A software processing chain for evaluating thesaurus quality

Javier Lacasta, Gilles Falquet, Javier Nogueras-Iso, and Javier Zarazaga-Soria
Cluj-Napoca Romania, 8-9 September 2016

Computer Science and Systems Engineering Dept., Universidad de Zaragoza, Spain.
Centre Universitaire d’Informatique, Universite de Geneve, Switzerland
Quality in thesauri

- The “quality” is a measure of excellence or a state of being free from defects, deficiencies and significant variations (ISO 8402).
- ISO 25964 defines the structure, properties and relations of thesauri.
  - Mandatory and optional properties (preferred labels, definitions).
  - Structure of the content (charset, acronyms use,…).
  - Rules to obtain homogeneity along the thesaurus.
  - Proper use of properties and relations.
- Detecting the fulfilment of these features requires lexical, syntactic and semantic analysis of the content of the thesaurus.
- We have developed a tool that identifies problems in any of these elements and it generates a report detailing the problems found.
Validations performed

- **Property analysis:**
  - Detection of incomplete preferred labels and definitions.
  - Detection of non-alphabetic characters, adverbs, initial articles, and acronyms (in preferred labels).
  - Detection of duplicated labels and inconsistencies in the use of uppercase and plurals.
  - Detection of syntactically complex labels (analysis of the use of prepositions, conjunctions and adjectives).

- **Relation analysis:**
  - Detection of BT/NT cycles.
  - Detection of non-informative RT relations (in the same BT/NT hierarchy).
  - Detection of semantically invalid BT/NT relations (without a subordinate-superordinate meaning).
Validation tool

- Modular architecture
  - Composition of validation modules, each one focused on reviewing a single feature of the thesaurus.
  - Adding a new validation only requires to define a new component that does the task.
  - Independent tasks can be executed in parallel.

- Different types of validators
  - Thesaurus level: Analyze the thesaurus as a whole, each reviewed element requires the others as context to determine its correctness.
  - Concept level: The analysis requires information of multiple properties inside the processed concept to determine the correctness. It is independent of other concepts.
  - Label level: Focused on a label, the result is independent of the rest of the thesaurus.
Thesaurus level validators

- BT/NT cycle analysis.
  - Tarjan's strongly connected components algorithm.
- RT relevance analysis.
  - Reviewing BTs of concepts in RT.
- Preferred label uniqueness analysis
  - Using a set structure.
Concept level validators

- Definition, BT/NT, Preferred label completeness.
  - Simple existence check.

![Diagram showing the flow of concept level validators]

- Def Completeness analysis
- BT/NT Completeness analysis
- Preferred label Completeness analysis

Analysis result integration → Annotated Thesaurus
Label level validators

- Detection of non-alphabetic characters, acronyms, and uppercase.
  - Regular expressions.
- Plural detection: Adapted Porter stemming algorithm.
- Conjunction, adverb, article, prepositional phrase, complexity: POS tagging.
- Alignment to WordNet: String match ignoring plurals and case (multiple synsets).
Label level validators, result integration

- Plural and uppercase analysis
  - Detection of inconsistencies.
- BT/NT correctness analysis.
  - Disambiguation of WordNet senses.
  - Alignment to DOLCE ontology to identify subordinate/superordinate meaning.
BT/NT correctness analysis

- Language and structure filtering: Selection of WordNet senses in base to the concept labels and the context of previously aligned ones.
- BT/NT analysis: Match with DOLCE ontology and identification of the relation meaning.
  - Subclass, participation, location relations have a subordinate meaning compatible with BT/NT relation.
Tool implementation

- Use of Spring framework.
  - Facilitates the use of the dependency-injection pattern to define decoupled components.
  - Facilitates the parallel execution of the decoupled components.

```xml
<batch:step id="bt_nt_correctnessAnalysis">
  <batch:tasklet task-executor="simpleTaskExecutor" throttle-limit="20">
    <batch:chunk reader="modelMemory_reader" processor="detectIncorrect_bt_nt_processor"
                writer="empty_writer" commit-interval="1" />
  </batch:tasklet>
</batch:step>
<beans:bean id="simpleTaskExecutor" class="org.springframework.core.task.SimpleAsyncTaskExecutor"/>
```

- Sequential implementation:
  - Urbamet: 85 seconds, Gemet: 261 seconds

- Parallel implementation:
  - Urbamet: 41 seconds, Gemet: 133 seconds
## Experiments

<table>
<thead>
<tr>
<th>Thesaurus</th>
<th>Concepts</th>
<th>PL(en)</th>
<th>PL(es)</th>
<th>PL(fr)</th>
<th>AL(fr)</th>
<th>Def(en)</th>
<th>BT-NT</th>
<th>RT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urbamet</td>
<td>3844</td>
<td>3844</td>
<td>3845</td>
<td>3846</td>
<td>504</td>
<td>0</td>
<td>3821</td>
<td>0</td>
</tr>
<tr>
<td>Gemet</td>
<td>5244</td>
<td>5244</td>
<td>5244</td>
<td>5244</td>
<td>0</td>
<td>4909</td>
<td>5332</td>
<td>1043</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measure name</th>
<th>Urbamet value</th>
<th>Gemet value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Property completeness analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complementness and uniqueness of preferred labels</td>
<td>99.97%</td>
<td>100%</td>
</tr>
<tr>
<td>Completeness of definitions</td>
<td>0%</td>
<td>31.20%</td>
</tr>
<tr>
<td>Property content analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-existence of non-alphabetic characters in labels</td>
<td>99.94%</td>
<td>99.92%</td>
</tr>
<tr>
<td>Non use of adverbs / initial articles in labels</td>
<td>97.54%</td>
<td>98.45%</td>
</tr>
<tr>
<td>Non use of acronyms in preferred labels</td>
<td>99.24%</td>
<td>99.28%</td>
</tr>
<tr>
<td>Property context analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-existence of duplicated labels</td>
<td>99.77%</td>
<td>98.83%</td>
</tr>
<tr>
<td>Consistent use of uppercase in labels</td>
<td>99.35%</td>
<td>97.60%</td>
</tr>
<tr>
<td>Consistent use of plurals in labels</td>
<td>94.50%</td>
<td>88.72%</td>
</tr>
<tr>
<td>Property complexity analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non use of prepositional phrases in labels(en)</td>
<td>90.01%</td>
<td>94.58%</td>
</tr>
<tr>
<td>Non use of too long noun phrases / conjunctions</td>
<td>99.30%</td>
<td>99.24%</td>
</tr>
<tr>
<td>Relation coherence analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Informative RTs</td>
<td>NA</td>
<td>93.67%</td>
</tr>
<tr>
<td>Completeness of BT/NT</td>
<td>99.40%</td>
<td>100%</td>
</tr>
<tr>
<td>Non-existence of BT/NT cycles</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Semantic correctness of BT/NT</td>
<td>69.70%</td>
<td>75.76%</td>
</tr>
</tbody>
</table>
Validation of results

- Manual review of a branch to detect false positives and negatives
  - Urbamet: 208 Concepts
  - Gemet: 310 concepts

### Urbamet quality measure

<table>
<thead>
<tr>
<th>Feature</th>
<th>tp</th>
<th>tn</th>
<th>fp</th>
<th>fn</th>
<th>na</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonuse of adverbs / initial articles in labels</td>
<td>99.67%</td>
<td>0%</td>
<td>0%</td>
<td>0.23%</td>
<td></td>
<td>99.67%</td>
</tr>
<tr>
<td>Consistent use of plurals in labels</td>
<td>97.75%</td>
<td>2.08%</td>
<td>0%</td>
<td>0.16%</td>
<td></td>
<td>99.83%</td>
</tr>
<tr>
<td>Nonuse of prepositional phrases in labels(en)</td>
<td>87.01%</td>
<td>12.98%</td>
<td>0%</td>
<td>0%</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>Nonuse of too long noun phrases / conjunctions</td>
<td>99.16%</td>
<td>0.50%</td>
<td>0.33%</td>
<td>0%</td>
<td></td>
<td>99.66%</td>
</tr>
<tr>
<td>Semantic correctness of BT/NT</td>
<td>64.73%</td>
<td>12.56%</td>
<td>4.83%</td>
<td>14.49%</td>
<td>3.38%</td>
<td>77.29%</td>
</tr>
</tbody>
</table>

### Gemet quality measure

<table>
<thead>
<tr>
<th>Feature</th>
<th>tp</th>
<th>tn</th>
<th>fp</th>
<th>fn</th>
<th>na</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonuse of adverbs / initial articles in labels</td>
<td>99.73%</td>
<td>0.24%</td>
<td>0%</td>
<td>0.02%</td>
<td></td>
<td>99.98%</td>
</tr>
<tr>
<td>Consistent use of plurals in labels</td>
<td>92.91%</td>
<td>6.23%</td>
<td>0%</td>
<td>0.86%</td>
<td></td>
<td>99.14%</td>
</tr>
<tr>
<td>Nonuse of prepositional phrases in labels(en)</td>
<td>94.58%</td>
<td>5.42%</td>
<td>0%</td>
<td>0%</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>Nonuse of too long noun phrases / conjunctions</td>
<td>99.90%</td>
<td>0.10%</td>
<td>0%</td>
<td>0%</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>Semantic correctness of BT/NT</td>
<td>88.39%</td>
<td>1.30%</td>
<td>9.67%</td>
<td>0.64%</td>
<td>0%</td>
<td>89.69%</td>
</tr>
</tbody>
</table>
Conclusions

- We have developed a tool to validate thesauri.
- Its modular architecture facilitates extension and use:
  - The addition of new validation components is simple.
  - Independent validations are executed in parallel.
  - It can be used as a final application, but it is easy to integrate in other applications or services.
    - Each validation module can be used individually.
- The results obtained in the experiments have shown a suitable behavior with a reasonable number of false positives and negatives.