Word Alignment Update

Xuansong Li, Stephanie Strassel, Stephen Grimes
Linguistic Data Consortium
{xuansong, strassel, sgrimes}@ldc.upenn.edu

Chinese Word Alignment

- Existing annotation task emphasizes finding *minimum equivalent translation unit* and *translation rules* between source and target language
- Per site request, LDC has been working to define an enhanced annotation task with input from Niyu Ge and WA advisory committee
  - Step towards structure alignment (align symmetric structures and describe asymmetric structures) based on semantic or contextual equivalence between target and source

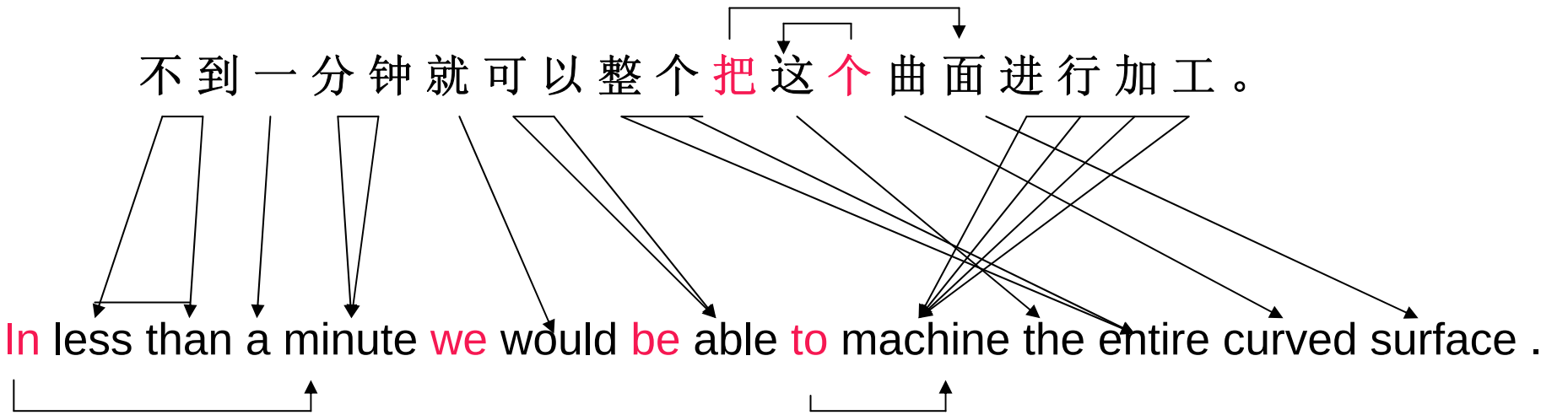
Chinese WA Tagging Task Overview

- Distinguishes three categories of links
  - Terminal links: exact match between source & target
    - Semantic, function, three types of 的 (DE) constructions (clause, modifier, possessive)
  - Composite links: semantic equivalence based on asymmetric structures
    - Grammatically or contextually-inferred
  - Empty links: one side (source or target) empty
- Adds tags for unaligned words

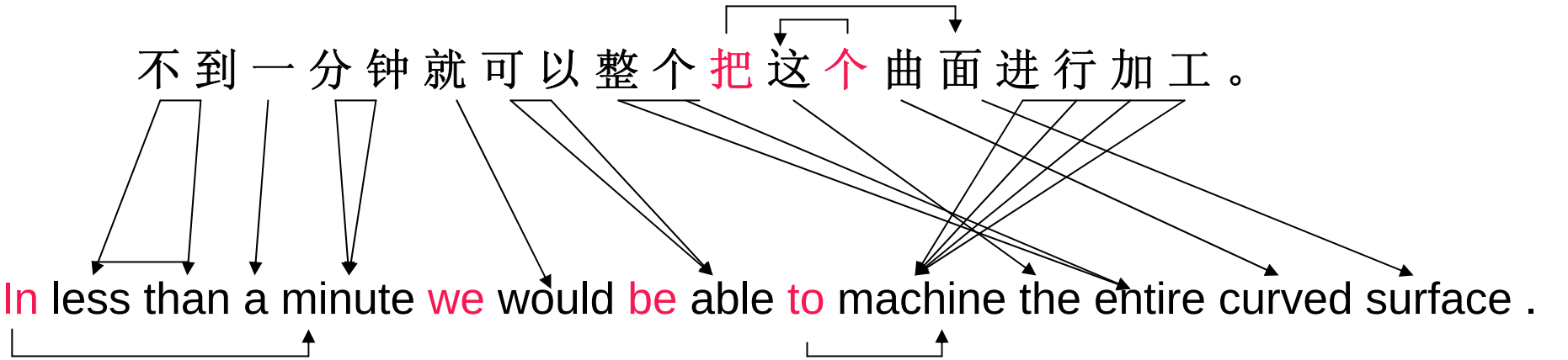
  Tense/passive marker
  Omni-funciton-preposition marker
  DE-modifier maker
  Possessive marker
  To-infinitive marker
  Sentence marker
  Measure-word marker
  Determiner/demonstrative marker
  Relative-clause marker
  Anaphoric reference marker
  Local context marker
  Context obligatory marker
  Context non-obligatory marker
- Additional feature tag types
  - Word-order change
  - Discontinuity feature

Current Alignment Task

Morphological level alignment (*char-based* alignment)
Word level (semantic equivalent alignment:
translated-correct/incorrect,
not-translate-correct/incorrect)



New Alignment Task



把 is tagged as omni-function-preposition (phrase level)
个 is tagged as measure-word (phrase-level)
"In" is tagged as omni-function-preposition (phrase-level)
"we" is tagged as context obligatory (sentence/discourse level)
"be" is tagged as context obligatory (sentence level)
"to" is tagged as to-infinitive (phrase level)
把 is related to word order change (feature tagging)

(9 types of links, 13 types of unaligned word tags, 2 types of feature tags)

Progress and Timeline

- Guidelines completed this month
  - Awaiting endorsement from committee
- IBM/LDC joint pilot study to begin later this month
  - Guidelines validity, IAA, efficiency need to be established
- LDC production annotation expected to begin in April
- Throughput for tagging task not yet known
  - Requires enhancements to annotation toolkit
  - Requires new annotator training
- Thanks to WA committee especially Niyu, Rich and Daniel