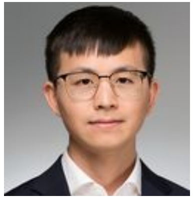
Max Planck Institute
for Innovation and Competition

Europäisches Patentamt
European Patent Office
Office européen des brevets

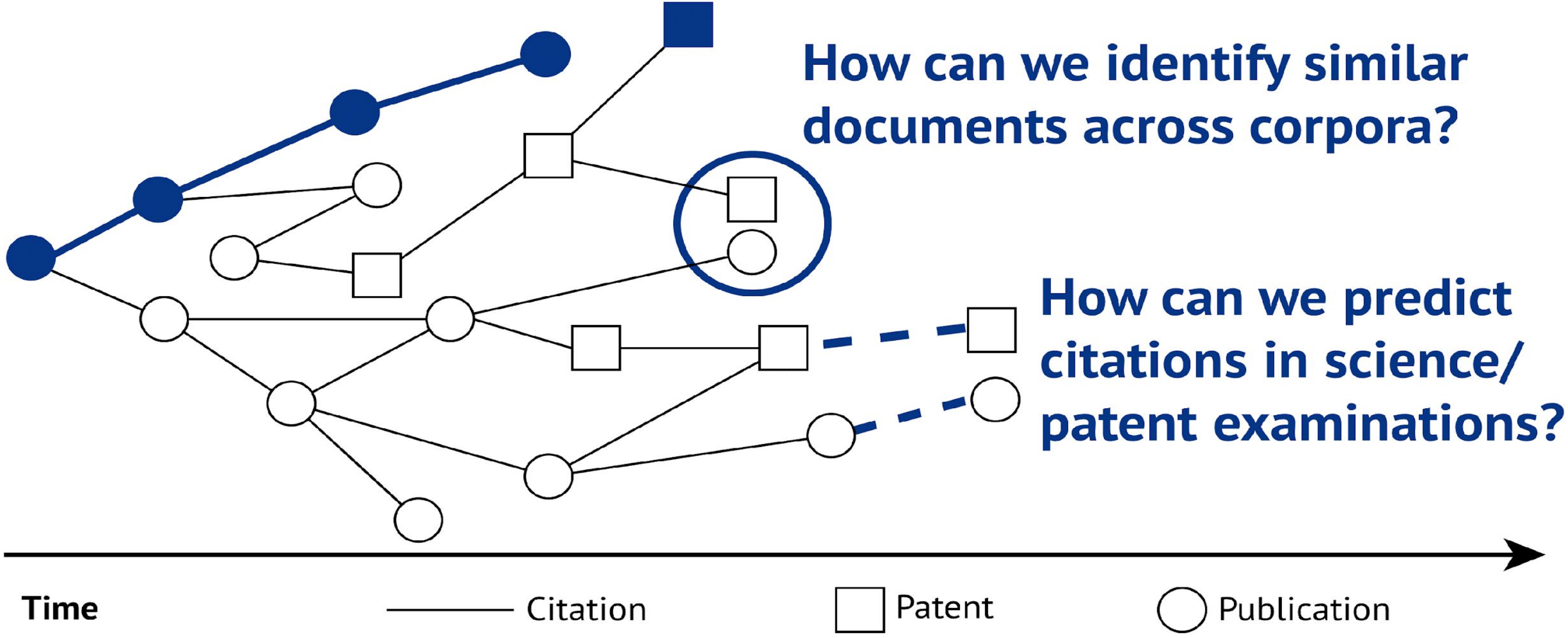
Tracing the Flow of Knowledge from Science to Technology Using Deep Learning

Michael Rose - Erik Buunk - Sebastian Erhardt - Cheng Li - Mainak Ghosh- Dietmar Harhoff



How can we trace the flow of knowledge across corpora?



How can we identify similar documents across corpora?

How can we predict citations in science/ patent examinations?

Project Overview

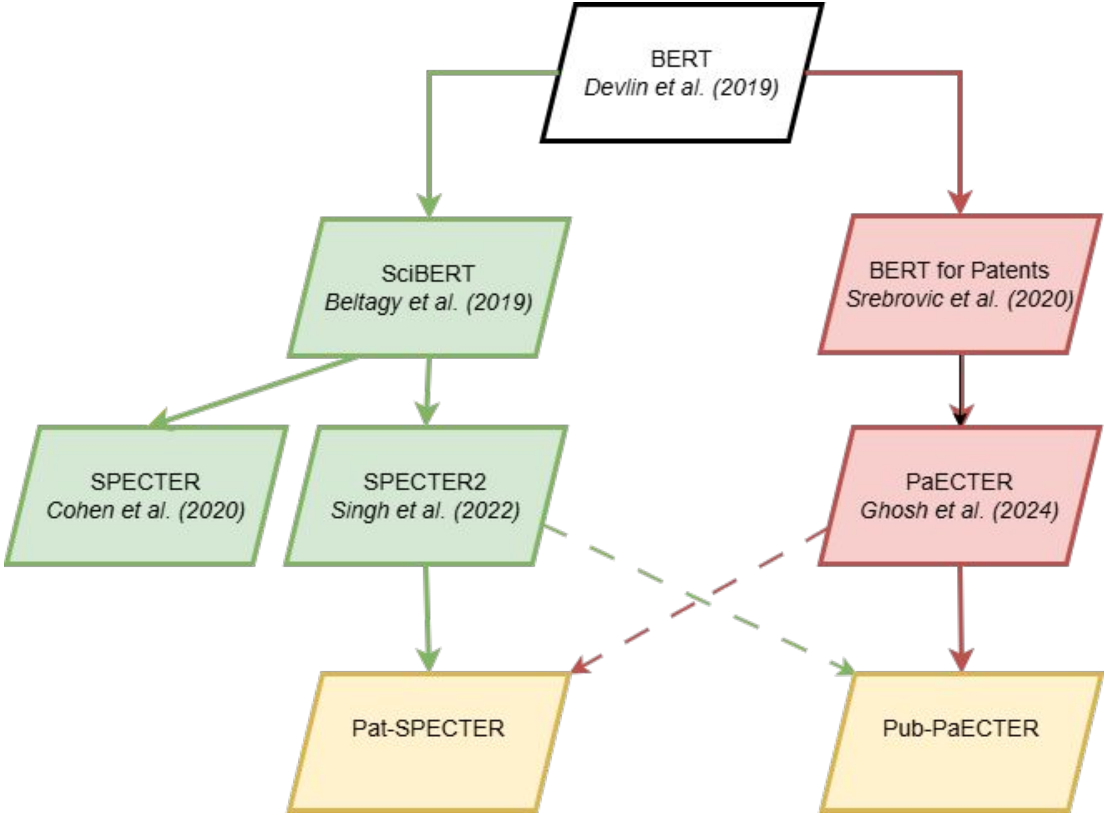
Problem: How to identify patents and publications based **textual content?**

- Current solutions
  - Not scalable
  - Outdated methods
  - Domain-specific
  - proprietary

Solution: **Compare** cross-corpus machine learning models, trained on **patents <u>and</u> scientific publications** as horse race

- Represent texts numerically as **embeddings**
- Apply in a real world application
- Store embeddings in vector database

The horses


BERT
Devlin et al. (2019)
SciBERT
Beltagy et al. (2019)
BERT for Patents
Srebrovic et al. (2020)
SPECTER
Cohen et al. (2020)
SPECTER2
Singh et al. (2022)
PaECTER
Ghosh et al. (2024)
Pat-SPECTER
Pub-PaECTER

Training the Pub-PaECTER

Goal: Fine-tune PaECTER model on dataset of SPECTER

600k random documents from SemanticScholar

1. (focal paper | random cited paper | random negative)
2. (focal paper | random cited paper | random negative)
3. (focal paper | random cited paper | random negative)
4. (focal paper | random cited paper | random indirectly cited negative)
5. (focal paper | random cited paper | random indirectly cited negative)

Training the Pat-SPECTER

Goal: Fine-tune SPECTER model on dataset of PaECTER

300k patent families from PATSTAT with Euro-PCT application

```
1.   ( focal patent | random cited X/Y patent | random negative with same CPC )
2.   ( focal patent | random cited X/Y patent | random negative with same CPC )
3.   ( focal patent | random cited X/Y patent | random negative with same CPC )
4.   ( focal patent | random cited X/Y patent | random indirectly cited negative )
5.   ( focal patent | random cited A patent   | random indirectly cited negative )
```

The racing ground

Goal: For a given patents, **rank 30 publications** by semantic similarity

- 1000 randomly selected patents
- 5 actually cited publications (*Reliance on Science*)
- 25 randomly non-cited publications

(provided they all have an English abstract)

Ranking metrics:

- Rank First Relevant
- Mean Average Precision
- Mean Reciprocal Rank among 10

Results of the horse race

PatSPECTER statistically dominates any other model

	Avg. RFR		MAP		MRR@10	
Model	CLS	Mean	CLS	Mean	CLS	Mean
BERT	2.52	1.29	46.76	79.05	69.86	91.52
SciBERT	2.48	1.37	49.23	71.14	71.75	90.55
BERT for Patents	1.20	1.10	78.51	85.99	93.14	96.97
SPECTER	1.08	1.13	**91.64**	86.57	97.72	96.23
SPECTER2	1.11	1.35	88.29	76.51	95.94	91.62
Our Models						
Pat-SPECTER	**1.05**	1.12	91.38	87.24	**98.04**	96.06
PaECTER	1.13	1.07	86.45	89.72	96.12	97.55
Pub-PaECTER	1.32	1.25	76.49	79.57	92.51	94.15

Applications

1. Separating patent paper pairs (PPP) from patent paper citations (PPC)
2. Predicting patent paper pairs (PPP)
3. Prior art search for publications

Leverage **Logic-Mill.net**: Database for approximate nearest neighbor searches for patents and publications

- 53M **DocDB** patent families w/ English abstract
- 120M **OpenAlex** works w/ English abstract

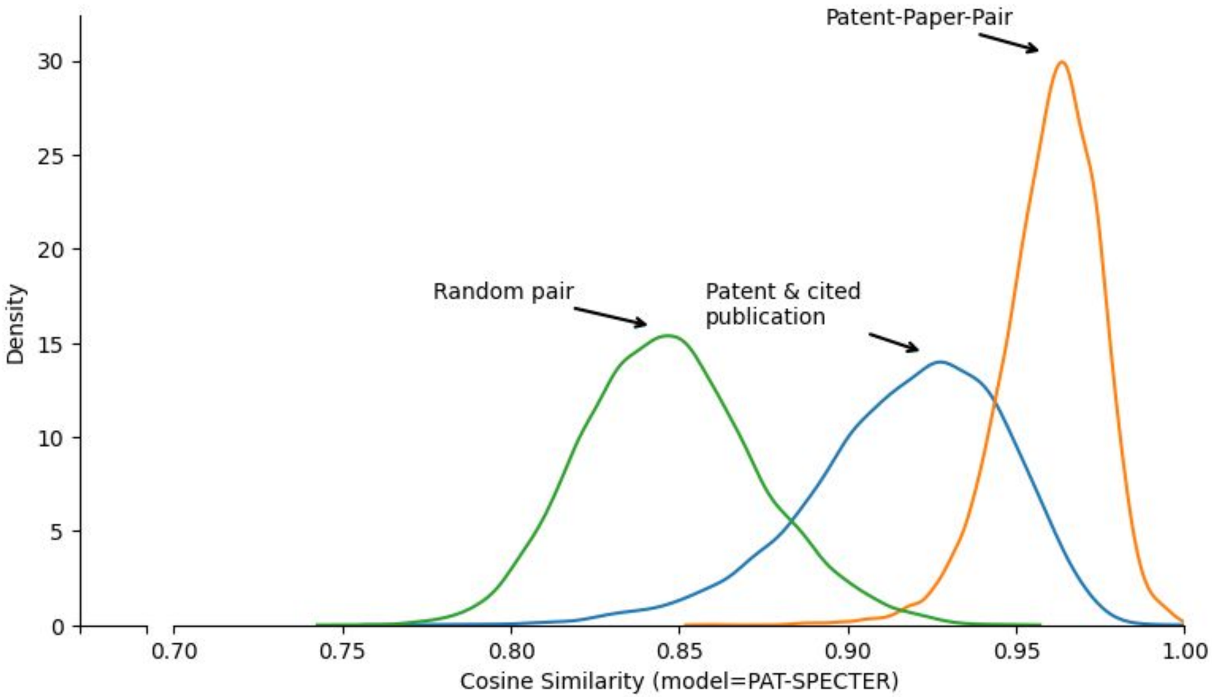
Caveat of OpenAlex: Many publications missing abstract; duplicates

1. Separating PPP from PPC

Goal: Based on textual similarity, separate

1. Patent Paper Pairs
2. Patent Paper Citations - Paper/Non patent literature (NPL)
3. Patent with Random papers

Do the **distributions of cosine similarities** overlap?



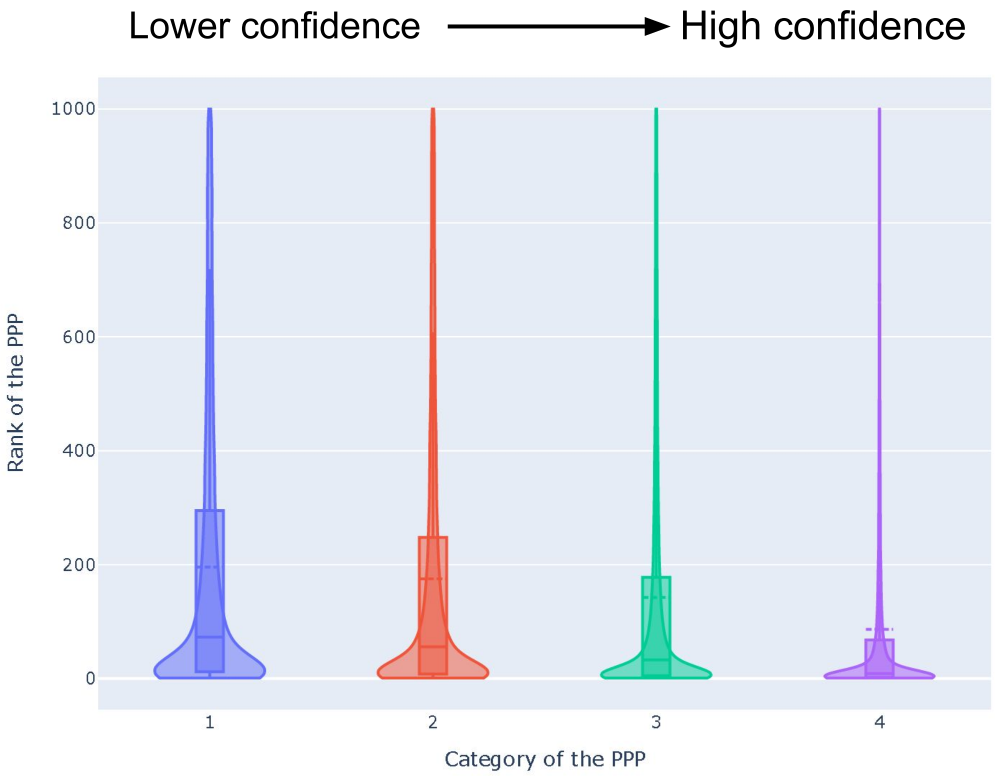
- Distributions clearly separable
- Optimal similarity cutoff (based on F1 score): 0.949

2. Predicting patent-paper-pairs

Goal: Identify **papers that are very similar** or identical to **a patent**

- PPP has 4 confidence scores:
  - very high for category 4 (prediction greater than 0.99) to
  - category 1 (prediction between 0.70 and 0.80)
- For each of the 550k patents, we search for the 1000 approximate nearest neighbors of OpenAlex and calculate the rank of the actual connected paper

- Able to match: 342,252 pairs (≈62%) within the first 1000 ANNs
- For the matched PPPs, about 65% have a rank between 1 and 100, and nearly 90% have a rank between 1 and 500.
- Predictions with higher confidence is consistent with our findings.

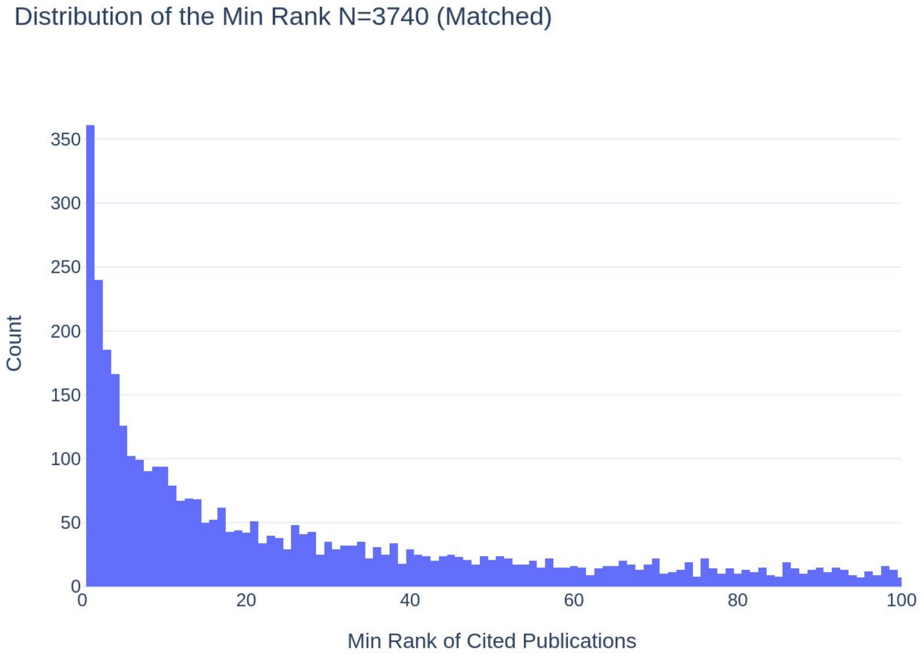


3. Prior Art Search for publications

Goal: **predict actual patent citations** among **all of OpenAlex**

10k patents randomly from the PPC dataset

Consider only patent paper citations with highest confidence score (3.7k)

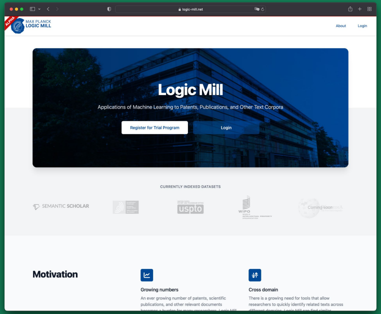


k	MRR	MAP
5	0.162932	0.052162
10	0.179573	0.066367
20	0.190259	0.080288
50	0.197886	0.093623
100	0.200598	0.100609

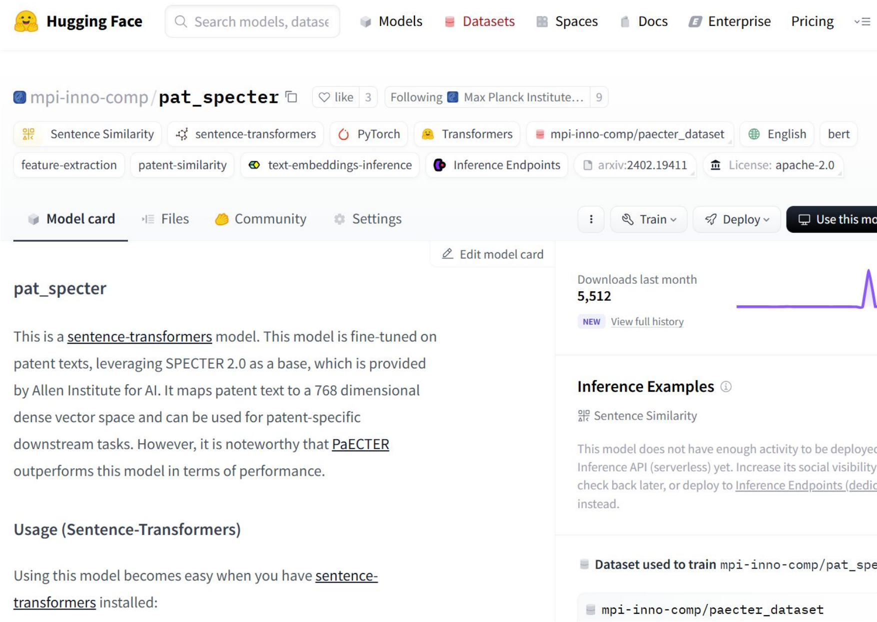
- In **37%** at least 1 result in top 100
- Distribution rather steep

Use **Logic Mill**

- Register at https://logic-mill.net/
- Check out documentation
- Use Python, R, Stata, … to pull data through our API (website generates code snippets)



Use **Pat-SPECTER**



huggingface.co/mpi-inno-comp/pat_specter