Open Collaboration for Developing and Using Asian Language Treebank Wrap-up of 2016-2018

Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah (BPPT)
Aw Ai Ti, Nabilah Binte Md Johan (I2R)
Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thái (IOIT/UET)
Soky Kak, Kea Sorn, Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng (NIPTICT)
Khin Mar Soe, Khin Thandar Nwet (UCSY)
Masao Utiyama, Chenchen Ding (NICT)
Chai Wutiwiwatchai, Thepchai Supnithi, Prachya Boonkwan(NECTEC)
Ria A. Sagum, Michael B. dela Fuente, Carlo G. Inovero, Janelle Kyra A. Sagum (PUP)

Objective of Asian Language Treebank

- Provide Asian Language Treebank for free for research

→Cover many under-resourced Asian languages

→Facilitate the rapid development of Asian NLP

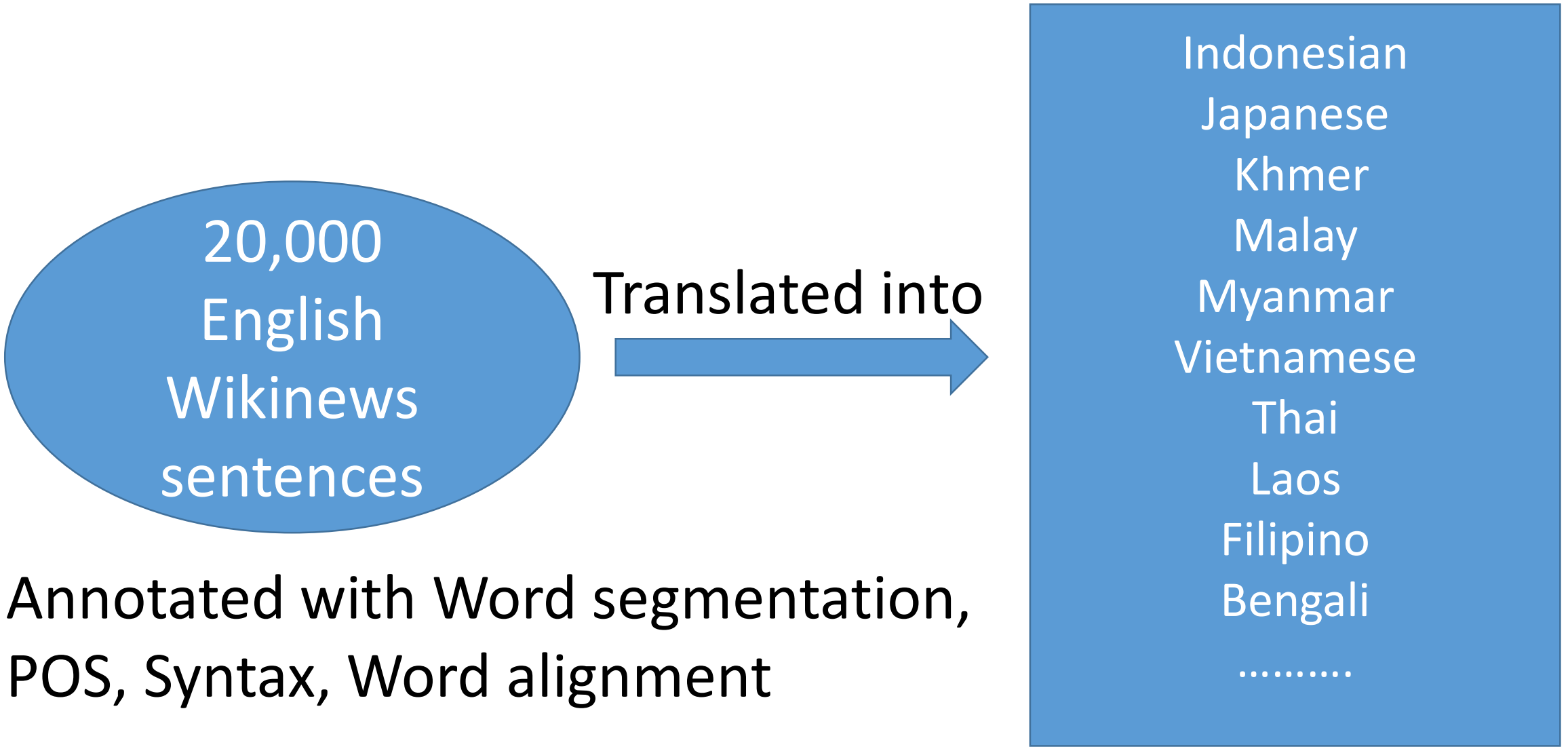
→Provide the common ground for comparison/evaluation of Asian NLP

→We will release ALT with a

Creative Commons

Attribution-NonCommercial-ShareAlike

What will be the Asian Language Treebank (ALT)



Annotated with Word segmentation,
POS, Syntax, Word alignment

Samples (en, id, ja, km, ms, my, vi, th, fli)

- Italy have defeated Portugal 31-5 in Pool C of the 2007 Rugby World Cup at Parc des Princes, Paris, France.
- Italia berhasil mengalahkan Portugal 31-5 di grup C dalam Piala Dunia Rugby 2007 di Parc des Princes, Paris, Perancis.
- 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。
- អ៊ីតាលីបានឈ្នះលើព័រទុយហ្គាល់ 31-5 ក្នុងប៉ូលCនៃពីធីប្រកួតពានរង្វាន់ពិភពលោកនៃកីឡាបាល់ឱបឆ្នាំ2007ដែលប្រព្រឹត្តនៅប៉ាសឌេសប្រីន ក្រុងប៉ារីស បារាំង។
- Itali telah mengalahkan Portugal 31-5 dalam Pool C pada Piala Dunia Ragbi 2007 di Parc des Princes, Paris, Perancis.
- ပြင်သစ်နိုင်ငံ ပါရီမြို့ ပါဒက်စ် မရင့်စက် ၌ ၂၀၀၇ခုနှစ် ရပ်ဘီ ကမ္ဘာ ဖလား တွင် အီတလီ သည် ပေါ်တူဂီ ကို ၃၁-၅ ဂိုး ဖြင့် ရေကူးကန် စီ တွင် ရှုံးနိမ့်သွားဖြီသည် ။
- 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.
- อิตาลีได้เอาชนะโปรตุเกสด้วยคะแนน31ต่อ5 ในกลุ่มc ของการแข่งขันรักบี้เวิลด์คัพปี2007 ที่สนามปาร์กเดแพร็งส์ ที่กรุงปารีส ประเทศฝรั่งเศส
- Natalo ng Italya ang Portugal sa puntos na 31-5 sa Grupong C noong 2007 sa Pandaigdigang laro ng Ragbi sa Parc des Princes, Paris, France.

Project Goal

- NICT will develop and release the parallel corpus for ALT
- Each member institute shall develop and release ALT for each language
- Each member institute shall decide the amount of ALT, which will be developed and released by that institute
- ALT will be used for research and development on Asian NLP

Results

- Meetings hosted by NIPTICT, BBPT, UCSY, and NECTEC
- ALT resources are available at the project page

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

- Corporation with U-STAR

  ALT parallel corpus has been used for U-STAR

  Khmer SMT has been released to U-STAR
- 12 papers. One paper got a best paper award. Another paper was published at the most competitive international conference.
- Myanmar ALT is used in a machine translation workshop, improving MT over a well-known online machine translation service.

Indonesian ALT Results by BPPT

2016

- Indonesian POS Tagger
- Simple Syntax Tree Builder
- Indonesian POS Tagset
- Indonesian Syntactic Tagset

2017

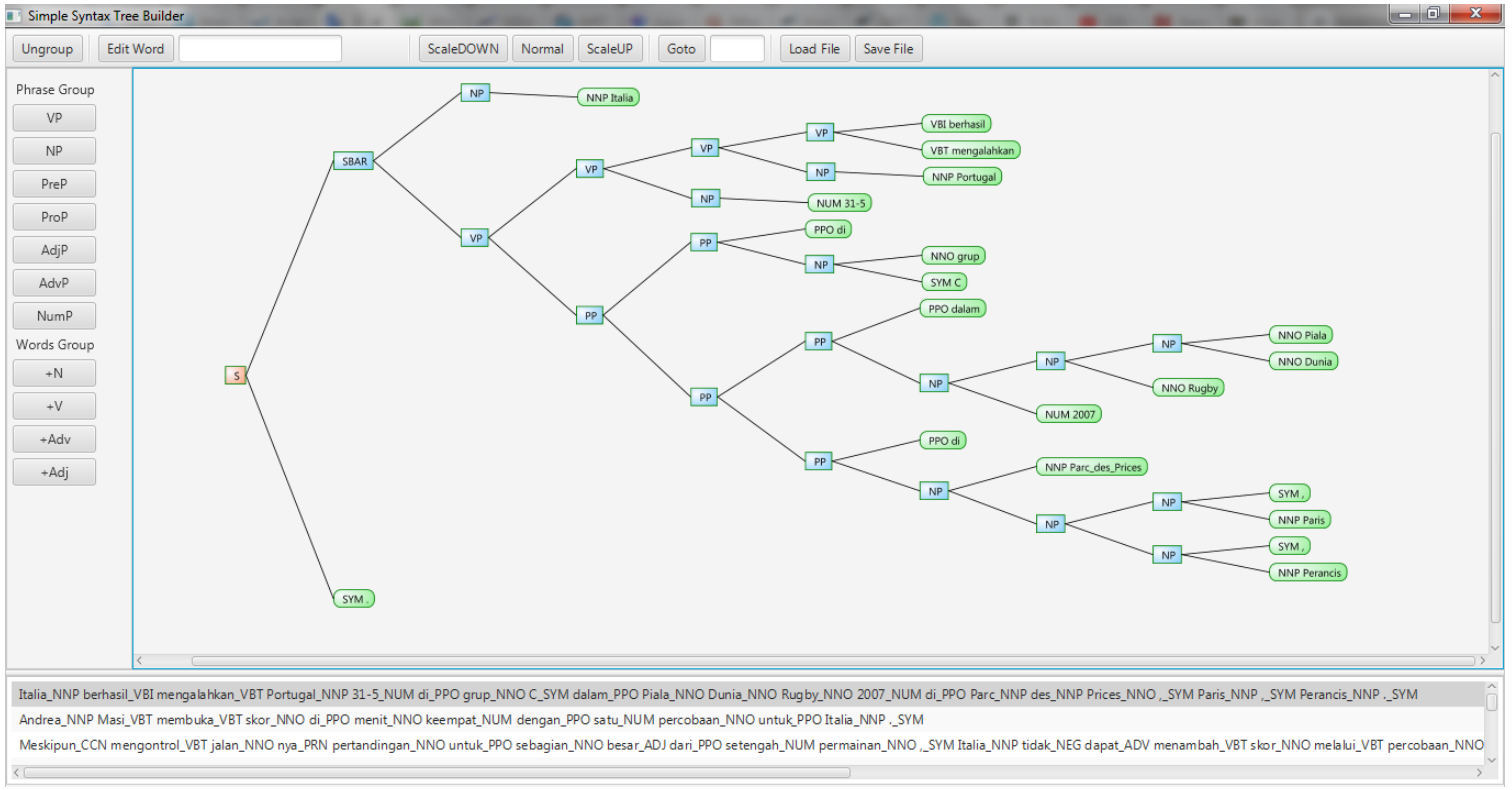
- Re-Translation EnWN to IdWN : 20 K sentences
- IdWN Tagging : 20 K Sentences
- Indonesian Syntax Tree : 20 K Sentences
- Word alignment between EnWN and IdWN using MGIZA automatically : 20 K Sentences

2018

- Re-Translation EnWN to IdWN : 10 K sentences
- Re-tagging sentences : 10 K Sentences
- Rebuild Tree : 10 K Sentences -→ October 2018
- Re-Processing word alignment between EnWN and IdWN using MGIZA and Check manually : 10 k Sentences → October 2018

Simple Syntax Tree Builder

Resulting Tree



- Syntactic Tree text output :

(S (SBAR (NP (NNP Italia))(VP (VP (VP (VP (VBI berhasil)(VBT mengalahkan))(NP (NNP Portugal)))(NP (NUM 31-5)))(PP (PP (PPO di)(NP (NNO grup)(SYM C)))(PP (PP (PPO dalam)(NP (NP (NP (NNO Piala)(NNO Dunia))(NNO Rugby))(NUM 2007)))(PP (PPO di)(NP (NNP Parc_des_Prices)(NP (NP (SYM ,)(NNP Paris))(NP (SYM ,)(NNP Perancis)))))))))(SYM .))

Malay ALT by I2R

1. Word segmentation and word alignment for 20k parallel sentences have been done automatically in August 2016.
2. The automatic aligned sentences are checked manually by linguists who know both languages.
3. The verified and aligned sentences are uploaded to ALT server in April 2017.
4. The 20k Malay sentences are automatically tagged and verified manually.
5. Propose the Malay dependency relations categories and annotation scheme.

English/Japanese ALT by NICT

- 20,000 sentences were tokenized and parsed
- The parsing style mostly follows the Penn Treebank style
- These are available to the public from the Project Web site

(S (S (BASENP (NNP Italy)) (VP (VBP have) (VP (VP (VP (VBN defeated) (BASENP (NNP Portugal))) (ADVP (RB 31-5))) (PP (IN in) (NP (BASENP (NNP Pool) (NNP C)) (PP (IN of) (NP (BASENP (DT the) (NN 2007) (NNP Rugby) (NNP World) (NNP Cup)) (PP (IN at) (NP (BASENP (NNP Parc) (FW des) (NNP Princes)) (COMMA ,) (BASENP (NNP Paris) (COMMA ,) (NNP France))))))))))) (PERIOD .))

(S (S (PP (NP (PP (NP (S-REL-NSBJ (VP (PP (NP (PP (BASENP (NNP フランス)) (IN の)) (NP (BASENP (NNP パリ)) (COMMA 、) (BASENP (NNP パルク) (NNP ・) (NNP デ) (NNP ・) (NNP プランス)))) (IN で)) (VP (VP (VBO 行わ) (VP (VBV れ))) (VP (MD た))))) (BASENP (NNP ２００７) (NNP 年) (NNP ラグビー) (NNP ワールドカップ))) (IN の)) (BASENP (NN プール) (NN Ｃ))) (IN で)) (COMMA 、) (S (PP-SBJ (BASENP (NNP イタリア)) (IN は)) (VP (PP (BASENP (NN ３１) (CC 対) (NN ５)) (IN で)) (VP (PP-OBJ (BASENP (NNP ポルトガル)) (IN を)) (VP (VBV 下し) (VP (MD た))))))) (PERIOD 。))

Named Entity Tags based on OntoNotes

- 23 types of tags
- Person: the President (Kennedy)
- Job Title: (American President) Bush
- Non human creature: (Flipper) the dolphin
- Nationality: an (American) publishing group

(S (S (BASENP (PDT All)(DT the)(NNS suspects))(VP (VBP are)(NP (BASENP (JJ male)(NE-NRP (NNPS Finns)))(VP (VBG residing)(PP (IN in)(BASENP (JJ southern)(NE-GPE (NNP Finland))))))))(PERIOD .))

Myanmar ALT by UCSY with NICT

- Finished and released resources at the project Web site
  - Refined word alignment and syntactic tree
  - Improved tokenization and POS tagging
  - Get recommendation from Myanmar Language Commission
- Resources under Cleaning
  - Syntactic tree data
  - Planned up to March, 2019

Khmer ALT by NIPTICT with NICT

- Annotation Guidelines for Khmer
  - Released on ALT home page
  - Updated along with the data preparing
  - A temporary stable version
- Final outcome of FY 2018
  - Word segmentation : 20,106 sentences
  - POS tagging : 20,106 sentences
  - Word segmentation : 9,000 sentences

Orthographic Errors

- កណ្តាល → ក ណ ◌្ដ ◌ា ល gone dal (correct pronunciation)
- កណ្តាល → ក ណ ◌្ត ◌ា ល gone tal

1. ◌្ដ → ◌្ក

- បន្ត → ប ន ◌្ដ bone dor (correct pronunciation)
- បន្ត → ប ន ◌្ត bone tor

2. ◌្រ ◌្ (con.) → ◌្ (con.) ◌្រ

- ត្រី → ត ន ◌្ត ◌្រ ◌ី (correct)
- ត្រី → ត ន ◌្រ ◌្ត ◌ី
- ត្រី → ត ន ◌្រ ◌ី ◌្ត

3. (d. vowel) ◌្ (con.) → ◌្ (con.)(d. vowel)

Vietnamese ALT by UET && IOIT

- Difficulties in Annotation
  - Isolating property
  - Training annotators
  - Long sentences
- Results
  - Word segmentation 10,000 sentences
  - POS tagging 7,000 sentences
  - Syntax annotation 4,000 sentences
  - Word alignment 3,000 sentences within FY 2018

Vietnamese language

- There is no word separator

Cô ấy$_{she}$ giữ gìn$_{takecare}$ **sắc đẹp**$_{beauty}$.
She takes care of her beauty.

Bức$_{picture}$ này$_{this}$ màu **sắc**$_{color}$ **đẹp**$_{beautiful}$ hơn$_{more}$.
The color of this picture is more beautiful.

- Isolating property: Vietnamese does not use the morphological marking of case, gender, number, …

1	**Anh ấy**$_{he}$ **ra**$_{come}$ Hà Nội$_{Hanoi}$.
1e	**He comes** to Hanoi.
2	Giám đốc$_{manager}$ bảo$_{ask}$ **anh ấy**$_{him}$ **ra**$_{come}$ Hà Nội$_{Hanoi}$.
2e	The manager asks **him to come** to Hanoi.

- Word order typology: SVO

Thai ALT by NECTEC (Just beginning)

Tag names		Description	Example
1.	ADJP	Adjective phrase	ขั้นตอนต่อไป (next steps)
2.	ADVP	Adverb phrase	อย่างแข็งขัน (staunchly); อย่างดี (completely); น่ากังวลใจอย่างหนัก (gravely concerning)
3.	NP	Noun phrase	การสั่งห้ามการสูบบุหรี่ในพื้นที่สาธารณะ (a ban on smoking in enclosed public spaces); คะแนนเสียงส่วนน้อย 33 เสียง (the slim margin of 33 votes)
4.	PP	Prepositional phrase	ในพื้นที่สาธารณะ (in enclosed public spaces); ในการเลือกตั้งปี 2001 (in the 2001 election)
5.	S	Simple declarative clause	พวกตนจะได้รับการพิจารณาเหมาะสมที่ศาล (They are properly dealth with by the courts.)
6.	SBAR	Subordinate clause	นักการเมือง "ผู้มีจิตวิญญาณและความกล้าหาญ" (a "spirited and courageous" politician); วิทยาลัยที่เกี่ยวข้อง (the college involved)
7.	VP	Verb phrase	ถูกจับ (was arrested); ทำหน้าที่เป็นโฆษกพรรคด้านสุขภาพ (Acting as the party's health spokesman)
8.	CONJP	The combinations of a comma and a CC, and a colon/ semicolon and a CC	หลังจาก (after); แล้วก็ (and); นอกจากนั้น (then); แม้แต่ (even)

Thai tree complexity

- Long sentence with many words
  - The longest sentence of Thai trees is composed of 98 words!

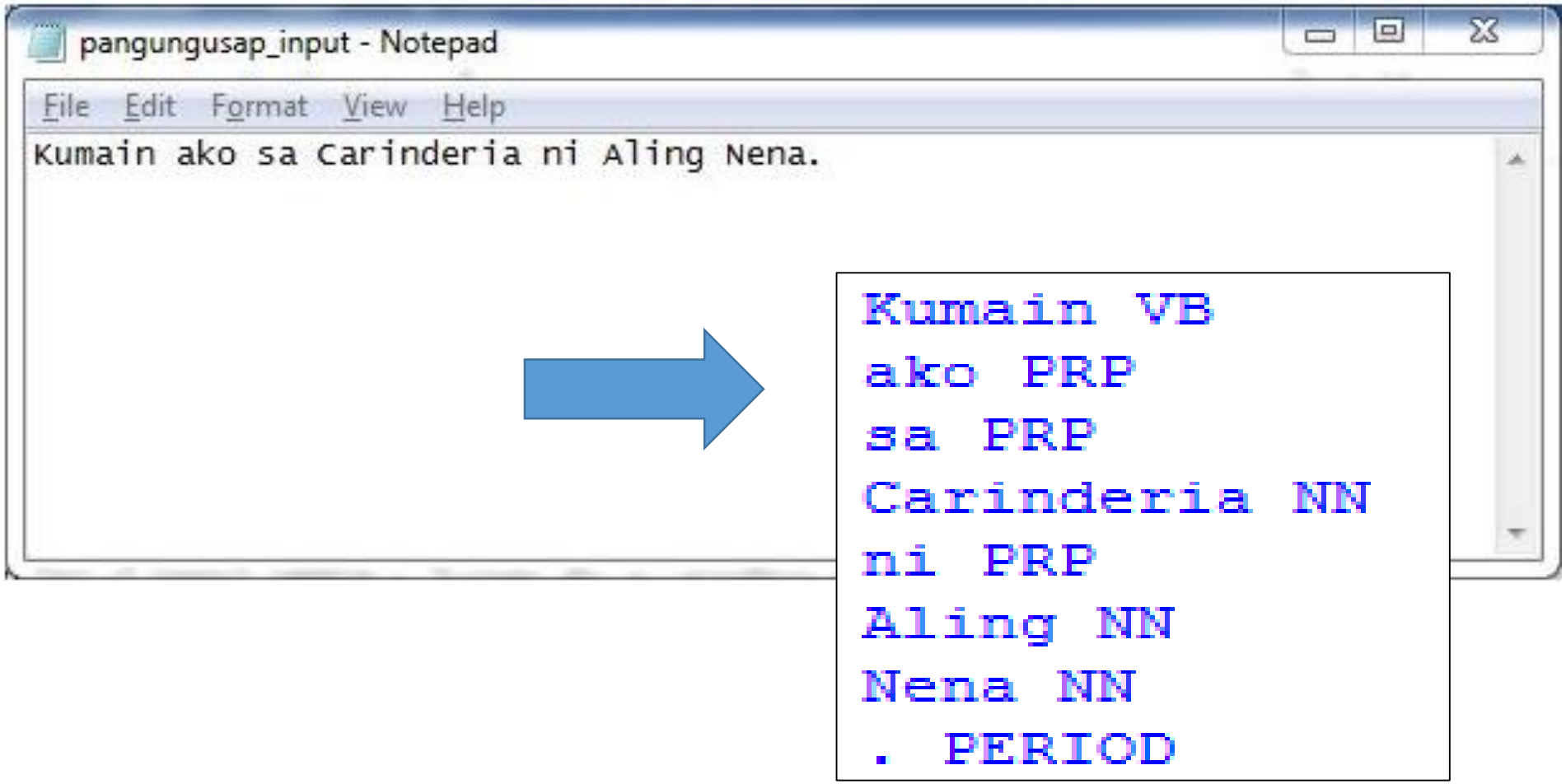
วุฒิสมาชิก|Joe| |Ludwig|(|สังกัด|พรรคแรงงาน|, รัฐควีนส์แลนด์|)|ได้|ซักถาม|รัฐมนตรี|กระทรวงยุติธรรม|และ|วุฒิสมาชิก|ฝ่าย|ศุลกากร|Chris| |Ellison ใน|รายการ|ควิสชั่น ไทม์|ว่า|เขา|ยืนยัน|คำ|กล่าวอ้าง|ของ|เขา|เมื่อ|วาน|นี้|ที่|ว่า การ|ค้น|พบ|เรือ|เจียน|เส็ง|เป็น|แสดง|ให้|เห็น|ว่า|รัฐบาล|ออสเตรเลีย|ได้|เข้า|"|ตรวจตรา|ทาง|อากาศ|และ|ทาง|ทะเล|เพื่อ|สกัด|เรือ|ใน|สภาวะ|การณ์|เช่น|นี้ ซึ่ง|ภารกิจ|ดัง|กล่าว|ได้|สำเร็จ|แล้ว|" รวม|ทั้ง|ซักถาม|ว่า เหตุ|ใด|รัฐบาล|จึง|ใช้|เวลา|ถึง|สอง|สัปดาห์|ใน|การ|สกัด|เรือ|ลำ|ดัง|กล่าว|หลัง|จาก|ที่|เข้า|มา|ใน|เขต|น่านน้ำ|ออสเตรเลีย|แล้ว|

Senator Joe Ludwig (Labor, Queensland) asked of the Minister for Justice and Customs Senator Chris Ellison in Question Time whether he stood by his claim that he made yesterday that the discovery of the Jian Seng demonstrated that the Australian Government had in place "aerial and maritime surveillance to intercept a vessel in these circumstances, and that was done", and asked why it took the Government two weeks to intercept the vessel after it entered Australian waters.

Filipino by PUP (Just beginning)

- Development of Filipino ALT in contribution for the project Asian Language Treebank that can be used in translating the Filipino Language to other Asian Languages.

Sample input in Part-of-Speech tagging is “Kumain ako sa Carinderia ni Aling Nena.”



Conclusion

- Thank ASEAN IVO for the great opportunity to work together!
- The goal has been achieved.
- We have provided Asian Language Treebank for free for research
- Many research papers has been published
- ALT has been used for the development of machine translation technology

Publications

- Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Rapid Sun, Vichet Chea, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, Chenchen Ding. (2016) "Introduction of the Asian Language Treebank" Oriental COCOSDA.
- Chenchen Ding, Masao Utiyama, Eiichiro Sumita. (2016) Similar Southeast Asian Languages: Corpus-Based Case Study on Thai-Laotian and Malay-Indonesian. WAT.
- Gunarso Gunarso, Hammam Riza. (2016) An Overview of BPPT's Indonesian Language Resources. ALR12.

Publications

- Hsu Myat Mo , KhinThandar Nwet , Khin Mar Soe, "CRF-Based Named Entity Recognition for Myanmar Language", The Proceedings of the International Conference on Genetic and Evolutionary Computing (ICGEC2016), pages 204–211, Fuzhou, China, November 7-9 2016

- KhinThandar Nwet , Khin Mar Soe, "Myanmar-English Machine Translation Model", The Proceedings of the International Conference on Genetic and Evolutionary Computing (ICGEC2016), pages 195-203, Fuzhou, China, November 7-9 2016

- Hnin Thu Zar Aye, Chenchen Ding, Win Pa Pa, Khin Thandar Nwet, Masao Utiyama, Eiichiro Sumita, "English-to-Myanmar Statistical Machine Translation Using a Language Model on Part-of-Speech in Decoding", The Proceedings of 15th International Conference on Computer Application(ICCA2017), pages 409-414. Yangon, Myanmar, 16-17 February

Publications

- Chenchen Ding, Vichet Chea, Masao Utiyama, Eiichiro Sumita, Sethserey Sam and Sopheap Seng.(2017) Statistical Khmer Name Romanization. PACLING. (Best Paper Award)
- Chenchen Ding, Win Pa Pa, Masao Utiyama and Eiichiro Sumita. (2017) Burmese (Myanmar) Name Romanization: A Sub-Syllabic Segmentation Scheme for Statistical Solutions. PACLING
- Chenchen Ding, Masao Utiyama and Eiichiro Sumita. (2018) Simplified Abugidas. ACL

Publications

- Agung Santosa, Asril Jarin, Made Gunawan, Teduh Uliniansyah, Gunarso, Elvira Nurfadhilah, Lyla Ruslana, Fara Ayuningtyas, Harnum Annisa, and Hammam Riza (2018) "Utilizing Indonesian Data Resources for Text-to-Speech Using End-to-End Method" O-Cocosda.
- Yi Mon Shwe Sin, Khin Mar Soe. (2018) Large Scale Myanmar to English Neural Machine Translation System. IEEE-GCCE.
- Minh-Thuan Nguyen, Van-Tan Buiy, Huy-Hien Vuz, Phuong-Thai Nguyen, Chi-Mai Luong. (2018) Enhancing the quality of Phrase-table in Statistical Machine Translation for Less-Common and Low-Resource Languages. IALP.