Collaborative-AI Knowledge Graph Generation: Taxonomization of IATE, the EU Terminology

Alena Vasilevich
Computational Linguist@Coreon

alena@coreon.com
https://www.linkedin.com/company/coreon-gmbh
Agenda

- Structured data and IATE as a resource: why taxonomize?
- Manual Taxonomization
- Automatic Taxonomization, powered by Machine Learning
- Benefits of Collaborative-AI Approach
Taxonomization in a Nutshell:

An ordered scheme

Haystack of concepts
Taxonomy Auto-generation: Others vs. Coreon

SOTA:
- information extraction and pattern-based methods
- combinations of tags and Wikipedia Category Hierarchy
- already existing KG triple representations
- strict domain-specific properties (corpora-based; assumes corpora represent the domain)

Coreon:
- a flat list of concepts to taxonomize, without established relations
- no domain corpora at hand
Interactive Terminology for Europe, IATE

• Introduced in 2004, used by most EU Institutions, covers all EU domains
• Recent focus on healthcare, financial crisis, environment, fisheries, and migration
• EuroVoc for domain classification system

• Number of concepts: 961,116
• Number of terms: 7,992,325
• New terms last week: 1,646
Fortes of Structured Data

- A Powerful Resource for AI/ML projects
- Cross-lingual Data Analysis
- Enterprise Search
- Actionable intelligence
- Cross-border Interoperability
Two Tested Approaches

semi-automatic
- study domain
- taxonomize automatically
- name group concepts
- curate data

manual
- study domain
- load into Coreon
- build taxonomy from scratch

measure and compare
- time
- edit actions
- taxonomies
Manual Taxonomization

- Top level nodes, temporary helper buckets, and lots to do...

- Concept card displaying important metadata
Auto-Taxonomization:
Data + fastText WE + Louvain Algorithm
Human Revision of AI-Drafted Taxonomy

initial situation after automatic taxonomization
Good Clustering, Bad Clustering?

- 55 clusters, majority pretty accurate
- some clusters are off, and we blame WE:
  - ‘interstitial space’ and ‘hospital pharmacy’
  - spaces appearing in similar “semantic neighborhoods”
- some existing IATE concepts became parents of concept clusters
## Effort and Performance

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Manual</th>
<th>Semi-Automatic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curator’s recorded time (hours)</td>
<td>40h</td>
<td>8h</td>
</tr>
<tr>
<td>Relations created / changed</td>
<td>1,147</td>
<td>432</td>
</tr>
<tr>
<td>Concepts created</td>
<td>115</td>
<td>28</td>
</tr>
<tr>
<td>Intermediate structural nodes renamed</td>
<td>—</td>
<td>45</td>
</tr>
<tr>
<td>Overall relations</td>
<td>679</td>
<td>470</td>
</tr>
</tbody>
</table>
Resulting Taxonomies

- **Load into Coreon**
- **Build taxonomy from scratch**

**Automatic taxonomization using ML algorithm**
- **Name auto concepts**
- **Move wrong concepts**
Why Taxonomize?

- Effective way to add structure to data
- Improve data quality
  - avoid duplicates and overlapping concepts
  - associative relations
- Easier and safer data maintenance
- Formalize multilingual knowledge, make it machine-digestible
- Boost performance of AI algorithms, priming them with structured data
Future Work: Improving Performance

- Domain-specific data to improve WE (tuning/re-training)
- Leverage metadata
- Exploit concept definitions if available

**Modified Live-Virus Vaccine:**

**Definition:**

vaccine made from an isolate of an attenuated virus

**Definition reference:**

COM-EN based on:

**Note:**

Attenuated means the virus cannot cause disease but it can reproduce in the body cells and stimulate immunity.