

# Leveraging ASEAN Economic Community through Language Translation Services



Badan Pengkajian dan Penerapan Teknologi

Hamмам Riza

Center for Information and Communication Technology  
Agency for the Assessment and Application of Technology  
(BPPT)



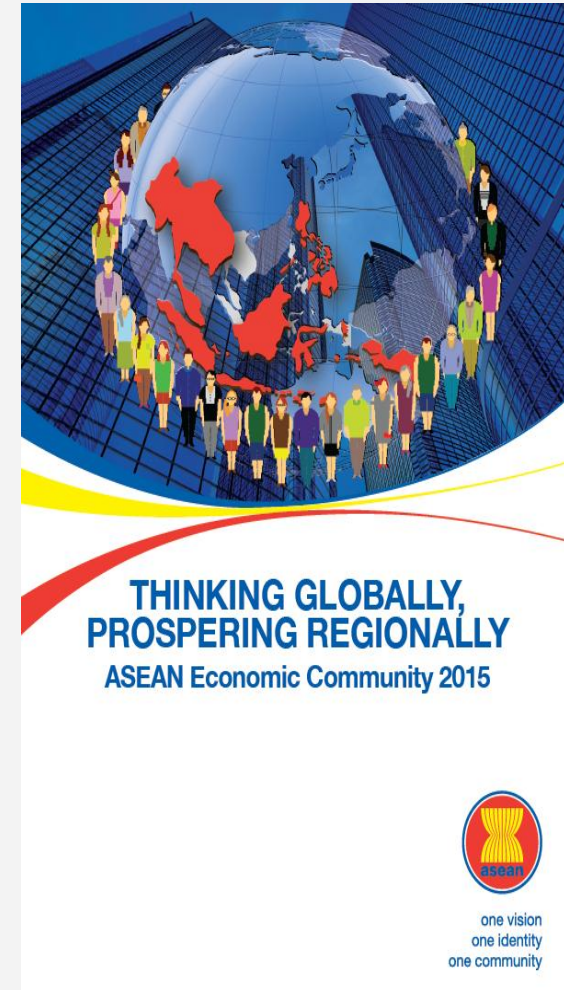
# Abstract

In this talk, I will present the activities of BPPT's Speech and Translation Research Group (PERISALAH), including the latest bidirectional Indonesian-English statistical machine translation, and our contribution toward developing network-based ASEAN Languages Translation Public Services (ASEAN-MT). This collaboration work involving 7 ASEAN languages: Thai, Malay, Indonesian, Lao, Cambodian, Vietnamese, Myanmar and English as predominant language in ASEAN countries. As part of the inauguration of ASEAN Economic Communities 2015, the ASEAN-MT public service in the travel domain will be brought online with acceptable response time with at least 3 of 5 score of user satisfaction on the service is achieved.

This joint effort not only building a practical network-based service on ASEAN languages text translation in the tourism domain, but most important to strengthen ASEAN collaboration in science and technology, sharing languages resources and knowledge of the translation technology among ASEAN member states and other countries. As part of improving Indonesian-English statistical machine translation, I will introduce the new 2014 project within Indonesian Language Development Agency (Badan Pengembangan Bahasa Indonesia) on building large scale parallel corpora based on ANTARA News corpora and e-KBBI (the official Indonesian Large Electronic Dictionary)

# Introduction

- ❖ **READY:** This is ASEAN's time. In the geographic heart of the world's premier growth corridor, ASEAN is poised to "seize the moment."
- ❖ **SET:** With a market of **over 600 million consumers** and combined **GDP of nearly US\$3 trillion**, ASEAN is offering a future of prosperity and stability. The AEC is one of the foundations of that future.
- ❖ **GO:** Agreements on trade, services and investment are changing the economic landscape and allowing the freer flow of goods, services and people across the region.



# 1



## ASEAN Economic Community ประเทศ ประชาคมเศรษฐกิจอาเซียน



**THAILAND**  
ราชอาณาจักรไทย

ธงชาติประเทศไทย  
ผู้ปกครอง  
เมืองหลวง  
พื้นที่  
ประชากร  
ภาษา  
สกุลเงิน

ประชาธิปไตยอันมีพระมหากษัตริย์ทรงเป็นประมุข  
นายกรัฐมนตรี  
คุณสมชาย พริยกุล  
512,115 ตารางกิโลเมตร  
66 ล้านคน (พ.ศ. 2553)  
ไทย  
บาท

**INDONESIA**  
สาธารณรัฐอินโดนีเซีย

ธงชาติอินโดนีเซีย  
ผู้ปกครอง  
เมืองหลวง  
พื้นที่  
ประชากร  
ภาษา  
สกุลเงิน

ประชาธิปไตยที่มีประธานาธิบดีเป็นประมุข  
ประธานาธิบดี  
จาการ์  
1,804,443 ตารางกิโลเมตร  
242 ล้านคน (พ.ศ. 2553)  
อินโดนีเซีย  
รูเปียห์

**VIET NAM**  
สาธารณรัฐสังคมนิยมเวียดนาม

ธงชาติเวียดนาม  
ผู้ปกครอง  
เมืองหลวง  
พื้นที่  
ประชากร  
ภาษา  
สกุลเงิน

ลัทธิคอมมิวนิสต์  
เป็นระบอบการปกครอง  
นายกรัฐมนตรี  
ฮานอย  
331,890 ตารางกิโลเมตร  
89.37 ล้านคน (พ.ศ. 2553)  
เวียดนาม  
ดอง

**CAMBODIA**  
ราชอาณาจักรกัมพูชา

ธงชาติกัมพูชา  
ผู้ปกครอง  
เมืองหลวง  
พื้นที่  
ประชากร  
ภาษา  
สกุลเงิน

ประชาธิปไตยแบบราชาธิปไตยที่มีกษัตริย์เป็นประมุข  
นายกรัฐมนตรี  
พนมเปญ  
181,035 ตารางกิโลเมตร  
14.14 ล้านคน (พ.ศ. 2553)  
เขมร  
เรียลกัมพูชา

**LAO PDR**  
สาธารณรัฐประชาธิปไตยประชาชนลาว

ธงชาติลาว  
ผู้ปกครอง  
เมืองหลวง  
พื้นที่  
ประชากร  
ภาษา  
สกุลเงิน

ลัทธิคอมมิวนิสต์ (ระบอบการปกครอง)  
โดยมีพรรคประชาชนปฏิวัติลาว  
นายกรัฐมนตรี  
เวียงจันทน์  
236,800 ตารางกิโลเมตร  
6.0 ล้านคน (พ.ศ. 2553)  
ลาว  
กีบ

**SINGAPORE**  
สาธารณรัฐสิงคโปร์

ธงชาติสิงคโปร์  
ผู้ปกครอง  
เมืองหลวง  
พื้นที่  
ประชากร  
ภาษา  
สกุลเงิน

สาธารณรัฐแบบรัฐสภา (มีกษัตริย์)  
นายกรัฐมนตรี  
สิงคโปร์  
710 ตารางกิโลเมตร  
5.08 ล้านคน (พ.ศ. 2553)  
อังกฤษ, มาเลย์, จีนกลาง  
ดอลลาร์สิงคโปร์

**MALAYSIA**  
มาเลเซีย

ธงชาติมาเลเซีย  
ผู้ปกครอง  
เมืองหลวง  
พื้นที่  
ประชากร  
ภาษา  
สกุลเงิน

สหพันธรัฐ โดยมีกษัตริย์ทรงเป็นประมุข  
นายกรัฐมนตรี  
กัวลาลัมเปอร์  
329,738 ตารางกิโลเมตร  
28.9 ล้านคน (พ.ศ. 2553)  
มาเลย์  
ริงกิตมาเลย์

**Brunei Darussalam**  
รัฐบรูไนดารุสซาลาม

ธงชาติบรูไน  
ผู้ปกครอง  
เมืองหลวง  
พื้นที่  
ประชากร  
ภาษา  
สกุลเงิน

สมบูรณาญาสิทธิราชย์  
โดยมีกษัตริย์ทรงเป็นประมุข  
นายกรัฐมนตรี  
บันดาร์เซาะบายวัน  
5,765 ตารางกิโลเมตร  
414,000 คน (พ.ศ. 2553)  
มาเลย์  
ดอลลาร์บรูไน

**MYANMAR**  
สาธารณรัฐแห่งสหภาพพม่า

ธงชาติพม่า  
ผู้ปกครอง  
เมืองหลวง  
พื้นที่  
ประชากร  
ภาษา  
สกุลเงิน

ระบอบรัฐสภา  
ประธานาธิบดี  
เนปีดอ  
657,140 ตารางกิโลเมตร  
58.38 ล้านคน (พ.ศ. 2553)  
พม่า  
จ๊าด

**PHILIPPINES**  
สาธารณรัฐฟิลิปปินส์

ธงชาติฟิลิปปินส์  
ผู้ปกครอง  
เมืองหลวง  
พื้นที่  
ประชากร  
ภาษา  
สกุลเงิน

สาธารณรัฐ โดยมีประธานาธิบดีเป็นประมุข  
ประธานาธิบดี  
มะนิลา  
298,170 ตารางกิโลเมตร  
101.8 ล้านคน (พ.ศ. 2553)  
อังกฤษ, สเปน, ฟิลิปปิน  
เปโซ

The ASEAN Economic Community (AEC) shall be the goal of regional economic integration by 2015.

AEC envisages the following key characteristics:

- (a) a single market and production base,
- (b) a highly competitive economic region,
- (c) a region of equitable economic development,
- (d) a region fully integrated into the global economy.

Total Population: 600 million+  
GDP: USD \$3 trillion





# ASEAN Languages



# Network-based ASEAN language translation for public services (ASEAN-MT)



- **ASEAN-MT Background**

ASEAN languages translation is increasingly important to support the coming AEC 2015

- **The ASEAN-MT Project**

- Endorsed by SCMIT in May 2011
- Approved for ASF partial support in May 2012
- Start in July 2012
- Launch in Oct 2015

# ASEAN-MT Objectives

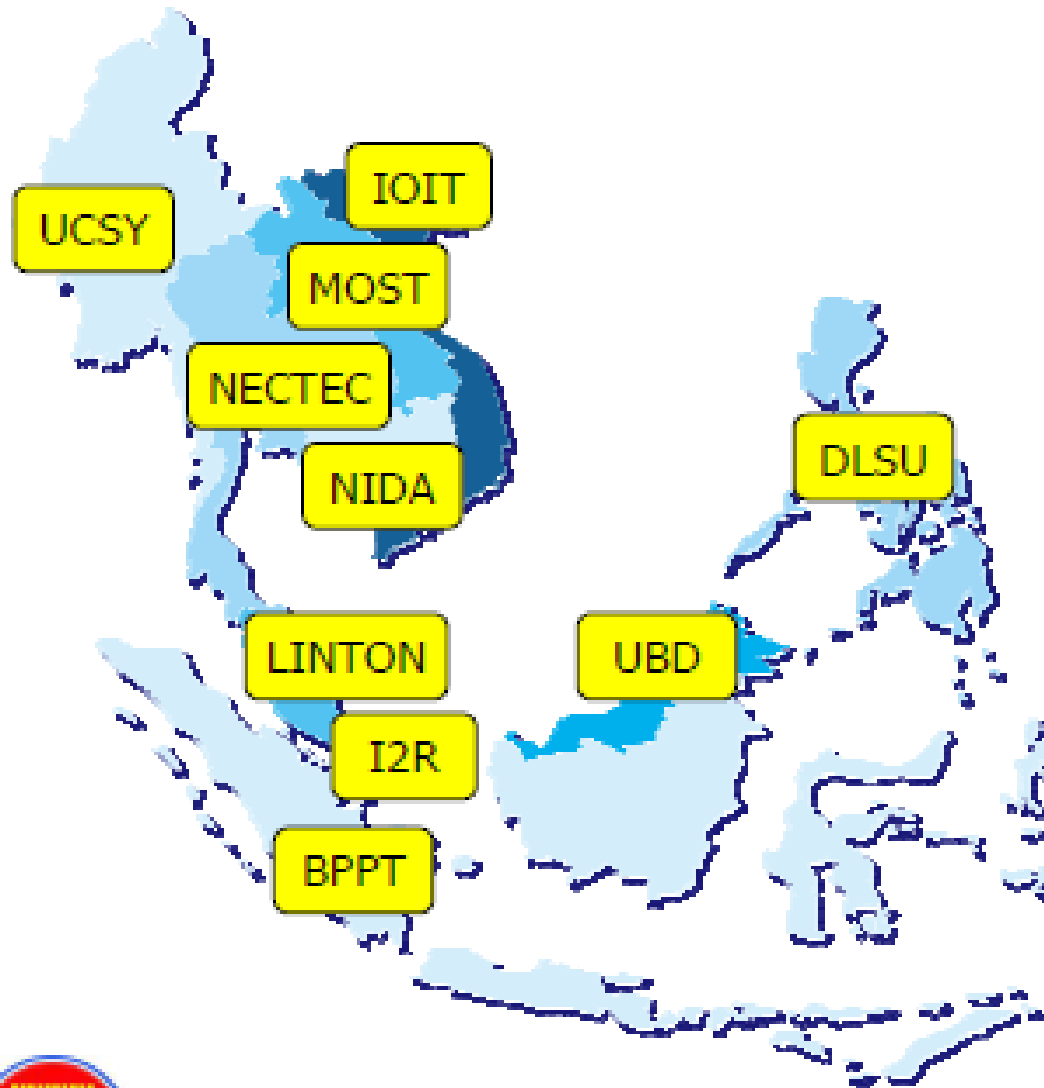
- R&D on network-based ASEAN languages machine translation for the public use
- Sharing language translation knowledge and resources among ASEAN countries
- Reducing the language barrier among and outside ASEAN after the beginning of ASEAN Community in 2015.

Create

Share

Sustain

# ASEAN-MT Collaboration



- *Universiti Brunei Darussalam (UBD)*
- *National Information Communications Technology Development Authority (NiDA)*
- *Agency for the Assessment and Application of Technology (BPPT)*
- *Computer Technology and Electronic Institute, Ministry of Science and Technology*
- *Linton University College*
- *University of Computer Studies, Yangon (UCSY)*
- *De La Salle University (DLSU)*
- *Institute for Infocomm Research (I2R)*
- *Institute of Information Technology (IOIT)*
- *National Electronics and Computer Technology Center (NECTEC)*

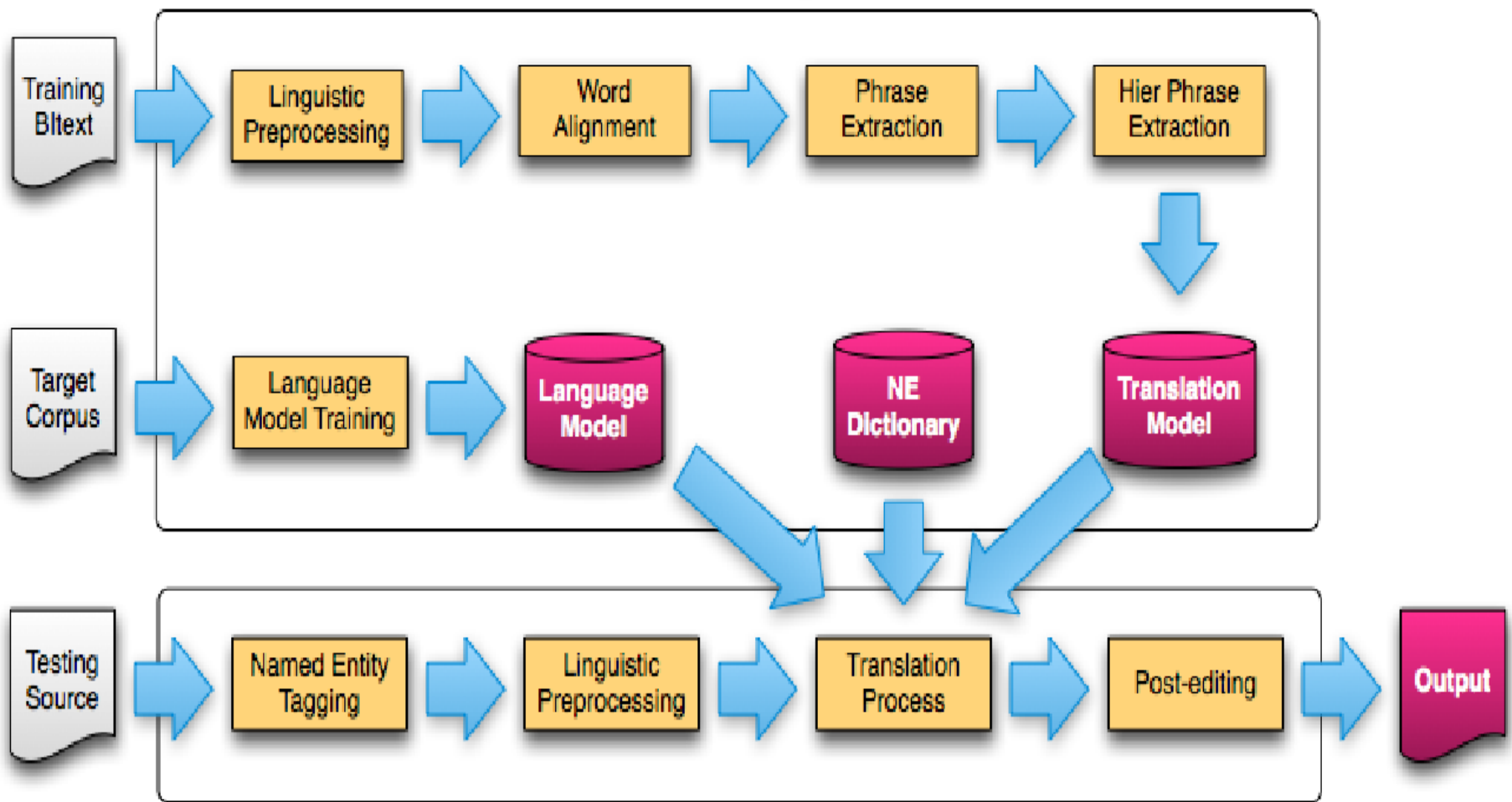


ASEAN-MT – NAC 2013

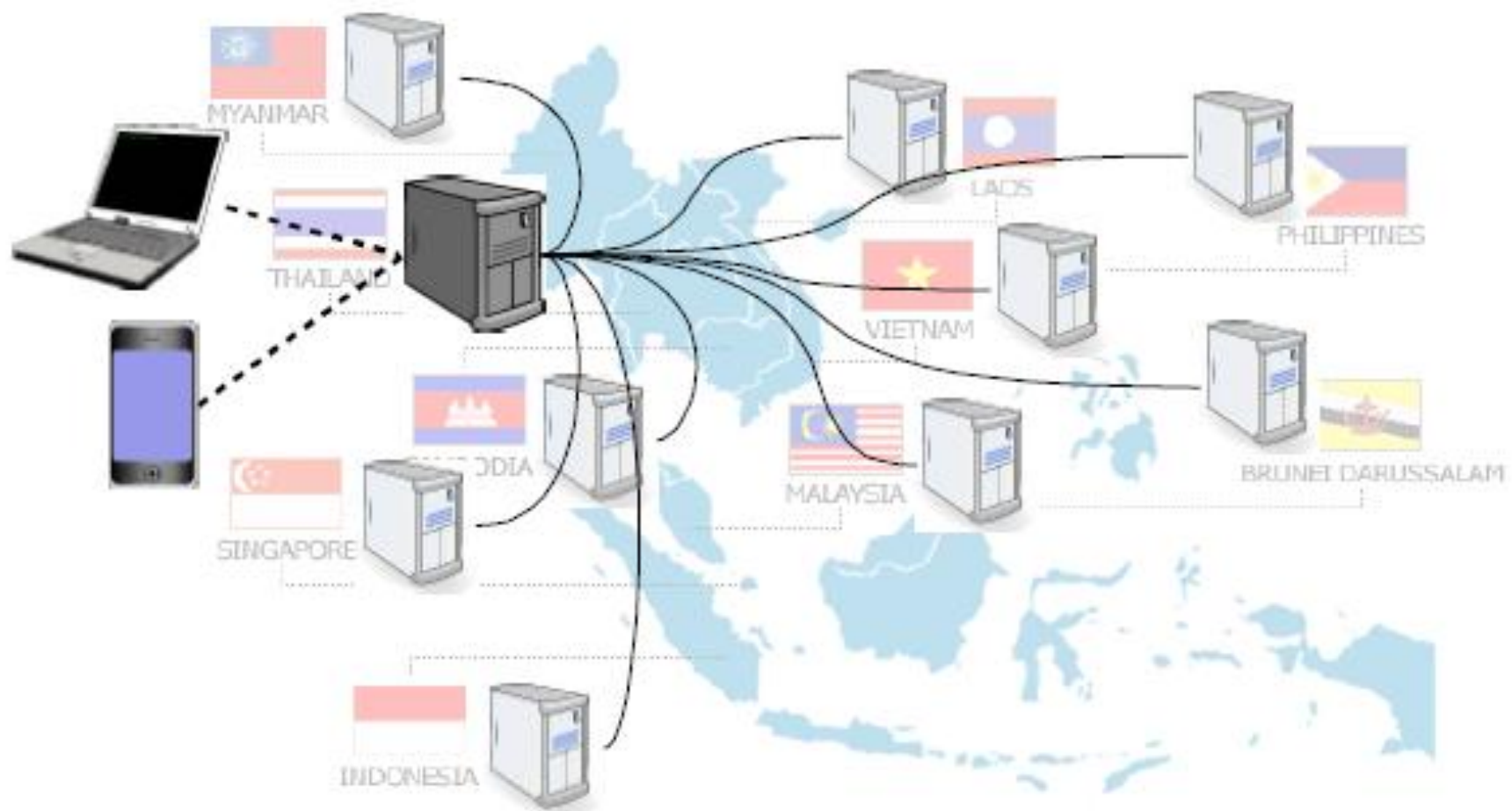


# ASEAN-MT Approach

## Statistical based Machine Translation



# ASEAN-MT Architecture



# Demonstration at SCMIT and COST meeting May 23, 2013



Firefox | Problem loading page | WELCOME TO THE KANTARY HOTEL... | ASEAN Machine Translation Demo | +

www.aseanmt.org/demo/ | Google

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

Flags of ASEAN member states: Indonesia, Philippines, Thailand, Laos, Malaysia, Brunei, Singapore, Vietnam, Myanmar.

**Translation 1:**  
From: English | To: Thai  
Input: I want to go to the restaurant  
Output: ฉัน ต้องการ ไป ภัตตาคาร

**Translation 2:**  
From: Thai | To: Vietnamese  
Input: ฉัน ต้องการ ไป ภัตตาคาร  
Output: Tôi muốn đi đến nhà hàng.

**ASEAN Logo and Member States:**  
PHILIPPINES, MALAYSIA, THAILAND, MYANMAR, VIETNAM, SINGAPORE, LAOS, BRUNEI, CAMBODIA, INDONESIA

Vision, One Identity, One Community.

One

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

Windows Taskbar: Desktop, 8:59, 19/12/2556



# ASEAN-MT Activities

	Activity	1 <sup>st</sup> year		2 <sup>nd</sup> year		3 <sup>rd</sup> year	
		2012	2013	2014	2015		
Y1.1	Kick-off meeting (WC meeting 1)						
Y1.2	Resource development		5				
Y1.3	Technology workshop						
Y2.1	Individual MT engine development		5				
Y2.2	Service system development						
Y2.3	Client application development						
Y2.4	Demonstration in ASEAN COST						
Y2.5	WC meeting 2						
Y3.1	Resource extension						
Y3.2	Improvement and evaluation						
Y3.3	WC meeting 3 and press						
Y3.4	Conclusion and report						



# MT Public Service



Thailand developers developed an application which will help the Association of Southeast Asian Nation in terms of language barriers. The app is called an ASEAN One App. Right now the app is on its first phase. It will translate a local language into 11 different ASEAN language which includes English.

Source: <http://philnews.ph/2012/03/09/asean-one-app-plans-speak-11-languages-developed-thailand/>

# Extending ASEAN+ICJK Languages

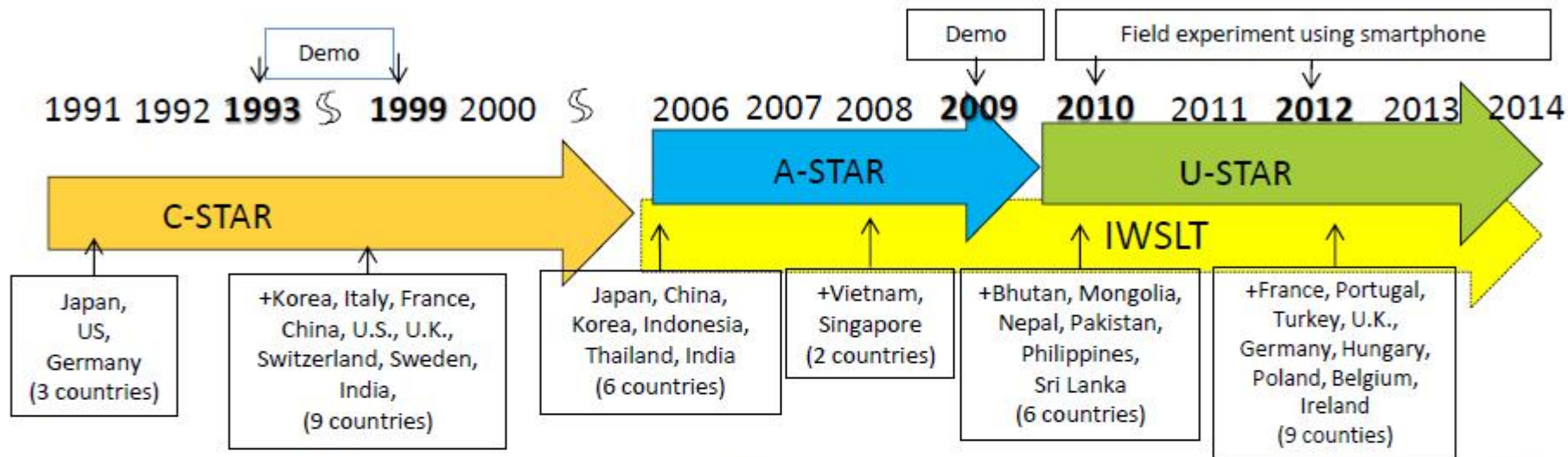


WAT 2015?



# International Research Consortia

for Network-based Speech-to-Speech Translation Technology



## U-STAR

Universal  
Speech

Translation

Advanced Research Consortium

**30 institutes from  
25 countries/regions**



# SMT Indonesian-English



[Home](#) [About Us](#) [Partners](#) [Projects](#) [Activities](#) [Outputs](#) [Linguistic Data Bank](#) [Localization Policy Data Bank](#) [Contact Us](#)

## Indonesia

[\[Publications\]](#) [\[Software\]](#) [\[Linguistic Resources\]](#)

### Partners



### Log In

Username

Password

Remember Me

[Lost your password?](#)  
[Register](#)

### Search



### Publications

- [Research Report on Local Language Computing: Development of Indonesian Language Resources and Translation System](#) [License](#)
- [Research Report on Corpus Design and Collection and Cleaning Tools English to Bahasa Indonesia](#) [License](#)
- [Research Report on Translation, Alignment, and Other Issues Related to Parallel Corpus Development from English to Bahasa Indonesia](#) [License](#)
- [Research paper on Probabilistic Part of Speech Tagging for Bahasa Indonesia](#) [License](#)
- [Initial Design Report on Statistical Machine Translation Framework](#) [License](#)
- [Final Design Report on Statistical Machine Translation Framework](#) [License](#)
- [Research Report on SMT for English to Bahasa Indonesia](#) [License](#)



### Software

- [SMT from English to Bahasa Indonesia \(Online\)](#) [License](#)  
[Tools Corpus](#)

- [Part of Speech Tagger for Bahasa Indonesia](#) [License](#)

[↑TOP](#)



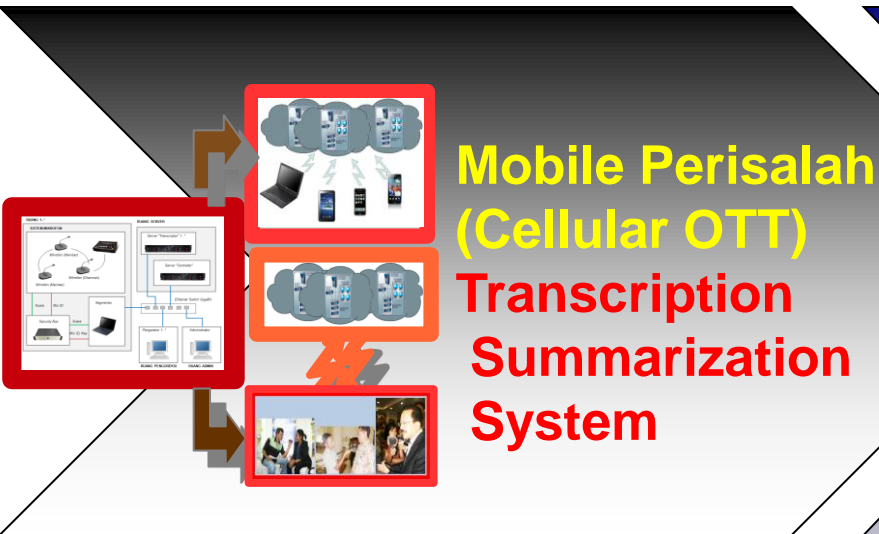
### Linguistic Resources

- [500,000 Word Bahasa Indonesia Parallel Corpus with Penn Treebank](#) [License](#)
- [500,000 Word Bahasa Indonesia Corpus and Parallel English Translation](#) [License](#)
- [One Million POS Tagged Corpus of Bahasa Indonesia](#) [License](#)

[↑TOP](#)



# Translation Technology Application Roadmap



**2013-2014**



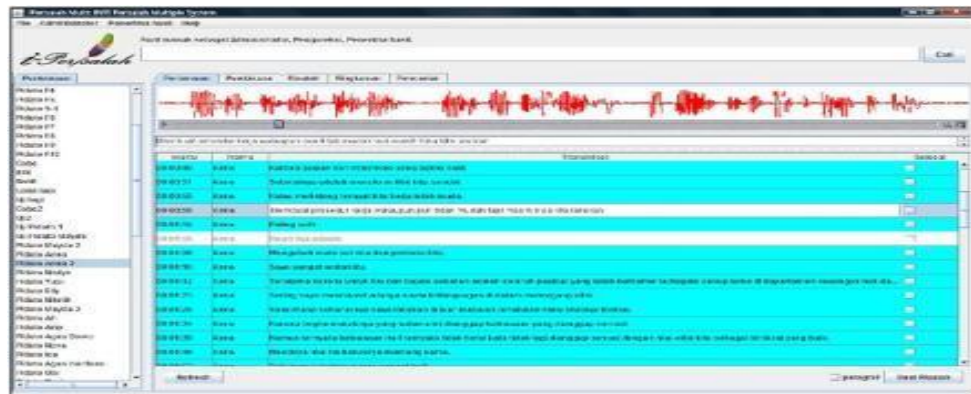
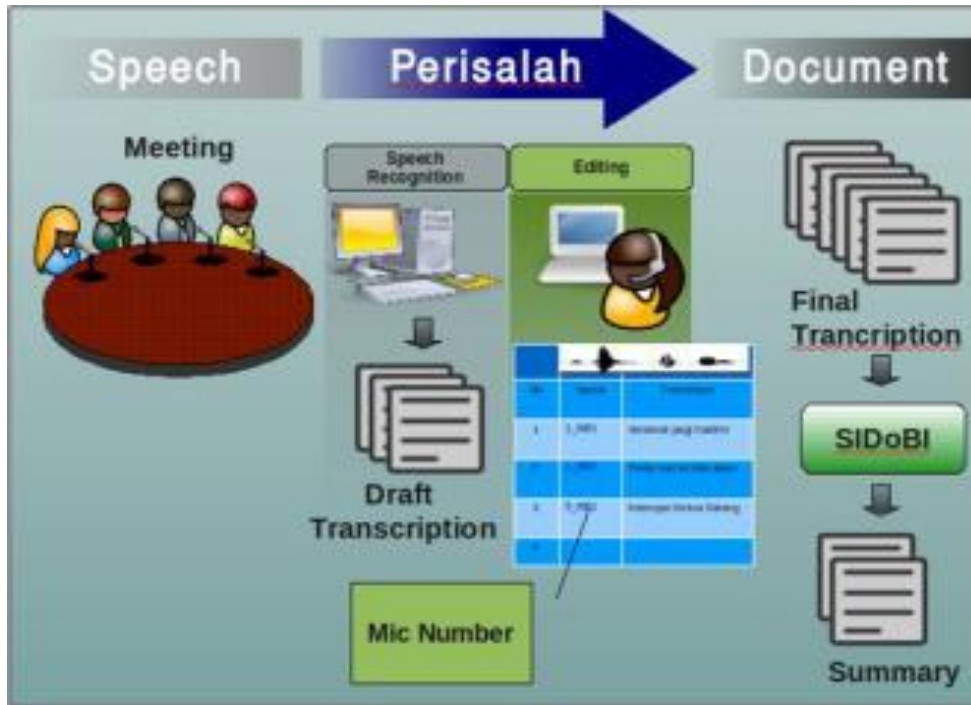
**2015-2016**



**Network-based ASEAN Language Translation Public Service (ASEAN-MT)**  
**Universal Speech Translation Advance Research (U-STAR)**  
**Speech Corpora, Parallel Text Corpora ,TTS, ASR, NLP Tools**

# BPPT - Digital Signal Processing (DSP) Laboratory

## Speech Processing



- **Language Processing Tools**

- Stemmer, POS Tagger
- Named Entity Tagger, Phrase Chunker
- Statistical Constituent Parser and Dependency Parser
- Indonesian Reference Resolution and Semantic Analysis
- Indonesian MindMap  
Generator: <http://mindmap.kataku.org>
- Game for learning Japanese-Indonesia: <http://honyaku.kataku.org>
- Purchase pattern on social media: <http://elysis.kataku.org>

- **Text Mining**

- Indonesian Question Answering System
  - Open Domain
  - Closed Domain with ontology
  - Factoid, List Factoid, Non Factoid
- Indonesian Information Extraction

- **Speech**

- Indonesian Automatic Speech Recognizer
- Indonesian Speech Synthesizer
- Preliminary Research on Indonesian TTS based on “Unit Selection” approach
- Rebuild of Indonesian Diphone Database for Diphone Concatenation based Indonesian TTS
- Building Indonesian TTS based on MARY platform
- Improvement of Indonesian Prosody
- Indonesian Syllable TTS for special purpose application

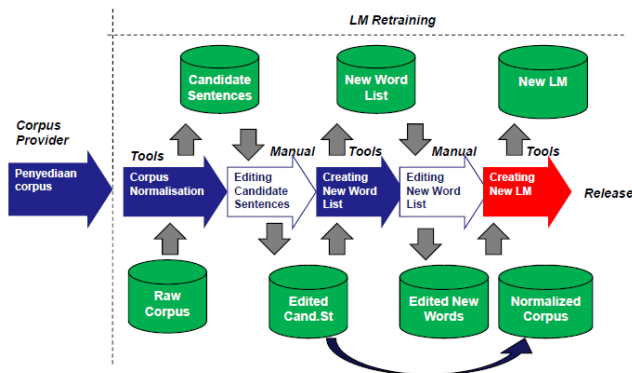
# Corpus Development 2014



Institut Teknologi Bandung



- Perisalah v.3 Speech Corpora
- Perisalah v.3 POS-Tagged Corpus (with Indonesian Language Development Center)
- Corpus Management System



- Indonesian POS Tagged Corpus
- **Extended Word Clustering Algorithm**
- Indonesian Question Answering Using of Indonesian
- TTS for blind operator who work in Call Center
- Indonesian Named Entity Tagged Corpus
- Regional Language Lexical Database
- Indonesian-Japanese Parallel Corpus

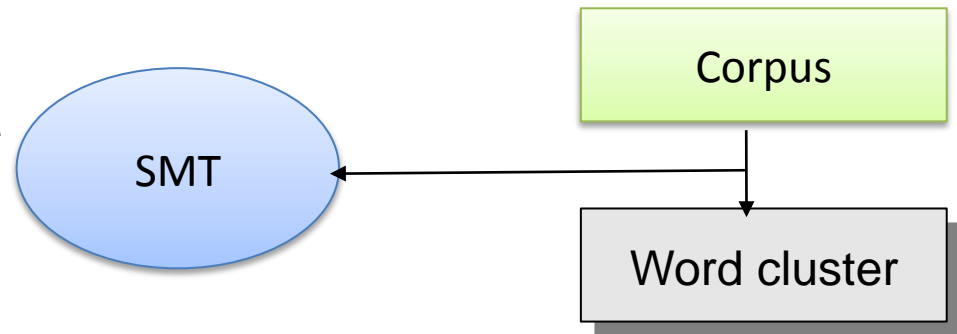


# Indonesian ASEAN-MT: Improving Statistical Machine Translation with POS Tagged Corpus

Sujaini et.al, Comparison of POS Tagset for improving English-Indonesian SMT, O-COCOSDA, 2014

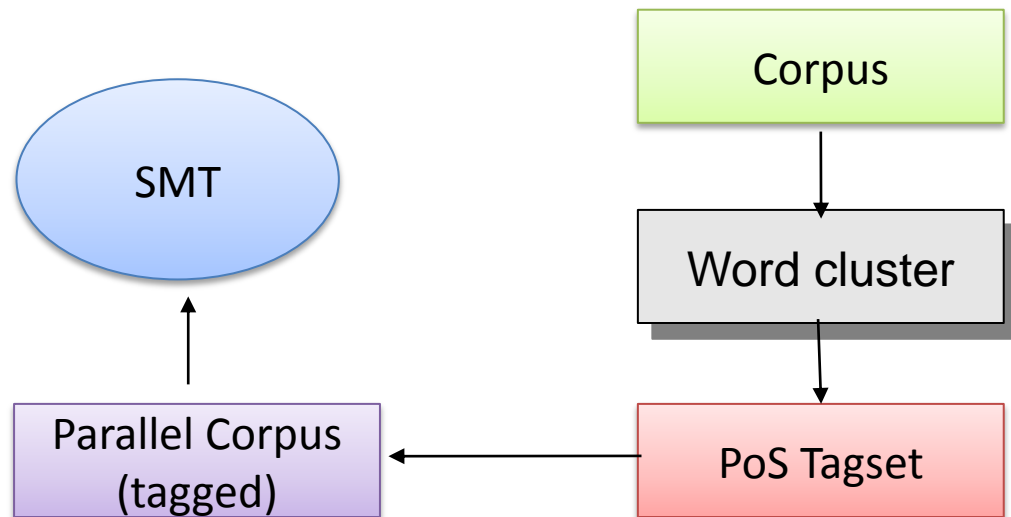
## 1. Algorithmic improvement

Extended WSB vs WSB  
vs MKCLS

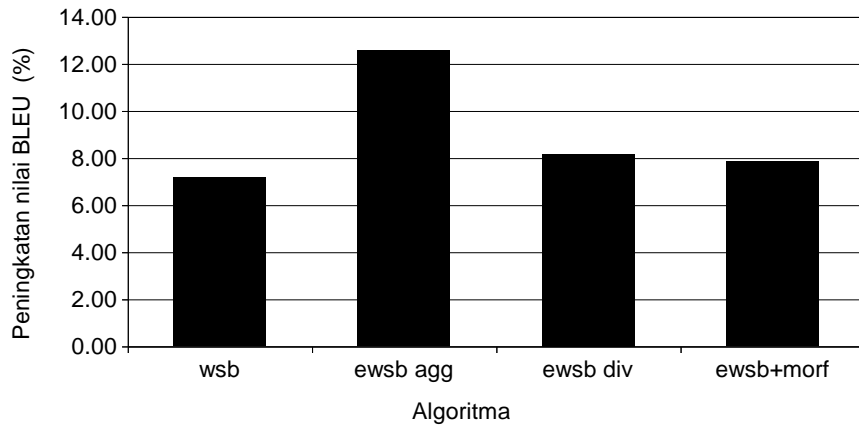


## 2. POS Tagset

Computed POS  
vs Grammar POS  
vs without POS



# Indonesian ASEAN-MT: Improving Statistical Machine Translation with POS Tagged Corpus



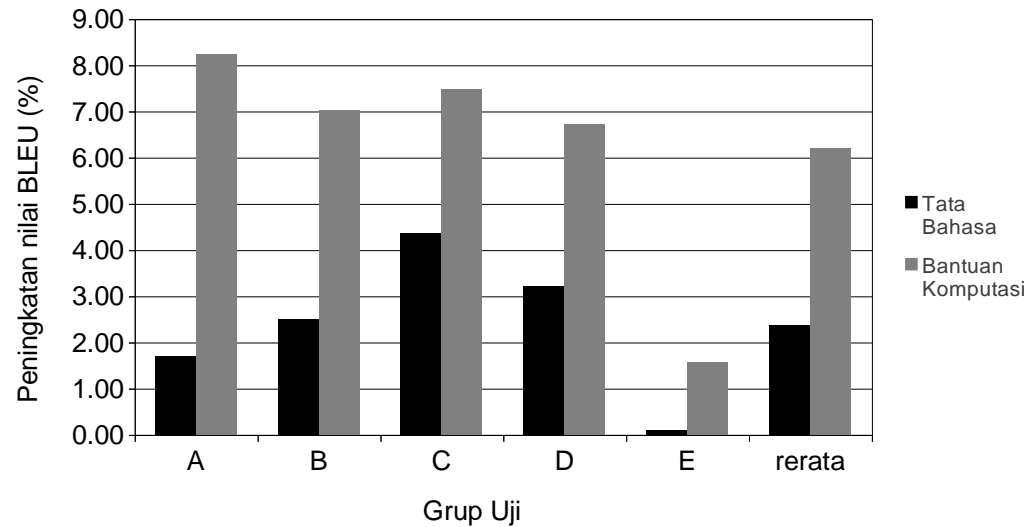
## Experiment on WSB Algorithm

Improving BLEU score using WSB Algorithm against mkcls

Base value mkcls = 38,31%

## Improving SMT with POS

BLEU score of SMT with 6 test groups



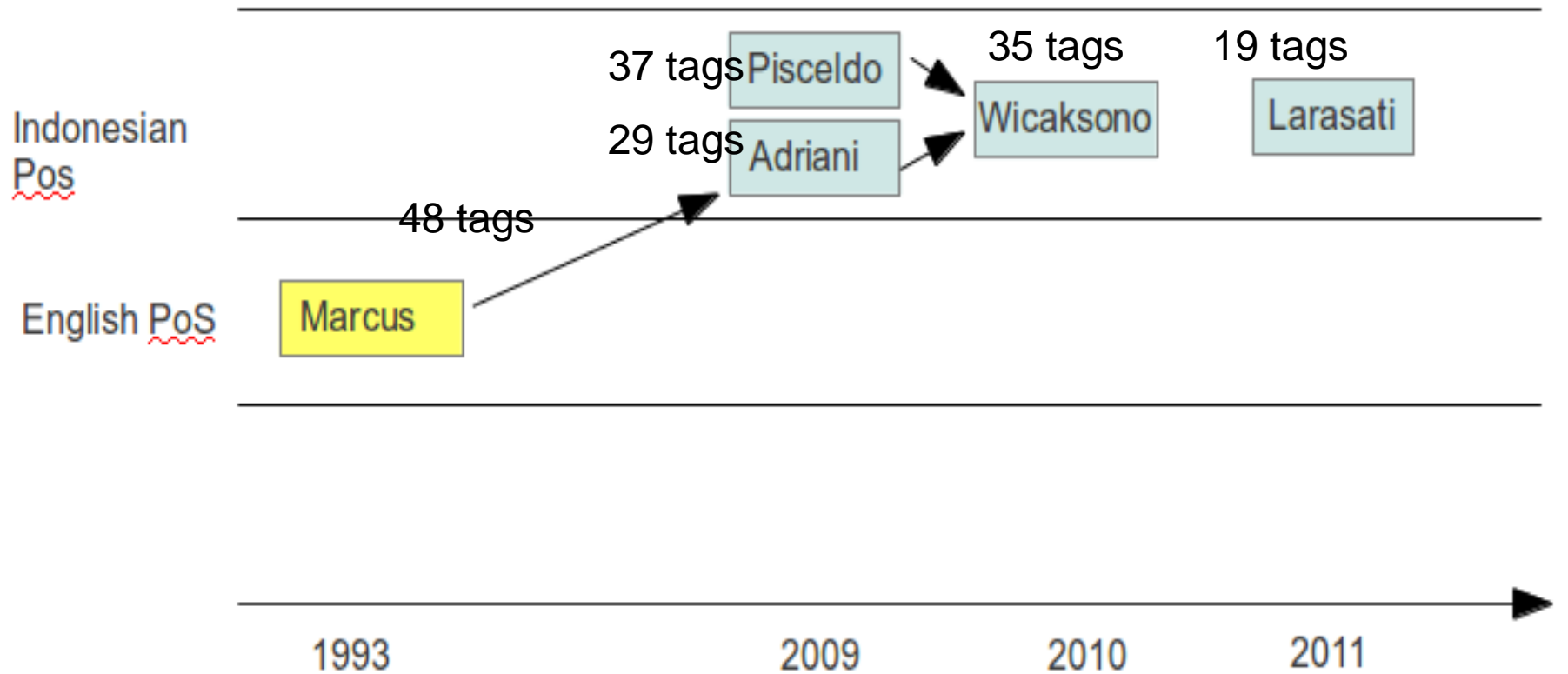
**EXTENDED WORD SIMILARITY BASED CLUSTERING  
ON UNSUPERVISED POS INDUCTION TO IMPROVE  
ENGLISH-INDONESIAN  
STATISTICAL MACHINE TRANSLATION**

# State of the Art

## Factored based SMT

No	Peneliti	Tahun	Model	Bahasa	Peningkatan nilai BLEU (%)
1	Koehn dan Hieu Hoang	2007	PoS	English–German	0,61
	Koehn dan Hieu Hoang	2007	PoS % morfologis	English–Spanish	3,59
2	Bojar	2007	morfologis	English-Czech	7,75
3	Youssef dkk.	2009	PoS	English–Arabic	4,92
4	Razavian dkk.	2010	Suffix LM	English–Iraqi	5,06
	Razavian dkk.	2010	Suffix LM	Spanish-English	0,95
	Razavian dkk.	2010	Suffix LM	Arabic-English	2,44
5	Wuebker dkk.	2013	WC	French-German	1,40
	Wuebker dkk.	2013	WC	German-English	0,30

# PoS Tagset for Bahasa Indonesia





# Objectives

- We present the unsupervised Part-of-Speech (PoS) induction algorithm to improve translations quality on statistical machine translation.
- The proposed algorithm is an extension of the algorithm Word-Similarity-Based (WSB) clustering.
- In the clustering, the similarity between words is measured by its grammatical relation with other words. The grammatical relation is represented as the n-gram relation.
- We extend the WSB clustering by take into account for the previous words in measuring the grammatical relation. The clustering results are then used in the English-Indonesia statistical machine translation.

## EWSB

$$I(t, w_1, r, w_2) = \log \frac{\text{Cnt}(t, w_1, r, w_2) \cdot \text{Cnt}(t, *, r, *)}{\text{Cnt}(t, w_1, r, *) \cdot \text{Cnt}(t, *, r, w_2)}$$

$$S(w_1, w_2) = \frac{\sum_{(t, r, w) \in T, (w_1) \cap T(w_2)} [I(t, w_1, r, w) \cdot I(t, w_2, r, w)]}{\sum_{(t, r, w) \in T, (w_1)} I(t, w_1, r, w) + \sum_{(t, r, w) \in T, (w_2)} I(t, w_2, r, w)}$$

$$\text{sim}(C_1, C_2) = \frac{1}{N_1 * N_2} \sum_{w_1 \in C_1} \sum_{w_2 \in C_2} \text{sim}(w_1, w_2) + \frac{\lambda}{N_1 + N_2}$$

# Experiments

The experiments were conducted using seven sets of 2,000 English-Indonesian parallel sentences as the training data for the machine translation and 300 sentences as the testing data.

The machine translation tools used in this research are Moses as decoding tool, SRILM as language model processor, and GIZA++ as phrase translation model processor. BLEU (Bilingual Evaluation Understudy) score is used to evaluate the quality of translation result

There were four clustering algorithms employed in the experiment: mkcls (word clustering provided in Moses), WSB by Jeff, EWSB-Agglomerative and EWSB-Divisive.

# The BLEU score

TABLE I ACCURACY OF MACHINE TRANSLATION

<b>Corpus</b>	<b>MKCLS(%)</b>	<b>WSB(%)</b>	<b>EWSB A(%)</b>	<b>EWSB D (%)</b>
A	68.13	66.66	69.95	68.69
B	69.69	69.65	70.39	70.80
C	64.28	67.35	67.97	67.35
D	34.82	36.84	36.66	35.57
E	28.12	28.08	28.22	28.62
F	38.26	38.68	38.56	37.99
G	77.36	78.70	78.60	79.30
Average	67.37	67.89	69.44	68.95

Notes :

EWSB A = EWSB Agglomerative

EWSB D = EWSB Divisive



# Conclusion

Based on Indonesian language characteristics, we extended the word similarity based clustering to enhance the quality of English-Indonesian machine translation. Using 14,000 English-Indonesian sentence pair as the training data and 300 sentences as the testing data, the experimental result showed that our extension gave higher BLEU score compared to the original WSB clustering, and it increased the translation accuracy 2.07%.

# Asean Economic Community (AEC)



<http://www.aec2008.itg.govt.com>

[asean@itg.govt.com](mailto:asean@itg.govt.com)

Harigato Gozaimasu

# THANK YOU