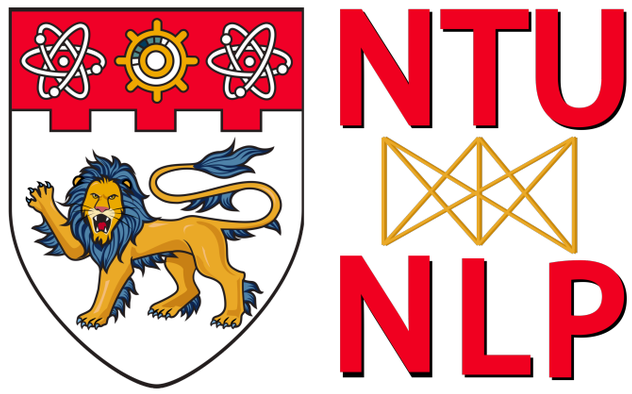
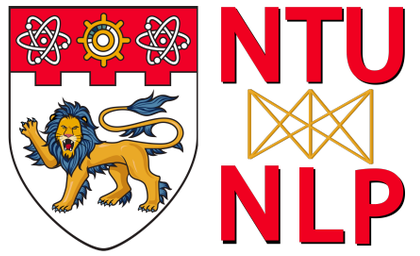
Deep Learning for Natural Language Processing

Shafiq Joty



Lecture 13: Multilingual NLP & Recap

Overview



Models/Algorithms

- Linear models
- Feed-forward Neural Nets (FNN)
- Window-based methods
- Convolutional Nets
- Recurrent Neural Nets
- Recursive Neural Nets

NLP tasks/applications

- Word meaning
- Language modelling
- Sequence tagging
- Sequence encoding
- Parsing
- Hierarchical encoding

Overview

NTU NLP

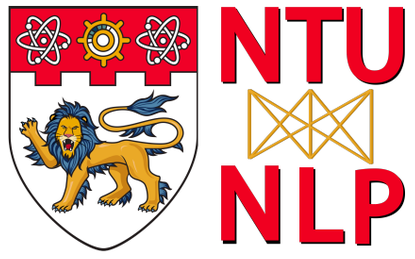
Models/Algorithms

- Seq2Seq
  - + Attention
  - + Subword
- Seq2Seq Varints
- Transformer Seq2Seq

NLP tasks/applications

- Machine Translation
- Summarization
- Parsing
- Dialogue generation

Overview

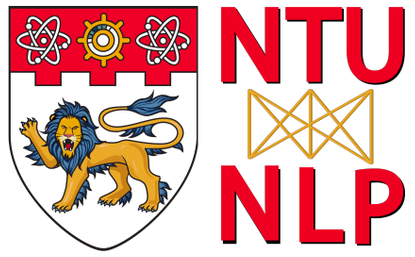


Models/Algorithms

NLP tasks/applications

- Multilingual NLP
- Self-supervised pretraining
- Robust & adversarial NLP

Today: Multilingual NLP



- Why Multilingual NLP?
- Cross-lingual Models

Why Multilingual NLP?

NTU NLP

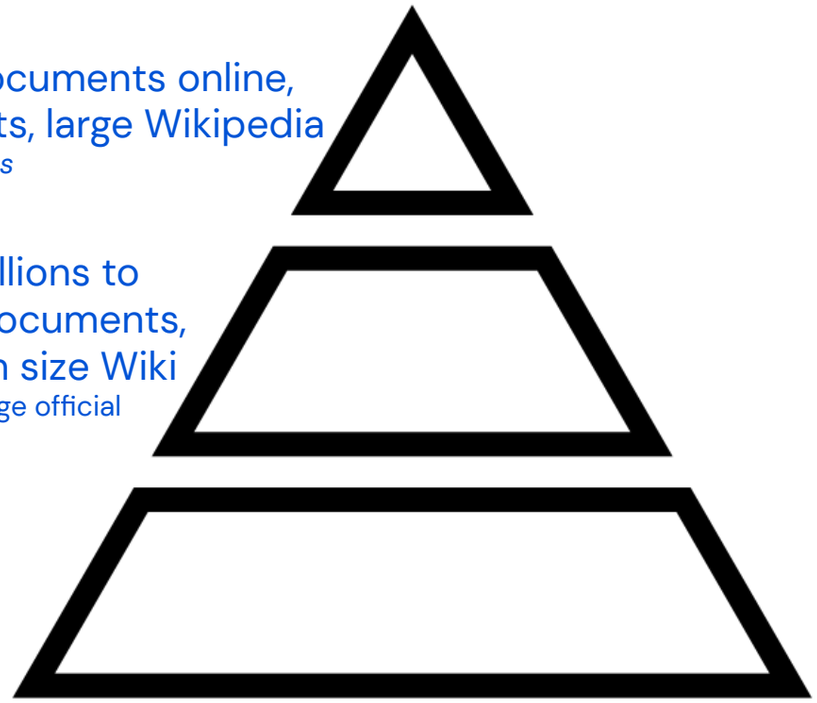
NLP not just for English but for the rest of the world's 7000 languages

Why Multilingual NLP?

High-resource
100s of millions of documents online, large labelled datasets, large Wikipedia
English, major world languages

Medium-resource
Few labelled data, millions to 100,000s of online documents, parallel data, medium size Wiki
Most European languages, large official languages

Low-resource
No labelled data, few data online, small or no Wikipedia
Most languages in the world

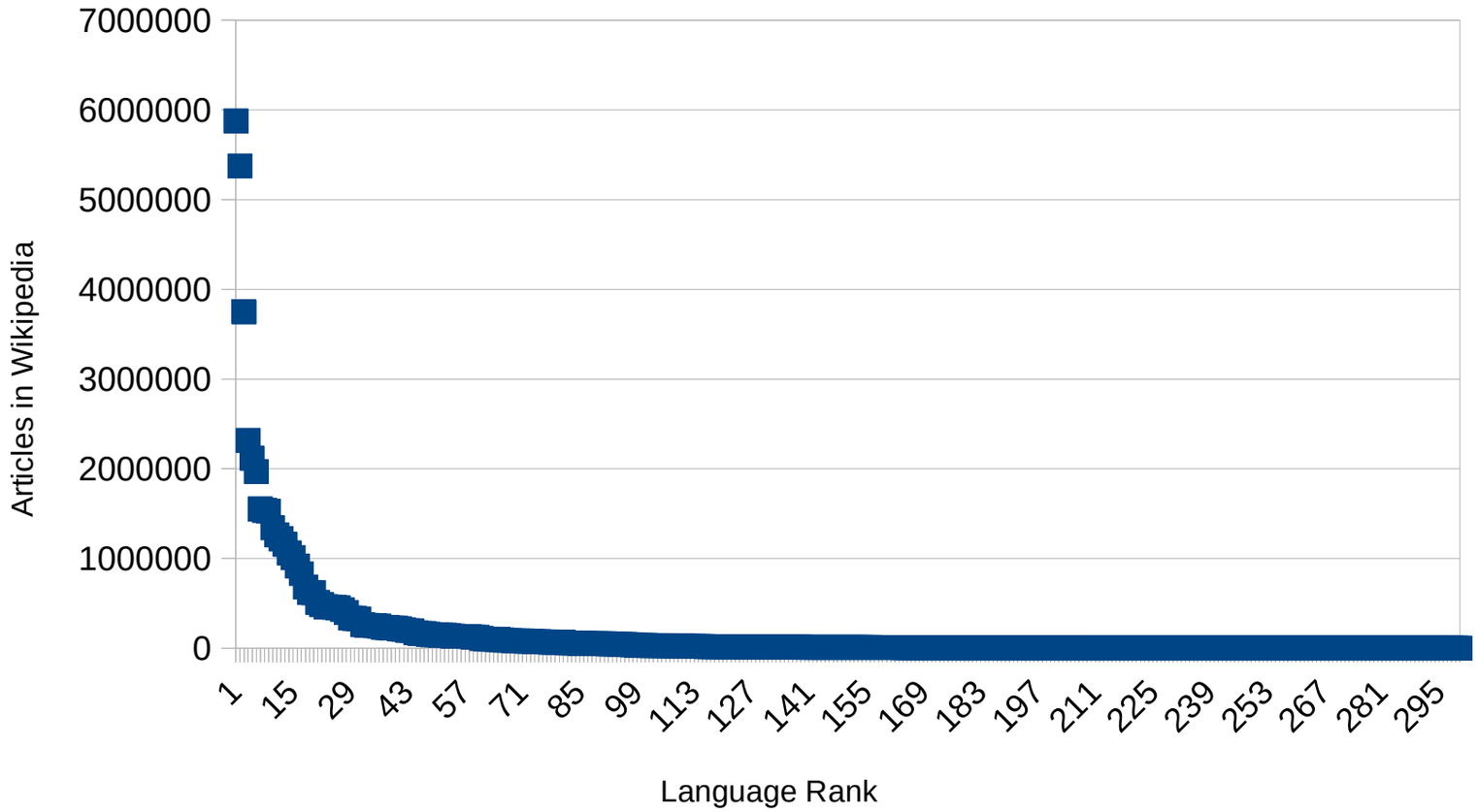


NLP research has mostly focused on these

We need to apply NLP to these

source: (Ruder, 2019)

Why Multilingual NLP?



- There is not enough monolingual data for many languages
- Even less annotated data for tasks (MT, sequence labels)

source: (Ruder, 2019)

Old Paradigm

Old paradigm:

- Language-specific NLP
- **Language-specific** feature computation and preprocessing

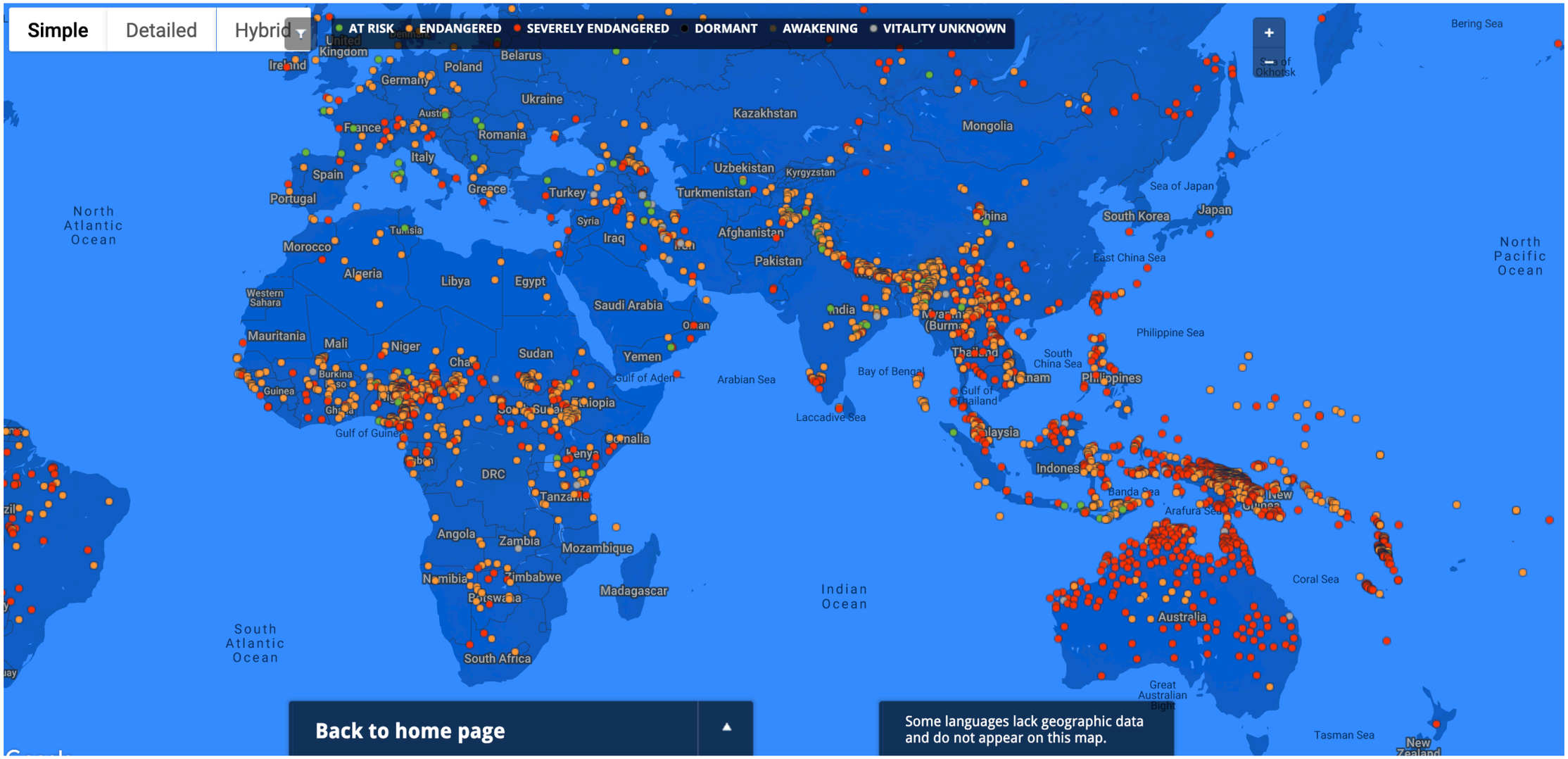
-	*p.currentSense + p.lemma*
-	*p.currentSense + p.pos*
-	*p.currentSense + a.pos*
-	p_{-1}*.FEAT1*
-	*p.FEAT2*
-	p_1*.FEAT3*
-	*p.semrm.semdprel*
-	*p.lm.dprel*
-	*p.form + p.children.dprel.bag*
	$p.lemma_n$ $(n = -1, 0)$
-	*p.lemma + p.lemma*$_1$
-	$p.pos_{-1}$ *+ p.pos*
-	$p.pos_1$
-	*p.pos + p.children.dprel.bag*

- Manual curation and annotation of large-scale resources for thousands of languages in infeasible or prohibitively expensive

source: (Ruder, 2019)

Why Multilingual NLP?

NTU NLP



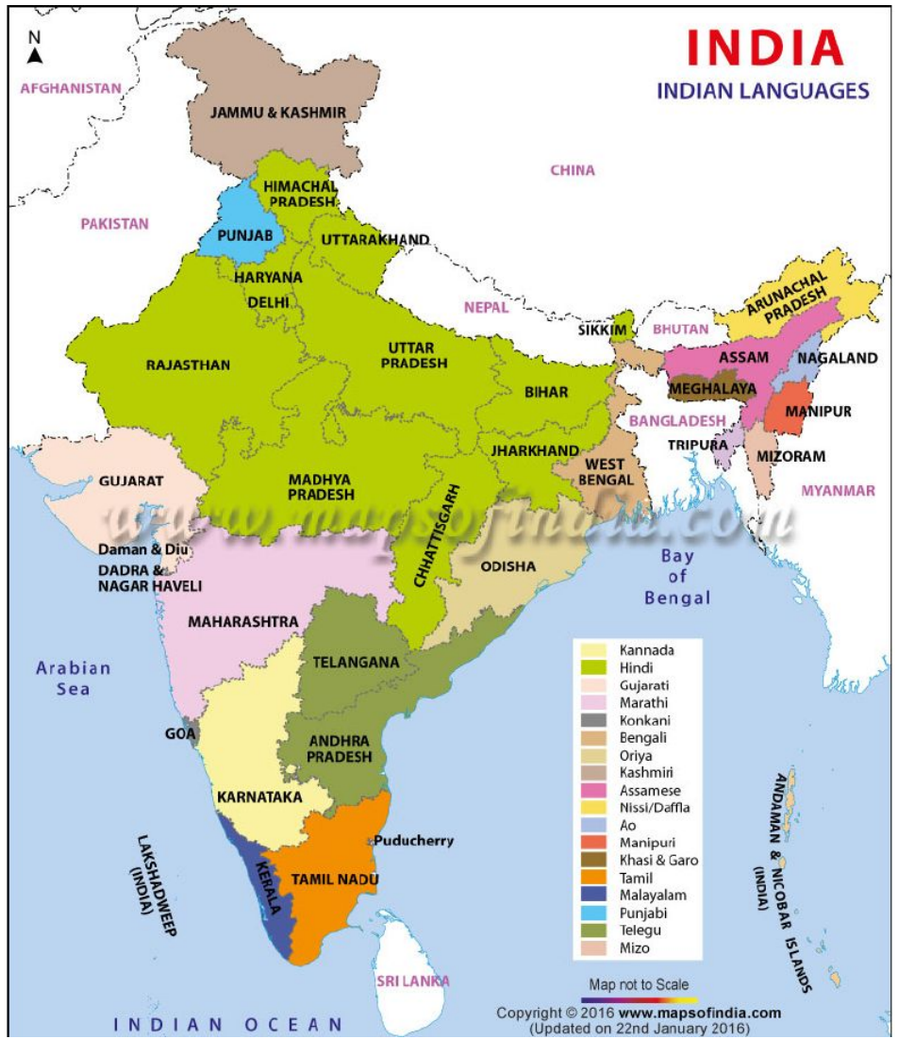
http://endangeredlanguages.com/

Low Resource Languages

NTU NLP

There are about 460 languages in India.

1.38 billion people



source: CS11-737

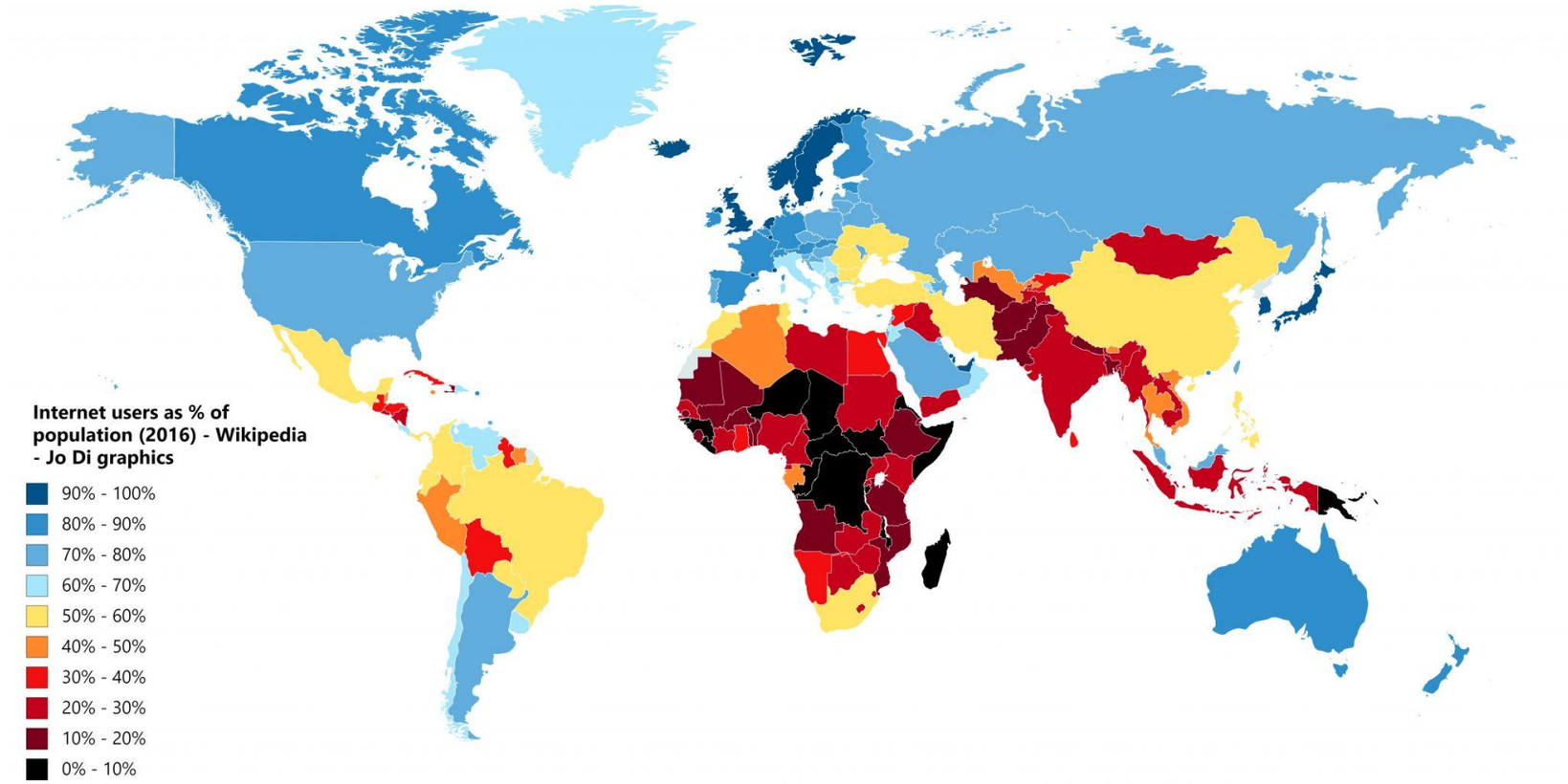
Low Resource Languages

NTU NLP

Africa is a continent with a very high linguistic diversity:
there are an estimated 1.5-2K African languages from 6 language families.
1.33 billion people

source: CS11-737

Low Resource Languages

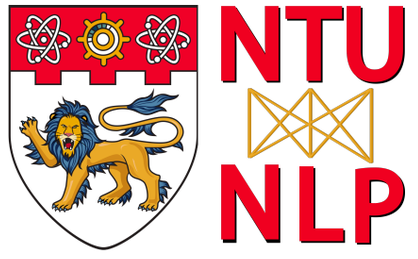


https://jodi.graphics/2018/05/11/internet-users-as-of-population/

40% of world's population: South Asia - 1.75 billion, Africa - 1.3 billion, etc.

source: CS11-737

How to define similarity across languages



Word overlap and sub-word overlap

- Russian – Русский
- Ukraininan – Українська
- Chinese – 中文
- Korean – 한국어
- Vietnamese – Tiếng Việt
- Georgian – ქართული
- Japanese – 日本人
- Turkish – Türk
- Hebrew – עִבְרִית
- Arabic – عربی
- Hindi – हिन्दी
- Xhosa – isiXhosa

Areal similarity

Demographic similarity

source: CS11-737

How to define similarity across languages

Genealogical similarity

1. Niger–Congo (1,542 languages) (21.7%)
2. Austronesian (1,257 languages) (17.7%)
3. Trans–New Guinea (482 languages) (6.8%)
4. Sino-Tibetan (455 languages) (6.4%)
5. Indo-European (448 languages) (6.3%)
6. Australian [dubious] (381 languages) (5.4%)
7. Afro-Asiatic (377 languages) (5.3%)
8. Nilo-Saharan [dubious] (206 languages) (2.9%)
9. Oto-Manguean (178 languages) (2.5%)
10. Austroasiatic (167 languages) (2.3%)
11. Tai–Kadai (91 languages) (1.3%)
12. Dravidian (86 languages) (1.2%)
13. Tupian (76 languages) (1.1%)

www.ethnologue.com

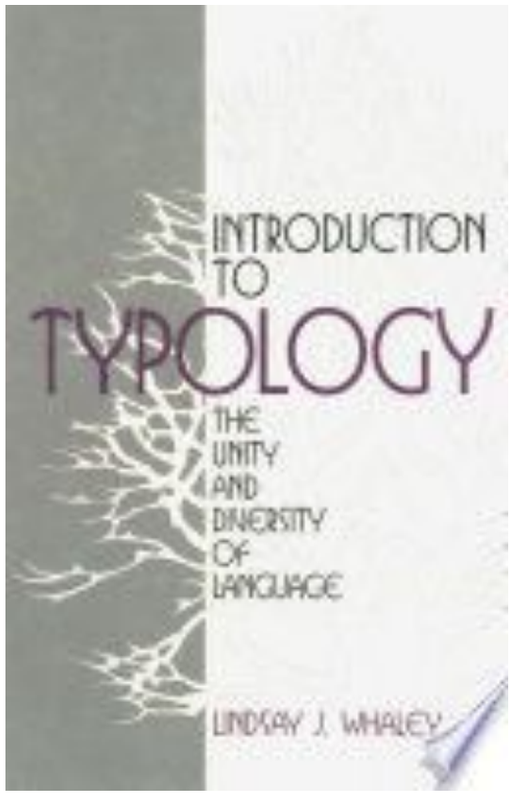
source: CS11-737

How to define similarity across languages

Typological similarity

- Linguistic typology: classification of languages according to their functional and structural properties
  - explains common properties across languages
  - explains structural diversity across languages

"The classification of languages or components of languages based on shared formal characteristics."



source: CS11-737

How to define similarity across languages

NTU NLP

Typological similarity

- 2,676 languages, 192 attributes

ID#	Feature Name	Category	Feature Values
1	Consonant Inventories	Phonology (19)	{1:Large, 2:Small, 3:Moderately Small, 4:Moderately Large, 5:Average}
23	Locus of Marking in the Clause	Morphology (10)	{1:Head, 2:None, 3:Dependent, 4:Double, 5:Other}
30	Number of Genders	Nominal Categories (28)	{1:Three, 2:None, 3:Two, 4:Four, 5:Five or More}
58	Obligatory Possessive Inflection	Nominal Syntax (7)	{1:Absent, 2:Exists}
66	The Perfect	Verbal Categories (16)	{1:None, 2:Other, 3:From 'finish' or 'already', 4:From Possessive}
81	Order of Subject, Object and Verb	Word Order (17)	{1:SVO, 2:SOV, 3:No Dominant Order, 4:VSO, 5:VOS, 6:OVS, 7:OSV}
121	Comparative Constructions	Simple Clauses (24)	{1:Conjoined, 2:Locational, 3:Particle, 4:Exceed}
125	Purpose Clauses	Complex Sentences (7)	{1:Balanced/deranked, 2:Deranked, 3:Balanced}
138	Tea	Lexicon (10)	{1:Other, 2:Derived from Sinitic 'cha', 3:Derived from Chinese 'te'}
140	Question Particles in Sign Languages	Sign Languages (2)	{1:None, 2:One, 3:More than one}
142	Para-Linguistic Usages of Clicks	Other (2)	{1:Logical meanings, 2:Affective meanings, 3:Other or none}

Example from Georgi, Xia and Lewis (2010)

Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013.
The World Atlas of Language Structures Online.
Leipzig: Max Planck Institute for Evolutionary Anthropology.

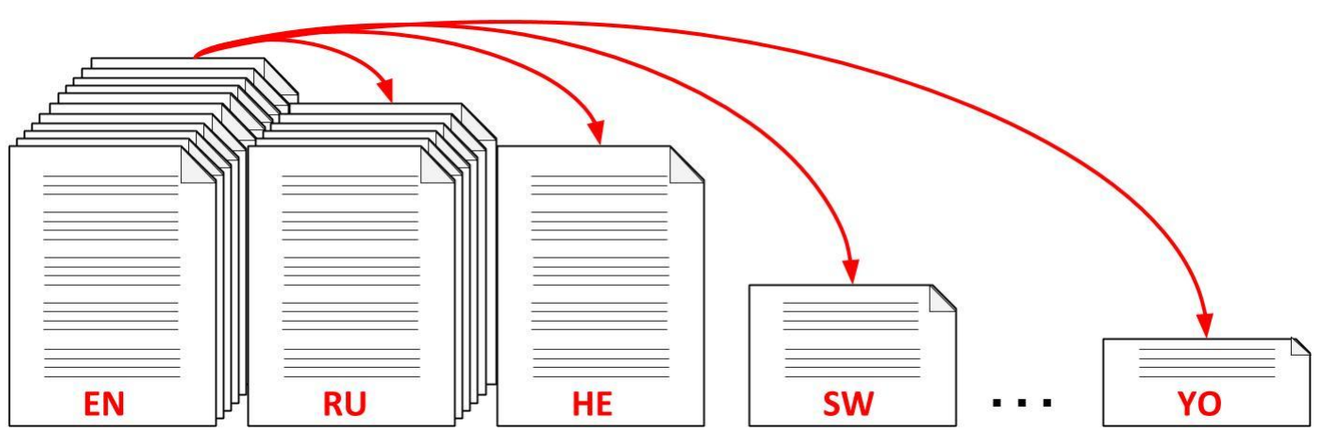
source: CS11-737

Approaches to Low-resource NLP

NTU NLP

1. Cross-lingual **transfer** learning

- Transfer of resources and models from resource-rich source to resource-poor target languages

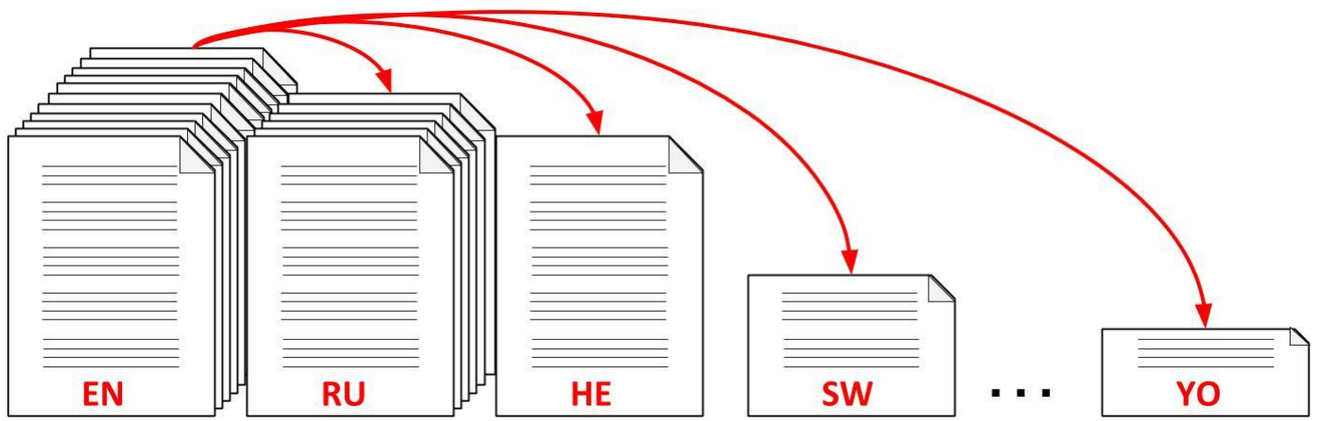


source: CS11-737

Approaches to Low-resource NLP

NTU NLP

1. Cross-lingual **transfer** learning



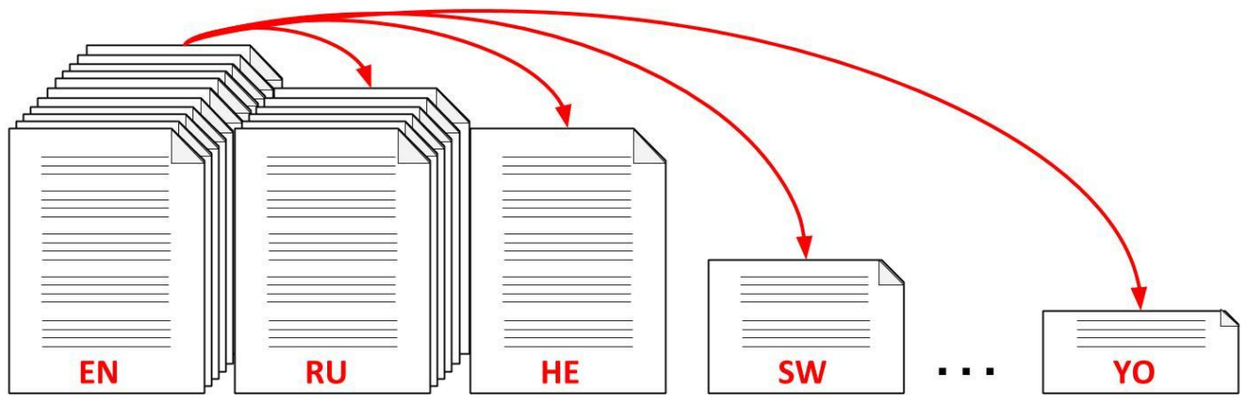
- **Instance** level transfer: transfer annotations (e.g., NER, POS tags) via cross-lingual bridges (e.g., word or phrase alignments)
- **Model** level transfer: train a model in a resource-rich language and adapt (e.g. fine-tune) it in a resource-poor language

source: CS11-737

Approaches to Low-resource NLP

NTU NLP

1. Cross-lingual **transfer** learning

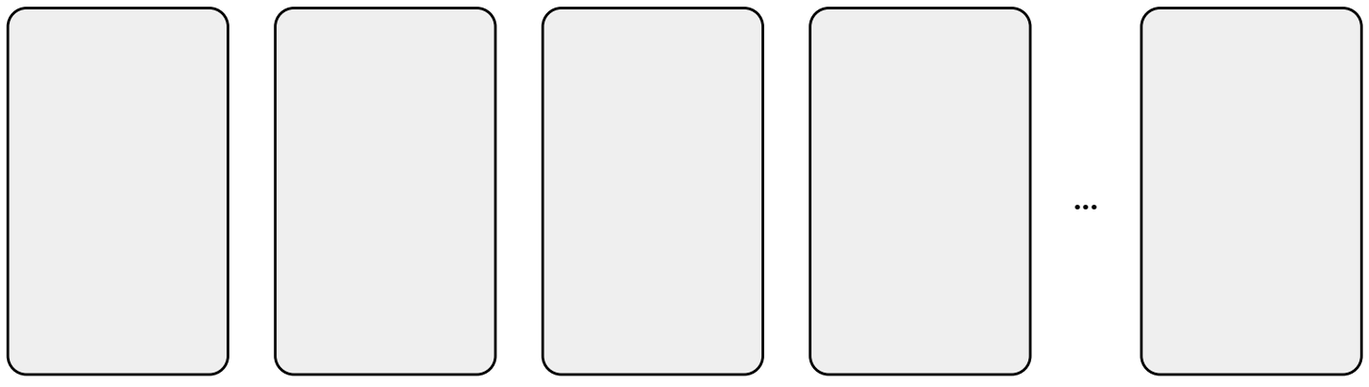


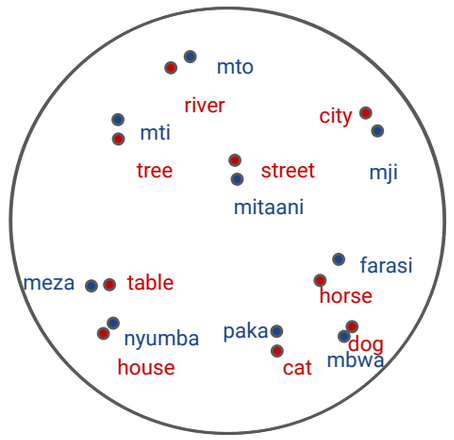
- **Model** level transfer: train a model in a resource-rich language and adapt (e.g. fine-tune) it in a resource-poor language
  - Zero-shot learning – train a model in one domains and assume it generalizes more or less out-of-the-box in a low-resource domain
  - Few shot learning – train a model in one domain and use only few examples from a low-resource domain to adapt it

source: CS11-737

Cross-lingual Model Learning Paradigm

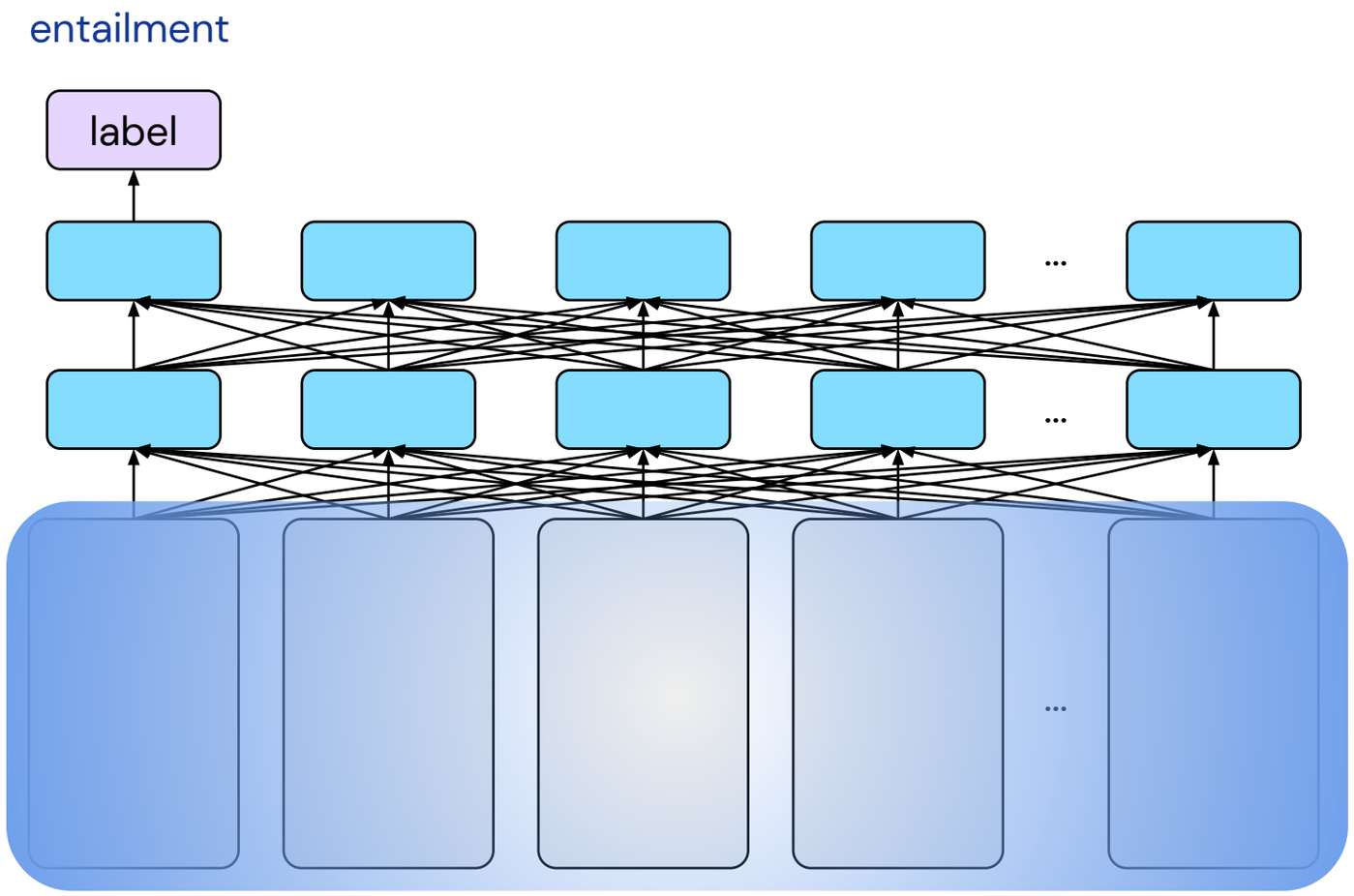
- Step 1: Learn cross-lingual representations





source: (Ruder, 2019)

Cross-lingual Model Learning Paradigm

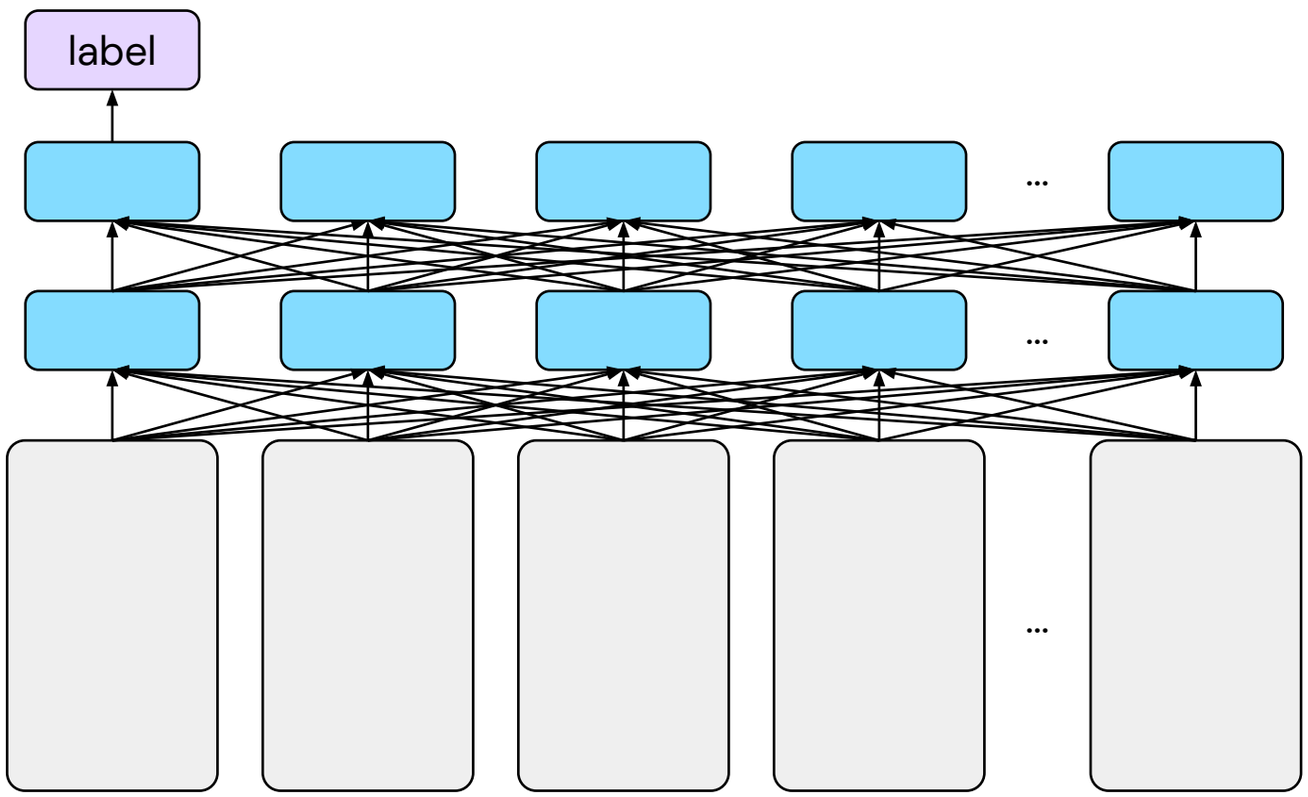


- Step 1: Learn cross-lingual representations
- Step 2: Fine-tune model on labelled data of a high-resource language (mostly English)

  cross-lingual parameters are often kept fixed

source: (Ruder, 2019)

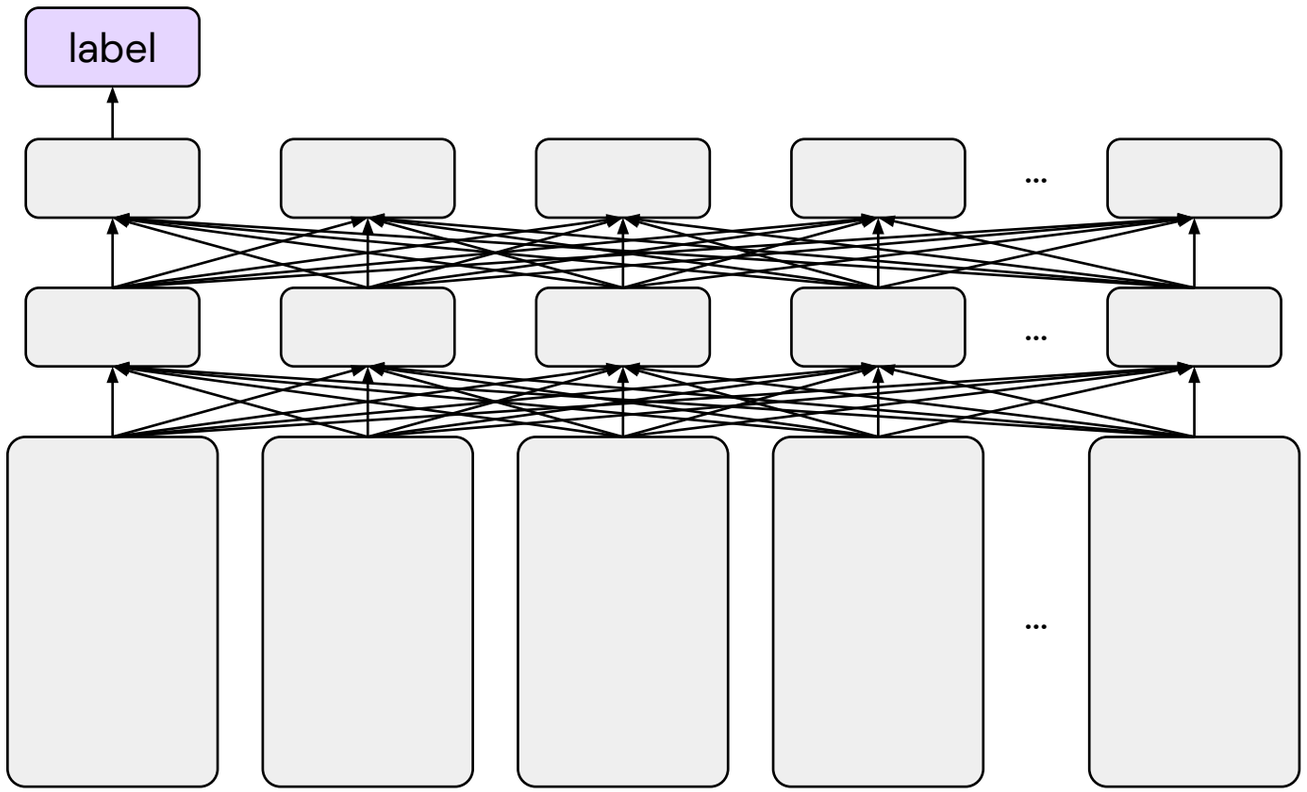
Cross-lingual Model Learning Paradigm



- Step 1: Learn cross-lingual representations
- Step 2: Fine-tune model on labelled data of a high-resource language
- Step 3: Zero-shot transfer to a low-resource language

source: (Ruder, 2019)

Cross-lingual Learning Paradigm



- Cross-lingual representations can be learned at different levels: from the word level to the sentence level

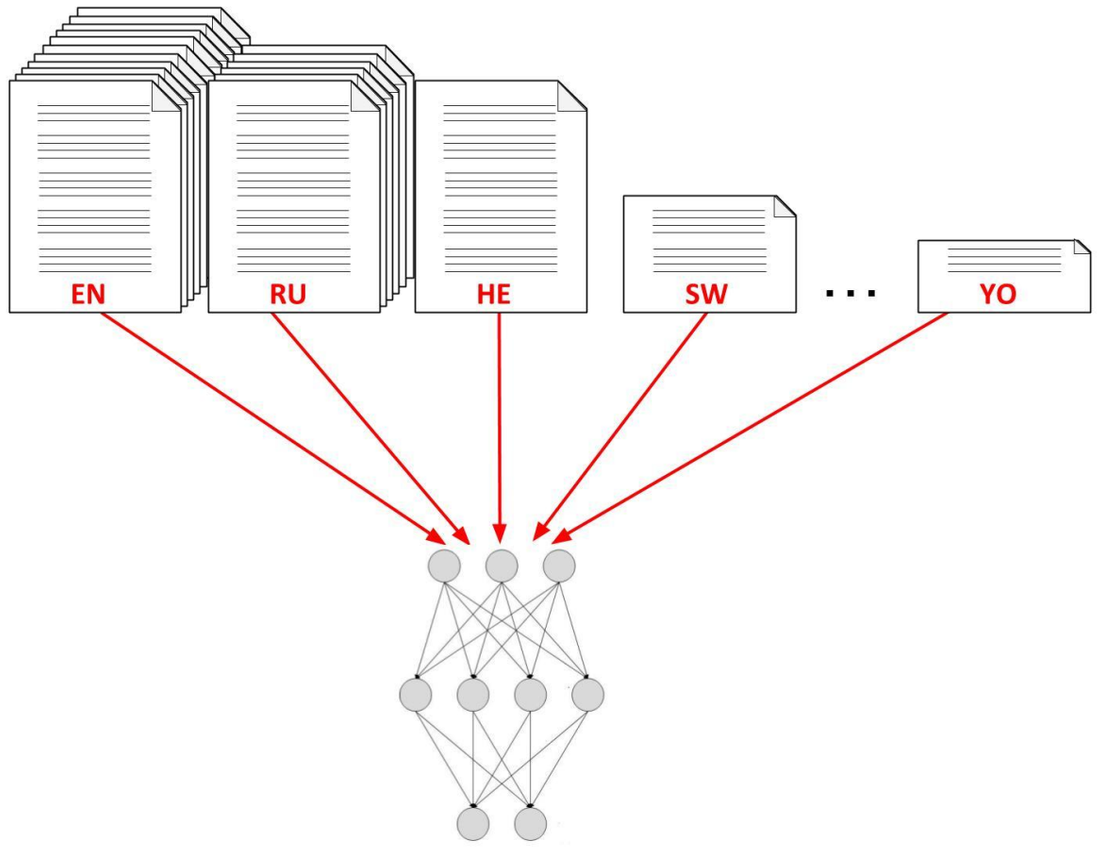
E.g., multi-lingual BERT

source: (Ruder, 2019)

Approaches to Low-resource NLP

NTU NLP

2. Joint Multi-lingual learning

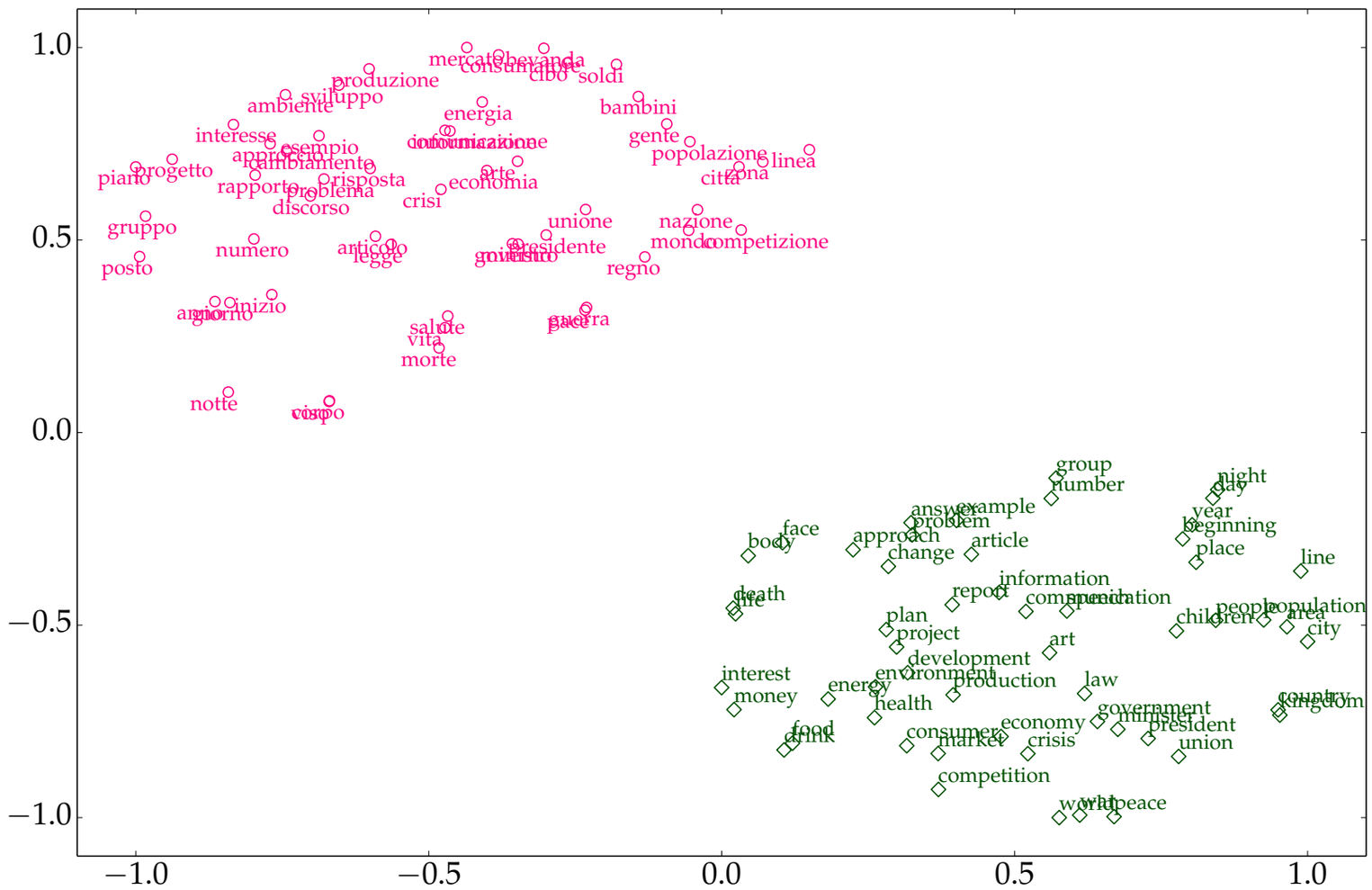


Train a single model on a mix of datasets in all languages, to enable data and parameter sharing where possible

source: CS11-737

Cross-lingual Word Embeddings

Cross-lingual representations of words in a joint embedding space

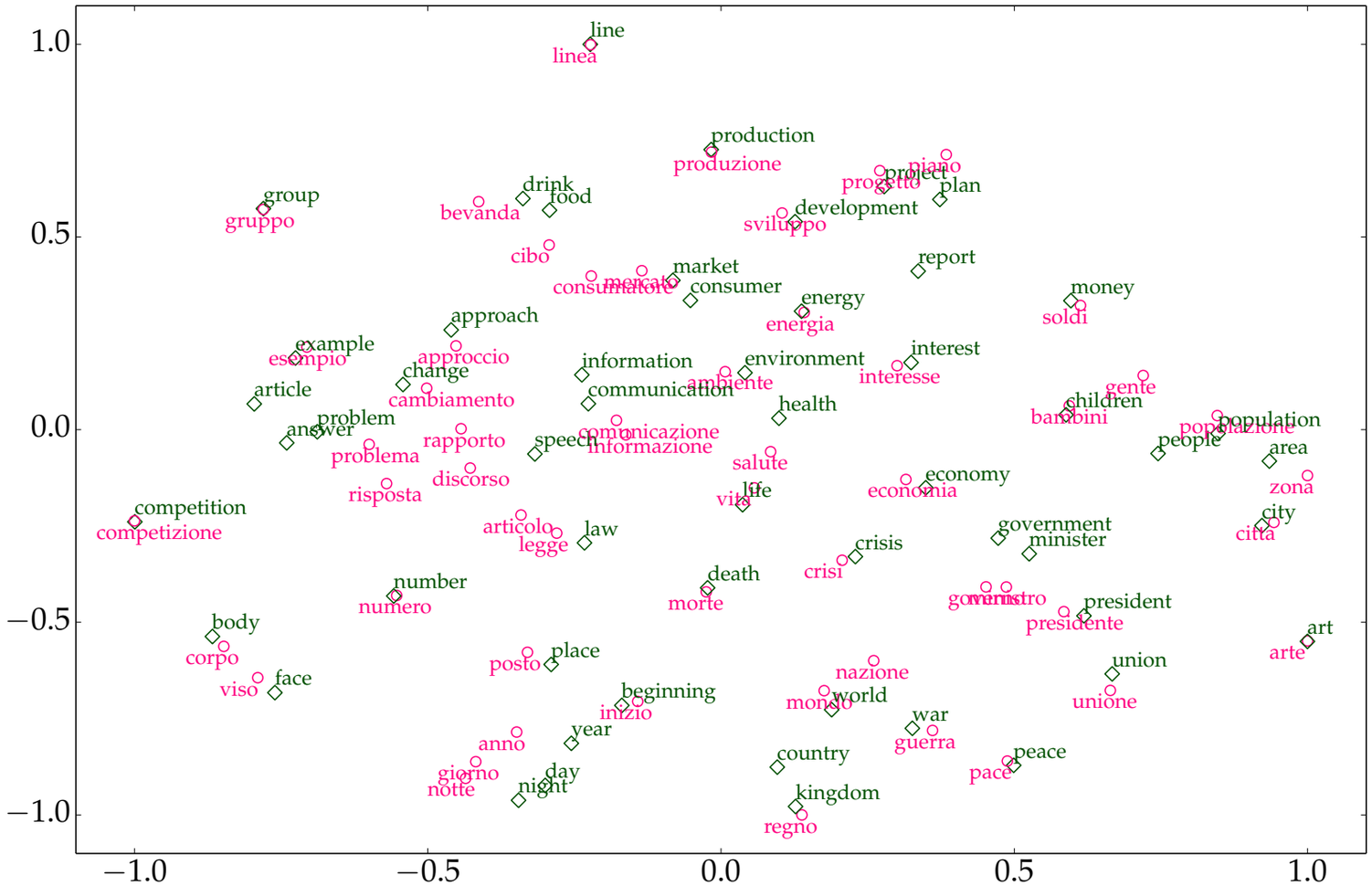


Two monolingual word embeddings (Ruder, 2019)

Cross-lingual Word Embeddings

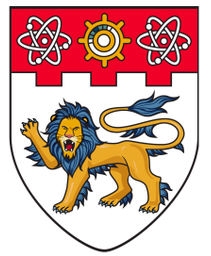
NTU NLP

Cross-lingual representations of words in a joint embedding space



Cross-lingual word embeddings (Ruder et al, 2019)

Cross-lingual Word Embeddings



NTU NLP

Why do we need cross-lingual embeddings?

- Compare word vectors (meaning of words) across languages.
  - Key to bilingual lexicon induction and machine translation
- Transfer of knowledge (semantic & syntactic) between languages.
  - Cross-lingual NER, QA, Sentiment Analysis

Cross-lingual Word Embeddings

What kind of data is used to learn cross-lingual embeddings?

	Parallel	Comparable
Word	Dictionaries	Images
Sentence	Translations	Captions
Document	-	Wikipedia

(a) Word, par. (b) Word, comp. (c) Sentence, par. (d) Sentence, comp. (e) Doc., comp.

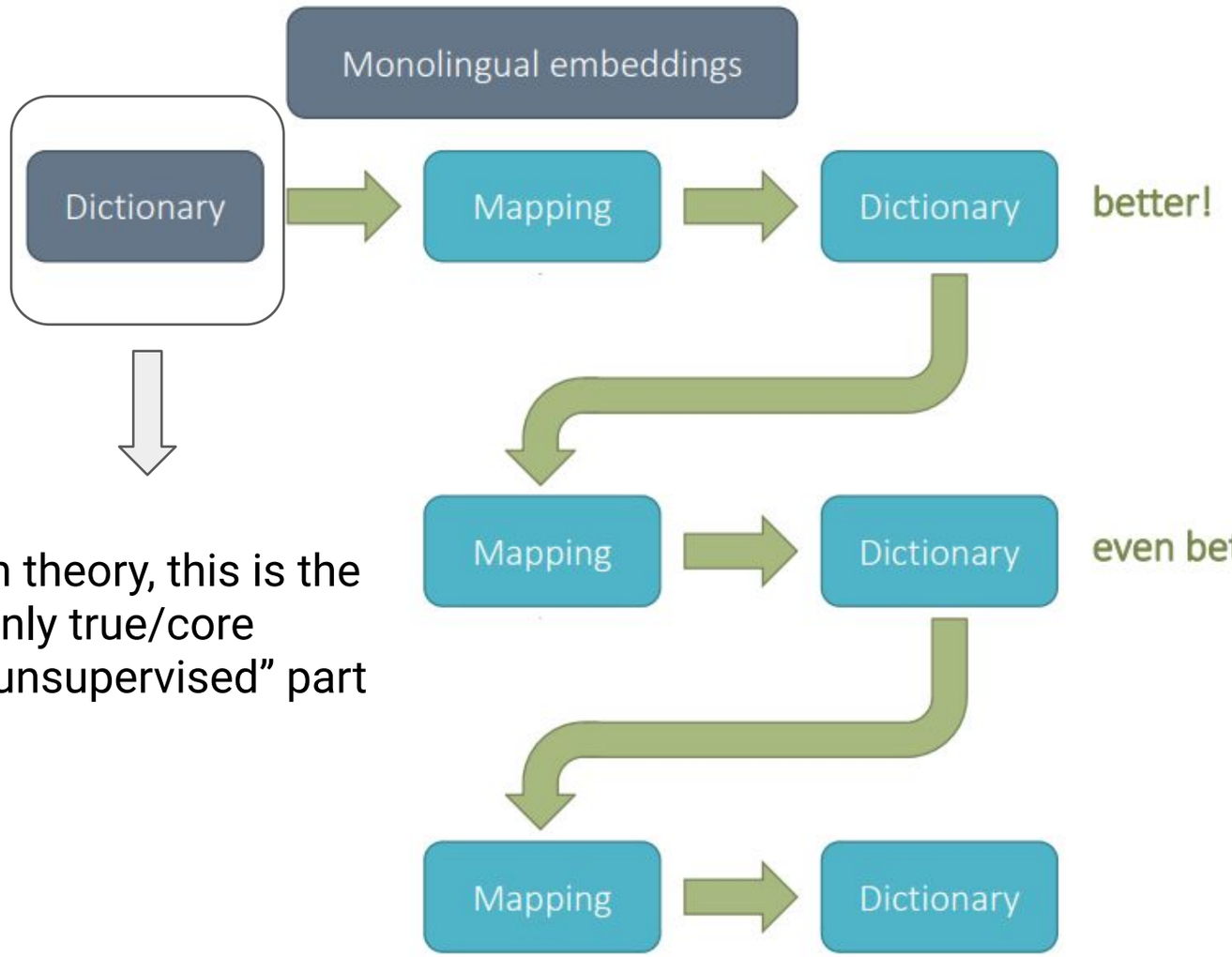
source: (Ruder, 2019)

Cross-lingual Embedding Learning Methods

	Parallel	Comparable
Word —Mapping	Mikolov et al. (2013b) Faruqui and Dyer (2014) Lazaridou et al. (2015) Dinu et al. (2015) Xing et al. (2015) Lu et al. (2015) Vulić and Korhonen (2016) Ammar et al. (2016b) Zhang et al. (2016b, 2017ab) Artexte et al. (2016, 2017, 2018ab) Smith et al. (2017) Hauer et al. (2017) Mrkšić et al. (2017b) Conneau et al. (2018a) Joulin et al. (2018) Alvarez-Melis and Jaakkola (2018) Ruder et al. (2018) Glavaš et al. (2019)	Bergsma and Van Durme (2011) Kiela et al. (2015) Vulić et al. (2016)
Word —Pseudo-bilingual	Xiao and Guo (2014) Gouws and Søgaard (2015) Duong et al. (2016) Adams et al. (2017)	Duong et al. (2015)
Word —Joint	Klementiev et al. (2012) Kočiský et al. (2014)	
Sentence —Matrix factorization	Zou et al. (2013) Shi et al. (2015) Gardner et al. (2015) Guo et al. (2015) Vyas and Carpuat (2016)	
Sentence —Compositional	Hermann and Blunsom (2013, 2014) Soyer et al. (2015)	
Sentence —Autoencoder	Lauly et al. (2013) Chandar et al. (2014)	
Sentence —Skip-gram	Gouws et al. (2015) Luong et al. (2015) Coulmance et al. (2015) Pham et al. (2015)	
Sentence —Other	Levy et al. (2017) Rajendran et al. (2016)	Calixto et al. (2017) Gella et al. (2017)
Document		Vulić and Moens (2013a, 2014, 2016) Søgaard et al. (2015) Mogadala and Rettinger (2016)

Check also more recent work including ours (see suggested readings)

Supervised/Unsupervised Mapping based methods

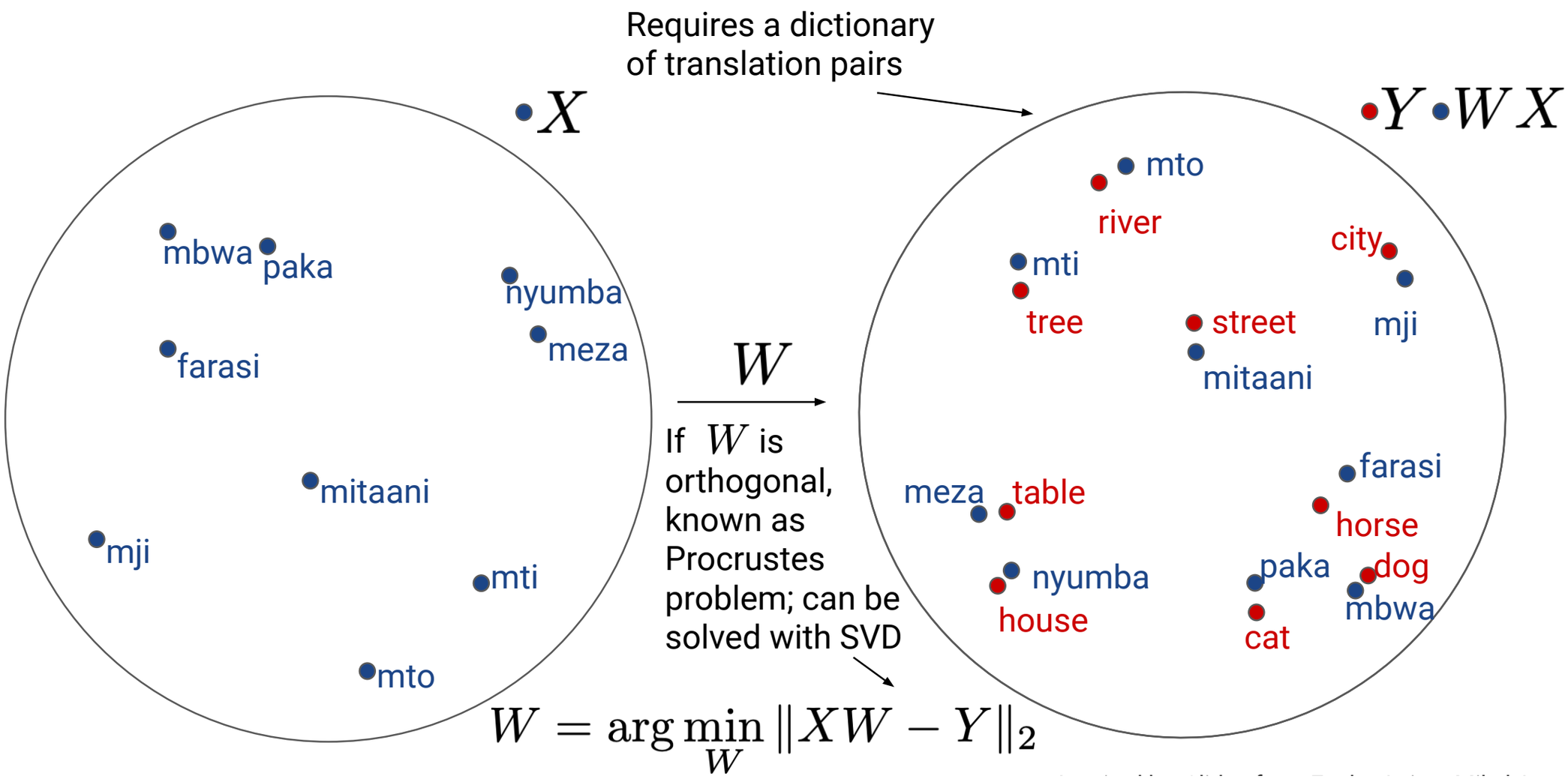


In theory, this is the only true/core “unsupervised” part

- Most current approaches use self-learning
- Particularly useful with limited supervision
- The seed dictionary improves over time
- Initial seed dictionary is either given (supervised) or learned (unsupervised)
- Many ways to learn seed dictionary
- One common approach: adversarial mapping

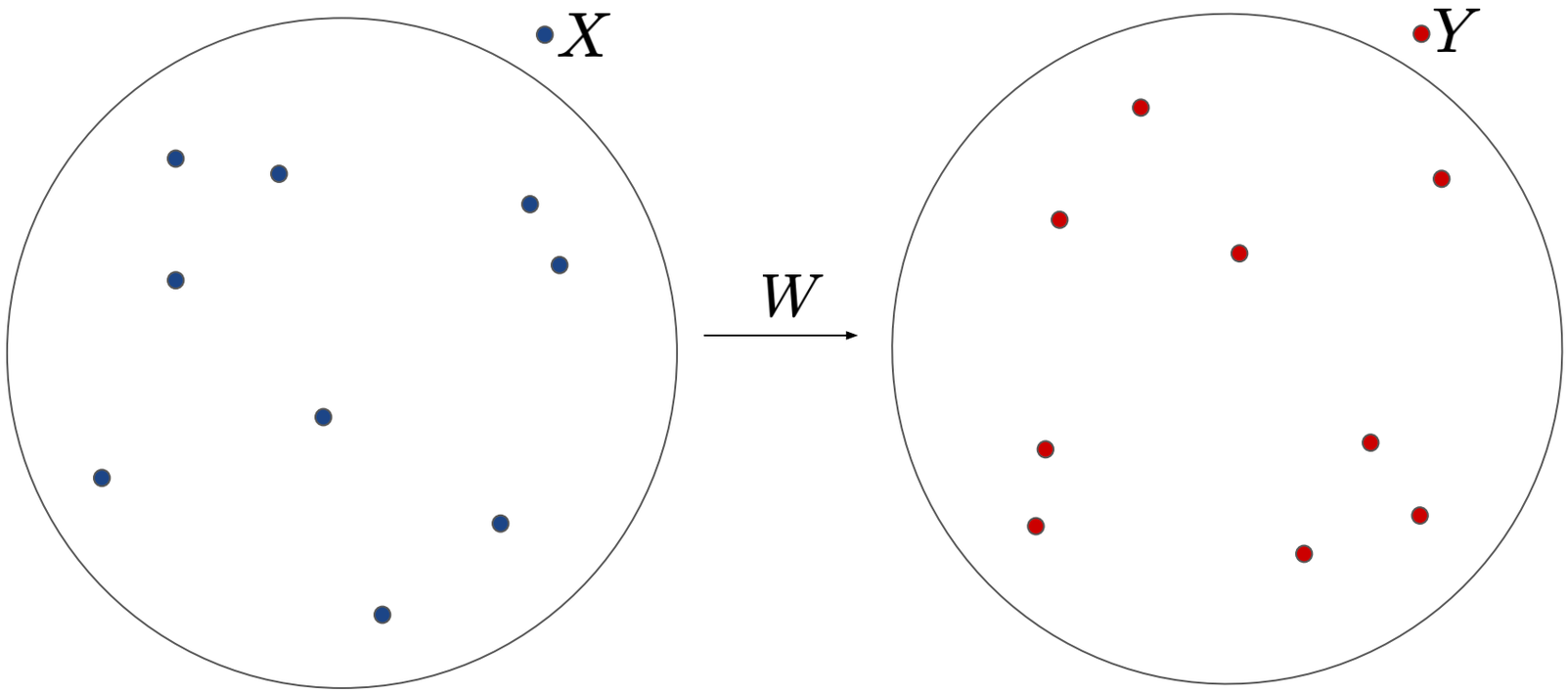
source: (Ruder, 2019)

Supervised Mapping based methods



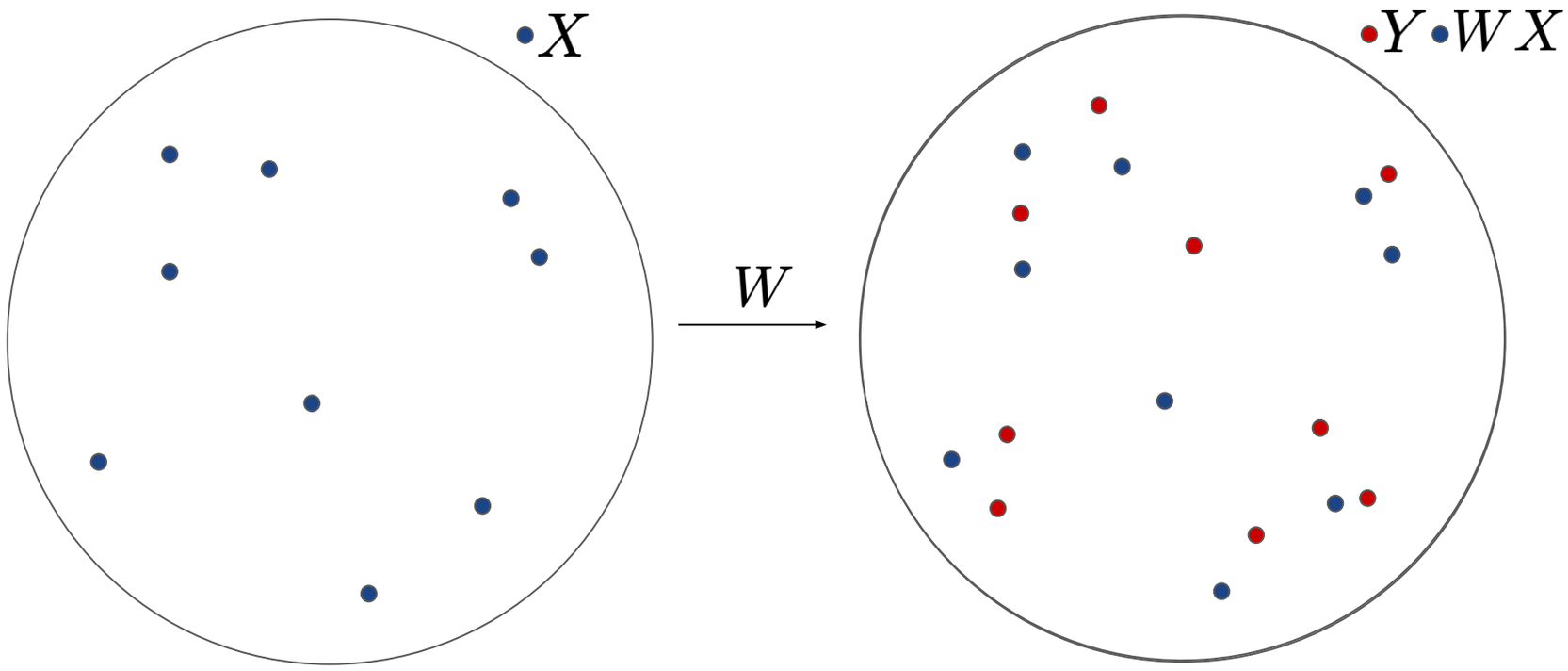
source: (Ruder, 2019)

Unsupervised Mapping



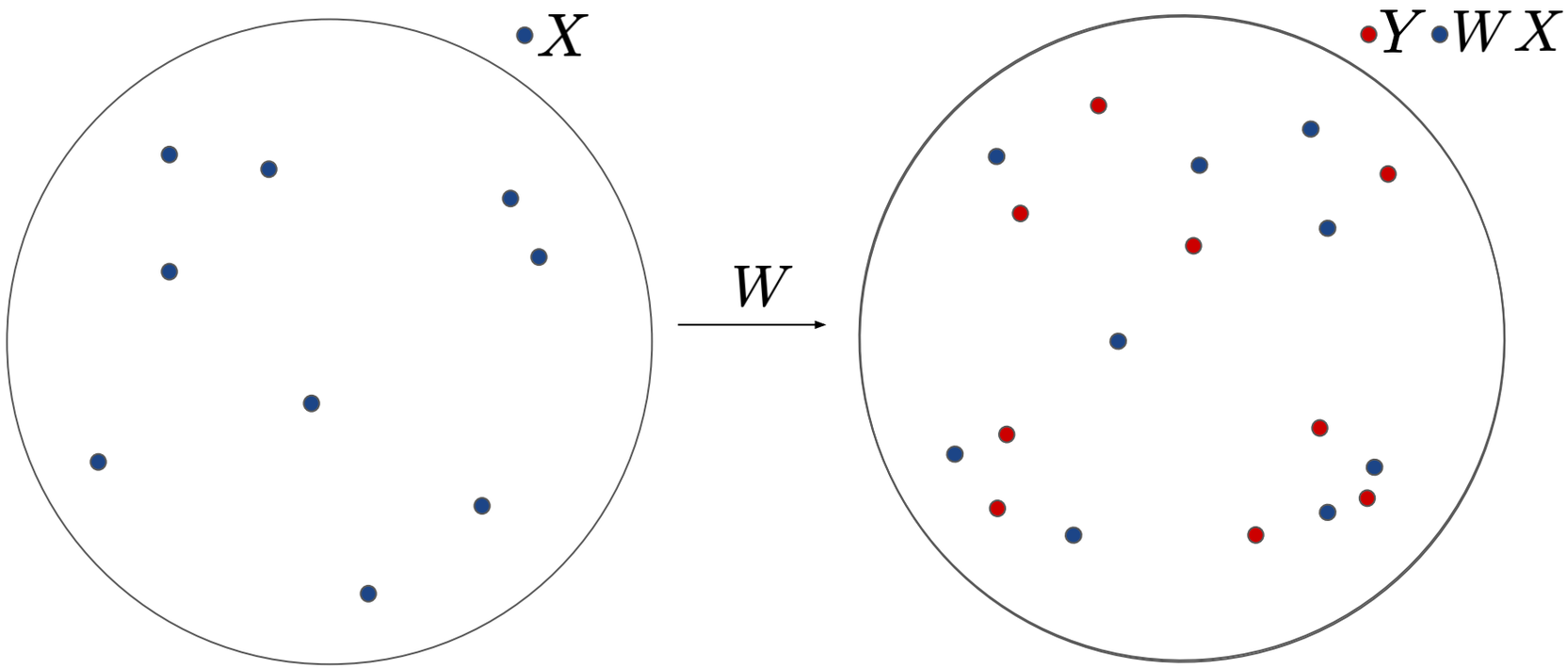
source: (Ruder, 2019)

Unsupervised Mapping



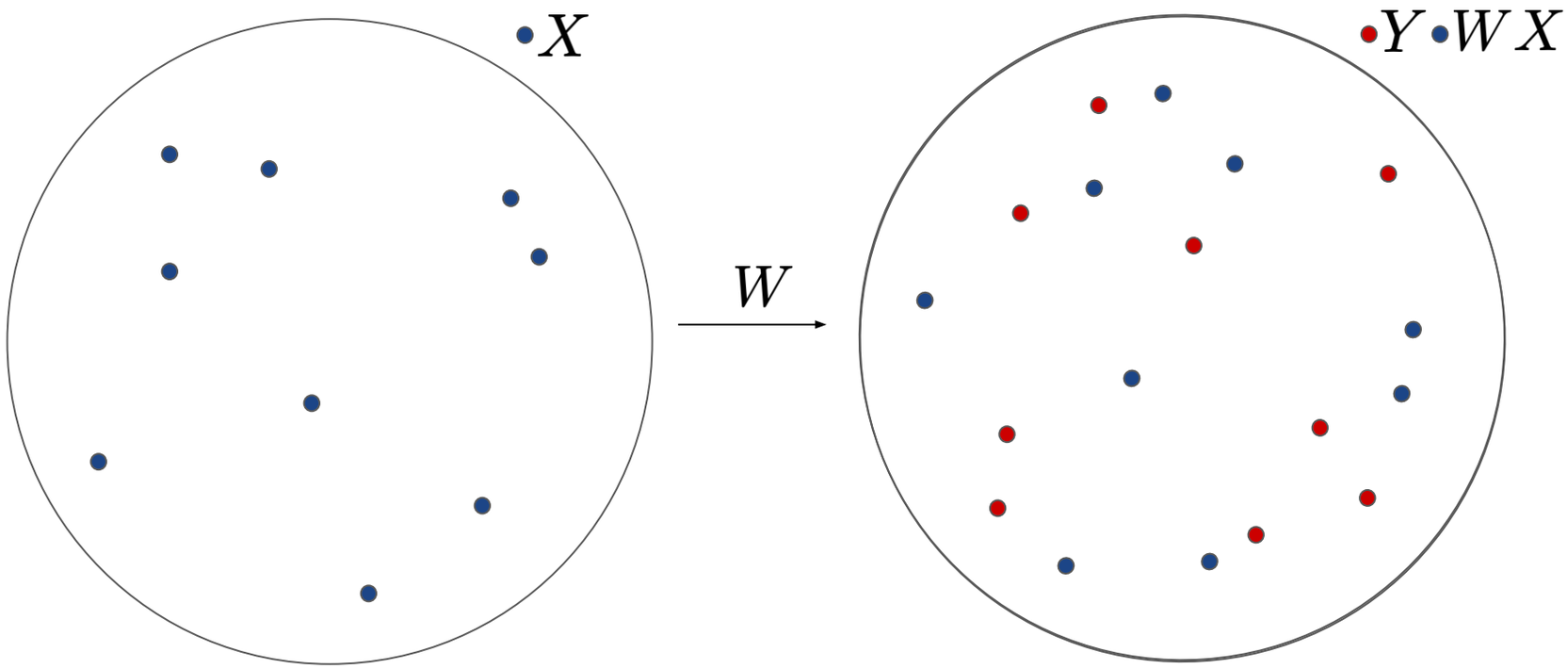
source: (Ruder, 2019)

Unsupervised Mapping



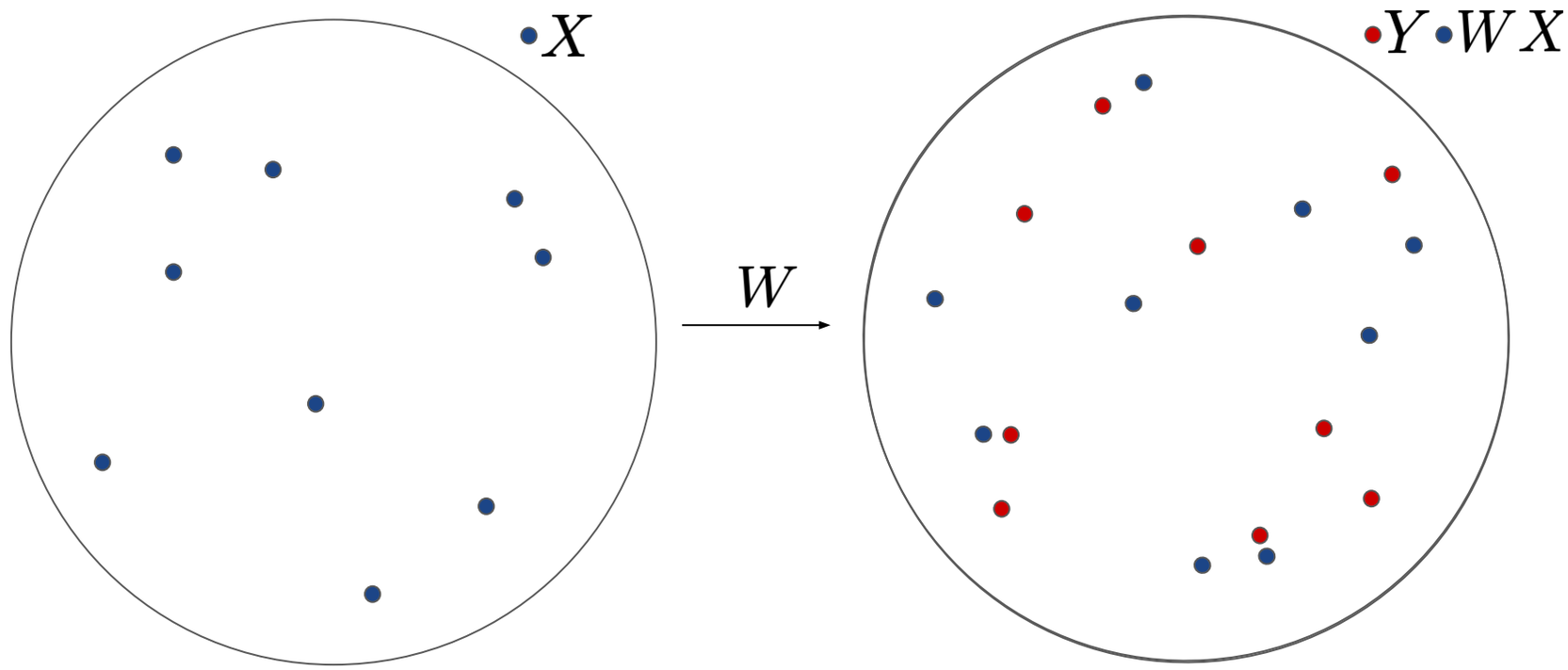
source: (Ruder, 2019)

Unsupervised Mapping



source: (Ruder, 2019)

Unsupervised Mapping

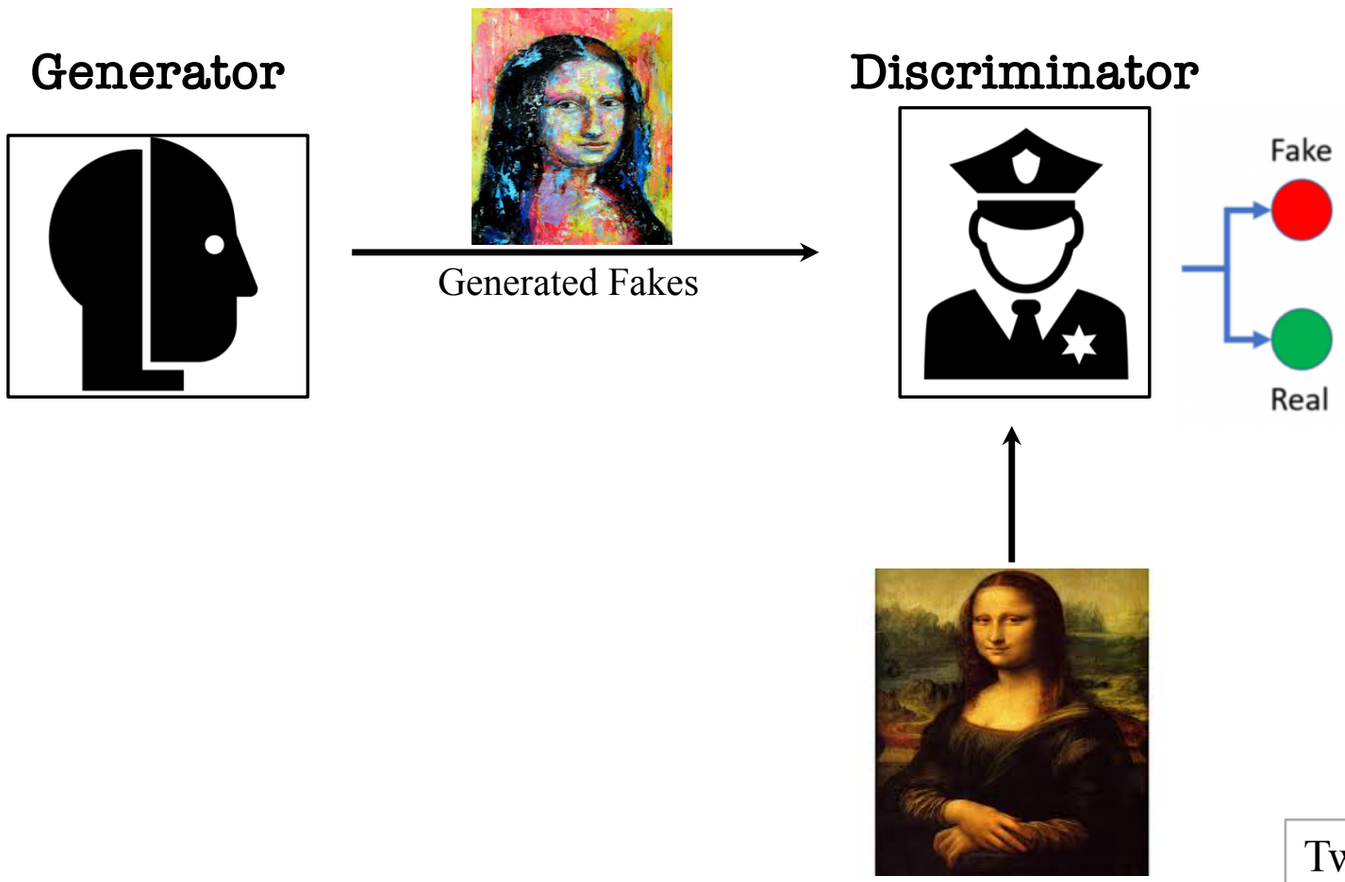


source: (Ruder, 2019)

Generative Adversarial Nets

NTU NLP

[Goodfellow et al., 2014]

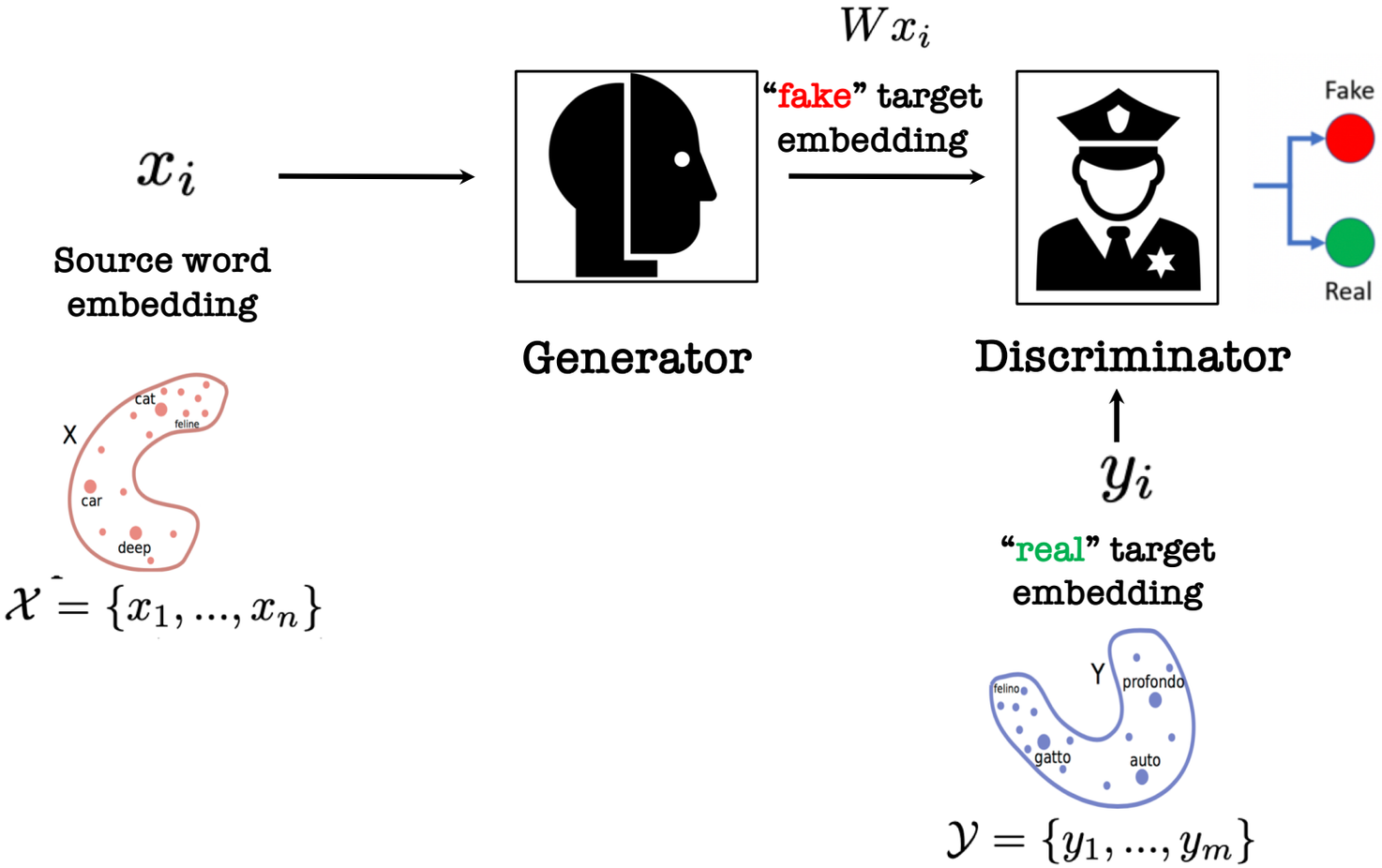


Two player game, where
Generator: generates fake examples
Discriminator: differentiate between real and fake

Applications of GANs in NLP

NTU NLP

Unsupervised Word Translation



- Generator needs to match distribution of target language in order to fool discriminator consistently.
- *Hypothesis*: Best way to do this is to align words with their translations.

Generator: projects source word embedding x_i into the target language using linear mapper W ⇒ Wx_i

Discriminator: differentiate between "fake" projected embeddings and "true" target language embeddings y_i

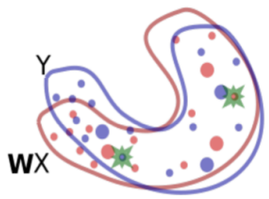
Conneau et al. (2018) Popularly known as **MUSE**

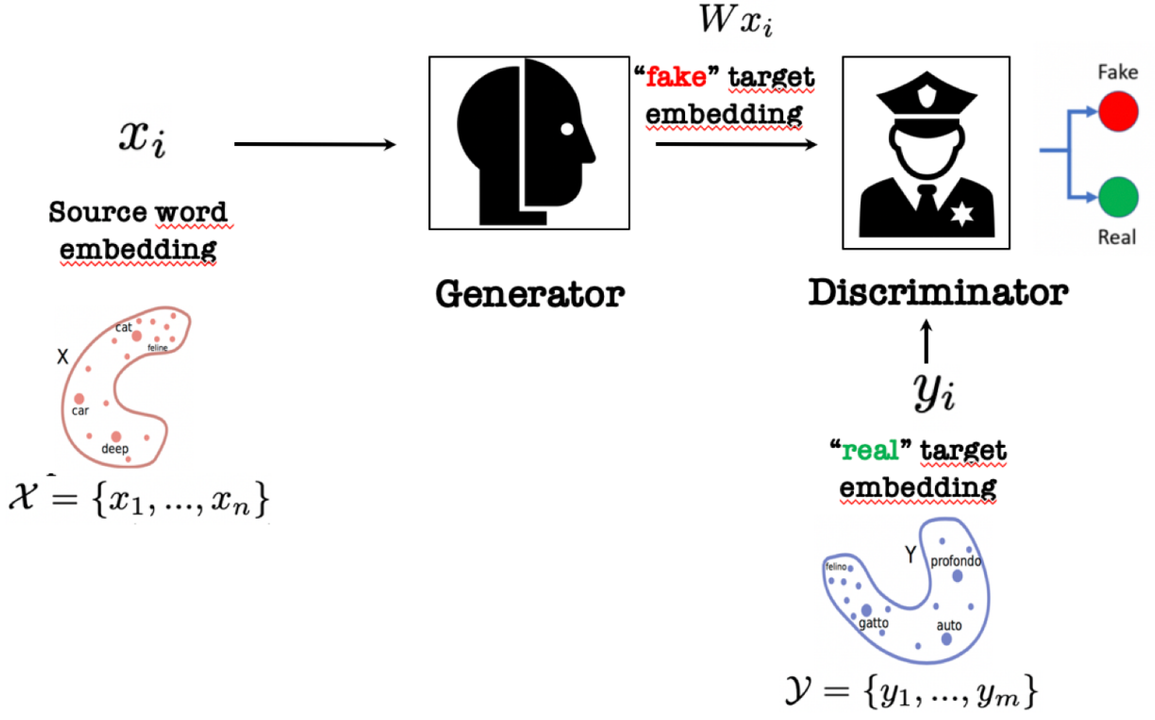
Source: [NAACL Tutorial]

Applications of GANs in NLP

NTU NLP

Unsupervised Word Translation





Objective

$\theta_D \Rightarrow$ Discriminator parameters

$P_{\theta_D}(\text{source} = 1|z) \Rightarrow$ Probability that z is the mapping of a source embedding

Discriminator Loss

An embedding in X — An embedding in Y

$$\mathcal{L}_D(\theta_D|W) = -\frac{1}{n}\sum_{i=1}^{n} \log P_{\theta_D}(\text{source} = 1|Wx_i) - \frac{1}{m}\sum_{i=1}^{m} \log P_{\theta_D}(\text{source} = 0|y_i)$$

Maximize probability of predicting correct source

Generator/Mapper Loss

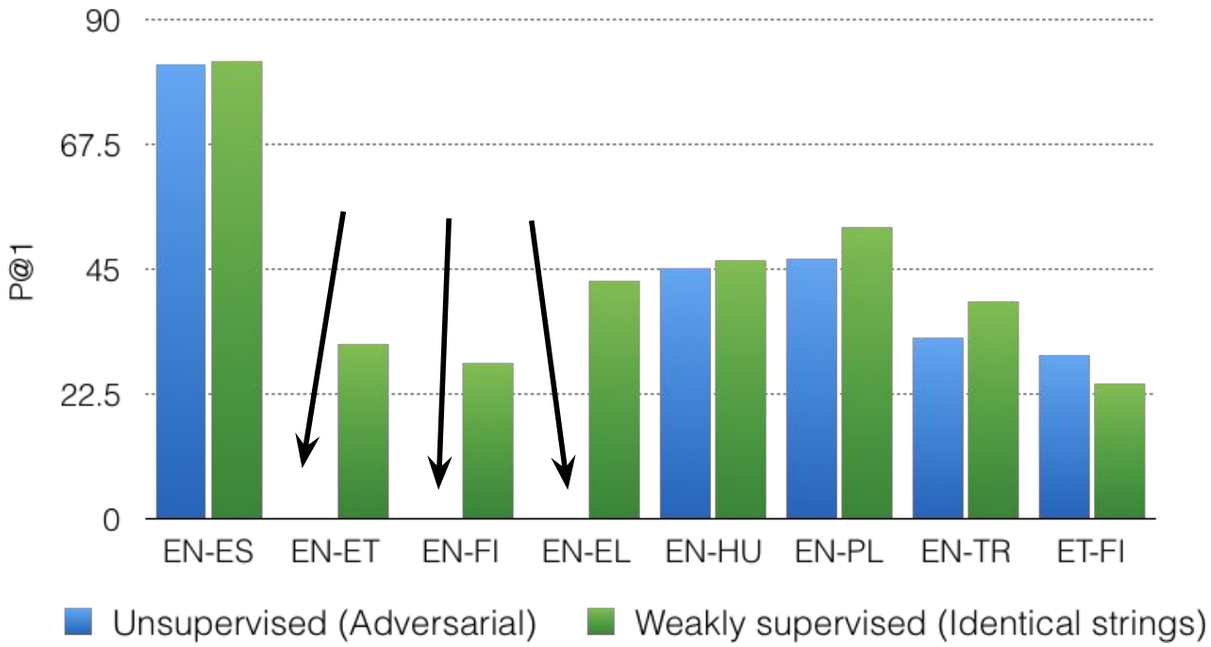
$$\mathcal{L}_W(W|\theta_D) = -\frac{1}{n}\sum_{i=1}^{n} \log P_{\theta_D}(\text{source} = 0|Wx_i) - \frac{1}{m}\sum_{i=1}^{m} \log P_{\theta_D}(\text{source} = 1|y_i)$$

Maximize probability of fooling discriminator

Conneau et al. (2018) Popularly known as **MUSE**

Source: [NAACL Tutorial]

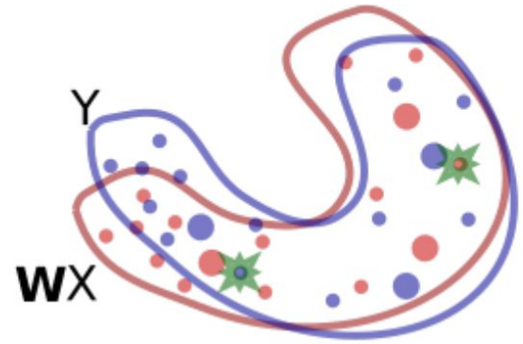
Adversarial Mapping



Søgaard et al. (2018)

- More specifically, monolingual embedding spaces should be ***approximately isomorphic***, i.e. same number of vertices, connected the same way
- **Does not strictly hold** even for related languages

- **Fail to map** between distant languages
  → Why?
- Remember the unsupervised alignment step:

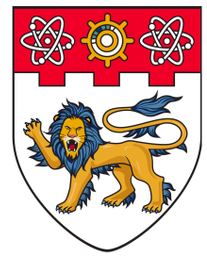


- Embedding spaces need to be **similar** for this to work

Our recent work (ACL, 2020) is independent of this assumption

source: (Ruder, 2019)

Evaluation of Cross Lingual Vectors



NTU NLP

Word Translation /Dictionary Induction (standard/intrinsic task):

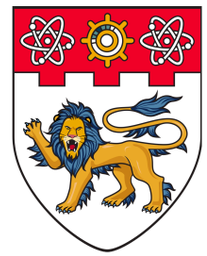
	En-Ms		En-Fi		En-Et		En-Tr		En-El		En-Fa		En-He		En-Ta		En-Bn		En-Hi	
	→	←	→	←	→	←	→	←	→	←	→	←	→	←	→	←	→	←	→	←
GH Distance	0.49		0.54		0.68		0.41		0.46		0.39		0.45		0.29		0.49		0.56	
Unsupervised Baselines																				
Artetxe et al. (2018b)	49.0	49.7	49.8	63.5	33.7	51.2	52.7	63.5	47.6	63.4	33.4	40.7	43.8	57.5	0.0	0.0	18.4	23.9	39.7	48.0
Conneau et al. (2018)	46.2	0.0	38.4	0.0	19.4	0.0	46.4	0.0	39.5	0.0	30.5	0.0	36.8	53.1	0.0	0.0	0.0	0.0	0.0	0.0
Supervision With **"1K Unique"** Seed Dictionary																				
Supervised Baselines																				
Artetxe et al. (2017)	36.5	41.0	40.8	56.0	21.3	39.0	39.5	56.5	34.5	56.2	24.1	35.7	30.2	51.7	5.4	12.7	6.2	19.9	22.6	38.8
Artetxe et al. (2018a)	35.3	34.0	30.8	40.8	21.6	32.6	33.7	43.3	32.0	46.4	22.8	27.6	32.27	39.1	7.3	11.9	11.3	15.7	26.2	30.7
Conneau et al. (2018)	46.2	44.7	46.0	58.4	29.3	40.0	44.8	58.5	42.1	56.5	31.6	38.4	38.3	52.4	11.7	16.0	14.3	19.7	32.5	42.3
Joulin et al. (2018)	31.4	30.7	30.4	41.4	20.1	26.0	30.7	36.5	28.8	43.6	18.7	23.1	33.5	34.3	6.0	10.1	7.6	11.3	20.7	25.7
LNMAP	**50.1**	**50.6**	**54.3**	**64.3**	**39.1**	**51.1**	**52.9**	**64.0**	**48.3**	**62.1**	**35.3**	**41.3**	**45.3**	**54.9**	**18.3**	**24.3**	**19.7**	**30.3**	**36.9**	**49.3**
Supervision With **"5K Unique"** Seed Dictionary																				
Supervised Baselines																				
Artetxe et al. (2017)	36.5	42.0	40.8	57.0	22.4	39.6	39.6	56.7	37.2	56.4	26.0	35.3	31.6	51.9	6.2	13.4	8.2	21.3	23.2	38.3
Artetxe et al. (2018a)	54.6	52.5	48.8	65.2	38.2	54.8	52.0	65.1	47.5	64.6	38.4	42.4	47.4	57.4	18.4	25.79	21.9	31.8	40.3	49.5
Conneau et al. (2018)	46.4	45.7	46.0	59.2	31.0	41.7	45.9	60.1	43.1	56.8	31.6	37.7	38.4	53.4	14.27	19.1	15.0	22.6	32.9	42.8
Joulin et al. (2018)	50.0	49.3	53.0	66.1	39.8	52.0	**54.0**	61.7	47.6	63.4	**39.6**	42.2	**53.0**	56.3	16.0	24.25	21.3	27.0	38.3	47.5
LNMAP	**50.2**	**54.6**	**54.7**	**69.2**	**41.8**	**57.2**	52.0	**66.6**	**49.7**	**65.5**	36.8	**44.5**	47.3	**58.7**	**20.7**	**30.7**	**22.2**	**36.2**	**39.1**	**51.6**
Supervision With **"Whole"** ("5K Unique" Source Words) Seed Dictionary																				
Supervised Baselines																				
Artetxe et al. (2017)	37.0	41.6	40.8	57.0	22.7	39.5	38.8	56.9	37.5	57.2	25.4	36.3	32.2	52.1	5.9	14.1	7.7	21.7	22.4	38.3
Artetxe et al. (2018a)	55.2	51.7	48.9	64.6	37.4	54.0	52.2	63.7	48.2	65.0	39.0	42.6	47.6	58.0	19.6	25.2	21.1	30.6	**40.4**	50.0
Conneau et al. (2018)	46.3	44.8	46.4	59.0	30.9	42.0	45.8	59.0	44.4	57.4	31.8	38.8	39.0	53.4	15.1	18.4	15.5	22.4	32.9	44.4
Joulin et al. (2018)	**51.4**	49.1	**55.6**	65.8	40.0	50.2	**53.8**	61.7	**49.1**	62.8	**40.5**	42.4	**52.2**	57.9	17.7	24.0	20.2	26.9	38.2	47.1
LNMAP	50.4	**56.0**	55.1	**71.5**	**41.9**	**57.5**	52.6	**66.3**	**49.1**	**67.1**	36.5	**43.7**	47.4	**61.0**	**19.9**	**32.0**	**22.5**	**36.6**	38.7	**52.7**

Table 1: Translation accuracy (P@1) on **low-resource** languages on **MUSE dataset** using fastText embeddings.

MUSE: https://github.com/facebookresearch/MUSE

source: (Mohiuddin et al, 2020)

Evaluation of Cross Lingual Vectors



NTU NLP

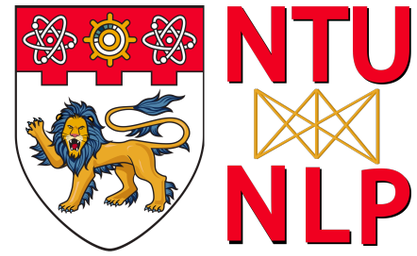
Extrinsic tasks: Cross-lingual Natural Language Inference (XNLI)

Supervised	Dict	EN–DE	EN–FR	EN–TR	EN–RU	Avg
PROC	1K	0.561	0.504	0.534	0.544	0.536
PROC	5K	0.607	0.534	0.568	0.585	**0.574**
PROC-B	1K	0.613	0.543	0.568	0.593	**0.579**
PROC-B	3K	0.615	0.532	0.573	0.599	**0.580**
DLV	5K	0.614	0.556	0.536	0.579	**0.571**
RCSLS	1K	0.376	0.357	0.387	0.378	0.374
RCSLS	5K	0.390	0.363	0.387	0.399	0.385
Unsupervised						
VECMAP		0.604	0.613	0.534	0.574	**0.581**
MUSE		0.611	0.536	0.359*	0.363*	0.467
ICP		0.580	0.510	0.400*	0.572	0.516
GWA		0.427*	0.383*	0.359*	0.376*	0.386

Other Extrinsic tasks:

- Cross-lingual Document Classification
- Cross-lingual Information Retrieval
- Cross-lingual Paraphrase Identification
- Cross-lingual Question Answering

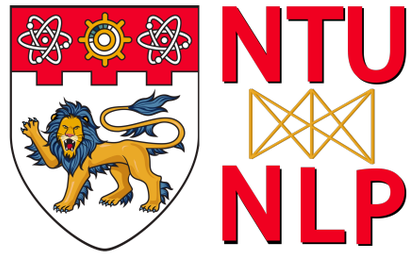
source: (Gravas et al, 2019)



mBERT & XLM

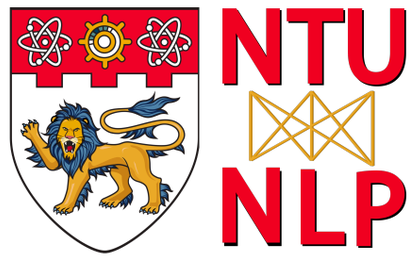
Credit: Some of the base slides were prepared by folks in NTU-NLP

XLM (Lample & Conneau)



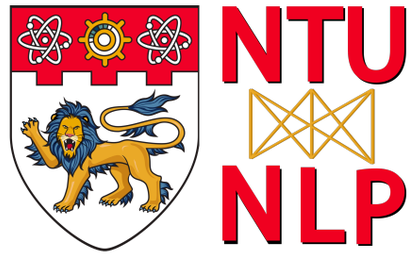
- Extend approach to multiple languages and show the effectiveness of cross-lingual pretraining
- Introduce a new unsupervised method for learning cross-lingual representations using XLM and investigate two monolingual pretraining objectives
- Introduce a new supervised learning objective that improves cross-lingual pretraining when parallel data is available
- Significantly outperform the previous state of the art on cross-lingual classification, unsupervised MT and supervised MT

XLM (Lample & Conneau)



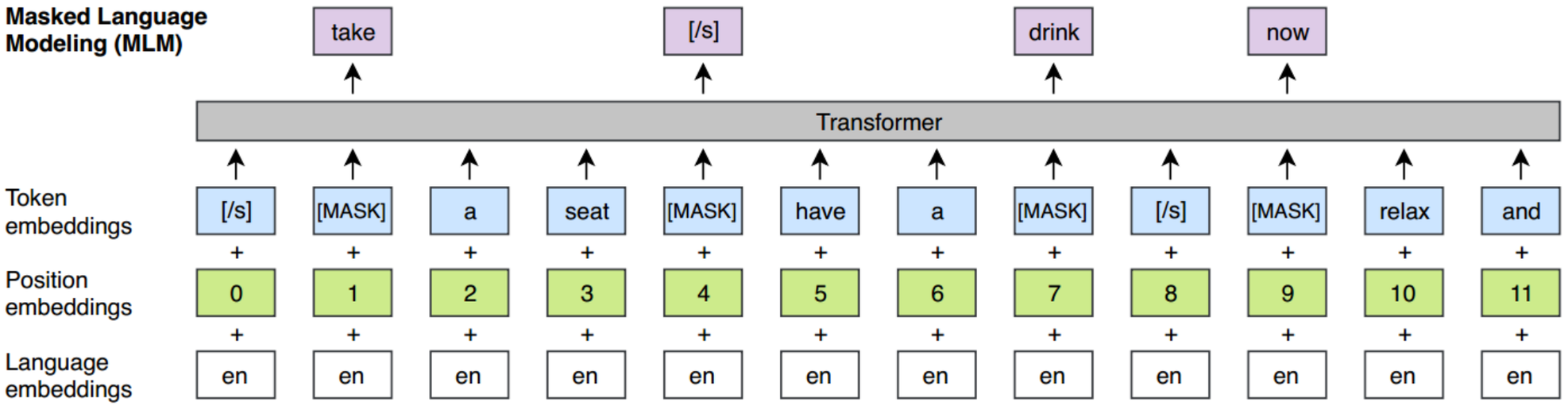
- Proposed three language modeling objectives
- Two of them only require monolingual data (unsupervised), while the third one requires parallel sentences (supervised)

XLM (Lample & Conneau)



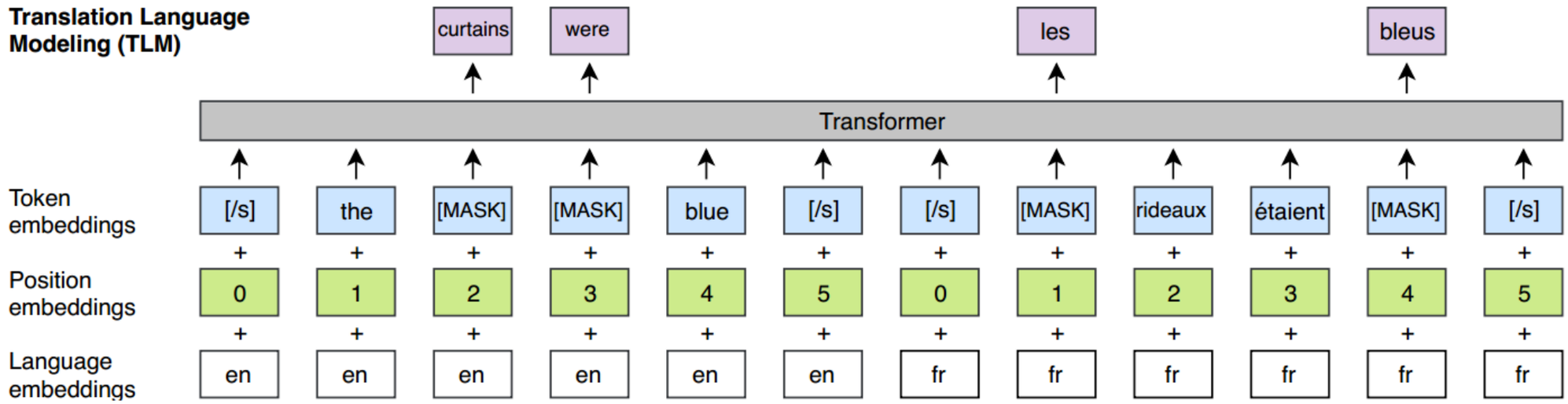
- Process all languages with the same shared vocabulary created through BPE
- Learn BPE splits on the concatenation of sentences sampled randomly from monolingual corpora.
- Sampling increases the number of tokens associated to low-resource languages and alleviates bias towards high-resource languages

XLM: Pre-training



- Similar to BERT
- Instead of using pairs of sentences, use text stream of arbitrary number of sentences (truncated at 256 tokens)
- To counter the imbalance between rare and frequent tokens, they subsample the frequent outputs

XLM: Pre-training



- Extension of MLM, where instead of considering monolingual text streams, concatenate parallel sentences
- Randomly mask words in both the source and target sentences
- For prediction, model can either attend to surrounding same language words or to the other language words
- This encourages the model to align the English and French representations

XLM: Finetuning

Cross-lingual classification

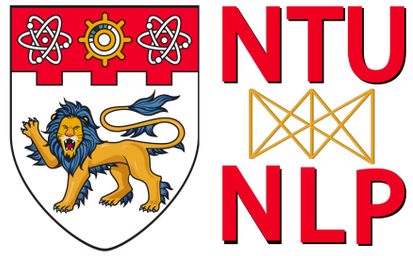
- XLM works as a better initialization of sentence encoders for zero-shot cross-lingual classification
- Add a linear classifier on top of the first hidden state of the pretrained Transformer, and fine-tune all parameters on the English NLI training dataset
- Evaluate the capacity of the model to make correct NLI predictions in the 15 XNLI languages

XLM: Finetuning

Machine Translation

- XLM works as a better initialization of supervised and unsupervised NMT systems
- Pretrain entire encoder and decoder with a cross-lingual language model

XLM: Classification Results



- MultiNLI & XNLI

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

XLM: UNMT Results

	en-fr	fr-en	en-de	de-en	en-ro	ro-en
Previous state-of-the-art - Lample et al. (2018b)						
NMT	25.1	24.2	17.2	21.0	21.2	19.4
PBSMT	28.1	27.2	17.8	22.7	21.3	23.0
PBSMT + NMT	27.6	27.7	20.2	25.2	25.1	23.9
Our results for different encoder and decoder initializations						
EMB	EMB	29.4	29.4	21.3	27.3	27.5	26.6
-	-	13.0	15.8	6.7	15.3	18.9	18.3
-	CLM	25.3	26.4	19.2	26.0	25.7	24.6
-	MLM	29.2	29.1	21.6	28.6	28.2	27.3
CLM	-	28.7	28.2	24.4	30.3	29.2	28.0
CLM	CLM	30.4	30.0	22.7	30.5	29.0	27.8
CLM	MLM	32.3	31.6	24.3	32.5	31.6	29.8
MLM	-	31.6	32.1	**27.0**	33.2	31.8	30.5
MLM	CLM	**33.4**	32.3	24.9	32.9	31.7	30.4
MLM	MLM	**33.4**	**33.3**	26.4	**34.3**	**33.3**	**31.8**

XLM: SNMT Results

Pretraining	-	CLM	MLM
Sennrich et al. (2016)	33.9	-	-
ro $\rightarrow$ en	28.4	31.5	35.3
ro $\leftrightarrow$ en	28.5	31.5	35.6
ro $\leftrightarrow$ en + BT	34.4	37.0	**38.5**

mBART: Multilingual BART

NTU NLP

mBART

- A Fully Trained
- Sequence-to-Sequence
- Denoising Auto-Encoder

pre-trained on large-scale monolingual corpora in many languages using the BART objective.

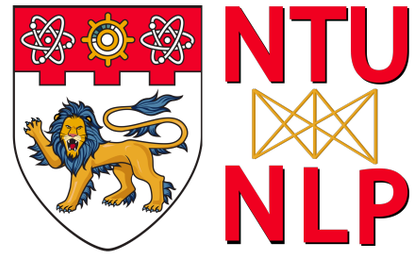
It can finetune both supervised and unsupervised settings.

mBART: Multilingual BART

Large-Scale common crawl (CC) corpus.

- Short Infos
- Pre-train on a subset of 25 languages CC25
- Balance the corpus by $\lambda_i = \frac{1}{p_i} \cdot \frac{p_i^\alpha}{\sum_i p_i^\alpha}$, $\alpha = .7$
- 250000 sub-word tokens.
- token covers more lang. than the CC25 for inference.
- No additional pre-processing (true- casing or normalizing punctuation/characters)

mBART: Multilingual BART

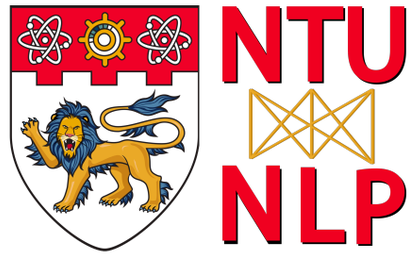


- Regular Transformer (Vaswani et al. 2017)
- 12 layer enc-dec, 1024 dim represenation, 16 heads.
- 680M Params.

Noise function g,

- Span masking.
- Mask 35% of the words in each instance.
- Random sampling a span length according to poison ($\lambda = 3.5$) dist.
- permute the order of sentences within each instance.
- Decoder input starts with a <LID>

mBART: Multilingual BART



- Sample a Language ID <LID>
- Mask 35% of the words in each instance.
- Pack as many consecutive sentences as possible.
- Ends with either doc boundary or 512 token.

mBART: Multilingual BART

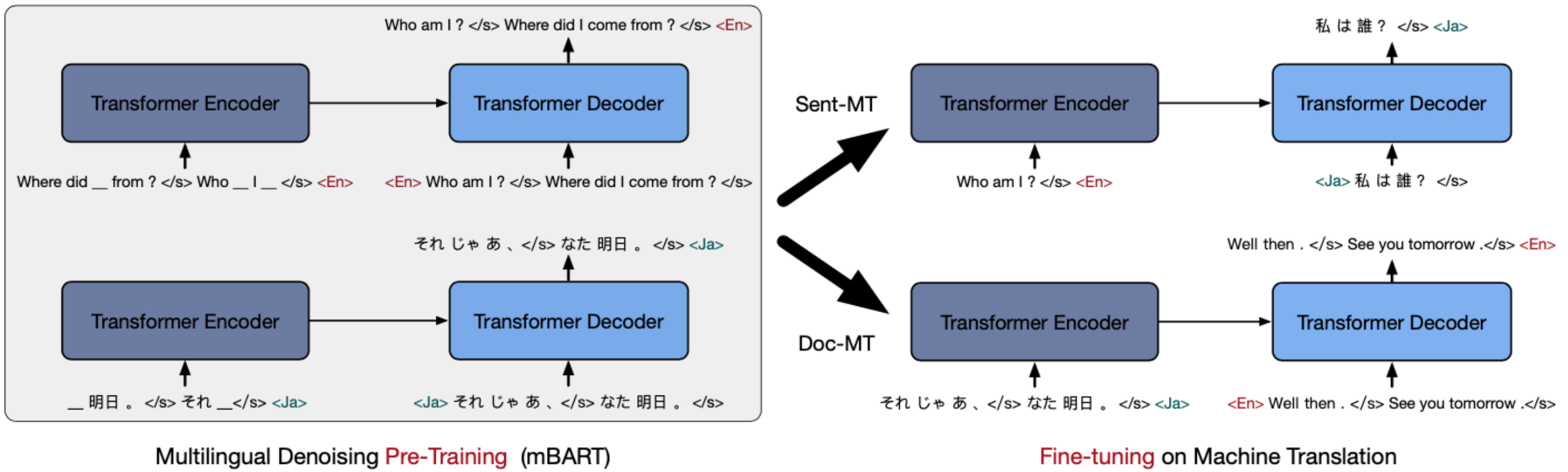
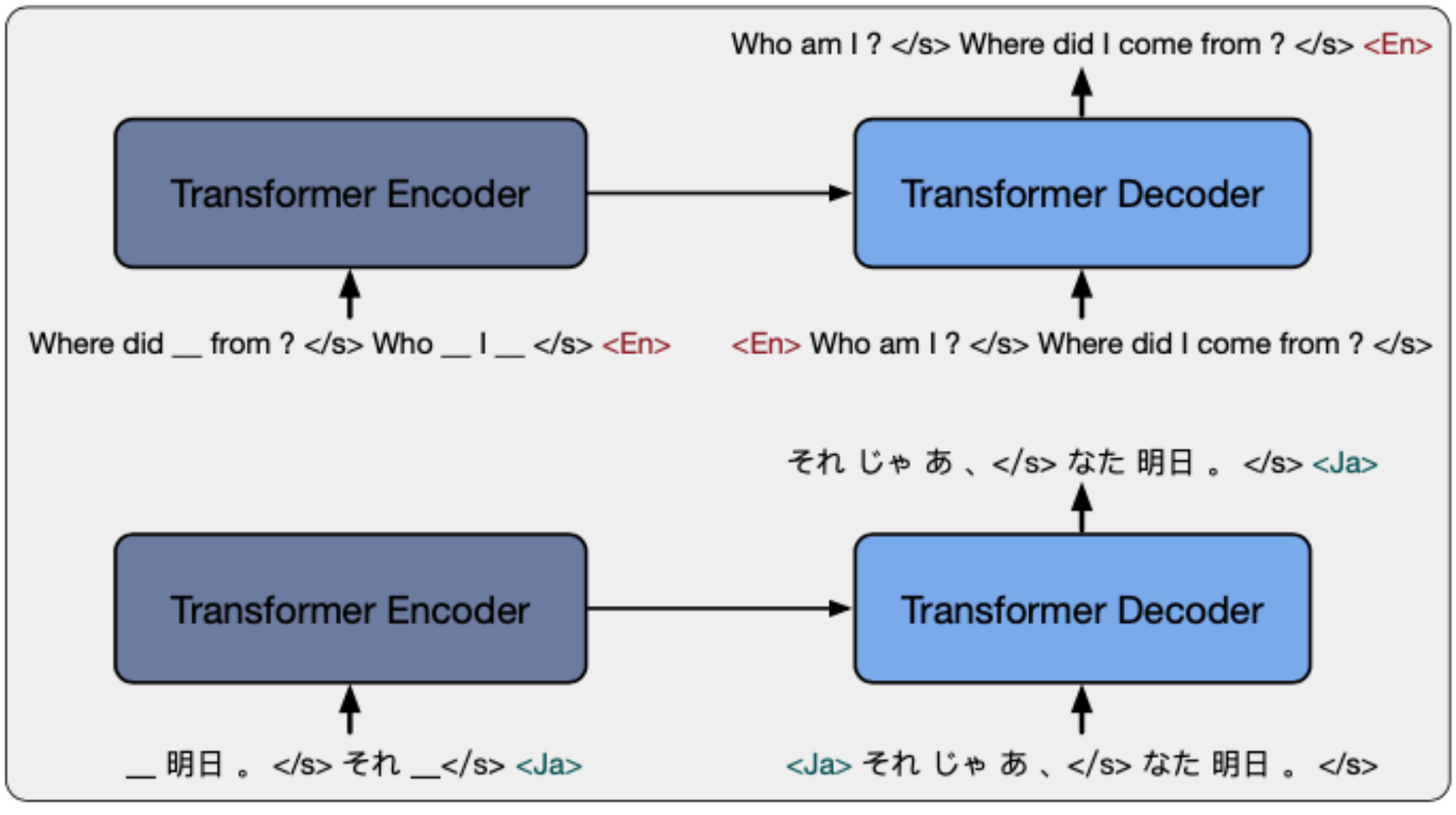


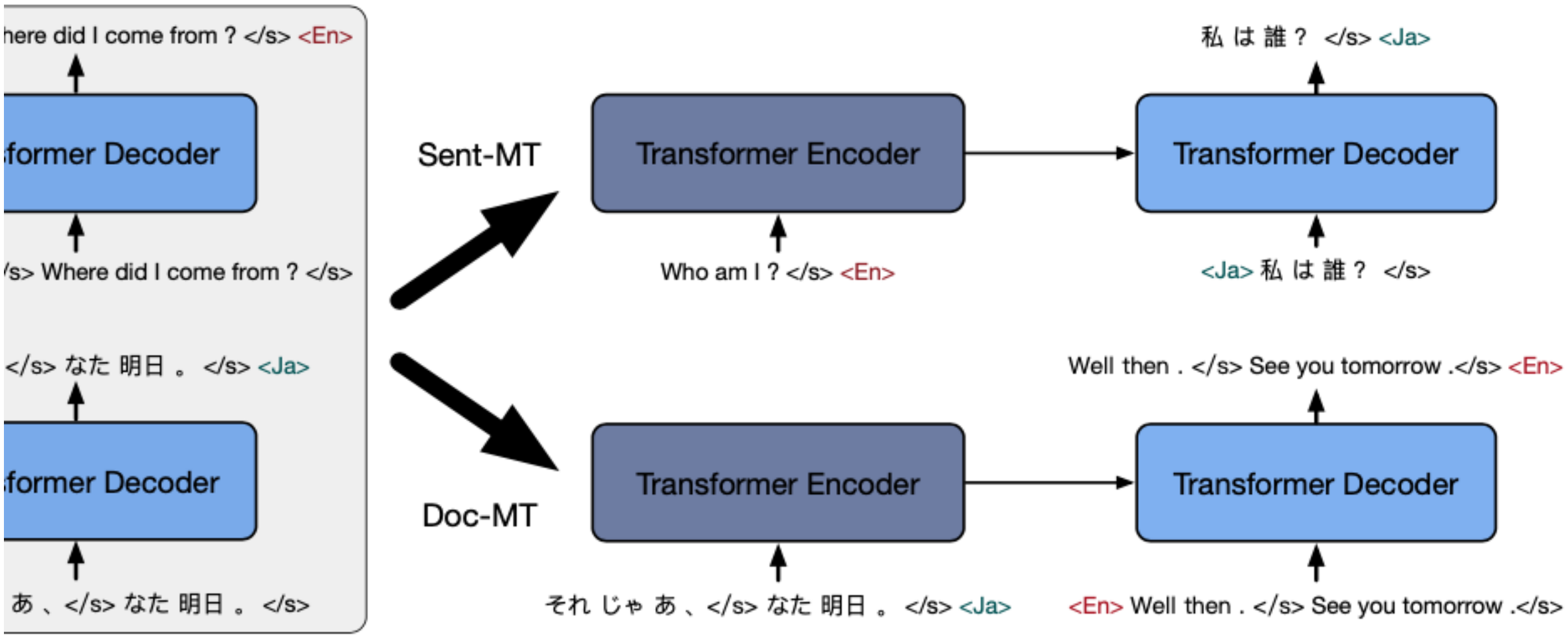
Figure 1: Framework for our Multilingual Denoising Pre-training (left) and fine-tuning on downstream MT tasks (right), where we use (1) sentence permutation (2) word-span masking as the injected noise. A special language id token is added at both the encoder and decoder. One multilingual pre-trained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

mBART: Multilingual BART



Multilingual Denoising Pre-Training (mBART)

mBART: Multilingual BART



(mBART)

Fine-tuning on Machine Translation

mBART: Results (Translation)

Languages	En-Gu		En-Kk		En-Vi		En-Tr		En-Ja		En-Ko	
Data Source	WMT19		WMT19		IWSLT15		WMT17		IWSLT17		IWSLT17	
Size	10K		91K		133K		207K		223K		230K	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	0.0	0.0	0.8	0.2	23.6	24.8	12.2	9.5	10.4	12.3	15.3	16.3
mBART25	**0.3**	**0.1**	**7.4**	**2.5**	**36.1**	**35.4**	**22.5**	**17.8**	**19.1**	**19.4**	**24.6**	**22.6**

Languages	En-Nl		En-Ar		En-It		En-My		En-Ne		En-Ro	
Data Source	IWSLT17		IWSLT17		IWSLT17		WAT19		FLoRes		WMT16	
Size	237K		250K		250K		259K		564K		608K	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	34.6	29.3	27.5	16.9	31.7	28.0	23.3	34.9	7.6	4.3	34.0	34.3
mBART25	**43.3**	**34.8**	**37.6**	**21.6**	**39.8**	**34.0**	**28.3**	**36.9**	**14.5**	**7.4**	**37.8**	**37.7**

Languages	En-Si		En-Hi		En-Et		En-Lt		En-Fi		En-Lv	
Data Source	FLoRes		ITTB		WMT18		WMT19		WMT17		WMT17	
Size	647K		1.56M		1.94M		2.11M		2.66M		4.50M	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	7.2	1.2	10.9	14.2	22.6	17.9	18.1	12.1	21.8	20.2	15.6	12.9
mBART25	**13.7**	**3.3**	**23.5**	**20.8**	**27.8**	**21.4**	**22.4**	**15.3**	**28.5**	**22.4**	**19.3**	**15.9**

mBART: Results (Translation)

Languages	Cs	Es	Zh	De	Ru	Fr
Size	11M	15M	25M	28M	29M	41M
Random	16.5	33.2	**35.0**	**30.9**	**31.5**	**41.4**
mBART25	**18.0**	**34.0**	33.3	30.5	31.3	41.0

Table 3: **High Resource Machine Translation** where all the datasets are from their latest WMT competitions. We only evaluate our models on En-X translation.

mBART: Results (Translation)

Languages	Cs	Es	Zh	De	Ru	Fr
Size	11M	15M	25M	28M	29M	41M
Random	16.5	33.2	**35.0**	**30.9**	**31.5**	**41.4**
mBART25	**18.0**	**34.0**	33.3	30.5	31.3	41.0

Table 3: **High Resource Machine Translation** where all the datasets are from their latest WMT competitions. We only evaluate our models on En-X translation.

mBART: Results (Translation)

Pre-training		Fine-tuning		
Model	**Data**	**En→Ro**	**Ro→En**	**+BT**
Random	None	34.3	34.0	36.8
XLM (2019)	En Ro	-	35.6	38.5
MASS (2019)	En Ro	-	-	39.1
BART (2019)	En	-	-	38.0
XLM-R (2019)	CC100	35.6	35.8	-
BART-En	En	36.0	35.8	37.4
BART-Ro	Ro	37.6	36.8	38.1
mBART02	En Ro	**38.5**	**38.5**	**39.9**
mBART25	CC25	37.7	37.8	38.8

Table 4: **Comparison with Other Pre-training Approaches** on WMT16 Ro-En.

mBART: Results (Translation)

	Monolingual	Nl-En	En-Nl	Ar-En	En-Ar	Nl-De	De-Nl
Random	None	34.6 (-8.7)	29.3 (-5.5)	27.5 (-10.1)	16.9 (-4.7)	21.3 (-6.4)	20.9 (-5.2)
mBART02	En Ro	41.4 (-2.9)	34.5 (-0.3)	34.9 (-2.7)	21.2 (-0.4)	26.1 (-1.6)	25.4 (-0.7)
mBART06	En Ro Cs It Fr Es	43.1 (-0.2)	34.6 (-0.2)	37.3 (-0.3)	21.1 (-0.5)	26.4 (-1.3)	25.3 (-0.8)
mBART25	All	**43.3**	**34.8**	**37.6**	**21.6**	**27.7**	**26.1**

Table 7: **Generalization to Unseen Languages** Language transfer results, fine-tuning on language-pairs without pre-training on them. mBART25 uses all languages during pre-training, while other settings contain at least one unseen language pair. For each model, we also show the gap to mBART25 results.