Looking Inside the Box: Context-Sensitive Translation for CLIR

Ferhan Ture, Jimmy Lin, and Douglas W. Oard
University of Maryland
{fture|jimmylin|oard}@umd.edu

**State-of-the-art approaches in statistical MT**
- model translation with synchronous context-free grammars (SCFGs)
- allow reordering of non-consecutive phrases
- can handle 1-to-many and many-to-1 translations

**Decoding**
- searching for top N translations of a given sentence
- takes < 1 sec with modern decoders

\[ t^* = \text{argmax}_{t \in \text{Hyp}} \text{score}(t) \]

\[ \text{score}(t) = \lambda \text{score}(t_{\text{base}}) + (1-\lambda) \text{score}(t_{\text{interp}}) \]

**Example derivation**
\[ s = \{ \text{"maternal leave in Europe"} \} \]
- \[ R_1 : \{ X \} ||| \{ X \} ||| \{ X \} \]
- \[ \sum_{t \in \text{Hyp}} \text{score}(t) \]
- \[ R_2 : \{ X \} ||| \{ X \} \]
- \[ \text{score}(t) = \lambda \text{score}(t_{\text{base}}) + (1-\lambda) \text{score}(t_{\text{interp}}) \]

\[ t = \{ \text{"congé de maternité en Europe"} \} \]

**Our Approach**

- CLIR literature \( \rightarrow \) MT systems have been treated as a **"black box"** \( \rightarrow \) (Maged et al, 2011) show improvements when this is changed.
- Our approach \( \rightarrow \) Looking inside the box:

**Efficiency**
Processes in all CLIR models:
- word-aligning bitext to train the translation model
- initialization
- query generation
- document retrieval

For MT-based CLIR approaches:
- extra processes: grammar extraction + decoding
- compact representation \( \rightarrow \) retrieval takes less time

10-best MT approach saves over 40% time!

**Effectiveness**
Experimental results on TREC-S6 (54 English topics and 164,778 Chinese documents):

- Differences are not statistically significant.
- Results are consistent with theoretical expectations:
  10-best > 1-best
  10-best > bitext
  interp > 10-best

Grid search to find best-performing interpolation:
\[ \lambda^{*} = 0.85 \]
\[ \text{MAP} = 0.3404 \]

\[ \text{Pr}_{\text{interp}} > \text{Pr}_{\text{base}} \] in 36 out of 54 topics.