Parallel Aligned Treebanks at LDC: Update on Current Efforts

Ann Bies and Haejoong Lee
Linguistic Data Consortium
{bies,haejoong}@ldc.upenn.edu
Goal and Challenges

- **Goal:** Sentence-aligned parallel Treebanks

- **Challenges**
  - Desire to correct sentence/SU boundaries during TB annotation
  - Different data priorities/schedules for WA vs. TB

- **Proposal**
  - If downstream annotation can change SU boundaries, need to use WA annotation as intermediary to get token-aligned parallel TB alignment → proposal following
  - If SU boundaries are inviolable for downstream annotation, then SUs will already be parallel
Parallel Aligned Treebanks: LDC Proposal

- Unsegmented Foreign Text → Segment → Segmented Foreign Text → Translate → Segmented English Text
- Segmented Foreign Text → Treebank → Foreign TB → Token Mapping
- Segmented Foreign Text → Word Alignment → WA Annotation → Token Mapping
- Segmented English Text → Treebank → English TB → Token Mapping

Alignment Algorithm → Tree-to-tree Mapping → Foreign TB → English TB
Step 1: LDC provides sentence segmented source text to TB, WA teams

- Source text is locked, immutable
Character stream is pre-defined and locked
Step 2: Treebank, WA teams produce annotation based on LDC source text

- Provide mapping table from TB, WA tokens to strings in source text
Depiction of TB Mapping Tables

Mapping from English tokens to strings of source text

Mapping from Non-English tokens to strings of source text
Depiction of WA Mapping Tables

- Mapping from English words to strings of source text
- Mapping from non-English words to strings of source text
- WA annotation

Word

English sentence

Non-English sentence
**Step 3:** LDC runs alignment algorithm

- Incorporate mapping tables provided by TB and WA teams to create mapping between English and non-English trees

![Diagram of the alignment process](image)
Depiction of Tree-to-Tree Mapping
Progress & Milestones

- May 2008: Discussion begins in BAC committee
- Jun-Aug 2008: LDC develops, tests proposal
- Aug-Sep 2008: BBN, LDC discuss and refine proposal
- Nov 2008: Final proposal circulated to BAC group
- Dec 2008: Sample package circulated to BAC group
- Current: Waiting for feedback from BAC
- Pending BAC buy-in, plan will be implemented for all future releases
  - Full implementation requires both TB and WA annotation on same data
  - Requires TB, WA teams to use LDC source text and produce mappings
- Past data sets can be addressed as needed, over time
- Need additional site input to synchronize data set priorities for WA, TB teams