ASEAN IVO

ICT Virtual Organization of ASEAN Institutes and NICT (ASEAN IVO)

Final Project Report (Executive Summary) Form

I. Title of Proposed Project:

Open Collaboration for Developing and Using Asian Language Treebank

II. Project Leader:

Full name: Masao Utiyama
Institution: National Institute of Information and Communications Technology, Japan
Address: 3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289 Japan
Phone: +81-774-98-6343
E-mail: mutiyama@nict.go.jp

III. Project Members:

Name	Position/Degree	Department, Institution,Country	Email Address
Hammam Riza	Deputy Chairman IT,Energy and Material / PhD	BPPT, Indonesia	hammam.riza@bppt.go.id
Aw Ai Ti	Unit Head, Human Language Technologies / Ms.	HLT, I2R, Singapore	aaiti@i2r.a-star.edu.sg
Luong Chi Mai	Assoc. Prof. / PhD	Multimedia Human-Machine Language Technology, IOIT, Vietnam	lcmai@ioit.ac.vn
Sethserey Sam	Vice President of Research / PhD	NIPTICT, Cambodia	sethserey.sam@niptict.edu.kh
Khin Mar Soe	Professor / PhD	NLP Lab,UCSY, Myanmar	khinmarsoe@ucsy.edu.mm
Masao Utiyama	Executive Researcher / PhD	UCRI, NICT, Japan	mutiyama@nict.go.jp
Thepchai Supnithi	Research Team Leader/ Principal Researcher	NECTEC, Thailand	Thepchai.Supnithi@nectec.or.th
Ria A. Sagum	Assoc. Prof. / CCIS Faculty Researcher	PUP, Philippines	rasagum@pup.edu.ph

IV. Total Amount (US$):

87,000USD

V. Duration (6-36 Months):

36 months starting from April 1st, 2016

VI. Executive Summary

NLP is one of the core technologies in ICT. This is because the contents of information are conveyed by natural languages. The state-of-the-art technologies in NLP are based on treebanks. A treebank is a linguistic knowledge representation of natural language texts.

The main problem of the creation of a treebank is that it needs a lot of linguistic knowledge for the language. Each language treebank needs each language expertise. In particular, there has been no publicly available POS-tagged and constituency tree corpora for most of Asian languages before our project.

This background has made us propose this project for developing Asian Language Treebank (ALT). The objective of ALT is developing a parallel treebank for Asian languages. The benefits of ALT to the society is immense. ALT will accelerate research of NLP for Asian languages, such as Indonesian, Vietnamese, Japanese, Khmer, Laos, Malay, Myanmar, Philippine, Thai, and so on. This will result in the better communication in the ASEAN region and the world.

ASEAN IVO is an ideal organization for developing ALT, because it consists of top-level NLP research institutes for Asian languages. Without ASEAN IVO, it would be impossible to corporate and cover main Asian languages for building treebanks.

One of the characteristics of ALT is it uses the parallel sentences for creating treebanks. This means that each sentence will be translated into several languages and be annotated with several languages. In this project, BPPT, I2R, IOIT, NIPTICT, UCSY, NECTEC, PUP and NICT have developed ALT for Indonesian, Malay, Vietnamese, Khmer, Myanmar, Thai, Filipino and Japanese languages, respectively. (NICT has also developed English ALT). Those different language treebanks have been built from the translated Wikinews sentences (about 20,000 sentences).

NICT built an annotation server for developing ALT using the English texts above. The member institutes used this ALT annotation server or their original tools for building ALT. The annotation server helps all steps of the development; translation, word segmentation, word alignment, POS tagging, and syntax annotation.

After three years of the ALT project, we have developed treebanks and software, which are available from the project website:

http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/index.html

In contrast, usual treebanks are very hard to share, because the original texts in usual treebanks have strict copyrights, which do not let researchers share treebanks.

ALT has already been used by other researchers. For example, “The 5th Workshop on Asian Translation” (WAT 2018) used ALT project data in their translation task.

The ALT project data is the only data that cover wide ASEAN languages. Consequently, NLP researchers working with ASEAN languages will use the ALT project data. In other words, our project contributes the development of NLP and ICT in ASEAN.