

# *Cross-Lingual Lexical Triggers in Statistical Language Modeling*

July 11-12, 2003

Woosung Kim and Sanjeev Khudanpur  
woosung@cs.jhu.edu khudanpur@jhu.edu

Presentation by Peng Xu

Center for Language and Speech Processing  
The Johns Hopkins University, Baltimore, MD 21218, USA

# Introduction

- ❑ Motivation :
  - ✓ Success of statistical modeling techniques
    - ❖ Development of modeling and automatic learning techniques
    - ❖ A large amount of data for training is available
    - ❖ Most resources on English, French and German
- ❑ 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)
  - ✓ Language modeling (this talk)

# Introduction

- We present:
  - ✓ An approach to sharpen an LM in a resource-deficient language using comparable text from resource-rich languages

# Introduction

- We present:
  - ✓ An approach to sharpen an LM in a resource-deficient language using comparable text from resource-rich languages
  - ✓ Story-specific language models from contemporaneous text

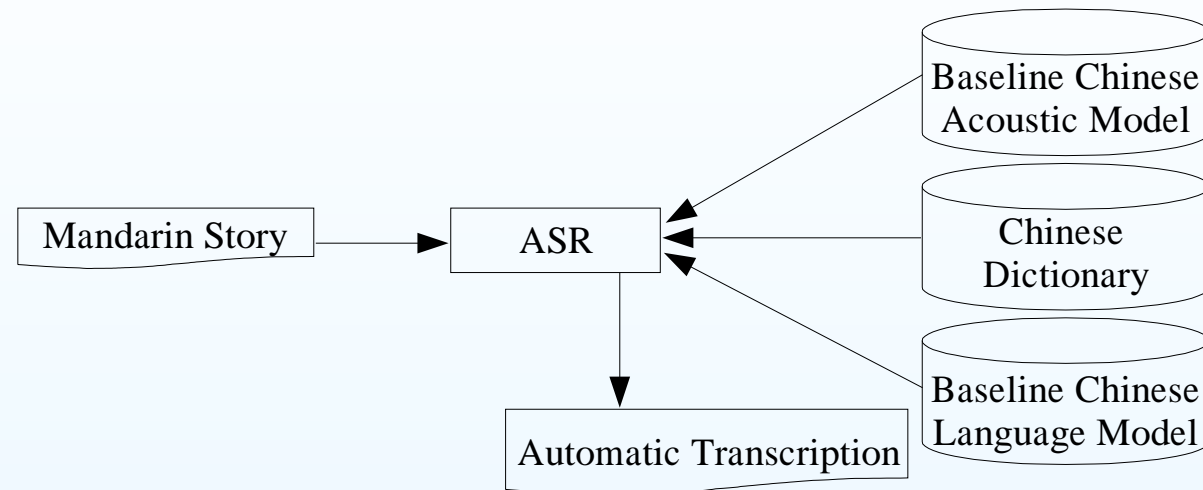
# Introduction

- We present:
  - ✓ An approach to sharpen an LM in a resource-deficient language using comparable text from resource-rich languages
  - ✓ Story-specific language models from contemporaneous text
  - ✓ Integration of machine translation (MT), cross-language information retrieval (CLIR), and language modeling (LM)

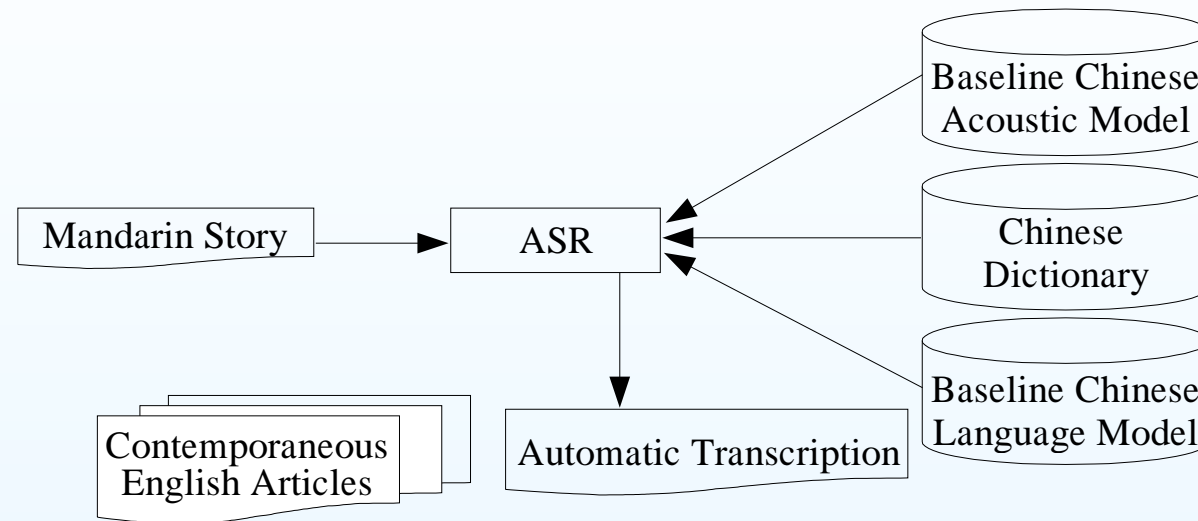
# Introduction

- ❑ We present:
  - ✓ An approach to sharpen an LM in a resource-deficient language using comparable text from resource-rich languages
  - ✓ Story-specific language models from contemporaneous text
  - ✓ Integration of machine translation (MT), cross-language information retrieval (CLIR), and language modeling (LM)
  - ✓ Cross-lingual lexical triggers instead of MT dictionary → Document-aligned corpus

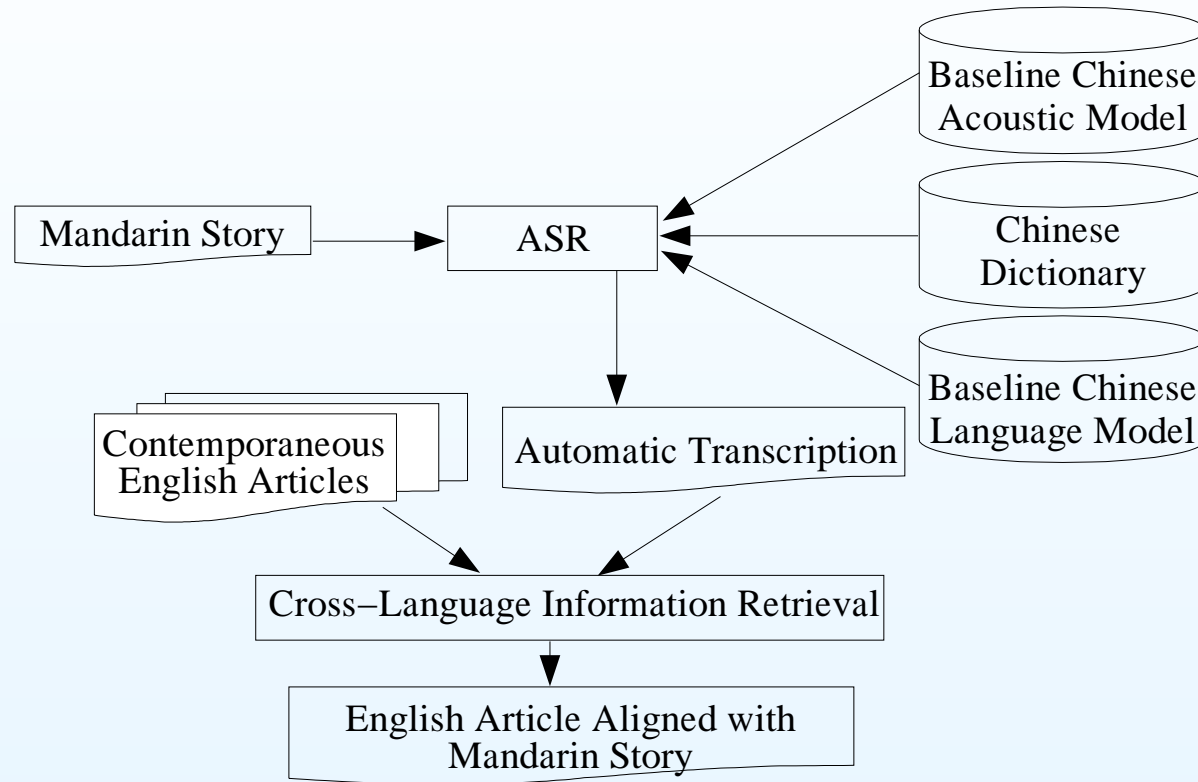
# Cross-Lingual LM for ASR



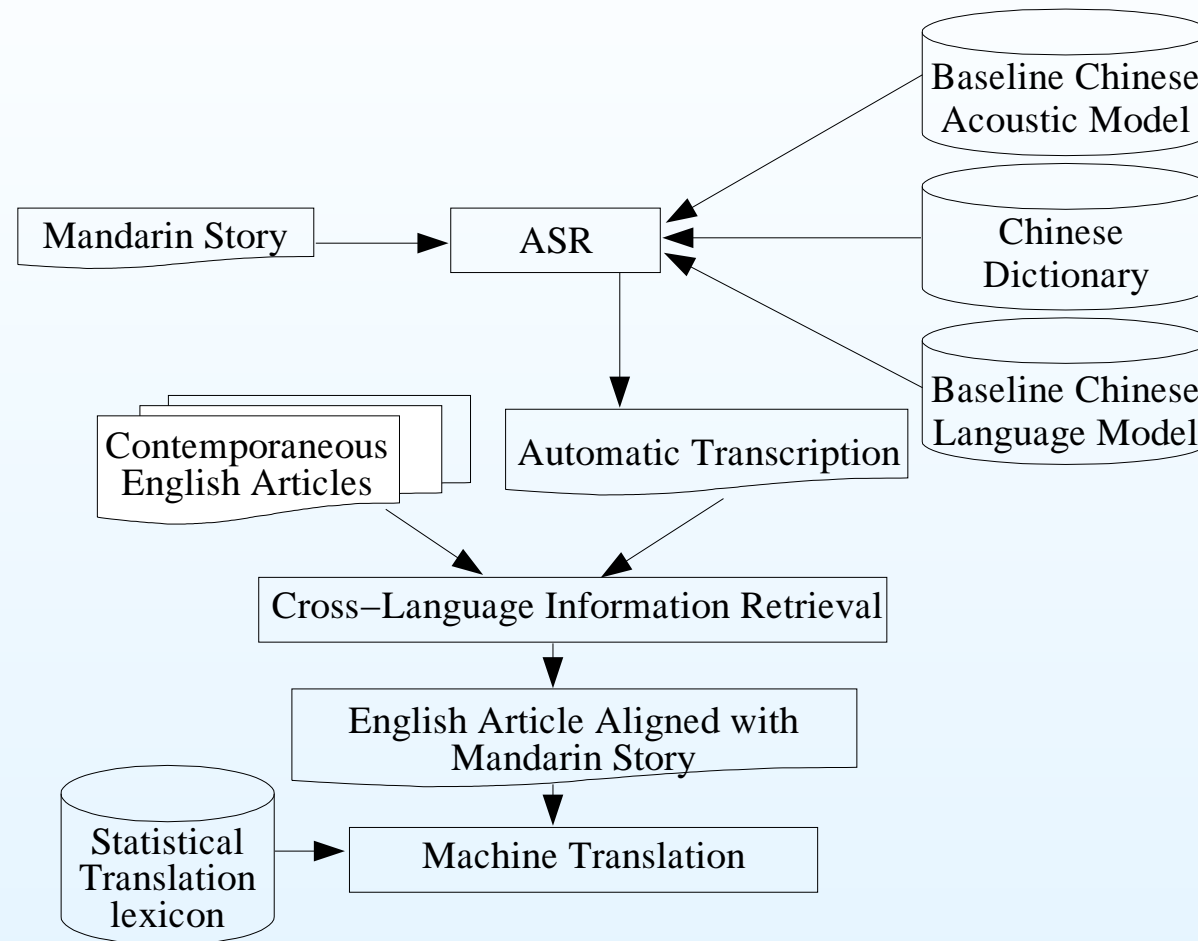
# Cross-Lingual LM for ASR



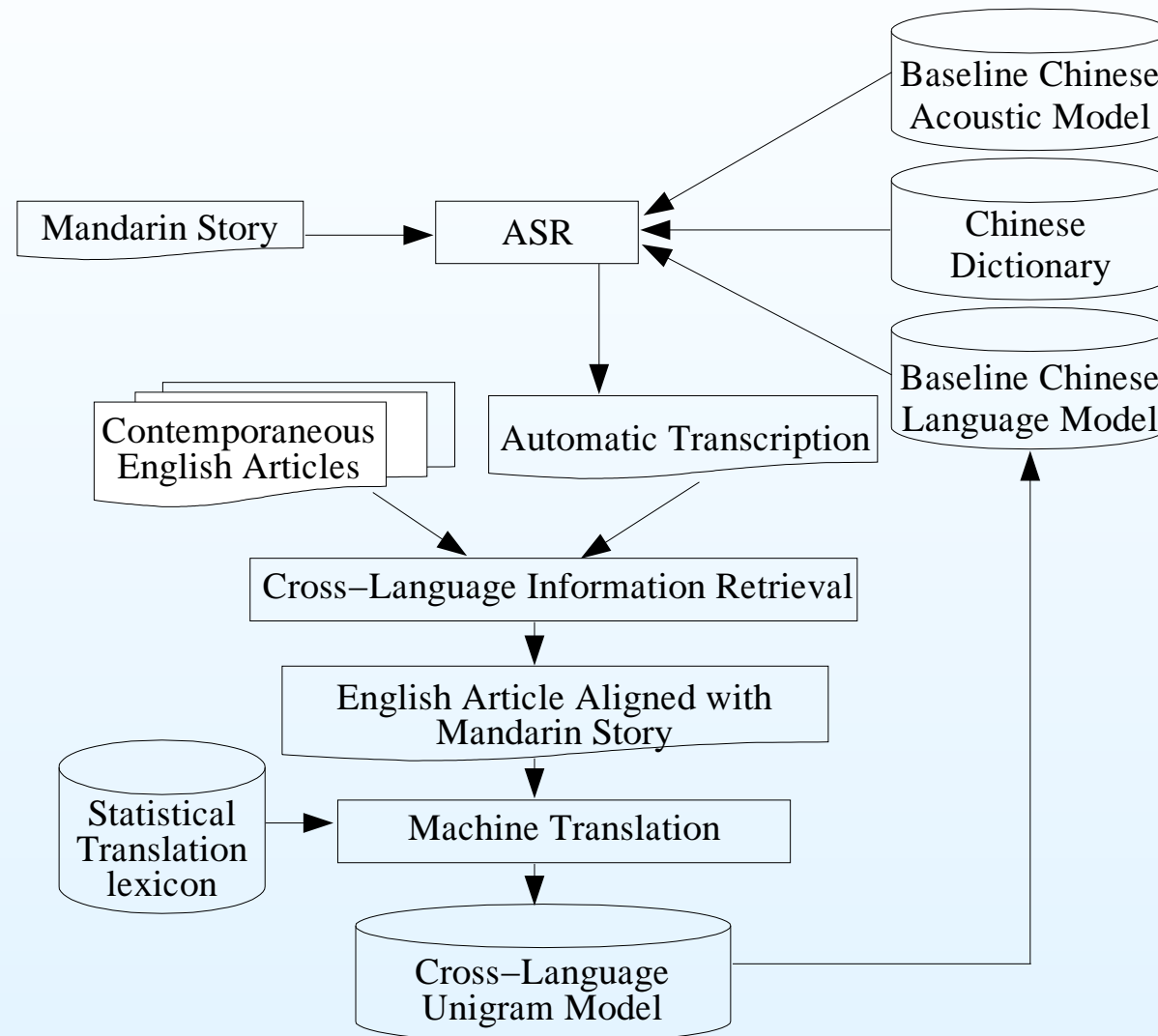
# Cross-Lingual LM for ASR



# Cross-Lingual LM for ASR



# Cross-Lingual LM for ASR



## Model Estimation

- Assume document correspondence,  $d_i^E \leftrightarrow d_i^C$ , is known for Chinese test doc  $d_i^C$ ,

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

- Cross-Language LM construction
  - ✓ Build story-specific cross-language LMs,  $P(c|d_i^E)$
  - ✓ Linear interpolation with the baseline trigram LM

$$\begin{aligned} P_{\text{CL-interpolated}}(c_k|c_{k-1}, c_{k-2}, d_i^E) \\ = \lambda P_{\text{CL-unigram}}(c_k|d_i^E) + (1 - \lambda) P(c_k|c_{k-1}, c_{k-2}) \end{aligned}$$

## Model Estimation

- Document correspondence → obtained by CLIR
  - ✓ For each Chinese test doc  $d_i^C$ , create English bag-of-words

$$P_{\text{CL-unigram}}(e|d_i^C) = \sum_{c \in \mathcal{C}} P_T(e|c) \hat{P}(c|d_i^C)$$

- ✓ Use it to find the English doc with the highest cosine similarity (TF-IDF weight)

$$d_i^E = \operatorname{argmax}_{d_j^E} \operatorname{sim}(P_{\text{CL-unigram}}(e|d_i^C), \hat{P}(e|d_j^E)).$$

## Translation Lexicons

- ❑ Estimation of  $P_T(c|e)$  and  $P_T(e|c)$  → GIZA++ translation table
  - ✓ GIZA++ : statistical MT tool based on IBM model-4
  - ✓ Input : Hong Kong news Chinese-English parallel corpus
    - 18K docs, 200K sents, 4M wds each
  - ✓ Output : MT system with several tables
    - Only translation tables are used :  $P(e|c)$  and  $P(c|e)$
  - ✓ Sentence-aligned parallel corpus is needed

## Cross-Lingual Lexical Triggers

- ❑ Translation lexicons : expensive to acquire
- ❑ Mutual information-based lexical triggers : successful in a monolingual setting (Tillmann & Ney, 97)
- ❑ Cross-lingual triggers : a trigger pair consists of  $(e, c)$
- ❑ Identification of CL triggers : Average Mutual Information

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

where  $P(e, c) = \frac{\#d(e,c)}{N}$  and  $P(e, \bar{c}) = \frac{\#d(e,\bar{c})}{N}$

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

## Estimation of Trigger LM Probs

- Maximum likelihood estimation vs. ad hoc estimation

$$P_{\text{Trig}}^{\text{MLE}}(c|e) = \frac{\sum_{i : d_i^E \ni e} N_{d_i^C}(c)}{\sum_{c' \in \mathcal{C}} \sum_{i : d_i^E \ni e} N_{d_i^C}(c')}$$

$$P_{\text{Trig}}^{\text{ad hoc}}(c|e) = \frac{I(e; c)}{\sum_{c' \in \mathcal{C}} I(e; c')}$$

- Interpolation with trigram model :

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

$$P_{\text{Trig-interpolated}}(c_k | c_{k-1}, c_{k-2}, d_i^E) =$$

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

## Topic-Dependent LMs

- ❑ Topic clustering : standard K-means clustering
- ❑ 100 classes for target lang : Chinese
- ❑ Interpolation with baseline trigram LM

$$\begin{aligned} P_{\text{Topic-trigram}}(c_k | c_{k-1}, c_{k-2}, t_i) \\ = \lambda P_{t_i}(c_k | c_{k-1}, c_{k-2}) + (1 - \lambda) P(c_k | c_{k-1}, c_{k-2}) \end{aligned}$$

- ❑ Interpolation with baseline trigram LM + CL-unigram

$$\begin{aligned} P_{\text{Topic+CL}}(c_k | c_{k-1}, c_{k-2}, t_i) \\ = \lambda_1 P_{t_i}(c_k | c_{k-1}, c_{k-2}) + \lambda_2 P_{\text{CL-unigram}}(c_k | d_i^E) + \\ (1 - \lambda_1 - \lambda_2) P(c_k | c_{k-1}, c_{k-2}) \end{aligned}$$

## Training and Test Corpora

- ❑ 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
  - ✓ NAB (1997 LA, WP) + TDT-2 (1998 APW, NYT)
  - ✓ 45K docs, 30M wds
- ❑ Vocab : 51K for Chinese

## ASR Experimental Results : Baseline

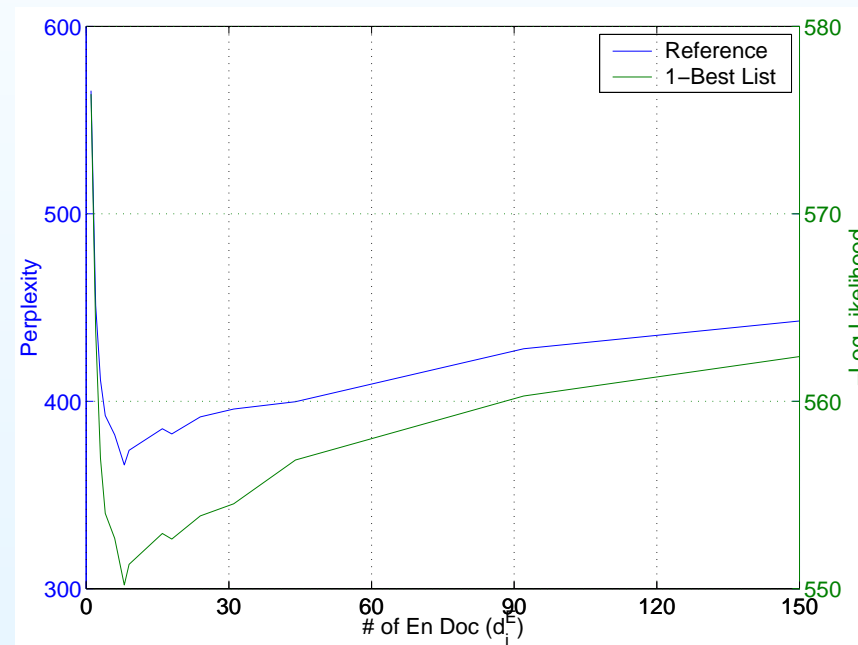
- ❑ Bigrams LMs used for for each lattice generation
- ❑ 300-best list rescoring
- ❑ Single English document and global  $\lambda$  are used

Language model	Perplexity	WER	$p$ -value
XINHUA trigram	426	49.9%	–
CL-interpolated	375	49.5%	0.208
HUB-4NE trigram	1195	60.1%	–
CL-interpolated	750	59.3%	< 0.001

- ❑ Trigger probability estimation
  - ➔ Ad hoc ( $P_{\text{Trig}}^{\text{ad hoc}}$ ) is slightly better than MLE ( $P_{\text{Trig}}^{\text{MLE}}$ ) (367 vs. 370 on XINHUA, 727 vs. 736 on HUB4-NE)

## Likelihood-Based Selection of English Doc and $\lambda$

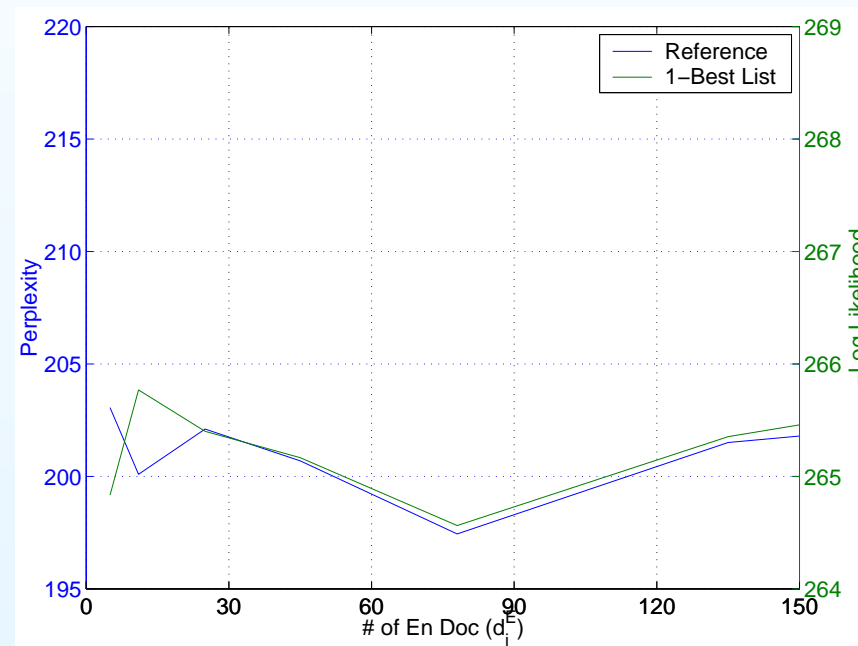
$$P_{\text{CL-interpolated}}(c_k | c_{k-1}, c_{k-2}, d_i^E) \\ = \lambda P_{\text{CL-unigram}}(c_k | d_i^E) + (1 - \lambda) P(c_k | c_{k-1}, c_{k-2})$$



Perplexity (Likelihood) v/s # En Doc ( $d_i^E$ ) for Sample Story 1

## Likelihood-Based Selection of English Doc and $\lambda$

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



Perplexity (Likelihood) v/s # En Doc ( $d_i^E$ ) for Sample Story 2

## ASR Experimental Results

- ❑ LM text : XINHUA corpus (13M wds)
- ❑ Lattices generated from XINHUA bigram LM
- ❑ Topic clustering with XINHUA corpus

Language model	Perplexity	WER	CER	<i>p</i> -value
XINHUA trigram	426	49.9%	28.8%	–
Topic-trigram	381	49.1%	28.4%	0.003
Trig-interpolated	367	49.1%	28.6%	0.004
CL-interpolated	346	48.8%	28.4%	< 0.001
Topic + Trig-interp.	340	48.7%	28.4%	< 0.001
Topic + CL-interp.	326	48.5%	28.2%	< 0.001
Topic + Trig- + CL-interp.	320	48.3%	28.1%	< 0.001

## Experiments in a Resource-Deficient Setting

- ❑ Extreme case : HUB-4 acoustic model training data
- ❑ Lattices generated from HUB-4 bigram LM
- ❑ Topic clustering with HUB-4 corpus → topic doesn't help

Language model	Perplexity	WER	CER	<i>p</i> -value
HUB4-NE trigram	1195	60.1%	44.1%	–
Topic-trigram	1122	60.0%	44.1%	0.660
Trig-interpolated	727	58.8%	43.3%	< 0.001
CL-interpolated	630	58.8%	43.1%	< 0.001
Topic + Trig-interp.	730	59.2%	43.5%	0.002
Topic + CL-interp.	631	59.0%	43.3%	< 0.001
Topic + Trig- + CL-interp.	627	59.0%	43.3%	< 0.001

## Conclusions

- ❑ Exploits side-information from contemporaneous articles  
→ useful for resource-deficient languages
- ❑ Statistically significant improvements in ASR WER (1.4% absolute)
- ❑ Comparable improvements over within-language topic-dependant LMs; further gains from combination
- ❑ Use of CL trigger pairs → sentence-aligned corpus is no longer needed
- ❑ Future work
  - ✓ Maximum entropy models
  - ✓ Application to other languages : e.g. Arabic
  - ✓ Application to other tasks : e.g. machine translation

## References

- ❑ T. Schultz and A. Waibel. 1998. Language independent and language adaptive large vocabulary speech recognition. *Proceedings of the International Conference on Spoken Language Processing*, 5:1819-1822, Sydney, Australia.
- ❑ W. Byrne, P. Beyerlein, J. Huerta, S. Khudanpur, B. Marathi, J. Morgan, N. Peterek, J. Picone, D. Vergyri and W. Wang. 2000. Towards language independent acoustic modeling. *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, 2:1029 - 1032, Istanbul, Turkey.
- ❑ D. Yarowsky, G. Ngai and R. Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. *Proceedings of the Human Language Technologies Workshop*, pages 109-116, Santa Monica, CA.
- ❑ Hong Kong News parallel text corpus. 2000. Available through the Linguistic Data Consortium. <http://www ldc.upenn.edu/Catalog/LDC2000T46.html>
- ❑ Matched Pairs Sentence-Segment Word Error (MAPSSWE) Test. <http://www.nist.gov/speech/tests/sigttests/mapsswe.htm>
- ❑ C. Tillmann and H. Ney. 1997. Word trigger and the EM algorithm. In *Proceedings of the Workshop Computational Natural Language Learning (CoNLL 97)*, pages 117–124, Madrid, Spain.