The MIT-LL/AFRL IWSLT-2009 MT System

Wade Shen, Brian Delaney, A. Ryan Aminzadeh
Tim Anderson and Ray Slyh
2 December 2009

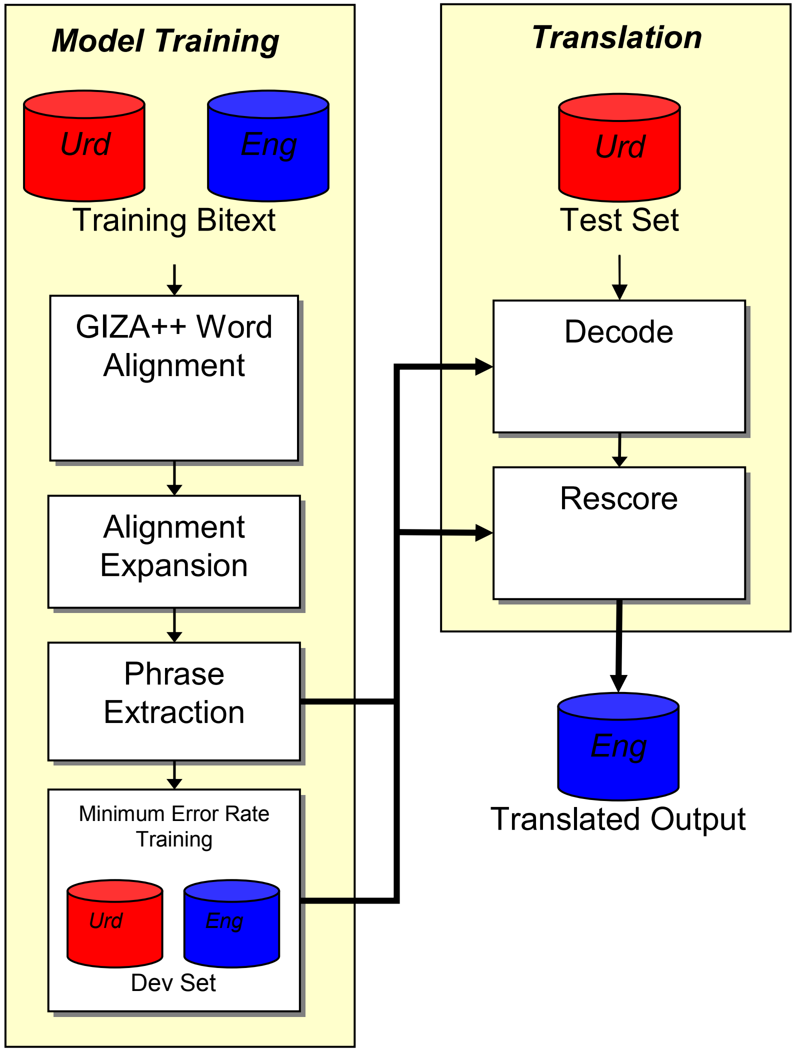
This work is sponsored by the United States Air Force Research Laboratory under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government.

Outline

- **IWSLT-2009 System Architecture**
- Better Arabic Morphology Processing
  - CoMMA
- Domain Adaptation Overview
  - Unsupervised and Semi-supervised Adaptation
  - Human-in-the Loop Adaptation

Statistical Translation System
Experimental Architecture

- **Standard Statistical Architecture**
- **Developed in-house to support SMT experiments**
  - **Framework for experiments with low-resource languages**
  - **Test-bed for S2S MT system**
- **Most components are home-grown**
  - **Phrase Training/Minimum Error Rate Training**
  - **Moses and FST decoders used, comparable performance**
- **Participated in Arabic/Turkish ⇨ English BTEC Data track**



Phrase Based FST Decoder

- Based on MIT FST toolkit: http://people.csail.mit.edu/ilh/fst/

- The target language hypothesis is the best path through the following transducer:

$$E = I \circ P \circ D \circ T \circ L$$

- where,
  - I = source language input acceptor
  - P = phrase segmentation transducer
  - D = weighted phrase swapping transducer
  - T = weighted phrase translation transducer (source phrases to target words)
  - L = weighted target language model acceptor

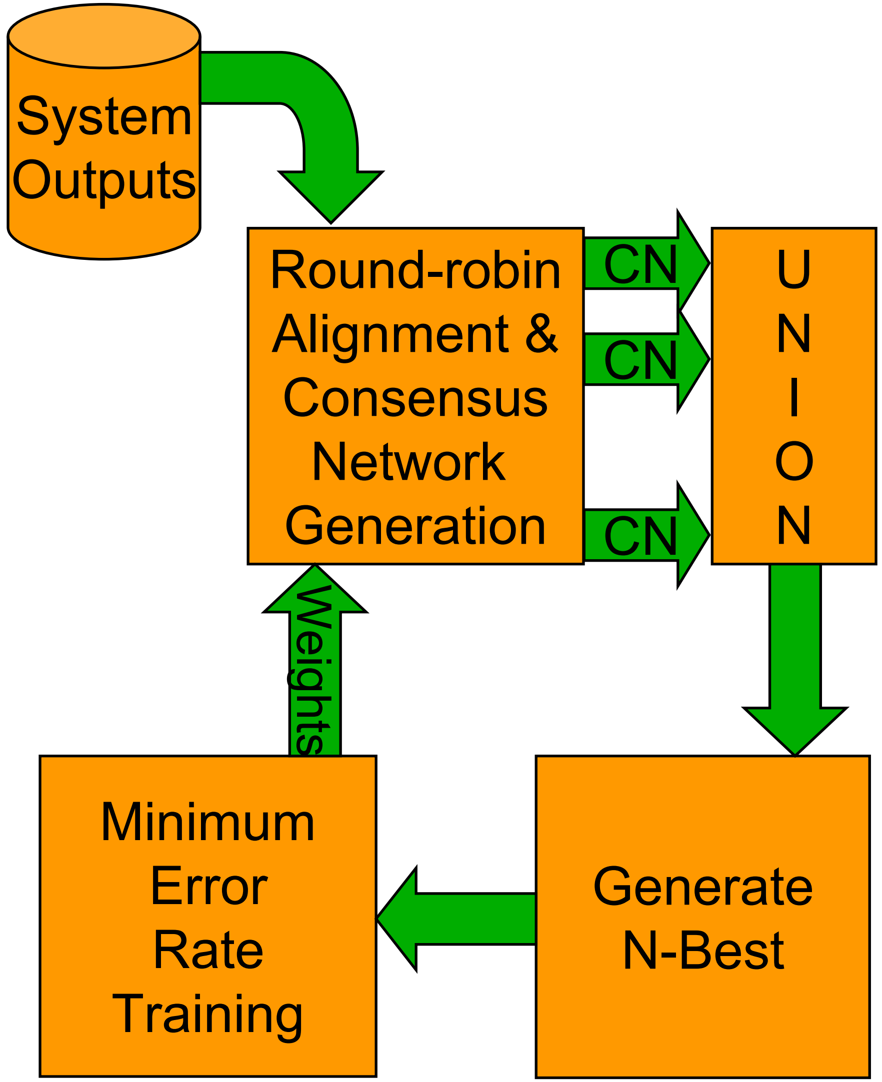
- Apply phrase swapping twice for long distance reordering

- OOV words are inserted during decoding as parallel links to P, D, T, and L models.

- Allows for direct decoding on pruned ASR lattices

System Combination

- Generate consensus networks using round-robin alignment, where each system gets to be the skeleton alignment
- Take union of all consensus networks and apply a language model
- Weight optimization on a development set using n-best lists
- Final combination on unseen data using optimized system weights



Outline

- IWSLT-2009 System Architecture
- Better Arabic Morphology Processing
  - CoMMA
- Domain Adaptation Overview
  - Unsupervised and Semi-supervised Adaptation
  - Human-in-the Loop Adaptation

Arabic Preprocessing

AP5 Review

Preprocessing Method	Mean BLEU on dev6
Baseline (No normalization or AP5)	42.06
Remove all diacritics except tanween, no AP5	49.40
Remove all diacritics, no AP5	50.39
Remove all diacritics, apply AP5	53.55

- “Diacritics” removed:
  - Short vowels
  - **Sukuun:** Marks absence of sort vowel
  - **Shadda:** Marks consonant gemination (i.e., doubling)
  - **Tanween:** Case markers for indefinite forms & other uses
  - **Tatweel:** Stretches letters in Arabic typography (not a true diacritic)
- AP5 segments the following from stems:
  - **Prefixes:** al-, bi-, fa-, ka-, li-, wa-
  - **Suffixes:** Attached pronouns

CoMMA Processing for Arabic

- **Observation:** *With limited training data more morphological processing seems to help, less with more training data*

- **Count Mediated Morphological Analysis**
  - **Modification to AP5: decide segmentation based on counts**

- **Given a count threshold t, and a vocabulary W**

- **Foreach w in |W|**
  - **Apply AP5 diacritic normalization procedure**
  - **If count(w) < t**

    Apply AP5 segmentation of clitics, etc.
  - **Else don't segment**

CoMMA Experiments

Baseline (No Tokenization)

COMMA Threshold	BLEU Score	
	Dev6	Dev7
0	50.00	51.94
20	53.92	54.29
200	53.14	54.64
2,000	54.02	54.57
10,000	53.33	54.48

CoMMa

AP5 (all tokens segmented)

- AP5 and CoMMa results in 7-8% relative improvement
- CoMMa only slightly better than AP5, +0.5–1.5 BLEU in system combination

Outline

- IWSLT-2009 System Architecture
- Better Arabic Morphology Processing
  - CoMMA
- **Domain Adaptation Overview**
  - **Unsupervised and Semi-supervised Adaptation**
  - **Human-in-the Loop Adaptation**

Cross Domain Adaptation Overview

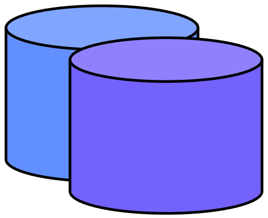
- **Observations from past work**
  - **SMT performs best when training and test data are matched**
  - **Adding large volumes of out-of-domain data to training does not improve performance**

- **Adaptation**
  - **GOAL:** *Optimally port general purpose (out-of-domain) models to specific domain with limited in-domain data*

GP Model

+

Adaptation Data



→

Adapted Model

- **NOTE:** *Adapted Systems not used in IWSLT BTEC submissions*

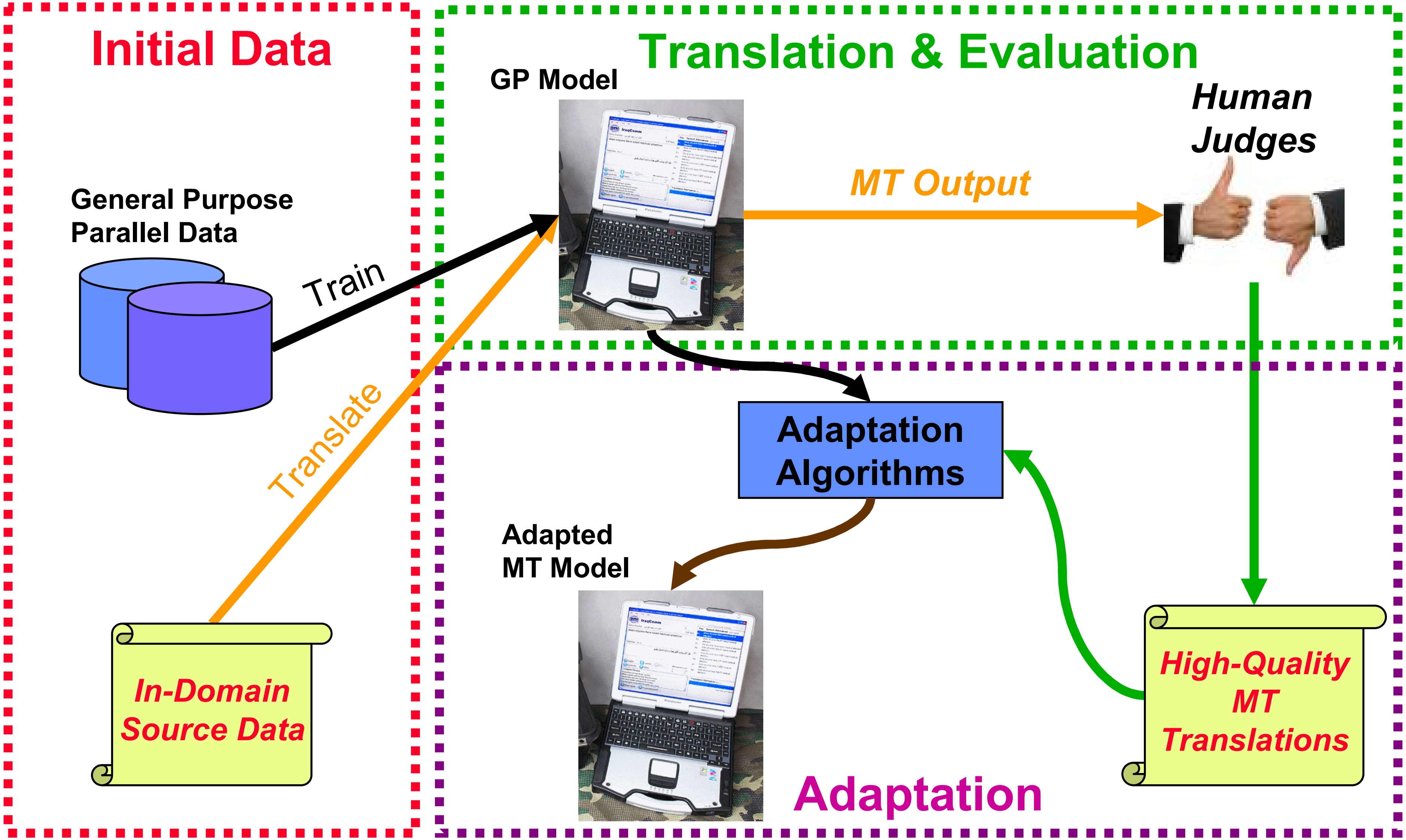
Data

- General purpose data:
  - 500k Arabic-English parallel data from ISI automatically extracted parallel corpus
  - Domain: newswire data

- In-domain (adaptation) data:
  - 20k IWSLT-2009 BTEC Arabic-English training set
  - Domain: travel

Adaptation of Phrase-based MT Models

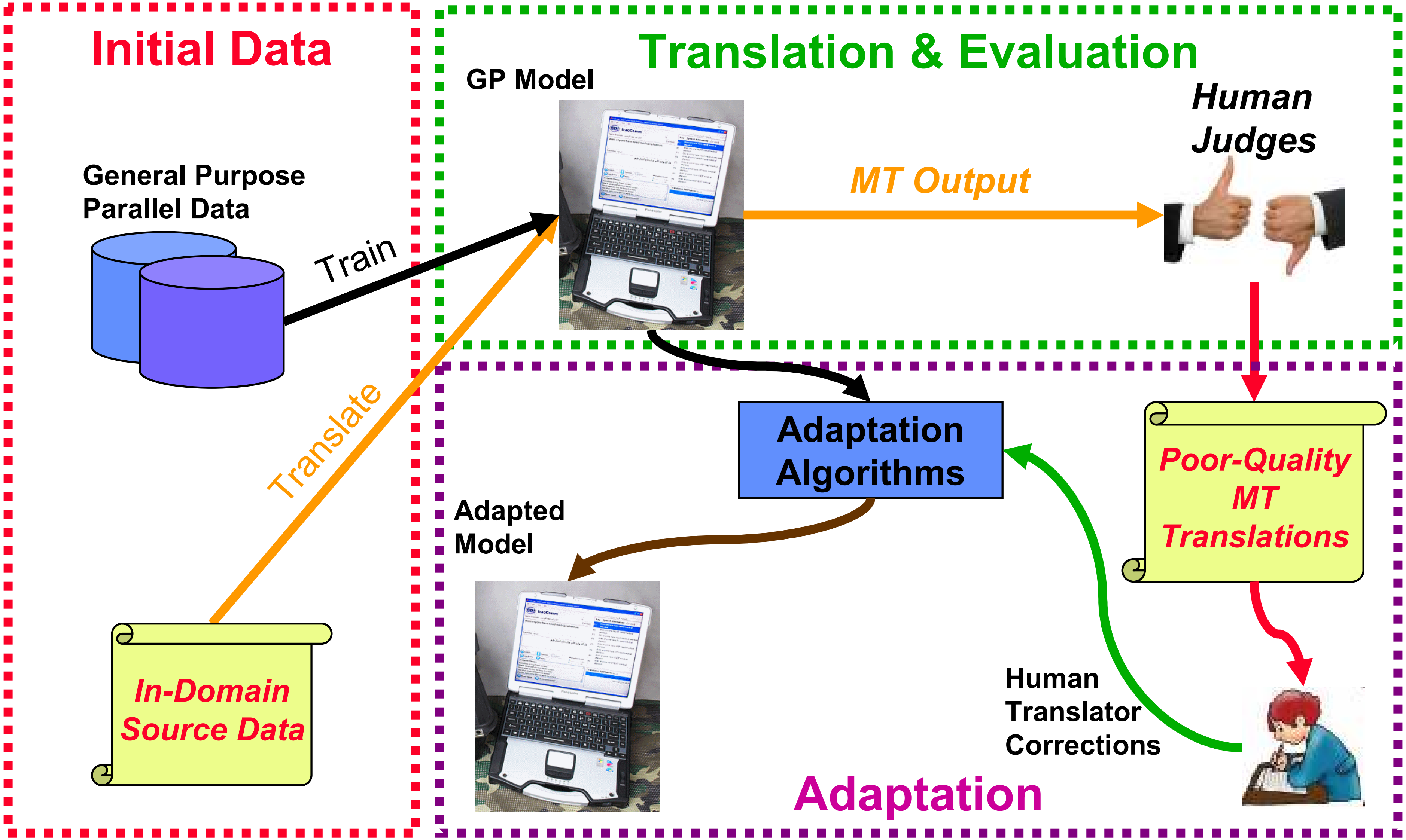
Semi-supervised



999999-13
TRA 12/2/2009

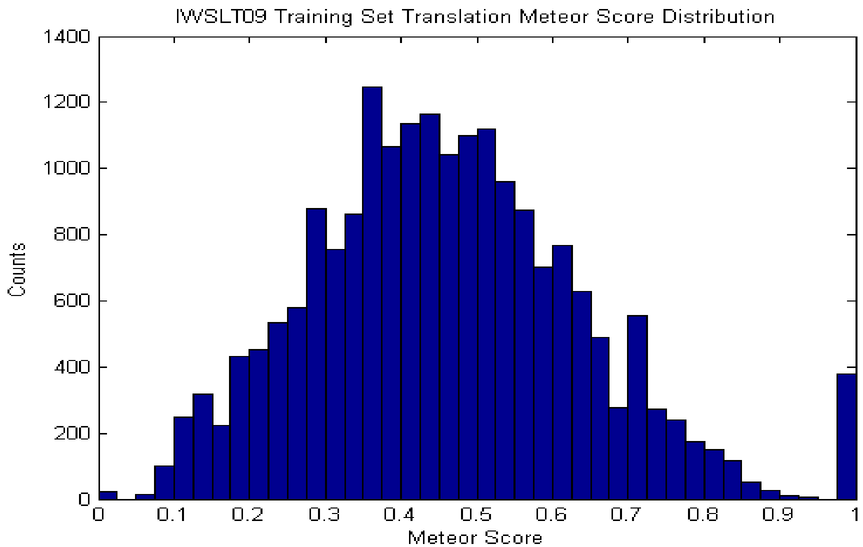
Adaptation of Phrase-based MT Models

Human-in-the-Loop



Selection of In-domain Adaptation Data

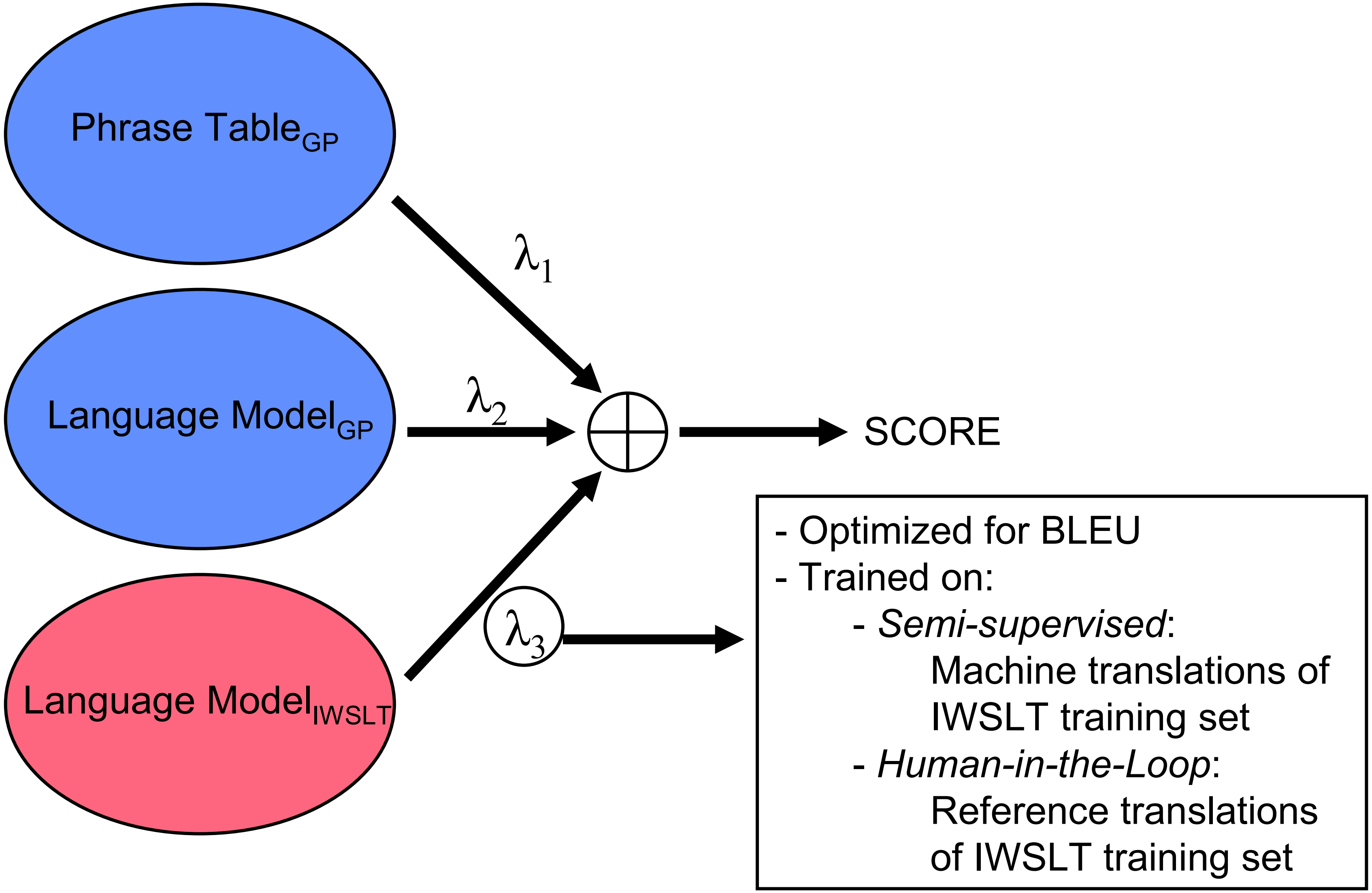
- **General purpose models used to translate the IWSLT '09 training set**
- **Translations ranked using METEOR as a proxy for a human judge**
- **Ranked sentences divided into octiles and used for experiments:**
  - **Semi-supervised adaptation:** *Use top scoring octiles for adaptation*
    Goal: is to use best in-domain target data
  - **Human-in-the-loop adaptation:** *Use bottom scoring octiles for adaptation*
    Goal: is to correct worst in-domain target data (active learning paradigm)



Octiles	1	2	3	4	5	6	7	8
METEOR	0.66	0.57	0.51	0.45	0.40	0.34	0.26	0.00

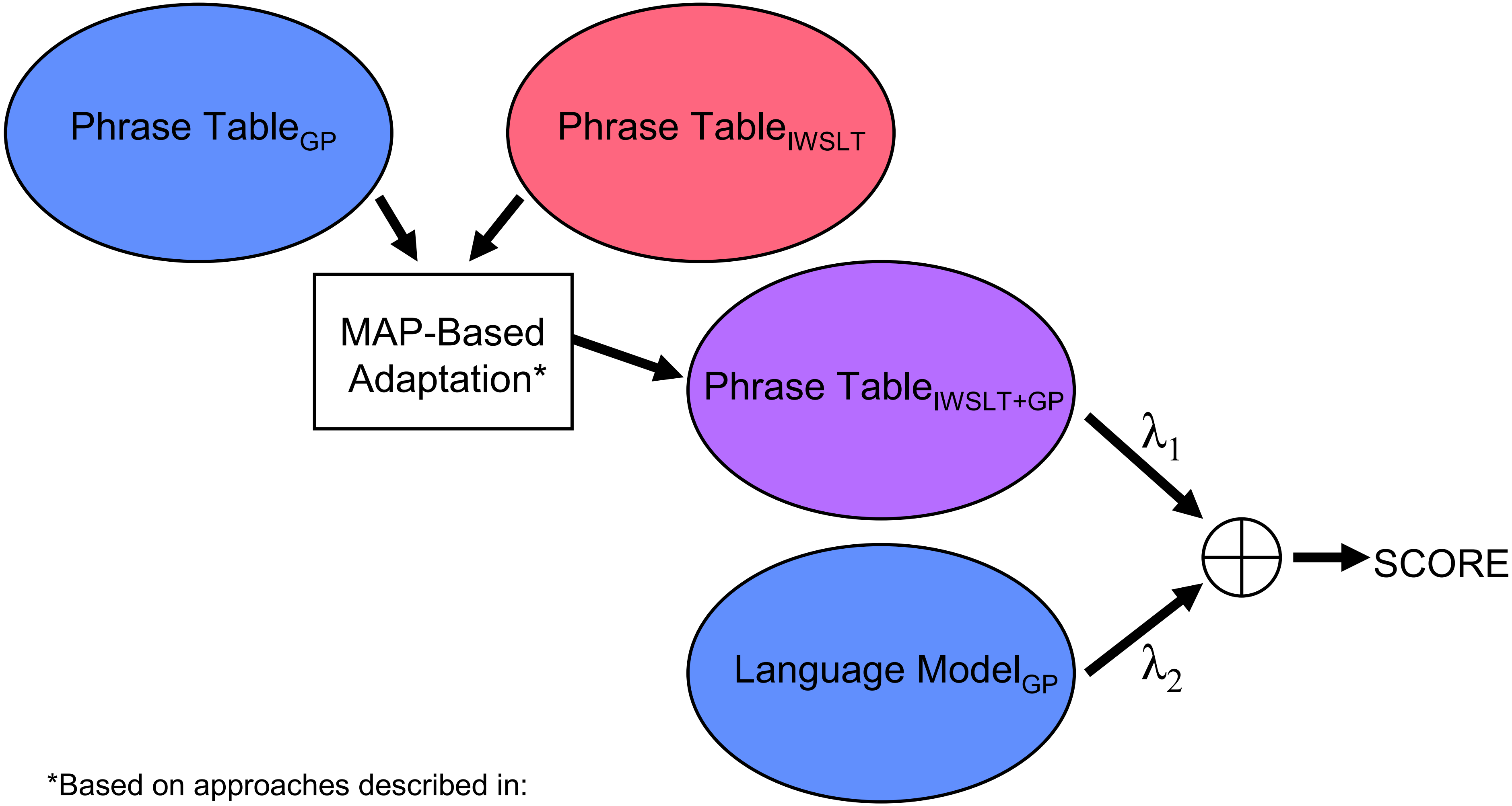
Adaptation Approaches

Language Model Adaptation



Adaptation Approaches

Phrase Table Adaptation



*Based on approaches described in:

[1] C. Lee & J. Gauvain, "*Speaker adaptation based on MAP estimation of HMM parameters*," ICASSP 1993.
[2] M. Federico, "Bayesian estimation methods for n-gram language model adaptation", ICSLP 1996.
[3] M. Bacchiani and B. Roark, "*Unsupervised Language Model Adaptation*," ICASSP 2003.

Phrase Table MAP Adaptation

- Interpolated phrase table probabilities are computed using the following equation:

$$\hat{p}(s \mid t) = \lambda p_{in-domain}(s \mid t) + (1-\lambda)\, p_{gp}(s \mid t)$$

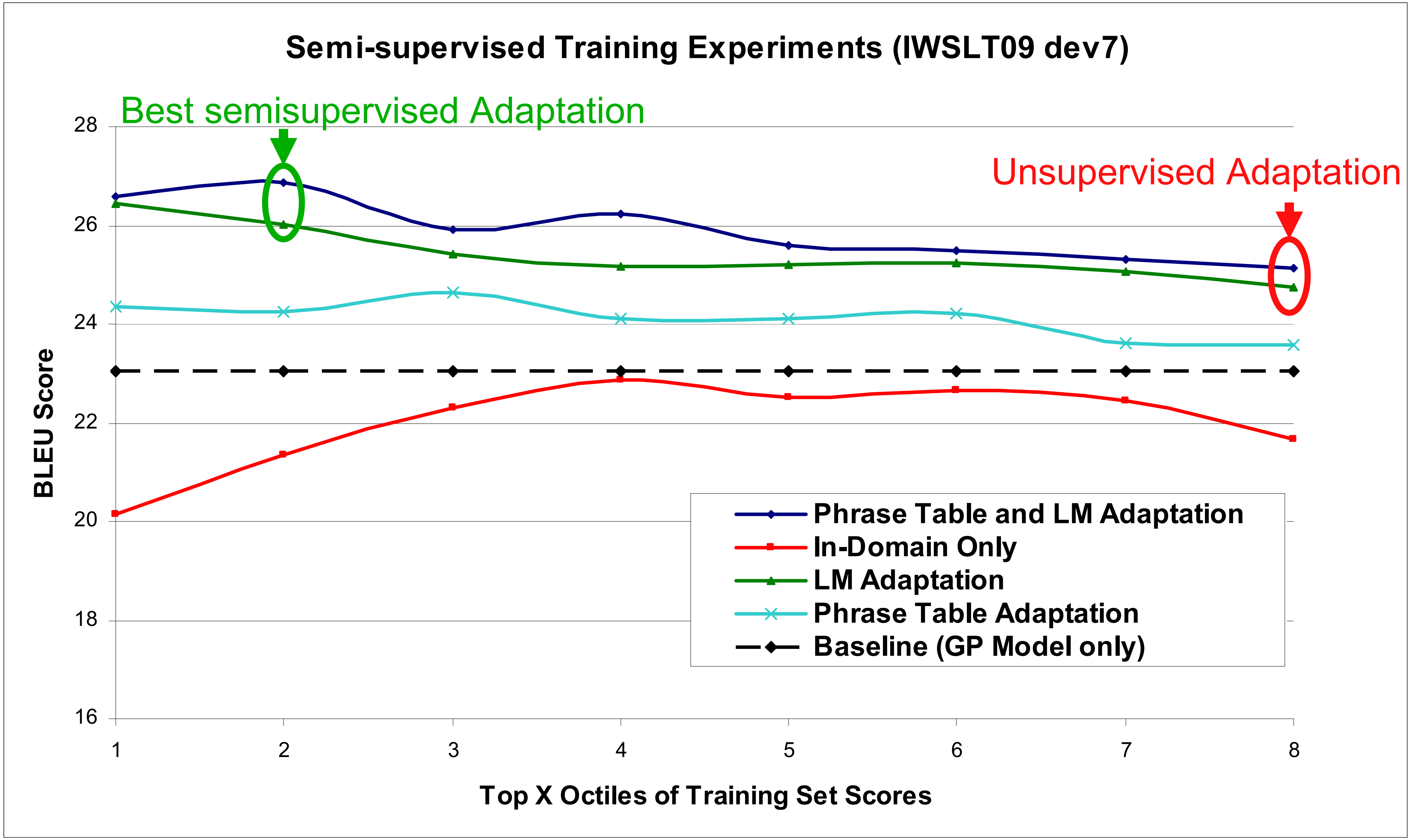
- $p_{in\text{-}domain}$: probability estimate from in-domain models
- p_{gp}: probability estimate from general purpose models
- λ: interpolation coefficient computed using the following equation:

$$\lambda = \frac{N_{in-domain}(s,t)}{N_{in-domain}(s,t) + \tau}$$

- τ: Fixed-value MAP relevance factor
- $N_{in\text{-}domain}$(s,t): observed count of phrase pair (s,t)

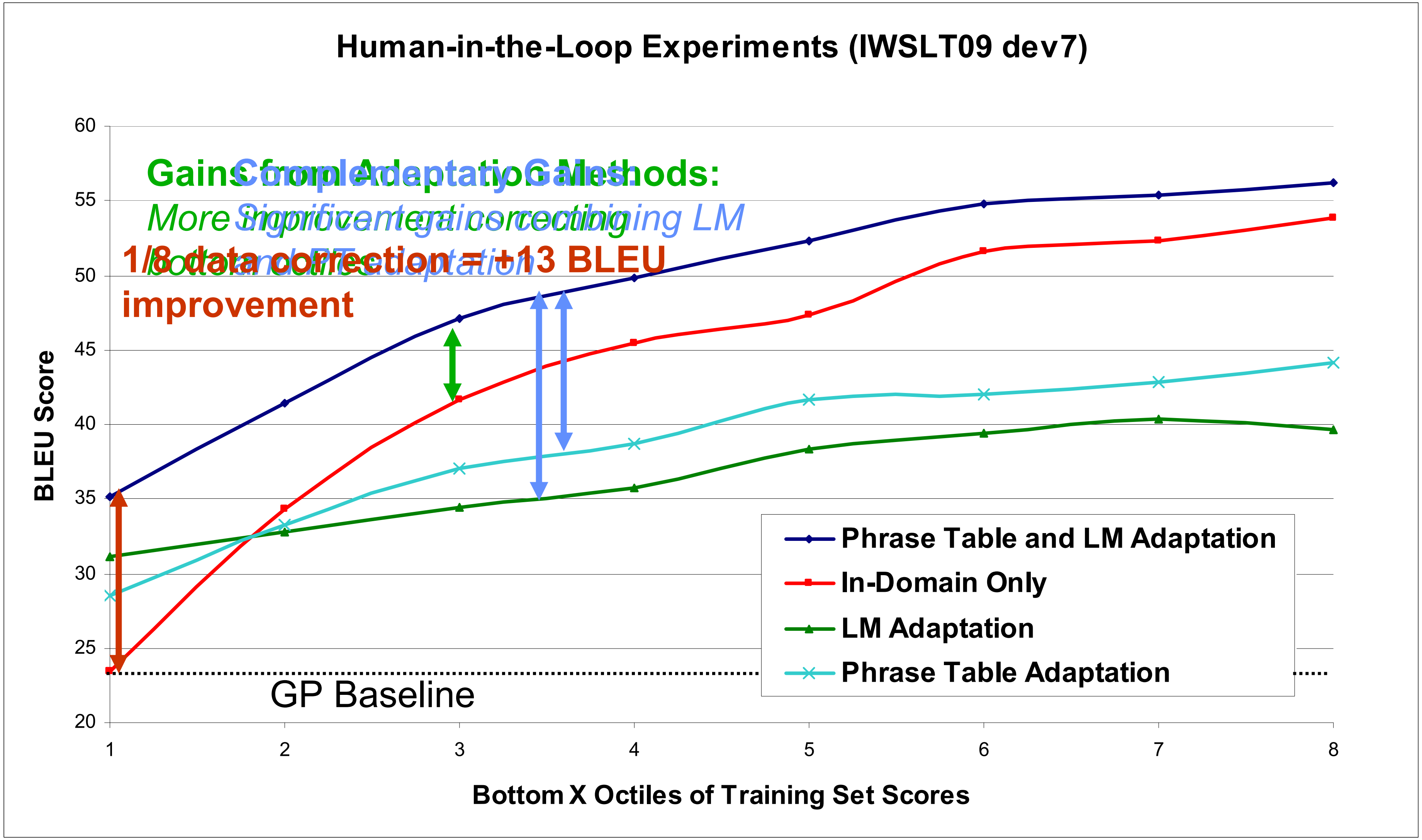
Experimental Results

Semi-supervised Adaptation



Experimental Results

Human-in-the-Loop Adaptation



Experimental Results

Best System Scores

System	dev7	eval
GP	23.06	21.35
GP + Unsupervised LM + PT Adaptation	25.74	23.86
GP + Semi-supervised LM + PT Adaptation (Top quartile)	27.19	25.89
IWSLT ’09 Baseline	54.63	52.69
GP + Human-in-the-Loop LM + PT Adaptation	56.57	56.11

Conclusions

- Morphological processing is critical
  - +4 BLEU for Turkish using Bilkent Analyzer
  - +3.5-4 BLEU for Arabic using AP5
- CoMMa gains in system combination
  - Multiple CoMMa systems (20, 200, 2000): +0.5-1.5 BLEU over AP5
- Unsupervised Adaptation
  - LM: +1.5 BLEU, PT: +0.5 BLEU
  - Combined: +2.5-3.0 BLEU (15% relative) compared to GP only
- Semi-supervised Adaptation
  - Gains +1.5-2 BLEU over Unsupervised, only ¼ of total data
  - But requires human judgement
- Human-in-the-Loop Adaptation
  - +2-3.5 BLEU using all IWSLT data
  - +13 BLEU using 1/8th of total data
  - Gains from LM and PT are non-additive