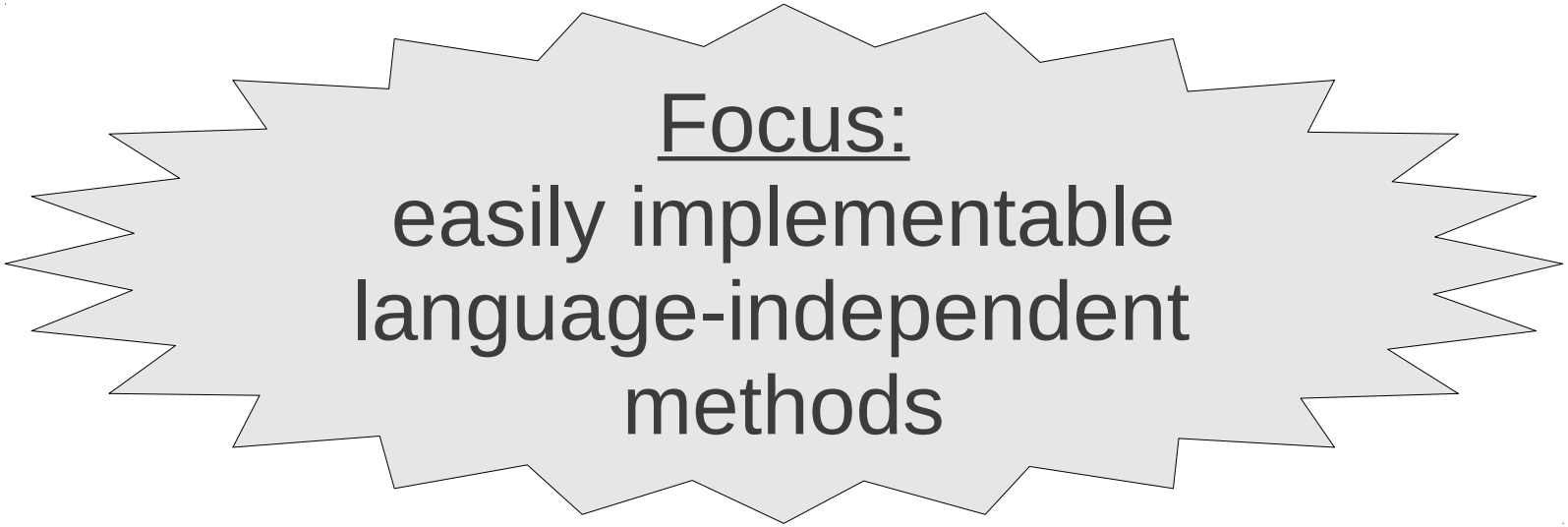
The NAIST Machine Translation System for IWSLT 2012

Graham Neubig, Kevin Duh, Masaya Ogushi, Takamoto Kano
Tetsuo Kiso, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura

Nara Institute of Science and Technology
12/6/2012

Overview

- Phrase-based machine translation
- Built on Moses (experiment management system)
- Evaluated on TED Translation:
  - English → French official track
  - XXX → English other tracks



English-French

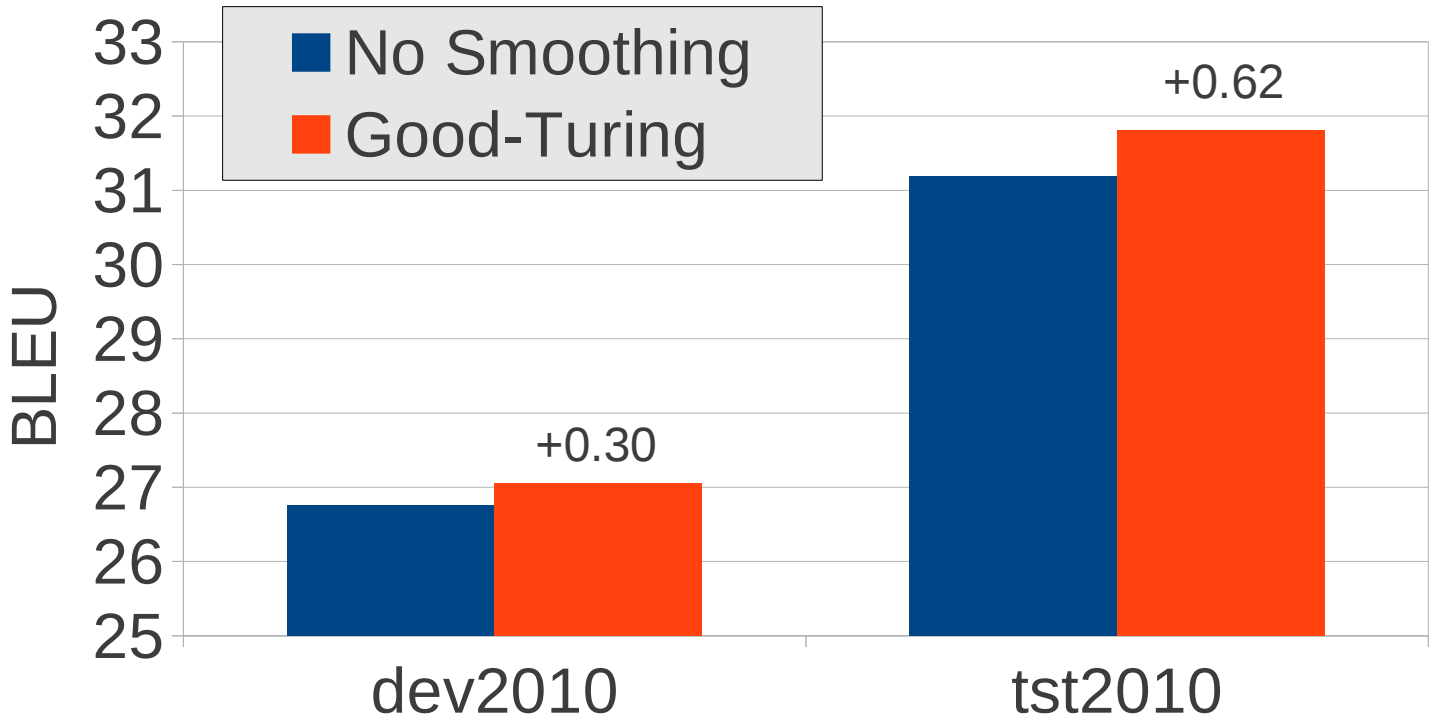
Summary of English-French

- Four successful statistical methods:
  - Phrase-table smoothing
  - Language model interpolation
  - Calibrated minimum Bayes risk decoding
  - Large-scale data with filtering
- Combination raises BLEU 29.75 → 31.81
- Ablation tests to examine the factors

Phrase Table Smoothing

- Phrase probabilities for rare phrases over-fit the training data
- Smoothing discounts observed counts when calculating probabilities
- Here we test Good-Turing smoothing [Foster 06]

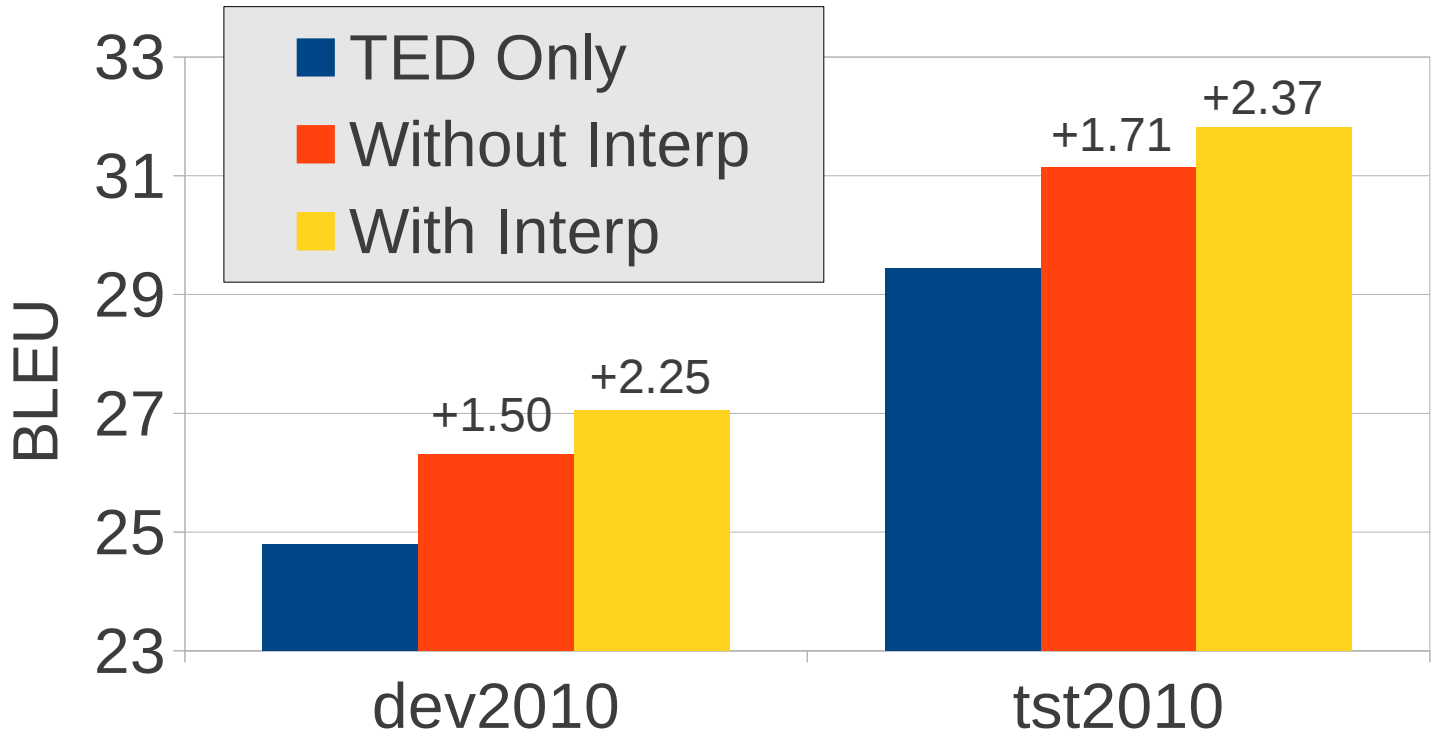
Smoothing Results



Language Model Interpolation

- LM data from heterogeneous sources
  - TED, News Commentary, Europarl, Giga
- Combine using simple linear interpolation
- Maximize likelihood of development set [Jelinek 80]

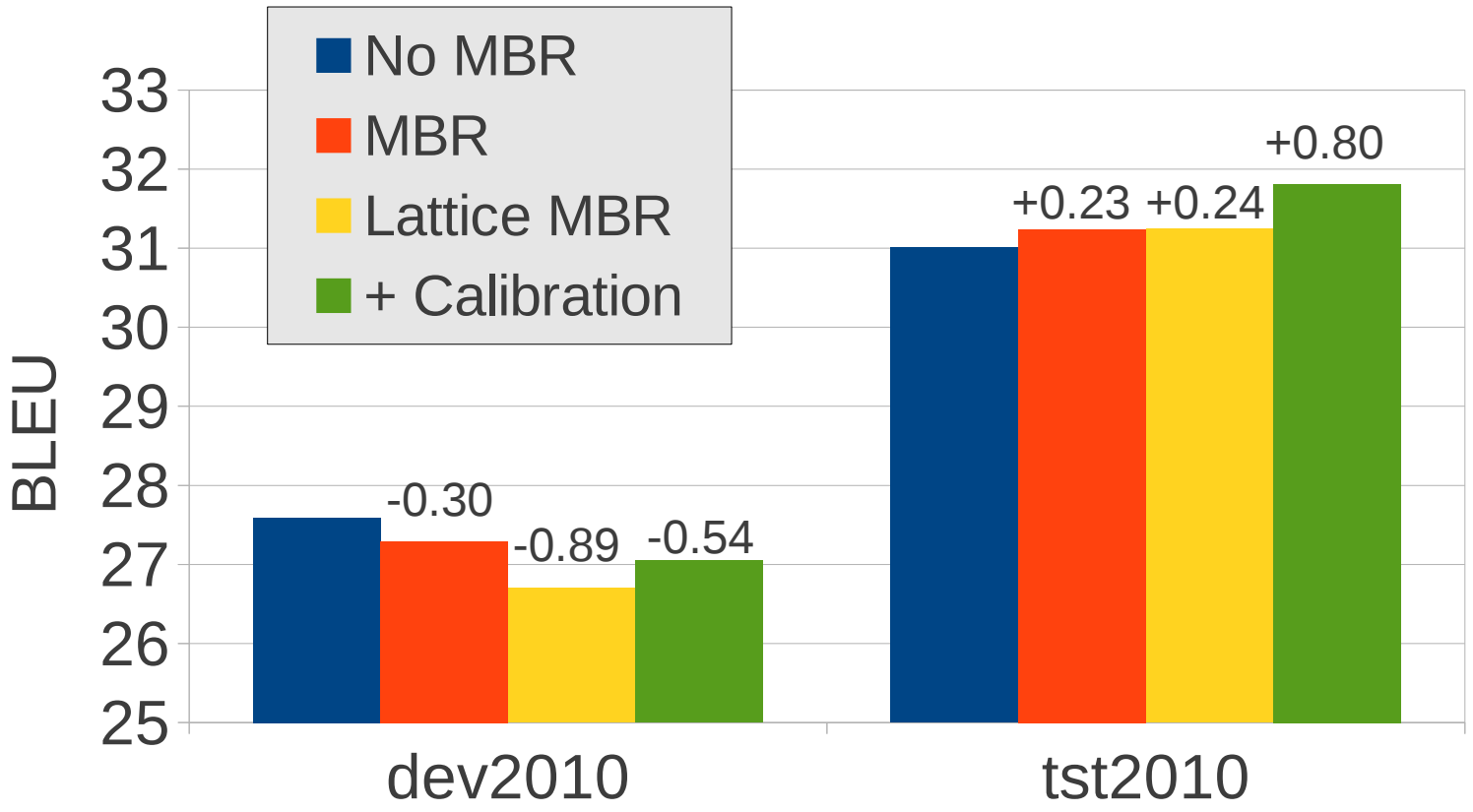
Language Model Interpolation Results



Minimum Bayes Risk Decoding

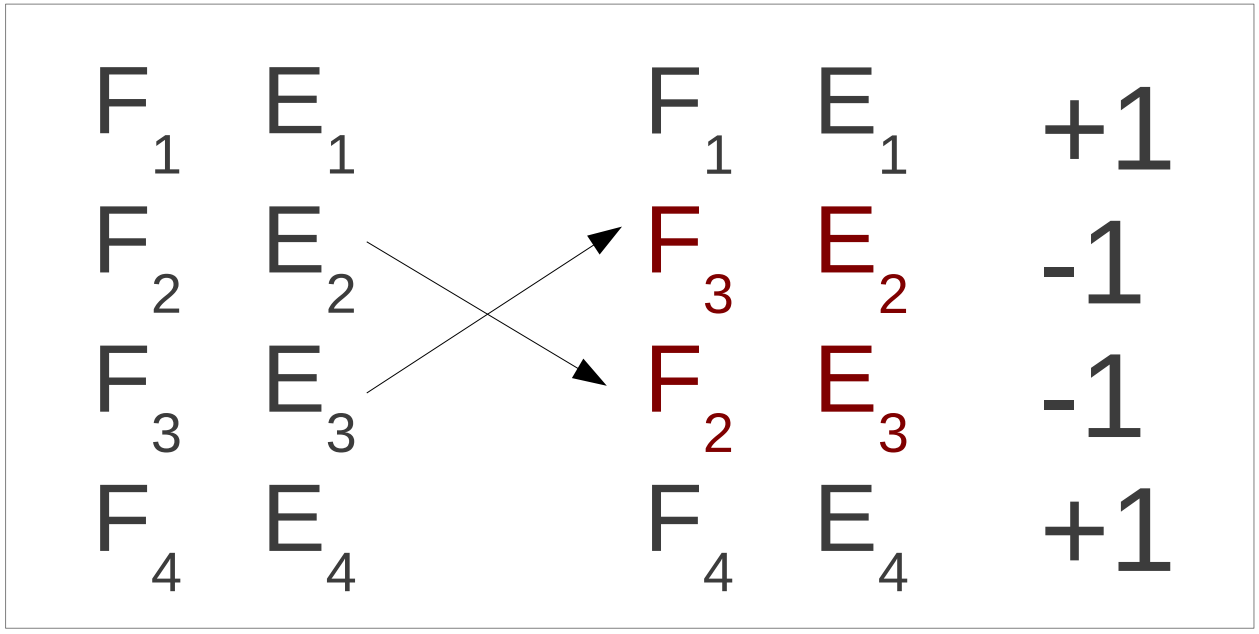
- Normal Decoding: Choose the translation with highest probability
- MBR Decoding: From an n-best list, choose the translation with the lowest expected loss [Kumar 04]
- Lattice MBR Decoding: MBR over lattices [Tromble 08]
- Also tested calibrating the probability distribution

Minimum Bayes Risk Results

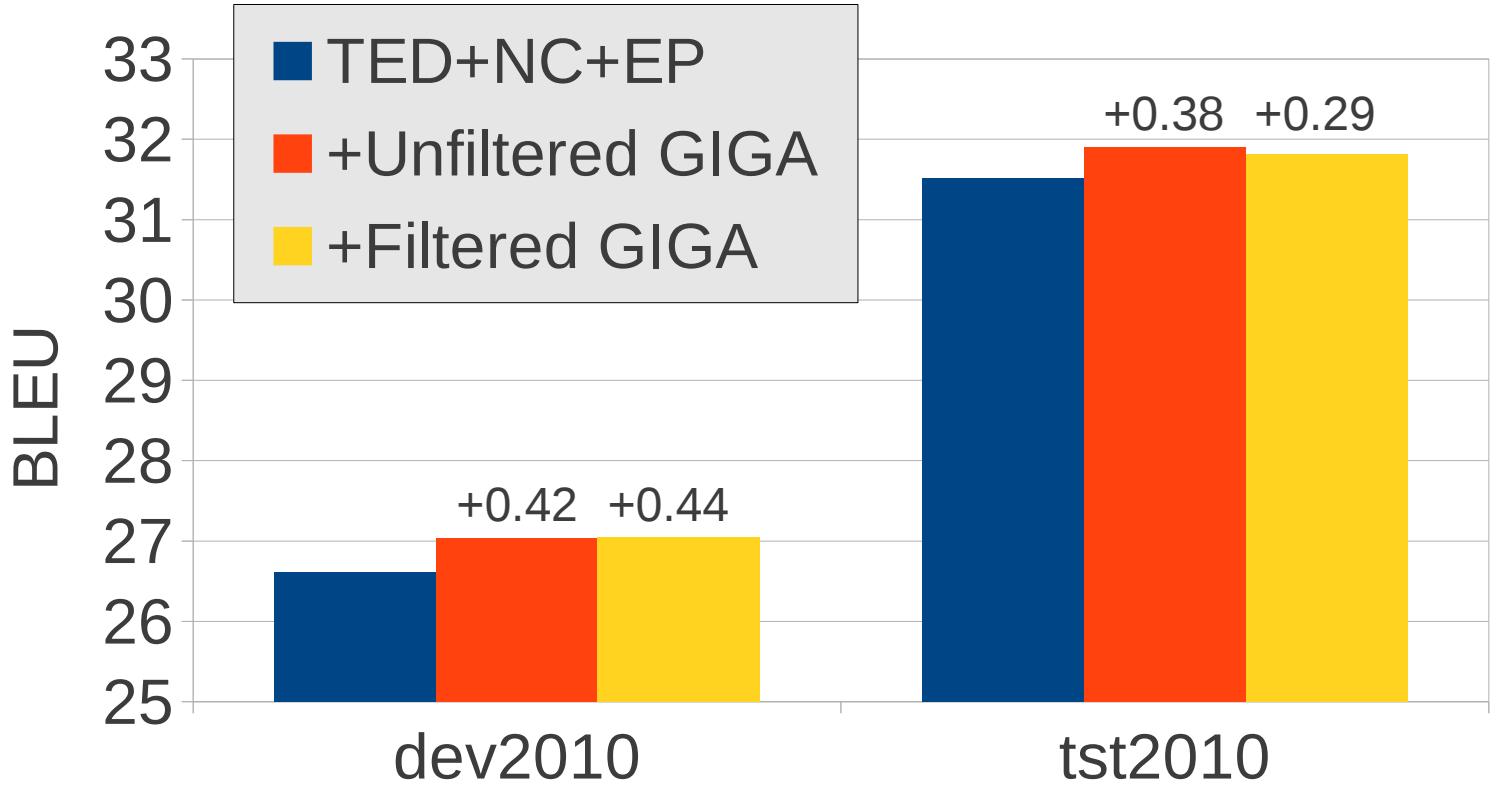


Large-Scale Data with Filtering

- Giga-word English-French corpus is large, but noisy
- Train a classifier to detect noisy sentences
  - Features: Model 1, Alignment, Length Ratio, Same Word
- Use pseudo-negative training examples by swapping 30% of sentences [Mediani 2011]



Data Filtering Results



Other Methods Investigated

- Out of domain TM data
- Word alignment methods + combination
- Lexical reordering models
- MERT vs. PRO tuning

See the paper for more details!

XXX-English Language Pairs

Linguistic Family Tree

- Indo-European Family:
  - Germantic: German (de), Dutch (nl), English (en)
  - Italic: **Portuguese (pt), Romanian (ro)**
  - Slavic: **Polish (pl), Russian (ru)**, Slovak (sk)
- Afro-Asiatic Family: Arabic (ar)
- Altaic Family: Turkish (tr)

MT Issues

- Morphology:
  - **pl**/**ru**/sk (fusional)
  - tr (agglutinative)
  - de/nl (compounding)
  - **pt**/**ro** (some inflection)
- Word order:
  - de/nl (SOV, V2)
  - ar (VSO)

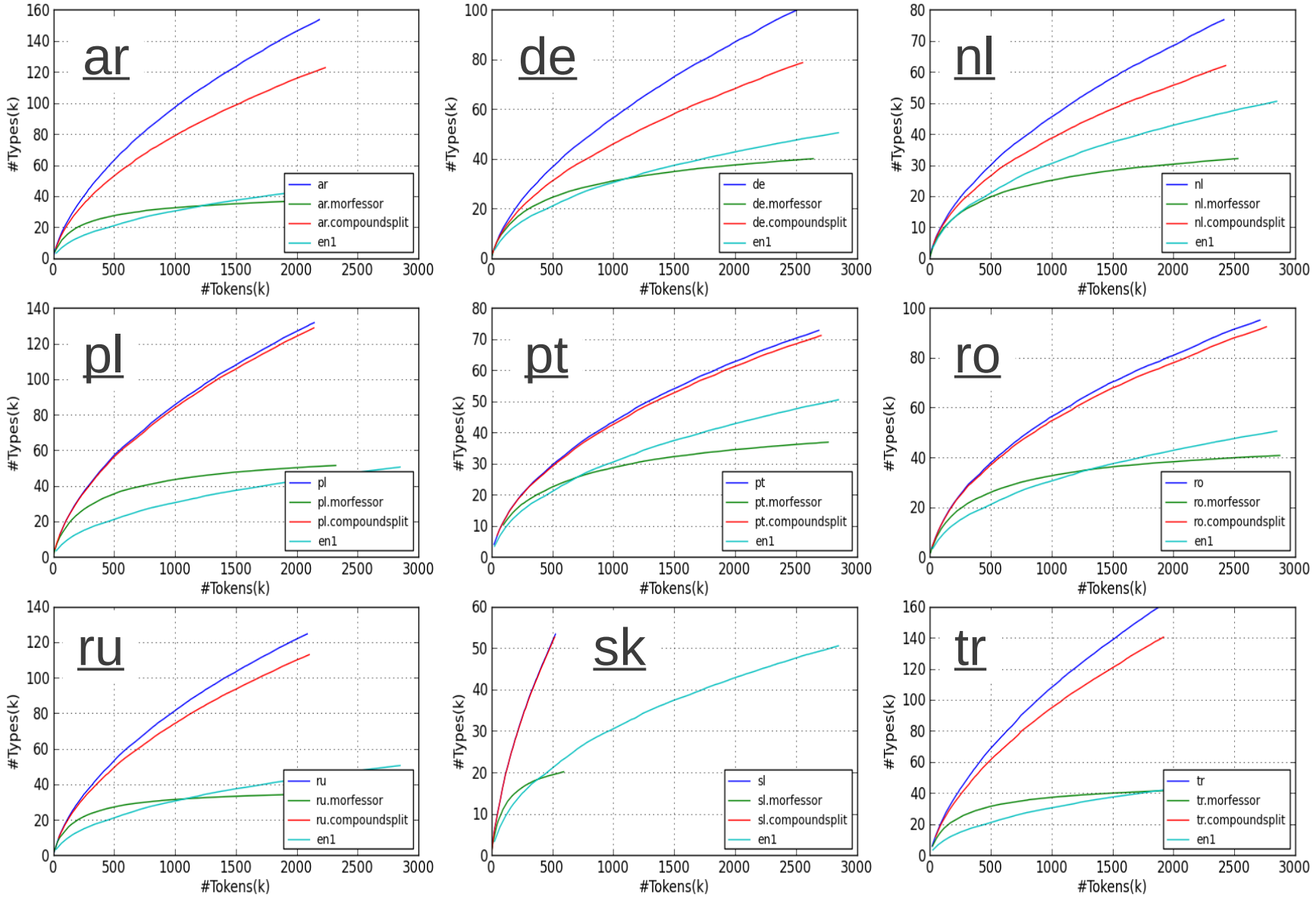
Summary of XXX-English Systems

- Common EMS setup: compare performance of existing techniques cross-linguistically
- What worked generally:
  - Unsupervised Morphology
    - Using Morfessor and compound-splitter.perl
  - Gigaword LM

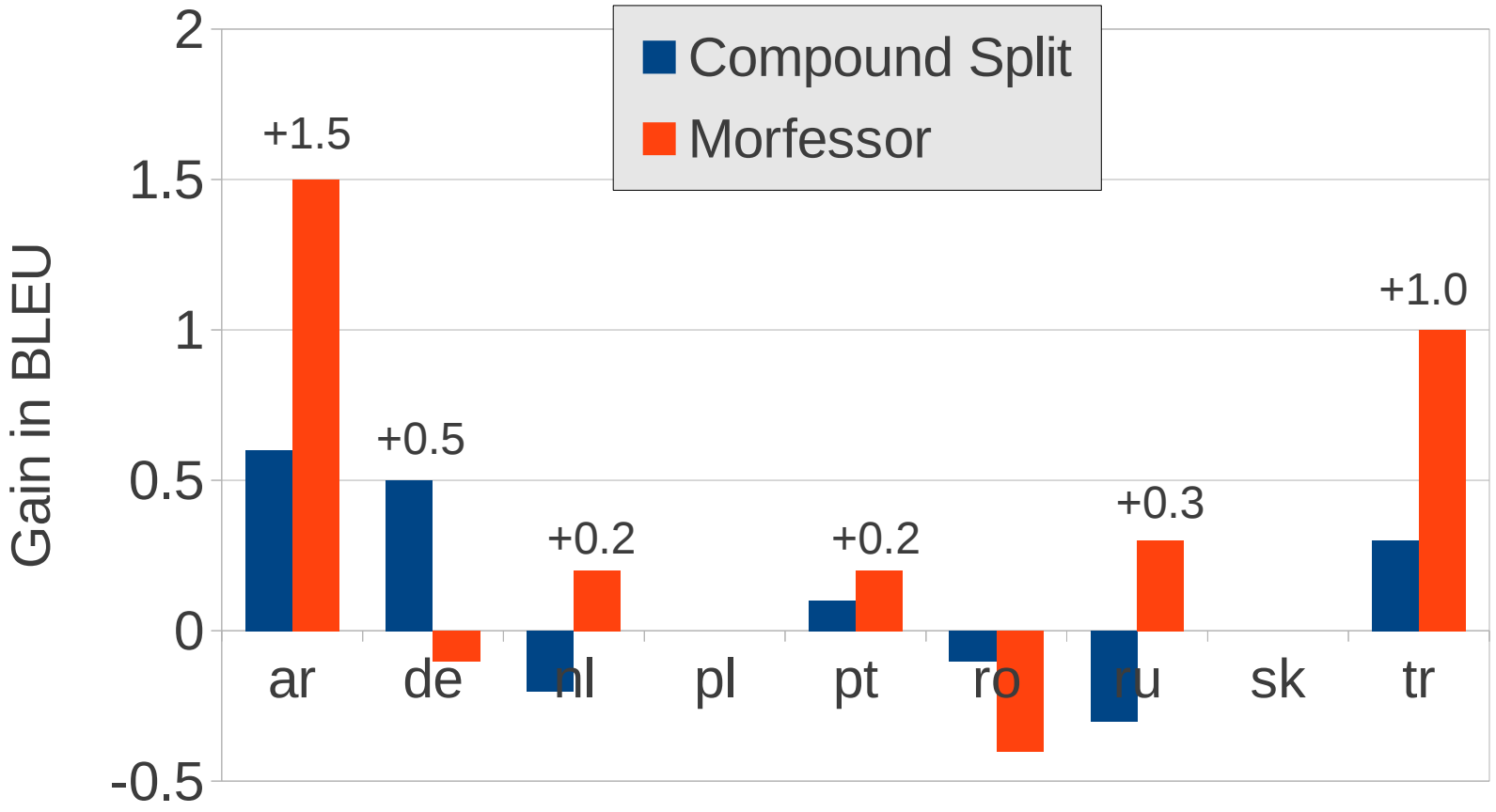
Unsupervised Morphology

- Compound-splitter.perl [Koehn 03]
  - Breaks apart words if subparts are seen in training data over a certain frequency
- Morfessor 1.0 [Creutz 02]
  - Use Minimum Description Length principle to find a small set of morphemes that covers the training words
  - Discovers both free & bound morphemes
  - Small modification: Morfessor segments too aggressively for unknown words, so keep OOV as is

Vocabulary Growth Rate

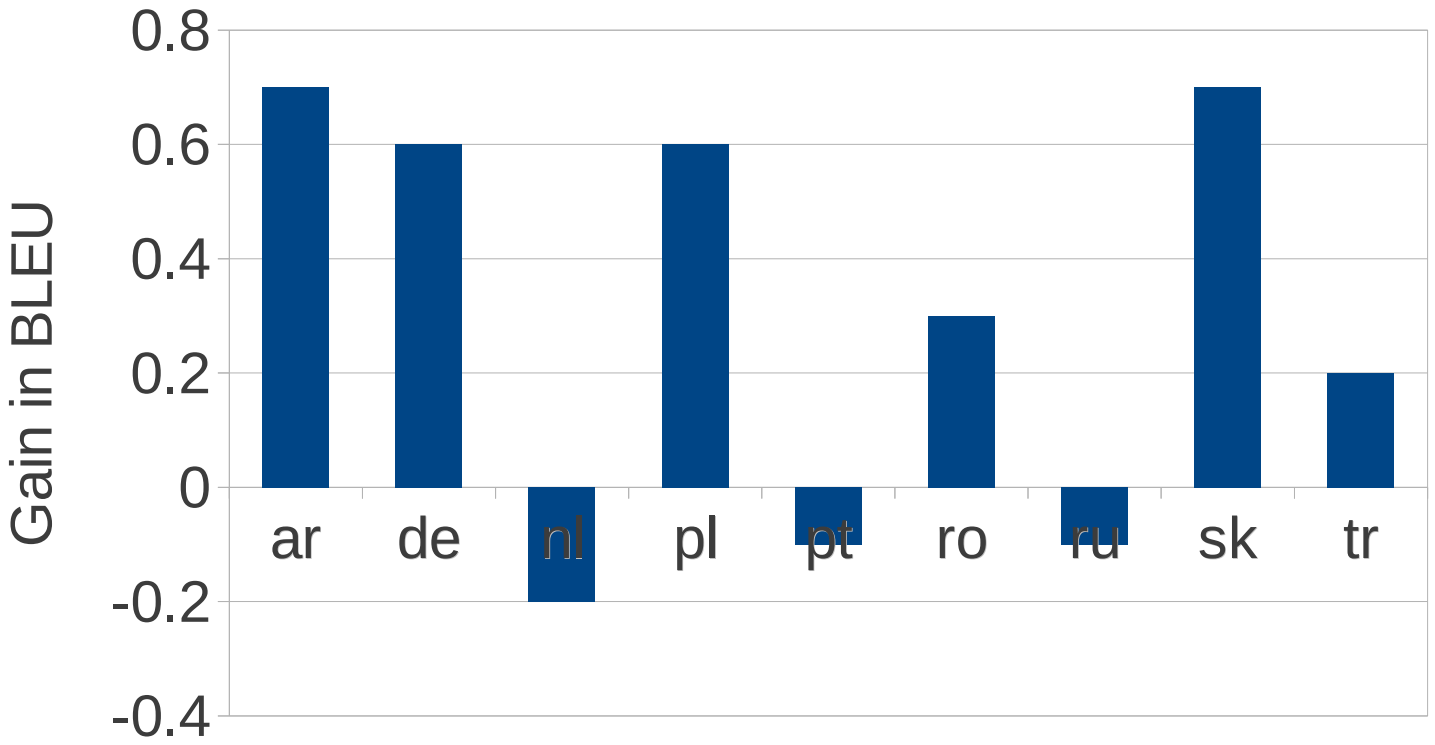


Morphology Results



Language Model Addition

- Added additional Giga-Word language model



Other Methods Investigated

- Out of domain TM data
- Lattice-based MBR

See the paper for more details!

Thank You!