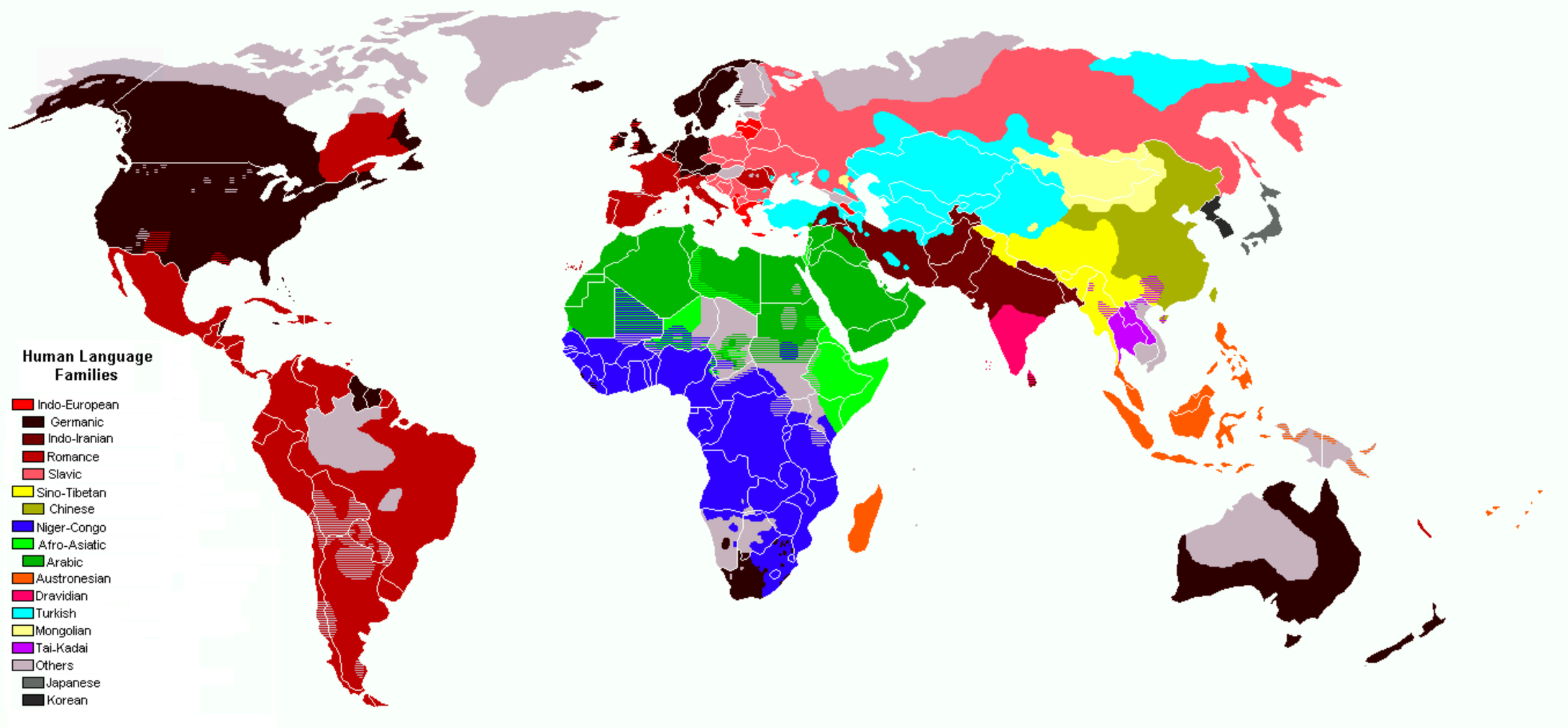
Multilinguality

COMP3361 — Week 11

Lingpeng Kong

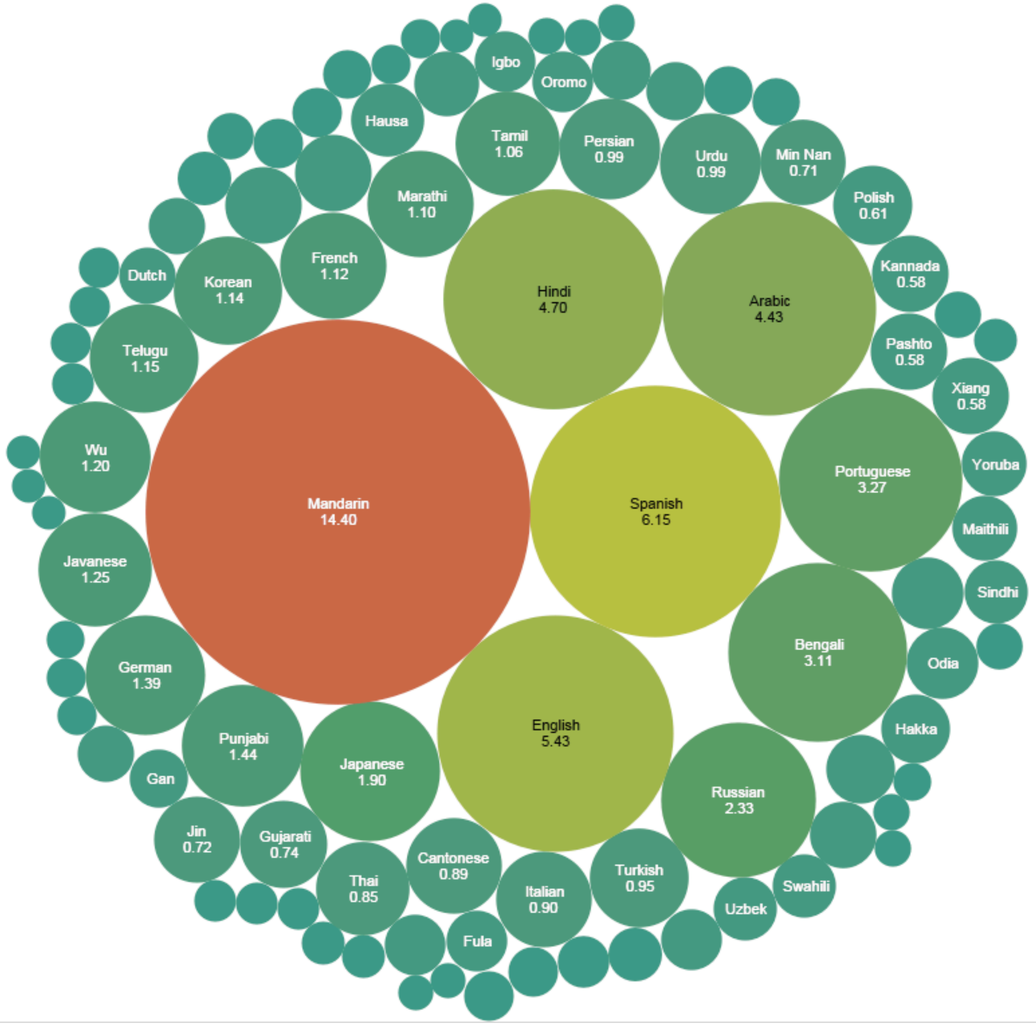
Department of Computer Science, The University of Hong Kong

Languages in the World

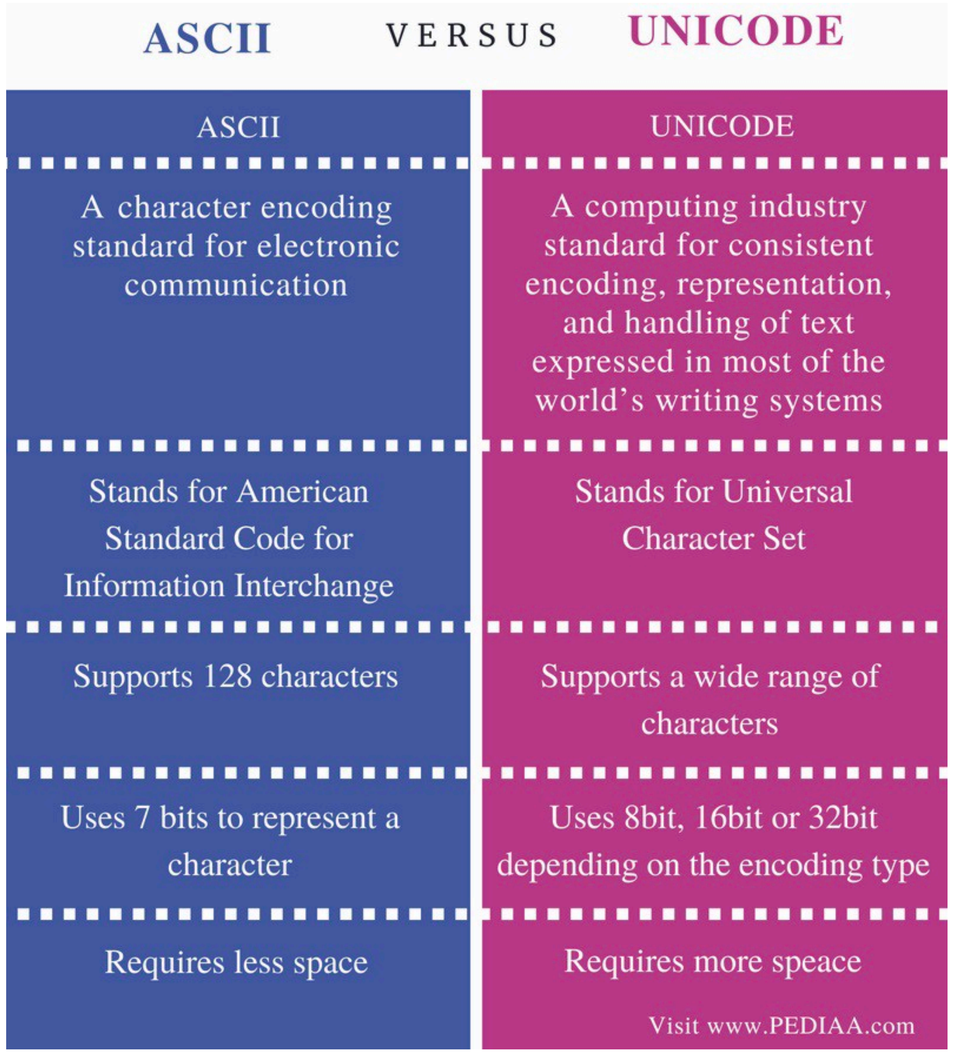


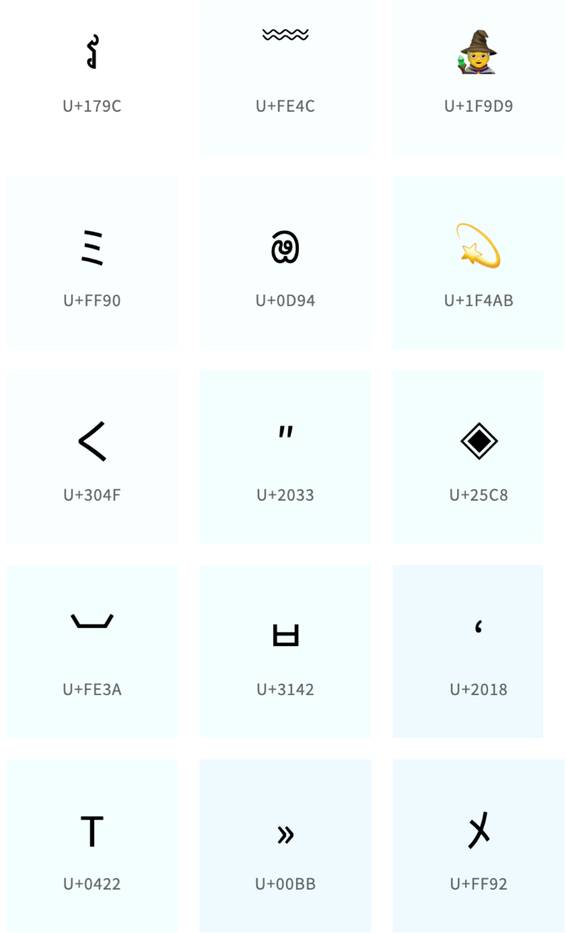
Languages in the World

Rank	Language	Speakers (millions)	Percentage of world pop. (March 2019)[8]	Language family	Branch
1	Mandarin Chinese	918	11.922%	Sino-Tibetan	Sinitic
2	Spanish	480	5.994%	Indo-European	Romance
3	English	379	4.922%	Indo-European	Germanic
4	Hindi (sanskritised Hindustani)[9]	341	4.429%	Indo-European	Indo-Aryan
5	Bengali	300	4.000%	Indo-European	Indo-Aryan
6	Portuguese	221	2.870%	Indo-European	Romance
7	Russian	154	2.000%	Indo-European	Balto-Slavic
8	Japanese	128	1.662%	Japonic	Japanese
9	Western Punjabi[10]	92.7	1.204%	Indo-European	Indo-Aryan
10	Marathi	83.1	1.079%	Indo-European	Indo-Aryan
11	Telugu	82.0	1.065%	Dravidian	South-Central
12	Wu Chinese	81.4	1.057%	Sino-Tibetan	Sinitic
13	Turkish	79.4	1.031%	Turkic	Oghuz
14	Korean	77.3	1.004%	Koreanic	language isolate
15	French	77.2	1.003%	Indo-European	Romance
16	German (only Standard German)	76.1	0.988%	Indo-European	Germanic
17	Vietnamese	76.0	0.987%	Austroasiatic	Vietic
18	Tamil	75.0	0.974%	Dravidian	South
19	Yue Chinese	73.1	0.949%	Sino-Tibetan	Sinitic
20	Urdu (Persianised Hindustani)[9]	68.6	0.891%	Indo-European	Indo-Aryan
21	Javanese	68.3	0.887%	Austronesian	Malayo-Polynesian
22	Italian	64.8	0.842%	Indo-European	Romance
23	Egyptian Arabic	64.6	0.839%	Afroasiatic	Semitic
24	Gujarati	56.4	0.732%	Indo-European	Indo-Aryan
25	Iranian Persian	52.8	0.686%	Indo-European	Iranian



Languages in Computers





Language Identification

That is life	English
C’est la vie	Portuguese
Yahee jeevan hai	Hindi
Bubomi obo	isiXhosa
それが人生です	Japanese

Source: O’Sullivan, towards data science

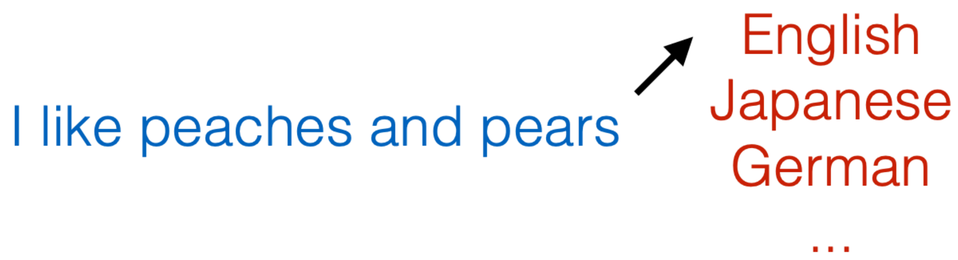
Language Identification

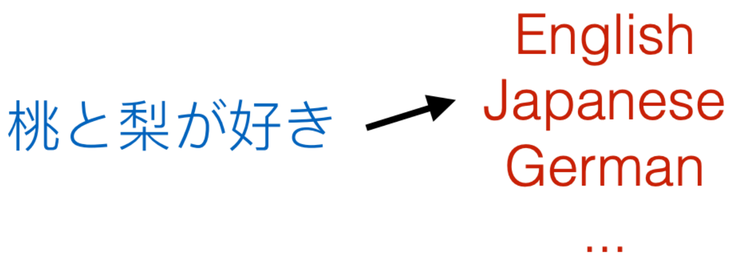
Lexically inspired methods

Language model based methods

勉强　　勉強

Text classification (Machine Learning)





Code-Switching

我真的很appreciate上次你的跟我讲的ideas啊！

to amplify & emphasize

quote

to show identity

habitual expressions

.....

Code-Switching

Newpapers and Wikipedia will not be code-switched

Why care about NLP of non-standard language?

Code-switching is how people actually communicate

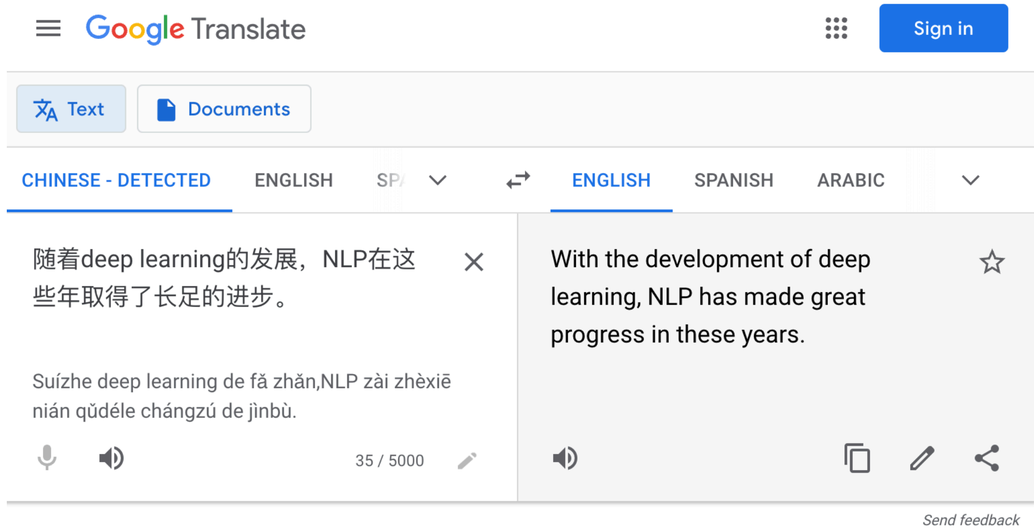
People type in questions to Google/Bing

They talk to call centers

They write their opinions

Use of code-switching can define group membership

People trust code-switched communication better



(Not that it is fake, but its their language and someone developed communication in their language)

Facebook, Amazon, Microsoft, Apple all want to understand Code-switch more

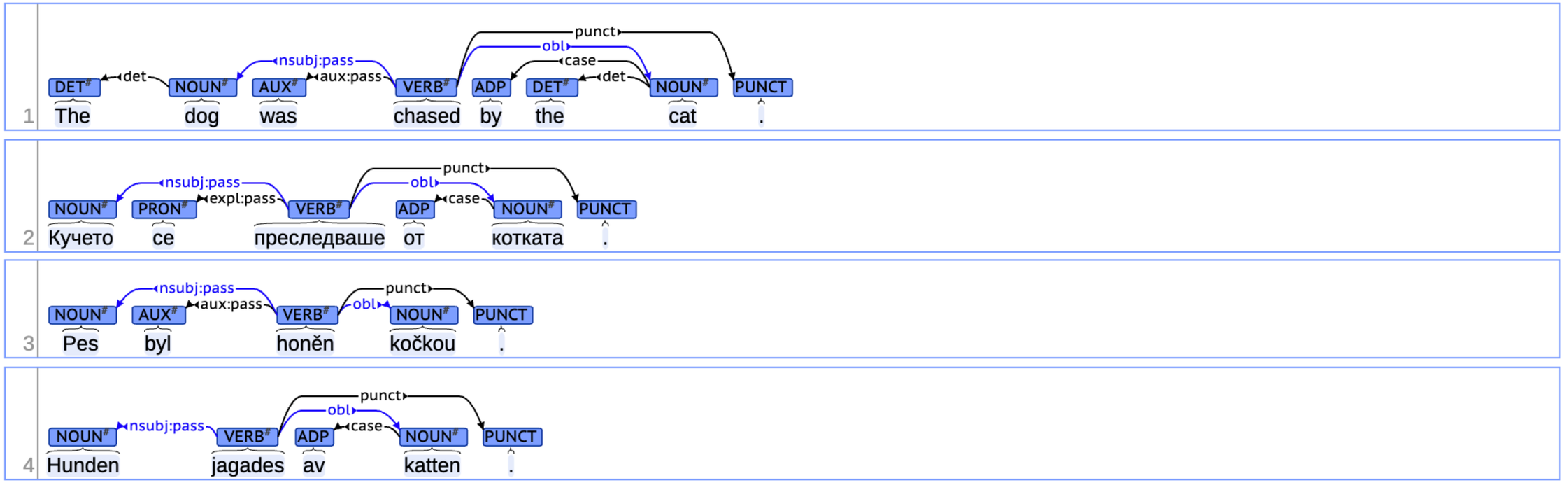
Code-Switching

Techniques:

Very similar to general low-resource language issues:

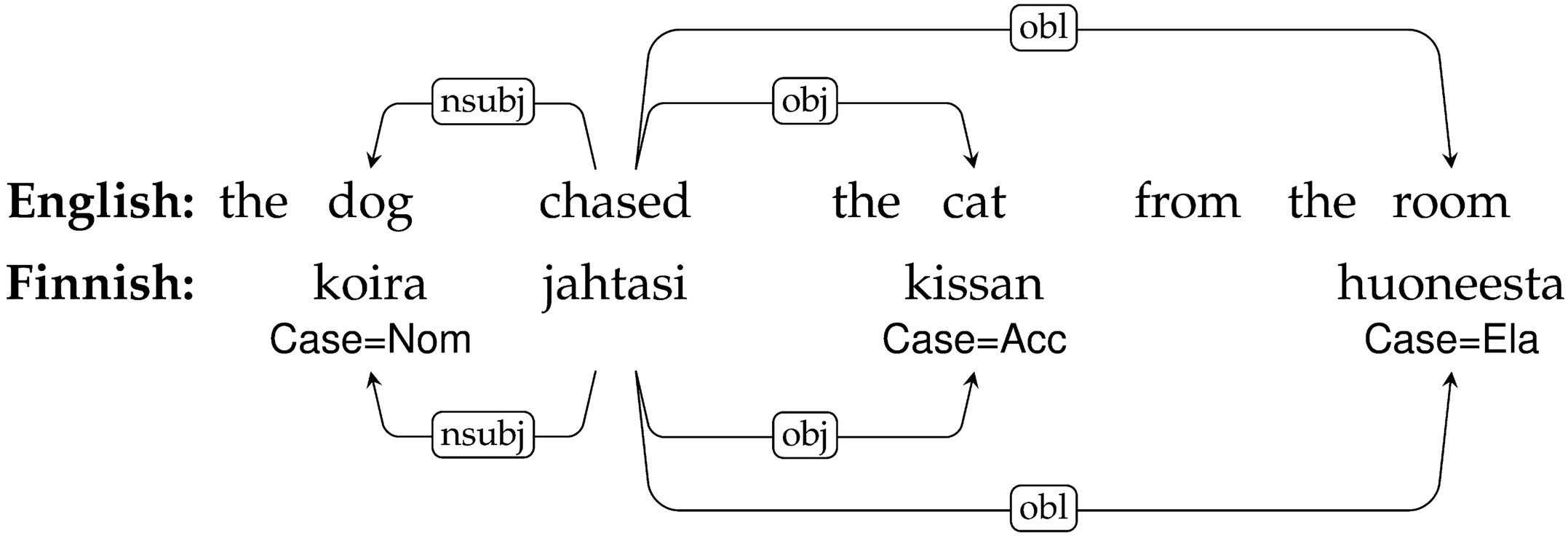
Find appropriate data
Bootstrap labeling
Data argumentation / generation techniques
Find new (reliable) evaluation techniques

Universal Dependencies



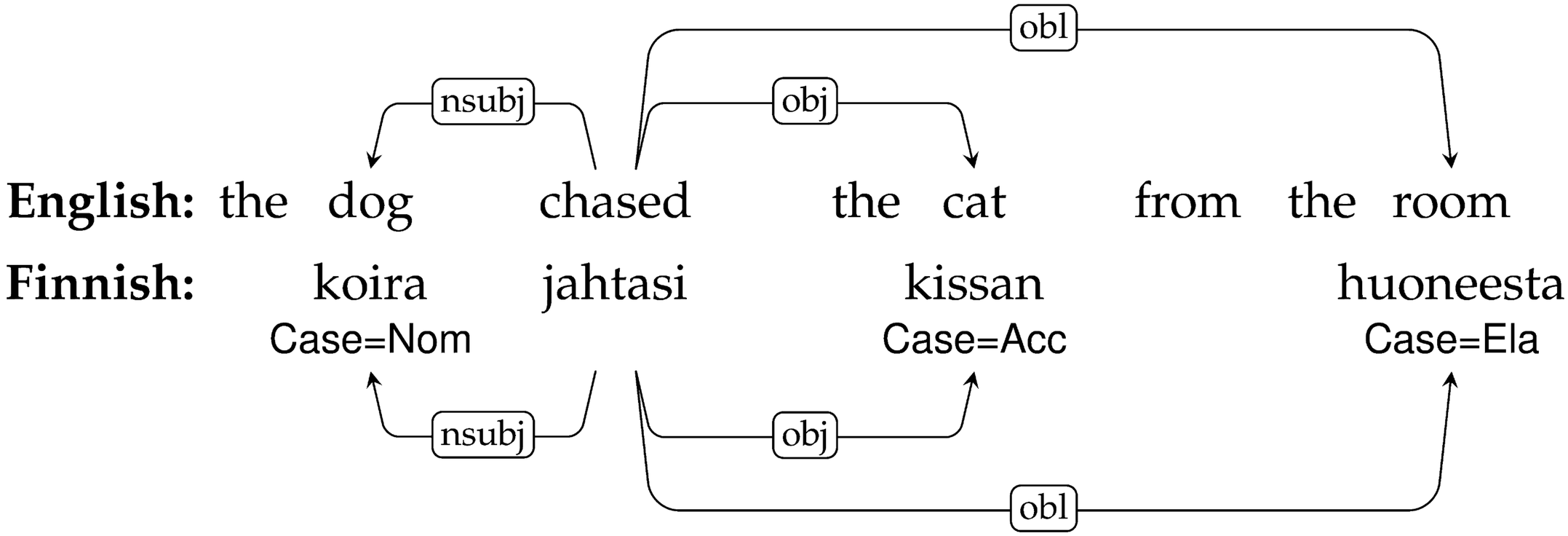
(Marie-Catherine de Marneffe et al., 2021)

Universal Dependencies



(Marie-Catherine de Marneffe et al., 2021)

Universal Dependencies



(Marie-Catherine de Marneffe et al., 2021)

Universal Dependencies

LAS	target language							average
	de	en	es	fr	it	pt	sv	
monolingual	**79.3**	**85.9**	83.7	81.7	88.7	85.7	83.5	84.0
MALOPA	70.4	69.3	72.4	71.1	78.0	74.1	65.4	71.5
+lexical	76.7	82.0	82.7	81.2	87.6	82.1	81.2	81.9
+language ID	78.6	84.2	83.4	**82.4**	**89.1**	84.2	82.6	83.5
+fine-grained POS	78.9	85.4	**84.3**	**82.4**	89.0	**86.2**	**84.5**	**84.3**

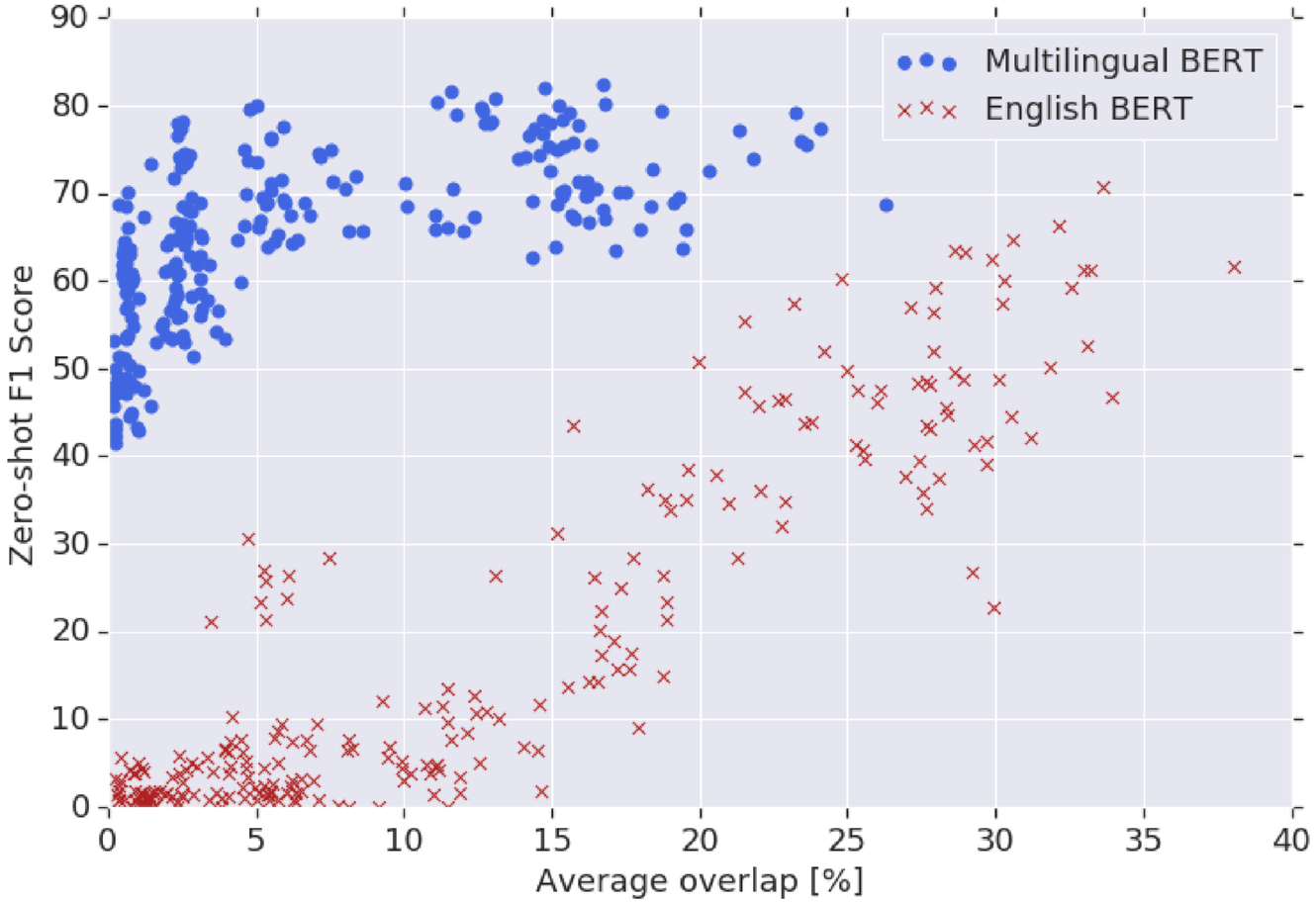
(Many Languages One Parser, Ammar et al., 2016)

Multilingual BERT

Pre-trained from monolingual corpora in 104 languages

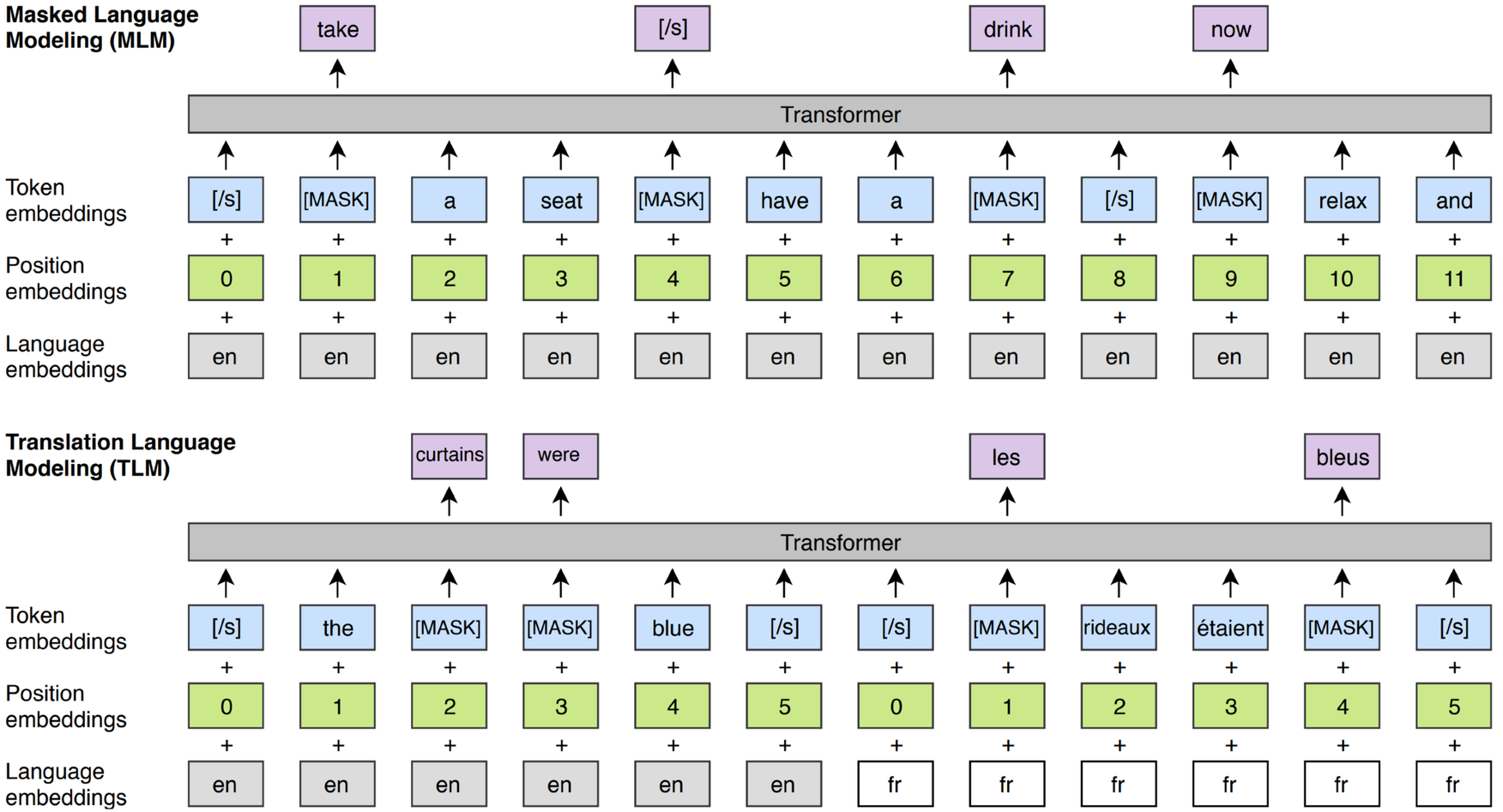
Good at zero shot cross-lingual model transfer:

Fine-tune the model using task-specific supervised training data from one language, and evaluate that task in a different language!



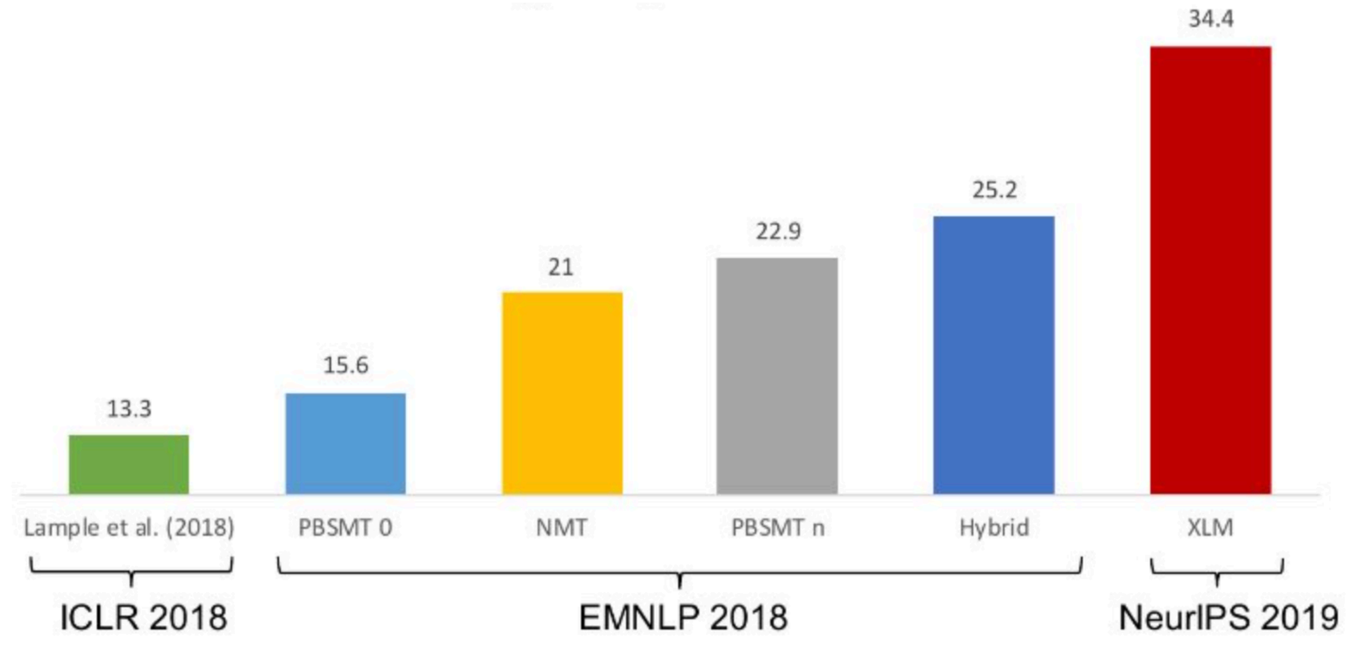
(Devlin et al., 2019; Pires et al., 2019)

Cross-lingual Language Model Pretraining



(Lample and Conneau, 2019)

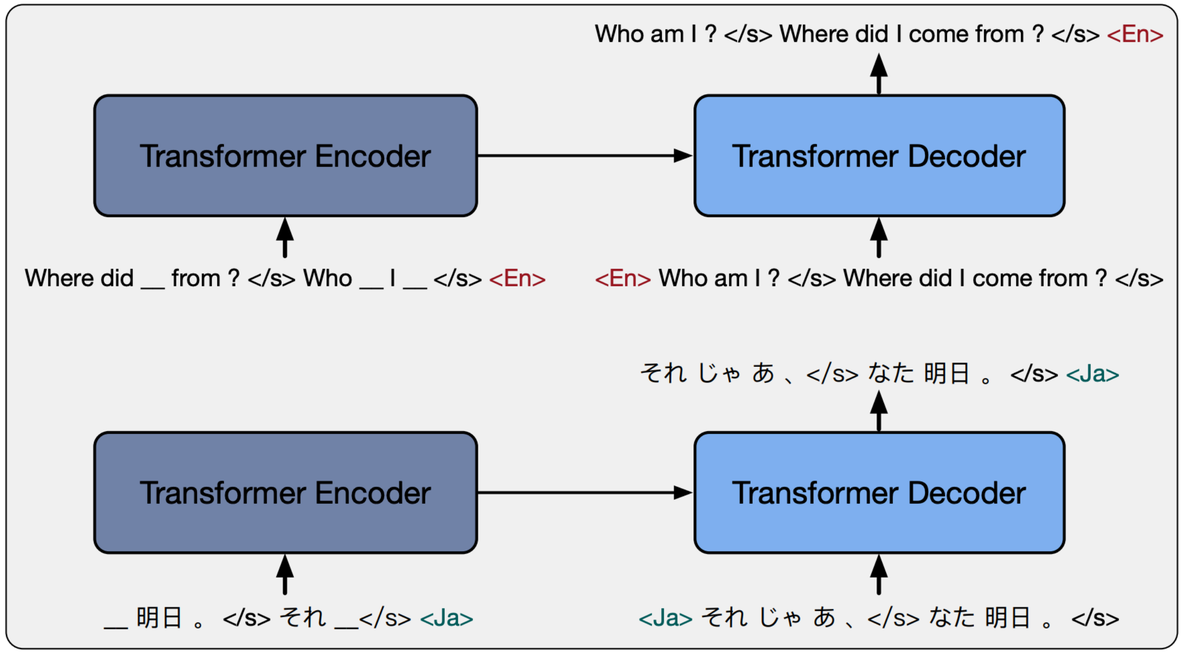
Cross-lingual Language Model Pretraining



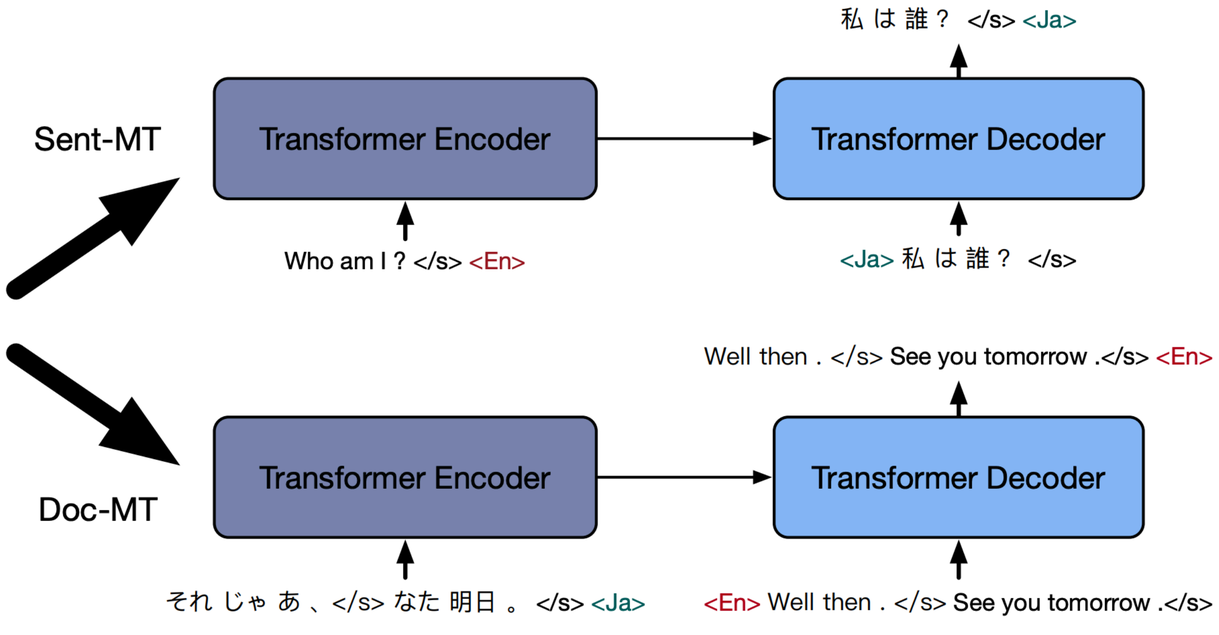
Evolution of Unsupervised MT BLEU score on German-English - newstest 2016

(Lample and Conneau, 2019)

mBART: Multilingual Denoising Pretraining



Multilingual Denoising Pre-Training (mBART)



Fine-tuning on Machine Translation

(Liu et al, 2020)