Open Collaboration for Developing and Using Asian Language Treebank (ALT)

NICT
Masao Utiyama, Eiichiro Sumita
What are parallel corpus and treebank?

- Parallel corpus → same sentences with different languages
- Treebank → Linguistic knowledge annotated onto texts

- Machine translation (MT) needs parallel corpus
- Natural language processing (NLP) needs treebanks to develop tools
  - Word segmentation → fundamental tool for NLP
  - Parsing → needed in advanced NLP
Why we need Asian Language Treebank (ALT)?

- **Accelerates research of NLP** for Asian languages
  - Indonesian, Vietnamese, Myanmar, Japanese, Khmer, Laos, Malay, Philippine, Thai, ....

- **No** publicly available POS-tagged and constituency tree corpora for most of Asian languages. (Though, some corpora are available for some languages)

- **No parallel corpora** among all Asian languages

- Expected members
  - NICT, BPPT, IOIT, UCSY, and other research bodies
  - NICT and UCSY have already started making parts of ALT
  - NICT, BPPT, IOIT and USCY agreed to propose the ALT project to ASEAN IVO
Current progress in FY 2015

• NICT developed an annotation server for ALT

• NICT is translating English Wikinews (460,000 words, 20,000 sentences) into Indonesian, Vietnamese, Japanese, Thai, Khmer, Laos, Malay, Philippine

• NICT is making the Japanese and English treebanks and word alignment data

• UCSY is translating the English into Myanmar and making the Myanmar treebank and word alignment data
Development Steps

• NICT provides English and translation data
• NICT provides an web-based annotation system (if needed)
• Step 1: Word segmentation and alignment
• Step 2: POS tagging
• Step 3: Tree building
<table>
<thead>
<tr>
<th>English</th>
<th>Vietnamese</th>
<th>Indonesian</th>
<th>Thai</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andrea Masi opened the scoring in the fourth minute with a try for Italy.</td>
<td>Andrea Maisi đã mở tỉ số cho Ý ở phút thứ tư với một quả try.</td>
<td>Andrea Masi membuka skor di menit keenam dengan satu try untuk Italia.</td>
<td>Andrea Masi ได้เปิดฉากทำคะแนนในนาทีที่สี่ ด้วยการทำคะแนนจากจิตาลี</td>
</tr>
<tr>
<td>Despite controlling the game for much of the first half, Italy could not score any other tries before the interval but David Bortolussi kicked three penalties to extend their lead.</td>
<td>Chiếm thế áp đảo trong hầu hết hiệp đầu nhưng Ý đã không thể ghi thêm try nào trước khi nghỉ giữa giờ, tuy nhiên David Bortolussi đã sút ba quả phạt đền kéo dài thế dẫn đầu của họ.</td>
<td>Meskipun mengontrol jalannya pertandingan untuk sebagian besar dari setengah permainan, Italia tidak dapat menambah skor melalui try lainnya sebelum istirahat, namun David Bortolussi menendang tiga penalti untuk memperlebar keunggulan mereka.</td>
<td>ทั้งที่เป็นฝ่ายคุมเกมส์ในช่วงแรกของการแข่งขัน แต่ก่อนถึงเวลาพักครึ่งนักล่าสุดก็ไม่สามารถทำคะแนนได้จิบเลย แต่ David Bortolussi ได้ทำคะแนนนำให้กับทีมด้วยการทำลูกโทษสามลูก</td>
</tr>
<tr>
<td>Portugal never gave up and David Penalva scored a try in the 33rd minute, providing their only points of the match.</td>
<td>Bồ Đào Nha chưa bao giờ từ bỏ và David Penalva đã ghi một try ở phút thứ 33, điểm duy nhất của họ trong trận đấu.</td>
<td>Portugal tidak pernah menyerah dan David Penalva mencetak skor melalui try di menit ke-33 yang menjadi skor akhir mereka.</td>
<td>โปรตุเกสไม่ยอมแพ้ แม้ David Penalva ได้ทำคะแนนล่าสุดในนาทีที่ 33 ซึ่งเป็นเพียงคะแนนเดียวของพวกเขาในการแข่งขันครั้งนี้</td>
</tr>
</tbody>
</table>
UI for Word Alignment
UI for Tagging
UI for Tree building
How we use ALT?

• Word segmentation ➔ developing Japanese word segmenter
• POS and tree ➔ developing Japanese/English parser
• Word alignment ➔ Preordering SMT
English/Chinese ➔ Japanese MT

• Parse the input English/Chinese sentence
• Reorder the English/Chinese sentence according to pre-ordering rules (automatically obtained from word alignment)
• Translate the reordered sentences

• Parsing and word alignment accuracies are crucial
FIG. 3C is a graph illustrating a simulation that includes the effects of resonance, cyclic clocks, and a change in logic current.

Pre-ordering: FIG. 3C _va1_ resonance of effects, cyclic clocks, and logic current in change _va2_ includes that simulation _va2_ illustrating graph is.

MT: 図3Cは、共振による効果、環状のクロック、および論理電流の変化を含むシミュレーションを示すグラフである。
Performance of Chinese-Japanese patent MT

- A company uses the NICT CJ SMT engine for translating the first claims of Chinese patents

![Graph showing accuracy with more than 3 times improvement]

2015/11/26
Conclusion

• Treebank and Parallel corpus are important language resources
• ALT makes parallel treebank for Asian languages
• Corpora and tools are shared by the project members
• They will be available to the public