

Cross-Lingual Priming of Language Models for Speech Recognition

February 7, 2003

Woosung Kim and Sanjeev Khudanpur

woosung@cs.jhu.edu khudanpur@jhu.edu

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

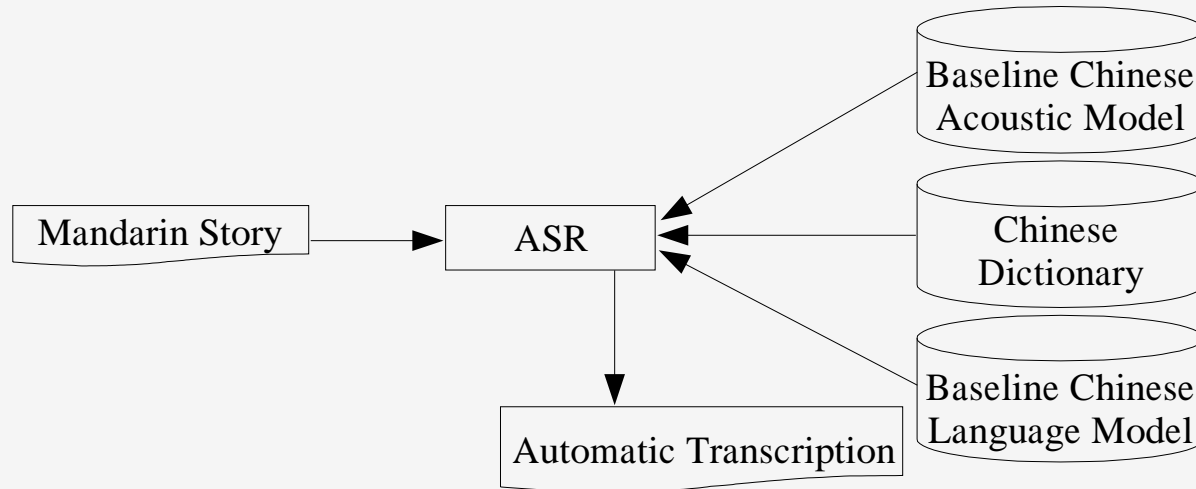
Introduction

- 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

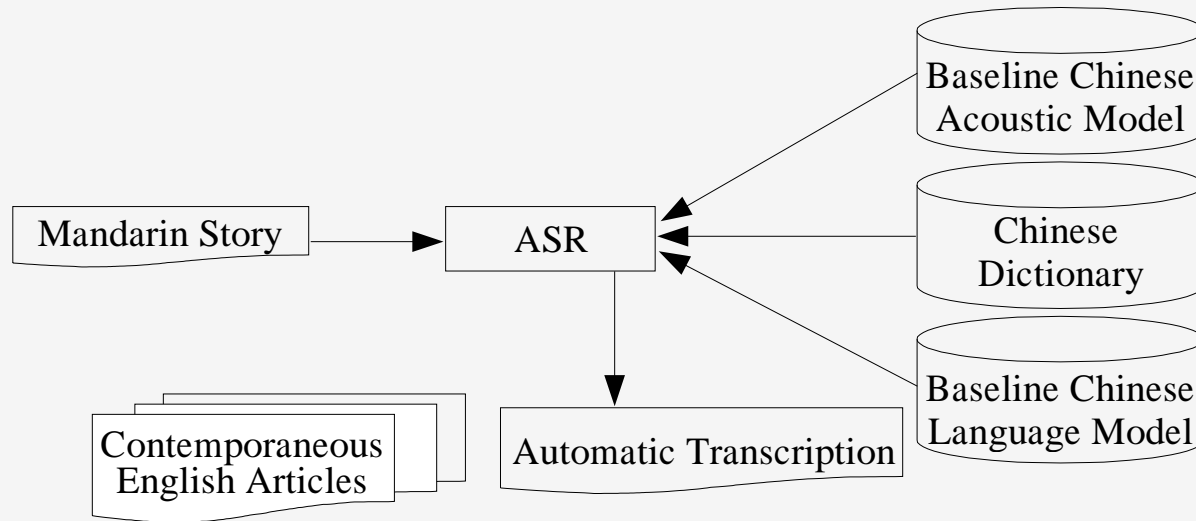
Introduction

- 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)

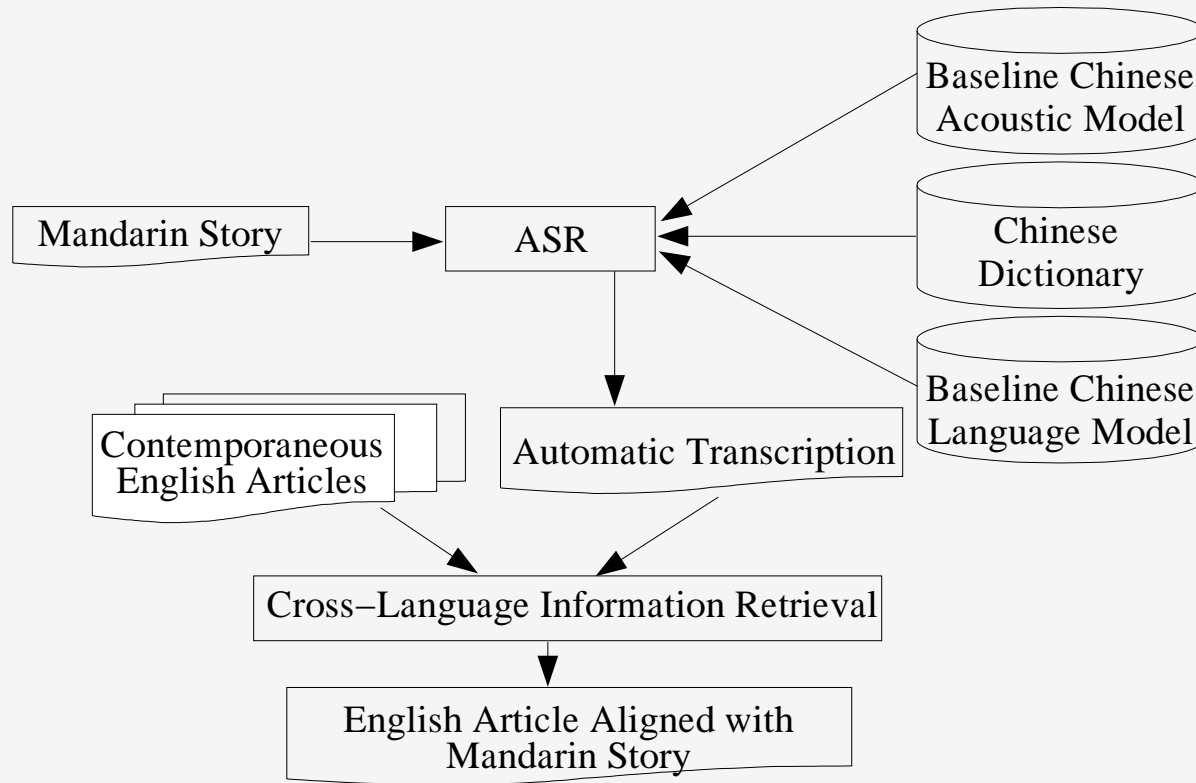
Cross-Lingual LM for ASR



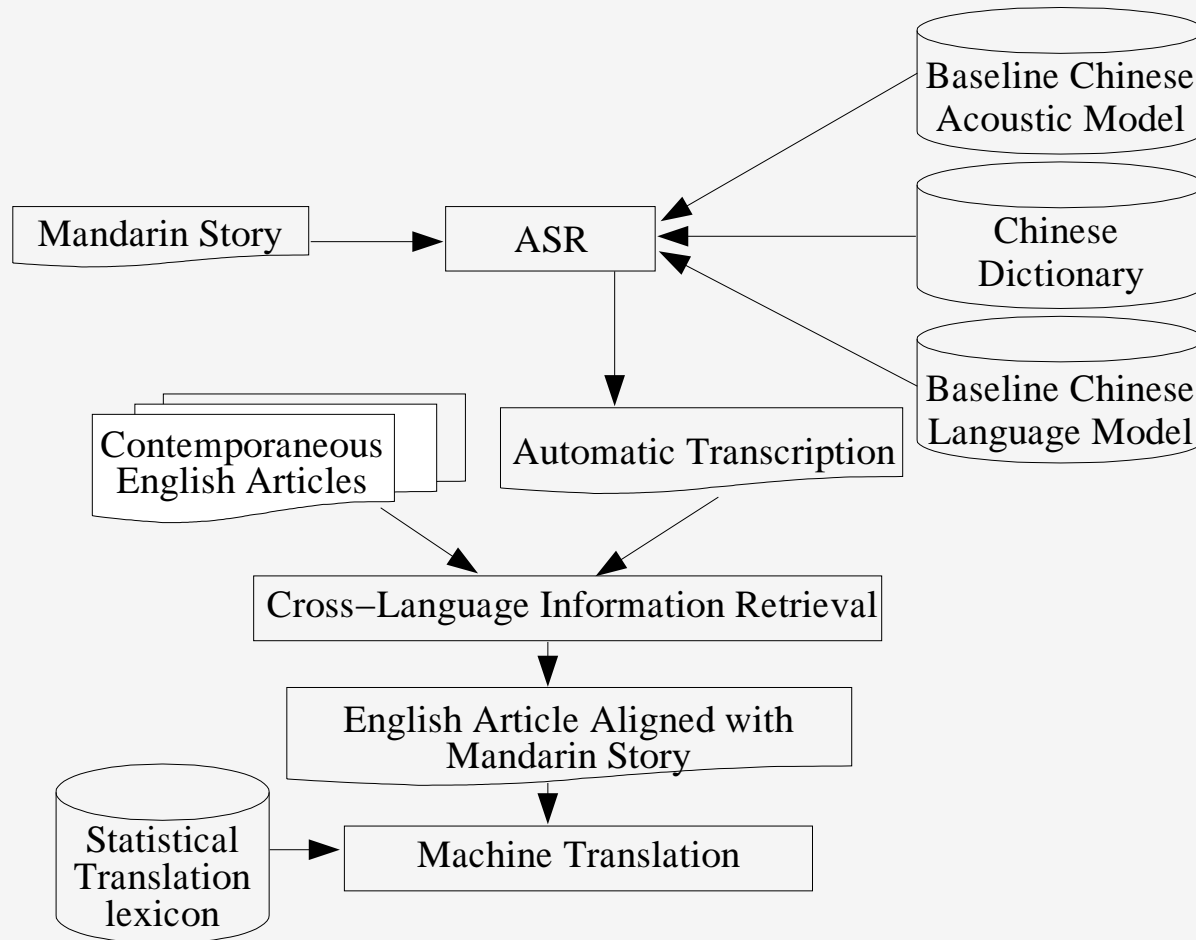
Cross-Lingual LM for ASR



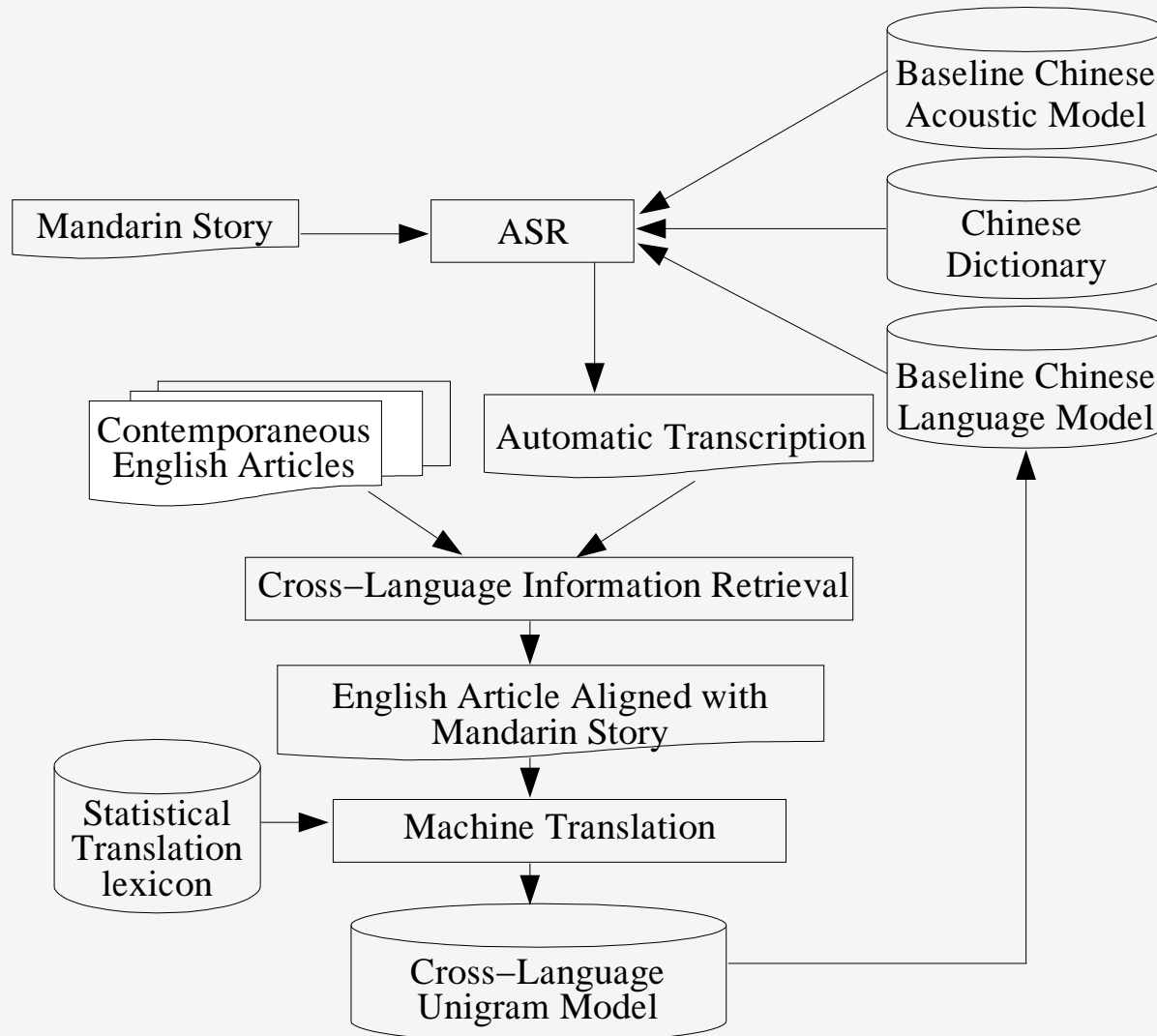
Cross-Lingual LM for ASR



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 ,

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

- ❑ 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
 - ✓ Use it to find the English doc with the highest cosine similarity

$$d_i^E = \operatorname{argmax}_{d_j^E \in \mathcal{D}^E} \operatorname{sim}_{CL}(P(e|d_i^C), \hat{P}(e|d_j^E))$$

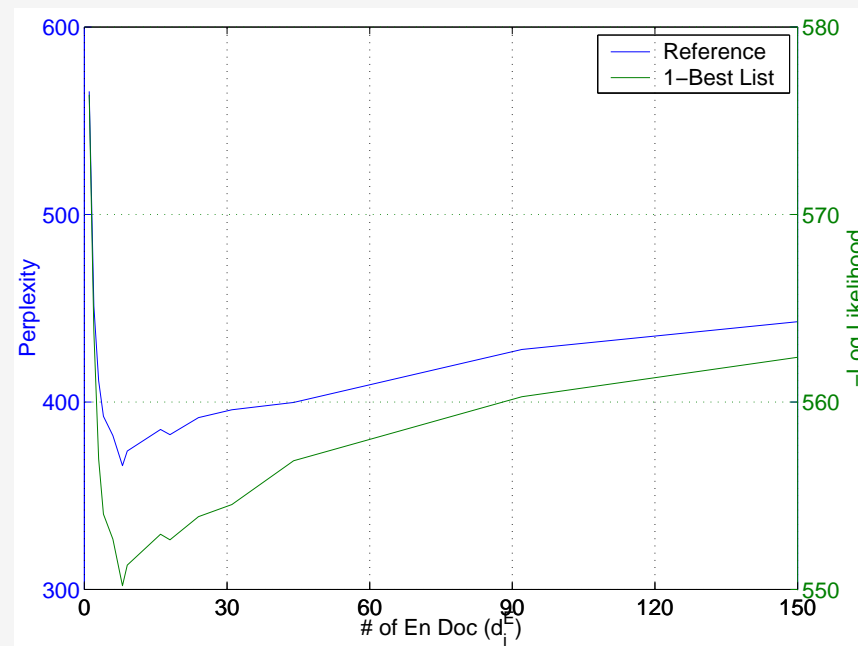
- ❑ Estimation of $P_T(c|e)$ and $P_T(e|c)$ → GIZA++ translation table
 - ✓ GIZA++ : statistical MT tool based on IBM model-4
 - ✓ Input : Hong Kong news Chinese-English parallel corpus
 - 18K docs, 200K sents, 4M wds each
 - ✓ Output : MT system with several tables
 - Only translation tables are used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

Training and Test Corpora

- ❑ Acoustic model training
 - ✓ HUB4-NE Mandarin training data (96K wds) ~ 10 hours
- ❑ Chinese monolingual language model training
 - ✓ PDXR : 290M wds
= People's Daily + Xinhua news + China Radio news int'l
 - ✓ 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

Likelihood-Based Selection of English Doc and λ

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

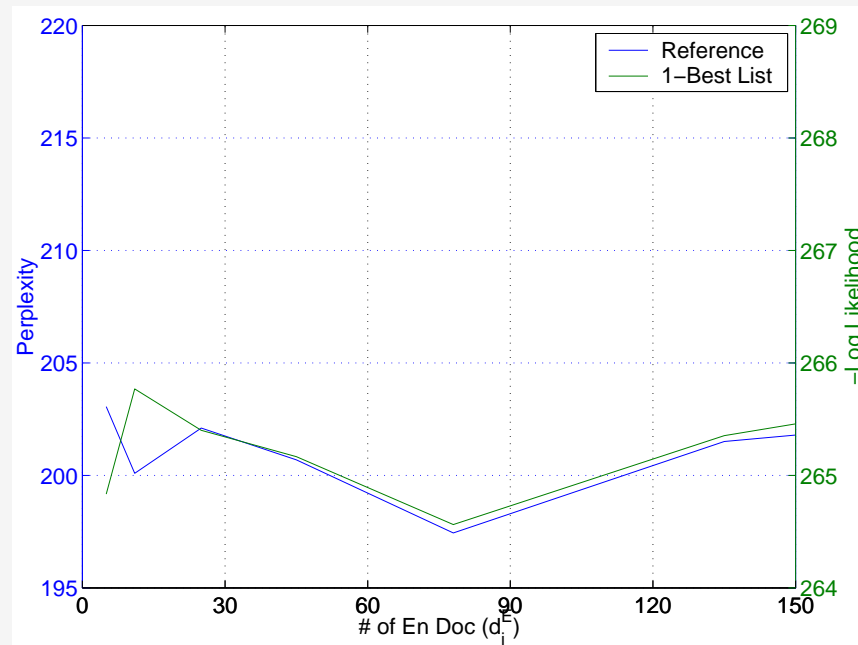


Perplexity (Likelihood) v/s # En Doc (d_i^E) for Sample Story 1

Likelihood-Based Selection of English Doc and λ

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

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



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

ASR Experimental Results

- ❑ Vocab : 51K for Chinese
- ❑ PDXR bigram LM for lattice generation (50.7% WER, 29.6% CER)
- ❑ 300-best list rescoring
- ❑ Oracle worst/best WER (CER) : 92.3/32.2 (59.4/15.2) %

Language model	Perplexity	WER	CER	<i>p</i> -value
Baseline PDXR trigram	283	47.6%	26.9%	—
Topic-trigram	247	47.3%	26.7%	0.174
CL-interpolated	248	47.1%	26.8%	0.028
Topic-trigram + CL	225	46.8%	26.5%	0.003

Experiments in a Resource-Deficient Setting

- ❑ Reduced to reasonable size : XINHUA only (13M wds)
- ❑ Lattices generated from XINHUA bigram LM
- ❑ Topic clustering with XINHUA corpus

Language model	Perplexity	WER	CER	<i>p</i> -value
Baseline XINHUA trigram	426	49.9%	28.8%	–
Topic-trigram	381	49.1%	28.4%	0.003
CL-interpolated	346	48.8%	28.4%	< 0.001
Topic-trigram + CL	326	48.5%	28.2%	< 0.001

Experiments in a Resource-Deficient Setting

- ❑ Extreme case : HUB-4 acoustic model training data (96K wds)
- ❑ 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
Baseline HUB-4 trigram	1195	60.1%	44.1%	–
Topic-trigram	1122	60.0%	44.1%	0.660
CL-interpolated	630	58.8%	43.1%	< 0.001
Topic-trigram + CL	631	59.0%	43.3%	< 0.001

Conclusions

- ❑ Exploits side-information from contemporaneous articles
 - ➔ useful for resource-deficient languages
- ❑ Statistically significant improvements in ASR WER
- ❑ Comparable improvements over within-language topic-dependant LMs; further gains from combination
- ❑ Future work
 - ✓ Cross-language triggers, maximum entropy model
 - ✓ 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/sigtests/mapsswe.htm>
- ❑ Technical Report