Transfer Learning Grammar for Multilingual Surface Realisation

Atif Khurshid[1], Seemab Latif[1], Rabia Latif[2]

[1]National University of Sciences & Technology, Islamabad

[2]Prince Sultan University, Riyadh

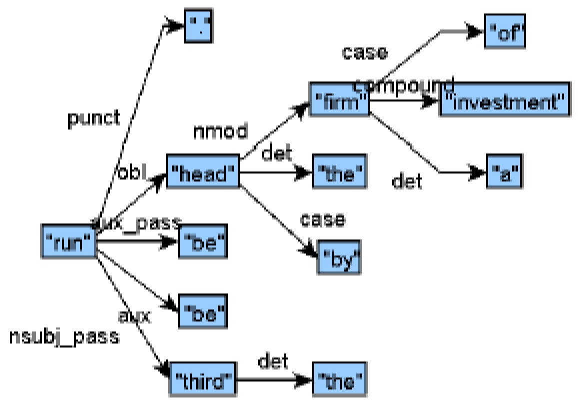
Surface Realisation

The final step in a traditional natural language generation pipeline

Formation of words into grammatically correct sentences by reordering and inflection

Source	dark be . room the grow
Target	The room was growing dark.

Requires specialized datasets unavailable for low-resource languages



Research Questions

1. Cross-lingual generalization

Can a model trained on multiple languages familiarize itself with their common features?

2. Performance

How does multilingual training affect the generated text for each language?

Universal Grammar

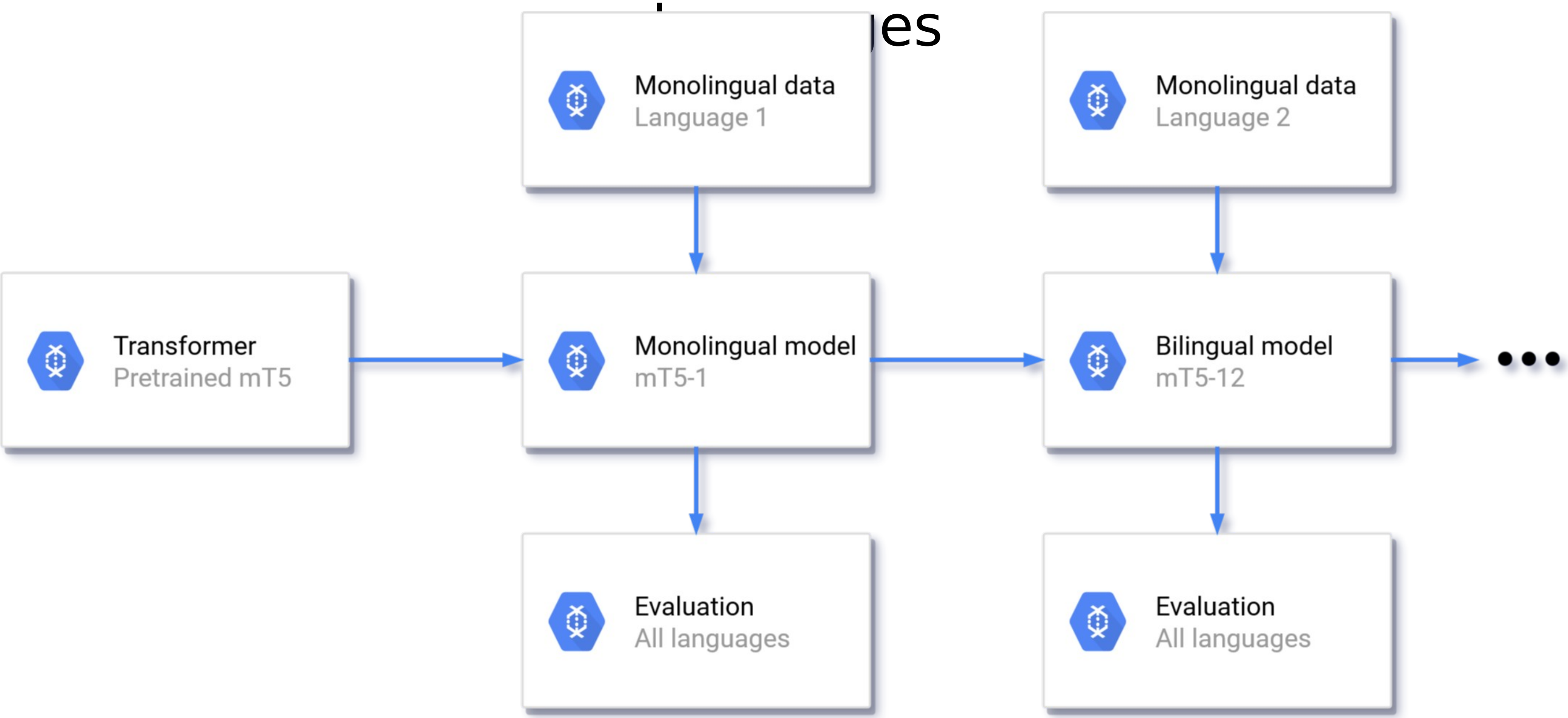
Languages share grammatical features innate to humans

Allows children to learn languages despite ‘poverty of stimulus’

> IN FACT, BY UNIVERSAL GRAMMAR I MEAN JUST THAT SYSTEM OF PRINCIPLES AND STRUCTURES THAT ARE THE PREREQUISITES FOR ACQUISITION OF LANGUAGE, AND TO WHICH EVERY LANGUAGE NECESSARILY CONFORMS.
>
> - NOAM CHOMSKY -

Methodology

Incrementally fine-tune an mT5 transformer model for multiple languages



Evaluation

- Cross-entropy as language modelling loss
  - Language and task comprehension during training
- Bilingual Evaluation Understudy (BLEU-4)
  - N-gram comparison of generated sentences with target sentences
- Inverse normalized character-based string-edit distance (DIST)
  - Edits required to convert generated sentences to target sentences

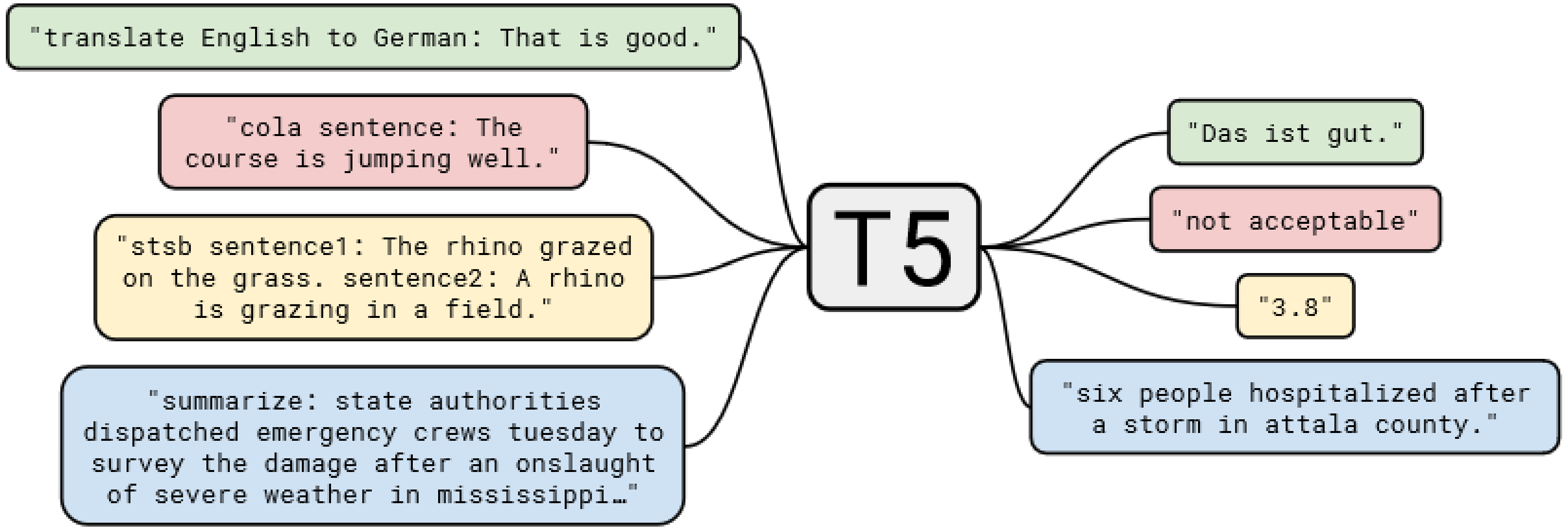
Experiments

- Multilingual Surface Realisation Shared Task Dataset
  - Annual task for development of multilingual realisation engines
  - Determination of word order and inflection
  - Universal Dependency structures designed by expert linguists

1	The	the	DET	DT	Definite=Def|PronType=Art	2	det
2	third	third	ADJ	JJ	Degree=Pos|NumType=Ord	5	nsubj_pass
3	was	be	AUX	VBD	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin	5	aux
4	being	be	AUX	VBG	VerbForm=Ger	5	aux_pass
5	run	run	VERB	VBN	Tense=Past|VerbForm=Part|Voice=Pass	0	root
6	by	by	ADP	IN	_	8	case
7	the	the	DET	DT	Definite=Def|PronType=Art	8	det
8	head	head	NOUN	NN	Number=Sing	5	obl
9	of	of	ADP	IN	_	12	case
10	an	a	DET	DT	Definite=Ind|PronType=Art	12	det
11	investment	investment	NOUN	NN	Number=Sing	12	compound
12	firm	firm	NOUN	NN	Number=Sing	8	nmod
13	.	.	PUNCT	.	_	5	punct

Experiments

- mT5: Multilingual Text-to-Text Transfer Transformer
  - Trained on Text-to-text denoising like T5
  - Multilingual dataset of 101 languages
  - mT5-small: 300M parameters



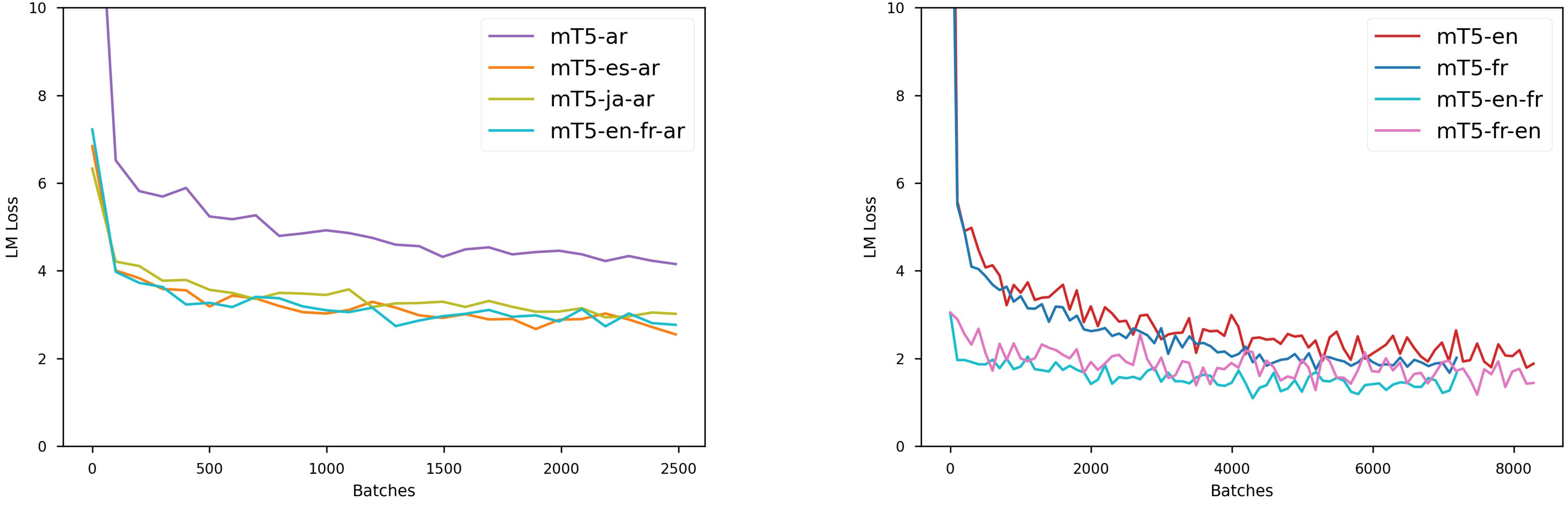
Experiments

- Data Processing
  - Extract source and target text from UD structures
- Model Processing
  - Train monolingual models and evaluate
    - Arabic, English, Spanish and French mode
  - Train multilingual models and evaluate
    - English-French, Spanish-Arabic etc.
  - Identical hyperparameters
    - Epochs: 5, Batch Size: 12, Adam (lr = 5e-5
    - Beam size: 5

Models	Dataset	Training Samples	Training Time
mT5-ar	Arabic	6075	25 min
mT5-en	English	19976	45 min
mT5-es	Spanish	28492	60 min
mT5-fr	French	17484	45 min
mT5-ja	Japanese	7133	30 min
mT5-en-fr	French	17484	75 min
mT5-es-ar	Arabic	6075	25 min
mT5-fr-en	English	19976	80 min
mT5-ja-ar	Arabic	6075	25 min
mT5-en-fr-ar	Arabic	6075	25 min

Results

Multilingual training results in reduced language modelling loss



Results

Multilingual training increases BLEU and DIST scores for similar languages

Model	English		French	
	BLEU	DIST	BLEU	DIST
mT5-en	7.40	34.40	1.56	28.91
mT5-en-fr	9.20	38.10	7.86	37.32
mT5-fr	3.30	27.20	3.22	30.31
mT5-fr-en	13.93	41.38	4.91	35.60

Results

Multilingual training decreases BLEU and DIST scores for different languages

Model	English		French		Spanish		Arabic	
	BLEU	**DIST**	**BLEU**	**DIST**	**BLEU**	**DIST**	**BLEU**	**DIST**
mT5-es	4.44	31.48	2.82	33.70	7.83	35.0	0.01	11.06
mT5-es-ar	2.26	31.70	0.87	29.83	2.39	30.90	0.39	24.69
mT5-en-fr	9.20	38.10	7.86	37.32	2.71	31.94	0.01	10.47
mT5-en-fr-ar	5.63	35.32	3.05	33.50	0.82	28.37	0.39	25.83

Results

The effects of multilingual training on English text generation

- Multilingual models generate better sentences
  - More training data

- Sentences contain some syntactic and semantic errors
  - Incomplete training

Input	observation make the of a on he some . few pic' good
Reference	He makes some good observations on a few of the pic's.
mT5-en	He made some pictures on a few pic's on a pic.
mT5-en-fr	They make a good observation on the pic' of a pic'.
mT5-fr-en	He made some good observations on the pic' of a good pic'.
mT5-en-fr-ar	They make a good observation on pic'.
mT5-fr	L' observation of a pic' on a pic'.

Conclusion

- Languages with similar grammar benefit from iterative training
  - Different languages disorient the model
- Overall performance is not comparable to actual realisation engines
  - Text-to-text approach
  - Model complexity
  - Models not trained to completion
- Better performance with more complex models, longer training
  - Trade-off between computational resources and human effort

Questions