Language Model Adaptation for Automatic Speech Recognition and Statistical Machine Translation

Woosung Kim
Advisor: Prof. Sanjeev Khudanpur

Center for Language and Speech Processing
Dept. of Computer Science
The Johns Hopkins University

August 12, 2004

woosung@cs.jhu.edu
Outline

1. Introduction
2. Basics of Language Modeling
3. Language Model Adaptation
4. Cross-Lingual LM for Automatic Speech Recognition
5. Cross-Lingual Lexical Triggers
6. Cross-Lingual Latent Semantic Analysis
7. LM Adaptation for Statistical Machine Translation
8. Conclusions
Natural Language Applications

- Cross-Lingual IR/Adaptive MT
- Information Retrieval
- Machine Translation
- Speech Recognition
- Adaptive SR/Adaptive SR
- Speech to Speech Translation
Success of statistical modeling techniques: under two conditions

1. A reasonable size of training data exists
2. The training data comes from the same population as the test data

Ex1: How to construct stochastic models in resource-deficient languages? ➔ Bootstrap from other languages, e.g.
   - Universal phone-set for ASR (Schultz & Waibel, 98, Byrne et al, 00)
   - Exploit parallel texts to project morphological analyzers, POS taggers, etc. (Yarowsky, Ngai & Wicentowski, 01)
   - Cross-Lingual LM adaptation (this talk)

Ex2: If we have an extensive amount of data, is there a better way? ➔ Monolingual LM adaptation
Overview

- An approach to sharpen an LM
  - In a resource-deficient language: based on contemporaneous texts from resource-rich languages
  - In a resource-rich language: based on a selected small amount of texts
- Story-specific language modeling from comparable texts
- Integration of
  - machine translation (MT)
  - cross-language information retrieval (CLIR)
  - language modeling (LM)
- Experiments on automatic speech recognition (ASR) & statistical machine translation (SMT)
Ex1: Optical Character Recognition

All
I am a student

Ex2: Speech Recognition

It’s [tuː] cold
He is [tuː] years old
too
two

- Speech & NLP applications = Search problems
- Need to suppress ungrammatical outputs
- Assign a probability to given a sequence of word strings

\[ P(“\text{He is two years old”}) \gg P(“\text{He is too years old”}) \]
ASR problem: Finding word string $\hat{W} = w_1, w_2, \ldots, w_n$ given acoustic evidence $A$

$$\hat{W} = \arg \max_W P(W|A)$$

Bayes’ Rule

$$P(W|A) = \frac{P(W)P(A|W)}{P(A)}$$

$$\hat{W} = \arg \max_W P(W)P(A|W)$$

Language Model
Acoustic Model
$P(w_1, w_2, \cdots, w_n)$

\[ P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) \cdots P(w_n|w_1, \cdots, w_{n-1}) \]

\[ \approx P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) \cdots P(w_n|w_{n-2}, w_{n-1}) \]

\[ = P(w_1)P(w_2|w_1) \prod_{i=3}^{n} P(w_i|w_{i-2}, w_{i-1}) \] (4)

\[ P(w_i|w_{i-2}, w_{i-1}) = \frac{N(w_{i-2}, w_{i-1}, w_i)}{N(w_{i-2}, w_{i-1})} \implies \text{Trigram} \] (5)

\[ P(w_i|w_{i-1}) = \frac{N(w_{i-1}, w_i)}{N(w_{i-1})} \implies \text{Bigram} \] (6)

\[ P(w_i) = \frac{N(w_i)}{\sum_{w \in V} N(w)} \implies \text{Unigram} \] (7)

where $N(\ast)$ = the number of times (\ast) appears and $V$ is the vocabulary.
Word Error Rate (WER): Performance of ASR given LM

REF: This great machine can recognize speech
HYP: This machine can wreck nice beach

\[
\text{WER} = \frac{\text{DEL} + \text{SUB} + \text{INS}}{\text{No. words in REF}} \times 100 = \frac{1 + 2 + 1}{6} \times 100 = 66.7\%
\]

Perplexity (PPL): based on cross entropy of test data \( D \) w.r.t. LM \( M \) \( \rightarrow \) Geometric Mean

\[
H(P_D; P_M) = -\frac{1}{N} \sum_{i=1}^{N} \log_2 P_M(w_i|w_{i-2}, w_{i-1}) \quad (8)
\]

\[
PPL_M(D) = 2^{H(P_D; P_M)} = [P_M(w_1, \cdots, w_N)]^{-\frac{1}{N}} \quad (9)
\]

For both of WER and PPL, the lower, the better!
Language Model Adaptation

1. The are to know the issues necessary.
2. This will have this problems data.
3. One understand these problems the information.
4. Two would do problems above any a time.
5. A also get any other problems people operators tools.
6. Three do the a time all the problem.
7. Please need provide them resolve.
8. In insert all.
9. We.

Figure: IBM Trigram Example
Language Model Adaptation

✔ Handicaps of N-gram LMs:
  - Good at predicting function words, but not so good at predicting content words
  - Not able to capture long distance dependencies due to the Markov assumption
  - Static: cannot capture the dynamic nature of texts such as topic, domain, genre, etc.

LM adaptation

1. Build LMs according to topic, domain, genre, etc.
2. Adapt topic, domain, or genre specific LMs to the N-gram LM via linear interpolation or ME/MDI

Cross-Lingual LM adaptation: Flowchart

- Mandarin Story
- Automatic Speech Recognition
- Contemporaneous English Articles
- Automatic Transcription
- Cross-Language Information Retrieval
- English Article Aligned with Mandarin Story
- Statistical Machine Translation
- Cross-Language Unigram Model
- Baseline Chinese Acoustic Model
- Chinese Dictionary (Vocabulary)
- Baseline Chinese Language Model

$p_t(c \mid e)$

$P(c \mid d_i^E)$

$P_{CL \text{ unigram}}(c \mid d_i^E)$

$P(d_i^c)$
Finding document correspondence between $d_i^C \leftrightarrow d_j^E$ by CLIR (e.g. based on Cosine similarity)

Translation dictionary $P_T(c|e)$ (e.g. by GIZA++)

Given the document correspondence and $P_T(c|e)$,

$$P_{\text{CL-unigram}}(c|d_i^E) = \sum_{e \in E} P_T(c|e) \hat{P}(e|d_i^E), \quad \forall c \in C \quad (10)$$

Cross-Language LM construction

- Build story-specific cross-language LMs
- Linear interpolation with the baseline trigram LM

$$P_{\text{CL-interpolated}}(c_k|c_{k-1}, c_{k-2}, d_i^E) = \lambda P_{\text{CL-unigram}}(c_k|d_i^E) + (1 - \lambda) P(c_k|c_{k-1}, c_{k-2}) \quad (11)$$
## Database

**Hong Kong News (Ch-En) parallel text corpus**

<table>
<thead>
<tr>
<th>Language</th>
<th>Chinese</th>
<th></th>
<th>English</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus Partition</td>
<td>Train</td>
<td>Dev</td>
<td>Eval</td>
<td>Train</td>
</tr>
<tr>
<td># of Documents</td>
<td>16010</td>
<td>750</td>
<td>682</td>
<td>16010</td>
</tr>
<tr>
<td># of Word Tokens</td>
<td>4.2M</td>
<td>255K</td>
<td>177K</td>
<td>4.3M</td>
</tr>
<tr>
<td># of Characters</td>
<td>6.2M</td>
<td>376K</td>
<td>260K</td>
<td>–</td>
</tr>
<tr>
<td>Vocab Size</td>
<td>41K</td>
<td>–</td>
<td>39K</td>
<td>–</td>
</tr>
<tr>
<td>OOV Rate</td>
<td>–</td>
<td>–</td>
<td>0.4%</td>
<td>–</td>
</tr>
</tbody>
</table>

**Table:** Partition of the Hong Kong News corpus into training (Train), cross-validation and development (Dev) and evaluation (Eval) sets.
## Perplexity Results

<table>
<thead>
<tr>
<th>Language model (# Words)</th>
<th>C-dev PPL</th>
<th>C-eval PPL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Word</td>
<td>Char</td>
</tr>
<tr>
<td>Baseline trigram (4.2M)</td>
<td>106</td>
<td>23.7</td>
</tr>
<tr>
<td>CL-intpl w/ $d_i^E$ from CLIR</td>
<td>90.1</td>
<td>21.2</td>
</tr>
<tr>
<td>CL-intpl w/ true $d_i^E$</td>
<td>89.7</td>
<td>21.1</td>
</tr>
</tbody>
</table>

**Table:** Performance of story-specific language models with cross-lingual cues.
ASR Experimental Setup

- Acoustic model training
  - HUB4-NE Mandarin training data (96K wds) ~ 10 hours
- Chinese monolingual language model training
  - XINHUA: 13M wds
  - HUB4-NE: 96K wds
- ASR test set: NIST HUB4-NE test data (only F0 portion)
  1263 sents, 9.8K wds (1997 ~ 1998)
- English CLIR corpus: NAB-TDT
  - 45K docs, 30M wds
- N-best list rescoring

![Flowchart Diagram]

- Speech Recognizer (Bigram LM)
- 300-Best List
- N-Best List Rescoring w/ New LM

Examples:

- I am going to ...
- I am gonna ...
- I'm gonna ...
- My going to ...

1 hypothesis
Figure: Perplexity of the Reference Transcription and the Likelihood of the ASR Output v/s Number of $d_i^E$ for Typical Test Stories.

\[
P_{\text{CL-interpolated}}(c_k|c_{k-1}, c_{k-2}, d_i^E) = \lambda_{d_i^E} P_{\text{CL-unigram}}(c_k|d_i^E) + (1 - \lambda_{d_i^E}) P(c_k|c_{k-1}, c_{k-2})
\]
### ASR Results

<table>
<thead>
<tr>
<th>Language model</th>
<th>Perp</th>
<th>WER</th>
<th>CER</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>XINHUA trigram</td>
<td>426</td>
<td>49.9%</td>
<td>28.8%</td>
<td>–</td>
</tr>
<tr>
<td>CL-interpolated 1</td>
<td>375</td>
<td>49.5%</td>
<td>28.7%</td>
<td>0.208</td>
</tr>
<tr>
<td>CL-interpolated 2</td>
<td>346</td>
<td>48.8%</td>
<td>28.4%</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>HUB-4NE trigram</td>
<td>1195</td>
<td>60.1%</td>
<td>44.1%</td>
<td>–</td>
</tr>
<tr>
<td>CL-interpolated 1</td>
<td>750</td>
<td>59.3%</td>
<td>43.7%</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>CL-interpolated 2</td>
<td>630</td>
<td>58.8%</td>
<td>43.1%</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

CL-interpolated 1: Top 1 En doc, Single fixed $\lambda$
CL-interpolated 2: Top $N$ En docs, Doc-Specific $\lambda$
Recap

Cross-Lingual LM adaptation

- Data deficiency problem ➔
  1. Identify contemporaneous texts in a resource-rich language
  2. Extract useful stats
  3. Project into our target language

- Translation dictionary is required
- Proposed method: use statistical MT toolkit (GIZA++)
- A sentence-aligned parallel corpus is needed ➔ Expensive to obtain (esp. in resource-deficient languages)

We present methods to use:

1. cross-lingual lexical triggers
2. cross-lingual latent semantic analysis (LSA)

- Document-aligned parallel corpus is enough
- No explicit MT dictionary is needed
Triggers: Monolingual vs. Cross-Lingual

- Monolingual Triggers: e.g. “either ... or”
- Cross-Lingual Setting: Translation lexicons
- Based on Average Mutual Information ($I(e; c)$)

Monolingual Trigger Pair Candidates

Cross-Lingual Trigger Pair Candidates

Figure: Monolingual Triggers

Figure: Cross-Lingual Triggers
Cross-Lingual Lexical Triggers: Identification

Let

\[ P(e, c) = \frac{\#d(e, c)}{N} \quad \text{and} \quad P(e, \bar{c}) = \frac{\#d(e, \bar{c})}{N} \]  

(12)

where \( \#d(e) \) denote the number of English articles in which \( e \) occurs, and let

\[ P(e) = \frac{\#d(e)}{N} \quad \text{and} \quad P(c|e) = \frac{P(e, c)}{P(e)} \]  

(13)

\[ I(e; c) = P(e, c) \log \frac{P(c|e)}{P(c)} + P(e, \bar{c}) \log \frac{P(\bar{c}|e)}{P(\bar{c})} + P(\bar{e}, c) \log \frac{P(c|\bar{e})}{P(c)} + P(\bar{e}, \bar{c}) \log \frac{P(\bar{c}|\bar{e})}{P(\bar{c})} \]  

(14)
Cross-Lingual Lexical Triggers: Estimation

We estimate the trigger-based CL unigram probs with

\[ P_{\text{Trig}}(c|e) = \frac{I(e; c)}{\sum_{c' \in \mathcal{C}} I(e; c')} , \]  

(15)

Analogous to (10),

\[ P_{\text{Trig-unigram}}(c|d_i^E) = \sum_{e \in \mathcal{E}} P_{\text{Trig}}(c|e) \hat{P}(e|d_i^E) \]  

(16)

Again, we build the interpolated model

\[ P_{\text{Trig-interpolated}}(c_k|c_{k-1}, c_{k-2}, d_i^E) = \lambda P_{\text{Trig-unigram}}(c_k|d_i^E) + (1 - \lambda) P(c_k|c_{k-1}, c_{k-2}) \]  

(17)
Cross-Lingual Latent Semantic Analysis

✔ Singular Value Decomposition (SVD) of the parallel corpus
  - Input: word-document frequency matrix, \( W \)
  - Reduce the dimension into the smaller but adequate subspace
    \( \rightarrow \) Singular Value Decomposition: \( U, V, \) and \( S \)
  - \( S \): diagonal matrix w/ diagonal entries \( \sigma_1, \cdots, \sigma_k \) where
    \( \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k (k \geq R) \)
  - Remove noisy entries by setting \( \sigma_i = 0 \) for \( i > R \)

\[
W = U S V^T
\]

\( M \times N \) \hspace{1cm} \( M \times R \) \hspace{1cm} \( R \times R \) \hspace{1cm} \( R \times N \)
Given a monolingual corpus, $\overline{W}$, in either side
Use the same matrices $U, S$
Project into low-dimensional space, $\overline{V}^T = S^{-1}U^T\overline{W}$
Compare a query and a document in the reduced dimensional space
Word vs. word comparison

Each row in $W$ corresponds to a word (either $e_i$ or $c_j$)

Compare $c_j \in C$ and $e_i \in E$ to find most similar entries

Estimation of the translation probability $\Rightarrow$ similarity

$$P_{LSA}(c|e) = \frac{Sim(c, e)^\gamma}{\sum_{c' \in C} Sim(c, e)^\gamma} \quad \text{where} \quad \gamma \gg 1 \quad (18)$$
## ASR Results

<table>
<thead>
<tr>
<th>Language model</th>
<th>Perp</th>
<th>WER</th>
<th>CER</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>XINHUA trigram</td>
<td>426</td>
<td>49.9%</td>
<td>28.8%</td>
<td>–</td>
</tr>
<tr>
<td>CL-interpolated 2</td>
<td>346</td>
<td>48.8%</td>
<td>28.4%</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Trigger-interpolated</td>
<td>367</td>
<td>49.1%</td>
<td>28.6%</td>
<td>0.004</td>
</tr>
<tr>
<td>LSA-interpolated</td>
<td>364</td>
<td>49.3%</td>
<td>28.9%</td>
<td>0.043</td>
</tr>
<tr>
<td>Trig+LSA-interpolated</td>
<td>351</td>
<td>49.0%</td>
<td>28.7%</td>
<td>0.002</td>
</tr>
<tr>
<td>HUB-4NE trigram</td>
<td>1195</td>
<td>60.1%</td>
<td>44.1%</td>
<td>–</td>
</tr>
<tr>
<td>CL-interpolated 2</td>
<td>630</td>
<td>58.8%</td>
<td>43.1%</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Trigger-interpolated</td>
<td>727</td>
<td>58.8%</td>
<td>43.3%</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>LSA-interpolated</td>
<td>695</td>
<td>58.6%</td>
<td>43.1%</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Trig+LSA-interpolated</td>
<td>686</td>
<td>58.7%</td>
<td>43.2%</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
**Language Models: MT problem**

- **MT problem**: Finding English sentence $\hat{E} = e_1, e_2, \cdots, e_n$ given foreign sentence $F$

  $$\hat{E} = \arg \max_{E} P(E|F)$$  \hspace{1cm} (19)

- **Bayes’ Rule**

  $$P(E|F) = \frac{P(E)P(F|E)}{P(F)}$$  \hspace{1cm} (20)

  $$\hat{E} = \arg \max_{E} P(E)P(F|E)$$  \hspace{1cm} (21)
IBM Model for SMT

SMT as string rewriting

Mary did not slap the green witch
\[ \Downarrow \]
Mary not slap slap slap the the green witch
\[ \Downarrow \]
Mary no daba una botefada a la verde bruja
\[ \Downarrow \]
Mary no daba una botefada a la bruja verde

\[ n(\#|e) \]
\[ t(f|e) \]
\[ d(\#|\#) \]

\[ \Rightarrow \] All parameters can be automatically estimated from a sentence-aligned parallel corpus (training data)!!
BLEU (BiLingual Evaluation Understudy) measure

- Automatic measure approximates human judgement
- Compare MT outputs to human reference translations
- Inspired by Precision in IR: What percentage of hypothesis N-grams appears in the references?

REF: the blue dog barks
HYP: the the blue the

- Unigram precision: $4/4$
- Modified unigram precision: $p_1 = 2/4$
- Modified bigram precision: $p_2 = 1/3$

BP = \[
\begin{cases} 
1, & \text{if } c > r \\
\exp(1 - r/c), & \text{if } c \leq r 
\end{cases}
\]

BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right),
typically $N = 4$ and $w_n = 1/N$. 
SMT Experimental Setup

- 20004 Johns Hopkins Chinese to English MT Setup for NIST evaluation
- Translation model training: Chinese-English parallel corpus (Ch:175M words, En:207M words)
- Baseline LM: mixture of four LMs
  - AFP (200.8M words)
  - XINHUA (155.7M words)
  - FBIS (10.5M words)
  - People’s Daily (16.2M words)
- IR toolkit: LEMUR
- 1000-best list rescoring
  - Use the 3gram equally mixed LM for 1st pass N-best list generation
  - Use better LMs for the 2nd pass rescoring
## SMT Results

### BLEU scores of NIST MT Evaluation 01–03 sets

<table>
<thead>
<tr>
<th>Model</th>
<th>Eval01</th>
<th>Eval02</th>
<th>Eval03</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3g</td>
<td>4g</td>
<td>3g</td>
</tr>
<tr>
<td>Baseline LM</td>
<td>28.7</td>
<td>29.7</td>
<td>27.7</td>
</tr>
<tr>
<td>Adaptive LM 1</td>
<td>29.7</td>
<td>30.0</td>
<td>28.7</td>
</tr>
<tr>
<td>Adaptive LM 2</td>
<td>29.7</td>
<td>30.2</td>
<td>28.6</td>
</tr>
</tbody>
</table>

Adaptive LM 1: Top 1000 En docs
Adaptive LM 2: Top $N$ En docs
Conclusions

- Language model adaptation
  - If no data: cross-lingual adaptation from other language
  - If no dictionary: cross-lingual triggers, cross-lingual LSA
  - If enough data: select useful data

- Story-specific likelihood-based optimization

- Significant improvements in
  - LM perplexity
  - ASR WER
  - SMT BLEU

- Future work: maximum entropy/minimum discrimination information