KEYSTONE Conference 2016

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

IAAA
Advanced Information Systems Laboratory
Universidad Zaragoza

UNIVERSITÉ DE GENÈVE

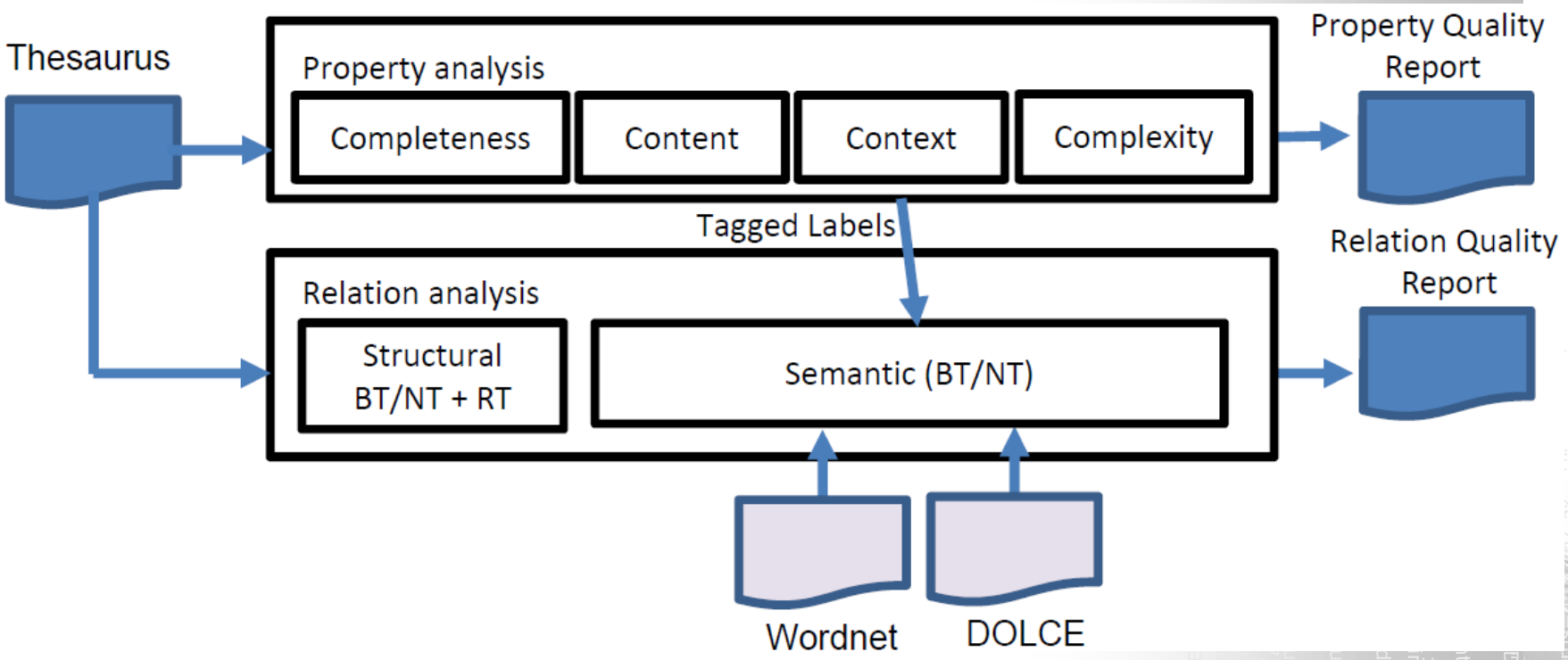
Quality in thesauri

- 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 process



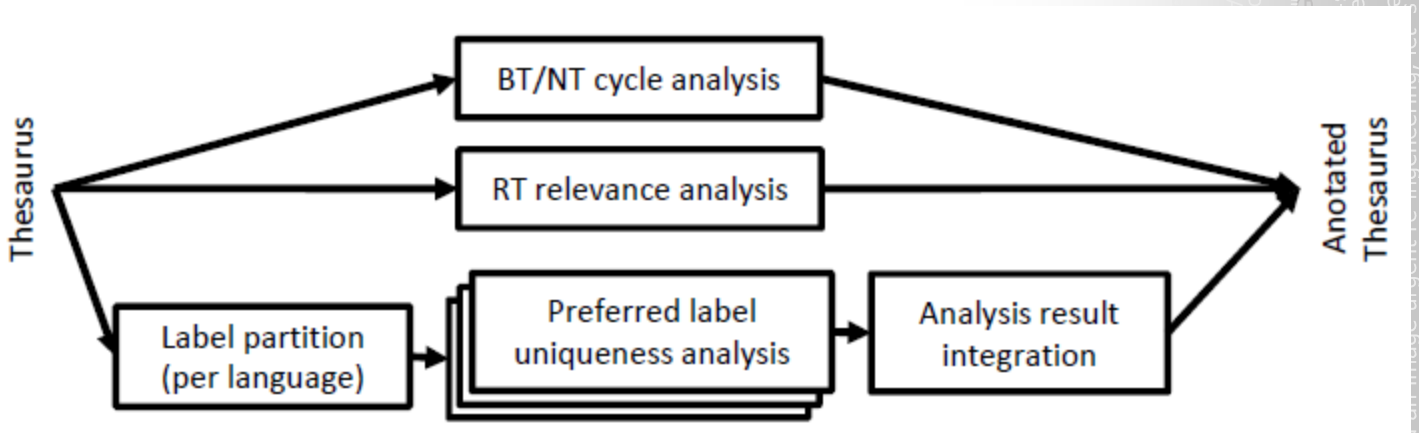
- An automatic method for reporting the quality of thesauri. Data & Knowledge Engineering Volume 104, July 2016, Pages 1–14.

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.

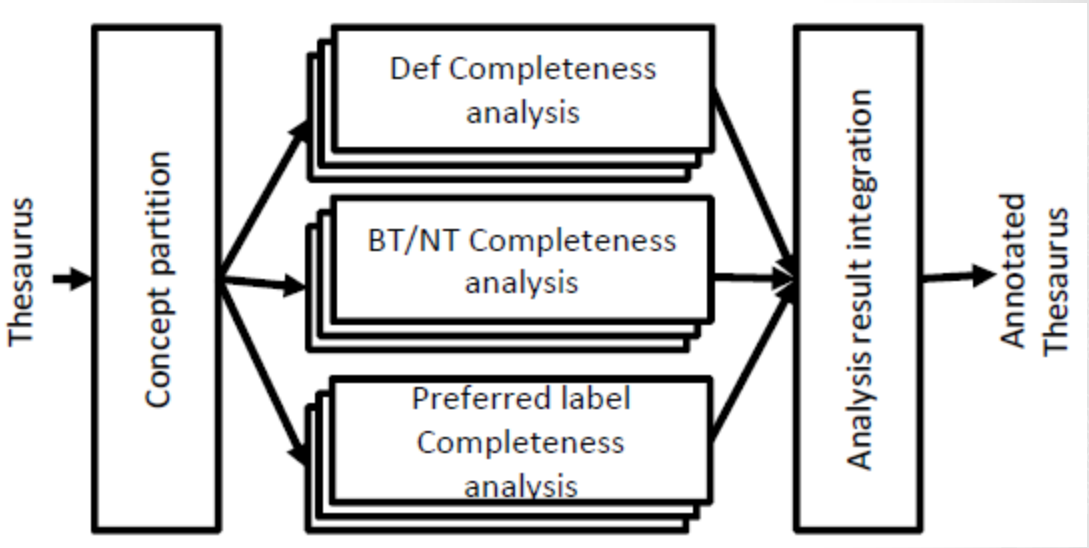
Thesaurus level validators

- 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.



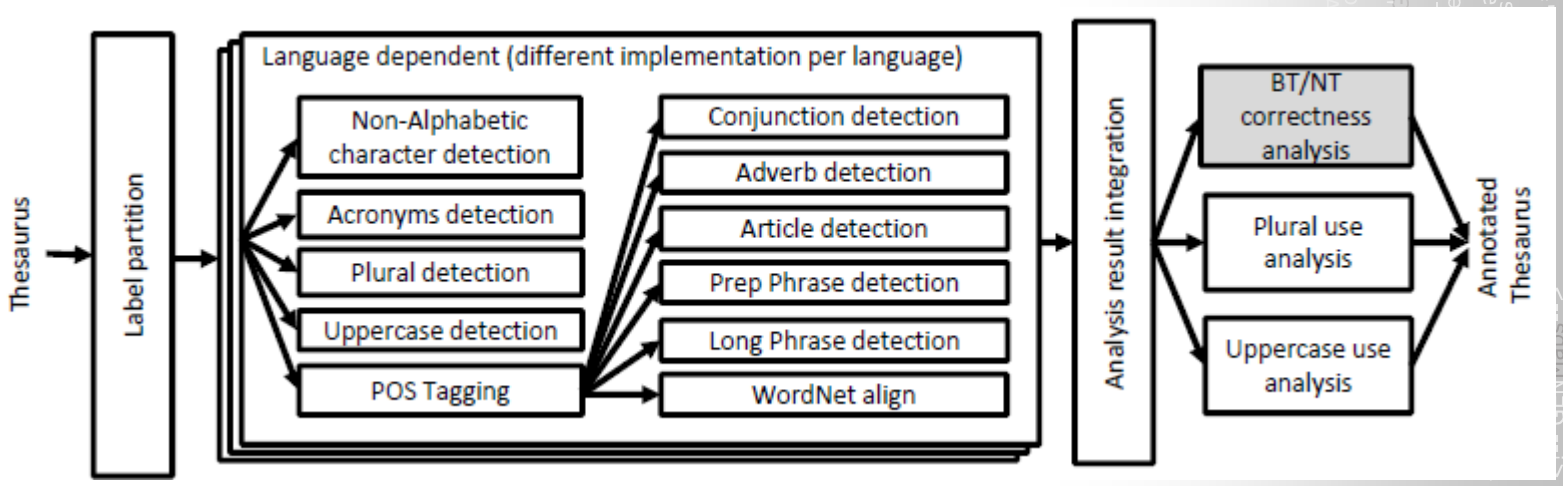
Concept level validators

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



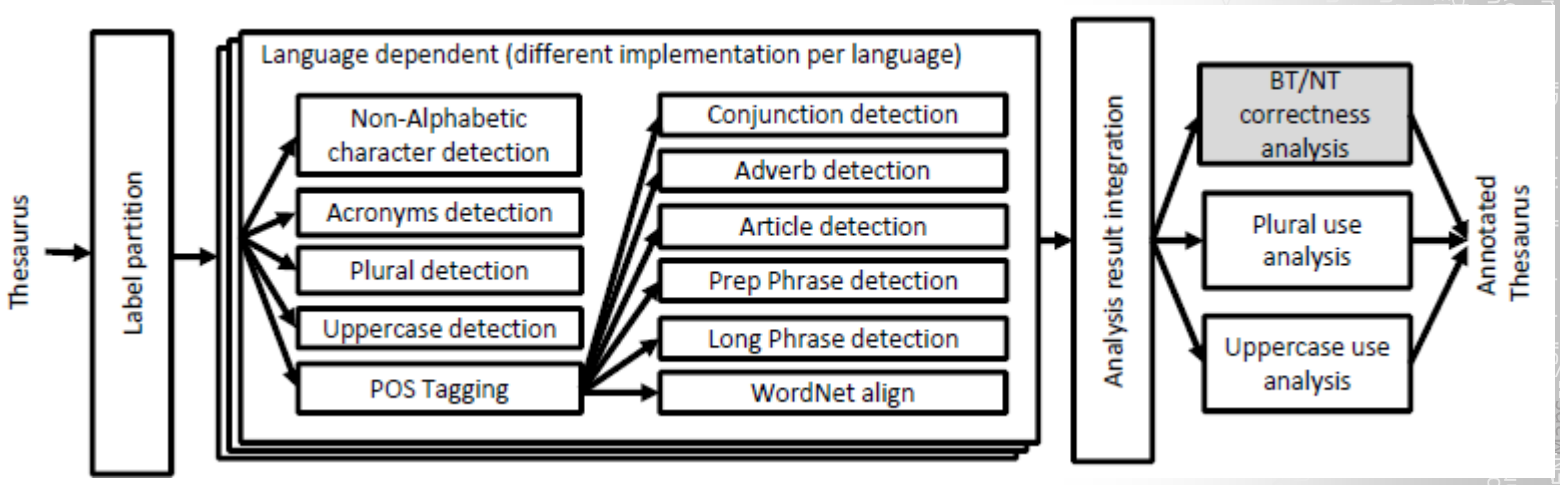
Label level validators

- Detection of non-alphabetic characters, acronyms, and upercase.
  - 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).



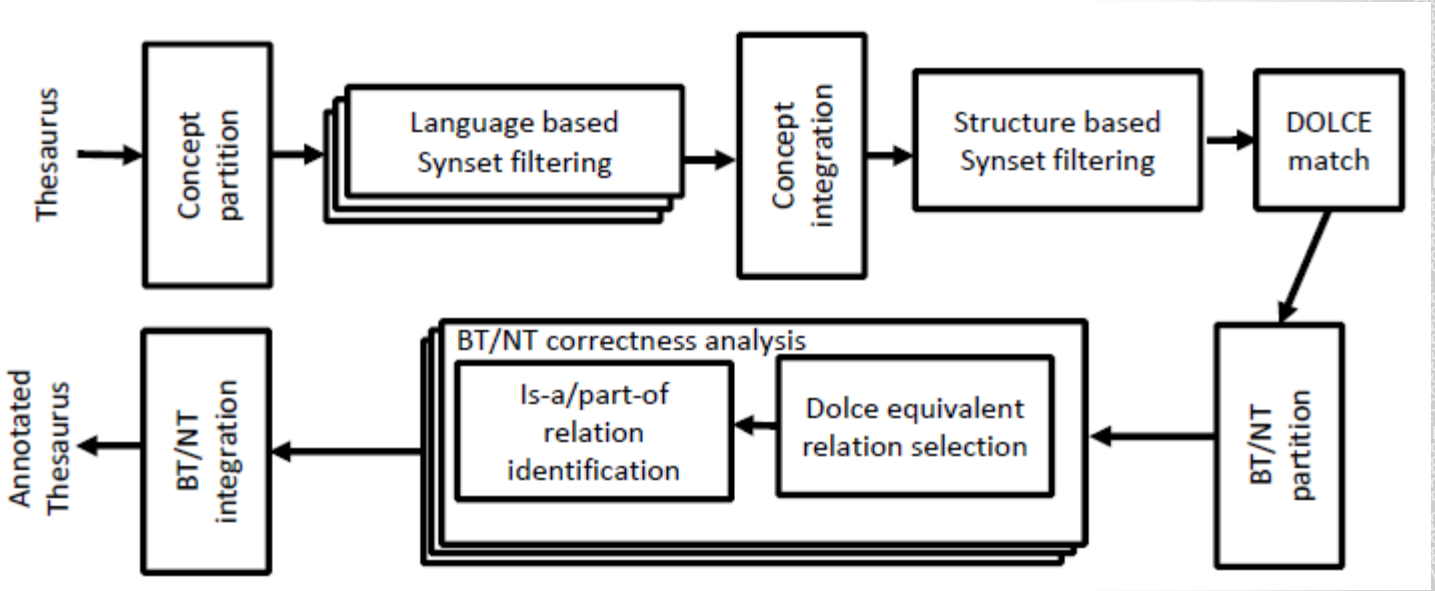
Label level validators, result integration

- Plural and uppercase analysis
  - Detection of inconsistences.
- 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.

```
<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

Thesaurus	Concepts	PL(en)	PL(es)	PL(fr)	AL(fr)	Def(en)	BT-NT	RT
Urbamet	3844	3844	3845	3846	504	0	3821	0
Gemet	5244	5244	5244	5244	0	4909	5332	1043

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

Validation of results

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

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

Gemet quality measure	tp	tn	fp	fn	na	accuracy
Nonuse of adverbs / initial articles in labels	99.73%	0.24%	0%	0.02%		99.98%
Consistent use of plurals in labels	92.91%	6.23%	0%	0.86%		99.14%
Nonuse of prepositional phrases in labels(en)	94.58%	5.42%	0%	0%		100%
Nonuse of too long noun phrases / conjunctions	99.90%	0.10%	0%	0%		100%
Semantic correctness of BT/NT	88.39%	1.30%	9.67%	0.64%	0%	89.69%

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.