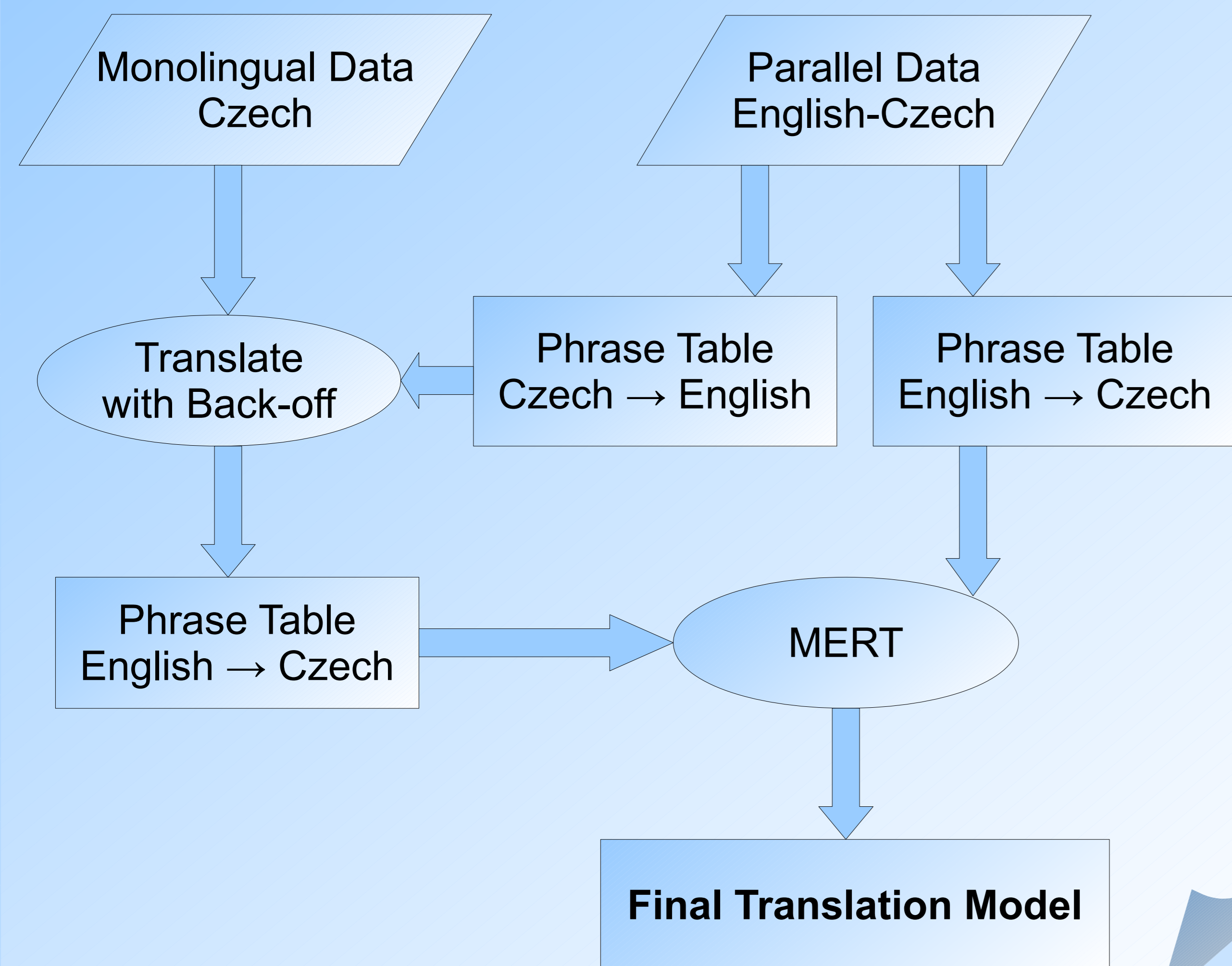


Overview

Target-side monolingual data help in LM. Can we use it also in TM?
 The Trick: Reverse self-training with back-off

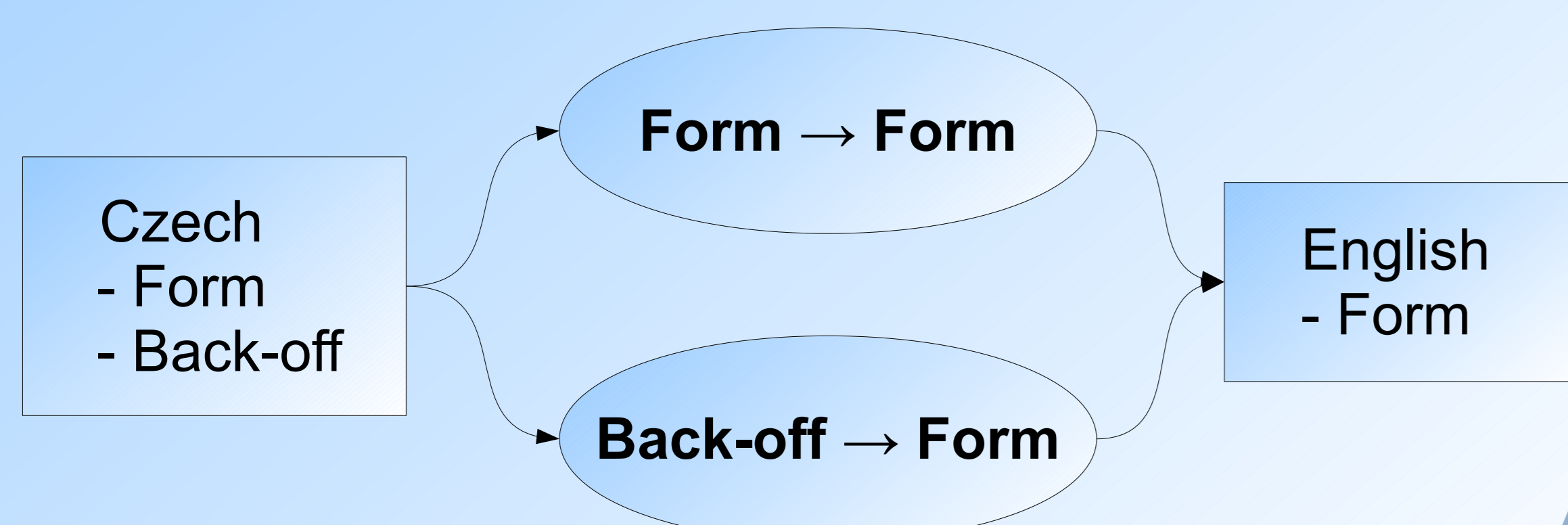
- Helps:
- in small data setting and
 - into morphologically rich languages.

Reverse Self-training



Translation with Back-off

- Reverse translation must handle unknown forms (these will become the newly learned forms).
- Factored model with alternative decoding paths.
- The back-off factor (e.g. lemmas) unifies different word forms.



Learning Unseen Forms

Small Parallel Data:

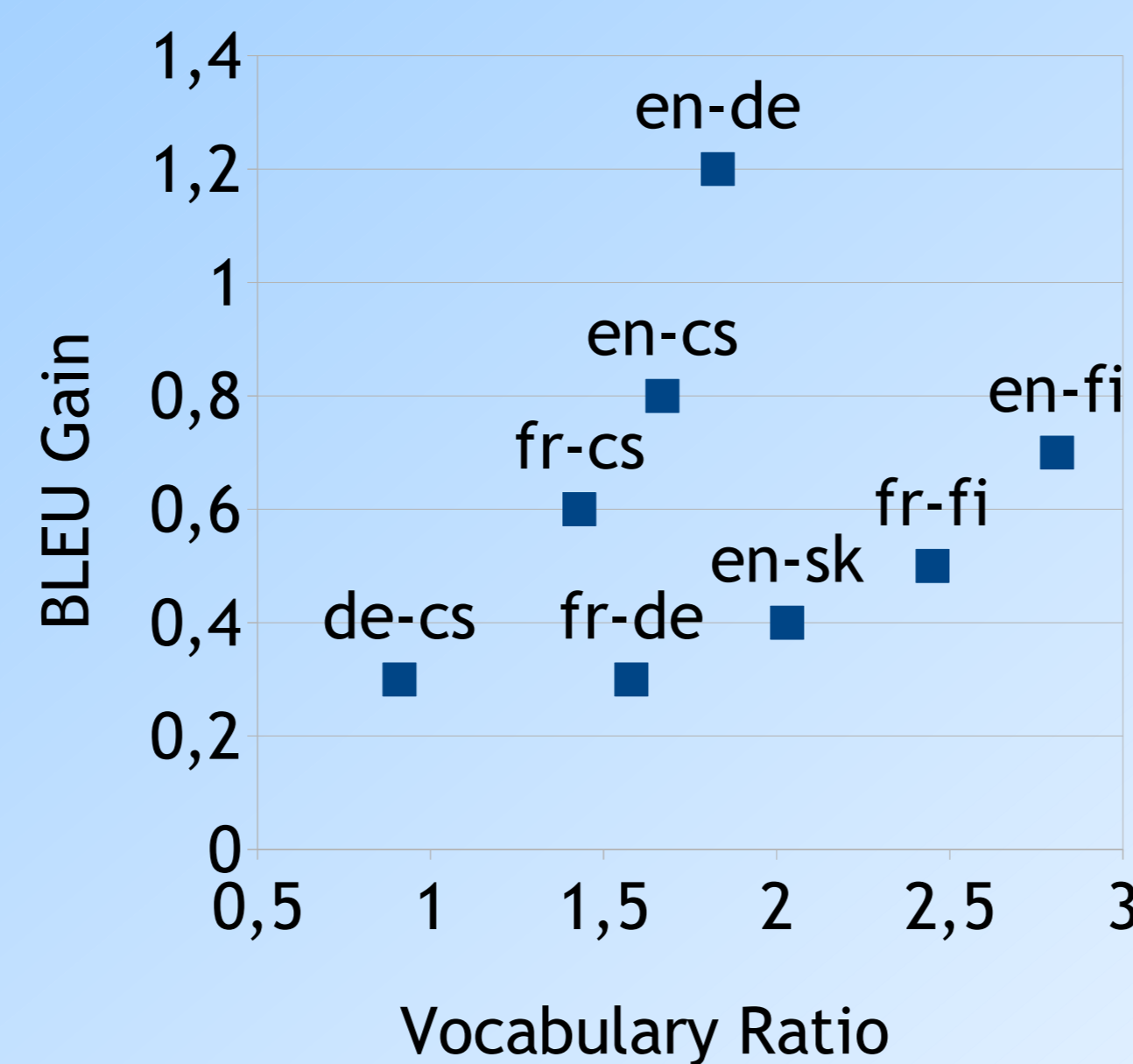
| Source English | Target Czech | Czech Lemmatized |
|--------------------|-------------------|------------------|
| a cat chased... | kočka honila... | kočka honit... |
| I saw a cat | viděl jsem kočku | vidět být kočka |
| I read about a dog | četl jsem o psovi | číst být o pes |

Large Monolingual Data:

? četl jsem o kočce číst být o kočka
 I read about a cat ← Use reverse translation backed-off by lemmas

⇒ Learned a new phrase (o kočce) including a form never seen in parallel data (kočce).

Comparison Across Languages

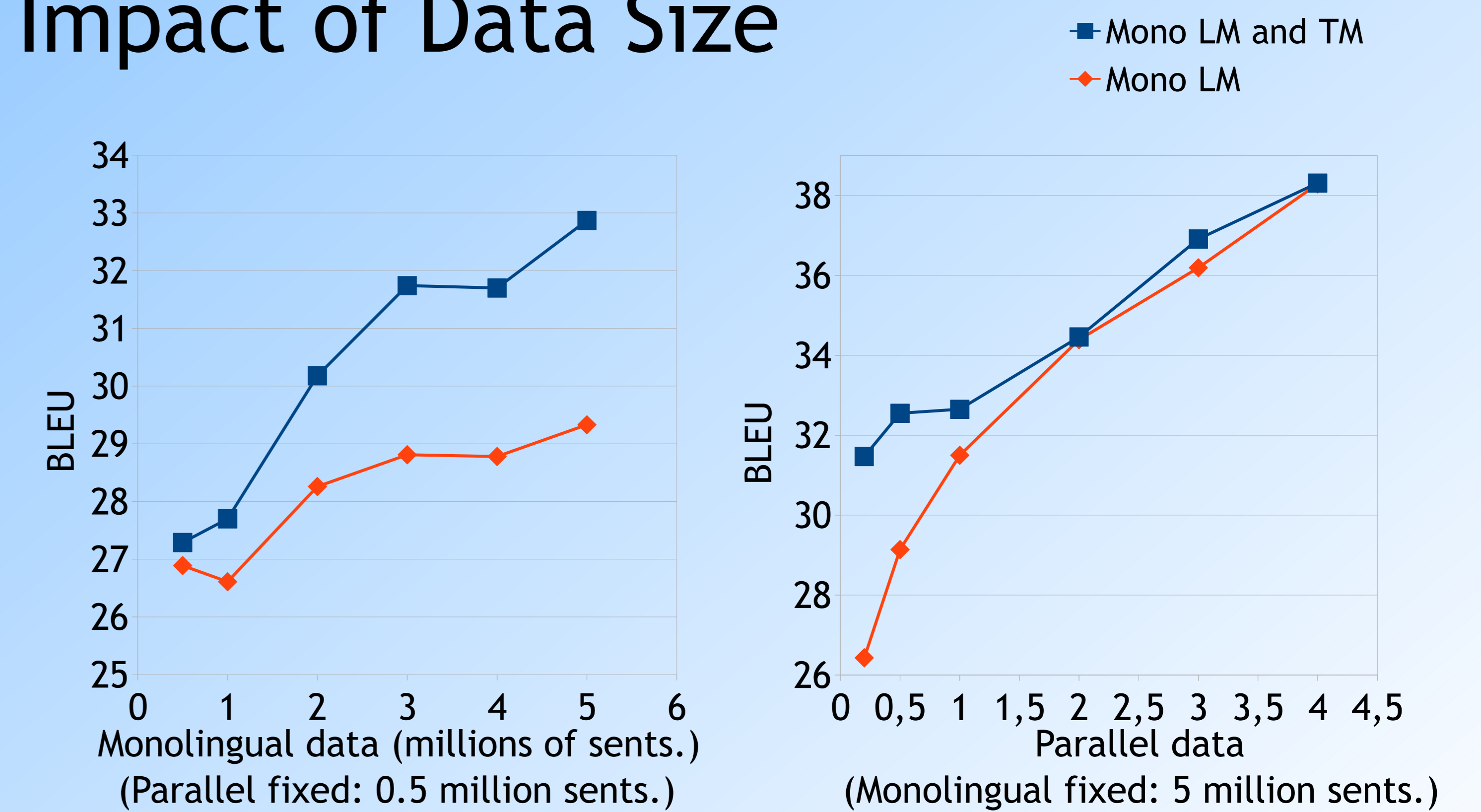


- Absolute gain over the +Mono LM baseline
- Parallel Data: 90-125 thousand sentences
- Monolingual Data: 0.6-0.9 million sentences

Problems

- MERT had to optimize many weights of two very similar models.
- Many derivations lead to the same hypothesis
- 100-best list contained only ~6 unique strings, compared to ~35 in the baseline setup
 ⇒ unstable,
 ⇒ diverging runs had to be repeated.
- Lattice MERT did not help.
- Possible solution: "Better Hypothesis Testing..." (Clark et al., 2011).

Impact of Data Size



- English→Czech translation.
- All data from CzEng 0.9.

Our WMT11 System Submissions

| Target | Mono LM | +Mono TM |
|--------|---------|----------|
| German | 14.8 | 14.8 |
| Czech | 15.7 | 15.9 |

German - constrained
 Czech - constrained except LM

Case-insensitive BLEU scores

German

- No improvement in BLEU score.
- Parallel data already sufficiently large
- Not all available data used in reverse self-training.

Czech

- Achieved a small improvement.
- Only 2010 and 2011 News data used in reverse self-training.

Conclusions

- Reverse self-training learns to produce forms not seen in parallel data.
- Greater effect for language pairs with very different vocabulary sizes.
- More monolingual data => greater effect.
- More parallel data => the effect diminishes.
- Good back-off: forms with last 3 characters removed.