



# NICT's Neural and Statistical Machine Translation Systems for the WMT18 News Translation Task

Benjamin Marie Rui Wang

Atsushi Fujita Masao Utiyama Eiichiro Sumita

{bmarie, wangrui, atsushi.fujita, utiyama, eiichiro.sumita}@nict.go.jp

National Institute of Information and Communications Technology (NICT), Kyoto, Japan

## SUMMARY

### The shared tasks

- News translation, *constrained* (only provided data are used)
- Estonian-English (et-en), Finnish-English (fi-en), Turkish-English (tr-en), and Chinese-English (zh-en)

### Key points

- No specific linguistic pre-processing
- SMT and NMT (transformer) systems
- Incremental training for NMT
- Combination of SMT and NMT through  $n$ -best lists reranking

### The results

- Best systems (BLEU) for Estonian-English and Finnish-English

## DATA

### Bilingual data

Language pair	#sent. pairs	#tokens	
et-en	1.9M	29.4M (Et)	36.0M (En)
fi-en	3.1M	52.9M (Fi)	72.8M (En)
tr-en	207.4k	4.4M (Tr)	5.1M (En)
zh-en	24.8M	509.9M (Zh)	576.2M (En)

### Monolingual data

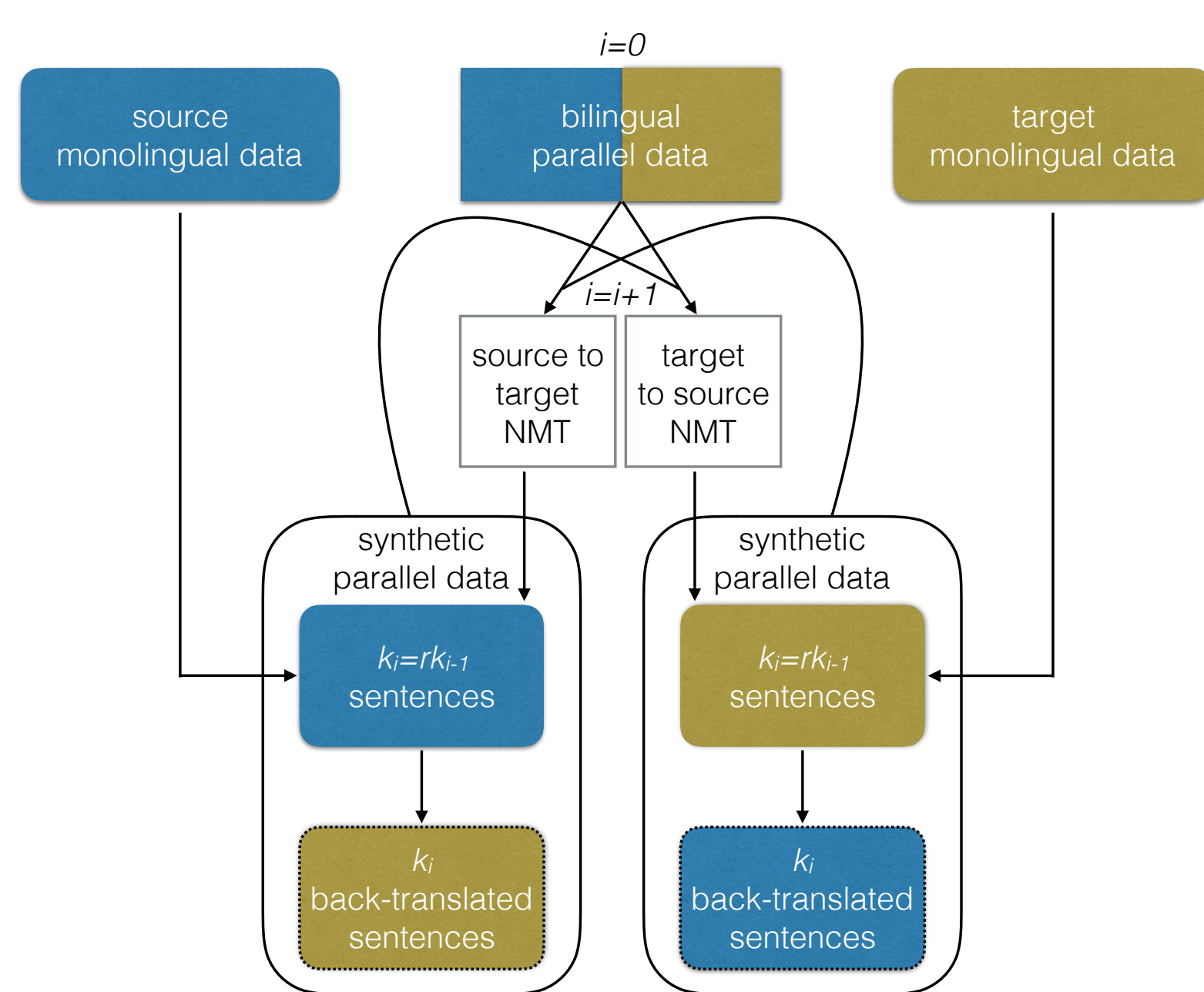
Language	#lines	#tokens
en	338.7M	7.5B
et	146.1M	3.6B
fi	177.1M	3.2B
tr	105.0M	1.8B
zh	130.5M	2.3B

### Pre-processing

- Tokenizer:
  - Moses: en, et, fi and tr
  - Jieba: zh
- Truecased (en, et, fi and tr)
- No data filtering

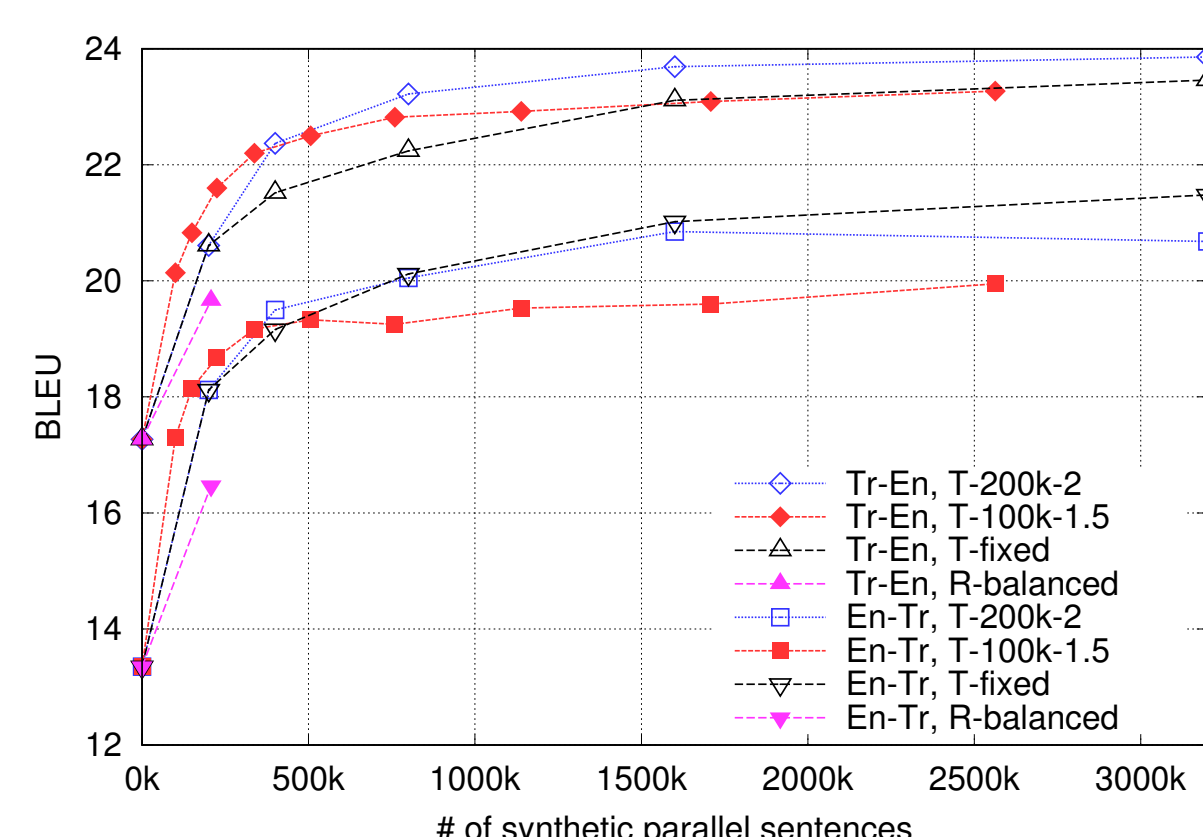
## INCREMENTAL TRAINING FOR NMT

### Training on more and better back-translated data



### Experiments with tr-en

- T-200k-2:  $r = 2, i_0 = 200k$
- T-100k-1.5:  $r = 1.5, i_0 = 100k$
- T-fixed: no incremental training
- R-balanced: backtr. with RNN



## COMBINATION OF SMT AND NMT

### Method (Marie+, 18)

1. Produce  $n$ -best lists with NMT and SMT
2. Compute SMT and NMT features for all the hypotheses
3. Merge the  $n$ -best lists
4. Rerank the hypotheses given the feature scores

### Features

- Left-right and right-left NMT forward and backward models
- 4-gram language models
- SMT lexical translation probabilities
- Averaged word posterior probabilities
- Hypothesis length features
- System flag
- Phrase-based forced decoding score (for tr-en only) (Zhang+, 18)
- MBR decoding for  $n$ -best list (for tr-en only)

## EXPERIMENTS

### Systems

- SMT (#3, #4, #5, #6): Moses (hierarchical for et-en and fi-en)
- NMT (#1, #2, #6): Marian (transformer model)
- Reranker (#2, #6): KB-MIRA

### Human evaluation

- First rank for all tasks (except for zh-en)
- Many other systems are also ranked first
- Strong correlation with BLEU

### Results

#	System	et→en	en→et	fi→en	en→fi	tr→en	en→tr	zh→en	en→zh
1.	Moses	18.2	15.1	15.8	10.7	12.1	8.4	16.9	28.0
2.	Moses NMT-reranked	20.2	17.6	17.5	12.2	14.2	10.1	19.0	29.9
3.	Marian single (w/o backtr)	22.9	18.5	17.6	13.2	20.2	12.2	23.7	33.0
4.	Marian single (w/ backtr)	28.6	24.0	23.1	16.8	25.2	18.0	24.7	37.2
5.	Marian ensemble (w/ backtr)	29.1	24.3	23.6	17.3	25.8	18.3	25.9	37.9
6.	Moses + Marian	30.7	25.2	24.9	18.2	26.9	19.2	26.7	39.7

- Moses (#1) underperforms Marian (#3) by several BLEU points
- Moses NMT-reranked (#2) is slightly below Marian (#3) for fi-en
- Large improvements with back-translation (#4) ( $> +5$ BLEU)
- Small improvements with ensemble decoding (6 models) (#5)
- Best results when combining SMT and NMT (#6)
- Best systems (#6) of WMT for et-en and fi-en (BLEU)