Exploring Enhanced Code-Switched Noising for Pretraining in Neural Machine Translation

Vivek Iyer, Arturo Oncevay, Alexandra Birch
School of Informatics, University of Edinburgh
{**vivek.iyer**, a.oncevay, a.birch}@ed.ac.uk

I. INTRODUCTION

SOTA Recipe for training multilingual NMT models
Aligned Augmentation (Pan et al., 2021)

Synthesize Code-Switched (CS) sentences → Pretrain MT models to "denoise" CS sentences → TA DA! Better cross-lingual representations Superior MT performance

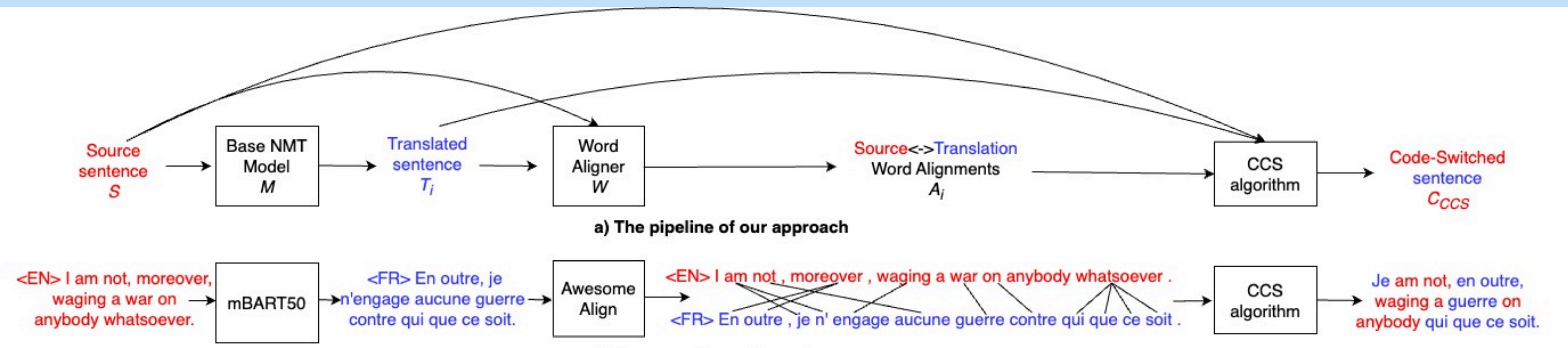
- For synthesising code-switched sentences, Pan et al. (2022) use bilingual **MUSE dictionaries**
- These only provide **non-contextual**, **one-to-one word-level** translations
- This leads to **significant noise** in the pretraining corpus (polysemes, multi-word expressions, lack of linguistic agreement etc)
- This in turn might potentially harm downstream MT performance!

II. RESEARCH QUESTIONS

RQ1) Does synthesise higher quality CS text lead to better downstream MT performance?
RQ2) How does CS pretraining scale to "more challenging" language families, such as agglutinative and low-resource languages?
RQ3) What are the key factors to consider when pretraining on CS text, and what role do they play in performance enhancement?

III. APPROACH

To answer this, we propose CCS (Contextual Code-Switching) that extracts **contextual, many-to-many substitutions** for generating high-quality CS text.



a) The pipeline of our approach

b) An example run-through

IV. EXPERIMENTS

We conduct experiments on 3 different language families, comparing the dictionary-based (AA) and contextual (CCS) approaches. Results display spBLEU scores.

	En - Es		En - Fr		En - It		En - Ro		Avg.	
	→	←	→	←	→	←	→	←	→	←
AA	25.0	26.2	28.8	28.7	23.8	26.8	18.7	24.1	25.7	27.0
CCS	**30.7**	**29.1**	**33.1**	**30.9**	**29.1**	**29.0**	**25.4**	**30.4**	**29.6**	**29.9**
Δ	**+5.7**	**+2.9**	**+4.3**	**+2.2**	**+5.3**	**+2.2**	**+6.7**	**+6.3**	**+3.9**	**+2.9**

a) Romance (High-Resource)

	En - Fi		En - Et		Avg.	
	→	←	→	←	→	←
AA	15.6	19.3	20.5	23.3	18.1	21.3
CCS	**21.2**	**21.2**	**25.6**	**25.7**	**23.4**	**23.5**
Δ	**+5.6**	**+1.9**	**+5.1**	**+2.4**	**+5.3**	**+2.2**

b) Uralic (Agglutinative)

	En - Hi		En - Gu		Avg.	
	→	←	→	←	→	←
AA	**28.4**	**24.6**	10.2	11.5	19.3	18.1
CCS	28.0	24.0	**12.9**	**12.9**	**20.5**	**18.5**
Δ	-0.4	-0.6	**+2.7**	**+1.4**	**+1.2**	**+0.4**

c) Indo-Aryan (Low-Resource)

V. ANALYSIS

i. Importance of Context

Source	45-year-old man has been remanded in custody on a firearms charge following a disturbance at a travellers ' site on Monday when six people were arrested .
AA	A 45-year-old humano tem been remanded in gardes sobre one arms débit following una necazuri at una travellers' site habilitado Monday when six people stavano arrested.
CCS	A 45-anos-old homme ha stato reţinut en custodia on a firearms charge urma a disturbión en un viajante 'site del manhã when six persone fueron arrestadas

ii. Importance of Many-to-Many substitutions

	Romance		Uralic		Indo-Aryan	
	En-X	X-En	En-X	X-En	En-X	X-En
CCS (1-1)	28.53	28.98	21.50	21.95	18.95	18.00
CCS (m-n)	**29.58**	**29.85**	**23.40**	**23.45**	**20.45**	**18.45**

iii. Importance of CS Language Count

	Romance		Uralic		Indo-Aryan	
	En-X	X-En	En-X	X-En	En-X	X-En
CCS (MLCS)	**29.6**	**29.9**	21.50	21.95	18.95	18.00
CCS (BLCS)	28.2	28.6	**23.40**	**23.45**	**20.45**	**18.45**

iv. Importance of Fine-Tuning

	Romance		Uralic		Indo-Aryan	
	En-X	X-En	En-X	X-En	En-X	X-En
CCS (mono. + parallel CS pretrain)	29.58	**29.85**	23.40	23.45	20.45	18.45
CCS (mono. CS pretrain) + parallel BLFT	28.65	28.43	23.55	23.80	16.35	14.50
CCS (mono. CS pretrain) + parallel MLFT	**30.00**	29.68	**25.20**	**25.85**	**23.55**	**22.35**

KEY TAKEAWAYS!

✓ **Improvements in quality of synthetic CS text can lead to huge improvements in MT performance, even beating massive models**
✓ **While improvements are observed across the board (even for low-resource langs), highest gains are for agglutinative languages**
✓ **Context, many-to-many substitutions, language count etc. play a key role in enhancing performance across various families**