Online appendix for: Boosting the signal: Contextualization, extraction, and exploration of in-text patent citations

January 16, 2023

In this Online Appendix, we describe in detail our data sources, the extraction tasks, the validation tasks, the processing pipeline, and the resulting data set, alongside associated performance metrics and close examinations of shortcomings. We aim to provide extensive insights into our technical choices in order to provide full transparency for any who wish to use the data, as well as to stimulate and enable future extensions or improvements.

Contents

1 Citation extraction and cleaning

1.1 Data

The processing pipeline starts with the full-text of 16,781,144 patent documents filed at the U.S. Patent and Trademark Office (USPTO) since 1790.[1] The primary data source for this task was the full-text data of the IFI CLAIMS dataset, made available by Google Patents as part of its public datasets.[2]

The text we consider is the specification section of the patent. The specification is a written description of the invention which details the manner and process of making and using the invention. It also includes information about related applications and government interest statements (de Rassenfosse et al., 2019), when they exist. The specification does *not* include the patent's claims or the information on the front-page.

1.2 Extraction task

Our starting point is effectively a long chain of characters without any obvious structure, nor any indication about which characters might refer to a patent citation. As such, the first step involves identifying the strings of characters that refer to a patent citation in the full text. An early attempt to do so dates back to Galibert et al. (2010), who combined a set of regular expressions to identify the cited patent number itself (e.g., country codes followed by a series of digits) in conjuntion with neighbouring text cues (e.g. "herein described by"). A similar approach was implemented by Berkes (2018) for U.S. patents published before 1947. Although intuitive, these approaches lead to only moderately satisfying results. Galibert et al. (2010) report a precision of 64.4 percent, a recall of 61 percent and a f1-score of 62.9 percent while Berkes (2018) does not report performance metrics. The fundamental reason behind these low scores is the lack of formatting requirements for in-text citations, which in turn leads to a large number of both citation cues and patent number formatting which are difficult to capture with regular-expression-based extraction techniques. On this point, Adams (2010) warned the community about the complexity of the extraction task. Using a random sample of USPTO patents, he found an "alarming" (p. 26) range of in-text patent citation formats. This problem severly limits citation extraction exercises that are dependent on lists of predefined rules, generally leading to mixed results and, above all, a lack of generalisability.

In order to overcome this limitation, natural-language processing (NLP) researchers have

[1] However, the first extracted citation is in 1846.

[2] https://console.cloud.google.com/marketplace/partners/patents-public-data

developed statistical models that can learn to find and tag entities, such as cited patents, using a training set of annotated documents, wherein a researcher has labeled the presence (or not) of the entities of interest. Although an in-depth presentation of the related Named Entity Recognition (NER) literature is outside the scope of this paper, we summarize the general working principles of these models below and direct interested readers to the excellent survey conducted by Li et al. (2020) for further reading.

The key to this approach is to view a text as two sequences: a sequence of tokens and a corresponding sequence of latent labels (for examples in this supplement, we label patent citations "PATCIT" and label everything else "O"). The task is, therefore, to predict the sequence of labels with the sequence of tokens as the input. The algorithm is trained on an annotated set of documents: a set of documents for which we know both the sequence of tokens and the sequence of labels. The probability that each token belongs to a given label is a recursive function of the token itself and its features (digits, capital letters, etc), as well as the neighbouring tokens (its context) and the *neighbouring labels.* The overall goal of the algorithm is to predict, correctly, the full sequence of latent labels for a given sequence of tokens. If a token (or a sequence of tokens) is unknown or otherwise deviates from the examples in the annotated set, the algorithm can still leverage the other attributes to decide which sequence of labels is the most probable for the whole sentence, leading to a considerable improvement in generalisability relative to rule-based approaches.

For example, let us assume that the algorithm has been trained on a corpus of texts where citations come in the following form (with d denoting any digit): "described by patent *d,ddd,ddd*" and where the corresponding sequence of labels is [O, O, O, PATCIT]. Let us further assume that the algorithm is supplied a new text with a slightly different form of citation, such as "described by Pat 9,535,657". Although the algorithm has never seen the token "Pat", it has learnt from the training data that the sequence of token "described by" frequently precedes a PATCIT label by two tokens. Combined with the fact that the token "9,535,657" exhibits the features frequently associated with a PATCIT (digits and commas), then the algorithm is expected to override the absence of the "patent" token and still to predict the right sequence of labels, [O, O, O, PATCIT].

The aforementioned limitations and improvement opportunities are well-known to the machine-learning community. In particular, Lopez (2010) developed the Grobid library in 2008 with the goal of overcoming the limitations of rule-based approaches by using a statistical approach. Grobid has now become an open-source project leveraging modern NLP to efficiently structure scientific documents in general, but retains a specific focus on patents. It includes models trained at extracting and structuring bibliographical references (scientific articles, books, proceedings, etc.) and patents from full-text documents. The algorithmic backbone of Grobid is the Conditional Random Fields (CRF) model, first introduced in 2001 (Lafferty et al., 2001) and belonging to the family of sequence-labeling models described above. The CRF model has been widely used in various fields and applications.[3]

Grobid's patent citations' extraction model was originally trained on 200 annotated full-text patents.[4] The specific features entering the CRF model that is implemented by Grobid to support patent citation detection include the relative position of the current token in the document, the matching of a common country code indicating the issuing office (e.g., US, EP, WO, etc.) and the matching of a common kind code indicating the document type (e.g., A1, A2, B1, B2, etc.).

The output of the extraction tasks is a set of text spans that were tagged as patent citations (e.g., "United States Patent 9,535,657"). The information extracted at this stage is not structured and, therefore, requires significant post-processing before it is usable.

1.3 Parsing task

The next step involves parsing the extracted patent citation strings. We take the raw span of the extracted citation as an input, with the goal of obtaining the following attributes: the country code of the patent authority, the patent document number, and the type of the patent document. This task is challenging due to the many formats in which patent citations occur in the text. Typically, the patent authority can appear as a code or a name (e.g "US Patent 9,535,657" or "United States Patent 9,535,657") either immediately next to the patent number or relatively far from it (e.g., "US Patent number 9,535,657" or "US Patents 9,911,050, 9,607,328, 9,535,657").

Lopez (2010) proposes an efficient solution for tackling this task. The fundamental idea is that both the sets of possible inputs and the sets of possible outputs for each patent attribute are finite (e.g., the list of patent organisation names and the list of their codes respectively). In addition, each element of the input vocabulary should be mapped to a unique element of the output vocabulary (e.g. "United States" with "US" or "European Patent Office" with "EP"). In short, for any given patent attribute, the parsing operation can be thought of as a translation operation between two languages with a finite vocabulary. In practive, this task is implemented in Grobid by a Finite State Transducer (FST), a process which appeared early in the history of automated translation.[5]

The output of this task is a well-structured set of attributes describing the cited patent.

[3] See Sutton and McCallum (2006) for a survey.

[4] The training set was enriched since that time and now includes 270 patents, comprising 51 percent EPO patents, 33 percent WIPO patents and the remaining 26 percent USPTO patents.

[5] See Roche and Schabes (1997) for an in-depth review of Finite State Transducers.

1.4 Matching task

The final task matches each extracted patent citation to a unique and consolidated identifier, in order to connect each cited patent document to commonly-used patent datasets. For patents, an identifier common to many patent datasets is the DOCDB publication number.[6] At this point, we depart from Grobid, which relies on the European Patent Office (EPO) search API[7] to perform the matching process and uses the EPO document number as its target and consolidation device.

Unfortunately, in the large majority of cases, in-text patent citations do not report the kind code of a patent, or report the original patent number rather than the version used in the DOCDB publication number, making it impossible to assemble the DOCDB publication number using the parsed attributes only. In order to overcome this limitation, we have relied on the Google Patents Linking Application Programming Interface (API).[8] Taking various inputs, such as the patent office code, the patent number and kind code, the API returns the associated DOCDB publication number. At a high level, the internal mechanism of this service is the following.[9] First, a large number of variations of each publication number are generated. For each variation, the original patent office and DOCDB formatted versions are indexed. Variations include adding and removing padding zeroes, two and four digit year dates inside of patent number, Japanese era variants, and different combinations of country code, patent number and kind code. Altogether, these variations constitute a large lookup table linking many variations of a publication number to its DOCDB formatted version. Then, at the time of lookup, punctuation is stripped and the country code, number and kind code are searched for before being checked for matches in the large lookup table. Note that there are two distinct services, one for applications and one for patents.[10] We decide which one to call based on the status attribute parsed by Grobid which can take four values: "application", "provisional", "patent" and "reissued". The first two trigger the application service, while the last two trigger the patent service.

Using the unique publication number returned by the Google Patents Linking API we were able to connect each cited document with richer information from patent datasets generally used by researchers (e.g., PATSTAT, PatentsView, IFI CLAIMS, etc.). We enriched each cited patent with the following attributes: publication date, application identifier, patent

[6] For simplicity, we use the term "publication number" for both the publication number (for published patents) and the application number (for patent applications).

[7] http://v3.espacenet.com/publicationDetails/biblio

[8] https://patents.google.com/api/match

[9] We thank Ian Wetherbee from Google Patents for his kind explanation.

[10] Applications: https://patents.google.com/api/match?appnum
Patents: https://patents.google.com/api/match?pubnum

publication identifier, INPADOC and DOCDB family identifiers.

1.5 Pipeline Summary

To illustrate the above process in its entirity, consider the following excerpt from the description of US-9606907-B2, which cites two U.S. patents:

> "Examples of circuits which can serve as the control circuit ... are described in more detail by U.S. Pat. Nos. 7,289,386 and 7,532,537, each of which is incorporated in its entirety by reference herein."

After the Grobid processing, we know that the patent US-9606907-B2 cites two patents from the U.S. patent office ("US" patent authority code) and that their original numbers are 7,289,386 and 7,532,537. Using the Google Patents Linking API, we find that the two patent citations embedded in the text can be uniquely identified by their publication numbers, namely US-7532537-B2 and US-7289386-B2.

The above pipeline was deployed remotely on a large-size compute engine from Amazon Web Services.[11] In order to increase speed, we used multi-processing, a technique that runs multiple processes in parallel at the same time. This technique is especially useful for 'cpu-bound' rather than 'io-bound' operations; that is, when computation is the main limiting factor rather than internal communication. Overall, from the 16,781,144 patent documents that we processed, we were able to extract 63,854,733 in-text patent citations, of which 49,409,629 were matched with a publication number comprising 13,611,323 unique patent documents.

2 Validation of extracted citations

In order to assess the quality of the citation dataset, we undertook a thorough validation exercise of the data and the extraction, parsing, and matching tasks. To do so, we relied on Prodigy, a scriptable annotation tool.[12] Lopez (2010) reports performance metrics for all these tasks, however the set of documents we are considering differs somewhat from the corpus in that work. In particular, a significant number of patents in our corpus are much older than any document considered for Grobid training and evaluation. Lastly, we also carried out detailed error analyses to support future improvement efforts.

[11] We used a t2.xlarge (4 cores and 16Gb of Ram) located in the "USA East Ohio" computing zone.

[12] Prodigy (2018-2020) https://prodi.gy/.

2.1 Data consistency

USPTO patent documents' format and scan quality (for older patents) has changed throughout the years. Before 1971, patents were largely unstructured with no clear delimitation between the metadata and the specification text that is of interest to us (see Figure 1). The modern patent format was introduced in 1971 and progressively replaced the old format before becoming the only format published from 1976. This new format is semi-structured and clearly distinguishes between the metadata sections and the specification section, *inter alia* (see Figure 2). These peculiarities of the source data have some notable implications on our output data.

First, the text of patents published in the old format includes the header of the patent. The header summarizes the main attributes of the patent, including its technological classes, title and, most importantly, its number. In this case, the extraction algorithm is likely to extract a patent citation which does not correspond to the kind of object we are looking for. Fortunately, this specific pitfall is relatively easy to spot as the citation appears very early in the text. Figure 3 reports the distribution of the rank of the first character of the extracted citations before and after 1976.[13] We observe a clear excess mass between 0 and 50 characters before 1976. To understand this pattern, we randomly drew 50 citations from pre-1971 patents that started before character 50. We found that 88 percent were patent-self references, 8 percent were technological classes and 4 percent were dates. To address this issue, we chose to flag all citations detected in a patent published before 1976 and starting before character 50 to facilitate their exclusion from analysis.

Second, in the old format, what we now call 'front-page citations' were printed *after* the patent specification, and these are also sometimes mistakenly included in our source data as part of the full-text of the patent. Since all patents have a different number of characters, we consider the relative location of these citations. Figure 4 shows their distribution as a function of their relative place (expressed in percentile) in the full text. Comparing the distribution before and after 1976 reveals a sizable excess mass in the pre-1976 distribution in the last four percent of the full-text characters. To check the nature of these citations, we examine a random sample of 100 citations extracted from patents published before 1976 and occurring in the last four percent of the characters, finding that 99 percent belong to what we would now call the 'front-page' citations section. Hence, for patents published before 1976, we flag all citations detected in the last 4 percent of the full-text for easy exclusion.

Third, during the transition period between the old and new formats (approximately throughout 1971–1975) there were two patent formats being published, complicating the

[13] The spike observed in the post-1976 graph is due to the 'Related Application' section of the patent, which can be addressed with patent family information.

delineation of the specification text section during this time period. As a result, we observed that 'full-texts' from this time mistakenly include the front-page of patents that are in the modern format. This can lead to the incidental extraction of 'in-text' citations that are actually front-matter, including front-page citations and references to the patent itself (including priority filings). Unfortunately, there is no straightforward solution to this problem. We encourage data users to systematically ignore patents that are both in text and front page citations during this time span.

All figures, with the exceptions of 3 and 4, exclude flagged patent citations as they are unlikely to correspond to real in-text patent citations (unless explicitly specified).

2.2 Extraction task

Lopez (2010) reports performance metrics for the extraction task. Using cross-validation, a technique consisting in training the model ten times using 80 percent of the sample and testing it on the remaining 20 percent, the author reports the following average performance metrics: 94.66 percent of precision, 96.16 percent of recall and a f1-score of 95.4 percent. As far as we know, these are the best performances reported in the literature to date. Although this motivated our choice to use Grobid, we are fully aware that our dataset partly differs from the Grobid training set and thus performance could be affected.

In order to evaluate the quality of our citation extraction, we randomly sampled 160 U.S. patents and annotated them by hand. As previously discussed, the citation of a patent can come in various ways. For instance, the country of the patent office can be reported as a code preceding the patent number, as a name anywhere in the surrounding of the patent number, etc. In this context, the only stable element of a patent citation is the patent number itself. That is why Grobid returns the first and the last character of the patent number of detected patent citations. Hence, our validation exercise consisted in comparing the spans detected by Grobid as a patent number and the spans labelled by humans as a patent number. Each patent was annotated by a single human annotator using the platform featured by Figure 5a.[14] The body of the text is displayed together with annotations from Grobid predictions and the annotator goes through the text to correct missing and wrong annotations. The tagged spans are saved upon exit. As depicted by Figure 6, the validation sample and the universe of citing patents display similar distributions by publication year.

From the 160 random U.S. patents in the validation set, human annotators found that 103 (64.4 percent) patents cited at least one patent for a total of 470 in-text patent citations. Table 4 reports the extraction performance that we obtained, together with the Galibert et al.

[14] "Human annotators" are the coauthors of this paper.

(2010) and Lopez (2010) benchmarks. Comparing 'gold' annotations from human annotators with the predictions obtained from Grobid, we find that Grobid exhibits a satisfying 97 percent precision and 82 percent recall (f1-score nearing 90 percent). Importantly, these results significantly outperform Galibert et al. (2010), who used regular expressions to extract citations; they reported a precision of 64.4 percent and a recall of 61 percent. This result clearly confirms that a statistical approach is much more effective than a regular expression approach in the context of in-text patent citations' extraction. Interestingly, the performance obtained by Grobid on our extended corpus is very similar to the benchmark reported by Lopez (2010) regarding precision (97.44% vs 97%) but lower in terms of recall (97.74% vs. 82%). This difference means that, applied to our extended corpus, Grobid is as reliable as reported in Lopez (2010) when it has detected a patent citation. However, it misses patent citations more often in our extended corpus; this is due to older forms of citations appearing in early-twentieth century patents.

The error analysis suggests that both false positives and false negatives exhibit patterns that could be specifically addressed by future improvements of the Grobid training set. Table 5 provides examples for each category of errors that we were able to identify. Starting with false negatives (real citations that were not detected by Grobid), we find three categories of context generating this type of errors: 1) the context does not clearly mention "patent" or "application" but rather implicitly suggests a patent citation; 2) the patent is cited in the form "inventor (date) <PATCIT>" and 3) the patent is cited as "Serial Number <PATCIT>". While category 1) could have been expected and would certainly be hard to correct without generating a large number of false positives, categories 2) and 3) might certainly be partly addressed by augmenting the training dataset with older patents that tend to adopt this form of citations more often. Looking at false positives (text spans that were wrongly identified by Grobid as citations) we find three categories of errors as well: 1) technological classes reported as "dd/ddd", 2) date and 3) docket number. Note that the categories 2 and 3 have only one occurrence each.

2.3 Parsing task

Grobid's FST implementation was built manually based on 1,500 patent citation examples. It was then evaluated on 250 references which were unseen before. Lopez (2010) reports a 97.2 percent accuracy for the full parsing task (patent organisation code, number and kind code).

In order to validate the quality of the parsing task that we conducted, we randomly sampled 300 extracted citations alongside their parsed attributes. Within the text, attributes can be relatively far from the patent number that serves as the citation anchor. Hence, it was necessary to provide the human annotators with a contextualized citation; using the patent number reported by Grobid as an anchor, we extracted a chunk of text containing a window of ten tokens on the right and left of the detected patent. This text and the tagged patent were then displayed to the annotator together with the Grobid parsed attribute as illustrated by Figure 5b. The annotator would then accept or reject the attribute depending on what they actually found in the text. Each example was validated by a single annotator whose decisions were saved upon exit.

Since the attributes can be used independently, a detailed understanding of the performance and errors for each attribute may be valuable for the community. Hence, we performed three distinct validation exercises, one for each attribute. Our results are summarized in Table 7.

We first checked for sample representativeness with respect to the parsing of the patent organisation. Table 6 reports the distribution of the patent organisations in the validation sample. It appears that two-thirds of the citations in the sample were mapped to the U.S. patent office. This result is consistent with the results that we report for the full dataset in the main body of this article. Similarly, the patent organisations in the remaining third of the validation sample are also the most represented organisations at scale, including the Japan Patent Office, the World Intellectual Property Organisation, the European Patent Office and the German Patent Office. Of the 300 examples that we validated, we found only five errors, leading to a 98.3 percent accuracy score. Errors were spread over five distinct patent offices and we do not observe any systematic confusion between patent offices, which suggest that errors generate noise rather than a systematic bias.[15]

When it comes to the parsing of the patent number, there is no specific way to checking sample representativeness. Overall, of the 300 examples that we validated, we found thirteen errors, leading to a 95.7 percent accuracy score. Among the errors, we find two recurring cases. First, patent citations in their Paris Cooperation Treaty (PCT) form (e.g., PCT/EP2005/008238) generate patent numbers mixing part of the letters in the prefix and the patent number itself (e.g., PTEP2005008238). Second, as already reported in Lopez (2010), we found that Grobid removes the first letter of the patent number of Japanese applications with date prior to 2000 (e.g., H08-193210, where H stands for the Heisei era that spanned from 1989 to 2019). However, this indication is key to uniquely identify the application. This letter refers to the era and acts as the time marker. Note that this specific issue is partially fixed by the Google Patent matching API.

[15] The five offices were: SA (Saudi Arabia), AL (Albania), CH (Switzerland), DE (Germany) and BE (Belgium).

Lastly, we validated the parsing of the kind code, which indicates the specific kind of document the citation refers to (granted patent, application, reissue, design, etc.). For 502 random samples, we obtain an accuracy of 97.6 percent. Note, however, that this measure includes a large proportion of null results as the kind code is rarely reported in the text. In order to further characterize the quality of the parsing, we drew a sample of 50 citations where the parsed kind code was not null. We found 7 mistakes, giving a 'conditional' accuracy of 86 percent. Specifically, we found three groups of parsing errors: errors due to unconventional formatting, issues with optical character recognition on scanned documents, and Grobid mistakenly interpreting 'Cl' (abbreviation of 'class') for the 'C' kind code. Importantly, every instance in standard form was correctly parsed.

2.4 Matching task

The matching task involves associating the extracted attributes with a unique identifier, that is, the DOCDB publication number. In order to validate this step of the process, we randomly sampled 200 citations from our final dataset and compared the concatenation of the parsed attributes with the publication number provided by the Google Patent's Linking API. The annotator's task was to answer the following questions: i) if there is a matched publication number, is it the right one? ii) if there is no match, would it be possible to find one for a human reasonably well trained in the task? A single human annotator fulfilled this validation exercise. Based on that, we can assign each annotated example to a standard classification outcome category and derive the associated performance metrics. Table 8 summarizes these categories, their contents and the results from the validation exercise.

On the 200 examples in the validation sample, we find that 147 were matched and 53 remained unmatched. Among the 147 matches, 137 were correct (True positives) and 10 were incorrect (False positives) including six patents that could have been matched and four non-patent items that should not have been matched. Among the 53 unmatched examples, we found that 17 could have been matched (False negatives) while no match could be found for the remaining 36 (True negatives). Overall, we find that the matching procedure achieves a 93.2 percent precision and a 88.96 percent recall, leading to a 91.06 percent f1-score.

Next, we explored the nature of the errors and non-matches. Tables 9 and 10 respectively detail errors occurring during matching and cases classified as unmatchable by the human annotator. We find that errors arising at this final step of the processing pipeline are often inherited from upstream steps. Among the ten incorrect matches, half are due to either a parsing error or an extraction error. In the same way, among the thirty-six unmatched citations that were judged unmatchable, 56 percent were directly related to either a parsing error or an extraction error. Another group of errors seems to arise from the specificities of in-text citations and their intrinsic ambiguities. This group includes citations of provisional patent applications[16] (which may never appear in standard patent datasets) and partial citations that even a human cannot match. This family of errors represent 41 percent of the thirty-six unmatchable detected citations in our validation sample. Finally, focusing on the unmatched citations that a human can match reveals some blind spots of the Linking API. Over the seventeen cases in this category, 52 percent are caused by missing zeros after the country code/year or a Japanese publication number reporting the year after the serial number rather than before it as is usually expected.

While the previous step can characterize the performance of the matching procedure generally, the small size of the validation sample means we are unlikely to uncover rare irregularities that might nonetheless be numerous at large scale.

Considering the full dataset, Figure 7 shows the yearly number (7a) and share (7b) of citing patents according to the matching status of the extracted in-text citations.[17] Patents with all in-text citations matched to a publication number represent 42.7 percent of the total, whereas those with only some in-text citations matched represent 32.7 percent. Patents with no in-text citations matched account for the remaining 24.6 percent. From 1947 to 1964, patents with all in-text citations matched report an increasing yearly share, from around 40 percent to almost 70 percent. For patents published between 1965 and 1975, the performance of our matching procedure worsens, as the proportion of patents with some or no citations matched grows markedly. From 1976 onwards, the share of patents with all citations matched jumps to 77 percent in 1976 before slowly decreasing since that time until the current day.

These aggregate figures mask high variation between the patent offices associated with the cited patent documents. Table 11 reports the number of extracted in-text citations alongside the share of those that were matched, for the top five patent offices in our dataset. More than half of the extracted in-text citations are made to patents filed at the USPTO (about 58% of the total). We are able to match 89 percent of them to their correct publication number. Patents filed at the World Intellectual Patent Organisation (WIPO) and the Japan Patent Office (JPO) with, respectively, around 6.5 million (10% of the total) and 5.7 million (9% of the total) citations are the second and third largest groups. We match almost 82 percent of the citations to WIPO patent filings and around 77 percent to JPO patent filings. We obtain a similar match rate (73%) for the 1.4 million extracted citations to patents filed at the German Patent and Trade Mark Office (DPMA). Lastly, we obtain less satisfactory

[16] A provisional application is a legal document filed at the patent office that establishes an early filing date, but does not mature into an issued patent unless the applicant files a regular non-provisional patent application within one year.

[17] We consider only citing patents with at least one extracted in-text citation.

match rates for citations to EPO patent filings; of 2.2 million citations, we match only 51 percent.

3 Data description

Data generation and validation reproducibility is guaranteed by the codebase hosted on the project repository. Validation data are supported by Data Version Control (DVC). Since the project is open-source and continuously improving, exact replication of the data and results detailed above requires the user to choose the tag '0.3.1' of the code.[18]

The data are reported as a nested table that is structured as follows:

- Each entry corresponds to the patent document from which we extracted patent citations. Each such patent is identified by a publication number (primary key). In addition to the publication number, we also report its publication date, application identifier, and patent publication identifier. We also include DOCDB and INPADOC family codes, which identify the constellation of inter-related patents that protect the same invention across jurisdictions.
- Each entry has a citation variable in which cited patents are listed and their attributes are nested. Any detected patent is represented by the two attributes parsed by Grobid, the code of its patent office and its original number. When these two attributes can be matched with a publication number, we also report the publication date, application identifier, patent publication identifier and the DOCDB and INPADOC family identifiers. We also report a flag indicating whether the extracted citation is likely to belong to the front matter or the header.

The schema of the table is detailed below.

Name	**Description**	**Type**	**Nb non null**
publication_number	Publication number.	STR	16781144
publication_date	Publication date (yyyymmdd).	INT	15862299
appln_id	PATSTAT application identification. Surrogate key: Technical unique identifier without any business meaning	INT	15862299

[18] [Weblink redacted to preserve anonymity.]

Name	Description	Type	Nb non null
pat_publn_id	PATSTAT Patent publication identification. Surrogate key for patent publications.	INT	15862299
docdb_family_id	Identifier of a DOCDB simple family. Means that most probably the applications share exactly the same priorities (Paris Convention or technical relation or others).	INT	15862299
inpadoc_family_id	Identifier of an INPADOC extended priority family. Means that the applications share a priority directly or indirectly via a third application.	INT	15862299
citation		REC	16781144
___.country_code	Country code of the cited patent. Parsed by Grobid.	STR	64185636
___.original_number	Original number of the cited patent. Parsed by Grobid.	STR	64185636
___.kind_code	Kind code of the cited patent. Parsed by Grobid.	STR	6096368
___.status	The status of the cited patent. Parsed by Grobid.	STR	64185636
___.pubnum	Concatenation of country code, original number and kind code of the cited patent. Based on attributes parsed attributes.	STR	64185636
___.publication_number	Publication number of the cited patent. Obtained from the google patent linking API.	STR	49542360
___.publication_date	Publication date (yyyymmdd) of the cited patent based on the matched publication_number.	INT	49231609

Name	Description	Type	Nb non null
____.appln_id	PATSTAT application identification of the cited patent. Based on the matched publication_number. Surrogate key: Technical unique identifier without any business meaning.	INT	49231609
____.pat_publn_id	PATSTAT Patent publication identification of the cited patent. Based on the matched publication_number. Surrogate key for patent publications.	INT	49231609
____.docdb_family_id	Identifier of a DOCDB simple family of the cited patent. Based on the matched publication_number. Means that most probably the applications share exactly the same priorities (Paris Convention or technical relation or others).	INT	49231609
____.inpadoc_family_id	Identifier of an INPADOC extended priority family of the cited patent. Based on the matched publication_number. Means that the applications share a priority directly or indirectly via a third application.	STR	49231609
____.flag	Flag detected citations which are likely to be in the header rather than in the specification itself. Flag is True for citations extracted from patents published in the pre-1976 format and with all occurrences detected before character 50 or in the last 4 percent of the text. It is recommended to exclude those citations from most analyses.	BOOL	71407446
____.char_start	First character of the detected cited patent. Refers to description_localized.text in patents-public-data.patents.publications.	INT	71407446

Name	**Description**	**Type**	**Nb non null**
___.char_end	Last character of the detected cited patent. Refers to description_localized.text in patents-public-data.patents.publications.	INT	71407446

Notes: Nested variables are denoted by a dot. For instance, ___.country_code is the country code of a cited patent nested in the citation variable.

References

Adams, S. (2010). The text, the full text and nothing but the text: Part 1–standards for creating textual information in patent documents and general search implications. *World Patent Information*, 32(1):22–29.

Berkes, E. (2018). Comprehensive universe of US patents (CUSP): Data and facts. *Unpublished manuscript, Ohio State University*.

de Rassenfosse, G., Jaffe, A., and Raiteri, E. (2019). The procurement of innovation by the US government. *PLOS ONE*, 14(8):e0218927.

Galibert, O., Rosset, S., Tannier, X., and Grandry, F. (2010). Hybrid citation extraction from patents. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, pages 17–23.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Li, J., Sun, A., Han, J., and Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1.

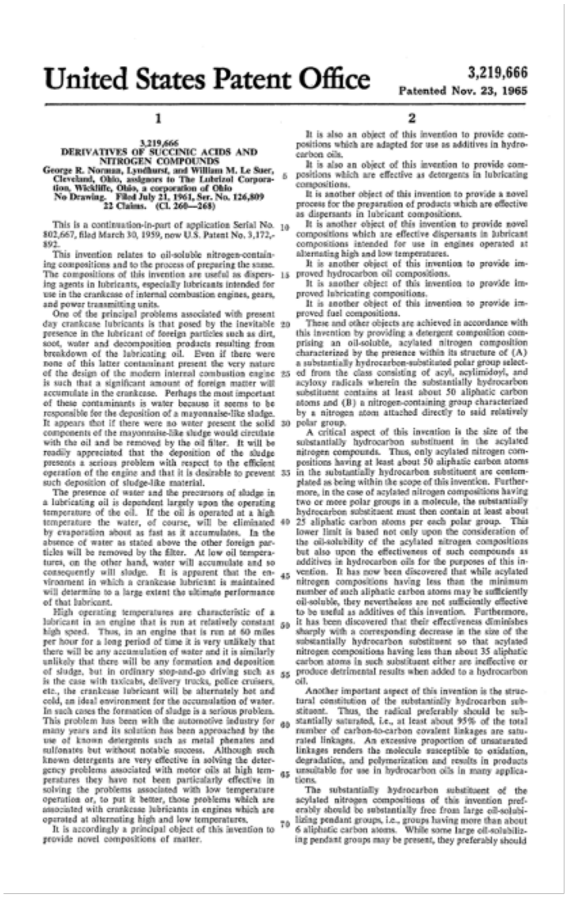
Lopez, P. (2010). Automatic extraction and resolution of bibliographical references in patent documents. In *Information Retrieval Facility Conference*, pages 120–135. Springer.

Roche, E. and Schabes, Y. (1997). *Finite-state language processing*. MIT press.

Sutton, C. and McCallum, A. (2006). An introduction to conditional random fields for relational learning. *Introduction to statistical relational learning*, 2:93–128.

Figures

Figure 1: Example of the USPTO “old” patent format (US-3219666-A)



United States Patent Office **3,219,666**

Patented Nov. 23, 1965

1

3,219,666
DERIVATIVES OF SUCCINIC ACIDS AND NITROGEN COMPOUNDS
George R. Norman, Lyndhurst, and William M. Le Suer, Cleveland, Ohio, assignors to The Lubrizol Corporation, Wickliffe, Ohio, a corporation of Ohio
No Drawing. Filed July 21, 1961, Ser. No. 126,809
22 Claims. (Cl. 260—268)

This is a continuation-in-part of application Serial No. 802,667, filed March 30, 1959, now U.S. Patent No. 3,172,892.

This invention relates to oil-soluble nitrogen-containing compositions and to the process of preparing the same. The compositions of this invention are useful as dispersing agents in lubricants, especially lubricants intended for use in the crankcase of internal combustion engines, gears, and power transmitting units.

One of the principal problems associated with present day crankcase lubricants is that posed by the inevitable presence in the lubricant of foreign particles such as dirt, soot, water and decomposition products resulting from breakdown of the lubricating oil. Even if there were none of this latter contaminant present the very nature of the design of the modern internal combustion engine is such that a significant amount of foreign matter will accumulate in the crankcase. Perhaps the most important of these contaminants is water because it seems to be responsible for the deposition of a mayonnaise-like sludge. It appears that if there were no water present the solid components of the mayonnaise-like sludge would circulate with the oil and be removed by the oil filter. It will be readily appreciated that the deposition of the sludge presents a serious problem with respect to the efficient operation of the engine and that it is desirable to prevent such deposition of sludge-like material.

The presence of water and the precursors of sludge in a lubricating oil is dependent largely upon the operating temperature of the oil. If the oil is operated at a high temperature the water, of course, will be eliminated by evaporation about as fast as it accumulates. In the absence of water as stated above the other foreign particles will be removed by the filter. At low oil temperatures, on the other hand, water will accumulate and so consequently will sludge. It is apparent that the environment in which a crankcase lubricant is maintained will determine to a large extent the ultimate performance of that lubricant.

High operating temperatures are characteristic of a lubricant in an engine that is run at relatively constant high speed. Thus, in an engine that is run at 60 miles per hour for a long period of time it is very unlikely that there will be any accumulation of water and it is similarly unlikely that there will be any formation and deposition of sludge, but in ordinary stop-and-go driving such as is the case with taxicabs, delivery trucks, police cruisers, etc., the crankcase lubricant will be alternately hot and cold, an ideal environment for the accumulation of water. In such cases the formation of sludge is a serious problem. This problem has been with the automotive industry for many years and its solution has been approached by the use of known detergents such as metal phenates and sulfonates but without notable success. Although such known detergents are very effective in solving the detergency problems associated with motor oils at high temperatures they have not been particularly effective in solving the problems associated with low temperature operation or, to put it better, those problems which are associated with crankcase lubricants in engines which are operated at alternating high and low temperatures.

It is accordingly a principal object of this invention to provide novel compositions of matter.

2

It is also an object of this invention to provide compositions which are adapted for use as additives in hydrocarbon oils.

It is also an object of this invention to provide compositions which are effective as detergents in lubricating compositions.

It is another object of this invention to provide a novel process for the preparation of products which are effective as dispersants in lubricant compositions.

It is another object of this invention to provide novel compositions which are effective dispersants in lubricant compositions intended for use in engines operated at alternating high and low temperatures.

It is another object of this invention to provide improved hydrocarbon oil compositions.

It is another object of this invention to provide improved lubricating compositions.

It is another object of this invention to provide improved fuel compositions.

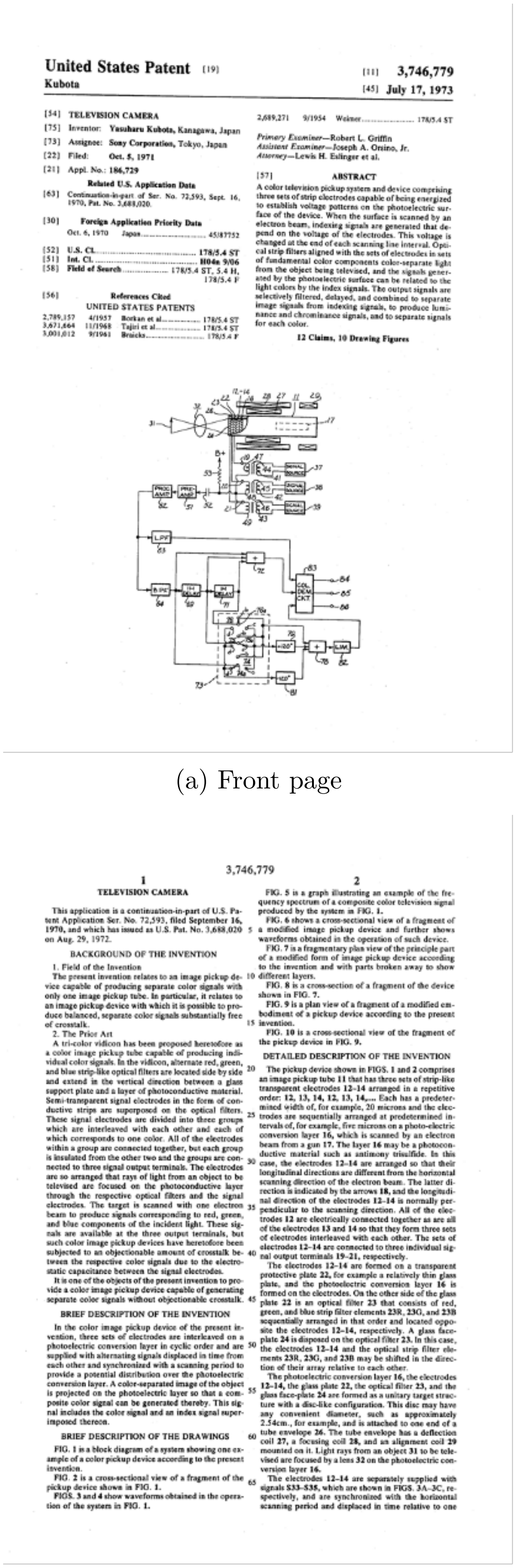
These and other objects are achieved in accordance with this invention by providing a detergent composition comprising an oil-soluble, acylated nitrogen composition characterized by the presence within its structure of (A) a substantially hydrocarbon-substituted polar group selected from the class consisting of acyl, acylimidoyl, and acyloxy radicals wherein the substantially hydrocarbon substituent contains at least about 50 aliphatic carbon atoms and (B) a nitrogen-containing group characterized by a nitrogen atom attached directly to said relatively polar group.

A critical aspect of this invention is the size of the substantially hydrocarbon substituent in the acylated nitrogen compounds. Thus, only acylated nitrogen compositions having at least about 50 aliphatic carbon atoms in the substantially hydrocarbon substituent are contemplated as being within the scope of this invention. Furthermore, in the case of acylated nitrogen compositions having two or more polar groups in a molecule, the substantially hydrocarbon substituent must then contain at least about 25 aliphatic carbon atoms per each polar group. This lower limit is based not only upon the consideration of the oil-solubility of the acylated nitrogen compositions but also upon the effectiveness of such compounds as additives in hydrocarbon oils for the purposes of this invention. It has now been discovered that while acylated nitrogen compositions having less than the minimum number of such aliphatic carbon atoms may be sufficiently oil-soluble, they nevertheless are not sufficiently effective to be useful as additives of this invention. Furthermore, it has been discovered that their effectiveness diminishes sharply with a corresponding decrease in the size of the substantially hydrocarbon substituent so that acylated nitrogen compositions having less than about 35 aliphatic carbon atoms in such substituent either are ineffective or produce detrimental results when added to a hydrocarbon oil.

Another important aspect of this invention is the structural constitution of the substantially hydrocarbon substituent. Thus, the radical preferably should be substantially saturated, i.e., at least about 95% of the total number of carbon-to-carbon covalent linkages are saturated linkages. An excessive proportion of unsaturated linkages renders the molecule susceptible to oxidation, degradation, and polymerization and results in products unsuitable for use in hydrocarbon oils in many applications.

The substantially hydrocarbon substituent of the acylated nitrogen compositions of this invention preferably should be substantially free from large oil-solubilizing pendant groups, i.e., groups having more than about 6 aliphatic carbon atoms. While some large oil-solubilizing pendant groups may be present, they preferably should

Figure 2: Example of the USPTO “new” patent format (US-3746779-A)



United States Patent [19] [11] **3,746,779**

Kubota [45] **July 17, 1973**

[54] TELEVISION CAMERA

[75] Inventor: **Yasuharu Kubota,** Kanagawa, Japan

[73] Assignee: **Sony Corporation,** Tokyo, Japan

[22] Filed: **Oct. 5, 1971**

[21] Appl. No.: **186,729**

Related U.S. Application Data

[63] Continuation-in-part of Ser. No. 72,593, Sept. 16, 1970, Pat. No. 3,688,020.

[30] **Foreign Application Priority Data**

Oct. 6, 1970 Japan........................ 45/87752

[52] **U.S. Cl.**.. **178/5.4 ST**

[51] **Int. Cl.**.. **H04n 9/06**

[58] **Field of Search**.................. 178/5.4 ST, 5.4 H, 178/5.4 F

[56] **References Cited**

UNITED STATES PATENTS

2,789,157	4/1957	Borkan et al...............	178/5.4 ST
3,671,664	11/1968	Tajiri et al..................	178/5.4 ST
3,001,012	9/1961	Braicks......................	178/5.4 F
2,689,271	9/1954	Weimer.....................	178/5.4 ST

Primary Examiner—Robert L. Griffin
Assistant Examiner—Joseph A. Orsino, Jr.
Attorney—Lewis H. Eslinger et al.

[57] **ABSTRACT**

A color television pickup system and device comprising three sets of strip electrodes capable of being energized to establish voltage patterns on the photoelectric surface of the device. When the surface is scanned by an electron beam, indexing signals are generated that depend on the voltage of the electrodes. This voltage is changed at the end of each scanning line interval. Optical strip filters aligned with the sets of electrodes in sets of fundamental color components color-separate light from the object being televised, and the signals generated by the photoelectric surface can be related to the light colors by the index signals. The output signals are selectively filtered, delayed, and combined to separate image signals from indexing signals, to produce luminance and chrominance signals, and to separate signals for each color.

12 Claims, 10 Drawing Figures

(a) Front page

3,746,779

1

TELEVISION CAMERA

This application is a continuation-in-part of U.S. Patent Application Ser. No. 72,593, filed September 16, 1970, and which has issued as U.S. Pat. No. 3,688,020 on Aug. 29, 1972.

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates to an image pickup device capable of producing separate color signals with only one image pickup tube. In particular, it relates to an image pickup device with which it is possible to produce balanced, separate color signals substantially free of crosstalk.

2. The Prior Art

A tri-color vidicon has been proposed heretofore as a color image pickup tube capable of producing individual color signals. In the vidicon, alternate red, green, and blue strip-like optical filters are located side by side and extend in the vertical direction between a glass support plate and a layer of photoconductive material. Semi-transparent signal electrodes in the form of conductive strips are superposed on the optical filters. These signal electrodes are divided into three groups which are interleaved with each other and each of which corresponds to one color. All of the electrodes within a group are connected together, but each group is insulated from the other two and the groups are connected to three signal output terminals. The electrodes are so arranged that rays of light from an object to be televised are focused on the photoconductive layer through the respective optical filters and the signal electrodes. The target is scanned with one electron beam to produce signals corresponding to red, green, and blue components of the incident light. These signals are available at the three output terminals, but such color image pickup devices have heretofore been subjected to an objectionable amount of crosstalk between the respective color signals due to the electrostatic capacitance between the signal electrodes.

It is one of the objects of the present invention to provide a color image pickup device capable of generating separate color signals without objectionable crosstalk.

BRIEF DESCRIPTION OF THE INVENTION

In the color image pickup device of the present invention, three sets of electrodes are interleaved on a photoelectric conversion layer in cyclic order and are supplied with alternating signals displaced in time from each other and synchronized with a scanning period to provide a potential distribution over the photoelectric conversion layer. A color-separated image of the object is projected on the photoelectric layer so that a composite color signal can be generated thereby. This signal includes the color signal and an index signal superimposed thereon.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a block diagram of a system showing one example of a color pickup device according to the present invention.

FIG. 2 is a cross-sectional view of a fragment of the pickup device shown in FIG. 1.

FIGS. 3 and 4 show waveforms obtained in the operation of the system in FIG. 1.

2

FIG. 5 is a graph illustrating an example of the frequency spectrum of a composite color television signal produced by the system in FIG. 1.

FIG. 6 shows a cross-sectional view of a fragment of a modified image pickup device and further shows waveforms obtained in the operation of such device.

FIG. 7 is a fragmentary plan view of the principle part of a modified form of image pickup device according to the invention and with parts broken away to show different layers.

FIG. 8 is a cross-section of a fragment of the device shown in FIG. 7.

FIG. 9 is a plan view of a fragment of a modified embodiment of a pickup device according to the present invention.

FIG. 10 is a cross-sectional view of the fragment of the pickup device in FIG. 9.

DETAILED DESCRIPTION OF THE INVENTION

The pickup device shown in FIGS. 1 and 2 comprises an image pickup tube **11** that has three sets of strip-like transparent electrodes **12–14** arranged in a repetitive order: **12, 13, 14, 12, 13, 14,...** Each has a predetermined width of, for example, 20 microns and the electrodes are sequentially arranged at predetermined intervals of, for example, five microns on a photo-electric conversion layer **16**, which is scanned by an electron beam from a gun **17**. The layer **16** may be a photoconductive material such as antimony trisulfide. In this case, the electrodes **12–14** are arranged so that their longitudinal directions are different from the horizontal scanning direction of the electron beam. The latter direction is indicated by the arrows **18**, and the longitudinal direction of the electrodes **12–14** is normally perpendicular to the scanning direction. All of the electrodes **12** are electrically connected together as are all of the electrodes **13** and **14** so that they form three sets of electrodes interleaved with each other. The sets of electrodes **12–14** are connected to three individual signal output terminals **19–21**, respectively.

The electrodes **12–14** are formed on a transparent protective plate **22**, for example a relatively thin glass plate, and the photoelectric conversion layer **16** is formed on the electrodes. On the other side of the glass plate **22** is an optical filter **23** that consists of red, green, and blue strip filter elements **23R, 23G,** and **23B** sequentially arranged in that order and located opposite the electrodes **12–14**, respectively. A glass face-plate **24** is disposed on the optical filter **23**. In this case, the electrodes **12–14** and the optical strip filter elements **23R, 23G,** and **23B** may be shifted in the direction of their array relative to each other.

The photoelectric conversion layer **16**, the electrodes **12–14**, the glass plate **22**, the optical filter **23**, and the glass face-plate **24** are formed as a unitary target structure with a disc-like configuration. This disc may have any convenient diameter, such as approximately 2.54cm., for example, and is attached to one end of a tube envelope **26**. The tube envelope has a deflection coil **27**, a focusing coil **28**, and an alignment coil **29** mounted on it. Light rays from an object **31** to be televised are focused by a lens **32** on the photoelectric conversion layer **16**.

The electrodes **12–14** are separately supplied with signals **S33–S35**, which are shown in FIGS. 3A–3C, respectively, and are synchronized with the horizontal scanning period and displaced in time relative to one

(b) Specification

Figure 3: Empirical probability distribution function of citation detection as a function of the starting character

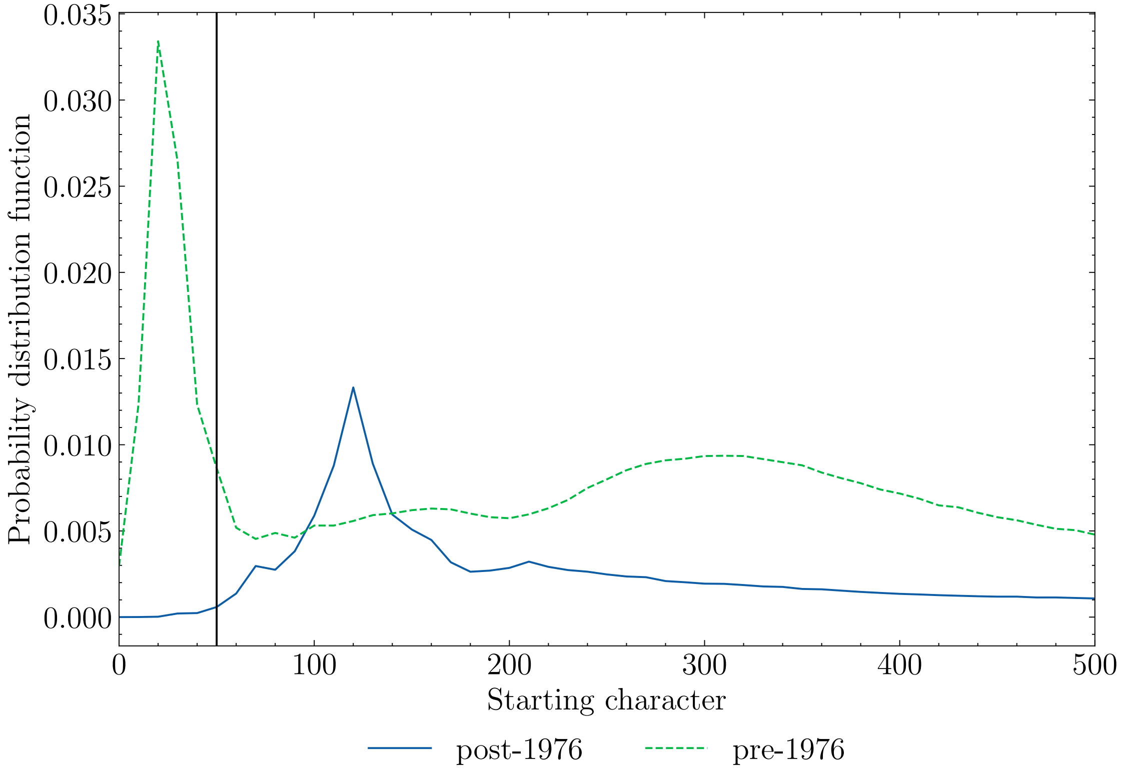


Figure 4: Empirical probability distribution function of citation detection as a function of the relative place of the starting character

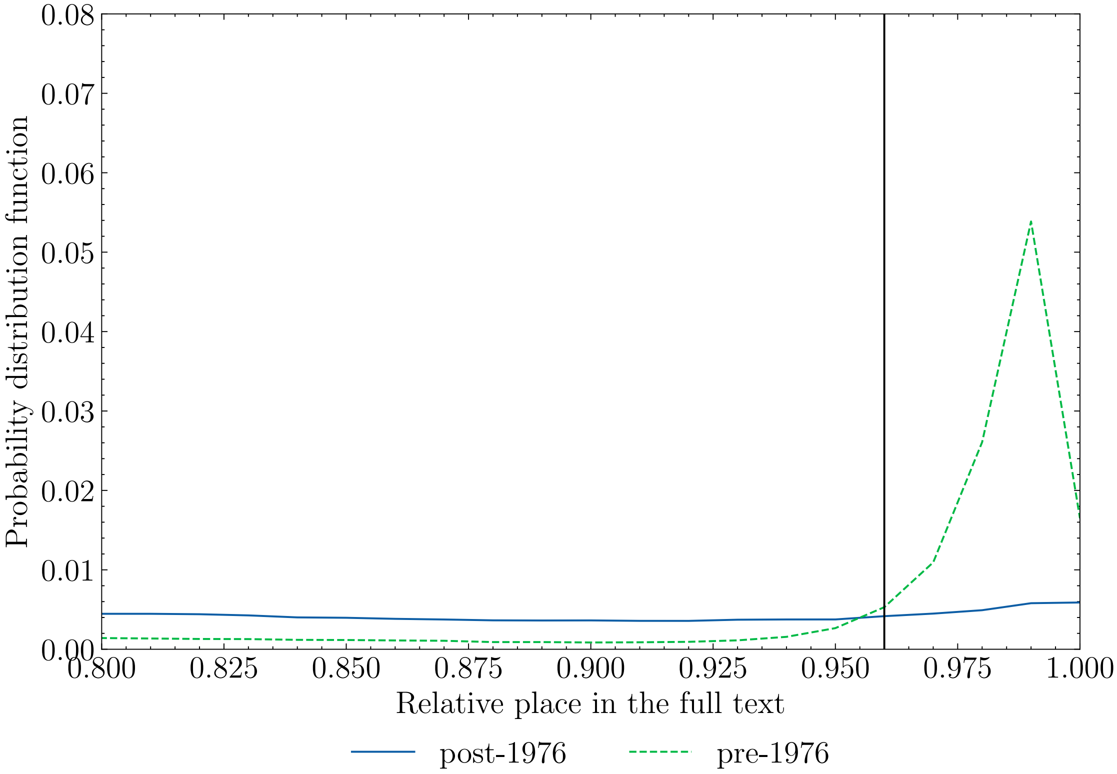
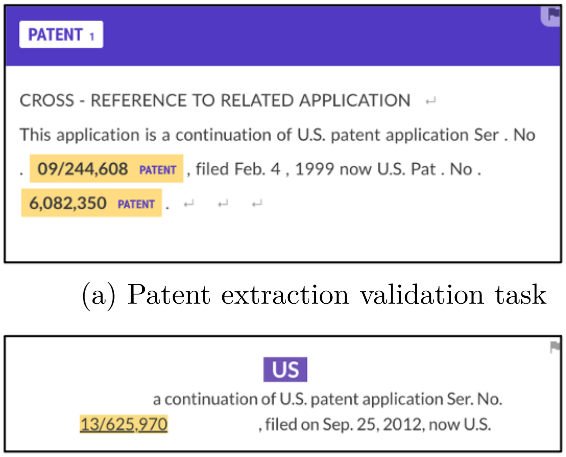


Figure 5: Preview of the annotation platform



CROSS - REFERENCE TO RELATED APPLICATION ↵

This application is a continuation of U.S. patent application Ser . No . 09/244,608 PATENT , filed Feb. 4 , 1999 now U.S. Pat . No . 6,082,350 PATENT . ↵ ↵ ↵

(a) Patent extraction validation task

a continuation of U.S. patent application Ser. No. 13/625,970 , filed on Sep. 25, 2012, now U.S.

(b) Patent parsing validation task (organisation name)

Figure 6: Empirical cumulative distribution function of patents in the validation sample and in the universe of U.S. patents (by decade)

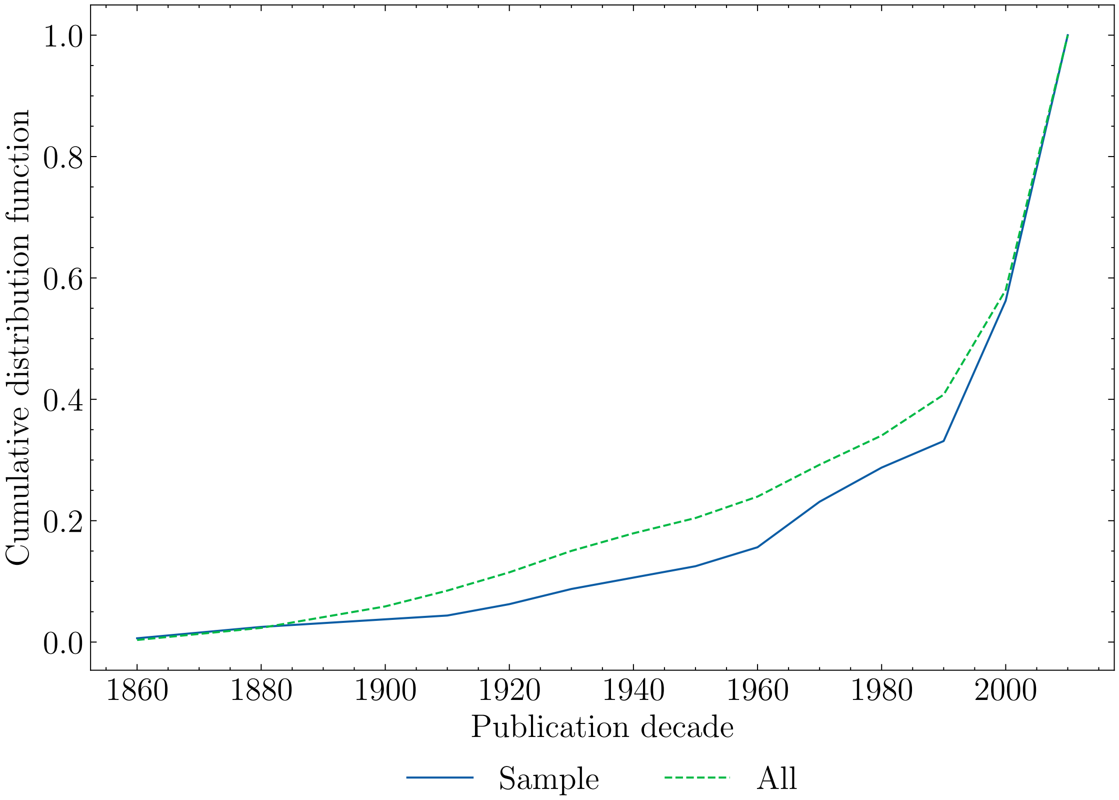
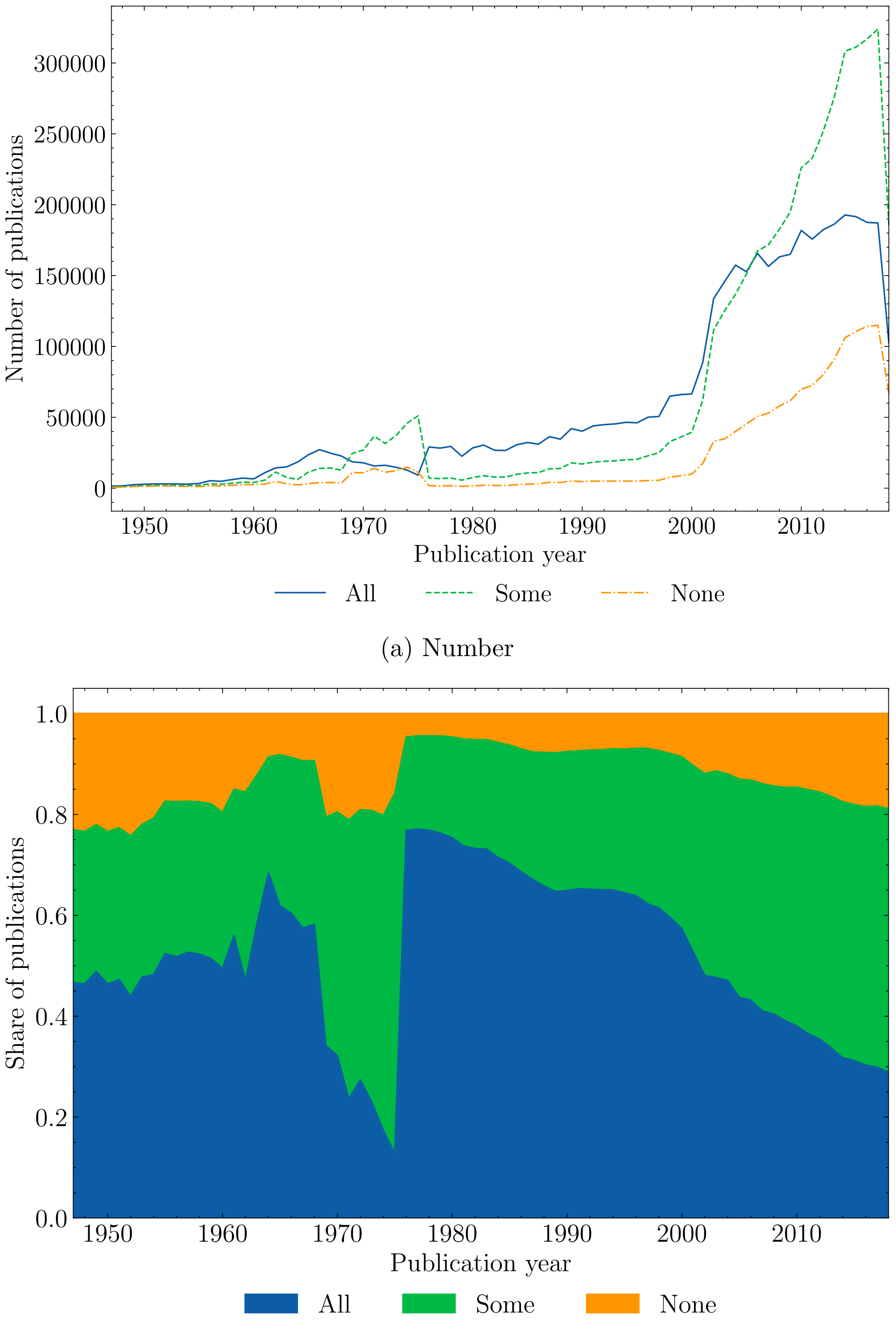


Figure 7: Citing patents over time by in-text citation match status



(a) Number

(b) Share

Notes: "All" (blue solid line) refers to patent publications for which it was possible to match all extracted in-text citations. "Some" (orange dashed line) refers to patent publications for which it was possible to match only some extracted in-text citations. "None" (green dash-dot line) depicts patent publications for which we could not match any extracted in-text citation.

Tables

Table 2: Composition of the dataset

Kind code	**Kind of document**		**Number**	**Share**
	Pre 2001	*Post 2001*		
A	Patent	Patent application	11,909,035	0.71
B	Reexamination certificate	Patent	4,188,597	0.25
S	-	Design patent	613,050	0.04
P	Plant patent	Plant patent & Plant patent application	34,852	2.00E-3
E	-	Reissued patent	32,226	2.00E-3
H	-	Statutory invention registration (SIR)	2,255	1.00E-4
I	-	-	1,129	6.00E-5

Table 3: Composition of the dataset: focus on patents and applications

Kind code	**Kind of document**		**Number**	**Share**
	Pre 2001	*Post 2001*		
A	Patent	-	6,145,197	0.37
A1	-	Patent application publication	5,753,613	0.34
A2	-	Patent application publication (republication)	1,742	1.00E-4
A9	-	Patent application publication (corrected publication)	8,483	5.00E-4
B1	-	Patent (no pre-grant publication)	776,074	0.04
B2	-	Patent	3,412,523	0.2

Table 4: In-text patent citations extraction performance

	Number of patents in the test set	**Avg number of patent tags per patent**	**Precision**	**Recall**	**F1-score**
Galibert et al. (2010)	760	12.75	64.4%	61.0%	62.6%
Lopez (2010)	20	9.96	97.44%	97.74%	97.68%
Verluise et al (2020)	160	2.93	97%	82%	89.2%

Table 5: In-text patent citations extraction error analysis

Error type	**Category**	**Example**
False negative	1	"introduced into a mold (as in Example 1 of 2,154,639) wherein it is polymerized to form a A"
	2	"Aug. 20, 1935 2,255,030 Tholstrup Sept. 2, 1941 2,394,733 Wittenrnyer Feb. 12, 1946 2,433,349 Drewell Dec. 30"
	3	"Filed May 25, 1973, Ser. No. 364,196 Int. Cl. Blk 1/00, 3/06; C01b"
False positive	1	"US. Cl ..29/492, 29/497, 29/498, 29/502, 29/589, 29/628 [51] lnt.Cl."
	2	"Aug. 12, 1941. ALKAN&#39; emoumnmrc COMPASS I iled July 15, 1936 3"
	3	"No. 09/808,790, (Attorney Docket No. 20468-000110), previously incorporated herein by reference. FIG"

Notes: The underlined span of text triggered the error. In the false negative case, it was not detected by Grobid as a patent citation while it should have been the case. In the false positive case, it was detected by Grobid as a patent citation while it is not.

Table 6: Distribution of U.S. patent citations by patent office

Patent office	**Number of occurrences in validation sample**	**Share in validation sample**	**Share in universe of U.S. patents**
US	203	0.67	0.61
JP	52	0.17	0.09
WO	18	0.06	0.10
DE	9	0.03	0.02
EP	5	0.02	0.03
KR	4	0.01	7.00E-3
FR	4	0.01	6.00E-3
BE	2	7.00E-03	3.00E-3
SA	1	3.00E-03	3.00E-3
CH	1	3.00E-03	3.00E-3
AL	1	3.00E-03	0.02

Table 7: In-text patent citations parsing accuracy

	Number of examples in the test set	**Organisation name**	**Original number**	**Kind code**	**All**
Lopez (2010	250	-	-	-	97.2%
Verluise et al (2020)	300	98.4%	95.7%	97.6%	-

Notes: Lopez (2010) does not distinguish between the accuracy on the three attributes and reports the overall accuracy of the Finite State Transducers to translate the natural language citation into a fully structured citation represented by its three attributes.

Table 8: In-text patent citations matching performance

	True		**False**	
	Content	*Number*	*Content*	*Number*
Positive	A publication number was correctly matched	137	A publication number was incorrectly matched	10
Negative	No matched publication number and no match found by the annotator	36	No matched publication number but a match was found by the annotator	17

Table 9: In-text patent citations matching error analysis

Error type	**Category**	**Sub-category**	**Example**	**Number of occurrences**
False match	Incorrect patent	Badly formatted pre-2000 Japanese patent	JP5064281 instead of JPS5064281	5
		Incorrect extraction of pre-1970 U.S. patent due to bad OCR	CA-8465T-T (from 2,936,846 5/60 Tyler et al, in reference list)	1
	Non patent	Garbled table	-	2
		Technology class	US-32537 extracted from "... U.S. Cl. 325/392, 325/37..."	1
		Date	US-312012 extracted from "...filed Aug. 31, 2012, . . . "	1
False no-match	Formatting	Missing leading zeros after country code or date	EP592106 instead of EP0592106	6
		Year reported after instead of before patent number	JP3518222000 instead of JP2000351822	3
		Incorrect extraction of country code	SU-14553625 extracted from "U.S. Utility application Ser. No. 14/553,625"	1
	Wrong service call	-	-	7

Notes: Error analysis based on 200 random examples.

Table 10: Extracted citations judged unmatchable by the annotator

Category	**Example**	**Number of occurrences**
Garbled tables	AL-1226-C extracted from "...AL C 257 75.108 67.122 6.016 1..."	11
Provisional patent applications	US-60723639 extracted from "U.S. provisional application Ser. No. 60/723,639"; provisional patent applications are not public information	8
Incorrect and ambiguous number formats	EP-87309853 extracted from "European patent specification No 87309853.7" (non-standard format of a non-searchable application number)	4
Incorrect parsed attributes	WO-PTS0767103 instead of WO-PTUS07067103	5
Non searchable	DE-19654649 (not indexed by Google Patents)	3
Non patents (technological class, dates, etc)	US-32128 extracted from "... U.S. Cl. 322/79, 310/68 D, 321/28, ..."	10

Notes: The number of occurrences includes both matched and unmatched examples.

Table 11: Number and share of citations matched by patent organisation (selected)

Patent organisation	**Total number of citations**	**Share of citations matched**
USPTO	37,072,526	89.14
WIPO	6,453,099	81.89
JPO	5,659,300	77.22
EPO	2,228,096	51.27
DPMA	1,371,114	73.46