An Experimental Management System

Philipp Koehn
presented by Barry Haddow

16 September 2010
How do you run experiments?
Executing a Lot of Scripts

tokenize < corpus.en > corpus.en.tok
_tokenize < corpus.fr > corpus.fr.tok
lowercase < corpus.en.tok > corpus.en.lc
_lowercase < corpus.fr.tok > corpus.fr.lc
...
mert.perl ....
moses ...
mteval-v13.pl ...
Executing a Lot of Scripts

Oh wait, a mistake!

```bash
tokenize < corpus.en > corpus.en.tok
tokenize -l fr < corpus.fr > corpus.fr.tok
lowercase < corpus.en.tok > corpus.en.lc
lowercase < corpus.fr.tok > corpus.fr.lc
...
mert.perl ....
moses ...
mteval-v13.pl ...
```
Variations

• For instance, varying the distortion limit

```perl
for(my $dl=3;$dl<=10;$dl++) {
   'moses -dl $dl ... > output.dl-$dl';
   'wrap-xml.perl < output.dl-$dl > output.dl-$dl.sgm';
   'mteval-v13.pl -t output.dl-$dl ... > score.dl-$dl'
}
```

• But:
  – needs to be customized for every case
  – what if some of the steps crash?
  – how schedule in parallel on different machines / cluster?
A New Student Arrives

Hey, how do I build a German-English system?
A New Student Arrives

Hey, how do I build a German-English system?

blah blah europarl blah newstest2009 blah blah
One Week Later

Hey, results suck!

Did you use a 5-gram LM?

What tuning set did you use?
One Week Later

Hey, results suck!

Did you use a 5-gram LM?
What tuning set did you use?
One Week Later

Hey, results suck!

Did you use a 5-gram LM?
What tuning set did you use?

???
One Year Later

Huh, how did I build the system?
One Year Later

Huh, how did I build the system?

What is output.dl5.foo99-better.redone.fix3?
There must be a better way...
Experiment.perl

- One configuration file for all settings: record of all experimental details

- Scheduler of individual steps in pipeline
  - automatically keeps track of dependencies
  - on single machine, multi-core machines, GridEngine clusters
  - parallel execution
  - crash detection
  - automatic re-use of prior results

- Fast to use
  - set up a new experiments in minutes
  - set up a variation of an experiment in seconds
Workflow automatically generated by experiment.perl
How does it work?

• Write a configuration file (typically by adapting an existing file)

• Execute: experiment.pl -config config -exec
Components

- **experiment.perl**
  - executable that schedules all steps
  - has functions that define more complex steps (e.g., tuning)

- **experiment.meta**
  - meta-configuration file
  - defines all steps and their dependencies
  - template definitions for simpler steps

- **config**
  - includes settings for one experimental run
**Jargon**

**experiment:** consists of a number of **runs** that share the same data, same processing **steps**, etc.
example: WMT 2010 German-English system

**run:** individual experimental instance with specific settings and one outcome
example: run with higher distortion limit

**step:** processing step of a **run**
examples: tokenization, decoding

**module:** logical block of processing **steps**
examples: corpus preparation, tuning

**set:** corpus for training or language modeling
examples: Europarl, News Commentary

**setting:** specific parameter in the configuration of a experimental **run**
example: decoder-setting = "-dl 8"
experiment.meta

- Definition of potential steps of an experiment

- Grouped in modules:
  - CORPUS: preparing a parallel corpus
  - INPUT-FACTOR and OUTPUT-FACTOR: commands to create factors
  - TRAINING: training a translation model
  - LM: training a language model
  - INTERPOLATED-LM: interpolate language models
  - SPLITTER: training a word splitting model
  - RECASING: training a recaser
  - TRUECASING: training a truecaser
  - TUNING: running minimum error rate training to set component weights
  - TESTING: translating and scoring a test set
  - REPORTING: compile all scores in one file
Step Definition

[LM]
get-corpus
  in: get-corpus-script
  out: raw-corpus
  [...] 

tokenize
  in: raw-corpus
  out: tokenized-corpus
  [...] 

- Input and outputs establish dependencies between steps (as in a Makefile)
  - tokenize requires as input raw-corpus
  - get-corpus produces as output raw-corpus
  - when tokenize needs to be run, first raw-corpus needs to be executed
Configuration File

- Input to a step may be specified in configuration file (config):
  
  ```
  [LM:europarl]
  
  ### raw corpus file
  #
  raw-corpus = $europarl-v3/training/europarl-v3.en
  ```

- May limit which steps need to be executed
  
  - tokenize requires as input raw-corpus
  - raw-corpus is specified in the config file
  - no need to run get-corpus
Elements of Step Definitions

- Several parameters for step definitions are used in experiment.meta:
  - in and out: established dependencies between steps
  - default-name: file name of output
  - template: template for the command that executes step
  - pass-unless: only execute if the given setting is used
  - error: if STDERR contains specified key words, step has failed
  - rerun-on-change: limits re-use if specified settings are changed

- There are more (see paper or documentation)
Definition of LM:tokenize

```
_tokenize
 in: raw-corpus
 out: tokenized-corpus
 default-name: lm/tok
 pass-unless: output-tokenizer
 template: $output-tokenizer < IN > OUT
 parallelizable: yes
```
Configuration File

- List of settings

- Comments and empty lines for better readability

- Organized in sections for each module
  - start of section indicated by module (and set) name
  - examples: [TRAINING] or [CORPUS:europarl]

- Syntax of setting definition: setting = value
Configuration File: Syntax

- Settings can be used as variables to define other settings:
  
  ```bash
  working-dir = /home/pkoehn/experiment
  wmt10-data = $working-dir/data
  ```

- Variable names may be placed in curly brackets for clearer separation:
  
  ```bash
  wmt10-data = ${working-dir}/data
  ```

- References to output of other steps
  
  ```bash
  [RECASING]
tokenized = [LM:europarl:tokenized-corpus]
  ```
Step Files

- Command to execute is stored in a file

- After execution, other files are created:

  steps/1/LM_europarl_tokenize.1
  steps/1/LM_europarl_tokenize.1.DONE
  steps/1/LM_europarl_tokenize.1.INFO
  steps/1/LM_europarl_tokenize.1.STDERR
  steps/1/LM_europarl_tokenize.1.STDERR.digest
  steps/1/LM_europarl_tokenize.1.STDOUT

  - meta information (INFO, DONE)
  - output (STDERR, STDOUT)
  - digest of output for indicators of crash (STDERR.digest)
Re-Use of Steps

• Example:
  – run 1: baseline
  – run 2: change order of language model
  → tokenization and truecasing of language model training data can be re-used

• Files in directory for language model data:

  % ls -tr lm/*
  lm/europarl.tok.1
  lm/europarl.truecased.1
  lm/europarl.lm.1
  lm/europarl.lm.2
# Web Interface

## All Experimental Setups

<table>
<thead>
<tr>
<th>ID</th>
<th>User</th>
<th>Task</th>
<th>Directory</th>
</tr>
</thead>
<tbody>
<tr>
<td>97</td>
<td>pkoehn</td>
<td>Acquis Truecased</td>
<td>/group/project/statmt2/pkoehn/acquis-truecase</td>
</tr>
<tr>
<td>96</td>
<td>pkoehn</td>
<td>Chinese-English AGILE 2008</td>
<td>/group/project/statmt2/pkoehn/agile08-chinese</td>
</tr>
<tr>
<td>95</td>
<td>miles</td>
<td>Random testing</td>
<td>/group/project/statmt7/miles/experiments /ep-enfr/work</td>
</tr>
<tr>
<td>94</td>
<td>joseph</td>
<td>Proj2008 Impl.Adapted experiment(fr-en)for News Comm.</td>
<td>/group/project/statmt2/joseph/experimentJo/task6</td>
</tr>
<tr>
<td>93</td>
<td>joseph</td>
<td>Proj2008 Impl.Baseline experiment(fr-en)for News Comm.</td>
<td>/group/project/statmt2/joseph/experimentJo/task5</td>
</tr>
<tr>
<td>92</td>
<td>jschroe1</td>
<td>FR-EN System Combination Components</td>
<td>/group/project/statmt9/josh/experiments /fr-syscomb/work</td>
</tr>
</tbody>
</table>

List of experiments

Philipp Koehn  
Experimental Management System  
16 September 2010
# List of Runs

**Task: WMT10 German-English (pkoehn)**

<table>
<thead>
<tr>
<th>ID</th>
<th>start</th>
<th>end</th>
<th>avg</th>
<th>newstest2009</th>
<th>newstest2010</th>
</tr>
</thead>
</table>
| [1042-16] 11+analysis | 16 May | 16 May | BLEU-c: 21.74  
BLEU: 22.91 | 21.03 (1.002) | 22.45 (1.041) |
| [1042-15] 11+Internal emplus test set | 21 Apr | crashed | - | - |
| [1042-14] 9+interpolated-tm.lm-weighted | 21 Feb | 21 Feb  
9: 0.239258 -> 0.239296 | - | 20.81 (1.003) | - |
| [1042-13] 9+only-ep | 21 Feb | 21 Feb  
13: 0.235046 -> 0.235053 | - | 20.42 (1.002) | - |
| [1042-12] 9+only-nc | 21 Feb | 21 Feb  
7: 0.2222237 | - | 18.96 (1.002) | - |
Analysis: Basic Statistics

- Basic statistics
  - n-gram precision
  - evaluation metrics
  - coverage of the input in corpus and translation model
  - phrase segmentations used

<table>
<thead>
<tr>
<th>Coverage</th>
<th>model</th>
<th>corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2047</td>
<td>1708</td>
</tr>
<tr>
<td>1</td>
<td>738</td>
<td>518</td>
</tr>
<tr>
<td>2-5</td>
<td>1483</td>
<td>818</td>
</tr>
<tr>
<td>6+</td>
<td>61745</td>
<td>62969</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phrase Segmentation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4+</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 to</td>
<td>26897</td>
<td>2145</td>
<td>278</td>
<td>90</td>
</tr>
<tr>
<td>2 to</td>
<td>4144</td>
<td>14414</td>
<td>2518</td>
<td>432</td>
</tr>
<tr>
<td>3 to</td>
<td>639</td>
<td>3522</td>
<td>4821</td>
<td>1272</td>
</tr>
<tr>
<td>4+ to</td>
<td>158</td>
<td>855</td>
<td>1693</td>
<td>2135</td>
</tr>
</tbody>
</table>

by token / by type / details

by word / by phrase
## Analysis: Unknown Words

grouped by frequency in test set

<table>
<thead>
<tr>
<th>Unknown Words</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eatonville</td>
<td>18</td>
</tr>
<tr>
<td>Hurston</td>
<td>16</td>
</tr>
<tr>
<td>Barrick</td>
<td>12</td>
</tr>
<tr>
<td>Hema</td>
<td>12</td>
</tr>
<tr>
<td>Stewards</td>
<td>12</td>
</tr>
<tr>
<td>Gebrselassie</td>
<td>11</td>
</tr>
<tr>
<td>Flamenco</td>
<td>10</td>
</tr>
<tr>
<td>Mango</td>
<td>10</td>
</tr>
<tr>
<td>Glitter</td>
<td>9</td>
</tr>
<tr>
<td>ÚOHS</td>
<td>9</td>
</tr>
<tr>
<td>ČTÚ</td>
<td>9</td>
</tr>
<tr>
<td>Coles</td>
<td>8</td>
</tr>
<tr>
<td>Deka</td>
<td>8</td>
</tr>
<tr>
<td>Garci</td>
<td>8</td>
</tr>
<tr>
<td>ITV</td>
<td>8</td>
</tr>
<tr>
<td>Ózdi</td>
<td>8</td>
</tr>
<tr>
<td>Eatonvilles</td>
<td>4</td>
</tr>
<tr>
<td>Együtt</td>
<td>4</td>
</tr>
<tr>
<td>Garver</td>
<td>4</td>
</tr>
<tr>
<td>Harmadik</td>
<td>4</td>
</tr>
<tr>
<td>Hurstons</td>
<td>4</td>
</tr>
<tr>
<td>Jobl, Jol</td>
<td>4</td>
</tr>
<tr>
<td>Jos, Jövőrt</td>
<td>4</td>
</tr>
<tr>
<td>Kovaljev</td>
<td>4</td>
</tr>
<tr>
<td>Krever</td>
<td>4</td>
</tr>
<tr>
<td>Lados</td>
<td>4</td>
</tr>
<tr>
<td>Mercandelli</td>
<td>4</td>
</tr>
<tr>
<td>Stehplätze</td>
<td>4</td>
</tr>
<tr>
<td>Tauro</td>
<td>4</td>
</tr>
<tr>
<td>Tórtola</td>
<td>4</td>
</tr>
<tr>
<td>Zenobia</td>
<td>4</td>
</tr>
<tr>
<td>fön</td>
<td>4</td>
</tr>
<tr>
<td>Ózdi</td>
<td>4</td>
</tr>
<tr>
<td>3: Anmil,</td>
<td>3</td>
</tr>
<tr>
<td>Atlasz, BR23C</td>
<td>3</td>
</tr>
<tr>
<td>BSA, Bayón</td>
<td>3</td>
</tr>
<tr>
<td>Biztos, Bt.</td>
<td>3</td>
</tr>
<tr>
<td>Butch, Casado</td>
<td>3</td>
</tr>
<tr>
<td>Dal, Embraer</td>
<td>3</td>
</tr>
<tr>
<td>FT, Faymann</td>
<td>3</td>
</tr>
<tr>
<td>Fiatal, Gregg</td>
<td>3</td>
</tr>
<tr>
<td>Gélineau, HSV</td>
<td>3</td>
</tr>
<tr>
<td>Hanzelka</td>
<td>3</td>
</tr>
<tr>
<td>Helläusern, Iván</td>
<td>3</td>
</tr>
<tr>
<td>Janssen, Jančura</td>
<td>3</td>
</tr>
<tr>
<td>Chemical, Chigi</td>
<td>3</td>
</tr>
<tr>
<td>Cineast, Comics</td>
<td>3</td>
</tr>
<tr>
<td>Commerzbank, Cappola</td>
<td>3</td>
</tr>
<tr>
<td>Corker, Cowon, DF</td>
<td>3</td>
</tr>
<tr>
<td>Dinks, Download</td>
<td>3</td>
</tr>
<tr>
<td>Drehbewegung, Drzewiecki, Drápal</td>
<td>3</td>
</tr>
<tr>
<td>Drüsseldorfer, Ella</td>
<td>3</td>
</tr>
<tr>
<td>albums, alondra, andoh, anm., armiñon, ashford, bzo, baloldal</td>
<td>3</td>
</tr>
<tr>
<td>bani, baugesellschaften, bedienkomfort, bento, bentos, bingleys, bojen, bowens, bowery, boyd, bringley, browser, bělholávek, CBGB,</td>
<td>3</td>
</tr>
<tr>
<td>8,25, 8,81, 9,14, 99,80, AAC, ADQ, ART, Aareal, abbremsens, abhöraktion, absenzen, abwesenheiten, abwiegen, abwärtsso, achronot, actor, adSense, adWords, aday, adobe, adressverzeichnisses, adwords, adelard, Agazio, Akku, Akron, aktuálné.cz, alameda, alatrste, alcolock, aleš, alhambra, alleinregierer, Amazonengebiet, amil, aminei, amministrazione, amway, andalusierin, andik, Android, andél, angeklagtem, ansa, anthology, anti statistika, apnoe, aquel, arabija, arbeiternehmers, arcandor, arriaga, asiana, askale, astronomen, aufeislegen, augäpfel, ausdrückstärke, ausführungs-, ausgeruhter, ausscheidungsspiele,</td>
<td>3</td>
</tr>
</tbody>
</table>
Analysis: Output Annotation

Color highlighting to indicate n-gram overlap with reference translation

darker bleu = word is part of larger n-gram match
Analysis: Input Annotation

100 occurrences in corpus, 52 distinct translations, translation entropy: 3.08447

• For each word and phrase, color coding and stats on
  - number of occurrences in training corpus
  - number of distinct translations in translation model
  - entropy of conditional translation probability distribution $\phi(e|f)$ (normalized)
Analysis: Alignment

Phrase alignment of the decoding process

(red border, interactive)
**Analysis: Tree Alignment**

Uses nested boxes to indicate tree structure
(red border, yellow shaded spans in focus, interactive)
for syntax model, non-terminals are also shown
Analysis: Comparison of 2 Runs

annotated sentences
sorted by order worse display fullscreen showing more all
identical same better worse
2348 51 57 69
93% 2% 2% 3%

[2143:0.2974] In Austria, Haider and Co. are ready to govern to prevent a red and black coalition.
[2143:0.1754] In Austria, Haider and Co. are prepared to rule to prevent a red and black coalition.
[ref] Haider and his party are ready to govern Austria in order to avoid red @-@ black coalition.

[2165:0.3174] The SPÖ wants to show that the cooperation of both parties is possible - in some countries and in the social partnership that is already the case.
[2165:0.2061] The SPÖ wants to show that a cooperation of both parties is possible - in some countries and in the social partnership that is already the case.
[ref] SPÖ would like to show that the cooperation of the two parties is possible - it does exist in some of the provinces as well as in social partnership.

Different words are highlighted
sortable by most improvement, deterioration
Conclusion

- Experiment.perl makes life easier
  - setting up complex experiments with one configuration file
  - permanent record of parameter settings
  - easily distributed (Edinburgh’s WMT 2010 system configs available)

- Analysis allows insight into model performance
  - basic stats
  - inspect derivations and options of decoder
  - differences between two runs

- Future plans
  - integrate more tools (also yours, help wanted!)
  - scheduling jobs on Hadoop
  - more analysis