Low resource machine translation and NLP - new advances

Alexandra Birch

Global Under-Resourced MEdia Translation

2019-2022

THE UNIVERSITY OF EDINBURGH

gourmet

Machine Learning for NLP

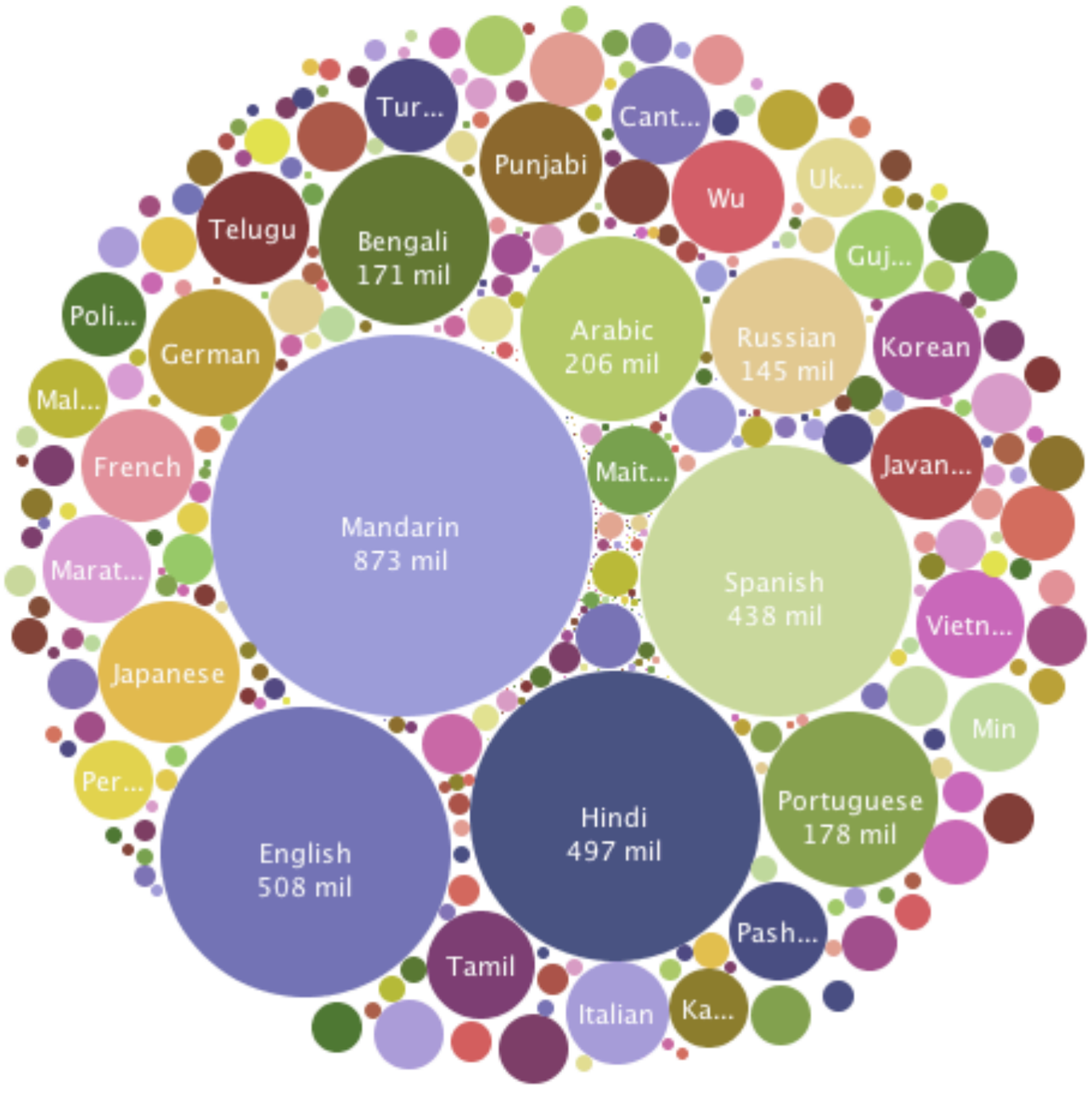
	<Input>	<Output>
Translation	**How are you?**	**Bawo ni? (Yoruba)**
Dialogue	**Can I book a flight?**	**What is your destination?**
Question Answering	**How old is Trump?**	**74 years**

Supervised learning: lots of labelled data <Input, Output>

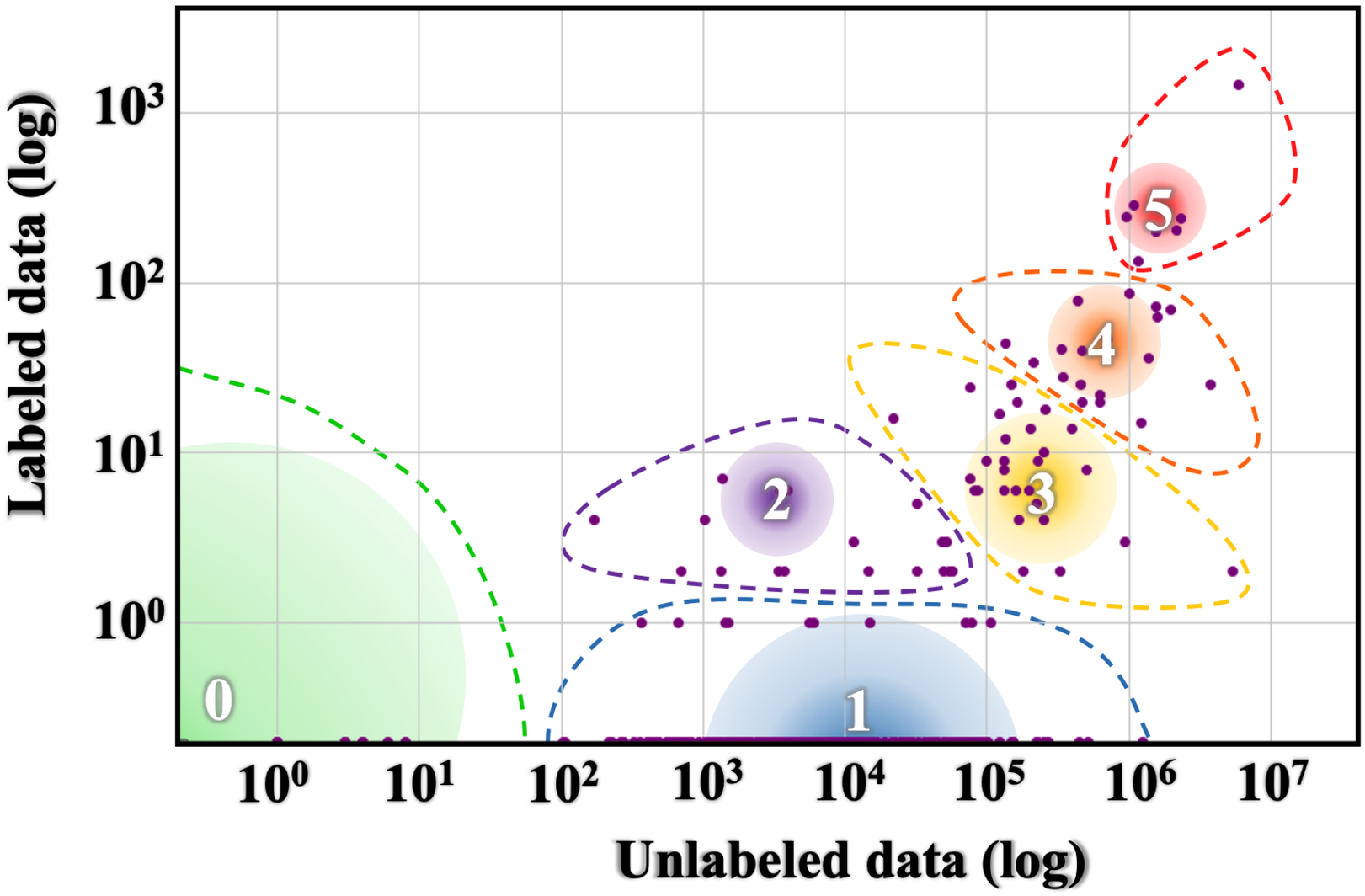
Unsupervised learning: lots of unlabelled data <Input> <Output>

gourmet

Diversity of Languages



What is low-resource?



The State and Fate of Linguistic Diversity and Inclusion in the NLP World Joshi et al. 2020

gourmet

What is low-resource?

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

Most languages are low-resource
Almost all language pairs are low-resource

gourmet

Low Resource MT

- Creating More Data
- Monolingual Data
- Multilingual Data
- Model Centric Techniques
- Research Community

gourmet

Low Resource MT

- **Creating More Data**
- Monolingual Data
- Multilingual Data
- Model Centric Techniques
- Research Community

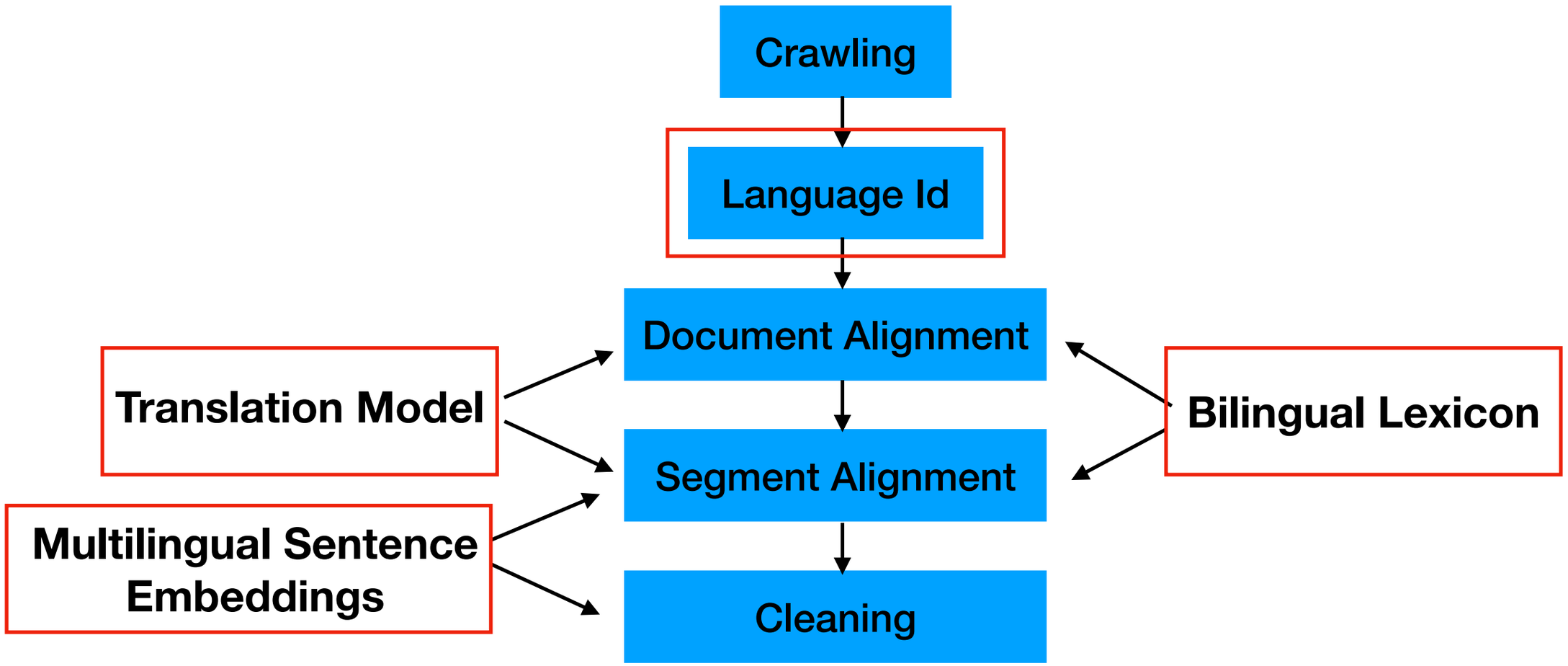
gourmet

Data

- OPUS > 500 languages Tiedemann et al. 2012
    - Bible, GNOME
- Paracrawl: large scale crawling, and internet archive Bañón et al. 2020
- WikiMatrix: 85 lang, using sentence embeddings Schwenk et al. 2019
- JW300: 54k lang pairs Agic and Vulic 2019

gourmet

Data - Crawling



Data - Crawling

		Parallel		
		CCAligned	**ParaCrawl v7.1**	**WikiMatrix**
	#langs audited / total	65 / 119	21 / 38	20 / 78
	%langs audited	54.62%	55.26%	25.64%
	#sents audited / total	8037 / 907M	2214 / 521M	1997 / 95M
	%sents audited	0.00089%	0.00043%	0.00211%
macro	C	29.25%	76.14%	23.74%
	X	29.46%	19.17%	68.18%
	WL	9.44%	3.43%	6.08%
	NL	31.42%	1.13%	1.60%
	offensive	0.01%	0.00%	0.00%
	porn	5.30%	0.63%	0.00%

Caswell et al 2021

gourmet

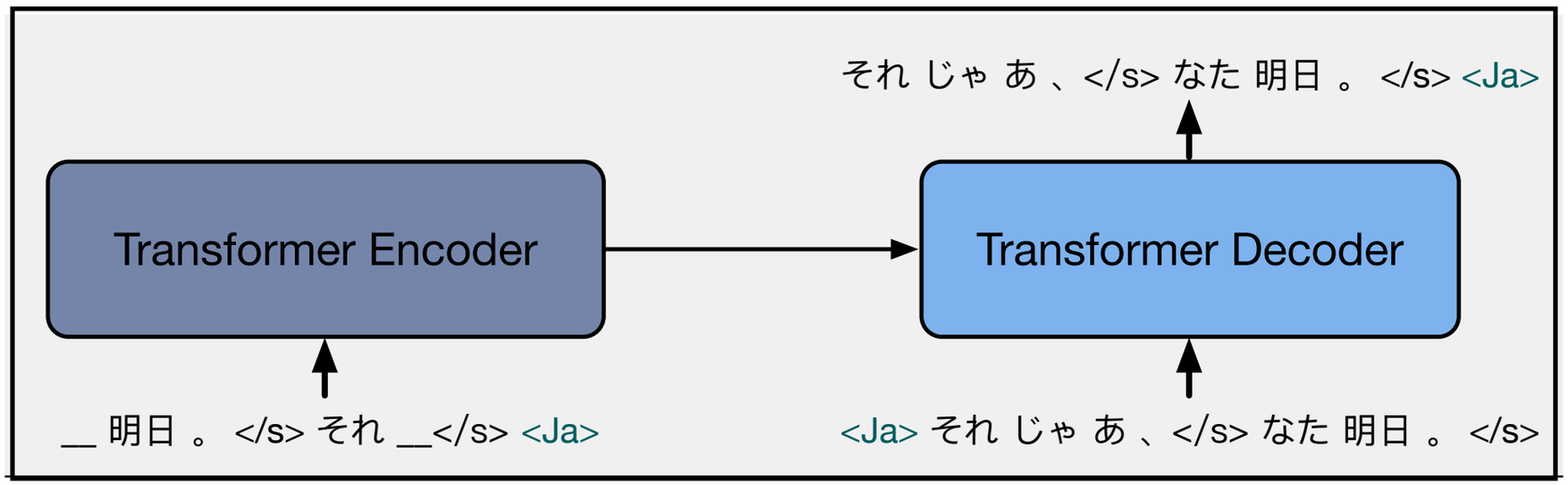
Low Resource MT

- Creating More Data
- **Monolingual Data**
- Multilingual Data
- Model Centric Techniques
- Research Community

gourmet

Monolingual Data - Pre-training

Use unlabelled <input> and/or <output> data, **pre-train** the model to predict the next or missing word

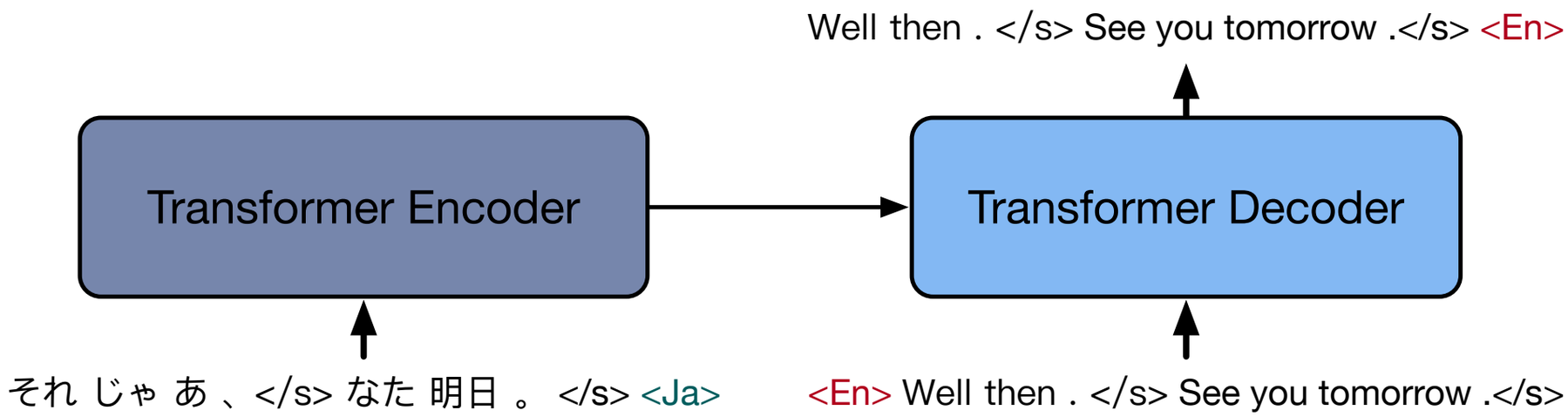


Multilingual denoising pre-training (mBART) Liu et al 2020

BERT, GPT2

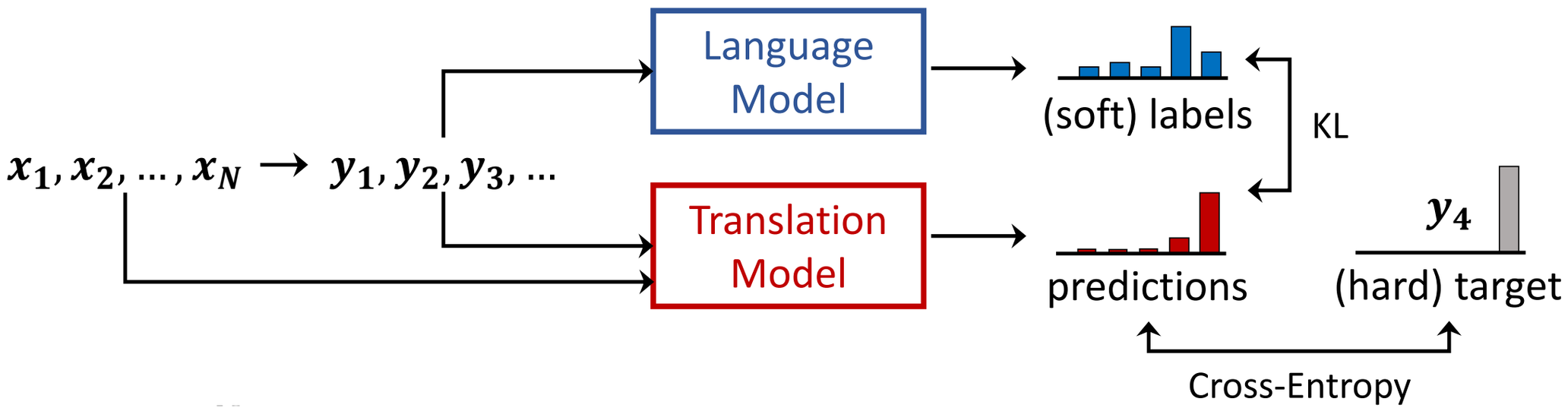
Monolingual Data - Pre-training

Use labelled <input, output> data, **fine-tune** the model to predict the translation



Monolingual Data - Pre-training

Baziiotis et al 2020



Monolingual Data - Synthetic

- back translation — Sennrich et al 2016
  - use English->Yoruba system to translate English
  - Train Yoruba->English system on <*Yoruba*, English>
- iterative translation - self learning — Hoang et al. 2018
  - Train Yoruba->English, English->Yoruba, English->Yoruba etc.
- Unsupervised MT — Lample et al. 2018

Monolingual Data - Synthetic

- Data augmentation using LMs

Arthaud et al 2021

=> s_1 = Cette centrale nucléaire menace d'exploser à tout moment !

This nuclear power plant could explode at any time!

=> masked context: This _____ power plant could explode at any time!

=> top matches = by context

A **coal** power plant produces carbon-intensive power.	That's a **solar**-powered ship.	They threatened to drop a **thermonuclear** bomb.
Sweden largely relies on **hydroelectric** power.	They have called for subsidies for **cleaner** electricity.	He's drinking **apple** juice. ···

=> s_2 = Une centrale charbon produit une électricité carbonée.

alignments

A coal power plant produces carbon-intensive power.

s_3 = Une centrale nucléaire produit une électricité carbonée.

A nuclear power plant produces carbon-intensive power.

gourmet

Low Resource MT

- Creating More Data
- Monolingual Data
- **Multilingual Data**
- Model Centric Techniques
- Research Community

gourmet

Multilingual Data

French -> English

↓

Hausa -> English

English
German
Mandarin
Arabic
French
Hausa

→

English
German
Mandarin
Arabic
French
Hausa

Transfer Learning

Zoph et al. 2016

Multilingual Models

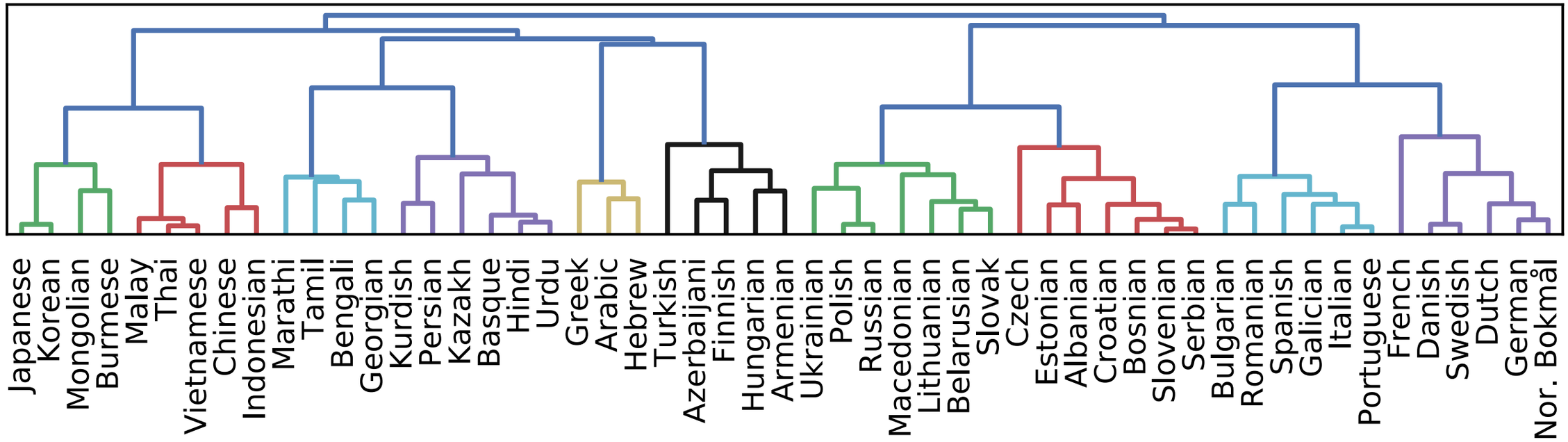
Johnson et al. 2016

gourmet

Multilingual Data

What languages to train together? Oncevay et al. 2020

Robust language representation:
World Atlas of Language Structure + language embedding

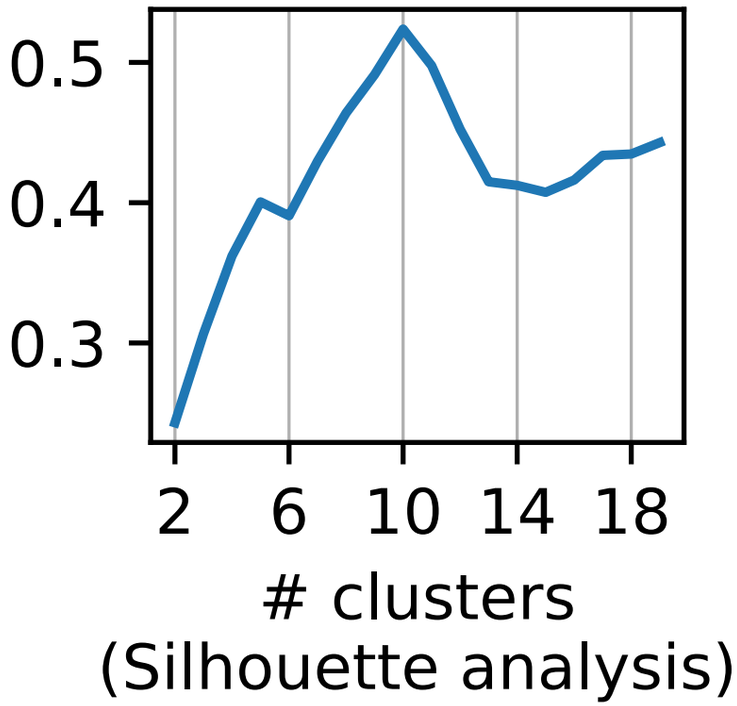


THE UNIVERSITY OF EDINBURGH

gourmet

Multilingual Data

How many languages to train together?



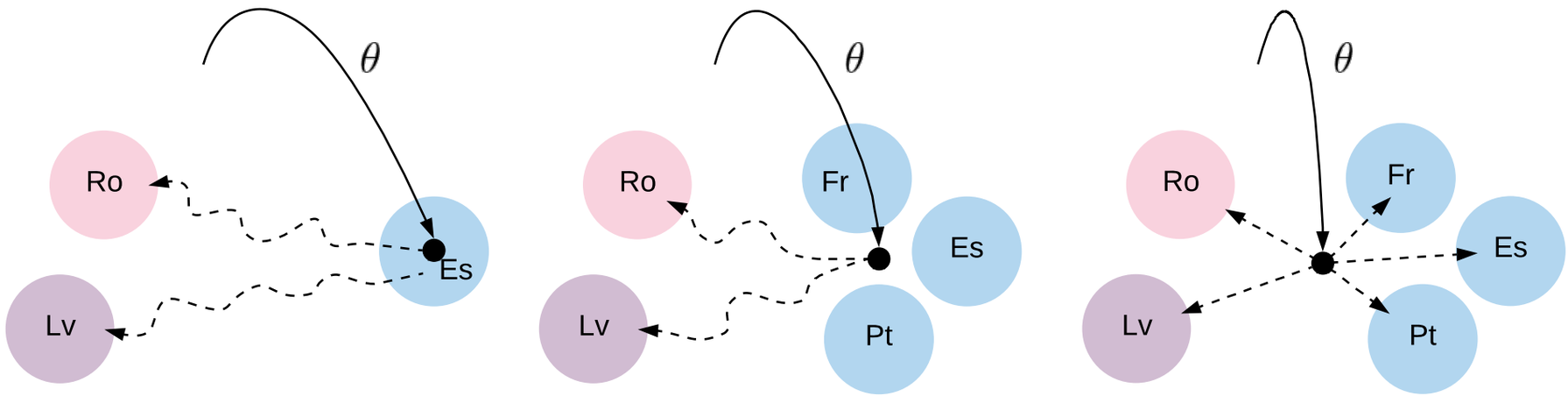
Low Resource MT

- Creating More Data
- Monolingual Data
- Multilingual Data
- **Model Centric Techniques**
- Research Community

gourmet

Model Centric Techniques

- Model-agnostic meta-learning (MAML) for machine translation



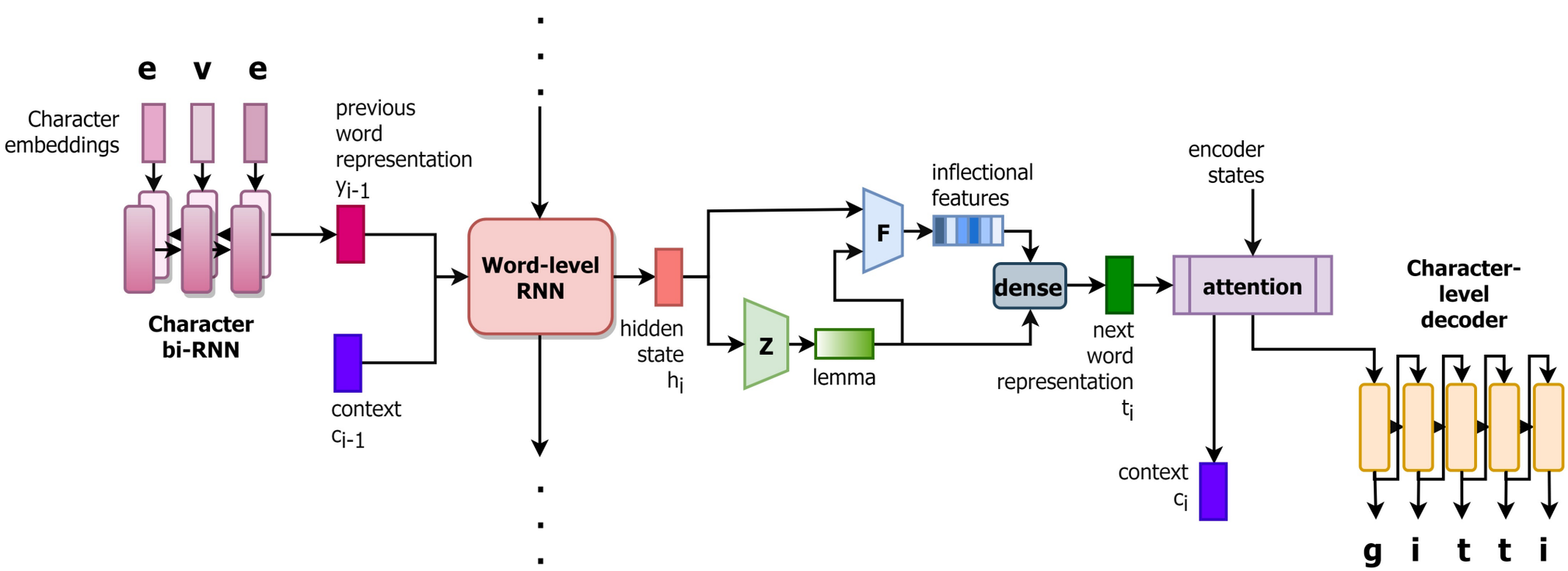
(a) Transfer Learning (b) Multilingual Transfer Learning (c) Meta Learning

Gu et al. 2018

Model Centric Techniques

- Latent Variable Models

Ataman et al. 2020



Research Community

Conference on Machine Translation (WMT): news shared task

Finnish	2015-2018
Romanian	2016
Latvian	2017
Estonian	2018
Turkish	2016-2018
Kazakh	2019
Gujarathi	2019
Tamil	2020
Inuktitut	2020
Pashto	2020
Khmer	2020

gourmet

Research Community

- LoResMT
- Workshop for Asian Translation
- African NLP
- Masekane

  https://github.com/masakhane-io/masakhane-mt/blob/master/MT4LRL.md
- Gourmet!

gourmet

Summary

- Find, Clean, Create Data
- Use all available resources:
    - monolingual
    - multilingual
- Better learning
- Build community interest and capability!

Thank you!

Barry Haddow

Rachel Bawden

Antonio Valerio Miceli Barone

Jindřich Helcl

gourmet