Automatic Speech Recognition and Statistical Machine Translation under Uncertainty

Lambert Mathias
Advisor: Prof. William Byrne
Thesis Committee: Prof. Gerard Meyer, Prof. Trac Tran and Prof. Frederick Jelinek

Center for Language and Speech Processing
Department of Electrical and Computer Engineering
Johns Hopkins University

December 7, 2007

Uncertainty Reduction in Language Processing

- Two applications of statistical methods in language processing
  - Automatic Speech Recognition (ASR)
  - Statistical Machine Translation (SMT)
- Statistical learning approaches have to deal with *uncertainty*
  - Data sparsity, noise, incorrect modeling assumptions etc.
- Uncertainty in SMT
  - Ambiguity in translation from speech
  - Using a cascaded approach to translate speech
- Uncertainty in ASR
  - Training acoustic models in the absence of reliable transcripts
  - Lightly supervised discriminative training in the medical domain[1]

[1] Juergen Fritsch, Girija Yegnarayanan, Multimodal Technologies Inc.

Uncertainty Reduction in Language Processing

- Two applications of statistical methods in language processing
  - Automatic Speech Recognition (ASR)
  - Statistical Machine Translation (SMT)
- Statistical learning approaches have to deal with *uncertainty*
  - Data sparsity, noise, incorrect modeling assumptions etc.
- Uncertainty in SMT
  - Ambiguity in translation from speech
  - Using a cascaded approach to translate speech
- Uncertainty in ASR
  - Training acoustic models in the absence of reliable transcripts
  - Lightly supervised discriminative training in the medical domain[1]

[1] Juergen Fritsch, Girija Yegnarayanan, Multimodal Technologies Inc.

Uncertainty Reduction in Language Processing

- Two applications of statistical methods in language processing
  - Automatic Speech Recognition (ASR)
  - Statistical Machine Translation (SMT)
- Statistical learning approaches have to deal with *uncertainty*
  - Data sparsity, noise, incorrect modeling assumptions etc.
- Uncertainty in SMT
  - Ambiguity in translation from speech
  - Using a cascaded approach to translate speech
- Uncertainty in ASR
  - Training acoustic models in the absence of reliable transcripts
  - Lightly supervised discriminative training in the medical domain[1]

[1] Juergen Fritsch, Girija Yegnarayanan, Multimodal Technologies Inc.

Outline Part I: Speech Translation

Outline Part II: Discriminative Training for MT

Part I

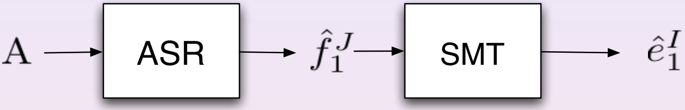
Speech Translation

Motivation

- Applications of Speech Translation
  - Facilitate international business communication
  - Computer aided learning
  - Language acquisition aid
- Statistical Speech Translation components
  - Automatic Speech Recognition (ASR)
  - Statistical Machine Translation (SMT)
- Integrating the ASR and SMT components
  - Varying levels of interaction - N-best list, Lattices, Confusion Networks
  - Coupling of ASR with the SMT - maximum information transfer

Outline

Loosely-Coupled Speech Translation



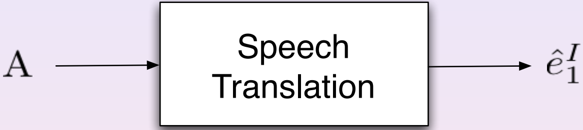
- Translate 1-best[2]
  - sub-optimal - no interaction between ASR and SMT
- Translate multiple hypotheses instead[3]
  - Does not exploit sub-sentential language information

---

[2] Liu et al, Noise Robustness in Speech to Speech Translation, IBM Technical Report 2003

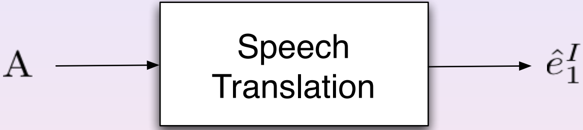
[3] Black et al, Rapid Development of Speech to Speech Translation Systems, ICSLP 2002

Integrated Speech Translation Architecture



- Objective: Tight integration of ASR with SMT system
- Unified modeling framework
  - No *a priori* decisions on candidates for translation
  - Integrated search over a larger space of translation candidates
  - SMT system robust to ASR errors
- Problem: How to translate ASR Lattices ?

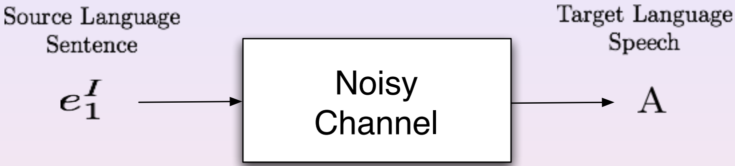
Integrated Speech Translation Architecture



- Objective: Tight integration of ASR with SMT system
- Unified modeling framework
  - No *a priori* decisions on candidates for translation
  - Integrated search over a larger space of translation candidates
  - SMT system robust to ASR errors
- Problem: How to translate ASR Lattices ?

Outline

Noisy Channel Model



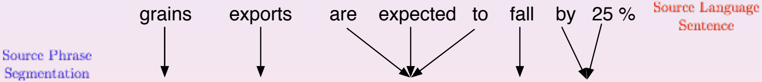
- Translation MAP decoder

$$\hat{e}_1^{\hat{I}} = \operatorname*{argmax}_{I, e_1^I} \Big\{ \max_{f_1^J} \quad \underbrace{P(e_1^I)}_{\substack{\textit{Language}\\\textit{Model}}} \quad \underbrace{P(f_1^J | e_1^I)}_{\substack{\textit{Translation}\\\textit{Model}}} \quad \underbrace{\mathbf{P(A | f_1^J)}}_{\substack{\textit{Acoustic}\\\textit{Model}}} \Big\}$$

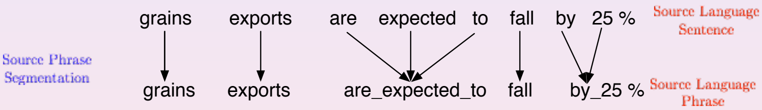
The Generative Process

grains exports are expected to fall by 25 % Source Language Sentence

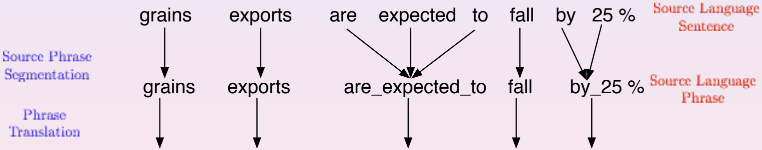
The Generative Process



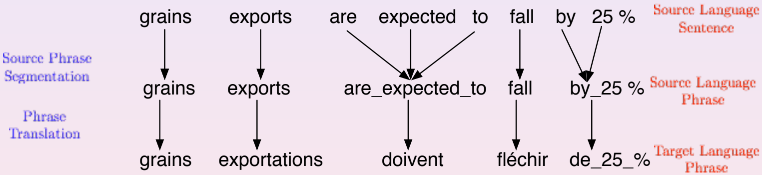
The Generative Process



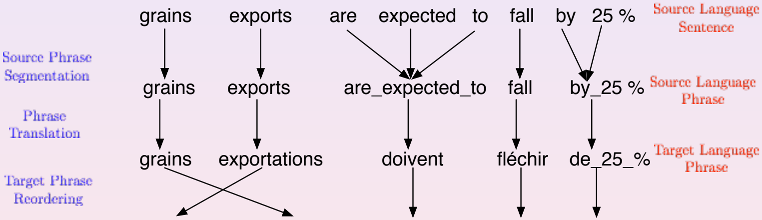
The Generative Process



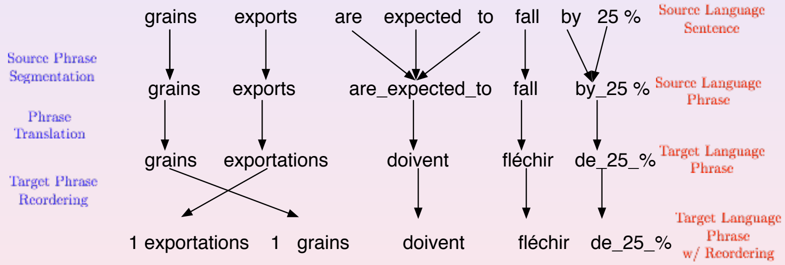
The Generative Process



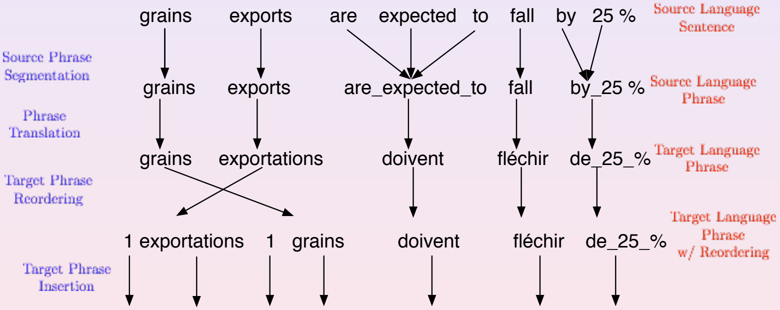
The Generative Process



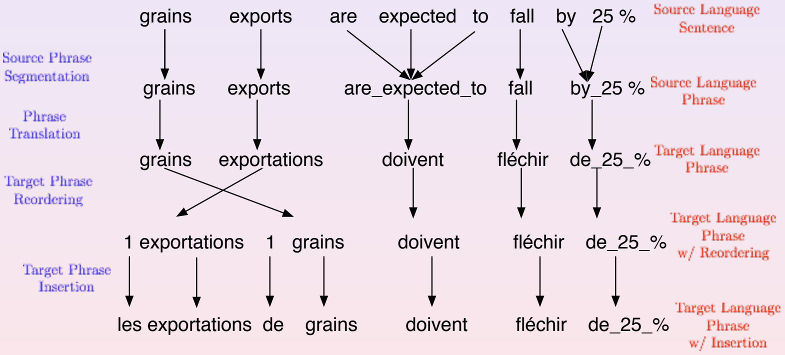
The Generative Process



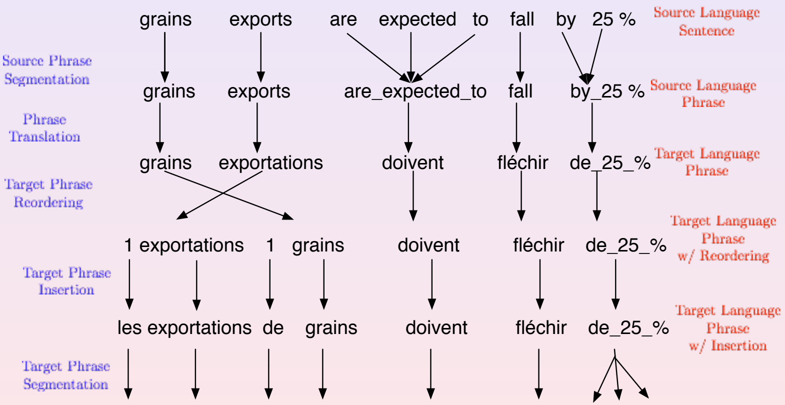
The Generative Process



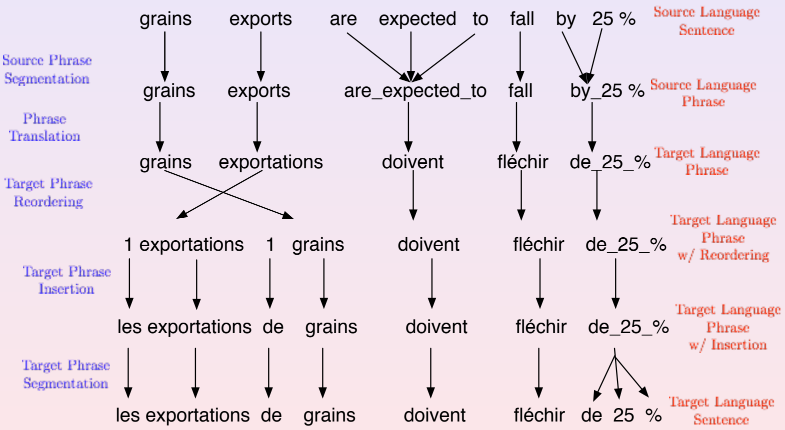
The Generative Process



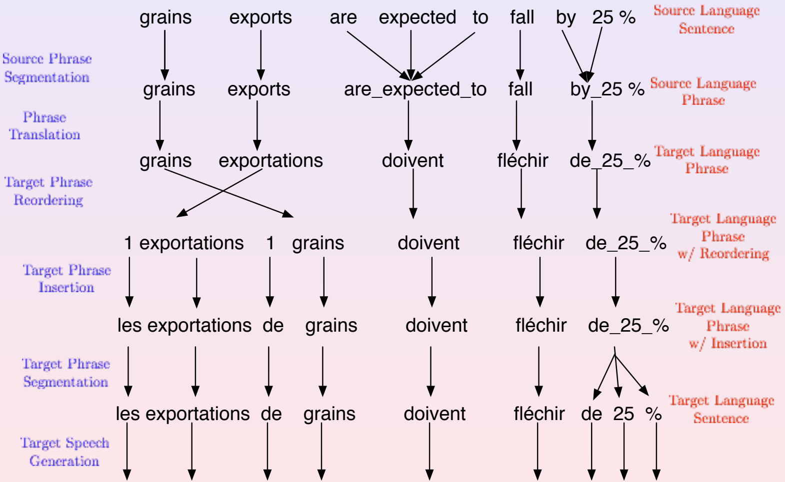
The Generative Process



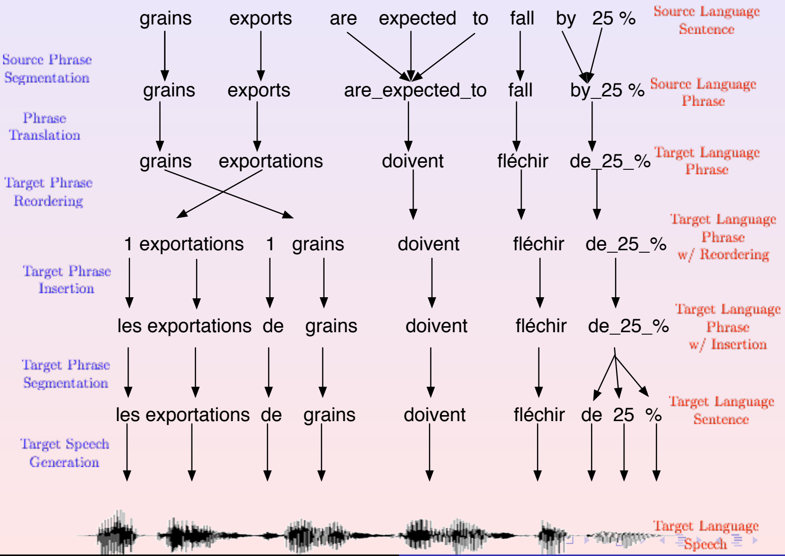
The Generative Process



The Generative Process



The Generative Process



Target Speech		*Target Sentence*		*Target Phrase with Insertion*		*Reordered Target Phrase*		*Target Phrase*		*Source Phrase*		*Source Sentence*
$\mathbf{A}$	$\longleftarrow$	$\mathbf{f_1^J}$	$\longleftarrow$	$\mathbf{v_1^R}$	$\longleftarrow$	$\mathbf{y_1^K}$	$\longleftarrow$	$\mathbf{x_1^K}$	$\longleftarrow$	$\mathbf{u_1^K}$	$\longleftarrow$	$\mathbf{e_1^I}$
Models	$\mathbf{P(A \vert f_1^J)}$		$\mathbf{P(f_1^J \vert v_1^R)}$		$\mathbf{P(v_1^R \vert y_1^K)}$		$\mathbf{P(y_1^K \vert x_1^K)}$		$\mathbf{P(x_1^K \vert u_1^K)}$		$\mathbf{P(u_1^K \vert e_1^I)}$	$\mathbf{P(e_1^I)}$
FSMs	$\mathcal{L}$		$\mathbf{\Omega}$		$\mathbf{\Phi}$		$\mathbf{R}$		$\mathbf{Y}$		$\mathbf{W}$	$\mathbf{G}$
	Target Word Acoustic Lattice		*Target Phrase Segmentation Transducer*		*Target Phrase Insertion Transducer*		*Target Phrase Reordering Transducer*		*Source – Target Phrase Translation Transducer*		*Source Phrase Segmentation Transducer*	*Source Language Model*

- Transformations via stochastic models implemented as WFSTs
- Built with standard WFST operations - composition and best path search
  - Translation graph $G \circ W \circ Y \circ R \circ \Phi \circ \Omega \circ \mathcal{L}$
- Straightforward extension of the text based SMT system

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}} \{ \max_{f_1^J \in \mathcal{L}} \max_{v_1^R, y_1^K, x_1^K, u_1^K, K} \underbrace{P(A|f_1^J)}_{\text{Target Lattice}} \underbrace{P(f_1^J, v_1^R, y_1^K, x_1^K, u_1^K, e_1^I)}_{\text{Text Translation}} \}$$

Target Speech	*Target Sentence*	*Target Phrase with Insertion*	*Reordered Target Phrase*	*Target Phrase*	*Source Phrase*	*Source Sentence*
$\mathbf{A}$	$\mathbf{f_1^J}$	$\mathbf{v_1^R}$	$\mathbf{y_1^K}$	$\mathbf{x_1^K}$	$\mathbf{u_1^K}$	$\mathbf{e_1^I}$

$$\mathbf{A} \longleftarrow \mathbf{f_1^J} \longleftarrow \mathbf{v_1^R} \longleftarrow \mathbf{y_1^K} \longleftarrow \mathbf{x_1^K} \longleftarrow \mathbf{u_1^K} \longleftarrow \mathbf{e_1^I}$$

Models	$\mathbf{P}(\mathbf{A}\vert\mathbf{f_1^J})$	$\mathbf{P}(\mathbf{f_1^J}\vert\mathbf{v_1^R})$	$\mathbf{P}(\mathbf{v_1^R}\vert\mathbf{y_1^K})$	$\mathbf{P}(\mathbf{y_1^K}\vert\mathbf{x_1^K})$	$\mathbf{P}(\mathbf{x_1^K}\vert\mathbf{u_1^K})$	$\mathbf{P}(\mathbf{u_1^K}\vert\mathbf{e_1^I})$	$\mathbf{P}(\mathbf{e_1^I})$
FSMs	$\mathcal{L}$	$\mathbf{\Omega}$	$\mathbf{\Phi}$	$\mathbf{R}$	$\mathbf{Y}$	$\mathbf{W}$	$\mathbf{G}$
	Target Word Acoustic Lattice	*Target Phrase Segmentation Transducer*	*Target Phrase Insertion Transducer*	*Target Phrase Reordering Transducer*	*Source – Target Phrase Translation Transducer*	*Source Phrase Segmentation Transducer*	*Source Language Model*

- Transformations via stochastic models implemented as WFSTs
- Built with standard WFST operations - composition and best path search
  - Translation graph $G \circ W \circ Y \circ R \circ \Phi \circ \Omega \circ \mathcal{L}$
- Straightforward extension of the text based SMT system

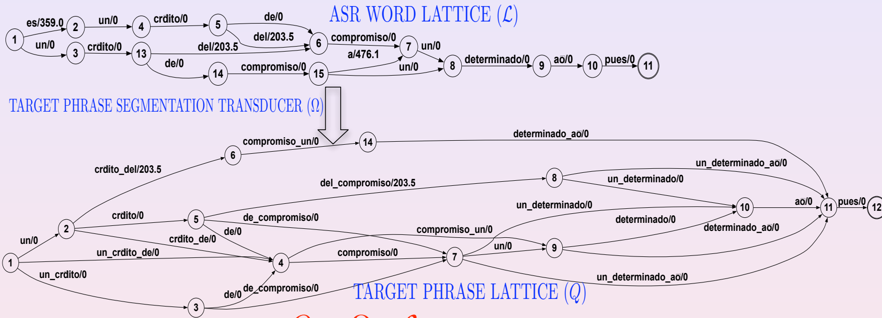
$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}} \{ \max_{f_1^J \in \mathcal{L}} \max_{v_1^R, y_1^K, x_1^K, u_1^K, K} \underbrace{P(A|f_1^J)}_{\text{Target Lattice}} \underbrace{P(f_1^J, v_1^R, y_1^K, x_1^K, u_1^K, e_1^I)}_{\text{Text Translation}} \}$$

Translation under ASR Posterior

- Translation under the proper ASR posterior distribution
- Uncertainty reduction: strong target LM can help guide the translation process

$$\hat{e}_1^{\hat{I}} = \operatorname*{argmax}_{I, e_1^I} \{ \max_{f_1^J \in \mathcal{L}} \max_{v_1^R, y_1^K, x_1^K, u_1^K, K} \underbrace{P(A|f_1^J) P(f_1^J)}_{\textit{Target Lattice}} \underbrace{P(f_1^J, v_1^R, y_1^K, x_1^K, u_1^K, e_1^I)}_{\textit{Text Translation}} \}$$

Translation is from a Lattice of Phrase Sequences



$$Q = \Omega \circ \mathcal{L}$$

- Phrase sequence lattice has foreign phrase sequences in the ASR lattice
  - Phrase sequence corresponds to translatable word sequences in lattice
  - Lattice contains the ASR weights
  - Translation of phrase lattice is MAP translation of ASR word lattice

Direct Translation of ASR word lattice

- Original Problem: How to translate ASR word lattice ?
- New Problem: How to extract translatable phrases from ASR lattice ?
- Speech Translation recast as an ASR analysis problem:
  1. Perform foreign language ASR to obtain speech lattice $\mathcal{L}$
  2. Analyze foreign language word lattice and extract translatable phrases
  3. Build the translation component models and convert the word lattice to a phrase lattice $\Omega \circ \mathcal{L}$
  4. Translate the foreign language phrase lattice
- Extracting translatable phrases[4]

$$C(w_1^k) = \sum_{\pi \in \mathcal{L}} \#_{w_1^k}(\pi)\ [[\mathcal{L}]](w_1^k)$$

- Filtering low-confidence phrases

$$p(w_1^k) = \frac{C(w_1^k)}{\sum_{w_1^k \in \mathcal{L}} C(w_1^k)}$$

[4] Count automaton -GRM Library

Direct Translation of ASR word lattice

- Original Problem: How to translate ASR word lattice ?
- New Problem: How to extract translatable phrases from ASR lattice ?
- Speech Translation recast as an ASR analysis problem:
  1. Perform foreign language ASR to obtain speech lattice $\mathcal{L}$
  2. Analyze foreign language word lattice and extract translatable phrases
  3. Build the translation component models and convert the word lattice to a phrase lattice $\Omega \circ \mathcal{L}$
  4. Translate the foreign language phrase lattice
- Extracting translatable phrases[4]

$$C(w_1^k) = \sum_{\pi \in \mathcal{L}} \#_{w_1^k}(\pi)\ [[\mathcal{L}]](w_1^k)$$

- Filtering low-confidence phrases

$$p(w_1^k) = \frac{C(w_1^k)}{\sum_{w_1^k \in \mathcal{L}} C(w_1^k)}$$

[4] Count automaton -GRM Library

Direct Translation of ASR word lattice

- Original Problem: How to translate ASR word lattice ?
- New Problem: How to extract translatable phrases from ASR lattice ?
- Speech Translation recast as an ASR analysis problem:
  1. Perform foreign language ASR to obtain speech lattice $\mathcal{L}$
  2. Analyze foreign language word lattice and extract translatable phrases
  3. Build the translation component models and convert the word lattice to a phrase lattice $\Omega \circ \mathcal{L}$
  4. Translate the foreign language phrase lattice
- Extracting translatable phrases[4]

$$C(w_1^k) = \sum_{\pi \in \mathcal{L}} \#_{w_1^k}(\pi)\ [[\mathcal{L}]](w_1^k)$$

- Filtering low-confidence phrases

$$p(w_1^k) = \frac{C(w_1^k)}{\sum_{w_1^k \in \mathcal{L}} C(w_1^k)}$$

[4] Count automaton -GRM Library

Outline

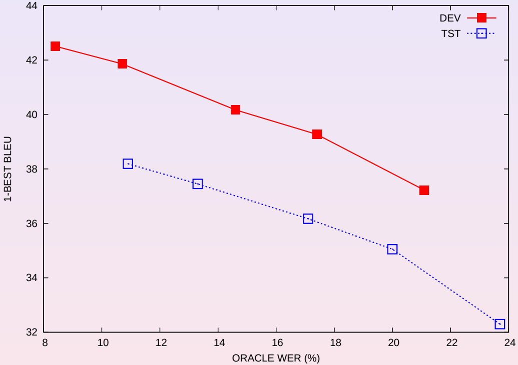
Experiment Setup

- Parallel Text : documents in the two languages aligned at sentence level - 1.4M
- Word alignments using MTTK[5]
- Phrase pair inventory of phrasal translations[6]
- Component distributions under current phrase pair inventory
- Automated evaluation measure
  - BLEU[7]: geometric mean of *n*-gram overlap with penalization for short sentences

---

[5] Deng, Y. and Byrne W, 2005
[6] Och 2002
[7] Papineni et al, 2001

ASR Lattice Oracle WER Translation Experiments



- Oracle WER path - path in lattice closest to reference transcript under edit distance
- Decrease in WER correlates with increase in BLEU

TCSTAR 2005 EPPS Lattice Translation performance

Spanish Source	DEV BLEU	EVAL BLEU
Reference Transcription	48.6	42.4
ASR 1-best	39.5	32.5
ASR Lattice	40.7	**33.6**

- Lattice translation better than ASR 1-best hypothesis

Speech Translation N-best list translation quality

Translation Input	oracle-best BLEU	
	DEV	TST
Reference Transcription	68.4	60.3
ASR 1-best	55.6	48.1
ASR Lattice	57.8	49.7

Table: Oracle-best BLEU for EPPS development and test set measured over a 1000-best translation list

- Oracle-best BLEU: N-best list candidate with maximum sentence level BLEU score
- Quantifies the best possible performance under BLEU given the current inventory of phrases

Input	Translation
Reference	1. in accordance with the committee on budgets the period for tabling amendments for the second reading of the european union budget will end on wednesday the first of december at twelve noon. 2. in agreement with the committee on budgets the deadline for the presentation of projects of amendment for the second reading of the european union budget will finish on wednesday first of december at twelve noon.
ASR 1-best	according to the committee on budgets of the deadline for the submission of projects amendment concerning the second reading of the budget union will end on wednesday , one of the twelve noon
ASR Lattice	**in accordance with** the committee on budgets of the deadline for the submission of projects amendment concerning the second reading of the budget union will end on wednesday , one of the twelve noon

Part II

Discriminative Training for MT

Outline

Motivation

- Automatic evaluation measures for improving translation performance
- Need to optimize MT parameters for a particular task or evaluation metric automatically
- Efficient procedures for parameter tuning are needed
  - Capable of handling many features
  - Robust
  - State-of-the-art translation performance

Outline

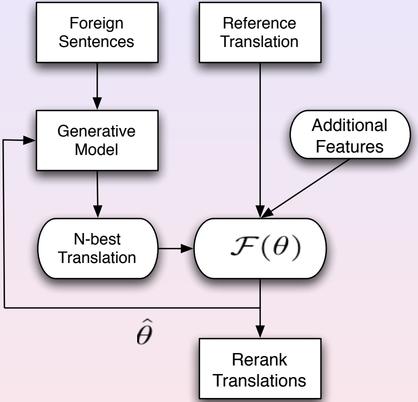
Learning Problem Definition

- Objective: Discriminative training for improved translation
  - Introduce growth transformations for MT parameter optimization
  - Compare translation performance with MET line search optimization

- Training problem:
  - Given a parallel training corpus - foreign sentences and their translations
  - A set of parameters: $\theta = \{\theta_1, \ldots, \theta_Q\}$
  - A joint distribution $p_\theta(\mathbf{e}_s, \mathbf{f}_s) = \prod_q \Phi_q(\mathbf{e}_s, \mathbf{f}_s)^{\theta_q}$
  - An objective function: $F(\theta) = f(p_\theta(\mathbf{e}_s, \mathbf{f}_s))$
  - Optimization problem:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \; \mathcal{F}(\theta)$$

- Translation evaluated on a blind test set using the optimized parameters



- 2-pass approach - decoding followed by optimization
- Incorporate additional features

Outline

Minimum Error Training

$$F_{MET}(\theta) = \sum_{s=1}^{S} BLEU(\operatorname*{argmax}_{\mathbf{e}_s} p_\theta(\mathbf{e}_s|\mathbf{f}_s), \mathbf{e}_s^+)$$

$\mathbf{e}_s$ = english translation hypothesis
$\mathbf{e}_s^+$ = english translation reference
$\mathbf{f}_s$ = foreign sentence

- Not smooth and not differentiable
- Piecewise constant
- Multidimensional gradient free search - line search subroutine[8]
- Current state-of-the-art

[8] F. Och, Minimum Error Training in Statsitical Machine Translation, 2002

Expected BLEU Maximization

$$F_{MBR}(\theta) = \sum_{s=1}^{S} \sum_{k=1}^{N_s} BLEU(\mathbf{e}_{sk}, \mathbf{e}_s^{+}) \frac{p_\theta(\mathbf{e}_{sk}, \mathbf{f}_s)}{\sum_{k=1}^{N_s} p_\theta(\mathbf{e}_{sk}, \mathbf{f}_s)}$$

$\mathbf{e}_s =$ english translation hypothesis
$\mathbf{e}_s^{+} =$ english translation reference
$\mathbf{f}_s =$ foreign sentence

- Partial credit to all hypotheses in N-best list
- Non-convex but differentiable
- Related to Minimum Bayes Risk decoding

Maximizing the posterior

$$F_{MMI}(\theta) = \sum_{s=1}^{S} \log \frac{p_\theta(\mathbf{e}_s^+, \mathbf{f}_s)}{\sum_{k=1}^{N_s} p_\theta(\mathbf{e}_{sk}, \mathbf{f}_s)}$$

$\mathbf{e}_s$ = english translation hypothesis
$\mathbf{e}_s^+$ = english translation reference
$\mathbf{f}_s$ = foreign sentence

- Maximum Mutual Information (MMI) - separating truth from competing classes
- Smooth and differentiable
- Labeling the true class - reference vs oracle BLEU hypothesis

Outline

Growth Transformations in ASR

- Locally maximize a rational function[9] $F(\theta) = \frac{N(\theta)}{D(\theta)}$
- Assumes $N(\theta)$ and $D(\theta)$ are polynomials in θ, $D(\theta) > 0$
- Parameter set $\theta = \{\theta_q | \theta_q \geq 0, \quad \sum_q \theta_q = 1\}$
- Define a transformation

$$T(\theta)_q = \frac{\theta_q \left(\frac{\partial P_\theta(\Theta)}{\partial \theta_q} + C \right)}{\sum_{q=1}^{Q} \theta_q \left(\frac{\partial P_\theta(\Theta)}{\partial \theta_q} + C \right)}$$

  where,

$$P_\theta(\Theta) = N_\theta(\Theta) - F(\theta) D_\theta(\Theta) + C$$

- For sufficiently large $C > 0$, then $T(\theta)$ is a growth transform if

$$F(T(\theta)) \geq F(\theta)$$

[9] Gopalakrishnan 1991

Growth Transformations in ASR

- Locally maximize a rational function[9] $F(\theta) = \frac{N(\theta)}{D(\theta)}$
- Assumes $N(\theta)$ and $D(\theta)$ are polynomials in θ, $D(\theta) > 0$
- Parameter set $\theta = \{\theta_q | \theta_q \geq 0, \quad \sum_q \theta_q = 1\}$
- Define a transformation

$$T(\theta)_q = \frac{\theta_q \left(\frac{\partial P_\theta(\Theta)}{\partial \theta_q} + C \right)}{\sum_{q=1}^{Q} \theta_q \left(\frac{\partial P_\theta(\Theta)}{\partial \theta_q} + C \right)}$$

where,

$$P_\theta(\Theta) = N_\theta(\Theta) - F(\theta) D_\theta(\Theta) + C$$

- For sufficiently large $C > 0$, then $T(\theta)$ is a growth transform if

$$F(T(\theta)) \geq F(\theta)$$

[9] Gopalakrishnan 1991

Growth Transformations in ASR

- Locally maximize a rational function[9] $F(\theta) = \frac{N(\theta)}{D(\theta)}$
- Assumes $N(\theta)$ and $D(\theta)$ are polynomials in θ, $D(\theta) > 0$
- Parameter set $\theta = \{\theta_q | \theta_q \geq 0, \quad \sum_q \theta_q = 1\}$
- Define a transformation

$$T(\theta)_q = \frac{\theta_q \left(\frac{\partial P_\theta(\Theta)}{\partial \theta_q} + C \right)}{\sum_{q=1}^{Q} \theta_q \left(\frac{\partial P_\theta(\Theta)}{\partial \theta_q} + C \right)}$$

where,

$$P_\theta(\Theta) = N_\theta(\Theta) - F(\theta) D_\theta(\Theta) + C$$

- For sufficiently large $C > 0$, then $T(\theta)$ is a growth transform if

$$F(T(\theta)) \geq F(\theta)$$

[9] Gopalakrishnan 1991

Enumerating the joint distribution

- Note that the objective $F(\theta) = f(p_\theta(\mathbf{e}_s, \mathbf{f}_s))$ not a polynomial in θ

$$p_\theta(\mathbf{e}_s, \mathbf{f}_s) = \prod_q \Phi_q(\mathbf{e}_s, \mathbf{f}_s)^{\theta_q} \approx \underbrace{\prod_{q=1} \sum_{k=0}^{n} \frac{\left(\theta_q \log \Phi_q(\mathbf{e}_s, \mathbf{f}_s)\right)^k}{k!}}_{p_\theta^{(n)}(\mathbf{e}_s, \mathbf{f}_s)}$$

- Growth transform for polynomials: $F^{(n)}(T^{(n)}(\theta)) \geq F^{(n)}(\theta)$
- If $\lim_{n\to\infty} T^{(n)}(\theta) \to T(\theta)$ and $\lim_{n\to\infty} F^{(n)}(\theta) \to F(\theta)$ then

$$\lim_{n\to\infty} F^{(n)}(T^{(n)}(\theta)) \geq \lim_{n\to\infty} F^{(n)}(\theta) \leftrightarrow F(T(\theta)) \geq F(\theta)$$

Result

Growth transformations can be extended to any function $f(\theta)$ that is differentiable and is analytical[a]

[a] Kanevsky 1995

Enumerating the joint distribution

- Note that the objective $F(\theta) = f(p_\theta(\mathbf{e}_s, \mathbf{f}_s))$ not a polynomial in θ

$$p_\theta(\mathbf{e}_s, \mathbf{f}_s) = \prod_q \Phi_q(\mathbf{e}_s, \mathbf{f}_s)^{\theta_q} \approx \underbrace{\prod_{q=1} \sum_{k=0}^{n} \frac{\left(\theta_q \ \log \Phi_q(\mathbf{e}_s, \mathbf{f}_s)\right)^k}{k!}}_{p_\theta^{(n)}(\mathbf{e}_s, \mathbf{f}_s)}$$

- Growth transform for polynomials: $F^{(n)}(T^{(n)}(\theta)) \geq F^{(n)}(\theta)$
- If $\lim_{n\to\infty} T^{(n)}(\theta) \to T(\theta)$ and $\lim_{n\to\infty} F^{(n)}(\theta) \to F(\theta)$ then

$$\lim_{n\to\infty} F^{(n)}(T^{(n)}(\theta)) \geq \lim_{n\to\infty} F^{(n)}(\theta) \leftrightarrow F(T(\theta)) \geq F(\theta)$$

Result

Growth transformations can be extended to any function $f(\theta)$ that is differentiable and is analytical[a]

[a] Kanevsky 1995

Enumerating the joint distribution

- Note that the objective $F(\theta) = f(p_\theta(\mathbf{e}_s, \mathbf{f}_s))$ not a polynomial in θ

$$p_\theta(\mathbf{e}_s, \mathbf{f}_s) = \prod_q \Phi_q(\mathbf{e}_s, \mathbf{f}_s)^{\theta_q} \approx \underbrace{\prod_{q=1} \sum_{k=0}^{n} \frac{\left(\theta_q \ \log \Phi_q(\mathbf{e}_s, \mathbf{f}_s)\right)^k}{k!}}_{p_\theta^{(n)}(\mathbf{e}_s, \mathbf{f}_s)}$$

- Growth transform for polynomials: $F^{(n)}(T^{(n)}(\theta)) \geq F^{(n)}(\theta)$
- If $\lim_{n\to\infty} T^{(n)}(\theta) \to T(\theta)$ and $\lim_{n\to\infty} F^{(n)}(\theta) \to F(\theta)$ then

$$\lim_{n\to\infty} F^{(n)}(T^{(n)}(\theta)) \geq \lim_{n\to\infty} F^{(n)}(\theta) \leftrightarrow F(T(\theta)) \geq F(\theta)$$

Result

Growth transformations can be extended to any function $f(\theta)$ that is differentiable and is analytical[a]

[a] Kanevsky 1995

Enumerating the joint distribution

- Note that the objective $F(\theta) = f(p_\theta(\mathbf{e}_s, \mathbf{f}_s))$ not a polynomial in θ

$$p_\theta(\mathbf{e}_s, \mathbf{f}_s) = \prod_q \Phi_q(\mathbf{e}_s, \mathbf{f}_s)^{\theta_q} \approx \underbrace{\prod_{q=1} \sum_{k=0}^{n} \frac{\left(\theta_q \ \log \Phi_q(\mathbf{e}_s, \mathbf{f}_s)\right)^k}{k!}}_{p_\theta^{(n)}(\mathbf{e}_s, \mathbf{f}_s)}$$

- Growth transform for polynomials: $F^{(n)}(T^{(n)}(\theta)) \geq F^{(n)}(\theta)$
- If $\lim_{n\to\infty} T^{(n)}(\theta) \to T(\theta)$ and $\lim_{n\to\infty} F^{(n)}(\theta) \to F(\theta)$ then

$$\lim_{n\to\infty} F^{(n)}(T^{(n)}(\theta)) \geq \lim_{n\to\infty} F^{(n)}(\theta) \leftrightarrow F(T(\theta)) \geq F(\theta)$$

Result

Growth transformations can be extended to any function $f(\theta)$ that is differentiable and is analytical[a]

[a] Kanevsky 1995

Growth Transform Iterative Training Procedure

1. Initialize the parameter vector $\theta = \{\theta_1, \cdots \theta_Q\}$, such that $\sum_{q=1}^{Q} \theta_q = 1$, and $i = 0$.
2. For each parameter $\theta_q^{(i)}$, calculate the gradient $\nabla F(\theta^{(i)})|_{\theta=\theta_q^{(i)}}$
3. For each parameter $\theta_q^{(i)}$, calculate the parameter update

$$\theta_q{}^{(i+1)} = \frac{\theta_q^{(i)} \left(\nabla F(\theta^{(i)})|_{\theta=\theta_q^{(i)}} + C \right)}{\sum_{q=1}^{Q} \theta_q^{(i)} \left(\nabla F(\theta^{(i)})|_{\theta=\theta_q^{(i)}} + C \right)}$$

4. If $F(\theta^{(i+1)}) \leq F(\theta^{(i)})$ or if $i == MAXITER$, then terminate. Else, $i \leftarrow i + 1$, goto Step 2.

Implementation Details

- Posterior scaling

$$p_{\theta,\alpha}(\mathbf{e}_{sk}|\mathbf{f}_s) = \frac{p_\theta(\mathbf{e}_{sk},\mathbf{f}_s)^\alpha}{\sum_{k=1}^{N_s} p_\theta(\mathbf{e}_{sk},\mathbf{f}_s)^\alpha}$$

- Convergence Rate

$$C = N_c * \left[\max\left\{\max_q\{-\nabla F(\theta)|_{\theta=\theta_q}\}, 0\right\} + \epsilon\right]$$

- Entropy Regularization

$$G(\theta) = F(\theta) + T\ H(p_{\theta,\alpha})$$

Outline

Hyper-Parameter Tuning

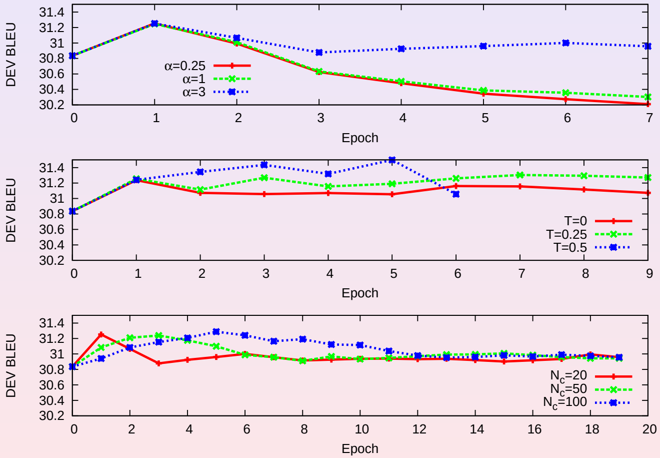


Figure: Chinese-English Text Translation Task: Expected BLEU over a 1000-best N-best list

MMI choosing the true class - oracle vs reference

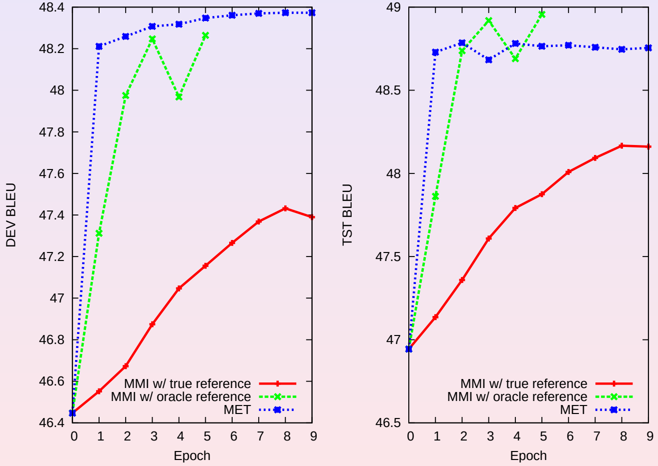


Figure: Arabic-English Text Translation Task: MMI over a 1000-best N-best list

Speech Translation Experiments

Translation Input	Training Criterion	Optimization Method	BLEU DEV	BLEU TST
ASR 1-best	MET	Line Search	39.5	32.5
	MMI	Growth Transform	35.8	**33.2**
	Expected BLEU	Growth Transform	38.1	32.3
ASR Lattice	MET	Line Search	40.7	33.6
	MMI	Growth Transform	36.8	34.3
	Expected BLEU	Growth Transform	37.2	**34.6**

Table: Comparing MMI and Expected BLEU and MET training criteria for the Spanish-English ASR translation task

- MET overfits the training data
- Expected BLEU and MMI outperform MET for lattice translation

Conclusions: Machine Translation

- Novel weighted finite state approach to translation of speech
  - Noisy channel formulation direct extension of text translation systems
  - Efficient phrase extraction from lattices
  - ASR phrase pruning to control ambiguity
  - Improved performance over ASR 1-best
  - Confusion network decoding - word and phrase confusion networks
- Improved discriminative training for SMT
  - Iterative growth transformation based updates for the MT parameters
  - Comparable to MET line search
  - Principled approach to the optimization of MT objective functions
  - Extensions to lattice based training using WFSTs
  - Anticipate further gains by increasing the number of features

Conclusions: Machine Translation

- Novel weighted finite state approach to translation of speech
  - Noisy channel formulation direct extension of text translation systems
  - Efficient phrase extraction from lattices
  - ASR phrase pruning to control ambiguity
  - Improved performance over ASR 1-best
  - Confusion network decoding - word and phrase confusion networks
- Improved discriminative training for SMT
  - Iterative growth transformation based updates for the MT parameters
  - Comparable to MET line search
  - Principled approach to the optimization of MT objective functions
  - Extensions to lattice based training using WFSTs
  - Anticipate further gains by increasing the number of features

Thank You !