AGILE's Progress in Speech to Text

Participating Sites:

BBN

Cambridge University (CU)

LIMSI

Overview

- Evaluation results and progress (Long Nguyen)
- Recognition units for Arabic STT (Long Nguyen)
- Recent progress on Arabic STT (Lori Lamel)
- Development of AGILE Chinese STT (Phil Woodland)

AGILE's Progress on Arabic STT

- **Significant reduction in word error rate (WER) for all development test sets**
  - 25% relative for broadcast news (BN)
  - 30% relative for broadcast conversation (BC)

System	tn6	tc6	dn6	dc6	eval06	dev07	eval07
Eval06	19.4	30.6	18.1	28.6	23.8	--	16.0
Eval07	14.4	21.0	13.3	18.6	17.0	10.3	11.8
Rel. Gain	*25.7%*	*31.4%*	*26.5%*	*35.0%*	*28.6%*	--	*26.3%*

- **Notes:**
  - tn6 & tc6: BN and BC subsets of the main AGILE tuning set
  - dn6 & dc6: BN and BC subsets of AGILE dev06

AGILE's Progress on Arabic STT (cont)

- Team's STT final output is ROVER combination of outputs from BBN, LIMSI, and CU
- Significant progress due to:
  - Multiple complementary systems
  - Improved acoustic models based on either graphemes or phonetics and word- or morpheme-based lexical units
  - Dual audio segmentations to accommodate mixed BN and BC testing material
  - Utilization of all available training data

AGILE's Progress on Mandarin STT

- **About 25% relative reduction in character error rate (CER) for both Phase-2 development (dev07) and evaluation (eval07) test sets**

System	eval06	dev07	eval07	retest
Eval06	17.5	12.0	11.4	--
Eval07	16.1	10.0	9.3	--
ReTest	15.3	9.2	8.5	7.8

- **Final output produced by CU's system after cross-adapting to BBN's output**

Key Contributions for Mandarin STT

GALE
The AGILE
Team

- **Improved pitch feature extraction algorithm**
- **Developed complementary systems for better system combination**
- **Utilized all available training data**
- ***(further details of progress to be presented later by Phil Woodland)***

Summary

- Made significant progress in STT for both Arabic and Mandarin for Phase-2 Evaluation
- Made more progress for Mandarin during the Re-Test
- Still need to improve STT performance further to achieve better MT results to hopefully attain the challenging Phase-3 Evaluation targets

Recognition Units for Arabic STT

Introduction

- **Arabic vocabulary is very large due to its morphological complexity**
  - **Estimated to be about 60 billion unique words (or surface forms)** [K. Darwish, "Building a shallow Arabic morphological analyzer in one day," *Proc. ACL workshop on computational approaches to semitic languages,* 2002]
- **Decent Arabic STT lexicons using surface forms have to be sufficiently large, but…**
  - **Obtaining phonetic pronunciations is not straight forward**
  - **High out of vocabulary rate is an inherent problem**
- **Explored using words or morphemes as STT recognition units**
  - **For word-based system, use either real phonetic pronunciations or just graphemes**

Phonetic System

- Use words as recognition units
- Each word is modeled by one or more sequences of phonemes of its phonetic pronunciations
- Pronunciations are derived from Buckwalter morphological analyzer or looked up in fully-vowelized Arabic Treebank corpus
  - Only about 800K of the 1.3M words of the STT language model data can have pronunciations obtained by this procedure
- Recognition lexicon consists of 333K words (filtered from the 400K most frequent words)

Graphemic System

- **Also use words as recognition units**
- **Each word is modeled by a sequence of letters of its spelling**
  - ***Pronunciations*** **are deterministic (hence automatic)**
- **Recognition lexicon consists of 350K most frequent words**
- **Performance almost as good as that of a comparable phonetic system**

Morphemic System

- **Use morphemes as recognition units**
- **Morphemes determined by a simple morphological decomposition using a set of affixes and a few rules**
  - Details can be found in our ICASSP06 paper *'Morphological Decomposition for Arabic Broadcast News Transcription'*
- **Morpheme's pronunciations are derived from words' pronunciations during the decomposition process**
- **Recognition lexicon consists of 65K morphemes**
- **Performance almost as good as that of a comparable phonetic system**

Comparison and Combination of Results

- **Comparable performance individually but they all seem to complement each other pretty well such that combination of all three provides substantial reduction in WER**

System	eval06	dev07	eval07
Phonetic	20.0	11.8	14.0
Graphemic	19.8	12.8	14.6
Morphemic	20.7	12.4	14.4
Combination	*18.5*	*11.1*	*12.9*

Dictionary Expansion

- **Since Buckwalter morphological analyzer does not cover all possible words, some automatic approach to generate phonetic pronunciations is required**
- **Developed simple multi-gram-like rules based on graphemes and existing phonetic dictionary to derive new pronunciations**
  - Details in CU's ICASSP08 paper "Phonetic pronunciations for Arabic speech-to-text systems" [Diehl2008]
- **Obtained consistent gains when expanding recognition lexicons from 260K to 350K words**

Single Phonetic Pronunciation

- In addition to phonetic system (MPron) and graphemic system (Graph), a single-phonetic-pronunciation system (SPron) was developed at CU
  - Used either explicit or implicit short vowels and nunation modeling
  - Single pronunciations are derived from probabilistic rules based on multiple-pronunciation phonetic dictionary
  - Details also in [Diehl2008]
- Quite effective in multi-pass adaptation framework
  - Used in early pass (P2) to generate lattices for later rescoring and combination

System Combination

System P2 → P3		bcad06	bnad06	dev07
P3a	Graph → Graph	24.1	18.5	14.6
P3b	Graph → MPron	23.6	17.9	13.9
P3c	SPron → MPron	23.8	17.8	13.6
P3d	SPron → SPron	25.0	18.9	14.5
P3a + P3b		22.5	17.2	13.7
P3a + P3c	CNC	22.5	17.0	13.1
P3a + P3d		23.0	17.6	13.6

- Cross-adapting SPron → MPron (P3c) best individual system
- Consistent gains from combining Graph and MPron (P3a + P3b)
- Best gains from combining Graph and cross-adapted MPron (P3a + P3c)
  - 3-way CNC gave no additional gains (often slight degradation)

Summary

- **Word-based systems, either phonetic or graphemic, and morpheme-based systems can have comparable performance individually but combine effectively**
- **Automatic generation of Arabic phonetic *pronunciations* is possible for STT**
- **Even though the underlying STT technologies are language independent, more language-specific developments, such as morphological decomposition and automatic generation of phonetic pronunciations, are required to improve Arabic STT**

Update on Arabic STT at LIMSI

Lori Lamel, Abdel. Messaoudi, Jean-Luc Gauvain, Petr Fousek

Gale PI meeting
Tampa
April 7-8, 2008

GALE
The AGILE
Team

Objective: Improve Arabic STT

- Improve acoustic, lexical and language models
- Morphological decomposition
- Probabilistic features
- Results
- Summary and some other research directions

Morphological Decomposition

- Several sites have been investigating morphological decomposition to address the huge lexical variety in Arabic
- Initial decomposition experiments with a rule-based approach
  - Based on Buckwalter analysis with heuristics
  - If multiple decompositions are possible, keep the longest prefix
  - Residual root word must not be a compound word
  - Root must contain at least 3 letters and be in lexicon
  - Only one decomposition is allowed for a given word
- Extensions: affixes for dialect, limiting decomposition

Morphological Decomposition - Dialect Affixes

- Decomposition rules typically fail on words in dialect
- Some of the differences are due to dialectal affixes
- Set of dialectal affixes added to the Bulkwalter prefix table
  - hAl (*this + the*): 45%
  - EAl (*over + the*): 25%
  - bhAl (*with/by + this + the*): 9%
  - E (*over*): 7%
  - whAl (*and + this + the*): 6%
  - wEAl (*and + over + the*): 5%
  - lhAl (*to/for + this + the*): 3%
- MSA may have several possible final vocalized forms, in dialect the final vowel is usually absent (a sekoun)

Morphological Decomposition - 3 Variants

- Version 1: Decompose the following affixes based on Buckwalter:
  - 12 prefixes with 'Al': Al wAl fAl bAl wbAl fbAl ll wll fll kAl wkAl fkAl
  - 11 prefixes without 'Al': w f b wb fb l wl fl k wk fk
  - 6 negation prefixes: mA wmA fmA lA wlA flA
  - 3 prefixes future tense: s ws fs
  - suffixes (possessive pronouns): y, ny, nA, h, hm, hmA, hn, k, kmA, km, kn
  - 7 dialect affixes
- Version 2: forbid decomposition of the most frequent 65k words
- Version 3: restrict decomposition of 'Al' preceding solar consonants (t, v, d, g, r, z, s, $, S, D, T, Z, l, n), since 'l' is often assimilated with consonant

  V2: wbAlslAm = w+b+Al+slAm $\longrightarrow$ wbAl + slAm
  V3: $\longrightarrow$ wb + AlslAm

Morphological Decomposition - Results

bnat06	*Vocab. size*	*WER (%)*
Reference word based	200k	22.0
Decomposition version 1	270k	24.0
Decomposition version 2, LM	300k	22.3
Decomposition version 2, LM + AM	300k	22.1
Decomposition version 3, LM	320k	21.6

- Jun07 acoustic model training set
- Small language model training set: 100M words
- 1 pass decoder

Morphological Decomposition - Results

Conditions	*bnat06*	*bnad06*	*bcat06*	*bcad06*	*eval06*	*dev07*	*eval07*
Baseline	16.7	15.5	22.8	20.4	19.3	12.4	13.7
Decomp.	16.7	15.3	23.1	20.6	19.4	12.2	13.8
Combin.	16.1	14.9	22.3	19.7	18.5	11.8	13.2

- 1200 hour acoustic model training
- Same AMs for both conditions (sub-optimal)
- Full language model training (1.1B words), NN LM, 290K
- Full training/testing does not validate earlier results
- 3 pass decoder
- Combination gives 0.6% gain across test sets

MLP Features

- PLP9 – 9 frames of PLP (wider context 150ms)
- LP-TRAP features [Hermansky & Sharma, TRAPs - classifiers of TempoRAl Patterns, *ICSLP'98*; Fousek, Extraction of Features for Automatic Recognition of Speech Based on Spectral Dynamics, 2007]
- Bottle-neck MLP [Grézl, Karafiát, Kontár & Černocký, Probabilistic and Bottle-Neck Features for LVCSR of Meetings, *ICASSP'07*]
- Feature vs system combination
  - combine raw features at the MLP input
  - concatenate MLP features (78 fea)
  - cross-adaptation
  - ROVER combination

MLP training

MLP targets	MLP train data	WER(%)
phones	1.5 hrs	27.3
	17 hrs	25.3
	170 hrs	25.0
states	17 hrs	24.7
	63 hrs	24.2
	301 hrs	23.4
	1168 hrs	22.2

- 400 hour HMM training
- MLP training from 1.5 hours to 1168 hours
- 1 pass decoder, MLP_{9xPLP}, 39 features

Feature combination

- Feature concatenation (39+39 $\longrightarrow$ 78)
- MLP combination (39+39 $\longrightarrow$ 39)
- MLP trained on 63 hrs, HMM trained on 400 hours

Features	Pass 1 WER
PLP	25.1
MLP_{9xPLP}	24.2
MLP_{wLP}	25.8
MLP_{comb}	23.8
PLP + MLP_{9xPLP}	22.7
PLP + MLP_{wLP}	21.7
MLP_{9xPLP} + MLP_{wLP}	22.2

Raw features:

9xPLP: 9 frames of PLPs, 9 $\times$39 = 351 features

wLP: wLP-TRAP, 19 bands $\times$ 25 features = 475 features

comb: concatenation of wLP and 9xPLP = 826 features

- Best results obtained with feature vector concatenation

Experimental Results (1)

- 1200 hour acoustic model training, MMI training
- Full language model training (1.1B words)
- 1 pass decoder

Conditions	*bnat06*	*bnad06*	*bcat06*	*bcad06*	*eval06*	*dev07*	*eval07*
Baseline	18.8	17.5	25.3	22.4	21.6	14.5	16.1
MLP	18.1	17.0	24.2	21.9	21.4	13.9	15.6
PLP+MLP	16.7	15.7	22.6	20.0	19.9	12.8	14.2

Experimental Results (2)

- 1200 hour acoustic model training, MMI training
- Full language model training (1.1B words), NN LM, 290K
- 2/3 pass decoder

Conditions	*bnat06*	*bnad06*	*bcat06*	*bcad06*	*eval06*	*dev07*	*eval07*
Baseline	16.7	15.5	22.8	20.4	19.3	12.4	13.7
PLP+MLP	15.4	14.3	21.1	18.6	18.4	11.6	13.0
Comb.	15.0	13.8	20.7	18.3	17.7	11.2	12.4
+ Decomp	14.5	13.2	20.2	17.9	17.1	10.6	11.9

- MLLR and SAT work with MLP features, but the gain is less than for PLP features.

Summary

- Explored different ways to combine MLP and PLP features
- ROVER and feature concatenation better than feature combination and cross-adaptation
- Morphological decomposition system performance close to word based system, and combines well with word-based system
- Other ongoing research:
  - Reducing supervision for acoustc model training
  - Using generic vowel model in recognition lexicon
  - Pitch and duration modeling
  - Continuous space language modeling

Development of AGILE Chinese STT

Andrew Liu, Kai Yu, Mark Gales, Phil Woodland,
Tim Ng, Bing Zhang, Kham Nguyen, Long Nguyen

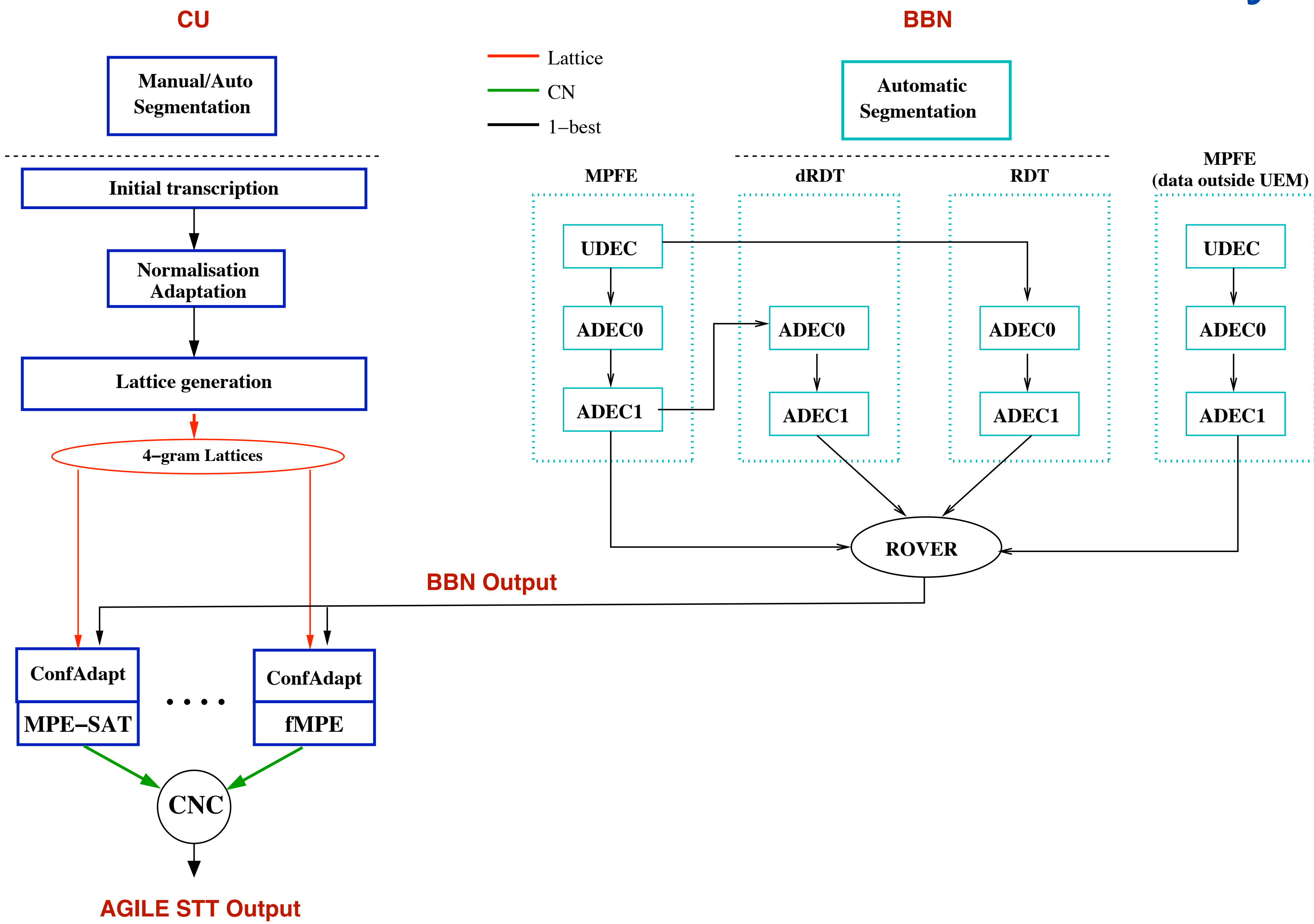
April 8th 2008

Cambridge University Engineering Department
BBN Technologies

Overview of AGILE Chinese STT

- Progress since June 2006, to June 2007 evaluation and 2007 retest.
- Overall system architecture remains the same.
- Cross-adaptation of CU system using BBN hypotheses (BBN→CU)
- Optimized for STT-MT integration.
- Preserving STT character to word tokenization for translation.
- Analysis of the effects of manual segmentation

Overall Architecture of AGILE Chinese Retest STT System



Chinese STT Improvements (Automatic Segmentation)

System	eval06	dev07	eval07	dev08
Jun'06 CU	17.4	12.9	12.3	—
Jun'06 BBN	19.3	14.3	13.2	—
Jun'06 BBN→CU$^\diamond$	16.6	12.0	11.4	—
Jun'07 CU	16.1	10.9	10.4	—
Jun'07 BBN	16.1	10.4	9.5	—
Jun'07 BBN→CU*	15.1	10.0	9.3	—
Dec'07 CU	15.1	9.8	9.2	9.1
Dec'07 BBN	15.7	9.5	8.8	8.8
Dec'07 BBN→CU†	14.4	9.2	8.5	8.3

CER improvements of of AGILE Chinese STT systems using automatic segmentation,
$\diamond$ 2006 evaluation system, $*$ 2007 evaluation system, $\dagger$ 2007 retest system

- Dec'07 improvements up to 2.9% CER reduction on eval07.
- 25% relative improvement over Jun'06 system, 9% over Jun'07.

Improved Acoustic Models (BBN)

AM	Pitch	eval06	dev07	eval07	dev08
520hr	Old	19.0	14.0	-	-
1370hr		17.2	11.6	-	-
1370hr	New	16.6	10.3	9.8	9.4
1567hr		16.6	10.4	9.6	9.0

BBN Jun'06(520hr), Jun'07(1370hr) and Dec'07(1567hr) acoustic models using auto seg.

- Improved CER by 2.4% (eval06) to 3.6% (dev07) [tuning set]
  - additional 1047 hours more of speech training data;
  - improved pitch feature extraction: 0.6% CER reduction (eval06)
    - ESPS style new pitch detection algorithm (RAPT);
    - linear interpolation of log pitch values across unvoiced regions.
- refined audio segmentation: up to 0.2% CER reduction.
- using data outside UEM time boundaries for adaptation: up to 0.1% gain.

Improved Language Models (BBN)

- Language modelling: 0.7% CER reduction on dev07, 0.4% on dev08.

- additional 1.1G characters of texts and more data sources: 0.1% gain on both dev07 and dev08.

Data Source	#Char
GALE releases up to P3R1	25M
LDC Giga Word version 3	238M
CU web data collection	381M
IBM Sina data	450M
Total	1094M

- improved grouping of data sources: 0.2% gain on dev07, 0.1% on dev08.

- using dev07 as tuning set in training and decoding: 0.4% gain on dev07, 0.2% on dev08.

Improved Acoustic Models (CU)

AM	bnmdev06	bcmdev05	dev07
Jun'06	10.5	21.5	17.6
Jun'07	9.5	20.3	14.3
Dec'07	9.3	19.7	13.5

CU Jun'06, Jun'07 and Dec'07 acoustic models on auto seg and Jun'07 LM

- Acoustic modelling: improved CER by up to 4.1% (23% relative) on dev07.
  - additional 1120 hours more of speech training data;
  - refined processing of audio transcriptions;
  - improved pitch feature extraction: 0.3% to 0.5% CER reduction.
    - PCHIP interpolation on both voiced and unvoiced regions;
    - 5-point average smoothing of log pitch using a Gaussian window.
- Also added fMPE branch which gives small gains in combination: up to 0.1%
- Multiple segmentations can yield further gains (typically 0.2-0.3% but not used due to impact on translation).

Improved Language Models (CU)

LM	bnmdev06	bcmdev05	dev07
Jun'06	7.9	18.0	11.7
Jun'07	7.9	17.6	11.4
Dec'07	7.5	17.8	10.6

Pass 2 CER performance using automatic segmentation and Dec'07 MPE AMs

- Language modelling: improved CER by up to 1.1%.
  - additional 1.7G words of texts and more text sources
  - significant increase in model size, e.g., 4-grams from 7M to 56M.
  - expanded vocabulary including more English acronyms.
  - improved training/interpolations configurations.

- Other LM techniques investigated:
  - character to word segmentation: increased word-list, no gain.
  - language model adaptation: discriminative/perplexity based, no gain.

Manual vs Automatic Segmentation

- Overlapping speech introduces issues in reference transcriptions
  - multiple segment references exist in overlapped speech
  - possible to select a single reference (e.g. longest or first) in overlap regions
    - used in automatic segmentation system evaluation

Ref. in Overlap regions	auto	manual
Single	9.6	9.0
Multiple	—	10.3

CU Dec'07 AM with Jun'07 LM, BBN→CU system performance on dev07 test set

- Large CER reductions possible using manual segmentation: 0.6%
  - sensitive to performance of automatic segmenter on particular test set
- Scoring all manual segments significantly worse performance
  - overlapped data performance 25%-50% CER depending on % overlap
  - performance excluding all overlapping data 7.9% CER

Manual vs Automatic Segmentation (cont)

- Schemes investigated for using manual segmentation/overlapped speech
  - further segmenting manual segmentation into "sentences": no gain
  - use unadapted (initial pass) output for single character words in overlap regions (used in retest): small gain in overlapped region

Performance Gains on eval07sub

- eval07sub is a 1 hour subset of `eval07` re-used in the December retest
  - data not used for tuning of any systems

System	Segmentation	eval07sub
Jun'07 CU	auto	9.5
Jun'07 BBN		8.3
Jun'07 BBN→CU*		8.4
Dec'07 CU	auto	8.4
Dec'07 BBN		7.9
Dec'07 BBN→CU		7.7
Dec'07 CU	manual	7.8
Dec'07 BBN	auto	7.9
Dec'07 BBN→CU†	manual	7.3

System Combination performance using a single reference in overlap regions,
∗ 2007 evaluation system, † retest system.

Performance Gains on eval07sub (cont)

- Significant gains from evaluation (Jun'07) to retest (Dec'07) system:
  - 13% relative (1.1% absolute) reduction in CER
- Gains from manual segmentation less than on dev07
  - 0.6% using CU-only system, 0.4% using cross-adaptation

Conclusion/Summary

- Overall 25% relative reduction in CER from Jun'06 to Dec'07
- Same overall cross-adaptation architecture for system combination
  - CER improvements from ROVER possible but impact on translation
- Significant improvements at both BBN and CU in both Acoustic Models and Language Models
- Improved processing procedures and algorithms (e.g. pitch processing, fMPE, adaptation etc)
- Used new GALE (LDC+contributed) training data resources
- Discussed impact of manual segmentation