

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

Translation examples

English	Vietnamese	Indonesian	Thai
Italy have defeated Portugal 31-5 in Pool C of the 2007 Rugby World Cup at Parc des Princes, Paris, France.	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.	Italia berhasil mengalahkan Portugal 31-5 di grup C dalam Piala Dunia Rugby 2007 di Parc des Princes, Paris, Perancis.	อิตาลีได้เอาชนะโปรตุเกสด้วยคะแนน31ต่อ5 ในกลุ่มC ของการแข่งขันรักบี้เวิลด์คัพปี2007 ที่สนามปาร์กเดแพร็งส์ ที่กรุงปารีส ประเทศฝรั่งเศส
Andrea Masi opened the scoring in the fourth minute with a try for Italy.	Andrea Maisi đã mở tỉ số cho Ý ở phút thứ tư với một quả try.	Andrea Masi membuka skor di menit keempat dengan satu try untuk Italia.	Andrea Masi ได้เปิดฉากทำคะแนนในนาทีที่สี่ ด้วยการทำคะแนนจากฝั่งอิตาลี
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.	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.	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.	ทั้งที่เป็นฝ่ายคุมเกมส่ในครั้งแรกของการแข่งขัน แต่ก่อนถึงเวลาพักครึ่งอิตาลีก็ไม่สามารถทำคะแนนได้สักเลย แต่David Bortolussi ได้ทำคะแนนนำให้กับทีมด้วยการยิงลูกโทษสามลูก
Portugal never gave up and David Penalva scored a try in the 33rd minute, providing their only points of the match.	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.	Portugal tidak pernah menyerah dan David Penalva mencetak skor melalui try di menit ke-33 yang menjadi skor akhir mereka.	โปรตุเกสไม่ละความพยายาม และDavid Penalva ได้ทำคะแนนด้วยลูกโทษในนาทีที่33 ซึ่งเป็นเพียงคะแนนเดียวของพวกเขาในการแข่งขันครั้งนี้

UI for Word Alignment

Site Title 表示言語/language: 日本語 help

タスク 掲示板 ?

ワークスペース

Task Title help 作業中 ID:290

Translation Alignment Tagging Tree Building

<http://www.google.co.jp>


task description XXX

カナダは、2014年世界女子車椅子バスケットボール選手権に勝利。 No.1

Translation → Alignment → Tagging → Tree Building ノート 1 スキップ

Alignment リセット

カナダ は 、 2014 年 世界 女子 車椅子 バスケットボール 選手権 に 勝利 。



Canada wins 2014 Women 's World Wheelchair Basketball Championship

保存 → 次の作業へ

カナダは2014年世界女子車椅子バスケットボール選手権において、トロントにあるマツミアスレチックセンターでドイツを54対50で破り、金メダルを獲得した。 No.2

Translation → Alignment → Tagging → Tree Building ノート スキップ

試合は接戦で、両チームのミスショットと転倒によって勝敗が決まった。 No.3

Translation → Alignment → Tagging → Tree Building ノート スキップ

✓ 提出する

担当者

• help Translation Alignment Tagging Tree Building

UI for Tagging

Site Title 表示言語/language: 日本語 help

タスク 掲示板 ?

ワークスペース

Task Title help 作業中 ID:290

Translation Alignment **Tagging** Tree Building

<http://www.google.co.jp>
task description XXX

カナダは、2014年世界女子車椅子バスケットボール選手権に勝利。 No.1

Translation → Alignment → **Tagging** → Tree Building ノート 1 スキップ

Tagging リセット

Canada wins 2014 NUM Women 's World Wheelchair Basketball Championship

保存

カナダは2014年世界女子車椅子バスケットボール選手権において、トロントにあるルを獲得した。 No.2

Translation → Alignment → **Tagging** → Tree Building ノート スキップ

試合は接戦で、両チームのミスショットと転倒によって勝敗が決まった。 No.3

Translation → Alignment → **Tagging** → Tree Building ノート スキップ

✓ 提出

担当者 help Translation Alignment **Tagging** Tree Building

VERB VERB description

NOUN NOUN description

PRON PRON description

ADJ ADJ description

ADP ADP description

CONJ CONJ description

DET DET description

NUM NUM description

X X description

. . description

UI for Tree building

Site Title 表示言語/language: 日本語 ヘルプ help

タスク 掲示板

ワークスペース

Task Title help 作業中

Translation Alignment Tagging Tree Building ID:290

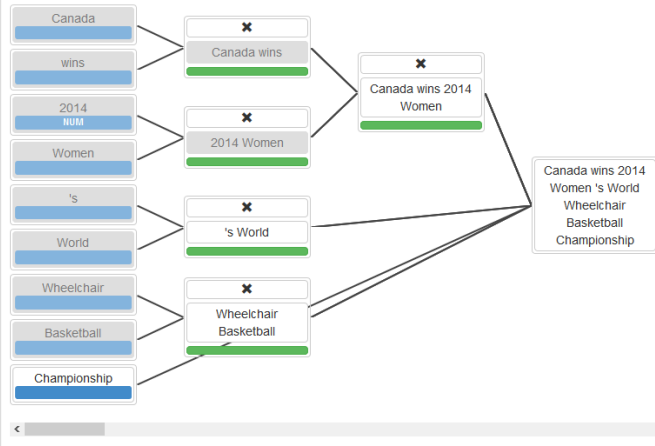
<http://www.google.co.jp>

task description XXX

カナダは、2014年世界女子車椅子バスケットボール選手権に勝利。 No.1

Translation Alignment Tagging Tree Building ノート

Tree Building リセット



Canada wins 2014 Women's World Wheelchair Basketball Championship

保存 次の作業へ

カナダは2014年世界女子車椅子バスケットボール選手権において、トロントにあるマツミアースレチックセンターでドイツを54対50で破り、金メダルを獲得した。 No.2

Translation Alignment Tagging Tree Building ノート

試合は接戦で、両チームのミスショットと転倒によって勝敗が決まった。 No.3

Translation Alignment Tagging Tree Building ノート

提出する

担当者 help Translation Alignment Tagging 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**

Example Translation

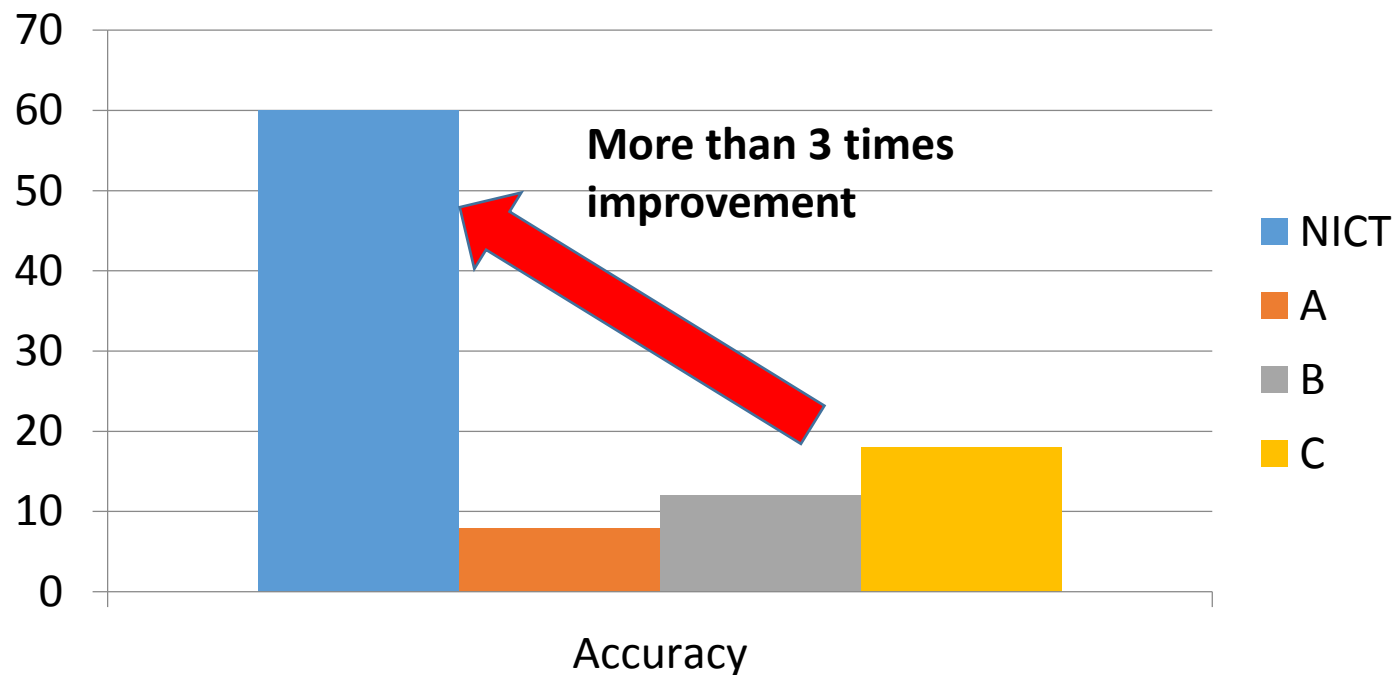
Input: FIG. 3C is a graph illustrating a simulation that includes the effects of resonance, cyclic clocks, and a change in logic current.

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



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