IWSLT 2007

Overview of this year's Evaluation Campaign

Cameron Shaw Fordyce
fordyce@celct.it

IWSLT: from past to future*

- IWSLT 2003-2007

campaign	CSTAR03	IWSLT04	IWSLT05	IWSLT06	IWSLT07
description	closed eval. of MT engines	feasibility of MT tech., metrics	feasibility of speech trans.	robustness of speech trans. tech.	spontaneous dialogues
translation	text	text	read speech	spontaneous speech	dialogues
input data	text	text	ASR output	text, speech input	ASR output, text
language pairs	CE JE IE KE	CE JE	CE, EC JE AE KE	CE JE AE IE	CE JE AE IE
participants	6	14	19	19	24
submissions	6	28	65	73	75

Outline of Overview

- Evaluation Campaign
  - Challenge and Classical Tasks
  - Data Preparation
  - Input and Data Track Conditions
  - Evaluation Specifications
- Evaluation Results
  - Human Evaluations
  - Automatic Evaluations
- Conclusions
  - Challenge and Classical Tasks
  - Human Evaluation

Challenge and Classical Tasks

- Travel Domain with speech inputs
- Challenge Tasks for 2007
  - Two language directions:
    - Chinese to English
      - Semi-spontaneous speech input (IWSLT06)
    - Italian to English
      - Spontaneous speech in dialogues
- Classical Tasks for 2007
  - Read speech inputs
  - Two language directions:
    - Japanese to English
    - Arabic to English
  - Repetition of IWSLT05 Task

Challenge and Classical Tasks

- Travel Domain
- Challenge Tasks for 2007
  - Two language directions:
    - Chinese to English
      - Semi-spontaneous speech input (IWSLT06) Text input
    - Italian to English
      - Spontaneous speech in dialogues
- Classical Tasks for 2007
  - Read speech inputs
  - Two language directions:
    - Japanese to English
    - Arabic to English
  - Repetition of IWSLT05 Task

Schedule

- Training/Dev Data Release: 30 April 2007
- LR Links Due: 09 June 2007
- Test Data Release: 09 July 2007 -- 23 July 2007
- Translations Due: 13 July 2007 -- 27 July 2007
- Technical Papers Due: 06 August 2007
- System Reports Due: 13 August 2007

BTEC Corpus

- BTEC Corpus: Basic Travel Expression Corpus
  - Travel Expressions like those found in traveler's phrasebooks
    - “Where is the restroom?”
  - 172k sentence pairs collected/translated by C-STAR partners
  - Languages available: Arabic, Chinese, English, Italian, Japanese, Korean

Adam Corpus

- SITAL/ADAM: for Challenge Task Italian to English
  - Domain: tourism and train transportation, reservations
  - Recorded speech of simulated travel agent – client dialogues
  - Transcriptions contain no punctuation or case information
  - Transcriptions include speech artifacts such as
    - repetitions: ***il il*** *costo*
    - restarts: *è proprio in centro* ***a no scusi nei dintorni*** *di verona*
    - fillers: *no no no ma otto e quindici* ***va beh*** *sì prima*
  - 200 human-human dialogues/250 human-machine dialogues
  - 58377 total words in human/human

Adam Corpus

IT: *buongiorno io vorrei prenotare un volo da roma a verona*

EN: *Good morning. I would like to book a flight from Rome to Verona.*

IT: *è proprio in centro a no scusi nei dintorni di verona*

EN: *It's right in the centre. No, I'm sorry. On the outskirts of Verona.*

IT: *beh se non c' è altro potrebbe prenotarmi questo anche*

EN: *Well, if there nothing else available, you can also book me on this flight.*

Adam Corpus: Translation Instructions

- Keep all meaning possible
- Make translation as natural and fluent as possible in target sentence
- Add capitalization and punctuation only when it is clearly necessary.
- Maintain tone of Italian
- Remove sparingly speech artefacts such as excessive repetitions
- Translator provided with standard glossary to normalize translations of some terms (city names, train names, etc.)

Data

- Training, Dev Data:
  - BTEC for all tasks from previous campaigns
  - SITAL for Italian to English task
  - All other "publicly" available and affordable resources
- Additional Data for Challenge Task for IE
  - ~996 sentences IT, 1 Reference Translation
- Test Data:
  - Classical Tasks
    - JE, AE, CE: 489 Sentences (6 references)
  - Challenge Tasks
    - IE: 724 Sentences (4 reference translations)

Evaluation Specifications

- Input Test Data requirement:
  - True-cased, and punctuation inserted
- Automatic Evaluation
  - all primary runs / contrastive runs evaluated/ranked
  - metric: BLEU (6/4)
- Human Evaluation:
  - Ranking(All primary runs)
  - Adequacy/Fluency(Top 3 ASR , CE runs)
  - 3 or more human evaluations for each system
  - 9 Evaluators: 8 paid with experience in MT evaluation
  - 300 randomly selected sentences from test sets

Top Systems: BLEU

Language	Task	BLEU	Group
AE	ASR	0.4555	UPC
	Clean	0.4933	TUBITAK-UEKAE
JE	ASR	0.4386	CMU-UKA
	Clean	0.4841	TUBITAK-UEKAE
IE	ASR	0.4229	FBK
	Clean	0.4531	RWTH
CE	Clean	0.4077	I2R

Human Evaluation: Ranking

- Evaluators are instructed to “rank” system outputs presented in groups of 3-5
- Each system is presented in groups with every other system
- Scale is relative to the other systems
- All systems/all tasks/all input conditions
- Easier task for Assessors
- Higher Levels of Intra/Inter-Annotator correlation

Human Evaluation:Ranking

Source: IWSLT07_CE_TEST_159\这个 电子 邮件 地址 能 收 邮件 吗 ？

Reference: IWSLT07_TEST_159\1\Can I receive email at this address ?

Translation	Rank
IWSLT07_CE_TEST_159\Can you mail this mail address ?	○ 1 Worst ○ 2 ○ 3 ○ 4 ○ 5 Best
Will this e mail mail ?	○ 1 Worst ○ 2 ○ 3 ○ 4 ○ 5 Best
IWSLT07_CE_TEST_159\Can you accept mail this mail address ?	○ 1 Worst ○ 2 ○ 3 ○ 4 ○ 5 Best
IWSLT07_TEST_159\This address E mail, will you accept mail.	○ 1 Worst ○ 2 ○ 3 ○ 4 ○ 5 Best
IWSLT07_CE_TEST_159\Can you mail this e-mail ?	○ 1 Worst ○ 2 ○ 3 ○ 4 ○ 5 Best
Annotator: cam **Task:** IWSLT07 Chinese English Clean	Annotate
Instructions: Rank each **whole sentence** translation from Best to Worst relative to the other choices (ties are allowed).	

Human Evaluation: Adequacy/Fluency

- Top 3 systems judged for all ASR input conditions, CE task
- Adequacy/Fluency judged together on same screen
- 5 systems presented together
- Scales from 1-5
- Scores normalized to account for evaluator variation

Human Evaluation: Adequacy/Fluency

Source: IWSLT07_TEST_463\أين الحمّام؟

Reference: IWSLT07_TEST_463\1\Where is the restroom ?

Translation	Adequacy	Fluency
IWSLT07_TEST_463-spk24_3\Where is the lavatory ?	○ ○ ○ ○ ○ 1 2 3 4 5	○ ○ ○ ○ ○ 1 2 3 4 5
IWSLT07_TEST_463\Where's the toilet?	○ ○ ○ ○ ○ 1 2 3 4 5	○ ○ ○ ○ ○ 1 2 3 4 5
IWSLT07_AE_TEST_463\Where is the lavatory ?	○ ○ ○ ○ ○ 1 2 3 4 5	○ ○ ○ ○ ○ 1 2 3 4 5
Annotator: cam **Task:** IWSLT07 Arabic English ASR NIST		Annotate
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Human Evaluation: Results

IE ASR		IE Clean	
SYSTEM	% BETTER	SYSTEM	% BETTER
FBK	**48.5**	FBK	**52.05.00**
RWTH	42.4	RWTH	50.06.00
ATR	40.2	ATR	45.09.00
UEDIN	29.0	MIT	33.01.00
UW	27.8	NTT	32.05.00
MIT	24.6	INESCID	28.09.00
NTT	24.2	HKUST	23.03.00
RALI	24.2	ITI	19.06.00

Human Evaluation: Results

JE ASR		JE Clean	
SYSTEM	% BETTER	SYSTEM	% BETTER
ATR	**27.03.00**	CMU	**32.7**
CMU-UKA	26.8	ATR	30.5
UEKAE	24.2	FBK	30.5
NTT	23.5	TOTTORI	28.0
FBK	23.3	UEKAE	27.4
DCU	19.2	NTT	27.3
HKUST	18.3	HKUST	21.9
		DCU	21.2
		GREYC	21.0

Participation in IWSLT 2007

- Group Participation

Language	Task	Groups
AE	ASR	10
	Clean	11
JE	ASR	7
	Clean	9
IE	ASR	10
	Clean	10
CE	Clean	15

- Continued and increased participation this year.

Groups

Group	Country	Type*	MT System	Tasks
NiCT/ATR SLC	JP	Phrase-based	NICT/ATR	IE,CE,JE
Ch. AS, Inst. of Comp. Tech.	CN	Syntax-based	ICT	CE
Ch. AS, Inst. of Automation,	CN	Phrase-based	CASIA	CE
Xiamen U., Sch. Inf. Sci. & Tech.	CN	Phrase-based	XMU	CE
U. J. Fourier, LIG Lab., GETALP	FR	SMT	LIG	AE
Tottori U., Faculty of Eng.,	JP	SMT	TOTTORI	JE
U. de Montréal, U. of Avignon	CA/FR	Phrase-based	MISTRAL	IE
GREYC, U. of Caen	FR	EBMT	GREYC	JE,AE
Inst. Infocomm Res., Dept. HLT	SG	SMT	I2R	CE
FBK - Fondazione Bruno Kesler	IT	SMT	FBK	IE,CE,JE

Groups(2)

Dublin City Univ.	IE	EBMT	DCU	CE,JE,AE
RWTH Aachen Univ.	DE	Phrase-based	RWTH	IE,CE
INESC-ID, SL Lab (L2F)	PT	SMT	INESC-ID	IE
MIT Lincoln Lab., AF Res. Lab.	US	SMT	MIT-LL	IE,CE,AE
NTT Comm. Science Labs	JP	Phrase-based	NTT	IE,CE,AE, JE
U. of Washington	UK	SMT	UW	IE,AE
InterACT, CMU, U. of Karlsruhe	US,DE	Syntax-augmented	CMU-UKA	CE,JE,AE
HK U. of Science and Tech.	CN	Phrase-based	HKUST	IE,CE,AE, JE
Institut Tecnològic d'Informàtica	ES	SMT	ITI/UPV	IE
Nat. Res. Inst. of Elec. and Crypt. & Sci., Tech. Res. Council	TR	Phrase-based	TUBITAK-UEKAE	JE,AE
U. of Maryland	US	Phrase-based	UMD	CE,AE

Reflections

- The Good
  - High levels of participation
  - Sharing of Resources
  - Focus on research goals rather than competition
  - Human Evaluation for All Tasks

- The Bad?
  - Rescheduling due to other important campaigns
  - Drop-out of CE Challenge Task due to problems with data preparation

Acknowledgements

- C-Star Partners. In particular,
  - ATR: Micheal Paul
  - FBK: Roldano Cattoni, Mauro Cettolo, Nicola Bertoldi
  - CMU: Matthias Eck, Mark Fuhs
- Others
  - Josh Schroeder
  - Chris Callison-Burch
  - John Henderson
- All the participants