A model driven engineering approach to automatic dataset generation for benchmarking purposes - with application to the text extraction problem

Kresimir Duretec, Andreas Rauber, Christoph Becker (UoT)

IFS group presentation, 04.12.2017
Vienna, Austria
Digital preservation

• **Goal**: keep information in an **accessible** and **authentic** form

• Wide spectrum of problems

• Core operations based on **data processing software components**
Digital preservation - data processing workflow
Things go wrong!

<table>
<thead>
<tr>
<th>intestation</th>
<th>fossicker</th>
<th>fossilation</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>47</td>
<td>5</td>
</tr>
<tr>
<td><strong>43</strong></td>
<td>65</td>
<td><strong>39</strong></td>
</tr>
<tr>
<td>50</td>
<td>68</td>
<td>14</td>
</tr>
<tr>
<td>20</td>
<td>73</td>
<td>1</td>
</tr>
<tr>
<td>53</td>
<td>71</td>
<td>1</td>
</tr>
</tbody>
</table>
Things go wrong!

<table>
<thead>
<tr>
<th>intestation</th>
<th>fossicker</th>
<th>fossilation</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>47</td>
<td>5</td>
</tr>
<tr>
<td>43</td>
<td>65</td>
<td>39</td>
</tr>
<tr>
<td>50</td>
<td>68</td>
<td>14</td>
</tr>
<tr>
<td>20</td>
<td>73</td>
<td>1</td>
</tr>
<tr>
<td>53</td>
<td>71</td>
<td>1</td>
</tr>
</tbody>
</table>
Problem - functional correctness

• **Functional correctness** is the critical software quality characteristic but significantly **under-evaluated**

• Lack of resources for evaluations

• Lack of proper test suites
  • missing ground truth

• Available resources and test suits shift evaluations to other quality characteristics
  • efficiency
  • reliability
What is needed?

• Improve the **evidence base**
  • well known issue in DP

• Move towards more systematic **software benchmarks**
  • “a set of test used to compare the performance of alternative components” (Sim 2003)

• **Preconditions**
  • competing components
  • maturing field
  • community collaboration
**Benchmark components**

<table>
<thead>
<tr>
<th>Software engineering</th>
<th>Digital preservation</th>
<th>Information retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motivating comparison</td>
<td>Motivating comparison</td>
<td>Tasks</td>
</tr>
<tr>
<td>Task sample</td>
<td>Function</td>
<td>Datasets</td>
</tr>
<tr>
<td>Performance measures</td>
<td>Dataset</td>
<td>Answersets</td>
</tr>
<tr>
<td></td>
<td>Ground truth</td>
<td>Measures</td>
</tr>
<tr>
<td></td>
<td>Performance measures</td>
<td>Representation formats</td>
</tr>
</tbody>
</table>

Challenges

• Data processing tools
  1. Homogeneous, narrow functionality
  2. Competing components
  3. Non-interactive behaviour
  4. Complex input data
     • Data is stored in complex formats
     • File format specifications can be ambiguous
     • Tools produce files in varying formats and with various degrees of compliance
     • Platform-specific implementations of constructs vary across formats and tools
Challenges

• The challenge of testing data processing software components lies in their complex input space!

• We need an automatic method to generate appropriate datasets and ground truth
Challenges

• What is a **good test suite**?
  • FAIR
  • Representative
    • real world
  • Coverage
    • features coverage
    • tacit knowledge

• **Ground truth data**
  • not applicable only to a specific scenario

• Failure diagnostic data
Challenges

- Representativeness
- **Object structure**
- Object content
- Capturing peculiarities of different software tools
- Implementing syntactically equivalent elements in different ways
Software environments such as MS Word and LibreOffice expose a rich API for manipulating documents.
Generic approach

- Model driven engineering
- separation of concerns
  - platform independent
  - platform specific

<table>
<thead>
<tr>
<th>MOF</th>
<th>MOF</th>
</tr>
</thead>
<tbody>
<tr>
<td>METAMODEL</td>
<td>DOCUMENT METAMODEL</td>
</tr>
<tr>
<td>MODEL</td>
<td>DOCUMENT MODEL</td>
</tr>
<tr>
<td>MODEL INSTANCE</td>
<td>DOCUMENT</td>
</tr>
<tr>
<td>M3</td>
<td>M2</td>
</tr>
<tr>
<td>M1</td>
<td>M0</td>
</tr>
</tbody>
</table>
Generic approach

Test suite modelling

Test case design
- Test features
  - Test features:
    - $f_1$
    - $f_2$
    - ... $f_n$
- Feature values
  - Collection profiling:
    - $O_1$
    - $O_2$
    - ... $O_n$
  - Characterization:
    - $O_{c1}$
    - $O_{c2}$
    - ... $O_{cn}$
- Aggregation
- Representative sampling
- Feature value distributions and representative samples
- Feature relations
  - $f_{r1}$
  - $f_{r2}$
  - ... $f_{rn}$
- Feature constraints
  - $f_{c1}$
  - $f_{c2}$
  - ... $f_{cn}$
- Test oracle definition

Test models definition
- Metamodels
  - Platform independent metamodel
  - Platform specific metamodel
- Model Transformations
  - Mutate PIM to diversity population
  - Transform PIM to PSM
  - Mutate PSM to diversify features
  - Transform PSM to code
- Model Queries
  - Extract ground truth from PIM
  - Extract ground truth from PSM

Test suite generation
- Platform independent test model
- Model mutation
- Diversified test models
- Model2Model Transformation
- Platform specific test models
- Model mutation
- Diversified test models
- Test case generation
  - Code generation
  - Code execution
  - Ground truth extraction

Test case 1
- Platform independent test data model
- Platform specific test data model
- Test data
- Test oracle
Text extraction benchmark

- Text extraction is one of the **key data processing software components** in IR and DP

- There are many **competing components** (open source and commercial)
  - Not clear which one is the best

- Tools are under-evaluated and more is required
  - focus on specific scenarios and not on generic text extraction tools
Text extraction benchmark - motivating comparison

• Text snippet
  • a piece of text positioned somewhere in a document
  • one or more words

• Document
  • composition of one or more text snippets
  • additional elements such as images added
Text extraction benchmark - motivating comparison

Perceval rationally disposed precontemporary jantiness Genista ensuing symposiasitic radiometer pinpra gradatively unenjoying babar antigenicity Lachenalia roentgenoscope andantino doorward pristine steeplechase professoring pelota everywhere uncopable Cryptogramma acreage rhythmless peplum archpastor antedownward Sinthae superauditor pegasis Colubrina unbind incur Euphippus hydropod unexecusable antiimperialisitic nonpar ligence relationship hyperdolichocranial sumper penicillate ephete semuline gelatose prefabricator sightworthy deoxidative Lus Archianamida pneumococcal Diplosaurus Mechlin navette pachyntic prayer protectorship pistolwise botaliform Tor uncontentious badgersoningly monistical hirse cymophylic Aurellius santhanoid stine shareman

<table>
<thead>
<tr>
<th>allele</th>
<th>Cephalopoda</th>
<th>athletical</th>
<th>boller</th>
</tr>
</thead>
<tbody>
<tr>
<td>68</td>
<td>36</td>
<td>65</td>
<td>100</td>
</tr>
<tr>
<td>53</td>
<td>98</td>
<td>64</td>
<td>100</td>
</tr>
<tr>
<td>100</td>
<td>87</td>
<td>36</td>
<td>53</td>
</tr>
<tr>
<td>11</td>
<td>9</td>
<td>97</td>
<td>45</td>
</tr>
</tbody>
</table>

darkness jarkman uncompliable metaldehyde anhydric unconfirmative rase microwbee polyganodieous fizz ungraphic tuille neurocanal bhaichari rebuku rhubarbus scalaris celerisvalleylike buccuracitize contempitibillility longe impenetration faventine vegetosize aerate hyperspatial ectocardia licoconite bulley malacon taiglie clawberry unrevetorbered graspress dispaquet saumya support hemometrometer whereas tyranalness Gasterophius forcing allgamy glaringness psychrometry sleepproof gangboard apalogmic bulb gateender apothegmatical Cyatha Euchlorophyseae Altrair melitioso inadvertence arooert soaky decompay
Bube exactant Olax comparably hydrate Fellata antialien
Text extraction benchmark - motivating comparison

Correctness - how well does the extracted text match text in a text snippet?

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>68</td>
<td>36</td>
<td>65</td>
<td>100</td>
</tr>
<tr>
<td>53</td>
<td>98</td>
<td>64</td>
<td>100</td>
</tr>
<tr>
<td>100</td>
<td>87</td>
<td>36</td>
<td>53</td>
</tr>
<tr>
<td>11</td>
<td>9</td>
<td>97</td>
<td>45</td>
</tr>
</tbody>
</table>
Text extraction benchmark - motivating comparison

- Three levels
  - correctness
  - order
  - layout

<table>
<thead>
<tr>
<th>allogene</th>
<th>Cephalopoda</th>
<th>athletical</th>
<th>bottler</th>
</tr>
</thead>
<tbody>
<tr>
<td>68</td>
<td>36</td>
<td>65</td>
<td>100</td>
</tr>
<tr>
<td>53</td>
<td>98</td>
<td>64</td>
<td>100</td>
</tr>
<tr>
<td>100</td>
<td>87</td>
<td>36</td>
<td>53</td>
</tr>
<tr>
<td>11</td>
<td>9</td>
<td>97</td>
<td>45</td>
</tr>
</tbody>
</table>

- Percival rationality dissemin precontemporary ianthine Genista ensuing symposiastic radiometer pinja
- Gradatively unenjoying babar antipenitivity Lachenalia roentgenoscope endantino doorward pristine steeplechaser
- Professoring pelota everywhere uncomptable Cryptograms acreage rhythmites peplum archopterix antirendowment Slimbe superauditor pegasi Colubrina unbind incur Epiphippus hydropod unexcusable antimeristaltic nonpar ligense relationship hyperdoliocranial suumer penicilate epette semilune gelatose
- Prefabricator sightworthy deoxidative Lug Archaeannelida pneumococcal Diposaurus Mcchlin navette pachytic
- Prayer preductorship pistolwise bolutiform Tor uncontenment badgeringly monistical hirse tymphatic Aurelius
- Xanthimide stime sharesman

- Misplose Cocystis guanamine plagiarization Beragonism deliveress Apatea existentialism geology pneumatic
- Bragi rabbinish Angean valerate watchwork irreality trochleoid nonwetter neoglosissia mottilike
- Tachometer protospermifer diplostichous Herpetomana angiomya anunciaracy roky sub sequential
- Macrodactylis beastship athanae Archidiskodon Sastean symphony Paraguayian parodiabel diploy washthroat
- Cerberic hambry overissue trussel sociocrat epicylidal probate unforecasted snoozy sardine pachycephary
- Needs slogan mercantile fictioneer metromony radiational nimblebrained pyrazolyl workable overleave
- Hydroflate impress uninjuarous operativity sandbagger querten compositograph necrotyle

- Darkness Jarkman uncompilable metdehyde anhydric unconfirmative rase microcrube polymedscriptious fizz
- Ungraphic tuille necronaul bhaichari rebucker rhadosphere celocentesis valleylike bureaucratic
- Contemplobility longe impeachment Favorunite vegetomize areset hyperspatial ectocardia licroconite bullyt
- Malacan taglike slabbary unreverberated grampus disquiet Saumya suppot hemoterinometer whereas
- Tyrannicalness Gasterophillus forcing allegomy glaringness psychometry sleepproof gangboard apologetic bulb
- Gatetender apothematisal Cystaha Euchlorophycae altair meltriose inadvertence alter-root soaky decouman
- Bube exactment Diax comparably Hydriote Fellata antialien
Text extraction benchmark - motivating comparison

Order - do text snippets appear in the same order?
Text extraction benchmark - motivating comparison

mispolc Qocysl guanine pleiotropic Berenson deliveress Apatela existentialism geology pneumonic Bragi rabbobition Augen valerate watchwork irrealty trochoidal nemuterine hengholmstia mollusk tachometer proteohemipener diplodichous Herpetonomas angioyma annunciation roky subsistent macrodactylic beetship athanas Archidiskodon Sastean symphony Paraguayean parodible diapoy washthrough Cerebratic harbury overissue trussel socicrat epicycoidal probate unforecated snozy sardine pachychaphy needs slogan mercantile fictioneer metrometry radiatinal nimblebrained pyrazolyl workable overleaves hydroflate impress uninjurious operativity sandbagger querten cocomanogap necrotypes

Percival rational disensu pulmonary precontemporary jakthine Genista ensuising sympathmatic radiometer pinja graduating unwinning babar antiquiticy Lachanella roentgenoscope andantino doornard pristine steeplechaser professoring pelota everywhere uncatable Cryptopasia acrege rhythymes peplum orchestron art:params stibias superauditor pegaoid Colbrina unbind incur Epiphippus hydrogen unexcusable antimaleriatric nonpar ligance relationship hyperbolohicranal sumpnic penctillate ephele semilune gelatose prefabricator sightworthy deoxidative lug Archianellids pneuromococal Dispensaurus Meclin navette pachytic prayer pretectorship pistolwise botulism Tor uncontentious badgeringly monstalonic hircine cynophylic Aurellius xanthanide stime sharesman

<table>
<thead>
<tr>
<th>allogene</th>
<th>Cephalopoda</th>
<th>athletical</th>
<th>bottler</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>36</td>
<td>65</td>
<td>100</td>
</tr>
<tr>
<td>53</td>
<td>98</td>
<td>64</td>
<td>100</td>
</tr>
<tr>
<td>100</td>
<td>87</td>
<td>36</td>
<td>53</td>
</tr>
<tr>
<td>11</td>
<td>9</td>
<td>97</td>
<td>45</td>
</tr>
</tbody>
</table>

darkness jarkman uncompilable metaldehyde anhydric unconferrable rase microbeber polyamodioeous fizz ungovicus tuille neurocranal bhiachari rebuka rhadobiophere celenticelis valleylike bureuractize contempitability longe impenetration Faventine vegetatize aereate hyperspatial ectocardia licoconite bullety malcon tagliky sallbery unperverberated grampus diapoy Saunsa support hematoanotis whereas tyrannicalness Gasterophilus forcing allogamy glariness psychrometry sleepproof gangboard apollagric bulb gatetender apothegmatical Cyatha Euclorophyceae Altair melitriose inadvertence aloeroot soaky decocyan Bube exactman Olax comparably Hydrateellata antabilen

dermision Qocysl guanine pleiotropic Berenson deliveress Apatela existentialism geology pneumonic Bragi rabbobition Augen valerate watchwork irrealty trochoidal nemuterine hengholmstia mollusk tachometer proteohemipener diplodichous Herpetonomas angioyma annunciation roky subsistent macrodactylic beetship athanas Archidiskodon Sastean symphony Paraguayean parodible diapoy washthrough Cerebratic harbury overissue trussel socicrat epicycoidal probate unforecated snozy sardine pachychaphy needs slogan mercantile fictioneer metrometry radiatinal nimblebrained pyrazolyl workable overleaves hydroflate impress uninjurious operativity sandbagger querten cocomanogap necrotypes

Percival rational disensu pulmonary precontemporary jakthine Genista ensuising sympathmatic radiometer pinja graduating unwinning babar antiquiticy Lachanella roentgenoscope andantino doornard pristine steeplechaser professoring pelota everywhere uncatable Cryptopasia acrege rhythymes peplum orchestron art:params stibias superauditor pegaoid Colbrina unbind incur Epiphippus hydrogen unexcusable antimaleriatric nonpar ligance relationship hyperbolohicranal sumpnic penctillate ephele semilune gelatose prefabricator sightworthy deoxidative lug Archianellids pneuromococal Dispensaurus Meclin navette pachytic prayer pretectorship pistolwise botulism Tor uncontentious badgeringly monstalonic hircine cynophylic Aurellius xanthanide stime sharesman

<table>
<thead>
<tr>
<th>allogene</th>
<th>Cephalopoda</th>
<th>athletical</th>
<th>bottler</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>36</td>
<td>65</td>
<td>100</td>
</tr>
<tr>
<td>53</td>
<td>98</td>
<td>64</td>
<td>100</td>
</tr>
<tr>
<td>100</td>
<td>87</td>
<td>36</td>
<td>53</td>
</tr>
<tr>
<td>11</td>
<td>9</td>
<td>97</td>
<td>45</td>
</tr>
</tbody>
</table>

darkness jarkman uncompilable metaldehyde anhydric unconferrable rase microbeber polyamodioeous fizz ungovicus tuille neurocranal bhiachari rebuka rhadobiophere celenticelis valleylike bureuractize contempitability longe impenetration Faventine vegetatize aereate hyperspatial ectocardia licoconite bullety malcon tagliky sallbery unperverberated grampus diapoy Saunsa support hematoanotis whereas tyrannicalness Gasterophilus forcing allogamy glariness psychrometry sleepproof gangboard apollagric bulb gatetender apothegmatical Cyatha Euclorophyceae Altair melitriose inadvertence aloeroot soaky decocyan Bube exactman Olax comparably Hydrateellata antabilen
Structure - is the structure of a text snippet preserved?
### Text extraction benchmark - performance measures

<table>
<thead>
<tr>
<th>Tool level measures</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Correctness</strong></td>
<td>Average proportion of correct words (per document)</td>
</tr>
<tr>
<td><strong>Order</strong></td>
<td>Percentage of documents with correct order of text snippets</td>
</tr>
<tr>
<td><strong>Structure</strong></td>
<td>Average proportion of text snippets with correct structure (per document)</td>
</tr>
</tbody>
</table>
Text extraction benchmark - test case design (test features)

- Building a feature tree
- **15 features** identified
Text extraction benchmark - test case design
(feature value distributions)

- Govdocs collection
  - 76,616 doc files
  - 75,405 processed
  - features:
    - number of pages
    - number of paragraphs
    - number of lines
    - number of words
    - number of characters
    - number of tables

(a) page distributions
(b) paragraph distributions
(c) table distributions
(d) word distributions
Text extraction benchmark - test case design (representative samples)

- Representative samples:
  - number of pages, paragraphs, words and tables
  - **345 samples** selected
Text extraction benchmark - test models definitions (PIMM)