# Making Patent Citations Uncool Again

Gaétan de Rassenfosse
EPFL - IIPP

Cyril Verluise
PSE and Collège de France[1]

Research Retreat IIPP-CEMI, March 2020

[1]We are thankful to F. Gerotto for excellent research assistance.

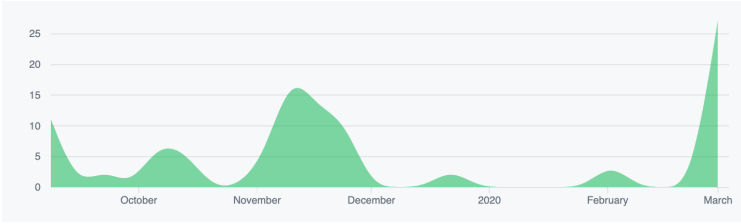
**Motivation**

- Patent-to-patent citations: device for major breakthrough in Innovation, Growth, etc
- Many more Citations: Non Patent Literature, In-text $\rightarrow$ potential to shed new light on old and new issues
- Improved ability to extract data from unstructured content $\rightarrow$ Patent field: [Marx and Fuegi, 2019] (in-text), [Bryan et al., 2020] (front-page)
- Renewed interest for Patent Citations

**Approach**

- Open source (MIT-2)
- Comprehensive
- Leverage (existing) Machine Learning Solutions

# Contributions




- **October, 19th**: Release of the "Worldwide Patent-to-*NPL* Citations" database - `Beta`
- **November, 21st**: `__` - `v0.1`
- **November, 26th**: Release of the "US Contextual Patent-to-Patent and Patent-to-NPL" database - `Beta`
- **December, 28th**: `__` - `v0.1`
- **March, 3rd**: Release of the "Worldwide Patent-to-*NPL* Citations" database - `v0.2`

# How it works

# A stack of industrial-strength open source software

**Extraction, Parsing and Consolidation**

- **GROBID** ([GRO, 2020]): a machine learning library for extracting, parsing and re-structuring raw documents (...) into structured documents with a particular focus on technical and scientific publications.
  - Started in 2008 and open sourced in 2011 by Patrice Lopez
  - Leverage Conditional Random Fields to label sub-parts of the text
  - State-of-the-art in bibliographic reference parsing [Tkaczyk et al., 2018]
- **Biblio-Glutton**: a bibliographical reference matching service.

**Labeling, Validation, Text Classification**

- **Doccano**: Open source text annotation tool for machine learning practitioner.
- **spaCy**: Industrial-strength Natural Language Processing - `textCategorizer`.

# In practice

**Architecture**

- 1 machine with input data, GROBID and Biblio-Glutton + PubMed and Unpaywall
- 1 machine with Crossref database

**Efficiency**

- NPL Citations: $\sim$ 2 million NPL citations per day $\rightarrow$ 100 USD for DOCDB
- Full-text: $\sim$ 300k full-texts per day $\rightarrow$ 600 USD for USPTO

# Dataset

# Worldwide Patent-to-NPL Citations - `v0.2-npl`

**Input**: 40 million DOCDB NPL citations
**What we do**

- *Classify* in 9 classes[2]
- *Parse*[3]
- *Consolidate* bibliographical references

**Results**

1. Classify with 90% accuracy $\rightarrow$ 27 million bibliographic references
2. Match a DOI for 11 million+ NPL citations with 99% precision
3. Parsing precision of the main bibliographical attributes ranges above 70%

[2] Bibliographical reference, office action, patent..., search report, webpage, product documentation, norm and standard, database, litigation

[3] Retrieve structured attributes `title_a`, `title_j`, `year`, etc

# v0.2-npl - Classification

NPL citations are messy

**ML text classification**

- Define 9 classes
- Label $\approx$ 3k NPL citations by hand
- Learn a multi-class text classifer (`cnn+bow`), evaluate on development set
- Results:

| accuracy | precision | recall | f1 |
|---|---|---|---|
| 0.9 | 0.89 | 0.88 | 0.88 |

Table: Classifier average performance over all classes More

Last but not least, apply at scale More

# v0.2-npl - Consolidation

Number of DOI matches: 11,005,114

**Evaluation**

- Random draw
- Label 300 by hand (gold)
- Results:

| match doc | version discrepancy | year discrepancy |
|---|---|---|
| 0.99 | 0.0 | 0.02 |

Table: Share of DOI matches with ...

# v0.2-npl - Parsing

Parsing of the remaining (non DOI-matched) bibliographical references

**Evaluation**

- Random draw
- Label main attributes by hand for 150 bibliographical references (gold)
- Results:

| | **True** | **False** | **Accuracy** |
|---|---|---|---|
| year | 138 | 12 | 0.92 |
| volume | 135 | 15 | 0.9 |
| issue | 132 | 18 | 0.88 |
| title article | 116 | 34 | 0.77 |
| title journal | 107 | 43 | 0.71 |
| title meta (non journal) | 120 | 30 | 0.8 |

# US contextual Patent-to-Patent and Patent-to-NPL

**Input**: 16 million USPTO full-text patents

**What we do?**

- Extract
- Parse
- Consolidate patent citations and bibliographical reference

**Results**

1. 70+ million contextual bibliographical reference extracted
2. 13+ million contextual bibliographical reference matched with a DOI
3. 80+ million contextual patent citations extracted

No validation yet

# What's next?

*"ASK NOT WHAT YOUR COUNTRY CAN DO FOR YOU..."*

Happy for contribution!

# Last words

# Data Access

- **License**: Open access, CC4 (very permissive)
- **Google Cloud Bigquery**: Publicly Available at `https://console.cloud.google.com/bigquery?project=npl-parsing&p=npl-parsing&d=patcit&page=dataset`. Easy to query, adpated to the scale
- **Bulk Download**: Coming soon!
- **List-based download (API-like)**: Depending on community request

# Stay up-to-date

We are constantly improving the database, make sure that you have the latest news

- Star the project GitHub repo: `https://github.com/cverluise/PatCit`
- Follow us on twitter @gderasse and @CyrilVerluise
- Drop us a mail so that we add you the loop

# Contribute!

This project is by and for the community → sandbox for experimenting a new way of doing.

- Raise issues
- Request features
- Tackle issues (`good first issue`)
- Fork, extend, merge - Take credit for your work

Thank you!

# References I


(2008 — 2020).
Grobid.
https://github.com/kermitt2/grobid.

Bryan, K. A., Ozcan, Y., and Sampat, B. (2020).
In-text patent citations: A user's guide.
*Research Policy*, 49(4):103946.

Marx, M. and Fuegi, A. (2019).
Reliance on science: Worldwide front-page patent citations to scientific articles.
*Boston University Questrom School of Business Research Paper*, (3331686).

Tkaczyk, D., Collins, A., Sheridan, P., and Beel, J. (2018).
Evaluation and comparison of open source bibliographic reference parsers: a business use case.
*arXiv preprint arXiv:1802.01168*.

# Classifier performance by class

| | BIBL REF | SEARCH REPORT | OFFICE ACTION | PAT | PROD DOC. | NORM STD | WEB PAGE | DATA BASE | LITI-GATIO |
|---|---|---|---|---|---|---|---|---|---|
| precision | 0.92 | 1.0 | 0.99 | 0.91 | 0.44 | 0.86 | 0.53 | 0.89 | 0.25 |
| recall | 0.95 | 0.92 | 0.93 | 0.94 | 0.43 | 0.6 | 0.53 | 0.73 | 0.11 |
| f1 | 0.93 | 0.96 | 0.96 | 0.93 | 0.44 | 0.71 | 0.53 | 0.8 | 0.15 |
| support | 370.0 | 86.0 | 76.0 | 68.0 | 37.0 | 20.0 | 17.0 | 11.0 | 9.0 |

# Classifier confusion matrix

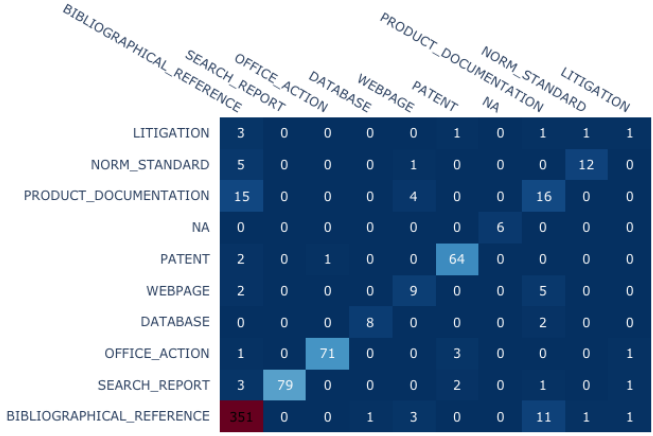



Figure: Confusion matrix on the development set

# Number of NPL citation by class

| npl_class | count |
|---|---|
| BIBLIOGRAPHICAL_REFERENCE | 27478140 |
| OFFICE_ACTION | 3247676 |
| PATENT | 2698016 |
| SEARCH_REPORT | 2279862 |
| WEBPAGE | 1293687 |
| PRODUCT_DOCUMENTATION | 734606 |
| NORM_STANDARD | 600323 |
| Unknown | 260030 |
| DATABASE | 246578 |
| LITIGATION | 191017 |