**Main Contributions**

- **BTTC 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 phrase language models, obtained by clustering training data
  - Mixture weight estimation at the level of single source sentence or complete test set

**Linguistic Pre-Processing for Morphologically Rich Languages**

- Morphological segmentation of Arabic:
  - Mixture of 2-gram language models, obtained by clustering training data
  - Mixture weight estimation at the level of single source sentence or complete test set

- Morphological segmentation of Turkish:
  - word harmony (i.e. other phonological phenomena)
  - systematic stem and suffix allomorphy
  - agglutinative language
  - large variety of possible segmentation schemes
  - tag notation abstracts from suffix allomorphy

- Morphological segmentation of Arabic:
  - specific tokenization (e.g. for Arabic-to-English translation)
  - removal of short words and normalization of UTF-8 characters and digits
  - comparison of two-state-of-the-art segmenters: MADA and AMIRA

- Baseline:
  - AR (2006)
  - Turkish: 48.6/48.3
  - Arabic: 30.1

- Lexical approximation:
  - replace OOV words in the test with morphologically similar words of the training
  - automatic choice of the best replacement
  - Turkish: choose word sharing lemma and largest number of suffix tags
  - Arabic: progressively remove prefixes and suffixes from the OOV word until a replace is found

- Example:
  - Arabic: "..."..."..."
  - Turkish: "..."..."..."

**Online Language Model Adaptation for Spoken Dialog Translation**

- Model adaptation
  - LM score is given by either single LM (baselines) or mixture of (smaller) LMs: $p(e) = \sum w_i p_i(e)$

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

- On-line weight optimization:
  - Set specific weights (over complete source side of test set)
  - Set specific weights (over complete source side of test set)
  - Two-step weight optimization: See figure

**Evaluation Results**

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

- **BTTC Arabic-English:**
  - Best segment scheme (MADH) dramatically lowers test's OOV, minimizes differences in word granularity between TR and EN, reduces training dictionary size and data sparseness.
  - MADA on test with gold reference only
  - Distortion limit (DL) set to 10, due to high word order mismatch
  - Morphological segmentation yields 5 points BLEU improvement
  - Lexical approach: does not improve above-unseen conditions
  - Unlimited distortion results inconsistent across test sets

- **BTTC English-Arabic:**
  - Training data: train + dev2 2 and 3 (with gold reference only)
  - MERT on dev3 using all references
  - Specific tokenization alone yields around half point BLEU improvement (51.36 to 51.75 on test)
  - Morphological segmentation through MADA yields additional 2.3 points on dev3, but only 0.5 on test
  - AMRA results inconsistent across test sets
  - Lexical approach also does not improve: improvement only on the official test

- **CT English-Chinese:**
  - Development set of CT task used for MERT, then included into training corpus
  - Development sets of previous campaigns not included, only their vocabulary
  - Improvements in terms of perplexity are only partially mirrored into translation quality
  - Primary run: six dialog clusters, 24-top weight optimization

- **CT Chinese-English:**
  - Same setup as for English-to-Chinese

**Summary and Future Work**

- Specific linguistic preprocessing is crucial for morphologically rich languages
- **TOPS:** refine our Turkish segmentation schemes by addressing verbal suffixation in a better way
- **TOPS:** feed Moses with multiple options for lexical approximation

**Acknowledgements**

This work was supported by the EuroMatrixPlus project (IST-21720), which is funded by the European Commission under the Seventh Framework Programme for Research and Technological Development, and by the Spanish MURI under scholarship AP2005-4023 and grant CONSIDER Ingenio-2010 CSD2007-00018.