Parallel Text Processing
Resources in GATE

Niraj Aswani
In next 30-40 minutes...

- GATE Demo
- Requirements
- Parallel text resources
- Demos
  - Alignment Editor
  - Cross-document co-reference resolution
GATE Demo

- Limitations
  - PRs process one document at a time
  - A Corpus Pipeline passes only one document at a time to the PRs in the pipeline
Requirements

● Cross-lingual Text Processing
  – Requires more than one document
  – e.g. A Sentence Aligner

● Cross-document Co-reference resolution
  – Access to multiple documents
A Possible Solution?

- Re-write Pipelines to feed more than one document
  - how many docs?
  - which docs?
Parallel Text Resources

• Compound Document
  – A set of multiple documents
  – Similar to a corpus and yet presented as a single GATE Document (implements gate.Document)

• SO...
  – Corpus pipelines would work
  – Question: which document?
Parallel Text Resources

- Switch Members PR
  - Sets the focus on a particular member document
Parallel Text Resources

- Switch Members PR
  - e.g. given a compound document with an English and its parallel Hindi document
  - set the focus on English document
  - Process with some PRs
  - set the focus on Hindi document
  - Process with some Prs

- Demo
Alignment Editor

- Allows aligning texts manually at any annotation level
  - e.g. section, paragraph, sentence, word etc.
  - one pair is shown at a time
- Customizable
  - PreDisplay Action
  - Actions With the Align/Unalign Action
  - PostDisplay Action
- Demo
Parallel Text Resources

• Composite Document
  – merge texts of members of compound document
    • e.g. processing only abstracts from publications
  – yet keeps the merged text linked to the source documents
More Info: http://www.gate.ac.uk/sale/tao/index.html

Chapter 12 Tools for Alignment Tasks

Thank You!