UdS-DFKI@WMT20: Unsupervised MT and Very Low Resource Supervised MT for German↔Upper Sorbian

Sourav Dutta[1] Jesujoba O. Alabi[1,3] Saptarashmi Bandyopadhyay[2]
Dana Ruiter[1] Josef van Genabith[1,3]

[1]Saarland University (UdS)

[2]University of Maryland

[3]DFKI GmbH

WMT 2020

Data

Corpus	# Sentences
Unsupervised	
Sorbian Institute$_{hsb}$	339k
Witaj Sprachzentrum$_{hsb}$	222k
Web-scraped data$_{hsb}$	134k
Europarl$_{de}$	2,107k
News Commentary$_{de}$	422k
Wikipedia$_{de}$	833k
Wikipedia$_{hsb}$	76k
Supervised	
Parallel	60k
Europarl	568k
JW300	1,179k

Preprocessing

1. Normalization, Deduplication, Tokenization, Truecasing
2. **Factorization**: `word token` | `lemma` | `POS`
   ... hsb: *UDPipe* lemmatizer and POS-tagger
   ... de: *Snowball* stemmer + *spacy* POS-tagger
3. Byte-Pair Encoding (BPE) of $5k$ (default)
4. Quality token prefix (`<low>`, `<medium>`, `<high>`)
   per sentence per corpus

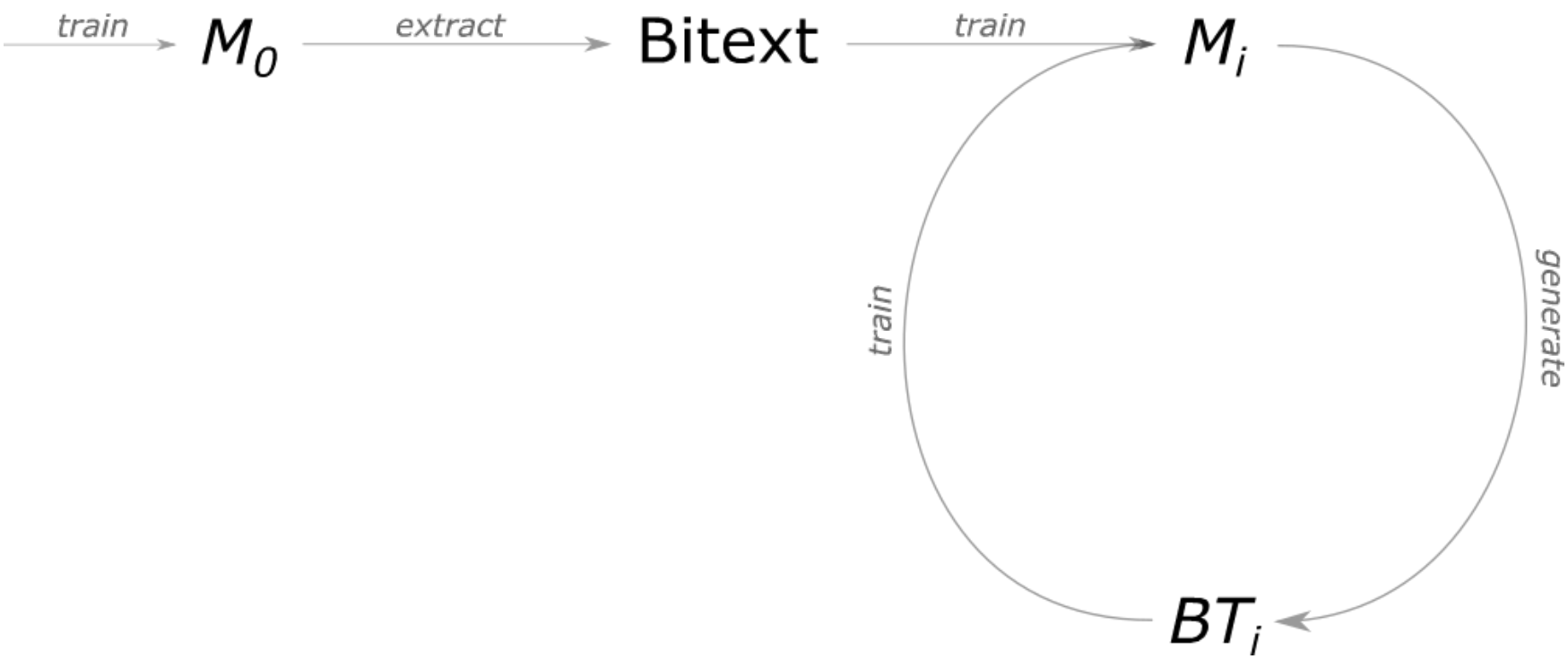
Supervised Task

1. German (de) -> Upper Sorbian (hsb)
2. Factorization of source text (German)

Model	**BLEU**		
	Dev20	DevTest20	Test20
de(fac.) $\rightarrow$ hsb	44.34	41.15	40.90

Unsupervised Task

Method



Unsupervised Task

Results

Model	**BLEU**		
	Dev20	DevTest20	Test20
de $\rightarrow$ hsb	13.60	9.90	10.30
hsb $\rightarrow$ de	11.30	8.10	8.90
hsb(fac.) $\rightarrow$ de	9.80	8.70	9.00

Thank you

Sourav Dutta
Saarland University, Germany
souravd@coli.uni-saarland.de