T7: Tech-Taxonomy with a Text To Text Transfer Transformer

You Zuo (Almanach, Inria), Kim Gerdes (Lisn, Upsay, ISS, Inria)

ALMAnaCH project-team Seminar - 22/10/2021
Outline

1. INTRODUCTION
   - what is tech-taxonomy?
   - what does qatent.com do?
   - why do we need tech-taxonomy?

2. METHODS
3. EVALUATION
4. CONCLUSION
5. Q & A
Introduction

- Definition of taxonomy
- Taxonomy vs. ontology/ knowledge graph
Introduction

- Web application for drafting patents
- Patent description generator
- Patent application validator

Why do we need tech-taxonomy?
- patent

Technical Terms

General (Hypernym)

Specific (Hyponym)
Problems

- Currently existing taxonomies / ontologies
  WordNet ⇒ too general!
  Wikidata ⇒ unrelated items: celebrities, locations, events
  or no information about hypernym
Problem
- Currently existing taxonomies / ontologies
  WordNet, Wikidata
- Patent domains

| A | HUMAN NECESSITIES                        |
| B | PERFORMING OPERATIONS; TRANSPORTING     |
| C | CHEMISTRY; METALLURGY                   |
| D | TEXTILES; PAPER                         |
| E | FIXED CONSTRUCTIONS                     |
| F | MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING |
| G | PHYSICS                                 |
| H | ELECTRICITY                             |
Solutions

- Cooperative Patent Classification (CPC) classification system + Wikipedia Category trees
Solutions

CPC classification system

Hierarchy
- Section (one letter A to H and also Y)
  - Class (two digits)
    - Subclass (one letter)
      - Group (one to three digits)
        - subgroup (at least two digits)

Totally 262K categories in CPC!!
Solutions

CPC classification system

example "A01B33/02"

- **Section A** => Human Necessities
  - **Class 01** => Agriculture; forestry; animal husbandry; trapping; fishing
  - **Subclass B** => Soil working in agriculture or forestry, parts, details, or accessories of agricultural machines or implements, in general (making or covering furrows or holes for sowing, planting, or manuring A01C5/00; soil working for engineering purposes E01, E02, E21; measuring areas for agricultural purposes G01B)
  - **Group 33** => tilling implements with rotary driven tools (e.g. in combination with fertiliser distributors or seeders, with grubbing chains, with sloping axles, with driven discs)
    - **Subgroup 02** => ... with rigid tools
Solutions

example "A01B33/02"

- Section A => Human Necessities
  - Class 01 => Agriculture; forestry; animal husbandry; trapping; fishing
Solutions
example "A01B33/02"

- **Group 33** => tilling implements with rotary driven tools (e.g. in combination with fertiliser distributors or seeders, with grubbing chains, with sloping axles, with driven discs)
  - **Subgroup 02** => ... with rigid tools

308K nodes retrieved finally!

Parse analysis by Alma Parías García
Solutions

- Wikipedia Category trees

Extracted Wikipedia Category Trees under eight domains:
- Physics
- Biology
- Mathematics
- Engineering
- Computer Science
- Health
- Chemistry
- Software Engineering
## Solutions

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of terms</th>
<th>Number of cleaned terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physics</td>
<td>57752</td>
<td>1032</td>
</tr>
<tr>
<td>Health</td>
<td>51137</td>
<td>2121</td>
</tr>
<tr>
<td>Software Engineering</td>
<td>47262</td>
<td>1307</td>
</tr>
<tr>
<td>Computer Science</td>
<td>21098</td>
<td>991</td>
</tr>
<tr>
<td>Chemistry</td>
<td>33854</td>
<td>3246</td>
</tr>
<tr>
<td>Mathematics</td>
<td>65019</td>
<td>2935</td>
</tr>
<tr>
<td>Biology</td>
<td>76470</td>
<td>737</td>
</tr>
<tr>
<td>Engineering</td>
<td>46932</td>
<td>4816</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>399521</strong></td>
<td><strong>17185</strong></td>
</tr>
</tbody>
</table>

Number of terms for each Wikipedia Category Tree

Distribution by topic of Wikipedia Category Trees
Solutions

- Cooperative Patent Classification (CPC) classification system + Wikipedia Category trees
  ⇒ 339k nodes in our tech-taxonomy
Evaluation

- Human evaluation

- models for hypernym prediction
  - CRIM (Bernier-Colborne & Barrière, 2018)
  - TransE (Bordes et al., 2013)
  - hypo2hyper (Cho et al., 2020)
  - T5 (Raffel et al., 2020)
Models for hypernym prediction

**CRIM** *(Bernier-Colborne & Barrière, 2018)*
Ranked 1st in SemEval 2018 Task 9

- Learning Projection matrices from query embedding ⇒ target hypernym embeddings
- fastText as term representation (OOV, rare/ unfamiliar words)
Link prediction model

TransE \((Bordes \text{ et al.}, 2013)\)
- Relationships are represented as translations in the embedding space
- if \((h, r, t)\) holds, then \(h + r \approx t\)
Hypernym prediction as sequence to sequence task

hypo2hyper \cite{Cho2020} - LSTM-based seq2seq model with Luong attention \cite{Luong2015}

<table>
<thead>
<tr>
<th>Hyponym</th>
<th>Generated hypernym path</th>
<th>Gold hypernym</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ pizza.n.01</td>
<td>dish.n.02 → nutriment.n.01 → food.n.01 → … → entity.n.01</td>
<td>dish.n.02</td>
</tr>
<tr>
<td>✓ alps.n.01</td>
<td>range.n.04 → geological_formation.n.01 → … → entity.n.01</td>
<td>range.n.04</td>
</tr>
<tr>
<td>✓ whisper.v.01</td>
<td>talk.v.02 → communicate.v.02 → interact.v.01 → act.v.01</td>
<td>talk.v.02</td>
</tr>
<tr>
<td>✗ proletarian.n.01</td>
<td>*worker.n.01 → person.n.01 → causal_agent.n.01 → … → entity.n.01</td>
<td>commoner.n.01</td>
</tr>
<tr>
<td>✗ austerity.n.01</td>
<td>*punishment.n.01 → social_control.n.01 → … → entity.n.01</td>
<td>self-discipline.n.01</td>
</tr>
<tr>
<td>✗ compulsive.n.01</td>
<td>*sick_person.n.01 → unfortunate.n.01 → person.n.01 → … → entity.n.01</td>
<td>person.n.01</td>
</tr>
</tbody>
</table>

Table 1: We frame hypernym prediction as a sequence generation problem. Given a query hyponym (e.g., pizza.n.01), the hypo2path rev model generates its taxonomy path, from its direct hypernym (dish.n.02) to the root node (entity.n.01). ✓ and ✗ indicate a correct and an incorrect prediction, respectively. In each example, an underlined synset corresponds to what the model predicted as a direct hypernym.
Evaluation

T5 (Raffel et al., 2020)

- Text to Text Transfer Transformer

predict hypernym: support vector machine

T5

classifier/ algorithm/ supervised learning
Evaluation

Metrics
- Hits@k score
  percentage of correct true labels that appeared in the top k (k=1, 3, 10) ranked predictions

- Mean reciprocal rank (MRR)
  the position of the first correct results in ranked list of outcomes

\[ MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \]
Results

Table 12. Results on test set of Qatent Taxonomy (incomplete)

<table>
<thead>
<tr>
<th>Model</th>
<th>H@1</th>
<th>H@3</th>
<th>H@10</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closet Vector</td>
<td>9.49</td>
<td>17.61</td>
<td>28.26</td>
<td>15.57</td>
</tr>
<tr>
<td>CRIM</td>
<td>10.99</td>
<td>20.48</td>
<td>31.57</td>
<td>17.68</td>
</tr>
<tr>
<td>TransE</td>
<td>37.82</td>
<td>55.25</td>
<td>69.64</td>
<td>48.39</td>
</tr>
<tr>
<td>Hypo2Hyper</td>
<td>11.04</td>
<td>21.03</td>
<td>30.18</td>
<td>17.45</td>
</tr>
<tr>
<td>T5</td>
<td>61.00</td>
<td>68.35</td>
<td>79.10</td>
<td>70.43</td>
</tr>
</tbody>
</table>
Results

- many potential terms that have not been included in our taxonomy

⇒ Term recognizer used in qatent:

8. The method of 1, further comprising: in response to displaying the scrollable refresh trigger, providing first audio feedback; and in response to determining that the scrollable

- 45 % of patent terms are not in our taxonomy
Results

- many potential terms that have not been included in our taxonomy

- 45% of patent terms are not in our taxonomy ⇒ T5
Ongoing and Future Work

1. Semi-supervised method to correct taxonomy extracted from CPC
   Using model trained on wikipedia category trees as a filter ⇒
   eg. keep terms if in the top 10

2. Try different pre-trained transformers: GPT-NeoX
3. Expand patent terms by implementing Hearst patterns in patent and technical wikipedia pages:
   Y such as X, Y other than X, not all Y are X, Y including X, Y especially X, Y like X, Y for example X, Y which includes X, X are also Y, X are all Y
Conclusion

- Taxonomies are not dead, still useful for paraphrase generation, intelligent editors
- Non-trivial task to develop domain-specific taxonomies

- Our attempt: Term recognizer (span Categorizer + NER from spaCy) + T5 \(\Rightarrow\) taxonomy completion
Q & A

Thank you for your attention!
Reference


Reference


- Tung Tran and Ramakanth Kavuluru. Supervised approaches to assign cooperative patent classification (cpc) codes to patents. In MIKE, 2017.
