Image tasks @ CLEF-IP 2012

Mihai Lupu, Florina Piroi, Allan Hanbury, Igor Filippov, Alan Sexton
Overview

- Chemical structure recognition results
- Flowchart recognition results
- Next steps
Chemical Structure Evaluation

Mihai Lupu
Igor Filippov
Alan Sexton
Chemical structure recognition

• Good results for both tasks
  – Segmentation

<table>
<thead>
<tr>
<th>Tolerance</th>
<th>Precision</th>
<th>Recall</th>
<th>F_1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.70803</td>
<td>0.68622</td>
<td>0.69696</td>
</tr>
<tr>
<td>10</td>
<td>0.79311</td>
<td>0.76868</td>
<td>0.78070</td>
</tr>
<tr>
<td>20</td>
<td>0.82071</td>
<td>0.79543</td>
<td>0.80787</td>
</tr>
<tr>
<td>40</td>
<td>0.86696</td>
<td>0.84025</td>
<td>0.85340</td>
</tr>
<tr>
<td>55</td>
<td>0.88694</td>
<td>0.85962</td>
<td>0.87307</td>
</tr>
</tbody>
</table>

– Recognition

<table>
<thead>
<tr>
<th></th>
<th>Automatic Set</th>
<th></th>
<th>Manual Set</th>
<th></th>
<th>Total</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#Structures Recalled</td>
<td>%</td>
<td>#Structures Recalled</td>
<td>%</td>
<td>#Structures Recalled</td>
<td>%</td>
</tr>
<tr>
<td>saic</td>
<td>865</td>
<td>761</td>
<td>88%</td>
<td>95</td>
<td>38</td>
<td>40%</td>
</tr>
<tr>
<td>uob-1</td>
<td>865</td>
<td>832</td>
<td>96%</td>
<td>95</td>
<td>44</td>
<td>46%</td>
</tr>
<tr>
<td>uob-2</td>
<td>865</td>
<td>821</td>
<td>95%</td>
<td>95</td>
<td>56</td>
<td>59%</td>
</tr>
<tr>
<td>uob-3</td>
<td>865</td>
<td>821</td>
<td>95%</td>
<td>95</td>
<td>44</td>
<td>46%</td>
</tr>
<tr>
<td>uob-4</td>
<td>865</td>
<td>832</td>
<td>96%</td>
<td>95</td>
<td>54</td>
<td>57%</td>
</tr>
</tbody>
</table>
Chemical structure recognition

• Good results for both tasks
  – Segmentation
  – Recognition

Caveat

The test structures have been pre-selected to have a INCHI representation (i.e. no Markush, no ‘fancy stuff’)

<table>
<thead>
<tr>
<th></th>
<th>said</th>
<th>uob-1</th>
<th>uob-2</th>
<th>uob-3</th>
<th>uob-4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>968</td>
<td>865</td>
<td>865</td>
<td>865</td>
<td>865</td>
</tr>
<tr>
<td></td>
<td>862</td>
<td>852</td>
<td>821</td>
<td>832</td>
<td>832</td>
</tr>
<tr>
<td></td>
<td>96%</td>
<td>95%</td>
<td>95%</td>
<td>96%</td>
<td>96%</td>
</tr>
<tr>
<td></td>
<td>96</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>41</td>
<td>56</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>46%</td>
<td>59%</td>
<td>57%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>91%</td>
<td>90%</td>
<td>90%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16%</td>
<td>28%</td>
<td>3%</td>
<td>17%</td>
<td>2%</td>
</tr>
</tbody>
</table>
Flowchart Recognition Evaluation

Mihai Lupu
Florina Piroi
Allan Hanbury
Results (so far)

- Most common sub-graph
Results (so far)

- Node type match
Results (so far)

- Text recognition (Edit Distance) - average
Results (so far)

• Text recognition (Edit Distance) - sum

![Bar chart showing text recognition results for various datasets.](chart.png)
Results (so far)

- Text recognition (Edit Distance) - normalized
Caveat

- 44 of 100 topics evaluated
  - Only those where a score was computed for all participating runs
  - ‘selected’ by the algorithm (by not finishing in reasonable time or crashing)

<table>
<thead>
<tr>
<th></th>
<th>Flowcharts in results</th>
<th>Flowcharts not in results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Min</strong></td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td><strong>Max</strong></td>
<td>21</td>
<td>51</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>13</td>
<td>22</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>12.41</td>
<td>24.98</td>
</tr>
<tr>
<td><strong>std.dev</strong></td>
<td>3.62</td>
<td>8.9</td>
</tr>
</tbody>
</table>
Examples of finished evaluations

[FIG. 2]

START
Enter album pictures. $100
Enter registered person's information. $102
Enter picture place information. $104
Enter subject information. $106
Generate additional information according to result of input process. $108
Register album pictures related to additional information. $110

END

FIG. 14A

START
CHECK PAPER SIZE
CHANGE?
YES
NO
SET PAPER SIZE
CHANGE BIT

CHECK PRESENCE OR ABSENCE OF PAPER
CHANGE?
YES
NO
SET PAPER PRESENCE/ABSENCE CHANGE BIT

CHECK FUNCTION OF PAPER FEEDING UNIT
CHANGE?
YES
NO
SET FUNCTION CHANGE BIT OF PAPER FEEDING UNIT

FIG. 6

END

FIG. 2

START
READ ENGINE OPERATION PARAMETERS
S10
Determine basic ignition timing GMAP
S12
Determine various correction values including heat
S14
ENGINE UNDER PREFERRED COLD STARTING CONDITION
S16
Y
S18
HEAT=0

Determine output ignition timing GLOG
S20
OUTPUT GLOG
S22
END

OPERATION AT ACD CONTROLLER

CALL INITIATED BY REMOTE TELEPHONE STATION?
S100
YES
NO
ANSWER CALL Top via ITL 323 PROTOCOL S112
SEND STORRED GREETING MESSAGE TO REMOTE TELEPHONE STATION S124
QUERY REMOTE TELEPHONE STATION CAPABILITIES S128
SEND CONTROL SIGNALS TO REMOTE TELEPHONE STATION USER INTERFACE S132
SEND WELCOME SCREEN MESSAGE TO REMOTE TELEPHONE STATION S136
FORWARD CALL TO ATTENDANT S138

IS ATTENDANT AVAILABLE?
S140
YES
A
NO
S144

FIG. 6
Measures

• Idea / Motivation
  – The process depicted is important
    • Links between nodes
      – Graph structure
        » Most Common Sub-graph
  – Node types are less important
    • Only evaluate after structure recognition
  – Text is important
    • Provided that the structure is recognized
Measures (cont.)

• Most common sub-graph (MCS)
  – It’s the largest sub-graph common to all graphs in a set of graphs
  – McGregor algorithm
    • Backtracking
    • High computational costs
  – Modified to find all variants of the most common sub-graph
    • Because (we think) filtering on node type would be too restrictive
    • Even more complex
• **Node type match & Edit distance**
  – Taken separately as the best of all different variants of largest common sub-graphs
  – E.g. if for topic X, 5 different ways to match nodes of run Y were found (all having a score of 0.7 in the MCS), compute the node type match & edit distance for each
    – Node-type match : 0.5 0.5 0.9 0.9 0.5
    – Edit distance: 100 100 90 120 90
  • Then the result scores are:
    – Node match: 0.9
    – Edit distance: 90
Edit Distance

• Smaller is better
  – Smaller?
    • no nodes were returned
    • Actually good match
  – Only comparable in relation with MCS
What’s up next?

• But first...
Information Retrieval Special Issue on IR in the Intellectual Property Domain

• Goal: present cutting-edge research results on open topics related to IR in the Intellectual Property domain in order to advance the current state-of-the-art

• Submissions encouraged to make use of evaluation campaign datasets: CLEF-IP 2011/2012, TREC-CHEM 2011 and NTCIR

• Submissions due: **15 March 2013**

• More info: [https://sites.google.com/site/sipatentir](https://sites.google.com/site/sipatentir)
What’s up next?

• Overview and plans
Image retrieval in patents

- Must be approached by image type
- Step 1:
  - classification

Figure 1: Examples of types of figures in patents
Figure 1: Examples of types of figures in patents
Mathematical formulas

\[
\frac{\Delta \omega}{\omega_t} = \frac{\omega_t - \omega_{so}}{\omega_t} \approx \frac{ms^2}{ms^2 + \frac{km}{b} s + k},
\]
Chemistry
Genetic data

HindIII
5' GATCGATTAAGCTTCGGAACCGCTCGATCTCGCCGCC 3' Oligonukleotid A

AGCCCCACGGCCGACGGAACCCCTCGATCTCGCCGCACTGGTAGACATGAG

110

TCGGGTTGCCCACCGTGTGCTTGGCGAGCTAGAGCGGCGTTGACCATCTGTACCTC

MetGlu

5'-untranslatiert  Beginn

Leserahmen
(Signalpeptid)
Plots/Graphs

FIG. 7
402 class HasTable$Checked : IDictionary {
404    void HasTable$Invariant() {
405        ASSERT([constraint clause from Fig. 3]);
406    }
407    void set$Pre (Object key, Object val) { ASSERT(true); }
408    void set$Post (Object key, Object value, Object result) {
410        ASSERT([ensure clause from Fig. 3]);
412    }
414    Object[] keys; Object[] values;
418    Object set (Object key, Object value) {
420        Object result ;
422        HasTable$Invariant();
424        set$Pre(key, value);
426        try {
428            [body of the set method from the implementation code]
432            <return value / result = value; break END;>
434            } catch (Exception e) {
436                result = e;
438            }
440        } END :
444        HasTable$Invariant();
446        set$Post(key, value, result );
448        if (result is Exception) throw result ; else return result ;
450    } }
Characters/symbols
**Tables**

<table>
<thead>
<tr>
<th>Sample No.</th>
<th>Total amount of network-forming oxides (ppm)</th>
<th>Calcining temp. (°C)</th>
<th>Firing temp. (°C)</th>
<th>Mechanical strength (Kgf/cm²)</th>
<th>Resistance to chemicals (%)</th>
<th>Moisture resistance (%)</th>
<th>Glass phase at grain boundary</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>230</td>
<td>970</td>
<td>1210</td>
<td>1663</td>
<td>.3</td>
<td>-5</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>190</td>
<td>950</td>
<td>1170</td>
<td>1444</td>
<td>.5</td>
<td>-8</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>630</td>
<td>810</td>
<td>1100</td>
<td>1559</td>
<td>.1</td>
<td>-6</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>460</td>
<td>1000</td>
<td>1280</td>
<td>1695</td>
<td>.1</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>400</td>
<td>840</td>
<td>1210</td>
<td>1396</td>
<td>.8</td>
<td>-8</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>120</td>
<td>850</td>
<td>1230</td>
<td>2039</td>
<td>0</td>
<td>0</td>
<td>No</td>
</tr>
</tbody>
</table>