AUTOMATED KOS-BASED SUBJECT INDEXING IN INIS

Zaven Hakopov
Dmitry Mironov
Dobrica Savic
Yulia Svetashova
ATOMS FOR PEACE AND DEVELOPMENT
How the IAEA supports the Sustainable Development Goals
INTERNATIONAL NUCLEAR INFORMATION SYSTEM (INIS)

155 INIS members

131 Member States
24 international organizations

4 objectives
Acquire, process, preserve literature on peaceful uses of nuclear science and technology
Provide free and open access to nuclear information
Develop and maintain a nuclear knowledge organization system
Assist IAEA Member States in building their scientific information management capacities

INIS repository
4.1 million bibliographic records
104,000 in 2017

540,000 full-text documents
8,000 in 2017

2.9 million web page views
1.6 million unique searches
1 million unique visitors
Resulting subject analysis often contains:

- Too broad descriptors
- Misleading suggestions
- Incorrect interpretations

Years of experience:

- 45 years
- 4 millions indexed records

Expert system:

- INIS Thesaurus
TWO-PASS INDEXING, TIER 1

840 unique rules
- Add
- Replace
- Remove
- Keep

KOS integration
- Thesaurus
- INIS subjects
- Expert knowledge
- Quality control

840

Frequent corrections
- Specialist experience
<table>
<thead>
<tr>
<th>ID</th>
<th>Number of records</th>
<th>Descriptors</th>
<th>Records where more than one descriptor were</th>
<th>Operations per record performed by human after…</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Total after…</td>
<td>added by human after…</td>
<td>removed by human after…</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2280</td>
<td>955</td>
<td>864</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2280</td>
<td>955</td>
<td>864</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2280</td>
<td>955</td>
<td>864</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>2280</td>
<td>955</td>
<td>864</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>2280</td>
<td>955</td>
<td>864</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>2280</td>
<td>955</td>
<td>864</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>2280</td>
<td>955</td>
<td>864</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>2280</td>
<td>955</td>
<td>864</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>2280</td>
<td>955</td>
<td>864</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>2280</td>
<td>955</td>
<td>864</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>2280</td>
<td>955</td>
<td>864</td>
</tr>
</tbody>
</table>

Average:

- Descriptors:
  - Total after…: 2280
  - added by human after…: 955
  - removed by human after…: 864
  - added after…: 114
  - removed after…: 63
- Operations per record performed by human after…:
  - CAI: 1530
  - RUBAI: 155
  - Human: 18.93
  - CAI: 23.58
  - RUBAI: 1.79
  - Human: 24.18
  - CAI: 20.41
  - RUBAI: 1.66
  - Human: 14.22
  - CAI: 16.44
  - RUBAI: 2.27
  - Human: 16.44

Average:

- Descriptors:
  - Total after…: 2280
  - added by human after…: 955
  - removed by human after…: 864
  - added after…: 114
  - removed after…: 63
- Operations per record performed by human after…:
  - CAI: 1530
  - RUBAI: 155
  - Human: 18.93
  - CAI: 23.58
  - RUBAI: 1.79
  - Human: 24.18
  - CAI: 20.41
  - RUBAI: 1.66
  - Human: 14.22
  - CAI: 16.44
  - RUBAI: 2.27
  - Human: 16.44
Operations per record needed by human after CAI

- Set of records
- Number of operations per record

- CAI
- RUBAI
ARE WE DOING WELL?

TRIGA TYPE REACTORS
REACTORS
NEUTRONS
NUCLEAR FUELS
MAGNESIUM 20
REACTOR DESIGN
RESEARCH REACTORS
LEAD
FISSION PRODUCTS
NEUTRON BEAMS
RADIOGRAPHY
MELANOMAS

TRIGA TYPE REACTORS
NUCLEAR FUELS
REACTOR DESIGN
RESEARCH REACTORS
FISSION PRODUCTS
NEUTRON BEAMS
INDUSTRIAL RADIOGRAPHY
MELANOMAS
NEUTRON CAPTURE THERAPY
ISOTOPE PRODUCTION
TRIGA-2-MUSASHI REACTOR
### FROM A SINGLE DECISION TO A DATASET

**Dataset**

<table>
<thead>
<tr>
<th>Category</th>
<th>CMV</th>
<th>Weight</th>
<th>Title Match</th>
<th>Operation</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>RUBAI</td>
<td>60</td>
<td>-0.0211</td>
<td>0</td>
<td>2: remove</td>
<td>1: approve</td>
</tr>
<tr>
<td>Human Action</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human Operation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human Decision</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRIGA TYPE REACTORS</td>
<td>Keep</td>
<td>Keep</td>
<td>Keep</td>
<td>Approve</td>
<td></td>
</tr>
<tr>
<td>REACTORS</td>
<td>Remove</td>
<td>Remove</td>
<td>Keep</td>
<td>Approve</td>
<td></td>
</tr>
<tr>
<td>RESEARCH REACTORS</td>
<td>Keep</td>
<td>Keep</td>
<td>Approve</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LEAD</td>
<td>Remove</td>
<td>Remove</td>
<td>Approve</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RADIOGRAPHY</td>
<td>Add</td>
<td>Remove</td>
<td>Reject</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MELANOMAS</td>
<td>Add</td>
<td>Remove</td>
<td>Reject</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEUTRON CAPTURE THERAPY</td>
<td>Add</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISOTOPE PRODUCTION</td>
<td>Add</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INDUSTRIAL RADIOGRAPHY</td>
<td>Add</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Features</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Label</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
TWO-PASS INDEXING, TIER 2

CAI → RUBAI → RUBAI-ML → Human

840 unique rules

Add
Replace
Remove
Keep

84% accuracy
37,000 unique decisions

840 unique decisions

DATA analysis
Machine Learning
KOS
Modeling human cognition

DATA set
5,600 documents
37,000 unique decisions

INIS subjects
Expert knowledge
Quality control
Integration
KOS
Human

CAI
RUBAI
RUBAI-ML
Human
### Model evaluation results

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>Specificity</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.837</td>
<td>0.838</td>
<td>0.793</td>
<td>0.873</td>
<td>0.815</td>
</tr>
</tbody>
</table>

### Total descriptors after...

<table>
<thead>
<tr>
<th>ID</th>
<th>Number of records</th>
<th>Total after RUBAI-ML</th>
<th>Operations per record performed by human after RUBAI-ML</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CAI</td>
<td>RUBAI</td>
<td>CAI</td>
</tr>
<tr>
<td>1</td>
<td>120</td>
<td>2519</td>
<td>1032</td>
</tr>
<tr>
<td>2</td>
<td>111</td>
<td>2104</td>
<td>859</td>
</tr>
<tr>
<td>3</td>
<td>109</td>
<td>2217</td>
<td>916</td>
</tr>
<tr>
<td>4</td>
<td>22</td>
<td>474</td>
<td>169</td>
</tr>
</tbody>
</table>

... CLOSER TO THE HUMAN CHOICES
Identifies missing descriptors in case of:

- Specific semantic relations
- Narrower descriptors
- Broader document scope
840 unique rules
Add
Replace
Remove
Keep

Integration
Thesaurus
INIS subjects
Expert knowledge
Quality control

KOS

Machine Learning
KOS
Modeling human cognition

Work experience
Expert knowledge
KOS integration
Data analysis
Decision simulation

DATA set
5,600 documents
84% accuracy
37,000 unique decisions
Main outcomes:
• Model human decision-making strategies
• Gradual optimization of the indexing quality

Future work:
• Extend the coverage and consistency of rules
• Improve ML validation component
• OR: replace rules and ML validation with a single ML solution
• DREAM: completely data-driven classification algorithm based on (convolutional) neural networks