The MITLL/AFRL TC-Star System

Experiments with Large Vocabulary Speech Translation

Wade Shen, Brian Delaney, and Tim Anderson

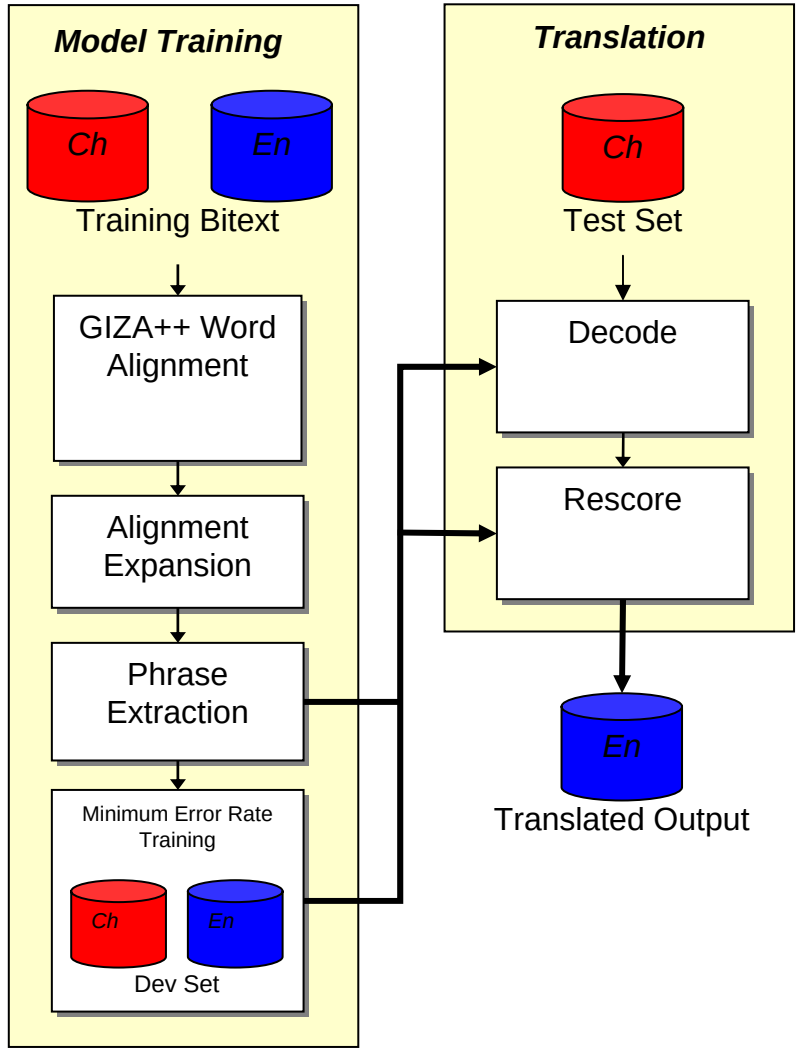
29 March 2006

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.

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**
  - **Internal Viterbi n-best decoder and Pharaoh used.**
- **Recent work**
  - **Scaling to large vocabulary**
  - **FST Decoder**



The MITLL/AFRL MT System

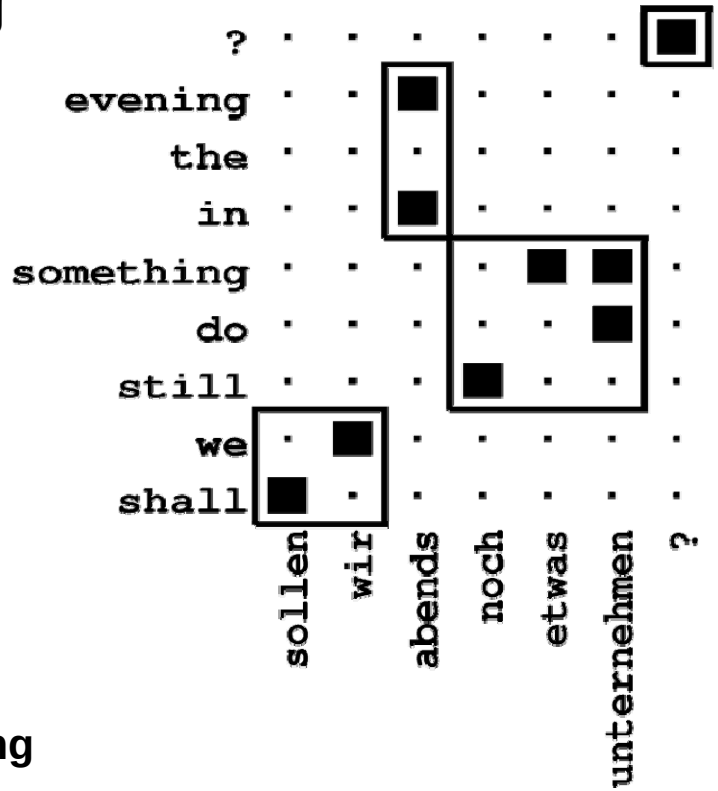
Overview

- Translation Model
- Minimum Error Rate Training
- Decoder
- Simple and Effective Truecasing
- Evaluation Results
  - Text Evaluation
  - ASR Evaluation
- Next Steps

Translation Model

Phrase Extraction

- **Basic Alignment Template Model Proposed by Och & Ney 2000**
  - **Expand word alignments interpolating between the intersection and union of bidirectional GIZA++ alignments**
  - **Extract consistent *phrase* pairs from expanded alignments**



- **Modifications**
  1. **Minor modifications result in +2 BLEU point gain**
  2. **Multiple parallel models:**
     1. **Tag-based PT**
     2. **Stemmed PT**
     3. **Can be combined in decoding/rescoring for optimization**

Translation Model

Distortion, Lexical and Language Models

- **Distortion**
  - **We used Pharaoh's simple model:**

$$P_D(e|f) = exp(-\sum_i |FinalW_{i-1} + 1 - FirstW_i|)$$

- **Lexical Weighting**
  - **Both model 4 and expanded alignment lexical translation models tried**
  - **Expanded alignments ⇨ 1.5 BLEU point gain**

- **Language Model**
  - **Trained with SRILM**
  - **Interpolated with Knesser-Ney discounting used for decoding**
  - **Bigram, 4-gram LM and 5-gram LM used during rescoring**
  - **2-8 gram LM also tried**
  - **No cleanup or normalization performed on base models**

Minimum Error Rate Training

- **Log-linear Model Combination**

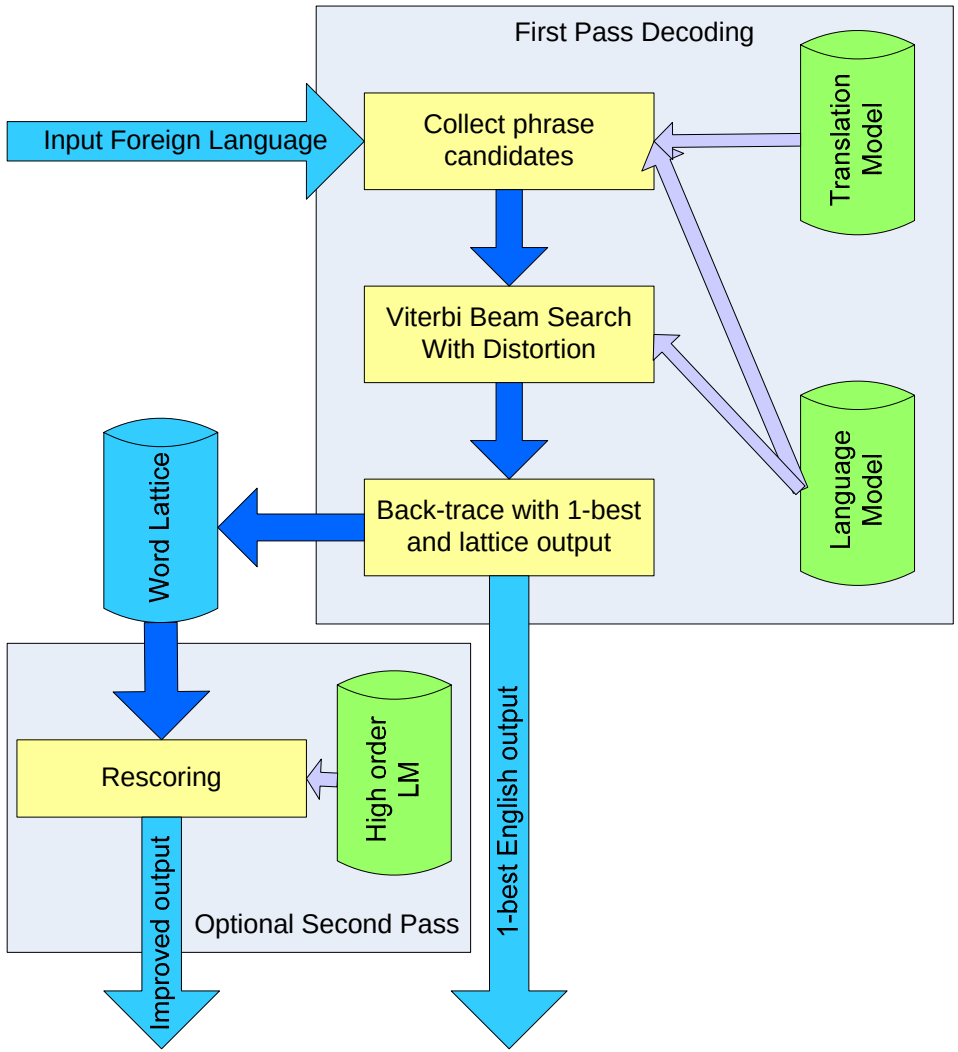
$$\hat{P}(e|f) = \frac{\sum_M \lambda_M h_M(e,f)}{\sum_{e'} exp(\sum_M \lambda_M h_M(e',f))}$$

- **Additional Language models applied during rescoring**
- **N-best lists of 2k used**
  - **Minor gain with 8k n-best**
- **15-20% relative improvement over hand optimized parameters**
- **Insignificant differences from beam-width relaxation**

Model Weight Parameters	
1	*P(f|e)* – Forward Translation Model
2	*P(e|f)* – Backward Translation Model
3	*LexW(f|e)* – Forward Lexical Weight
4	*LexW(e|f)* – Backward Lexical Weight
5	*PPen* – Constant, per-phrase Penalty
6	*WPen* – Constant, per-word Penalty
7	*Dist* – Distortion Model
8	Bi-LM – Bigram Language Model
9	*Tri-LM* – Trigram Language Model
10	*4-LM* – Four-gram Language Model
11	5-LM – Five-gram Language Model
12	*ClassLM* – Nine-gram class-based LM
13	*WordPost – Word Posterior*

Decoder Development

- **A phrase-based Viterbi beam search decoder has been implemented**
- **Decoder can account for word movement between source and target languages (*distortion*)**
  - **With distortion, search complexity approaches $O(2^n)$**
- **Decoding speed:**
  - **Monotone search (without distortion) can exceed 500 words per second**
  - **With distortion, search slows to 10 words per second but can be improved with limits on distortion**
- **Decoder can produce word lattice output for optional second pass rescoring with higher order language models**



TrueCasing

Problem

- **Reduced vocabulary size of lower cased training data often results in better models**
- **However, lower cased output has reduced readability and may affect other down stream processing such as named entity detection**
- **TrueCasing is the process of reintroducing case information to the translated output**
- **Example:**
  - **the it department at intel is hiring cs majors.**
  - **The IT department at Intel is hiring CS majors.**

TrueCasing
Design (cont)

- **Problem can be solved by searching through all possible capitalizations seen in the training set and choosing the most probable result**

$$\hat{W} = \arg\max_{W_i} P_{lm}(W_i) \prod_{w \in W_i} P_m(w_{uc}|w_{lc})$$

  - **P_{lm} is the case-sensitive language model probability of candidate word sequence W_i**
  - **P_m is the unigram probability of each possible case mapping given a lower cased input word**

- **Can be easily implemented using the SRILM toolkit**

TrueCased Output

Results

BLEU scores with lower and mixed case references

Input \ Reference	Lower case	Mixed Case
Lower Case	54.45	40.65
Mixed Case		52.51

- **Used 5-gram LM and case mapping probabilities based on relative frequencies from training data**
- **Upper bound on performance is 45.04 when scoring against lower cased reference**
- **TrueCasing the MT output increases the BLEU score by more than 10 points when scored against the mixed case reference**

Text Evaluation Results

What Worked

Parameters Varied	
Minimum Error Rate Training	Additional Language Models (4-gram and 5-gram)
Truecasing	Word Posteriors

Configurations	BLEU
Base: UTF-8, no MER, trigram LM	*46.03*
+ Additional LMs	47.00
+ Additional LMs + MER + Default Tuning	54.07
+ Additional LMs + Word Posteriors + MER	54.23
+ Additional LMs + MER + Best Tuning	**54.45**
+ True-cased	52.51

Text Evaluation Results (Cont')

What Didn't Worked

Parameters Varied	
Stemmed LMs	Other LMs (bigram, 6-8 gram, skip LMs)
Class-based LMs	

Configurations	BLEU
Best Combination + MER	***54.45***
+ Other LMs + MER	54.20
+ Class LMs + MER	54.00
+ Stemmed LMs + MER	54.09
+ Skipped LMs + MER	53.85

ASR Transcription Results

- **ASR**
  - **Baseline Verbatim Results w/LIMSI ASR (9.8% WER): 45.10**
  - **Scored N-best using confusion net posteriors**
    - **Renormalized posterior mass to 1.0, ASR weight = 3.0**
  - **Compared 1-best vs. N-best**
    - ***Using Nbest ⇨ 2-4% relative improvement***
  - **Used system optimized with verbatim transcription and cleaned LMs**
  - **N-best and confusion network (FST-based) decoders were used**

N-best	Decoder Used	BLEU
1	Viterbi	40.33
1	FST	39.40
50 from conf-net	Viterbi	41.37
Full conf-net	FST	39.94

Continuing Experiments

Tag-based Translation Models

- **Motivation**
  - **Word-to-word translation cases generate poor output**
  - **Need “templates” in back-off conditions**
  - **Example:**

	Source	MT	Reference
Test Example	`el día 9 hemos celebrado igualmente`	`the | day | 9 we | held | equally`	`on the 9 we likewise celebrated`
Test Tag	**DA NC Z VAI VMP RB**	**DT NN CD PRP VBZ RB**	**IN DT CD PRP RB VBZ**

Can we learn this generalization of the example?

Continuing Experiments
Tag-based Translation Models II

- **Proposed method**
  - **Tag-based TMs decoded in parallel with text TMs**
  - **Related to Factored Translation Model Approach**
  - **Models combined in log-linear way (MER optimization)**

- **Status**
  - **Retagged train, dev and test data. Optimized tag models independent of text models.**
  - **Training on Openlab data tagged yields good patterns**

	MT	Reference
Text Translation	`The day 9 we held equally`	`on the 9 we likewise celebrated`
Text Tags	DT NN CD PRP VBZ RB	IN DT CD PRP RB VBZ
Tag Translation Result	IN DT CD PRP RB VBZ	

Summary

- The MIT/AFRL MT system is capable of state-of-the-art performance on a Spanish-English task in both ASR and manual transcription conditions
- Many in-house components were built, but we also rely on the existence of freely available components such as GIZA++ and SRILM to accelerate development
- Further research into error mitigation techniques for speech to speech machine translation is needed

Next Steps

- **ASR Lattice rescoring and joint optimization**
  - **Further FST Decoder engineering needed**
- **Decoder development and evaluation**
- **Further optimization to large vocabulary tasks**
  - **LM normalization**
  - **Specialized entity/number translation modules**
- **Hybrid Interlingual efforts with MIT/CSAIL**