The TALP-UPC Machine Translation System for the WMT19 News Translation Task: Pivoting Techniques for Low Resource MT

Noe Casas noe.casas@upc.edu
José A. R. Fonollosa jose.fonollosa@upc.edu
Carlos Escolano carlos.escolano@upc.edu
Christine Basta christine.raouf.saad.basta@tsc.upc.edu
Marta R. Costa-jussà marta.ruiz@upc.edu

Task Description

- News translation: English $\longleftrightarrow$ Kazakh
- Low resources ($\sim$ 100K parallel segments)
- Large amount of Russian-English data ($\sim$ 6M).
- Large amount of Russian-Kazakh data ($\sim$ 4M).

Corpora (after cleaning)

- Kazakh-English training data:

	sents.	words	vocab.	L_{max}	L_{mean}
Kazakh	99.6K	1.2M	139.6K	85	11.7
English		1.5M	85.3K	102	14.9

- English-Russian training data:

	sents.	words	vocab.	L_{max}	L_{mean}
Russian	6.1M	125.6M	3.2M	80	20.7
English		144.9M	2.0M	80	23.9

- Russian-Kazakh training data:

	sents.	words	vocab.	L_{max}	L_{mean}
Russian	4.2M	78.8M	1.4M	96	18.9
Kazakh		75.3M	1.6M	70	18.0

SMT Pivot Systems

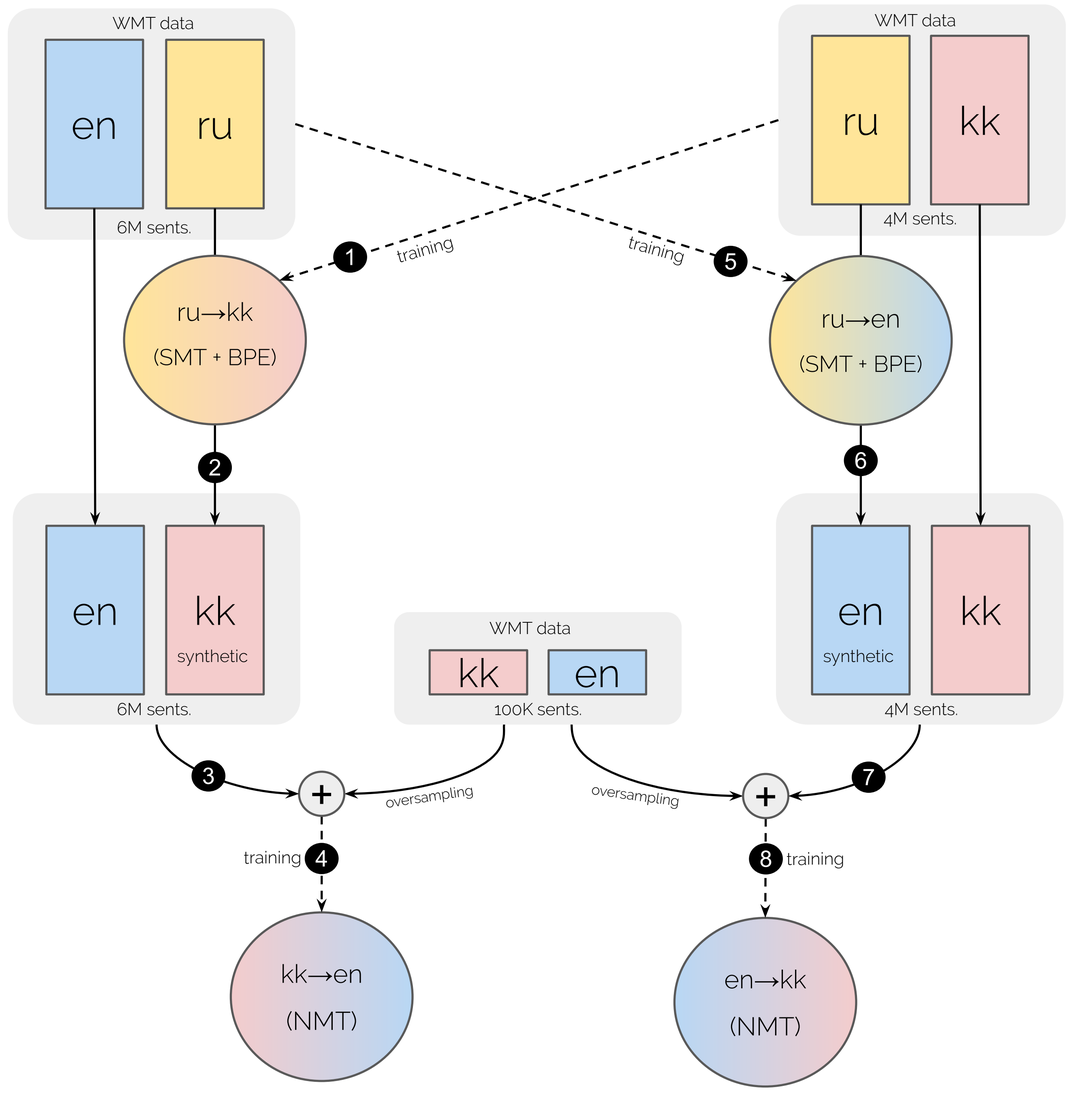
- BPE to cope with morphologically-rich langs (32K merge operations).
- 5-gram order LM to compensate for subword segmentation.
- Phrase-based SMT with Moses.

NMT Systems

- BPE vocabulary with 32K merge operations.
- Joint source-target attention model (He et al., 2018; Fonollosa et al., 2019).
- Hyperparams: 14 layers, embedding dimensionality of 1024, feedforward expansion of dimensionality 4096 and 16 attention heads.

This work is supported by Lucy Software / United Language Group (ULG) and the Catalan Agency for Management of University and Research Grants (AGAUR) through an Industrial PhD Grant, by a Google Faculty Research Award, by AGAUR through the FI PhD Scholarship and by the Spanish Ministerio de Economía y Competitividad, the European Regional Development Fund and the Agencia Estatal de Investigación through the postdoctoral senior grant Ramón y Cajal, contract TEC2015-69266-P (MINECO/FEDER,EU) and contract PCIN-2017-079 (AEI/MINECO).

Strategy: pivoting through Russian



❶ Train ru→kk SMT engine (BPE vocab).

❷ Create synthetic Kakakh-English corpus: use ru→kk engine to translate into Kazakh the Russian side of the Russian-Kazakh data.

❸ Combine synthetic and parallel data.

❹ Train NMT system on the combined data.

❺ Train ru→en SMT engine (BPE vocab).

❻ Create synthetic English-Kazakh corpus : use ru→en engine to translate into English the Russian side of the Russian-Kazakh data.

❼ Combine synthetic and parallel data.

❽ Train NMT system on the combined data.

Evaluation (BLEU on held-out subset of dev)

	RBMT	SMT	SMT+BPE	NMT	NMT
Kazakh→English	1.51	6.34	7.48	2.32	21.00
English→Kazakh	1.46	3.53	3.82	1.42	15.47
		parallel training data			augmented data

Competition Results

Kazakh→ English

Ave.	Ave. z	System
72.2	0.270	online-B
70.1	0.218	NEU
69.7	0.189	rug-morfessor
68.1	0.133	online-G
67.1	0.113	talp-upc-2019
67.0	0.092	NRC-CNRC
65.8	0.066	Frank-s-MT
65.6	0.064	NICT
64.5	0.003	CUNI-T2T-transfer
48.9	−0.477	UMD
32.1	−1.058	DBMS-KU

English→ Kazakh

Ave.	Ave. z	System
81.5	0.746	HUMAN
67.6	0.262	UAlacant-NMT
63.8	0.243	online-B
63.8	0.222	UAlacant-NM
63.8	0.222	RBMT
63.3	0.126	NEU
63.3	0.108	MSRA-CrossBERT
60.4	0.097	CUNI-T2T-transfer
61.7	0.078	online-G
55.2	−0.049	rug-bpe
49.0	−0.328	talp-upc-2019
41.4	−0.493	NICT
11.6	−1.395	DBMS-KU

Linguistic Background

Kazakh:

- Turkic language
- Cyrillic script.
- Agglutinative.

English:

- West Germanic lang.
- Latin script.
- Simple morphology.

Russian:

- East Slavic language.
- Cyrillic script.
- Fusional morphology.