Multilingual / Cross-lingual Methods

Wei Xu

(many slides from Greg Durrett)

Alpaca and Llama

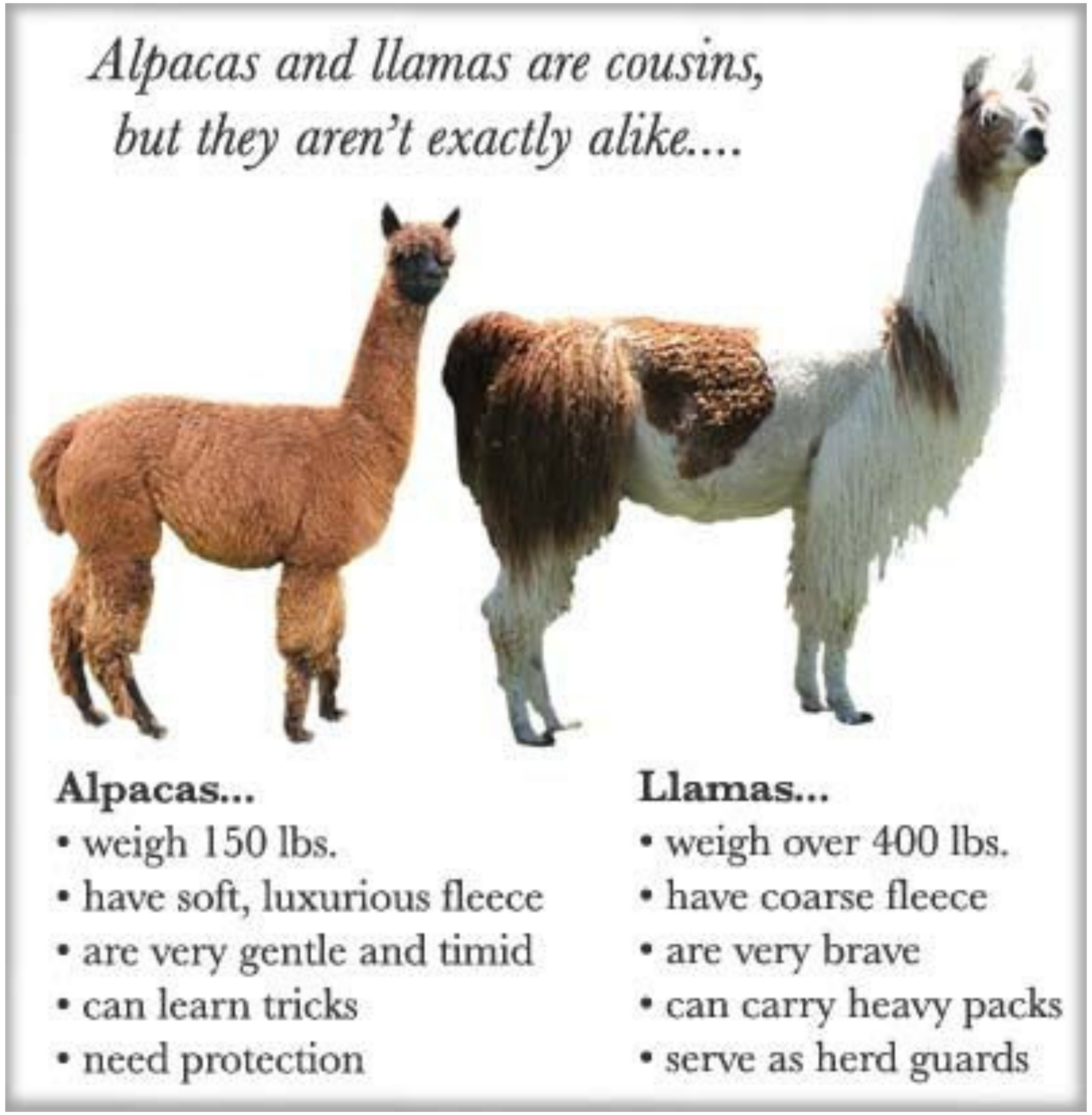


Image Credit: Unknown

This Lecture

- Morphology
- Word Segmentation
- Cross-lingual Tagging and Parsing
- Cross-Lingual Word Representations (Multilingual LLMs)
- Extras: Pairwise Ranking Model, Class-balanced Focal Loss

Morphology

What is morphology?

- Study of how words form
- Derivational morphology: create a new *lexeme* from a base

  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

Neologism

- Semantic shift, lexical derivation, dialectal variation, blending, or compounding, etc.

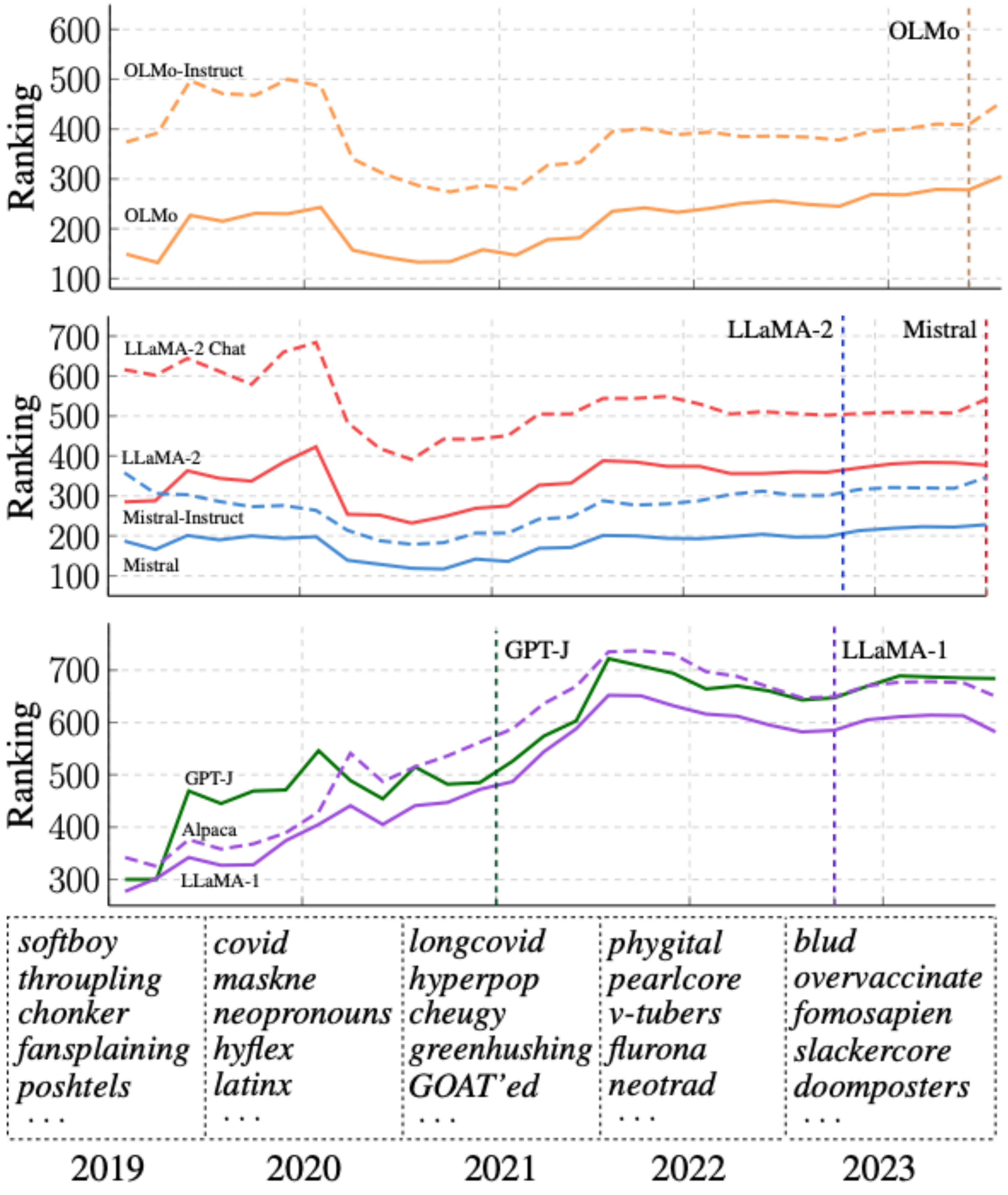


Figure 6: Rankings of neologisms over time compared to 5000 common words. Newer models are plotted separately. Dashed lines show model knowledge cutoffs[5]. Example neologisms from each year are provided, and neologisms without trendlines are reported at the end.

Zhang et al. (2024)

Morphological Inflection

- In English: 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 <u>die Kinder</u>

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

Irregular Inflection

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

Turkish	English
Muvaffak	Successful
Muvaffakiyet	Success
Muvaffakiyet**siz**	**Un**successful ('**without** success')
Muvaffakiyetsiz**leş**(*-mek*)	*(To)* **become** unsuccessful
Muvaffakiyetsizleş**tir**(*-mek*)	*(To)* **make one** unsuccessful
Muvaffakiyetsizleştiri**ci**	**Maker of** unsuccessful ones
Muvaffakiyetsizleştirici**leş**(*-mek*)	*(To)* **become** a maker of unsuccessful ones
Muvaffakiyetsizleştiricileş**tir**(*-mek*)	*(To)* **make one** a maker of unsuccessful ones
Muvaffakiyetsizleştiricileştiri**ver**(*-mek*)	*(To)* **easily/quickly** make one a maker of unsuccessful ones
Muvaffakiyetsizleştiricileştirivere**bil**(*-mek*)	*(To)* **be able to make** one easily/quickly a maker of unsuccessful ones
Muvaffakiyetsizleştiricileştirivere**meye**bil(*-mek*)	**Not** *(to)* be able to make one easily/quickly a maker of unsuccessful ones
Muvaffakiyetsizleştiricileştiriveremeyebil**ecek**	**One who is** not able to make one easily/quickly a maker of unsuccessful ones
Muvaffakiyetsizleştiricileştiriveremeyebilecek**ler**	**Those** who are not able to make one easily/quickly a maker of unsuccessful ones
Muvaffakiyetsizleştiricileştiriveremeyebilecekleri**miz**	Those who **we** cannot make easily/quickly a maker unsuccessful ones
Muvaffakiyetsizleştiricileştiriveremeyebileceklerimiz**den**	**From** those we can not easily/quickly make a maker of unsuccessful ones
Muvaffakiyetsizleştiricileştiriveremeyebileceklerimizden**miş**	**(Would be)** from those we can not easily/quickly make a maker of unsuccessful ones
Muvaffakiyetsizleştiricileştiriveremeyebileceklerimizdenmiş**siniz**	**You** would be from those we can not easily/quickly make a maker of unsuccessful ones
Muvaffakiyetsizleştiricileştiriveremeyebileceklerimizdenmişsiniz**cesine**	**Like** you would be from those we can not easily/quickly make a maker of unsuccessful ones

- 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 (2005-now): >100 languages
- Word piece / byte-pair encoding models for MT are pretty good at handling these if there's enough data

NLP performance inequalities

- NLP progress has been restricted to a minuscule subset of the world's 6,500 languages

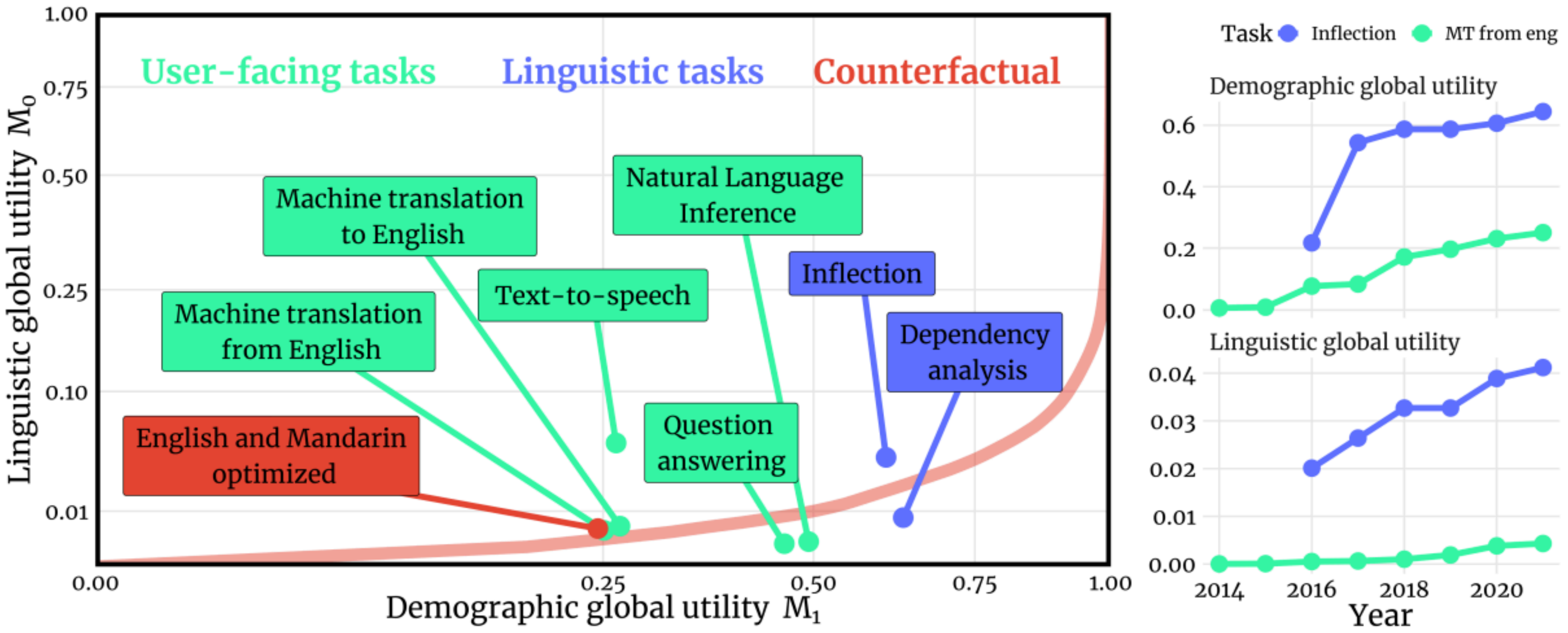
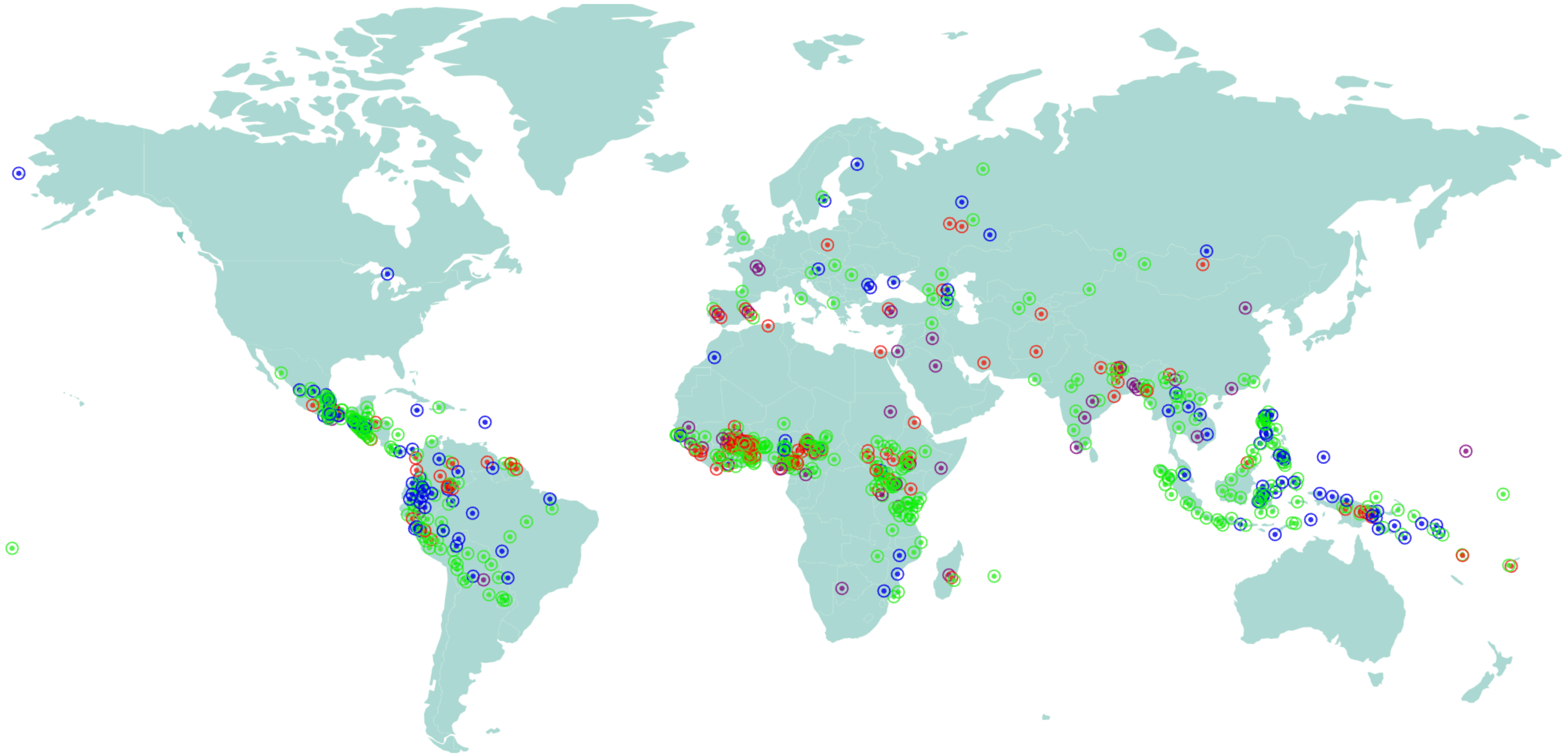


Figure 1: Left panel: linguistic and demographic global utility metrics for a number of language technology tasks. The red curve corresponds to the sequence where first the language with the largest number of users is set to utility 1, then the second, and so on. Right panel: recent historical progression of two language technology tasks: Inflection and Machine Translation from English.

Blasi (2021)

CMU Wilderness Multilingual Speech

- 650+ Languages
- 20 hours of aligned speech per language
- Data from read New Testaments (http://www.bible.is/)



http://festvox.org/cmu_wilderness/map.html Black (2019)

Universal Dependencies

- Over 100 languages

Current UD Languages

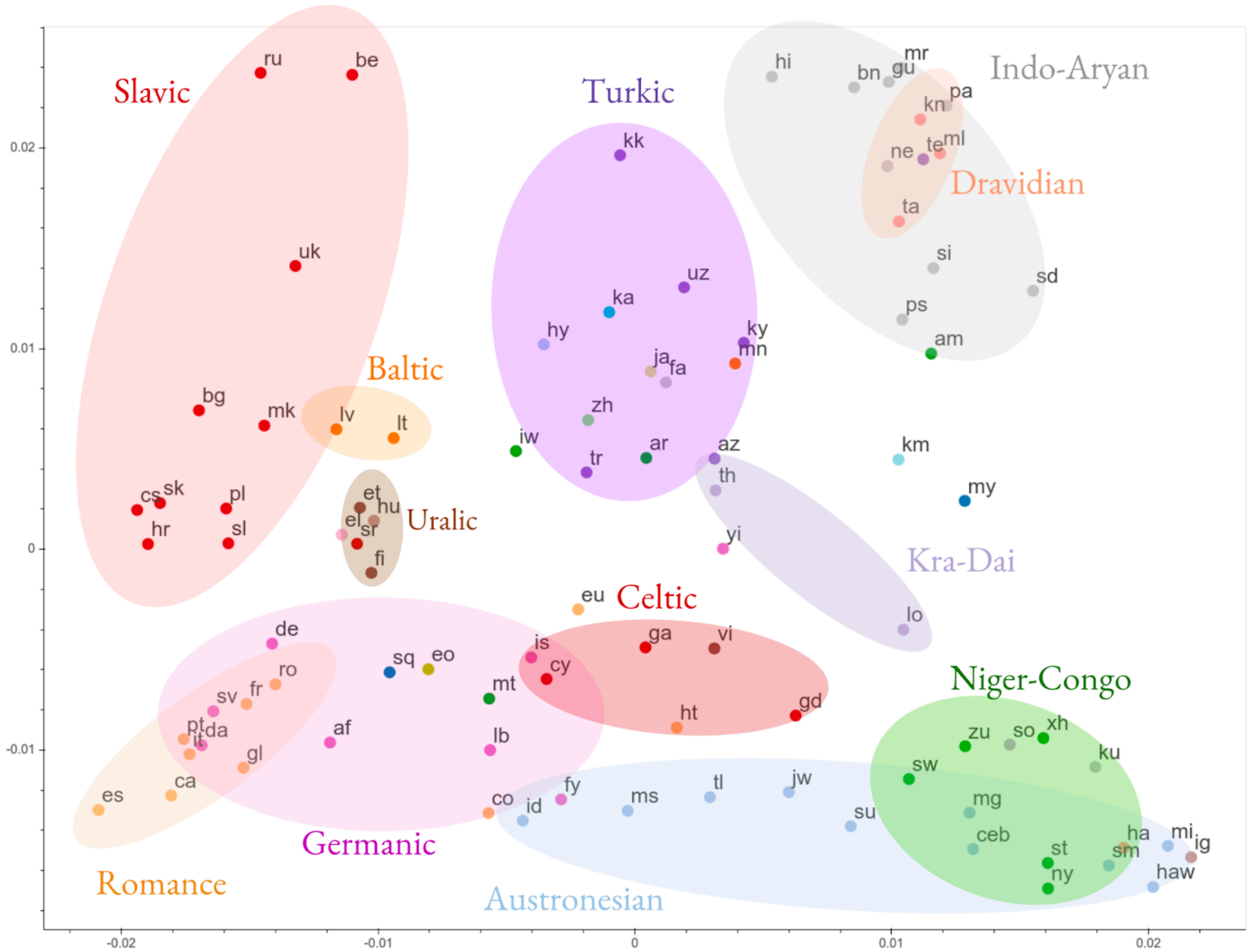
Information about language families (and genera for families with multiple branches) is mostly taken from WALS Online (IE = Indo-European).

Abaza	1	<1K	Northwest Caucasian
Afrikaans	1	49K	IE, Germanic
Akkadian	2	25K	Afro-Asiatic, Semitic
Akuntsu	1	1K	Tupian, Tupari
Albanian	1	<1K	IE, Albanian
Amharic	1	10K	Afro-Asiatic, Semitic
Ancient Greek	2	416K	IE, Greek
Ancient Hebrew	1	39K	Afro-Asiatic, Semitic
Apurina	1	<1K	Arawakan
Arabic	3	1,042K	Afro-Asiatic, Semitic
Armenian	2	94K	IE, Armenian
Assyrian	1	<1K	Afro-Asiatic, Semitic
Bambara	1	13K	Mande
Basque	1	121K	Basque
Beja	1	<1K	Afro-Asiatic, Cushitic
Belarusian	1	305K	IE, Slavic
Bengali	1	<1K	IE, Indic
Bhojpuri	1	6K	IE, Indic
Breton	1	10K	IE, Celtic
Bulgarian	1	156K	IE, Slavic
Buryat	1	10K	Mongolic
Cantonese	1	13K	Sino-Tibetan
Catalan	1	553K	IE, Romance
Cebuano	1	1K	Austronesian, Central Philippine
Chinese	6	287K	Sino-Tibetan
Chukchi	1	6K	Chukotko-Kamchatkan
Classical Chinese	1	310K	Sino-Tibetan

https://universaldependencies.org/

Massively Multilingual MT

- For 103 languages



Arivazhagan et al. (2019), Kudugunta et al. (2019)

Massively Multilingual MT

- For 200 languages (54B parameters)
  - Mixture of Expert (BOE) model. With more low-resource language pairs in the training data, the multilingual systems start to overfit.
  - Solutions: regularization, curriculum learning, self-supervised learning, and diversifying back-translation.

	eng_Latn-xx			xx-eng_Latn		
	MMTAfrica	M2M-100*	NLLB-200	MMTAfrica	M2M-100*	NLLB-200
hau_Latn	-/-	4.0/-	**33.6**/**53.5**	-/-	16.3/-	**38.5**/**57.3**
ibo_Latn	21.4/-	19.9/-	**25.8**/**41.4**	15.4/-	12.0/-	**35.5**/**54.4**
lug_Latn	-/-	7.6/-	**16.8**/**39.8**	-/-	7.7/-	**27.4**/**46.7**
luo_Latn	-/-	13.7/-	**18.0**/**38.5**	-/-	11.8/-	**24.5**/**43.7**
swh_Latn	**40.1**/-	27.1/-	37.9/**58.6**	28.4/-	25.8/-	**48.1**/**66.1**
wol_Latn	-/-	8.2/-	**11.5**/**29.7**	-/-	7.5/-	**22.4**/**41.2**
xho_Latn	27.1/-	-/-	**29.5**/**48.6**	21.7/-	-/-	**41.9**/**59.9**
yor_Latn	12.0/-	13.4/-	**13.8**/**25.5**	9.0/-	9.3/-	**26.6**/**46.3**
zul_Latn	-/-	19.2/-	**36.3**/**53.3**	-/-	19.2/-	**43.4**/**61.5**

Table 31: **Comparison on FLORES-101 devtest on African Languages.** We compare to two

Fan et al. (2022), NLLB Team (2022)

Word Segmentation

Chinese Word Segmentation

- Word segmentation: some languages including Chinese do not have white spaces between words.
- LSTMs over character embeddings / character bigram embeddings to predict word boundaries

没有登录 讨论 贡献 创建账号 登录

维基百科
自由的百科全书

条目 讨论 大陆简体 汉 漢 阅读 编辑 查看历史 搜索维基百科

首页
分类索引
特色内容
新闻动态
最近更改
随机条目
资助维基百科

帮助

帮助
维基社群
方针与指引
互助客栈
知识问答
字词转换
IRC即时聊天
联络我们
关于维基百科

工具

链入页面
相关更改
上传文件

中文维基百科Facebook粉丝专页正式上线，邀请大家一同关注。 [关闭]

佐治亚理工学院 [编辑]

维基百科，自由的百科全书

坐标：33°46′33″N 84°23′41″W

此条目需要编修，以确保文法、用词、语气、格式、标点等使用恰当。 (2018年1月5日)
请按照校对指引，帮助编辑这个条目。（帮助、讨论）

佐治亚理工学院（英语：Georgia Institute of Technology，简称**Georgia Tech**或**Tech**，常缩写为**Gatech**），是美国一所顶尖公立研究型大学，始建于1885年，是美国大学协会、大学研究协会的成员之一。学校总部位于美国佐治亚州首府亚特兰大市，与艾文理大学同处于一个城市。除了亚特兰大主校区，也在佐治亚州沙瓦纳、法国洛林大区的首府梅斯开设分校，在爱尔兰共和国的阿斯隆市、新加坡国立大学设有联合研究所。

佐治亚理工以工程学、计算机、材料科学、建筑学、管理学闻名于世，与东北部的麻省理工、西部的加州理工并称美国三大理工学院[5][6][7]。其下属的GTRI航空系统设计实验室承担了美国政府的一些机密的重大科研项目。校本部拥有143幢建筑物，校园曾是1996年夏季奥林匹克运动会的奥运村，拥有奥运级体育场馆、酒店、医院、科技广场等其他设施。

佐治亚理工工曾与上海交通大学合作开设双学位项目，现于中国深圳开办天津大学佐治亚理工深圳学院（GTSI）。

佐治亚理工学院

老校名	Georgia School of Technology
校训	Progress and Service
创办时间	1885年10月13日
IPEDS编码	139755

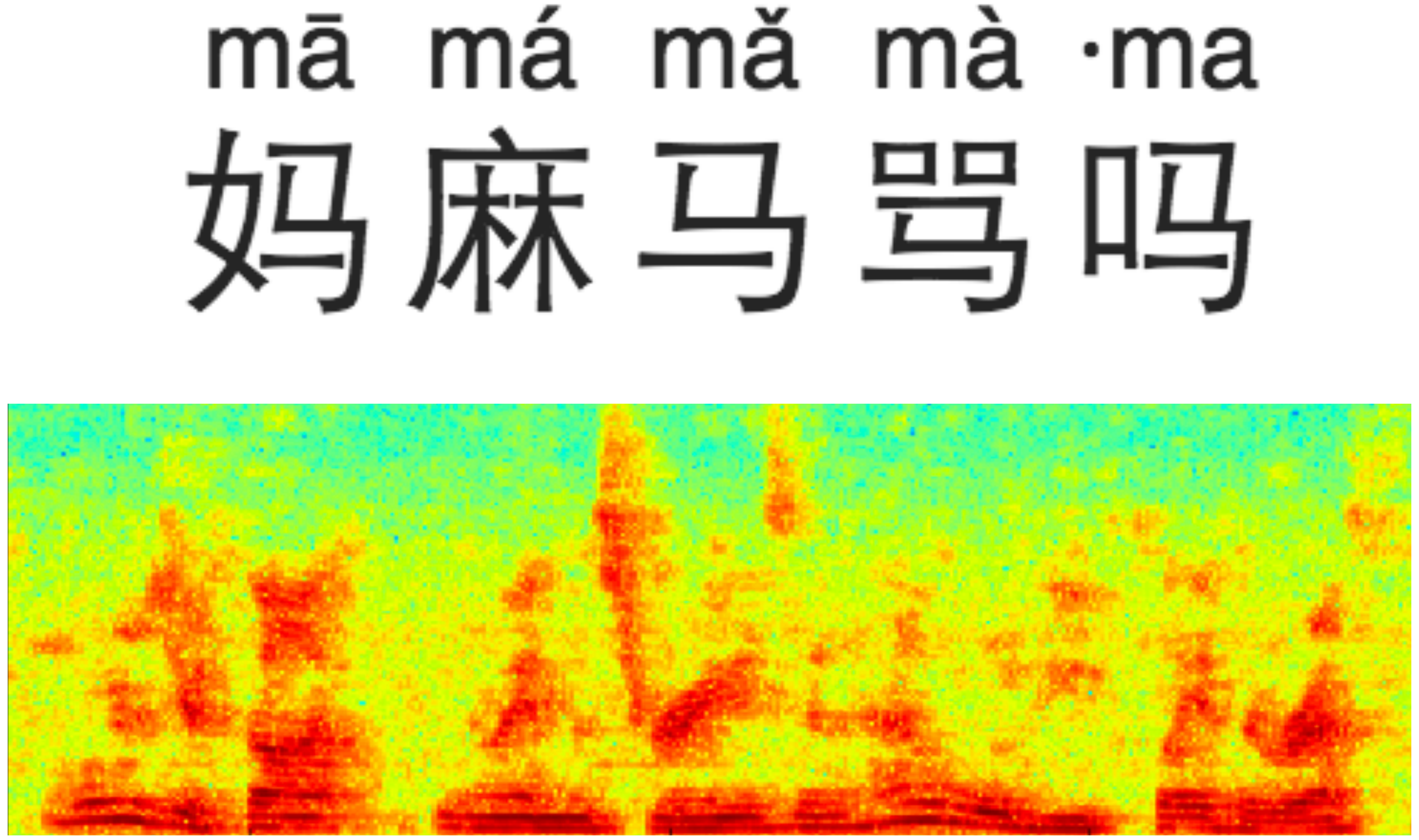
Challenges of Chinese

- Thousands of characters!   >80K

的	不	用	一	新	时	吗	所	被	说	男	事
人	在	之	二	市	也	家	后	只	着	女	得
上	于	要	三	与	还	可	分	都	位	几	真
中	前	好	四	内	出	件	种	做	把	各	对
下	者	了	五	本	去	里	将	已	吧	谁	看
大	会	年	六	地	到	最	很	长	难	找	见
小	号	月	十	这	他	回	而	行	来	子	加
是	我	日	个	此	性	万	数	等	站	字	更
没	和	为	次	建	就	能	天	再	每	那	多
有	你	名	元	全	部	爱	无	以	起	哪	少

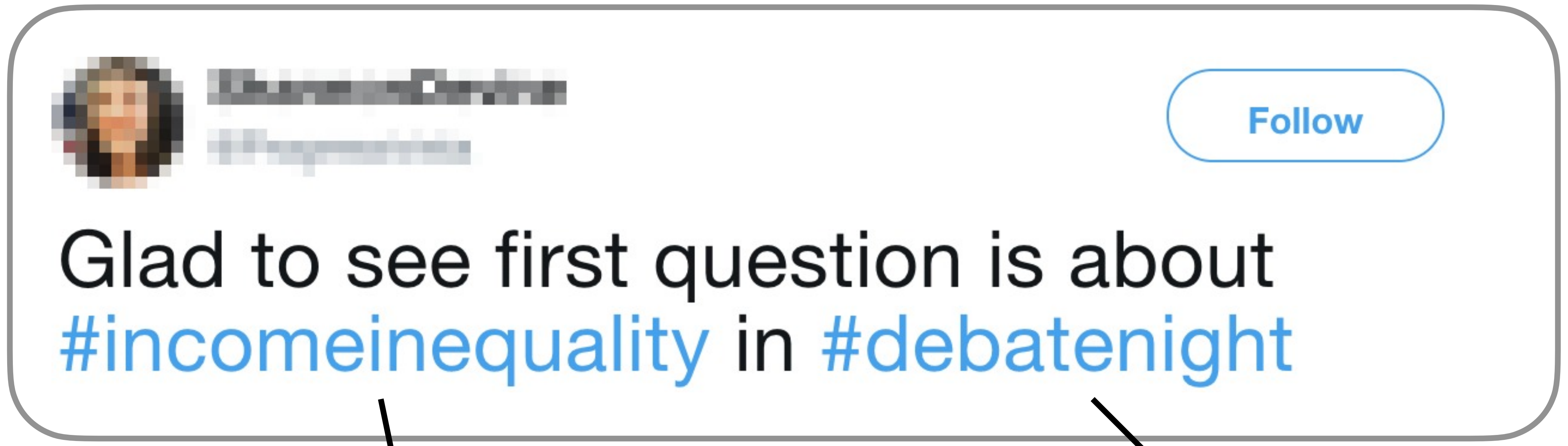
Challenges of Chinese

- Mandarin and Cantonese are both *tonal* languages.
- The homophone problem is ubiquitous.



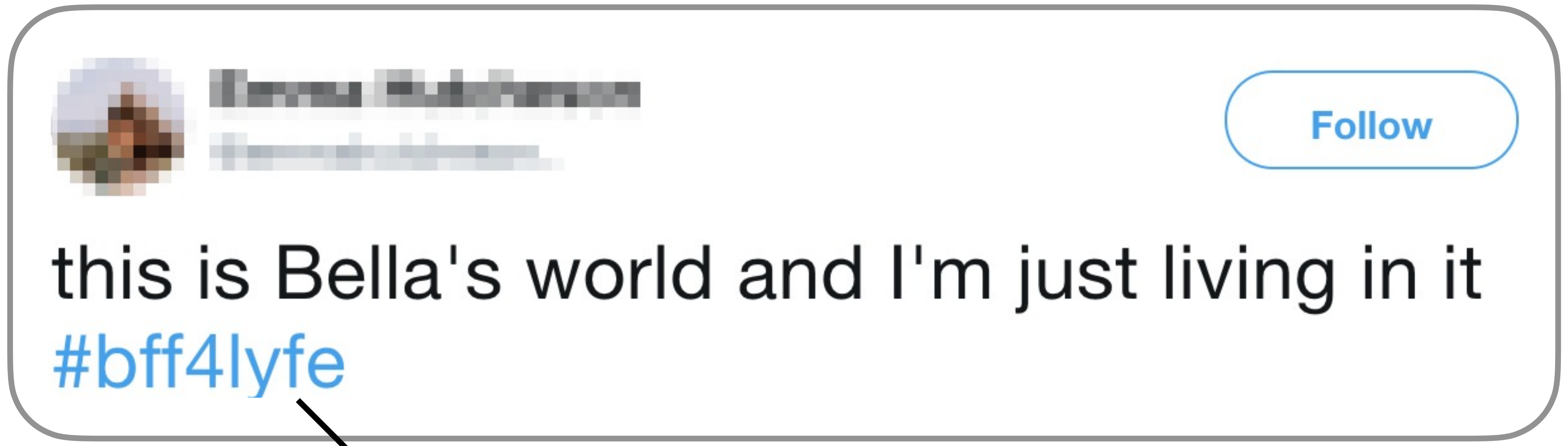
English Word Segmentation?

A case study: Hashtag Segmentation



income inequality # debate night

conveys the **topic** of the tweet



bff 4 lyfe

conveys the **sentiment** of the tweet

Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. “Multi-task Pairwise Neural Ranking for Hashtag Segmentation” in ACL (2019)

Hashtag Segmentation

- Challenges: entities, abbreviations, non-standard spellings, slang …

\# pawpawty
Microsoft's WordBreaker
(Wang et al., 2011)

\# pawpaw ty
GATE's hashtag tokenizer
(Maynard & Greenword, 2014)

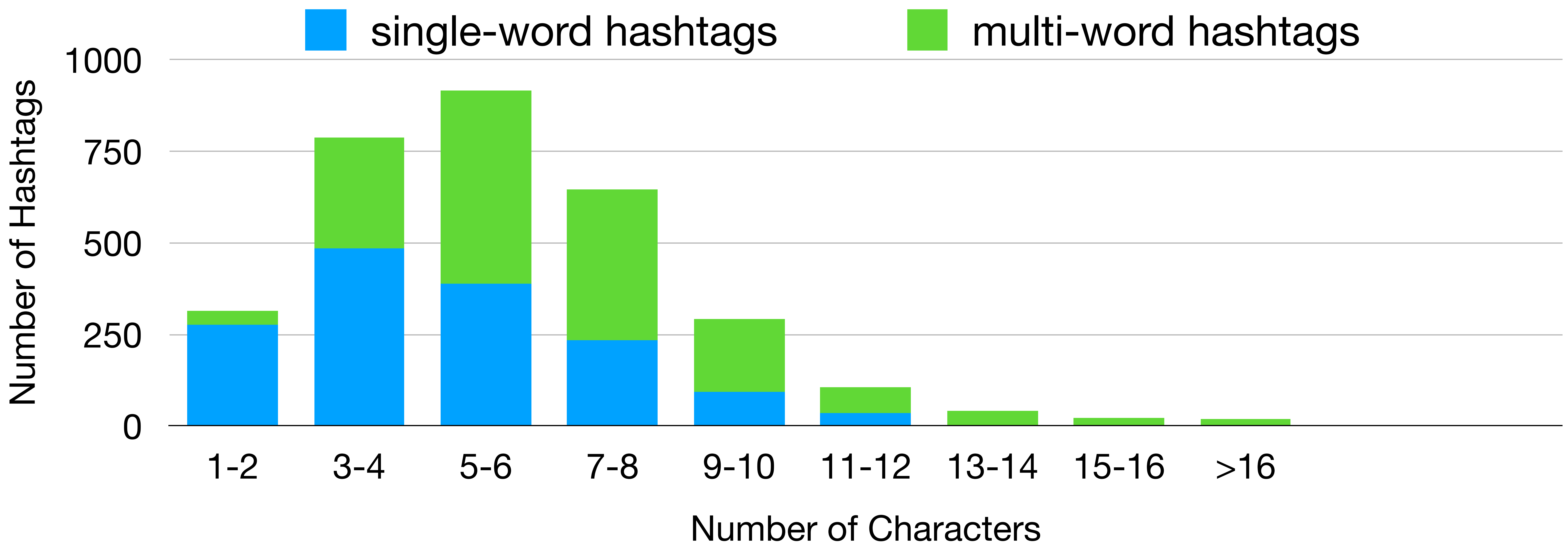
\# pawpawty
(Çelebi & Özgür, 2017)

\# paw pawty
HashtagMaster
(Our Work) ✓

Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. "Multi-task Pairwise Neural Ranking for Hashtag Segmentation" in ACL (2019)

Hashtag Segmentation

- Most hashtags have <15 characters. We can (almost) enumerate all 2^(1-len) possible segmentations.



Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. “Multi-task Pairwise Neural Ranking for Hashtag Segmentation” in ACL (2019)

Hashtag Segmentation

- It's also very hard to tell apart the top-ranked ones.

input hashtag

h: #songsongaddafisitunes

s_1: # song song addafis itunes

s_2: **# songs on gaddafi s itunes**

s_3: # songs on gaddaf is itunes

....

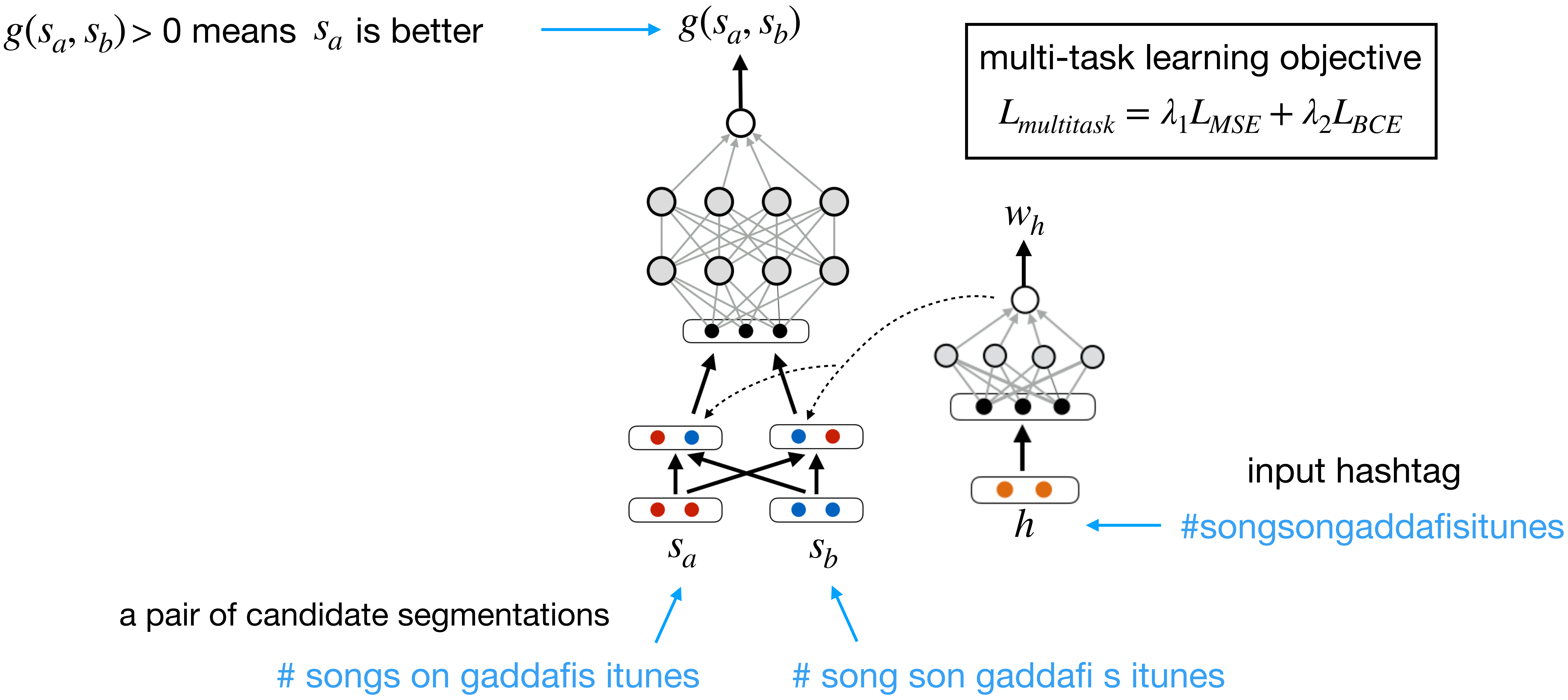
s_k: # song son gaddafis itunes

candidate segmentations (top-k)

Mounica Maddela, Wei Xu, Daniel Preoțiuc-Pietro. "Multi-task Pairwise Neural Ranking for Hashtag Segmentation" in ACL (2019)

Hashtag Segmentation

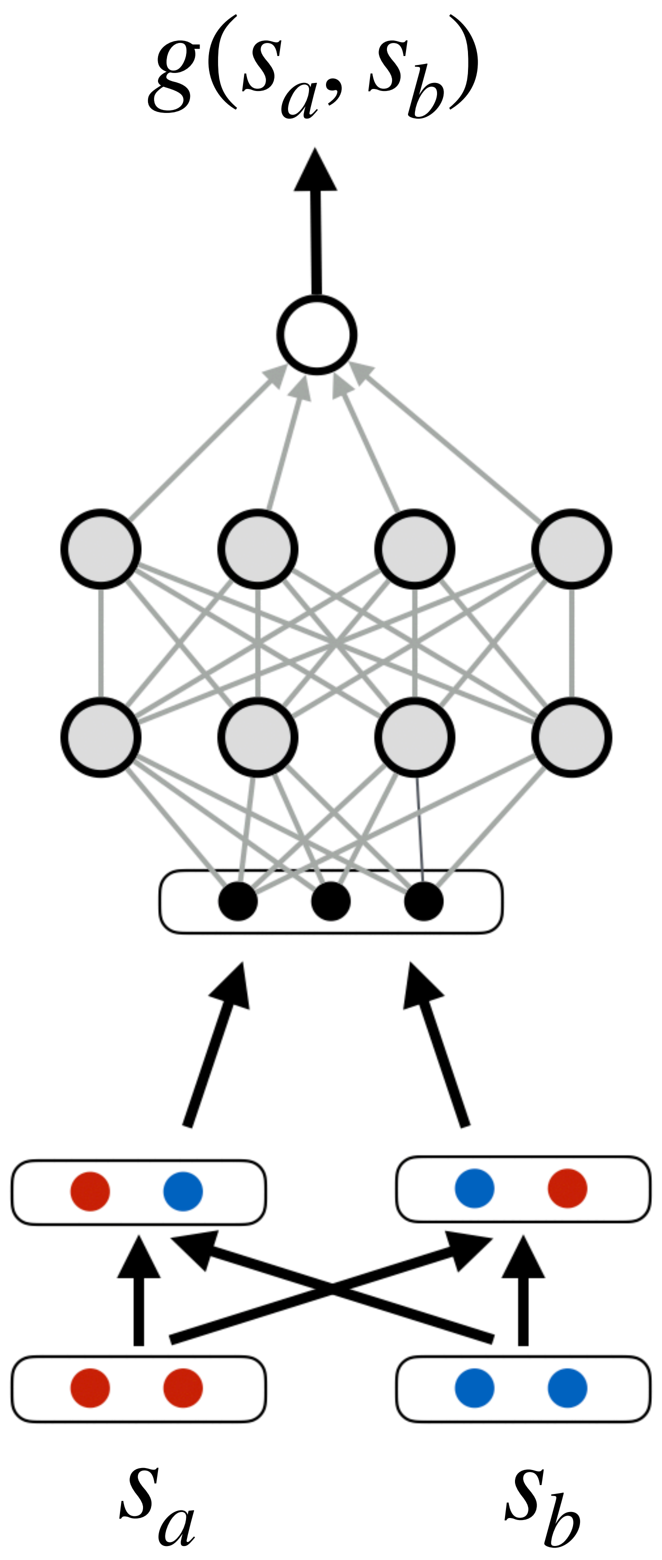
- Solution: pairwise ranking!



Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. “Multi-task Pairwise Neural Ranking for Hashtag Segmentation” in ACL (2019)

Hashtag Segmentation

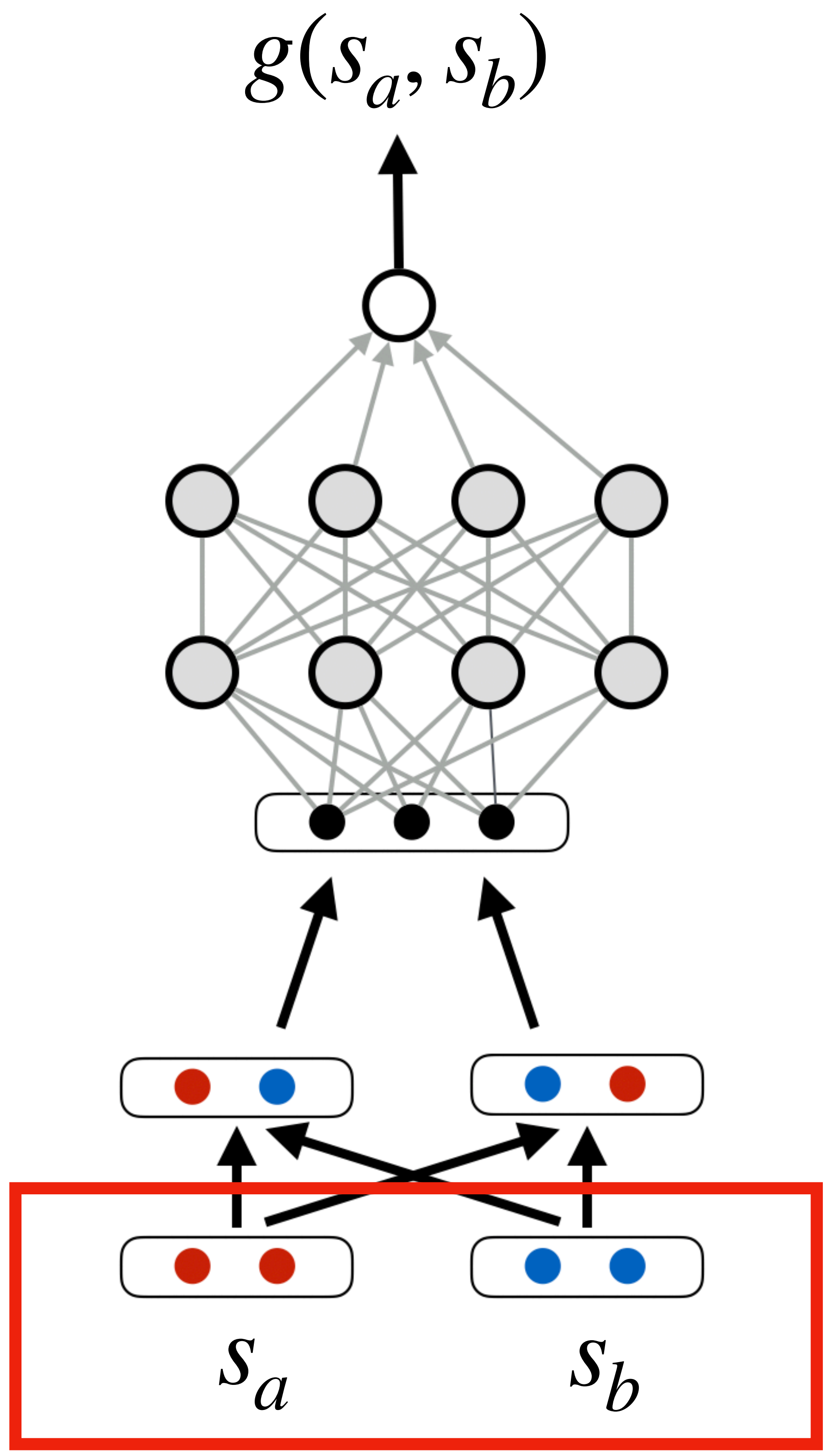
- So we can more easily compare very similar segmentations. We rerank the top-k candidates.



Mounica Maddela, Wei Xu, Daniel Preoțiuc-Pietro. "Multi-task Pairwise Neural Ranking for Hashtag Segmentation" in ACL (2019)

Hashtag Segmentation

- The neural pairwise ranking model uses a small number of numerical/ binary features.



Good Turing Smoothing
- Twitter
- Gigaword

Kneser-Ney Smoothing
- Twitter
- Gigaword

Ngram Language Model Probabilities

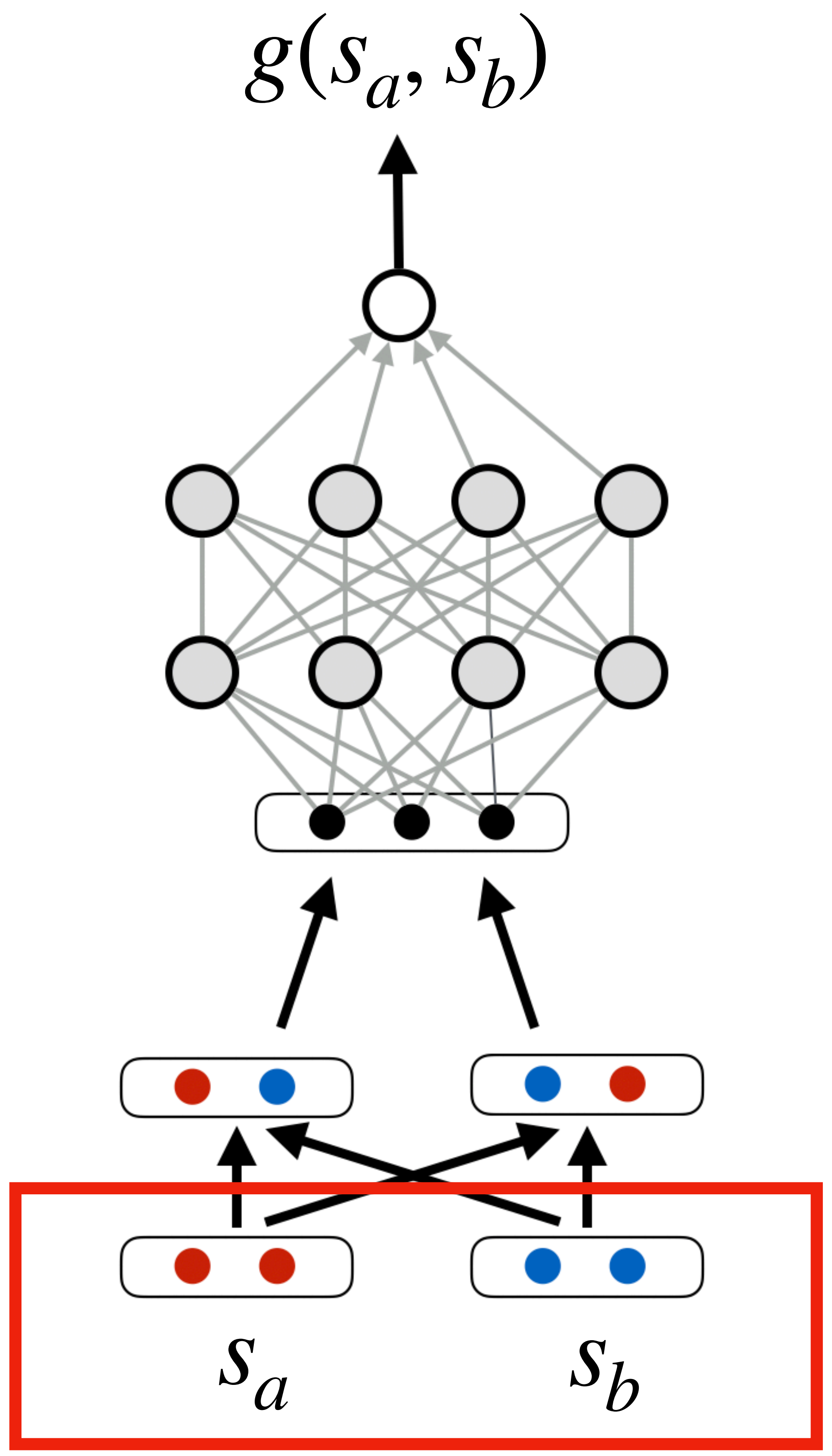
Word length
Number of words
Word shapes
Urban Dictionary
Named entities
Google counts

Linguistic Features

Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. “Multi-task Pairwise Neural Ranking for Hashtag Segmentation” in ACL (2019)

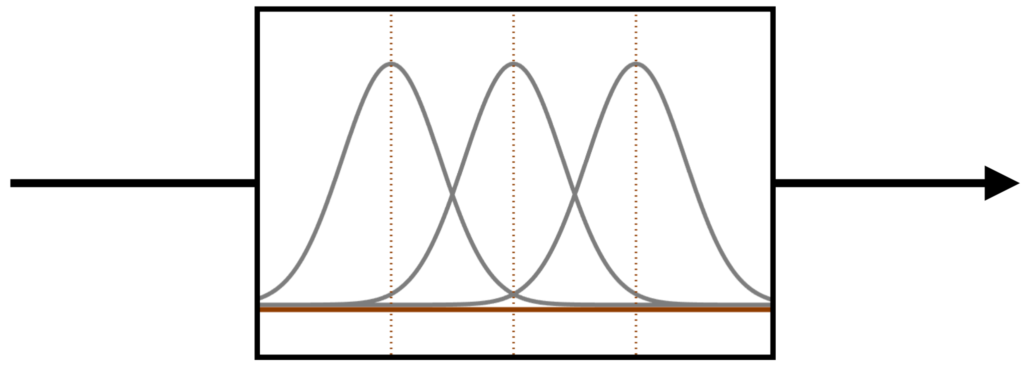
Hashtag Segmentation

- Vectorize numerical/binary features.



Gaussian Vectorization

$f_1(s_a) = 0.41$



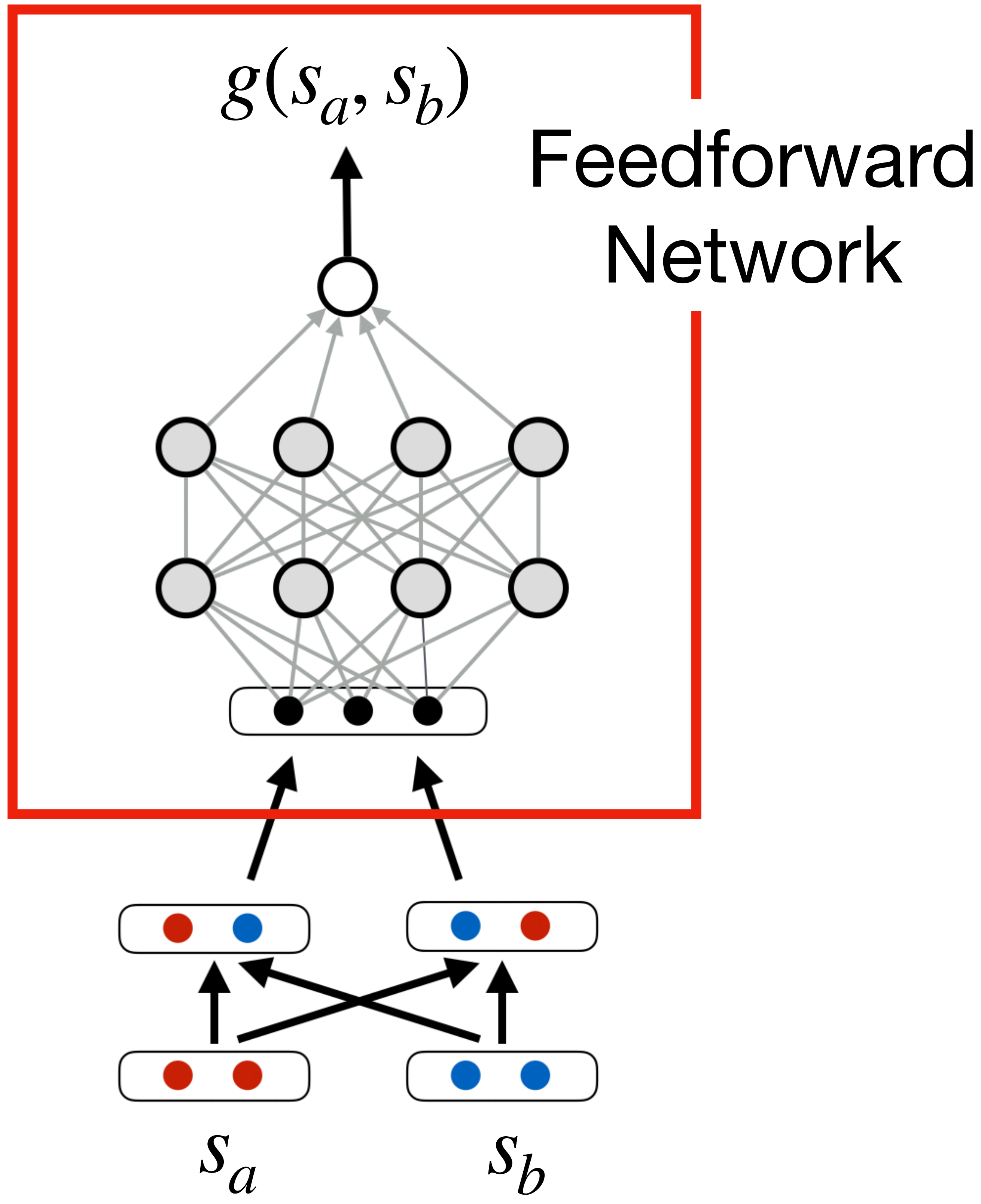
$\overrightarrow{f_1(s_a)}$ = [~0.0, **0.44, 0.54,** ~0.02, ~0.0]

$$d_j(f(\cdot)) = e^{-\frac{(f(\cdot)-\mu_j)^2}{2\sigma^2}}$$

Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. “Multi-task Pairwise Neural Ranking for Hashtag Segmentation” in ACL (2019)

Hashtag Segmentation

- Trained with mean squared error (MSE) or margin ranking loss.



Predicted Pairwise Score

$$L_{MSE} = \frac{1}{m} \sum_{i=1}^{m} (g^{*(i)}(s_a, s_b) - g^{(i)}(s_a, s_b))^2$$

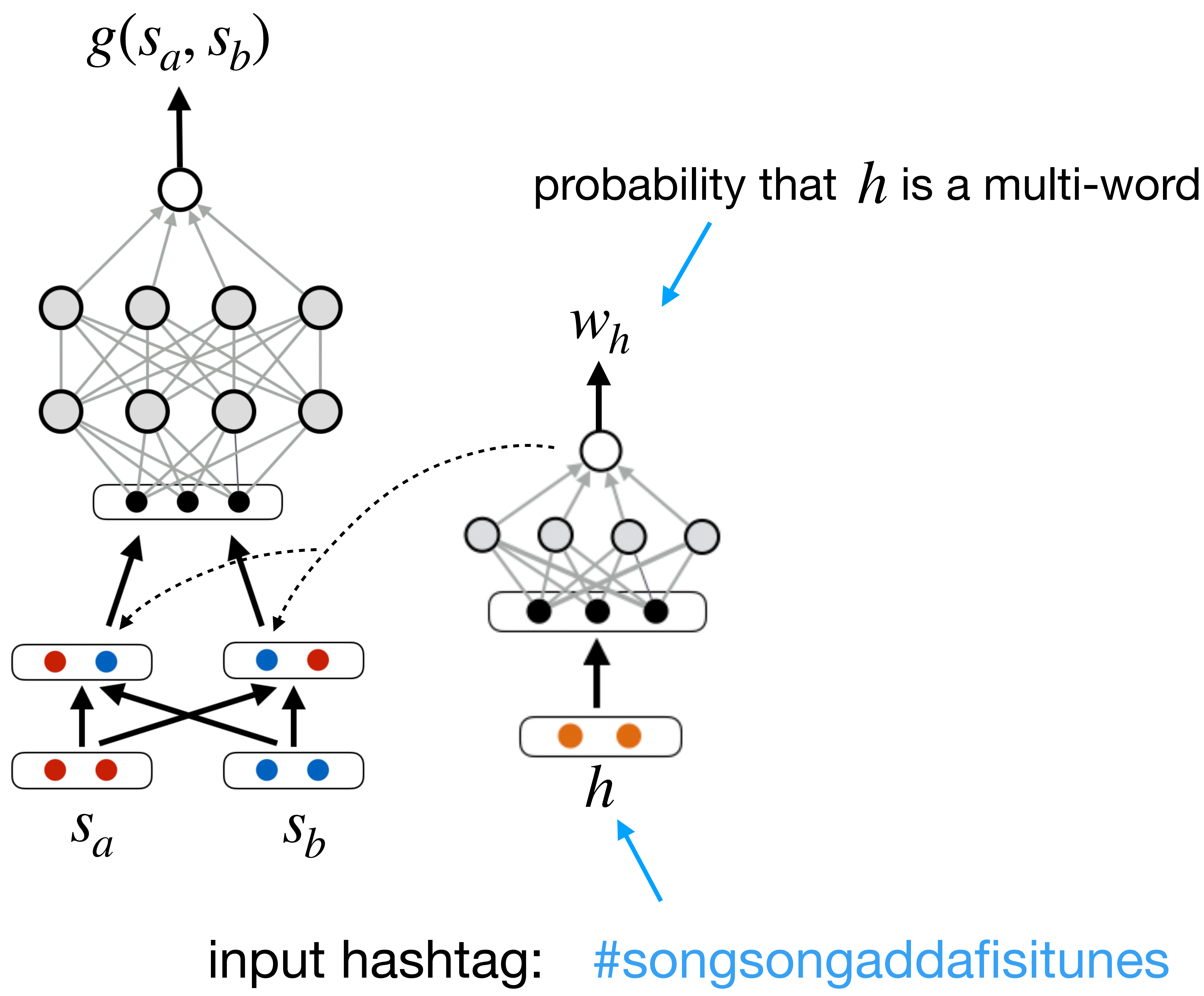
Gold Pairwise Score

$g^*(s_a, s_b) = sim(s_a, s^*) - sim(s_b, s^*)$, where s^* is the gold segmentation.

Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. "Multi-task Pairwise Neural Ranking for Hashtag Segmentation" in ACL (2019)

Hashtag Segmentation

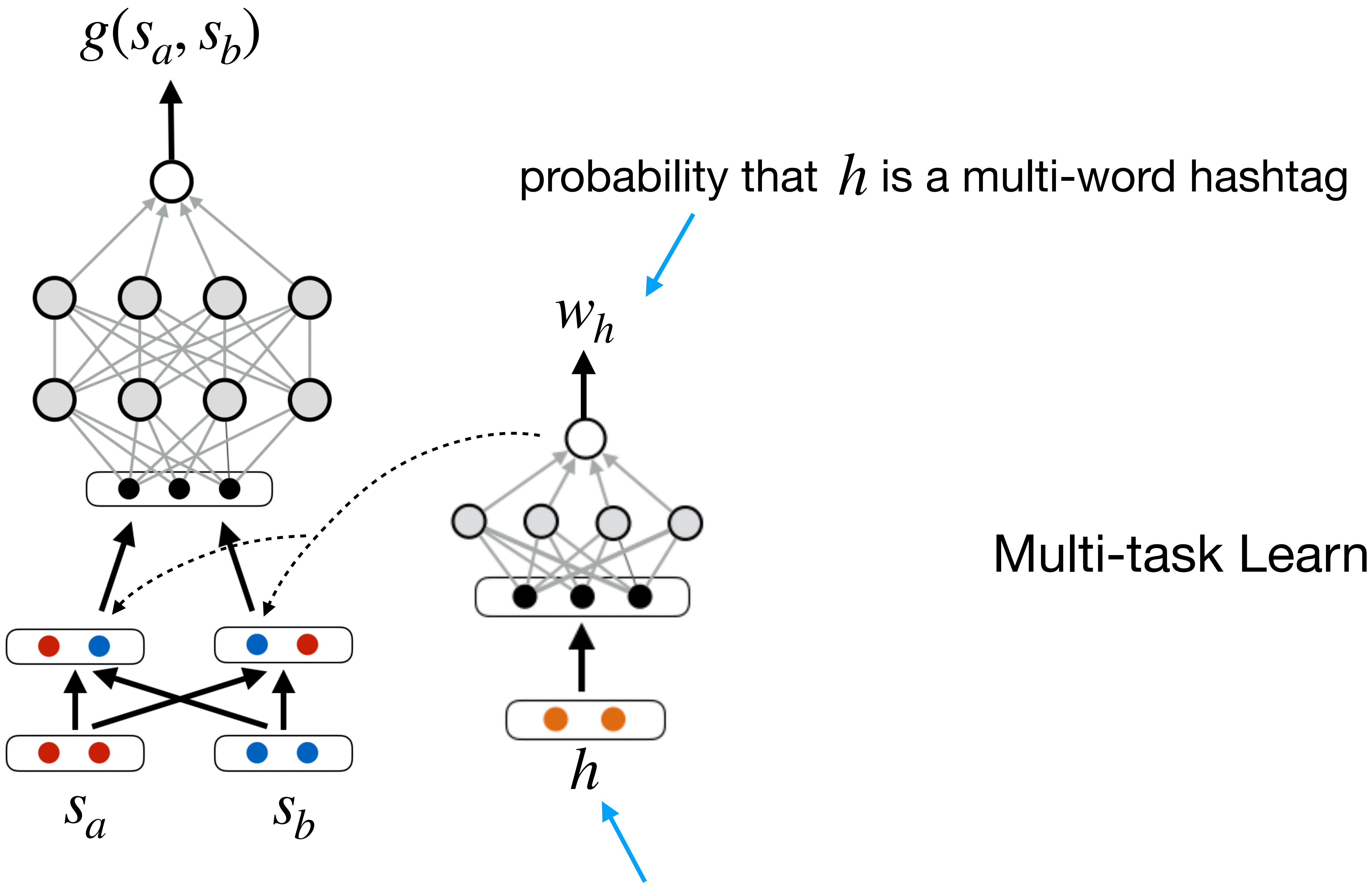
- Adaptive multi-task learning: as different features work for single- vs. multi-word hashtags, we introduce a binary classification task.



Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. "Multi-task Pairwise Neural Ranking for Hashtag Segmentation" in ACL (2019)

Hashtag Segmentation

- Adaptive multi-task learning: as different features work for single- vs. multi-word hashtags, we introduce a binary classification task.



Pairwise Ranking (main)

Multi-task Learning Objective: $L_{multitask} = \lambda_1 L_{MSE} + \lambda_2 L_{BCE}$

Binary Classification (auxiliary)

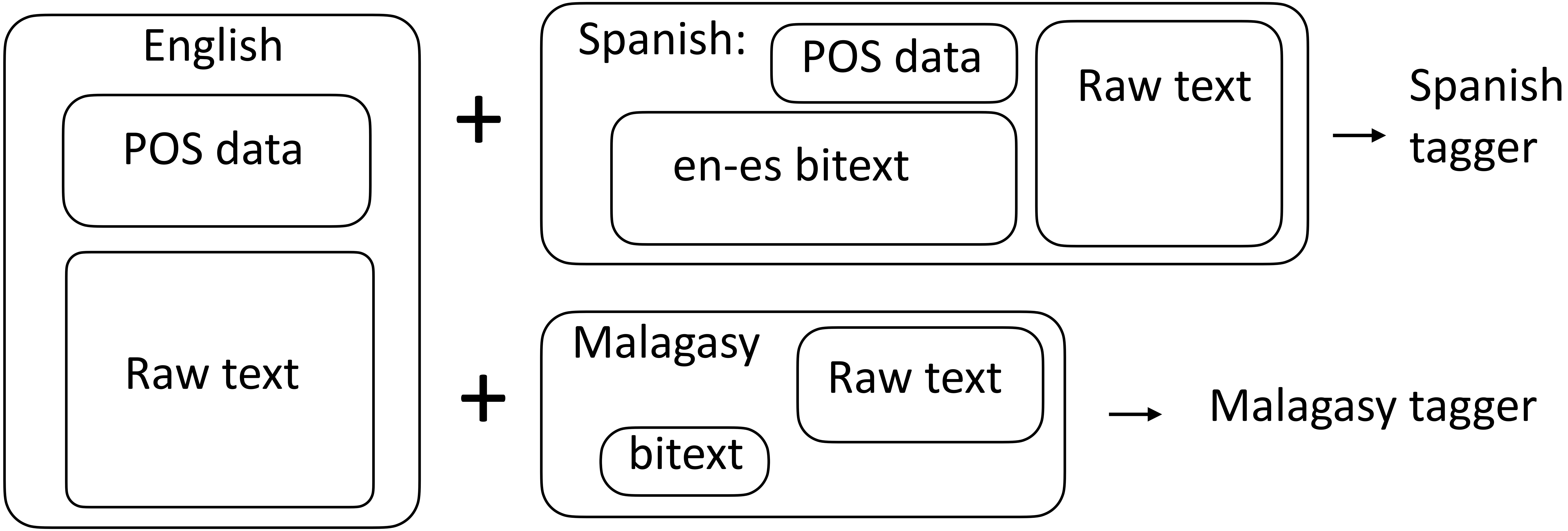
input hashtag: #songsongaddafisitunes

Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. "Multi-task Pairwise Neural Ranking for Hashtag Segmentation" in ACL (2019)

Cross-Lingual Tagging and Parsing

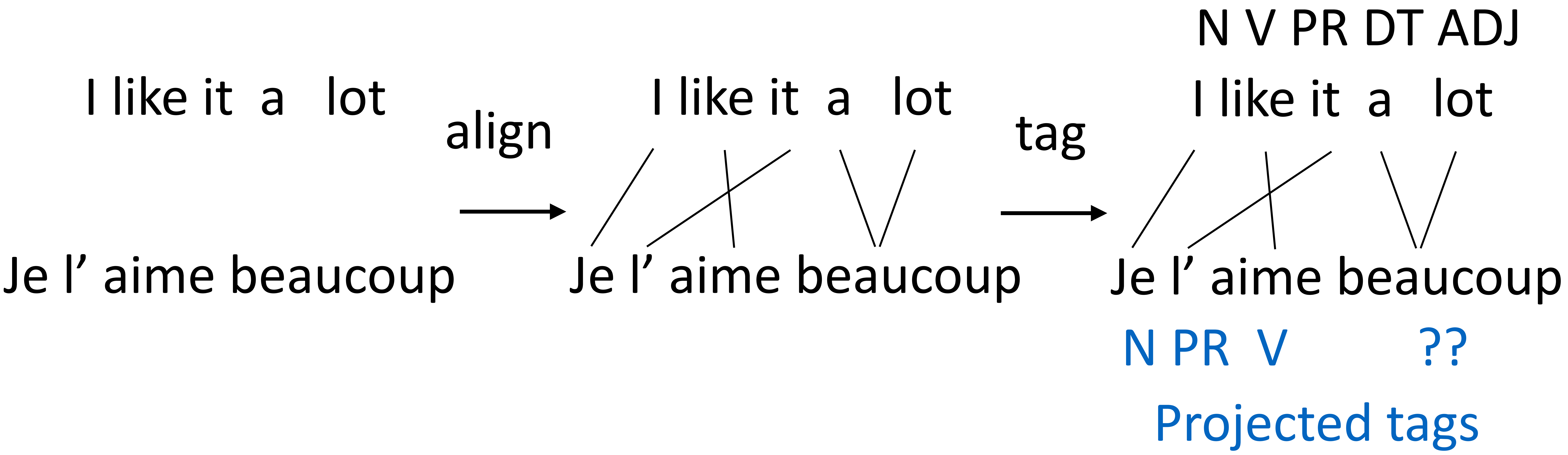
Cross-Lingual Tagging

- 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?

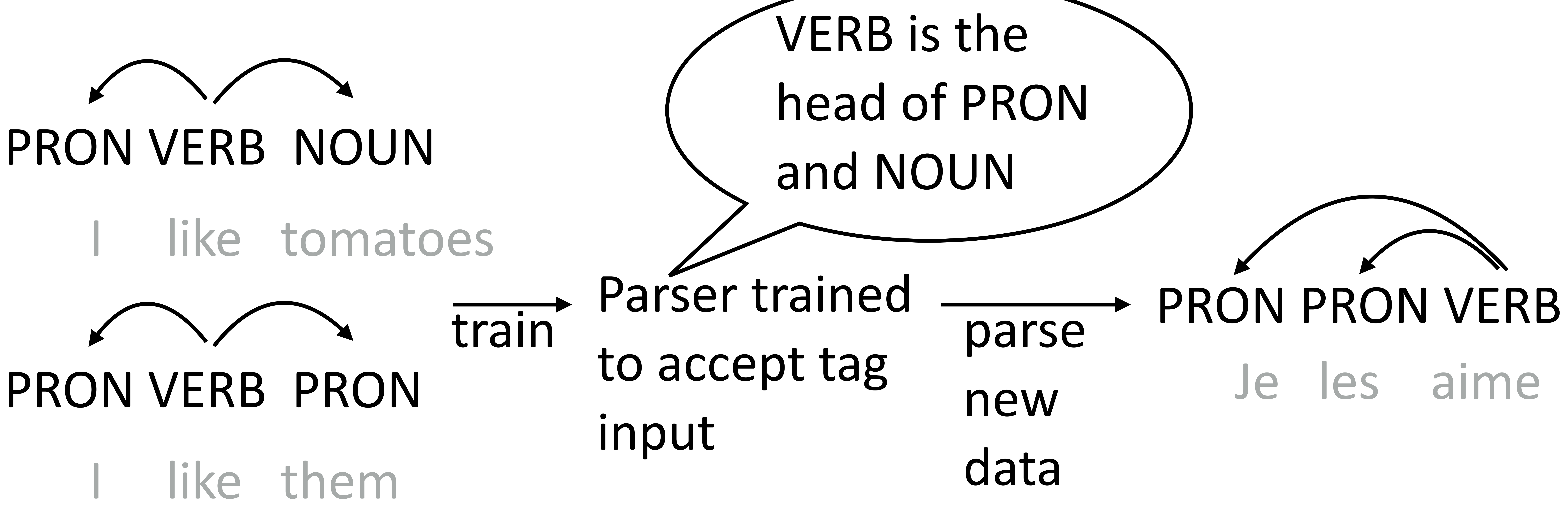


- Tag with English tagger, project across bitext, train French tagger? Works pretty well

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		pred-POS	
	source	gold-POS	gold-POS	multi-dir.	multi-proj.	multi-dir.	multi-proj.
da	it	48.6	46.3	48.9	49.5	46.2	47.5
de	nl	55.8	48.9	56.7	56.6	51.7	52.0
el	en	63.9	51.7	60.1	65.1	58.5	63.0
es	it	68.4	53.2	64.2	64.5	55.6	56.5
it	pt	69.1	58.5	64.1	65.0	56.8	58.9
nl	el	62.1	49.9	55.8	65.7	54.3	64.4
pt	it	74.8	61.6	74.0	75.6	67.7	70.3
sv	pt	66.8	54.8	65.3	68.0	58.3	62.1
avg		63.7	51.6	61.1	63.8	56.1	59.3

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

McDonald et al. (2011)

EasyProject

- Rely on a robust MT system (w/ or w/o word alignment) to do label projection:

Figure 1: An example of translating and projecting English ACE event triggers and named entities to Chinese. (a) Label projection pipeline starts with machine translation of the English sentence to Chinese, followed by word-to-word alignment. Then, label spans are projected based on word alignments. (b) Markers are inserted around entity and event trigger spans in the text. The modified sentence with markers inserted is then fed as input to a machine translation system, projecting the label span markers to the target language as a byproduct of translation.

Chen et al. (2023)

Cross-Lingual Word Representations (Multilingual LLMs)

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

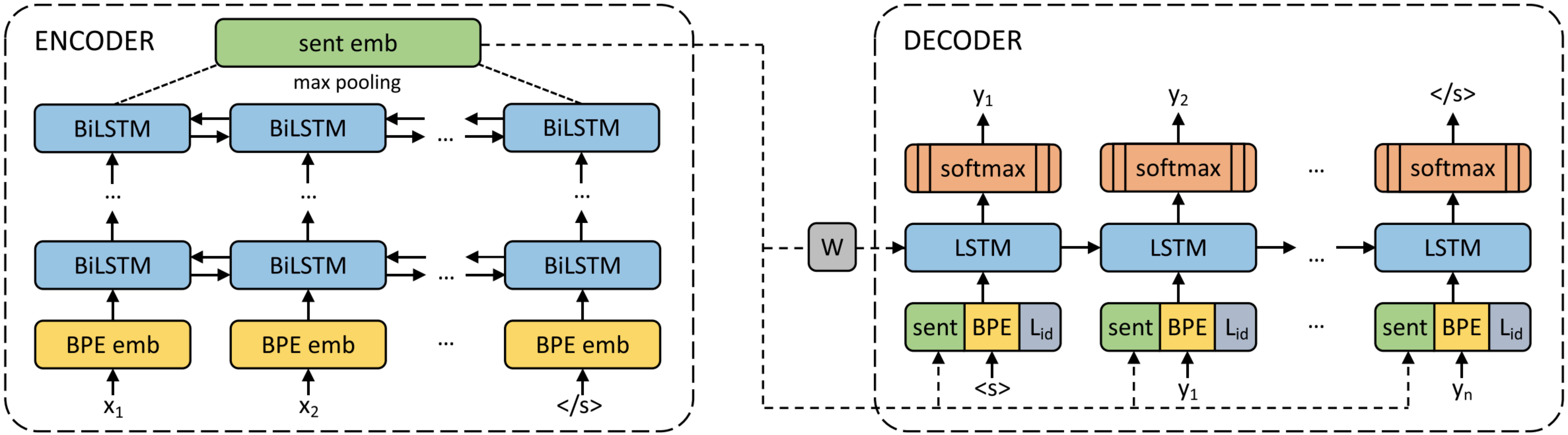
ID: 47
I Je J'

- 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 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

Devlin et al. (2019)

Multilingual BERT: Results

Fine-tuning \ Eval	EN	DE	NL	ES
EN	**90.70**	69.74	77.36	73.59
DE	73.83	**82.00**	76.25	70.03
NL	65.46	65.68	**89.86**	72.10
ES	65.38	59.40	64.39	**87.18**

Table 1: NER F1 results on the CoNLL data.

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)

XLM-RoBERTa (XLM-R)

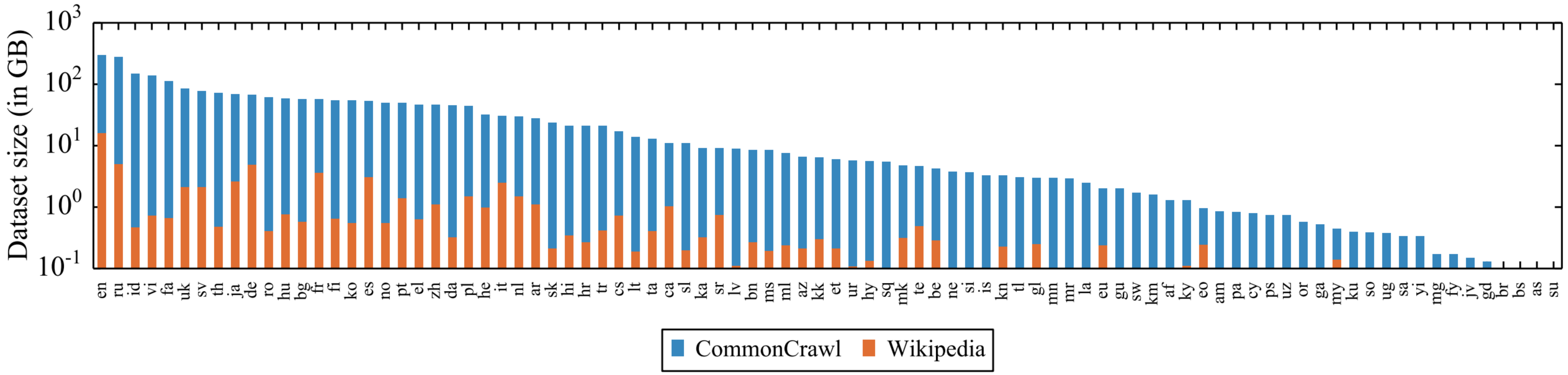


Figure 1: Amount of data in GiB (log-scale) for the 88 languages that appear in both the Wiki-100 corpus used for mBERT and XLM-100, and the CC-100 used for XLM-R. CC-100 increases the amount of data by several orders of magnitude, in particular for low-resource languages.

- 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)

mT5

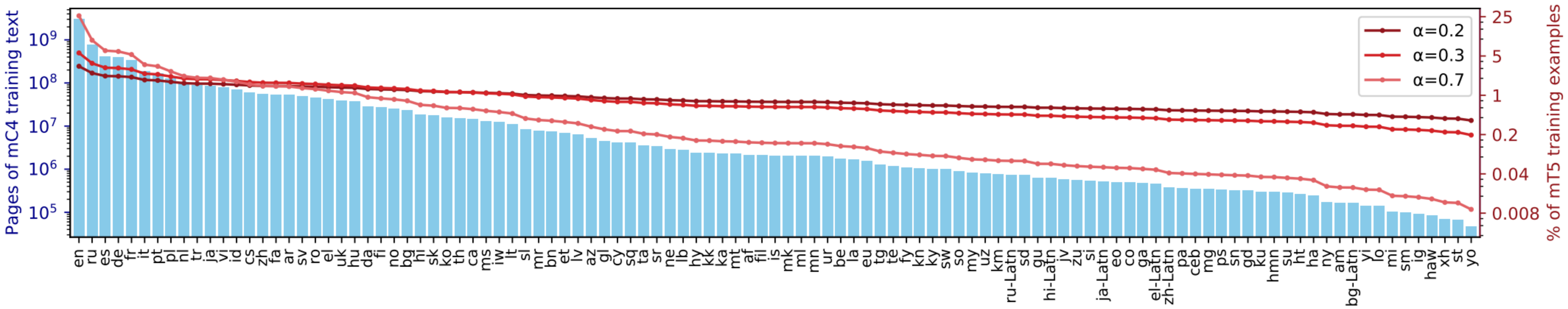


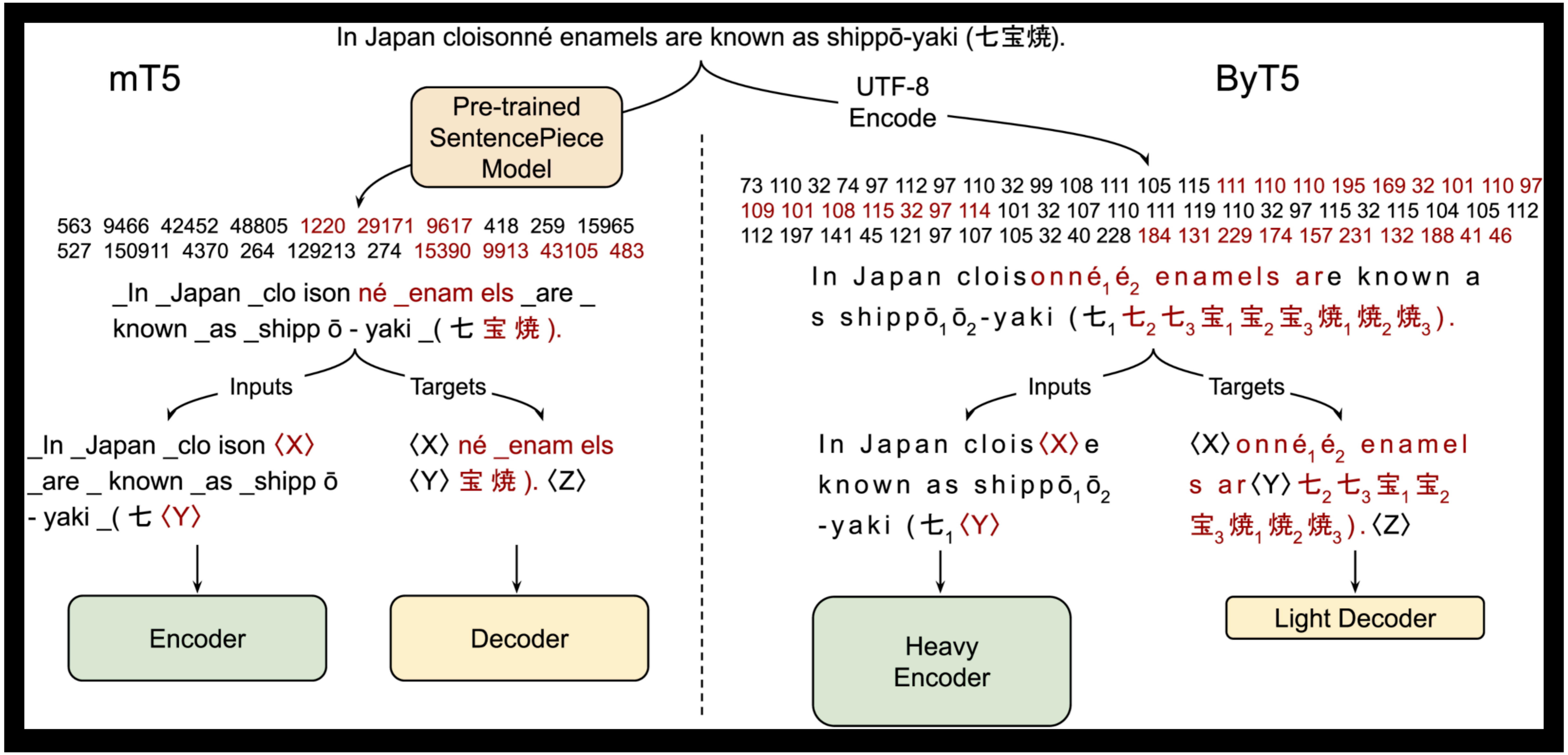
Figure 1: Page counts per language in mC4 (left axis), and percentage of mT5 training examples coming from each language, for different language sampling exponents α (right axis). Our final model uses α=0.3.

Model	Architecture	Parameters	# languages	Data source
mBERT (Devlin, 2018)	Encoder-only	180M	104	Wikipedia
XLM (Conneau and Lample, 2019)	Encoder-only	570M	100	Wikipedia
XLM-R (Conneau et al., 2020)	Encoder-only	270M – 550M	100	Common Crawl (CCNet)
mBART (Lewis et al., 2020b)	Encoder-decoder	680M	25	Common Crawl (CC25)
MARGE (Lewis et al., 2020a)	Encoder-decoder	960M	26	Wikipedia or CC-News
mT5 (ours)	Encoder-decoder	300M – 13B	101	Common Crawl (mC4)

Table 1: Comparison of mT5 to existing massively multilingual pre-trained language models. Multiple versions of XLM and mBERT exist; we refer here to the ones that cover the most languages. Note that XLM-R counts five Romanized variants as separate languages, while we ignore six Romanized variants in the mT5 language count.

Xue et al. (2021)

Scaling Up: mT5 vs. ByT5



Pre-training example creation and network architecture of mT5 (Xue et al., 2021) vs. ByT5 (this work). **mT5**: Text is split into SentencePiece tokens, spans of ~3 tokens are masked (red), and the encoder/decoder transformer stacks have equal depth. **ByT5**: Text is processed as UTF-8 bytes, spans of ~20 bytes are masked, and the encoder is 3 × deeper than the decoder. ⟨X⟩ , ⟨Y⟩ , and ⟨Z⟩ represent sentinel tokens.

Xue et al. (2022)

mBART

- Multilingual extension of BART, a seq2seq denoising auto-encoder pre-trained on large-scale monolingual corpora in many languages.

- Works well for machine translation! (especially Chinese in comparison to M2M and NLLB — based on my student's experiments)

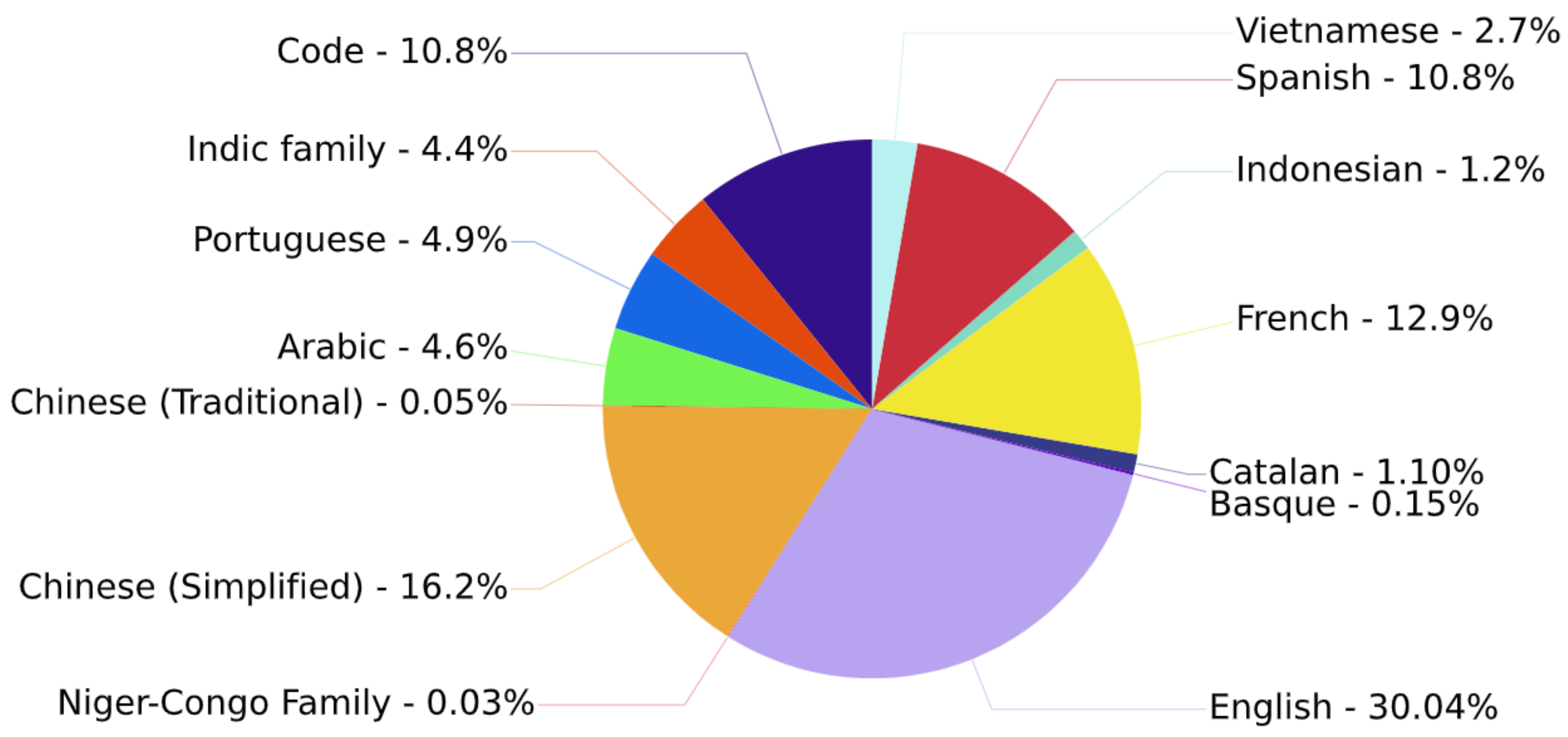
Code	Language	Tokens/M	Size/GB
En	English	55608	300.8
Ru	Russian	23408	278.0
Vi	Vietnamese	24757	137.3
Ja	Japanese	530 (*)	69.3
De	German	10297	66.6
Ro	Romanian	10354	61.4
Fr	French	9780	56.8
Fi	Finnish	6730	54.3
Ko	Korean	5644	54.2
Es	Spanish	9374	53.3
Zh	Chinese (Sim)	259 (*)	46.9
It	Italian	4983	30.2
Nl	Dutch	5025	29.3
Ar	Arabic	2869	28.0
Tr	Turkish	2736	20.9
Hi	Hindi	1715	20.2
Cs	Czech	2498	16.3
Lt	Lithuanian	1835	13.7
Lv	Latvian	1198	8.8
Kk	Kazakh	476	6.4
Et	Estonian	843	6.1
Ne	Nepali	237	3.8
Si	Sinhala	243	3.6
Gu	Gujarati	140	1.9
My	Burmese	56	1.6

Table 1: **Languages and Statistics of the CC25 Corpus.** A list of 25 languages ranked with monolingual corpus size. Throughout this paper, we replace the language names with their ISO codes for simplicity. (*) Chinese and Japanese corpus are not segmented, so the tokens counts here are sentences counts

Liu et al. (2020)

Bloom

- A BigScience initiative, open-access, 176B parameter (GPT-2 architecture)
- 59 languages (46 natural language + 13 programming language)
- 1.6TB of pre-processed text



https://huggingface.co/bigscience/bloom

XGLM

- Meta AI, 7.5B parameter (similar to GPT-3 architecture)
- 30 languages, 500B tokens of pre-processed text (CC100-XL of 68 Common Crawl snapshots (from Summer 2013 to March/April 2020))

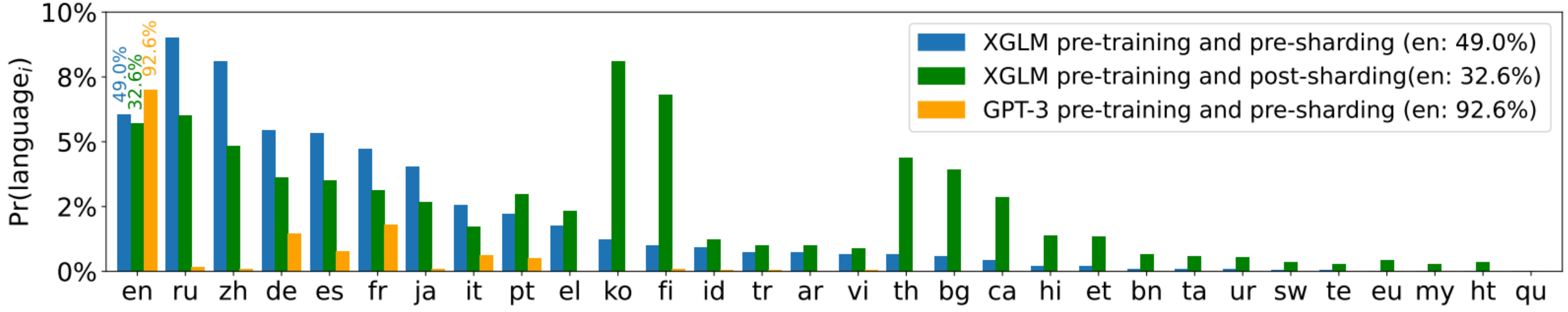
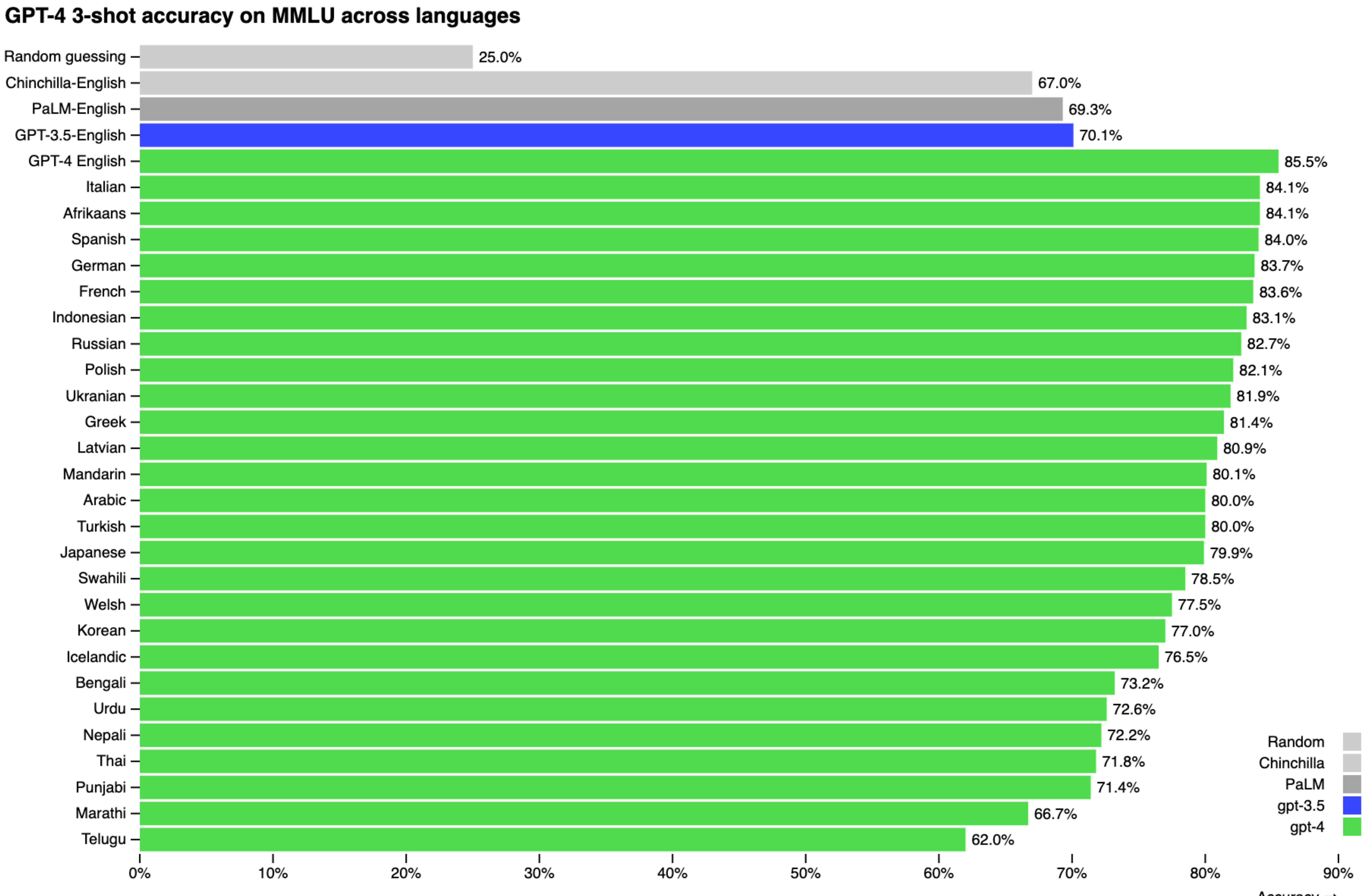


Figure 1: The % of each language l ($l = 1, 2, ..., 30$) in XGLM's pre-training data pre-upsampling (blue), post-upsampling (green), and its corresponding % in GPT-3's training data (orange). We truncate the y-axis at 10% to better visualize the tail distribution.

Lin et al. (2022)

GPT-4

- Tested on 26 languages, MMLU - Multiple-choice questions in 57 subjects



Prompt GPT-4 vs. Fine-tune mBERT

GPT-4 works pretty well on African languages, but fine-tuning a smaller BERT model (e.g., mDeBERTa-v3 of 276M parameters) with **label projection** methods can work even better.

Lang.	GPT-4†	FT_{En}	Translate-train			Translate-test	
			Awes-align	EasyProject	CODEC (Δ_{FT})	Awes-align	CODEC (Δ_{FT})
Bambara	46.8	37.1	45.0	45.8	45.8 (+8.7)	50.0	**55.6** (+18.5)
Ewe	75.5	75.3	78.3	78.5	**79.1** (+3.8)	72.5	**79.1** (+3.8)
Fon	19.4	49.6	59.3	61.4	**65.5** (+15.9)	62.8	61.4 (+11.8)
Hausa	70.7	71.7	72.7	72.2	72.4 (+0.7)	70.0	**73.7** (+2.0)
Igbo	51.7	59.3	63.5	65.6	70.9 (+11.6)	**77.2**	72.8 (+13.5)
Kinyarwanda	59.1	66.4	63.2	71.0	71.2 (+4.8)	64.9	**78.0** (+11.6)
Luganda	73.7	75.3	77.7	76.7	77.2 (+1.9)	**82.4**	82.3 (+7.0)
Luo	**55.2**	35.8	46.5	50.2	49.6 (+13.8)	52.6	52.9 (+17.1)
Mossi	44.2	45.0	52.2	53.1	**55.6** (+10.6)	48.4	50.4 (+5.4)
Chichewa	75.8	**79.5**	75.1	75.3	76.8 (-2.7)	78.0	76.8 (-2.7)
chiShona	66.8	35.2	69.5	55.9	72.4 (+37.2)	67.0	**78.4** (+43.2)
Kiswahili	82.6	**87.7**	82.4	83.6	83.1 (-4.6)	80.2	81.5 (-6.2)
Setswana	62.0	64.8	73.8	74.0	74.7 (+9.9)	**81.4**	80.3 (+15.5)
Akan/Twi	52.9	50.1	62.7	65.3	64.6 (+14.5)	72.6	**73.5** (+23.4)
Wolof	62.6	44.2	54.5	58.9	63.1 (+18.9)	58.1	**67.2** (+23.0)
isiXhosa	69.5	24.0	61.7	**71.1**	70.4 (+46.4)	52.7	69.2 (+45.2)
Yoruba	**58.2**	36.0	38.1	36.8	41.4 (+5.4)	49.1	58.0 (+22.0)
isiZulu	60.2	43.9	68.9	73.0	**74.8** (+30.9)	64.1	**76.9** (+33.0)
AVG	60.4	54.5	63.6	64.9	67.1 (+12.7)	65.8	**70.4** (+16.0)

[0] Awes-align - “Word Alignment by Fine-Tuning Embeddings on Parallel Corpora” Zi-Yi Dou, Graham Neubig (EACL 2021)
[1] EasyProject - “Frustratingly Easy Label Projection for Cross-lingual Transfer” Yang Chen, Chao Jiang, Alan Ritter, Wei Xu (ACL 2023 Findings)
[2] CODEC - “Constrained Decoding for Cross-lingual Label Projection” Duong Minh Le, Yang Chen, Alan Ritter, Wei Xu (ICLR 2024)

Le et al. (2024)

Language Contamination

- Models trained only on English text have been found to transfer surprisingly well to other languages.
- Common English pretraining corpora actually contain significant amounts of non-English text.

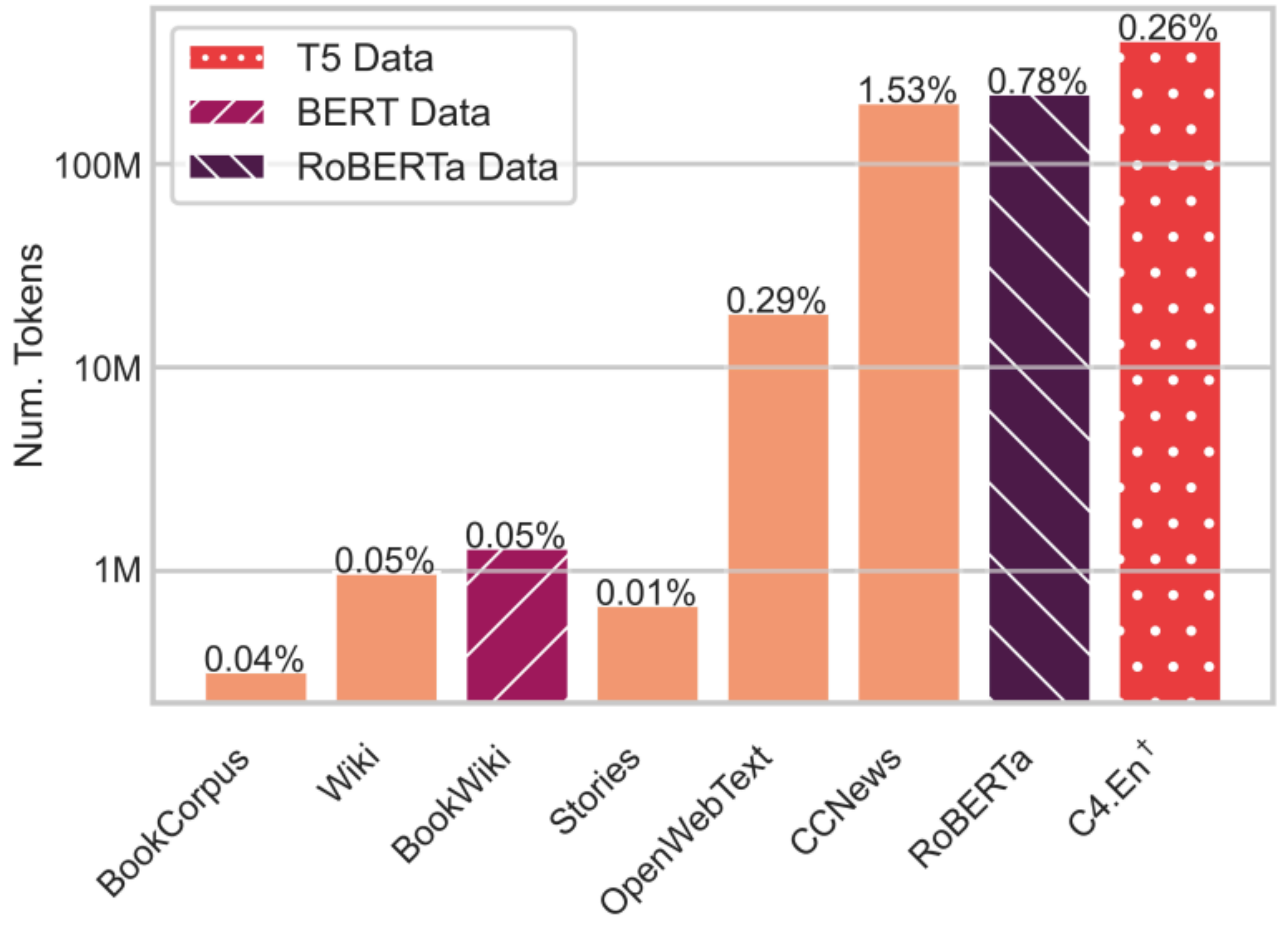


Figure 1: Estimated non-English data in English pre-training corpora (token count and total percentage); even small percentages lead to many tokens. C4.En (†) is estimated from the first 50M examples in the corpus.

Blevins & Zettlemoyer (2022)

Multilingual Datasets ++

Multilingual 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

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

Hu et al. (2020)

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

Language	Train (1-way)	Dev (3-way)	Test (3-way)
(English)	9,211	1031	1046
Arabic	23,092	1380	1421
Bengali	10,768	328	334
Finnish	15,285	2082	2065
Indonesian	14,952	1805	1809
Japanese	16,288	1709	1706
Kiswahili	17,613	2288	2278
Korean	10,981	1698	1722
Russian	12,803	1625	1637
Telugu	24,558	2479	2530
Thai	11,365	2245	2203
TOTAL	**166,916**	**18,670**	**18,751**

Clark et al. (2021)

TyDiQA

- Why not translate?

what topics will be discussed. For example, in TyDi QA, one Bengali question asks *What does sapodilla taste like?*, referring to a fruit that is unlikely to be mentioned in an English corpus, presenting unique challenges for transfer learning. Each of these issues makes a translated corpus more English-like, potentially inflating the apparent gains of transfer-learning approaches.

Clark 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
how far Uranus-SG.NOM from
Земл-и?
Earth-SG.GEN?

How far is Uranus from Earth?

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

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

Figure 3: Russian example of morphological variation across question-answer pairs due to the difference in syntactic context: the entities are identical but have different representation, making simple string matching more difficult. The names of the planets are in the subject (Уран, Uranus-NOM) and object of the preposition (от земли, from Earth-GEN) context in the question. The relevant passage with the answer has the names of the planets in a coordinating phrase that is an object of a preposition (между Ураном и Землёй, between Uranus-INSTR and Earth-INSTR). Because the syntactic contexts are different, the names of the planets have different case marking.

Clark et al. (2021)

Stanceosaurus

Claim: *“Raid at Tirupati temple priest’s house, 128 kg gold found”*

Keyword search: *“Tirupati”, “gold”*

- Lexically similar but irrelevant tweets are included to train robust stance classifiers;
- Finer-grained stance identification (i.e., lean supporting/refuting) to handle implicitness;
- Tweet threads are collected as misinformation often spreads through multi-turn conversations.

Zheng et al. (2022)

Stanceosaurus

More diverse (multicultural & multilingual) with 28k tweets towards 250 claims in English, Arabic, Hindi.

Zheng et al. (2022)

Stanceosaurus - Annotation Schema

Stance Categories

Supporting | Refuting | Discussing | Querying | Irrelevant

Implicit Leanings

Discussing-Supporting | Discussing-Refuting | Discussing-Neutral

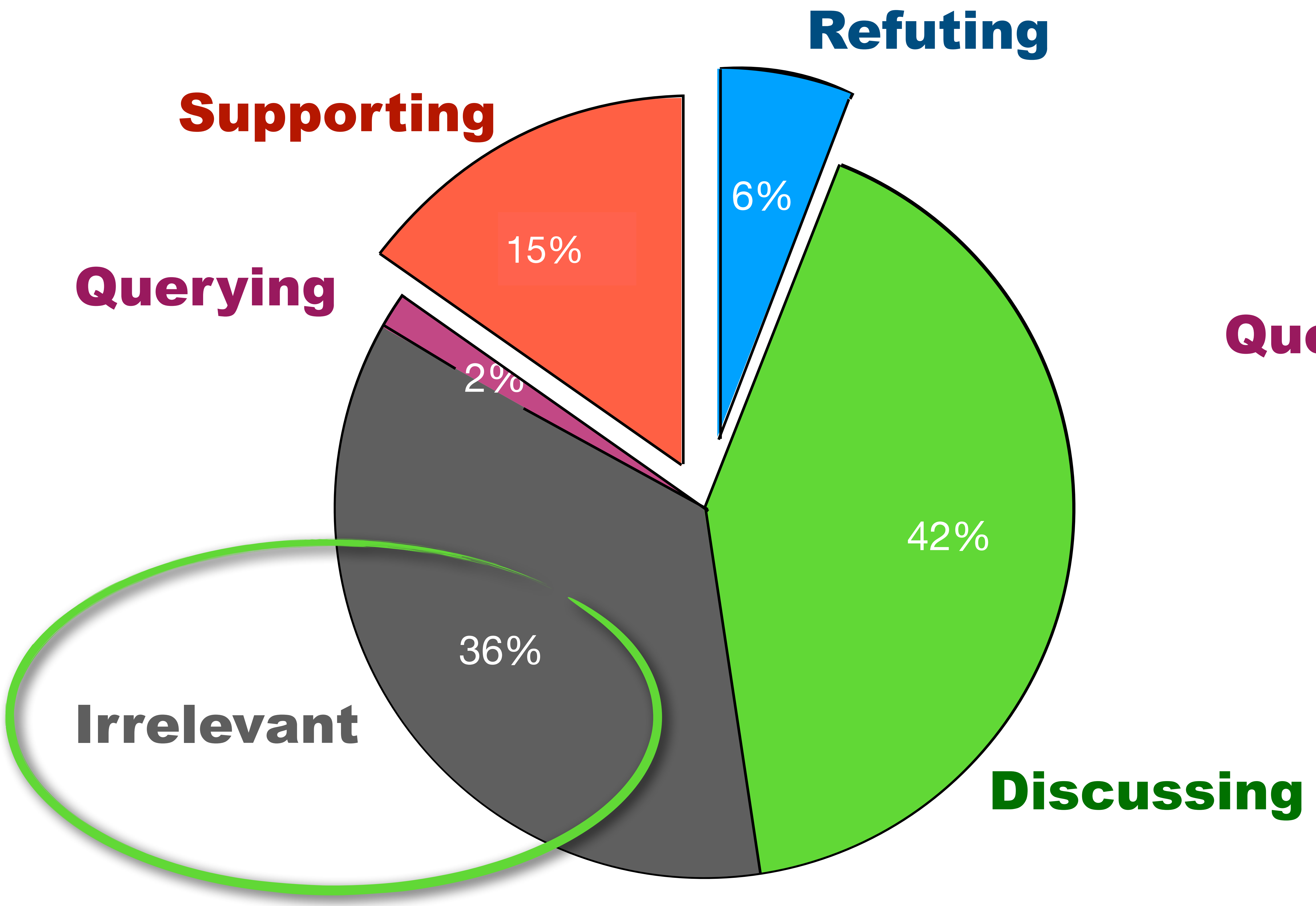
Stance categories include 4 relevant classes seen in previous work

Topically similar but irrelevant tweets are included to help build robust models

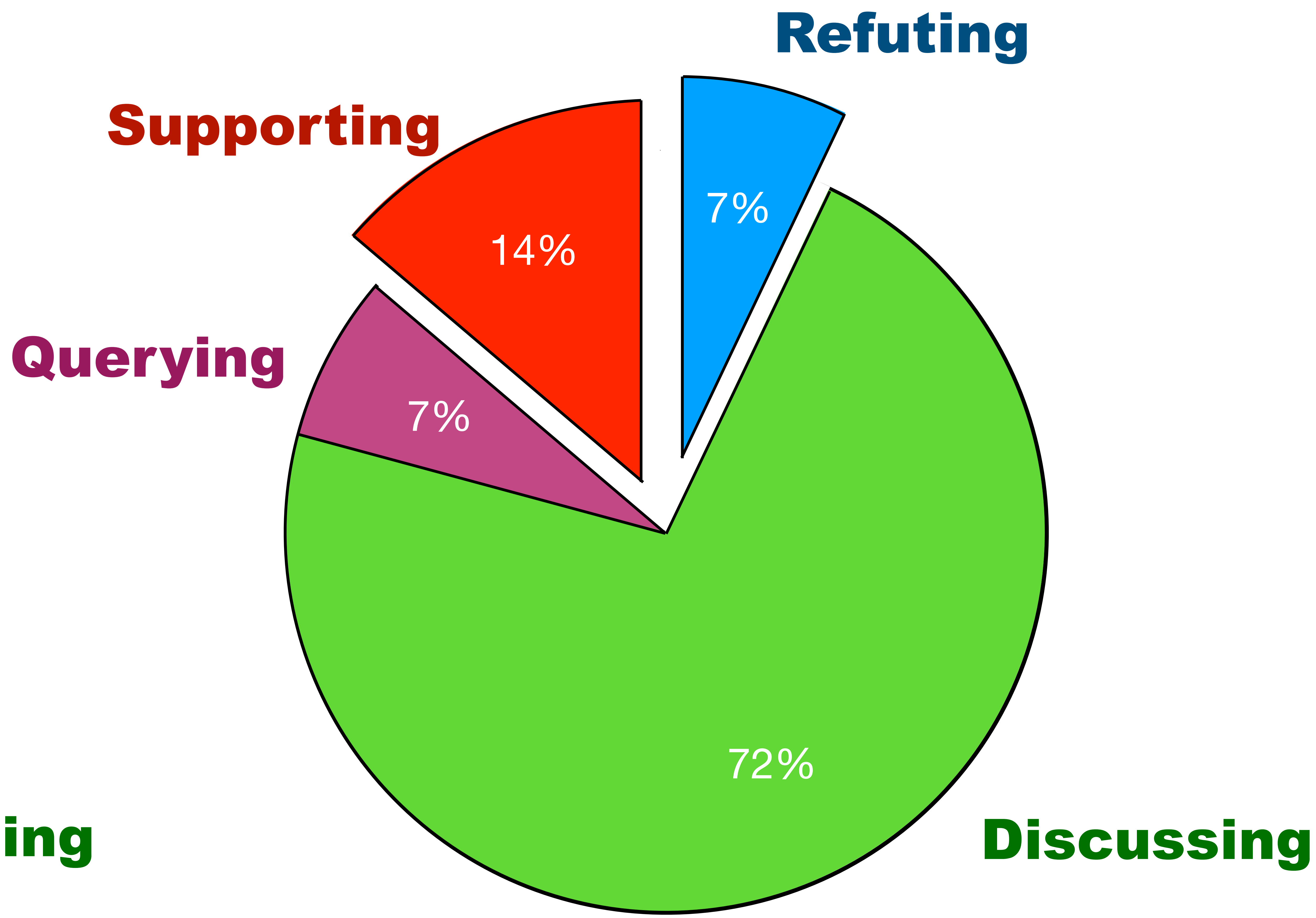
Neutral tweets may have an indirect bias

Stance Distributions

Supporting and Refuting make up only 21~22% of datapoints in Stanceosaurus and RumourEval.



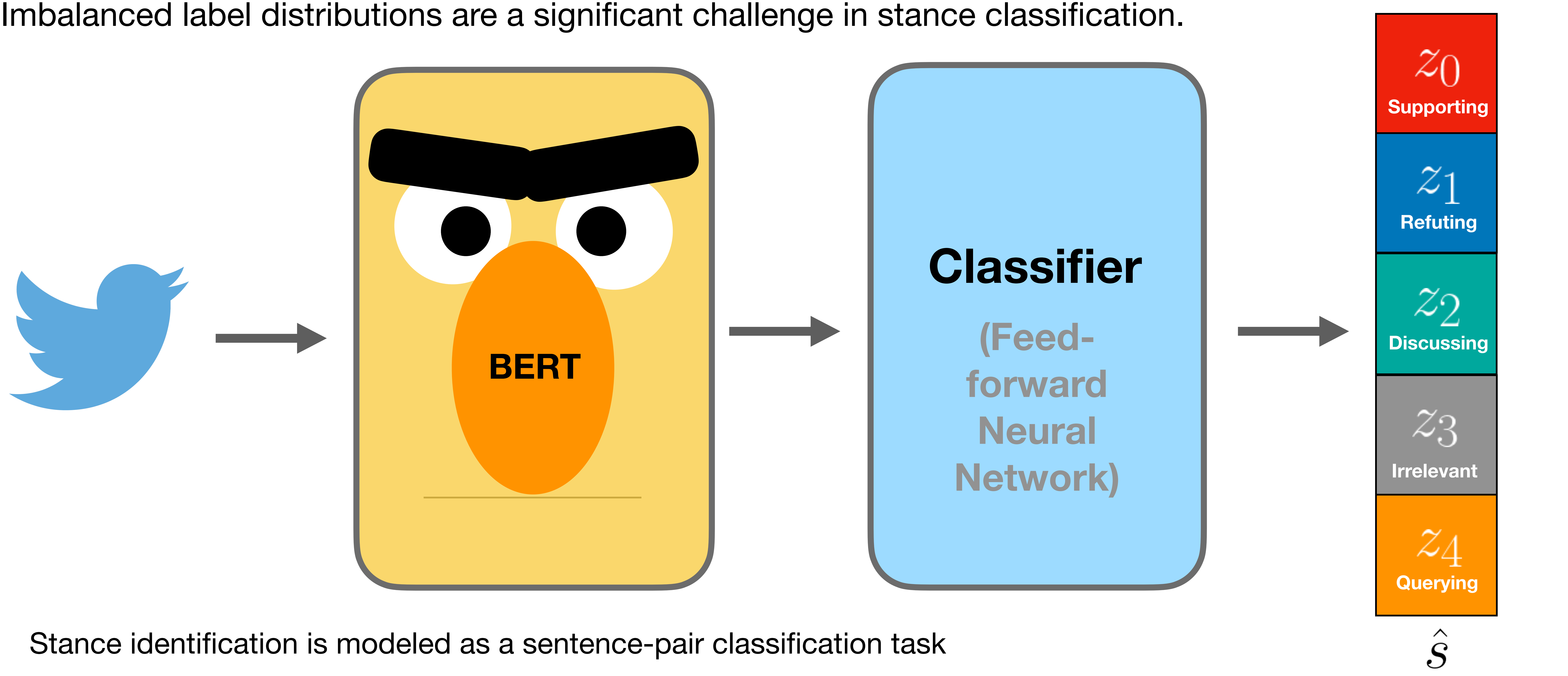
Stanceosaurus (this work)
Inclusion of similar but irrelevant datapoints



RumourEval
(Gorrell et al., 2018)

Automatic Stance Classification

Imbalanced label distributions are a significant challenge in stance classification.



Stance identification is modeled as a sentence-pair classification task

Tweets are encoded into contextualized word embeddings using transformer-based models

Baseline models are trained using cross-entropy loss

Class-Balanced Focal Loss

To address the imbalanced label distributions in stance classification:

Unnormalized scores

Re-weighting

$$\mathrm{CB}_{\mathrm{foc}}(\boxed{\hat{s}}, y) = -\boxed{\frac{1-\beta}{1-\beta^{n_y}}} \sum_{m \in C} (1-p_m)^{\gamma} \log(p_m)$$

$\hat{s} = (z_0, z_1, ..., z_{|C|-1})$

$y = \text{correct label}$

$n_y = \text{number of examples of label y}$

$C = \text{stance categories}$

β and γ hyperparameters

- Hyper-parameter β is tuned between [0.1, 1)
- Re-weighting term is inversely proportional with the number of data points in a class
- Weighs misclassifications of smaller classes more than larger classes
- During evaluation, the proportion of classes is estimated with the training set

* Class-Balanced Focal Loss (Lin et al. 2017) has been demonstrated to address imbalanced computer vision problems

Class-Balanced Focal Loss

$$p_m = (1 + e^{-z'_m})^{-1}$$

$$z'_m = \begin{cases} z_m & m = y \\ -z_m & \text{otherwise} \end{cases}$$

To address the imbalanced label distributions in stance classification:

Unnormalized scores

Re-weighting

Focal loss

$$\mathrm{CB}_{\mathrm{foc}}(\boxed{\hat{s}}, y) = -\boxed{\frac{1-\beta}{1-\beta^{n_y}}}\boxed{\sum_{m \in C} (1-p_m)^{\gamma} \log(p_m)}$$

$\hat{s} = (z_0, z_1, ..., z_{|C|-1})$

$y = \text{correct label}$

$n_y = \text{number of examples of label y}$

$C = \text{stance categories}$

β and γ hyperparameters

- The hyper-parameter γ is tuned between [0.1 1.1]
- Well-classified examples with high confidence reduce the focal loss term towards 0
- High confidence for incorrect classifications creates a larger focal loss
- Low confidence for correct classifications also generate a large focal loss

* Class-Balanced Focal Loss (Lin et al. 2017) has been demonstrated to address imbalanced computer vision problems

Pre-trained Models

Based on existing work, pre-trained Transformer models are used to establish baseline performance

English data

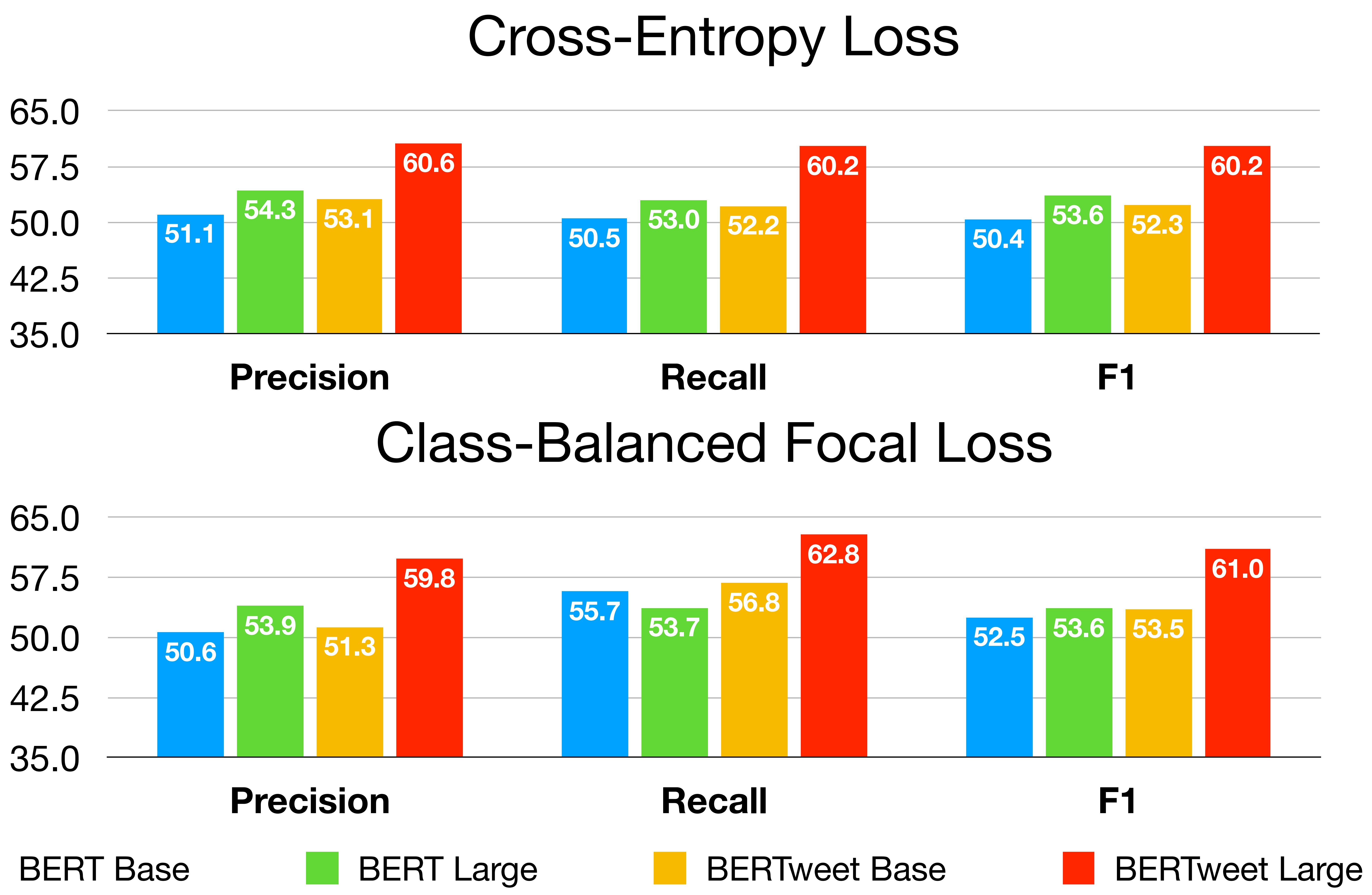
- BERT - Bidirectional Encoder Representations from Transformers
- BERTweet (Nguyen et al., 2020) - Pre-trained on a large Twitter corpus

Hindi and Arabic data

- Multilingual-BERT - Trained on a corpus of 104 different languages
- XLM-RoBERTa - Trained on CommonCrawl of 100 languages

Evaluation on English Dataset

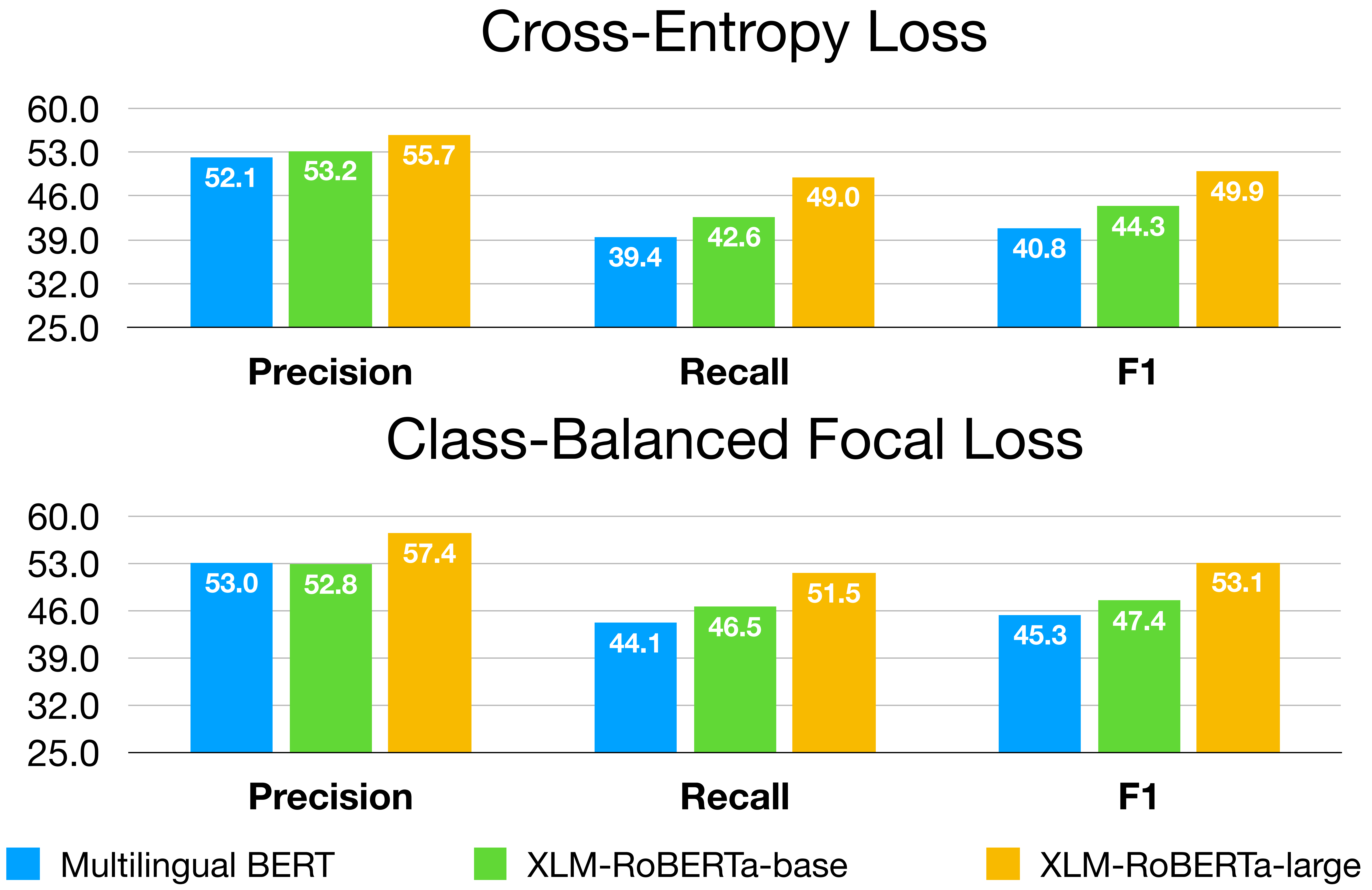
20707 English tweets for 190 total claims; 112 train / 34 dev / 44 test English claims; evaluate on individual tweet text.



BERTweet-large trained on Class-balanced Focal Loss outperforms all other models.

Evaluation on Hindi tweets

Cross-lingual transfer models classify stance trained on English data and evaluated on Hindi tweets.



XLM-RoBERTa models outperform Multilingual-BERT in cross-lingual transfer for Hindi tweets

Evaluation on Arabic tweets

Cross-lingual transfer models classify stance trained on English data and evaluated on Arabic tweets.

XLM-RoBERTa models outperform Multilingual-BERT in cross-lingual transfer for Arabic tweets

African Languages!

- AfroLID, a neural LID toolkit for 517 African languages and varieties.

Figure 1: All 50 African countries in our data, with our 517 languages/language varieties in colored circles overlayed within respective countries. More details are in Appendix E.

Word Order	Example Languages
SVO	Xhosa, Zulu, Yorùbá
SOV	Khoekhoe, Somali, Amharic
VSO	Murle, Kalenjin
VOS	Malagasy
No-dominant-order	Siswati, Nyamwezi, Bassa

Table 1: Sentential word order in our data.

Adebara et al. (2022)

Masakhane NER

- NER datasets for 20 African languages

Language	Sentence
English	The Emir of Kano turbaned Zhang who has spent 18 years in Nigeria
Amharic	የካኖ ኢምር በናይጀርያ ፩፰ ዓመት ያሳለፈውን ዣንግን ዋና መሪ አደረጉት
Hausa	Sarkin Kano yayi wa Zhang wanda yayi shekara 18 a Nigeria sarauta
Igbo	Onye Emir nke Kano kpubere Zhang okpu onye nke nọgoro afọ iri na asatọ na Naịjirịa
Kinyarwanda	Emir w'i Kano yimitse Zhang wari umaze imyaka 18 muri Nijeriya
Luganda	Emir w'e Kano yatikkidde Zhang amaze emyaka 18 mu Nigeria
Luo	Emir mar Kano ne orwakone turban Zhang ma osedak Nigeria kwuom higni 18
Nigerian-Pidgin	Emir of Kano turban Zhang wey don spend 18 years for Nigeria
Swahili	Emir wa Kano alimvisha kilemba Zhang ambaye alikaa miaka 18 nchini Nigeria
Wolof	Emiiru Kanó dafa kaala kii di Zhang mii def Nigeria fukki at ak juróom ñett
Yorùbá	Ẹ́mía ìlú Kánò wé láwàní lé orí Zhang ẹni tí ó ti lo ọdún méjìdínlógún ní orílè-èdè Nàìjíríà

Table 2: Example of named entities in different languages. PER, LOC, and DATE are in colours purple, orange, and green respectively. The original sentence is from BBC Pidgin.[3]

Adelani et al. (2021)

Where are we now?

- Universal dependencies: treebanks (+ tags) for 100+ 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 — but still many challenges for low-resource settings

Takeaways

- Many languages have richer morphology than English and pose distinct challenges
- Problems: how to analyze rich morphology, how to generate with it
- Can leverage resources for English using bitexts
- Multilingual models can be learned in a bitext-free way and can transfer between languages