Parameter Optimization for SMT: It Pays to **Learn** from **Hard Examples**

Preslav Nakov, Fahad Al-Obaidli, Francisco Guzmán, Stephan Vogel

Qatar Foundation — Qatar Computing Research Institute (معهد قطر لبحوث الحوسبة)

1. Overview

Statistical Machine Translation (SMT)

- tuning/testing: multiple references
- direction: X→English (e.g., NIST, IWSLT)

Reversing the direction to English→X

- single reference translation
- **multiple sources** for each tuning/testing sentence

Question: How to use the multiple tuning inputs?

Answer: Use the hardest input.

2. Method

A. Choosing an entire dataset

- **Baselines**
  - **select-first**: use the first input
  - **concat-all**: use all inputs
- **Choose BLEU / LEN**
  - **best-on-tuning**: tune using one input, then use the learned parameters to translate all inputs
  - **X-vs-all-but-X**: score each English input using the other English inputs as references
  - **backtranslate**: translate the reference to English, then compare it to each English input

B. Select the best input for each sentence

- **Compare to an "English" reference from**
  - our own system
  - Google Translate
- **Similarity measures:**
  - **BLEU+1 (B1)**: smooths n-gram counts only
  - **BLEU+1 BP smooth (B1-BP)**: also smooths length ratio
  - **BLEU+1 sigmoid length penalty (B1-SG)**: symmetric length penalty

$$LP(s_i, r) = 3 - 4 * \mathtt{sig}\left(\left[\frac{l(s_i) - l(r)}{\alpha}\right]^2\right)$$

  - **Length Difference (DL)**: minimize
  - **Minimum BLEU+1 (MIN-B1)**: instead of max
  - **Minimum Length (MIN-L).**

C. Synthesize input for each sentence

- **MEMT**: fuse all inputs into one new input, sub-sentence

3. Results

Tuning on NIST MT04x, testing on MT05x

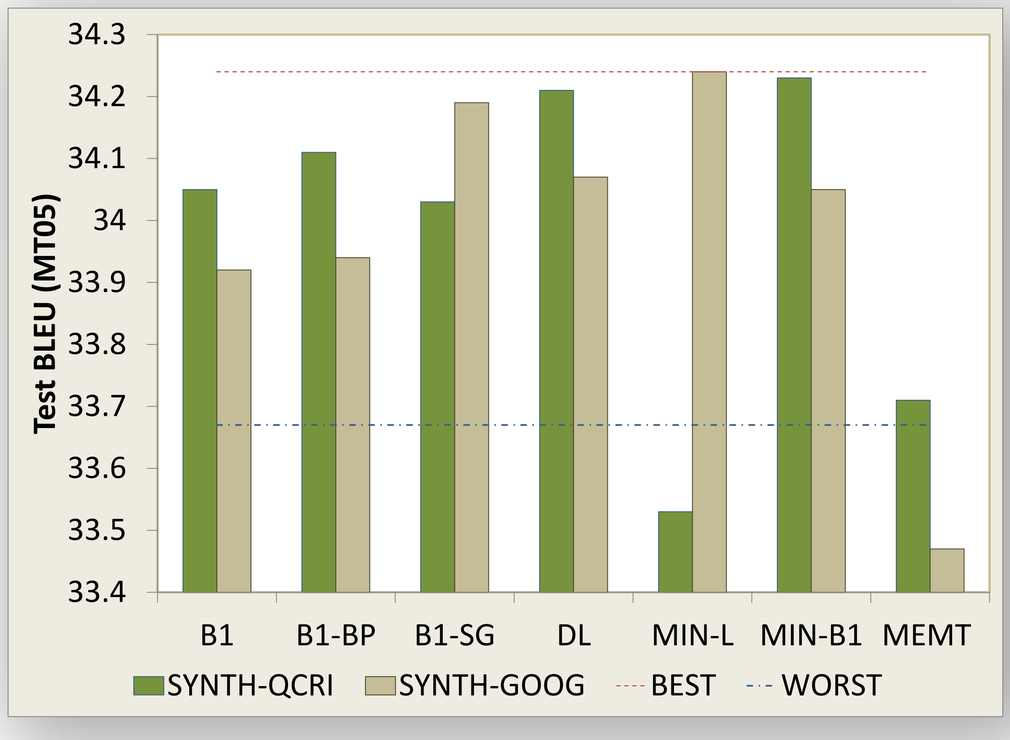
	MT050	MT051	MT052	MT053	MT054	AVG
MT040	**34.63**	**30.96**	**29.73**	40.40	**35.46**	**34.24**
MT041	34.37	30.59	29.44	**40.91**	35.31	34.12
MT042	34.34	30.57	29.08	40.64	35.12	33.95
MT043	33.99	30.23	29.06	40.62	34.81	33.74
MT044	33.87	30.18	28.96	40.51	34.82	33.67

Choosing an entire dataset

Selection strategy	MT040	MT041	MT042	MT043	MT044
BLEU					
AVG-BLEU	O				X
AVG-BLEU-noself	X				O
REF:ALL.vs.X-BLEU	O			X	
Backtranslate-QCRI:BLEU	O				X
Backtranslate-GOOG:BLEU	O			X	
LEN					
AVG-LR	O	X			
AVG-LR-noself	O	X			
REF:ALL.vs.X-LEN	O	X			
Backtranslate-QCRI:LEN		X			O
Backtranslate-GOOG:LEN	X				O
MT05-AVG-BLEU score	**34.24**	34.12	33.95	33.74	33.67

O = worst according to the selection criteria
X = best according to the selection criteria

Selecting/synthesizing input for each sentence



4. Conclusion

It is best to tune on the hardest available input

- Implications: (i) how we should choose a good translator, and (ii) how to find useful data for tuning
- Or maybe just an issue with BLEU?

5. Future Work

- Other *datasets*: from NIST, IWSLT
- Other *language pairs*, e.g., Chinese-English
- Other *evaluation measures*: TER, METEOR
- *Quality estimation* techniques
- Better *evaluation* with multiple inputs