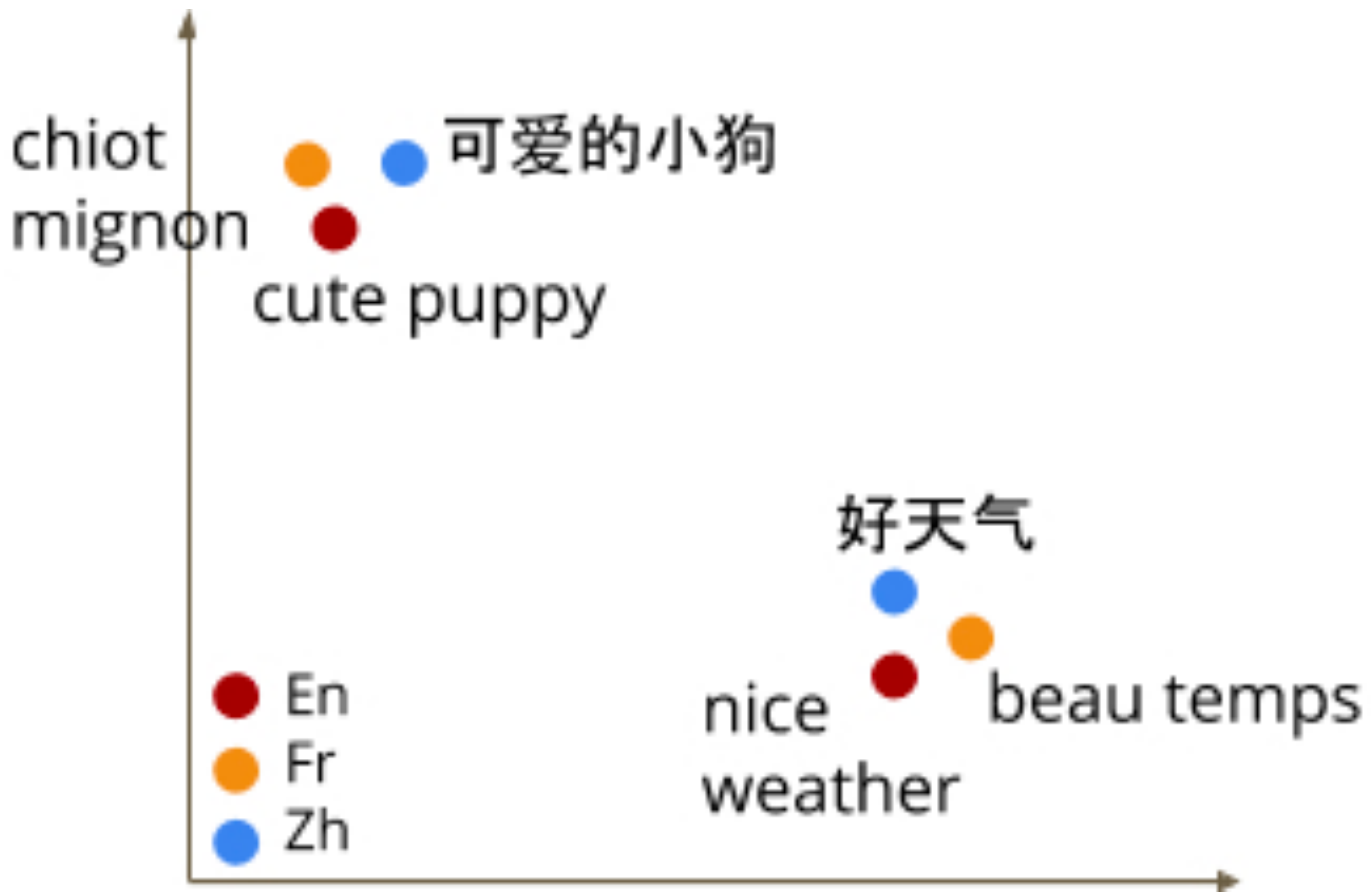
CS388: Natural Language Processing

Lecture 23: Multilingual Models



Credit: Google AI Blog

Greg Durrett

TEXAS
The University of Texas at Austin

Announcements

- FP due April 28
- **Presentations next week!**
  - Send Google slides by 6am the morning of your presentation.
  - 3 minutes is a firm limit!

NLP in other languages

- Other languages present some challenges not seen in English at all
- Some of our algorithms have been specified to English
  - Some structures like constituency parsing don't make sense for other languages
  - Neural methods are typically tuned to English-scale resources, may not be the best for other languages where less data is available
- This lecture: How can we leverage existing resources to do better in other languages without just annotating massive data?

This Lecture

- Morphology
- Cross-lingual tagging and parsing
- Multilingual pre-training
- Multilingual benchmarks
- Introduce ethical questions (for next time)

Morphology

What is morphology?

- Study of how words form
- Derivational morphology: create a new word from a root word

  estrange (v) => estrangement (n)

  become (v) => unbecoming (adj)

  - May not be totally regular: enflame => inflammable
- Inflectional morphology: word is inflected based on its context

  I become / she become**s**

  - Mostly applies to verbs and nouns

Morphological Inflection

- In English: I arrive / you arrive / he/she/it arrives / we arrive / you arrive / they arrive / [X] arrived

		I arrive	you arrive	he/she/it arrives		[X] arrived
		we arrive	you arrive	they arrive		

- In French:

		singular			plural		
		first	**second**	**third**	**first**	**second**	**third**
indicative		**je (j')**	**tu**	**il, elle**	**nous**	**vous**	**ils, elles**
(simple tenses)	**present**	arrive /a.ʁiv/	arrives /a.ʁiv/	arrive /a.ʁiv/	arrivons /a.ʁi.vɔ̃/	arrivez /a.ʁi.ve/	arrivent /a.ʁiv/
	imperfect	arrivais /a.ʁi.vɛ/	arrivais /a.ʁi.vɛ/	arrivait /a.ʁi.vɛ/	arrivions /a.ʁi.vjɔ̃/	arriviez /a.ʁi.vje/	arrivaient /a.ʁi.vɛ/
	past historic[2]	arrivai /a.ʁi.vɛ/	arrivas /a.ʁi.va/	arriva /a.ʁi.va/	arrivâmes /a.ʁi.vam/	arrivâtes /a.ʁi.vat/	arrivèrent /a.ʁi.vɛʁ/
	future	arriverai /a.ʁi.vʁɛ/	arriveras /a.ʁi.vʁa/	arrivera /a.ʁi.vʁa/	arriverons /a.ʁi.vʁɔ̃/	arriverez /a.ʁi.vʁe/	arriveront /a.ʁi.vʁɔ̃/
	conditional	arriverais /a.ʁi.vʁɛ/	arriverais /a.ʁi.vʁɛ/	arriverait /a.ʁi.vʁɛ/	arriverions /a.ʁi.və.ʁjɔ̃/	arriveriez /a.ʁi.və.ʁje/	arriveraient /a.ʁi.vʁɛ/

Morphological Inflection

- In Spanish:

		singular			plural		
		1st person	2nd person	3rd person	1st person	2nd person	3rd person
indicative		yo	tú vos	él/ella/ello usted	nosotros nosotras	vosotros vosotras	ellos/ellas ustedes
	present	llego	llegas tú llegás vos	llega	llegamos	llegáis	llegan
	imperfect	llegaba	llegabas	llegaba	llegábamos	llegabais	llegaban
	preterite	llegué	llegaste	llegó	llegamos	llegasteis	llegaron
	future	llegaré	llegarás	llegará	llegaremos	llegaréis	llegarán
	conditional	llegaría	llegarías	llegaría	llegaríamos	llegaríais	llegarían

Noun Inflection

- Not just verbs either; gender, number, case complicate things

Declension of Kind [hide ▲]

	singular			plural	
	indef.	def.	noun	def.	noun
nominative	ein	das	Kind	die	Kinder
genitive	eines	des	Kindes, Kinds	der	Kinder
dative	einem	dem	Kind, Kinde[1]	den	Kindern
accusative	ein	das	Kind	die	Kinder

- Nominative: I/he/she, accusative: me/him/her, genitive: mine/his/hers
- Dative: merged with accusative in English, shows recipient of something

  I taught the children <=> Ich unterrichte die Kinder

  I give the children a book <=> Ich gebe den Kindern ein Buch

Irregular Inflection

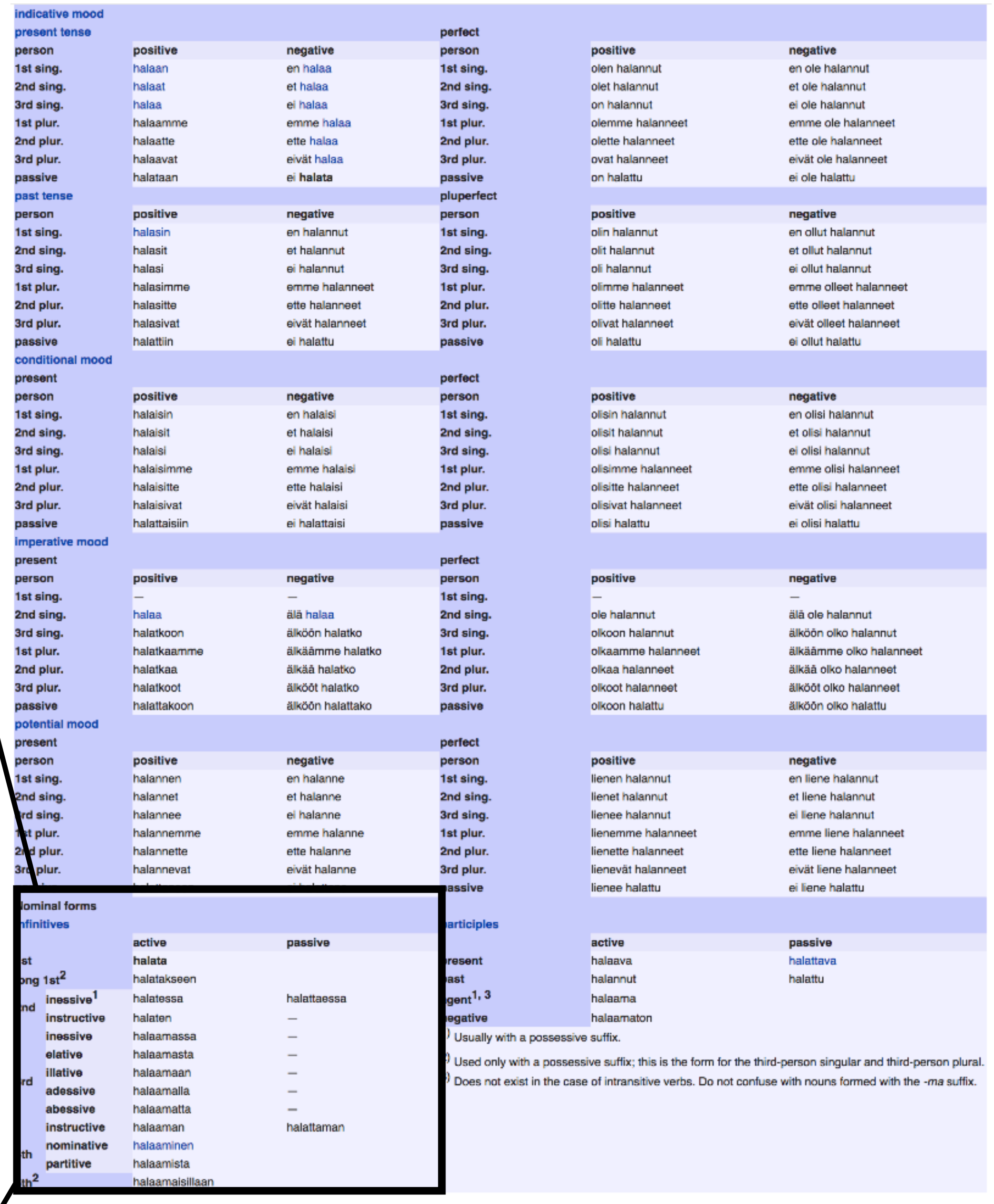
- Common words are often irregular
  - I am / you are / she is
  - Je suis / tu es / elle est
  - Soy / está / es

- Less common words typically fall into some regular *paradigm* — these are somewhat predictable

Agglutinating Langauges

- Finnish/Hungarian (Finno-Ugric), also Turkish: what a preposition would do in English is instead part of the verb (*hug*)

		active	passive
1st		**halata**	
long 1st[2]		halatakseen	
2nd	inessive[1]	halatessa	halattaessa
	instructive	halaten	—
3rd	inessive	halaamassa	—
	elative	halaamasta	—
	illative	halaamaan	—
	adessive	halaamalla	—
	abessive	halaamatta	—
	instructive	halaaman	halattaman
4th	nominative	halaaminen	
	partitive	halaamista	
5th[2]		halaamaisillaan	



halata: “hug”

illative: “into”

adessive: “on”

- Many possible forms — and in newswire data, only a few are observed

Morphologically-Rich Languages

- Many languages spoken all over the world have much richer morphology than English
  - CoNLL 2006 / 2007: dependency parsing + morphological analyses for ~15 mostly Indo-European languages
  - SPMRL shared tasks (2013-2014): Syntactic Parsing of Morphologically-Rich Languages
  - Universal Dependencies project
- Word piece / byte-pair encoding models for MT are pretty good at handling these if there's enough data

Morphological Analysis: Hungarian

But the government does not recommend reducing taxes.

Ám a kormány egyetlen adó csökkentését sem javasolja .

n=singular|case=nominative|proper=no

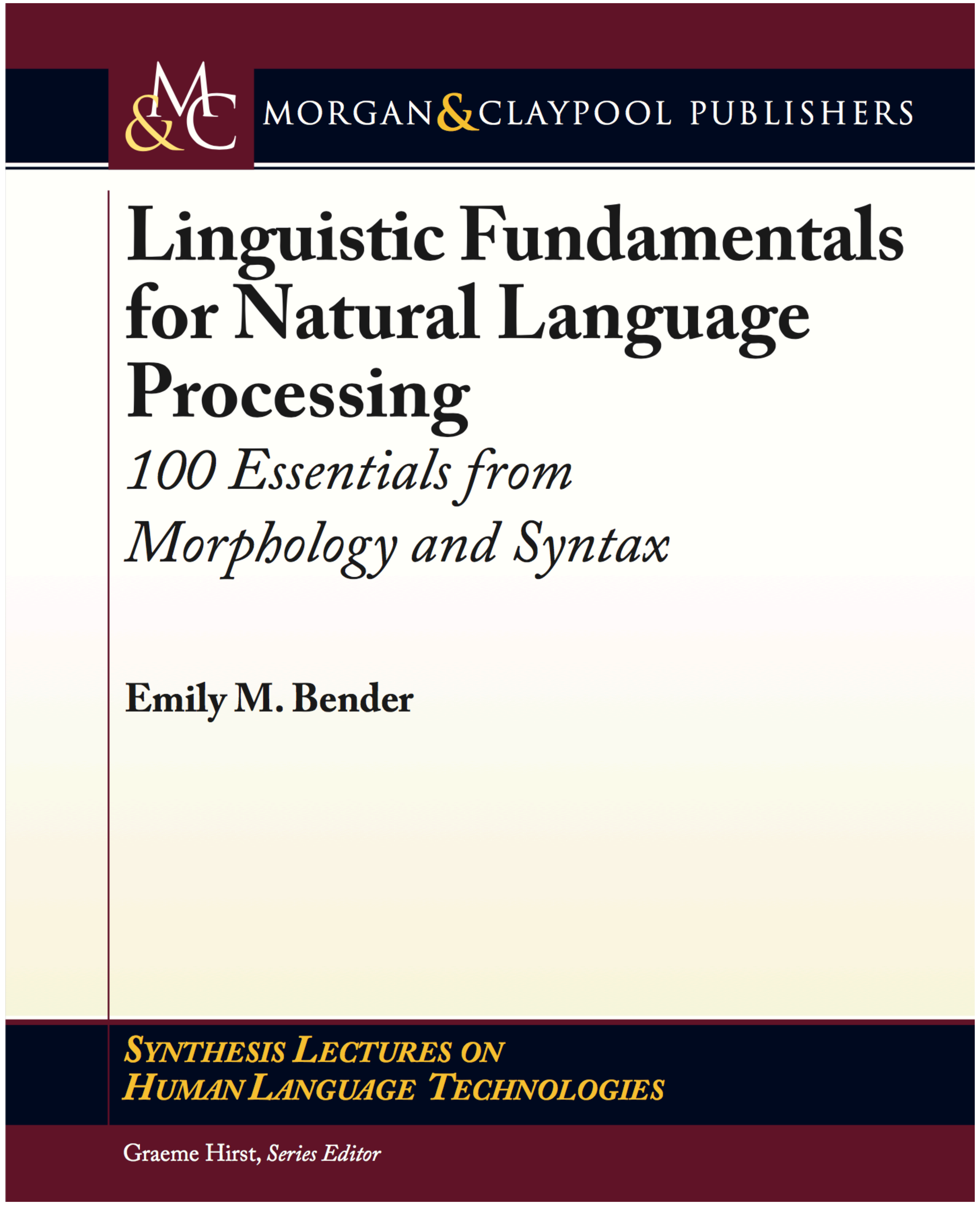
deg=positive|n=singular|case=nominative

n=singular|case=nominative|proper=no

n=singular|case=accusative|proper=no|pperson=3rd|pnumber=singular

mood=indicative|t=present|p=3rd|n=singular|def=yes

Morphologically-Rich Languages



- Great resources for challenging your assumptions about language and for understanding multilingual models!

Chinese Word Segmentation

- Word segmentation: some languages including Chinese are totally untokenized
- LSTMs over character embeddings / character bigram embeddings to predict word boundaries
- Having the right segmentation can help machine translation

冬天 (winter)，能 (can) 穿 (wear) 多少 (amount) 穿 (wear) 多少 (amount)；夏天 (summer)，能 (can) 穿 (wear) 多 (more) 少 (little) 穿 (wear) 多 (more) 少 (little)。

Without the word "夏天 (summer)" or "冬天 (winter)", it is difficult to segment the phrase "能穿多少穿多少".

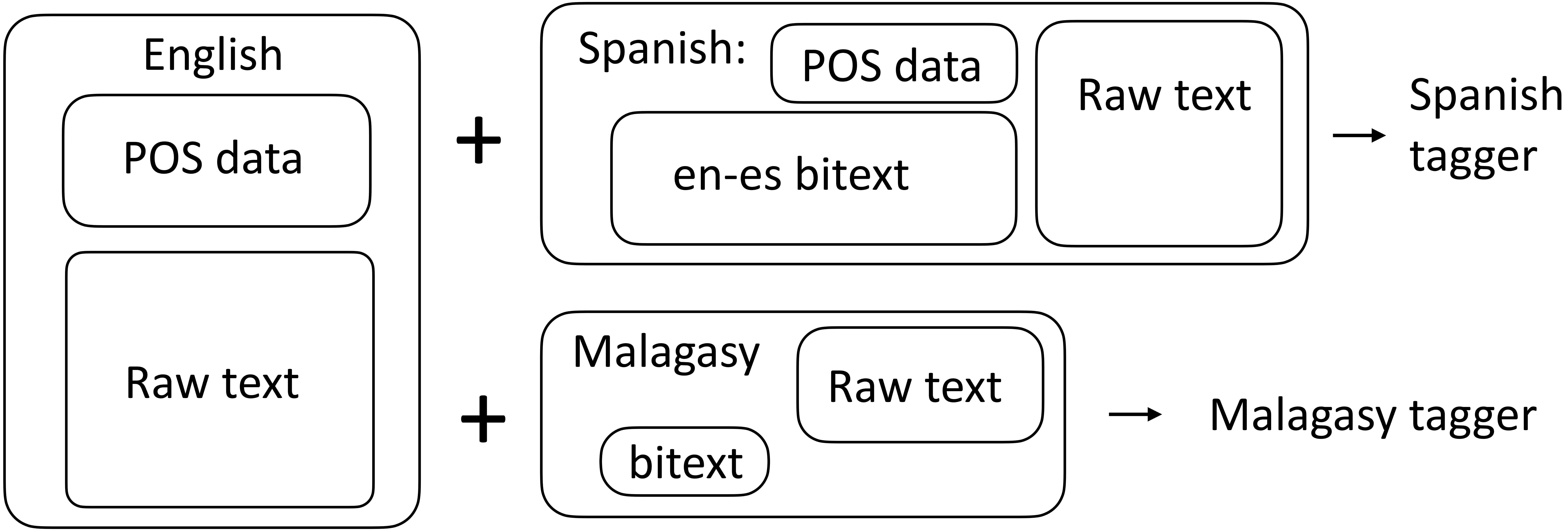
- separating nouns and pre-modifying adjectives:
  高血压 (*high blood pressure*)
  → 高(*high*) 血压(*blood pressure*)
- separating compound nouns:
  内政部 (*Department of Internal Affairs*)
  → 内政(*Internal Affairs*) 部(*Department*).

Chen et al. (2015)

Cross-Lingual Tagging and Parsing

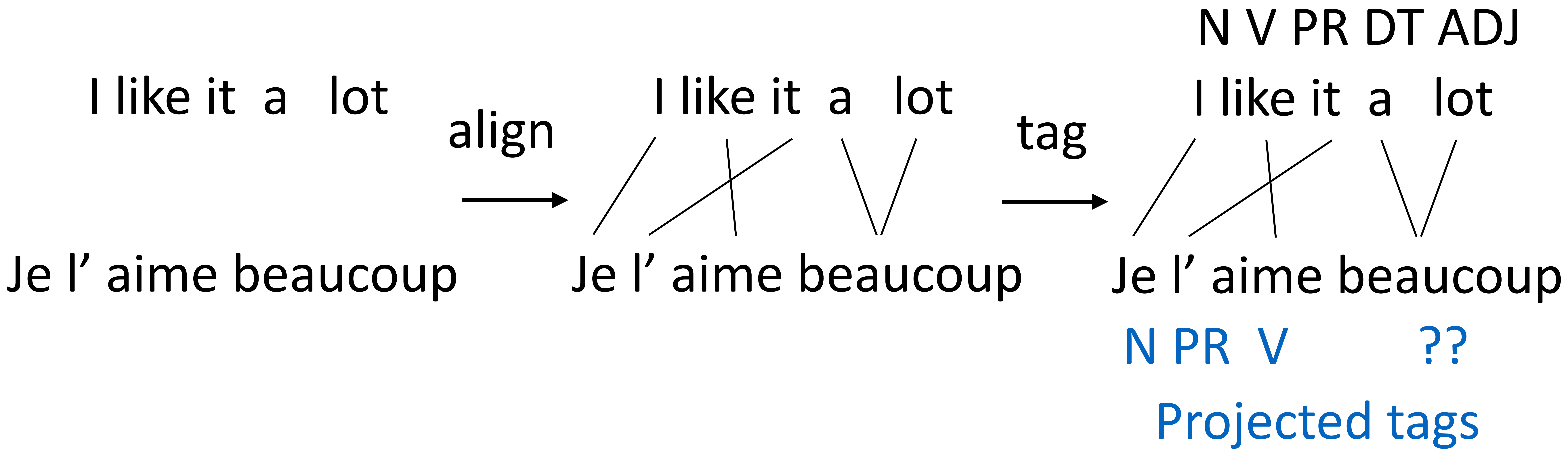
Cross-Lingual Tagging

- Labeling POS datasets is expensive
- Can we transfer annotation from *high-resource* languages (English, etc.) to *low-resource* languages?



Cross-Lingual Tagging

- Can we leverage word alignment here?

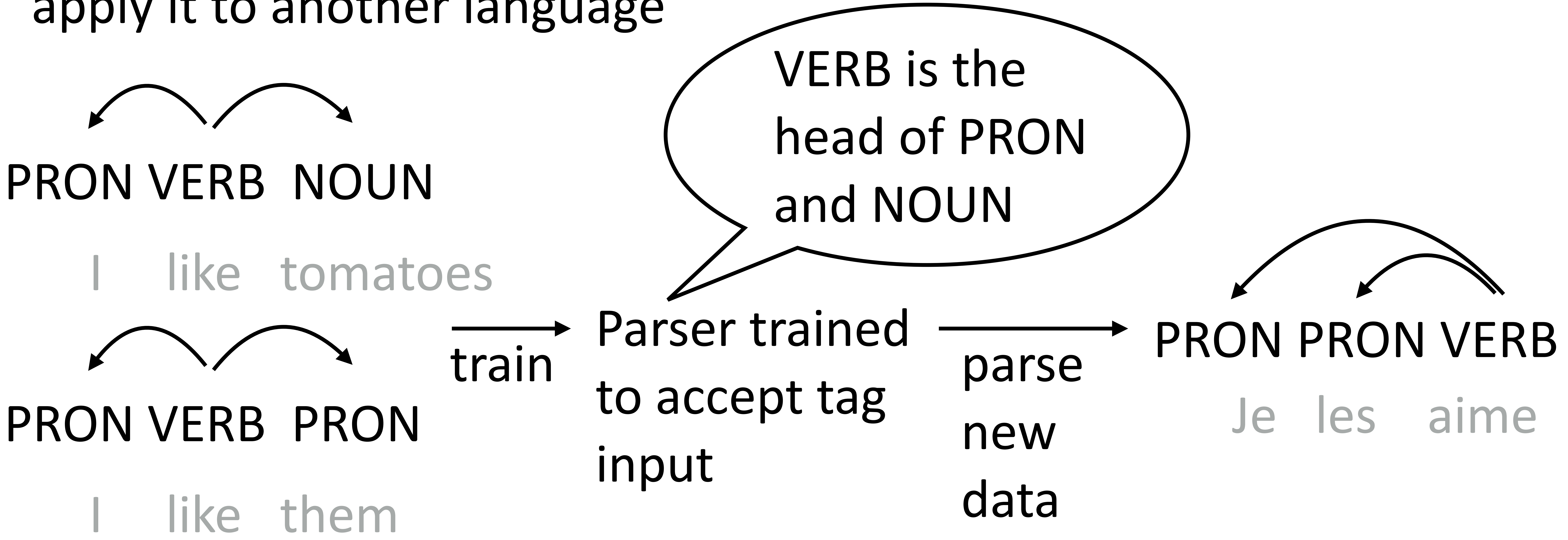


- Tag with English tagger, project across bitext, train French tagger? Works pretty well

Das and Petrov (2011)

Cross-Lingual Parsing

- Now that we can POS tag other languages, can we parse them too?
- Direct transfer: train a parser over POS sequences in one language, then apply it to another language



McDonald et al. (2011)

Cross-Lingual Parsing

	best-source		avg-source	gold-POS	
	source	gold-POS	gold-POS	multi-dir.	multi-proj.
da	it	48.6	46.3	48.9	49.5
de	nl	55.8	48.9	56.7	56.6
el	en	63.9	51.7	60.1	65.1
es	it	68.4	53.2	64.2	64.5
it	pt	69.1	58.5	64.1	65.0
nl	el	62.1	49.9	55.8	65.7
pt	it	74.8	61.6	74.0	75.6
sv	pt	66.8	54.8	65.3	68.0
avg		63.7	51.6	61.1	63.8

- Multi-dir: transfer a parser trained on a few source treebanks to the target language
- Multi-proj: more complex annotation projection approach

McDonald et al. (2011)

Cross-Lingual, Multilingual Word Representations

Multilingual Embeddings

- Input: corpora in many languages. Output: embeddings where similar words *in different languages* have similar embeddings

I have an apple
47 24 18 427

J' ai des oranges
47 24 89 1981

ID: 24
ai have

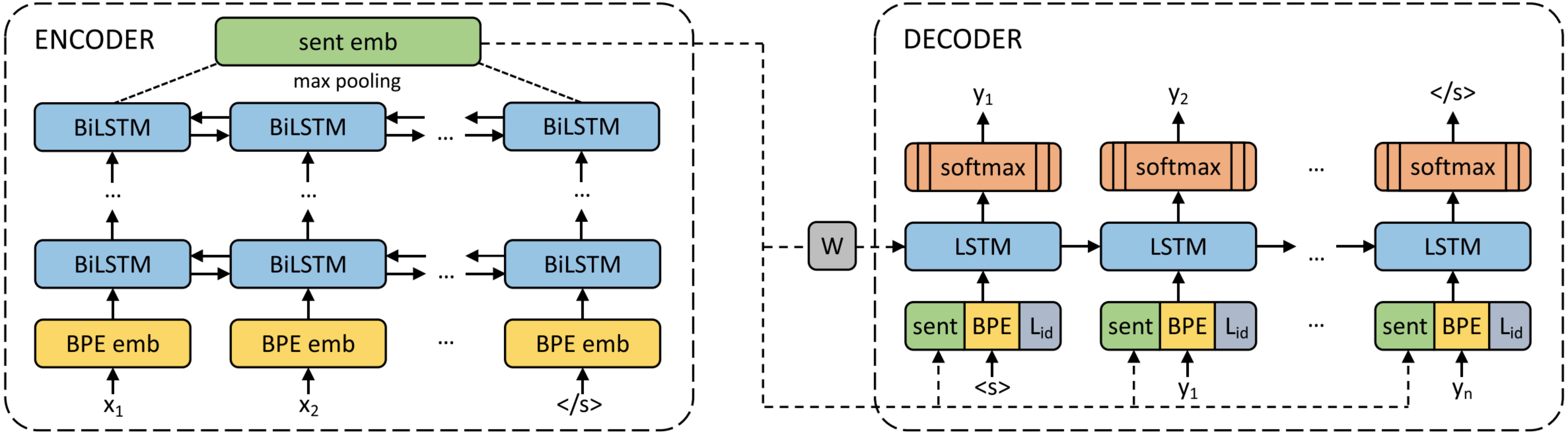
ID: 47
I Je J'

- multiCluster: use bilingual dictionaries to form clusters of words that are translations of one another, replace corpora with cluster IDs, train "monolingual" embeddings over all these corpora

- Works okay but not all that well

Ammar et al. (2016)

Multilingual Sentence Embeddings



- Form BPE vocabulary over all corpora (50k merges); will include characters from every script
- Take a bunch of bitexts and train an MT model between a bunch of language pairs with shared parameters, use W as sentence embeddings

Artetxe et al. (2019)

Multilingual Sentence Embeddings

		EN	fr	es	de	el	bg	ru
Zero-Shot Transfer, one NLI system for all languages:								
Conneau et al.	X-BiLSTM	73.7	67.7	68.7	67.7	68.9	67.9	65.4
(2018b)	X-CBOW	64.5	60.3	60.7	61.0	60.5	60.4	57.8
BERT uncased*	Transformer	81.4	–	74.3	70.5	–	–	–
Proposed method	BiLSTM	73.9	**71.9**	72.9	72.6	**72.8**	**74.2**	**72.1**

- Train a system for NLI (entailment/neutral/contradiction of a sentence pair) on English and evaluate on other languages

Artetxe et al. (2019)

Multilingual BERT

- Take top 104 Wikipedias, train BERT on all of them simultaneously
- What does this look like?

Beethoven may have proposed unsuccessfully to Therese Malfatti, the supposed dedicatee of "Für Elise"; his status as a commoner may again have interfered with those plans.

当人们在马尔法蒂身后发现这部小曲的手稿时，便误认为上面写的是“Für Elise”（即《给爱丽丝》）[51]。

Кита́й (официально — Кита́йская Наро́дная Респу́блика, сокращённо — КНР; кит. трад. 中華人民共和國, упр. 中华人民共和国, пиньинь: Zhōnghuá Rénmín Gònghéguó, палл.: Чжунхуа Жэньминь Гунхэго) — государство в Восточной Аз

Devlin et al. (2019)

Multilingual BERT: Results

Fine-tuning \ Eval	EN	DE	ES	IT
EN	**96.82**	89.40	85.91	91.60
DE	83.99	**93.99**	86.32	88.39
ES	81.64	88.87	**96.71**	93.71
IT	86.79	87.82	91.28	**98.11**

Table 2: POS accuracy on a subset of UD languages.

- Can transfer BERT directly across languages with some success
- …but this evaluation is on languages that all share an alphabet

Pires et al. (2019)

Multilingual BERT: Results

	HI	UR
HI	**97.1**	85.9
UR	91.1	**93.8**

	EN	BG	JA
EN	**96.8**	87.1	49.4
BG	82.2	**98.9**	51.6
JA	57.4	67.2	**96.5**

Table 4: POS accuracy on the UD test set for languages with different scripts. Row=fine-tuning, column=eval.

- Urdu (Arabic/Nastaliq script) => Hindi (Devanagari). Transfers well despite different alphabets!
- Japanese => English: different script and very different syntax

Pires et al. (2019)

Scaling Up: XLM-R

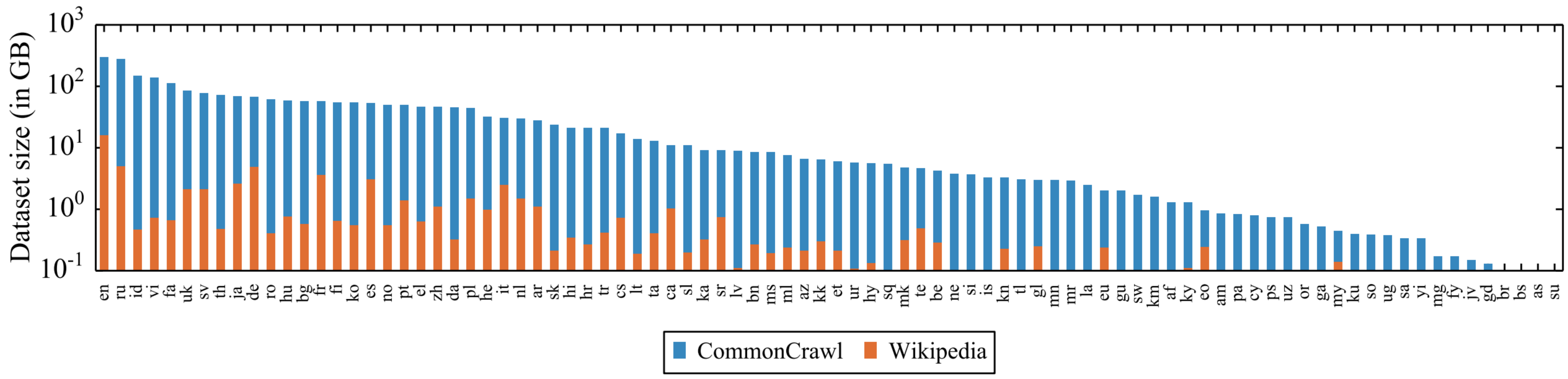
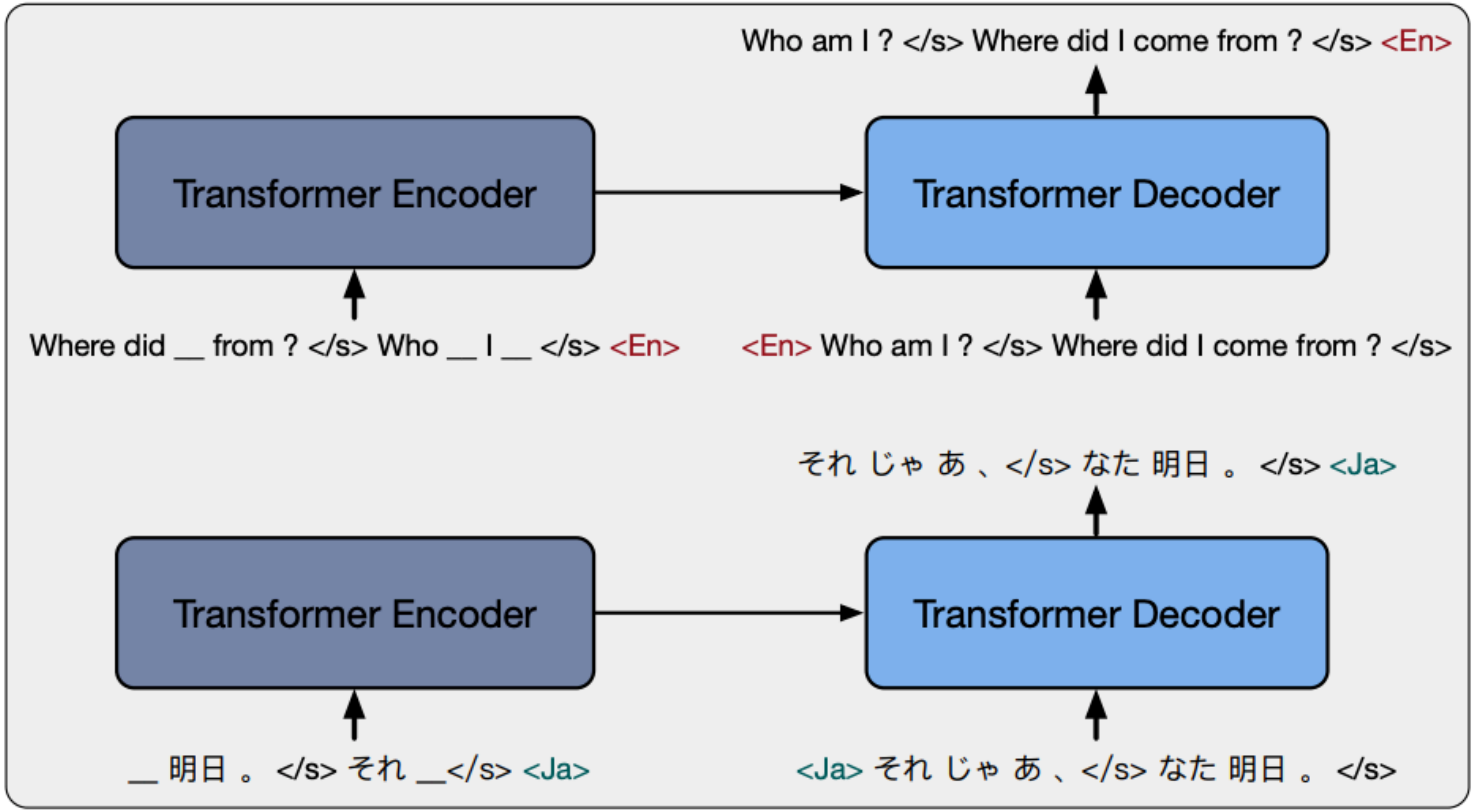


Figure 1: Amount of data in GiB (log-scale) for the 88 languages that appear in both the Wiki-100 corpus used for mBERT and XLM-100, and the CC-100 used for XLM-R. CC-100 increases the amount of data by several orders of magnitude, in particular for low-resource languages.

- Larger “Common Crawl” dataset, better performance than mBERT
- Low-resource languages benefit from training on other languages
- High-resource languages see a small performance hit, but not much

Conneau et al. (2019)

Scaling Up: mBART



Multilingual Denoising Pre-Training (mBART)

Liu et al. (2020)

Multilingual Benchmarks

Scaling Up: Benchmarks

Task	Corpus	|Train|	|Dev|	|Test|	Test sets	|Lang.|	Task
Classification	XNLI	392,702	2,490	5,010	translations	15	NLI
	PAWS-X	49,401	2,000	2,000	translations	7	Paraphrase
Struct. pred.	POS	21,253	3,974	47-20,436	ind. annot.	33 (90)	POS
	NER	20,000	10,000	1,000-10,000	ind. annot.	40 (176)	NER
QA	XQuAD	87,599	34,726	1,190	translations	11	Span extraction
	MLQA			4,517–11,590	translations	7	Span extraction
	TyDiQA-GoldP	3,696	634	323–2,719	ind. annot.	9	Span extraction
Retrieval	BUCC	-	-	1,896–14,330	-	5	Sent. retrieval
	Tatoeba	-	-	1,000	-	33 (122)	Sent. retrieval

- Many of these datasets are translations of base datasets, not originally annotated in those languages
- Exceptions: POS, NER, TyDiQA

Hu et al. (2021)

TyDiQA

- Typologically-diverse QA dataset
- Annotators write questions based on very short snippets of articles; answers may or may not exist, fetched from elsewhere in Wikipedia

Q: Как далеко Уран от Земл-и?
how far Uranus-SG.NOM from Earth-SG.GEN?

How far is Uranus from Earth?

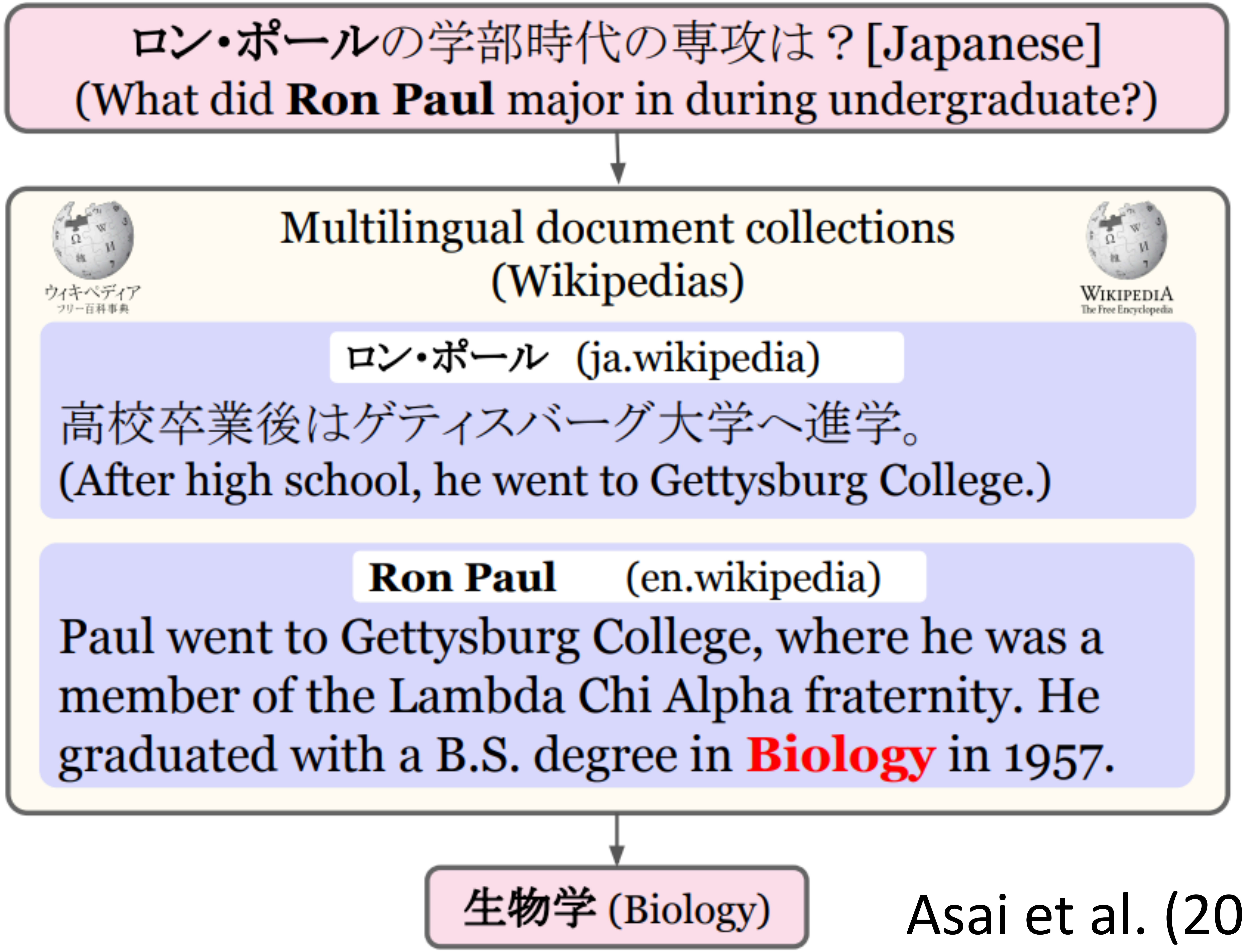
A: Расстояние между Уран-ом и Земл-ёй меняется от 2,6 до 3,15 млрд км...
distance between Uranus-SG.INSTR and Earth-SG.INSTR varies from 2,6 to 3,15 bln km...

The distance between Uranus and Earth fluctuates from 2.6 to 3.15 bln km...

Clark et al. (2021)

Xor QA

- Certain types of info may only be available in certain languages' Wikipedias — need to be able to answer questions multilingually



Asai et al. (2021)

Cross-Lingual Typing

- Train an mBERT-based typing model on Wikipedia data in English, Spanish, German and Finnish
- Achieves solid performance even on totally new languages like Japanese that don’t share a character set with these

Sequence: 菊池は <u>アメリカ大リーグ</u> への参戦も視野に進路が注目されていたが、10月25日に日本のプロ野球に挑戦することを表明していた。...

Translation: Kikuchi was considering <u>Major League Baseball</u> as his next career, but he announced that he would play professional baseball in Japan ...

Predictions: *baseball, established, establishments, in the united states, organizations, sports*

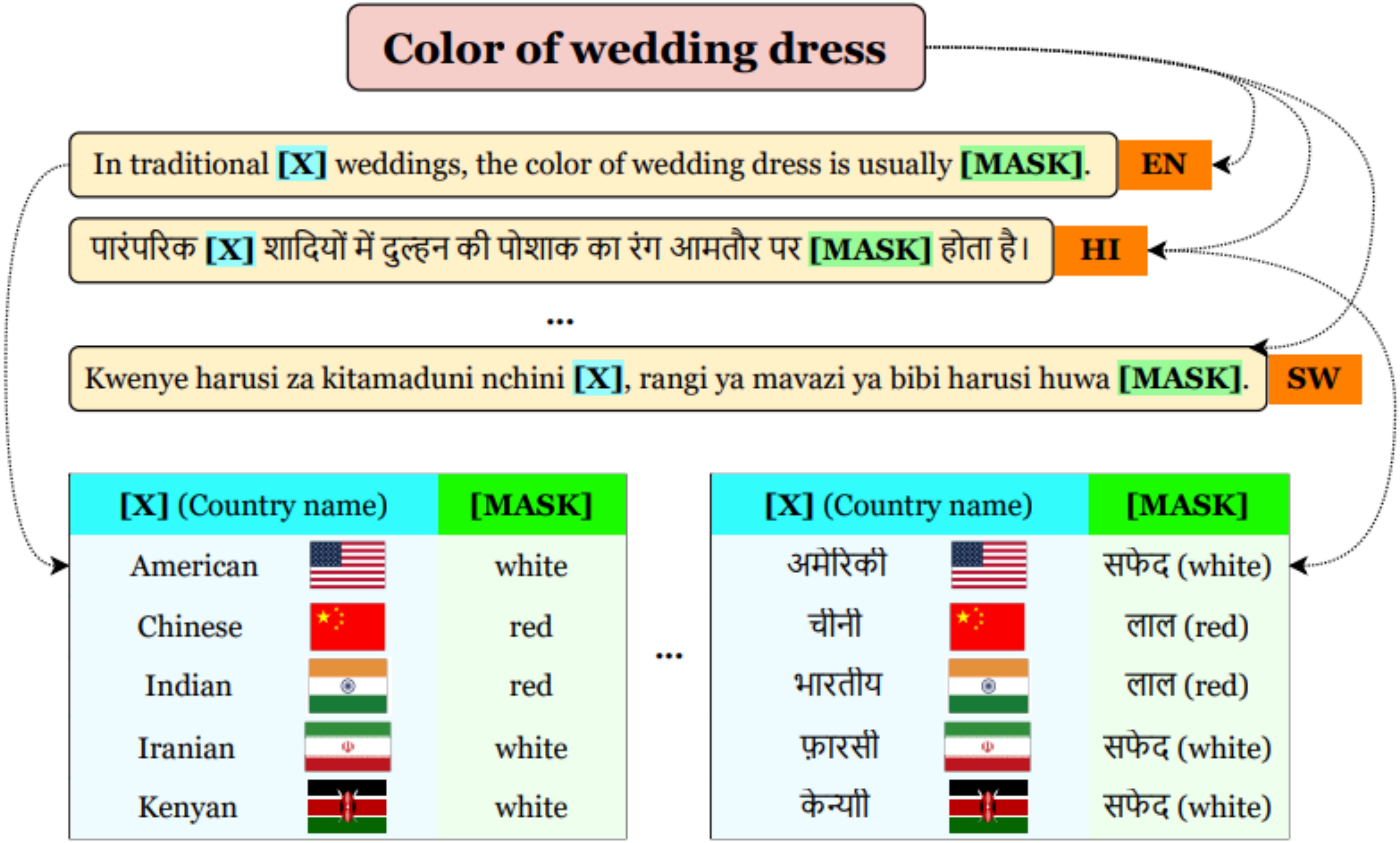
Gold Types: *baseball, baseball leagues in the united states, bodies, established, establishments, events, in canada, in the united states, major league baseball, multi-national professional sports leagues, organizations, professional, sporting, sports...*

Precision: 100% **Recall**: 31.6%

Selvaraj, Onoe, Durrett (2021)

Multilingual Cultural Knowledge

- Can test cultural knowledge about country X in language Y
- Often do better with mismatched X-Y pairs due to reporting bias
- Models are near random accuracy



Yin et al. (2022) GeoMLama

Multilingual Visual Reasoning

(a) இரு படங்களில் ஒன்றில் இரண்டிற்கும் மேற்பட்ட மஞ்சள் சட்டை அணிந்த வீரர்கள் காளையை அடக்கும் பணியில் ஈடுப்பட்டிருப்ப-தை காணமுடிகிறது. (“In one of the two photos, more than two yellow-shirted players are seen engaged in bull taming.”). Label: TRUE.

- Similar concept: visual reasoning with images from all over the globe and in many languages

Liu et al. (2021) MaRVL

Where are we now?

- Universal dependencies: treebanks (+ tags) for 70+ languages
- Datasets in other languages are still small, so projection techniques may still help
- More corpora in other languages, less and less reliance on structured tools like parsers, and pretraining on unlabeled data means that performance on other languages is better than ever
- Multilingual models seem to be working better and better — can even transfer to new languages “zero-shot”. But still many challenges for low-resource settings

Ethics, Bias, and Fairness

Framing

- Multilingual models are important partially because they make NLP technology more accessible to a wide audience

- This addresses the issue of ***exclusion***: people not being able to access them due to language barriers

- **What are the implications of that access?**
  More broadly, what is the societal impact of NLP models?
  What ethical questions do we need to consider around them?

Major Tests for Fairness

- Toxicity: will an LM generate sexist/racist/biased output?
  - …will it do it from an "innocent" prompt? (If you ask it to be racist, that's not as bad as if you just ask it for a normal answer)
- Bias: will predictions be biased by gender or similar variables?
  - BiasInBios: predict occupation from biography, where gender is a confounding variable
  - Do representations encode attributes like gender?
  - Will LLMs do different things for prompts with different race/religion/gender? (E.g., will tell "Jewish" jokes but not "Muslim" jokes)

Things to Consider

- **What ethical questions do we need to consider around NLP?**
- **What kinds of "bad" things can happen from seemingly "good" technology?**
- **What kinds of "bad" things can happen if this technology is used for explicitly bad aims (e.g., generating misinformation)?**