Research Activities for Translating *Asian* Languages

Eiichiro SUMITA

Associate Director General, Universal Communication Research Institute,
NICT, Japan

Coling 2016

PERIOD: **11-16 Dec. 2016, OSAKA, JAPAN**
VENUE: OSAKA INTERNATIONAL CONVENTION CENTER (OICC)
GENERAL CHAIR: **Dr. Nicoletta Calzolari** (ILC CNR)

Many hotels, including one high-quality hotel connected to the OICC venue.

Japanese food:

- 12 restaurants with **three stars**
- 52 with **two stars**
- 213 with **one star**

Tons of places to visit:

- Traditional puppetry
- Tunnel to Shrine
- Todaiji Temple
- Pop culture

Some important info

Homepage:

http://coling2016.anlp.jp/

Deadlines:

15 August 2016: Paper submission

30 September 2016: Notification

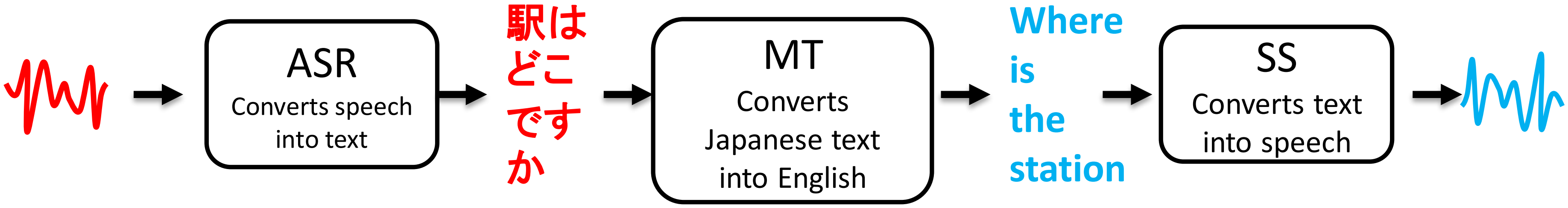
15 October 2016: Camera-ready versions due

Outline

1. **World**wide speech translation consortium, Universal Speech Translation Advanced Research (U-STAR)
2. Workshop on **Asian** Translation (WAT) & **Asian** Language Treebank (ALT)
3. Global Communication Program (GCP) in **Japan**
4. Recent research topics in the National Institute of Information Technology (**NICT), Japan**
5. Towards increasing **collaboration**

1. U-STAR (**U**niversal **S**peech **T**ranslation **A**dvanced **R**esearch) Consortium

Configuration of Speech Translation



ASR

Converts speech into text

駅はどこですか

MT

Converts Japanese text into English

Where is the station

SS

Converts text into speech

The world's first experiment:

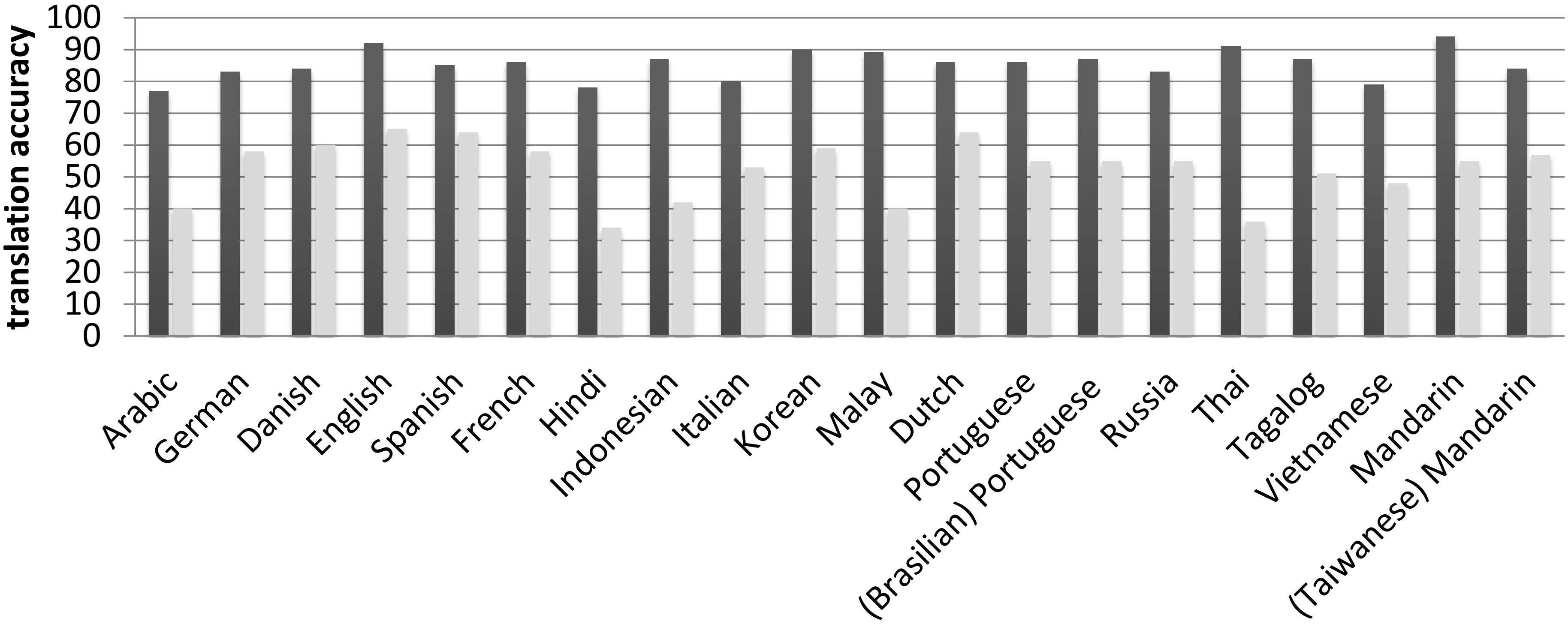
1992

Speech Translation System on a Smartphone
VoiceTra @ 2010.7

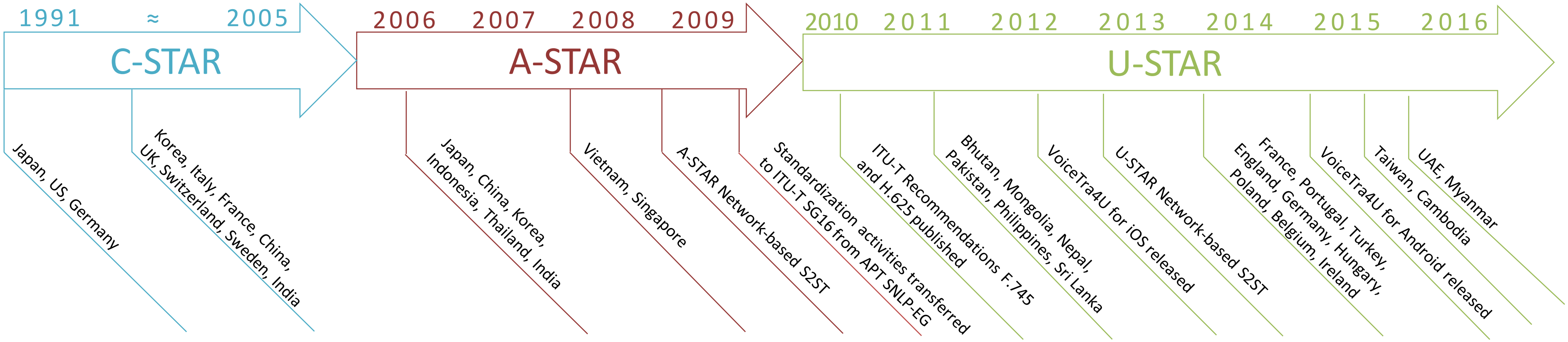
Part 1: How to use

Part 2: Don’ts

VoiceTra for the tourism domain demonstrates high-quality translation among **many languages**



Black: VoiceTra by NICT Gray: Other famous translator



1991 ≈ 2005

C-STAR

Japan, US, Germany

Korea, Italy, France, China,
UK, Switzerland, Sweden, India

2006 2007 2008 2009

A-STAR

Japan, China, Korea,
Indonesia, Thailand, India

Vietnam, Singapore

A-STAR Network-based S2ST

Standardization activities transferred
to ITU-T SG16 from APT SNLP-EG

2010 2011 2012 2013 2014 2015 2016

U-STAR

ITU-T Recommendations F.745
and H.625 published

Bhutan, Mongolia, Nepal,
Pakistan, Philippines, Sri Lanka

VoiceTra4U for iOS released

U-STAR Network-based S2ST

France, Portugal, Turkey,
England, Germany, Hungary,
Poland, Belgium, Ireland

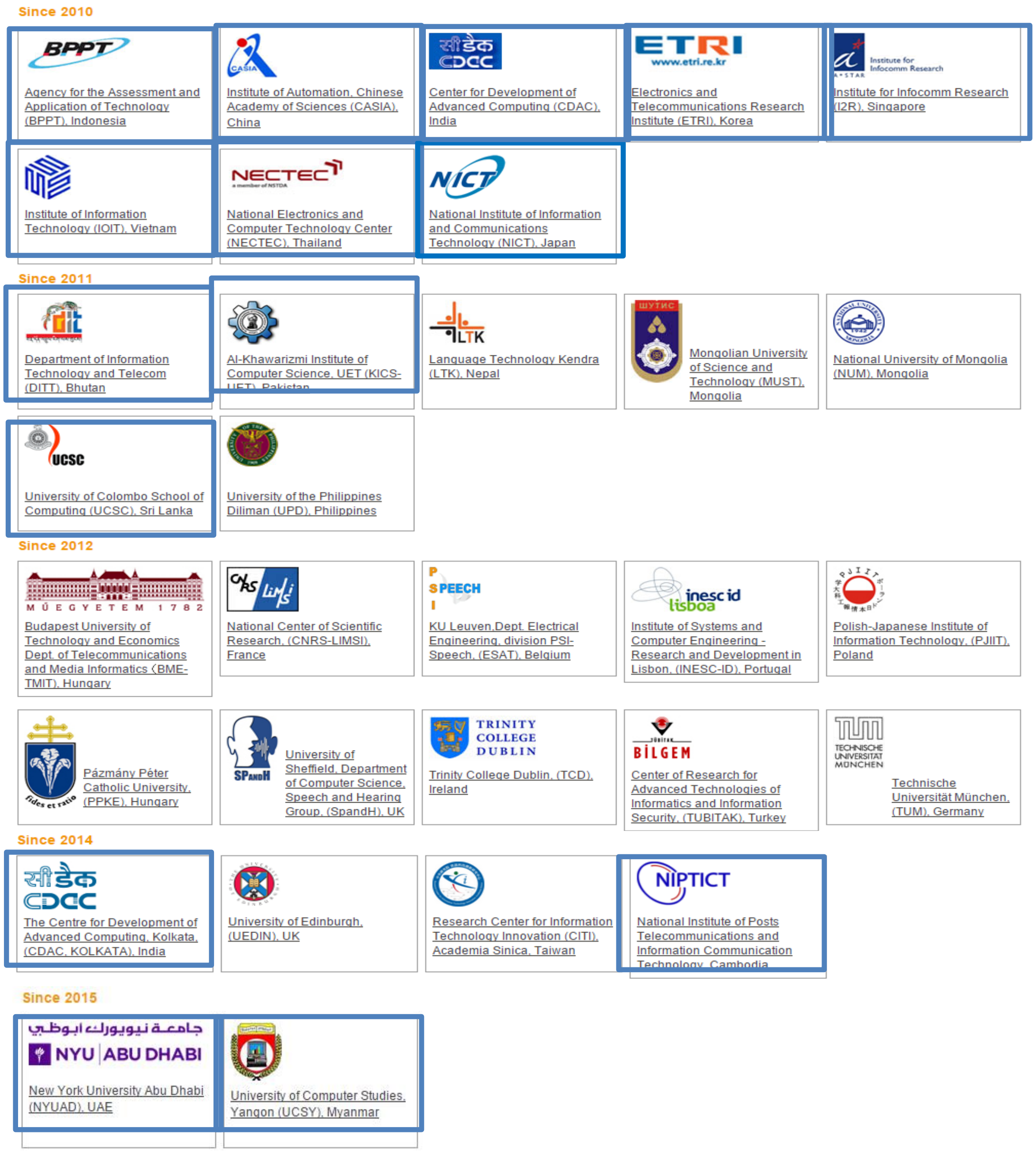
VoiceTra4U for Android released

Taiwan, Cambodia

UAE, Myanmar

1. **C-STAR started out over 24 years ago with 3 organizations.**
2. **Then, A-STAR extended the activity in Asia.**
3. **Now, with 32 institutes from 27 countries/regions, U-STAR has grown into one of the largest consortia in the world that conducts research on speech-to-speech translation.**

32 institutes



Since 2010

- Agency for the Assessment and Application of Technology (BPPT), Indonesia
- Institute of Automation, Chinese Academy of Sciences (CASIA), China
- Center for Development of Advanced Computing (CDAC), India
- Electronics and Telecommunications Research Institute (ETRI), Korea
- Institute for Infocomm Research (I2R), Singapore
- Institute of Information Technology (IOIT), Vietnam
- National Electronics and Computer Technology Center (NECTEC), Thailand
- National Institute of Information and Communications Technology (NICT), Japan

Since 2011

- Department of Information Technology and Telecom (DITT), Bhutan
- Al-Khawarizmi Institute of Computer Science, UET (KICS-UET), Pakistan
- Language Technology Kendra (LTK), Nepal
- Mongolian University of Science and Technology (MUST), Mongolia
- National University of Mongolia (NUM), Mongolia
- University of Colombo School of Computing (UCSC), Sri Lanka
- University of the Philippines Diliman (UPD), Philippines

Since 2012

- Budapest University of Technology and Economics Dept. of Telecommunications and Media Informatics (BME-TMIT), Hungary
- National Center of Scientific Research, (CNRS-LIMSI), France
- KU Leuven,Dept. Electrical Engineering, division PSI-Speech, (ESAT), Belgium
- Institute of Systems and Computer Engineering - Research and Development in Lisbon, (INESC-ID), Portugal
- Polish-Japanese Institute of Information Technology, (PJIIT), Poland
- Pázmány Péter Catholic University, (PPKE), Hungary
- University of Sheffield, Department of Computer Science, Speech and Hearing Group, (SpandH), UK
- Trinity College Dublin, (TCD), Ireland
- Center of Research for Advanced Technologies of Informatics and Information Security, (TUBITAK), Turkey
- Technische Universität München, (TUM), Germany

Since 2014

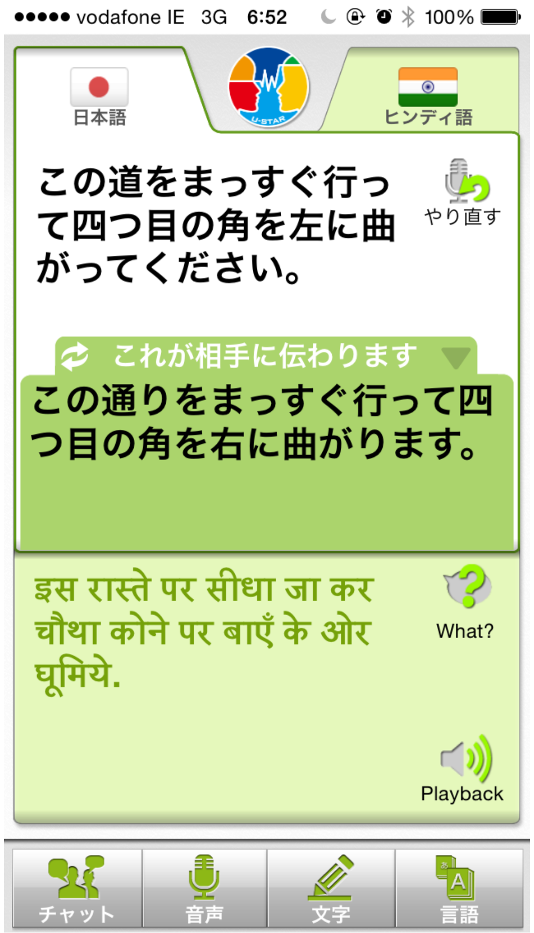
- The Centre for Development of Advanced Computing, Kolkata, (CDAC, KOLKATA), India
- University of Edinburgh, (UEDIN), UK
- Research Center for Information Technology Innovation (CITI), Academia Sinica, Taiwan
- National Institute of Posts Telecommunications and Information Communication Technology, Cambodia

Since 2015

- New York University Abu Dhabi (NYUAD), UAE
- University of Computer Studies, Yangon (UCSY), Myanmar

✓Smartphone applications for iOS and Android

✓Translates up to 30 languages including dialects



この道をまっすぐ行って四つ目の角を左に曲がってください。

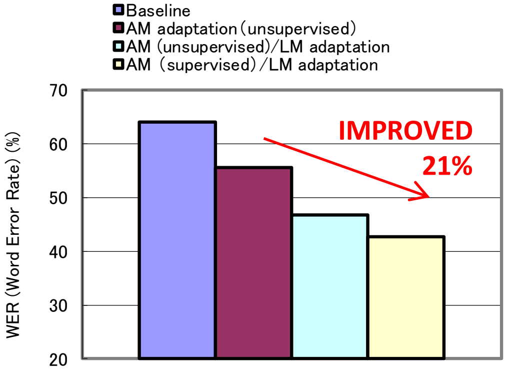
これが相手に伝わります

この通りをまっすぐ行って四つ目の角を右に曲がります。

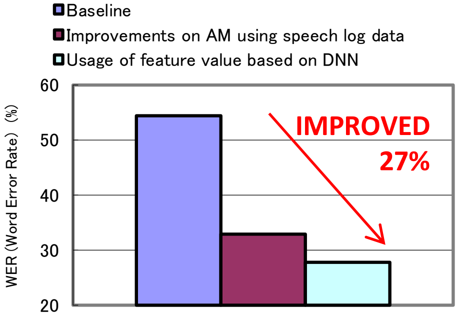
इस रास्ते पर सीधा जा कर चौथा कोने पर बाएँ के ओर घूमिये.

Speech Input (Available in 17 languages)	Text Input / Output (Available in 30 languages)		Speech Output (Available in 14 languages)
Dutch	Arabic	Malay	English (US)
English (US)	Br-Portuguese	Mandarin	Hindi
English (UK)	Danish	Mongolian	Hungarian
French	Dutch	Nepali	Indonesian
German	Dzongkha	Polish	Japanese
Hindi	English (US)	Portuguese	Korean
Hungarian	English (UK)	Russian	Malay
Indonesian	Filipino	Sinhala	Mandarin
Japanese	French	Spanish	Mongolian
Korean	German	Tw-Mandarin	Polish
Mandarin	Hindi	Thai	Portuguese
Malay	Hungarian	Turkish	Thai
Polish	Indonesian	Urdu	Turkish
Portuguese	Italian	Vietnamese	Vietnamese
Thai	Japanese		
Turkish	Korean		
Vietnamese			

Research collaborations among U-STAR Members



Improved about 20% in WER (Thai)
（Evaluated with 2,765 sentences of VoiceTra4U log）



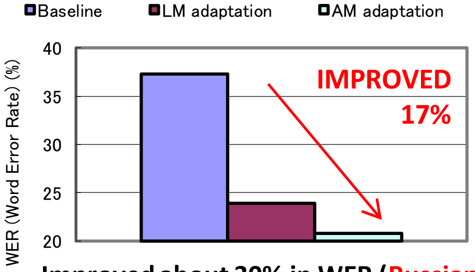
Improved about 30% in WER (Vietnamese)
（Evaluated with 803 sentences of VoiceTra4U log）

3 months from July 2013: 3 researchers from NECTEC (Thailand) visited NICT to work on Thai ASR.

1 month from August 2014 : one research student from UULM (Germany) visited NICT to work on Russian ASR.

2 months from September 2014 : one research student from IOIT (Vietnam) visited NICT to work on Vietnamese ASR

1 year from August 2014 : one researcher and two students from UCSY (Myanmar) visited NICT to work on Myanmar ASR



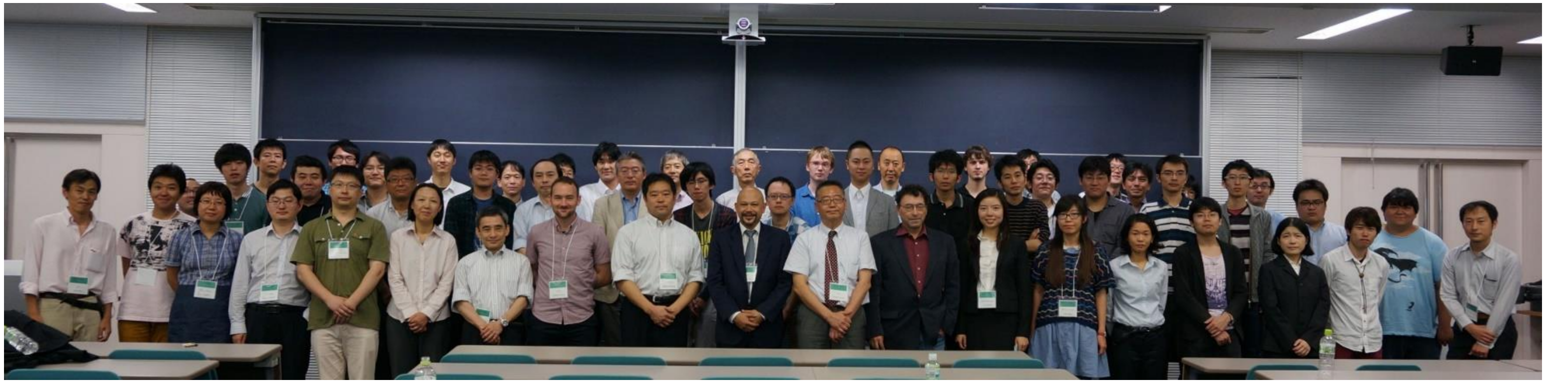
Improved about 30% in WER (Russian)
(Evaluated with read speech from 1 speaker)

Please join us @ http://www.ustar-consortium.com/contactinfo.html

2.1 Workshop on Asian Translation (WAT)

http://orchid.kuee.kyoto-u.ac.jp/WAT/

1st WAT (Workshop on Asian translation)



- http://orchid.kuee.kyoto-u.ac.jp/WAT/WAT2014/index.html
- Held in **Tokyo in October 2014**
- Had over **50 participants**
- Machine Translation evaluation campaign focusing on **scientific documents in Japanese-English/Japanese-Chinese**

2nd WAT

- http://orchid.kuee.kyoto-u.ac.jp/WAT/
- Will be held in Kyoto on 16 October 2015
- New Task: **translation of patent document** with bilingual Japanese-Chinese/Korean corpora (1 M) provided by Japan Patent Office (JPO)

3rd WAT

Please join us.

The third one will add new language pairs.

2.2 Asian Language Treebank (ALT)

1. Treebank accelerates research on NLP for each language
    - No publicly available POS-tagged and constituency tree corpora for many Asian languages.

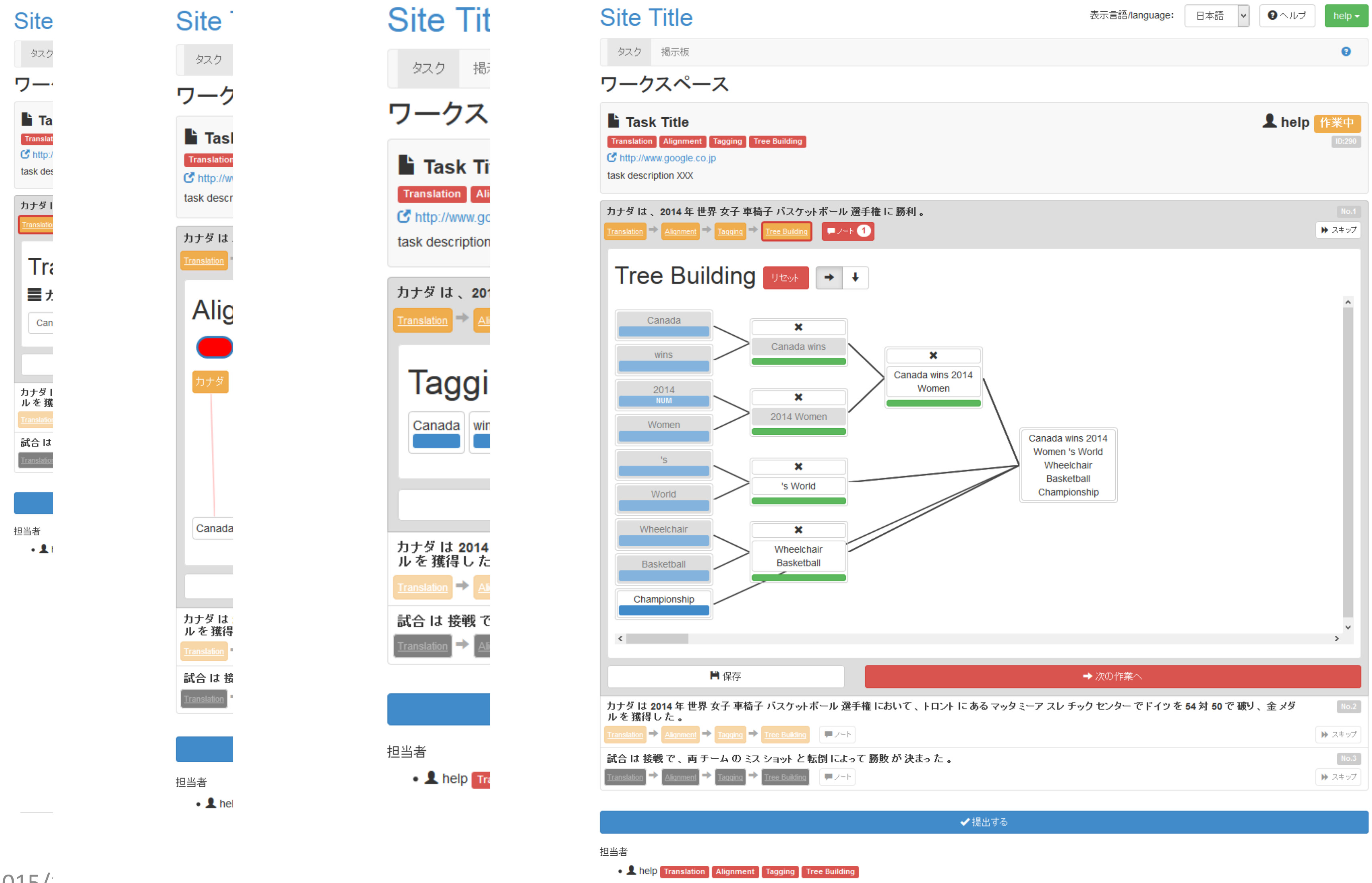
2. Parallel corpus accelerates research on Machine Translation between the languages
    - No big parallel corpora among many Asian languages

In order to solve these problems, let’s develop **treebanks** of **parallel corpora** together!

4 steps for developing treebanks

1. **Translating** the common English sentence into the Asian language
2. **Aligning words** between the sentence pairs of English and the Asian language
3. **Tagging** the **POS (part of speech)** for the sentence in the Asian language
4. **Building** the **tree** of the sentence in the Asian language

WEB-based UI for ALT's 4 steps



Site Title

表示言語/language: 日本語

ヘルプ

help

タスク 掲示板

ワークスペース

Task Title

help 作業中

Translation Alignment Tagging Tree Building

ID:290

http://www.google.co.jp

task description XXX

カナダ は 、 2014 年 世界 女子 車椅子 バスケットボール 選手権 に 勝利 。

No.1

Translation → Alignment → Tagging → Tree Building

ノート 1

スキップ

Tree Building

リセット

保存

次の作業へ

カナダ は 2014 年 世界 女子 車椅子 バスケットボール 選手権 において 、 トロント に ある マッタミーア スレ チック センター で ドイツ を 54 対 50 で 破り 、 金 メダル を 獲得 した 。

No.2

Translation → Alignment → Tagging → Tree Building

ノート

スキップ

試合 は 接戦 で 、 両 チーム の ミス ショット と 転倒 によって 勝敗 が 決まっ た 。

No.3

Translation → Alignment → Tagging → Tree Building

ノート

スキップ

提出する

担当者

- help Translation Alignment Tagging Tree Building

Progress so far

Language	Translate	Alignment	POS	Tree	completion
English	(-)	Under construction	Under construction	Under construction	March 2016
Japanese	Done	Under construction	Under construction	Under construction	
Myanmar	Done	Under construction	Under construction	Under construction	
Indonesian	Done	Planned	Planned	Planned	March 2017
Vietnamese	Done	Planned	Planned	Planned	
Others	Under consideration				?

3. **G**lobal **C**ommunication **P**rogram

The Global Communication Program (Japan)

- The Global Communication Program (GCP) is a Japanese government project announced on 11 April 2014 to develop a multi-lingual speech translation system to bridge the language barrier during the Olympic Games in 2020.
- http://www.soumu.go.jp/main_content/000285578.pdf (in Japanese)

Target

Real-time machine translation services

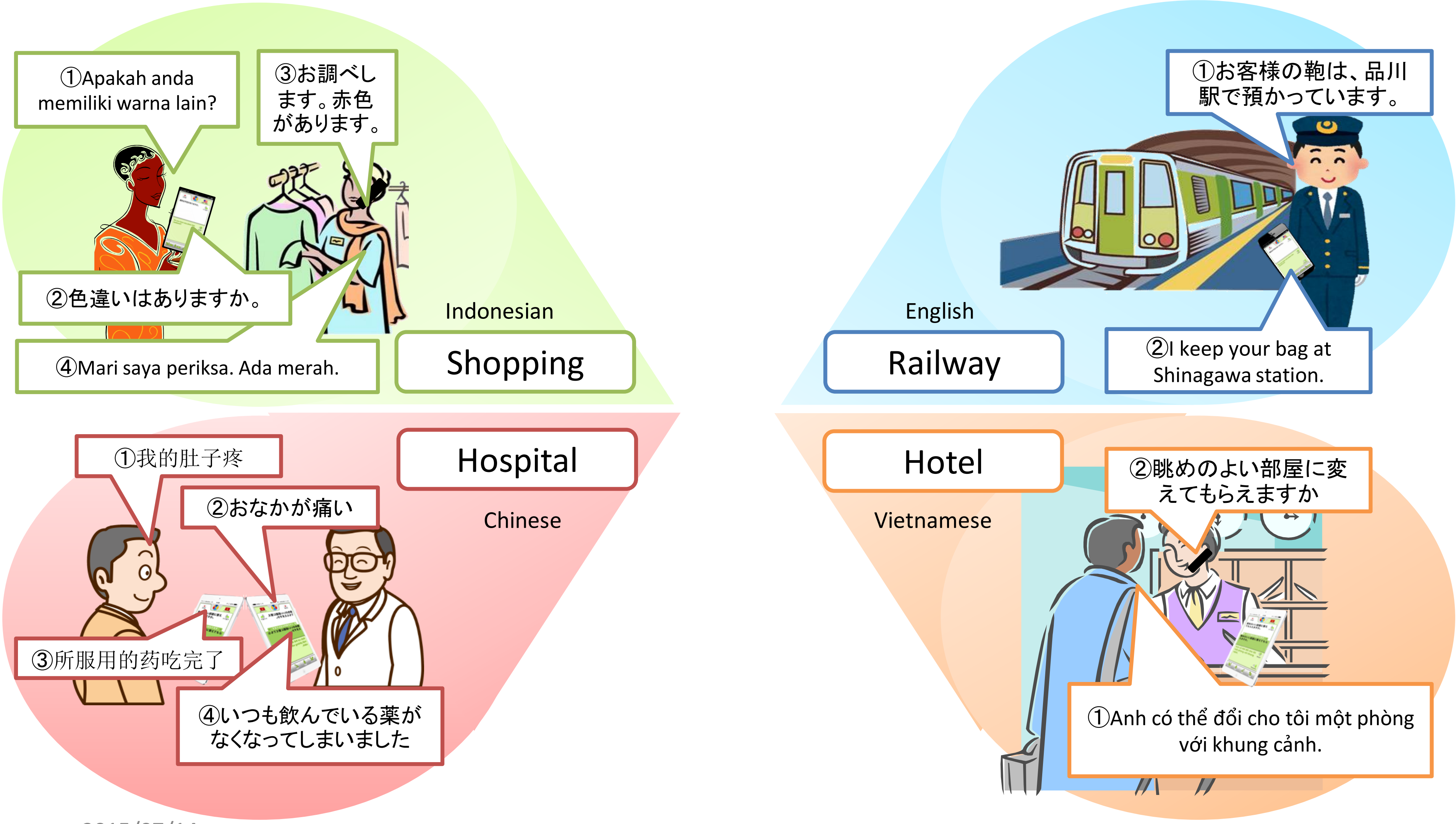
- covers **10** languages, including Asian ones such as Thai, Vietnamese, Indonesian and Myanmar
- using National Institute of Information and Communications Technology’s translation technology

NICT has established a new research center (16 Sept. 2014)

Panasonic FUJITSU NEC NTT ATR-Trek TOPPAN

HITACHI SONY TOSHIBA KDDI FEAT NHK

Future vision of Japanese society in 2020



Shopping

Indonesian

①Apakah anda memiliki warna lain?

②色違いはありますか。

③お調べします。赤色があります。

④Mari saya periksa. Ada merah.

Railway

English

①お客様の鞄は、品川駅で預かっています。

②I keep your bag at Shinagawa station.

Hospital

Chinese

①我的肚子疼

②おなかが痛い

③所服用的药吃完了

④いつも飲んでいる薬がなくなってしまいました

Hotel

Vietnamese

①Anh có thể đổi cho tôi một phòng với khung cảnh.

②眺めのよい部屋に変えてもらえますか

8 Reasons Why the Tokyo Olympics Will Be the Most Futuristic We've Ever Seen - **GIZMODO**

http://gizmodo.com/8-reasons-why-the-tokyo-olympics-will-be-the-most-futur-1728007440

Meanwhile, in the private sector, Panasonic is making a palm-sized gadget worn around the neck that will translate Japanese into 10 languages for the thousands of visitors set to descend on the metropolis. The electronics giant also plans to provide visitors with a smartphone app that scans Japanese signs and translates them on the spot. These are services that could be useful in countries across the globe.

https://www.youtube.com/watch?v=FluhQAXBX6E

4. Recent Research Topics at NICT

Translation of **Patent Claim**

Masaru Fuji, Atsushi Fujita, Masao Utiyama, Eiichiro Sumita, Yuji Matsumoto. Patent Claim Translation Based on Sublanguage-specific Sentence Structure. In Proceedings of Machine Translation Summit XV (**MT Summit 2015**), Miami, Florida, USA, October 30-November 3, 2015.

	EJ	JE
baseline	**23.7**	**22.3**
proposal	**28.8**	**27.5**

Big Gain in BLEU

Claim

The actuator according to claim 1, wherein an even number of notches are formed in said body, and the displacement of said rod in the axial direction is extracted.

Split source into patterns

PREA	the actuator according to claim 1
TRAP	wherein
PURP	an even number of notches are formed in said body, and the displacement of said rod in the axial direction is extracted

Convert to target patterns

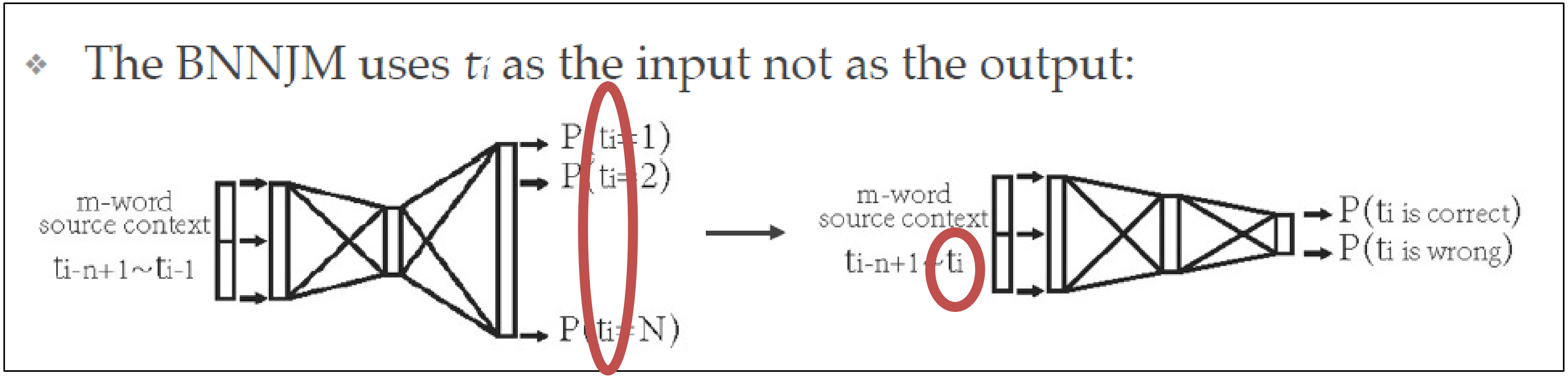
PURP	an even number of notches are formed in said body, and the displacement of said rod in the axial direction is extracted
TRAP	wherein
PREA	the actuator according to claim 1

Translate components by SMT with preordering

PURP	偶数個の切込みが形成されている前記本体であり、前記ロッドの変位には、軸方向 を抽出する
TRAP	ことを特徴とする
PREA	請求項1に記載のアクチュエータ

Binarized Neural Network Joint Model

Jingyi Zhang, Masao Utiyama, Eiichiro Sumita, Graham Neubig and Satoshi Nakamura (2015). A Binarized Neural Network Joint Model for Machine Translation. **EMNLP**.



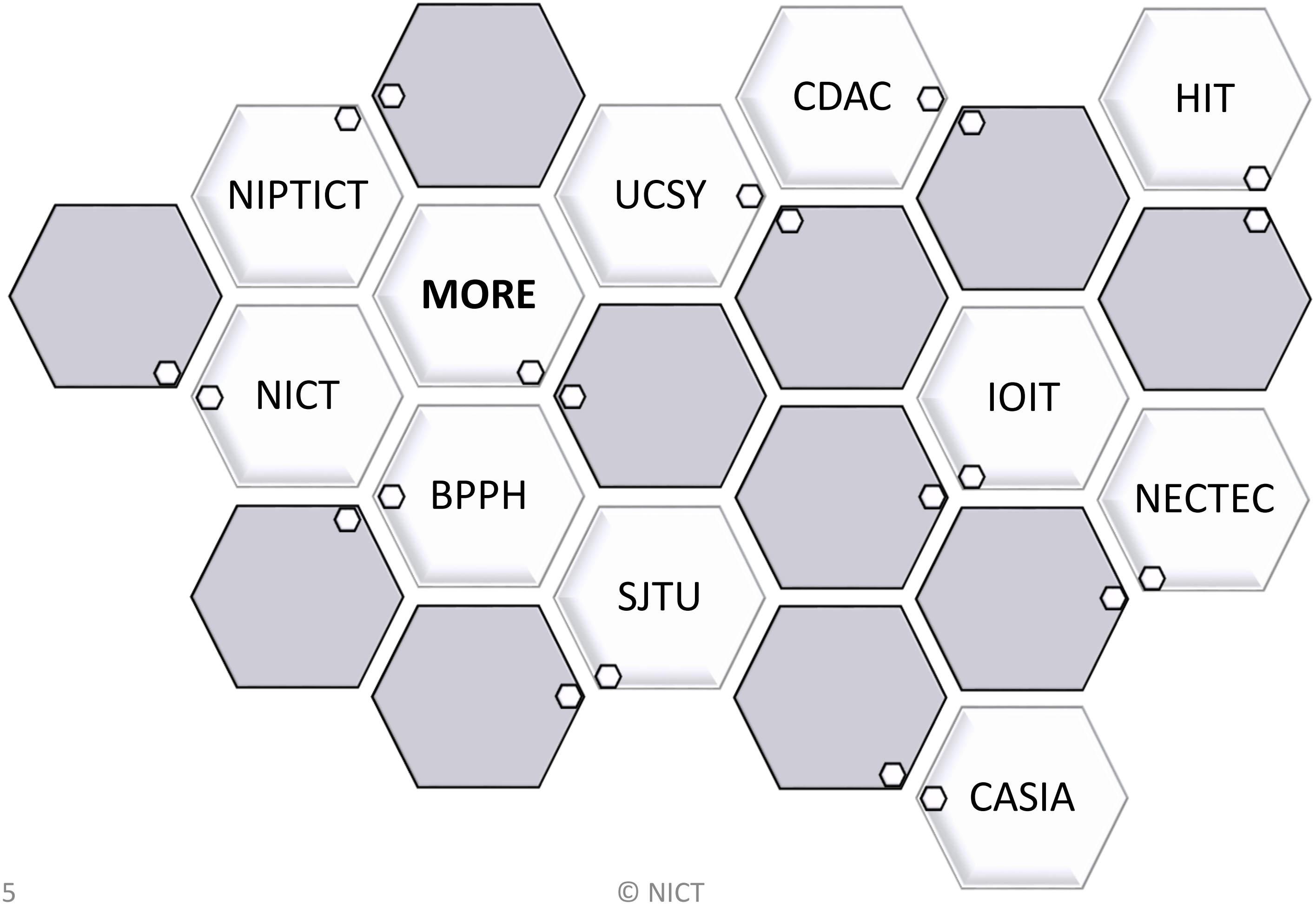
- The BNNJM uses t_i as the input not as the output:

Neural Network Joint Model

Binarized Neural Network Joint Model

The **binarized** model has achieved comparable performance and faster decoding/learning

5. Toward More and More Collaboration

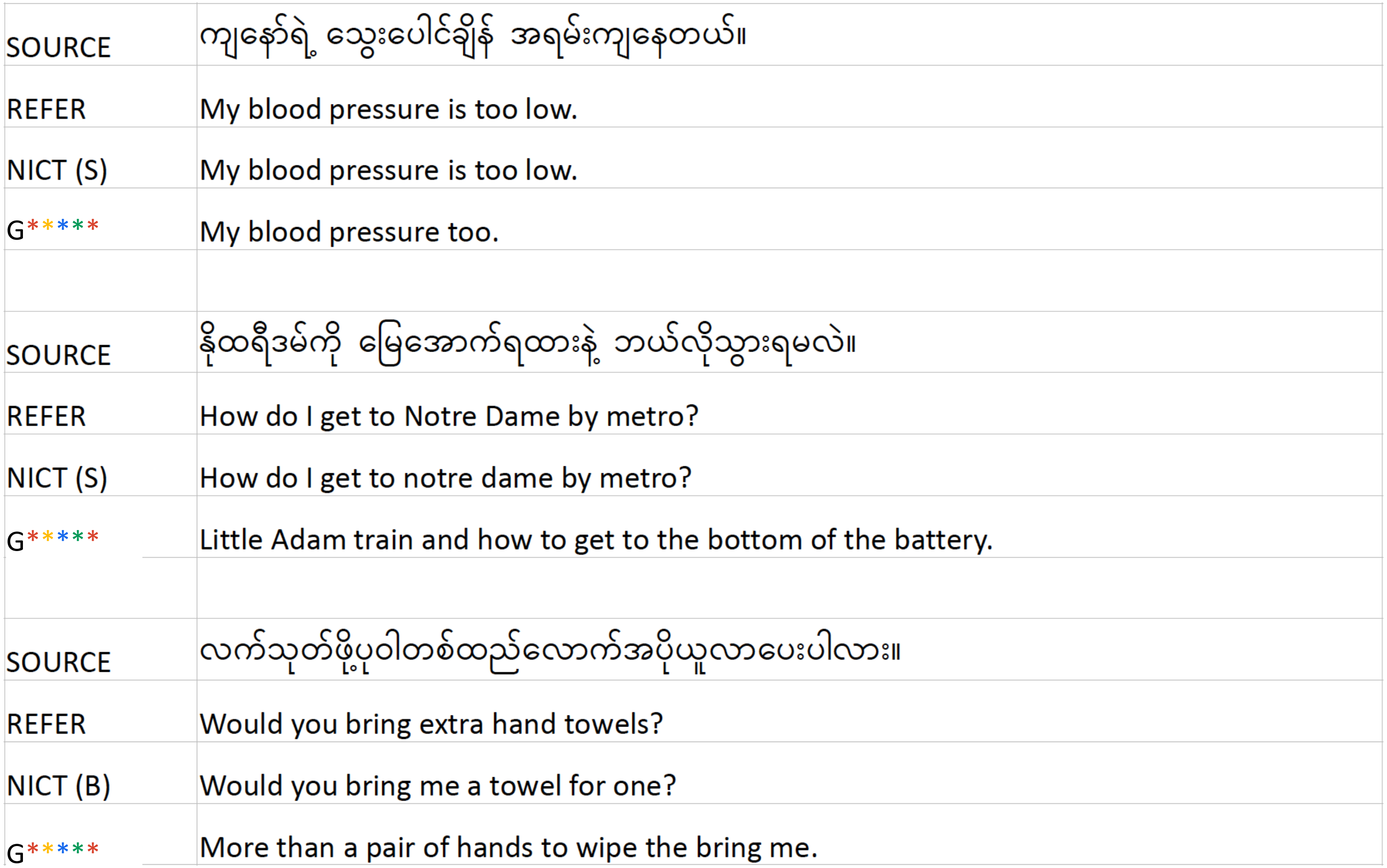


Collaboration between **Myanmar** and NICT(1/2)

- **Three researchers from UCSY studied S2S technologies from July 2014 to June 2015 at NICT.**
  - **They have developed the world's first Myanmar Speech Recognition system and Speech Synthesis system.**
  - **They are also working on the development of Machine Translation technology between Japanese, the Myanmar language and English.**

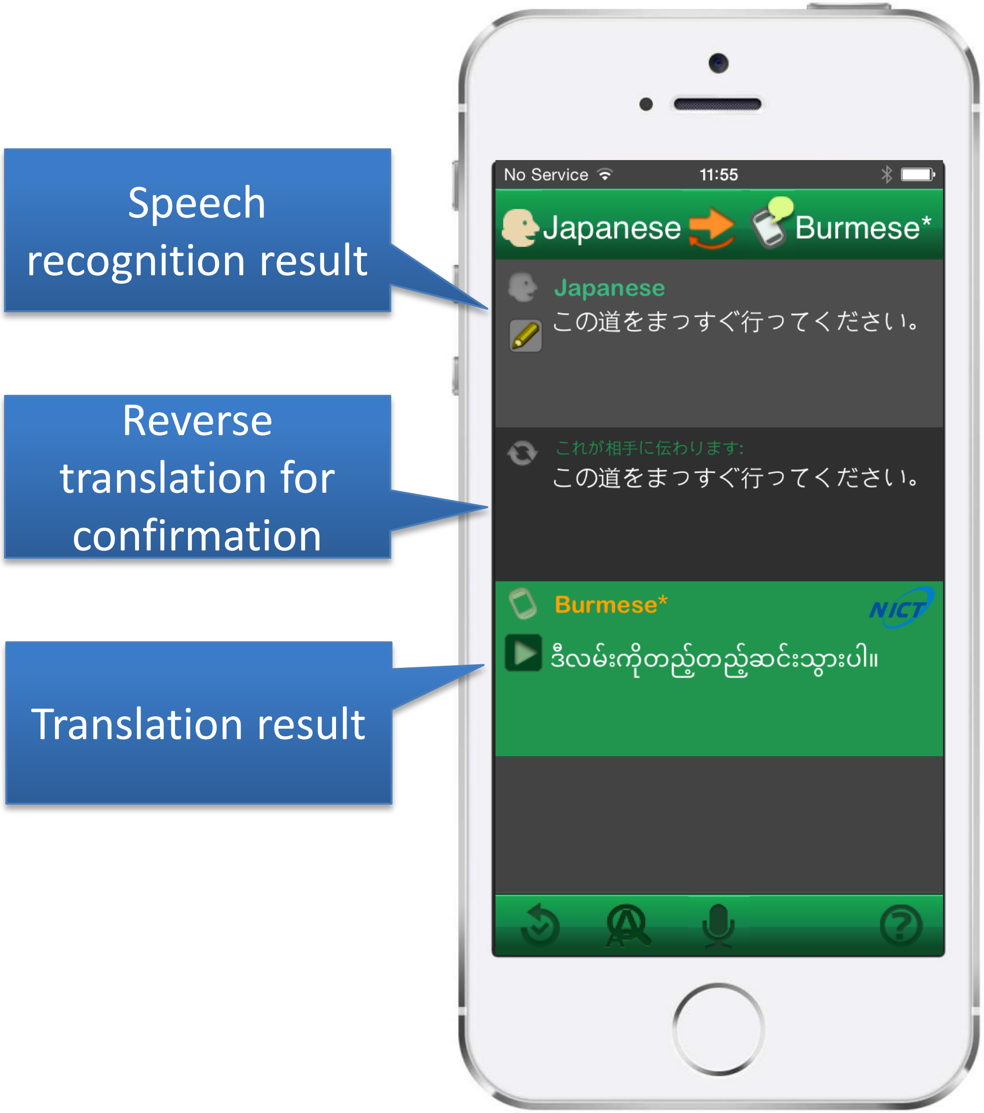
Collaboration between **Myanmar** and NICT (2/2)

- My-En (Translation results of NICT and G***** MT)



SOURCE	ကျနော်ရဲ့ သွေးပေါင်ချိန် အရမ်းကျနေတယ်။
REFER	My blood pressure is too low.
NICT (S)	My blood pressure is too low.
G*****	My blood pressure too.
SOURCE	နိုထရီဒမ်ကို မြေအောက်ရထားနဲ့ ဘယ်လိုသွားရမလဲ။
REFER	How do I get to Notre Dame by metro?
NICT (S)	How do I get to notre dame by metro?
G*****	Little Adam train and how to get to the bottom of the battery.
SOURCE	လက်သုတ်ဖို့ပုဝါတစ်ထည်လောက်အပိုယူလာပေးပါလား။
REFER	Would you bring extra hand towels?
NICT (B)	Would you bring me a towel for one?
G*****	More than a pair of hands to wipe the bring me.

Demonstration of speech translation from **English**/Japanese to **Myanmar**



Speech recognition result

Reverse translation for confirmation

Translation result

No Service 11:55

Japanese Burmese*

Japanese
この道をまっすぐ行ってください。

これが相手に伝わります:
この道をまっすぐ行ってください。

Burmese*
ဒီလမ်းကိုတည့်တည့်ဆင်းသွားပါ။

- Developed in collaboration with 3 researchers from Myanmar.
- Applicable for travel conversations such as: ‘Welcome’, ‘How about this one?’, ‘What is the price?’, ‘Thank you’, ‘Have a safe trip’.