بسم الله الرحمن الرحیم

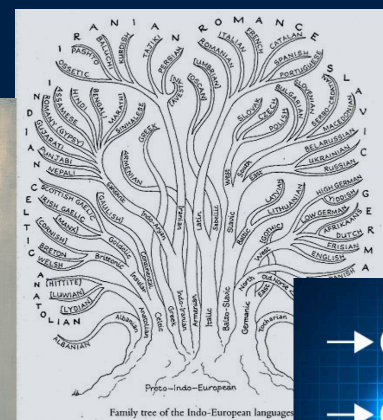
Multilingual Models & Data Processing (I)

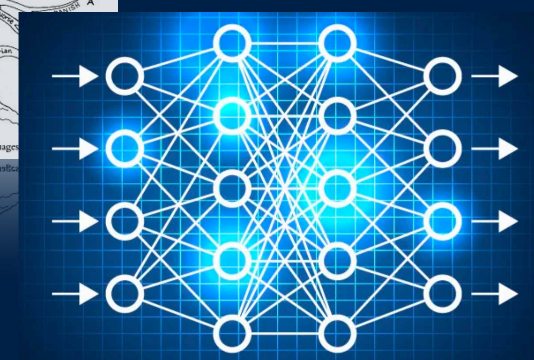
Ehsaneddin Asgari

Nov. 5th 2023

Artificial Intelligence Group
Computer Engineering Department, SUT

Why do we need to go beyond English?





Some important aspects of multilingual NLP?

Perspective	Among possible points ..
Fair Information Access	Language determines access to information and technologies
Linguistic values	Interesting typological features in resource-poor languages
Machine Learning	ML challenges in structure modeling, few-shot learning, inter-language transfer, etc.
Cultural values	Cultural legacies, values of specific countries or language communities

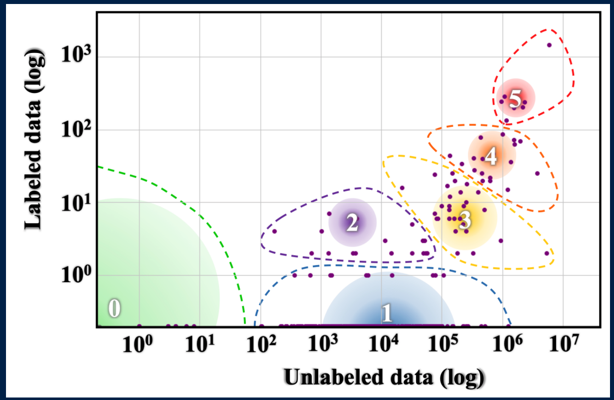
Languages in the world

There are around 7,000 languages in the world:

- Around 400 languages have more than 1M speakers.
- Around 1,200 languages have more than 100k speakers.
- Africa > 2000 languages & Indonesia 700 languages

Daan van Esch, Tamar Lucassen, Sebastian Ruder, Isaac Caswell, and Clara Rivera.
Writing System and Speaker Metadata for 2,800+ Language Varieties.
In Proceedings of the Thirteenth LREC. 2022.

Taxonomy of Languages (i)

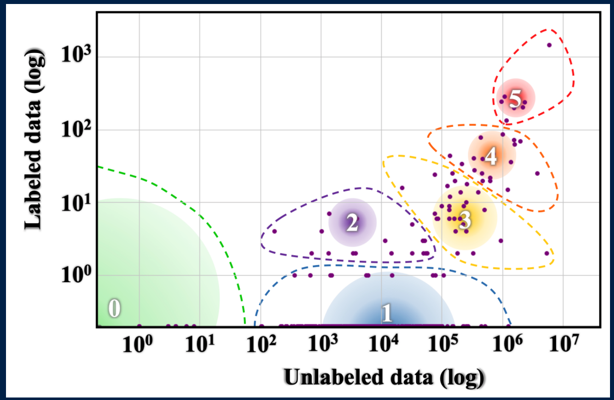


> LDC catalog and the ELRA Map for labeled datasets
> # of Wikipedia pages for unlabeled data resources

Class	Definition	Example Languages	#Langs	#Speakers	% of Total Langs
0 - The Left-Behinds	Ignored in language tech, limited resources, virtually no unlabeled data, digital upliftment unlikely	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1 - The Scraping-Bys	Some unlabeled data, potential improvement with organized effort, need for awareness and labeled dataset collection	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury.
The State and Fate of Linguistic Diversity and Inclusion in the NLP World.
In Proceedings of the 58th ACL, 2020.

Taxonomy of Languages (ii)

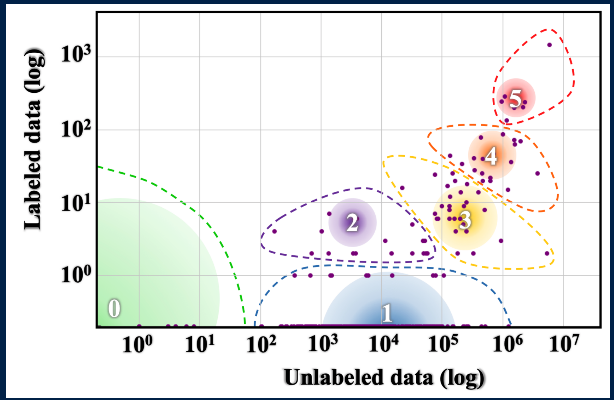


> LDC catalog and the ELRA Map for labeled datasets
> # of Wikipedia pages for unlabeled data resources

Class	Definition	Example Languages	#Langs	#Speakers	% of Total Langs
2 - The Hopefuls	Small labeled datasets, active research and support communities, promising future with more NLP tools	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3 - The Rising Stars	Benefited from unsupervised pre-training, strong web presence, cultural community online, need for labeled data collection	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury.
The State and Fate of Linguistic Diversity and Inclusion in the NLP World.
In Proceedings of the 58th ACL, 2020.

Taxonomy of Languages (iii)

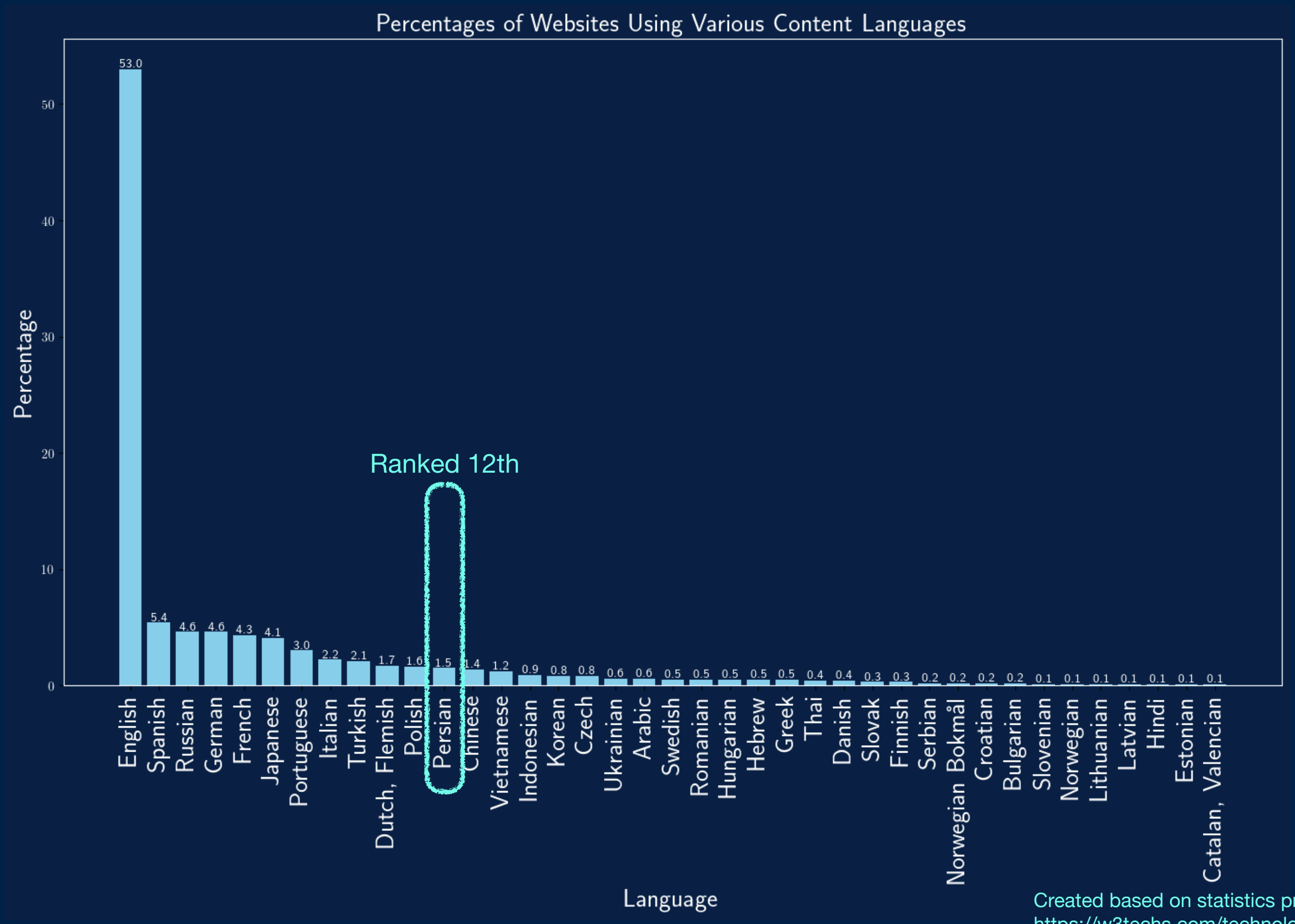


> LDC catalog and the ELRA Map for labeled datasets
> # of Wikipedia pages for unlabeled data resources

Class	Definition	Example Languages	#Langs	#Speakers	% of Total Langs
4 - The Underdogs	Large unlabeled data, strong resource firepower, active NLP research, potential to reach digital superiority	**Persian,** Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5 - The Winners	Dominant online presence, extensive industrial and government investment, rich resources and technologies	English, Spanish, German, Japanese, French	7	2.5B	0.28%

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury.
The State and Fate of Linguistic Diversity and Inclusion in the NLP World.
In Proceedings of the 58th ACL, 2020.

Web Content Language Distribution



Created based on statistics provided by w3techs on 3 November 2023
https://w3techs.com/technologies/overview/content_language

Language diversity of Iran

- Persian & dialects 58%
- Azeri & dialects 26%
- Kurdish 9%
- Luri 2%
- Balochi 1%
- Arabic 1%
- others 3%

https://www.internetworldstats.com

List of 66 languages extracted from
https://www.ethnologue.com/country/IR/

Iranian Languages in Progress @SUT

آذری

Marzia Nouri, Mahsa Amani, Reihaneh Zohrabi and Ehsaneddin Asgari
The Language Model, Resources, and Computational Pipelines for the Under-Resourced Iranian Azerbaijani
To be Appeared in AACL 2023.

Reihaneh Zohrabi, Mostafa Masumi, Omid Ghahroodi, Parham AbedAzad, Hamid Beigy, Mohammad Hossein Rohban and Ehsaneddin Asgari
Borderless Azerbaijani Processing: Linguistic Resources and a Transformer-based Approach for Azerbaijani Transliteration
To be Appeared in AACL 2023.

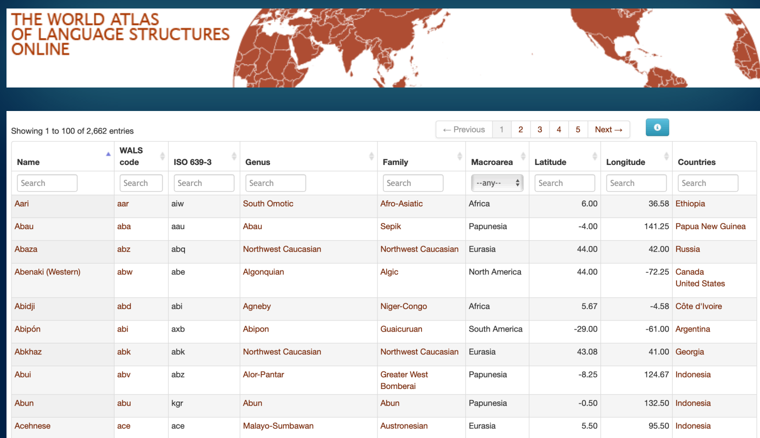
کردی

Borderless Kurdi Processing
To be submitted.

لری

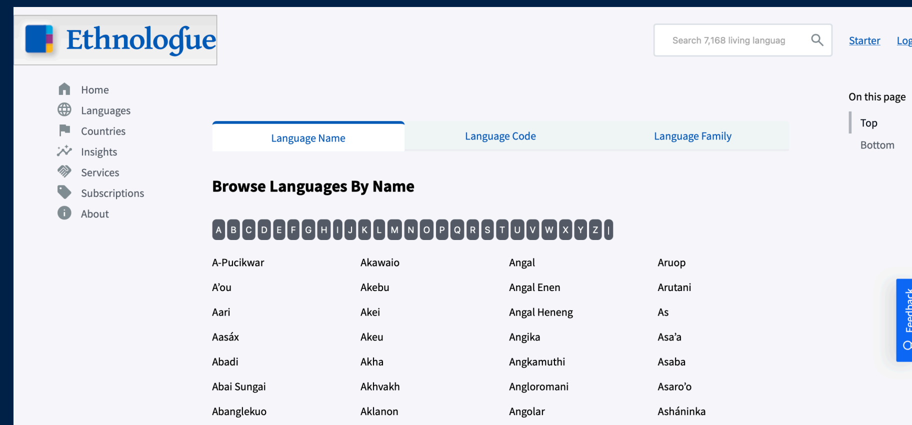
Luri Language Processing
In progress work.

Multilingual Model for Iranian Languages
In progress work.



Name	WALS code	ISO 639-3	Genus	Family	Macroarea	Latitude	Longitude	Countries
Aari	aar	aiw	South Omotic	Afro-Asiatic	Africa	6.00	36.58	Ethiopia
Abau	aba	aau	Abau	Sepik	Papunesia	-4.00	141.25	Papua New Guinea
Abaza	abz	abq	Northwest Caucasian	Northwest Caucasian	Eurasia	44.00	42.00	Russia
Abenaki (Western)	abw	abe	Algonquian	Algic	North America	44.00	-72.25	Canada United States
Abidji	abd	abi	Agneby	Niger-Congo	Africa	5.67	-4.58	Côte d'Ivoire
Abipón	abi	axb	Abipon	Guaicuruan	South America	-29.00	-61.00	Argentina
Abkhaz	abk	abk	Northwest Caucasian	Northwest Caucasian	Eurasia	43.08	41.00	Georgia
Abui	abv	abz	Alor-Pantar	Greater West Bomberai	Papunesia	-8.25	124.67	Indonesia
Abun	abu	kgr	Abun	Abun	Papunesia	-0.50	132.50	Indonesia
Acehnese	ace	ace	Malayo-Sumbawan	Austronesian	Eurasia	5.50	95.50	Indonesia

https://wals.info



https://www.ethnologue.com

https://languages.parsi.ai

- Introduction to

Multilingual Langauge Processing

Multilingual NLP

- Machine Translation
- Annotation Projection
- Collaborative Effort
- Multilingual LM

Machine Translation

- Encoder-decoder architectures
- Not relevant for low-resource languages

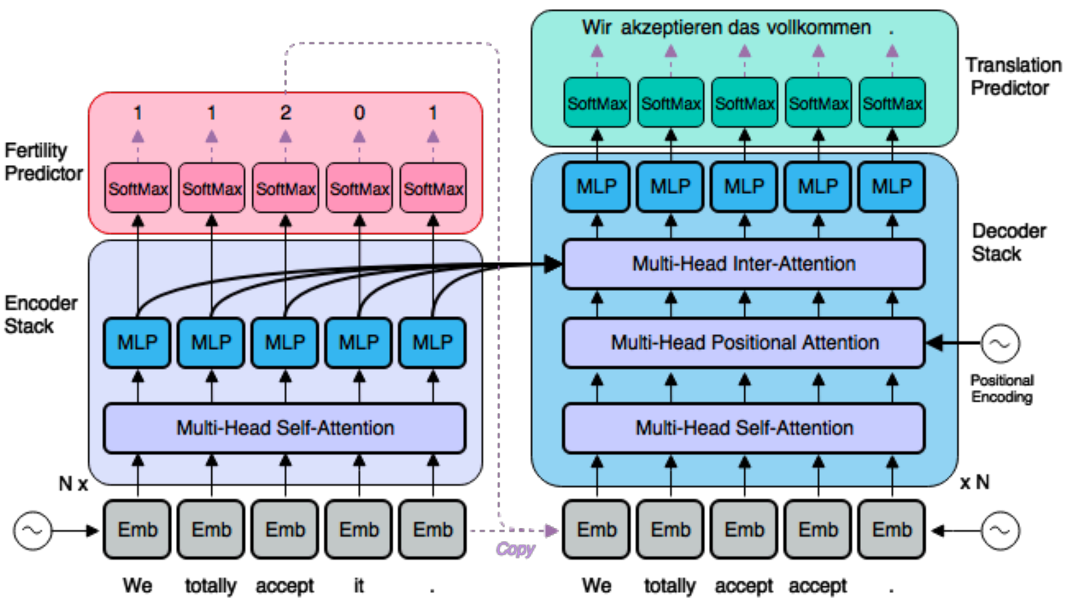
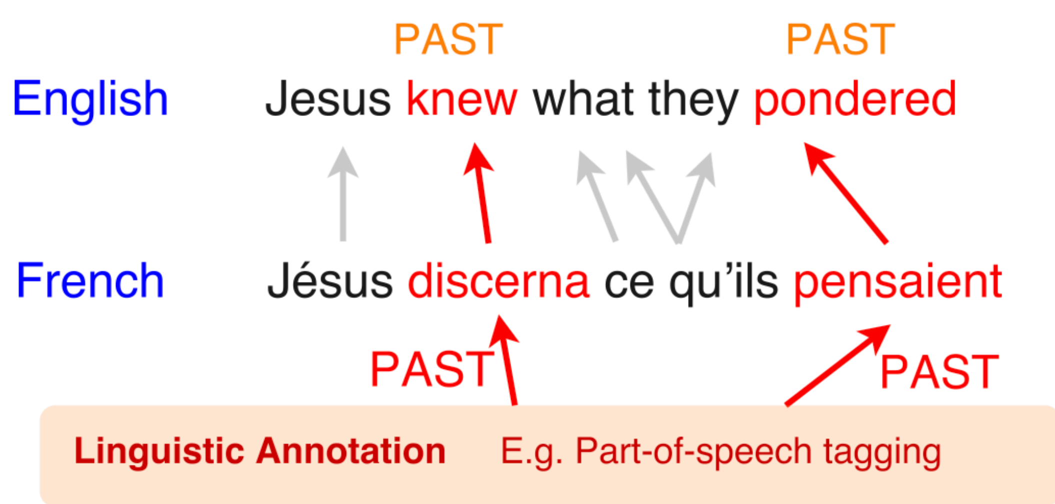


Figure adapted from https://blog.salesforceairesearch.com/

Annotation Projection

- Based on (statistical) word alignment inferred from parallel text.
- Resource creation for low-resource languages.

Important area of NLP research: Yarowsky et al. (2001); Spreyer and Frank (2008); Padó & Lapata (2009); Das and Petrov (2011); Agić et al. (2016).



1500+ languages crawler

Annotation Projection – Example

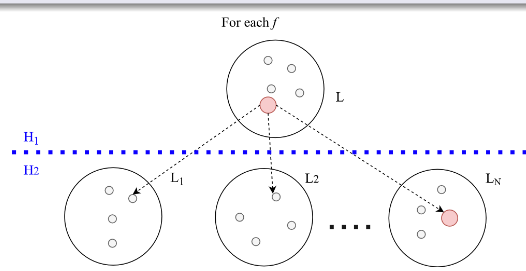
- Use word tokens in 1000+ languages as marker of linguistic distinction.
- But we need accurate alignments in 1000+ languages.

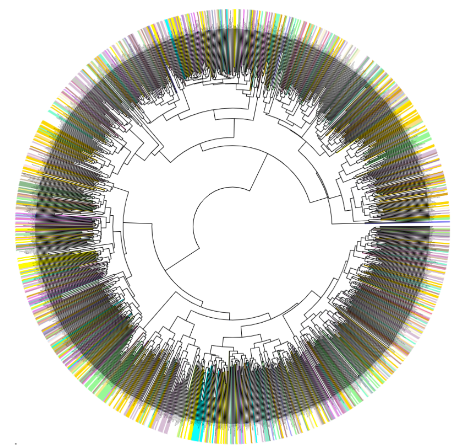
H1 Overt encoding exists.

$\forall$ linguistic distinction $f \rightarrow \exists$ few languages that encode f overtly

H2 Overt encoding can be projected.

Projection of f to a language $l' \rightarrow$ either overt or non-overt markers in l'.





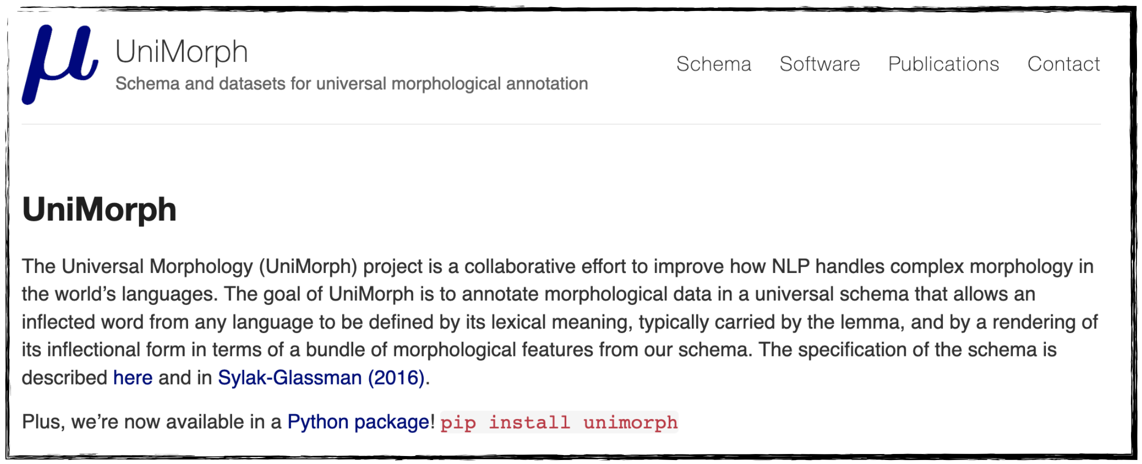
Ehsaneddin Asgari and Hinrich Schütze.
Past, Present, Future: A Computational Investigation of the Typology of Tense in 1000 Languages.
In Proceedings of the **EMNLP 2017**.

Multilingual NLP

- Machine Translation
- Annotation Projection
- Collaborative Efforts
- Multilingual LM

Collaborative Efforts in Resource Creation

- Annotation of morphological data in a universal schema


UniMorph
Schema and datasets for universal morphological annotation

Schema Software Publications Contact

UniMorph

The Universal Morphology (UniMorph) project is a collaborative effort to improve how NLP handles complex morphology in the world's languages. The goal of UniMorph is to annotate morphological data in a universal schema that allows an inflected word from any language to be defined by its lexical meaning, typically carried by the lemma, and by a rendering of its inflectional form in terms of a bundle of morphological features from our schema. The specification of the schema is described here and in Sylak-Glassman (2016).

Plus, we're now available in a Python package! `pip install unimorph`

https://unimorph.github.io/

Multilingual Language Model

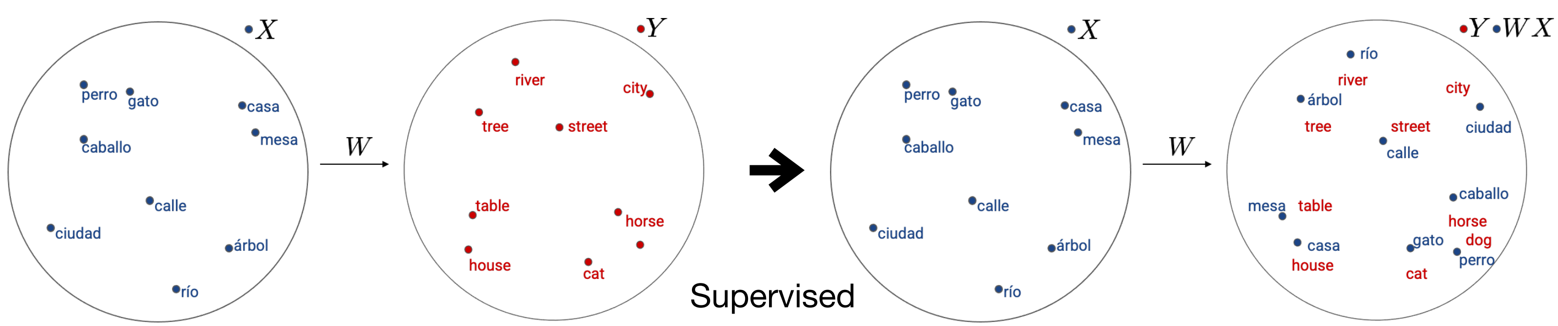
- Shared embedding spaces of language units among 1+ languages

Previous paradigm	New paradigm
Language-specific NLP models **Language-specific** feature computation and preprocessing	**Representation learning**: inputs are **semantic vectors which are multilingual (embeddings)**

Multilingual Language Model

- Shared embedding spaces of language units among 1+ languages

Previous paradigm	New paradigm
Language-specific NLP models	**Representation learning**: inputs are **semantic vectors which are multilingual (embeddings)**



Parallel sentences or words

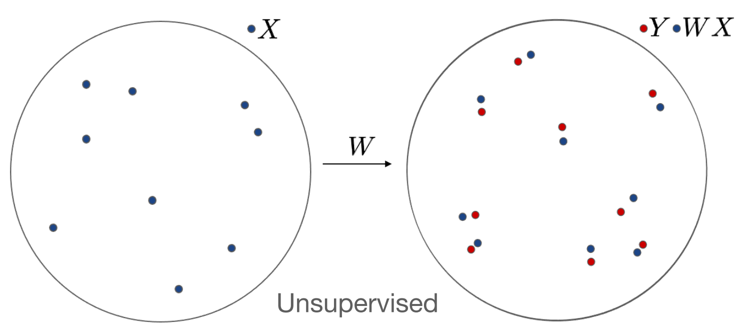
Multilingual NLP

- Machine Translation
- Annotation Projection
- Collaborative Effort
- Multilingual LM

Multilingual Language Model

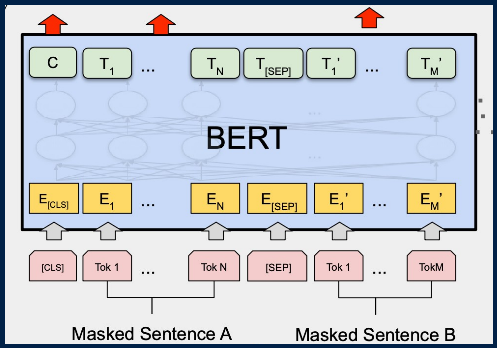
- Shared embedding spaces of language units among 1+ languages

Previous paradigm	New paradigm
Language-specific NLP models	**Representation learning**: inputs are **semantic vectors which are multilingual (embeddings)**



Multilingual Model

Ideas?

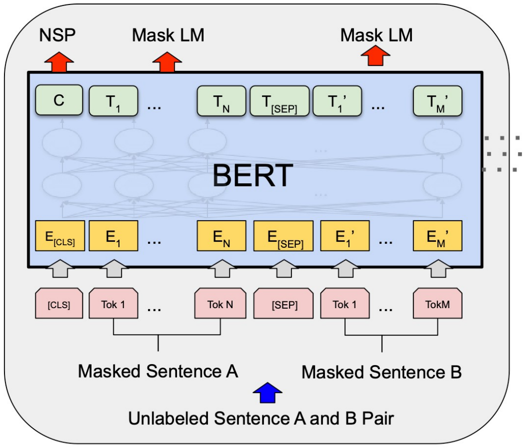


mBERT

- After multilingual MLM pretraining encodes text from any of the languages seen in pretraining

- Zero-shot language transfer for downstream NLP tasks

Telmo Pires, Eva Schlinger, and Dan Garrette.
How Multilingual is Multilingual BERT?
In Proceedings of the ACL 2019.

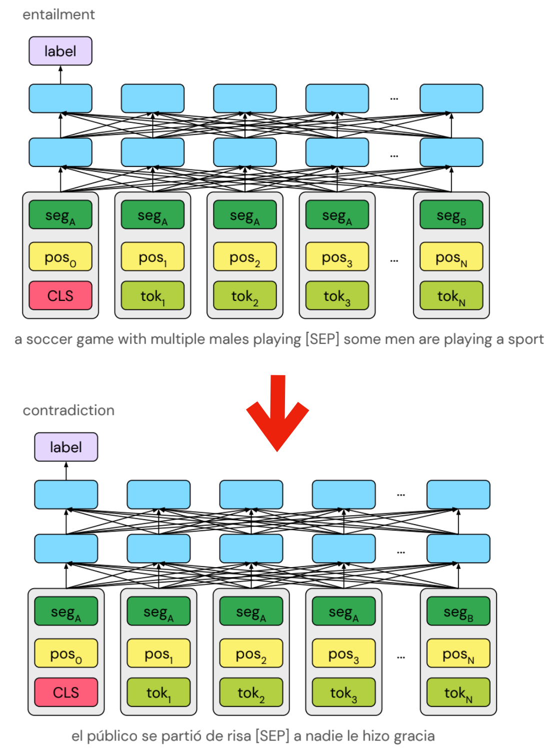


Pre-training

- mBERT Model
  - How Zeroshot?
  - Essentials?
- Word Piece Recap
- XLM-R Model
- XLM-V Model

mBERT

- Step 1: Combine corpora & learn joint subword vocab
  - Wikipedia pages of 104 languages with a shared vocabulary of 110K.
- Step 2: Joint pre-training
- Step 3: English fine-tuning
- Step 4: Zero-shot transfer



Telmo Pires, Eva Schlinger, and Dan Garrette.
How Multilingual is Multilingual BERT?
In Proceedings of the ACL 2019.

mBERT

- Shared Vocabulary?
  - Wikipedia pages of 104 languages with a shared vocabulary of 110K.

Do all languages need the same amount Vocab size?

- Zero-shot Transfer is the same for all pairs?

(1) from the same language family (subword overlap and word order)
(2) with large corpora in pretraining

Telmo Pires, Eva Schlinger, and Dan Garrette.
How Multilingual is Multilingual BERT?
In Proceedings of the ACL 2019.

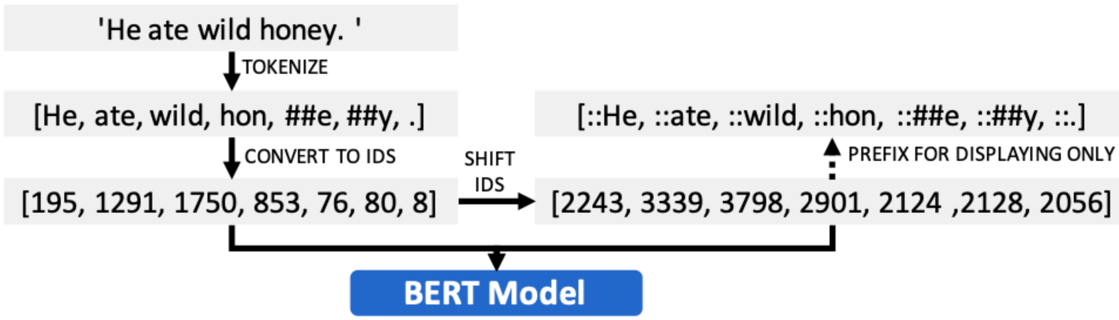
Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan.
When is BERT Multilingual? Isolating Crucial Ingredients for Cross-lingual Transfer.
In Proceedings of the NAACL 2022.

- mBERT Model
  - How Zeroshot?
  - Essentials?
- Word Piece Recap
- XLM-R Model
- XLM-V Model

Essentials for BERT multilinguality

Change in the surface form: English and Fake-English created by shifting unicode points

10k sentences of the Old Testament of the English King James Bible.



BERT-small model: We use the BERT-Base architecture modified to achieve a smaller model: we divide hidden sizes, etc intermediate size of the feed forward layer and number of attention heads by 12; thus, hidden size is 64 and intermediate size 256. While this leaves us with a single attention head,

Philipp Dufter and Hinrich Schütze.
Identifying Elements Essential for BERT's Multilinguality.
In Proceedings of the EMNLP, 2020.

- mBERT Model
  - How Zeroshot?
  - Essentials?
- Word Piece Recap
- XLM-R Model
- XLM-V Model

Essentials for BERT multilinguality

Evaluation of model multilinguality

(1) Sentence Retrieval (ρ)

$$R_{ij} = \text{cosine} - \text{sim}\left(e_i^{(\text{eng})}, e_j^{(\text{fake})}\right)$$

$$\rho = \frac{1}{2m} \sum_{i=1}^{m} \mathbb{1}_{\arg\max_l R_{il}=i} + \mathbb{1}_{\arg\max_l R_{li}=i}.$$

(2) Word Translation (τ)

word translation [CLS] {token} [SEP]

We use layer 0 (uncontextualized) and layer 8 (contextualized). Several papers have found layer 8 to work well for monolingual and multilingual tasks

$$\mu = 1/4\left(\tau_0 + \tau_8 + \rho_0 + \rho_8\right)$$

Philipp Dufter and Hinrich Schütze.
Identifying Elements Essential for BERT's Multilinguality.
In Proceedings of the EMNLP, 2020.

- mBERT Model
    - How Zeroshot?
    - Essentials?
- Word Piece Recap
- XLM-R Model
- XLM-V Model

Essentials for BERT multilinguality

Evaluation of model perplexity

$$x^{(1)}, x^{(2)}, \cdots, x^{(m)} \qquad length(x^{(i)}) = n_i \qquad M = \sum_{i=1}^{m} n_i$$

$$\prod_{1}^{m} P(x^{(i)}) \longrightarrow \prod_{1}^{m} \frac{1}{P(x^{(i)})} \longrightarrow \sqrt[M]{\frac{1}{\prod_1^m P(x^{(i)})}}$$

$$\sqrt[M]{\frac{1}{\prod_1^m P(x^{(i)})}} = 2^{log_2 \sqrt[M]{\frac{1}{\prod_1^m P(x^{(i)})}}} = 2^{-\frac{1}{M}\sum_{i=1}^{m} log_2 P(x^{(i)})}$$

Philipp Dufter and Hinrich Schütze.
Identifying Elements Essential for BERT’s Multilinguality.
In Proceedings of the EMNLP, 2020.

- mBERT Model
  - How Zeroshot?
  - Essentials?
- Word Piece Recap
- XLM-R Model
- XLM-V Model

Results and Conclusions

i) Shared position embeddings, shared special tokens, replacing masked tokens with random tokens (of the other language) and a limited amount of parameters are necessary elements for multilinguality.

ii) Word order is relevant: BERT is not multilingual with one language having an inverted word order.

iii) The comparability of training corpora contributes to multilinguality.

ID	Description	Mult.-score μ	Layer 0 Align. F_1	Layer 0 Retr. ρ	Layer 0 Trans. τ	Layer 8 Align. F_1	Layer 8 Retr. ρ	Layer 8 Trans. τ	MLM-Perpl. train	MLM-Perpl. dev
0	original	.70	1.00 $_{.00}$	.16 $_{.02}$	.88 $_{.02}$	1.00 $_{.00}$	.97 $_{.01}$	.79 $_{.03}$	9 $_{0.2}$	217 $_{7.8}$
1	lang-pos	.30	.87 $_{.05}$	.33 $_{.13}$	.40 $_{.09}$	.89 $_{.05}$	.39 $_{.15}$	.09 $_{.05}$	9 $_{0.1}$	216 $_{9.0}$
2	shift-special	.66	1.00 $_{.00}$	.15 $_{.02}$	.88 $_{.01}$	1.00 $_{.00}$	.97 $_{.02}$	.63 $_{.13}$	9 $_{0.1}$	227 $_{17.9}$
4	no-random	.68	1.00 $_{.00}$	.19 $_{.03}$	.87 $_{.02}$	1.00 $_{.00}$	.85 $_{.07}$	.82 $_{.04}$	9 $_{0.6}$	273 $_{7.7}$
5	lang-pos;shift-special	.20	.62 $_{.19}$	.22 $_{.19}$	.27 $_{.20}$	.72 $_{.22}$	.27 $_{.21}$	.05 $_{.04}$	10 $_{0.5}$	205 $_{7.6}$
6	lang-pos;no-random	.30	.91 $_{.04}$	.29 $_{.10}$	.36 $_{.12}$	.89 $_{.05}$	.32 $_{.15}$	.25 $_{.12}$	10 $_{0.4}$	271 $_{8.6}$
7	shift-special;no-random	.68	1.00 $_{.00}$	.21 $_{.03}$	.85 $_{.01}$	1.00 $_{.00}$	.89 $_{.06}$	.79 $_{.04}$	8 $_{0.3}$	259 $_{15.6}$
8	lang-pos;shift-special;no-random	.12	.46 $_{.26}$	.09 $_{.09}$	.18 $_{.22}$	.54 $_{.31}$	.11 $_{.11}$	.11 $_{.13}$	10 $_{0.6}$	254 $_{15.9}$
15	overparam	.58	1.00 $_{.00}$	.27 $_{.03}$	.63 $_{.05}$	1.00 $_{.00}$	.97 $_{.01}$	.47 $_{.06}$	2 $_{0.1}$	261 $_{4.5}$
16	lang-pos;overparam	.01	.25 $_{.10}$	.01 $_{.00}$	.01 $_{.00}$	.37 $_{.13}$	.01 $_{.00}$	.00 $_{.00}$	3 $_{0.0}$	254 $_{4.9}$
17	lang-pos;shift-special;no-random;overparam	.00	.05 $_{.02}$	.00 $_{.00}$	.00 $_{.00}$	.05 $_{.04}$	.00 $_{.00}$	.00 $_{.00}$	1 $_{0.0}$	307 $_{7.7}$
3	inv-order	.01	.02 $_{.00}$	.00 $_{.00}$	.01 $_{.00}$	.02 $_{.00}$	.01 $_{.01}$	.00 $_{.00}$	11 $_{0.3}$	209 $_{14.4}$
9	lang-pos;inv-order;shift-special;no-random	.00	.04 $_{.01}$	.00 $_{.00}$	.00 $_{.00}$	.03 $_{.01}$	.00 $_{.00}$	.00 $_{.00}$	10 $_{0.4}$	270 $_{20.1}$

Philipp Dufter and Hinrich Schütze.
Identifying Elements Essential for BERT's Multilinguality.
In Proceedings of the EMNLP, 2020.

بسم الله الرحمن الرحیم

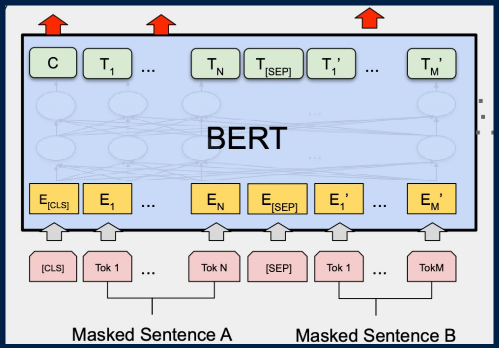
Multilingual Models & Data Processing (II)

Ehsaneddin Asgari

Nov. 7th 2023

Artificial Intelligence Group
Computer Engineering Department, SUT

Multilingual Model



- mBERT Model
- XLM Model
- MAD-X Model

Recap: mBERT

Training MLM BERT on multilingual data

Contributing factors to multilinguality

i) Shared position embeddings
ii) Shared special tokens
iii) Replacing masked tokens with random tokens (of the other language)
iv) limited amount of parameters are necessary elements for multilinguality.
v) Word order is relevant: BERT is not multilingual with inverted word order.
vi) The comparability of training corpora contributes to multilinguality.

Philipp Dufter and Hinrich Schütze.
Identifying Elements Essential for BERT's Multilinguality.
In Proceedings of the EMNLP, 2020.

Extending to 1000 languages?
Any Challenges?

- mBERT Model
- XLM Model
- MAD-X Model

Curse of multilinguality

Training a model on more languages means
it has less capacity to learn about each one.

- Parameter Allocation
- Language Interference
- Data Imbalance
- Language-Specific Features

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov.
Unsupervised Cross-lingual Representation Learning at Scale.
In Proceedings of the **ACL 2020**.

- mBERT Model
- XLM Model
- MAD-X Model

XLM Model

Model

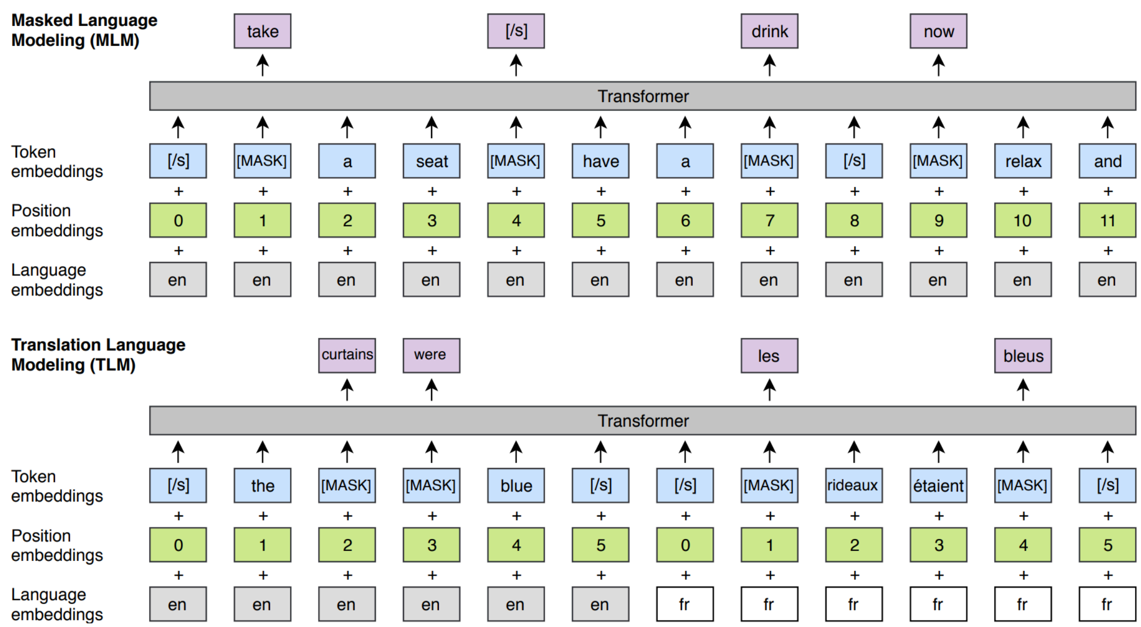


Figure 1: **Cross-lingual language model pretraining.** The MLM objective is similar to the one of Devlin et al. (2018), but with continuous streams of text as opposed to sentence pairs. The TLM objective extends MLM to pairs of parallel sentences. To predict a masked English word, the model can attend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

Conneau, Alexis, and Guillaume Lample.
Cross-lingual language model pretraining.
NeurIPS 2019.

- mBERT Model
- XLM Model
- MAD-X Model

XLM Model

BPE as subword tokenizer

- Sentences are sampled according to a multinomial distribution with probabilities q_i for language i.
- α = 0.5 (to promote low resource languages).

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^{N} p_j^\alpha} \qquad p_i = \frac{n_i}{\sum_{k=1}^{N} n_k}$$

Conneau, Alexis, and Guillaume Lample.
Cross-lingual language model pretraining.
NeurIPS 2019.

- mBERT Model
- XLM Model
- MAD-X Model

XLM Model

Model

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

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

Conneau, Alexis, and Guillaume Lample.
Cross-lingual language model pretraining.
NeurIPS 2019.

- mBERT Model
- XLM Model
- MAD-X Model

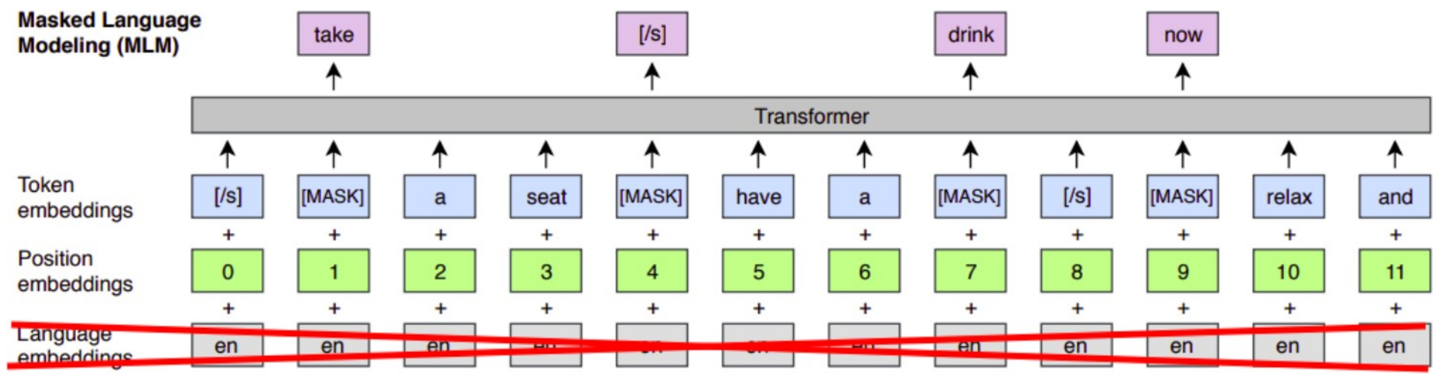
Advent of XMLR Model

Model	Objective	Pre-training data	Languages	Tokenizer & Vocab.	Model Size (Params)
BERT	MLM & NSP	Wikipedia	English	WordPiece & 30K	110M(base) & 335M(large)
mBERT	MLM & NSP	Wikipedia	104	WordPiece & 110K	172M
XLM	MLM & TLM	Wikipedia & Parallel sentences	100	BPE	?
RoBERTa	MLM	Wiki, CC-News, OpenWebText, CommonCrawl	English	bBPE & 50K	125M(base) & 355M(large)
XLM-R	MLM	CommonCrawl	100	Unigram & 250K	270M(base) & 550M(large)

- mBERT Model
- XLM Model
- MAD-X Model

XLMR Model

Model

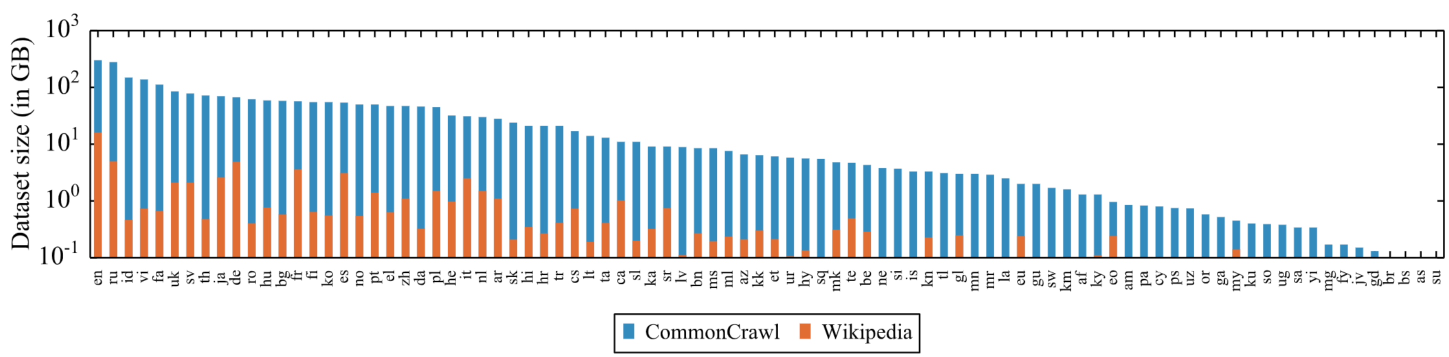


Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov.
Unsupervised Cross-lingual Representation Learning at Scale.
In Proceedings of the **ACL 2020**.

XLMR Model

Dataset

- CC-100, a clean CommonCrawl Corpus in 100 languages
- Use an internal language identification model in combination with the one from fastText
- Train language models in each language and use it to filter documents
- Significant dataset size increase, especially for low-resource languages



Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov.
Unsupervised Cross-lingual Representation Learning at Scale.
In Proceedings of the **ACL 2020**.

- mBERT Model
- XLM Model
- MAD-X Model

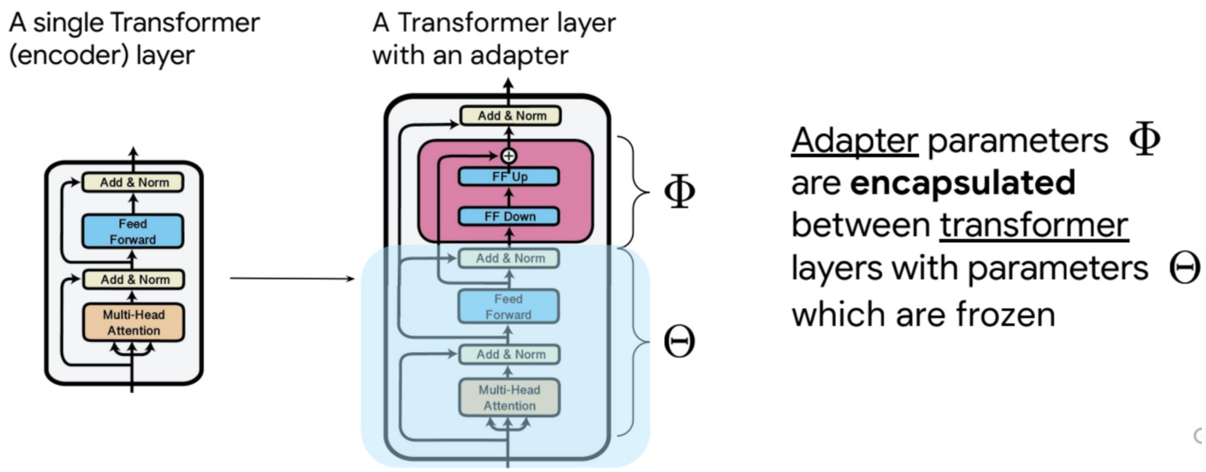


Figure 17: Transformer layer with an adapter [Ruder, 2022]

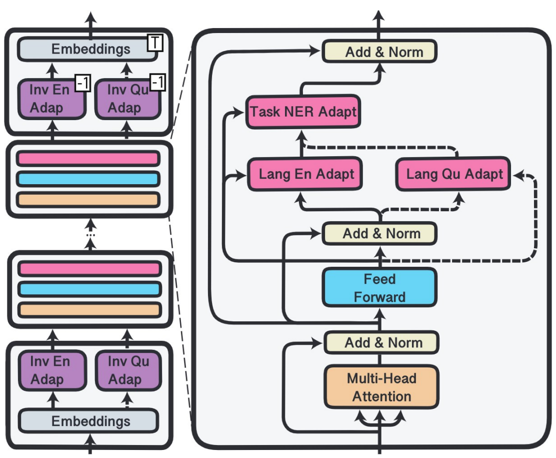
- Allocate additional capacity for each language using adapters
- Using a SOTA MLM as foundation, adapt the model to arbitrary tasks and languages by learning modular language- and task-specific representations via adapters
- Small bottleneck layers inserted between a pre-trained model's weights

- mBERT Model
- XLM Model
- MAD-X Model

- ***Step 1: Train Language Adapters***
  Train language adapters for the source language and the target language with MLM on Wikipedia

- ***Step 2: Train a Task Adapter***
  Train a task adapter in the source language stacked on top of the source language adapter. The language adapter and the transformer weights are frozen. Only the task adapter is trained

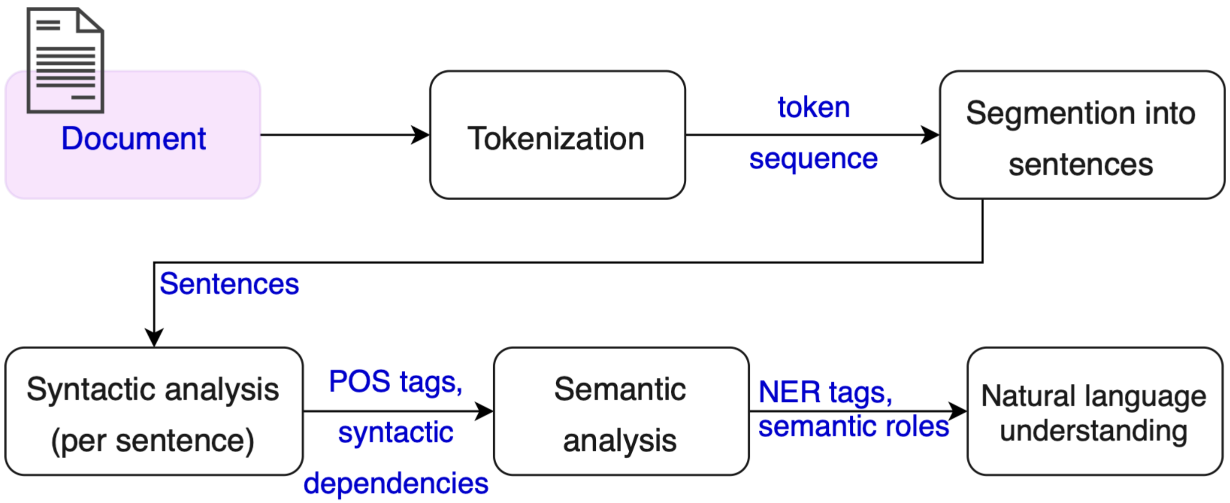
- ***Step 3: Zero-Shot Transfer to Target Language***
  Replace the source language adapter with the target language adapter, while keeping the "language agnostic" task adapter fixed



Is the tokenization important at all?

- **Why Subword?**
- Char-level
- BPE
- Subword Reg.
- BPE-Dropout
- Multi-granuality
- Multilinguality

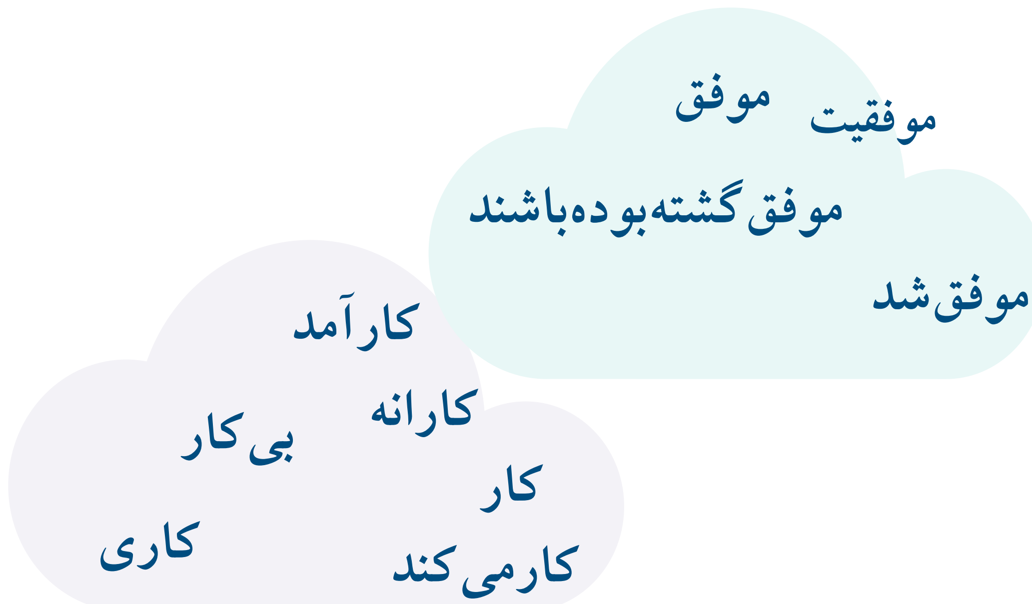
NLP Traditional Pipeline



Tokenization issue?

- **Why Subword?**
- Char-level
- BPE
- Subword Reg.
- BPE-Dropout
- Multi-granuality
- Multilinguality

Shared Morphemes within the Language



- **Why Subword?**
- Char–level
- BPE
- Subword Reg.
- BPE–Dropout
- Multi–granuality
- Multilinguality

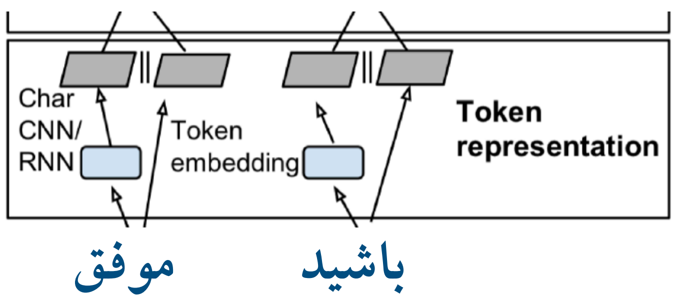
Shared Morphemes among Languages

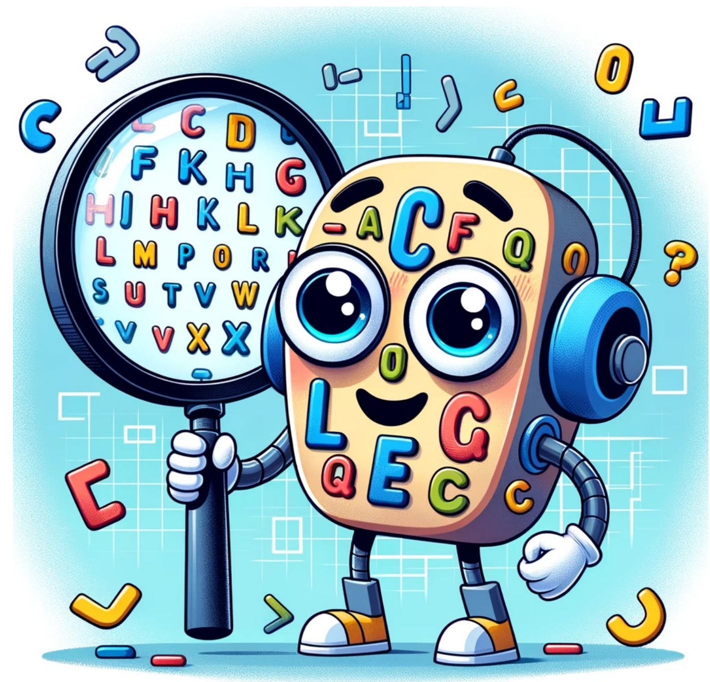
English Suffix	German	English	French	Italian	Spanish	Latin	Romanian
-tion	Information	Information	Information	Informazione	Información	Informatio	Informație
-ity	Qualität	Quality	Qualité	Qualità	Calidad	Qualitas	Calitate
-al	Global	Global	Global	Globale	Global	Globalis	Global
-ist	Spezialist	Specialist	Spécialiste	Specialista	Especialista	Specialistus	Specialist
-ism	Kapitalismus	Capitalism	Capitalisme	Capitalismo	Capitalismo	Capitalismus	Capitalism

- Why Subword?
- **Char-level**
- BPE
- Subword Reg.
- BPE-Dropout
- Multi-granuality
- Multilinguality

Character-level

مـ و فـ قـ يـ ت



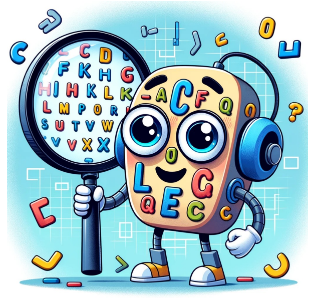


- Why Subword?
- Char–level
- BPE
- Subword Reg.
- BPE–Dropout
- Multi–granuality
- Multilinguality

Character–level

Advantages	Disadvantages
1. Smaller Vocabulary Size	1. Longer Sequences
2. Handles OOV Words	2. Limited Context Understanding
3. Captures Morphological Patterns	3. Training Difficulty
4. Language Agnosticism	4. Slower Processing Speed
5. Robustness to Noise	5. Suboptimal for Certain Tasks

Adel, Heike, Ehsaneddin Asgari, and Hinrich Schütze.
Overview of character-based models for natural language processing.
Computational Linguistics and Intelligent Text Processing 2017.



Byte-pair Encoding (BPE)

- Step 0: Set up vocabulary.
- Step 1: Represent words using characters
- Step 2: Count character pairs in vocabulary.
- Step 3: Merge highest frequency pairs, new symbol.
- Step 4: Continue merging until reaching desired vocab size.

Initial vocabulary: characters

↓

Split each word into characters

Words in the data:

word	count
c a t	4
m a t	5
m a t s	2
m a t e	3
a t e	3
e a t	2

Current merge table:

(empty)

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. ACL 2016

Gif from: https://tinyurl.com/22xk95hj

- Why Subword?
- Char-level
- **BPE (& BBPE)**
- Subword Reg.
- BPE-Dropout
- Multi-granuality
- Multilinguality

Byte-level Byte-pair Encoding (BBPE)

Original		質問して▁証明と証拠を求めましょう	Ask▁questions,▁demand▁proof,▁demand▁evidence.
Byte		E8 B3 AA E5 95 8F E3 81 97 E3 81 A6 E2 96 81 E8 A8 BC E6 98 8E E3 81 A8 E8 A8 BC E6 8B A0 E3 82 92 E6 B1 82 E3 82 81 E3 81 BE E3 81 97 E3 82 87 E3 81 86	41 73 6B E2 96 81 71 75 65 73 74 69 6F 6E 73 2C E2 96 81 64 65 6D 61 6E 64 E2 96 81 70 72 6F 6F 66 2C E2 96 81 64 65 6D 61 6E 64 E2 96 81 65 76 69 64 65 6E 63 65 2E
BBPE	1K	E8 B3 AA E595 8F しE381 A6 ▁E8 A8 BC 明 E381 A8 E8 A8 BC E6 8B A0 をE6 B1 82 めE381 BE しょう	A s k ▁quest ions , ▁dem and ▁pro of , ▁dem and ▁ev idence .
	2K	E8 B3 AA 問 しE381 A6 ▁E8 A8BC 明 E381 A8 E8 A8BC E68B A0 を E6 B1 82 めE381 BE しょう	A s k ▁qu est ion s , ▁d em and ▁pro o f , ▁d em and ▁e v id ence .
	4K	E8 B3 AA 問 しE381 A6 ▁E8 A8BC 明E381 A8 E8 A8BC 拠 をE6 B1 82 めE381 BE しょう	As k ▁quest ions , ▁d em and ▁pro of , ▁d em and ▁ev id ence .
	8K	E8 B3 AA問 しE381 A6 ▁E8 A8BC 明E381 A8 E8 A8BC 拠 をE6 B1 82めE381 BE しょう	As k ▁questions , ▁demand ▁pro of , ▁demand ▁evidence .
	16K	E8 B3 AA問 しE381 A6 ▁E8 A8BC 明E381 A8 E8 A8BC 拠 をE6 B1 82めE381 BE しょう	As k ▁questions , ▁demand ▁proof , ▁demand ▁evidence .
	32K	E8 B3 AA問しE381 A6 ▁E8 A8BC 明E381 A8 E8 A8BC 拠 をE6 B1 82 めE381 BE しょう	As k ▁questions , ▁demand ▁proof , ▁demand ▁evidence .
CHAR		質 問 し て ▁ 証 明 と 証 拠 を 求 め ま し ょ う	A s k ▁ q u e s t i o n s , ▁ d e m a n d ▁ p r o o f , ▁ d e m a n d ▁ e v i d e n c e .
BPE	16K	質問 して ▁ 証明 と 証拠 を求 め ましょう	As k ▁questions , ▁demand ▁pro of , ▁demand ▁evidence .
	32K	質問 して ▁証明 と 証拠 を求め ましょう	As k ▁questions , ▁demand ▁proof , ▁demand ▁evidence .

Wang, Changhan; Cho, Kyunghyun; Gu, Jiatao.
Neural machine translation with byte-level subwords.
In Proceedings of AAAI 2020.

- Why Subword?
- Char–level
- **BPE (& BBPE)**
- Subword Reg.
- BPE–Dropout
- Multi–granuality
- Multilinguality

Byte–level Byte–pair Encoding (BBPE)

- **Rare characters from noisy text** or **character-rich languages** such as Japanese and Chinese however can unnecessarily take up vocabulary slots and limit its compactness. Representing text at the level of bytes and using the 256 byte set as vocabulary is a potential solution to this issue.
- We claim that contextualizing BBPE embeddings is necessary, which can be implemented by a convolutional or recurrent layer. Our experiments show that BBPE has **comparable performance to BPE** while its size is only **1/8 of that for BPE**.
- In the multilingual setting, **BBPE maximizes vocabulary sharing across many languages and achieves better translation quality..**
  - **Maybe because of various token granularities in multilingual parallel sentences at the token level**
- BBPE enables transferring models between languages with non-overlapping character sets.

Wang, Changhan; Cho, Kyunghyun; Gu, Jiatao.
Neural machine translation with byte-level subwords.
In Proceedings of AAAI 2020.

Mengjiao Zhang and Jia Xu.
Byte-based Multilingual NMT for Endangered Languages.
In Proceedings of COLING 2022.

- Why Subword?
- Char-level
- **BPE (& BBPE)**
- Subword Reg.
- BPE-Dropout
- Multi-granuality
- Multilinguality

(B)?BPE

- Step 0: Set up vocabulary.
- Step 1: Represent words using characters / bytes
- Step 2: Count character/bytes pairs in vocabulary
- Step 3: Merge highest frequency pairs, new symbol.
- Step 4: Continue merging until reaching desired vocab size.

Issues?

Rico Sennrich, Barry Haddow, and Alexandra Birch.
Neural Machine Translation of Rare Words with Subword Units.
ACL 2016

- Why Subword?
- Char-level
- BPE (& BBPE)
- **Subword Reg.**
- BPE-Dropout
- Multi-granuality
- Multilinguality

Subword Regularization

Unigram language model, which is capable of outputing multiple subword segmentations with probabilities.

Given Vocabulary V, we want to estimate p(x_i)

$X^{(s)} \in D \rightarrow$ "sentence"
$\mathbf{x} = (x_1, \ldots, x_M) \rightarrow$ "subword sequence"

$$p(\mathbf{x}) = \prod_{i=1}^{M} p(x_i) \rightarrow \text{"unigram language model"}$$

Subwords (_ means spaces)	Vocabulary id sequence
_Hell/o/_world	13586 137 255
_H/ello/_world	320 7363 255
_He/llo/_world	579 10115 255
_/He/l/l/o/_world	7 18085 356 356 137 255
_H/el/l/o/_/world	320 585 356 137 7 12295

Table 1: Multiple subword sequences encoding the same sentence “Hello World”

Taku Kudo.
Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates.
In Proceedings of the ACL 2018.

- Why Subword?
- Char-level
- BPE (& BBPE)
- **Subword Reg.**
- BPE-Dropout
- Multi-granuality
- Multilinguality

Subword Regularization

1. Heuristically make a reasonably **big seed vocabulary V**
2. Repeat the following steps **until |V| reaches a desired vocabulary size**.
    (a) Fixing the set of vocabulary, optimize p(x) with the EM algorithm.

    $\mathscr{L} = \sum_{s=1}^{|D|} \log\left(P\left(X^{(s)}\right)\right) = \sum_{s=1}^{|D|} \log\left(\sum_{\mathbf{x}\in\mathscr{S}(X^{(s)})} P(\mathbf{x})\right)$ → Log likelihood

    $X^{(s)} \in D$ → "sentence"

    $|D|$ → "size of the dataset"

    $\mathscr{S}(X^{(s)})$ → "set of segmentation candidates built from the input sentence "$X^{(s)}$

    $\mathbf{x} = (x_1, \ldots, x_M)$ → "subword sequence"

    $p(\mathbf{x}) = \prod_{i=1}^{M} p(x_i)$ → "unigram language model"

    (b) Compute the $loss_i$ for each subword x_i , where $loss_i$ represents how likely the likelihood L is reduced when the subword x_i is removed from the current vocabulary.

    (c) Sort the symbols by $loss_i$ and keep top η % of subwords (η is 80, for example). Note that we always keep the subwords consisting of a single character to avoid out-of-vocabulary.

Taku Kudo.
Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates.
In Proceedings of the ACL 2018.

BPE–Dropout

BPE-dropout - simple and effective subword regularization method based on and compatible with conventional BPE.

It stochastically corrupts the segmentation procedure of BPE, which leads to producing multiple segmentations within the same fixed BPE framework.

Using BPE-dropout during training and the standard BPE during inference improves translation quality compared to the previous subword regularization.

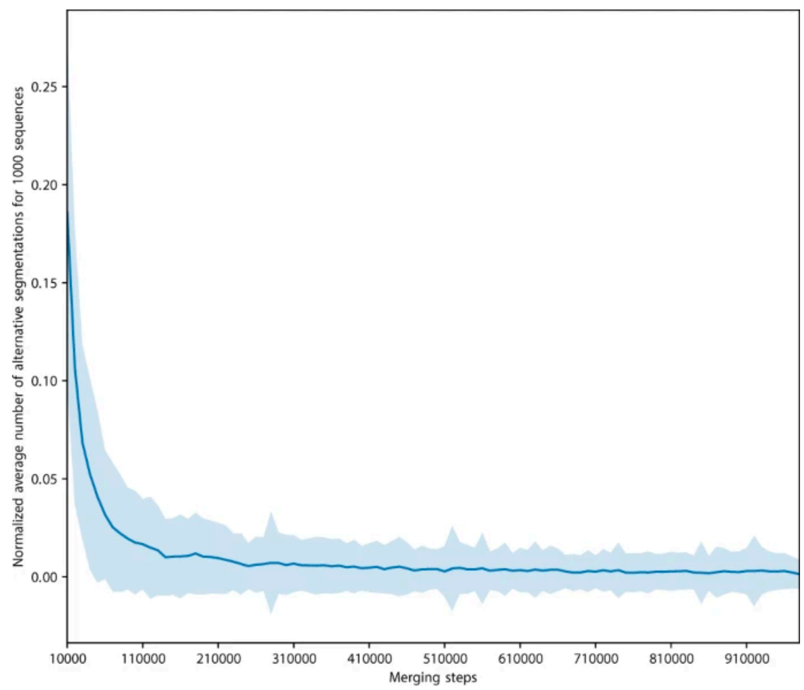
Algorithm 1: BPE-dropout

```
current_split ← characters from input_word;
do
    merges ← all possible merges[1] of tokens
      from current_split;
    for merge from merges do
        /* The only difference
           from BPE                    */
        remove merge from merges with the
          probability p;
    end
    if merges is not empty then
        merge ← select the merge with the
          highest priority from merges;
        apply merge to current_split;
    end
while merges is not empty;
return current_split;
```


Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita.
BPE-Dropout: Simple and Effective Subword Regularization.
In Proceedings of the **ACL 2020**.

- Why Subword?
- Char–level
- BPE (& BBPE)
- Subword Reg.
- BPE–Dropout
- **Multi–granuality**
- Multilinguality

Multi–Granulity BPE for Bioinformatics



- m-a-t-l-a-a-p-p-p-p-l-g-e-s-g-n-s-n-s-v-s-r
- ma t l aa pppp l g es g nsn svsr
- ma tlaa pppp l g es g nsn svsr
- ma tlaa ppppl g esg nsn svsr
- ma tlaa ppppl gesg nsn svsr
- matlaa ppppl gesg nsn svsr

Asgari, Ehsaneddin, Alice C. McHardy, and Mohammad RK Mofrad.
Probabilistic variable-length segmentation of protein sequences for discriminative motif discovery…
Scientific reports 2019.

- Why Subword?
- Char–level
- BPE (& BBPE)
- Subword Reg.
- BPE–Dropout
- Multi–granuality
- **Multilinguality**

XLM–V

- Large multilingual language models typically rely on a single vocabulary shared across 100+ languages.
- As these models have increased in parameter count and depth, vocabulary size has remained largely unchanged. This vocabulary bottleneck limits the representational capabil- ities of multilingual models like XLM-R.
- While multilingual language models have increased in parameter count and depth over time, vocabulary size has largely remained unchanged:
- 250K token vocabulary size as XLM-R base (Conneau et al., 2019), a 250M parameter model.

XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Masked Language Models
Arxiv Oct 2023

- Why Subword?
- Char-level
- BPE (& BBPE)
- Subword Reg.
- BPE-Dropout
- Multi-granuality
- **Multilinguality**

XLM-V

- Vocabulary bottleneck hinders the performance of multilingual models on question answering and sequence labeling where in-depth token-level and sequence-level understanding is essential (Wang et al., 2019).
- (1) vocabularies can be improved by de-emphasizing token sharing between languages with little lexical overlap
- (2) proper vocabulary capacity allocation for individual languages is crucial for ensuring that diverse languages are well-represented.

XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Masked Language Models
Arxiv Oct 2023

XLM-V

Finding language clusters

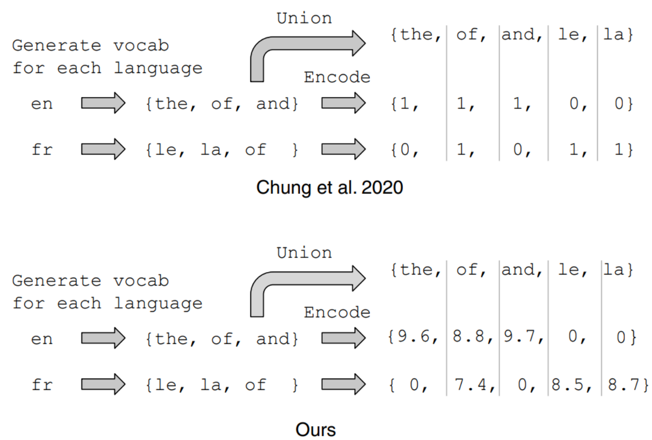


Figure 1: Similar to Chung et al. (2020), we also leverage the per-language sentencepiece vocabularies as a "lexical fingerprint" for clustering. However, instead of using binary vectors, we use the unigram log probability instead.

XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Masked Language Models
Arxiv Oct 2023

- Why Subword?
- Char-level
- BPE (& BBPE)
- Subword Reg.
- BPE-Dropout
- Multi-granuality
- **Multilinguality**

XLM-V

Results

Model	XNLI Acc.	NER Acc.	MLQA EM / F1	TyDiQA EM / F1	XQuAD EM / F1	ANLI F1	MNER F1	Average
XLM	69.1	-	32.6 / 48.5	29.1 / 43.6	44.3 / 59.8	-	-	-
XLM-R	76.2	-	46.3 / 63.7	- / -	- / -	38.5	-	-
XLM-R *reimpl.*	74.9	61.3	46.7 / 64.4	38.3 / 56.0	56.0 / 71.3	39.6	20.9	55.5
XLM-V	**76.0**	**64.7**	**47.7 / 66.0**	**39.7 / 56.9**	**56.3 / 71.9**	**45.4**	**32.1**	**59.0**

Table 2: Overall results across multiple multilingual datasets comparing our model against the XLM and XLM-R baselines. All results are based on crosslingual transfer after fine-tuning on English data. We computed the average result using the accuracy or F1 of each task. "*reimpl*" is our re-implementation of finetuning, used by both XLM-R and XLM-V. Please refer to the appendix for specific hyperparameters to reproduce each result. EM stands for exact match. ANLI refers to AmericasNLI and MNER refers to MasakhaNER.

XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Masked Language Models
Arxiv Oct 2023


תודה
Hvala
Teşekkür ederim
Хвала
متشكرم
Dankon
ขอบคุณ
Ευχαριστώ
謝謝
شكرا لك
Tak
Gracias
Grazie
Merci
Thank you
Sağol
Danke
Obrigado
감사합니다
谢谢
Tack
Спасибо
Спасибі
有り難う
Köszönöm
Dank u
धन्यवाद
ありがとう
Благодаря
Asante
Dank u
Multumesc
Terima kasih
Dziękuję
شكرا
Kiitos