IBM T. J. Watson Research Center

IBM Arabic-to-English Translation for IWSLT 2006

Young-Suk Lee

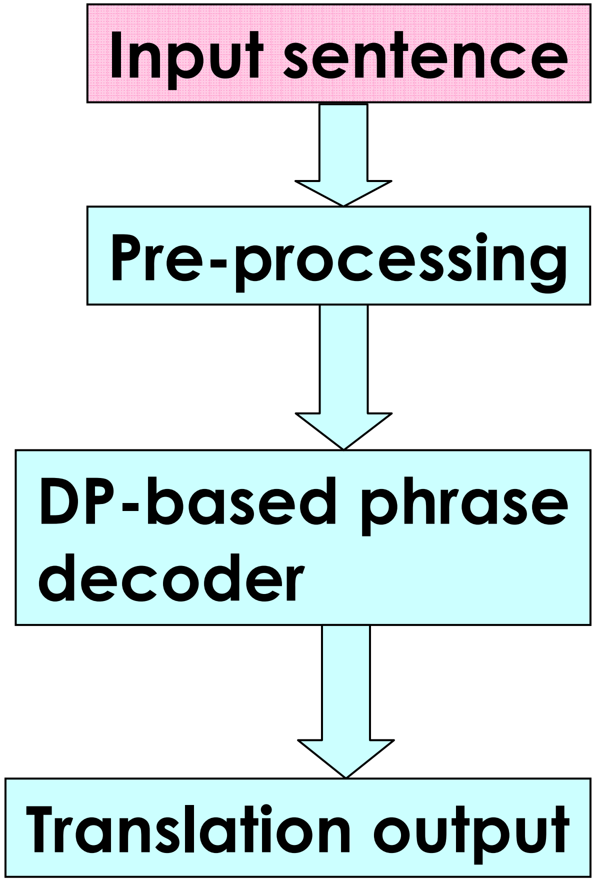
IWSLT 2006

November 28, 2006, Kyoto, Japan

Outline

- Baseline Decoding
- Technical Challenges and Solutions
- IWSLT 2006 Evaluation Results
- Conclusions

Baseline Decoding



Input sentence

Pre-processing

DP-based phrase decoder

Translation output

نفتح من السابعة بعد الظهر الى منتصف الليل

nftH mn AlsAbEp bEd AlZhr Aly mntSf Allyl

Phrase translation models

- Direct / source channel / unigram models
- [Tillmann 2003], [Lee et al. 2006]

Modified IBM Model 1 cost [Lee et al. 2006]

Word trigram language models

Distortion models [Al-Onaizan & Papineni 2006]

Word/block count penalty [Zens & Ney 2004]

We are open from seven p.m. to midnight .

IWSLT 2006 Challenge: High OOV Rate

Correct Recognition Result

	Token Cnt	OOV Cnt	Avg. seg length	OOV Rate
Eval06	5,229	609	10.5	11.65 %
Dev06	4,763	489	10.1	10.27 %
Eval05	3,164	157	6.3	4.96 %

Word Segmentation & Morphological Analysis

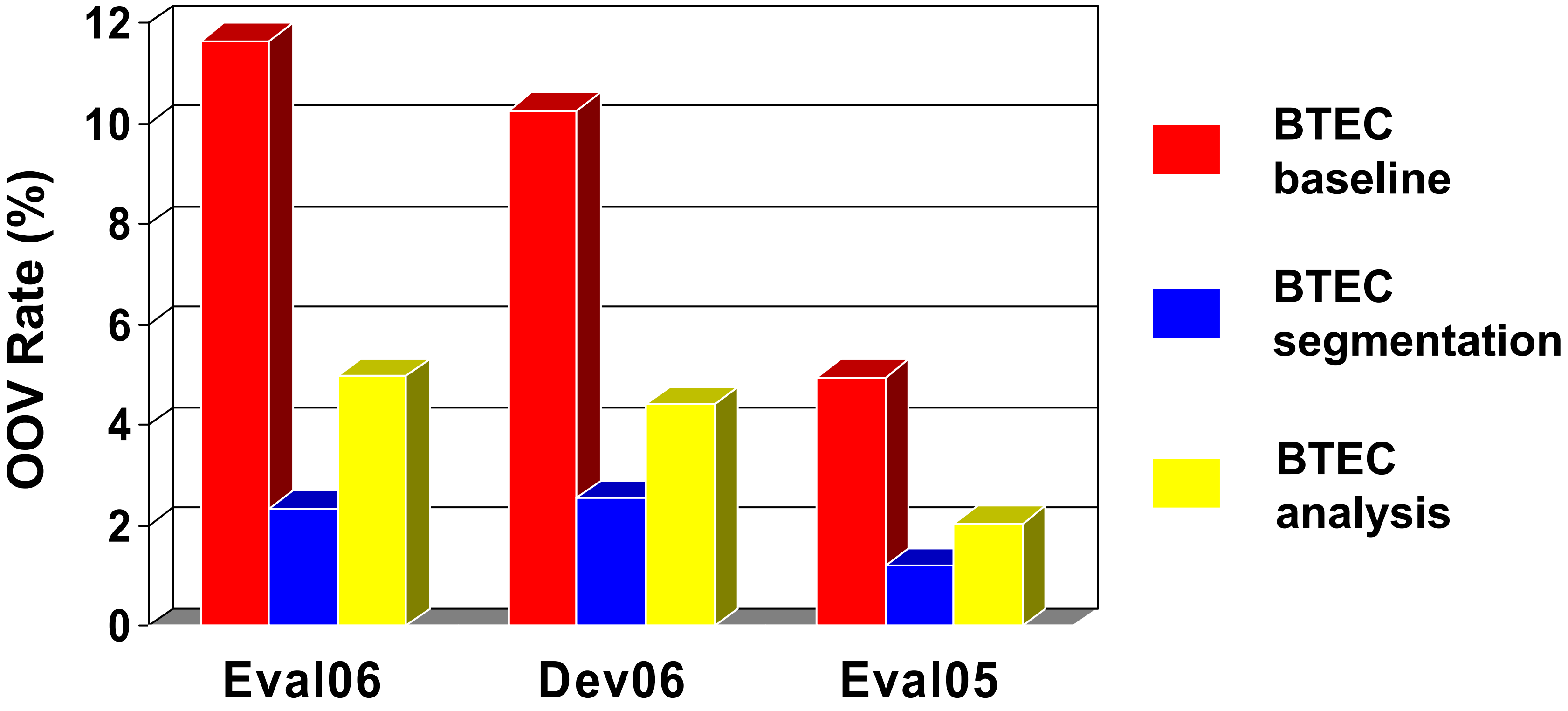
wsyHl sA}q AltjArb fy jAgwAr AlbrAzyly lwsyAnw bwrty mkAn AyrfAyn fy AlsbAq gdA AlAHd Al*y sykwn Awly xTwAth fy EAlm sbAqAt AlfwrmwlA

w# s# y# Hl sA}q Al# tjArb fy jAgwAr Al# brAzyly lwsyAnw bwrty mkAn AyrfAyn fy Al# sbAq gdA Al# AHd Al*y s# y# kwn Awly xTw +At +h fy EAlm sbAq +At AlfwrmwlA

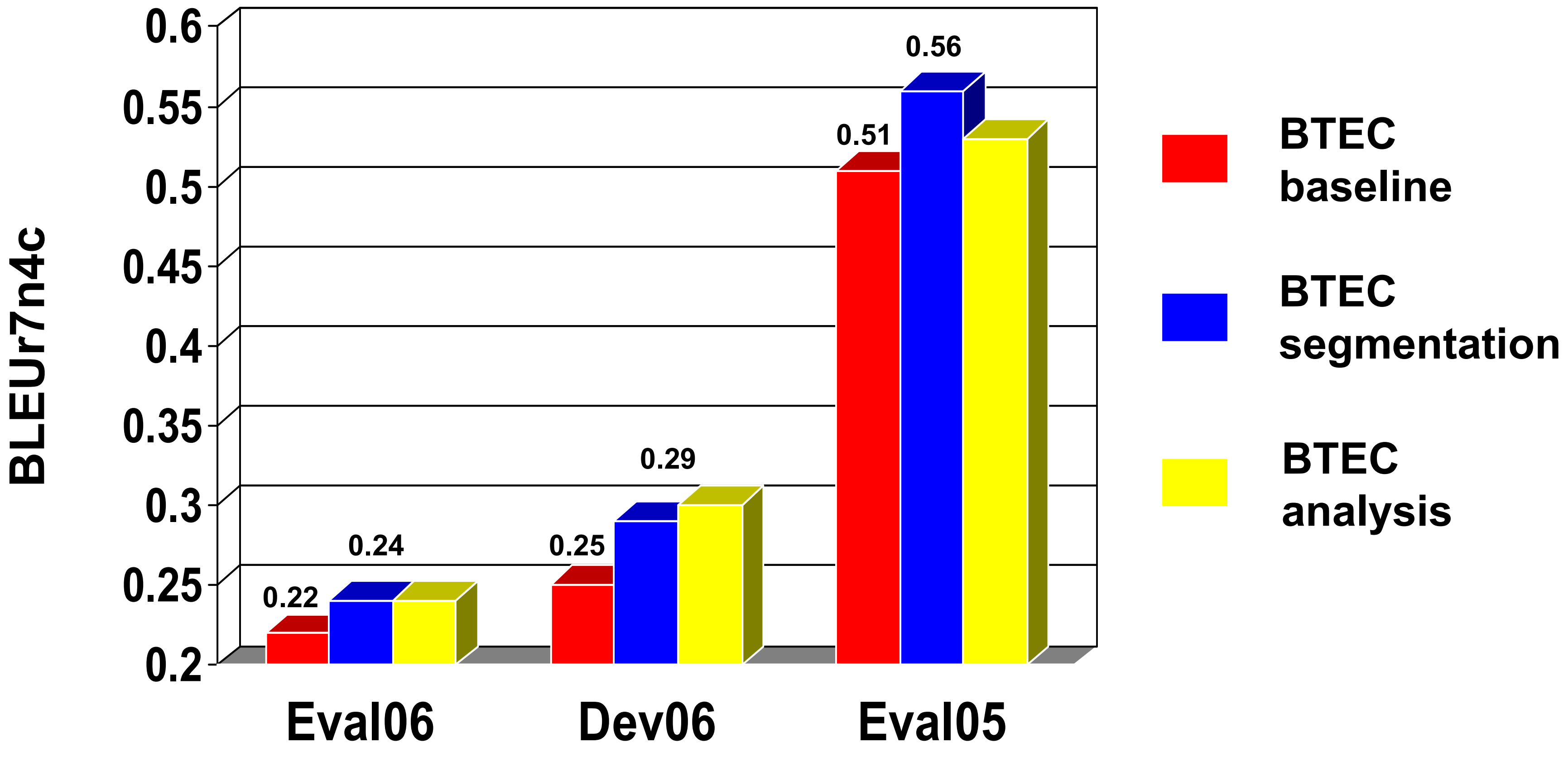
w# s# yHl sA}q ø tjArb fy jAgwAr Al# brAzyly lwsyAnw bwrty mkAn AyrfAyn fy Al# sbAq gdA ø AHd Al*y s# ykwn Awly xTwAt +h fy EAlm sbAqAt AlfwrmwlA

Word Segmentation [Lee et al. 2003], Morphological Analysis [Lee 2004]

OOV Rate Reduction



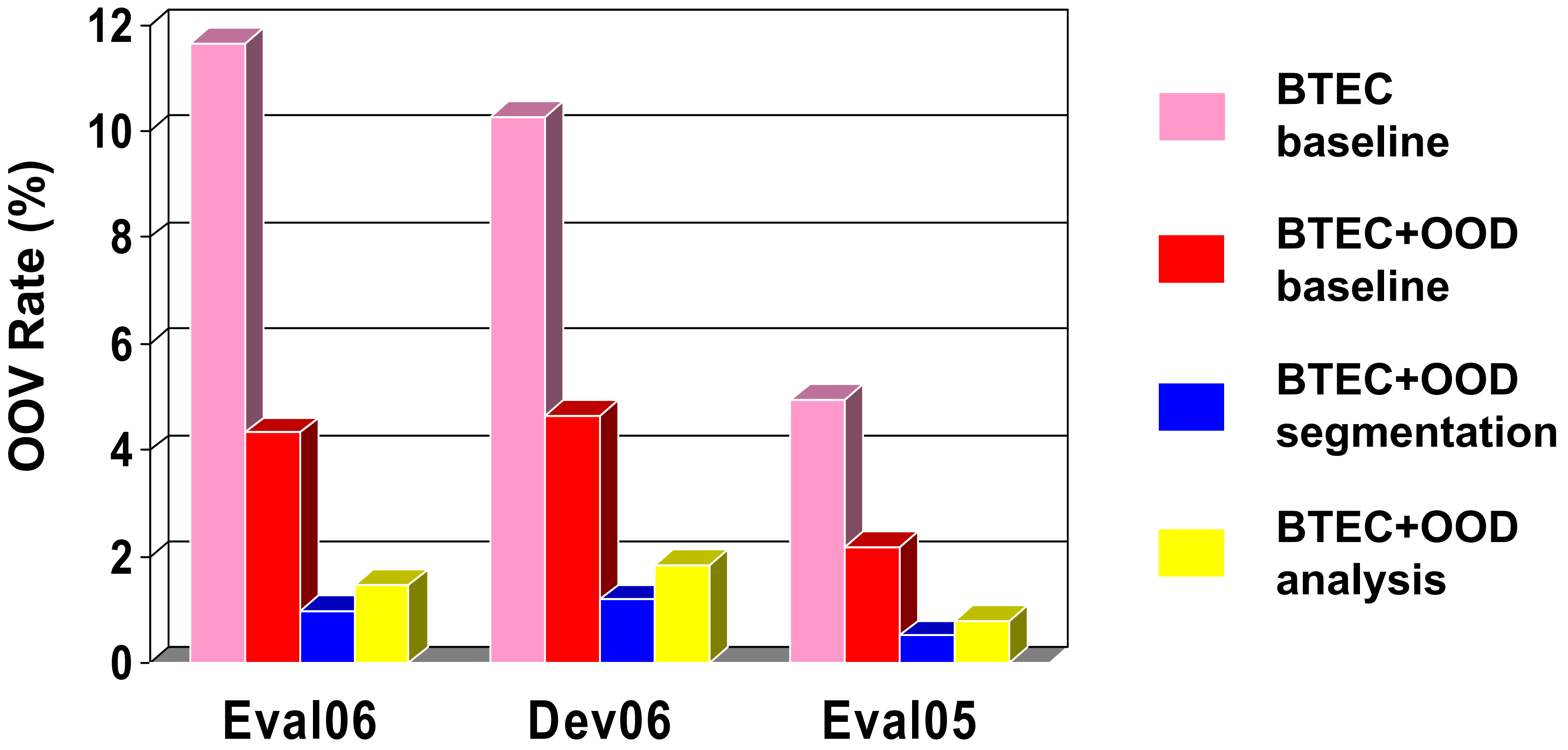
Translation Quality Improvement



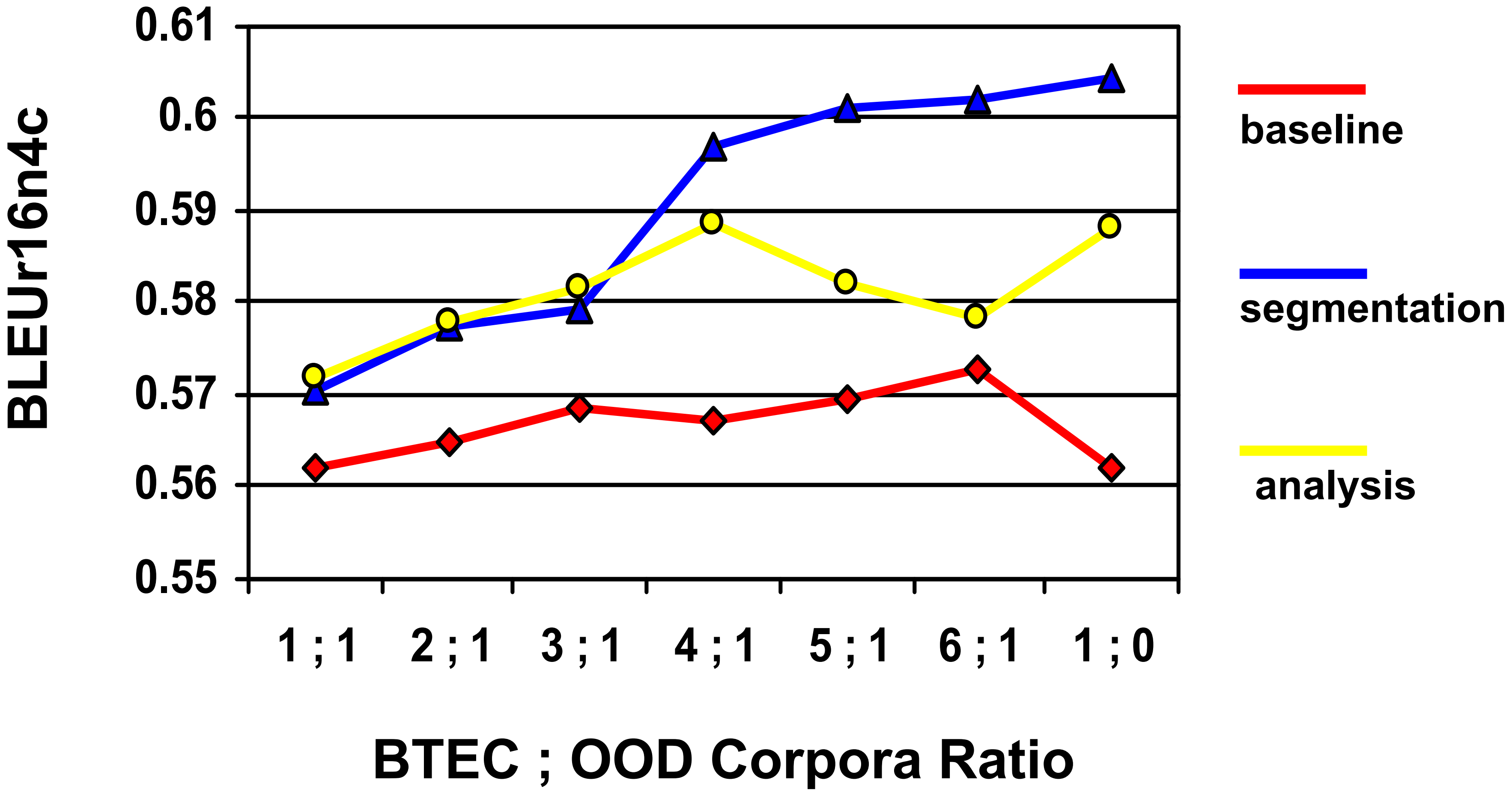
Out-of-Domain Corpora: Newswires

Source	# AR words	# EN words	# sent pairs
BTEC	159,213	189,239	20,000
LDC2003T18	26,146	33,869	1,043
LDC2003E05	103,717	129,181	4,235
LDC2003E09	123,505	150,865	5,003
LDC2004E07	520,971	681,613	20,358
LDC2004E11	227,792	310,079	8,576
LDC2004E08	1,771,893	2,207,934	52,042
LDC2005E46	616,879	819,354	24,874
LDC2001T55	70,183	80,354	2,346
FBIS	86,614	117,420	2,624
OOD Total	3,547,700	4,530,669	121,119

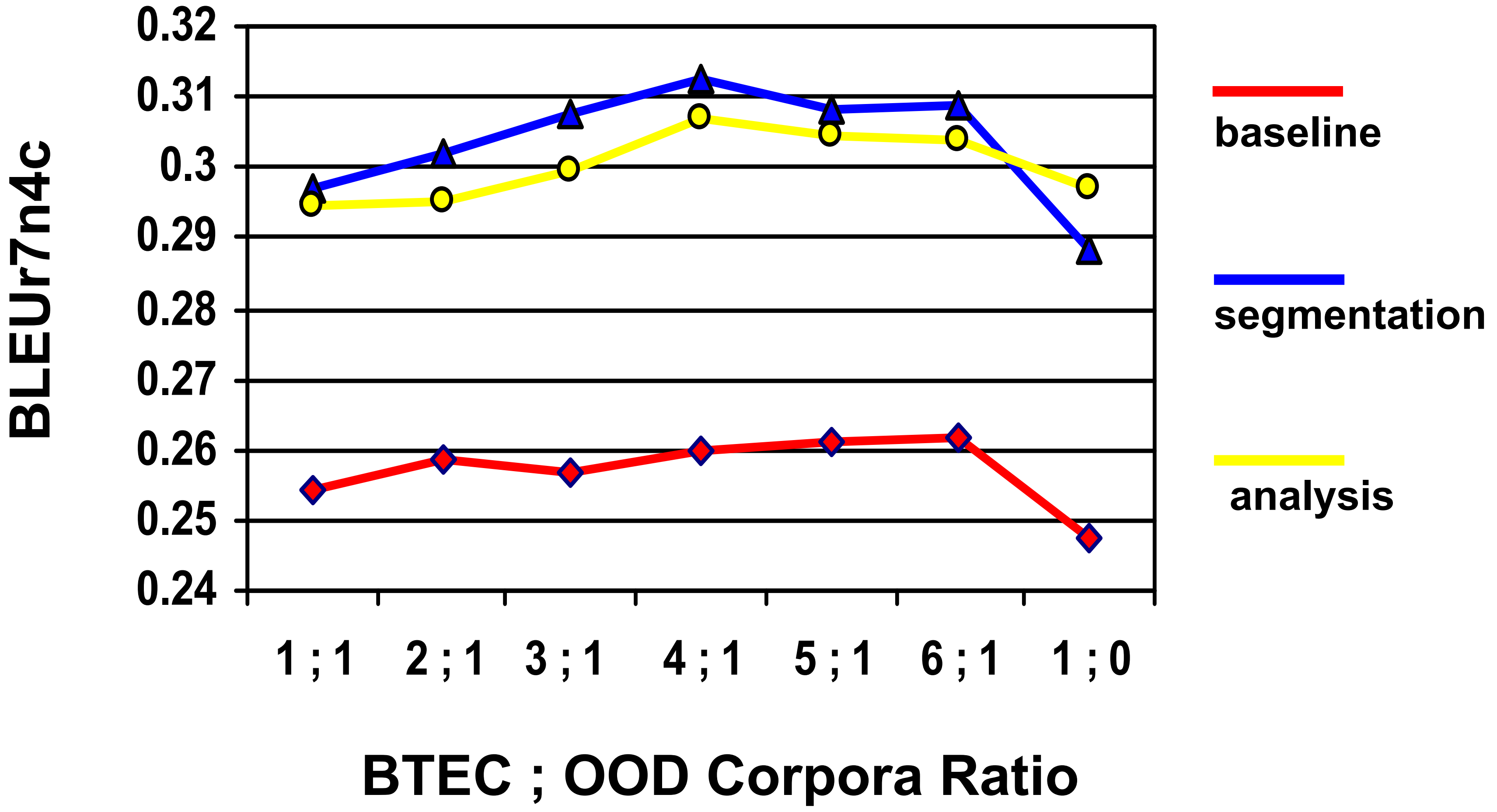
OOV Rate Reduction



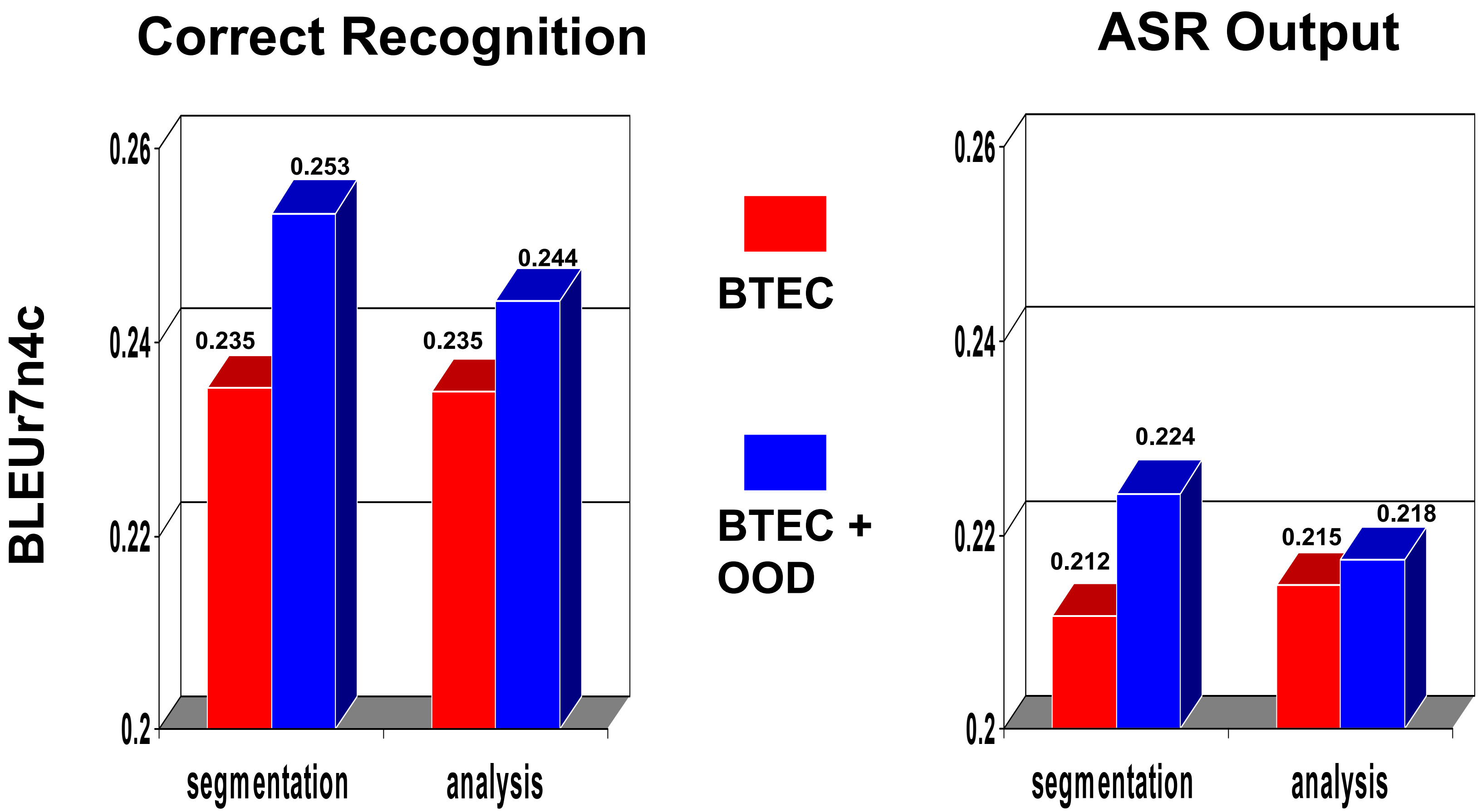
Eval05 Translation Quality Improvement



Dev06 Translation Quality Improvement



Eval06 Translation Quality Improvement



Correct Recognition

ASR Output

BTEC ; OOD Corpora Ratio = 4 ; 1

IWSLT 2006 Open Data Track Results

Correct Recognition Result					
	BLEU4	NIST	METEOR	WER	PER
Official	0.2549	6.3769	0.5316	0.5668	0.4825
Additional	0.2773	7.1681	0.5314	0.5593	0.4480
ASR Output					
	BLEU4	NIST	METEOR	WER	PER
Official	0.2274	5.8466	0.4845	0.6049	0.5198
Additional	0.2428	6.4867	0.4842	0.6035	0.4958

Scores in BLUE indicate the best scores under the given condition

Conclusions

- Techniques for improving translation quality & increasing vocabulary coverage
  - Word segmentation & morphological analysis
  - Proper combination of domain-specific and out-of-domain corpora for model training
- Effectiveness of the techniques
  - Demonstrated in the IBM Arabic-to-English translation system performances in the Open Data Track