The Johns Hopkins University 2003 Chinese-English Machine Translation System

July 21, 2003

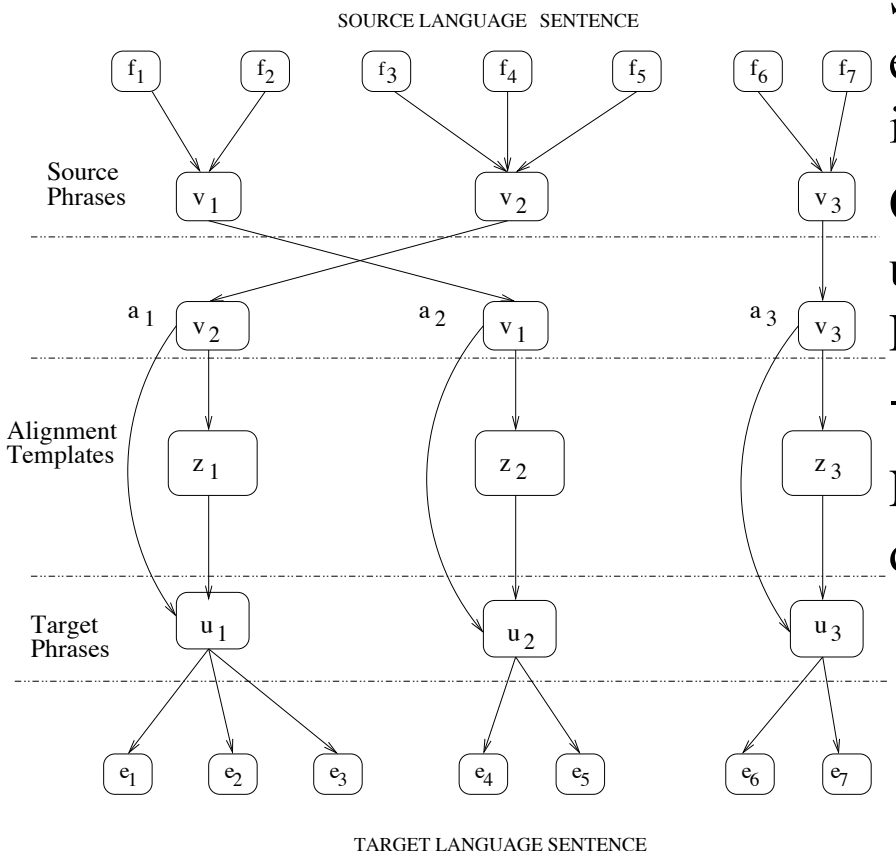
B. Byrne, S. Khudanpur, W. Kim, S. Kumar, P. Pecina,
P. Virga, P. Xu and D.Yarowsky

Presented by S. Kumar
Center for Language and Speech Processing
The Johns Hopkins University, Baltimore, MD, U.S.A.

Outline of Chinese-English Systems

- Weighted Finite State Transducer - Alignment Template Translation Model (ATTM) for Statistical Machine Translation
  - WFST-ATTM model (Kumar and Byrne : HLT-NAACL '03)
  - MT System Description (Byrne et.al.: MT Summit '03)
- Two MT systems
  - Baseline system
  - Document specific translation
    - Motivated by acoustic and language model adaptation in ASR
    - Goal is to build a translation system for each test document
    - Given a large heterogenous bitext collection, find the portion that is most similar to the test document; use this for testing
- Performance on Development and Evaluation Sets

Alignment Template Translation Model (ATTM) Architecture



Statistical models are trained for each model component and implemented as WFSTs

Overall decoding is implemented using standard and optimized FSM operations
- Uses the AT&T toolkit

No specialized search procedures

- No decoder!
- Easy generation of N-best Lists & lattices of translation hypotheses

Baseline System : Bitext and Word Alignments

- Training Data
  - Primary bitext: FBIS Chinese-English data
  - Sentence Alignment using Danyu Liu's aligner (JHU WS'01)
  - Word Segmentation of Chinese text using LDC segmenter
  - Aligner Output not uniformly good
    - Rank sentence-pairs using a lexical co-occurrence score
    - Select top-100k sentence pairs
- Word Alignments
  - Obtain word alignments using IBM-4 models (trained with GIZA++) in both translation directions (E $\rightarrow$ C and C $\rightarrow$ E), and form their union (following Och '02).

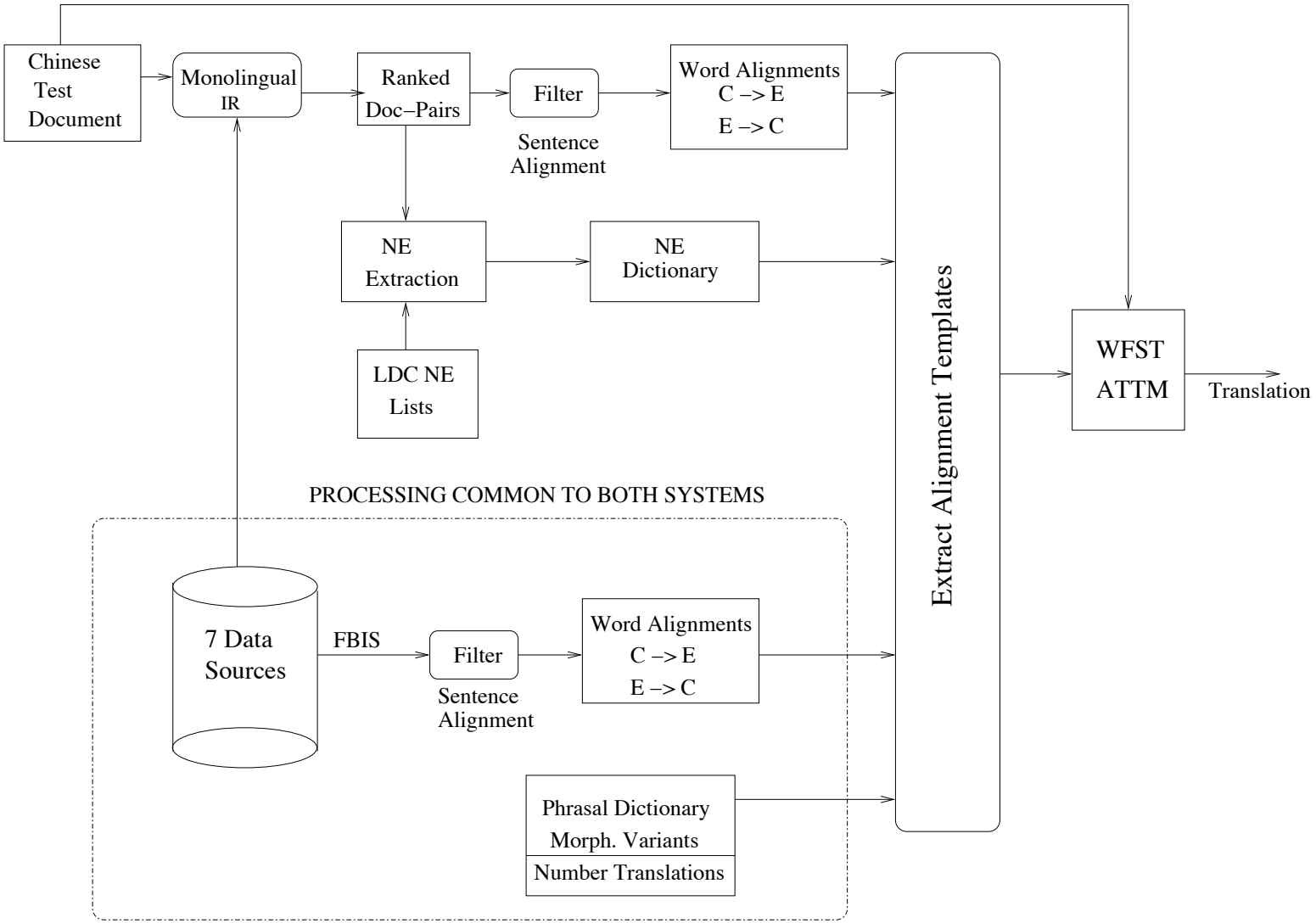
Baseline System : Translation and Language Models

- Alignment Templates
  - Use phrase-extract algorithm (Och '02) to extract an alignment template library from word alignments
  - Additional templates are added to those learnt from bitext
    - Phrasal entries from the LDC lexicon
    - Rule based Chinese-English translations for numbers, dates and times (E. Drabek - JHU and R. Hwa - UMD) to cover test set
    - Insertions of selected zero-fertility English words
  - Augment templates to produce inflected forms of English words (abide $\rightarrow$ abide,abiding,abided)
- Language Model
  - Training Data:
    - Online archives of *The People's Daily* - 16.9M words
    - English side of Xinhua corpus - 4.3M words
  - Trigram LM with Modified Kneser-Ney smoothing (SRILM toolkit)

Document Specific MT system

- Goal is to build Specialized Translation Models for each individual test document
- For each test document, create a
  - Document-specific training bitext
  - Refined set of name entities with translations
- The use of IR in MT to create a document specific training bitext
  For each test document
  - Use standard IR vector model to rank the training set docs using cosine similarity scores
  - Select the top 100k sentence-pairs using both IR scores and sentence-alignment quality measures
  - Generate word alignments using document-specific IBM-4 models.
- Document Specific Translation Models
  - Created at test time
  - Merge document specific word alignments with baseline word alignments
  - Extract document specific alignment template libraries (and translation vocabularies)

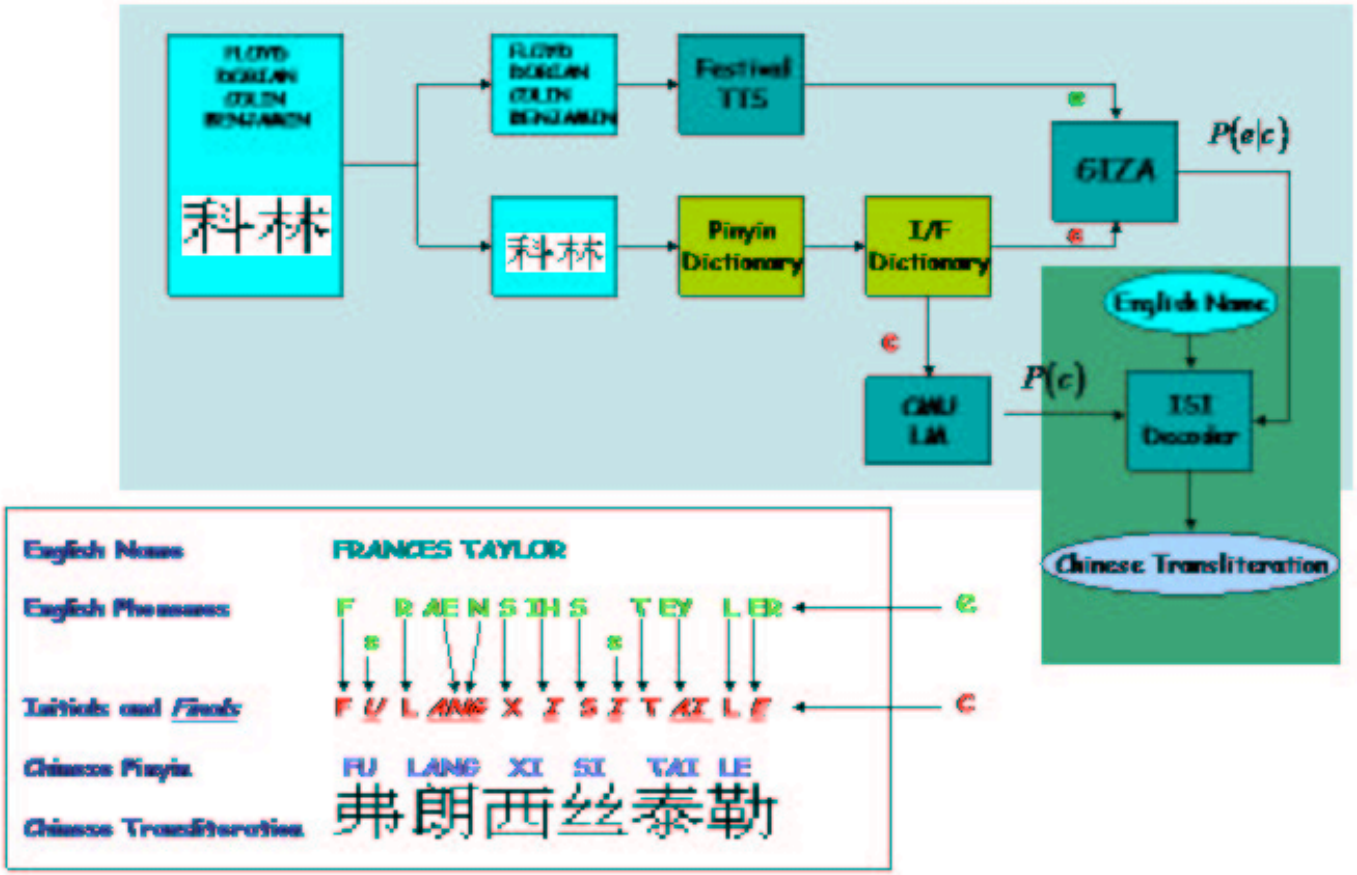
Block Diagram of the Baseline and Document Specific MT systems



Incorporation of Name Entities (NEs)

- Motivation: A straightforward experiment showed that adding the entire LDC NE list to IBM-4 training hurt translation performance
- Alternative Approach using ATTM & Document Specific training
  - Identify all English names that appear in the English side of the retrieved Chinese documents using LDC NE list
  - Filter the list by discarding any entry whose Chinese side does not appear in the test document
  - Add these NE translations to the alignment template library
- Allows picking NEs that were not segmented as a single Chinese word.

Transliteration of Proper Names



Translation Performance

- Test Sets:
  - ZBN-Eval02 (30 documents) : Zaobao-News portion of eval02
  - Eval03 (100 documents)

System	ZBN-Eval02		Eval03
	BLEU	NIST	NIST
Baseline (Contrast)	0.1600	6.6272	6.7892
Doc-Specific-Bitext	0.1660	6.8628	*
Doc-Specific-NE	0.1622	6.7081	*
Doc-Specific-Bitext-NE (Primary)	0.1758	7.0052	7.0519

- Document Specific approach for selecting MT training data & name entities improves performance over a baseline system
- Gains from Doc-Specific-Bitext and Doc-Specific-NE are more than additive

Conclusions and Future Work

- First use of our WFST model in an MT evaluation
  - System in development for less than 9 months
- Modular FSM approach allows for modular development of model components and plug-and-play evaluation
- In-house model training procedures based on GIZA++ alignments
- Decoder supports lattice/N-best list generation and rescoring.
- First large-scale integration of IR for MT in order to refine translation model training data and incorporate name entities
- Future Work
  - Full EM style retraining of WFST-ATTM model
  - Refinements to IR approach
  - Better integration of constituents in the ATTM framework

Hindi-English MT system

- Common Data Track
  - Word-Alignments from ISI
- MT Systems are "language-independent"
  - Hindi MT systems based on WFST-ATTM (similar to Chinese)
  - WFST-ATTM MT framework allowed rapid extension to Hindi with minimal re-engineering effort
  - Systems based on ITRANS and UTF-8 encodings (both acceptable)
  - 3rd System based on Minimum Bayes-Risk Rescoring of N-best Lists from UTF-8 system under NIST Loss function

System	ISI-Dev		Eval
	BLEU	NIST	NIST
UTF-8 (Primary)	0.1508	**4.7426**	**7.0333**
ITRANS	0.1439	4.6759	7.0442
MBR Rescoring	0.1506	**4.8432**	**7.0867**