2016 Conference of The Oriental Chapter of International Committee
for Coordination and Standardization of Speech Databases and Assessment Technique (O-COCOSDA)
26-28 October 2016, Bali, Indonesia

Introduction of the Asian Language Treebank

Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah
Badan Pengkajian dan Penerapan Teknologi, Indonesia

Aw Ai Ti, Sharifah Mahani Aljunied
Institute for Infocomm Research, Singapore

[†]Luong Chi Mai, [†]Vu Tat Thang, [‡]Nguyen Phuong Thái
[†]Institute of Information Technology, Vietnam
[‡]University of Engineering of Technology, Vietnam

Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng
National Institute of Posts, Telecoms and ICT, Cambodia

Khin Mar Soe, Khin Thandar Nwet
University of Computer Studies, Yangon, Myanmar

Masao Utiyama, Chenchen Ding
National Institute of Information and Communication Technology, Japan

Abstract— **This paper introduces our project for developing Asian Language Treebank (ALT). The ALT project aims to advance the state-of-the-art Asian natural language processing (NLP) techniques through the open collaboration for developing and using ALT. The project is a joint effort of six institutes for making a parallel treebank for seven languages: English, Indonesian, Japanese, Khmer, Malay, Myanmar, and Vietnamese. The process of building ALT began with sampling about 20,000 sentences from English Wikinews, and then these sentences were translated into the other six languages. ALT will have word segmentation, part-of-speech tags, syntactic analysis annotations, together with word alignment links among these languages.**

Keywords- Asian languages; parallel corpus; treebank

I. INTRODUCTION

Natural language processing (NLP) needs raw and annotated resources for research and development. Especially, the present state-of-the-art NLP techniques rely on treebanks, i.e., linguistically annotated corpora of natural language texts. The basic linguistic annotations in treebanks include word segmentation, part-of-speech (POS) tags, and syntactic information. The annotated data are used to produce NLP tools, such as word segmenters, POS taggers, and syntactic parsers. Those NLP tools are necessary to NLP applications, such as machine translation (MT), Web search engines, and summarization. Almost all NLP researches and tools are based on treebanks in a broad sense.

The main problem in creating a treebank is that it needs high-level linguistic competency for the language of interest. As a result, existing treebanks are limited in their sizes, annotation types, and languages. In particular, no publicly available POS tagged and syntactically annotated corpus is available for most of the Asian languages.

This background has motivated us to launch a project for building Asian Language Treebank (ALT). The purpose of the ALT project is to advance the state-of-the-art Asian NLP technologies through the open collaboration for developing and using ALT. At the initial stage of developing ALT, seven languages are targeted: English, Indonesian, Japanese, Khmer, Malay, Myanmar (Burmese), and Vietnamese. One of the characteristics of ALT is that it uses parallel sentences for creating treebanks. English sentences were first sampled from Wikinews, and then translated into the other six languages. Finally, the sentences in each language are linguistically annotated. Word alignment links between different languages are also provided using English as the pivot language.

Our approach is unique that no other treebank for Asian languages has this property. It has both scientific and engineering benefits. For the scientific benefits, we can compare various NLP techniques as applied to various languages on the same ground, which will reveal how NLP techniques can be applicable to different languages. For the engineering benefits, we can easily transfer the NLP techniques available in resource rich languages (for example, English and Japanese) into other languages, which will advance Asian NLP.

The structure of this paper is as follows. First, we survey Asian NLP resources for the six languages other than English. The purpose of this survey is two-fold: it provides the overview of current Asian NLP resources, so that one can readily access them, and it highlights the contributions of ALT to the community. Then, we describe the details of ALT. Finally, we discuss the future of ALT, especially the availability of ALT to the NLP community.

II. SURVEY OF ASIAN RESOURCE

ALT is a parallel treebank, which has word segmentation, POS tags, and syntax annotations with word alignment links. Hence, we mainly focus on NLP resources and tools related to these annotations. We are also aware of the importance of parallel corpora in developing NLP tools and applications, especially machine translation. Therefore, parallel corpora are also mentioned. For example, the 3rd Workshop on Asian Transla-

978-1-5090-3516-8/16/$31.00 ©2016 IEEE

tion focuses on machine translation among Asian languages and makes use of Chinese, English, Hindi, Indonesian, Japanese and Korean parallel texts [1].

A. Indonesian NLP resources

Indonesia has a long history on computational linguistics research. In 1987-1994, BPPT has been involved in CICC-Japan Project for Multilingual Machine Translation System (MMTS) based on inter-lingual approach. This project has an important role in the research and development of Indonesian computational linguistics. Many basic resources and tools were produced, such as Indonesian Basic Dictionary that contain morphological, syntactic, and semantic information (24,000 word roots), Indonesian Gazetteer, Indonesian Monolingual Corpus (1 million words), Indonesian Analysis System, etc. and those resources has become a foundation for other projects. Furthermore, Indonesia have also involved in the UNL project started in 1997.

Indonesia has joined PAN localization in 2010-2012 through the collaboration of BPPT and University of Indonesia with Pakistan, China, Nepal, Sri Lanka, Laos, Cambodia, Bhutan and Thailand. Indonesia also joined international research collaboration with A-STAR, U-STAR Consortium, Asian WordNet, ASEAN MT, and others. These projects have make use of the early Bahasa Indonesian resources, which resulted in other resources and improving existing resources. Up to now, BPPT has collected around 10 million sentences of Indonesian monolingual corpus, 250 thousand sentences of English-Indonesian parallel corpus, and enhanced the Indonesian Electronic Dictionary. Indonesian POS Tagger and Corpus Management System are also created to support various projects.

Two leading universities in Indonesia, Indonesian University and Bandung Institute of Technology, have also been very active in the field of computational linguistics. Many resources and tools have been developed such as statistical POS tagger, Indonesian WordNet, Indonesian POS Tagged Corpus, Indonesian Named Entity Tagged Corpus, Indonesian Syntactical Tree Tagged Corpus, Indonesian Dependency Tree Tagged Corpus, Indonesian-English Parallel Corpus and Indonesian-Japanese Parallel Corpus. Other universities have begun to be involved in this field such as Gadjah Mada University, Syarif Hidayatullah State Islamic University, Telkom University, Surabaya Institute of Technology, Electronics Engineering Polytechnic Institute of Surabaya, Trisakti University, Bina Nusantara University, Gunadarma University, Indonesia University of Education Petra University, etc. which has been organized into the new Indonesian Association of Computational Linguistics (INACL).

B. Japanese NLP resources

Japanese is one of the languages without spaces between words. Consequently, word segmentation is the first step for the most of NLP applications. Note that POS tagging is usually conducted together with word segmentation, because these are closely related and jointly determined in Japanese. Then, dependency analysis is often performed, while phrase structure analysis is considered as another option.

One of the well-known treebanks in Japanese is Kyoto University Text Corpus [2], which is *"a text corpus that is manually annotated with various linguistic information. It consists of approximately 40,000 sentences from Mainichi newspaper in 1995 with morphological and syntactic annotations. Out of these sentences, 5,000 sentences are annotated with predicate-argument structures including zero anaphora and co-references.*" However, "*note that this package does not include original sentences but include only annotation information. To recover the complete annotated corpus, it is necessary to obtain the Mainichi 1995 CD-ROM.*" Consequently, it is not easy to use this corpus outside Japan. Based on this corpus, JUMAN, a morphological analyzer, and KNP, a dependency, case structure, and anaphora analyzer, have been developed and made available to the public [3].

Other word segmentation or morphological analysis tools are ChaSen, MeCab and KyTea [4]. As stated above, in Japanese, the word segmentation and POS tagging is tightly coupled. Consequently, the above tools conduct both analyses simultaneously. However, because their models are trained on other corpora that are not available to the public, it is hard to compare a new method or tool with them in the same setting.

Besides KNP, CaboCha [5] and J.DepP [6] also conduct Japanese dependency analysis. Note that CaboCha is used with MeCab to conduct Japanese morphological and dependency analyses, while J.DepP alone carries out both analyses. J.DepP can be trained on the freely-available Kyoto-University and NTT Blog Corpus [7], comprising about 4,000 sentences. Another parser is Ckylark [8], a latent-annotated probabilistic context-free grammar parser.

A manually annotated word alignment data is available as "Kyoto Free Translation Task Japanese-English Alignment Data" [9]. This corpus is a derivative of "Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles" [10], which is also a derivative of Japanese Wikipedia. Because the copyright of Wikipedia is Creative Commons Attribution-Share-Alike 3.0 License, the translation of Japanese Wikipedia articles into English and the word alignment between Japanese and English are also available under that license. This is an example where the license can foster the development of NLP resources.

C. Khmer NLP resources

As an Asian Language, Cambodian language, which is referred to as Khmer, is also written continuously without using spaces between words. Thus, Khmer word segmenter is the most important tool for Khmer NLP. The recent paper about Khmer Word Segmention [13] demonstrated the highest accuracy up to 0.99 of F-score. The tool was trained with around 97,000 of manually annotated sentences. The corpus was collected from various domains. This corpus is currently not open to the public.

Because Khmer language is a low resource language in NLP, collecting and making data is time consuming. Public POS tagged data is only found at PAN localization's website [14]. The corpus is about 3,000 manually annotated sentences,

which were used to train for POS tagger for Khmer language. The tool is published by PAN localization [15].

Currently, NIPTICT has been manually annotating POS tagged data about 30,000 sentences, which is segmented by Khmer Word Segmenter [13]. The data is not open to the public.

Other useful textual resources are government documents such as law and declaration. Most documents are officially written in three languages, Khmer, English, and French, in parallel. These files are open to the public as PDF format only, and can be found freely on the government websites (for example, the constitution of the kingdom of Cambodia [16]). These are confidential documents to share as editable format. So, it is another time consuming task for converting them into text format. It needs good OCR and human corrections.

D. Malay NLP resources

Bahasa Malaysia is a language spoken mostly by the Malay people in Peninsular Malaysia, Brunei, Singapore and other areas in Southeast Asia. The language is considered as *low resource* for both NLP and MT tasks as there are not many annotated data or bilingual texts available. As a result, most works on Malay language processing is linguistically and/or heuristically driven.

One of the earliest works on Malay language processing was Othman's Stemming Algorithm [17] for information retrieval, which has 121 rules for prefixes, suffixes and infixes. Muhamad [18] further improved the work by Fatimah [19] using Rule Frequency Order with 426 rules and a root word dictionary. Similarly most Part-of-Speech tagging for Malay language was predicted based on the word morphology. The first Malay tagger developed by Mohamed [20] combined a hidden Markov model with morphological information to identify the POS of unknown words in the Malay language. Juhaida [21] further improved the POS tagging for unknown words using Decision Tree and Nearest Neighbor. However, the data used for the training algorithm was not mentioned

A Malay corpus database was created under Malaysia's Language and Literature Agency (DBP) for the storage and retrieval of large texts. Chuah [22] reported that the database has over 71,444,326 words over different sub-domains and is augmented with Malay Text Analysis system. Aw [23] developed a rule-based system that performs morphological analysis and syntactic parsing on the Malay text and use the linguistic information from the parsing trees to reorder the word, transfer the syntactic structure and generate the final English text. Vu [24] used morphological information of both Malay and English language to extract bilingual terms from Malay-English comparable corpora.

E. Myanmar NLP resources

Myanmar language does not use spaces between words. Therefore, word segmentation of Myanmar sentences is needed. The current Myanmar word segmenter achieves a precision of 97.9% [25]. The segmenter uses the annotated corpus containing 60,000 sentences currently. This corpus is not available to the public.

Myanmar POS tagging and parsing are also required for NLP tasks such as machine translation, name entity recognition and text summarization. Myanmar POS tagging and parsing are still under research at NLP Lab, University of Computer Study, Yangon.

Since Myanmar language is still low-resource for NLP tasks, the very first work for NLP tasks is creating the corpus. NLP Lab has created a corpus of 20,000 sentences (travel domain) in order to implement "the Network based ASEAN Languages Translation Public Service" (www.aseanmt.org) collaborating with the ASEAN countries. The lab has also prepared Myanmar part in ALT collaborating with NICT [12].

Myanmar-English-Myanmar bilingual Dictionary has been developed to obtain a structurally similar WordNet for Myanmar language. The dictionary consists of 30,000 words currently [26].

F. Vietnamese NLP resources

Like many other Asian languages such as Chinese, Japanese and Thai, there is no word delimiter in Vietnamese. The space is a syllable delimiter but not a word delimiter, so a Vietnamese sentence can often be segmented in many ways. Second, Vietnamese is an isolating language in which words do not change their forms according to their grammatical function in a sentence. Third, the Vietnamese syntax conforms to the subject verb-object (SVO) word order.

Vietnamese treebank (VTB) [27] is a branch of a national project which aims to develop basic resources and tools for Vietnamese Language and Speech Processing (VLSP) [28]. In addition to treebank, the VLSP project also develops other text-processing resources and tools including a Vietnamese machine readable dictionary, an English-Vietnamese parallel corpus, a word segmenter, a POS tagger, a chunker, and a syntactic parser. The Vietnamese treebank (VTB) and other resources and tools developed by the VLSP project have been posted on the VLSP web page [29] since 2010. The main goal was to build a corpus of 70,000 word segmented sentences, 20,000 POS tagged sentences, and 10,000 syntactic trees. As a resource, raw texts from the news domain focusing on social and political topics were extracted. VTB construction is an iterative process involving three phases: annotation, guideline revision, and tool upgrade. Annotators were supported by automatic-labeling tools, which are based on statistical machine learning methods, for sentence pre-processing. A tree editor is developed to support manual annotation. As a result, an annotation agreement of around 90% was achieved.

For word segmentation, vnTokenizer - a highly accurate segmenter (over 95%) which uses a hybrid approach to automatically tokenize Vietnamese text is used. This approach combines a finite-state automata technique, regular expression parsing, and a maximal-matching strategy augmented by statistical methods that resolve ambiguities of segmentation [30]. Also JVnTagger - a POS tagger based on Conditional Random Fields and Maximum Entropy is developed. The training data

size is 10,000 sentences. Experiments with 5-fold cross validation showed that F1 scores for CRFs and Maxent were 90.40% and 91.03%, respectively [28]. Another tool used was a syntactic parser based on lexicalized probabilistic context-free Grammars (LPCFGs). Bikel's parser was also customized which is well designed and easy to adapt to new languages. An entropy-based method is investigated for detecting errors and inconsistencies in the word-segmented and POS-tagged parts of VTB data.

Vietnam team joined international research collaboration with A-STAR, U-STAR Consortium, ASEAN MT, and others.

III. Proposal of Asian Language Tree Bank (ALT)

A. *Overview of ALT*

The ALT project was initiated by the National Institute of Information and Communications Technology, Japan (NICT) in 2014. NICT started to build Japanese and English ALT and worked with the University of Computer Studies, Yangon, Myanmar (UCSY) to build Myanmar ALT in 2014. Then, the Badan Pengkajian dan Penerapan Teknologi, Indonesia (BPPT), the Institute for Infocomm Research, Singapore (I^2R), the Institute of Information Technology, Vietnam (IOIT), and the National Institute of Posts, Telecoms and ICT, Cambodia (NIPTICT) joined to make ALT for Indonesian, Malay, Vietnamese, and Khmer in 2015 [11].

ALT comprises about 20,000 sentences originally sampled from the English Wikinews in 2014. These were already translated into the other six languages. These will be annotated with word segmentation, POS tags, and syntax information, in addition to the word alignment information. An example of parallel sentences of English, Indonesian, Japanese, Khmer, Malay, Myanmar, and Vietnamese is shown below.

[EN] Italy have defeated Portugal 31-5 in Pool C of the 2007 Rugby World Cup at Parc des Princes, Paris, France.
[ID] Italia berhasil mengalahkan Portugal 31-5 di grup C dalam Piala Dunia Rugby 2007 di Parc des Princes, Paris, Perancis.
[JA] フランスのパリ、パルク・デ・プランスで行われた2007年ラグビーワールドカップのプールCで、イタリアは31対5でポルトガルを下した。
[KH] អ៊ីតាលីបានឈ្នះលើព័រទុយហ្គាល់ 31-5 ក្នុងបូ៌លCនៃពិធីប្រកួតពានរង្វាន់ពិភពលោកនៃកីឡាបាល់ឱបឆ្នាំ2007ដែលប្រព្រឹត្តនៅប៉ាសឌេសប្រិន ក្រុងប៉ារីស ប៉ារាំង។
[MS] Itali telah mengalahkan Portugal 31-5 dalam Pool C pada Piala Dunia Ragbi 2007 di Parc des Princes, Paris, Perancis.
[MY] ပြင်သစ်နိုင်ငံ ပါရီမြို့ ပါ့ဒက်စ် ပရင့်စက် ၌ ၂၀၀၇ခုနှစ် ရပ်ဘီ ကမ္ဘာ ဖလား တွင် အီတလီ သည် ပေါ်တူဂီ ကို ၃၁-၅ ဂိုး ဖြင့် ရေကူးကန် စီ တွင် ရှုံးနိမ့်သွားပါသည် ။
[VI] Ý đã đánh bại Bồ Đào Nha với tỉ số 31-5 ở Bảng C Giải vô địch Rugby thế giới 2007 tại Parc des Princes, Pari, Pháp.

It is easily understood that development of ALT needs international collaboration of NLP researchers, because ALT needs high level of linguistic expertise for each language.

B. *ALT annotation Web server*

NICT has built an annotation server for developing ALT. It has been used to build Myanmar and Japanese ALT. The annotation server helps all steps of the development; translation, word segmentation, word alignment, POS tagging, and syntax annotation.

The data are represented in an XML format. Thus, the server is flexible enough to accept different POS tags for different languages, for example. It is also possible to mix automatic and manual annotations. That is, at first, automatic analysis is conducted by NLP tools. Then, it can be manually corrected using the server.

This server also supports users, groups and task management and has a multilingual help system. The administrator can assign particular tasks to users considering their linguistic expertise (such as assigning only translation tasks etc.). If the user is assigned all tasks he/she can continuously and efficiently work through all processes for the given sentence. For word alignment, POS tagging and syntax annotation, user input can be performed using only the mouse. Fig. 1 shows the user interface for word alignment annotation between an English sentence and the corresponding translated Myanmar sentence, and Fig. 2 shows the user interface for syntax annotation, or constituency tree building [12].

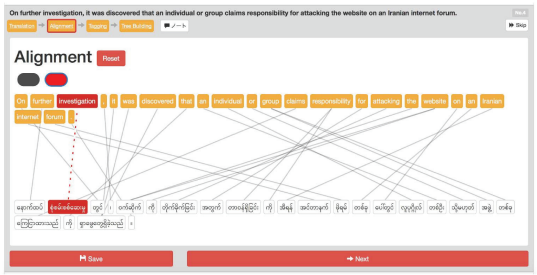

Fig. 1. Word alignment interface

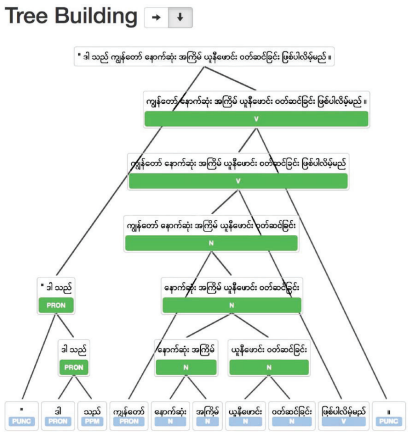

Fig. 2. Tree building interface

C. Current progress

The translation of the English sentences into the six languages has been completed. In addition to the six languages, the English sentences have also been translated into additional three languages, Laos, Thai, and Philippine. However, we currently have no institute to annotate them with linguistic information for ALT of these three languages.

Myanmar parts of ALT has been constructed within a year [12]. We will use it for Myanmar NLP. We are checking and improving the Myanmar data. The construction of Japanese and English ALT are ongoing. The construction of other ALT has begun in 2015.

IV. Discussion and Conclusion

We make ALT to support the advancement of Asian NLP. In order to reach this goal, the license of ALT is very important.

We selected English Wikinews as the original texts, because its license is Creative Commons Attribution 2.5 License. This is a very open license under which we can use it even commercially, provided that we give appropriate credit. This is in contrast with many treebanks whose original texts have stricter copyrights.

We are now discussing on the licensing of ALT. The license should help us promote the open collaboration for developing and using ALT to the community.

This paper has discussed the ALT project being conducted by the six institutes. ALT is intended to accelerate NLP development in low resource Asian languages. The corpus comprises about 20,000 sentences from the news domain consisting of Asian language translations from a shared English source text annotated with word segmentation, word alignment, POS tags, and syntax trees. ALT includes English, Indonesian, Japanese, Khmer, Malay, Myanmar and Vietnamese in the short term, and extend to other languages in the long term through collaboration with international research organizations.

The outcomes of the ALT project will be published at http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/

Acknowledgment

This work is partly supported by the ASEAN IVO Project "Open Collaboration for Developing and Using Asian Language Treebank".

References

[1] The 3rd Workshop on Asian Translation, http://orchid.kuee.kyoto-u.ac.jp/WAT/

[2] Daisuke Kawahara, Sadao Kurohashi and Koiti Hasida. "Construction of a Japanese Relevance-tagged Corpus," In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002), pp.2008-2013, 2002.

[3] JUMAN and KNP are available from Kurohasi and Kawahara Lab. http://nlp.ist.i.kyoto-u.ac.jp/EN/

[4] ChaSen is available at http://chasen-legacy.osdn.jp/. MeCab is available at http://taku910.github.io/mecab/. KyTea is available at http://www.phontron.com/kytea/.

[5] T. Kudo and Y. Matsumoto. Japanese Dependency Analysis using Cascaded Chunking. Proc. CoNLL, pp. 63--69. 2002. https://taku910.github.io/cabocha/

[6] J.DepP is available at http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jdepp/

[7] Kyoto-University and NTT Blog Corpus is available at http://nlp.ist.i.kyoto-u.ac.jp/kuntt/ (in Japanese)

[8] Ckylark is available at https://github.com/odashi/Ckylark

[9] Graham Neubig, "The Kyoto Free Translation Task," http://www.phontron.com/kftt, 2011.

[10] "Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles" is available at http://alaginrc.nict.go.jp/WikiCorpus/index_E.html

[11] Masao Utiyama, Eiichiro Sumita. "Open collaboration for developing and using Asian Language Treebank (ALT)," ASEAN IVO Forum 2015

[12] Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew Finch and Eiichiro Sumita. "Introducing the Asian Language Treebank (ALT)," LREC, 2016.

[13] Vichet Chea, Ye Kyaw Thu, Chenchen Ding, Masao Utiyama, Andrew Finch and Eiichiro Sumita,Khmer Word Segmentation Using Conditional Random Fields, In Khmer Natural Language Processing 2015, December 4, 2015

[14] PAN Localization's POS Tagged Corpus (http://www.panl10n.net/english/Outputs%20Phase%202/CCs/Cambodia/MoEYS/Software/2009/KhmerCorpus.zip)

[15] PAN Localization's POS Tagger (http://www.panl10n.net/english/Outputs%20Phase%202/CCs/Cambodia/MoEYS/Papers/2008/KhmerPOSTaggingV1.0.pdf)

[16] The Constitution Of the Kingdom of Cambodia in English (http://www.ccc.gov.kh/english/basic_text/Constitution%20of%20the%20Kingdom%20of%20Cambodia.pdf), France (http://www.ccc.gov.kh/french/textes_basic/Constitution%20du%20Royaume%20du%20Cambodge.pdf) and Khmer (http://www.ccc.gov.kh/khmer/basic_text/Constitution.pdf)

[17] Othman. "Pengakar perkataan melayu untuk sistem capaian dokumen," vol. MSc Thesis: National University of Malaysia, 1993.

[18] Muhamad Taufik Abdullah, Fatimah Ahmad, Ramlan Mahmod and Tengku Mohd Tengku Sembok. "Rules Frequency Order Stemmer for Malay Language," IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.2, February 2009.

[19] A. Fatimah. "A Malay Language Document Retrieval System: An Experimental Approach and Analysis," in Information Science, vol. PhD. Bangi: Universiti Kebangsaan Malaysia, 1995, pp. 322.

[20] H. Mohamed, N. Omar, and M. J. Ab Aziz. "Statistical Malay Part-of-Speech (POS) Tagger using Hidden Markov Approach," 2011 International Conference on Semantic Technology and Information Retrieval, 2011, pp. 231–236.

[21] Juhaida Abu Bakar1, Khairuddin Omar2, Mohammad Faidzul Nasrudin3, Mohd Zamri Murah. "MORPHOLOGY ANALYSIS IN MALAY POS PREDICTION," Proceeding of the International Conference on Artificial Intelligence in Computer Science and ICT(AICS 2013), 25 -26 November 2013, Langkawi, MALAYSIA.

[22] Choy-Kim Chuah, Zaharin Yusoff. "Computational Linguistics at Universiti Sains Malaysia," LREC 2002.

[23] AiTi Aw, Sharifah Mahani Aljunied, Lianhau Lee and Haizhou Li. "Piramid: Bahasa Indonesia and Bahasa Malaysia Translation System Enhanced through Comparable Corpora," TCAST 2009.

[24] Thuy Vu, AiTi Aw, Min Zhang. "Feature-Based Method for Document Alignment in Comparable News Corpora," European Chapter of the Association for Computational Linguistics (EACL), Athens, 2009.

[25] Chenchen Ding, Ye Kyaw Thu, Masao Utiyama, Eiichiro Sumita, "Word Segmentation for Burmese (Myanmar)", ACM Transactions on Asian and Low-Resource Language Information Processing, Vol. 15 Issue 4, Article No. 22, 2016.

[26] Myanmar-English-Manmar dictionary is available at www.nlpresearch-ucsy.edu.mm/NLP_UCSY/myanengmyan.html

[27] Phuong-Thai Nguyen, Anh-Cuong Le, Tu-Bao Ho, Van-Hiep Nguyen. Vietnamese Treebank Construction and Entropy-based Error Detection, Language Resources and Evaluation, 49 (3). pp. 487-519. ISSN 1574-0218, 2015.

[28] Luong Chi Mai, Ho Tu Bao, Project Report on National Project on Vietnamese Language and Speech Processing, Hanoi, March 2010.

[29] http://vlsp.vietlp.org:8080/demo/

[30] L. H. Phuong, N. T. M. Huyen, R. Azim, H. T. Vinh. A Hybrid Approach to Word Segmentation of Vietnamese Texts. In Proceedings of the 2nd International Conference on Language and Automata Theory and Applications. Springer LNCS 5196, Tarragona, Spain, 2008.