CSE Framework: A UIMA-based Distributed System for Configuration Space Exploration

Elmer Garduno\textsuperscript{2}, Zi Yang\textsuperscript{1}, Yan Fang\textsuperscript{3}, Avner Maiberg\textsuperscript{1}, Collin McCormack\textsuperscript{4}, Eric Nyberg\textsuperscript{1}

\textsuperscript{1} Carnegie Mellon University \quad \{ziy, amaiberg, ehn\}@cs.cmu.edu
\textsuperscript{2} Sinnia \quad elmerg@sinnia.com
\textsuperscript{3} Oracle Corporation \quad yan.fang@oracle.com
\textsuperscript{4} The Boeing Company \quad collin.w.mccormack@boeing.com
Which city in China has the largest number of foreign financial companies?

Keywords: China largest foreign financial company
Answer type: location (city)

<table>
<thead>
<tr>
<th>Document ID</th>
<th>Rank</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>FBIS3-58 (relevant)</td>
<td>1</td>
<td>0.7</td>
</tr>
<tr>
<td>AP880603-0268</td>
<td>2</td>
<td>0.65</td>
</tr>
<tr>
<td>WSJ920110-0013</td>
<td>3</td>
<td>0.64</td>
</tr>
<tr>
<td>FBIS3-45320 (relevant)</td>
<td>4</td>
<td>0.64</td>
</tr>
<tr>
<td>FT942-2016</td>
<td>5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Answer candidates:
- Beijing: 0.7, AP880603-0268
- Hong Kong: 0.65, WSJ920110-0013
- Shanghai: 0.64, FBIS3-58
- Taiwan: 0.5, FT942-2016
- Shanghai: 0.4, FBIS3-45320

Answer: Shanghai
CURRENT RESEARCH IN QA
What did we learn from Watson?

• QA systems can be fast enough, accurate enough, and confident enough to perform in the real world

• Key factors:
  – Scalable, parallel architecture
  – Agile, open advancement process

• Next big challenge: *rapid domain adaptation*
Automatic Optimization of QA for TREC Genomics Questions
Results of Automatic Optimization

<table>
<thead>
<tr>
<th></th>
<th>Participants</th>
<th>CSE</th>
<th>Scaled CSE</th>
</tr>
</thead>
<tbody>
<tr>
<td># Component</td>
<td>~1,000</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td># Configuration</td>
<td>~1,000</td>
<td>32</td>
<td>2,946</td>
</tr>
<tr>
<td># Trace</td>
<td>92</td>
<td>2,700</td>
<td>~1.426 x 10^{12}</td>
</tr>
<tr>
<td># Execution</td>
<td>~1,000</td>
<td>190,680</td>
<td>~6.050 x 10^{13}</td>
</tr>
<tr>
<td>Capacity (hours)</td>
<td>N/A</td>
<td>24</td>
<td>24</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DocMAP</th>
<th>Max</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.5439</td>
<td>.3083</td>
<td>.0198</td>
<td>.1486</td>
<td>.0345</td>
<td>.0007</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.5648</td>
<td>.4770</td>
<td>.1087</td>
<td>.1773</td>
<td>.1603</td>
<td>.0311</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.5072</td>
<td>.3509</td>
<td>.2679</td>
<td>.1181</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PsgMAP</th>
<th>Max</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.0198</td>
<td>.1087</td>
<td>.0311</td>
<td>.1181</td>
<td>.0713</td>
<td>.0164</td>
<td></td>
</tr>
</tbody>
</table>

Automatically Building an Information System by Another Meta-System?
Building an Information System Automatically

• CSE framework
The *benefit* of CSE framework

- Accelerate the system development cycle by automating the component selection and tuning!
- Save *cost*!

It requires

- Identify the tool, knowledge base, task algorithm candidates
- Provide information needs with known outcomes, e.g. answers to questions in the domain.
CSE - FRAMEWORK
Definition: Phase

• An information system

Phase $t$
The processing unit as the $t$-th step in a process
Definition: Component, Configuration

- Inside phase $t$
  - Component $f_t^c$
    - An instantiated processing unit in phase $t$
  - Configuration $\omega_t^c$
    - Consists of a set of parameters $\{\omega_t^{c,p}\}_p$

Configured component $f_t^c|\omega_t^c$
Definition: Trace

\[
\text{Trace } f^c|\omega^c = (f_1^{c_1}|\omega_1^{c_1}, f_2^{c_2}|\omega_2^{c_2}, ..., f_n^{c_n}|\omega_n^{c_n})
\]

An execution path that involves a single configured component for each phase
Exponential problem

The number of traces grows exponentially with the phases and the number of components.

Space should be pruned when possible to keep the space bounded.
Definition: Configuration space

- Pipeline

Configuration space $\mathcal{F}|\Omega = \{f^c|\omega^c\}_c$

Set of all configured components
UIMA - EXTENDED CONFIGURATION DESCRIPTOR
Extended Configuration Descriptor

• YAML format
• A simple yet complete configuration descriptor

```yaml
configuration:
  name: testqa-ziy-test
  author: ziy

persistence-provider:
  inherit: jdbc.db.persistence-provider

collection-reader:
  inherit: jdbc.db.collection-reader
  dataset: BIO-COMBINED
  sequence-start: 160
  sequence-end: 187
```
Extended Configuration Descriptor

• Phases and components *inherit* configuration properties or are declared as *classes*.

```python
pipeline:
  - inherit: jdbc.cse.phase
    name: keyterm-extractor
    options: |
      - inherit: default.keyterm.default
      - inherit: default.keyterm.faster
  - inherit: jdbc.cse.phase
    name: retrieval-stategist
    options: |
      - inherit: default.retrieval.default
      - inherit: default.retrieval.better
  - inherit: jdbc.cse.phase
    name: passage-extractor
    options: |
      - class: cmu.edu.default.ie.Default
```
Component configuration

class: edu.cmu.lti.oaqa.ecd.example.FirstPhaseAnnotatorA1
extract: true
cross-opts
  param-a: [value100,value200]
  param-b: [value300,value400]

This evaluates to the following Object[] param lists.

[extract: true, param-a: value100, param-b: value300]
[extract: true, param-a: value200, param-b: value300]
[extract: true, param-a: value100, param-b: value400]
[extract: true, param-a: value200, param-b: value400]
Extended Configuration Descriptor

- Evaluation metrics are pluggable, and can be specified at the local or global level.

- inherit: jdbc.eval.cse-retrieval-aggregator-consumer
- inherit: bioqa.eval.cse-passage-map-aggregator-consumer

post-process:
- inherit: jdbc.eval.cse-retrieval-evaluator-consumer
- inherit: report.csv-report-generator
builders: |
  - inherit: jdbc.report.f-measure-report-component

- inherit: bioqa.eval.cse-passage-map-evaluator-consumer
- inherit: report.csv-report-generator
builders: |
  - inherit: bioqa.report.map-report-component
In-phase pipelines

Example

#options = 13
IMPLEMENTATION
Implementation details

• Built on top of uimaFIT
• Combinatorial features are implemented using CAS Multiplier.
• CASes are persisted as compressed XMI
  – Once per trace at each phase.
  – Experiments can be restarted at any arbitrary point.
• Experimentation specific Type System
• Use UIMA-AS for external resources.
Distributed execution
Incremental improvement
## Per trace visibility

### Retrieval evaluation

<table>
<thead>
<tr>
<th>Trace</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>MAP</th>
<th>Binary recall</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DefaultKeytermExtractor[persistence-provider] inherit: internal.log-persistence-provider</td>
<td>0.0508452</td>
<td>0.383967</td>
<td>0.0898006</td>
<td>0.51594</td>
<td>0.0384515</td>
</tr>
<tr>
<td>2</td>
<td>LingPipeHmmPostagger[ModelFilePath/]pos-en-bio-medpost.HiddenMarkovModel[persistence-provider] inherit: internal.log-persistence-provider</td>
<td>0.0508452</td>
<td>0.383967</td>
<td>0.0898006</td>
<td>0.51594</td>
<td>0.0384515</td>
</tr>
</tbody>
</table>
## Error analysis

<table>
<thead>
<tr>
<th>Id</th>
<th>Question</th>
<th>Retrieved</th>
<th>Relevant</th>
<th>Average precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>187</td>
<td>How do mutations in familial hemiplegic migraine type 1 (FH-M1) gene affect calcium ion influx in hippocampal neurons?</td>
<td>1000</td>
<td>3</td>
<td>1.00000</td>
</tr>
<tr>
<td>188</td>
<td>How does BARD1 regulate BRCA1 activity?</td>
<td>1000</td>
<td>243</td>
<td>0.87096</td>
</tr>
<tr>
<td>175</td>
<td>How does L2 interact with L1 to form HPV11 viral capsids?</td>
<td>1000</td>
<td>33</td>
<td>0.84186</td>
</tr>
<tr>
<td>160</td>
<td>What is the role of PrP in mad cow disease?</td>
<td>1000</td>
<td>525</td>
<td>0.83522</td>
</tr>
<tr>
<td>181</td>
<td>How do mutations in the Huntingtin gene affect Huntington's disease?</td>
<td>1000</td>
<td>589</td>
<td>0.78279</td>
</tr>
<tr>
<td>170</td>
<td>How does COP2 contribute to CFTR export from the endoplasmic reticulum?</td>
<td>1000</td>
<td>36</td>
<td>0.76515</td>
</tr>
<tr>
<td>166</td>
<td>How do mutations in the Presenilin-1 gene affect Alzheimer's disease?</td>
<td>1000</td>
<td>388</td>
<td>0.71872</td>
</tr>
<tr>
<td>167</td>
<td>How does nucleoside diphosphate kinase (NM23) contribute to tumor progression?</td>
<td>1000</td>
<td>208</td>
<td>0.69712</td>
</tr>
<tr>
<td>164</td>
<td>What is the role of Nurr-77 in Parkinson's disease?</td>
<td>1000</td>
<td>7</td>
<td>0.66017</td>
</tr>
<tr>
<td>163</td>
<td>What is the role of APC (adenomatous polyposis coli) in colon cancer?</td>
<td>1000</td>
<td>262</td>
<td>0.62560</td>
</tr>
<tr>
<td>184</td>
<td>How do mutations in the Ras gene affect cell growth?</td>
<td>1000</td>
<td>5</td>
<td>0.65556</td>
</tr>
<tr>
<td>165</td>
<td>How do Cathepsin D (CDS1) and apolipoprotein E (ApoE) interactions contribute to Alzheimer's disease?</td>
<td>1000</td>
<td>17</td>
<td>0.54619</td>
</tr>
<tr>
<td>161</td>
<td>What is the role of IDE in Alzheimer's disease?</td>
<td>1000</td>
<td>68</td>
<td>0.48033</td>
</tr>
<tr>
<td>183</td>
<td>How do mutations in the NM23 gene affect tracheal development?</td>
<td>1000</td>
<td>19</td>
<td>0.47449</td>
</tr>
<tr>
<td>185</td>
<td>How do mutations in the hypocretin receptor 2 gene affect narcolepsy?</td>
<td>1000</td>
<td>25</td>
<td>0.46111</td>
</tr>
<tr>
<td>172</td>
<td>How does p53 affect apoptosis?</td>
<td>1000</td>
<td>587</td>
<td>0.46088</td>
</tr>
<tr>
<td>166</td>
<td>What is the role of Transforming growth factor-beta1 (TGF-beta1) in cerebral amyloid angiopathy (CAA)?</td>
<td>1000</td>
<td>34</td>
<td>0.42790</td>
</tr>
</tbody>
</table>
Other domains: QA4MRE

- Question Answering for Machine Reading
- Configuration space:
  - 12 UIMA components were first developed
  - Replace UIMA descriptors with ECD
- CSE
  - 46 configurations
  - 1,040 combinations
  - 1,322 executions

The best trace identified by CSE achieved 59.6% performance gain over the original pipeline.

[Building Optimal Question Answering System Automatically using Configuration Space Exploration (CSE) for QA4MRE 2013 Tasks Alkesh Patel, Zi Yang, Eric Nyberg and Teruko Mitamura]
FUTURE WORK AND COLLABORATION
Future work

• Advanced Configuration Space exploration and pruning (Bagpipes Framework).

• Run arbitrary UIMA pipelines on top of industry grade distributed systems (Spark, Mesos, HDFS).

• Further investigation on space, time, resources constraining.

• Use differential CAS storage.
Collaboration

http://oaqa.github.io
Thanks!