I. Introduction

- Motivation:
  - Multilingual, multi-domain, multi-source linguistic resources are available
  - Most resources are concentrated on popular languages such as English, French, and German
  - Relatively small resources for Chinese, Arabic, and other languages.
- Stochastic models require a large amount of data for training
- How to construct stochastic models in resource deficient languages?
  - Bootstrap methods by reusing models from resource rich languages, e.g.
  - Universal phone-set for ASR
  - Exploit parallel texts to project morphological analyzers, POS taggers, etc.
- 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 parallel text
  - Integration of machine translation (MT), cross-language information retrieval (CLIR), and language modeling (LM)

II. Story-Specific Cross-Language Language Models

III. Database

- Hong Kong news text corpus
- Duration: July 1997 to April 2000
- Chinese-English parallel text
- Document- and sentence-level alignment: CLSP WG1 text summarization
- 18K document pairs

IV. Baseline Chinese Language Model Estimation

- Word-based LM after automatic segmentation
- Standard trigram LM + Good-Turing discounting, Katz back-off
- Trained with Chinese portion of the Hong Kong news parallel text
- Estimation of character perplexity
- Calculate log probabilities based on words
- Divide cumulative log probabilities by the total number of characters

V. Cross-Language Language Model Estimation

- Assume document correspondence, \(d_1 \leftrightarrow d_2\), is known for Chinese test doc \(d_1\):
  \[
  P(d_2) = \sum_{d_1} P(d_1)P(d_2|d_1), \quad \forall d_1 \in \mathcal{D}
  \]
- Document correspondence \(\mathcal{D}\) obtained by CLIR
  - For each Chinese test doc \(d_1\), create English bag-of-words based on \(P(d_1)\)
  - Use it to find the English doc with the highest cosine similarity
  \[
  d^* = \arg \max_{d_2 \in \mathcal{D}} \text{cos}(d_1, d_2)
  \]
- Estimation of \(P(d_i|d_j)\) and \(P(d_j|d_i)\) \(\Rightarrow\) GIZA++ translation table
  - GIZA++: statistical machine translation tool based on IBM model-4
  - Input: 18K Chinese-English parallel text docs (sentence aligned)
  - Output: machine translation system consisting of several tables
  - Only translation tables are used: \(P(d_i|d_j)\) and \(P(d_j|d_i)\)
- Cross-Language LM Construction
  - Build story-specific cross-language LMs, \(P(d_i'|d_i)\)
  - Linear interpolation with the baseline trigram LM
    \[
    P(d_i'|d_i|d_j) = \lambda P(d_i'|d_j) + (1 - \lambda) P(d_i|d_i|d_j)
    \]
  - Comparison with topic-dependent LMs
    - Topic clustering: unsupervised K-means clustering
    - Topic-dependent unigram LM and trigram LM

VI. Experimental Results

- LM perplexity results

VII. Conclusions

- Use of resource rich languages to improve the estimation of stochastic models in resource deficient languages
- Improvements: 28.6% in word-level PPL reduction (combined with topic trigram LM)
- A successful integration of MT, CLIR and LM?
- Future work
  - Cross-language lexical triggers
  - Maximum entropy methods to combine cross-language LMs with other LMs
  - Applications to other tasks: MT, TDT
  - Applications to other resource deficient languages: Arabic

VIII. Ongoing Investigation: Chinese ASR Test

- Baseline system: CLSP workshop 2000, Mandarin pronunciation modeling
- LM training data: People’s Daily, Xinhua, China Radio \(\Rightarrow\) 291M words
- Test set: 1997, 1998 HUB-4NE test set \(\Rightarrow\) 1263 utterances, 12K words
- Lattice rescoring: 1st pass lattices generated from a bigram LM
- No parallel English text exists for this Chinese test set
- CLIR: find most similar doc from NAB’97 + TDT2 (NYT, APW) corpora (48K docs)

Presented in