

Toward Automatic Speech Interpretation

Nara Institute of Science and Technology
Data Science Center, and Graduate School of Science and Technology

Satoshi Nakamura

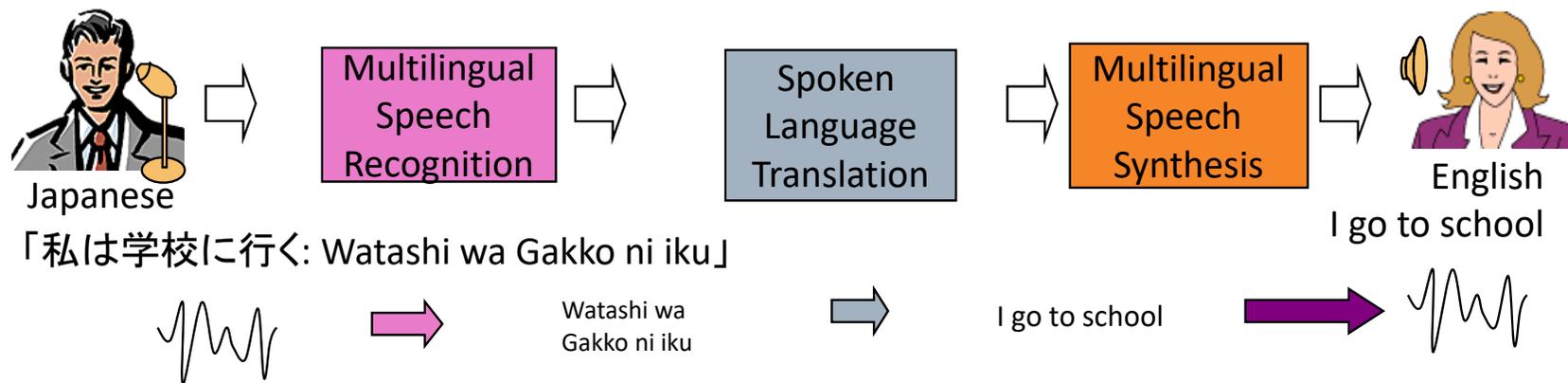
with

Katsuhito Sudo, Graham Neubig

Sakriani Sakti, Hiroki Tanaka,

Katsuki Chosa, Do Quoc Truong

Speech-to-Speech Translation System



Speech Translation and Text Translation

▶ Speech Translation

- Translation of spoken languages
- Speech recognition errors
- Translation from source language speech to target language speech (text)
- Short latency for real-time human communication

▶ Translation of Spoken Language

- Object is real-time communication and understanding
- Para-linguistic/non-linguistic information necessary
- Context dependent utterances, non syntactical utterances
- No punctuation
- No upper/lower case

Technical Background around 2000

▶ Corpus-based Approach

- Statistical modeling and large size training data

▶ Machine Translation

- Rule based:
Linguists created translation rules
- Corpus based :
 - Example-Based
Automatic extraction of translation rules [M.Nagao 1984 etc.]
 - Statistical MT (Statistical Machine Translation)
Extract rules statistically based on Noisy Channel Model
[P. F. Brown, et.al., 1993]

Contents

1. History of Automatic Speech Translation Research
2. Automatic Speech Interpretation Technologies
3. Current Project and Data Collection
4. Summary and Future Works

Speech Translation Projects

- ▶ **Japan**
 - ATR Speech-to-speech Translation (1986-2008)
 - NICT Speech-to-speech Translation (2008-2011, 2014-2020)

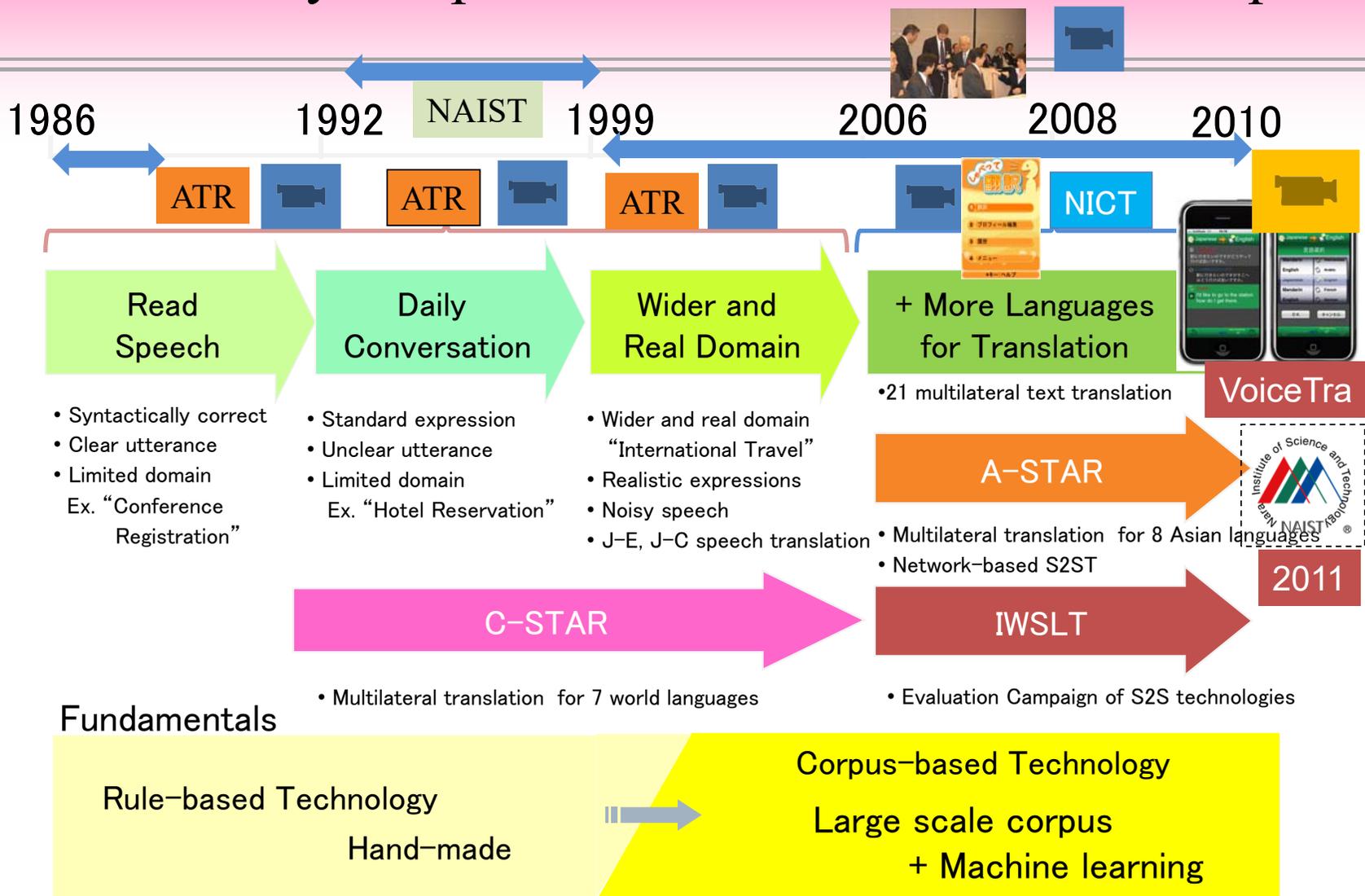
- ▶ **EU**
 - Verbmobile (1993-2000)
 - Nespole(2001-2003)
 - TC-Star(2004-2006)
 - EU-Bridge(2012-2014)

- ▶ **US**
 - DARPA TransTac, Communicator (2006-2010)
 - DARPA GALE(2006-2010)
 - DARPA BOLT(2011-2015)

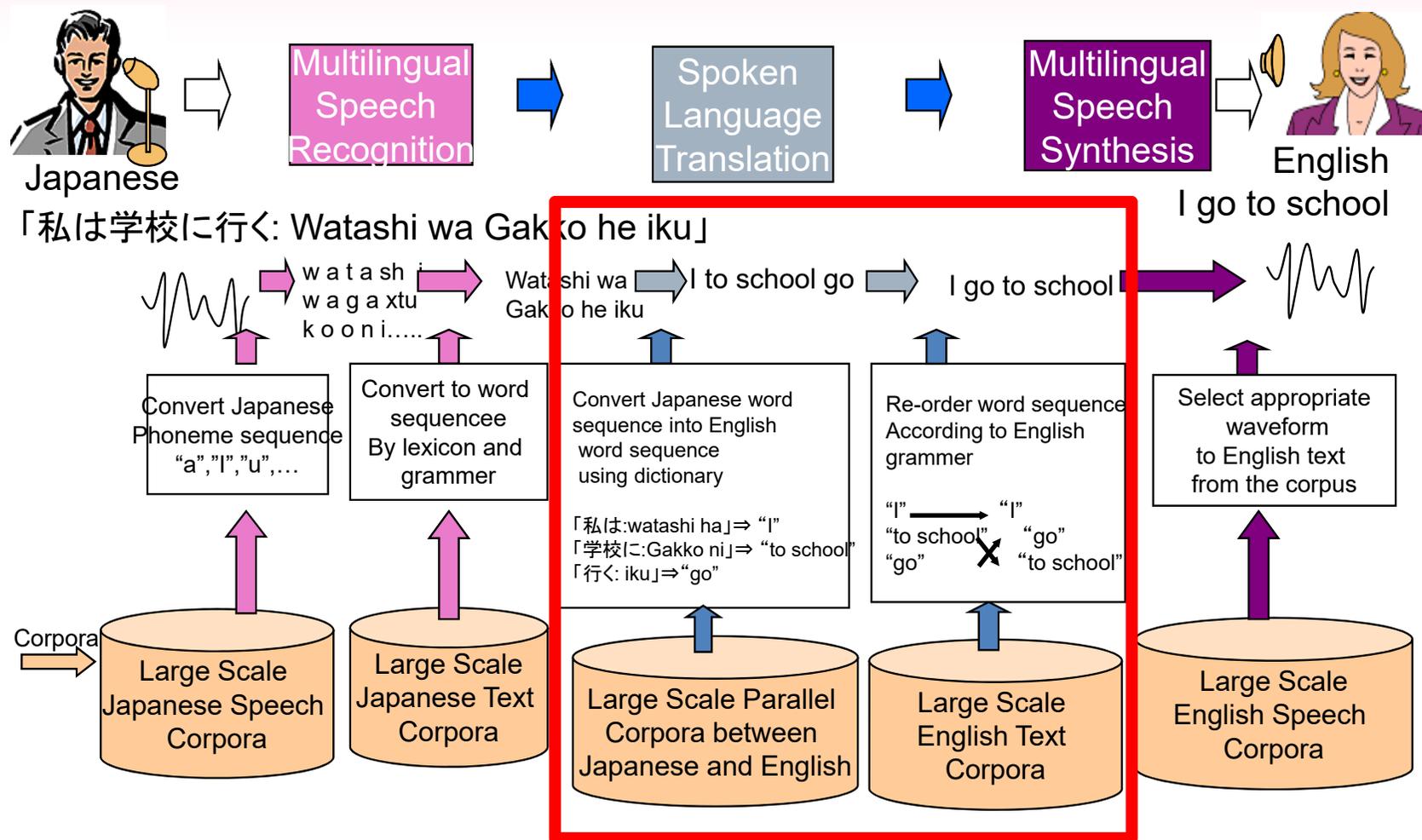
- ▶ **International**
 - C-Star Consortium (1991-2003)
 - IWSLT (2004-)

 - A-Star Consortium(2006-2008)
 - U-Star Consortium (2009-)

History of Speech Translation Research in Japan



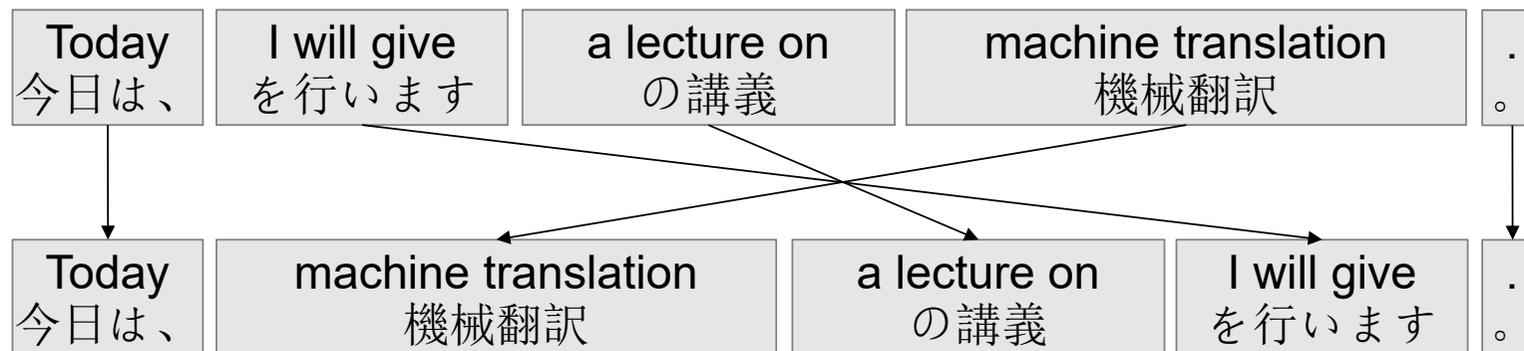
Mechanism of Speech Translation System



Phrase Based Machine Translation

- Divide the sentence into small phrases and translate

Today I will give a lecture on machine translation .



今日は、機械翻訳の講義を行います。
 kyowa kikaihonyaku no kogi wo okonaimasu

- Score translations with **translation model (TM)**, **reordering model (RM)**, and **language model (LM)**

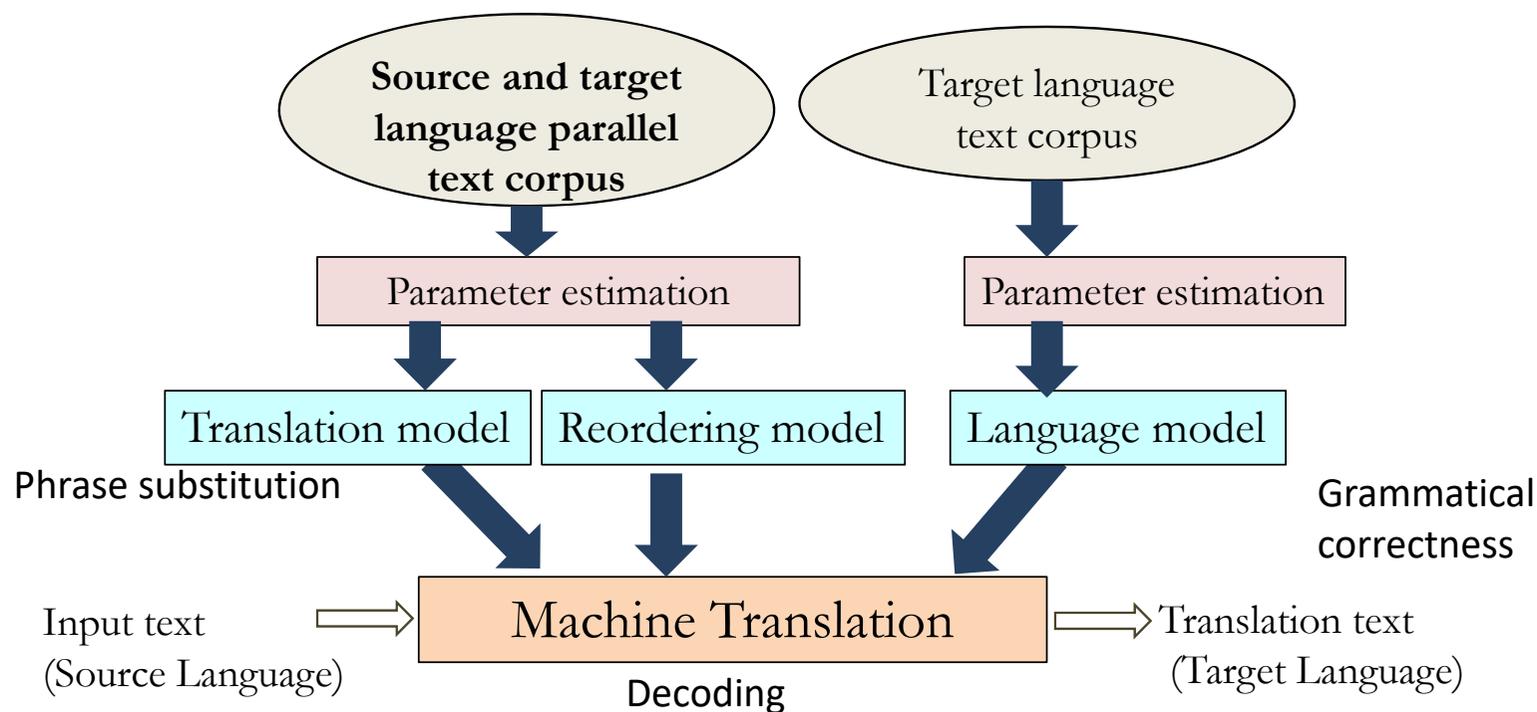
Translation Model Creation

- Perform **automatic alignment** of parallel text
- **Extract phrases** from the aligned text for translation



Statistical MT

- Translation Model, Reordering Model, Language Model



Parallel Corpus

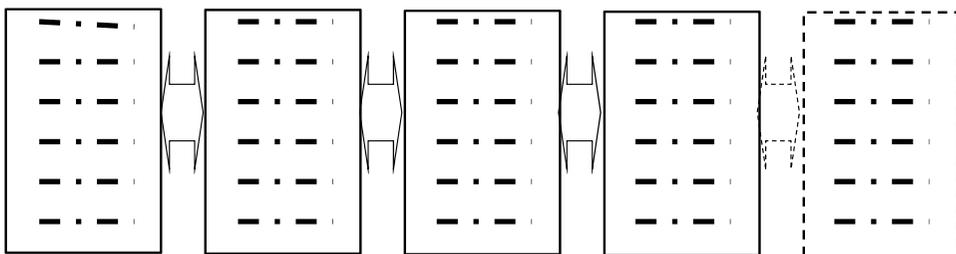
Japanese:

“mado wo aketemo iidesuka”

English:



1. may i open the window
2. ok if i open the window
3. can i open the window
4. could we crack the window
5. is it okay if i open the window
6. would you mind if i opened the window
7. is it okay to open the window
8. do you mind if i open the window
9. would it be all right to open the window
10. i'd like to open the window



ATR BTEC Corpus

Basic

12.2% (7)

- greet someone
- ask a question
- state one's purpose
- ...

Trouble

12.1% (20)

- luggage
- emergency
- medicine
- assistance
- ...

Shopping

10.0% (13)

- buy something
- gather information
- price
- wrapping
- ...

Move

8.4% (8)

- transportation
- buy a ticket
- rental car
- trouble
- ...

Stay

8.2% (11)

- make/change a reservation
- check-in
- trouble
- ...

Sightseeing 7.7% (11)

Restaurant 7.3% (11)

Communication 6.4% (6)

Airport 5.5% (14)

Business 5.3% (26)

Contact 4.0% (6)

Airplane 3.6% (11)

Homestay 2.3% (11)

Study Overseas 1.6% (14)

Drink 1.3% (4)

Exchange 1.2% (5)

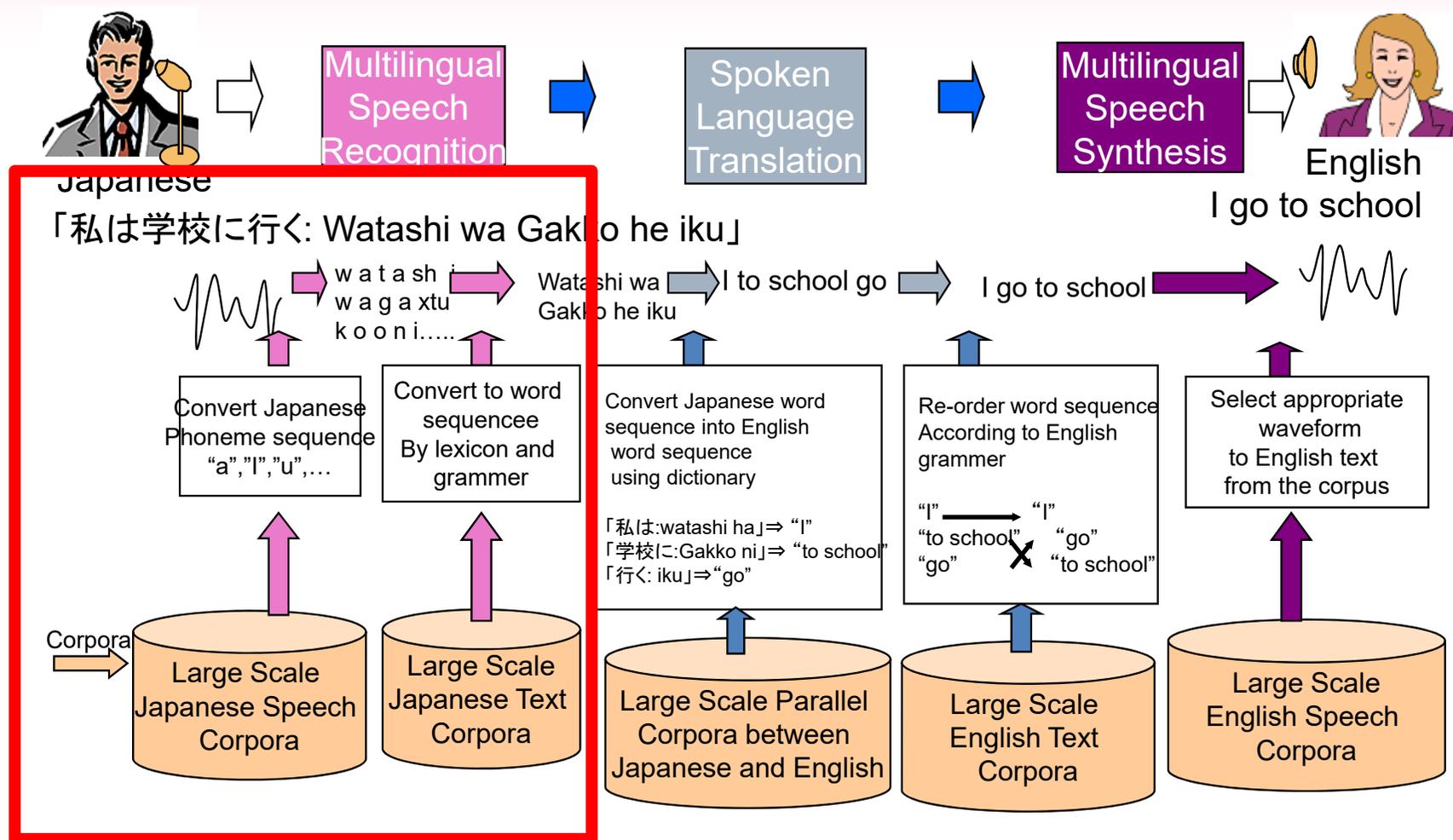
Snack 1.2% (4)

Beauty 0.8% (5)

Go Home 0.6% (4)

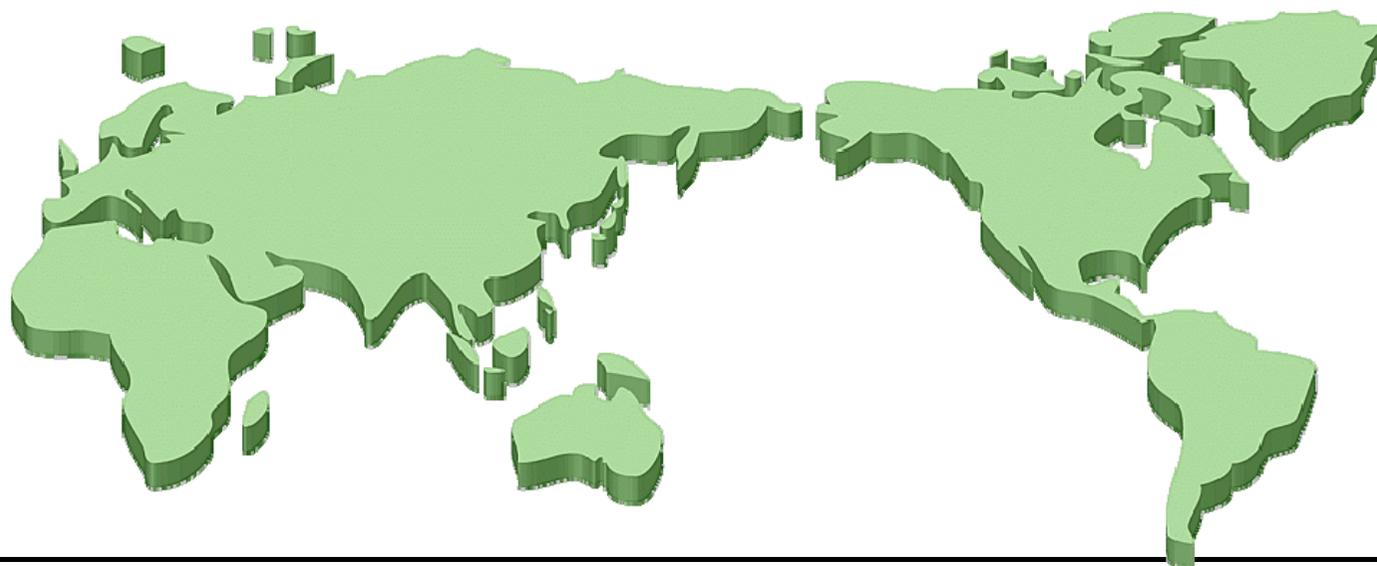
Research 0.1% (12)

Mechanism of Speech Translation System



Speech and Language Corpus for ASR

	Acoustic model	Language model
Japanese	4,200 speakers (271 hrs)	852k sentences
English	532 speakers (202 hrs) US, BRT, AUS	710k sentences
Chinese	536 speakers (249 hrs) Beijing, Shanghai, Canton, Taiwan	510k sentences



Speech to Speech Translation



Spoken Language
Communication
Research Laboratories

Launched in November 2007
The first network-based STS
translation service

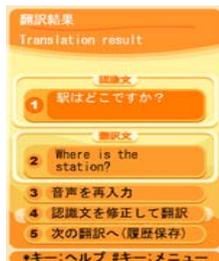
“Shabette Hon’yaku”
「しゃべって翻訳」



トップの画面



音声入力画面



翻訳結果出力画面

- Japanese-English
- NTTDocomo



905iシリーズ

- “VoiceTra” Network-based Speech Translation released on Jul. 2010
 - 21 language pair for Text I/O
 - 6 language pair for Speech I/O
- 800k download and 4M access worldwide as of 2011.3.

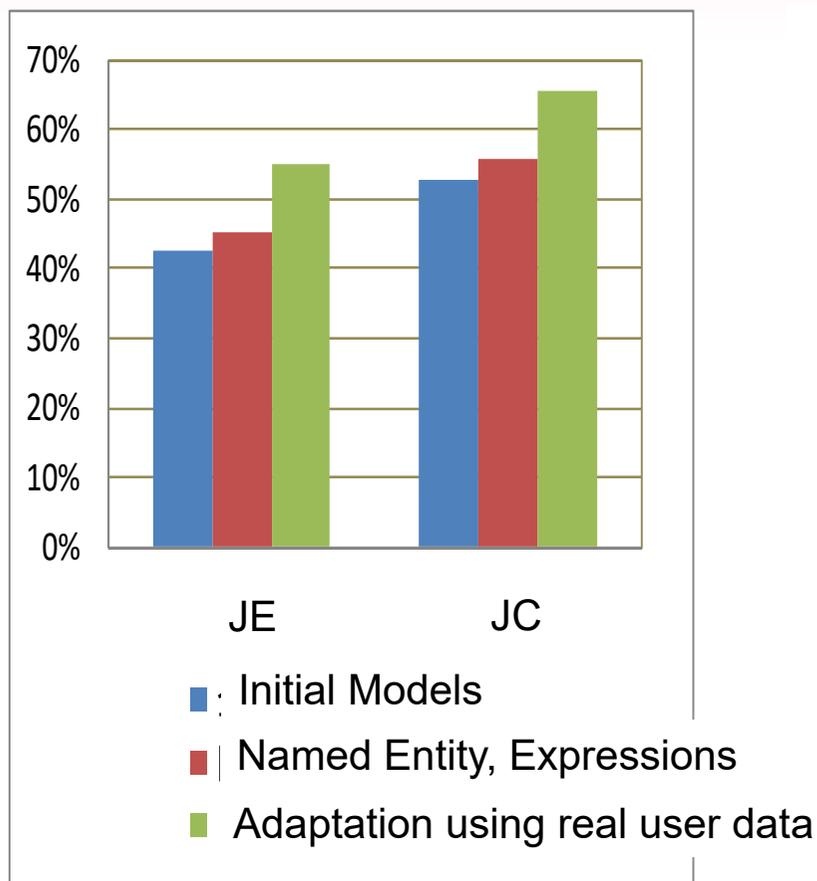


VoiceTra

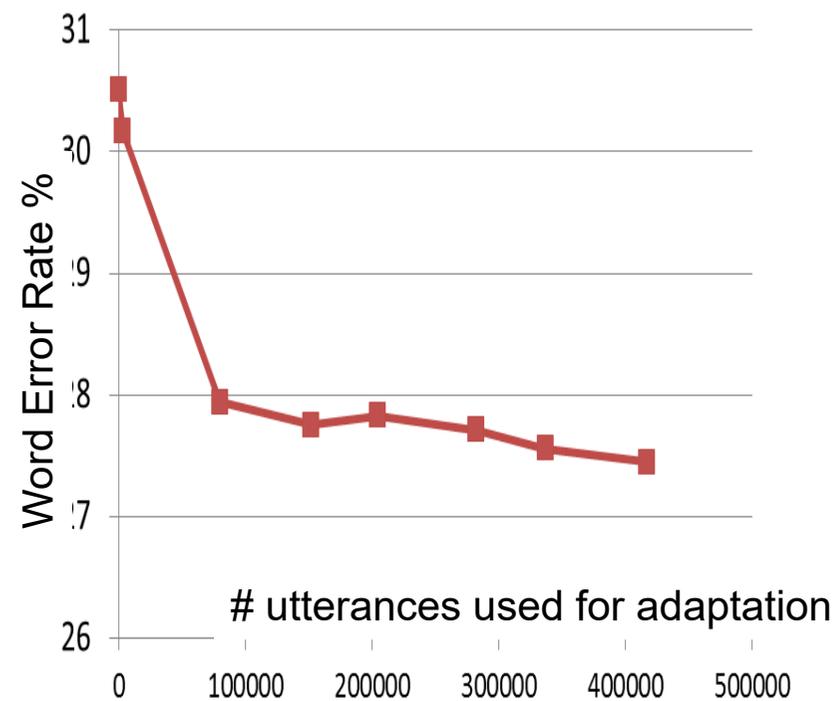


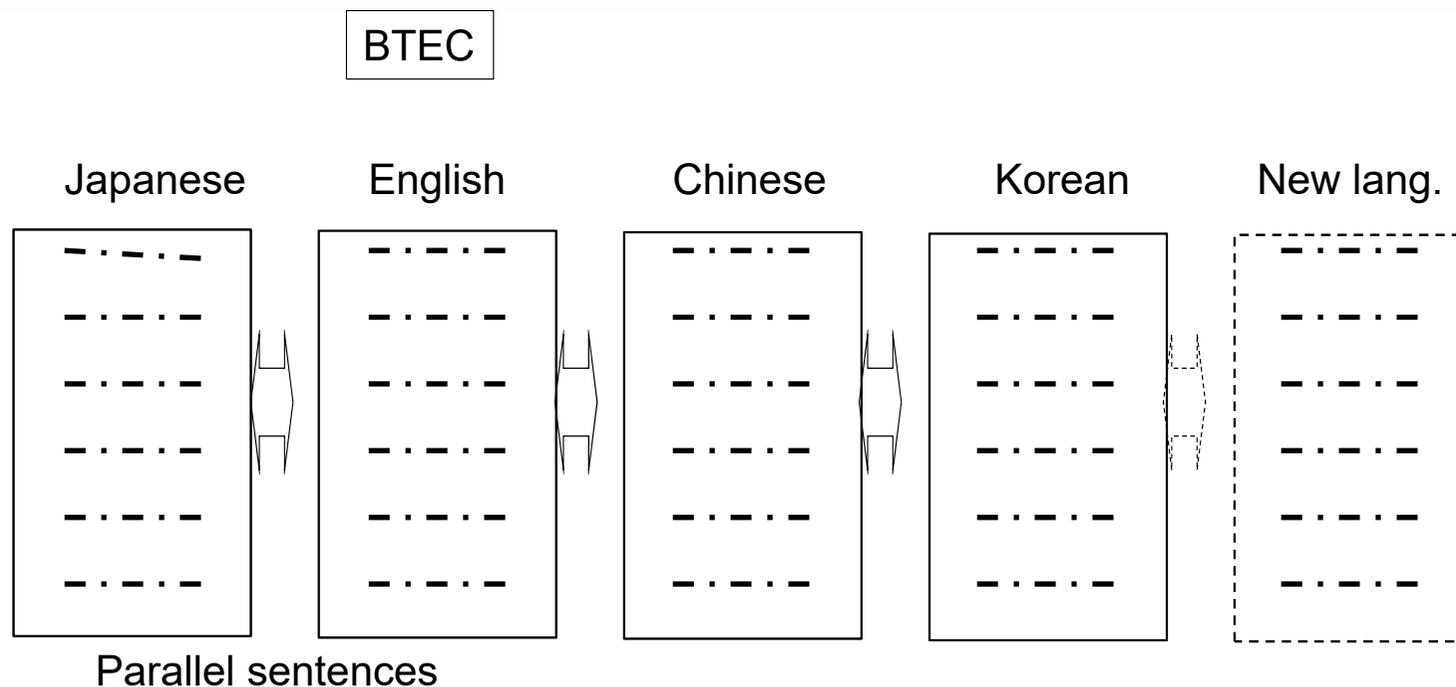
Japanese, English, Mandarin, Taiwanese Mandarin, German, French, Dutch, Danish, Italian, Spanish, Portuguese, Brazilian Portuguese, Russian, Arabic, Hindi, Indonesian, Malay, Thai, Tagalog, Vietnamese, Korean
 ※ Language in red can be input/output in voices.
 ※ There is no text input support for Hindi or Vietnamese.

A	Good
B	Fair
C	Acceptable
D	Nonsense
NIL	No Output

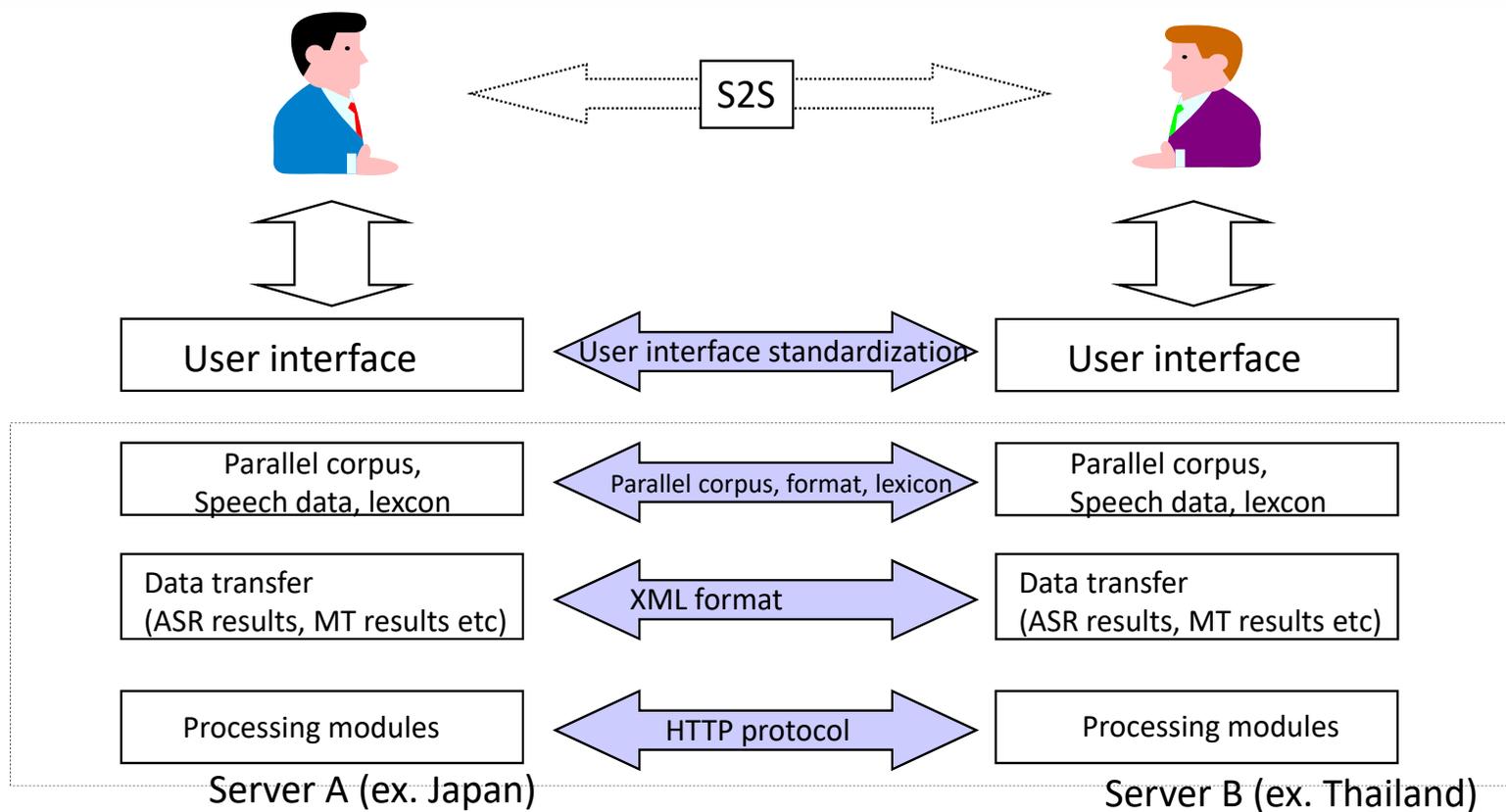


Subjective Evaluation % of ABC





Standardization Image

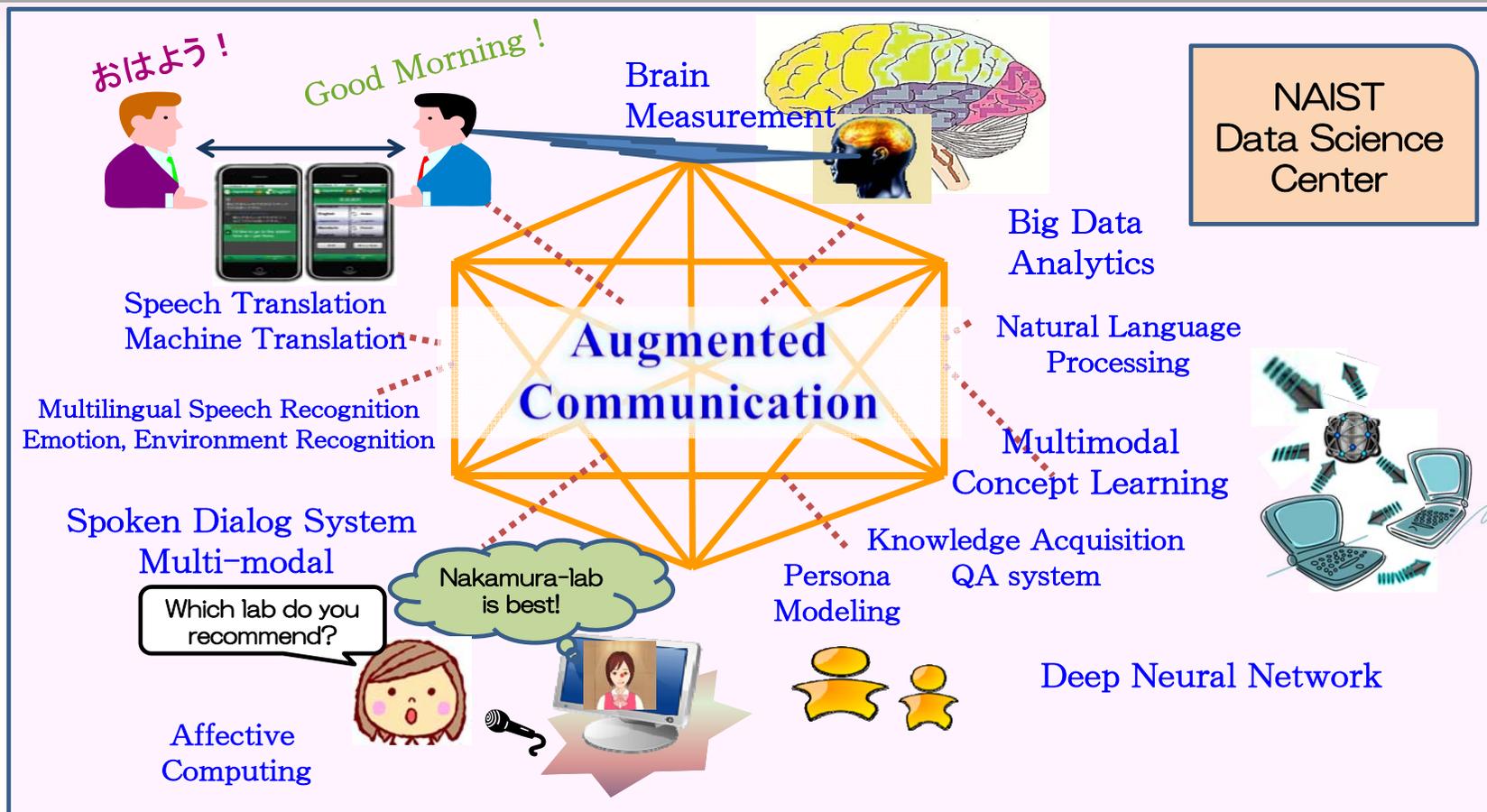


- ◆ Activity start for standardization of Network-based S2ST at ITU-T SG16
- ◆ Session period: October, 2009 to March, 2010
- ◆ NICT is the editor for S2ST standardization at ITU-T SG16, WP2 Q21/22

Document	Title	Scope
F.745	Functional Requirements for Network-based S2ST	- Definition of Network-based S2ST - Functions and service requirements of network-based S2ST
H.625	Architectural Requirements for Network-based S2ST	- Requirements of S2ST architecture - Definition of interface for Network-based S2ST

- ◆ Not only language conversion but also potentially added module like sign language are taken into account:
S2ST -> Modality conversion

Research Topics at NAIST



Integrating fundamental technologies into the augmented human-communication systems

Recent Progress of ASR after 2000

▶ Traditional Technologies

- Template Matching, Dynamic Programming [Sakoe 71]
- Hidden Markov Modeling, N-Gram Model [Mercer 83, etc]
- Neural Network, TDNN [Waibel 89], LSTM [Hochreiter 97]
- Weighted Finite State Transducer [Mohri 2006]
- Big Training Data, Data Collection through Trial Service

▶ Deep Learning (Hinton visited MSR)

- DNN-HMM [Hinton 2012]
 - Estimate State Posterior Probability by DNN
- Connectionist Temporal Classification [Graves 2013]
 - Predict Phoneme Label every frame
- Listen, Attend, and Spell [Chan 2016]
 - CTC + Attention: End-to-end modeling

Recent Speech Synthesis

▶ Traditional Technologies

- Formant-based Synthesis, Waveform Concatenation
- Statistical Speech Synthesis: HTS
 - Speech Synthesis by HMM
 - Tokuda, et al., “Speech parameter generation algorithms for HMM-based speech synthesis”, ICASSP 2000

▶ Deep Learning

- WaveNet
 - Waveform Convolution
 - van den Oord et al., “WAVENET: A GENERATIVE MODEL FOR RAW AUDIO”, arXiv:1609.03499v2 [cs.SD] 19 Sep 2016
- Tacotron
 - End-to-end speech synthesis with character input. Waveform generation by Griffin-Lim
 - Wang, et al., “TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS”, arXiv:1703.10135v2 [cs.CL] 6 Apr 2017
- Tacotron2:
 - Tacotron + WaveNet

Recent MT progress

▶ Traditional Technologies

- Rule-based MT :
Linguists generate translation rules
- Corpus-based MT:
 - Example-Based: Automatic rule extraction from corpus [M. Nagao84, Sato et.al.,89, Sumita et. al., 91]
 - Statistical MT: Statistical Modeling of MT. Extraction of model parameters from corpus and MT based on Noisy Channel Model [P. F. Brown, et.al. 93]
 - Phrase-base SMT
 - Tree-to-string
 - Statistical MT based on Tree Structure

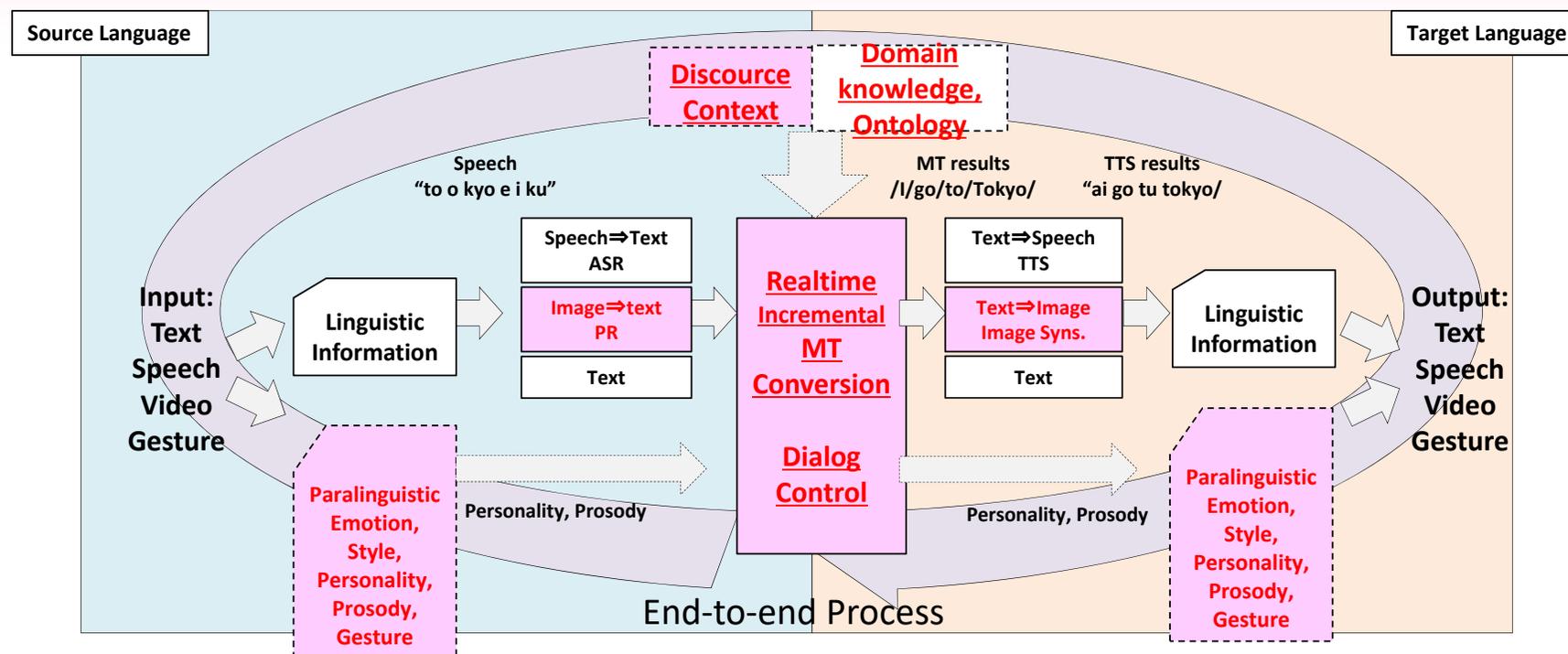
▶ Deep Learning

- Neural Machine Translation [2014]
 - Combination of Encoder and Decoder by LSTM
- Attention NMT [2015]
 - Add Attention to encoder and decoder
- Self Attention NMT [2017]
- Self attention by multiple heads. Transformer.

Contents

1. History of Automatic Speech Translation Research
2. Automatic Speech Interpretation Technologies
3. Speech Translation with Para-linguistic Information
4. Current Project and Data Collection
5. Summary and Future Works

Communication with Translation



Communication

- ① Simultaneity, Incremental, Latency,
- ② Para/non linguistic information

Human Interpreting [A.Mizuno 2016]

E-J Interpretation Example

(1) The relief workers (2) say (3) they don't have (4) enough food, water, shelter, and medical supplies (5) to deal with (6) the gigantic wave of refugees (7) who are ransacking the countryside (8) in search of the basics (9) to stay alive.

(1) 救援担当者は (9) 生きるための (8) 食料を求めて (7) 村を荒らし回っている (6) 大量の難民達の (5) 世話をするための (4) 十分な食料や水, 宿泊施設, 医療品が (3) ないと (2) 言っています.

(1) 救援担当者達の (2) 話では (4) 食料, 水, 宿泊施設, 医薬品が, (3) 足りず (6) 大量の難民達の (5) 世話が 出来ない のことです. (7) 難民達は 今村々を荒らし回って, (9) 生きるための (8) 食料を求めているのです.

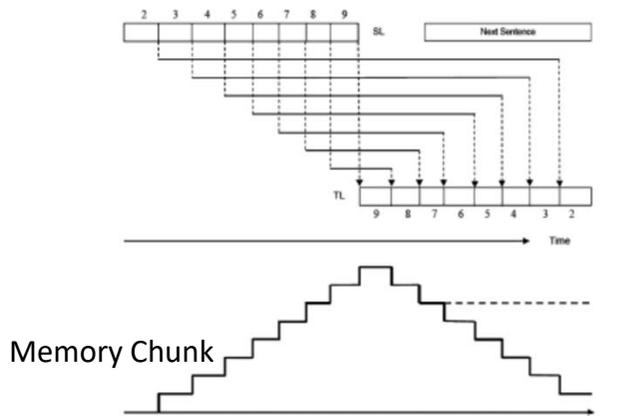
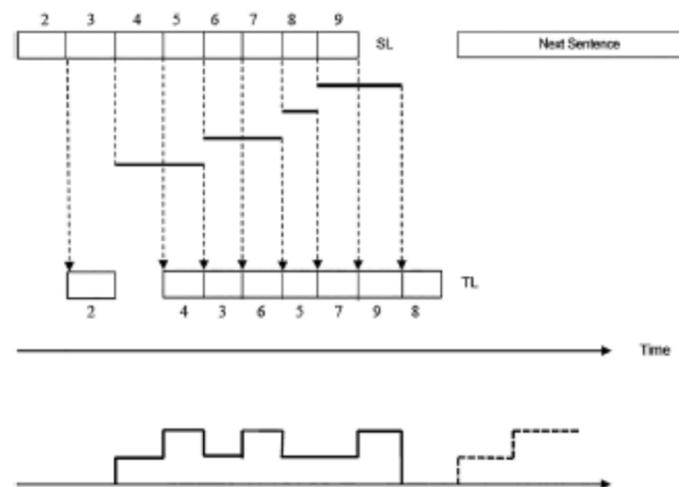


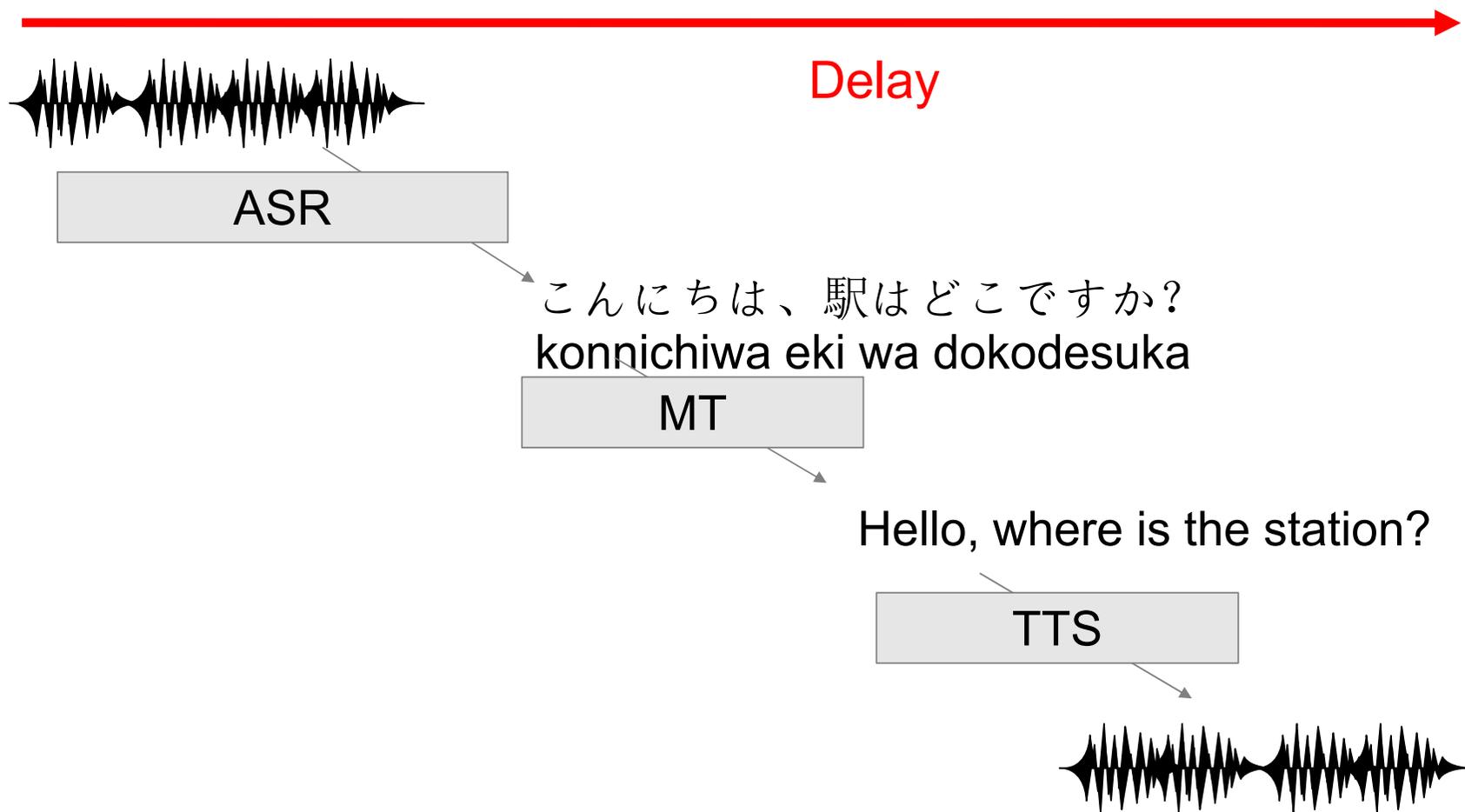
Fig.4 Translation to seek syntactic correspondence and its load
The dotted line of lower right indicates assumed load when next sentence comes in before the completion of translation of previous sentence.

Necessary #Chunk > 3 !

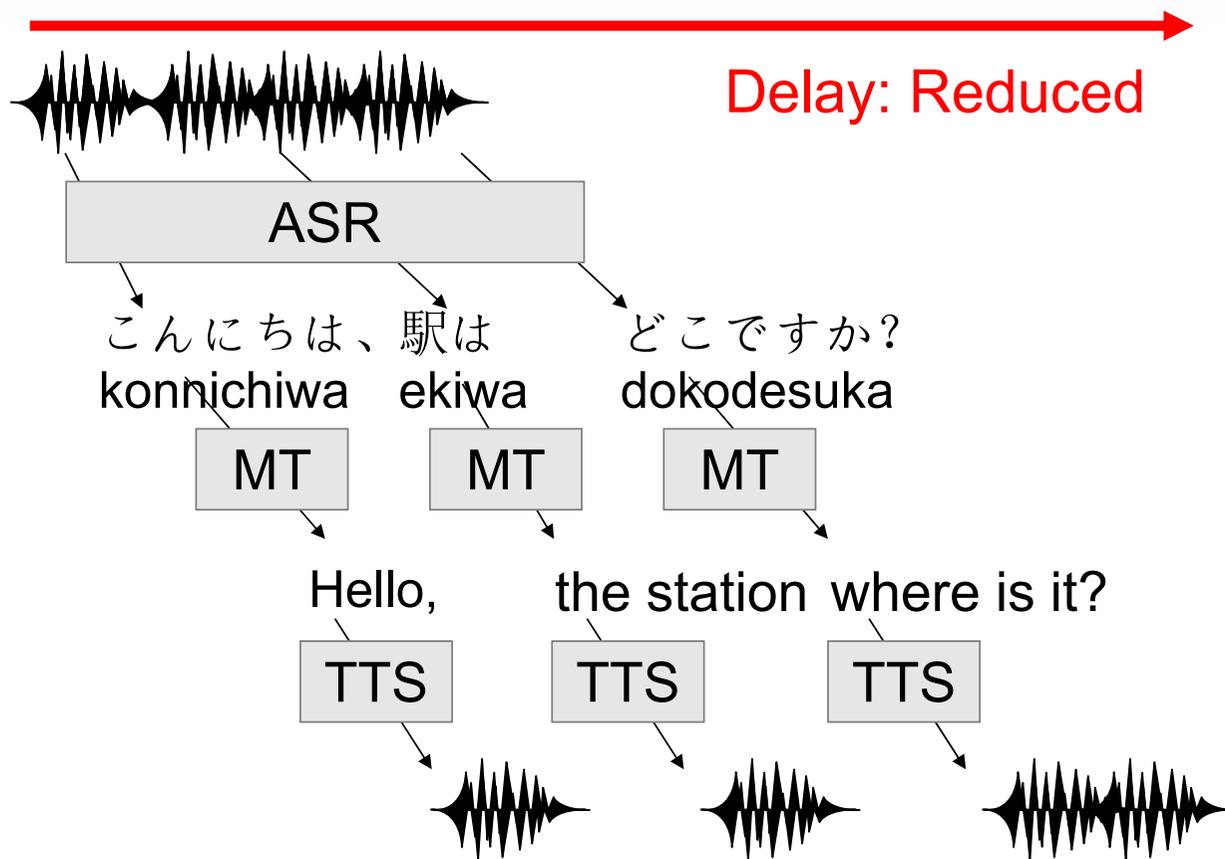


Necessary #Chunk < 3 !

Problem: Delay (Ear-Voice Span)



Simultaneous Incremental Speech Interpretation



But, this is not easy!

Four problems:

- **Segmentation:** When do we start interpretation?
- **Prediction:** Can we predict things that haven't been said?
- **Rewording:** Can we reword sentences to be conducive to simultaneous interpretation?
- **Evaluation:** How do we decide which results are better?

Re-ordering

- Crucial for translation accuracy:



Lexicalized Reordering Model

- Probabilistically models reordering for increased accuracy of translation
- Given current phrase and next phrase:

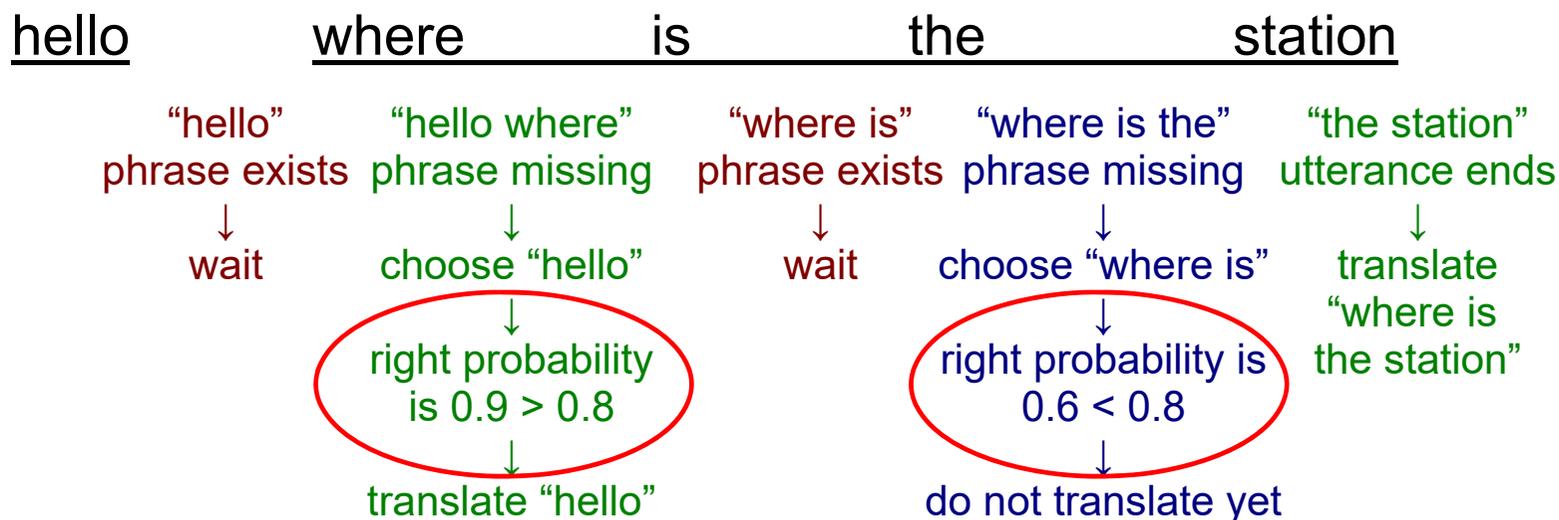


- “monotone” + “discontinuous right” = “right probability”

Adjusting Timing with Reordering Probabilities, 2012

- First, temporarily choose strings according to method one
- Next, if that phrase's **right probability** exceeds a threshold, actually translate the words in the cache

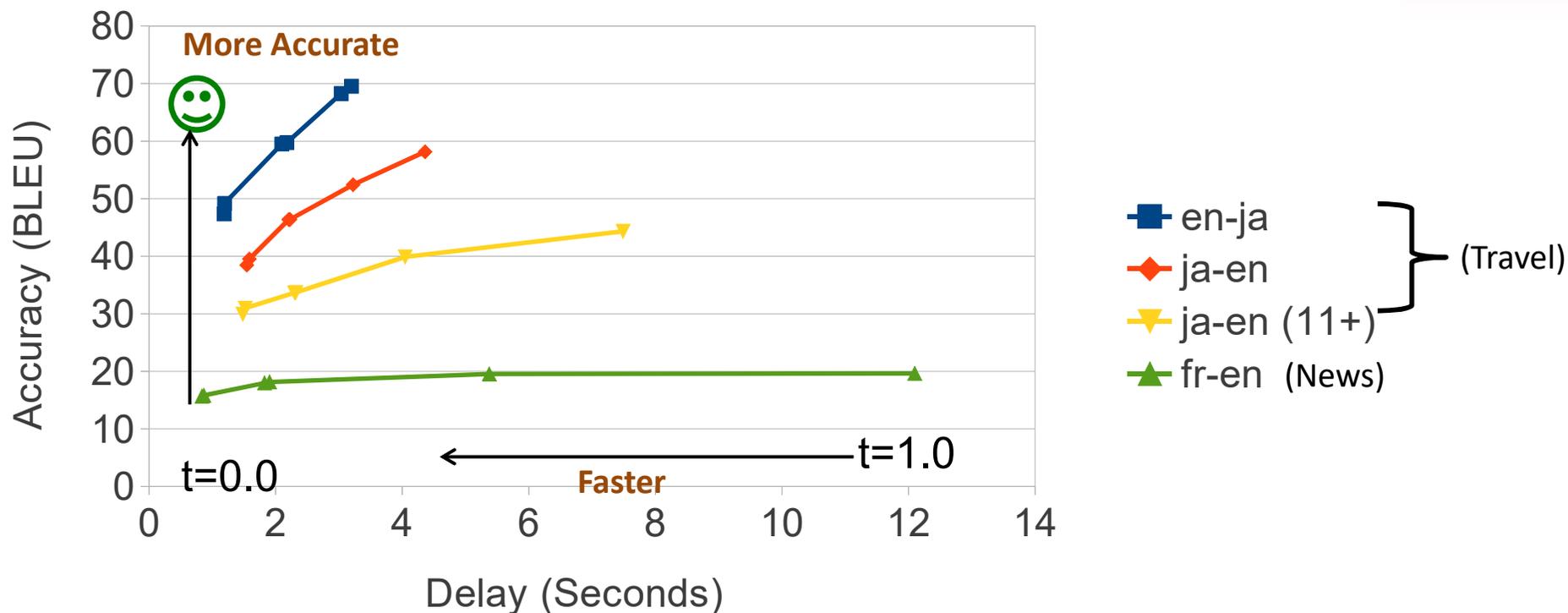
Example (threshold = 0.8):



Fujita, et. al., 2013

- Threshold 1.0 = traditional, 0.0 = method one

Comparison Across Settings



- Delay decreases in all settings
- Better delay/accuracy tradeoff for long sentences, similar languages

Experiments (IWSLT2013)

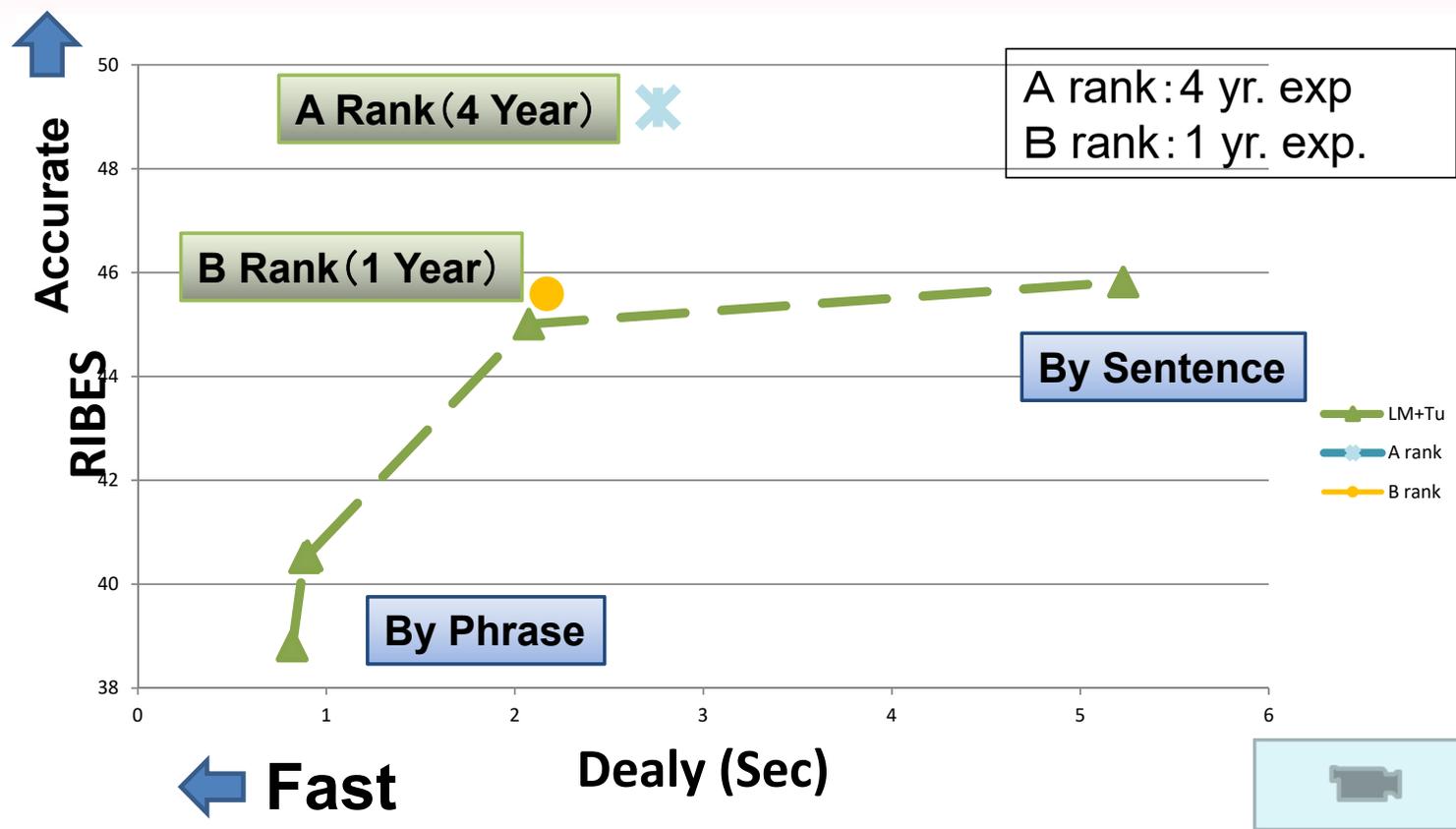
- ▶ Contents: TED Talk (English⇒Japanese)
 - Translation (Caption)
vs. Interpretation



- ▶ Human Interpreter
Three professionals with different skills

Skill Rank	# Years of Interpreter Experiences
S	15 years
A	4 years
B	1 year

SS2S vs. Human Interpreter Results on TED Talks

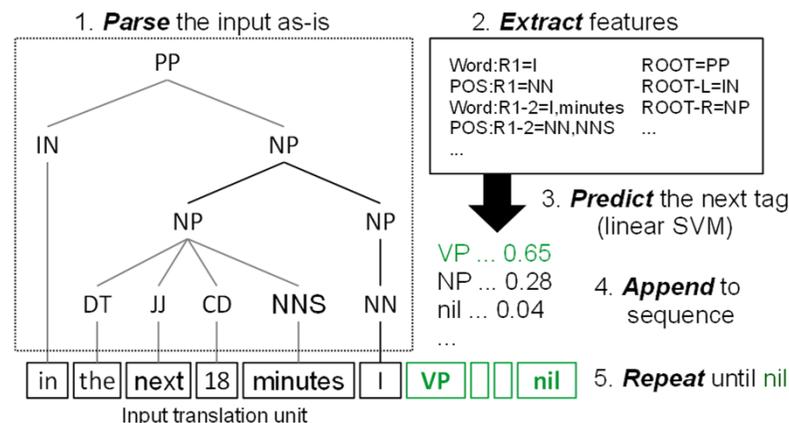


≡ B rank human interpreter with 1 year experience

Translation Timing Control by Syntactic Prediction, 2015

▶ Syntactic Prediction

- Incremental bottom up parsing
- Feature extraction and syntactic prediction



▶ Wait MT output when specific labels appear.

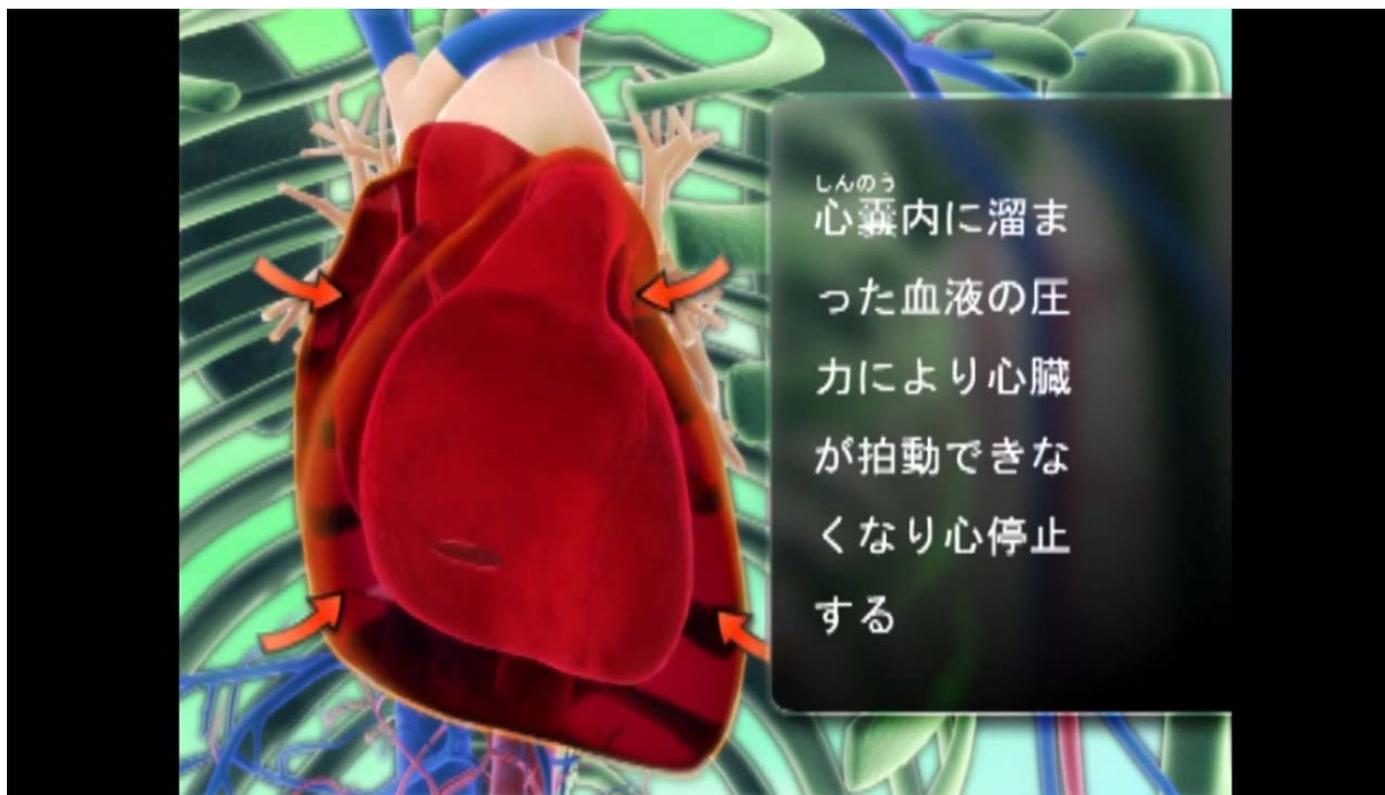
- Control MT output timing according reordering

Incremental parsing and syntactic prediction	in the next 18 minutes i 'm going to take [NP] (waiting) i 'm going to take you on a journey
MT results	18分である [NP] を行っています 皆さんを旅にお連れします

Oda, Yusuke *et al.*, Syntax-based Simultaneous Translation through Prediction of Unseen Syntactic Constituents, Proc. of ACL-IJCNLP 2015.

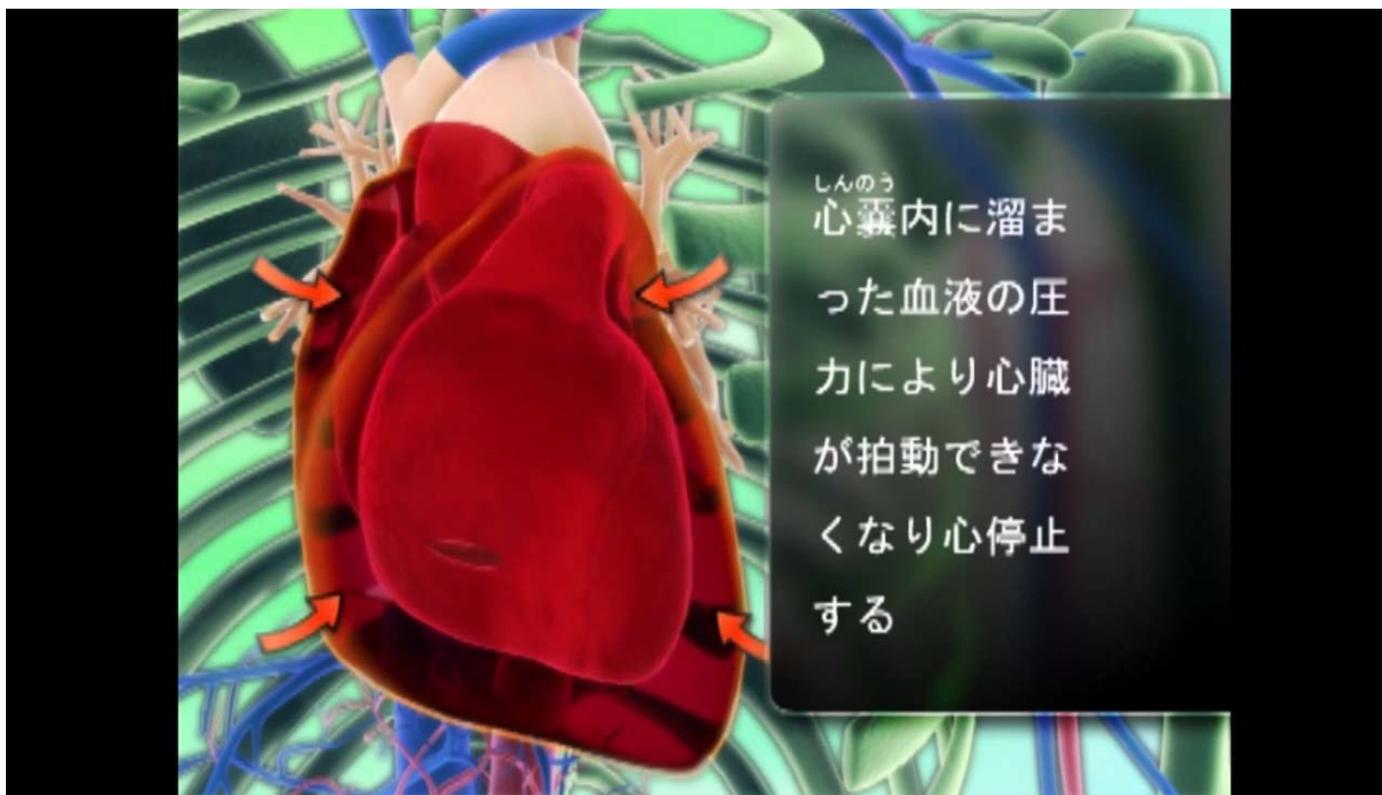
Sample 1 ,2015

- ▶ Conventional Automatic Speech Interpretation with Delay to Wait for Speech End (HirofumiSeo-trad.mp4)



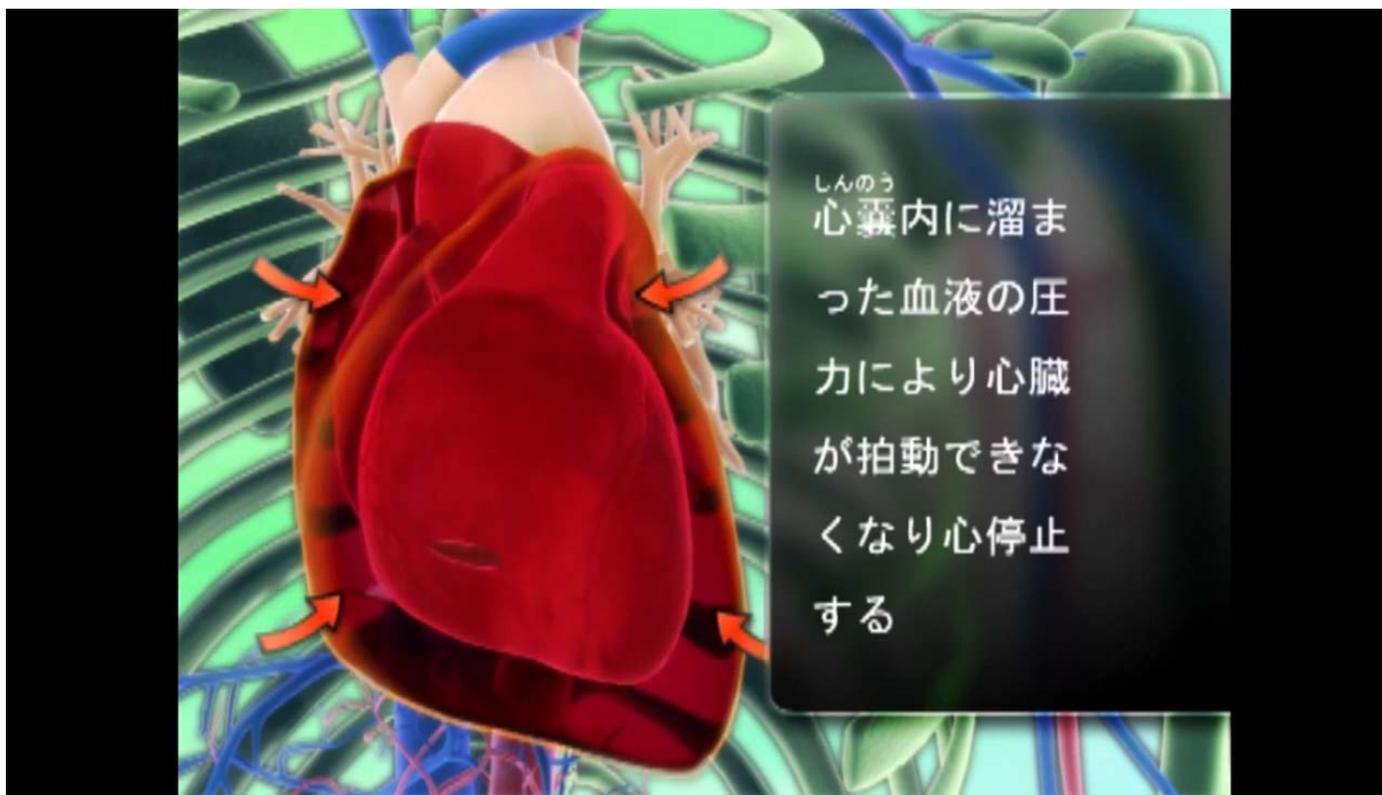
Sample 2 ,2015

- ▶ Actual Interpreter
(HirofumiSeo-interpreter.mp4)



Sample 3 ,2015

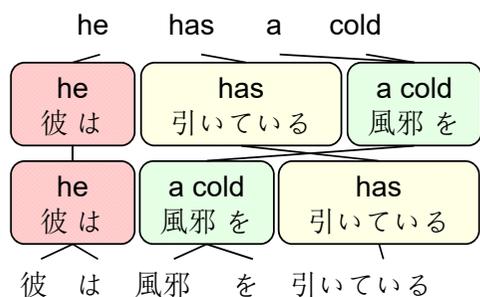
- ▶ Proposed Automatic Speech Interpretation (HirofumiSeo-simul.mp4:)



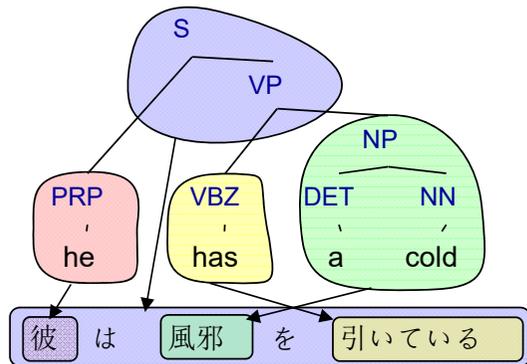
Statistical Translation Frameworks

Symbolic Models

Phrase-based MT [Koehn+ 03]

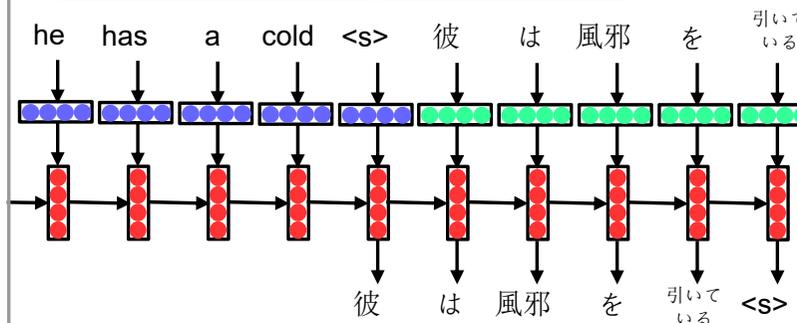


Tree-to-String MT [Liu+ 06]

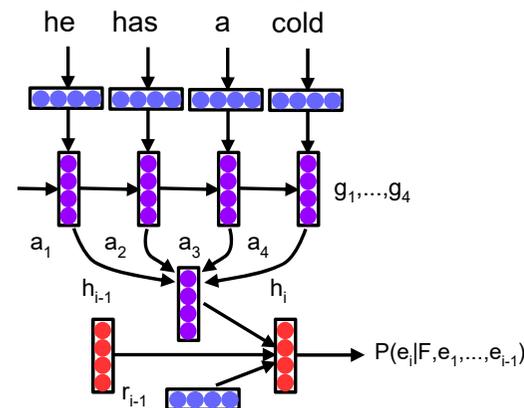


Continuous-space (Neural) Models

Encoder-Decoder [Sutskever+ 14]

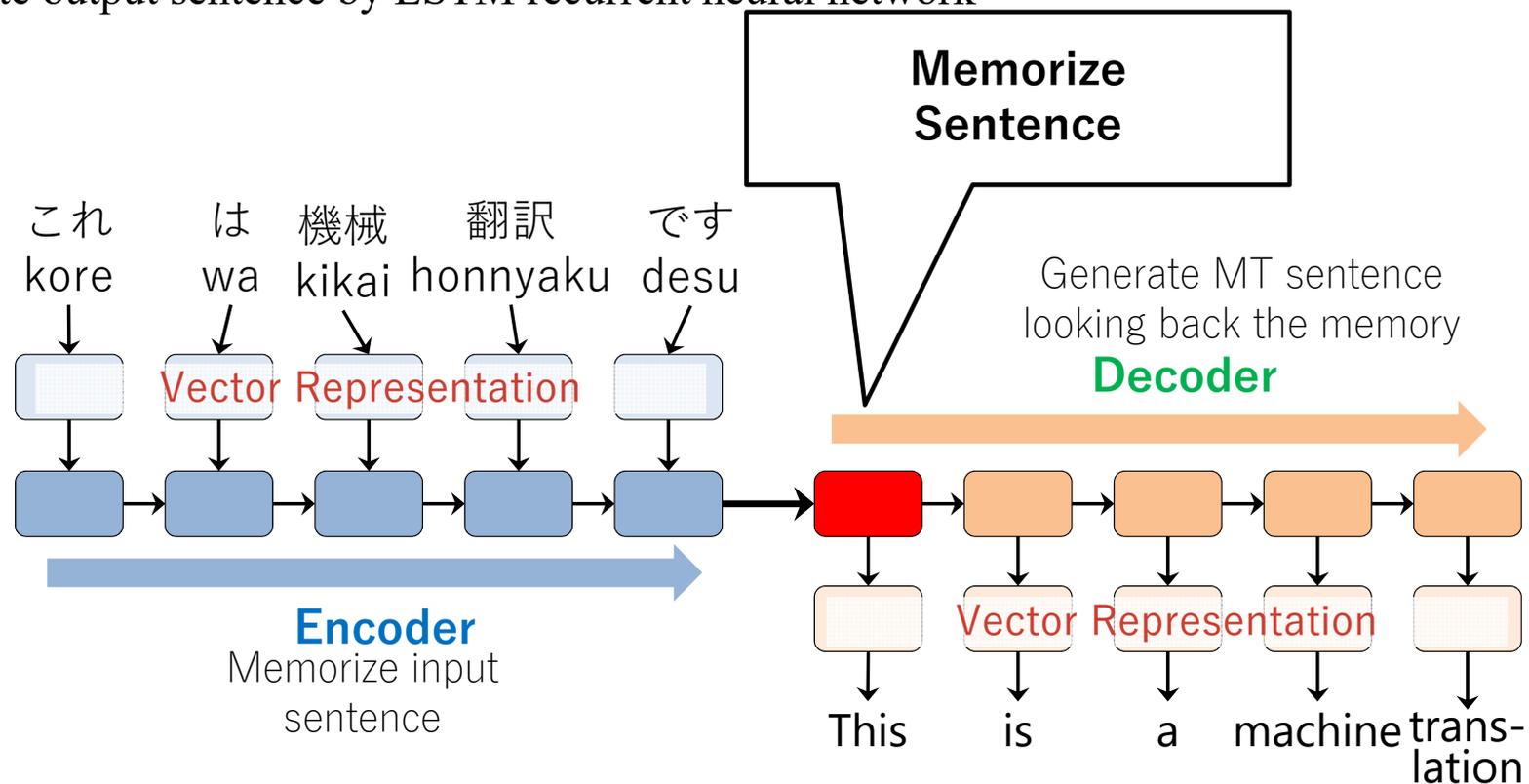


Attentional [Bahdanau+ 15]



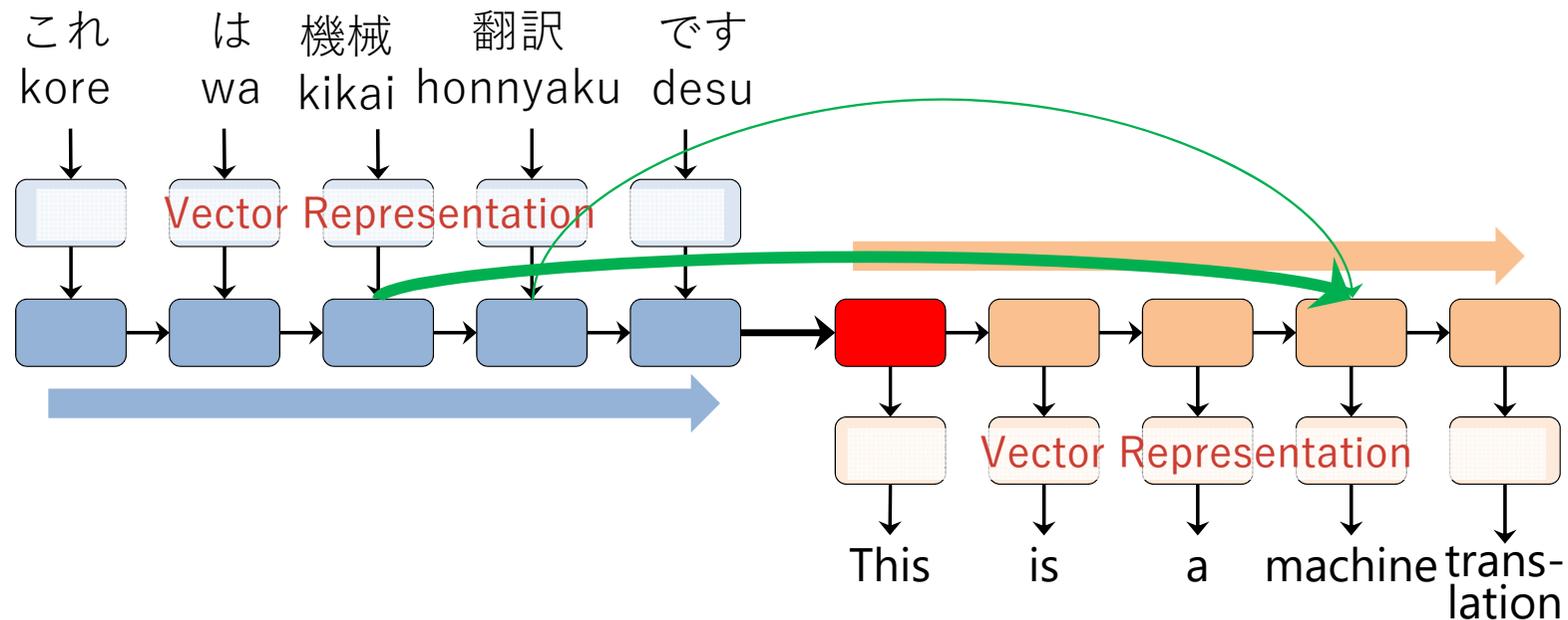
Encoder-decoder Model

- ▶ Memorize input sentence by LSTM recurrent neural network
- ▶ Generate output sentence by LSTM recurrent neural network

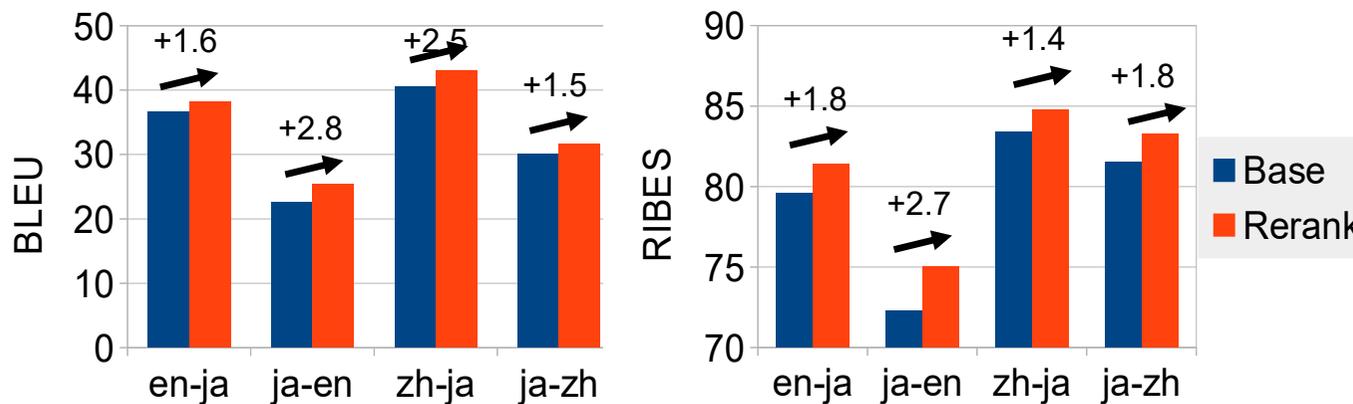


Attention Mechanism

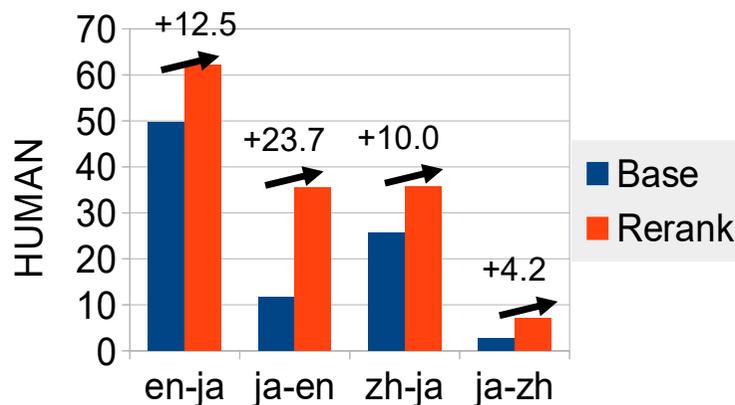
- ▶ Better Memorization of Sentence and Looking-back Mechanism
 - Weighted-sum by the attention



Results (Neubig, et.al, WAT2015)

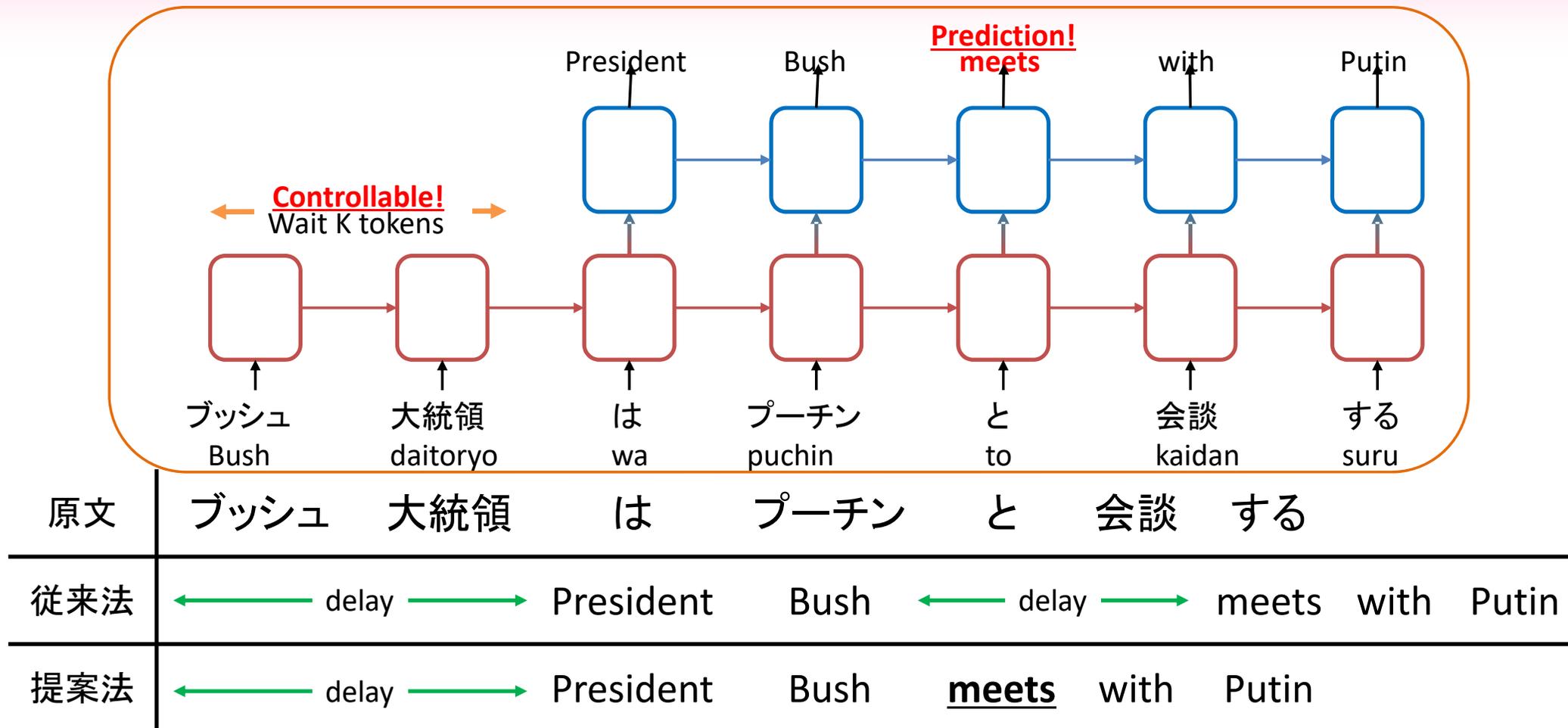


Confirm what we know: Neural reranking helps automatic evaluation.



Show what we didn't know: Also help manual evaluation.

Wait-k Algorithm



Mingbo Ma, et al., "STACL: Simultaneous Translation with Integrated Anticipation and Controllable Latency", arXiv:1810.08398v3 [cs.CL] 3 Nov 2018

Contents

1. History of Automatic Speech Translation Research
2. Automatic Speech Interpretation Technologies
3. Current Project and Data Collection
4. Summary and Future Works

JSPS Next Generation Speech Interpretation Research Project

▶ Objectives

- Incremental Automatic Speech Interpretation Algorithm
- Corpus Collection
- Evaluation Measure

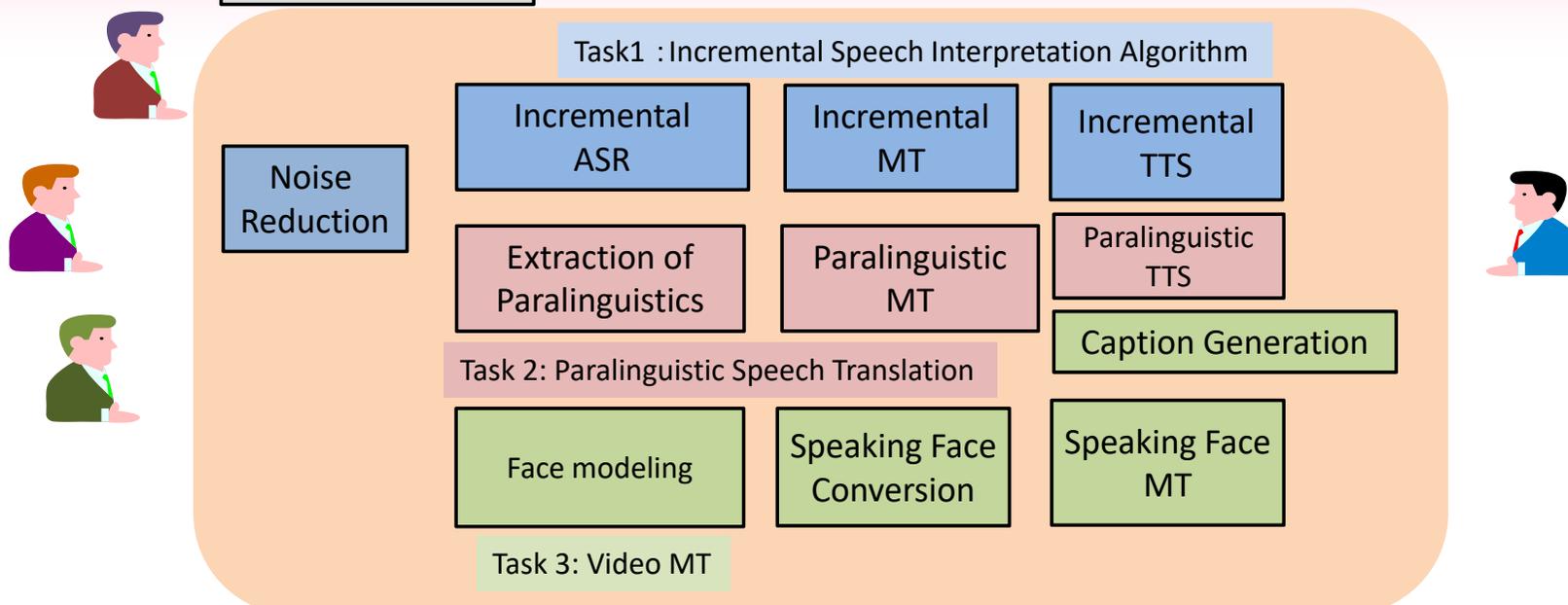
▶ Duration: 2017-2021, 5 years

▶ Member:

- Leader: Satoshi Nakamura (NAIST) Leader
- Acoustic Signal Processing: Hiroshi Saruwatari (U. Tokyo)
- Speech Recognition: Sakriani Sakti (NAIST), Tatsuya Kawahara (Kyoto U)
- Machine Translation: Katsuhito Sudo, Yuji Matsumoto (NAIST)
- Speech Synthesis: Tomoki Toda (Nagoya U), Shinnosuke Takamichi (U. Tokyo), Sakriani Sakti (NAIST)
- Audio-visual Translation: Shigeo Morishima (Waseda U)
- Cognitive Load Measurement: Hiroki Tanaka (NAIST)
- Corpus Collection: Katsuhito Sudo, Manami Matsuda (NAIST)

Project Overview

Noise, Reverberation



Task 4: Real Time Cognitive Load Measurement by Human Sensing

2x 32ch EEG, Gaze, Heart rate

Task 5: Corpus Collection and Prototyping

Collect 400 hours Data of Japanese and English Speech Interpretation

Building Prototype of the Incremental Speech Interpretation System

NAIST Interpreter Corpus

▶ 2012-2016

- Source speech: MP4 (TED), MP3 (CNN), PCM
- Interpreter speech: 24bit 48kHz PCM
 - Skill : S (10 years+), A(3 years+), B
 - Some data includes speech of multiple interpreters

Translation direction	Domain	Source Speech		Interpreter Speech	
		#files	#hours	#files	#hours
E->J	TED	74	15.2	58	12.3
	CNN	13	0.731	7	0.389
	Total	87	15.9	65	12.7
J->E	TED	60	11.9	60	11.9
	CSJ	31	5.51	31	5.51
	NHK	10	0.304	10	0.304
	Total	101	17.7	101	17.7

NAIST Interpreter Corpus 2018

▶ As of 2018

- Source speech: MP4 (TED, TEDx), PCM (CSJ)
- Interpreter speech: 16bit 16kHz PCM
 - Skill : S (10 years +), A (3 years +), B
 - For training set. Total 100 hours by the rank A interpreters
 - For test set. Total 24 hours by one from all rank interpreters

Translation direction	domain	Source speech		Interpreter speech	
		#files	#hours	#files	#hours
E->J	TED	302	66.8	302	66.8
	TED (test)	16	4	16	4
	total	318	70.8	318	70.8
J->E	CSJ	146	33	146	33
	TEDx (test)	19	4	19	4
	total	165	37	165	37

Book (Japanese version)



Contents

1. History of Automatic Speech Translation Research
2. Automatic Speech Interpretation Technologies
3. Current Project and Data Collection
4. Summary and Future Works

Summary

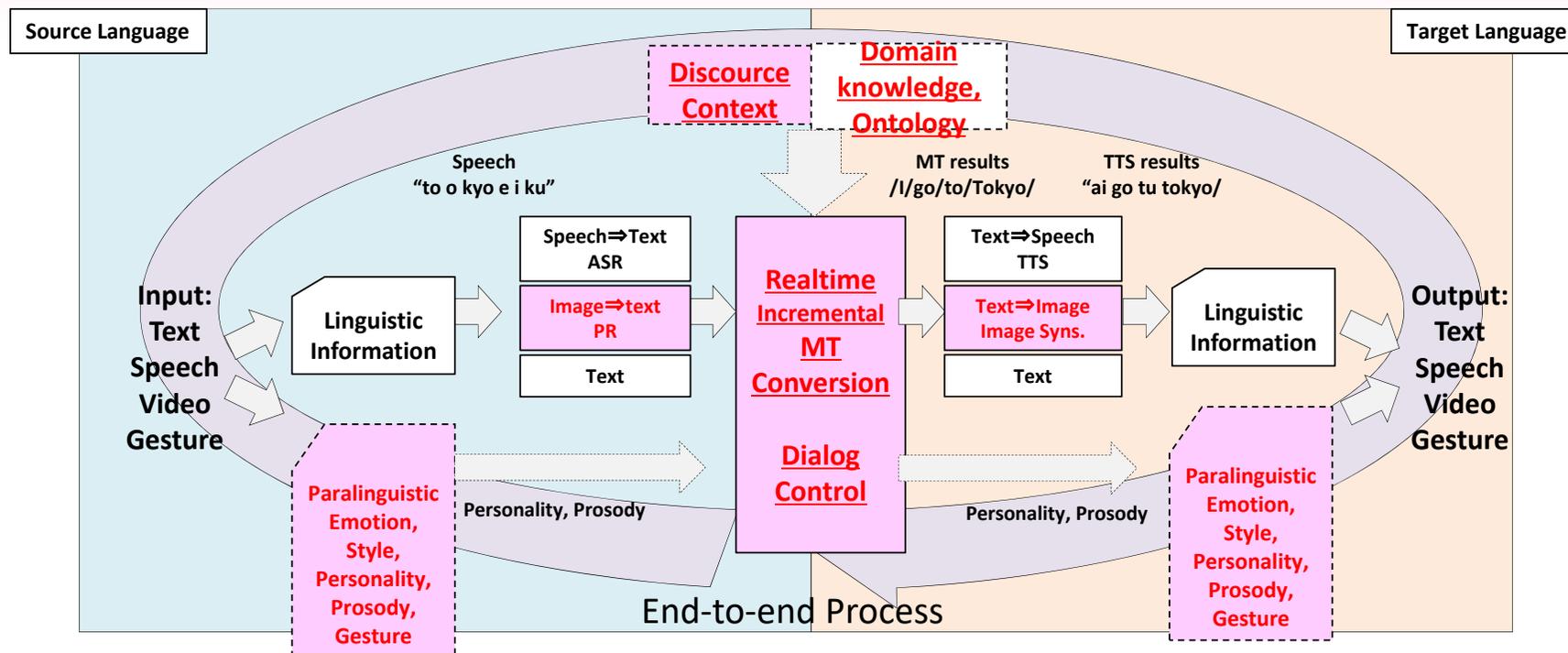
- ▶ Remarkable progress
 - By Statistical Machine Translation
 - Deep Neural Network
 - Progress in Speech Translation

- ▶ Automatic Speech Interpretation
 - Data Collection
 - Develop Algorithms both for Automatic Speech Interpretation and Interpreter Support System

- ▶ Further Research
 - Para-linguistics/ Multi-modal
 - Context/ Situation Dependency
 - Common Sense and Domain Knowledge
 - Semantics, Discourse Analysis
 - Towards Better Communication



Communication with Translation



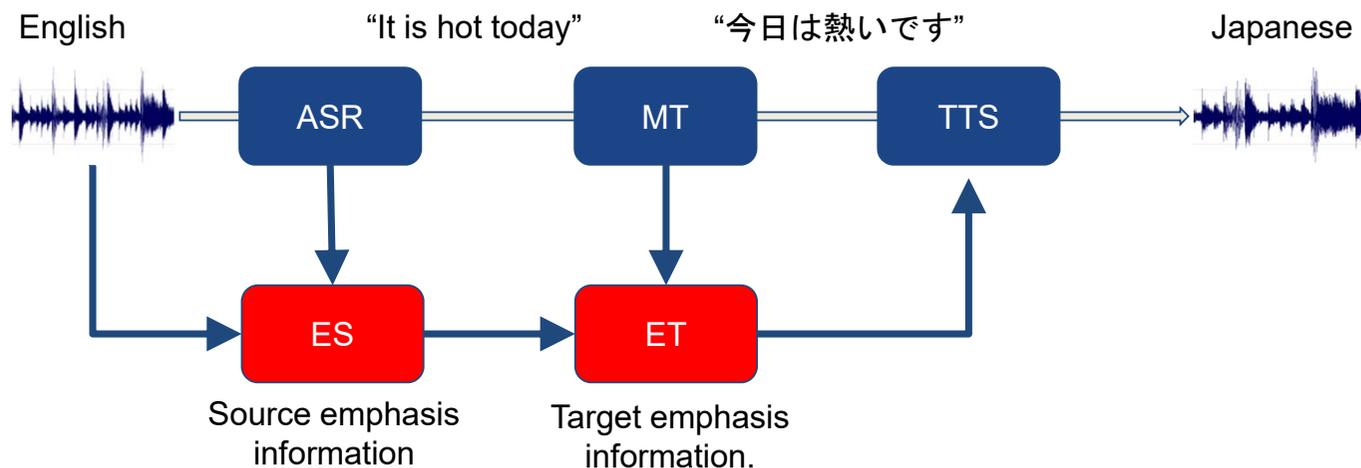
Communication

- ① Simultaneity, Incremental, Latency,
- ② Para/non linguistic information

Research Focus Up to Now

▶ Emphases Speech Translation

→ Translates speech while preserving emphasis information



(1) Emphasis estimation (ES) systems:

Estimate emphasis information given speech & a corresponding word sequence

(2) Emphasis translation (ET) systems:

Translate estimated emphasis information into another language

Speech Translation Samples

▶ English-Japanese Emphases Translation

