Selecting Data for English-to-Czech Machine Translation

Aleš Tamchyna, Petra Galuščáková, Amir Kamran, Miloš Stanojević, Ondřej Bojar
{tamchyna,galuscakova,kamran,bojar}@ufal.mff.cuni.cz, milosh.stanojevic@gmail.com
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics, Charles University in Prague

Parallel Data

CzEng 1.0

- 15 million parallel sentences.
- Various domains (law, fiction, web,...).
- Automatic ***filtering*** of bad sentences.

Clean Data ⇒ Better Translations?

Section	BLEU CzEng 0.9	BLEU CzEng 1.0	Vocab. Change
news (100k)	**14.34**	14.01	-9%
all (1M)	14.77	**15.23**	+10%

Filtering failed to distinguish between ***unusual*** and ***wrong*** sentence pairs.

⇒ Loss of vocabulary in some sections.

Conclusions

CzEng 1.0 vs. CzEng 0.9

- Overall, data from the new version lead to better translation quality.
- Filtering was beneficial but decreased vocabulary size for some domains.

Domain Adaptation Using IR

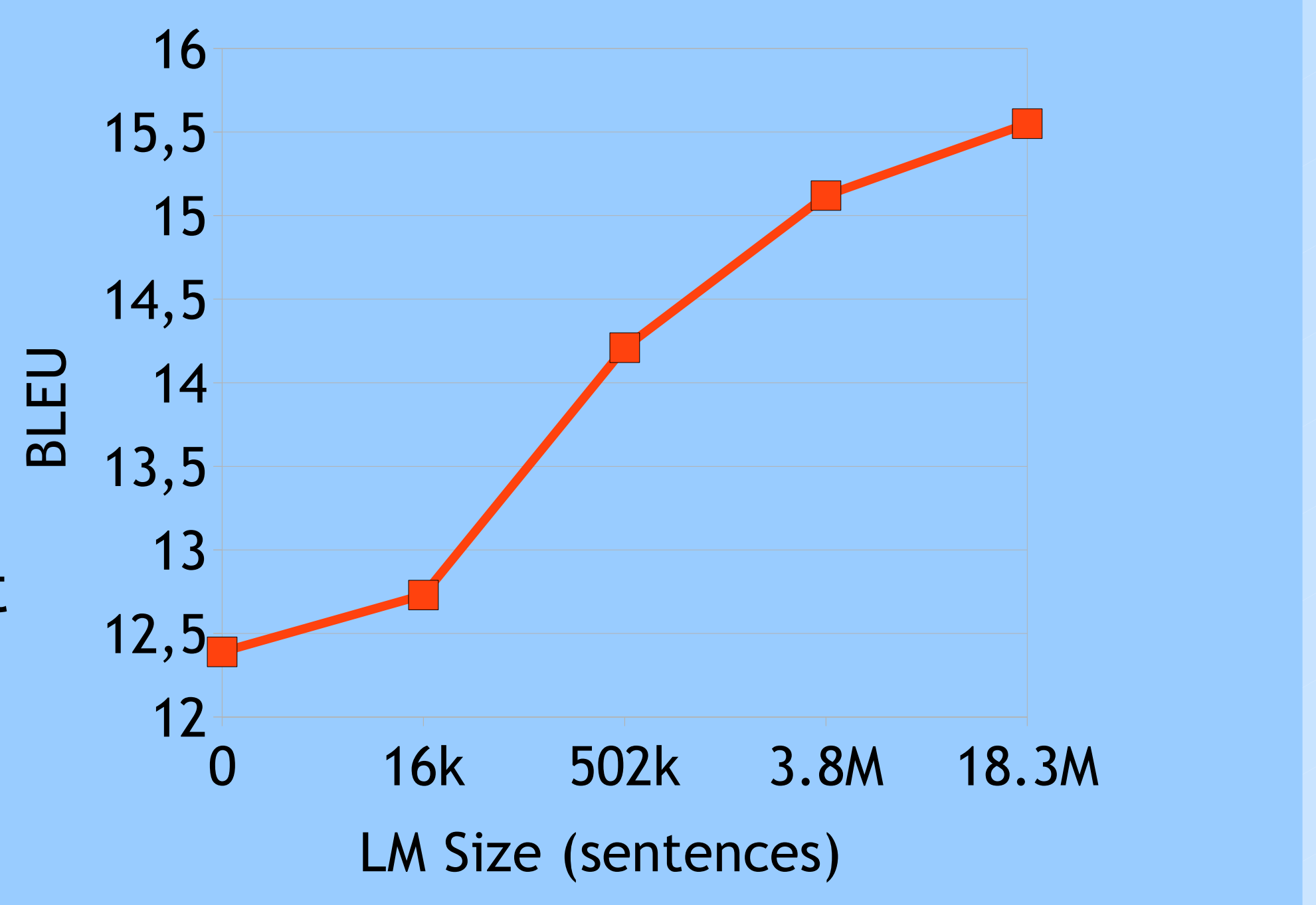
- Significant improvements in BLEU even with small additional selected data.
- Using a TM trained on reverse-translated data did not improve translations much further.
- Tuning to selected sentences helps, but good-quality in-domain data can outperform the Lucene-selected set.

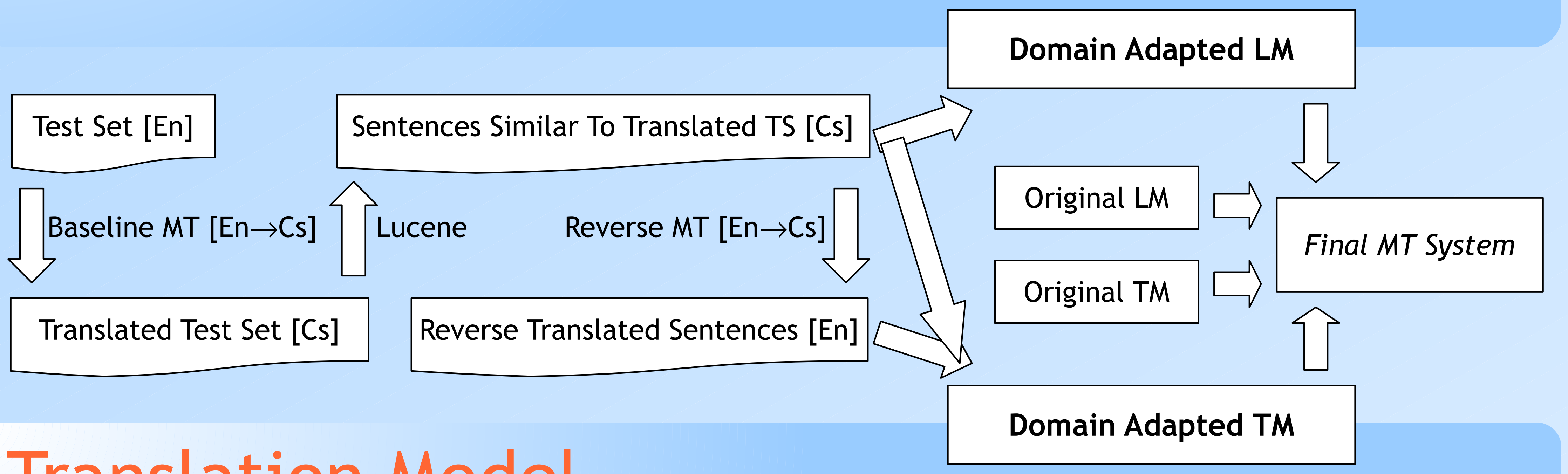
WMT Submission

- More reference translations in tuning lead to a better-rated system.

Language Model

- Domain adaptation for LM.
- Use ***Information Retrieval*** to select sentences from monolingual data.
- Criterion: ***Similarity to test set*** (source side).
- Train an additional LM on the selected sentences.
- Evaluation of various sizes of the tailored LM.
- Monolingual data from domain too similar to test set
  ⇒ Best performance when using the full corpus.
- Lucene queries use just bag-of-words:
  ⇒ Word frequencies ignored.
  ⇒ No regard for sentence structure.





Translation Model

- Use the selected sentences and their reverse translations as ***synthetic parallel data***.
- Final system uses 2 TMs (baseline + synthetic) and 2 LMs.
- Evaluated two sizes of n-best lists for tuning.

Additional Models	N-Best Size	Selected Sentences	BLEU
None	100	0	12.39
None	200	0	12.40
LM	100	502k	14.21
LM + TM	100	502k	14.32
LM + TM	200	502k	**14.36**

Submitted Systems

CU-TAMCH-BOJ

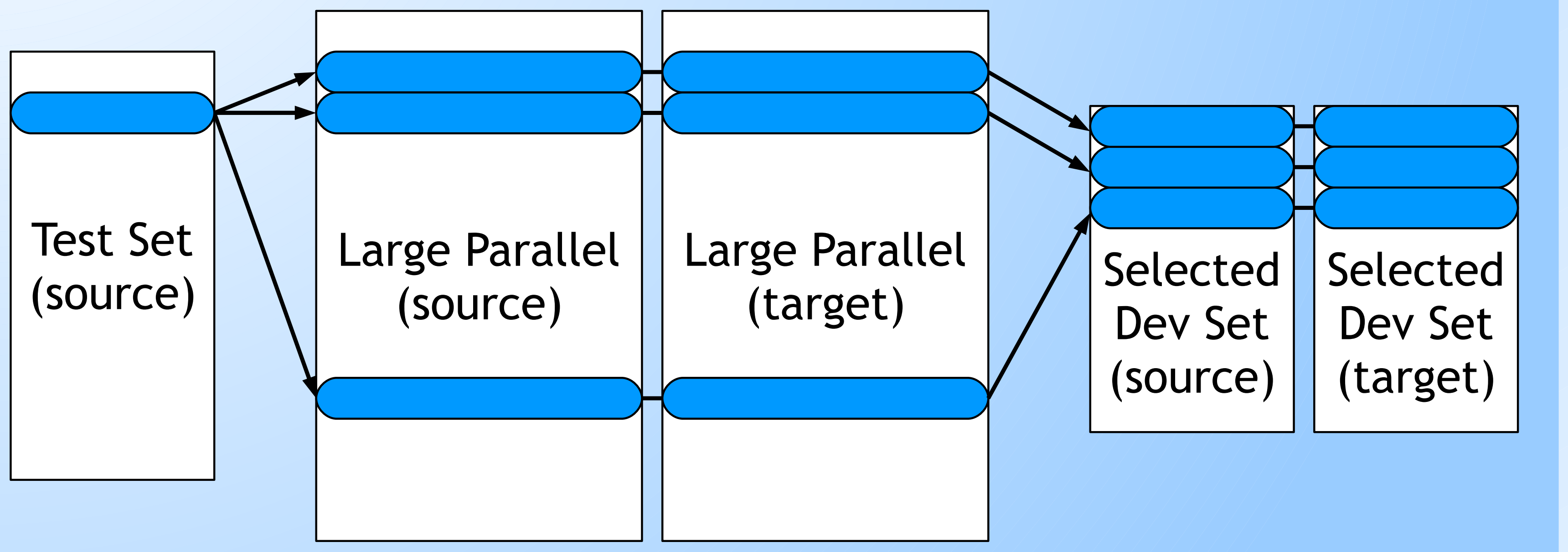
- English→Czech.
- Tuned on ***3 reference translations***.
- Parallel Data: CzEng 1.0.
- Monolingual: CzEng 1.0 + News Crawl.
- Factored setup: form|tag → form|tag.
- Target LMs on surface forms and tags.
- Contrastive baseline: 1 reference for tuning.

System	BLEU	TER	WMT Ranking
3 ref.	14.5	**0.765**	**4**
1 ref.	**14.6**	0.774	5

Tuning

Lucene Selection

Lucene used to select tuning sentences (devset) similar to test set.



Evaluation: different methods for devset selection.

- **Baseline** Random selection from parallel data.
- **Lucene** Lucene-selected similar sentences.
- **WMT10** Sentences for WMT10 evaluation.
- **Perfect** Test set (not a fair competitor).

Training data must not contain the development set
⇒ Selection is done before training.

System	BLEU
Baseline	11.41
Lucene	12.31
WMT10	**12.37**
Perfect	12.64

- The domain of WMT10 is identical to the test set.
- BLEU of Lucene-selected devset almost matches WMT10.
- Baseline is significantly worse.

Multiple References

Multiple reference translations of the development set.

Machine translated pseudo-reference

- Obtained using TectoMT, an English-Czech deep syntactic decoder.
- No improvement in BLEU.

Manually translated set

- WMT11 test data (1 reference) + 2 human translated references.
- BLEU similar to 1 reference but WMT12 ranking is better.

WMT 2012, June 7 – 8, Montreal

This work was supported by the project EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003+7E11051 of the Czech Republic) and the Czech Science Foundation grants P406/11/1499 and P406/10/P259.