Tracing science-technology-linkages: a machine learning pipeline for extracting and matching patent in-text references to scientific publications

Abbasiantaeb, Z.; Verberne, S.; Wang, J.

Information Processing and Management 62 (2025) 104264

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Tracing science-technology-linkages: A machine learning pipeline for extracting and matching patent in-text references to scientific publications

Zahra Abbasiantaeb [a,1], Suzan Verberne [b], Jian Wang [b,c,d,*]

[a] Information Retrieval Lab, University of Amsterdam, Science Park 900, 1098 XH, Amsterdam, The Netherlands
[b] Leiden Institute of Advanced Computer Science, Leiden University, Einsteinweg 55, 2333 CC, Amsterdam, The Netherlands
[c] Centre for Science and Technology Studies, Leiden University, Kolffpad 1, 2333 BN, Amsterdam, The Netherlands
[d] Management School, Lancaster University Leipzig, Nikolaistrase 10, 04109, Leipzig, Germany

ARTICLE INFO

Dataset link: https://zenodo.org/record/15756322, https://github.com/ZahraAbbasiantaeb/Patents.git

Keywords:
Text mining
Reference extraction
Science technology linkage
Citation analysis
Patent analysis

ABSTRACT

Patent references to science provide a valuable paper trail for investigating the knowledge flow from science to technological innovation. Research on patent–paper links has mostly concentrated on front-page references, often neglecting the more complex in-text references. Therefore, we developed a three-stage machine-learning pipeline to extract and match patent in-text references to scientific publications. Our pipeline performs the following tasks: (1) extracting reference strings from patent texts, (2) parsing fields from these reference strings, and (3) matching references to publications in the Web of Science (WoS) database. We developed a training dataset consisting of 3,900 (and 3,901) manually annotated references from 392 (and 319) randomly selected EPO (and USPTO) patents. The first stage, reference extraction, achieved almost perfect results with a precision of 98.9% and a recall of 97.7% at the reference level. Overall, the pipeline demonstrated robust performance, with a precision of 96.8% and a recall of 91.9% at the unique patent-paper-pair level. Applying this pipeline to EPO and USPTO patents granted between 1990 and 2022, we identified 5,438,836 (and 20,432,189) references from 492,469 (and 1,449,398) EPO (and USPTO) patents, 2,763,779 (and 11,069,995) of which are matched to WoS publications. This extensive dataset is a valuable resource for studying science-technology linkages. We offer open access to this dataset, along with the associated code and training data.

1. Introduction

Patent documents have long been recognized as valuable resources for understanding science and innovation, as well as informing policy-making and business intelligence (Liu et al., 2024; Yang et al., 2025; Zhang et al., 2025). While the metadata in patents provides essential insights, the full text offers an even richer source of information. This potential has garnered significant attention in academic research, particularly within the fields of information retrieval and natural language processing (Codina-Filbà et al., 2017; Fujii et al., 2007; Yun et al., 2022).

Since seminal works in the 1980s (Narin & Noma, 1985; Nunn & Oppenheim, 1980), patent references to scientific publications have served as a key resource for linking science and technology (Belderbos et al., 2024; Ke, 2020, 2023; Nagar et al., 2024; Poege

[*] Corresponding author at: Management School, Lancaster University Leipzig, Nikolaistrase 10, 04109, Leipzig, Germany.
E-mail address: jian.wang@lancaster.ac.uk (J. Wang).

[1] Work done while working at Leiden University.

https://doi.org/10.1016/j.ipm.2025.104264
Received 24 February 2025; Received in revised form 15 May 2025; Accepted 7 June 2025
Available online 2 July 2025

et al., 2019; Veugelers & Wang, 2019). However, most studies have focused primarily on front-page patent references, overlooking the richer and more complex in-text references embedded in the patent documents.

Patent front-page references are listed on the document's front page, serving as prior art to evaluate the novelty of a newly submitted patent. In contrast, patent in-text references appear within the body of the patent text and function similarly to references in scientific papers. These in-text references, however, are often unstructured, short, and noisy, complicating the extraction process. Despite their significant potential for improving our understanding of the link between scientific knowledge and technological innovation, in-text references remain underexplored in existing research (Abbasiantaeb et al., 2022; Bryan et al., 2020; Marx & Fuegi, 2022; Tamada et al., 2006; Verberne et al., 2019; Voskuil & Verberne, 2021). Furthermore, prior studies have shown that patent front-page references and in-text references encode different types of information, leading to divergent insights on the relationship between patent value and the characteristics of referenced science (Wang & Verberne, 2024).

The challenge in accurately extracting in-text references lies in their unstructured nature. For example, the USPTO patent on "CRISPR-Cas systems and methods for altering expression of gene products" cites four publications in its first paragraph through in-text references, each presented in a different format, and none of them appears on the front page:

- Goeddel, GENE EXPRESSION TECHNOLOGY: METHODS IN ENZYMOLOGY 185, Academic Press, San Diego, Calif. (1990).
- Boshart et al. Cell, 41:521-530 (1985)
- Mol. Cell. Biol. Vol. 8(1), p. 466–472, 1988
- Proc. Natl. Acad. Sci. USA., Vol. 78(3), p. 1527–31, 1981

These examples highlight the diverse and unstructured nature of in-text references, with each reference presented in a different format. This variability complicates the task of accurately extracting and matching these references to scientific publications—a challenge that this study addresses.

The objective of this work is twofold: (1) to develop a high-performing machine learning method for extracting patent in-text references and matching them with corresponding publications in the Web of Science (WoS) database, and (2) to apply this method to patents from the European Patent Office (EPO) and the United States Patent and Trademark Office (USPTO), creating a comprehensive dataset linking patents to scientific publications.

Our contributions are as follows: (1) we introduce and rigorously evaluate a novel machine learning-based pipeline for extracting and matching patent in-text references, addressing a critical gap in science-technology linkage studies; (2) we release two valuable resources: (a) a curated patent-paper-link dataset that facilitates the tracking of the translational process from scientific research to innovation, and (b) a high-quality, manually annotated dataset that can be leveraged for future text-mining tasks, particularly in the context of information retrieval and natural language processing. The patent-paper-link dataset and annotation data are available at: https://zenodo.org/record/15756322, and code are available at: https://github.com/ZahraAbbasiantaeb/Patents.git.

2. Related work

Only a handful of studies have addressed the task of in-text reference extraction and linking. Tamada et al. (2006) pioneered a systematic analysis of patent in-text references, employing regular expressions to extract them from patents at the Japan Patent Office (JPO). They successfully extracted 9,379 non-patent references from a sample of 1,500 patents across five technology fields, achieving a recall of 98.2% and precision of 98.1%.

Bryan et al. (2020) took a different approach, bypassing reference extraction to instead start with a set of scientific journals and then match publications in these journals with patent full texts. Their analysis covered 3,389,853 articles from 248 prominent journals, cited collectively in 342,667 U.S. Patent and Trademark Office (USPTO) patents.

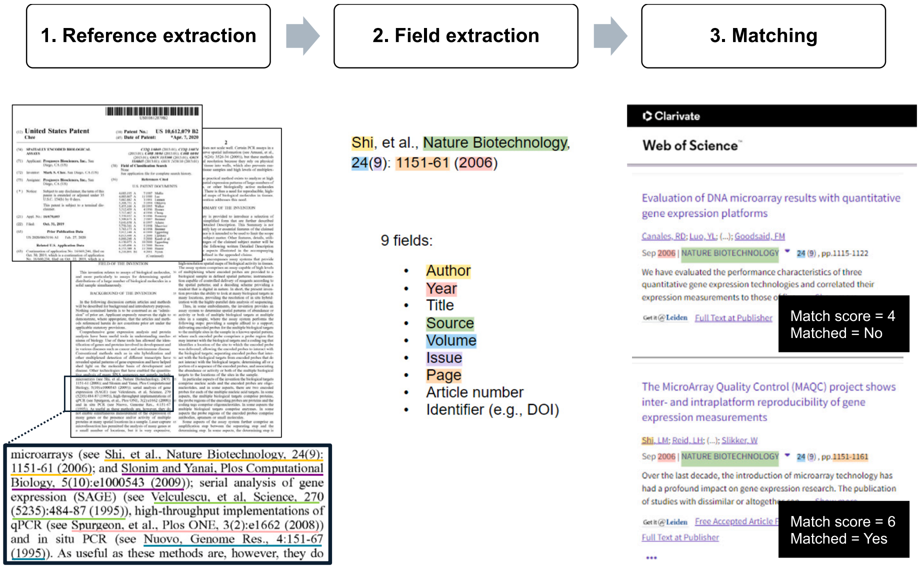
Marx and Fuegi (2022) combined rule-based methods with the GROBID model and processed the full corpus of USPTO and European Patent Office (EPO) patents, achieving precision ranging from 93.53% to 100% and recall from 82.05% to 57.70%.

Abbasiantaeb et al. (2022), Verberne et al. (2019), Voskuil and Verberne (2021) implemented machine learning techniques for sequence labeling, such as CRF, Flair, and BERT-based models that gave promising results despite the relatively small training dataset.

Building on these prior studies, this paper develops a more advanced machine-learning pipeline for extracting and matching patent in-text references to scientific publications and produces a high-quality and large-scale dataset tracing knowledge flows from science to innovation.

3. Methods

We developed a three-stage pipeline for extracting and matching the in-text scientific references from patents (Fig. 1). The first stage is *reference extraction*. In this stage, given the text of the patents, the in-text references are extracted using a sequence labeling model. The second stage is *field extraction*, where the fields of the references (author, journal, year, etc.) are extracted from the reference texts. The third stage is *matching*, where the extracted fields are matched with entries in the Web of Science (WOS) publication database to find the corresponding scientific publication. We used pre-trained sequence labeling models for the reference extraction and field extraction stages that we fine-tuned for our task. For this purpose, we collected a manually annotated benchmark dataset. We provide a detailed explanation of each of the stages in our pipeline in the following sections.



microarrays (see Shi, et al., Nature Biotechnology, 24(9): 1151-61 (2006); and Slonim and Yanai, Plos Computational Biology, 5(10):e1000543 (2009)); serial analysis of gene expression (SAGE) (see Velculescu, et al, Science, 270 (5235):484-87 (1995)), high-throughput implementations of qPCR (see Spurgeon, et al., Plos ONE, 3(2):e1662 (2008)) and in situ PCR (see Nuovo, Genome Res., 4:151-67 (1995)). As useful as these methods are, however, they do

Shi, et al., Nature Biotechnology, 24(9): 1151-61 (2006)

9 fields:

- Author
- Year
- Title
- Source
- Volume
- Issue
- Page
- Article number
- Identifier (e.g., DOI)

Evaluation of DNA microarray results with quantitative gene expression platforms

Canales, RD; Luo, YL; (...); Goodsaid, FM

Sep 2006 | NATURE BIOTECHNOLOGY 24 (9) , pp.1115-1122

We have evaluated the performance characteristics of three quantitative gene expression technologies and correlated their expression measurements to those o

Match score = 4
Matched = No

The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements

Shi, LM; Reid, LH; (...); Slikker, W

Sep 2006 | NATURE BIOTECHNOLOGY 24 (9) , pp.1151-1161

Over the last decade, the introduction of microarray technology has had a profound impact on gene expression research. The publication of studies with dissimilar or altogeth

Match score = 6
Matched = Yes

Fig. 1. Three-stage pipeline for extracting and matching patent in-text references.

3.1. Reference extraction

We approach the problem of reference extraction as a sequence labeling task. Given the text of the patent, the model labels each token with BIO labels, where 'B' means the beginning token of a reference, 'I' means a token inside a reference, and 'O' means a token outside of the Reference. (Ramshaw & Marcus, 1999). Note that when a patent cites multiple references in one group, for example "(Baneiji, et al. 1983. Cell 33: 729-740; Queen and Baltimore, 1983. Cell 33: 741-748)", the individual references are extracted and stored separately. Using human-labeled data (as explained in Section 4), we fine-tune and evaluate multiple pre-trained BERT-based language models for context modeling with a linear layer on top for classification. We experimented with BERT models that are trained on patent texts, i.e., PatentBert (Lee & Hsiang, 2019) and Bert for Patents (Srebrovic & Yonamine, 2020), expecting that they would be the most promising as we analyze patent texts. We also evaluated BERT-base (Devlin et al., 2018) and SciBERT which are trained on scientific texts (Beltagy et al., 2019).

To form the input sequences of the model, the text of the patent is segmented to make sure the length of the sequence does not exceed the maximum sequence length allowed by the model. To this aim, we first tokenize the patent's text using the tokenizer of the same language model, then segment the text into sequences of at most 512 tokens.

After labeling texts with the fine-tuned sequence labeling model, the references are identified based on the predicted BIO labels. Specifically, a reference string starts with a token with a 'B' label and includes subsequent tokens with 'I' labels after the initial 'B' label.

3.2. Field extraction

Similar to the first stage reference extraction model, the second stage field extraction model also uses a sequence labeling approach, based on three different BERT models: NER-BERT (Liu et al., 2021), SciBERT (Beltagy et al., 2019), and BERT-base (Devlin et al., 2018). Our method includes a pre-trained BERT-based model and a classification layer on top of each output.

The input of this model is the extracted reference string from the first stage. We do not give the context of the reference to the field extraction model and only rely on the reference text itself. We define 14 token-level labels that include: Year, Author-B, Author-I, Source-B, Source-I, Title-B, Title-I, Page-B, Page-I, Volume, Number, Issue, Identifier-B, Identifier-I. Title is the title of the publication, and Source is the name of the journal, conference, or title of the book when the reference is to a book chapter. The identifier can be the DOI, ISSN, or BSSN. Using these labels, we extract (at most) 9 fields from references and pass them to the next stage for matching.

3.3. Matching

In this stage, the fields extracted from the reference text in the second stage are used to identify the referenced scientific paper, using a rule-based approach. To cope with the large number of possible combinations between extracted strings and publications in the WoS database, we restrict the pool of candidates to WoS publications with the same publication year as indicated in the reference string. Note that requiring the exact matching of the year excludes the linking of publications with an erroneous year in the reference text. Making this matching more lenient would heavily increase the number of potential matches, especially in references of the form *Author, Year*. Also, we only observe 54 out of 6,956 unique references with missing year or wrong year that led to matching errors. Therefore, our matching script first takes the *Year* field from the array of fields and then compares all other fields of the references with all the publications in that year. We only process references that have a year field and the value of the year is not before 1980, because our WoS database only includes publications starting from 1980.

The field of *Author* may have multiple values, and in this case, we only use the first value, that is, we only match the first author. Because of variations in name use, we consider it a match when the author name of the record is a sub-string of the author name of the reference or vice versa (for example, ‘Flanagan’ and ‘Flanagan, ME’). The field of *Source* may also have multiple values, and we only use the first value. We consider it a match when the extracted source name is a substring of the full name, or standard abbreviations provided by the WoS database, or vice versa. For the *Page* field, we consider it a match if the first page matches and the end page either matches or is missing from one or both compared items. For the remaining numerical fields (issue, number), we only count a match when there is an exact match between the fields of the reference and the fields of that record. The issue and number in patent references can be safely assumed to be correct; we only observe 5 out 6,956 unique references with errors in page number leading to false positives (Table 7). The exact matching of numerical information also prevents the combinatorial explosion of including all possible errors made in the numeric strings.

We count how many fields are matched between the focal patent reference and each WoS publication with the same publication year, this information is stored as the match score s, with the maximum possible value of 9.

Our matching model adopts a multi-step approach: For each reference string, the algorithm starts with a more specific *rule* for identifying its correct WoS record (see the rules listed below). If this rule cannot identify a matched WoS record, then the algorithm moves on to the second, more general, rule, and then to the third, progressing from more deterministic to less deterministic criteria for matching.

We incorporate the following three rules for identifying matched WoS records:

1. The *Title* or *Identifier* (e.g., DOI) of the reference is exactly matched;
2. The maximum match score among all candidates is at least k (that is, at least $k-1$ other fields are matched in addition to *Year*), and there is only one candidate with this match score. We evaluate this with multiple values for k;
3. The maximum value of the match score is more than 3, and there is one or multiple candidates with the maximum match score. The first candidate for which the four fields, including *Year*, *Source*, *Volume*, and *Page* are all matched is selected.

Our first rule is the most deterministic, as it assumes that certain information, such as the title and DOI of a paper, is unique and sufficient for matching a reference. If a reference cannot be matched using these fields, we then consider records from WoS with the highest match score as candidates for matching. If there is only one record with a match score above k (Rule 2), we consider it a match. However, if multiple candidates share the highest match score (Rule 3), we examine specific fields to determine the correct match.

For example, the first stage reference extraction model extracted the following reference: “Sandborn, W . J ., et al.. New Engl . J . Med . 2012, 367, 616 - 624’. The second stage field extraction model extracted the following fields from this reference: ‘`Sandborn, W . J .`’ (Author), ‘`New Engl . J . Med .`’ (Source), ‘2012’ (Year), ‘616 − 624’ (Page), and ‘367’ (Volume). Then the third stage matching model matched this reference to the WoS publication with WoS Accession Number: `000307496600005` and DOI: `10.1056/NEJMoa1112168`. This match is identified through the second rule. More specifically, *Rule 1* could not identify a matched record in WoS from 2012 as no *Title* or *Identifier* were included in the reference string. Then *Rule 2* was activated, and this WoS publication has a match score of 5 (which is above the threshold of 4), while all other publications in WoS in the year 2012 have a match score lower than 5. More details about the performance of these rules will be reported in the performance evaluation section.

4. Dataset for training the pipeline

4.1. Sample

In this study, we focuse on English-language patents due to the open availability of the patent collections from the EPO and USPTO, as well as the availability of an English-language, domain-specific BERT model. We sampled a set of EPO and USPTO patents, separately, for manual annotation. The annotation is performed by paid local workers (master students). We gradually sampled 1000 random patents and released them for annotation, until we exhausted all our person-hours. Furthermore, only patents that have in-text references are useful for training the model, while the majority of patents do not have any in-text references. To prevent our sample from having many patents without references, for the sampled 1000 patents, we kept all patents that are predicted to have at least one in-text reference for annotation using the model developed by Verberne et al. (2019). For every four such patents, we added one randomly selected patent that is predicted not to have any in-text references. We discarded other patents that are predicted not to have any in-text references from the annotation work. Another point of consideration was to have a dataset with a balanced number of annotated EPO and USPTO patents. In the following, we provide further details regarding the sample.

EPO: We downloaded and processed “EP full-text data for text analytics” from the EPO website.[2] The version we download covers patents up to Week 30 of 2021, including the files up to ‘EP3800000.txt’. We kept patents that meet the following criteria: (1) are

[2] https://www.epo.org/searching-for-patents/data/bulk-data-sets.html

Table 1
Overview of the manually annotated dataset: number of annotated patents and references by patent office.

Type	Extracted references	Patents	Matched References (manually)
EPO	3,900	392	2,088 (53.5%)
USPTO	3,901	319	2,247 (57.6%)
All	7,801	711	4,335 (55.6%)

published between 1990–2022,[3] (2) are in English, (3) have title and description fields, and (4) are granted utility patents (i.e., type B1, B2, B3, and B9). When a patent has multiple published versions, we keep the most recent version. Then we randomly sampled 2000 EPO patents in two batches. As mentioned above, we first filtered out patents that are unlikely to have in-text references using the model of Verberne et al. (2019). According to the model's prediction, 1254 patents do not have any in-text references, while 746 patents have at least one. For annotation, we kept all these 746 patents and additionally sampled 187 patents from the rest of the patents, which are predicted not to have any references. We gradually released them for annotation. In the end, we annotated 725 patents (consisting of 580 predicted to have references and 145 predicted not to have any).

USPTO: We downloaded "Patent Grant Full Text Data (No Images) (JAN 1976 – PRESENT)" from the USPTO website.[4] The version we downloaded covers patents up to 12 March 2022, including the files up to 'ipg220412.xml'. We kept only B1 or B2 versions of utility patents published between 1990 and 2022. Then we randomly sampled 4000 patents in four batches. We used the same model for predicting whether the patent has any in-text references. According to the prediction of the model, 615 of them have at least one reference. For annotation, we kept all 615 patents and randomly sampled 154 from the rest of the patents that are predicted to have no references. We gradually released them for annotation. In the end, we annotated 650 USPTO patents, of which 520 are predicted to have references and 130 are not.

4.2. Annotation

In the pre-processing step, we added white spaces before and after each punctuation mark in the text of patents, because it helps the annotators to select the exact span more easily and to ensure that punctuation marks get their token label.

We had two rounds of annotation: (1) the first round for the first stage reference extraction model, and (2) the second for the second stage field extraction model and the third stage of matching.

In the first round, we hired 8 master students at Leiden University for 40 h each in one month, for annotating references in patent texts. We designed a guideline for annotators and trained them in one session. We had one pilot annotation step in which all annotators annotated the same 10 patents. We evaluated their performance in the pilot step and gave them feedback on their performance and additional instructions for the actual annotation. The annotators were given a total of 4 batches of data to annotate (one batch per week). We included some overlapping patents in each batch between different annotators to measure their inter-annotator agreement. For the overlapping patents, we included the union of the references annotated by the two annotators.

A total of 148 patents were independently annotated by two annotators. Across all BIO labels for the tokens in these patents, the annotators achieved a high inter-annotator agreement, with a Cohen's kappa (κ) score of 0.96. However, since the patents are lengthy and contain relatively few references, the majority of tokens are labeled as 'O' (non-reference), which increases the overall agreement score. In this case, all the patents had a total number of 1,158, 20,243, and 1,375,507 'B', 'I', and 'O' labels, respectively. To better capture the annotators' consistency, specifically on reference annotations, we focused on the union of all text spans marked as references by either annotator and recalculated the agreement restricted to these spans. In other words, we dropped the annotations where both annotators labeled a token as 'O'. In this case, the kappa agreement score (κ) was 0.54, indicating moderate agreement.

The annotated references were then checked by two senior annotators (i.e., the first and last authors of this paper) and were modified based on the agreement between them. In addition, some human errors were spotted during the second round of annotation (see the next paragraph) and corrected accordingly. Furthermore, in the evaluation step, we manually checked the prediction of the model for each fold of the dataset and modified the dataset if human annotation errors were spotted.

In the second round of annotation, for stages 2 and 3 of the pipeline, we again hired four master's students at Leiden University for 40 h each in one month. We asked them to (1) annotate the fields of the references, (2) find the corresponding scientific publication of the reference in the WoS database, and (3) determine the type of the reference (which can be journal, book, manual, and other). In addition, we improved the annotations of this part by manually comparing the final prediction of our pipeline with the annotations. For example, some of the "false positives" were correctly matched by our model while the annotators made an error. Also, some of the "false negatives" were due to the wrong annotation by annotators.

The statistics of the annotated dataset are shown in Table 1.

5. Performance evaluation

We first evaluate the performance of each stage model and then the whole pipeline, end-to-end.

[3] Recall that we only consider scientific publications from 1980 onward.

[4] https://bulkdata.uspto.gov/

Table 2
Performance comparison of different BERT-based models on the reference extraction task (step 1), evaluated separately for B- (beginning) and I- (inside) labels.

	B-labels		I-labels	
Model	Recall	Precision	Recall	Precision
SciBERT	95.3%	76.0%	98.5%	78.0%
BERT for patents (large)	95.2%	76.7%	98.3%	78.1%
PatentBERT	94.3%	75.2%	98.1%	77.5%
BERT	90.3%	72.5%	97.6%	76.0%

Table 3
Categorizing reference extraction outcomes at the reference level.

Outcome	N	Description
True positive	863	Exact extraction
True positive	38	Extracted with minimal differences, e.g., punctuation
True positive	28	Extracted with small differences
True positive	27	Extracted correctly but not in gold data
False negative	15	Extraction failed
False negative	8	Extracted with substantial difference
False positive	11	False extraction

5.1. Reference extraction

To choose a language model for the first stage task of reference extraction, we evaluated and compared different language models for predicting B- and I-labels, using five-fold cross-validation. While performance evaluation at the levels of B- and I-labels is standard practice for sequence labeling evaluation, it is more informative for users of the patent-paper-link dataset to look at the end-to-end performance at the reference level. To this end, we evaluated the performance of the reference extraction model using the best language model on a testing set at the reference level. We compared the extracted references and the human-labeled references in each patent. In the following, we provide a detailed explanation about experiments for choosing the best language model and reporting the performance of the reference extraction model at the reference level.

For extraction model selection, we fine-tuned and evaluated four pre-trained BERT-based language models, including BERT for patents, PatentBERT, SciBERT, and BERT-base for the reference extraction task. In this stage, a model with a higher recall is preferred. Because we aim to extract as many references as possible from the patent text, while the precision can be enhanced in subsequent stages. Specifically, any non-scientific or incorrect references would be discarded during the matching stage when the model fails to find a corresponding record in the WoS database. Results of evaluating the BERT models are shown in Table 2. We used five-fold cross-validation for evaluating the models. SciBERT achieved the highest recall. Therefore, we opted for SciBERT for the reference extraction task. We found the best parameter setting for the SciBERT model using five-fold cross-validation. We fine-tuned the sciBERT model using the best parameter setting on the entire labeled dataset for the main pipeline, which was then implemented on the full corpus of EPO and USPTO patent datasets.

For evaluating reference-level performance, we divided our labeled dataset into training- (80%) and testing-sets (20%). We fine-tuned our reference extraction model using SciBERT on the training set and evaluated the performance at reference-level on the testing set. We used the best parameter setting found using five-fold cross-validation for training the reference extraction model using SciBERT. We experimented with different values for batch size, number of training epochs, and learning rate. The best performance was achieved with a learning rate of 5×10^{-5}, 6 training epochs, a batch size of 8, and a maximum sequence length of 512. Table 3 shows different types of outcomes on our human-annotated testing set. In total, the reference extraction model extracted 863 reference strings that match exactly with human-annotated reference strings. 38 references were extracted with minimal differences (e.g., missing the punctuation marks at the end), and 28 with small differences (e.g., some author names were not included), which, however, would not affect further matching. Interestingly, the model extracted 27 references that were not annotated by human annotators, but our further investigation concluded that the model prediction was correct, while human annotators missed them. We classified all these four types of cases as true positives. Our model failed to extract 15 references and extracted 8 references with substantial differences (e.g., only include the author name) such that these references cannot be matched. We classified these two types of cases as false negatives. It is interesting to note that these cases coincide with difficult cases reported by our human annotators during the annotation process; our human annotators were not confident about whether they should be annotated or not. Our model extracted 11 references that are not actually references, which we labeled as false positives. Overall, the reference extraction model achieved a precision of 98.9% and a recall of 97.7%, which are very high (Table 4).

5.2. Field extraction

We fine-tuned and evaluated three BERT-based language models for the second stage task of field extraction: NER-BERT, SciBERT, and BERT-base. We evaluated the models using five-fold cross-validation. We selected the best model based on average precision and

Table 4
Performance of the reference extraction model evaluated at the reference level.

Precision	Recall	F1
98.9%	97.7%	98.3%

Table 5
Performance comparison of different language models on the field extraction task (step 2).

Model	Precision	Recall	F1
SciBERT	94.5%	95.8%	95.1%
NER-BERT	94.2%	95.6%	94.9%
BERT	94.1%	95.6%	94.8%

Table 6
Performance of different rules for the matching task (step 3). We report the number of true positives (TP), false positives (FP), precision, recall, and F1 in this table.

Rule 2	# TP	# FP	Precision	Recall	F1
k=2	3,548	166	95.5%	91.8%	93.6%
k=3	3,542	45	98.7%	91.6%	95.0%
k=4	3,398	12	99.6%	87.7%	93.3%
k=5	2,551	2	99.9%	65.9%	79.4%
k=6	406	1	99.8%	10.5%	19.0%
k=7	26	1	96.3%	0.7%	1.3%
k=8	0	0	NA	NA	NA
k=9	0	0	NA	NA	NA
Individual rules					
Rule 1	284	0	100.0%	7.3%	13.6%
Rule 2	3,542	45	98.7%	91.6%	95.0%
Rule 3	2,824	9	99.7%	73.0%	84.3%
Applying rules sequentially					
Rule 1+2	3,561	44	98.8%	92.2%	95.4%
Rule 1+2+3	3,573	50	98.6%	92.5%	95.5%

recall across the nine pre-defined labels. For extracting the fields, both precision and recall are important and involve a trade-off. While it is desirable to extract as many fields as possible, incorrectly extracted fields can lead to incorrect matching in the next stage.

Results of different models are reported in Table 5. The SciBERT model achieved the best performance for field extraction. We selected the best parameter setting for the SciBERT model based on five-fold cross-validation and used it for training the model for the main pipeline. We experimented with various values for batch size, learning rate, and number of training epochs. For the field extraction model, the best performance was obtained with a learning rate of 1×10^{-4}, a batch size of 16, a maximum sequence length of 200, and 10 training epochs.

5.3. Matching

As the matching model is not a supervised model, we evaluated its performance using the entire dataset. As we intend to study the performance of the matching model given a reference string, we evaluated the matching at the reference string level. We also removed duplicate reference strings and used the unique reference strings for evaluation. The annotated dataset has 6,956 unique reference strings, and 3863 of them can be matched to a scientific publication in WoS by human annotators.

We first evaluated the performance of three matching rules (see Methods section) individually. As *Rule 2* requires choosing a threshold, k, while *Rule 1* and *Rule 3* are more straightforward, we first present the results for *Rule 2* at different pre-specified values of k. Results are reported in Table 6 and Fig. 2. Results show that as we increase k from 2 to 3, we can substantially increase precision at a relatively small cost of recall, while further increases after 3 lead to substantial decreases in recall. This is also evident in the F1 score, which peaks when k equals 3. Therefore, we choose this value for matching.

Subsequently, we assessed the performance of these three rules, namely, we implemented each rule separately for matching reference strings to WoS records. Results are reported in Table 6. As expected, *Rule 1* achieved the highest precision, which is 1, but a very low recall. In other words, using the publication title or identifier together with the publication year can identify the correct WoS record without error (at least in our annotated dataset), but this rule can only retrieve a very small portion of matched references, as most references do not include the publication title or identifier. *Rule 2* achieved the highest recall, which confirms the expectation that considering the candidate with the highest match score is efficient. The precision of Rule 3 is 1 percentage

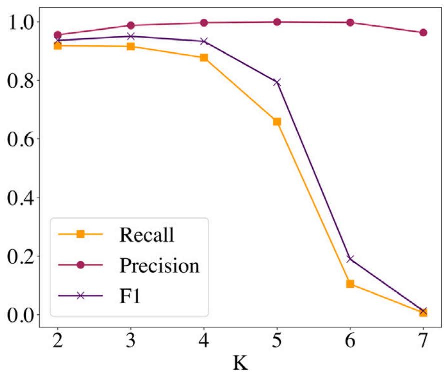


Fig. 2. Matching performance using only Rule 2, evaluated across different values of k.

Table 7
Identifying sources of matching errors.

Outcome	N	% of total	Description
True positive	3573	51.37%	
True negative	3043	43.75%	
False negative	169	2.43%	Multiple potential matches
False negative	59	0.85%	Field extraction error
False negative	48	0.69%	No year
False negative	13	0.19%	Variation in text
False negative	1	0.01%	Year wrong
False positive	34	0.49%	Not in WoS
False positive	11	0.16%	Error in reference
False positive	5	0.07%	Too little information

point higher than Rule 2, but its recall is 17 percentage points lower. The high precision of Rule 3 demonstrates its efficiency for the references that cannot be matched using Rule 2 because of having multiple candidates with the highest match score values.

The rationale of our matching model is to start with the most deterministic rules. We started with *Rule 1*, for reference strings that *Rule 1* cannot find a matched WoS record, we activated *Rule 2*, then for reference strings that *Rule 2* cannot find a match, we activated *Rule 3*. Therefore, Table 6 evaluated the performance of (a) *Rule 1* alone, (b): *Rule 1 + 2*, and (c): *Rule 1 + 2 + 3*. Results show that as we add *Rule 2*, recall increases by 0.848 while precision only decreases by 0.012, and F1 increases by 0.817. As we add *Rule 3*, recall increases by 0.003 while precision decreases by 0.002, and F1 increases by 0.001. Therefore, our final model adopted all three rules following the sequential approach. The recall, precision, and F1 of our final matching model are 0.986, 0.925, and 0.955, respectively.

To better understand the sources of matching errors, we manually examined all false positives and false negatives, with results summarized in Table 7. There are 6,956 unique references in total, of which 3,573 are true positives, cases where our model correctly identified a match. 3,043 are true negatives, where the reference is not in WoS and the model also returned no match. Among the false negatives: (1) 169 cases were ambiguous due to many WoS items having the same matching score; while human reviewers could identify the correct match, the model could not, (2) 59 cases failed due to errors in the second-stage field extraction, (3) 48 cases lacked year information, (4) 13 involved text variations such as in the title, and (5) 1 case had an incorrect year. Among the false positives: (1) 34 references were not in WoS, but the model incorrectly matched them based on partial field similarity, (2) 11 cases were due to reference errors, more specifically, 5 with the wrong year, 5 with the wrong page number, and 1 with the wrong author, and (3) 5 references contained too little information; although the model found a high-scoring match in WoS, human annotators were not confident it was correct.

5.4. End-to-end

For the end-to-end evaluation of the whole pipeline consisting of three models, we used 80% patents to train the models and the remaining 20% for testing. We used the best parameter setting found using five-fold cross-validation for training stages 1 and 2 models. The test set included 144 patents while our training set included 567 patents. The reference extraction model extracted a total of 980 references on the test set, while our gold dataset on the test set includes 942 references. The extracted references by the stage 1 reference extraction model were passed to the stage 2 field extraction model, and extracted fields from the stage 2 field extraction model were then passed to the stage 3 matching model.

We evaluated the end-to-end performance by comparing the set of unique patent-paper-pairs identified by our model and those identified by human annotators. The model in total extracted 401 unique pairs of patent-paper while the gold dataset contains 422

Table 8
End-to-end performance of the full pipeline evaluated at the unique patent–paper pair level over test set. We report the overall performance as well as the performance over patents with different technological fields. We report the precision, recall, F1, and number of gold patent–paper pairs (called Support) in the table.

IPC section	Precision	Recall	F1	Support
A	95.7%	89.8%	92.7%	49
B	100.0%	100.0%	100.0%	14
C	96.2%	91.9%	94.0%	246
G	98.8%	91.4%	95.0%	93
H	100.0%	92.3%	96.0%	13
Total	96.8%	91.9%	94.3%	422

Table 9
End-to-end performance of our model over patents from test set with different publication year. We report the precision, recall, and number of patent–paper pairs (called Support) from our gold set.

Year	1991	1992	1993	1994	1995	1996	1997	1999	2000
Precision	100.0%	95.0%	100.0%	100.0%	83.3%	95.5%	100.0%	95.8%	100.0%
Recall	100.0%	100.0%	100.0%	100.0%	100.0%	91.3%	100.0%	92.0%	100.0%
Support	3	19	1	3	5	23	8	25	6

Year	2001	2002	2003	2004	2006	2007	2009	2011	2012
Precision	100.0%	100.0%	97.8%	100.0%	93.8%	100.0%	96.8%	100.0%	100.0%
Recall	100.0%	100.0%	80.0%	100.0%	75.0%	100.0%	93.8%	100.0%	100.0%
Support	11	7	55	1	20	36	32	1	1

Year	2013	2014	2016	2017	2018	2019	2020
Precision	100.0%	100.0%	92.1%	100.0%	100.0%	100.0%	100.0%
Recall	75.0%	92.0%	93.5%	100.0%	95.3%	85.7%	100.0%
Support	4	25	62	1	64	7	1

unique pairs of patent-paper on the same set. Results are reported in Table 8, and the end-to-end precision is 96.8% and recall is 91.9%, which are very high. Using Marx and Fuegi (2022) as a comparison, which achieved precision ranging from 93.53% to 100% and recall from 57.70% to 82.05%, our pipeline obtained a remarkably higher recall without sacrificing much precision. As expected, the rule-based approach (which is the core of Marx & Fuegi, 2022) can achieve high precision at the expense of recall.

The randomly selected test patents cover five technological fields, namely IPC sections ‘A’, ‘B’, ‘C’, ‘G’, and ‘H’. We evaluate end-to-end performance separately for each section (Table 8). While there are some variations across fields, all perform consistently well, and no section shows any concerning deficiencies. We also assess performance across different publication years (Table 9). Although some variation exists, there is no clear trend indicating degradation or improvement over time.

6. Pipeline implementation and the resulting dataset

The whole pipeline was trained using the full dataset annotated manually, and then was implemented for extracting and matching in-text references from the corpus of EPO and USPTO patents.

EPO: The pipeline was executed on all utility patents of EPO that were published between 1990 and 2022. More specifically, we kept patents with types of B1, B2, B3, or B9, written in English, and then processed the most recent version with both description and title information. From 492,469 of these patents, our pipeline extracted at least one reference. The pipeline extracted a total number of 5,438,836 references. The average number of references for the patents that have at least one reference is 11. About 51% (2,763,779 references) of the extracted references are matched to scientific publications in the WOS database, while 2,675,057 of the extracted reference strings are not matched. The distribution of the matching score s for the extracted references is shown in Table 10. We do not have the record of the publications before the year 1980 in WOS database. The references with no year and the references that are published before 1980 include (1,123,513) 42% of the non-matched references and 21% of the total references.

USPTO: The pipeline was executed on all utility patents of USPTO that were published between 1990–2022. Our pipeline extracted 20,432,189 references from 1,449,398 unique patents. The average number of references for each USPTO patent (that has at least one reference) is 14. Among 20,432,189 extracted references, 11,069,995 of them (54%) can be matched and 9,362,194 of them (46%) cannot be matched with a publication in the WoS database. The distribution of the matching score for the extracted references is shown in Table 10. 41% of the non-matched references do not have publication year information or were published before 1980. These references account for 19% of the total references.

This final paper-patent-link dataset encompasses all EPO and USPTO utility patents published between 1990 and 2022. It includes 492,469 EPO patents with 438,836 in-text references, and 1,449,398 USPTO patents with 20,432,189 in-text references. The full dataset is publicly available at: https://zenodo.org/record/15756322.

We provide two primary data files: The first contains all patents and their extracted references, including metadata such as patent ID, full reference string, and extracted bibliographic fields (e.g., author, year, journal). The second file includes references that have been successfully matched to the Web of Science (WoS), along with the corresponding patent and paper IDs.

Table 10
Distribution of match scores for the extracted references from EPO and USPTO.

	EPO			USPTO		
s	Matched	Not Matched	Total	Matched	Not Matched	Total
No year	0	638,393	638,393	0	2,445,168	2,445,168
< 1980	0	485,120	485,120	0	1,351,659	1,351,659
1	0	226,341	226,341	0	728,867	728,867
2	10,698	1,114,559	1,125,257	26,429	4,007,685	4,034,114
3	124,969	174,209	299,178	473,496	686,416	1,159,912
4	580,262	35,247	615,509	2,072,405	137,154	2,209,559
5	1,649,672	1,188	1,650,860	6,503,339	5,245	6,508,584
6	348,116	0	348,116	1,766,368	0	1,766,368
7	49,765	0	49,765	225,776	0	227,776
8	297	0	297	2,182	0	2,182

‘*s*’ is the match score. ‘No year’ means that the reference does not have a publication year, and ‘< 1980’ means that the publication year of the reference was before 1980.

This dataset supports a wide range of research applications. For example: (1) From the perspective of patents, it enables analysis of the extent to which a patent (or inventor or firm) draws on scientific knowledge, and how such reliance shapes innovation outcomes. (2) From the perspective of scientific publications, it offers insights into which types of science are most influential for technological development, and how different scientific fields contribute to various innovation pathways. (3) By linking papers to patents, it allows for tracing knowledge flows between individual publications and inventions, scientists and inventors, and universities and firms, thereby deepening our understanding of the diffusion and adoption of scientific knowledge into innovation.

In addition, we publicly release our manually annotated training data through the same repository. This includes labeled patents for each step of the pipeline. It can serve as a valuable resource for researchers developing or benchmarking information extraction, reference parsing, or matching algorithms, particularly in the context of noisy, real-world patent text.

To ensure transparency and reproducibility, we also provide detailed annotation guidelines and the Python code used to implement the three-stage pipeline.

7. Discussion

We developed a three-stage automated pipeline to extract in-text references from the European Patent Office (EPO) and the United States Patent and Trademark Office (USPTO) patents and match them to publications in the Web of Science (WoS) database. The pipeline consists of three stages: (1) extracting reference strings from patent texts, (2) parsing these strings into relevant fields such as author names, journal titles, and publication years, and (3) matching the parsed references to corresponding records in WoS. For training and evaluation, we manually annotated 725 EPO patents and 650 USPTO patents with all relevant in-text references. From this annotated sample, we identified 3,900 references in 392 EPO patents, of which 2,088 were successfully manually matched to WoS publications, and 3,901 references in 319 USPTO patents, of which 2,247 were matched.

The performance of the reference extraction model was evaluated by reserving 20% of the annotated patents for testing. The model achieved a precision of 98.9% and a recall of 97.7% at the reference level. At the level of unique patent-paper-pairs, the pipeline demonstrated a precision of 96.8% and a recall of 91.9%. We then applied this pipeline to the full texts of EPO and USPTO patents granted between 1990 and 2022, yielding a total of 5,438,836 references from 492,469 EPO patents, of which 51% (2,763,779) were successfully matched to WoS publications. For USPTO patents, 20,432,189 references were identified across 1,449,398 patents, with 54% (11,069,995) matched to WoS.

While the pipeline demonstrates promising performance, there are several limitations that need to be addressed in future work. First, the current matching model relies on exact matches of publication year and specific text fields (e.g., author names, journal titles). This approach, though effective in ensuring precision, may exclude relevant references where there are data errors in these fields. Conducting a fuzzy match would lead to a combinatorial explosion of possible matches. Considering our high accuracy rate, we conclude that the performance costs of relying on exact matches are acceptable. Future research could explore fuzzy matching techniques and relax the year constraint to enhance the model’s recall. We do make the intermediary dataset of extracted reference strings public, so future research could directly build on such data.

Second, our model was trained on a random sample of patents from diverse fields and years. This broad sampling approach is best suited for our purpose of creating a dataset covering all fields and all years. However, research that focuses on a particular field or time period might benefit from a more targeted and homogeneous training set, however, it is also important to note that any field delineation will introduce errors and might systematically perform worse for patents at the periphery of the field or crossing fields.

Third, the current study is restricted to patents in English from EPO and USPTO, which are the most widely used for innovation studies. Expanding this research to include patents in other languages and from additional patent offices (e.g., China National Intellectual Property Administration (CNIPA), Korean Intellectual Property Office (KIPO), Japan Patent Office (JPO)) would enhance the generalizability and robustness of the pipeline. However, such an effort would also require language models that are trained on scientific and patent texts in these languages, as well as access to scientific publications in these languages.

Despite these limitations, our work makes several significant contributions to studies of science and innovation, especially the links between them. First, we introduce a high-performing machine learning-based methodology for extracting and matching patent in-text references, addressing a critical gap in science-technology-linkage studies. Notably, we apply sequence labeling — a technique typically used for extracting shorter named entities — to the more complex task of extracting reference strings, demonstrating the feasibility of adapting well-established sequence labeling methods to handle unstructured, domain-specific citation data in patent documents.

Second, our pipeline successfully tackles the challenge of extracting and matching unstructured in-text references to scientific publications, overcoming the variability and inconsistency of citation formats across patents. By automating this process, we provide a valuable tool for advancing the study of science-technology linkages.

Moreover, the dataset generated by our pipeline, which links patents to scientific publications via in-text references, offers a novel and high-quality resource for investigating the interaction between scientific research and innovation. This dataset enables the study of how scientific research translates into patented inventions. Our method contributes to the field by providing a useful way to trace the flow of scientific knowledge from publication to patent, offering fresh insights into the translational process.

Finally, the annotated training dataset we provide for reference extraction serves as a key resource for future advancements in reference mining and patent analysis, supporting the development of other supervised models and methods in analyzing patent texts.

CRediT authorship contribution statement

Zahra Abbasiantaeb: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Suzan Verberne:** Writing – review & editing, Methodology, Funding acquisition, Conceptualization. **Jian Wang:** Writing – review & editing, Writing – original draft, Funding acquisition, Data curation, Conceptualization.

Financial disclosure

This article has been funded by the Academic Research Programme of the European Patent Office.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT in order to improve the readability and language of the manuscript.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jian Wang reports financial support was provided by European Patent Office (EPO) Academic Research Programme. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Marika Bazelmans, Ellen Kageler, Vasiliki Kogia, Luuk van den Nouweland, Famke Nouwens, Emilio Sanchez Olivares, Federica Spiga, and Floris Tomassen for their diligent and meticulous annotation work.

Data availability

The patent-paper-link dataset and annotation data are available at: https://zenodo.org/record/15756322, and code are available at: https://github.com/ZahraAbbasiantaeb/Patents.git.

References

Abbasiantaeb, Z., Verberne, S., & Wang, J. (2022). Optimizing BERT-based reference mining from patents. In *PatentSemTech 2022* (p. 4).

Belderbos, R., Braito, N., & Wang, J. (2024). Heterogeneous university research and firm R&D location decisions: research orientation, academic quality, and investment type. *Journal of Technology Transfer*, *49*(7), 1959–1989. http://dx.doi.org/10.1007/s10961-024-10066-w.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676.

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

Codina-Filbà, J., Bouayad-Agha, N., Burga, A., Casamayor, G., Mille, S., Müller, A., Saggion, H., & Wanner, L. (2017). Using genre-specific features for patent summaries. *Information Processing & Management*, *53*(1), 151–174. http://dx.doi.org/10.1016/j.ipm.2016.07.002, URL: https://www.sciencedirect.com/science/article/pii/S0306457316302825.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Fujii, A., Iwayama, M., & Kando, N. (2007). Introduction to the special issue on patent processing. *Information Processing & Management*, *43*(5), 1149–1153. http://dx.doi.org/10.1016/j.ipm.2006.11.004.

Ke, Q. (2020). Technological impact of biomedical research: The role of basicness and novelty. *Research Policy*, *49*(7), Article 104071.

Ke, Q. (2023). Interdisciplinary research and technological impact: Evidence from biomedicine. *Scientometrics*, *128*(4), 2035–2077.

Lee, J.-S., & Hsiang, J. (2019). Patentbert: Patent classification with fine-tuning a pre-trained bert model. arXiv preprint arXiv:1906.02124.

Liu, Z., Jiang, F., Hu, Y., Shi, C., & Fung, P. (2021). NER-BERT: a pre-trained model for low-resource entity tagging. arXiv preprint arXiv:2112.00405.

Liu, W., Zhang, Y., Luo, X., Cao, Y., Gan, K., Ye, F., Tang, W., & Zhang, M. (2024). Patent transformation prediction: When a patent can be transformed. *Information Processing & Management*, *61*(6), Article 103872. http://dx.doi.org/10.1016/j.ipm.2024.103872.

Marx, M., & Fuegi, A. (2022). Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations. *Journal of Economics & Management Strategy*, *31*(2), 369–392.

Nagar, J. P., Breschi, S., & Fosfuri, A. (2024). ERC science and invention: Does ERC break free from the EU paradox? *Research Policy*, *53*(8), Article 105038.

Narin, F., & Noma, E. (1985). Is technology becoming science? *Scientometrics*, *7*(3), 369–381.

Nunn, H., & Oppenheim, C. (1980). *A patent journal citation network on prostaglandins*. Elsevier.

Poege, F., Harhoff, D., Gaessler, F., & Baruffaldi, S. (2019). Science quality and the value of inventions. *Science Advances*, *5*(12), eaay7323.

Ramshaw, L. A., & Marcus, M. P. (1999). Text chunking using transformation-based learning. In *Natural language processing using very large corpora* (pp. 157–176). Springer.

Srebrovic, R., & Yonamine, J. (2020). Leveraging the BERT algorithm for patents with TensorFlow and BigQuery. White Paper.

Tamada, S., Naito, Y., Kodama, F., Gemba, K., & Suzuki, J. (2006). Significant difference of dependence upon scientific knowledge among different technologies. *Scientometrics*, *68*, 289–302.

Verberne, S., Chios, I., & Wang, J. (2019). Extracting and matching patent in-text references to scientific publications. In *BIRNDL@ SIGIR* (pp. 56–69).

Veugelers, R., & Wang, J. (2019). Scientific novelty and technological impact. *Research Policy*, *48*(6), 1362–1372.

Voskuil, K., & Verberne, S. (2021). Improving reference mining in patents with BERT. In *Proceedings of the 11th international workshop on bibliometric-enhanced information retrieval* (pp. 78–88).

Wang, J., & Verberne, S. (2024). Comparing patent in-text and front-page references to science. *Journal of Informetrics*, *18*(4), Article 101564. http://dx.doi.org/10.1016/j.joi.2024.101564, URL: https://www.sciencedirect.com/science/article/pii/S1751157724000774.

Yang, X., Sun, B., & Liu, S. (2025). Study of technology communities and dominant technology lock-in in the internet of things domain - based on social network analysis of patent network. *Information Processing & Management*, *62*(1), Article 103959. http://dx.doi.org/10.1016/j.ipm.2024.103959, URL: https://www.sciencedirect.com/science/article/pii/S0306457324003182.

Yun, S., Cho, W., Kim, C., & Lee, S. (2022). Technological trend mining: identifying new technology opportunities using patent semantic analysis. *Information Processing & Management*, *59*(4), Article 102993. http://dx.doi.org/10.1016/j.ipm.2022.102993, URL: https://www.sciencedirect.com/science/article/pii/S030645732200108X.

Zhang, R., Yu, X., Zhang, B., Ren, Q., & Ji, Y. (2025). Discovering technology opportunities of latecomers based on RGNN and patent data: The example of huawei in self-driving vehicle industry. *Information Processing & Management*, *62*(1), Article 103908. http://dx.doi.org/10.1016/j.ipm.2024.103908, URL: https://www.sciencedirect.com/science/article/pii/S030645732400267X.