*₁: use the minus of ground-truth loss (CE) of student model as the reward function for teacher model selection.

(2)*reward*₂: besides (1), also introduce the minus of knowledge distillation loss (DL) into the reward function.

(3)*reward*₃: besides (2), also take the Accuracy metric of student model on dev set into account for better generalization

Advanced Approaches for CLMRC

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 prediction

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 Bai² Ming Gong² Yaobo Liang³
Nan Duan³ Yan Fu¹ Dixin Jiang²

3 CalibreNet: Calibration Networks for Multilingual Sequence Labeling

Shining Liang^{1,2*}, Linjun Shou³, Jian Pei⁴, Ming Gong³, Wanli Zuo^{1,2†}, Dixin Jiang^{3†}

4 Cross-lingual Machine Reading Comprehension with Language Branch Knowledge Distillation

Junhao Liu^{1,2*} Linjun Shou^{3*} Jian Pei⁴ Ming Gong³ Min Yang^{1†} Dixin Jiang^{3†}

5 Reinforced Multi-Teacher Selection for Knowledge Distillation

Fei Yuan^{1*} Linjun Shou² Jian Pei³ Wutao Lin² Ming Gong² Yan Fu¹ Dixin 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

New Tasks for CL Pretrained Models Fine-tuning

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

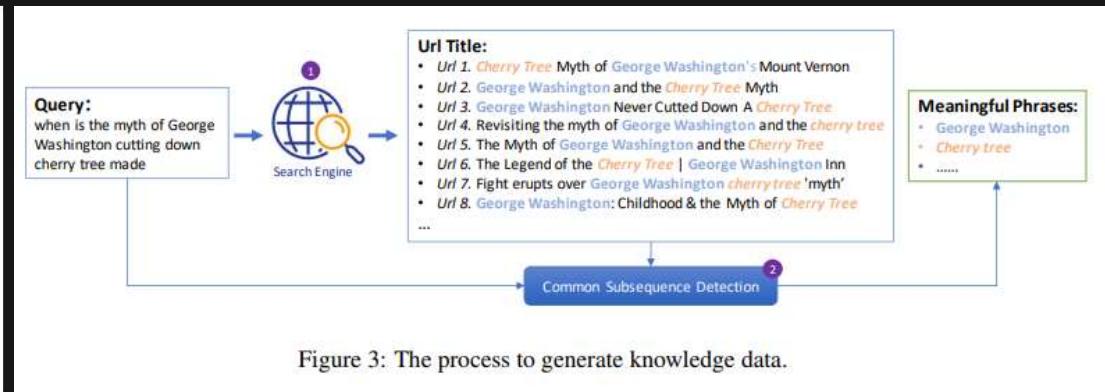
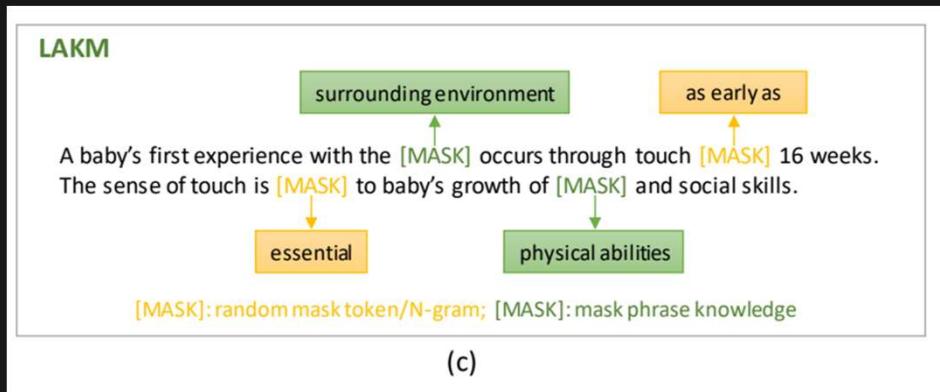


Figure 3: The process to generate knowledge data.

- As an auxiliary task of the main MRC task in the fine-tuning stage (i.e. multi-task training)
- Phrase Candidates Generation (for recall)
- Phrase Filtering (for precision)

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

New Tasks for CL Pretrained Models Fine-tuning

- Experiment Results

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

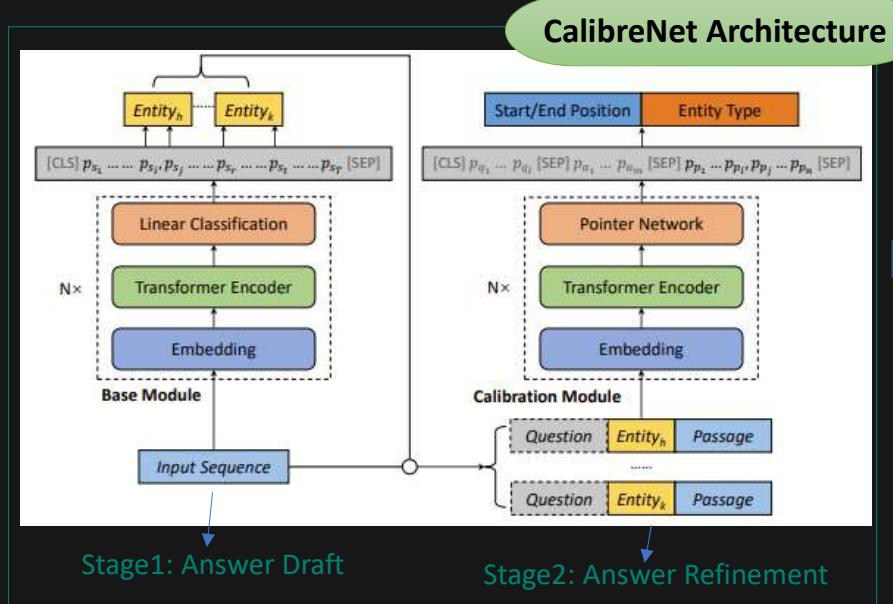
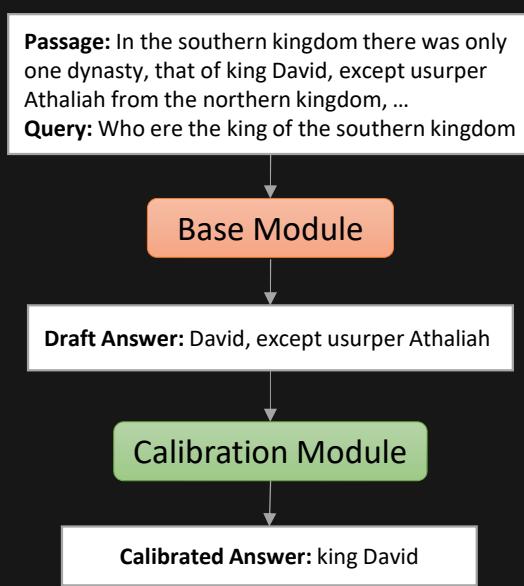
		MTQA (EM / F1)		
		en	fr	de
Baseline		65.8 / 86.6	41.3 / 70.9	50.7 / 76.2
LAKM		67.8 / 87.2	44.6 / 72.1	54.5 / 77.8

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

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

Post-processing for Boundary Refinement: CalibreNet

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

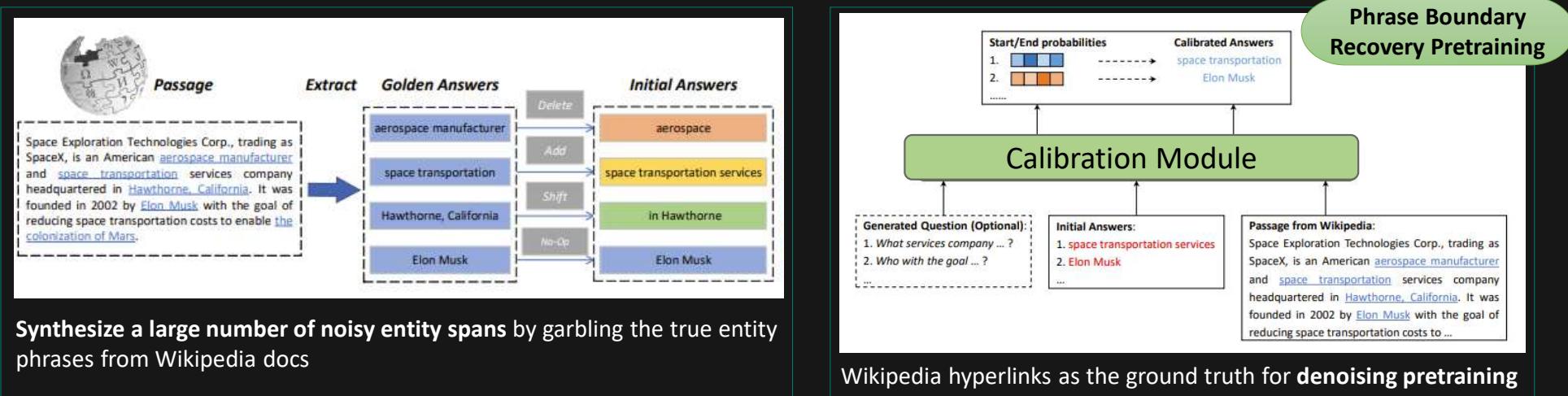


Challenge: how to collect training data for Calibration module, especially non-EN languages?

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

Post-processing for Boundary Refinement: CalibreNet

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



Results on MLQA
(F1/EM)

→

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

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

Key Take-aways

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

A

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

- New Mix-MRC task using MT data ②
- Dual encoders using MT data ① ⑥
- Adversarial training using MT data ⑦

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 ①
- Leverage language branch MT data to train multiple teacher models to reduce noise ④
- Leverage reinforcement learning to select best teacher per instance to reduce noise ⑤

C

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

- New LAKM task to infuse language specific boundary prior knowledge ②
- CalibraNet to add a new module to refine primary answer span ③

More Thinking & Ongoing Work

- Beyond CL pretrained model plus MT data as augmentation, there are more ongoing work to further address CLMRC challenges:
 - Large scale language specific unlabeled data + semi-supervised learning (i.e. real question passage pairs + predicted answer spans by self-learning)
 - Data Generation (i.e. seq2seq model to synthesize <question, passage, answer span>)

Outline

- Introduction [Dixin 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]



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



Linjun Shou



Xiubo Geng



Ming Gong

- ➡ • Future directions [Dixin Jiang]

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

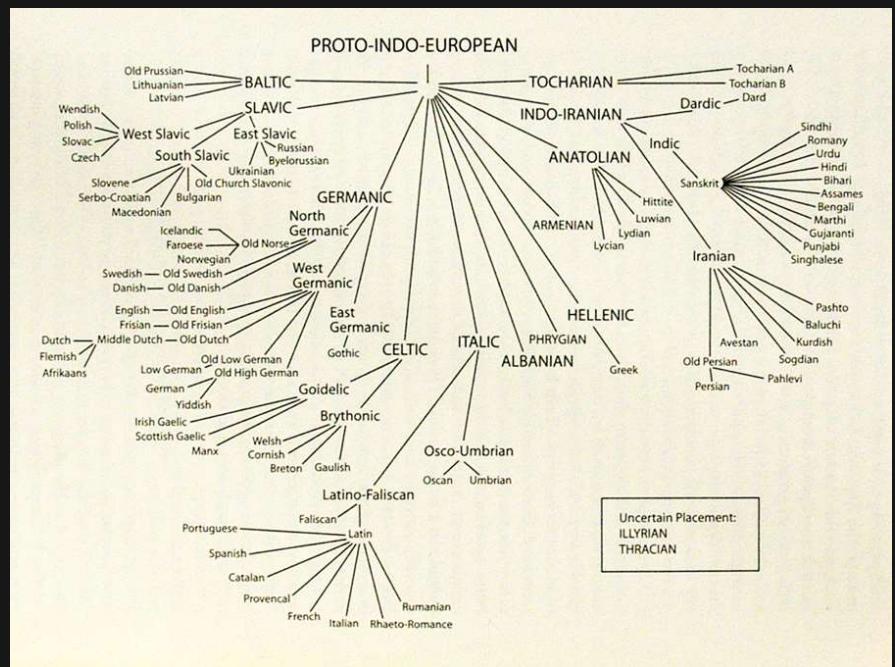
PHOIBLE: Steven Moran, Daniel McCloy, and Richard Wright. 2014. PHOIBLE Online. Max Planck Institute for Evolutionary Anthropology, Leipzig

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, Jena¹³²

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

