Minimum Bayes-Risk Techniques for Automatic Speech Recognition and Machine Translation

October 23, 2003

Shankar Kumar

Advisor: Prof. Bill Byrne

ECE Committee: Prof. Gert Cauwenberghs and Prof. Pablo Iglesias

Center for Language and Speech Processing and

Department of Electrical and Computer Engineering

The Johns Hopkins University

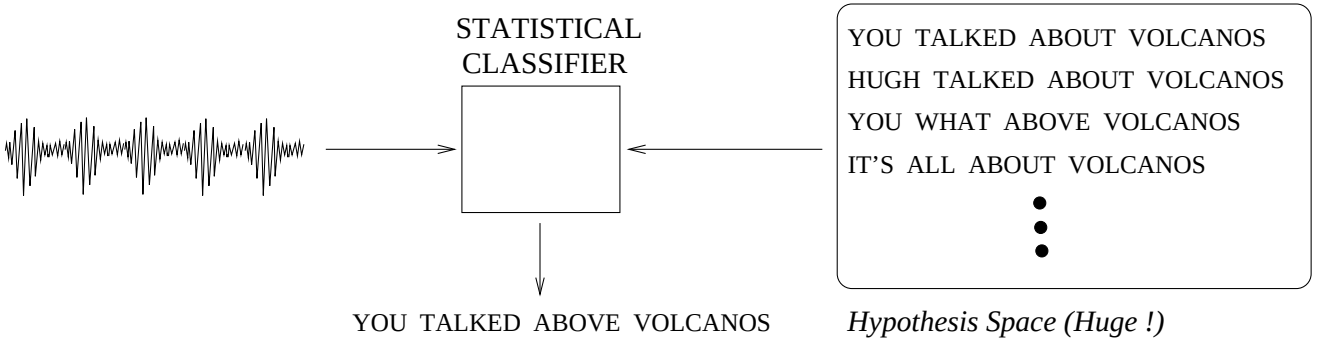
Motivation

- Automatic Speech Recognition (ASR) and Machine Translation (MT) are finding several applications
  - Examples: Information Retrieval from Text and Speech Archives, Devices for Speech to Speech Translation etc.
  - Usefulness is measured by Task-specific error metrics
- Maximum Likelihood techniques are used in estimation and classification of current ASR/MT systems
  - Do not take into account task-specific evaluation measures
- Minimum Bayes-Risk Classification
  - Building automatic systems tuned for specific tasks
  - Task-specific Loss functions
  - Formulation in two different areas - automatic speech recognition and machine translation

Outline

- Automatic Speech Recognition
  - Minimum Bayes-Risk Classifiers
  - Segmental Minimum Bayes-Risk Classification
  - Risk-Based Lattice Segmentation
- Statistical Machine Translation
  - A Statistical Translation Model
  - Minimum Bayes-Risk Classifiers for Word Alignment of Bilingual Texts
  - Minimum Bayes-Risk Classifiers for Machine Translation
- Conclusions and Future Work

Loss functions in Automatic Speech Recognition



- Loss function

Reference :	HUGH	TALKED	ABOUT	VOLCANOS	String Edit Distance (Word Error Rate)
Hypothesis :	YOU	TALKED	ABOUT	VOLCANOS	1/4 (25%)

- Loss-function is specific to the application of ASR system

Reference	:	HUGH	TALKED	ABOUT	VOLCANOS
Hypothesis	:	YOU	TALKED	ABOUT	VOLCANOS

	Sentences	Words	Keywords	Understanding
Loss(Truth,Hyp)	1/1	1/4	1/2	Large Loss

Minimum Bayes-Risk (MBR) Speech Recognizer

- Evaluate the expected loss of each hypothesis

$$E(W') = \sum_{W \in \mathcal{W}} L(W, W') P(W|A)$$

- Select the hypothesis with least expected loss

$$\delta_{MBR}(A) = \underset{W' \in \mathcal{W}}{\operatorname{argmin}} \sum_{W \in \mathcal{W}} L(W, W') P(W|A)$$

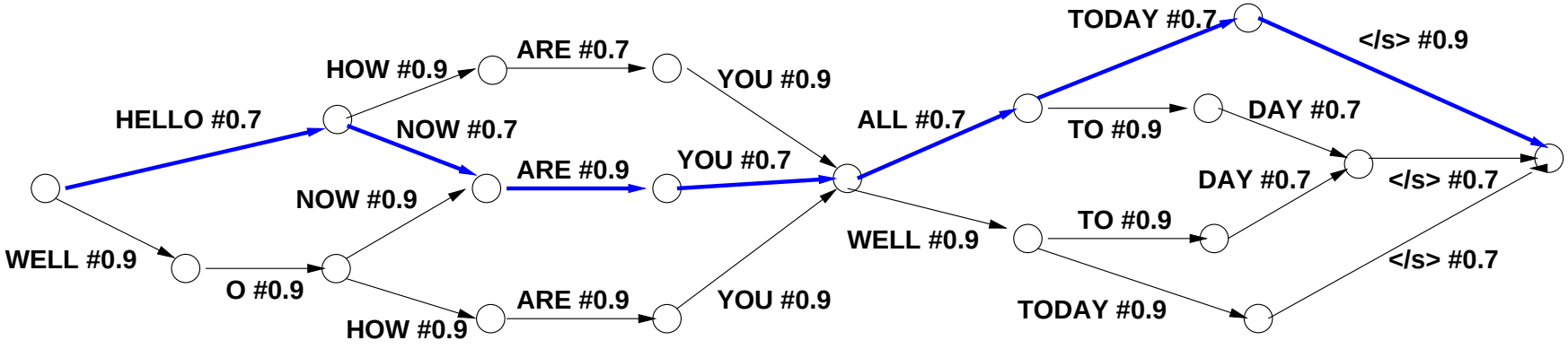
- Relation to Maximum A-posteriori Probability (MAP) Classifiers
  - Consider a sentence error loss function: $L(W, W') = \begin{cases} 1 \text{ if } W \neq W' \\ 0 \text{ otherwise} \end{cases}$

    Then, $\delta_{MBR}(A)$ reduces to the MAP classifier

$$\tilde{W} = \underset{W' \in \mathcal{W}}{\operatorname{argmax}} \; P(W'|A)$$

Algorithmic Implementations of MBR Speech Recognizers

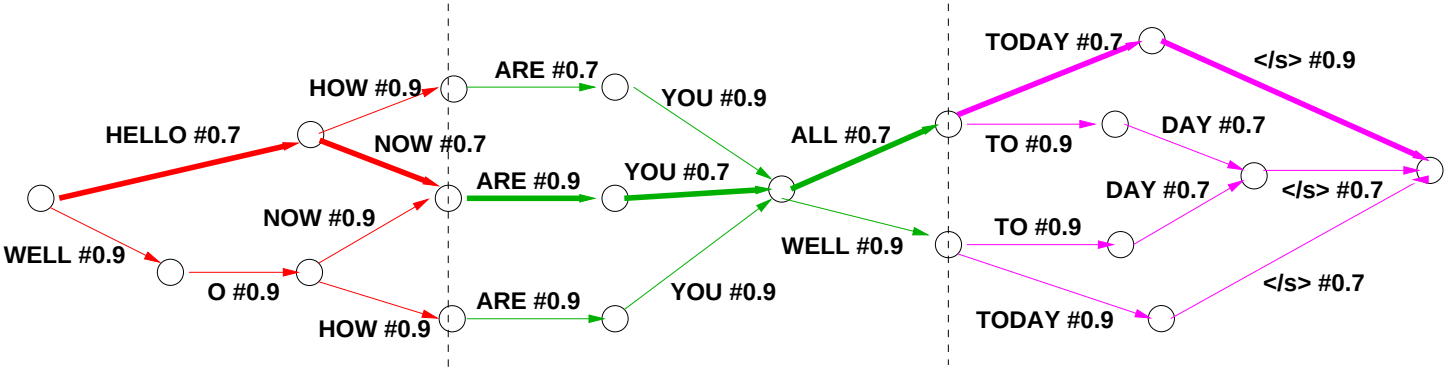
- Loss function of interest is String Edit distance (Word Error Rate)
- *Word Lattice*



  - Lattices are compact representation of the most likely word strings generated by a speech recognizer
- MBR Procedures to compute $\hat{W} = \underset{W' \in \mathcal{W}}{\operatorname{argmin}} \sum_{W \in \mathcal{W}} L(W, W') P(W|A)$
- Lattice rescoring via A^* search (Goel and Byrne: CSL '00)

Segmental Minimum Bayes-Risk Lattice Segmentation

- A^* search is expensive over large lattices
  - Pruning the lattices leads to search errors
  - Can we simplify the MBR decoder?
- Suppose we can segment the word lattice:



- Induced loss function: $L_I(W, W') = L(W_1, W_1') + L(W_2, W_2') + L(W_3, W_3')$
- MBR decoder can be decomposed into a sequence of segmental MBR decoders:

$$\hat{W} = \underset{W' \in \mathcal{W}_1}{\operatorname{argmin}} \sum_{W \in \mathcal{W}_1} L(W, W') P_1(W|A) \cdot \underset{W' \in \mathcal{W}_2}{\operatorname{argmin}} \sum_{W \in \mathcal{W}_2} L(W, W') P_2(W|A) \cdot \underset{W' \in \mathcal{W}_3}{\operatorname{argmin}} \sum_{W \in \mathcal{W}_3} L(W, W') P_3(W|A)$$

Trade-offs in Segmental MBR Lattice Segmentation

- MBR decoding on the entire lattice involves search errors
- Segmentation breaks up a single search problem into many simpler search problems
- An ideal segmentation: Loss between any two word strings unaffected by cutting
- Any segmentation restricts string alignments, and errors in approximating loss function between strings.

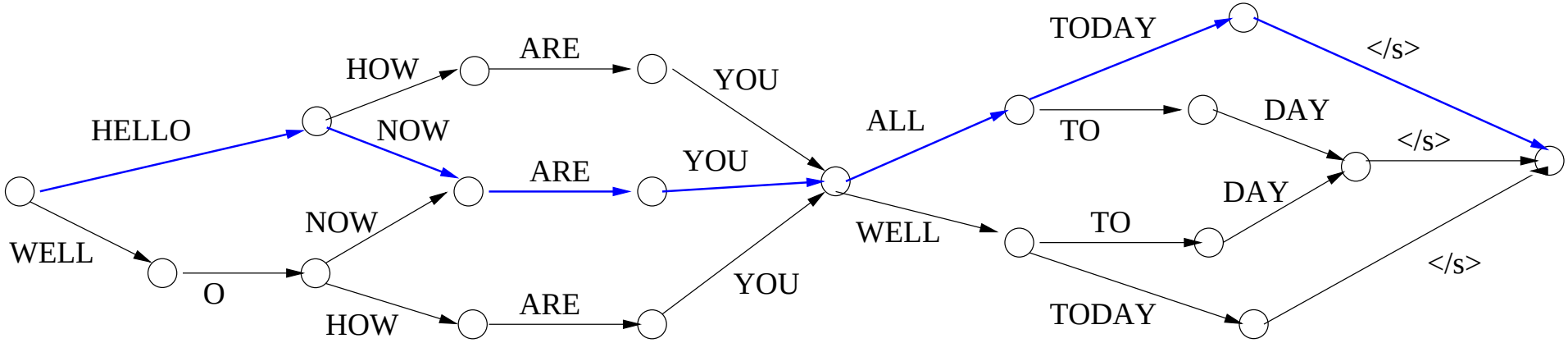
$$L(W, W') \leq \sum_{i=1}^{N} L(W_i, W_i')$$

- Therefore, segmentation involves tradeoff between search errors and errors in approximating the loss function
- Ideal segmentation criterion not achievable!
- Segmentation Rule: $L(\tilde{W}, W) = \sum_{i=1}^{K} L(\tilde{W}_i, W_i)$

Aligning a Lattice against a Word String

Motivation: Suppose we can align each word string in the lattice against $\tilde{W} = \tilde{w}_1^K$, we can segment the lattice into K segments

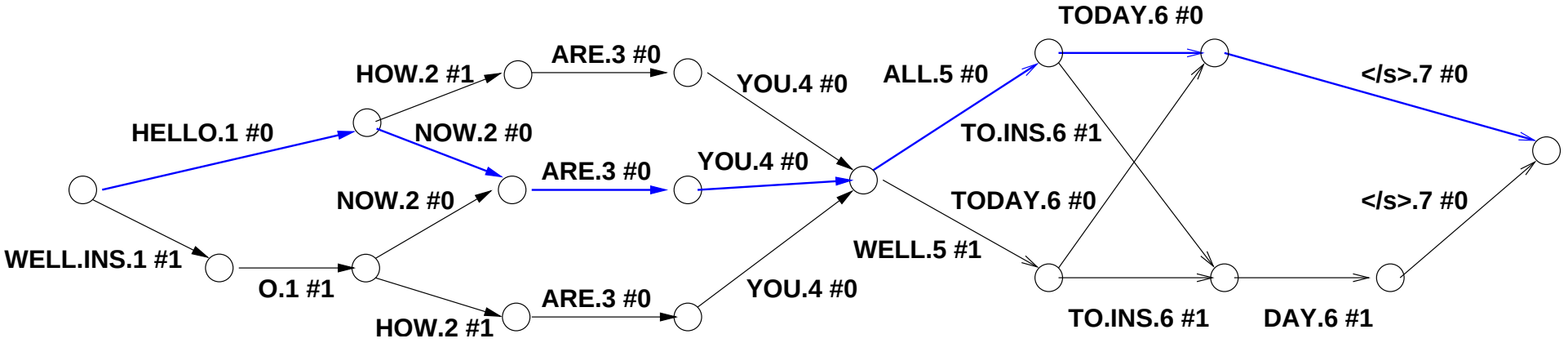
- Substrings in i^{th} set $\mathcal{W}_i$ will align with i^{th} word $\tilde{w}_i$
- We have developed an efficient (almost exact) procedure using Weight Finite State Transducers to generate the simultaneous string alignment of every string in the lattice wrt MAP hypothesis - this is encoded as an acceptor $\hat{\mathcal{A}}$
- Use alignment information from $\hat{\mathcal{A}}$ to segment the lattice into K sublattices



Aligning a Lattice against a Word String

Motivation: Suppose we can align each word string in the lattice against $\tilde{W} = \tilde{w}_1^K$, we can segment the lattice into K segments

- Substrings in i^{th} set $\mathcal{W}_i$ will align with i^{th} word $\tilde{w}_i$
- We have developed an efficient (almost exact) procedure using Weight Finite State Transducers to generate the simultaneous string alignment of every string in the lattice wrt MAP hypothesis - this is encoded as an acceptor $\hat{\mathcal{A}}$
- Use alignment information from $\hat{\mathcal{A}}$ to segment the lattice into K sublattices

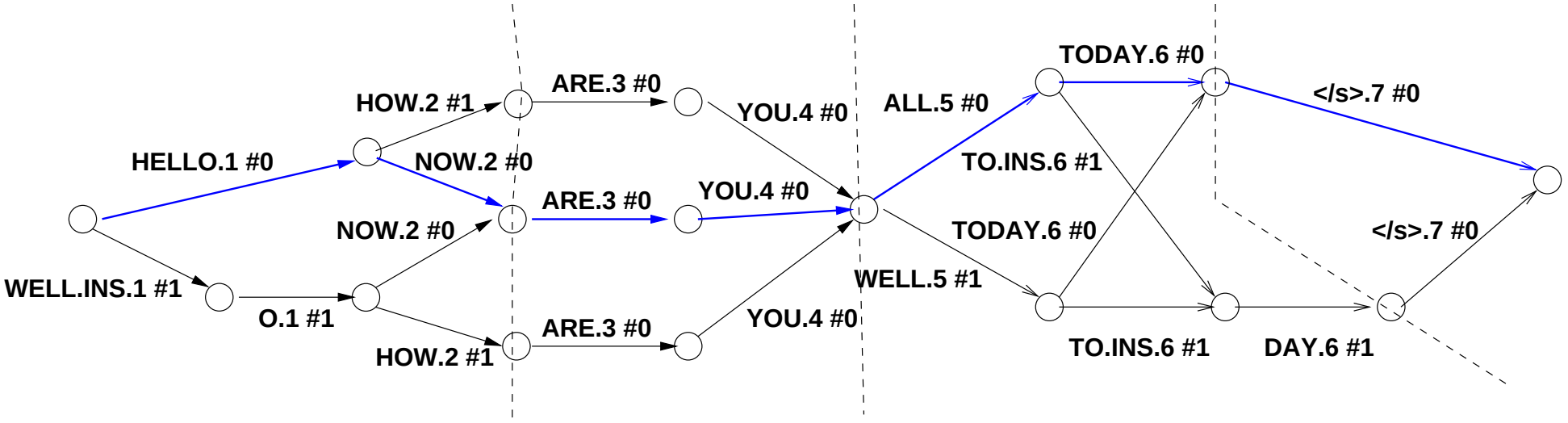


Periodic Risk-Based Lattice Cutting (PLC)

Segment the lattice into K segments relative to alignment against $\tilde{W} = \tilde{w}_1^K$

Properties

- Optimal wrt best path only : $L(W, W') \neq L_I(W, W')$ for $W \neq \tilde{W}$
- Segment the lattice along fewer cuts $\rightarrow$ Better approximations to loss function
- Solution: Segment Lattice into $< K$ segments by choosing cuts at equal periods

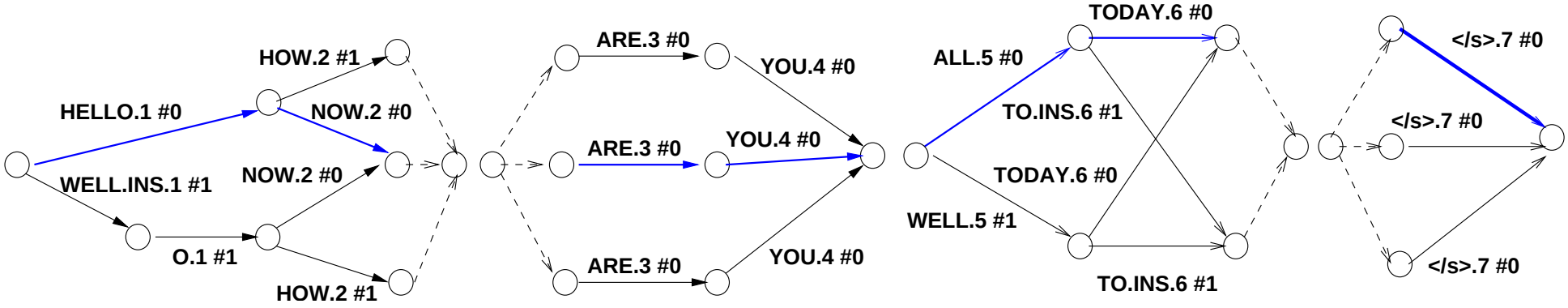


Periodic Risk-Based Lattice Cutting (PLC)

Segment the lattice into K segments relative to alignment against $\tilde{W} = \tilde{w}_1^K$

Properties

- Optimal wrt best path only : $L(W, W') \neq L_I(W, W')$ for $W \neq \tilde{W}$
- Segment the lattice along fewer cuts $\rightarrow$ Better approximations to loss function
- Solution: Segment Lattice into $< K$ segments by choosing cuts at equal periods



Recognition Performance of MBR Classifiers

- Task: SWITCHBOARD Large Vocabulary ASR (JHU 2001 Evaluation System)
- Test Sets: SWB1 (1831 utterances) and SWB2 (1755 utterances)
- MBR decoding strategy: A^* search on lattices

Decoder		WER(%)	
		SWB2	SWB1
MAP (baseline)		41.1	26.0
MBR Decoding			
Segmentation Strategy	Properties		
No Cutting (Period ∞)	search errors, no approx to loss function	40.4	25.5
PLC (Period 6)	intermediate	40.0	25.4
PLC (Period 1)	no search errors, poor approx to loss function	41.0	25.9

- Segmental MBR decoding performs better than MAP decoding or MBR decoders on unsegmented lattices
- Segmental MBR decoder performs better under PLC-6 compared to PLC-1

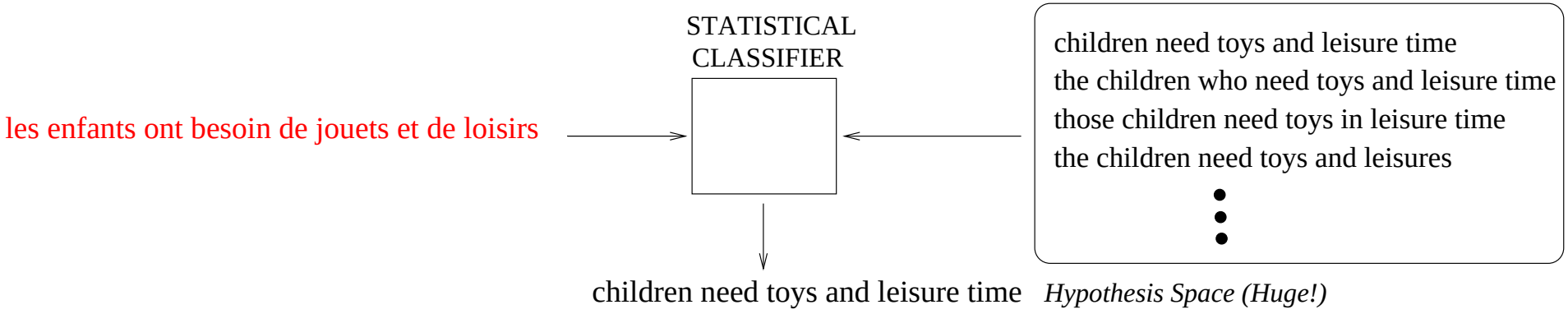
Outline

- Automatic Speech Recognition
  - Minimum Bayes-Risk Classifiers
  - Segmental Minimum Bayes-Risk Classification
  - Risk-Based Lattice Segmentation
- Statistical Machine Translation
  - A Statistical Translation Model
  - Minimum Bayes-Risk Classifiers for Word Alignment of Bilingual Texts
  - Minimum Bayes-Risk Classifiers for Machine Translation
- Conclusions and Future Work

Introduction to Statistical Machine Translation

Statistical Machine Translation :

Map a string of words in a source language (e.g. French) to a string of words in a target language (e.g. English) via statistical approaches



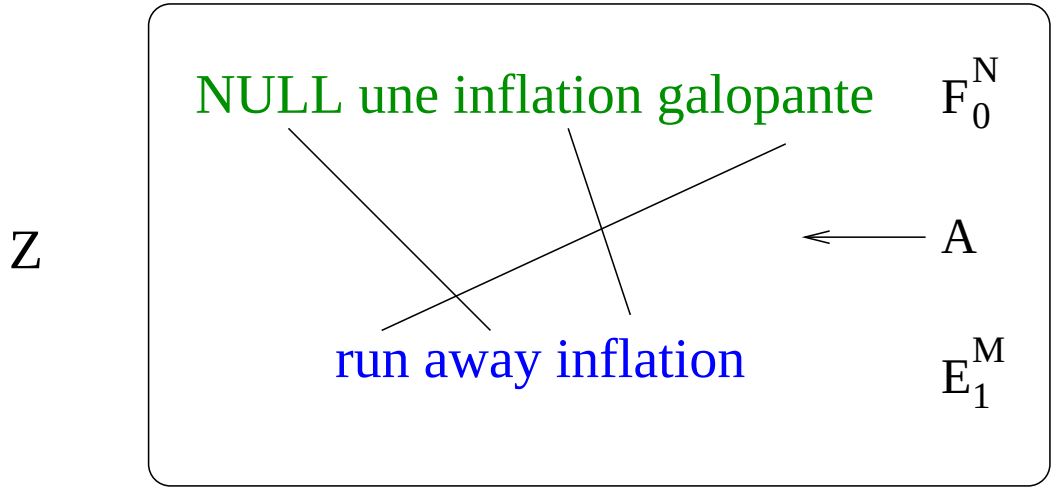
- Two sub-tasks of Machine Translation
  - Word-to-Word alignment of bilingual texts
  - Translation of sentences from source language to target language

Alignment Template Translation Model

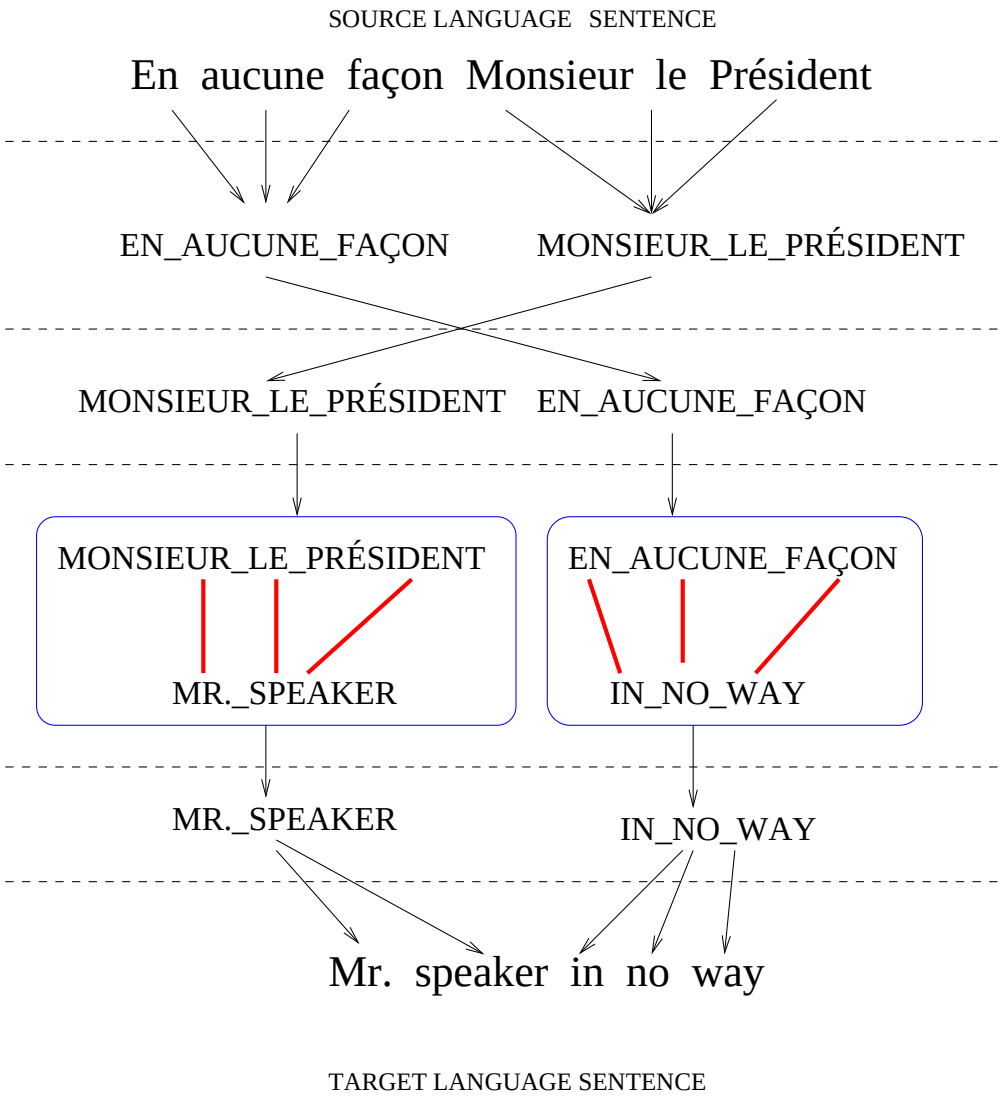
- Alignment Template Translation Model (ATTM) (Och, Tillmann and Ney '99) has emerged as a promising model for Statistical Machine Translation

What are Alignment Templates?

- Alignment Template $z = (E_1^M, F_0^N, A)$ specifies word alignments between word sequences E_1^M and F_0^N through a possible 0/1 valued matrix A.
- Alignment Templates map short word sequences in source language to short word sequences in target language



Alignment Template Translation Model Architecture



Component Models

Source Segmentation Model

Phrase Permutation Model

Template Sequence Model

Phrasal Translation Model

Weighted Finite State Transducer Translation Model

Reformulate the ATTM so that bitext-word alignment and translation can be implemented using Weighted Finite State Transducer (WFST) operations

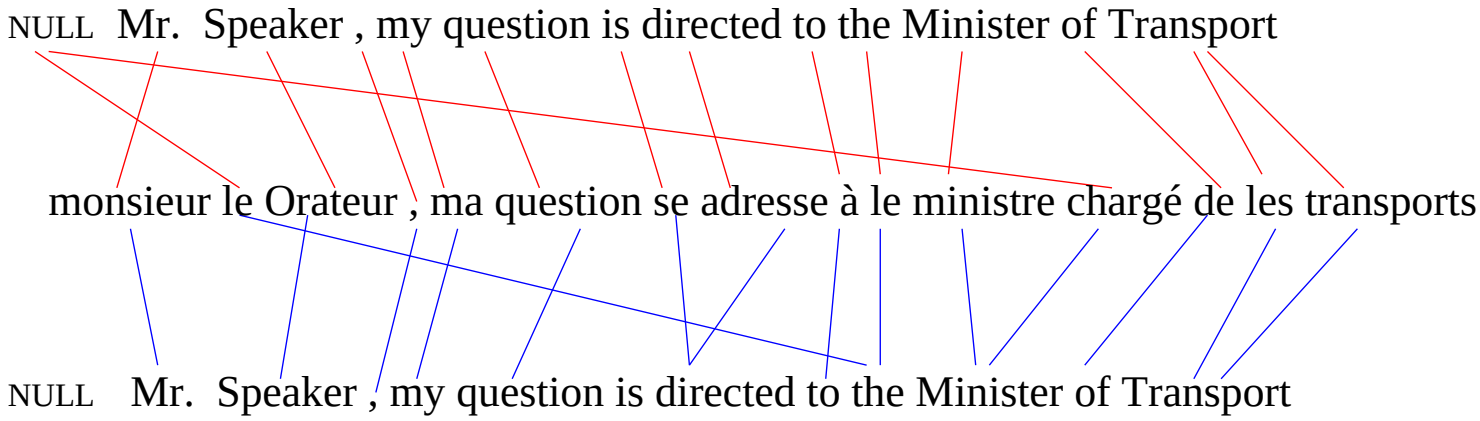
- Modular Implementation: Statistical models are trained for each model component and implemented as WFSTs
- WFST implementation makes it unnecessary to develop a specialized decoder
  - This decoder can even generate translation lattices and N-best lists
- WFST architecture provides support for generating bitext word alignments and alignment lattices
  - Novel approach!
  - Allows development of parameter re-estimation procedures
- Good performance in the NIST 2003 Chinese-English and Hindi-English MT Evaluations

Outline

- Automatic Speech Recognition
  - Minimum Bayes-Risk Classifiers
  - Segmental Minimum Bayes-Risk Classification
  - Risk-Based Lattice Segmentation
- Statistical Machine Translation
  - A Statistical Translation Model
  - Minimum Bayes-Risk Classifiers for Word Alignment of Bilingual Texts
  - Minimum Bayes-Risk Classifiers for Machine Translation
- Conclusions and Future Work

Word-to-Word Bitext Alignment

Competing Alignments for an English-French Sentence Pair



- Basic Terminology
  - (e_0^l, f_1^m) : An English-French Sentence Pair
  - *Alignment Links*: $b = (i, j)$: f_i linked to e_j
  - Alignment is defined by a *Link Set* $B = \{b_1, b_2, ..., b_m\}$
  - Some links are NULL links
- Given a candidate alignment B' and the reference alignment B, $L(B, B')$ is the *loss function* that measures B' wrt B.

MBR Word Alignments of Bilingual Texts

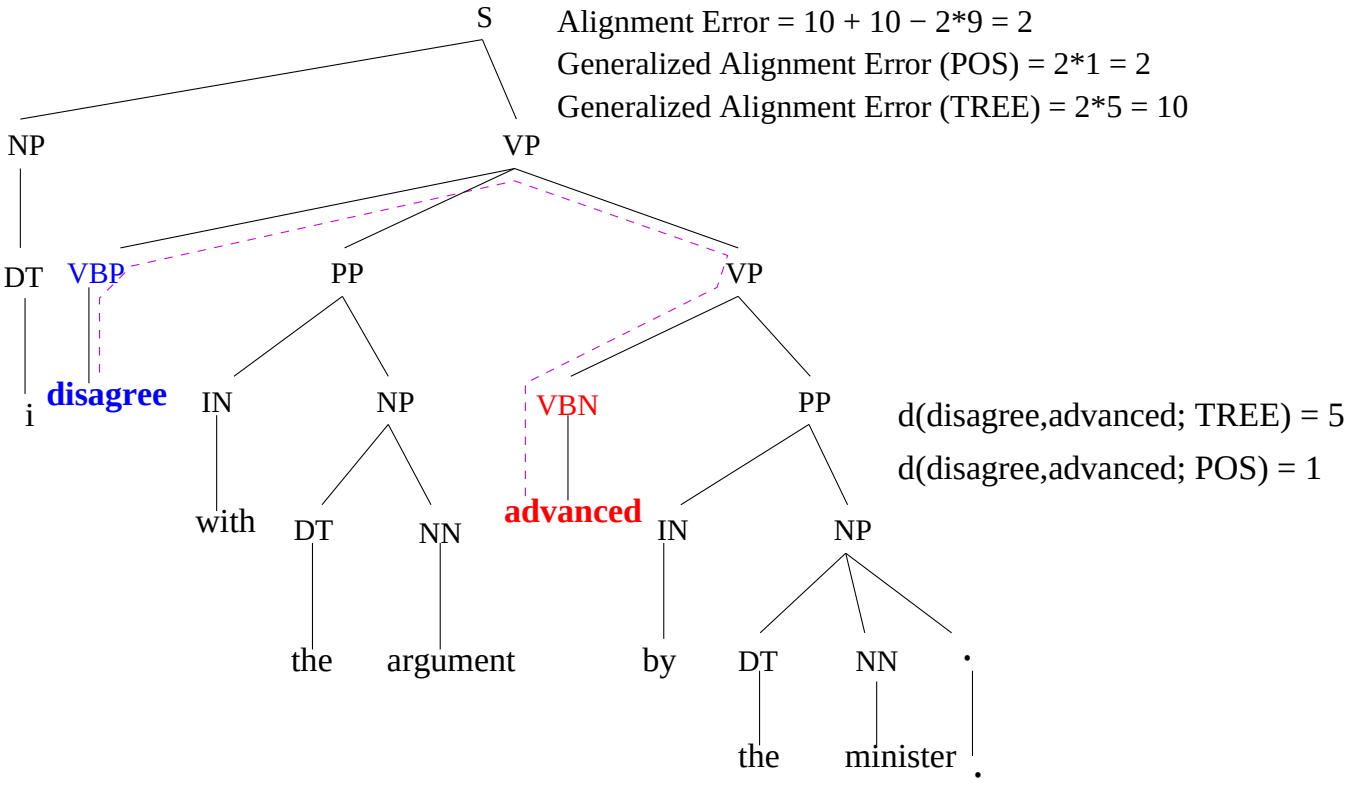
- Word-to-Word alignments of Bilingual texts are important components of an MT system
  - Alignment Templates are constructed from word alignments
  - Better alignments lead to better templates and therefore better translation performance
- Alignment loss functions to measure alignment quality
  - Different loss functions capture different features of alignments
  - Loss functions can use information from word-to-word links, parse-trees and POS tags - These are ignored by most of the current translation models
- Minimum Bayes-Risk (MBR) Alignments under each loss function
  - Performance gains by tuning alignment to the evaluation criterion

Loss functions for Bitext word alignment

- *Alignment Error* measures # of non-NULL alignment links by which the candidate alignment differs reference alignment
  - Derived from Alignment Error Rate (Och and Ney '00)
  - $L_{AE}(B, B') = |\bar{B}| + |\bar{B}'| - 2|\bar{B} \cap \bar{B}'|$
- *Generalized Alignment Error*: Extension of Alignment Error loss function to incorporate linguistic features
  $L_{GAE}(B, B') = 2 \sum_{b \in B} \sum_{b' \in B'} \delta_i(i') d_{ijj'}$ where $b = (i, j), b' = (i', j')$
  - Word-to-Word Distance Measure $d_{ijj'} = D((j, e_j), (j', e_{j'}); f_i)$ can be constructed using information from parse-trees or Part-of-Speech (POS) tags.
  - L_{GAE} can be almost reduced to L_{AE}
  - Example using Part-of-Speech Tags

$$d_{ijj'} = \begin{cases} 0 & \text{POS}(e_j) = \text{POS}(e_{j'}) \\ 1 & otherwise. \end{cases}$$

Examples of Word Alignment Loss Function



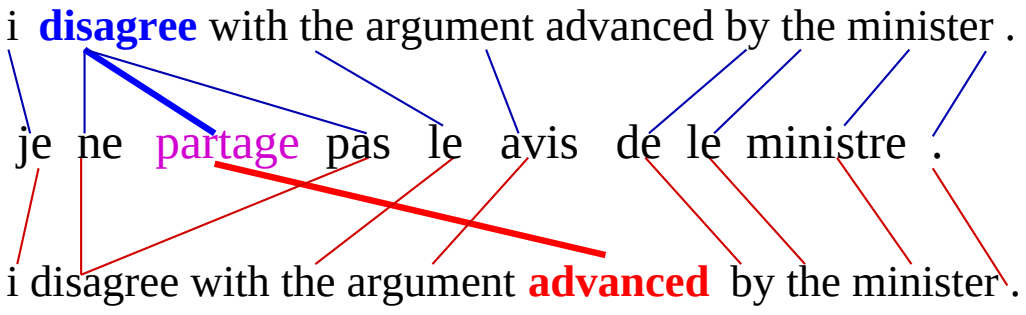
Alignment Error = 10 + 10 − 2*9 = 2

Generalized Alignment Error (POS) = 2*1 = 2

Generalized Alignment Error (TREE) = 2*5 = 10

d(disagree,advanced; TREE) = 5

d(disagree,advanced; POS) = 1



Minimum Bayes-Risk Decoding for Automatic Word Alignment

- Introduce a statistical model over alignments of a sentence pair (e, f) :$P(B|f, e)$
- MBR decoder

$$\hat{B} = \underset{B' \in \mathcal{B}}{\operatorname{argmin}} \sum_{B \in \mathcal{B}} L(B, B') P(B|f, e)$$

- $\mathcal{B}$ is the set of all alignments of (e, f)
  - This is approximated by the alignment lattice: the set of the most likely word alignments
- We have derived closed form expressions for the MBR decoder under two classes of alignment loss functions
  - Allows exact and efficient implementation of the lattice search

Minimum Bayes-Risk Alignment Experiments

Experiment Setup

- Training Data: 50,000 sentence pairs from French-English Hansards
- Test Data: 207 unseen sentence pairs from Hansards
- Evaluation: Measure error rates wrt human word alignments

			Generalized Alignment Error Rates	
	Decoder	AER (%)	TREE (%)	POS (%)
	ML	18.13	29.39	51.36
M	AE	14.87	19.81	36.42
B	GAE-TREE	23.26	14.45	26.76
R	GAE-POS	28.60	15.70	26.28

MBR decoder tuned for a loss function
performs the best under the corresponding error rate

Outline

- Automatic Speech Recognition
  - Minimum Bayes-Risk Classifiers
  - Segmental Minimum Bayes-Risk Classification
  - Risk-Based Lattice Segmentation
- Statistical Machine Translation
  - A Statistical Translation Model
  - Minimum Bayes-Risk Classifiers for Word Alignment of Bilingual Texts
  - Minimum Bayes-Risk Classifiers for Machine Translation
- Conclusions and Future Work

Loss functions for Machine Translation

- Automatic Evaluation of Machine Translation - Hard Problem!
- BLEU (Papineni et.al 2001) is an automatic MT metric - Shown to correlate well with human judgements on translation
- Other Metrics: Word Error Rate (WER) &
  Position Independent Word Error Rate (PER) : Minimum String edit distance between a reference sentence and any permutation of the hypothesis sentence

Loss function

Reference : mr. speaker , in absolutely no way .

Hypothesis : in absolutely no way , mr. chairman .

Sub-string Matches(Truth,Hyp)

1-word	2-word	3-word	4-word
7/8	3/7	2/6	1/5

Evaluation Metric(Truth,Hyp) (%)

BLEU	WER	PER
39.76%	6/8 = 75.0%	1/8 = 12.5%

BLEU computation: $\left(\frac{7}{8} \times \frac{3}{7} \times \frac{2}{6} \times \frac{1}{5}\right)^{\frac{1}{4}} = 0.3976$

Minimum Bayes-Risk Machine Translation

- Given a loss function, we can build Minimum Bayes-Risk Classifiers to optimize performance under the loss function.
- Setup
  - A baseline translation model to give the probabilities over translations: $P(E|F)$
  - A set $\mathcal{E}$ of N-Best Translations of F
  - A Loss function $L(E, E')$ that measures the the quality of a candidate translation E' relative to a reference translation E
- MBR Decoder

$$\hat{E} = \underset{E' \in \mathcal{E}}{\operatorname{argmin}} \sum_{E \in \mathcal{E}} L(E, E') P(E|F)$$

Performance of MBR Decoders for Machine Translation

- Experimental Setup: WS'03 - CLSP summer workshop
- Test Set: Chinese-English NIST MT Task (2002) , 878 sentences, 1000-best lists

		Performance Metrics		
		BLEU (%)	mWER(%)	mPER (%)
	MAP(baseline)	31.6	62.4	39.3
M	PER	31.7	62.2	38.5
B	WER	31.8	61.8	38.8
R	BLEU	31.9	62.5	39.2

- MBR Decoding allows translation process to be tuned for specific loss functions

Conclusions : Minimum Bayes-Risk Techniques

- Unified classification framework for two different tasks in speech and language processing
- Techniques are general and can be applied to a variety of scenarios
- Need design of various loss functions that measure task-dependent error rates
- Can optimize performance under task-dependent metrics

Conclusions : Segmental Minimum Bayes-Risk Lattice Segmentation

- Segmental MBR Classification and Lattice Cutting decompose a large utterance level MBR recognizer into a sequence of simpler sub-utterance level MBR recognizers
- Risk-Based Lattice Segmentation - robust and stable technique
- Basis for novel discriminative training procedures in ASR (Doumpiotis, Tsakalidis and Byrne ’03)
- Basis for novel classification schemes using Support Vector Machines for ASR (Venkataramani, Chakrabartty and Byrne ’03)
- Future Work: Investigate applications within the MALACH ASR project

Conclusions: Machine Translation

- The Weighted Finite State Transducer Alignment Template Translation Model
  - Powerful modeling framework for Machine Translation
  - A novel approach to generate word alignments and alignment lattices under this model
- MBR classifiers for bitext word alignment and translation
  - Alignment and translation can be tuned under specific loss functions
  - Syntactic features from English parsers and Part-of-Speech taggers can be integrated into a statistical MT system via appropriate definition of loss functions

Proposed Research

- Refinements to the Alignment Template Translation Model
  - Iterative parameter re-estimation via Expectation Maximization procedures
    - Model currently initialized from bitext word alignments
    - Alignment Lattices : Posterior Distributions over hidden variables
    - Expect improvements in alignment and translation performance
  - Reformulation as a source-channel model
  - New strategies for template selection
- MBR Classifiers for Bitext Word Alignment and Translation
  - Loss functions based on detailed models of translation
  - Extend search space to Translation Lattices

Thank you!

References

- V. Goel and W. Byrne 2000. Minimum Bayes-Risk Decoding for Automatic Speech Recognition, *Computer, Speech and Language*
- S. Kumar and W. Byrne 2002. Risk-Based Lattice Cutting for Segmental Minimum Bayes-Risk Decoding, *Proceedings of the International Conference on Spoken Language Processing*, Denver CO.
- V. Goel, S. Kumar and W. Byrne 2003. Segmental Minimum Bayes-Risk Decoding for Automatic Speech Recognition, *IEEE Transactions on Speech and Audio Processing, To appear*
- S. Kumar and W. Byrne 2002. Minimum Bayes-Risk Word Alignments of Bilingual Texts, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA
- S. Kumar and W. Byrne 2003. A Weighted Finite State Transducer Implementation of the Alignment Template Model for Statistical Machine Translation, *Proceedings of the Conference on Human Language Technology* , Edmonton, AB, Canada