Artificial Intelligence for Patent Prior Art Searching

Rossitza Setchi[1*], Irena Spasić [2], Jeffrey Morgan[2], Christopher Harrison[3], Richard Corken[3]

[1]Research Centre for AI, Robotics and Human Machine Systems, School of Engineering, Cardiff University, UK
[2]Data Innovation Research Institute, School of Computer Science and Informatics, Cardiff University, UK
[3]Intellectual Property Office (IPO), UK

*Corresponding author:
Professor Rossitza Setchi,
School of Engineering, Cardiff University, The Parade, Cardiff CF24 3AA, UK
Email: Setchi@cf.ac.uk

Artificial Intelligence for Patent Prior Art Searching

1. Introduction

Patent prior art searching is a needle-in-a-haystack challenge with the aim to find the most relevant prior art documents amongst the 100m+ published patent applications that exist worldwide. Its purpose is to find prior art that may bear impact on the patentability of an application. In its simplistic form, the prior art search involves the following steps:

- examining the *claims* and identifying terms/possible keywords,
- distilling what the defining part of the invention is and forming a *search statement*,
- identifying the most relevant *classifications* based on keywords and examiner's background knowledge,
- *optional background search* to identify the most suitable terms and synonyms,
- forming *search queries* using keywords, classification codes and Boolean functions,
- finding the patents that are *likeliest* to be relevant to the application,
- *sifting through* the retrieved documents, using colour coded highlights, drawers and sticky notes, to identify the most relevant patents,
- further *narrowing down* the search results, often using figures and manual disambiguation of concepts, to identify close conceptual similarities,
- *optional search* for published research/online materials,
- forming a *conclusion* (judgement) about the novelty and inventiveness of the application.

The *definition* of the search statement is one of the most important steps in the process. It requires clear understanding of the critical subject matter and the potential novelty of the application. Examiners often *modify* the search statement iteratively as their understanding of the prior art or the potential patentability of the application develops. The search statement may include words, which do *not* necessarily appear in the original claims. The most *time-consuming* step is sifting through the large number of patents retrieved.

The searching strategy employed by the examiners is very systematic due to the structured way patent literature is organised. The s*earch techniques currently used* include keywords, classifiers, Boolean logic, proximity operators, truncation operators (e.g. right word truncation), linking to full-text documents and patent families, linking to external and internal depositories, keyword and synonym selection, combining saved search queries appropriately, iterative modification of previously stored search queries in light of newly acquired phrases and terminology, citation search and multilingualism. Techniques currently used in p*ost-search analysis* include colour coding/highlighting, drawers and sticky notes. The key

linguistic and semantic challenges are legal wording, long sentences, acronyms, and the technical nature of claims.

Given the level of technical competence required, patent prior art-searching is considered ideally suited to investigating whether artificial intelligence (AI) can assist with improving the time efficiency of this information retrieval task. The study, funded through the Regulators' Pioneer Fund (RPF) of the UK Department for Business, Energy and Industrial Strategy (BEIS), aims to understand the feasibility, technical complexities and effectiveness of how AI solutions could benefit the work of the UK Intellectual Property Office (IPO) during the filing and prosecution of the 20,000 patent applications it receives annually. The goal is to reduce the time and cost of prior art searches / due-diligence checks and subsequently improve the quality of the examination process. This goal is in line with the vision to provide automated search tools that complement the examiners' knowledge and expertise.

The main research challenge from an academic perspective is in studying patent prior art searching as a highly interactive and complex human-centred process, requiring multiple searches, diverse search strategies and search management. This key academic challenge has several dimensions. Firstly, from AI perspective, prior art-searching involves several complex sub-processes including automated feature extraction, query expansion, document classification, document clustering and topic modelling. These subprocesses have been studied individually but there is not enough clarity if existing AI techniques could support efficiently all these aspects. Secondly, the research literature lacks structured approaches, which allow experimental comparison between different AI techniques and measures their efficiency for prior art searches.

Therefore, the objectives of the paper are to develop an *experimental platform* and *protocol,* which allow comparison of state-of-the-art AI techniques in terms of their suitability for automated feature extraction, query expansion, document classification, document clustering and topic modelling in patent searches.

The paper is organised as follows. The next section discusses the state-of-the-art in prior art searching in terms of challenges and AI techniques. Section 3 introduces the two main contributions of the paper: the experimental platform and experimental protocol developed for the study. Section 4 presents results and discussion. Section 5 concludes the paper.

2. Literature Review

This section highlights the challenges associated with prior art searching from the point of view of the patent examiner and introduces advanced AI approaches to address them.

2.1 Patent prior art-searching as a human-centred process

Patents are complex legal documents, which are associated with huge differences in length, strictly formalised semantic and syntactic structure, extensive use of standard and non-standard acronyms and domain terminology [1]. Patentees typically use their own lexicon in describing their inventive details [2] or use abstract or generic terms to maximise the protective scope. The diachronic aspect of the patent text genre contributes to changes in terminology, where one term may refer to a technical

concept during a certain time period and thereafter may switch to represent another [1, 3]. Patents often include different data artefacts (e.g. drawings, mathematical formulas, bio-sequence listings or chemical structures) that require specific techniques for effective search and analysis.

In addition to the standard metadata (e.g. title, abstract, publication date, applicants, inventors), patent offices typically assign classification coding to assist in managing (allocating) their examination workload and in searching patents, but these classification codes are not consistently applied across different patent offices [4].

The success of a prior art search relies upon the selection of relevant search queries [5]. An important component of a successful search process is the transformation of a human query (search request) into a query representation [6]. This process is influenced by the examiner's background and experience in the technical field, their knowledge, communication and presentation skills, the reputation for trust and reliability that they have built up and their approach to teamwork [7]. Typically, terms for prior art queries are extracted from the claim fields of patent applications. However, selecting relevant search terms is a difficult task due to the complex technical structure of patents and the presence of mismatched and vague terms; this often involves further research into the domain of the application.

When searching for prior art, patent examiners are currently mainly relying on keyword searches and Boolean logic. However, the consensus in the research literature in the information retrieval and patent domains is that a keyword-based search for prior art, even if done with most professional care, often produces suboptimal results [8]. This is particularly important considering the different consequences of false positive and false negative results in the patent domain. While false positives cause additional work for the patent examiner, who has to exclude the irrelevant documents from the report, false negatives may lead to an erroneous grant of a patent, which can have significant legal and financial implications [9].

The ongoing debate among patent professionals about the relative value of full-text versus controlled indexing [7] reveals open questions about search quality and whether full-text search strategies generate too much irrelevant material (low precision searching) or are more prone to miss relevant answers due to unexpected variation in terminology in the source documents (low recall). The existing diachronic nature and lexical diversity within part of the patent text genre make it more difficult to sample out data in order to establish a training set for text mining applications [10].

Another open issue is the feasibility of a fully automated system for prior art searching. Fully automated prior art retrieval systems are challenged by the technical content of the patents and the subtleties in interpretation of patent laws, which are influenced by recent court decisions [11]. Several recent studies advocate the development of user-centred information retrieval systems, which assist expert examiners in identifying relevant literature and making decisions in prior art. Such systems offer improved interactivity and transparency, which are critical in gaining the trust of the users. For example, a system called Sigma, currently piloted in the United States Patent and Trademark Office unit 2427 [11], not only performs basic

keyword searches but also allows the experts to create search strategies that are best suited to examining a particular application. Another study, which explored the use of word embeddings [12] concluded that no model by itself was sophisticated enough to match an expert's choice of keyword expansion.

2.2 A guided review of relevant AI techniques

A prior art search involves checking whether a similar idea has already been described in a filed patent. It is a specific example of information retrieval, i.e. the problem of locating relevant information of unstructured nature within a large collection of documents [13]. A thorough prior art search involves creating a search query involving different combinations of relevant search terms. Most commonly, the users express their information need using Boolean queries. To maximise the recall (i.e. the number of relevant documents retrieved) while minimising the users' effort in manually crafting an all-encompassing query, most general-purpose search engines rely on query expansion techniques. For example, given the original query (e.g. *automobile*), it can be expanded by including synonyms (e.g. *car*), meronyms (e.g. *engine*), hyponyms (e.g. *minivan*) and hypernyms (e.g. *vehicle*). Most commonly, WordNet [14] or another manually maintained thesaurus is used to effectively bridge a gap between the lexical space (text) and the semantic space (meaning) [15]. While such an approach may have been used successfully in domains such as biomedicine whose community has made strategic investments into re-usable lexico-semantic resources [16], the creation and maintenance of such resources across a wide range of domains covered by patents is not feasible.

To cope with chronic incompleteness of finite lexico-semantic resources, the field of natural language processing (NLP) has resorted to alternative approaches to learn lexico-semantic representations (called word embeddings) directly from text. Word embedding algorithms such as word2vec [17] and GloVe [18] successfully utilise the key assumption of distributional semantics that words with similar distributions tend to have similar meanings. Representing words in the form of real-valued vectors stands to bring multiple benefits for prior art search. First, words can be easily compared in terms of their similarity and other semantic relationships, which can be used to facilitate query expansion without the need for expensive lexico-semantic resources. For example, when trained on a corpus of clinical narratives the words *tablet* and *aspirin* will have similar representations. However, when trained on a corpus of instruction manuals, the word *tablet* will be instead more similar to *iPad*. Second, moving on from the term-document matrix to low-dimensional word embeddings avoids the curse of dimensionality, which is known to reduce the performance of machine learning algorithms [19].

Machine learning can facilitate rapid development of NLP tools by leveraging large amounts of text data. Given the abundant number of patents, machine learning represents a natural choice for developing such tools in the context of prior art search. Despite the abundance of raw data, supervised machine learning algorithms may suffer from the data annotation bottleneck. Nonetheless, they have been successfully whenever the codes are readily available, e.g. when clinical codes

integrated with free-text notes into electronic health records were utilised as class labels [20]. Upon approval, patents are routinely coded using the International Patent Classification (IPC) scheme [21], which provides an ideal dataset to train supervised classification of pending applications. To support fine-grained classification within IPC categories, topic modelling can be used to discover abstract topics within a collection of documents in an unsupervised way [22]. Its major benefit in the context of prior art searching is its interpretability, which provides the examiners with a means of quickly assessing the relevance of documents associated with that topic to support bulk processing of patents. To prioritise the processing of the remaining patents, whose number may still be too high, a similarity measure can be used to focus on the most relevant patents for a given application.

The potential of advanced AI technologies in IP analytics and management has been recognised by the WIPO community [23], [24]. A recently published paper [25] has adopted a road-mapping approach to review the use of 11 priority technologies in IP knowledge management, technology management, economic value, and extraction and effective management of information. However, prior art searches are not specifically discussed and the focus of the review is on machine learning and deep learning (e.g. semantic technologies are not included in the review). The potential of semantic technology is highlighted in another study [26], which proposes a method for analyzing the dependencies of independent and dependent claims using a semantic dependency analysis to define a patent scope indicator.

The literature review has highlighted a wide range of AI techniques, which are potential candidates to be tested experimentally. The existing knowledge gap, however, is methodological. This paper addresses this aspect by proposing a structured approach, which includes an experimental platform and an experimental protocol to test the feasibility of the candidate AI techniques. This is the first paper which maps the prior art search processes to candidate AI techniques and explores their feasibility.

3. Methods

3.1 Technical Requirements

To investigate the feasibility of AI for prior art search, the technical requirements (TRs) were first established through a series of eight interviews with experienced patent examiners with specialisms across a range of different technology areas. Each interview lasted approximately 2 hours.

TR1: Query expansion.

TR2: Document classification into existing coding scheme.

TR3: Identification of semantically similar documents.

TR4: Ranking of relevant documents based on document similarity.

TR5: Visualisation of the distinguishing characteristics of retrieved documents.

As a result of the requirements analysis, the relevant AI techniques discussed in the literature review have been assorted into five key areas (see Table 1). The choice of specific techniques within each area was based on the current state of the art.

Table 1. AI Techniques Considered

Techniques	TR1: Query expansion	TR2: Document classification	TR3: Document similarity	TR4: Ranking	TR5: Visualisation
Natural language processing: text segmentation, normalisation, lemmatisation, stemming, co-occurrences	x	x	x	x	x
Unsupervised learning: word embeddings, topic modelling	x	x	x		x
Supervised learning: support vector machine, naive Bayesian learning, decision tree induction, random forest, neural networks		x			
Similarity measures: Jaccard similarity, Euclidean distance, cosine similarity			x	x	
Semantic technologies: thesauri, ontologies, distributional semantics	x		x	x	x

The techniques listed in Table 1 were evaluated using 162,154 published patent applications from three domains: Civil Engineering, Computing and Transporting. These patents related to GB published patent applications in these three domains between 1979 and 2018. The three domains were chosen because they are the top three technology fields[1] at IPO over the past 10 years based on the number of GB patent applications received. Each domain was formally defined as the union of relevant inventions areas identified by their codes in the IPC scheme [21]. Table 2 lists the codes of the patents used in the experiments.

Table 2. Validation domains

Domain	IPC Codes
Civil Engineering	E01, E02, E03, E04, E05, E06, E21, E99
Computing	G06, G10L, G11C
Transporting	B60, B61, B62, B63, B64

[1] Of the 35 WIPO technology fields – see IPC concordance table at https://www.wipo.int/ipstats/en/

3.2 Experimental platform

The experimental platform was developed as a methodological tool for systematic experimentation with different algorithms. Figure 1 shows a conceptual diagram of the main processes involved in prior art search and the filtering of patent information. The platform allows testing of the following steps:

1. The examiner reads an application and *defines a search statement and a search query*.
2. The system *classifies* the application into one or more classes.
3. The system *extracts the most relevant keywords* from the application.
4. The system *suggests expanding the query* with other related words.
5. The examiner curates the search query.
6. The system launches a *search to retrieve documents* from the relevant classes.
7. The system *assorts the retrieved documents into topics*, each described by a set of keywords.
8. The examiner *selects the topic(s)* deemed most relevant to the application.
9. Documents from the relevant topic(s) are *ranked based on their similarity* to the application.
10. The content of each document is *colour-coded* to highlight its relevance to the application.

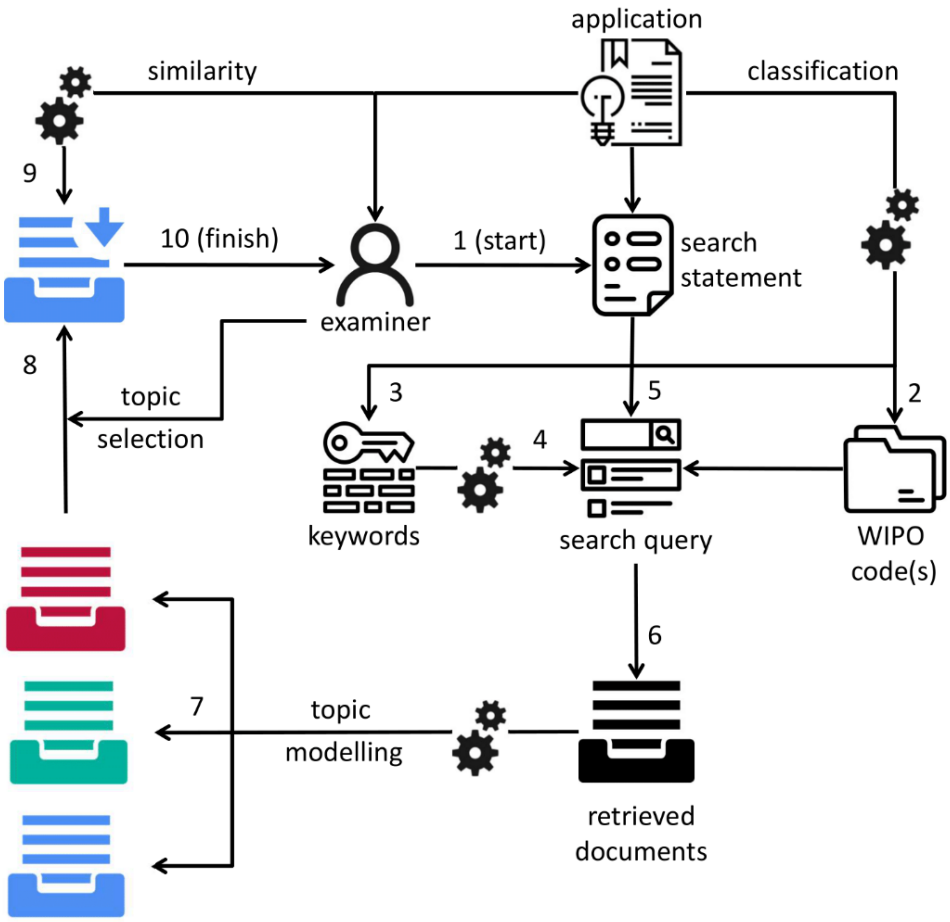


Figure 1. Experimental platform of prior art search and filtering of patent information

3.3 Experimental Protocol

Table 3 outlines the experimental protocol used to test different system functionalities. The associated technical requirements (TRs) are indicated in the first column. The specific AI technologies evaluated are listed in the right-most column. Their choice was informed by the current state of the art. For example, Elasticsearch is an open-source, distributed, scalable, real-time search engine [27]. It is currently the most popular search engine due to its performance especially when working with a large amount of data as is the case with patents. Similarly, WordNet [14], word2vec [17] and latent Dirichlet allocation [22] are the most prominent representatives of lexical databases, word embeddings approaches and topic modelling methods, respectively. However, a supervised learning method could not have been selected based on their prior performance. According to the "no free lunch" theorem, any two learning algorithms are equivalent when their performance is averaged across all possible problems [28]. In other words, there is no universally best learning algorithm, which suggests that the choice of an appropriate algorithm should be based on its performance for the particular problem at hand and the properties of data that characterise the problem. To that end, we systematically evaluated the performance of a wide range of supervised learning algorithms including support vector machines (SVMs), naive Bayesian learning, decision tree

induction, random forest and neural networks (specifically multilayer perceptron). A separate binary classifier was trained for each domain. Not surprisingly, the best performance was achieved by neural networks. However, they were not selected due to their tendency to overfit the data especially when the training dataset is small, which we concluded to be a strong possibility for some domains upon examining the distribution of patents across the IPC categories. Therefore, we selected the SVMs as the second best performing method. The results reported in this section are based on previously not seen test data.

The algorithms were trained on data provided by IPO with publication dates on or before 31 December 2018. Data provided includes the PATSTAT bibliographic database of worldwide patents (Autumn 2018 edition) [29], GB full-text patents (IPO, 1979-2018), EP full-text patents (1978-2018) [30] and US full-text patents (1976-2018) [31]. For data security reasons, IPO was unable to supply the accompanying search statements for either training or testing as their content is protected.

Table 3. Experimental protocol

TR	Step	Action	Algorithm
TR2	1	The system classifies the application into one of three domains: Civil Engineering, Computing, Transporting	Linear support vector machine (SVM) classifier with stochastic gradient descent (SGD) training
TR5	2	The system maps the application to the most relevant topics within the domain, each described by a set of keywords.	Latent Dirichlet allocation (LDA) [22]
TR1	3	The system extracts the most relevant keywords from the application.	Term frequency-inverse document frequency (TF-IDF)
TR1	4	The system suggests expanding the query with other related words, which were identified using: 1. general purpose lexical ontology 2. domain-specific word embeddings 3. topic modelling	1. WordNet [14] 2. word2vec [17] 3. LDA [22]
TR3	5	The user curates the search query.	N/A
TR3, TR4	6	The system launches a search to retrieve and rank at most 30 patents from the relevant domain and topics within.	Elasticsearch [27]

TR3, TR4	7	The retrieved patents are mixed with a set of 30 patents selected randomly from the same domain and then shuffled.	N/A
TR4	8	The system cross-references the query against each patent to colour-code its content.	N/A
TR3, TR4	9	The user assesses the relevance of each patent on a 3-point Likert scale.	N/A
TR5	10	Two users are presented with a set of keywords and asked to name independently the topic/abstract concept these keywords collectively represent. The users are asked to indicate their confidence lever on a 5-point Likert scale. In a second session the two users compare their results and agree a score using a 6-point similarity Likert scale.	N/A

The indicators used to measure the performance of the algorithms are summarised in Table 4.

Table 4. Indicators and measures

Aspect	Indicator	Measure
Classification	*Accuracy*	*F-measure* is the weighted harmonic mean of precision and recall. F-measure is a balanced way of measuring accuracy in terms of both precision and recall. *Precision* is often seen as a measure of exactness or quality, whereas *recall* is a measure of completeness or quantity.
Topic modelling	*Interpretability*	*Agreement* between two expert end-users.
Information retrieval and ranking	*Accuracy*	*Precision@k*, where k is a cut-off point in the ranked list of retrieved documents (e.g. k = 10, 20, etc.)
Usability	*User experience*	*Focus group* discussion

The testing was performed on a set of patents published since 1 January 2019 in each of the three technology sectors. For each sector, a set of another ten test patents were used to query this dataset. Each test patent was used to retrieve up to 30 patents deemed to be the most relevant ones by the system so that its performance can be evaluated. To blind the experiments, the retrieved patents were

combined with another 30 random patents selected automatically and then shuffled. Two patent examiners from each of the three test sectors assisted with the evaluation process. For each of the 10 test patents from their sector, each examiner was presented with an EpoqueNet working list [32] pre-populated with up to 60 patents described above and assessed their relevance on a 3-point Likert scale (Yes, Maybe, No). Although the examiners were not shown the rank at this point, this information was preserved nonetheless in order to calculate precision at point *k*. A supervisor from the IPO project board was present in the room to provide a quick overview of the testing process and to answer any questions. Each examiner was expected to complete the evaluation process within one day. The interface designed for the experiment is shown in Figure 2.

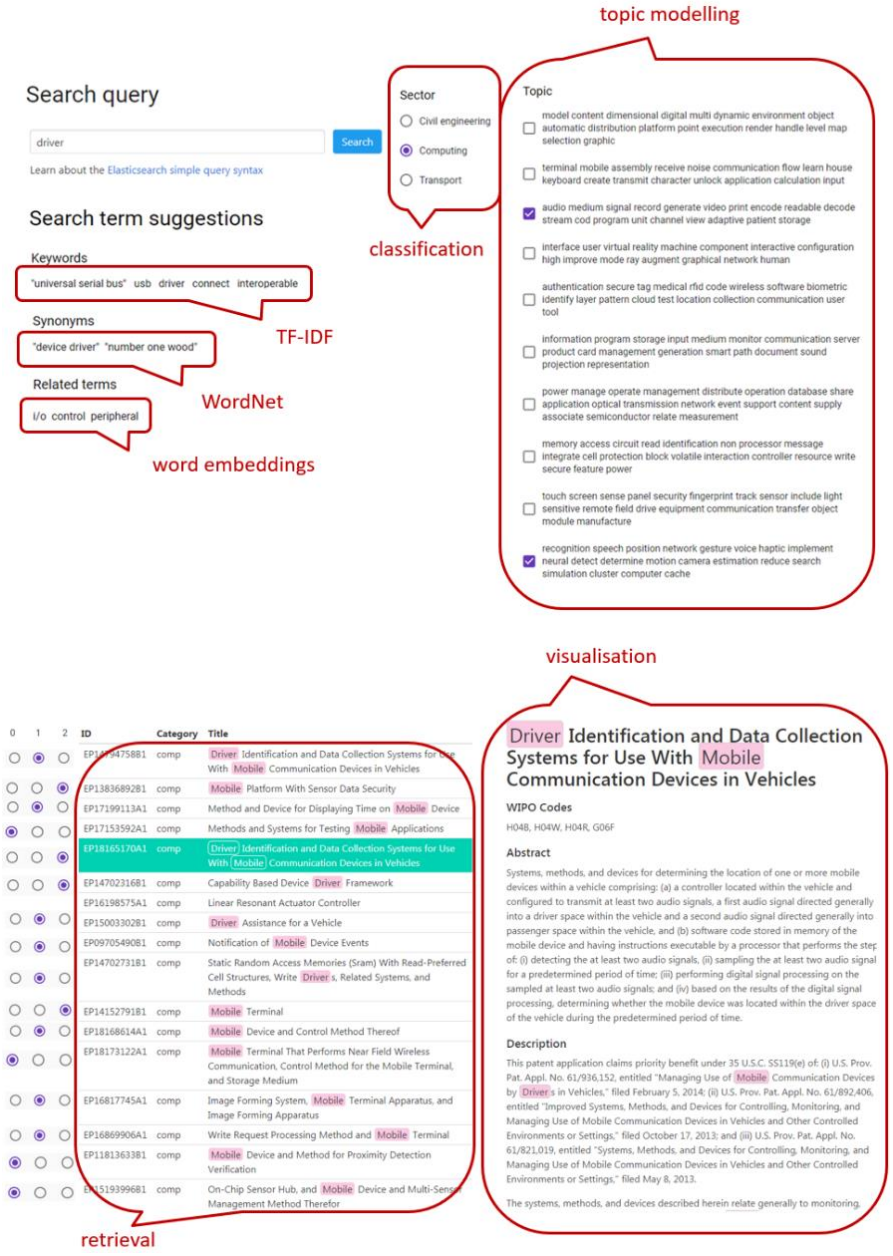


Fig. 2. Evaluation interface (search query and search results)

To evaluate the performance of information retrieval, the search results are presented back to the examiner who then annotate their relevance on a 3-point

Likert scale (Yes, Maybe, No) in line with the concept of the first and second drawer described before. The annotations are then used to calculate precision, which corresponds to the percentage of relevant documents among those retrieved by the system. The examiners are not shown the rank at this point, but this information is preserved nonetheless in order to calculate precision at point k.
The evaluation results are discussed in the next section.

4. Results and Discussion

4.1 Classification

To classify patents into one of the three domains, a single Support Vector Machine (SVM) classifier was trained with Stochastic Gradient Descent (SGD) using the Scikit-learn machine learning Python library [33]. Each patent was represented by text created by concatenating the title and the first 2,000 words of the description and its domain name. The patent text was the input variable to the classifier and the domain code of the patent was the output variable. The training data was divided into a training set and a testing set using the default setting of 75% training data and 25% test data. Table 5 shows the confusion matrix generated by the software, which presents the number of correct and incorrect predictions for the classifier. After training, the classifier was able to correctly predict the category of every patent in the test set. Table 6 shows the classification report for the SVM classifier that includes the precision and recall scores for a classifier. The 100% precision and recall scores in Table 6 confirm that the SVM classifier was able to correctly predict the category of every patent in the test set.

Table 5. Confusion matrix

		Predicted number of patents		
		Civil Engineering	Computing	Transporting
Actual	Civil Engineering	8,115	0	0
	Computing	0	12,422	0
	Transporting	0	0	12,560

Table 6. Classification performance

	Precision	Recall	F-measure	Support
Civil Engineering	100%	100%	100%	8,115
Computing	100%	100%	100%	12,422

Transporting	100%	100%	100%	12,560
Micro-average	100%	100%	100%	33,097
Macro-average	100%	100%	100%	33,097

4.2 Information Retrieval

Table 7 illustrates the degree of variation in the way search queries were formulated, with some taking full advantage of the search syntax (e.g. Examiner B in Computing) and others using the search syntax sub-optimally (e.g. Examiner A in Civil Engineering, not shown in the table, used AND to link synonyms such as *water*, *fluid* and *liquid* instead of OR).

Table 7. Search queries used in one of the domains (note the use of Boolean logic where | indicates an OR operator, multiple keywords suggest an AND operator)

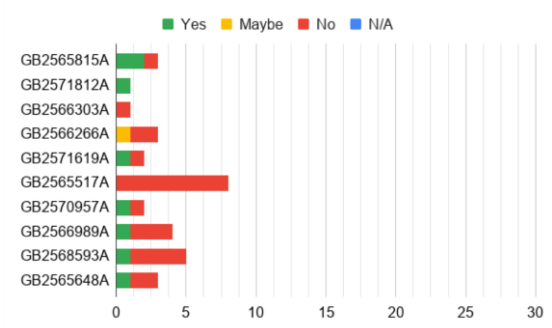
Domain	**Patent**	**Examiner C**	**Examiner D**
Computing	GB2568786 A	view plant gui configure theme	gui* | "user interface*" theme* | color* | colour* | dimension* | size* | font* | display* chang* | adjust* | modif* | adapt* | differ* measur* | control* | sens* | detect* | param*
	GB2571818 A	encoding neural network select interpolation	encode encoding encoded "neural network" "machine learning" choose (choice | (pick | selection)) option
	GB2570785 A	floorplan robot image	(robot* | automat* | autonom*) + (floor* | plan* | map*)
	GB2569804 A	authentication device service two second factor credential registered	authenticat* + (lan | "local area network") + (multiple | second* | devices | plural*) + (register* | subscrib* | join* | registrat*)
	GB2569223 A	feed paper printer display	(print* | paper* | sheet*) + (manag* | config* | control*)
	GB2570536 A	wearable ecg authentication temperature	(biometric* | heart* | ecg | pulse*) + (authenticat* | authori* | secur*) + (wearabl* | watch* | cloth*)

	GB2568779A	compare specie database	("imag* | object* | scene* | species | visual* recogn*") + (confidence* | threshold*) compar* | match* propert* | dimension* | attribute* | shape* | size* | characteristic* | parameter*
	GB2569426A	segmentation roi neural second	"medical imag*" | "medical diagnos*" | cade | cadx roi | loi | "region of interest" | locat* | posit* | area* | region*
	GB2570970A	sharp blur exposure virtual select region	("long-exposure" | "long exposure") virtual photography image (aggregate | combine | flatten | composite)

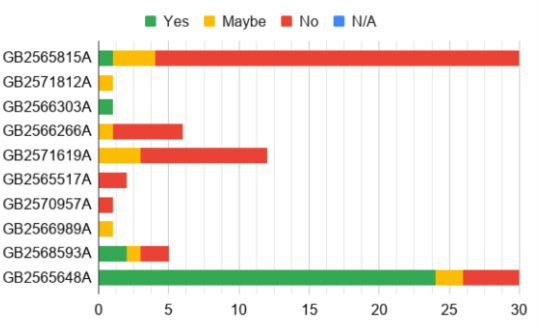
Fig. 3 provides the distribution of relevance annotations for each test patent and each examiner separately. As the examiners formulated their search queries independently, the search results differed significantly, hence the variation in the number of retrieved documents and their relevance. The human factor will always be a significant factor in prior art searching and there will always be natural variation because no two examiners will search in exactly the same way. Patent prior art searching at present is a fully manual process that relies on the experience and expertise of patent examiners to undertake the most effective and efficient search. There is no right or wrong way to search but some approaches may yield better and quicker results than others. There is also no right or wrong answer when it comes to the output of a prior art search because the nature of the patent prosecution process means that if a citation is sufficient to destroy the novelty or inventiveness of the application in question then the patent must be amended and re-examined. Any citation is ‘good’ citation if it means the applicant is forced to amend their application in light of its disclosure, even if it not the ‘most relevant’ citation; this is why it is impossible to measure recall in patent prior art searching because a patent examiner can in theory stop searching as soon as they find any novelty-destroying citation, although IPO examiners do strive to find the best citations before curtailing their searches. The examiner will then wait until the scope of the patent application is narrowed before undertaking additional searching.

Civil Engineering

Examiner A

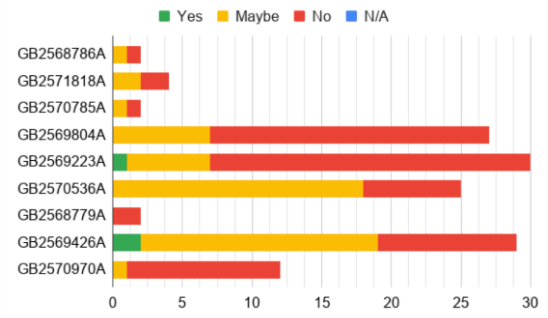


Examiner B

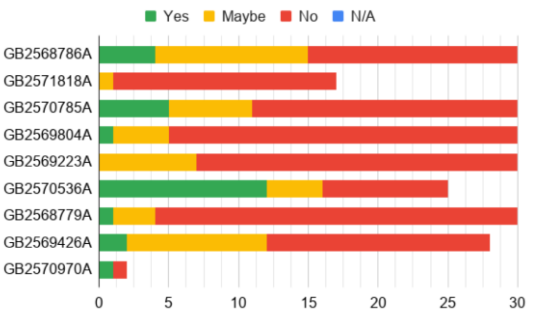


Computing

Examiner C

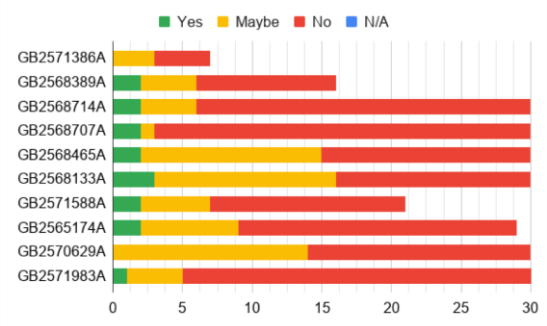


Examiner D



Transporting

Examiner E



Examiner F

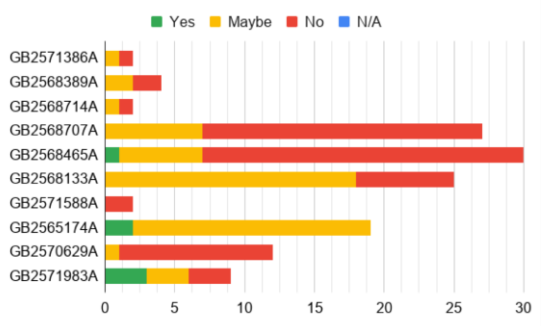


Figure 3. Distribution of the annotated results

The corresponding search results between the two examiners in each domain were compared to investigate the impact of using different search strategies. Table 8 provides the total number of patents retrieved by the examiner A but not examiner B (see column A – B) and vice versa (see column B – A). The overlapping set of patents (see column A □ B), i.e. those retrieved (and annotated) by both examiners, was used to calculate inter-annotator agreement using Cohen's kappa coefficient [34]. Strict agreement was applied using the original annotations (Yes, Maybe and No). For lenient agreement, the three labels were conflated into two, Relevant (Yes or Maybe) versus Irrelevant (No). The agreement observed in Civil Engineering and Computing was fair but unexpectedly low in Transporting, which indicates the need for more research. Ideally, in any future experiments, a third independent examiner should resolve any disagreements in order to establish ground truth.

Table 8. Differences in the search results and their interpretation

Domain	A – B	B – A	A □ B	Strict agreement	Lenient agreement
Civil Engineering	119	206	53	0.4135	0.6710
Computing	183	226	78	0.3221	0.5636
Transporting	31	80	34	0.1990	0.2446

Finally, precision was calculated using the two labels Relevant and Irrelevant. As shown in Table 9, precision varied between 34% and 50% across the six examiners, with overall average being 38%. Taking the ranking into account, these results were stratified across top 10, 20 and 30 documents (see Figure 4). Upon closer inspection, one can observe that precision at k = 10 varied between 30% and 50%, which indicates that the first page of search results contained between 3 and 5 relevant documents.

Table 9. Precision of information retrieval

	Civil Engineering		Computing		Transporting	
Patent	**Examiner A**	**Examiner B**	**Examiner C**	**Examiner D**	**Examiner E**	**Examiner F**
1	67%	13%	50%	50%	33%	43%
2	100%	100%	50%	6%	50%	38%
3	0%	100%	50%	37%	3%	20%

4	33%	17%	26%	17%	18%	10%
5	50%	25%	23%	23%	30%	50%
6	0%	0%	72%	64%	97%	53%
7	50%	0%	0%	13%	23%	33%
8	25%	100%	66%	43%	37%	31%
9	20%	60%	8%	50%	53%	47%
10	33%	87%	incomplete	incomplete	20%	16%
Average	**37.8%**	**50.2%**	**38.33%**	**33.66%**	**36.4%**	**34.1%**

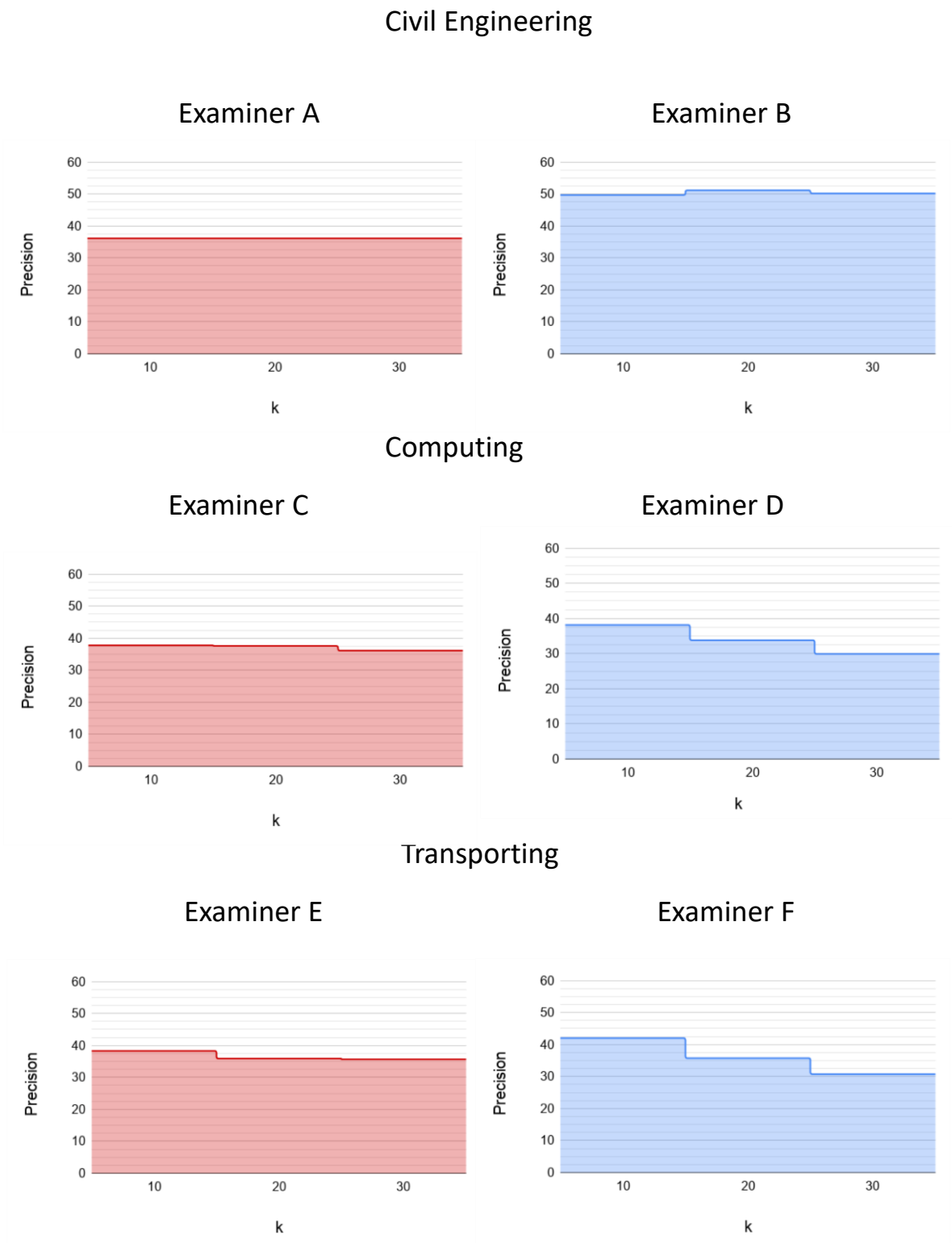


Figure 4. Precision at the top k retrieved documents (k = 10, 20, 30)

4.3 Topic Modelling

The topic modelling aspect of the study was evaluated using 10 topics; each topic represented by 15 keywords. Two examiners from each domain were asked to interpret independently a set of keywords and name a topic using a phrase that generalises the collective meaning of the keywords. No restrictions were imposed onto the choice of vocabulary or phrase format used by the examiners. The examiners were also asked to estimate the confidence in their final choice on a 5-point Likert scale. In the second phase of this experiment, both examiners gained access to the other examiner's choice of topic's name. They were then asked to independently estimate the similarity of the two names on a 6-point Likert scale. The average similarity was used to estimate the interpretability of the topics under the hypothesis that high similarity implies high interpretability and vice versa.

Table 10 shows three of the ten topics from Civil Engineering included in the topic modelling experiments. The experimental data show average confidence of 2.95, 3.00 and 3.10 in Civil Engineering, Computing and Transporting, respectively. Therefore, the confidence was consistently found to be moderate on the 5-point Likert scale. The average similarity was found to be 1.40 (slightly similar), 2.35 (moderately similar) and 2.65 (very similar) in Civil Engineering, Computing and Transporting, respectively on the 6-point Likert scale.

Table 10. The results of topic interpretability experiments in Civil Engineering

Topic ID	Keywords	Topic suggested by examiners 1 and 2	Confidence	Similarity
1	fluid drilling wellbore tool string valve downhole flow gas tubular oil injection sealing bore annular	well boring	very confident	very similar
		oil drilling, particularly bore linings and maintenance	moderately confident	very similar
2	sensor detection data power light unit signal electric information transmitted vehicle display electronic communication receiving	real time traffic signs	somewhat confident	slightly similar
		automated vehicles infrastructure	somewhat confident	slightly similar
3	tower platform post barry ladder vehicle	elevator	slightly confident	very dissimilar

	anchor concrete rail road frame track ground member cable	construction of transport infrastructure	slightly confident	very dissimilar

The inter-annotator agreement for both confidence and similarity was calculated using weighted Cohen's Kappa coefficient [21] to check whether the examiners were consistently finding some topics more difficult to interpret than others. The results are given in Tables 11- 14.

Although the confidence was found to be moderately high overall, it varied significantly across the topics in Computing. On the other hand, the judgement of similarity was found to be very consistent across all domains, albeit it was found to be low in Civil Engineering. The high similarity and high agreement obtained for Transporting illustrate the potential of using topic modelling to support prior art search. The preliminary results were obtained using a fixed number of topics and their keywords. Further experiments are needed to optimise the parameters of topic modelling for individual domains, as these can vary considerably in terms of their breadth and depth.

Table 11. Cohen's Kappa coefficient with linear weighting on confidence

Domain	Observed Kappa	Standard error	Confidence interval	Maximum possible	Proportion of maximum possible
Civil Engineering	0.5283	0.2025	0.1314-0.9252	0.9057	0.5833
Computing	0.1111	0.2267	0.0000-0.5554	0.7778	0.1428
Transporting	0.1667	0.1318	0.0000-0.4250	0.3750	0.4445

Table 12. Cohen's Kappa coefficient with quadratic weighting on confidence

Domain	Observed Kappa	Standard error	Confidence interval	Maximum possible	Proportion of maximum possible
Civil Engineering	0.7368	0.1558	0.4315-1.0000	0.9474	0.7777

Computing	0.0141	N/A	N/A	0.7183	0.0196
Transporting	0.3182	N/A	N/A	0.3182	1.0000

Table 13. Cohen's Kappa coefficient with linear weighting on similarity

Domain	Observed Kappa	Standard error	Confidence interval	Maximum possible	Proportion of maximum possible
Civil Engineering	0.6970	0.1729	0.3581-1.0000	0.6970	1.0000
Computing	0.5352	0.2542	0.0371-1.0000	0.5352	1.0000
Transporting	0.8024	0.1706	0.4680-1.0000	0.8024	1.0000

Table 14. Cohen's Kappa coefficient with quadratic weighting on similarity

Domain	Observed Kappa	Standard error	Confidence interval	Maximum possible	Proportion of maximum possible
Civil Engineering	0.8172	0.0872	0.6462-0.9882	N/A	N/A
Computing	0.6475	0.2749	0.1087-1.0000	N/A	N/A
Transporting	0.9231	N/A	N/A	0.9231	1.0000

5.4 Focus Group

The focus group discussion mainly focused on effectiveness, the ability of the system to retrieve the closest documents and their ranking. The group discussed relevance in the context of prior art searches and the different search strategies employed by examiners.

In general, the examiners were disappointed by the large number of irrelevant items on their list (note that the retrieved ‘results’ deliberately included a large number of

irrelevant patents). The examiners were not told about the 30/30 split at the time of the testing; they were under the impression that the purpose of the study was to generate search queries. The 30/30 split to remove positive bias may have inadvertently led to introducing negative bias.

In most cases, the examiners did not find the suggested keywords very helpful. They thought that topic modelling and visualisation could be potentially very useful. Ranking was in their opinion the most interesting aspect (note that the similarity scores were removed from the interface and the results were presented in random order). The examiners had different views about full text search strategies (most felt that the full text was full of misinformation) and which part of the patent provides the best starting point for their searches. The examiners were interested in the potential to discover new classifications and commented that incremental inventions are described using the existing taxonomy, but emerging disruptive technology and radically new inventions require evolving classifications.

The examiners made a number of suggestions about how the system performance could be improved. This included using flexible search strategies (e.g. using different parts of the patent text at different stages of the search process, selecting the most relevant paragraphs to the crux of the invention to make the retrieval task more focused, changing the weighting of the search parameters), hybrid search strategies (e.g. combining text and picture searches) and knowledge-based search strategies (e.g. enhancing the search with knowledge types such as method, process, methodology, etc.) and using domain-specific ontologies. The usability of the GUI interface and the impact of scrolling, especially on search term/synonym selection were also discussed. The focus group agreed that the best search tool should be one that supports a dynamic, iterative search process.

5. Conclusions

The main contributions of this study are the developed experimental platform and experimental protocol for comparisons between different AI techniques, which could be employed in patent prior art-searches. A wide range of state-of-the-art supervised and unsupervised machine learning approaches were considered that could support the tasks of feature extraction, query expansion, document classification, document clustering and topic modelling in patent searches.

The study concluded that it was not feasible to provide a fully automated solution as part of the application filing process. Nevertheless, the classification task produced very high classification accuracy, which shows potential to embed this function in the online patent pre-filing process to allow customers thinking of applying for a patent to undertake more easily due diligence checks.

The developed experimental platform for an AI-powered patent prior art search showed that AI has the potential to assist patent examiners in the future as part of

the prior art searching process. Different state-of-the-art AI algorithms can be used to retrieve the closest documents, rank relevant documents, suggest synonyms, suggest classifications, cluster and visualise the retrieved documents/concepts.

The study strongly suggests that the use of AI techniques to retrieve and rank documents could reduce the time and cost of prior art searches, and especially the process of sifting through a large number of patents retrieved. The experimental results for precision varied between 30% and 50%, which means that the first 10 search results contained between 3 and 5 relevant documents. However, AI is less effective in selecting relevant search queries. This was expected as the definition of the search statement is one of the most important and knowledge-intensive parts of the process. It requires clear understanding of the critical subject matter and the potential novelty of the application. Examiners often modify the search statement several times and often use words, which do not necessarily appear in the original claims. The construction of the search statement will remain a human task to suitably bound the AI search because of the wealth of specialist expertise and experience that an examiner has and is not something to be performed by AI.

Therefore, it could be feasible to provide examiners with a tool to aid searching but an AI-assisted search would require an examiner/expert to formulate a search statement; there are currently no effective AI algorithms, which can process the application and generate a search statement.

Another useful function could be topic modelling, i.e. the categorisation of patents into easily interpretable topics, each described by a set of keywords. It could be used by both applicants and patent examiners to visualise a domain but could be also utilised by data analysts to discover abstract topics, new terminology and trends in different domains emerging in different parts of the world.

The evaluation of the AI algorithms has clearly been challenging without separating the two aspects (search and retrieval). A better approach would have been to use the search statements formed by the examiners and focus on the retrieval and ranking aspects of the task only, although this was unfortunately out of the scope of this study because of IPO data sharing restrictions on the unpublished examiner search statements.

The study highlighted significant differences in the search strategies employed by the examiners and the need for innovative tools, which support more flexible search strategies. There are opportunities to enhance the current search process by developing new tools for retrieving image-based patents, collecting evidence of due diligence, spotting ambiguity, finding contradictions and visualising relationships among documents.

In conclusion, the study evaluated the viability of different AI technologies for patent prior art searching, including supervised and unsupervised machine learning, and found clear evidence that none of the available AI algorithms on their own can support every aspect of the prior art search process. The study identified the

potential of new approaches combining AI, NLP and computational semantics and highlighted the importance of human-centred decision and performance support tools. There is a need for a larger scale and more rigorous testing with more patents and more examiners, and more cutting-edge research on new algorithms supporting flexible search strategies and a dynamic, iterative search process.

Acknowledgements

This research was conducted with the financial support of the Department for Business, Energy and Industrial Strategy (BEIS) through the Regulators' Pioneer Fund (RPF). We would like to give special thanks to the IPO project board team led by Chris Harrison and Rich Corken (Maurice Blount, Peter Thomas-Keefe, Peter Evans, Kingsley Robinson, Stephen Otter, James Selway, Julia Leighton) and the IPO patent examiners (Kunal Saujani, Terence Newhouse, Alessandro Potenza, Chris Bennett, David Kirwin, Tom Simmonds, Caroline Bird, Manolis Rovilos) who were engaged with the core research team in examiner interviews, experimental testing and project evaluation.

Author contributions

Rossitza Setchi: Conceptualization, Methodology, Writing - original and final draft, Evaluation. Irena Spasić: Conceptualization, Methodology, Writing - original and final draft, Evaluation. Jeffrey Morgan: Data curation, Investigation, Coding, Validation. Christopher Harrison: Funding acquisition, Data curation, Evaluation, Writing – review and editing. Richard Corken: Funding acquisition, Evaluation, Writing – review and editing.

References

[1] Anderson L., Hanbury A., Rauber A. (2017). The Portability of Three Types of Text Mining Techniques into the Patent Text Genre. In: Lupu M, Mayer K, Tait J, Trippe AJ (eds) Current challenges in patent information retrieval. The information retrieval series, vol 29. Springer, Berlin/Heidelberg, pp. 287–304.

[2] Atkinson KH (2008). Toward a more rational patent search paradigm. In: Proceedings of the 1st ACM workshop on patent information retrieval, PaIR '08. ACM, New York, pp. 37–40.

[3] Harris CG, Arens R, Srinivasan P (2017) Using classification code hierarchies for patent prior art searches. In: Lupu M, Mayer K, Tait J, Trippe AJ (eds) Current challenges in patent information retrieval. The information retrieval series, vol 29. Springer, Berlin/Heidelberg, pp. 287–304.

[4] Alberts D., Yang C.B, Fobare-DePonio D., Koubek K., Robins S., Rodgets M., Simmons E., DeMarco D. (2017). Introduction to Patent Searching Practical Experience and Requirements for Searching the Patent Space. In: Lupu M, Mayer K, Tait J, Trippe AJ (eds) Current challenges in patent information retrieval. The information retrieval series, vol 29. Springer, Berlin/Heidelberg, pp. 287–304.

[5] Bashir S., Rauber A. (2010). Improving Retrievability of Patents in Prior art Search. In: Gurrin C. et al. (eds) Advances in Information Retrieval. ECIR 2010. Lecture Notes in Computer Science, vol 5993. Springer, Berlin, Heidelberg.

[6] Crestani, F. (2003). Combination of Similarity Measures for Effective Spoken Document Retrieval. Journal of Information Science, vol. 29(2), pp. 87-96.

[7] Adams, S. (2018). Is the Full Text the Full Answer? – Considerations of Database Quality. World Patent Information, vol. 54, pp. S66-S77.

[8] Helmers, L., Horn, F., Biegler, F., Oppermann, T., Muller, K.-R. (2019). Automating the Search for a Patent's Prior Art with a Full Text Similarity Search, PLOS ONE, 14(3): e0212103.

[9] Trippe A, Ruthven I. Evaluating Real Patent Retrieval Effectiveness. In: Lupu M, Mayer K, Kando N, Trippe AJ (eds.). Current Challenges in Patent Information Retrieval. Berlin, Heidelberg: Springer Berlin Heidelberg; 2017. p. 143–162.

[10] Oostdijk N, D'hondt E, van Halteren H, Verberne S (2010) Genre and domain in patent texts. In: Proceedings of the 3rd international workshop on patent information retrieval, PaIR '10. ACM, New York, pp. 39–46.

[11] Krishna, A., Feldman, B., Wolf, J., Gabel, G., Beliveau, S., Beach, T. (2016) Examiner Assisted Automated Patents Search. AAAI Fall Symposium Series: Cognitive Assistance in Government and Public sector Applications.

[12] Showkatramani G., Krishna A., Jin Y., Pepe A., Nula N., Gabel G. (2018) User Interface for Managing and Refining Related Patent Terms. In: Stephanidis C. (eds) HCI International 2018 – Posters' Extended Abstracts. HCI 2018. Communications in Computer and Information Science, vol 850. Springer, Cham.

[13] Manning CD, Raghavan P, Schütze H (2008) Introduction to Information Retrieval, Cambridge University Press.

[14] Fellbaum, C. (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press

[15] Spasic I, Ananiadou S, McNaught J, Kumar A (2005) Text mining and ontologies in biomedicine: Making sense of raw text. Briefings in Bioinformatics, vol. 6(3), pp. 239-251.

[16] Smith B, Ashburner M, Rosse C, et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology, vol. 25, pp. 1251-1255.

[17] Mikolov Tomas, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv, 1301.3781.

[18] Pennington J, Socher R, Manning C (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp. 1532-1543.

[19] Hughes, GF (1968) On the mean accuracy of statistical pattern recognizers. IEEE Transactions on Information Theory, vol. 14(1), pp 55-63.

[20] Spasic I, Nenadic G (2020) Clinical text data in machine learning: Systematic review. JMIR Medical Informatics, vol. 8(3):e17984.

[21] World Intellectual Property Organisation (2019) International Patent Classification. Available from: https://www.wipo.int/classifications/ipc/en/

[22] Blei DM, Ng A, Jordan MI (2003) Latent Dirichlet allocation. Journal of Machine Learning Research, vol. 3(4-5), pp. 993-1022.

[23] Trappey A.J.C., Lupu M., Stjepandic J. (2020) Embrace artificial intelligence technologies for advanced analytics and management of intellectual properties, World Patent Information, vol. 61, article 101970.

[24] Alderucci, D., & Sicker, D. (2019). Applying Artificial Intelligence to the patent system. Technology & Innovation, 20(4), 415-425.

[25] Aristodemou L. and Tietza F. (2018) The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data, World Patent Information, vol. 55, pp. 37-51.

[26] Wittfoth S. (2019) Measuring technological patent scope by semantic analysis of patent claims – An indicator for valuating patents, World Patent Information, vol. 58, paper 101906.

[27] Gormley C, Tony Z (2015) Elasticsearch: The Definitive Guide. Sebastopol, CA, USA: O'Reilly Media, Inc.

[28] Wolpert DH (1996) The lack of a priori distinctions between learning algorithms. Neural Computation, vol. 8(7), pp. 1341-1390.

[29] PATSTAT, European Patent Office, https://www.epo.org/searching-for-patents/business/patstat.html

[30] EP full-text, European Patent Office, https://www.epo.org/searching-for-patents/data/bulk-data-sets/data.html

[31] US full-text, US Patent and Trademark Office, https://bulkdata.uspto.gov/

[32] Andlauer, D. (2018). Automatic Pre-Search: An overview. World Patent Information, 54, pp. 559-565.

[33] Scikit-learn: Machine Learning in Python (2011), Pedregosa et al., JMLR 12, pp. 2825-2830.

[34] Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales, Educational and Psychological Measurement, vol. 20, pp. 37-46.