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

- Morphological segmentation of Turkish:
  - word harmony (as other phonological phenomena)
  - systematic stem and suffix allomorphy
  - agglutinative language
  - large variety of possible segmentation schemes

- Morphological segmentation of Arabic:
  - specific tokenization (for Talk) removal of short words and normalization of UTF8 characters and digits
  - comparison of two state-of-the-art segmenters: MADA and AMIRA

- Lexical approximation:
  - replace OOV words in the test with morphologically similar words of the training
  - deterministic choice of 2 best replacers
  - Turkish: choose word sharing lemma and longest number of suffix tags
  - Arabic: progressively remove prefixes and suffixes from the OOV word until a replacer is found

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

- Model adaptation
  - LM score is given by either single LM (baseline) or mixture of (small) LMs: \( p(e) = \sum w_n p_n(e) \)

- 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 HTK+CT data

- On-line weight optimization:
  - set specific weights (to 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
  - **Baseline** - BTEC Arabic-English
  - **Baseline** - CT English-Chinese

- **Comparison to state-of-the-art segmenters**:
  - MADA
  - AMIRA

- **BLEU and precision/recall**
  - CT English-Chinese results:
    - BLEU 51.36
    - Precision: 51.36

- **BLEU and distortion limit (DL)**
  - CT English-Chinese results:
    - BLEU 51.75
    - Precision: 51.75

**Summary and Future Work**

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

**Acknowledgements**

This work was supported by the EuroMatrixPlus project (ST2-24720), which is funded by the European Commission under the Seventh Framework Programme for Research and Technological Development, and by the Spanish MINECO under scholarship API2005-4023 and grant CONSOLIDER Ingenio-2010 CSD2007-00018.