The TÜBİTAK-UEKAE Statistical Machine Translation System for IWSLT 2010

Coşkun Mermer Hamza Kaya Mehmet Uğur Doğan
National Research Institute of Electronics and Cryptology (UEKAE) – TÜBİTAK, Turkey

www.multisaund.eu

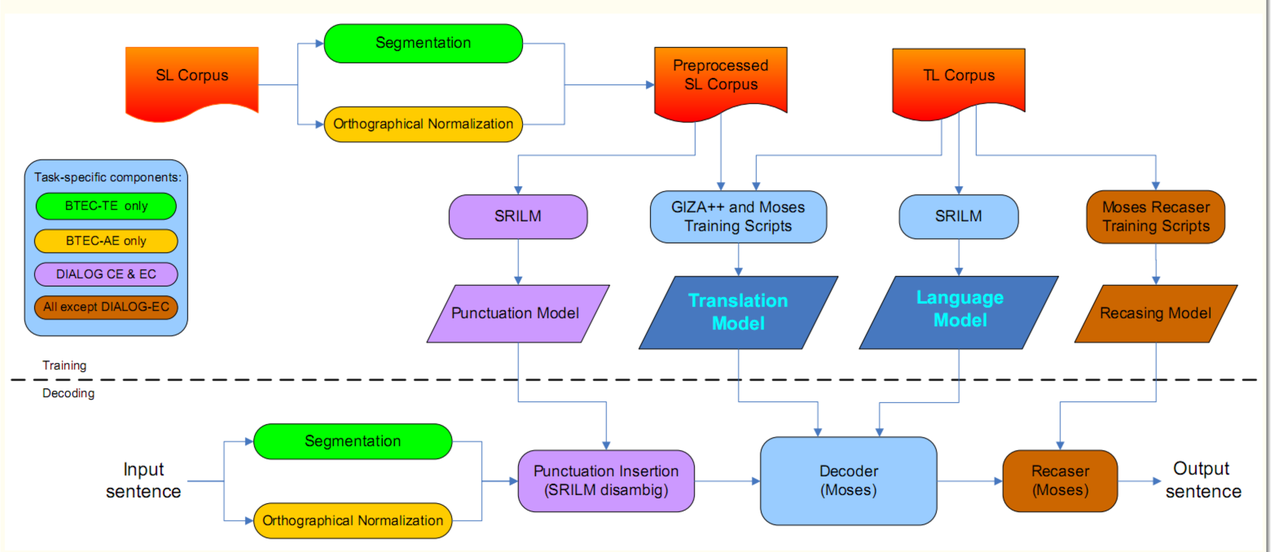
Introduction

- **Fourth** participation in IWSLT Evaluation Campaign (2007-2010)
- **Phrase-based SMT** system based on Moses (with GIZA++ & SRILM)
- Detailed investigation of segmentation methods for BTEC-TE task
  1. **Supervised**: Morphological analyzer (Oflazer) → Disambiguator (Sak) → Manually-crafted rules
  2. **Unsupervised**: Optimization of a probabilistic segmentation model (Morfessor)

What's New in 2010

- Minimum error rate training (via ZMERT)
- Hierarchical phrase-based translation (via Joshua)
- New techniques for unsupervised segmentation (via extensions to Morfessor)
- Participation in new tasks (DIALOG and TALK)

System Architecture



Morphology Problem in Turkish-English

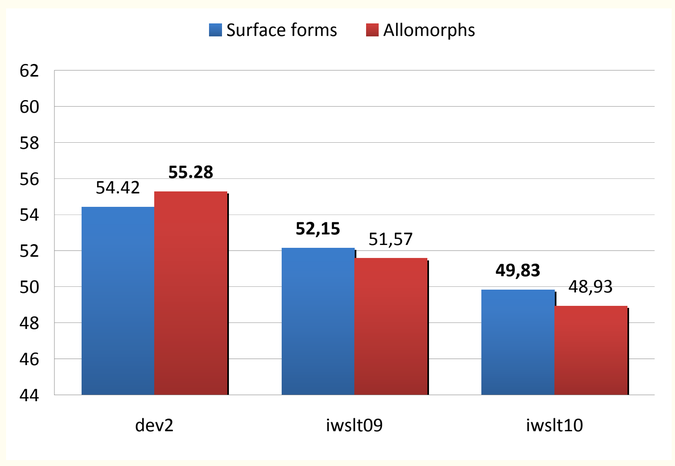
Morphological segmentation:

- Enables more accurate alignment
- Reduces data sparsity
- E.g., *yapamayacaksan* ↔ 'if you will not be able to do'

yap	+a	+ma	+yacak	+sa	+n
do	be able to	not	will	if	you

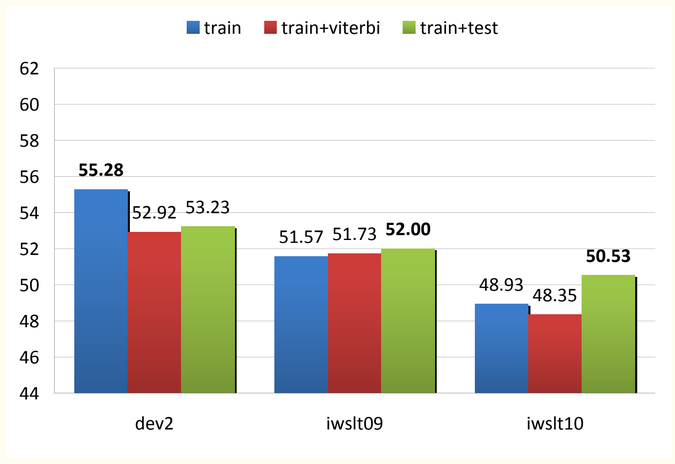
Allomorphs

- In Turkish, the same lexical morpheme can take on several surface forms depending on vowel harmony etc.
- E.g., {+dı,+di,+du,+dü,+tı,+ti,+tu,+tü} → +DH
- Map the most frequent allomorphic letters to their base letter
  - Applied only to suffixes in the segmented corpus
  - E.g., {d,t} → D, {ı,i,u,ü} → H



Segmentation Training Method

- What to do with test words unseen in training?
- Viterbi segmentation: Most likely segmentation of test words given model
  - Drawback: Forces segmentation, sometimes down to letters
- Including test set in training: Test word segmentations learned jointly with training words
  - Drawback: Not practical for "on-line" translation



Morfessor

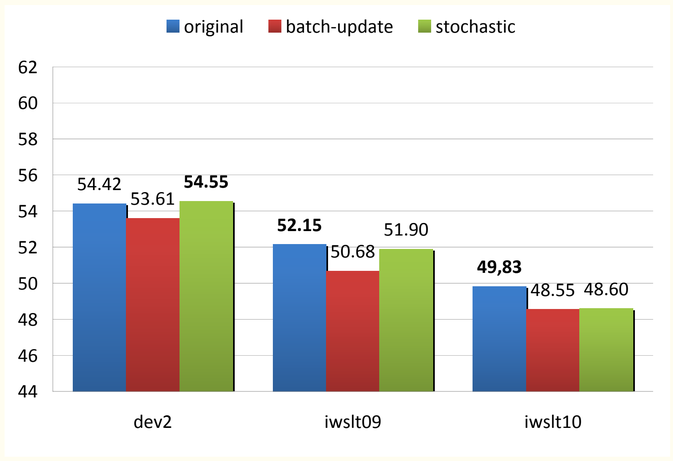
- Unsupervised segmentation (Creutz and Lagus, 2007)
- Prior over segmentation models, MAP search
- Equivalent to a minimum-description length (MDL) problem

$$\arg\max_M P(M|corpus) = \arg\max_M P(M)P(corpus|M)$$

corpus: Monolingual corpus to be segmented
M: Segmentation model

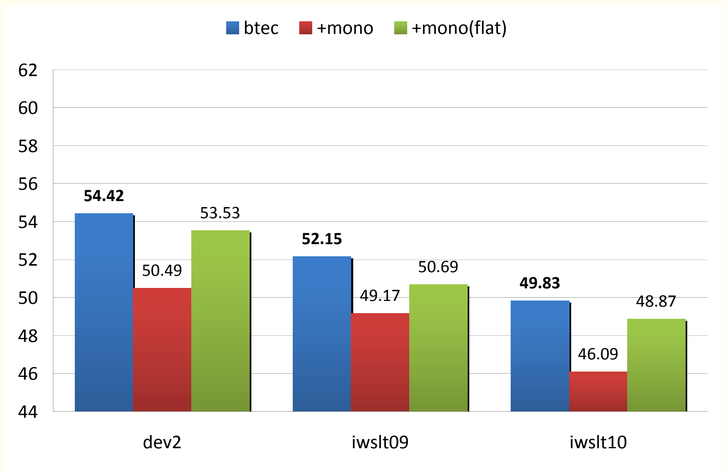
Segmentation Search

- Batch-update: Segmentation decisions can be calculated in parallel
- Gibbs: Instead of greedy decisions, randomly sample from the individual segmentation posteriors



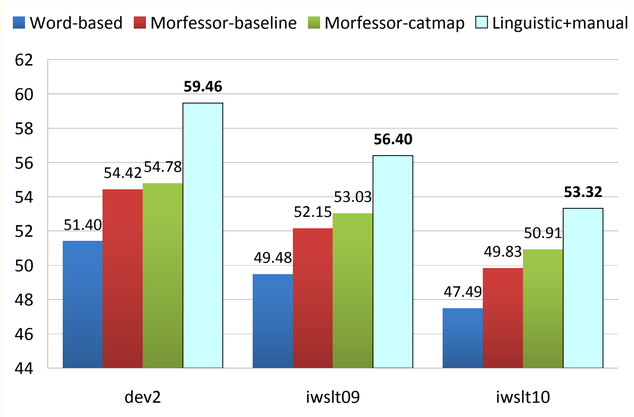
Out-of-Domain Segmentation Training

- Use large **monolingual** corpus
  - News domain, ∼400 M tokens & ∼500 K types
- Segmentation training on tokens (regular) vs. on types ("flat" vocabulary)
  - Regular: Cost of splitting a word type increases with its occurrence count
  - Flat: All word types are assumed to have a frequency of 1



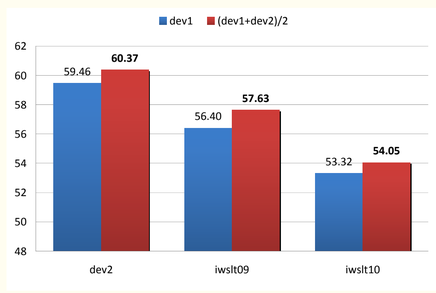
Summary of Segmentation Performances

- Comparison of different segmentation methods considered for IWSLT 2010
- Morfessor-catmap (Categories-MAP): An HMM model of transitions between hidden class assignments of segmented morphs
  - Prefix, Stem, Suffix



Tuning Method

- How to decide between *dev1* and *dev2* for tuning/testing?
- Tune separately on each, then average the weigth vectors



Conclusions

- Turkish-English system consistently **ranked #1** in METEOR and NIST scores (#2 in BLEU and TER scores) in IWSLT 2010 evaluations.
- Several innovative methods proposed with the goal of improving the **segmentation learning** in BTEC Turkish-English task, though most of them did not improve the BLEU scores.
- **Supervised segmentation** still yielded superior translation performance compared to the unsupervised methods we tried.

Acknowledgement

This work was supported in part by **MULTISAUND** (FP7-REGPOT-2008-1 project number 229861).