A Very Brief Introduction to **Multilingual and Cross-Lingual NLP**

Ivan Vulić
University of Cambridge & PolyAI

Online
May 23 2022

Why Cross-Lingual NLP?

“I’d like a ride to Russell Square“

אני רוצה מונית לתחנה המרכזית בתל אביב

“Posso fare un giro per sei persone a Roma Termini?“

“Један ауто до главне железничке молим Вас“

“یک کابین در ایستگاه اصلی اتوبوس لطفا“

“¿Puedo tomar un taxi hasta el aeropuerto?“

“Molim Vas jedno vozilo do Autobusnog“

هل يمكنني الحصول على سيارة أجرة من ميدان التحرير؟

“可以載我去故宮博物館嗎?“

“私は銀座にタクシーを手に入れることはできますか?“

Speaking more languages means communicating with more people…
...and reaching more users and customers...

Why Cross-Lingual NLP?

...but there are **more profound** and **democratic** reasons to work in this area:

- decreasing **the digital divide**
- dealing with **inequality of information**
- mitigating **cross-cultural biases**
- deploying language technology for **underrepresented** languages, dialects, minorities; societal impact
- understanding cross-linguistic differences

"95% of all languages in use today will never gain traction online" (Andras Kornai)

"The limits of my language *online* mean the limits of my world?"

Source: http://labs.theguardian.com/digital-language-divide/

Why Cross-Lingual NLP?

...but there are **more profound** and **democratic** reasons to work in this area:

- decreasing **the digital divide**
- dealing with **inequality of information**
- mitigating **cross-cultural biases**
- deploying language technology for **underrepresented** languages, dialects, minorities; societal impact
- understanding cross-linguistic differences

"95% of all languages in use today will never gain traction online" (Andras Kornai)

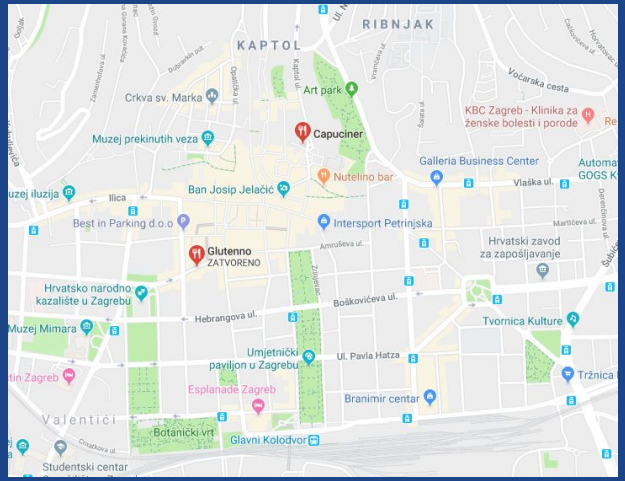
"The limits of my language *online* mean the limits of my world?"

Source: http://labs.theguardian.com/digital-language-divide/

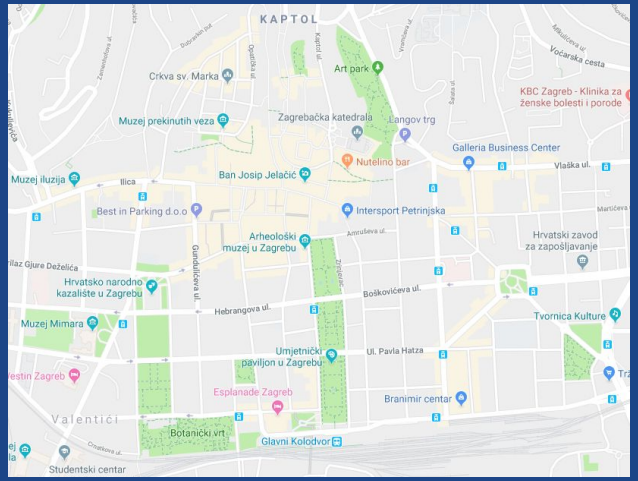
Why Multilingual NLP?

Inequality of information and representation can also affect how we understand places, events, processes...

We're in Zagreb searching for...



...éttermek (HU)



...jatetxea (EU)

...restaurants (EN)

Motivation (Very High-Level)

We want to understand and model the meaning of…

Source: dreamstime.com

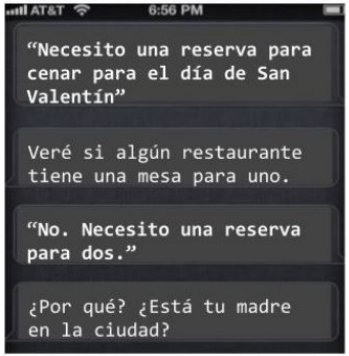
…without manual/human input and without perfect MT

English Conversational AI

A successful conversational agent must perform:

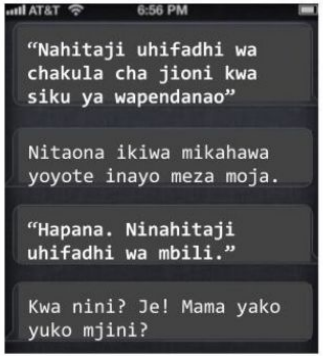
- Automatic speech recognition (ASR)
- Language analysis:
  - Language modeling, spelling correction
  - Syntactic analysis: POS tagging, parsing
  - Semantic analysis: named entity recognition, event detection, semantic role labeling, WSD
  - Coreference resolution, entity linking, commonsense reasoning, world knowledge
- Dialog modeling:
  - Natural language understanding, intent detection, language generation, dialog state tracking
- Information Retrieval and QA
- Text-to-Speech

Multilingual Conversational AI?

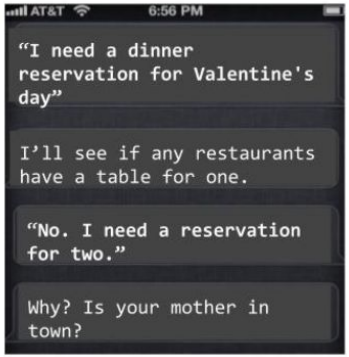


Spanish
534 million speakers

Hindi
615 million speakers



Swahili
100 million speakers



American English

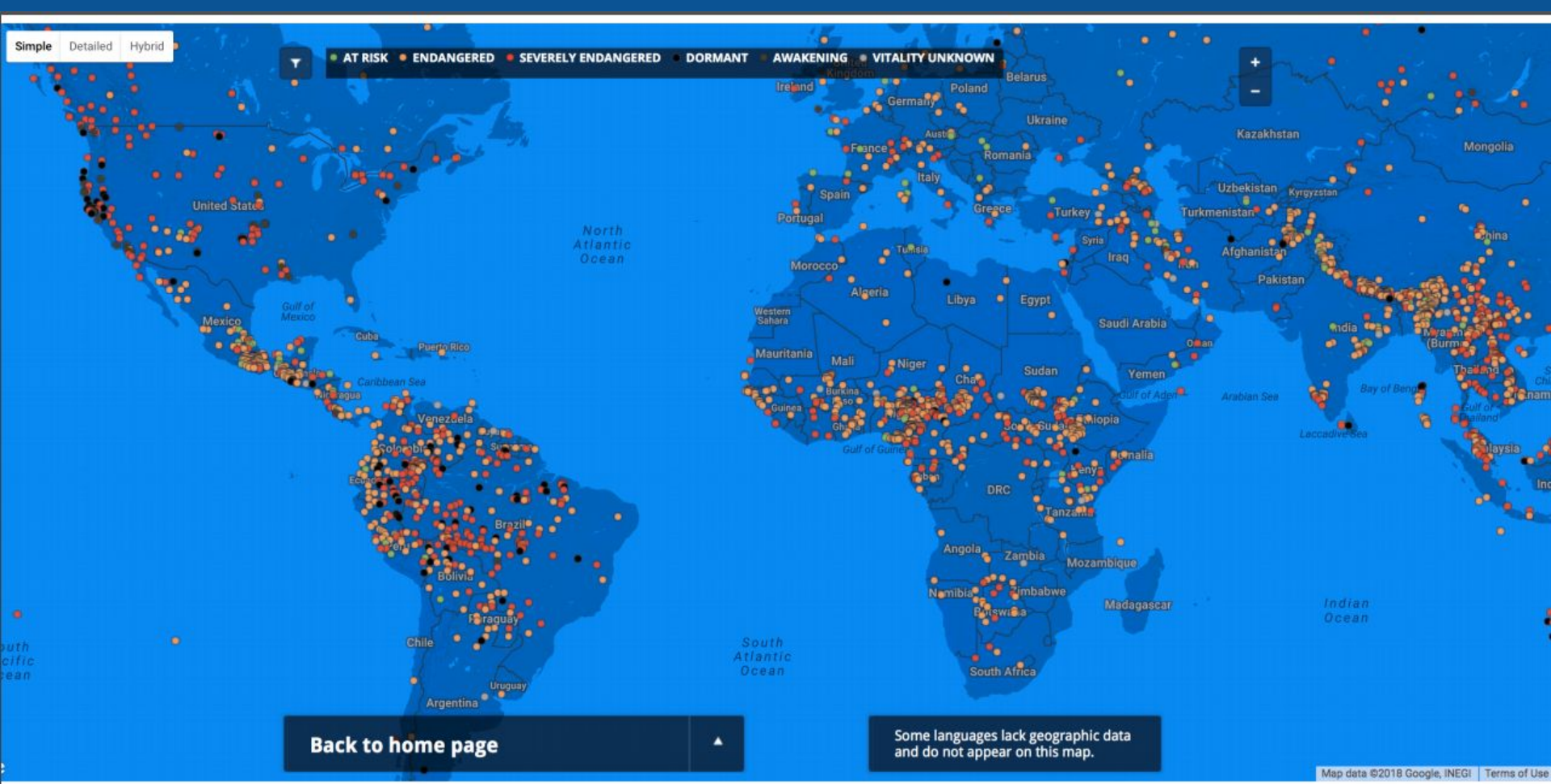
Scottish English

Hinglish

According to Ethnologue (2020) there are **7,117** living languages

What about varieties and dialects?

What about “social media” languages and slang?



http://endangeredlanguages.com/

How Do We Build NLP Systems?

- **Rule-based systems:** Work OK, but require lots of (expert) human effort for each language and task for where they are developed

- **Machine learning based systems:** Work really well when lots of data available, not at all in low-data scenarios

Machine Learning Models

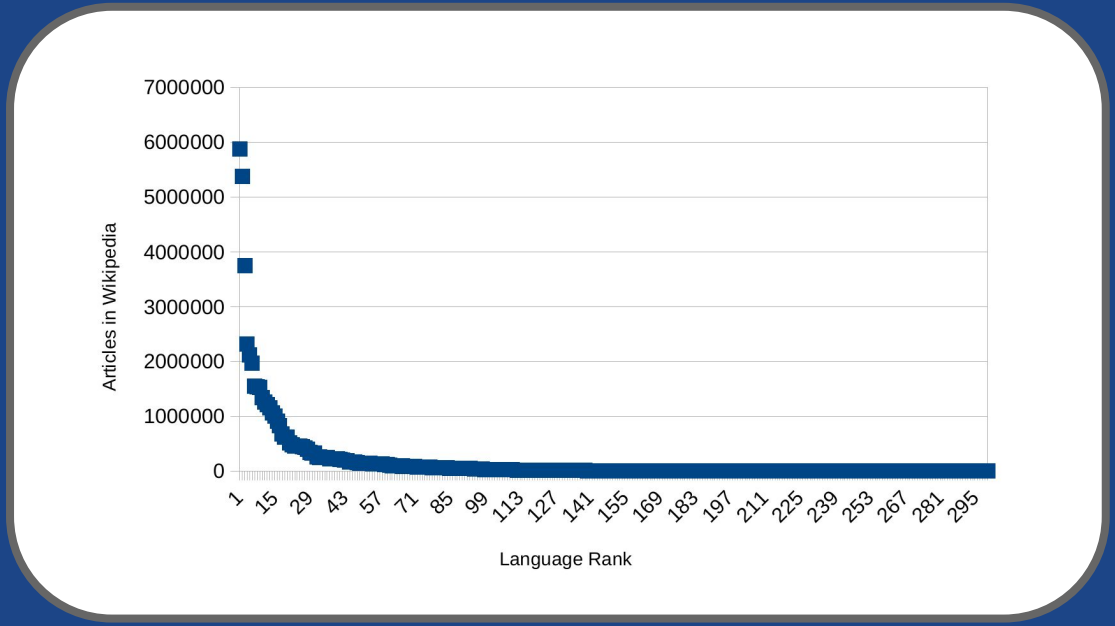
- Formally, map an input X into an output Y. Examples:

Input *X*	Output *Y*	Task
Text	Text in Other Language	Translation
Text	Response	Dialog
Speech	Transcript	Speech Recognition
Text	Linguistic Structure	Language Analysis

To learn, we can use:

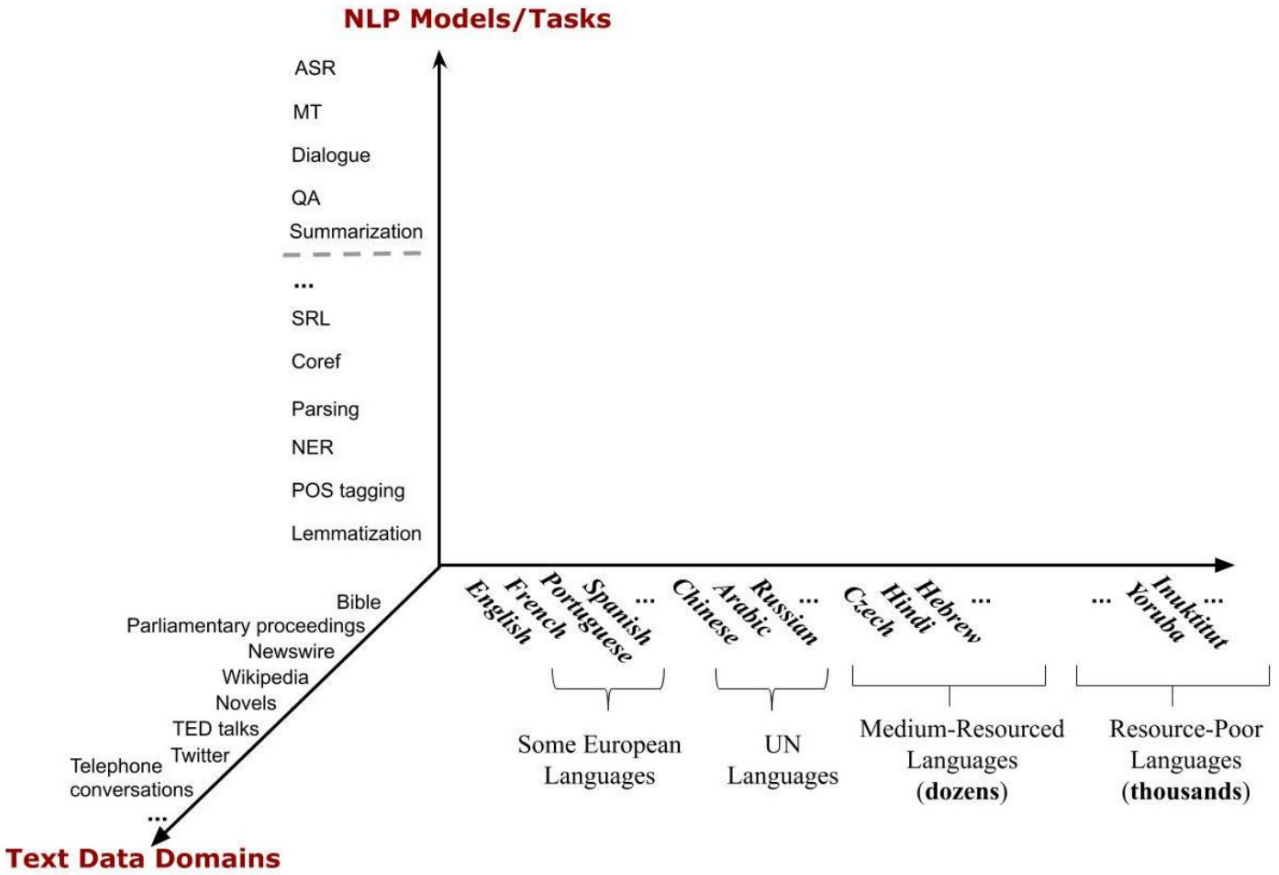
- Paired data <X, Y>, source data X, target data Y
- Paired/source/target data in similar languages

The Long Tail of Data



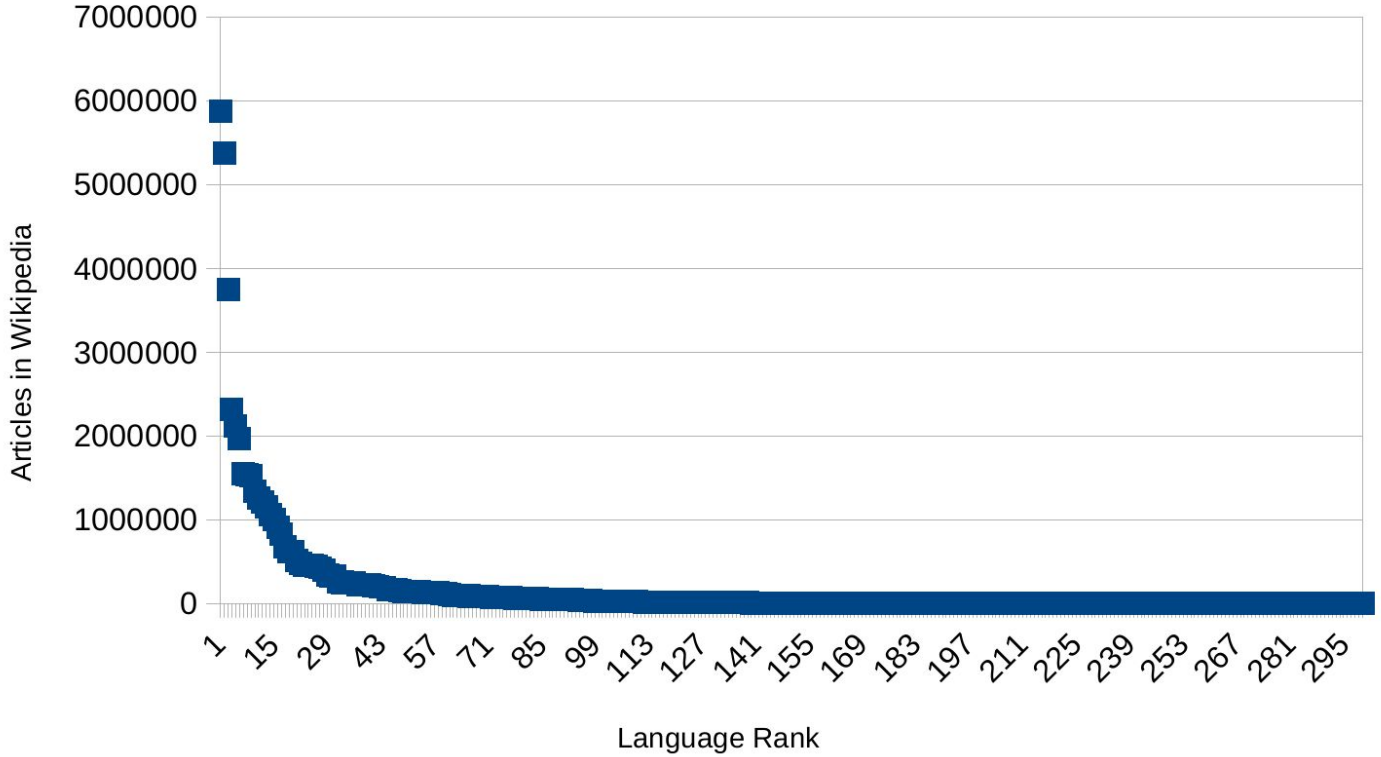
Even getting “raw” unannotated data is problematic for many languages…

How Do We Build NLP Systems?



The Long Tail of Data

Even getting raw monolingual data is problematic for many languages...



Multilingual and Cross-Lingual NLP: How to Cope

Better Models and Algorithms:

- sophisticated modeling/training methods - know NLP/ML!
- linguistically informed methods - know linguistics!

Better Data and Evaluation:

- every piece of relevant data can help - be resourceful!
- make data if necessary - be connected!
- track progress with challenging (and natural!) evaluation data

Better Adaptation:

- leverage similarity between languages, adapt quickly to low-data regimes and new domains

Universal Dependencies (UD)

Current UD Languages

Information about language families (and genera for families with multiple branches) is mostly taken from WALS Online (IE = Indo-European).

Abaza	1	<1K		Northwest Caucasian
Afrikaans	1	49K		IE, Germanic
Akkadian	2	23K		Afro-Asiatic, Semitic
Akuntsu	1	<1K		Tupian, Tupari
Albanian	1	<1K	W	IE, Albanian
Amharic	1	10K		Afro-Asiatic, Semitic
Ancient Greek	2	416K		IE, Greek
Apurina	1	<1K		Arawakan
Arabic	3	1,042K	W	Afro-Asiatic, Semitic
Armenian	1	52K		IE, Armenian
Assyrian	1	<1K		Afro-Asiatic, Semitic
Bambara	1	13K		Mande
Basque	1	121K		Basque
Belarusian	1	275K		IE, Slavic
Bhojpuri	2	6K		IE, Indic
Breton	1	10K	W	IE, Celtic
Bulgarian	1	156K		IE, Slavic
Buryat	1	10K		Mongolic
Cantonese	1	13K		Sino-Tibetan
Catalan	1	531K		IE, Romance
Chinese	5	285K	W	Sino-Tibetan
Chukchi	1	6K		Chukotko-Kamchatkan
Classical Chinese	1	233K		Sino-Tibetan
Coptic	1	48K		Afro-Asiatic, Egyptian
Croatian	1	199K	W	IE, Slavic
Czech	5	2,227K	W	IE, Slavic
Danish	2	100K		IE, Germanic
Dutch	2	306K	W	IE, Germanic
English	9	648K	W	IE, Germanic
Erzya	1	17K		Uralic, Mordvin

100+ languages

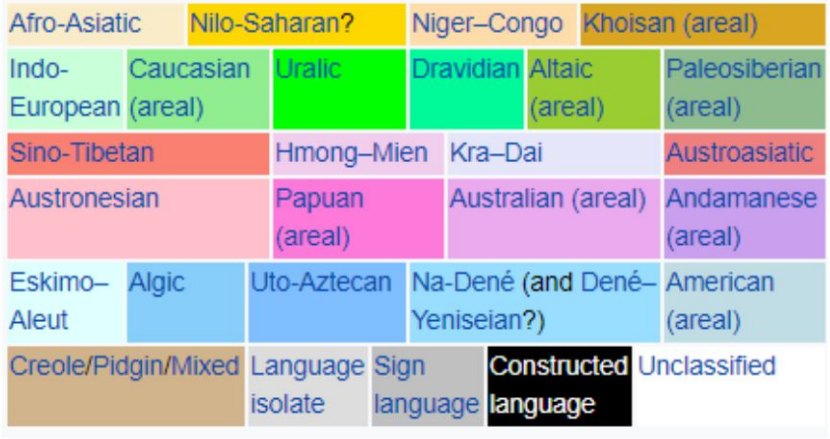
Manual curation and annotation of large-scale resources for thousands of languages is infeasible or prohibitively expensive

What about other tasks?

What about domains?

Language Variety and Varieties

Language family: a group of languages that originate from the same *ancestral/parental* language (proto-language)



[Image from: Wikipedia]

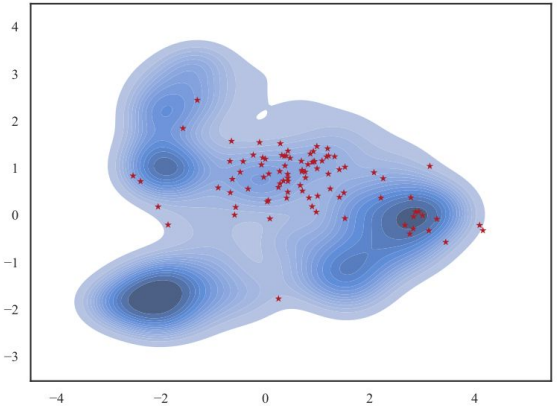


Figure 2: Density of WALS typological features of the world's languages reduced to 2 dimensions via PCA. Red dots are languages covered by UD. Darkness corresponds to more probable regions.

Image:
E.M. Ponti

Language isolates: no known/demonstrable genealogical relationship with any other language.

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 www.glottolog.org

Demographic similarity

How to Define Similarity Across Languages?

Typological similarity

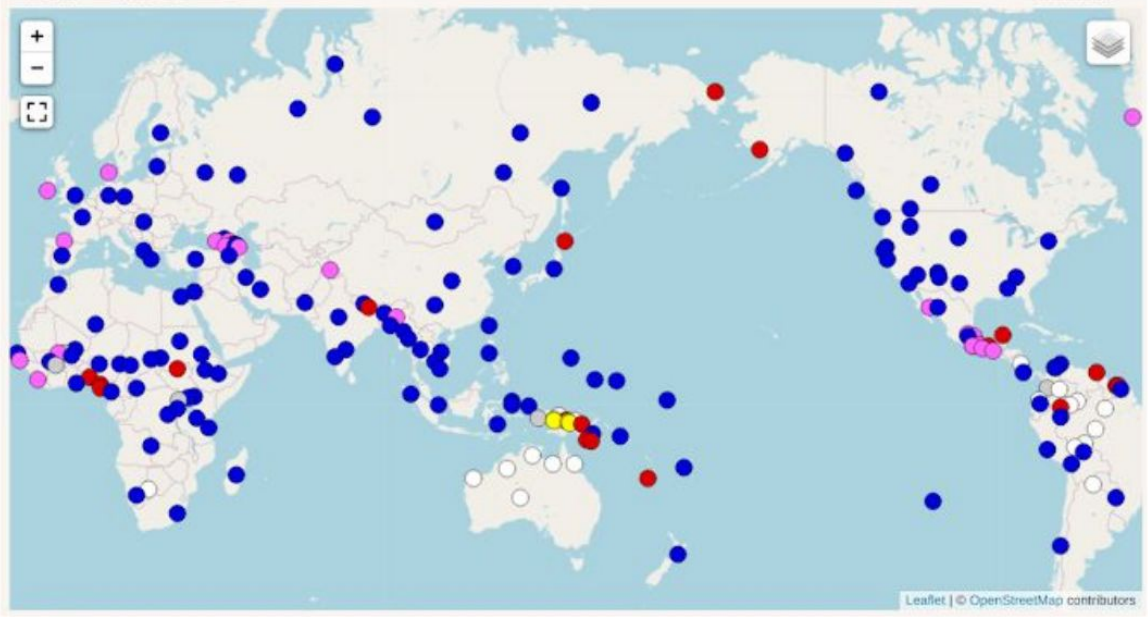
Linguistic typology: classification of languages according to their functional and structural properties:

- explains common properties across languages
- explains structural diversity across languages
- variation versus “universals”

[Ponti et al., CompLing-2019]

Linguistic Typology Example: Numerals

Feature 131A: Numeral Bases



Values

Decimal	125
Hybrid vigesimal-decimal	22
Pure vigesimal	20
Other base	5
Extended body-part system	4
Restricted	20

THE WORLD ATLAS OF LANGUAGE STRUCTURES ONLINE

wals.info/chapter/131

WALS

THE WORLD ATLAS OF LANGUAGE STRUCTURES ONLINE

wals.info

- 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.

Typological Databases

Useful for NLP?
URIEL and *lang2vec* representations
[Littel et al., EACL-2017]

https://pypi.org/project/lang2vec/

Name	Levels	Coverage	Feature Example
World Atlas of Language Structures (WALS)	Phonology, Morphosyntax, Lexical semantics	2,676 languages; 192 attributes; 17% values covered	ORDER OF OBJECT AND VERB Amele: OV (713) Gbaya Kara: VO (705)
Atlas of Pidgin and Creole Language Structures (APiCS)	Phonology, Morphosyntax	76 languages; 335 attributes	TENSE–ASPECT SYSTEMS Ternate Chabacano: purely aspectual (10) Afrikaans: purely temporal (1)
URIEL Typological Compendium	Phonology, Morphosyntax, Lexical semantics	8,070 languages; 284 attributes; ~439,000 values	CASE IS PREFIX Berber (Middle Atlas): yes (38) Hawaaian: no (993)
Syntactic Structures of the World's Languages (SSWL)	Morphosyntax	262 languages; 148 attributes; 45% values covered	STANDARD NEGATION IS SUFFIX Amharic: yes (21) Laal: no (170)
AUTOTYP	Morphosyntax	825 languages; ~1,000 attributes	PRESENCE OF CLUSIVITY !Kung (Ju): false Ik (Kuliak): true
Valency Patterns Leipzig (ValPaL)	Predicate–argument structures	36 languages; 80 attributes; 1,156 values	TO LAUGH Mandinka: 1 > V Sliammon: V.sbj[1] 1
Lyon–Albuquerque Phonological Systems Database (LAPSyD)	Phonology	422 languages; ~70 attributes	ɗ AND ʈ Sindhi: yes (1) Chuvash: no (421)
PHOIBLE Online	Phonology	2,155 languages; 2,160 attributes	m Vietnamese: yes (2053) Pirahã: no (102)
StressTyp2	Phonology	699 languages; 927 attributes	STRESS ON FIRST SYLLABLE Koromfé: yes (183) Cubeo: no (516)
World Loanword Database (WOLD)	Lexical semantics	41 languages; 24 attributes; ~2,000 values	HORSE Quechua: *kaballu* borrowed (24) Sakha: *silgi* no evidence (18)
Intercontinental Dictionary Series (IDS)	Lexical semantics	329 languages; 1,310 attributes	WORLD Russian: *mir* Tocharian A: *ärkiśoṣi*
Automated Similarity Judgment Program (ASJP)	Lexical semantics	7,221 languages; 40 attributes	I Ainu Maoka: *co7okay* Japanese: *watashi*

Ponti, E.M., O'horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., Shutova, E. and Korhonen, A., 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3), pp.559-601.

Multilingual Representation Learning

Cross-lingual Transfer Learning

Why Multilingual and Cross-Lingual Transfer Learning?

Rule-based systems and getting supervised data? Too expensive…

Unsupervised learning:

- Dispose of annotated data and unravel hidden structures within (large portions) of raw unannotated data
- Theoretically interesting, methodologically sophisticated…
- ...but yields subpar performance...

Crossing the Lexical Chasm

1. Full-Blown MT (SMT or NMT)
   - Parallel data needed, critical for under-resourced languages
   - Translate everything from the target language to the source language
2. Multilingual KBs
   - Texts represented using entities from a multilingual KB
   - Same entity ID for same concepts across languages
   - Issues: coverage, entity linking

BabelNet 2.0

A very large multilingual **encyclopedic dictionary** and **ontology**

Crossing the Lexical Chasm: Representation Learning

"The(ir) model, however, is only applicable to English, as large enough training sets do not exist for other languages…"

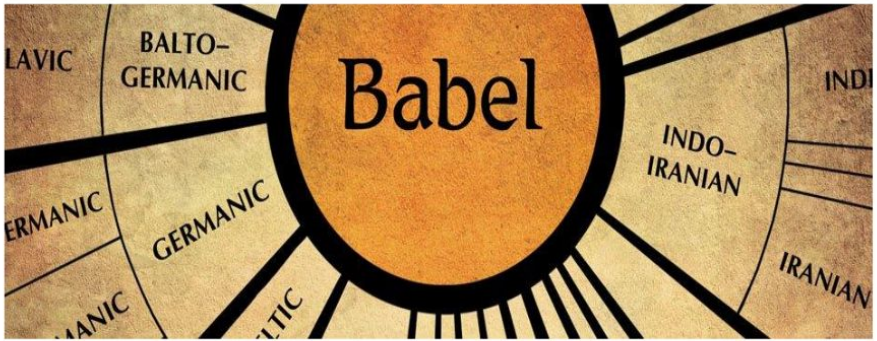
Old paradigm:

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

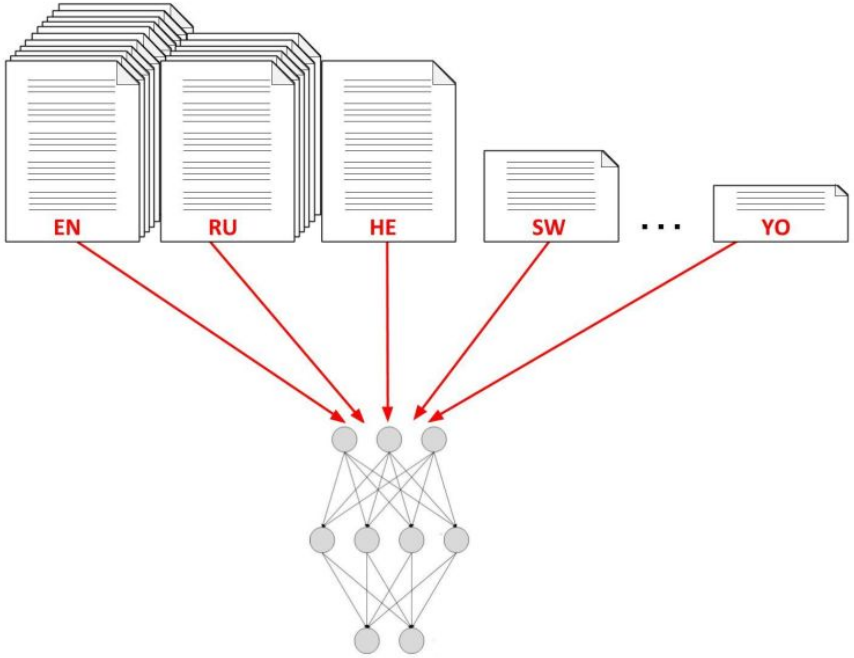
New paradigm:

- **Representation learning**: inputs are semantic vectors (embeddings)

Multilingual / cross-lingual representation and transfer learning



Approaches to Multilingual NLP



Joint multilingual learning – train a single model on a mix of datasets in all languages, to enable **data and parameter sharing** where possible

Approaches to Multilingual NLP

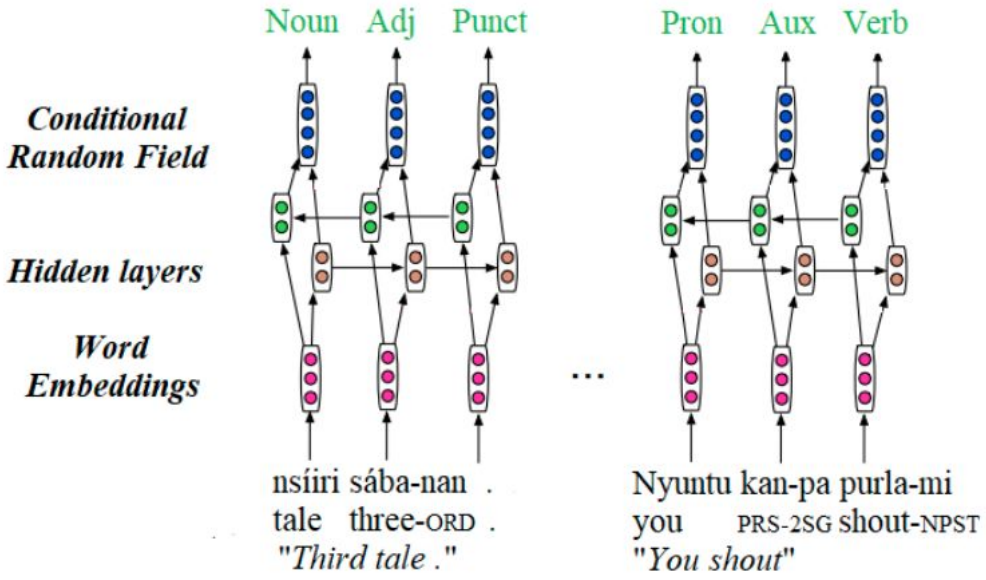


Image from: [Fang and Cohn, EMNLP-17]

Selective Sharing with joint multilingual learning – private versus shared language-specific parameters: define what to share

Cross-Lingual Transfer

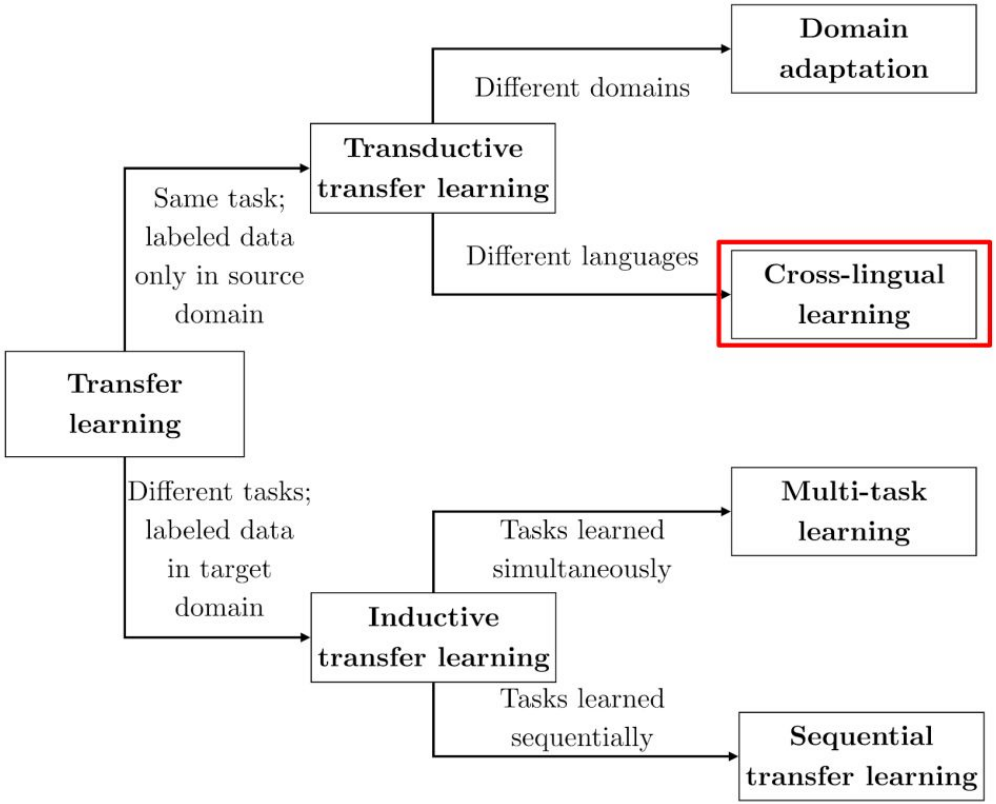
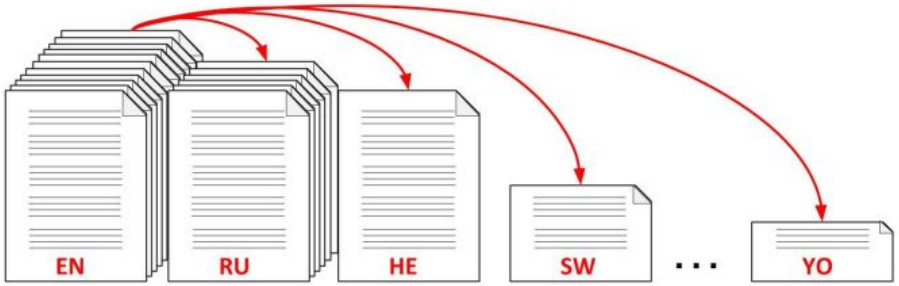


Image from: [Ruder, 2019]

Cross-Lingual Transfer



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

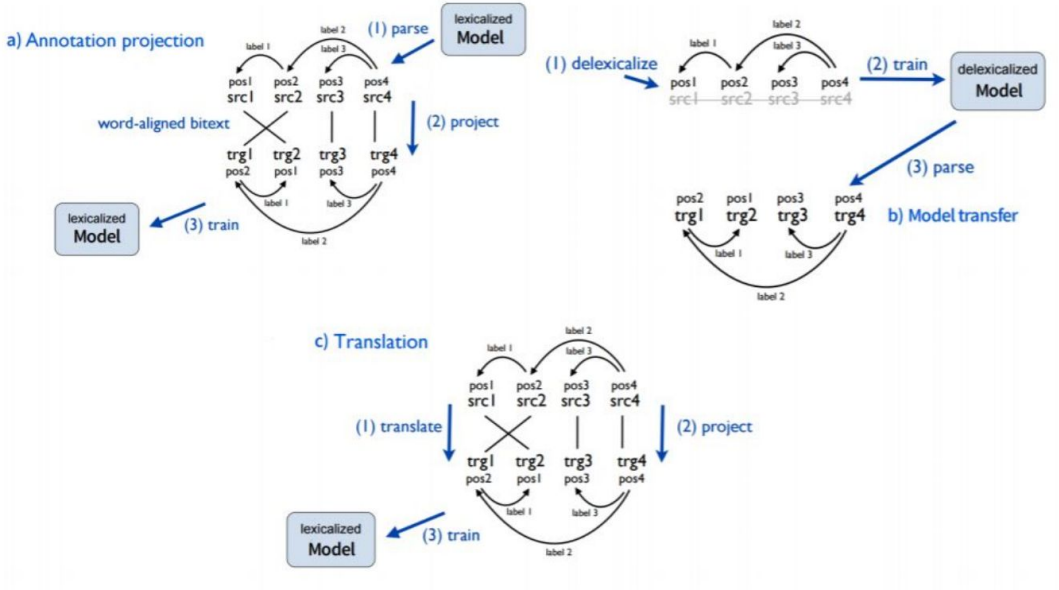
Zero-shot learning: train a model in one language/domain and assume it generalizes out-of-the-box in a low-resource language/domain

Few-shot learning: train a model in one language/domain and use only few examples from a low-resource language/domain to adapt it

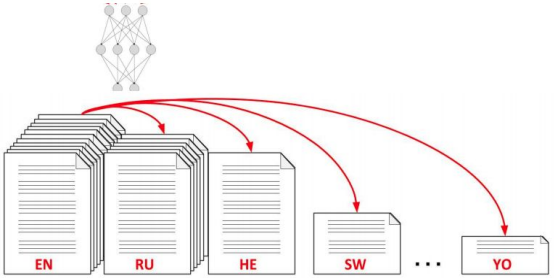
Cross-Lingual Transfer

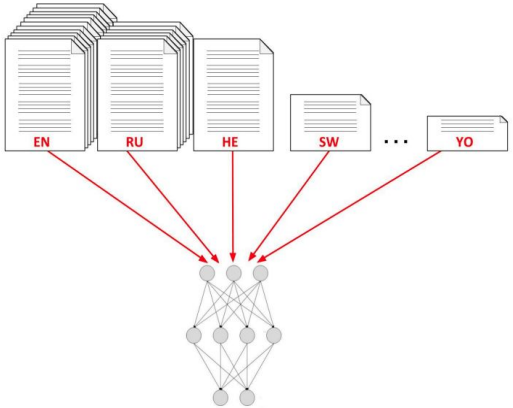
Transfer of annotations (e.g., POS tags, syntactic or semantic features) via cross-lingual bridges (e.g., word or phrase alignments)

Transfer of models – train a model in a resource-rich language and adapt (e.g. fine-tune) it in a resource-poor language



Choosing Transfer Languages

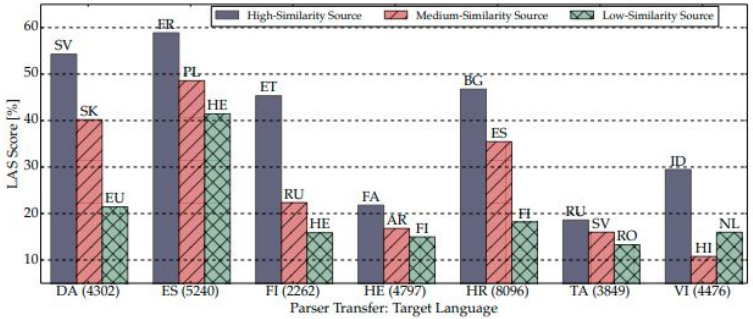




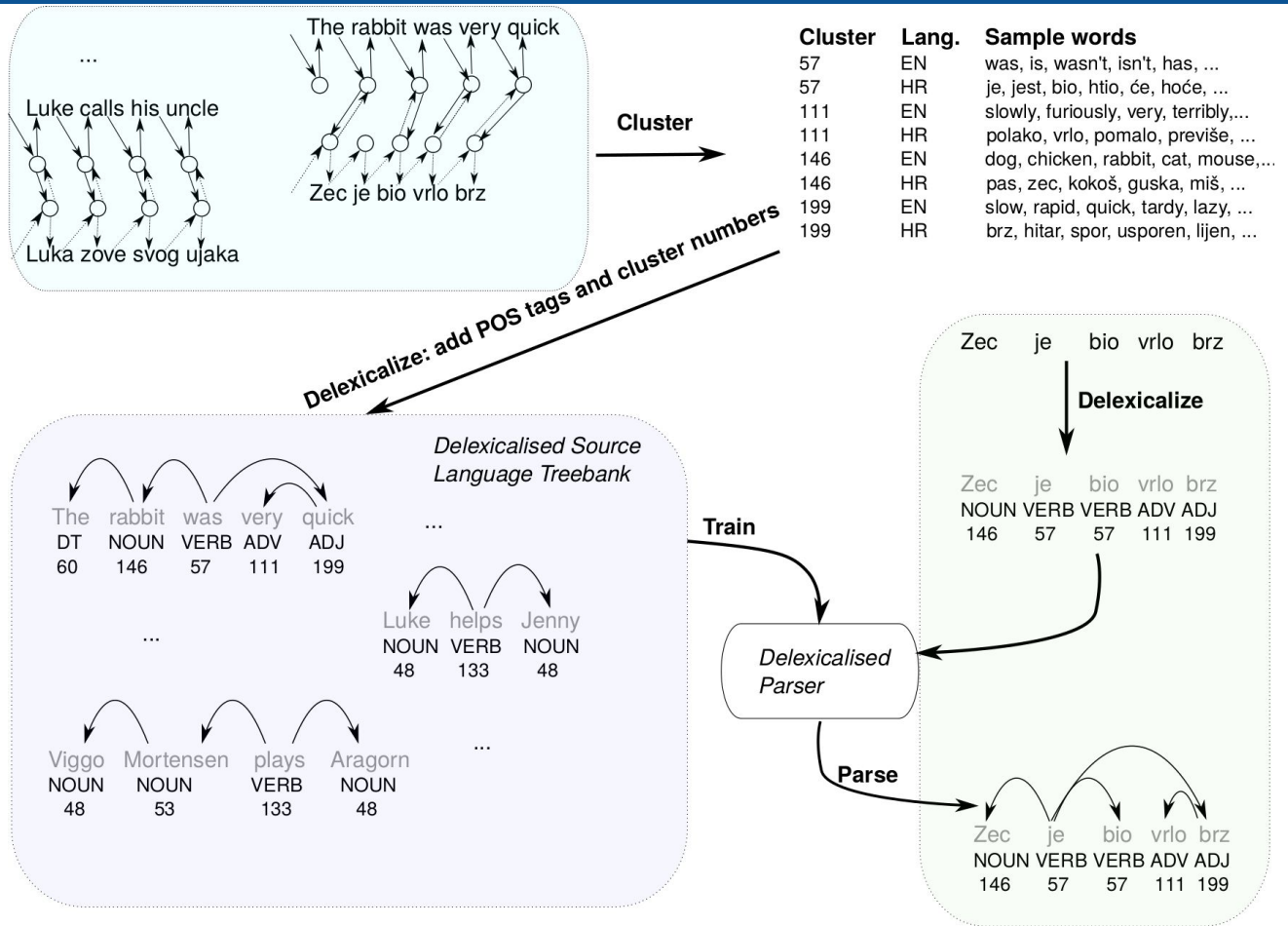
The choice of source should be informed (linguistic typology can help here)...

Structural and lexical dissimilarity -> suboptimal transfer

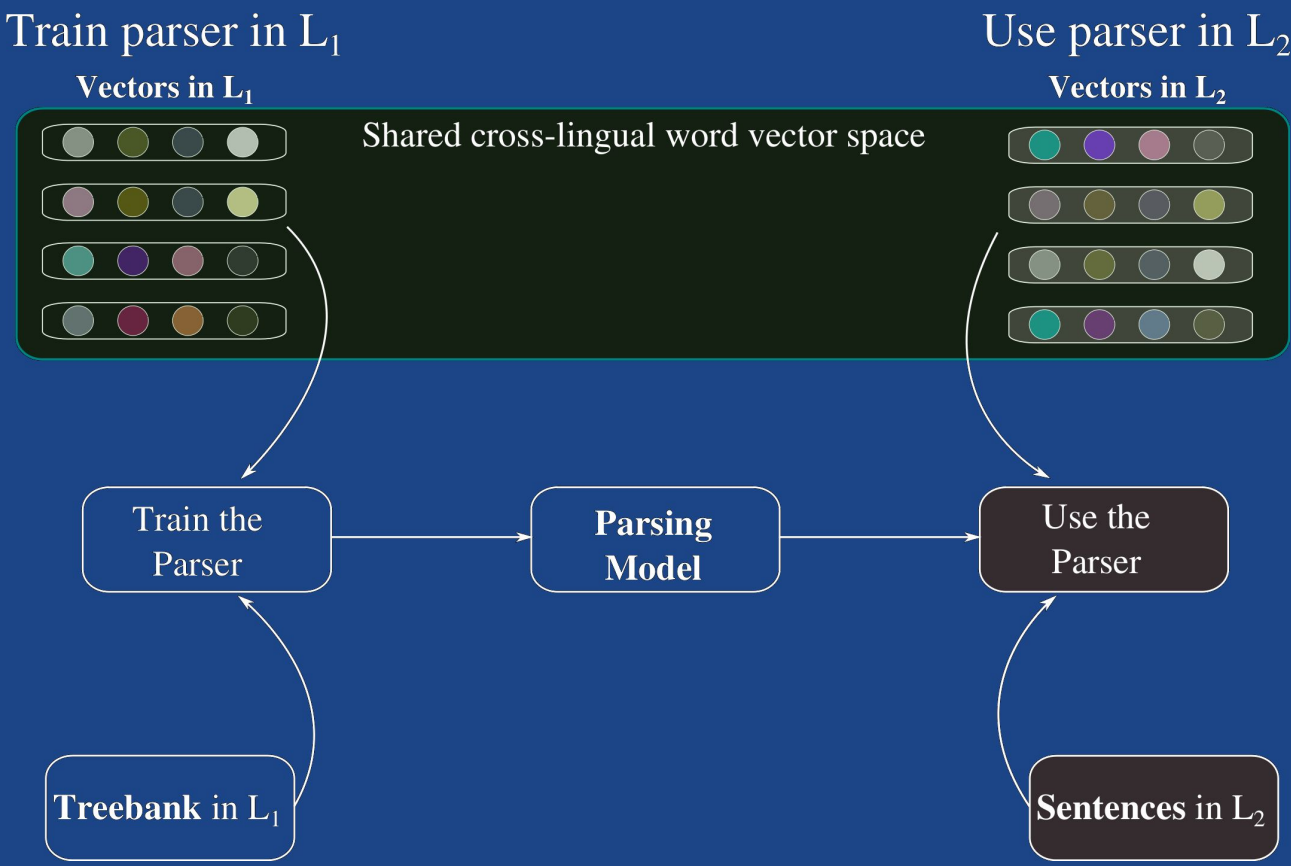
Lin, Y.H. et al. 2019. Choosing Transfer Languages for Cross-Lingual Learning. In Proc. ACL.
https://arxiv.org/pdf/1905.12688.pdf



Transfer in a Nutshell

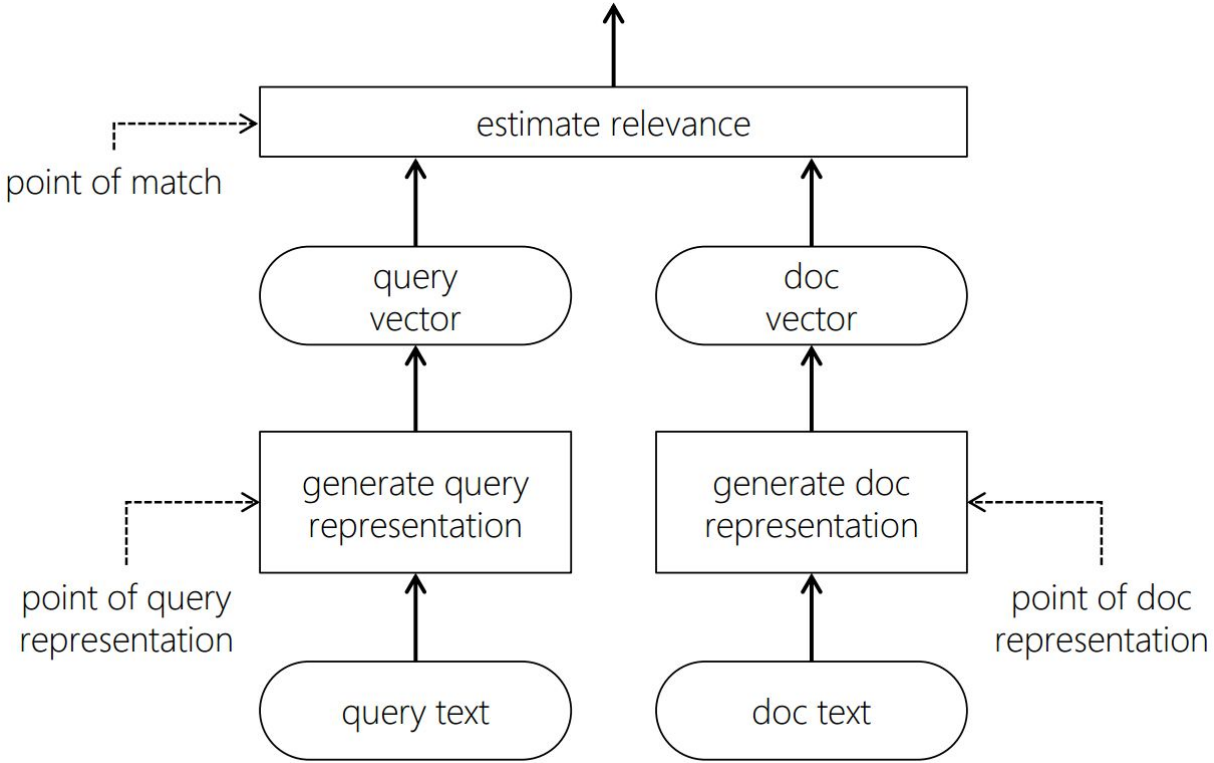


Cross-Lingual Representations Enable the Transfer



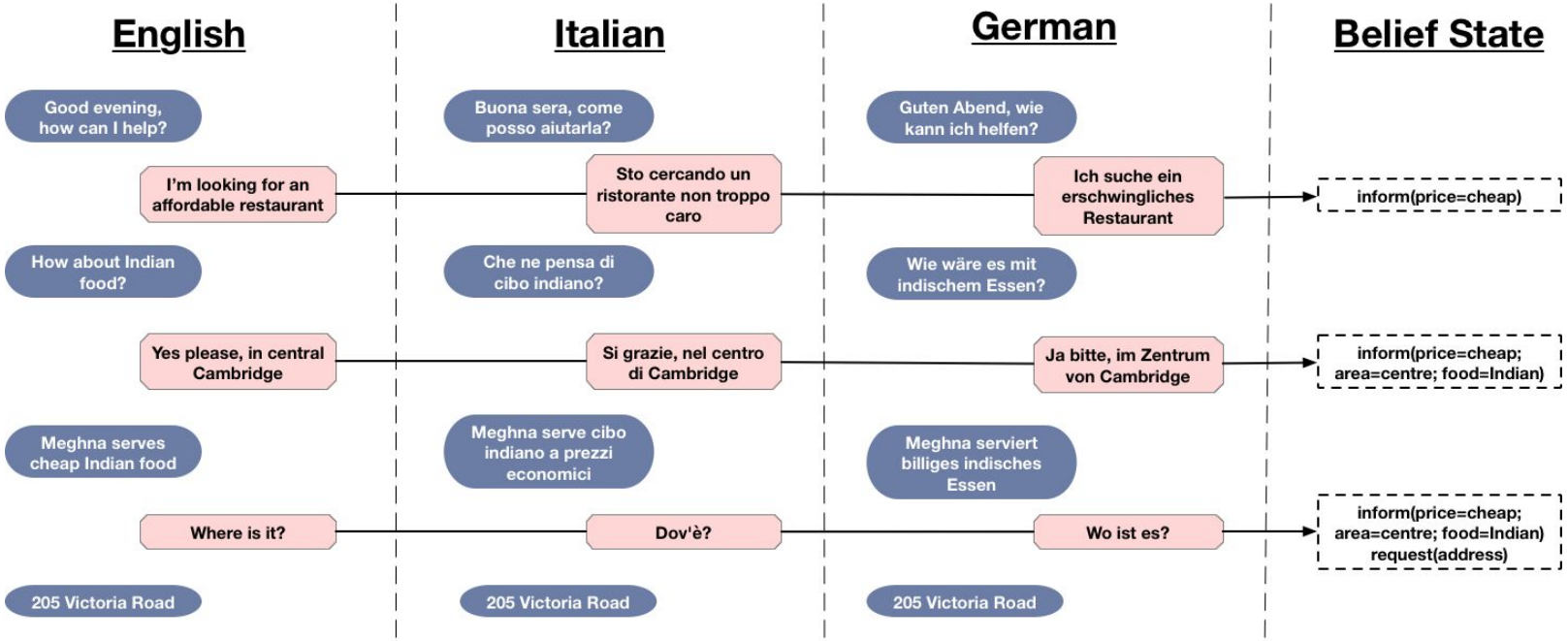
Cross-lingual (parser) transfer

Cross-Lingual Representations Enable Cross-Lingual IR and QA



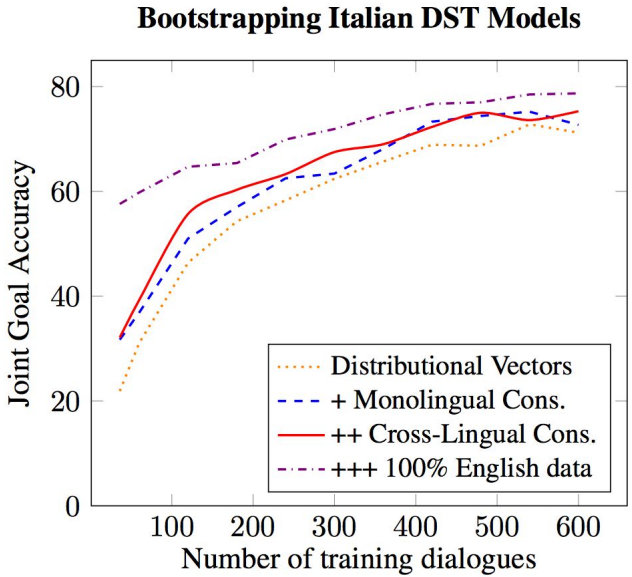
[Vulić and Moens, SIGIR-15; Mitra and Craswell arXiv-17; Litschko et al., SIGIR-18, SIGIR-19]

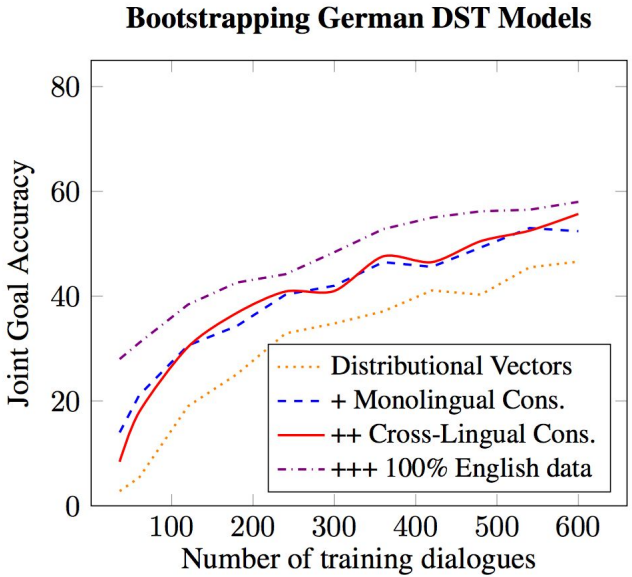
Cross-Lingual Representations Enable Conversational AI



Cross-lingual representations in a shared space: use training data from a resource-rich language?

Cross-Lingual Representations Enable Conversational AI





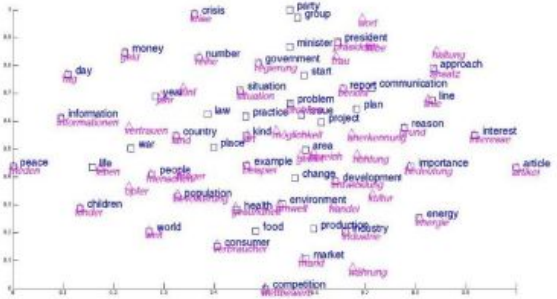
[Mrkšić et al., TACL-17]

Cross-lingual representations in a shared space: use training data from a resource-rich language.

Crossing the Lexical Chasm: Representation Learning

Multilingual / Cross-lingual representations of meaning

- **Word-level**
  - Cross-lingual word embeddings
  - Words with similar meaning across languages have similar vectors

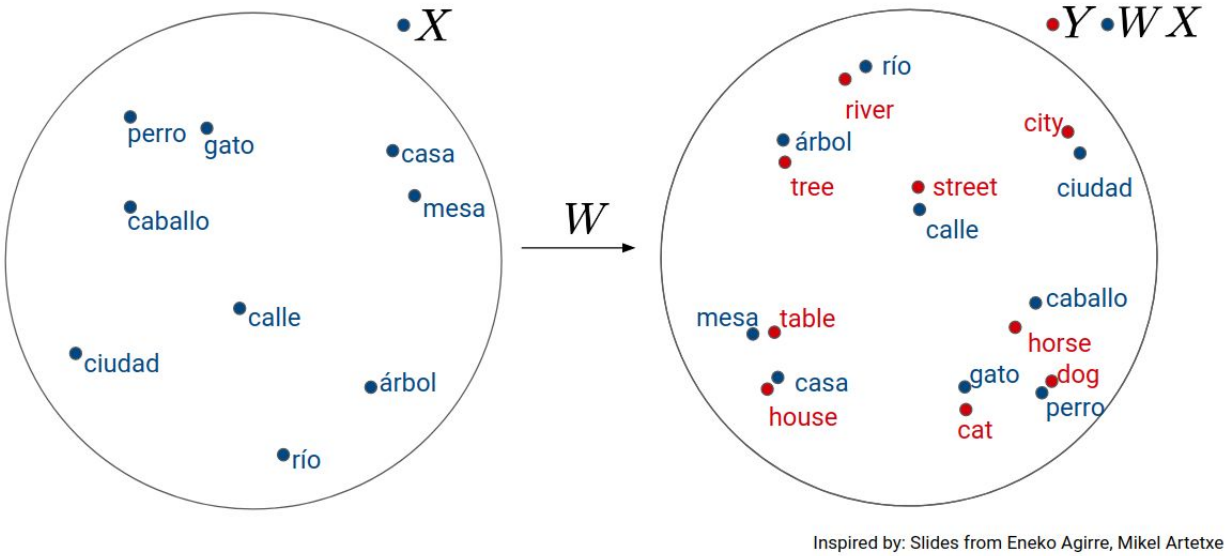


- **Text encoding**
  - Multilingual unsupervised pretraining
    - Multilingual BERT [Devlin et al., ‘19]
    - XLM(-R) [Conneau & Lample, ’19, Conneau et al., 2020]
    - mT5 [Xue et al., 2020]

Cross-Lingual Word Embeddings

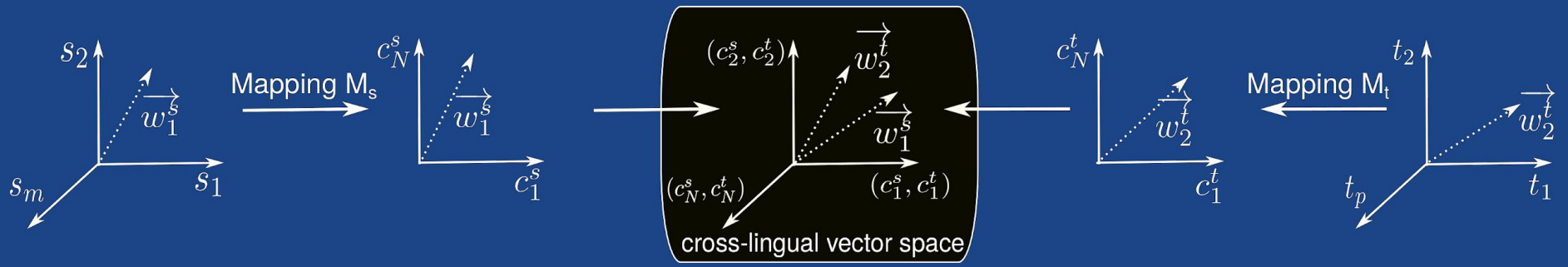
A range of different methods (with different data requirements), but the same **goal:**

Induce a semantic vector space in which words with similar meaning end up with similar vectors, regardless of whether they come from the same language or from different languages.



Inspired by: Slides from Eneko Agirre, Mikel Artetxe

The World Existed Before Embeddings and BERT



Traditional “count-based” cross-lingual vector vector spaces…

[Gaussier et al., ACL 2004; Laroche and Langlais, COLING 2010]

Cross-Lingual Word Embeddings (CLWEs)

Representation of a word $w_1^S \in V^S$:

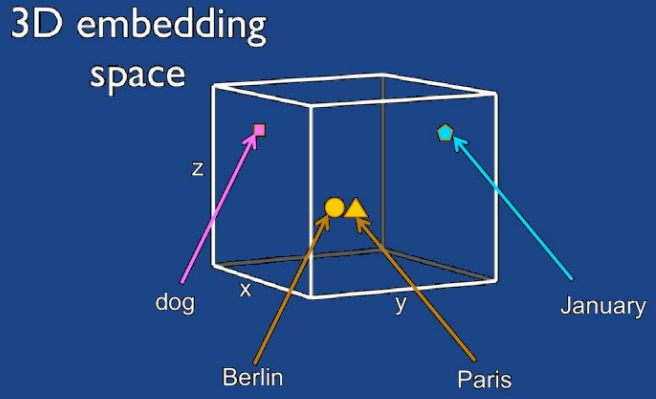
$$vec(w_1^S) = [f_1^1, f_2^1, \ldots, f_{dim}^1]$$

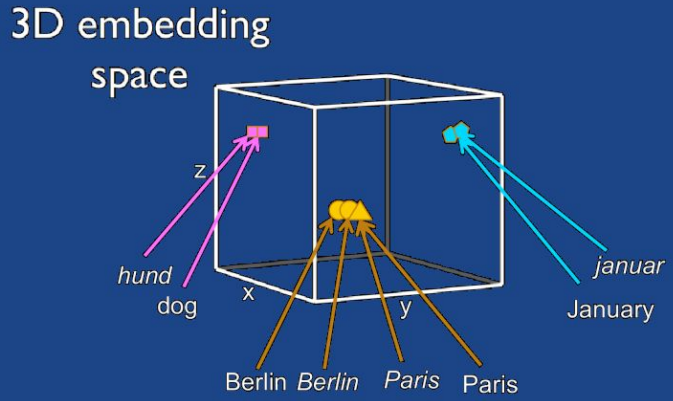
Exactly the same representation for $w_2^T \in V^T$:

$$vec(w_2^T) = [f_1^2, f_2^2, \ldots, f_{dim}^2]$$

Language-independent word representations in the same shared semantic (or *embedding*) space!

Cross-Lingual Word Embeddings





Monolingual vs. **Cross-lingual**

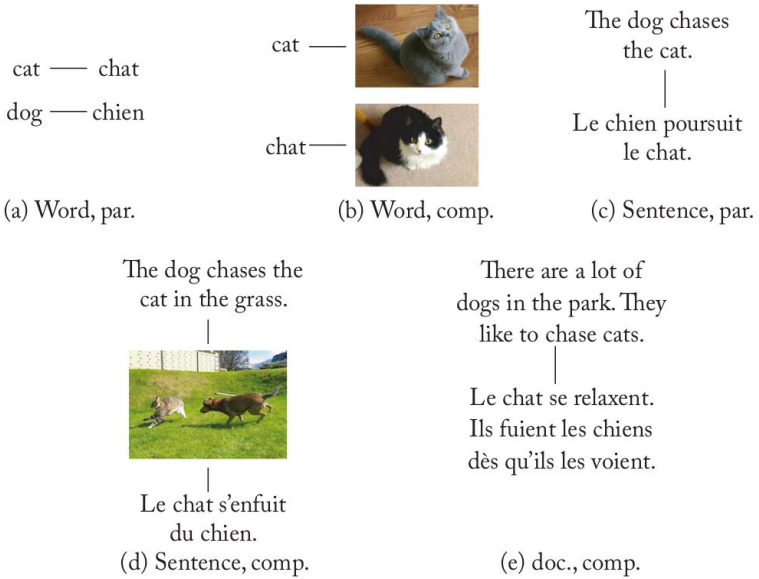
Q1 → Algorithm Design: How to align semantic spaces in two different languages?

Q2 → Data Requirements: Which **bilingual signals** are used for the alignment?

Cross-Lingual Word Embeddings

A large number of different methods, but **the same end goal**:

Induce a shared semantic vector space in which words with similar meaning end up with similar vectors, regardless of their actual language.



We need some bilingual supervision to learn CLWE-s.

Fully unsupervised CLWE-s: they rely only on monolingual data

Cross-Lingual Word Embeddings

A large number of different methods, but **the same end goal**:

Induce a shared semantic vector space in which words with similar meaning end up with similar vectors, regardless of their actual language.

Typology of methods for inducing CLWE-s [Ruder et al., JAIR-19; Søgaard et al., M&C Book]

- **Type of bilingual signal**
  - Document-level, sentence-level, **word-level, no signal (i.e., unsupervised)**
- **Comparability**
  - Parallel texts, comparable texts, non-comparable
- **Point/Time of alignment**
  - Joint embedding models vs. **Post-hoc alignment** vs. post-specialisation/retrofitting
- **Modality**
  - Text only vs. using images for alignment, e.g., [Kiela et al., EMNLP-15; Vulić et al., ACL-16]

General (Simplified) CLWE Methodology

- Most CLWE algorithms are formulated as:

$$\mathcal{L}_S(\mathbf{X}_S) + \mathcal{L}_T(\mathbf{X}_T) + \Omega(\mathbf{X}_S, \mathbf{X}_T)$$

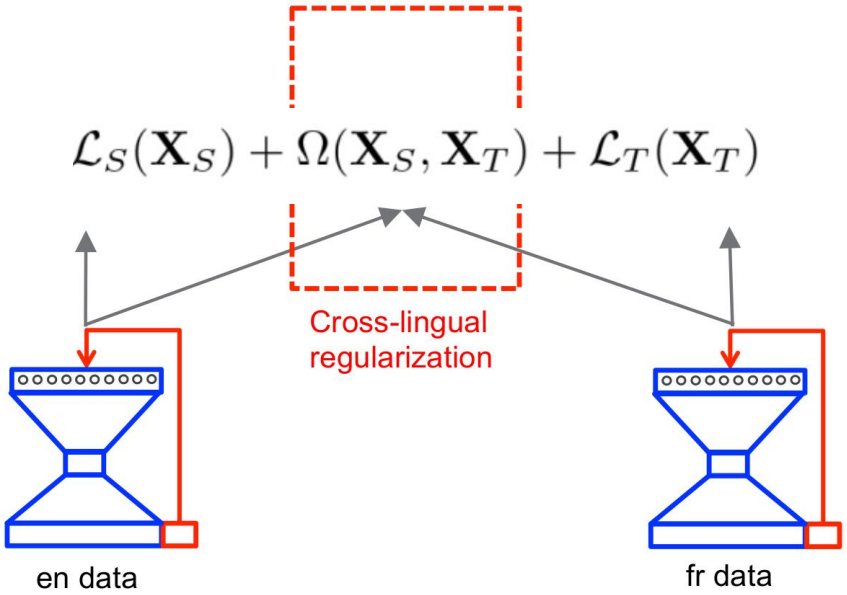


Image adapted from [Gouws et al., ICML-15]

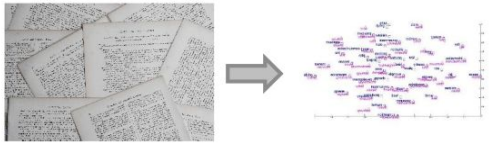
Joint CWLE Models *(selection)*

- **Using word-level cross-lingual signal: word alignments, bilingual dictionaries...**
  - Bilingual extensions of the monolingual skip-gram and CBOW models
    [Ammar et al., arXiv-15; Luong et al., NAACL-15; Guo et al., ACL-15; Shi et al., ACL-15]
  - Creating pseudo-bilingual corpora + training monolingual WE methods on such corpora
    [Gouws and Søgaard, NAACL-15; Duong et al., EMNLP-16; Adams et al., EACL-17]
- **Using sentence-level cross-lingual signal**
  - Compositional sentence model [Hermann and Blunsom, ACL-14]
  - Sentence-level bilingual skip-gram [Coulmance et al., EMNLP-15; Gouws et al., ICML-15]
  - Bilingual sentence autoencoders [Chandar et al., NeurIPS-14]
- **Using document-level cross-lingual signal**
  [Vulić and Moens, ACL-15, JAIR-16; Søgaard et al., ACL-15]

Joint versus Projection-Based Methods

Regardless of the source of supervision, there are two main strategies for inducing a bilingual/multilingual word embedding space:

1. **Joint embedding models**
   - Start from raw texts in two (or more) languages
   - Induce a bilingual (multilingual) space from scratch



2. **Post-hoc alignment models** (aka *projection* models)
   - Start from two independently pretrained monolingual embedding spaces
     - E.g., We apply word2vec on EN Wikipedia; then (independently) on ES Wikipedia
   - Learn the alignment/projection between the two monolingual spaces

CLWE Induction via Joint Models

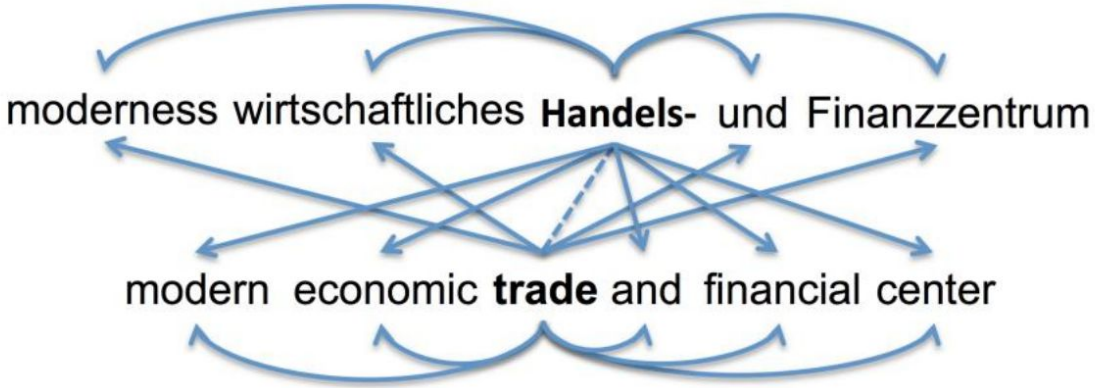
A number of models

Example: **Bilingual Skip-Gram**

Luong, M. T., Pham, H., & Manning, C. D. (2015, June). *Bilingual word representations with monolingual quality in mind*. In Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing (pp. 151-159).

Skip-Gram extended with cross-lingual context prediction

- Parallel data (mutual sentence translations) needed!
  - Automatic word alignment



CLWE Induction via Joint Models

Some shortcomings

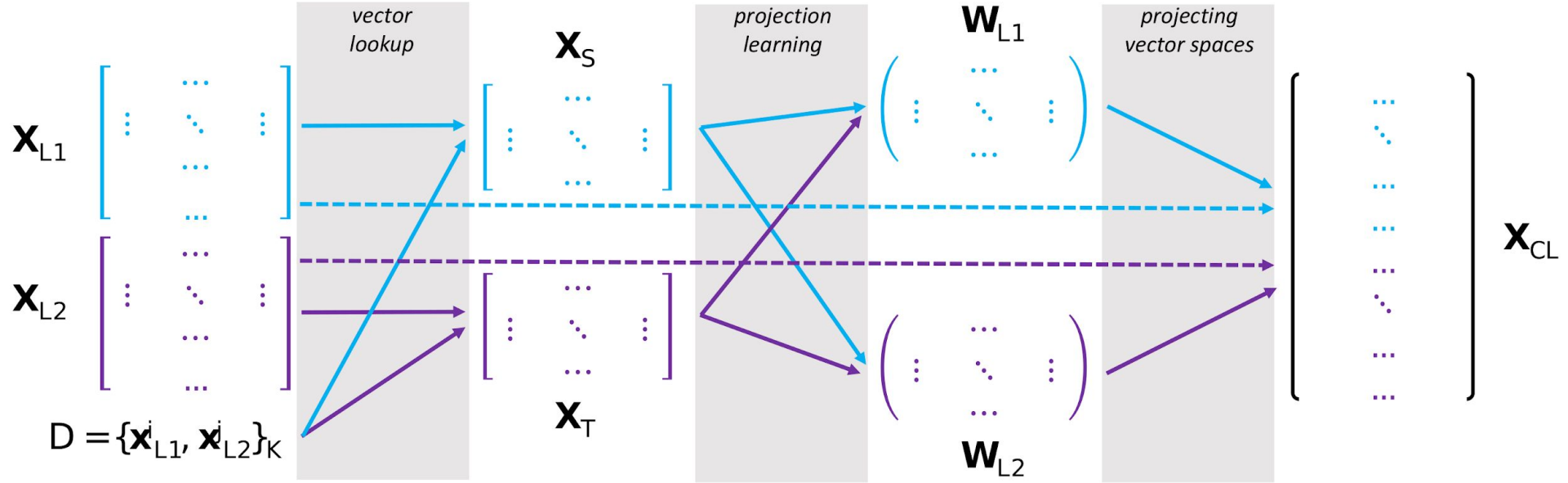
- Expensive model training for every pair of languages
- Bilingual models – not multilingual models
- Bilingual models not comparable, e.g., EN-DE vs. EN-ES
- Parallel sentences not so easily obtainable for all language pairs
  - Although there are extensions of Bilingual Skip-Gram that require only word-level supervision (i.e., word translations)

More elegant and less-resource demanding solution:

- Train monolingual vectors independently
- Light-weight post-hoc alignment between those spaces?
- Easy to induce truly multilingual spaces through post-hoc projections
- **Projection-based CLWE models**

Projection-based CLWE Learning

[Glavaš et al.; ACL 2019]



Post-hoc alignment of independently trained monolingual distributional word vectors

Alignment based on **word translation pairs** (dictionary **D**)

Projection-based CLWE Learning

Most models learn a single projection matrix $\mathbf{W}_{L1}$ (i..e., $\mathbf{W}_{L2} = \mathbf{I}$), but **bidirectional learning** is also common.

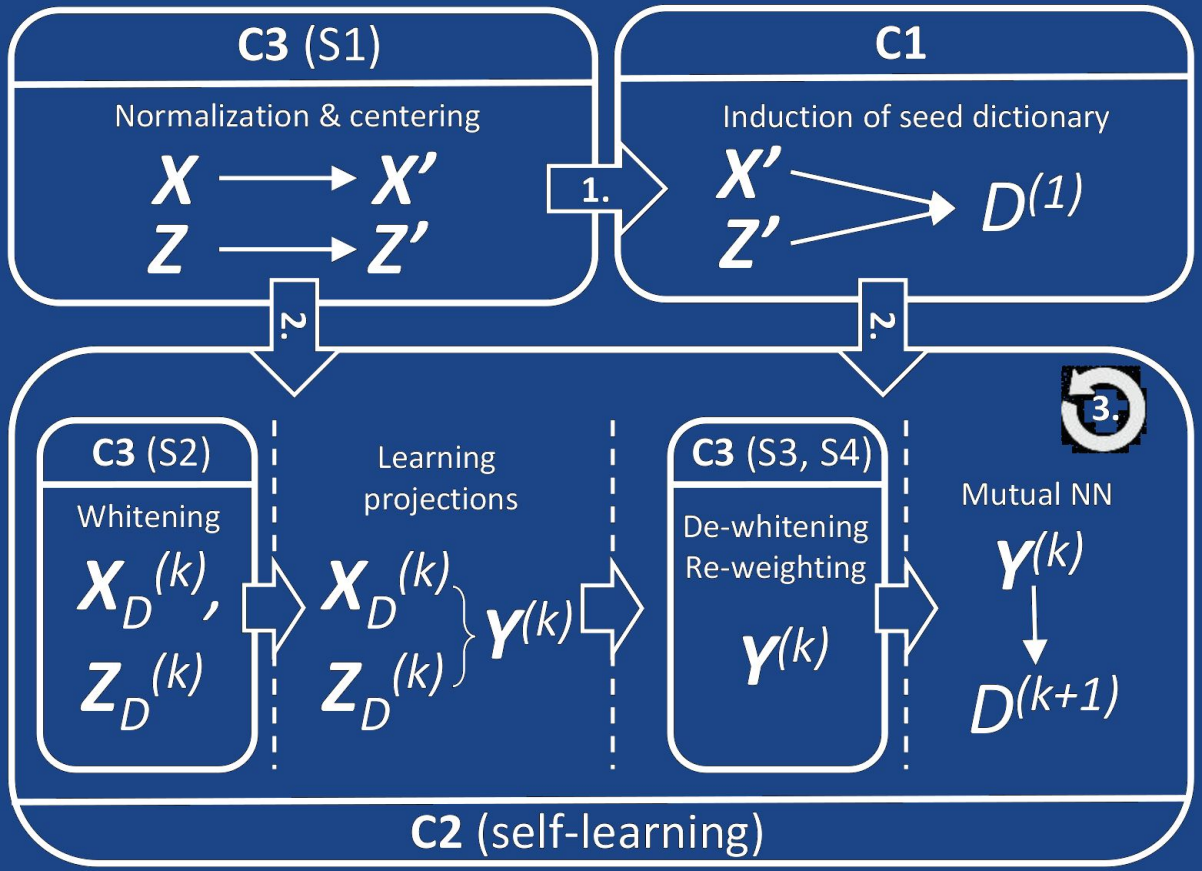
$$\begin{array}{r} \\ \text{bird} \\ \text{pretty} \\ \dots \\ \text{eat} \end{array} \begin{bmatrix} -1.18 & 0.21 & \dots & 0.11 \\ 0.23 & -0.53 & \dots & 0.34 \\ \dots & \dots & \dots & \dots \\ 0.78 & 1.33 & \dots & -0.47 \end{bmatrix} \mathbf{W}_{L1} = \begin{bmatrix} 0.59 & 1.01 & \dots & 0.37 \\ -0.34 & -0.27 & \dots & 0.41 \\ \dots & \dots & \dots & \dots \\ 0.81 & -0.31 & \dots & 0.29 \end{bmatrix} \begin{array}{l} \mathbf{X}_T \\ \text{Vogel} \\ \text{schön} \\ \dots \\ \text{essen} \end{array}$$

$\mathbf{X}_S$

How do we find the "optimal" projection matrix $\mathbf{W}_{L1}$?

- **Mean square error:** [Mikolov et al., arXiv-13] and most follow-up work
  ...except…
- **Canonical methods** [Faruqui et al., EACL-14; Lu et al., NAACL-15; Rotman et al., ACL-18]
- **Max-margin framework:** [Lazaridou et al., ACL-15; Mrkšić et al., TACL-17]
- **Relaxed Cross-Domain Similarity Local Scaling**: [Joulin et al., EMNLP-18]

Unsupervised and Weakly Supervised CLWE Induction



Typical focus: Seed dictionary induction and learning projections.

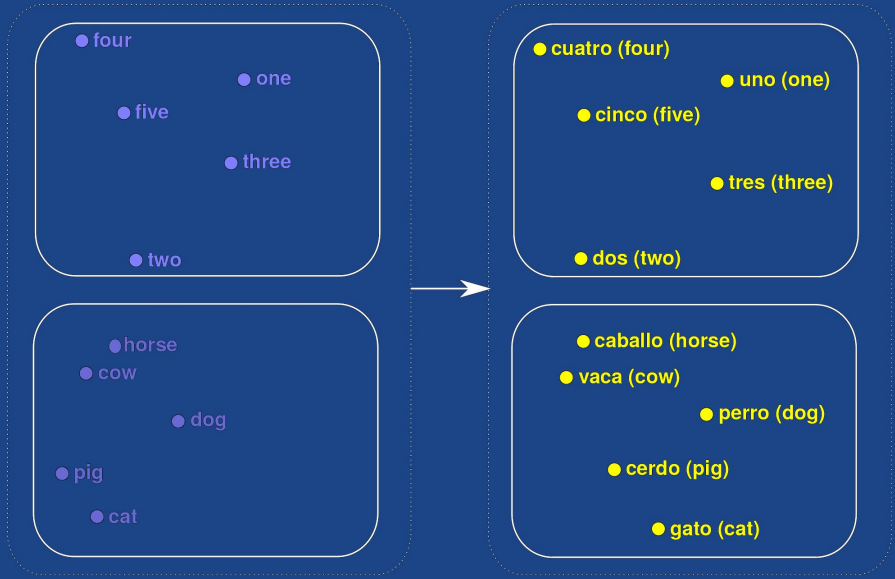
All we need are 100, 200, 500 word translation pairs… Or even nothing?

Many “tricks of the trade…”

(Approximate) Isomorphism

“. . . we hypothesize that, if languages are used to convey thematically similar information in similar contexts, these random processes should be approximately isomorphic between languages, and that this isomorphism can be learned from the statistics of the realizations of these processes, the monolingual corpora, in principle without any form of explicit alignment.”

[Miceli Barone, RepL4NLP-16]



Typical Evaluation: Bilingual Lexicon Induction (and Cross-Lingual Semantic Similarity)

en_morning			en_carpet		
Slavic+EN	**Germanic**	**Romance+EN**	**Slavic+EN**	**Germanic**	**Romance+EN**
en_daybreak	de_vormittag	pt_madrugada	en_rug	de_teppichboden	en_rug
en_morn	nl_krieken	it_mattina	bg_килим	nl_tapijten	it_moquette
bg_разсъмване	en_dawn	en_dawn	ru_ковролин	en_rug	it_tappeti
hr_svitanje	nl_zonsopkomst	pt_madrugadas	bg_килими	de_teppich	pt_tapete
hr_zore	sv_morgonen	es_madrugada	pl_dywany	en_carpeting	es_moqueta
bg_изгрев	de_tagesanbruch	it_nascente	bg_мокет	de_teppiche	it_tappetino
en_dawn	en_sunrise	en_morn	pl_dywanów	sv_mattor	en_carpeting
ru_утро	nl_opgang	es_aurora	hr_tepih	sv_matta	pt_carpete
bg_аврора	de_sonnenaufgang	fr_matin	pl_wykładziny	en_carpets	pt_tapetes
hr_jutro	nl_dageraad	fr_aurora	ru_ковер	nl_tapijt	fr_moquette
ru_рассвет	de_anbruch	es_amaneceres	ru_коврик	nl_kleedje	en_carpets
hr_zora	sv_morgon	en_sunrises	hr_ćilim	nl_vloerbedekking	es_alfombra
hr_zoru	en_daybreak	es_mañanero	en_carpeting	de_brücke	es_alfombras
pl_poranek	de_morgengrauen	fr_matinée	pl_dywan	de_matta	fr_tapis
en_sunrise	nl_zonsopgang	it_mattinata	ru_ковров	nl_matta	pt_tapeçaria
bg_зазоряване	nl_goedemorgen	pt_amanhecer	en_carpets	en_mat	it_zerbino

Semantic similarity ⟶ measuring distance in the induced cross-lingual space
Bilingual lexicons: *(en_morning, it_mattina), (en_carpet, hr_tepih), ...*

Bilingual Lexicon Induction (BLI)

Even BLI scores indicate that language similarity is quite important:

Model	Dict	EN–DE	IT–FR	HR–RU	EN–HR	DE–FI	TR–FR	RU–IT	FI–HR	TR–HR	TR–RU
PROC	1K	0.458	0.615	0.269	0.225	0.264	0.215	0.360	0.187	0.148	0.168
PROC	5K	0.544	0.669	0.372	0.336	0.359	0.338	0.474	0.294	0.259	0.290
PROC-B	1K	0.521	0.665	0.348	0.296	0.354	0.305	0.466	0.263	0.210	0.230
RCSLS	1K	0.501	0.637	0.291	0.267	0.288	0.247	0.383	0.214	0.170	0.191
RCSLS	5K	0.580	0.682	0.404	0.375	0.395	0.375	0.491	0.321	0.285	0.324

- Similar languages induce similar monolingual vector spaces
- It is easier to align such spaces
- As expected, cross-lingual transfer for more similar languages works more effectively...

Cross-Lingual Transfer with CLWEs

Use CLWEs for cross-lingual transfer of supervised NLP tasks?

Assumption: **zero-shot transfer**

- Only task- annotated data for the source language L_S, no annotated data in target language L_T

Steps:

1. Induce the bilingual shared word embedding space $\mathbf{X}_{TS}$
   - E.g., by projecting the target lang. space $\mathbf{X}_S$ to the source lang. Space $\mathbf{X}_T$
2. Train the (neural) model using the task-specific data in Ls
   - E.g., for *Named Entity Recognition*, train a *Bi-LSTM+classifier* using embeddings of source language words from the shared space $\mathbf{X}_{TS}$ as input
3. At prediction time, for texts in target language L_T, feed as input the embeddings of target language words from the same shared space $\mathbf{X}_{TS}$

New Paradigm: Multilingual Language Models

Multilingual and cross-lingual learning based on the **transfer learning formula**

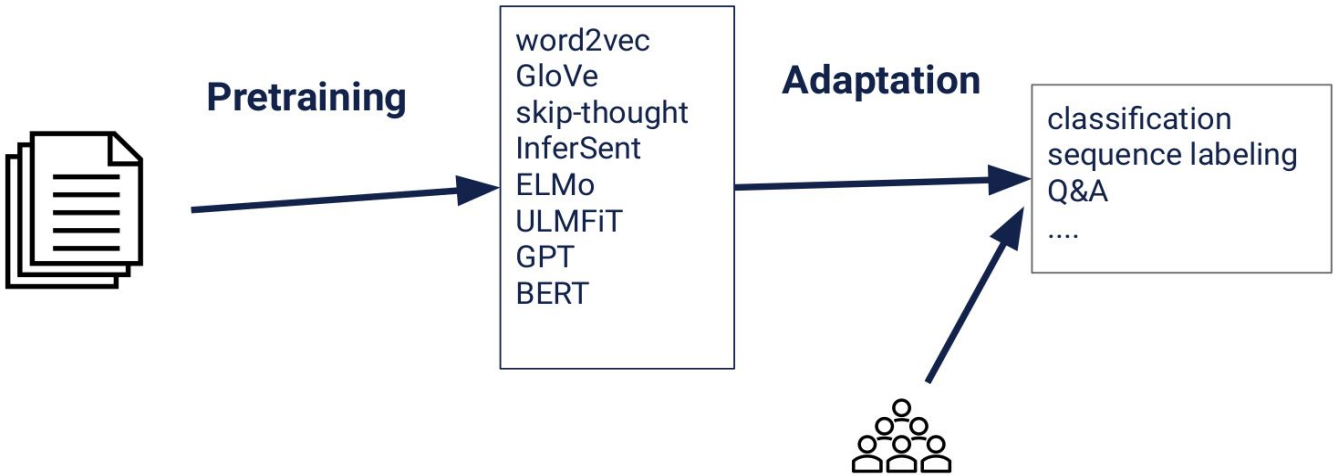


Image credit: NAACL 2019 Transfer learning tutorial

New Paradigm: Multilingual Language Models

Deep Transformer nets pretrained on large **multilingual corpora** via (masked) **language modeling** objectives

- Multilingual BERT, XLM-R, mT5

Automatically induces shared (subword) vocabulary across all languages

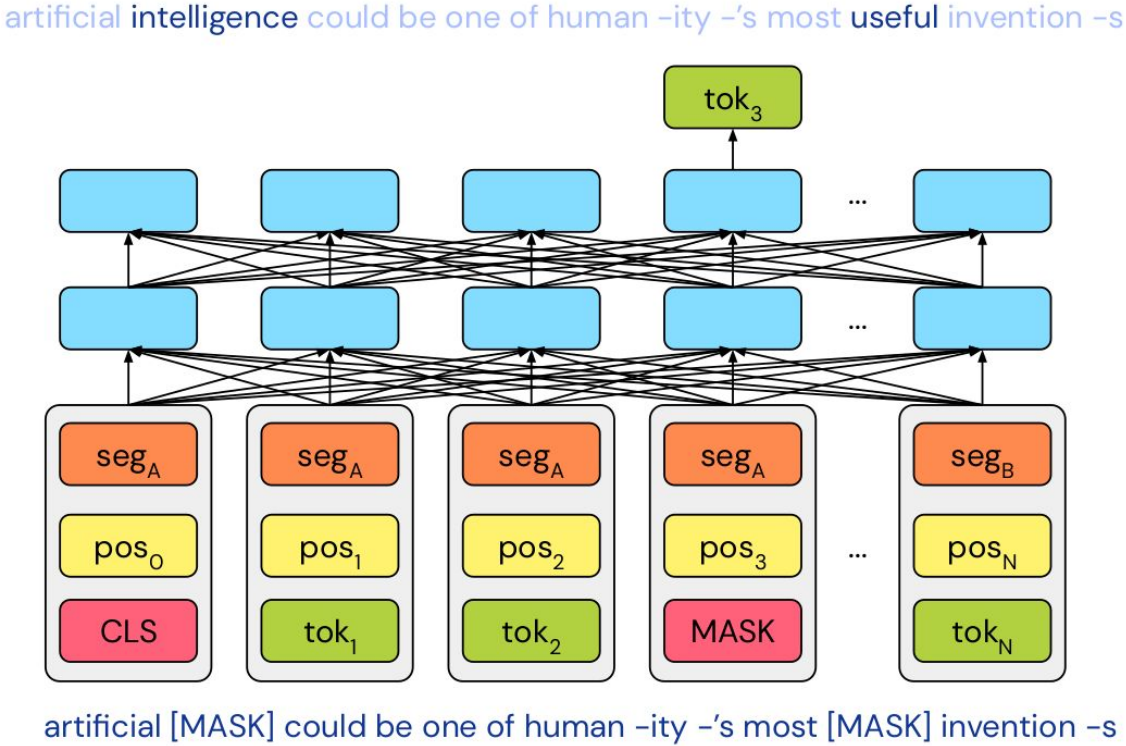
Unsupervised from the perspective of explicit cross-lingual signal

- Deemed very effective for zero-shot CL transfer

„Suprising cross-lingual effectiveness of BERT"

„mBERT surprisingly good at zero-shot CL model transfer"

How Do We Pretrain?



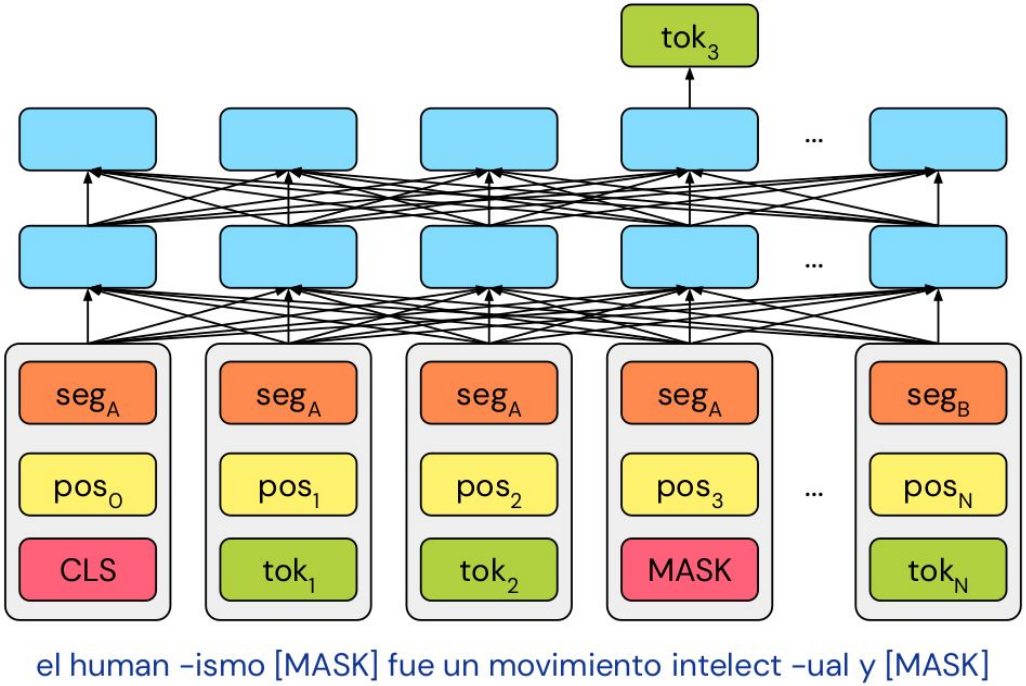
multilingual BERT (mBERT)

- Step 1: Combine corpora & learn joint subword vocab
- Step 2: Joint pre-training

Image courtesy of Sebastian Ruder

How Do We Pretrain?

el human -ismo renacentista fue un movimiento intelect -ual y filosófico



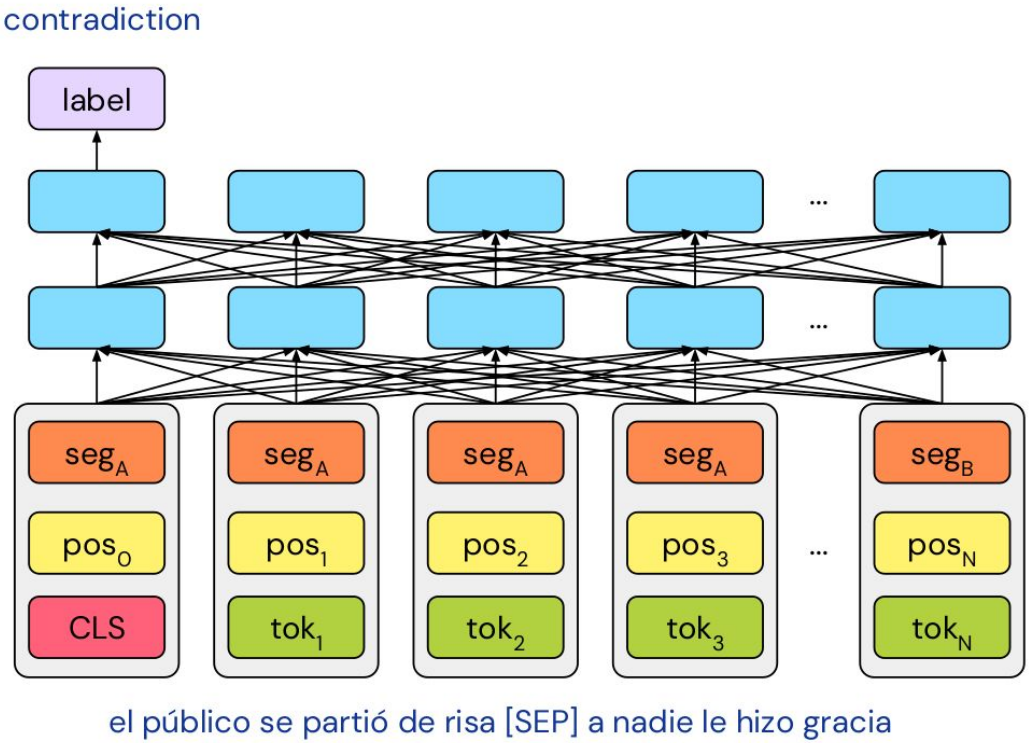
el human -ismo [MASK] fue un movimiento intelect -ual y [MASK]

multilingual BERT (mBERT)

- Step 1: Combine corpora & learn joint subword vocab
- Step 2: Joint pre-training

Image courtesy of Sebastian Ruder

How Do We Fine-Tune and Transfer?



- Step 1: Combine corpora & learn joint subword vocab
- Step 2: Joint pre-training
- Step 3: English fine-tuning
- Step 4: Zero-shot transfer

Image courtesy of Sebastian Ruder

Cross-Lingual Transfer with Multilingual LMs

Task: Zero-shot transfer to low-resource languages

Step 1:

Train a multilingual model.

Step 2:

Fine-tune model on a **task** in a high resource **source language.**

Animation Source: Google Research

Step 3:

Transfer and **evaluate** the model on a low resource **target language.**

Why?

Training **data** is **expensive** and not available for many languages, especially ones that are considered “low-resource”.

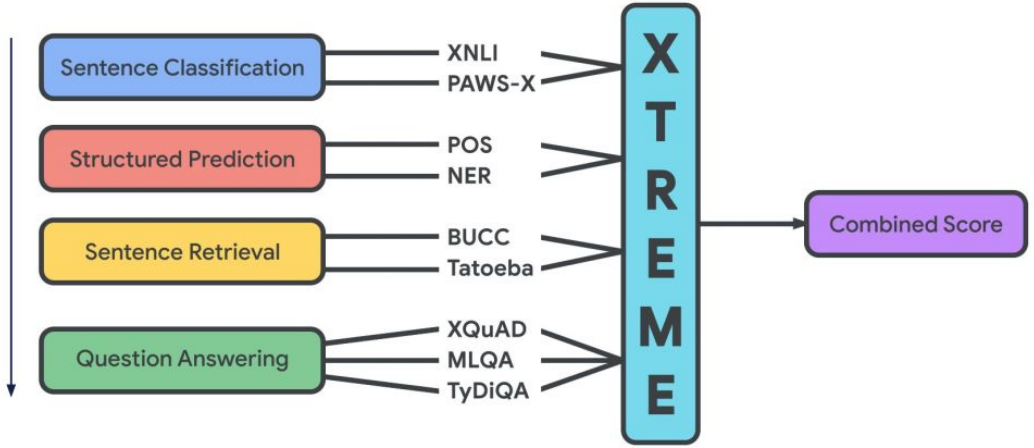
How Do We Evaluate Performance?

- Cross-lingual evaluation has mostly focused on a limited set of tasks under favourable conditions.
- A large-scale benchmark for evaluating cross-lingual generalisation has been missing.

XTREME (Cross-lingual Transfer Evaluation of Multilingual Encoders)

- Covers four core categories of NLP tasks on different levels of meaning across a large number of typologically diverse languages
- Provides a combined score as a measure of a model's **cross-lingual generalisation ability**

More complex natural language understanding required

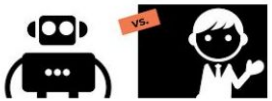


Slide courtesy of Sebastian Ruder

Besides XTREME, other cross-lingual tasks and data are available (e.g., XGLUE) - more recently XLM-R

How Do We Evaluate Performance?

It is not possible to evaluate on all tasks and on all languages...



Task difficulty: significant gap between human and model performance

Task diversity: transfer representations at different levels

Training efficiency: trainable with a modern GPU within one day

Multilinguality: tasks cover many languages and language families

Accessibility: a permissive license

Slide courtesy of Sebastian Ruder

How Do We Evaluate Performance?

It is not possible to evaluate on all tasks and on all languages… We need to sample a representative subset of languages (according to some well-defined criteria)...

Ponti, E. M., Glavaš, G., Majewska, O., Liu, Q., Vulić, I., & Korhonen, A. (2020). *XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 2362-2376.

Idea: selection according to the distribution of linguistic properties

- **Variety sampling** favors the inclusion of outlier languages

1. **Typological diversity**: entropy of distribution of linguistic properties
2. **Family index**: number of different families / sample size
3. **Geography index**: entropy of lang. distr. over 6 geographic macro-areas

	Range	XCOPA	TyDiQA	XNLI	XQUAD	MLQA	PAWS-X
Typology	[0, 1]	0.41	0.41	0.39	0.36	0.32	0.31
Family	[0, 1]	1	0.9	0.5	0.6	0.66	0.66
Geography	[0, ln 6]	1.67	0.92	0.37	0	0	0

So, Has mBERT Solved Zero-Shot Cross-Lingual Transfer?

– **No!** Settings in which they were evaluated were simply too favorable

„How multilingual is Multilingual BERT?" [Pires et al., ACL 19]

- **Tasks**: NER, POS; **Target languages**: DE, NL, ES

„Cross-lingual Ability of mBert: Empirical Study" [Karthikeyan et al., ICLR 20]

- **Tasks**: NER, NLI; **Target languages**: ES, HI, RU

– In most studies, the selected target languages were:

- (1) from the **same language family**,
- (2) with **large corpora in pretraining**

So, Has mBERT Solved Zero-Shot Cross-Lingual Transfer?

Lauscher, A., Ravishankar, V., Vulić, I., & Glavaš, G. (2020). *From Zero to Hero: On the Limitations of Zero-Shot Cross-Lingual Transfer with Multilingual Transformers*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4483-4499).

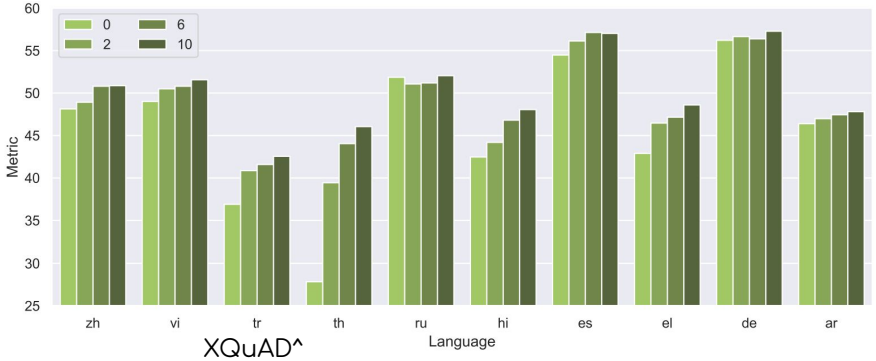
Task	Model	EN	ZH Δ	TR Δ	RU Δ	AR Δ	HI Δ	EU Δ	FI Δ	HE Δ	IT Δ	JA Δ	KO Δ	SV Δ	VI Δ	TH Δ	ES Δ	EL Δ	DE Δ	FR Δ	BG Δ	SW Δ	UR Δ
DEP	B	91.2	-43.9	-46.0	-28.1	-56.4	-36.1	-50.2	-30.7	-36.1	-17.1	**-60.1**	-56.1	-14.3	-	-	-	-	-	-	-	-	-
	X	92.0	**-85.4**	-44.2	-29.7	-54.6	-39	-49.5	-26.7	-39	-23.5	-80.5	-56.0	-16.3	-	-	-	-	-	-	-	-	-
POS	B	95.8	-38.0	-35.9	-16.0	-40.1	-33.4	-34.6	-21.9	-33.4	-19.8	**-46.1**	-42.0	-9.6	-	-	-	-	-	-	-	-	-
	X	96.3	-69.2	-27.7	-14.3	-37.1	-27.3	-31.9	-17.9	-27.3	-19.0	**-77.0**	-37.3	-10.7	-	-	-	-	-	-	-	-	-
NER	B	92.4	-23.3	-11.6	-10.7	**-31.7**	-11.1	-12.8	-3.8	-11.1	-2.6	-25.7	-13.8	-6.7	-	-	-	-	-	-	-	-	-
	X	91.6	**-34.8**	-6.2	-13.7	-24.6	-16.5	-8.0	-0.9	-16.5	-2.4	-30.1	-15.6	-2.2	-	-	-	-	-	-	-	-	-
XNLI	B	82.8	-13.6	-20.6	-13.5	-17.3	-21.3	-	-	-	-	-	-	-	-11.9	-28.1	-8.1	-14.1	-10.5	-7.8	-13.3	**-33.0**	-23.4
	X	84.3	-11.0	-11.3	-9.0	-13.0	-14.2	-	-	-	-	-	-	-	-9.7	-12.3	-5.8	-8.9	-7.8	-6.1	-6.6	**-20.2**	-17.3
XQuAD	B	71.1	-22.9	-34.2	-19.2	-24.7	-28.6	-	-	-	-	-	-	-	-22.1	**-43.2**	-16.6	-28.2	-14.8	-	-	-	-
	X	72.5	**-26.2**	-18.7	-15.4	-24.1	-22.8	-	-	-	-	-	-	-	-19.7	-14.8	-14.5	-15.7	-16.2	-	-	-	-

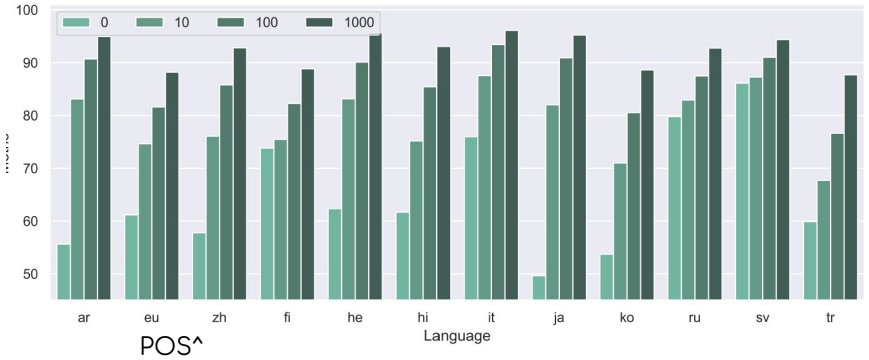
- B = mBERT (Base), X = XLM-R (Base)
- Drops **huge** for:
  1. Distant target languages and
  2. Target languages with small pretraining corpora

What about Few-Shot Transfer?

Should we focus more on few-shot transfer scenarios and quick annotation cycles?

Multilingual Pretraining → Massively Multilingual Transformer → Fine-tuning on target task in English → Few-shot fine-tuning on target task in target language → Prediction on target task in target language





[Lauscher et al., EMNLP-20; Zhao et al., ACL-21]

Few-Shot > (or >>) Zero-Shot

		K=0	K=1	K=2	K=4	K=8
MLDoc	EN	96.88	-	-	-	-
	DE	88.30	90.36 ± 1.48	90.77 ± 0.87	91.85 ± 0.83	91.98 ± 0.82
	FR	83.05	88.94 ± 2.46	89.71 ± 1.68	90.80 ± 0.88	91.01 ± 0.94
	ES	81.90	83.99 ± 2.35	85.65 ± 1.60	86.30 ± 1.85	88.46 ± 1.90
	IT	74.13	74.97 ± 2.04	75.29 ± 1.57	76.43 ± 1.41	78.12 ± 1.25
	RU	72.33	77.40 ± 4.27	80.57 ± 1.37	81.33 ± 1.33	81.91 ± 1.21
	ZH	84.38	87.18 ± 1.45	87.31 ± 1.53	88.33 ± 1.11	88.72 ± 1.05
	JA	74.58	76.23 ± 1.59	76.71 ± 2.12	78.60 ± 2.43	81.17 ± 1.72
MARC	EN	64.52	-	-	-	-
	DE	49.62	51.50 ± 1.58	52.76 ± 0.87	52.78 ± 1.00	53.32 ± 0.59
	FR	47.30	49.32 ± 1.34	49.70 ± 1.43	50.64 ± 0.94	51.23 ± 0.76
	ES	48.44	49.72 ± 1.24	49.96 ± 1.12	50.45 ± 1.22	51.25 ± 0.93
	ZH	40.40	43.19 ± 1.76	44.45 ± 1.36	45.40 ± 1.26	46.40 ± 0.93
	JA	38.84	41.95 ± 2.09	43.63 ± 1.30	43.98 ± 0.89	44.44 ± 0.69
XNLI	EN	82.67	-	-	-	-
	DE	70.32	70.58 ± 0.36	70.60 ± 0.34	70.61 ± 0.39	70.70 ± 0.50
	FR	73.57	73.41 ± 0.48	73.74 ± 0.46	73.57 ± 0.49	73.77 ± 0.44
	ES	73.71	73.84 ± 0.40	73.87 ± 0.44	73.74 ± 0.48	73.87 ± 0.46
	RU	68.70	68.81 ± 0.52	68.76 ± 0.54	68.87 ± 0.55	68.81 ± 0.77
	ZH	69.32	69.73 ± 0.94	69.75 ± 0.94	70.56 ± 0.76	70.62 ± 0.86
	AR	64.97	64.75 ± 0.36	64.82 ± 0.23	64.82 ± 0.23	64.94 ± 0.37
	BG	67.58	68.15 ± 0.69	68.19 ± 0.75	68.55 ± 0.67	68.32 ± 0.70
	EL	65.67	65.64 ± 0.40	65.73 ± 0.36	65.80 ± 0.41	66.00 ± 0.53
	HI	56.57	56.94 ± 0.82	57.07 ± 0.82	57.21 ± 1.14	57.82 ± 1.18
	SW	48.08	50.33 ± 1.08	50.28 ± 1.24	51.08 ± 0.62	51.01 ± 0.79
	TH	46.17	49.43 ± 2.60	50.08 ± 2.42	51.32 ± 2.07	52.16 ± 2.43
	TR	60.40	61.02 ± 0.68	61.20 ± 0.61	61.35 ± 0.49	61.31 ± 0.56
	UR	57.05	57.56 ± 0.85	57.83 ± 0.91	58.20 ± 0.93	58.67 ± 1.03
	VI	69.82	70.04 ± 0.59	70.14 ± 0.75	70.23 ± 0.63	70.41 ± 0.70
PAWSX	EN	93.90	-	-	-	-
	DE	83.80	84.14 ± 0.40	84.08 ± 0.42	84.04 ± 0.47	84.23 ± 0.66
	FR	86.90	87.07 ± 0.27	87.06 ± 0.37	87.03 ± 0.31	86.94 ± 0.41
	ES	88.25	87.90 ± 0.54	87.80 ± 0.56	87.84 ± 0.53	87.85 ± 0.75
	ZH	77.75	77.71 ± 0.37	77.63 ± 0.47	77.68 ± 0.51	77.82 ± 0.64
	JA	73.30	73.78 ± 0.75	73.71 ± 1.04	73.48 ± 0.69	73.79 ± 1.28
	KO	72.05	73.75 ± 1.30	73.11 ± 1.05	73.79 ± 0.92	73.31 ± 0.61

(...not only for token-level tasks)

Source Fine-Tuning Helps

	MLDoc		PAWSX			POS		NER	
	K=1	K=8	K=1	K=8		K=1	K=4	K=1	K=4
DE	-37.73	-7.67	-31.11	-30.82	RU	-15.89	-3.20	-48.19	-35.77
FR	-38.14	-13.21	-33.02	-32.34	ES	-9.51	-0.93	-63.98	-41.53
ES	-33.69	-14.38	-33.76	-33.97	VI	-7.82	-0.36	-54.41	-41.45
IT	-33.63	-12.62	-	-	TR	-15.05	-8.08	-54.35	-34.52
RU	-30.66	-11.08	-	-	TA	-13.72	-4.40	-34.70	-24.81
ZH	-37.31	-12.57	-23.74	-23.65	MR	-11.34	-3.63	-40.10	-25.68
JA	-29.82	-14.32	-20.97	-20.82	-	-	-	-	-
KO	-	-	-19.83	-19.68	-	-	-	-	-

Huge drops without source-language fine-tuning, using only target-language shots

Learning Even Better Representations...

- ...can be achieved with **bilingual supervision** (word translations of parallel data) [Wu & Conneau, ACL 20; Cao et al., ICLR 20; Hu et al., 2020]
- As with CLWEs: some bilingual/multilingual supervision ➔ better bilingual/multilingual representation space

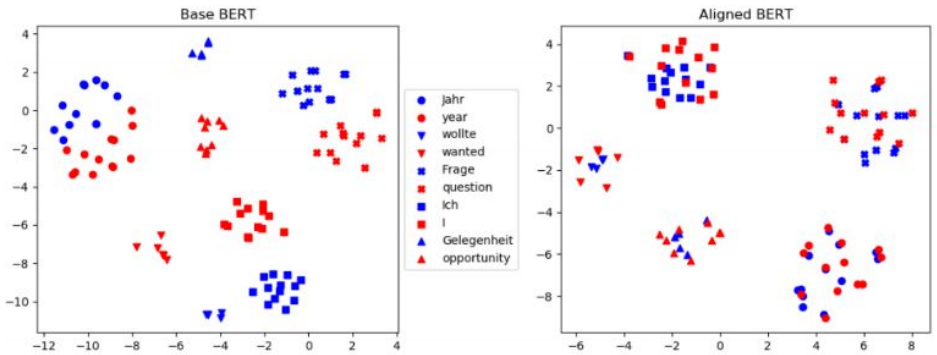
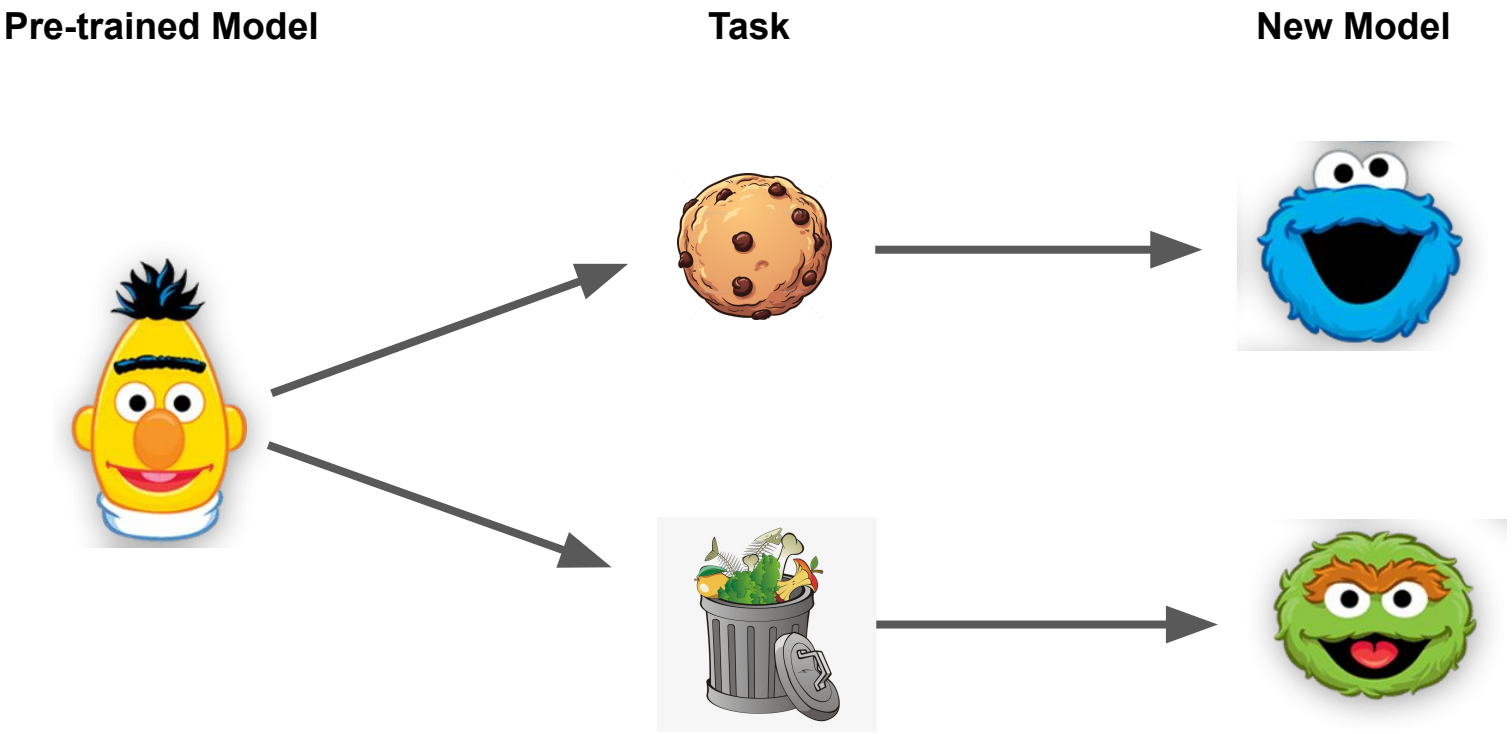


Image from [Cao et al., '20]

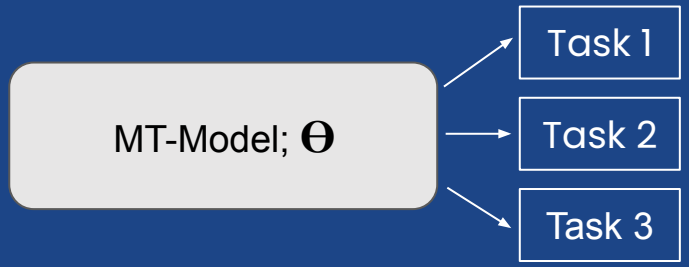
A Problem with the Standard Transfer Approach

Problems with **efficiency** and **modularity (reusability)** when fine-tuning the full model



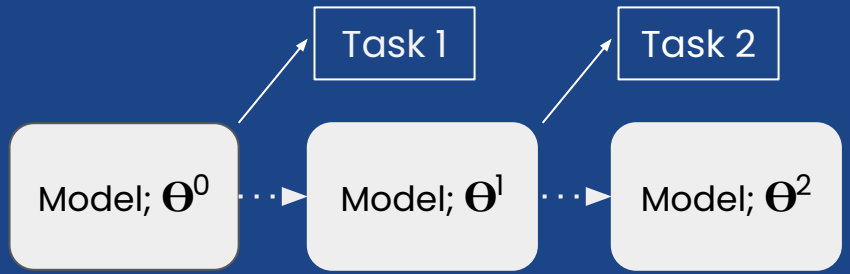
Problems of Multi-Task and Transfer Learning

Multi-Task Learning:



Catastrophic Interference:
Sharing all parameters Θ between tasks results in deterioration of performance for a subset of tasks.

Sequential Fine-Tuning:



Catastrophic Forgetting:
Sequential fine-tuning on tasks results in forgetting information learned in earlier stages of transfer learning.

Another Problem: The Curse of Multilinguality

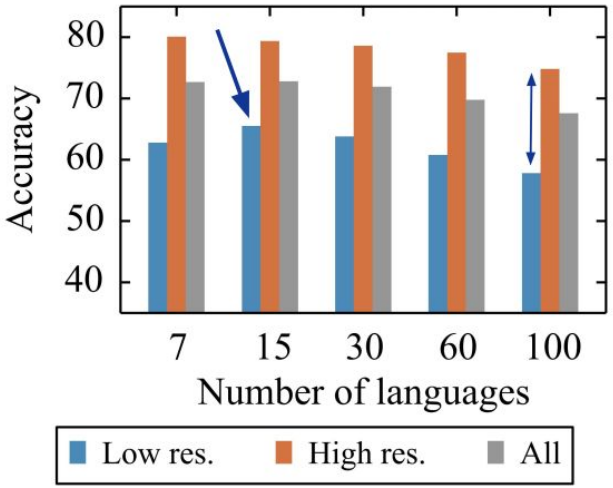
How do we “cram” all the languages into one single model?

The Curse of Multilinguality: Trade-off between model capacity and number of languages covered by the model

Languages compete for parameters in the model; for a fixed model capacity, **a model that covers fewer languages performs better**

Increasing the number of similar languages increases performance on low-resource languages (due to positive transfer) up to a point after which **dilution happens**

Adding more capacity helps to some extent
State-of-the-art models cover **about 100 languages** in their pre-training data

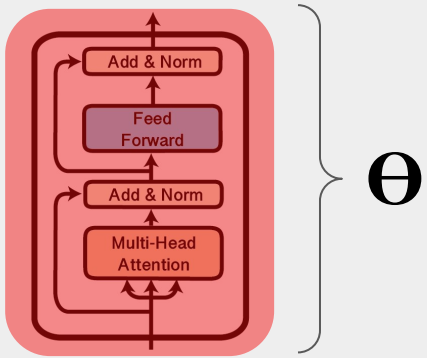


[Conneau et al., ACL-20]

Modular and Parameter-Efficient?

$$\theta \leftarrow \underset{\theta}{\operatorname{argmin}}\, L(D_{\text{NLI}}; \theta)$$

A single Transformer (encoder) layer



D_{NLI} = NLI Dataset

L = Loss function, e.g. cross entropy loss

θ = Parameters of the model

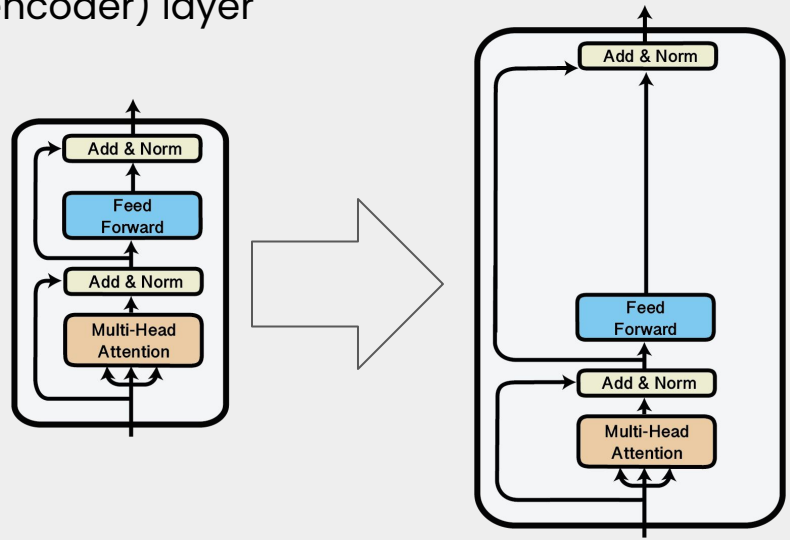
= Parameters are frozen

= Parameters are fine-tuned

Modular and Parameter-Efficient: Adapters

$$\theta \leftarrow \underset{\theta}{\operatorname{argmin}}\, L(D_{\text{NLI}}; \theta)$$

A single Transformer (encoder) layer



= Parameters are frozen

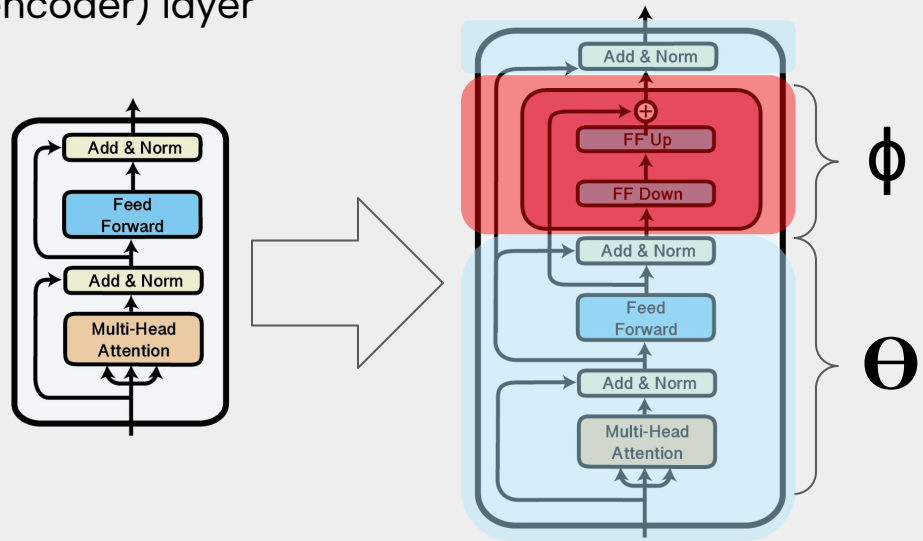
= Parameters are fine-tuned

Houlsby, Neil, et al. "Parameter-Efficient Transfer Learning for NLP." *International Conference on Machine Learning*. 2019.

Modular and Parameter-Efficient: Adapters

$$\phi \leftarrow \underset{\phi}{\operatorname{argmin}}\, L(D_{\text{NLI}}; \theta, \phi)$$

A single Transformer (encoder) layer



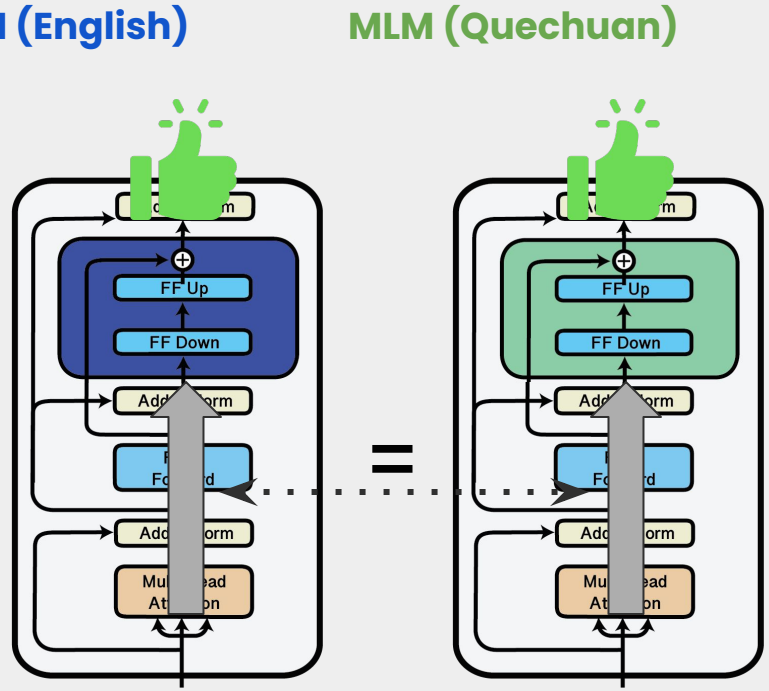
Adapter parameters ϕ are **encapsulated** between transformer layers with parameters θ which are frozen

= Parameters are frozen

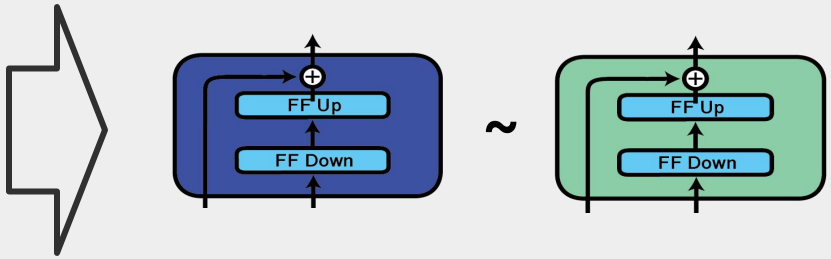
= Parameters are fine-tuned

Houlsby, Neil, et al. "Parameter-Efficient Transfer Learning for NLP." *International Conference on Machine Learning*. 2019.

Encapsulated Adapters?



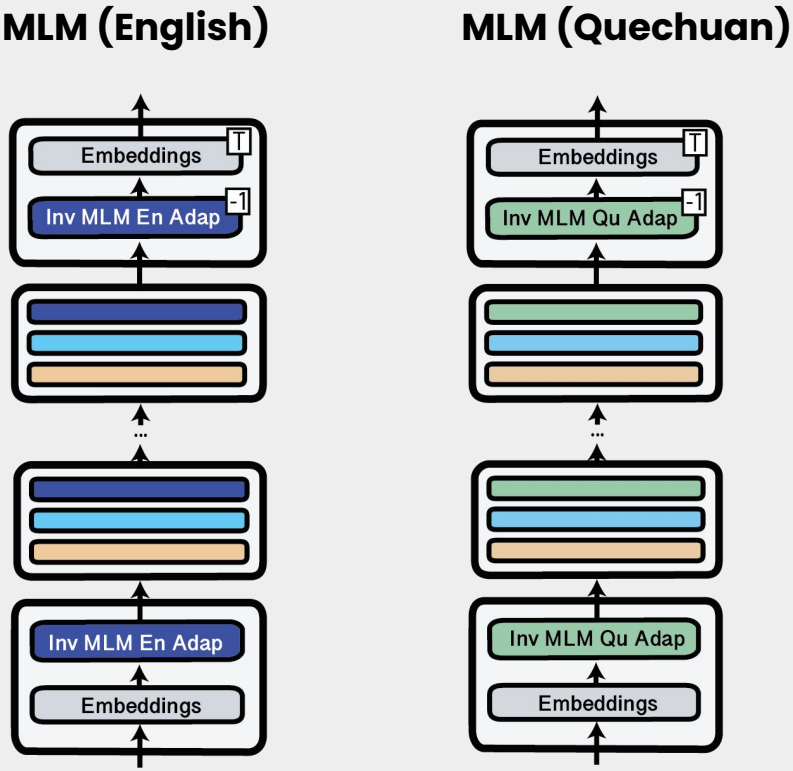
- Adapters **learn transformations** that make the underlying model **more suited** to a task or language.
- Using masked language modelling (MLM), we can learn **language-specific transformations** for e.g. **English** and **Quechua**.
- As long as the underlying model is kept fixed, these transformations are **roughly interchangeable**.



MAD-X: An Adapter-Based Framework for Transfer

Step 1: Train Language Adapters

We train **language adapters** for the **source language** and the **target language** with masked language modelling on Wikipedia.

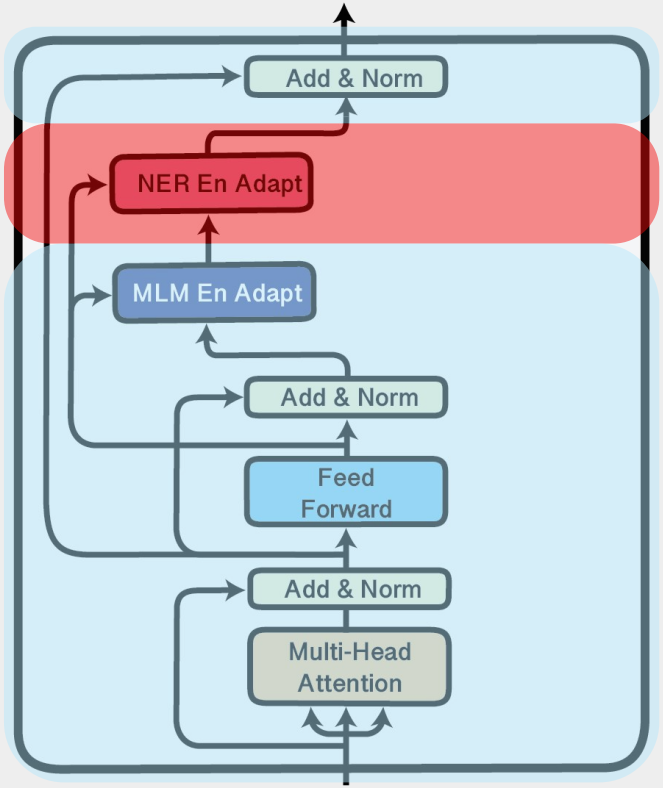


MAD-X

Step 2: Train a Task Adapter

We **train** **task adapters** in the source language **stacked** on top of the source **language adapter**.

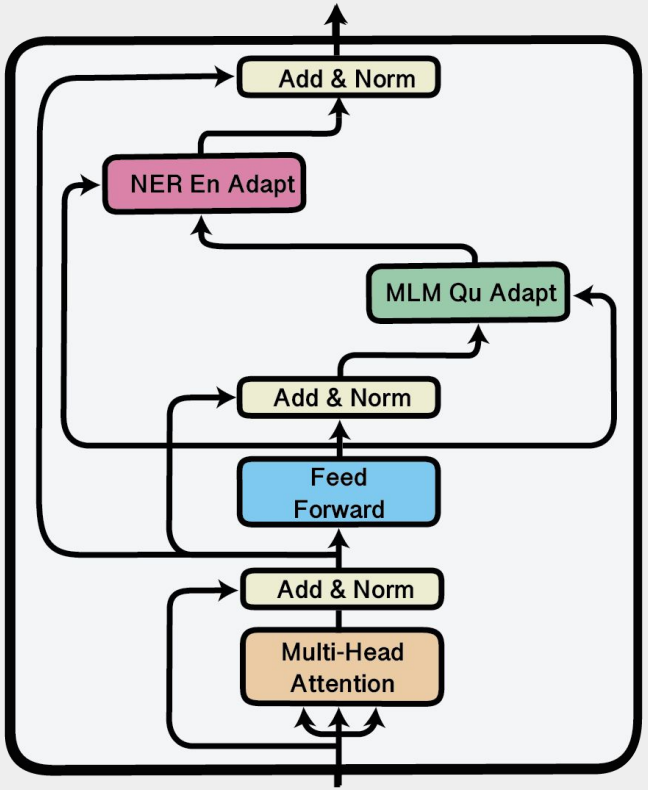
The language adapter ϕ_l as well as the transformer weights θ are frozen while only the task adapter parameter ϕ_t are trained.



MAD-X

Step 3: Zero-Shot transfer to unseen language

We **replace** the **source** language adapter with the **target** language adapter, while **keeping** the "language agnostic" **task adapter**.



Datasets: Inclusion of Diverse and Low-Resource

NER: WikiAnn Dataset We chose a diverse set of languages from **different language families**.

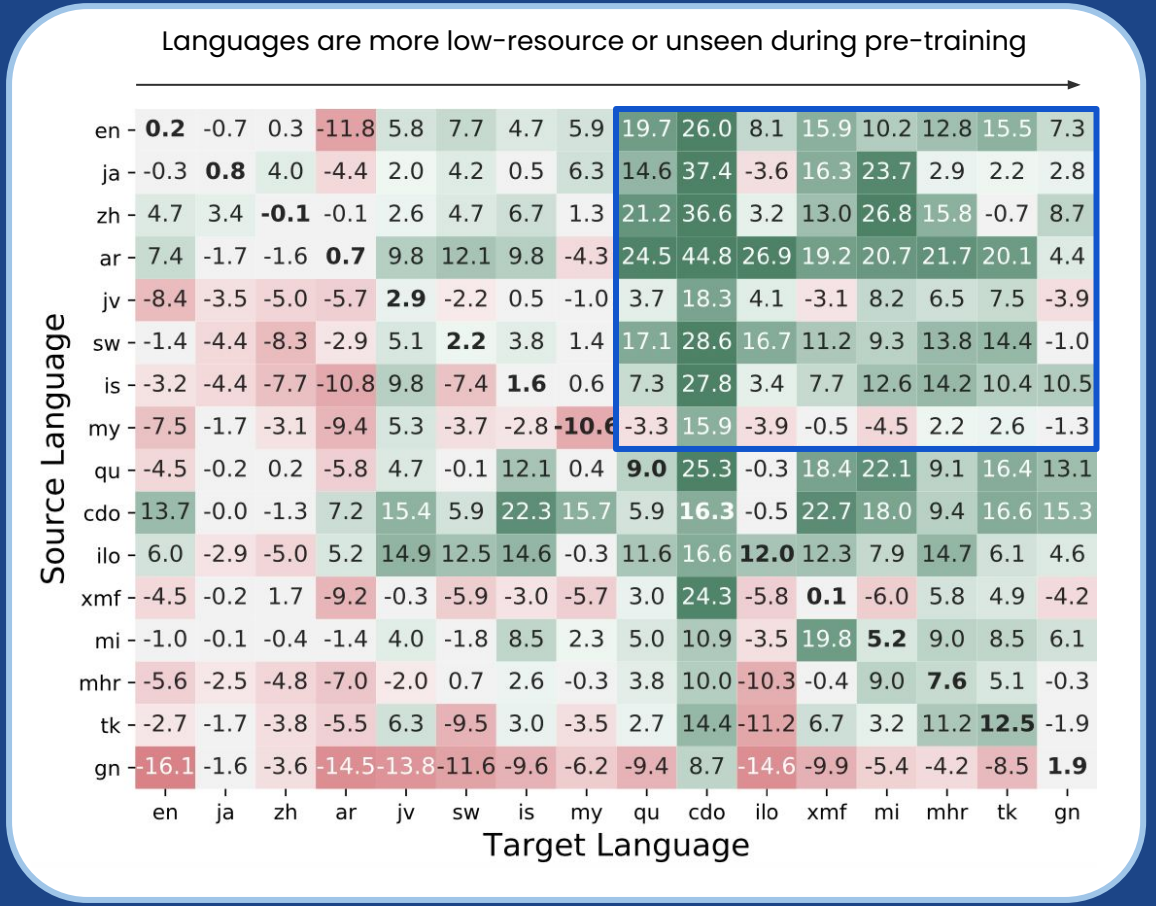
XQuAD (Cross-lingual Question Answering Dataset)

XCOPA (Ponti et al. 2020b)

Language	ISO code	Language family	# of Wiki articles	Covered by SOTA?
English	en	Indo-European	6.0M	✓
Japanese	ja	Japonic	1.2M	✓
Chinese	zh	Sino-Tibetan	1.1M	✓
Arabic	ar	Afro-Asiatic	1.0M	✓
Javanese	jv	Austronesian	57k	✓
Swahili	sw	Niger-Congo	56k	✓
Icelandic	is	Indo-European	49k	✓
Burmese	my	Sino-Tibetan	45k	✓
Quechua	qu	Quechua	22k	
Min Dong	cdo	Sino-Tibetan	15k	
Ilokano	ilo	Austronesian	14k	
Mingrelian	xmf	Kartvelian	13k	
Meadow Mari	mhr	Uralic	10k	
Maori	mi	Austronesian	7k	
Turkmen	tk	Turkic	6k	
Guarani	gn	Tupian	4k	

Relative **F1 improvement** of **MAD-X^{Large} over XLM-R^{Large}** in cross-lingual NER transfer.

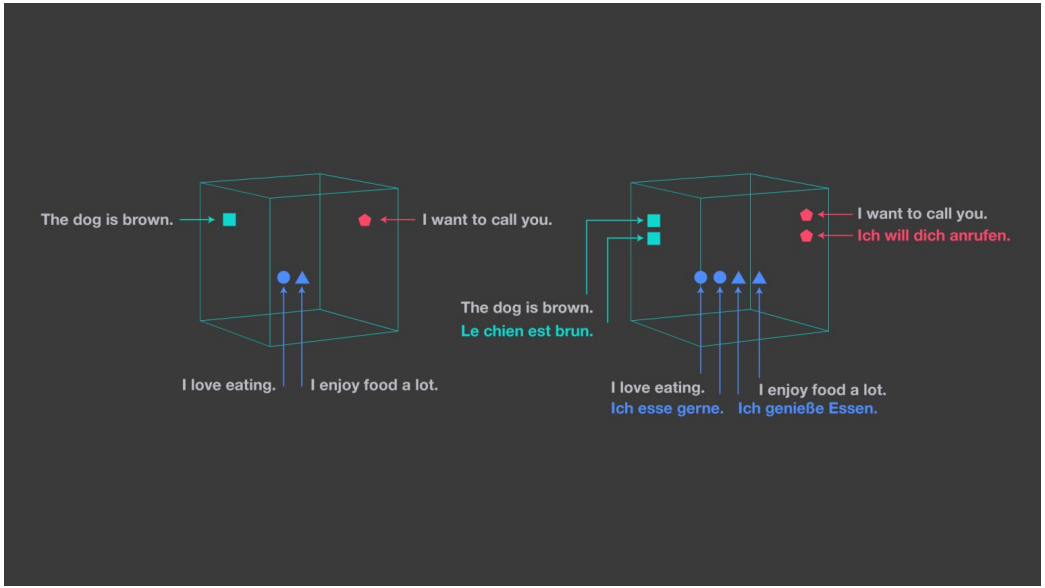
Top right corner represent the **realistic scenario** of **transfering from high resource to low resource**



Multilingual Sentence Encoders

Off-the-shelf multilingual Transformers (e.g., mBERT) are not good sentence encoders

Multilingual sentence encoders: multilingual and cross-lingual information search and semantic similarity



A shared **cross-lingual sentence space…**

Image courtesy of Meta AI

Multilingual Sentence Encoders

Multilingual language models can be transformed into sentence encoders, typically via **contrastive learning** (or the so-called **dual-encoder networks**) - this requires another lecture…

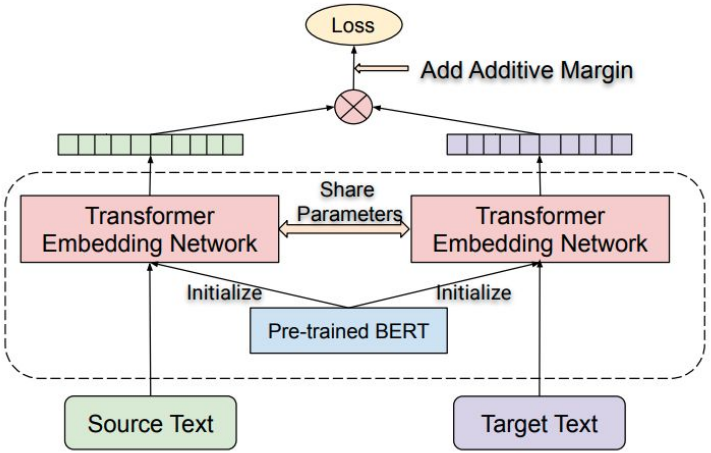


Image courtesy of Google Research

Some popular multilingual sentence encoders:

- LaBSE
- multilingual USE
- SBERT models

(some parallel data still needed…)

Read more about contrastive learning here:
https://lilianweng.github.io/posts/2021-05-31-contrastive/

Try multilingual sentence encoders yourself:
https://www.sbert.net/

Multilingual and Cross-Lingual NLP and IR: How to Cope?

Better Models and Algorithms:

- sophisticated modeling/training methods - know NLP/ML
- linguistically informed methods - know linguistics
- task knowledge - know your task

Better Data and Evaluation:

- every piece of relevant data can help - be resourceful
- make data if necessary - be connected
- track progress with challenging (and natural!) evaluation data

Better Adaptation:

- leverage similarity between languages
- adapt quickly to low-data regimes and new domains

Slide adapted from: CMU's Multilingual NLP Course

Take-Home Messages

(We only scratched the surface in this seminar…)

- Multilingual and cross-lingual NLP is a vibrant research field in the mission of democratising language technology
- Too many domains, too many languages, dialects -> we need general and adaptable solutions, we need to learn from whatever we've got...
- We have covered (at a very shallow level) high-level approaches as well as lower-level cutting-edge approaches to multilingual and cross-lingual NLP
  - cross-lingual word embeddings
  - massively multilingual language models
  - multilingual representation learning; cross-lingual transfer methods
  - some more advanced topics: few-shot learning, transfer via adapters, sentence encoders
- Despite positive trends, many languages are still left behind (and difficult to work with)

Advanced Topics

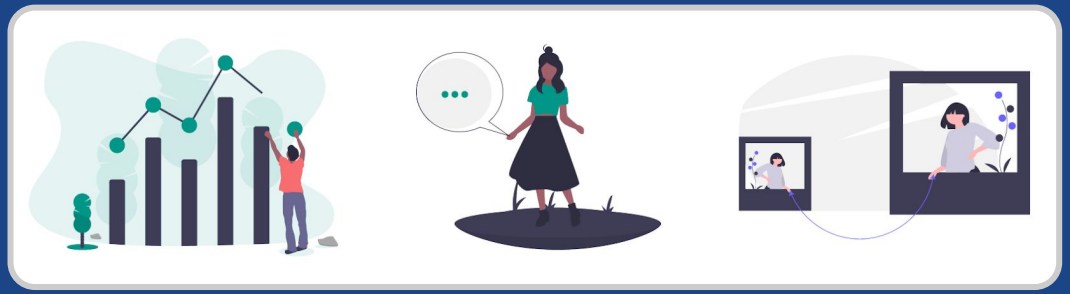
(We only scratched the surface in this seminar…)

- Active learning
- Meta-learning and few-shot adaptation strategies
- Data annotation and resource creation in low-resource languages
- Model adaptation to languages with unseen scripts
- Induction of linguistic structure from pretrained multilingual LMs
- Semantic specialisation of general-purpose models
- Learning multilingual word, sentence, and document encoders
- Unsupervised and weakly supervised Neural Machine Translation
- Injection of linguistic and world knowledge into multilingual text-based models
- Multi-modal multilingual modeling
- Creative applications of multilingual models
- Multilingual speech recognition
- Speech translation

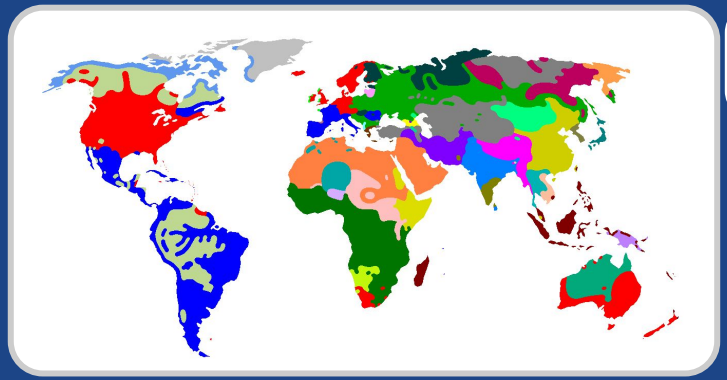
(Multilingual) Natural Language Processing

...and its many applications

Conversational Systems Virtual Assistants Information Search Question Answering



Digital Education Language Learning Assisted Translation Fact Checking



The grand challenge of multilinguality

Digital language divide versus equal opportunities

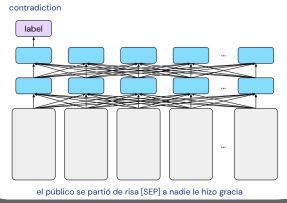
7,000+ languages; a wide spectrum of tasks and domains

- High-resource languages
- Medium-resource
- Low-resource
- Endangered

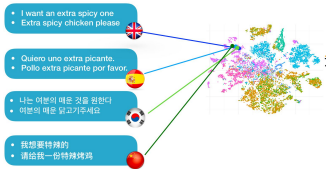
Towards Inclusive, Sustainable, Equitable Multilingual TOD

Widening the global reach of NLP: Far-reaching technological and socioeconomic consequences

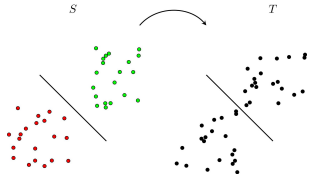
Deep (machine) learning



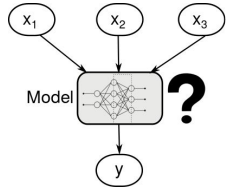
Multilingual representation learning



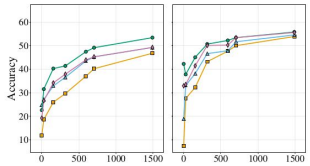
Cross-lingual knowledge transfer



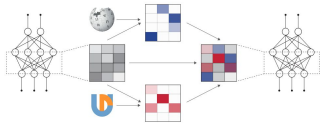
Transparency



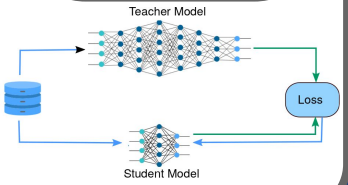
Sample Efficiency



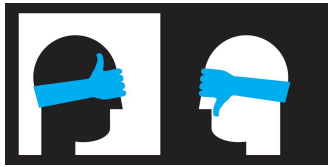
Modularity



Model Compactness



Fairness and Debiasing



Plus Other Crucial Aspects: *Cross-Cultural Adaptation, Multi-Modal Learning, Commonsense and World Knowledge, User Experience*

A Cross-Lingual Space of Thankyous!

ivi250@cam.ac.uk

Massive thanks and credits to Graham Neubig, Yulia Tsvetkov, Goran Glavaš, Sebastian Ruder, Edoardo Ponti, and Jonas Pfeiffer for sharing their amazing slides...

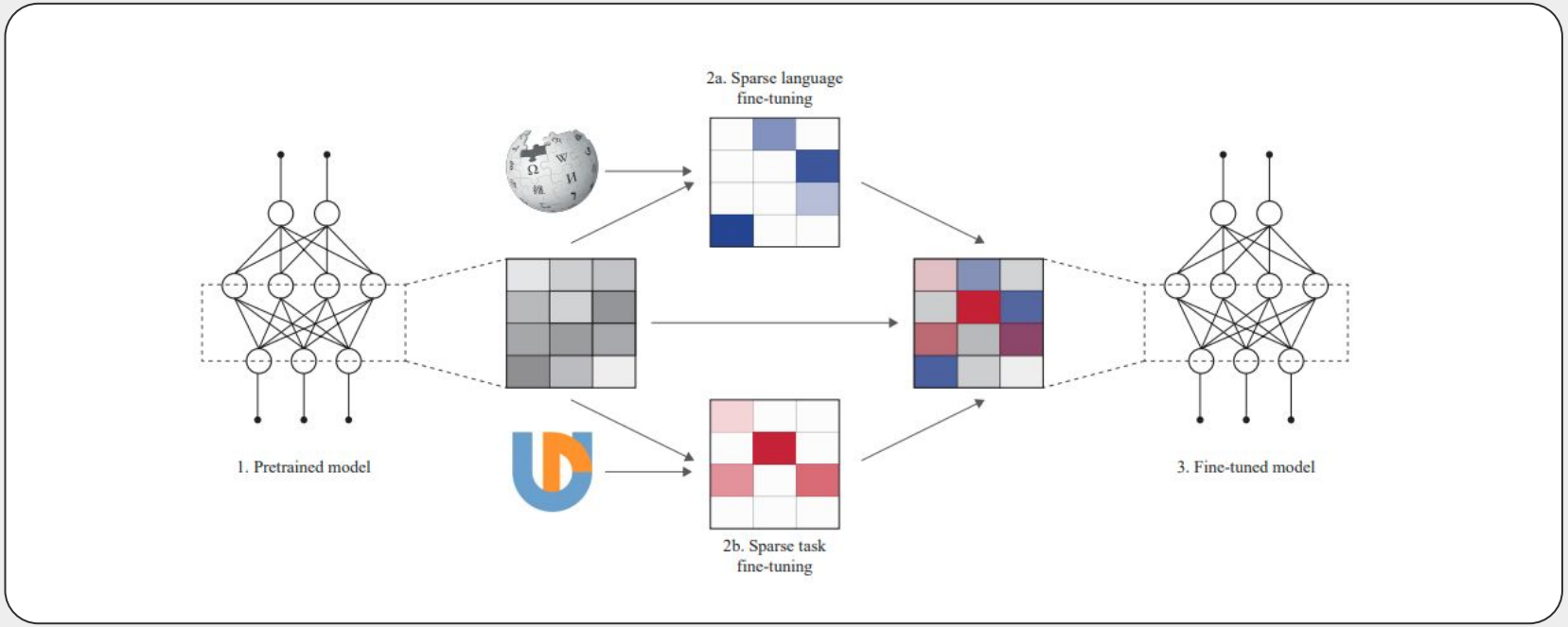
Further Reading

- CMU's course on multilingual NLP:
  http://demo.clab.cs.cmu.edu/11737fa20/ and http://phontron.com/class/multiling2022/
- A survey of cross-lingual word embedding models:
  https://jair.org/index.php/jair/article/view/11640/26511
- Modeling Language Variation and Universals: A Survey on Typological Linguistics for NLP
  https://arxiv.org/pdf/1807.00914.pdf
- Emerging Cross-Lingual Structure in Pretrained Language Models
  https://www.aclweb.org/anthology/2020.acl-main.536.pdf
- Choosing Transfer Languages for Cross-Lingual Learning
  https://www.aclweb.org/anthology/P19-1301.pdf
- From Zero to Hero: On the Limitations of Zero-Shot Language Transfer
  https://www.aclweb.org/anthology/2020.emnlp-main.363.pdf
- Decolonising Speech and Language Technology
  https://www.aclweb.org/anthology/2020.coling-main.313.pdf
- The State and Fate of Linguistic Diversity and Inclusion in the NLP World
  https://www.aclweb.org/anthology/2020.acl-main.560.pdf
- MAD-X: An Adapter-based Framework for Multi-task Cross-lingual Transfer
  https://www.aclweb.org/anthology/2020.emnlp-main.617.pdf

A Selection of More Advanced Topics

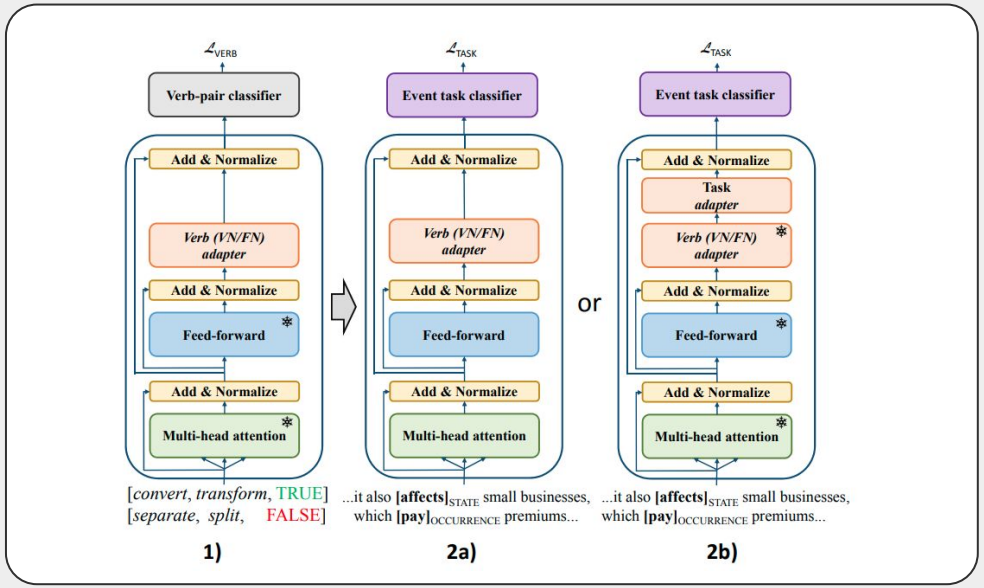
Research Branch 1: Other Parameter-Efficient Approaches

[Ansell et al., ACL-22] learn sparse and composable language and task masks that can be (re)combined for different tasks and languages - reusability and parameter efficiency



Research Branch 2: Storing External (Linguistic) Knowledge?

[Majewska et al., ACL-21] store VerbNet and FrameNet knowledge into dedicated VN/FN adapters



A more general approach:

Storing heterogeneous external (linguistic and world) knowledge into dedicated adapters, stacking and combining them...

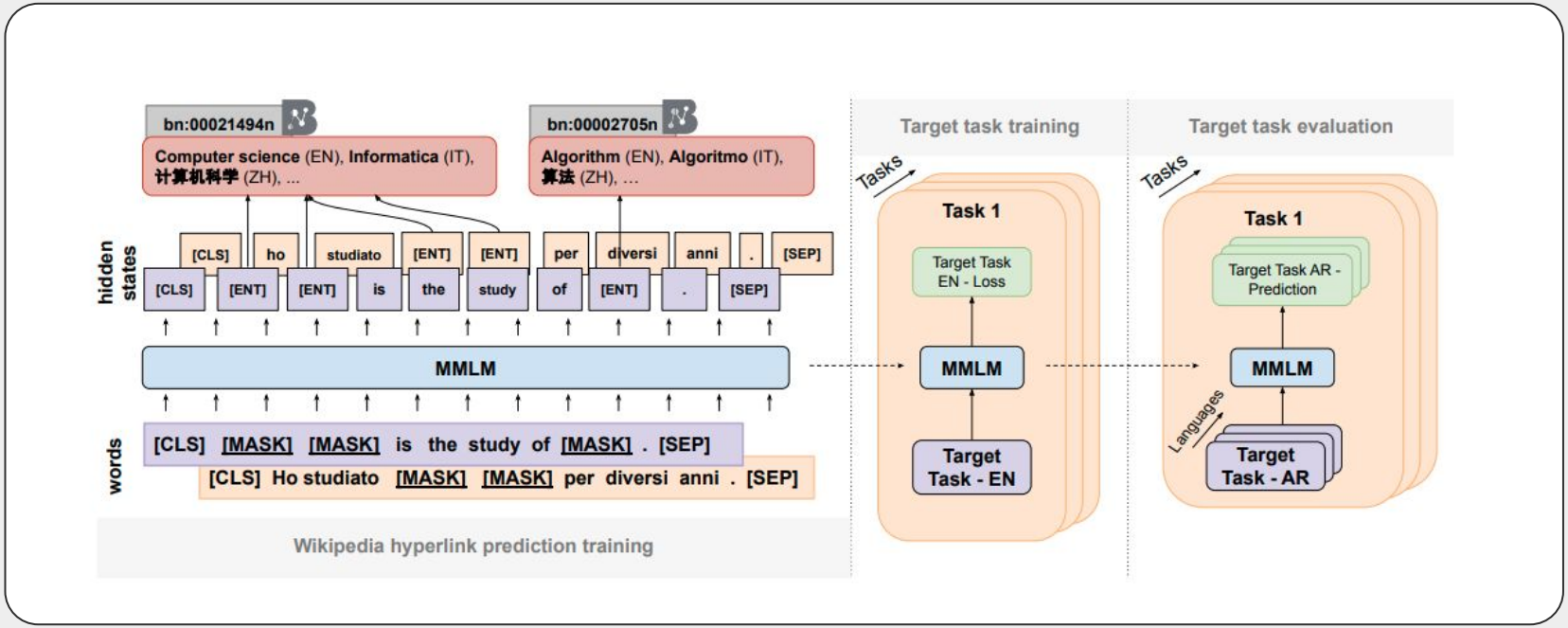
Gains on event processing tasks (trigger and argument identification and classification): *TempEval and ACE*

Transferring external knowledge from English to other languages

Our linguistic resources and knowledge bases do complement our (still) “distributional-only” models

Research Branch 3: Intermediate Tasks (Beyond MLM-ing)

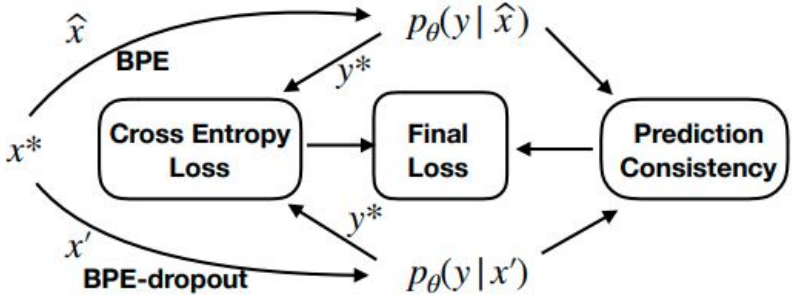
[Calixto et al., NAACL-21] propose to further fine-tune the multilingual Transformer on an intermediate task: Wikipedia hyperlink prediction -> a sort of cross-lingual STILT-ing (again using some additional external knowledge) -> Improvements in zero-shot transfer performance



Research Branch 4: Better Tokenization/Segmentation

[Wang et al., NAACL-21] propose a multi-view subword regularisation method which makes the models more robust to suboptimal segmentation during fine-tuning
-> Improvements in zero-shot transfer performance

en	excitement	**fr**	excita/tion
de	Auf/re/gung	**pt**	excita/ção
el	εν/ϑ/ουσι/ασμός	**ru**	волн/ение



[Rust et al., ACL-21] show that multilingual models with dedicated (i.e., monolingual) tokenisers close the gap to monolingually trained Transformers… It's not only about the dataset size - know your tokeniser -> more recently, get rid of language-specific and suboptimal tokenisers [Clark et al., arXiv-21]

Research Branch 5: (Creating New) Evaluation Data

It is not possible to evaluate on all tasks and on all languages…
(More research on this front recently - *MaskhaNER*, *AmericasNLI*…)

[Ponti et al., EMNLP-20] quantify some criteria for their **variety sampling** to include also outlier languages

- **Typological diversity:** entropy of distribution of linguistic properties
- **Family index:** number of different families / sample size
- **Geography index:** entropy of languages distributed over 6 macro-areas

	Range	XCOPA	TyDiQA	XNLI	XQUAD	MLQA	PAWS-X
Typology	[0, 1]	0.41	0.41	0.39	0.36	0.32	0.31
Family	[0, 1]	1	0.9	0.5	0.6	0.66	0.66
Geography	[0, ln 6]	1.67	0.92	0.37	0	0	0

We need training and task data beyond the usual suspects (NER, POS, parsing). What about massively multilingual NLU? What about dialogue? [Razumovskaiia et al., JAIR 2022]

Research Branch 6: More Modularity and Reusability

- Most combinations of NLP tasks and languages lack training examples, but we have *some* knowledge for *some* task-language combinations.
- Neural parameters for **unseen task-language combinations** can be approximated by transferring knowledge both from other tasks *and* from other languages, i.e., from **seen combinations**.
- For instance, training data for NER in Vietnamese and POS tagging in Wolof should lead to accurate NER predictions in Wolof.
- Core assumption: the neural parameter space is **structured** and can be factorised into (latent) task and language components -> factorising the **task x language x parameter tensor**.

[Ponti et al., TACL-21, arXiv-22]

Towards modularity of representations?

Research Branch 7: Task-Specialised Machine Translation

- ‘Translate-test’ and ‘translate-train’ are strong baselines for cross-lingual NLP
- Control the output of your NMT systems -> reward translations that yield better transfer performance (in few-shot setups)

[Ponti et al., arXiv-22]

Research Branch 8: Typology and Multilingual NLP

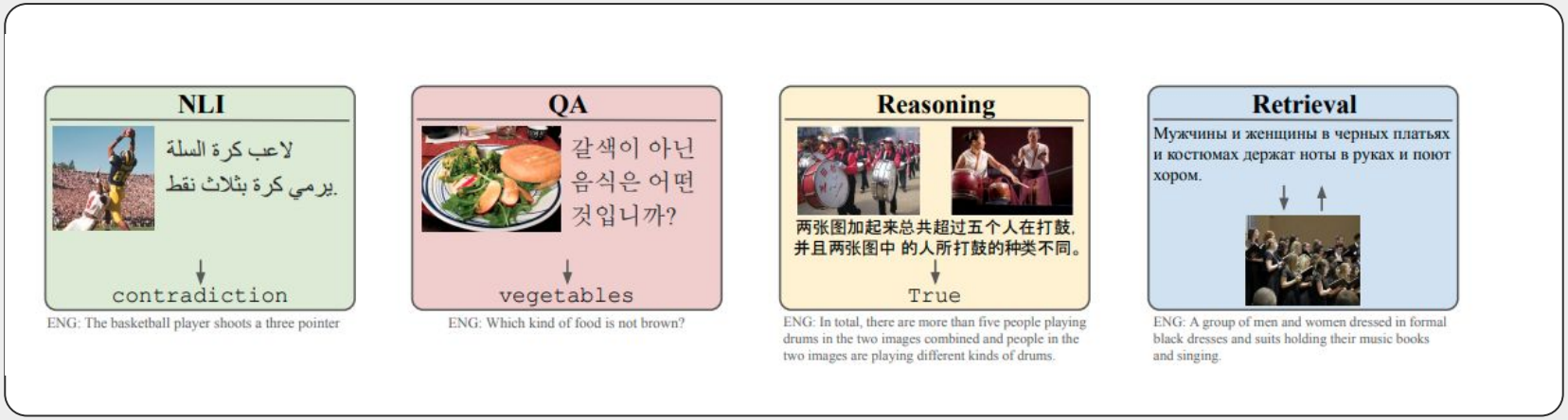
- What can linguistic typology do for multilingual NLP?

but also…

- What can multilingual NLP do for linguistic typology?
  - *Completion of typological databases?*
  - *Automatic linguistic structure induction? (BERT is actually quite good in encoding syntax)*
  - *Data-driven and instance-based typology?*

[Bjerva, Augenstein et al., multiple papers]

Research Branch 9: Multilingual Multimodal Learning



IGLUE: A Benchmark for Transfer Learning across Modalities, Tasks, and Languages

[Bugliarello et al., ICML-2022]