The Johns Hopkins University 2004 Chinese-English and Arabic-English MT Evaluation Systems

June 22, 2004

Shankar Kumar, Yonggang Deng, Charles Schafer

Woosung Kim, Paola Virga, Nizar Habash, David Smith, Filip Jurcicek

Bill Byrne, Sanjeev Khudanpur, Zak Shafran, David Yarowsky

Center for Language and Speech Processing

The Johns Hopkins University

Baltimore, MD 21218

The Development of the Translation Template Model

- Investigation of Weighted Finite State Transducer Frameworks for SMT
  - Interested in WFST techniques from work in ASR
  - WFST Formulation of IBM-3 model (Knight and Al-Onaizan '98)
    - Basis for Minimum Bayes-Risk word alignments of bitexts (Kumar and Byrne '02)
- Quest for a powerful SMT framework
  - Alignment Template Translation Model (Och, Tillmann and Ney '99)
    - Overcomes weakness of IBM-style models by phrase translations
  - Formulation and Implementation of WFST-ATTM (Kumar and Byrne '03)
    - Basis for JHU 2003 Chinese-English system
- Translation Template Model (Kumar and Byrne '04)
  - Correct integration of Language and Translation Models via Source-Channel Formulation
  - Ignores word alignments within phrase-pairs
  - Allow insertions of Target (French) phrases in the generative translation process
  - Easily scaled up to large bitexts

The Development of the 2004 MT Evaluation Systems

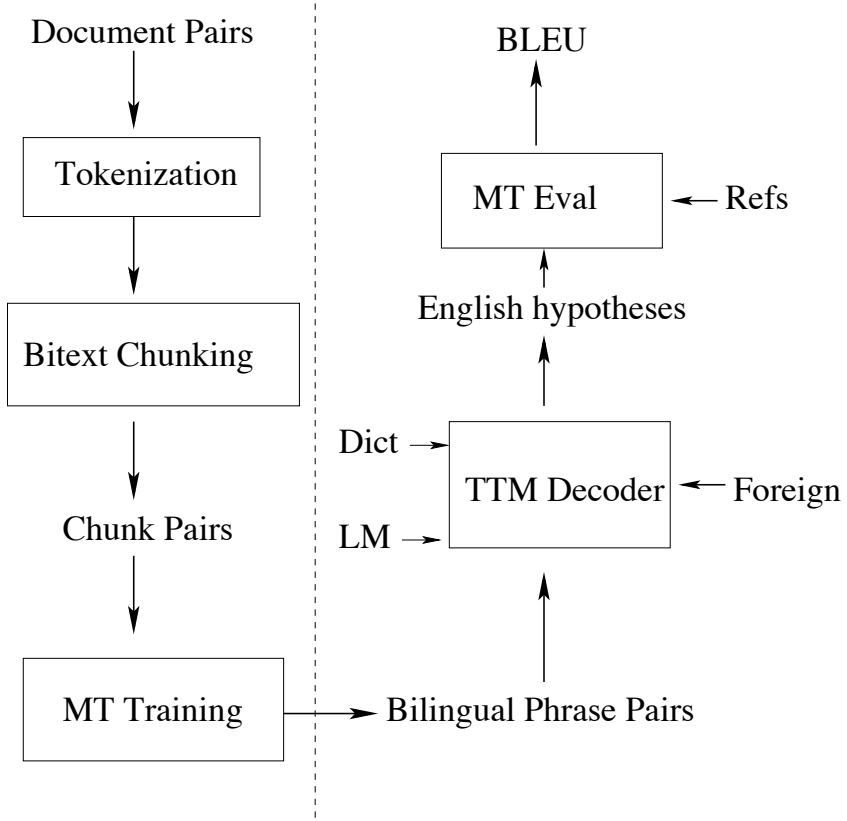
- Need to use all the available bitext for MT training
- Need better language modeling
  - More monolingual text
  - FSM implementations of arbitrary n-gram LMs (4-grams)
- Evaluation of all model components in terms of overall system BLEU
- We failed in all attempts to incorporate "linguistic knowledge" & improve BLEU
  - Morphology, Named entities, Syntax (Kumar and Byrne, HLT '04)
  - Modeling phrase movement did not help
- Multiple phrase segmentations of source sentence is essential in translation
- Eval '04 systems mainly developed by two graduate students (S. Kumar and Y. Deng)
- C-E system was developed in under a month
- A-E system was developed in under two weeks

Outline

- Bitext Chunking
- Translation Template Model
- Evaluation Systems

Bitext Chunking

- Goal
  - chop up document aligned bitext into smaller chunk pairs for Translation Model training.
  - improve chunking to improve translation performance
  - coarse Models of aligning documents in terms of subsentence segments
- Desired properties
  - High efficiency, discard as little as possible
  - Chunks are subsentences
  - Start from scratch, language independent, minimal linguistic knowledge required

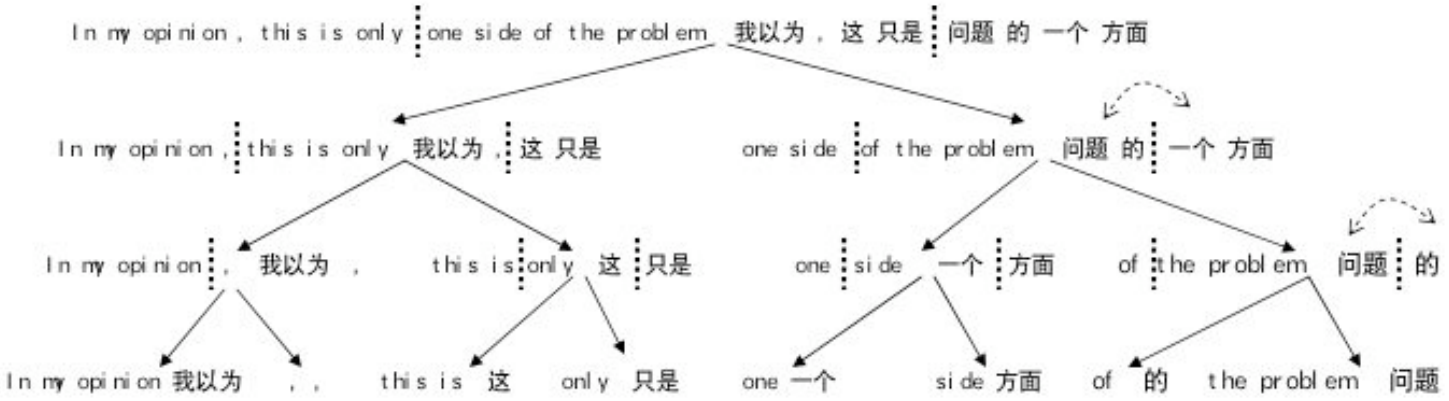


Stochastic Models of Bitext Chunking

- Coarse Models of aligning documents in terms of subsentence segments
- Define probability distribution over bitext documents and bitext chunks
- Incorporate any word translation model, e.g. IBM Model 1
- Different modeling assumptions lead to different alignment algorithms
  - Dynamic Programing (DP) [similar to Gale & Church '91]
  - Divisive Clustering (DC)

Divisive Clustering

- Modeling assumption
  - Smaller chunks are generated by binary splitting of bigger chunks
- Iteratively obtain a finer alignment by successive binary splitting
  - Reording enables complicated linking structures between segments
- Potential splitting points
  - English, Chinese: at punctuations; Arabic : all words
- Model assigns likelihood to chunks produced
  - Simple length and likelihood based stopping rule
- Not parsing
  - strictly text string segmentation driven by word-to-word translation probs (Model 1)

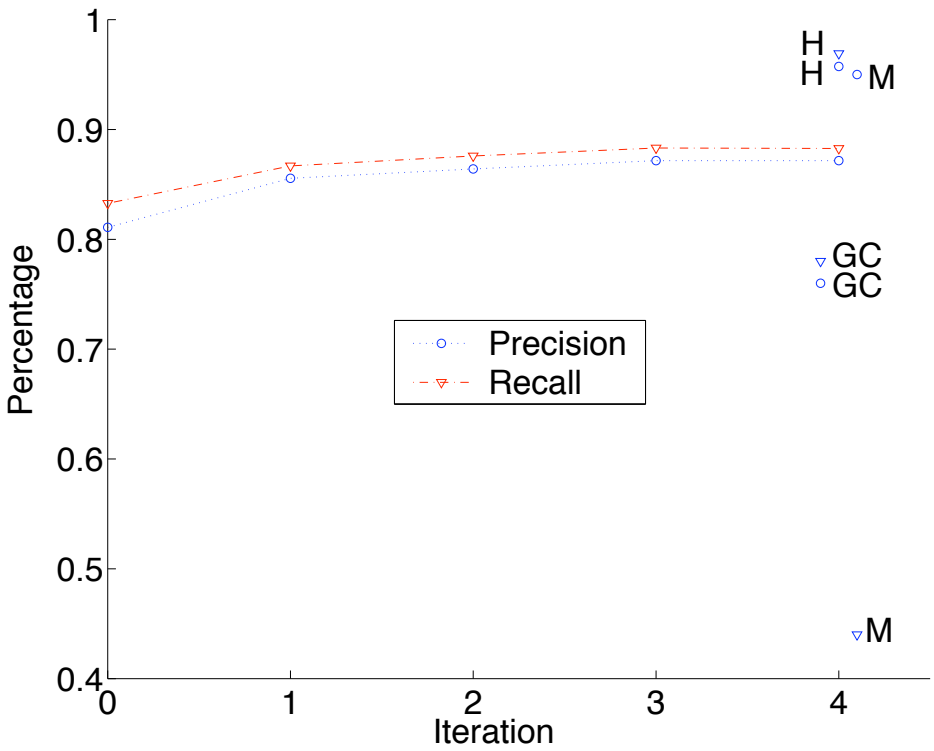


Why Divisive Clustering (DC)?

- Deriving short chunk pairs by going below sentence level
- Practical benefits
  - Shorter bitext segments can lead to quicker training of MT systems
  - Thorough exploitation of all available bitext
- DP vs. DC
  - DP: global scale, sequential links
  - DC: local optimal, swapping links
- Two level alignment procedure DP+DC
  - Initial steps DP - produce a coarse alignment by allowing big chunks
  - Refine by DC

Sentence Alignment Experiments

- If we restrict chopping points to sentence boundaries, we have a sentence aligner
- Unsupervised alignment of 122 Chinese/English documents from FBIS corpus
  - reference alignments were created manually
- Different algorithms have different operating points
- We benefit through iterative estimation of Model 1
- When Model 1 is estimated from manual sentence alignments, we get best Pre/Rec



- GC: (Gale & Church '91) aligner
- M: (Moore '02) aligner
- H: DP+DC w/ model trained from manually aligned pairs

DC Improves MT Performance

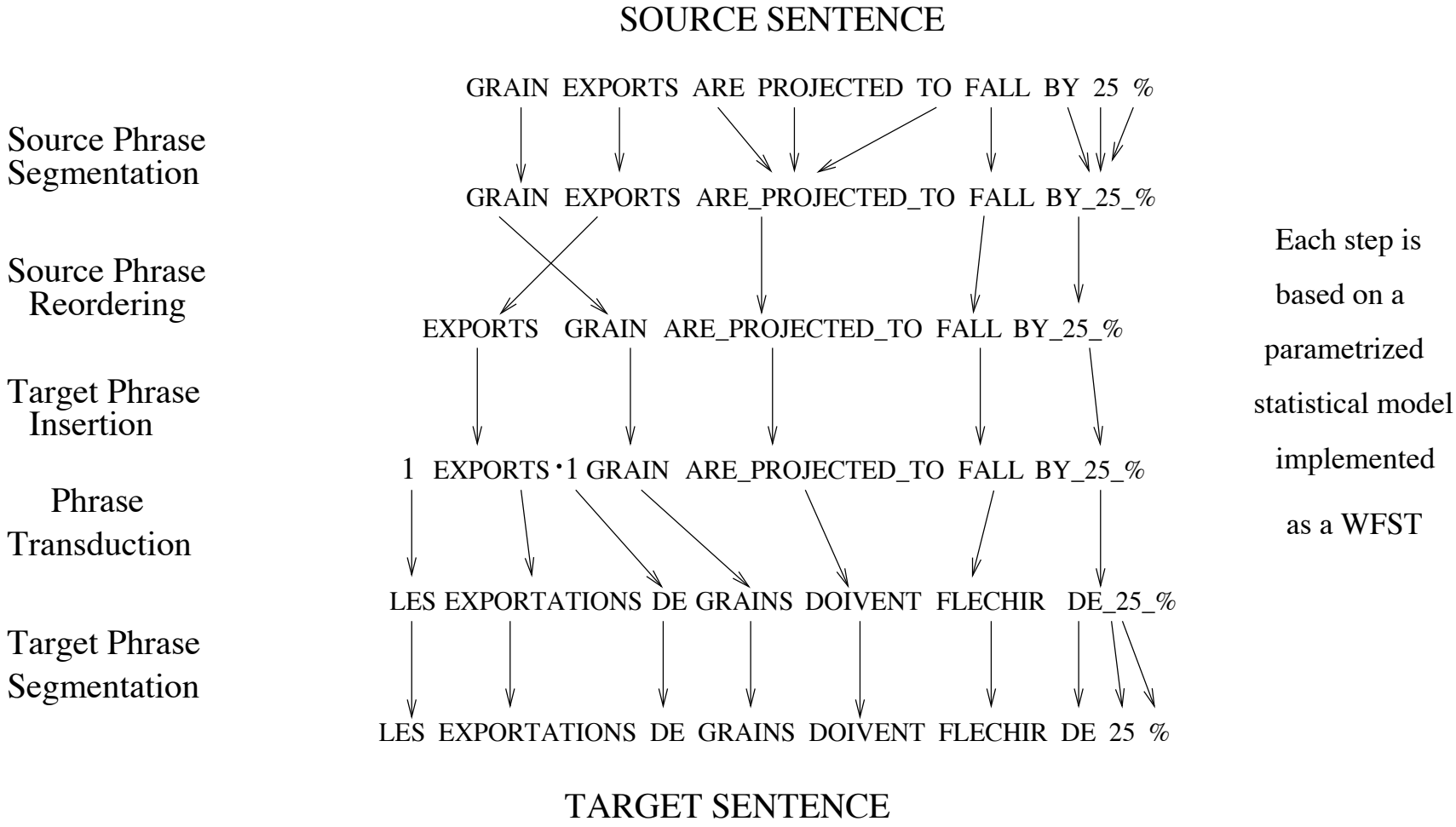
- Chinese/English bitext
  - start from scratch
  - processed by 2 iteration DP+DC procedures
- Arabic/English bitext
  - Keep existing sentence pairs (w/ <60 words) from LDC alignments
  - Apply DC to long sentence pairs
  - Percentage of usable Arabic bitext for MT training

	total English words(M)	Sent. Aligned Bitext	Sent. Aligned Bitext + DC
News	3.59	67%	98%
UN	131	74%	98%
BLEU with chunks from News: eval02		33.0	33.9
eval03		35.4	36.1

Outline

- Bitext Chunking
- Translation Template Model
- Evaluation Systems

Generative Translation Process Underlying the TTM



Each step is based on a parametrized statistical model implemented as a WFST

Investigative Experiments

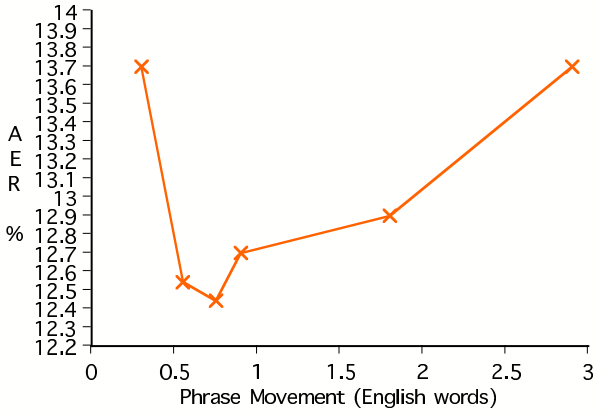
Goal: Identify contribution of each TTM component to the overall alignment and translation performance

- Effect of Phrase-level movement on Word Alignment
- Effect of Coverage of the Test Set (by the Phrase-Pair Inventory) on Translation
- For more experiments, see CLSP Research Note 48 (Kumar and Byrne '04)

Given the TTM, does allowing phrase-level movement help word alignment ?

- Task: French-English Hansards (Train: 48K pairs, Test: 500 sents)
- Generate an N-best list of reordered phrase sequences under TTM
- By varying Phrase Exclusion Probability (PEP), we can search over reorderings of the English phrase sequence to find optimal Alignment Error Rate (AER)
- Allows us to plot Phrase movement versus AER under the TTM (over the N-best list)

With a fixed number of reorderings (N=400), vary PEP



- At optimum AER, phrases move less than 1 word on average

With PEP fixed at optimal value, vary # of reorderings

# of reords.	AER (%)	Avg. Movement
None	12.8	0.0
1	12.8	0.2
200	12.5	0.7
1000	12.5	0.8

- Conclusion: Given current model, allowing phrases to move doesn't help AER
- Based on this, we construct our system to disallow phrase-level movement, allow only word-level movement within phrase-pairs

Coverage of the Test Set by the Phrase-Pair Inventory (PPI)

- Task: French-English Hansards (Train: 48K sent. pairs, Test: 500 sents)
- Train IBM-4 models on all bitext (48K sentence pairs) and obtain word alignments
- Collect Phrase Pair Inventories over 4 subsets of these word alignments (Och '02)
  - Alignment quality over these variable size inventories is held constant
- Coverage = % of phrases in the test set that exist in the PPI

# of sentence-pairs (K)	Test-Set Coverage (%)	BLEU (%)
5	20.79	19.6
12	26.75	20.8
24	31.35	21.5
48	36.02	22.3

- A higher coverage of the test set by PPI improves translation performance

Outline

- Bitext Chunking
- Translation Template Model
- Evaluation Systems

Text Processing and Bitexts

Text Processing

- Chinese Text segmented into words using LDC segmenter
- Arabic Text processing pipeline
  - Modified Buckwalter analyzer (D. Smith)
  - Post Processing to separate conjunctions, prepositions, Al- and pronouns
  - Al- and w- deletion (maybe wrong decision!)
- English Text processed using a simple tokenizer

Bitext for Translation Model Training

	Chinese-English	Arabic-English	
		Primary	Late Contrast
# of sent. pairs (K)	-	68.0	-
# of chunk pairs (M)	7.6	N/A	5.1
# of words (M)	175.7/207.4	2.0/2.1	123.0/132.5

English Language Models

Source	Xin	AFP	PD	FBIS	UN	AR-news	Total
C-E							
Small 3g	4.3	-	16.2	-	-	-	20.5
Big 3g	155.7	200.8	16.2	10.5	-	-	373.3
Big 4g	155.7	200.8	16.2	10.5	-	-	373.3
A-E							
Small 3g	63.1	200.8	-	-	-	2.1	266.0
Big 3g	83.0	210.0	-	-	131.0	3.6	428.0
Big 4g	83.0	210.0	-	-	131.0	3.6	428.0

- Baseline Language Models : Fixed linear interpolation weights
- Document-Specific 4-gram Language Models (W. Kim and P. Virga)
- For each test document
  - Decode using static LM to obtain 1-best hypothesis
  - Use 1-best hyp to query English text collection & retrieve 10k docs (≈ 3Mwords)
  - Interpolate 3-gram (doc. specific) with Big 4-gram
    - Compute interpolation weights to maximize likelihood of 1-best hypothesis

Translation Model Training

- Partition bitext and IBM-4 train models on each partition
- Merge word alignments & collect translation templates (phrase-pairs)

Chinese-English (Decode with Small3g)

Bitext Partition	Contribution by Source: En words (M)					BLEU (%)		
	FBIS	HKNews	XHTS	UN	Total	eval01	eval02	eval03
1	10.5	16.3	-	26.7	53.5	25.1	25.5	24.1
2	10.5	-	-	85.0	95.5	25.8	25.9	25.0
3	10.5	16.3	43.0	25.8	95.6	28.0	25.8	24.9
1+2+3						28.1	26.6	25.5

(XHTS = Xinhua, Hansards, Treebank, Sinorama)

Arabic-English (Decode with Small3g)

Bitext Partition	Contribution by Source: En words (M)			BLEU (%)	
	UN	News	Total	eval02	eval03
1	65.0	3.5	68.5	35.0	37.4
2	64.0	3.5	67.5	35.7	37.6
1+2				35.8	37.8

Performance of Evaluation Stages

System #		Decoding Method	BLEU%						
			Chinese-English				Arabic-English		
			e01	e02	e03	e04 (c)	e02	e03	e04 (c)
1	JHU-UMD '03	-	-	16.0	17.0	-	-	-	-
2	AE-Primary '04	-	-	-	-	-	-	38.0	30.6^{S}
3	Small3g		28.1	26.6	25.5	-	35.8	37.8	-
4	Big3g		28.7	27.7	27.1	26.5	38.1	40.1	-
5	Big4g	Nbest from 4	29.6	28.2	27.3	27.5	-	-	-
6		Lattice from 4	29.7	28.5	27.4	27.6	39.4	42.1	36.1^{S}
7	MBR	Nbest from 5	29.7	28.4	27.7	27.5	-	-	-
8		Nbest from 6	30.0	28.8	27.6	27.8	39.6	42.2	36.5^{S}

- Eval04 results are case-sensitive BLEU
  - Truecasing was performed using WS '03 capitalizer trained on in-domain English text
- A-E submitted systems: Primary (2), Very Late Contrast (6 and 8)
- System 6 (A-E) includes improvements from tuning word insertion penalty

Diagnostics: Development of Chinese-English Submitted Systems

All systems are MBR-BLEU rescoring over N-best lists generated by different LMs

System Name	LM	BLEU (%)			
		eval01	eval02	eval03	eval04 (Cased)
JHU-TTM	Big4g	29.6	28.2	27.3	27.5
	MBR	29.7	28.3	27.7	27.4
	MBR (posteval)	29.7	28.4	27.7	27.5
JHU-TTM-DSLM2	Doc-specific-Dyn 4g	29.8	28.9	27.6	27.5
	MBR	29.6	28.9	27.8	27.7
JHU-TTM-DSLM1	Doc-specific-Stat 4g	29.8	29.0	27.8	27.7
JHU-TTM-DSLM1+MBR	MBR	-	-	27.8	27.1
	MBR (posteval)	29.9	28.9	27.9	27.8

- Gains from Doc-specific LMs and MBR-BLEU rescoring are not additive!
  - Doc-specific LMs assign higher likelihood to MAP hypothesis
- Subsequent to evals, we found that performance of Doc.Specific LMs can also be achieved by Lattice Rescoring with a static 4gram LM followed by MBR rescoring on the new N-best list (See Page 20, system 9)

Conclusions

- A statistical MT system based on Bitext Chunking and TTM
  - Evaluation system benefitted from a careful study of contribution of each model component to overall performance
  - Bitext chunking allows use of almost all available bitext for model training
  - Large gains relative to JHU 2003 eval system (11% BLEU absolute on eval03)
  - Respectable performance in Chinese-English and Arabic-English MT (Very Late Contrast)
  - Similar Architectures for C-E and A-E systems
    - Primary C-E system developed in 1 month
    - Very Late Contrast A-E system developed in 10 days
- WFST Architecture supports Lattice/N-best List generation and rescoring
  - Gains from 4-gram LM rescoring and MBR-BLEU rescoring

References

- S. Kumar and W. Byrne 2004. A Weighted Finite State Transducer Translation Template Model for Statistical Machine Translation, *submitted to Natural language Engineering Available as CLSP Research Note 48*
- Y. Deng, S. Kumar and W. Byrne 2004. Bitext Chunk Alignment for Statistical Machine Translation, *CLSP Research Note*
- S. Kumar and W. Byrne 2004. Minimum Bayes-Risk Decoding for Statistical Machine Translation, *Proceedings of HLT-NAACL* , Boston MA
- W. Byrne et.al. 2003. The JHU Chinese-English 2003 Machine Translation System, *Proceedings of MT Summit IX* , New Orleans LA
- S. Kumar and W. Byrne 2003. A Weighted Finite State Transducer Implementation of the Alignment Template Model for Statistical Machine Translation, *Proceedings of HLT-NAACL*, Edmonton, AB, Canada
- S. Kumar and W. Byrne 2002. Minimum Bayes-Risk Word Alignments of Bilingual Texts, *Proceedings of EMNLP* , Philadelphia, PA