

Language and Speech Translation Activities in Thailand

Chai Wutiwiwatchai

**National Electronics and Computer Technology Center
National Science and Technology Development Agency
THAILAND**

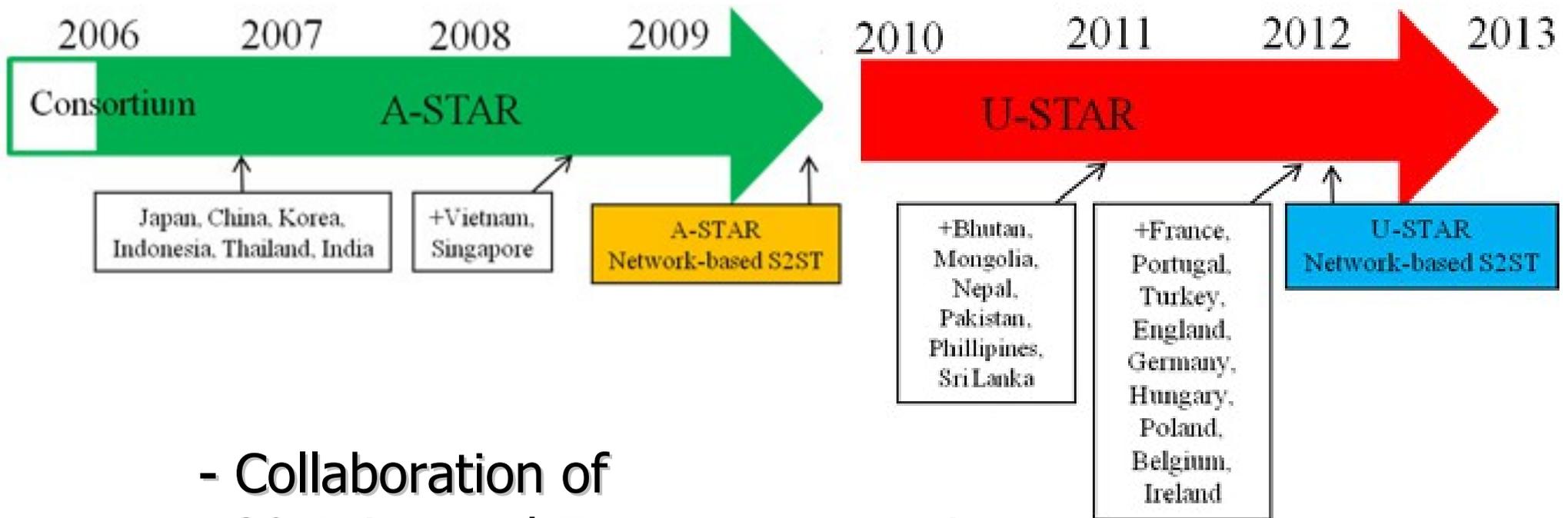
Outline

- **U-STAR Speech Translation (since 2007)**
 - Brief history
 - System architecture
 - Current status
 - Future plan
- **ASEAN Machine Translation (since 2012)**
 - Project overview
 - System architecture
 - Current status
 - Future plan

Outline

- **U-STAR Speech Translation (since 2007)**
 - Brief history
 - System architecture
 - Current status
 - Future plan
- **ASEAN Machine Translation (since 2012)**
 - Project overview
 - System architecture
 - Current status
 - Future plan

U-STAR History¹



- Collaboration of 30 Asian and European countries
- **Modality Conversion Marked-up Language (MCML)**, registered as an ITU-T recommendation standard

U-STAR History¹

- **2009 : A-STAR S2ST Live Demo**
 - Network-based Multilingual S2ST
 - 8 Asian languages and English
 - Peer-to-peer and Multi-party clients
 - Portable devices (UMPC)



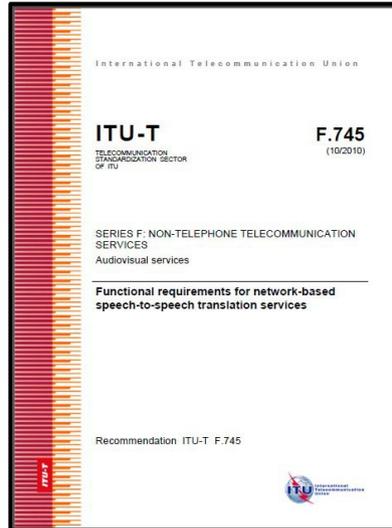
U-STAR History¹

- **2012 : U-STAR S2ST Public Service**

- Network-based Multilingual S2ST in the travel and sport domain launched in Jun 2012
- 23 Asian and European languages supported
- **VoiceTra4U-M**, an iPhone App available freely on the AppStore



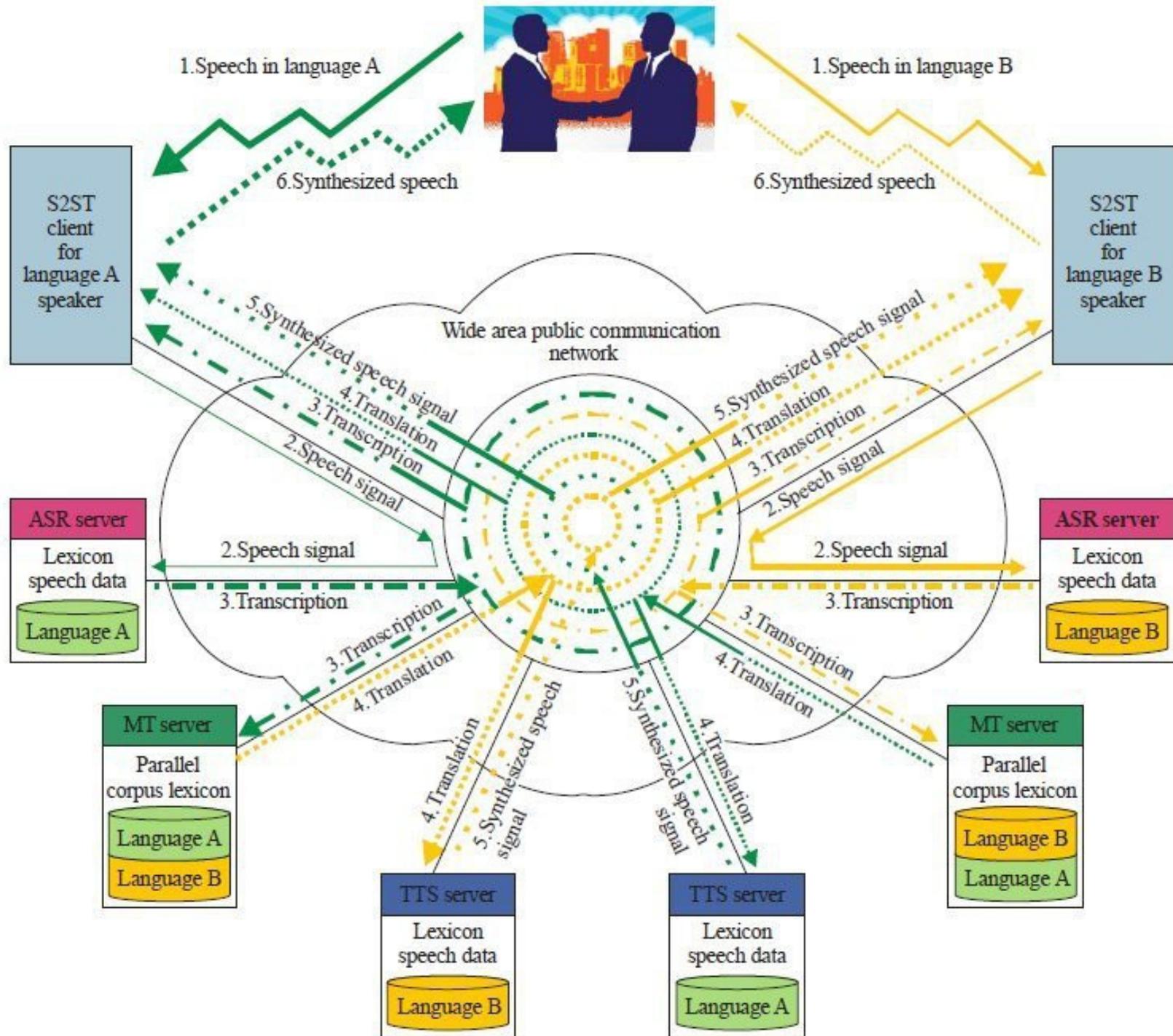
System Architecture²



- ITU-T F.745 – Functional requirements for network-based speech-to-speech translation services



- ITU-T H.625 – Architecture for network-based speech-to-speech translation services



Current Status

- **Language Resources**

- **Basic Travel Expression Corpus (BTEC)** has been used to translate to member languages since A-STAR
- To extend the service for users during London Olympic Games, an Olympic expression corpus by **Harbin Institute of Technology (HIT)** has been acquired and distributed to translate
- A **Named Entity (NE) list** of words related to Olympic expressions has also been collected from member countries

Current Status

- **Examples of Engines³**

Language	ASR	MT	TTS
English (En)	HMnet (SSS)		Concatenative
Hindi (Hi)		SMT (Cleopatra)	HMM
Indonesian (Id)	HMnet (SSS)	SMT (Moses)	HMM
Japanese (Ja)	HMnet (SSS)	SMT (Cleopatra)	Concatenative
Korean (Ko)	FST	RBMT (Parser)	HMM
Malay (Ms)	HMM	RBMT (Pyramid)	HMM
Thai (Th)	HMM	SMT (Moses)	HMM
Vietnamese (Vi)	HMM	SMT (Moses)	HMM
Chinese (Zh)	HMnet (SSS)	SMT (Cleopatra)	Concatenative

Current Status

- **No. of App Downloads**

 - 15,645 downloads (til 2012)

 - 5,179 Thai downloads (til 2012) - 2nd rank

 - 65,346 downloads (til 2014)

- **No. of Service Transactions**

 - 26,882 transactions (til 2012)

 - 5,179 transactions for Thai (til 2012) - 2nd rank

 - 514,552 transactions (til 2014)

 - 45,056 translations for Thai (til 2014)

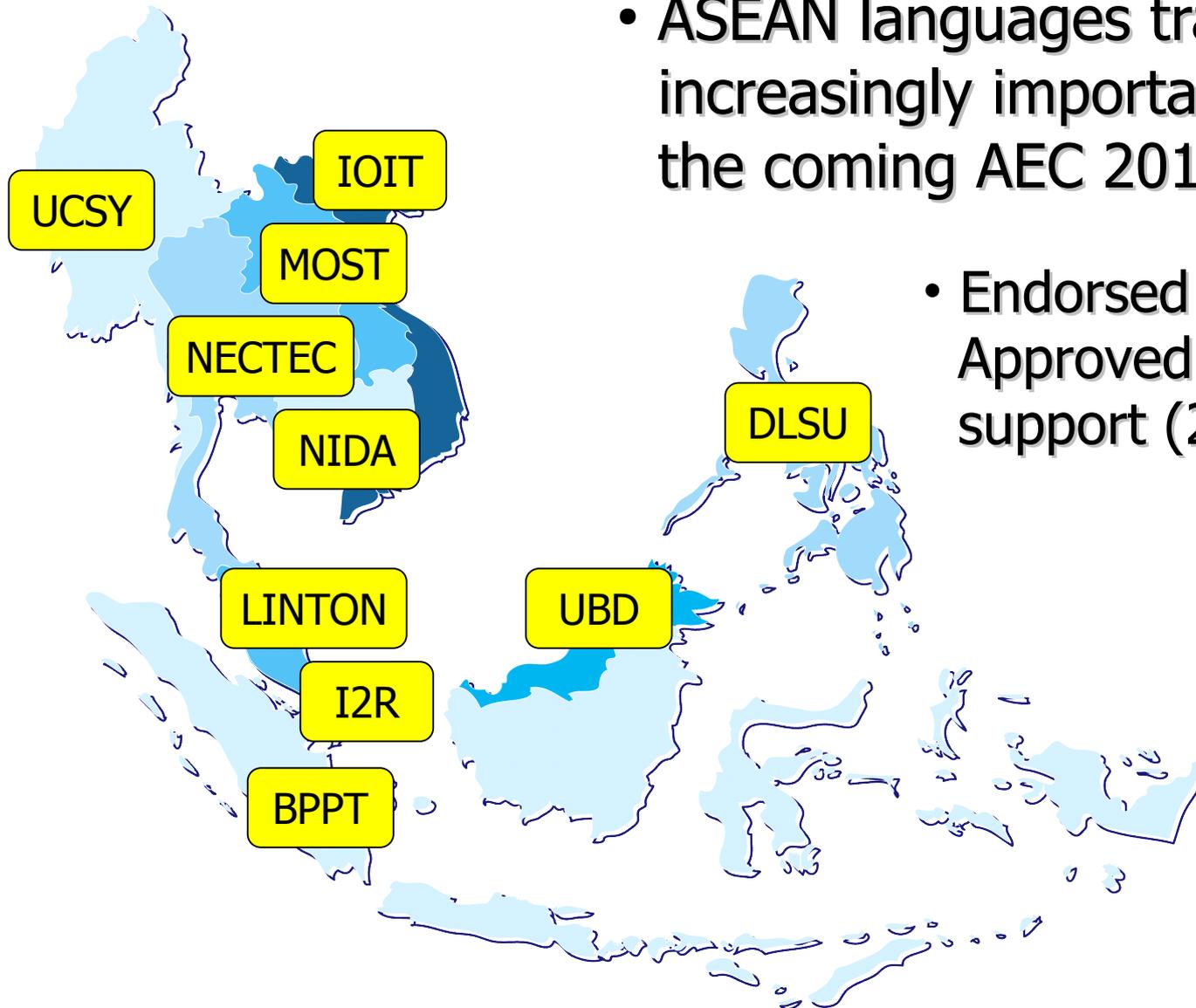
Future Plan

- **Named-Entities (NE)**
 - NE words are often language specific
 - Using a descriptive or transliterated word for unknown
- **Scalability and Extensibility**
 - Encouraging service maintenance from members
 - Improving service performance by using real data
 - Extending to new domains and languages
- **Service Latency**
 - The condition of network is the key
 - Setting communication mirror servers

Outline

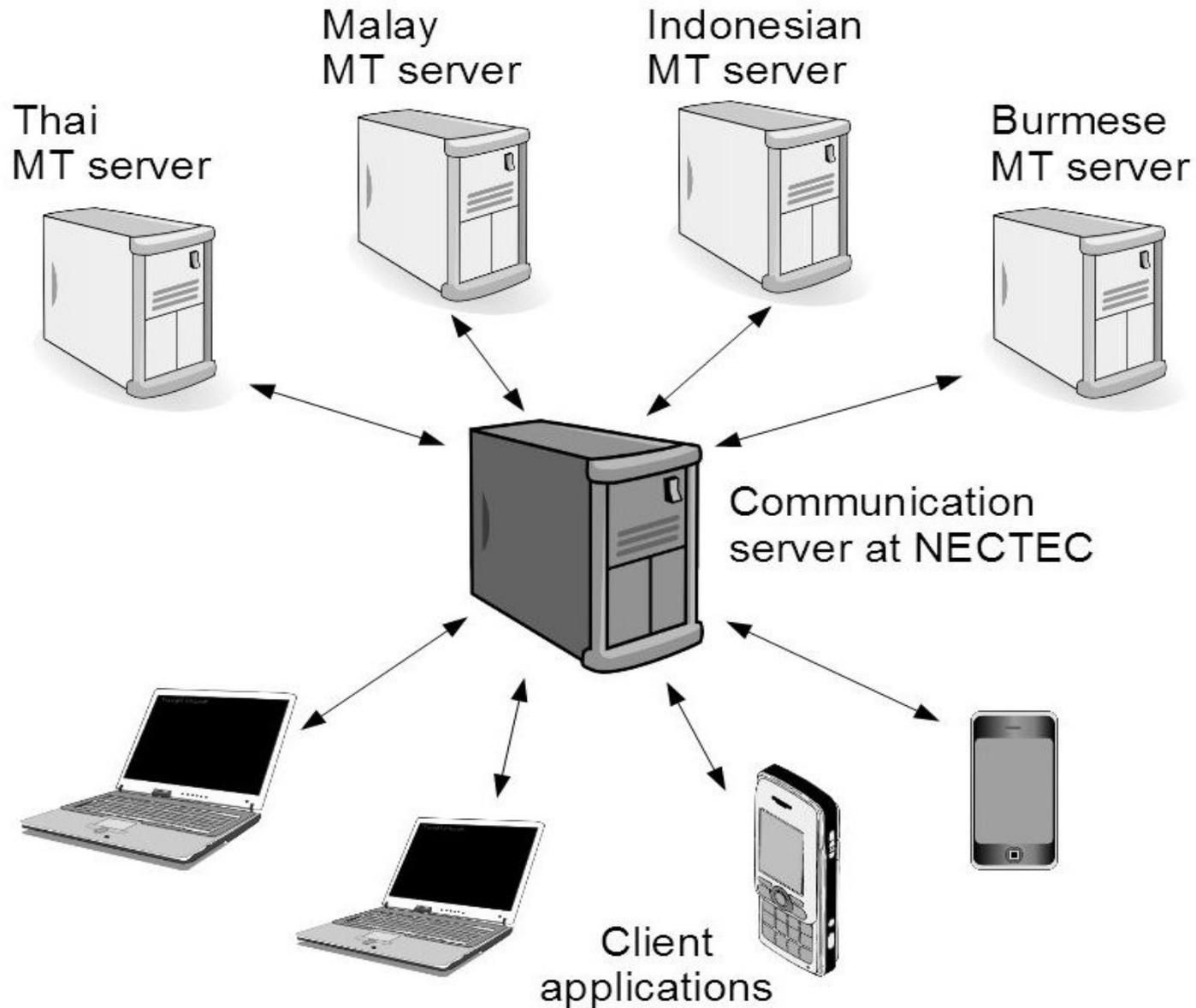
- **U-STAR Speech Translation (since 2007)**
 - Brief history
 - System architecture
 - Current status
 - Future plan
- **ASEAN Machine Translation (since 2012)**
 - Project overview
 - System architecture
 - Current status
 - Future plan

ASEAN-MT Project (Since 2012)

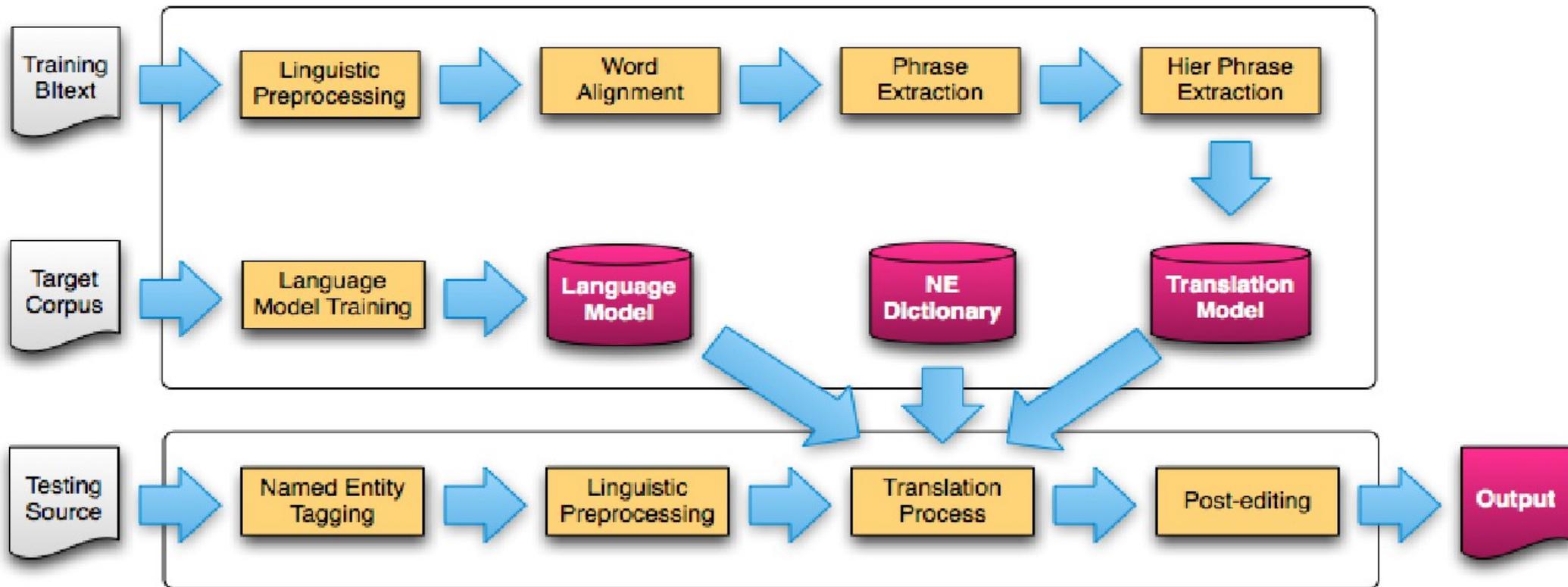


- ASEAN languages translation is increasingly important to support the coming AEC 2015
- Endorsed by ASEAN SCMIT
Approved for ASF partial support (2012-2014)

System Architecture



Statistical MT Approach



Current Status

- **Kick-off & The 1st Working Committee Meeting**
 - July 2012
 - Pattaya, Thailand
- **The 1st Technology Workshop**
 - January 2013
 - Pathumthani, Thailand
- **Progress Demonstration at ASEAN COST Meeting**
 - May 2013
 - Tagaytay City, Philippines
- **The 2nd Working Committee Meeting**
 - December 2013
 - Ayudhaya, Thailand

Current Status

Parallel text corpus: 20,000 sentences in travel domain

Country	Language	Translation	SMT	NE Tag
Brunei	Malay			
Cambodia	Cambodian	20,000	20,000	20,000
Indonesia	Indonesian	20,000	20,000	20,000
Laos	Lao	20,000	20,000	20,000
Malaysia	Malay	20,000	20,000	20,000
Myanmar	Myanmar	10,000	10,000	20,000
Philippines	Filipino	20,000	20,000	20,000
Singapore	Chinese	20,000	20,000	20,000
Thailand	Thai	20,000	20,000	20,000
Vietnam	Vietnamese	20,000	20,000	20,000

Current Status

Size	20,000
Domain	Travel
People	Greeting, Introduction, Communication
Survival	Transportation, Accommodation, Finance
Food	Food, Beverage, Restaurant
Fun	Recreation, Traveling, Shopping, Nightlife
Resource	Number, Time, Currency
Special Needs	Emergency, Health
NE types	17 types

4 <http://www.aseanmt.org/demo> Demonstration in ASEAN COST Meeting, May 2014

ASEAN Machine Translation Demo in Travel Domain (for the 65th COST Meeting only)

I want to go to the restaurant ฉัน ต้องการ ไป ภัตตาคาร Translate
From : English To : Thai Next

ฉัน ต้องการ ไป ภัตตาคาร Tôi muốn đi đến nhà hàng . Translate
From : Thai To : Vietnamese Next

PHILIPPINES MALAYSIA THAILAND
MYANMAR VIETNAM SINGAPORE LAOS
BRUNEI CAMBODIA INDONESIA

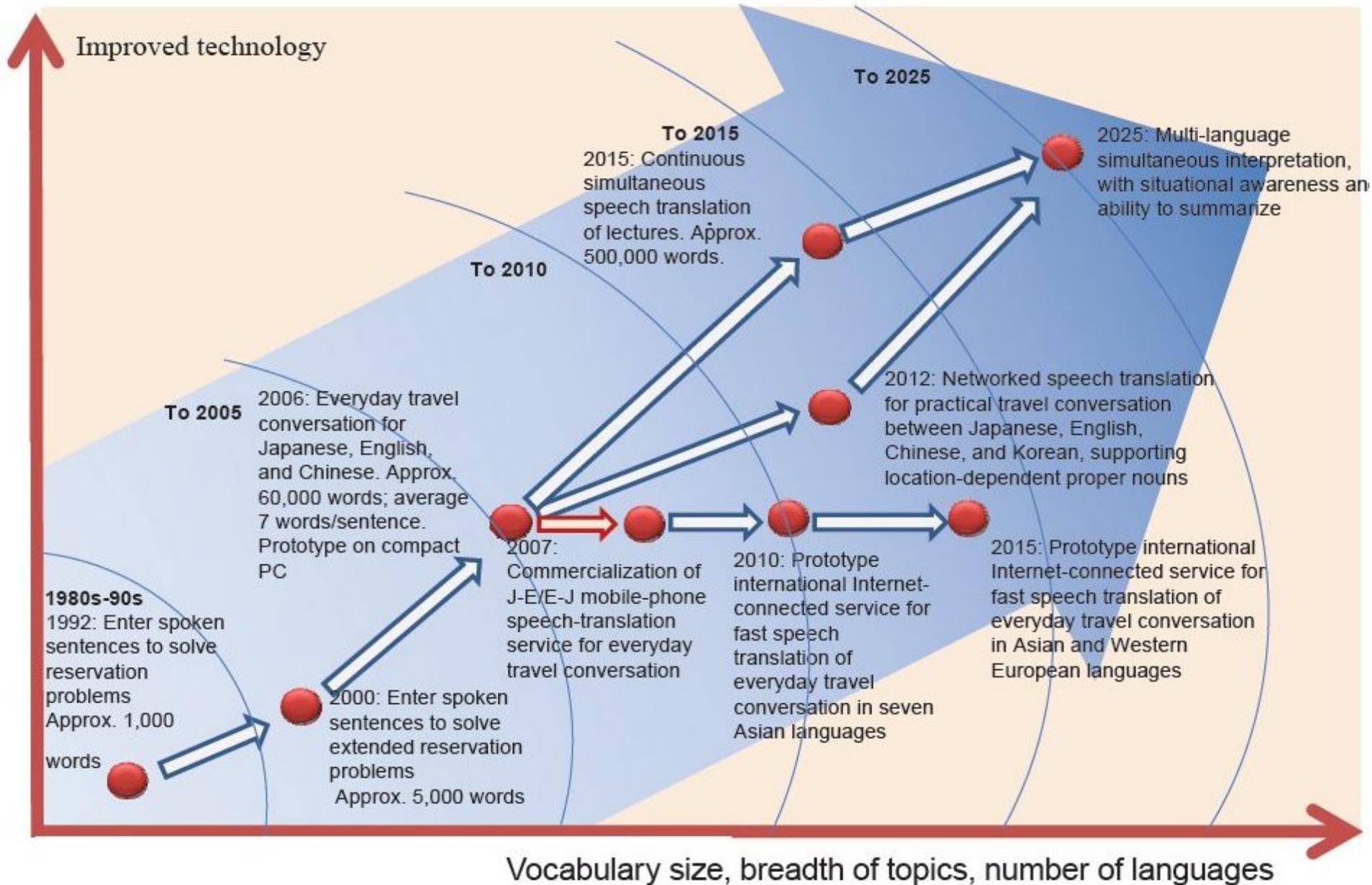
One
Vision, One Identity, One Community.

ASEAN Machine Translation Project.
National Electronics and Computer Technology Center
112 Thailand Science Park (TSP), Paholyothin Road,
Klong Nueng, Klong Luang, Pathumthani 12120, Thailand

Future Plan

- **Evaluation**
 - Overall system evaluation
- **Post-Editing Module**
 - R&D on a post-editing module
- **2015 Activities**
 - The 2nd technology workshop
 - The 3rd working committee meeting
 - Final demonstration at ASEAN COST meeting

• Future of S2ST⁵



References

- ¹ U-STAR consortium, <http://ustar-consortium.com/>
- ² ITU-T standard, <http://www.itu-t.int/>
- ³ S. Sakti, M. Paul, A. Finch, S. Sakai, T. T. Vu, N. Kimura, C. Hori, E. Sumita, S. Nakamura, J. Park, C. Wutiwiwatchai, B. Xu, H. Riza, K. Arora, C. M. Luong, H. Li, 2011. Toward translating Asian spoken languages. *Computer Speech and Language* (2011), doi:10.1016/j.csl.2011.07.001.
- ⁴ ASEAN-MT project, <http://www.aseanmt.org/>
- ⁵ S. Nakamura, 2009. Overcoming the language barrier with speech translation technology. *Science and Technology Trends – Quarterly Review* No. 31, Apr 2009, pp. 35-48.