make experiment

WMT 2010 workflow management
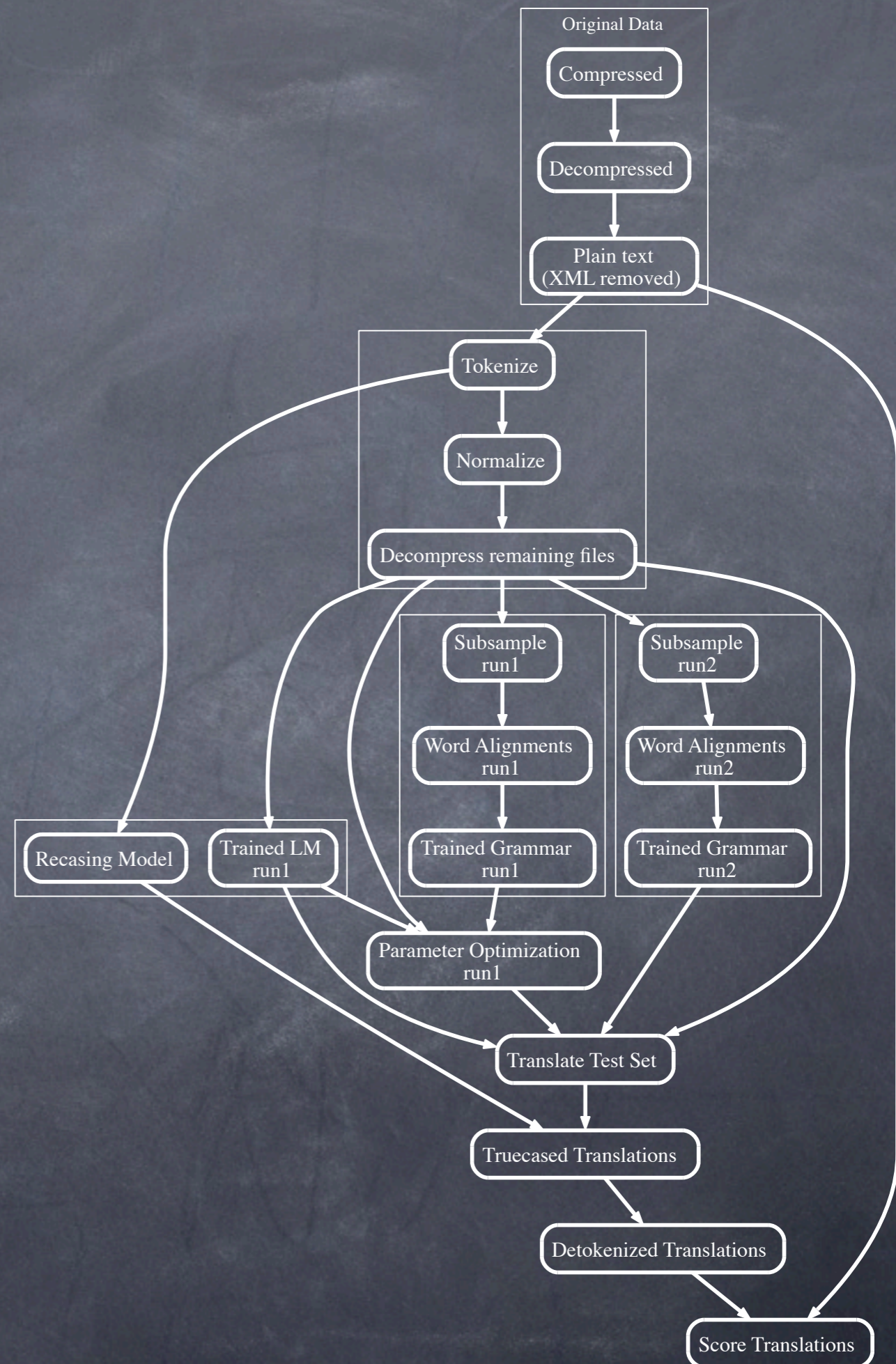
Goals
JHU Submission WMT 2010

- Running translation pipeline should be easy

  - Easy to understand

  - Easy to configure

  - Easy to monitor

  - Easy to run

- All results must be reproducible

Understand the Pipeline

- Workflow is complex

- Visualize using GraphViz

- Simple text format

- nodeName [label="text"]

- nodeA -> nodeB

- Output as graphics file

Original Data

- Compressed
- Decompressed
- Plain text (XML removed)

Tokenize

Normalize

Decompress remaining files

| Subsample run1 | Subsample run2 |
| Word Alignments run1 | Word Alignments run2 |

Recasing Model

Trained LM run1

Trained Grammar run1

Trained Grammar run2

Parameter Optimization run1

Translate Test Set

Truecased Translations

Detokenized Translations

Score Translations

Joshua Machine Translation Workflow

Configuration

- Configure each step

- Don't repeat yourself

- Explicitly mark dependencies

- Challenge: Should each step define variables for each input, or should can steps assume they know what their input is?

Monitor experiments

- Run results

  - Result dir gets name from its config file

  - Steps are numbered, named, & labelled

- Challenge: automatic naming of log files

- Challenge: visualize run status
  (via remote web interface?)

Dry run, run, re-run

- See what will be run:
  $ make --dry-run -f config/014.MERT.de-en.bleu.run1.mk

- Kick off the job:
  $ nohup make -f config/014.MERT.de-en.bleu.run1.mk &>
  999.logs/014.MERT.de-en.bleu.run1.log &

- Verify that everything finished:
  $ make --dry-run -f config/014.MERT.de-en.bleu.run1.mk
  make: Nothing to be done for `mert'.

Try it out

- Make scripts defining each logical step:
  svn co https://joshua.svn.sourceforge.net/svnroot/joshua/branches/pipeline/wmt10 000.makefiles

- Make scripts configuring each actual job:
  svn co https://joshua.svn.sourceforge.net/svnroot/joshua/branches/pipeline/wmt10-config configure-experiment

- Experiments to date: a01, a02, a03, a04, a05
  /mnt/data/wmt10.labelled