Exploring the patent landscape

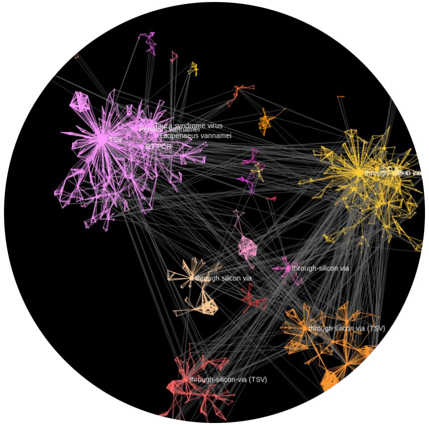
A NETWORK-BASED APPROACH FOR VISUALIZING AND ANALYZING HETEROGENEOUS PATENT GRAPHS

André Rattinger
Christian Gütl

Graz University of Technology

Overview

- Patents are a data source that is relatively easily and freely available
- There a some similarities to scientific papers
- Growing number of applications[1] (3.4million filed in 2021)
- How can we tackle this?
    - Extract information
    - Combine extracted infos with structured data
    - Visualize different facets of the data



[1] https://www.wipo.int/en/ipfactsandfigures/patents

Patents as research data

- Patents are
  - Long and well structured
  - Complex language and legal terms
  - Sometimes obfuscated in some ways to make them harder to find
- Patents are freely available from many sources
  - USPTO
  - EPO
  - Patentsview

Dataset

- EPO
- USPTO
    - Patentsview
- Smaller research datasets (CLEF, Patent collections)
- Our collection
    - A combination of the above (more than 8 million patents)
    - USPTO and Patentsview master classification files
    - Annotated patents for keyphrase extraction

Classification System

- Cooperative Classification System (CPC)
  - > 100 countries, hierarchical system, up to 14 levels deep
- Main classes which cover very broad topics:
  - A: Human Necessities B: Performing Operations, Transporting C: Chemistry, Metallurgy D: Textiles, Paper E: Fixed Constructions F: Mechanical Engineering, Lighting, Heating, Weapons G: Physics H: Electricity
- Work focuses on third level, because the system becomes very fine-grained after ~69,000 classes in total

Classification System

- Example Patent Class: H01F 1/01
- Section: H Electricity
    - Class: H01 Basic Electric Elements
    - Subclass: H01F Magnets
    - Main group: H01F 1/00 Magnets or magnetic bodies characterised by the magnetic materials therefore
    - Dot: 1/01 of inorganic materials

Preprocessing

- Tabular data:
  - Patent classes
  - Inventors
  - Countries
  - Citations
- Unstructured Data (complex, but similarities to publications)
  - Abstract
  - Long Description
  - Claims (important but technical)
  - Technical drawings

Extracting from unstructured data

- Examples
  - … Popular implementations of networks for MSS include network attached storage (NAS) and storage area networks (SAN). In NAS, MSS is typically accessed over known TCP/IP lines such as Ethernet …
  - … pixels of half of the total number of pixels are set as effective pixels, and the other half of the pixels are set as reference pixels that are references for correction, and a noise component is removed by subtracting …
- We use multiple text sections
- Keyphrase Boundary Infilling with Replacement (KBIR) model as a base for keyword extraction [1]
- Words are classified if they are part of a keyphrase or not

[1] https://aclanthology.org/2022.findings-naacl.67.pdf

Extracting from unstructured data

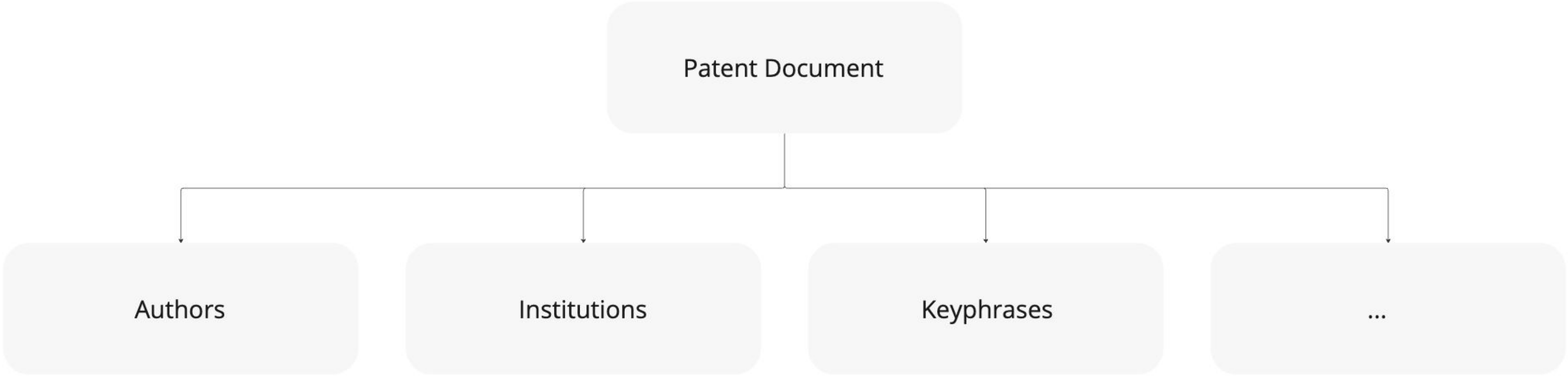
- Fine-tuned on our own annotated dataset of keyphrases specific to patents
  - 4600 unique keyphrases
  - Important because the language in patents can be quite specific
  - Alternatively models fine-tuned on scientific articles work also quite well and a combination would be promising
  - Other possible dataset for fine-tuning: inspec dataset (abstracts of scientific papers)
- English only for the moment (fine-tuning dataset is in english)
- Annotated all 8 million patents
  - Even subclasses have more than 50k keyphrases

Patent schema

- Getting structured data from text
  - Multiple applications outside of graph visualization
    - Training other models
    - Information Retrieval
    - Language Models
- We need to transform the data into tabular data to use in the Collaboration Spotting X graph visualization tool.

Patent schema

- Simple schema where the patent document is the “relation”
- Some ambiguity with multiple n-to-n matches like authors and institutions that could be disambiguated



Patent Document

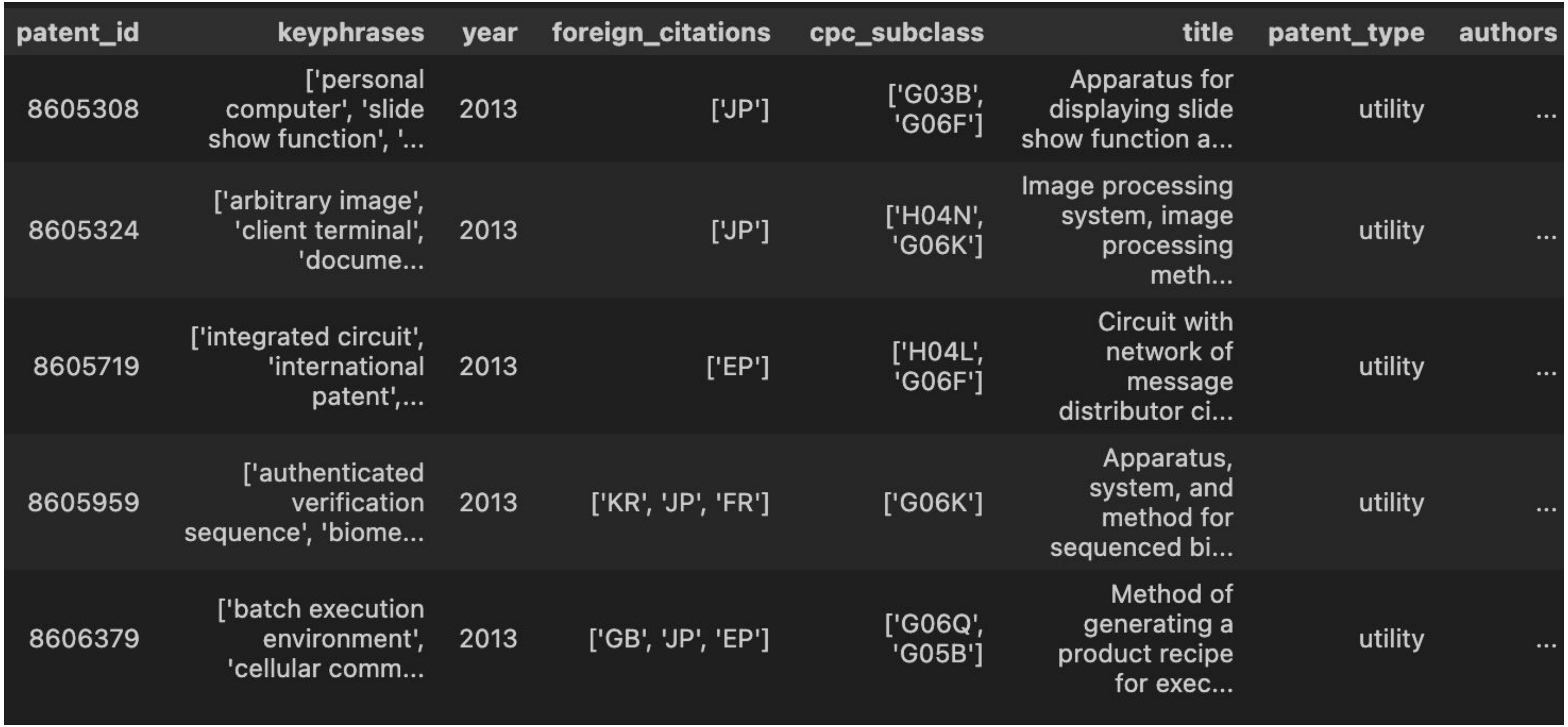
Authors

Institutions

Keyphrases

...

Patent DataFrame



patent_id	keyphrases	year	foreign_citations	cpc_subclass	title	patent_type	authors
8605308	['personal computer', 'slide show function', '...	2013	['JP']	['G03B', 'G06F']	Apparatus for displaying slide show function a...	utility	...
8605324	['arbitrary image', 'client terminal', 'docume...	2013	['JP']	['H04N', 'G06K']	Image processing system, image processing meth...	utility	...
8605719	['integrated circuit', 'international patent',...	2013	['EP']	['H04L', 'G06F']	Circuit with network of message distributor ci...	utility	...
8605959	['authenticated verification sequence', 'biome...	2013	['KR', 'JP', 'FR']	['G06K']	Apparatus, system, and method for sequenced bi...	utility	...
8606379	['batch execution environment', 'cellular comm...	2013	['GB', 'JP', 'EP']	['G06Q', 'G05B']	Method of generating a product recipe for exec...	utility	...

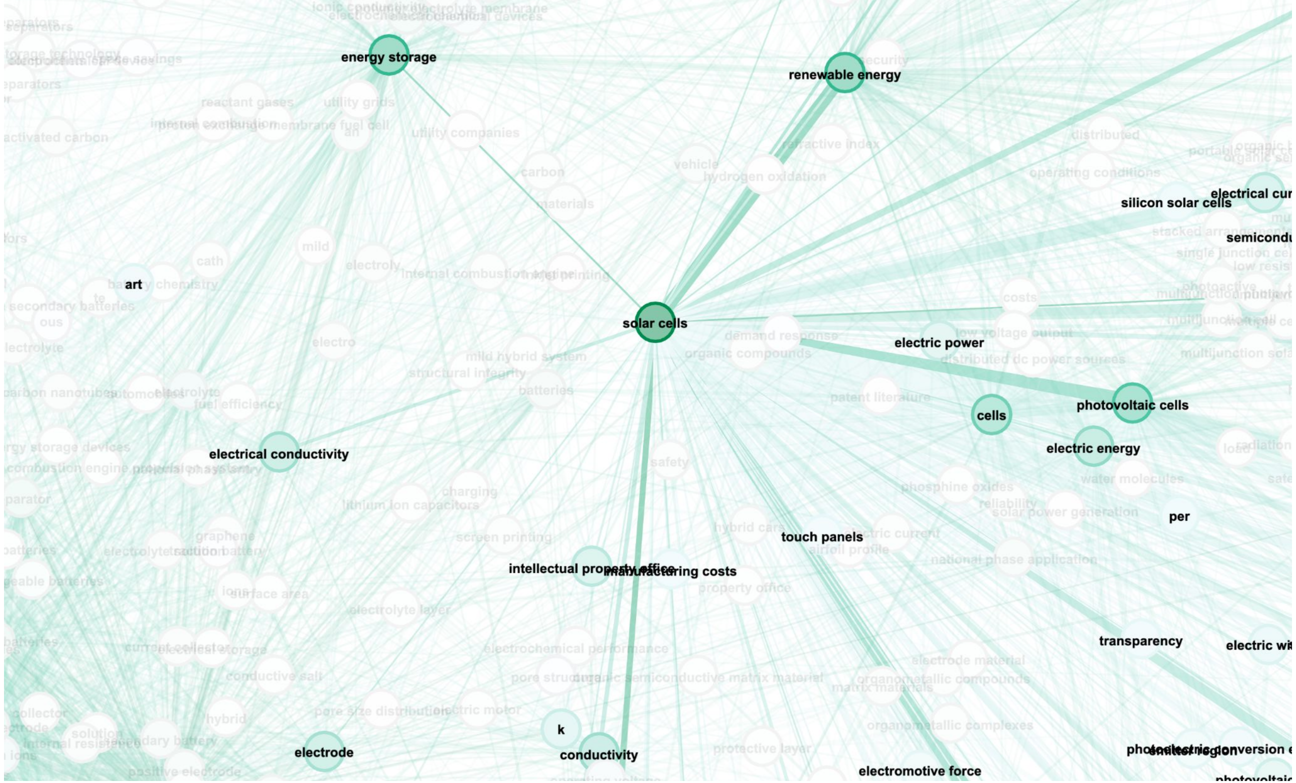
Classification System

- Focus on two patent classes
  - G06 Physics -> **COMPUTING; CALCULATING; COUNTING**
  - Y02 New Technological Developments -> **TECHNOLOGIES OR APPLICATIONS FOR MITIGATION OR ADAPTATION AGAINST CLIMATE CHANGE**

[1] https://www.uspto.gov/web/patents/classification/cpc/html/cpc-G06N.html

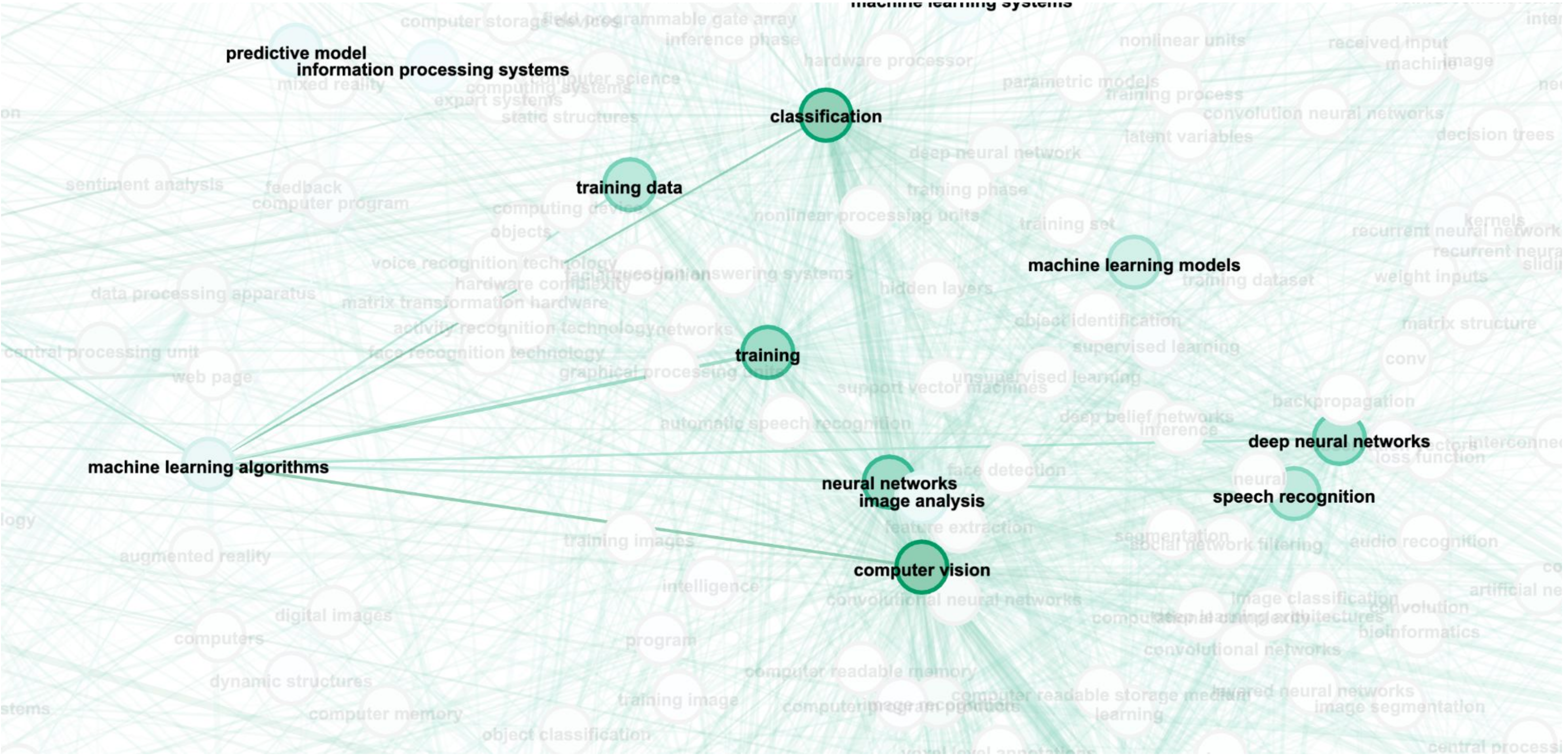
[2] https://www.uspto.gov/web/patents/classification/cpc/html/cpc-Y02E.html

Class Y02E Reduction of Greenhouse Gas Emissions



Examples

Class G06N Computing Arrangements based on specific computational models



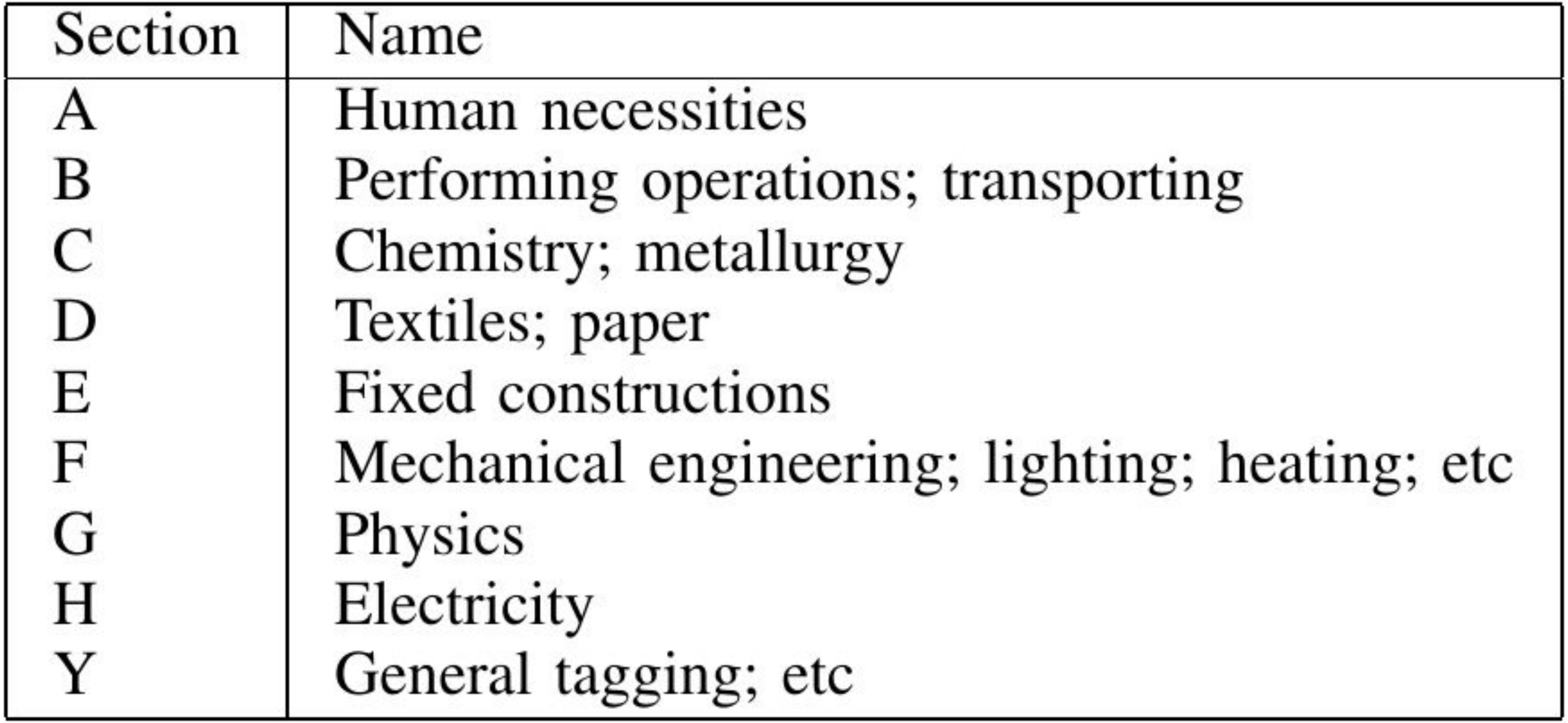
Conclusion

- Patents are easily available but complex
- Combining unstructured with structured data to create patent graphs
- Extracting data from patents is a hard problem but is made much easier by new models
    - Fine tuning is essential to get good results

Future Work

- Graph neural networks
    - Generalize unseen data, unexplored avenues in patents
    - Input for other patent or publication tasks
- Specialized knowledge-base chat systems (Language models)
    - Systems where it’s important to be exact
- Patent Retrieval and Classification

Thanks for your attention



Section	Name
A	Human necessities
B	Performing operations; transporting
C	Chemistry; metallurgy
D	Textiles; paper
E	Fixed constructions
F	Mechanical engineering; lighting; heating; etc
G	Physics
H	Electricity
Y	General tagging; etc