PhraseFix: Statistical Post-Editing of TectoMT

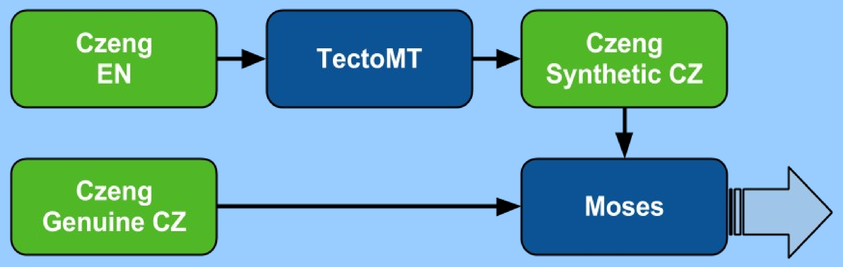
Petra Galuščáková, Martin Popel, and Ondřej Bojar
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{galuscakova,popel,bojar}@ufal.mff.cuni.cz

Given transfer-based system (**TectoMT**) and Moses (**SMT**).
Should we:

- Postedit? … statistical post-editing, SPE → **PhraseFix**
- Combine outputs?

PhraseFix

Moses is trained to translate from TectoMT output (i.e. poor Czech) into reference Czech.

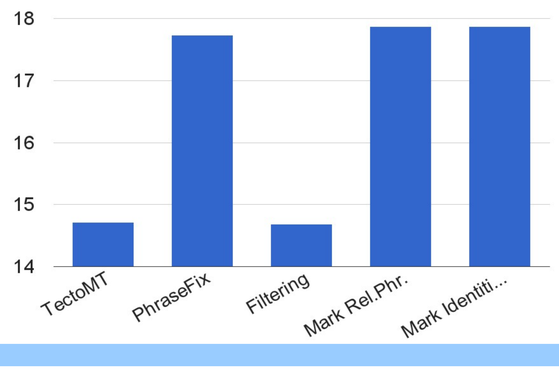


Variants

- **Phrase table filtering** - remove singletons and phrase pairs with forward probability ≤ 0.7
- **Marking of reliable phrases -** low-frequency phrase pairs are marked by a special feature
- **Marking of identities -** a special feature indicates the equality of the source and target phrases

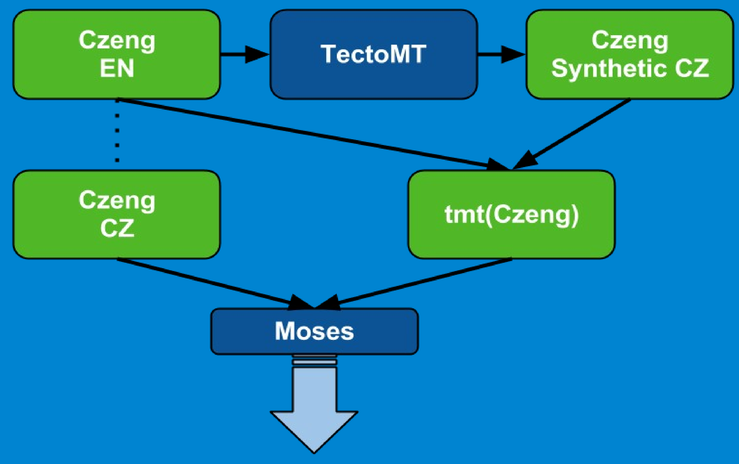
… MERT learns the weights of the additional features in Marking setups.

Results



Poor Man's System Combination

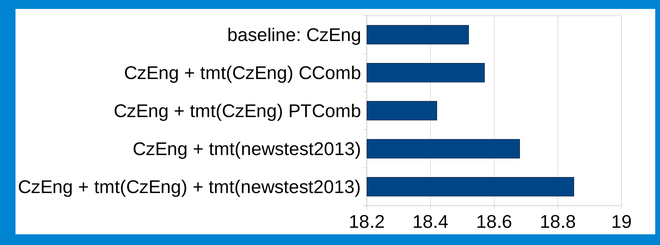
Output of TectoMT treated as additional parallel data for Moses.



Variants

- Corpus Combination (**CComb**) – output of TectoMT appended to the baseline parallel corpus
- Phrase Table Combination (**PTComb**) – output of TectoMT provides a secondary phrase table, MERT selects weights

Results



- Adding just *any* synthetic Czech is not very effective:
  CzEng + tmt(CzEng)
- Adding TectoMT output for **the test set** helps:
  CzEng + tmt(newstest2013)

Conclusion

- **SPE** works: **PhraseFix** better than **TectoMT**.
- **SPE** is further improved by marking identical and reliable phrases using a special feature.
- However, **SMT** still outperforms **SPE**.
- Using **TectoMT** output as additional data improves the **SMT** baseline.

This research is supported by the grants GAUK 9209/2013, FP7-ICT-2011-7-288487 (MosesCore) of the European Union and SVV project number 267 314.