Main Contributions

- HTCE Arabic-English and Turkish-English:
  - Special effort on **linguistic preprocessing** for morphologically rich source languages
  - In particular word segmentation and lexical approximation techniques
  - Dealing with mismatch in word granularity between source languages and English
- CT English-Chinese and Chinese-English:
  - Focus on **language model adaptation**
  - Mixture of in-domain language models, obtained by clustering training data
  - Weight estimation at the level of single source sentence or complete test set

Linguistic Pre-Processing for Morphologically Rich Languages

- Morphological segmentation of Turkish:
  - Word harmony (a rich phonological phenomenon)
  - Syntactic stem and suffix allomorphy
  - Agglutinative language
  - Huge variety of possible segmentation schemes

  - Tag notation abstracts from suffix allomorphy: Example: 
    - `future in the morning` → `morgen geben` → `morgen geben + future`
    - Our best segmentation scheme: MMT builds nominal case, possessive, copula and verb person suffixes

- Morphological segmentation of Arabic:
  - Specific tokenization (for TAs): removal of short words and normalization of UTF-8 characters and digits
  - Comparison of two-state-of-the-art segmenters: MADA and AMIRA

Baseline

- Example: had she tell us it to her colleague

  - Baseline: length + (new words + inflected words) * (in-domain / out-of-domain)

AMIRA

- Character/segment context

Lexical approximation:

  - Replace OOV words in the test with morphologically similar words of the training
  - Deterministic choice of 2-best replacement
  - Turkish: choose word sharing lemma and longest number of suffix tags

Example: 

  - `yap tamamla verdi kesin` → `yap tamamla verdi kesin`
  - Arabic: progressively remove prefix and suffixes from the OOV word until a replace is found

Example: 

  - `kustar` → `kustar` + `kustar` + `kustar`

Online Language Model Adaptation for Spoken Dialog Translation

- Model adaptation

  - LM score is given by either single LM (baseline) or mixture of (small) LMs:

    \[ p(e) = \sum_{j=1}^{L} \alpha_j \cdot \log(p_{e_j}(e)) \]

- Clustering using dialog annotations:

  - Each dialog is represented as a bag of both source and target words
  - CLUTO package was employed directly clustering, cosine-distance
  - 2, 4, 6 and 8 clusters
  - One set of LMs for each cluster + additional LM on HTCE+CT data

- On-line weight optimization:

  - Set specific weights (one complete source side of test set)
  - Sentence specific weights (one set of weights for each source sentence)
  - Two-step weight optimization: See figure

Evaluation Results

- **Baseline:** standard setup for Moses SMT toolkit

- **HTCE Arabic-English**:

  - Best segment scheme (M55) dramatically lowers test’s OOV, minimizes differences in word granularity between TR and EN, reduces training dictionary size and data sparseness
  - MER on devset using gold reference only
  - Distortion limit (DL) set to 10, due to high word order mismatch
  - Morph segmentation yields 5 points BLEU improvement
  - Lexical approach does not improve in -step-unbound conditions
  - Unlimited distortion results inconsistent across test sets

- **CT English-Chinese**:

  - Development set of CT task used for MER T, then included into training corpus
  - Best segment scheme yields 5 points BLEU improvement on devset1 using gold reference only
  - MER T on devset1 using gold reference only
  - Development set of previous campaigns not included, only their vocabulary
  - Morph segmentation yields 5 points BLEU improvement on devset2
  - AMIRA results inconsistent across test sets
  - Lexical approach also disappoints: improvement only on the official test

Summary and Future Work

- **Specific linguistic preprocessing** is crucial for morphologically rich languages
- **TDTD:** refine our Turkish segmentation schemes by addressing verbal suffusion in a better way
- **TDO3:** feed Moses with multiple options for lexical approximation
  - Adaptation yields limited gains in BLEU
  - Observed big gains in perplexity → room for improvement
  - TDO3: address larger tasks, involving unsupervised clustering and source-to-target weight map

Acknowledgements

This work was supported by the EuroMatrixPlus project (IST-231720), which is funded by the European Commission under the Seventh Framework Programme for Research and Technological Development, and by the Spanish MECD under scholarship AP2005-4023 and grant CONSOLIDER Ingenio-2010 CS02007-00018.