Expert Systems With Applications 177 (2021) 114910

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A non-factoid question answering system for prior art search

Morteza Zihayat *, Rochelle Etwaroo

Ted Rogers School of Management, Ryerson University, Toronto, ON M5G 2C5, Canada

ARTICLE INFO

Keywords:
Question answering
Prior art search
Pre-trained embeddings
Topic modeling
Search diversification
Sensemaking

ABSTRACT

A patent gives the owner of an invention the exclusive rights to make, use and sell their invention. Before a new patent application is filed, patent lawyers are required to engage in *Prior Art Search* to determine the likelihood that an invention is novel, valid or to make sense of the domain. To perform this search, existing platforms utilize keywords and Boolean Logic, which disregards the syntax and semantics of natural language and thus, making the search extremely difficult. Consequently, studies regarding semantics using neural embeddings exist, but these only consider a narrow number of unidirectional words. In this study, we propose an end-to-end framework to consider bidirectional semantics, syntax and the thematic nature of natural language for prior art search. The proposed framework goes beyond keywords as input queries and takes a patent as the input. The contributions of this paper is twofold; adapting pre-trained embedding models (e.g., BERT) to address the semantics and syntax of language, followed by the second component, which exploits topic modeling to build a diversified answer that covers all themes across domains of the input patent. We evaluate the performance of the proposed framework on the CLEF-IP 2011 benchmark dataset and a real-world dataset obtained from Google patent repository and show that the proposed framework outperforms existing methods and returns meaningful results for a given patent.

1. Introduction

Given today's rapid technological and innovative advances, trendy and en vogue ideas are born daily that can possibly garner large benefits for its creators. To ensure these ideas are indeed newly discovered and protected, the inventor could file a patent application to safeguard the right that they would solely reap the benefits of that invention. Thus, patents are granted to innovators to help them protect their creations from copyright infringement. However, before a new patent application is filed, a thorough research must be conducted by patent lawyers to not only check that the inventions are valid and patentable, but also to make sense of the invention in terms of its prospective user-base, potential new developments in the area, and how said idea or product could be produced and marketed to a greater extent. Additionally, competing inventors would often want to be aware of the state of a particular domain, by looking at current patent applications and patent documents to make sense of a particular invention space. This background research is essential for innovation, and it is called *Prior Art Search*.

To conduct Prior Art Research, patent agents often have to gather numerous patent applications, also referred to as documents in this paper, to sift through to find any piece of information that might jeopardize their client's idea or product. Moreover, they need to find answers to questions that thoroughly describe their client's inventions [16,20]. Most of these searches are conducted on public patent retrieval websites such as the *United States Trademark and Patent Office (USPTO* [1]*)* or *G*oogle Patents [2], as well as the other private subscription services which are available for use [30]. However, such platforms lack the ability to address all the challenges in prior art search. The challenges are discussed as follows:

1. **Limitations of keyword-based search:** Currently, to conduct prior art search, methods using keywords and Boolean Logics are exploited, which explicitly disregard inherently important details of natural language - that is, the semantics and syntax of words in combination. To devise keywords, this method of search requires the searcher to brainstorm words that would accurately and aptly describe the invention, which is quite a random process, as finding good keywords are never guaranteed. In Kim, Seo, and Croft (2011), the authors revealed that 87.5% of patent searchers use up to seven queries at their initial stage in prior art search to merely understand

* Corresponding author at: Room 2-032, 575 Bay Street, Toronto, ON M5G 2C5, Canada.
E-mail addresses: mzihayat@ryerson.ca (M. Zihayat), rochelle.etwaroo@ryerson.ca (R. Etwaroo).

[1] *https://www.uspto.gov/*

[2] https://patents.google.com/

https://doi.org/10.1016/j.eswa.2021.114910
Received 25 May 2020; Received in revised form 16 August 2020; Accepted 10 March 2021
Available online 18 March 2021

the information, then begin to develop well-formed queries to actually conduct the search. A closed-domain patent Question Answering (QA) system would be ideal in this case, where patent searchers would be able to ask questions in natural language about a certain area and be promptly returned an answer as opposed to spending time sifting through long patent applications to find relevant information. Furthermore, other approaches include the writing of complex and nested queries to conduct searches, which would require the patent searcher to take the time to learn the necessary query language and learn to perform robust searches. In many cases, neither keywords nor nested queries could be deemed sufficient in prior art search.

2. **Limitations of present embedding-based approaches.** From technical point of view, there are some studies that take word order in natural language into consideration via the use of neural embeddings, such as *word2vec* (Goldberg et al., 2014). However, these are still limited in the sense that they are unidirectional, and only consider words from either the left or right of a target word. Additionally, these approaches incorporate *L*atent Semantic Indexing, which does not take *homonyms* into consideration, nor can the model adapt well to new and unseen words. In a specialized domain such as the IP legal domain, this may pose an issue.
3. **Limitations of patent retrieval results.** When searching for similar inventions using the existing platforms, the results returned to the user is in the form of a patent application list, which can be in an unpredictable number of documents. A low number of result applications can lead to a lack in prior art search, which can threaten copyright infringement. However, a large number of documents would be time-consuming to cover for a human agent. These methods make sensemaking a mundane and time-consuming task since patent agents would have to take the time to sift to numerous documents to find the pieces of valuable information.

As such, to address previously mentioned issues, we propose a framework which will aid in prior art search. The main aim of this paper is to develop a question answering like system which helps patent lawyers to engage in sensemaking for prior art search in an effective manner. Through a question answering system, an agent would be able to ask non-traditional, non-factoid questions by means of using sentences of natural language, which considers the word order, word meaning and the thematic nature of language. Our system produces a combination of diversified results for patent agents to be used in the research and sensemaking processes. Our contributions are summarized as follows:

1. We propose a novel conceptual framework to aid in sensemaking for prior art search which uses bidirectional neural embeddings. We contend that models using bidirectional embeddings perform better than those that employ unidirectional embeddings, as we explore how the syntax and semantics of Natural Language in question and answering systems are affected by use of pre-trained neural embeddings. We propose three different versions of our framework, each employs a different approach of bidirectional neural embeddings and/or topic modeling.
2. We propose a framework which employs topic modeling to consider the thematic nature of natural language in results diversification. We argue that topic modeling would enable search results to be represented across different domains, thus, enabling sensemaking.
3. We evaluate the proposed framework using two real datasets collected from Google patents and CLEF-IP 2011. We consider three different query formats in terms of the length and nature of possible queries. We show that the proposed framework outperforms state-of-the-arts significantly.

The rest of the paper is organized as follows. In Section 2, we review and discuss the previous work done in the area of patent information retrieval for prior art search that concerns the semantics and syntax of natural language, and topic modeling for search retrieval diversification. We show where these areas are possibly flawed, which leads us to discuss our proposed conceptual framework and technical background in Section 3. In Section 4, we discuss the experimental design and evaluation metrics to be used to test the performance of the models, while in Section 4, we present the results from our experiments. Finally, in Section 5, we provide our conclusions and discuss our limitations and future work.

2. Literature review

In this section, we present an overview of existing work in patent information retrieval and related topics. .

2.1. Patent information retrieval

There have been considerable studies done in the area of patent retrieval, describing numerous approaches, varying from keyword search to language models (Shalaby et al., 2019). According to Jurgens et al., the patent retrieval process is quite elaborate (Jürgens et al., 2014). They present a framework which entails five iterative steps: *query formation*, *q*uery execution, *e*xamine results, *e*xtract results, and finally *r*eflect/stop. However, in the context of patent retrieval, one simple but extremely significant step is overlooked -keyword brainstorming. That is, the initial stage of the search, where patent agents devise ways to describe an invention to get similar patent applications to the said invention. A few of the different approaches to patent retrieval will be discussed in the following sections.

There are different ways in which questions about a certain topic or invention can be asked. While some use keyword approaches, where single words are used to describe inventions, there are studies that have used the entire full text of the patent document as a query (Helmers, Horn, Biegler, Oppermann, & Müller, 2019; Xue & Croft, 2009) and returns other patent applications based on similarity or proximity to the input document. However, the former approach may be problematic because inventions can sometimes be hard to describe since the language used in patent applications are often abstract, has specialized jargon or contain domain-specific words, which is difficult to represent a focused search area (Helmers et al., 2019; Bhavnani, Clarkson, & and Scholl, 2008). Additionally, in order to have high-quality searches, patent searchers need to have high expertise and experience (Jürgens et al., 2014).

Furthermore, the latter approach may have issues with complexity, as patent applications tend to be extremely wordy and complicated. To combat these issues, Query Expansion and Suggestion Techniques have been explored. A few studies deal with the expansion of a search question by utilizing catchphrase pair similarity between a search question and patent documents (Saraswat, Verma, & Gupta, 2019). However, in this case, catchphrases are referred to as noun phrases and there is neglect of the verb phrase which may be equally important in question and answering. Additionally, another study using query expansion techniques automatically inputs a list of keywords or relevant phases from an input document for a given question (Hristidis, Ruiz, Hernández, Farfán, & Varadarajan, 2010). However, these studies have proven to not work well due to noisy terms in initial questions. Also, computers tend to not know the ideal words for description, given the invention (Jürgens et al., 2014).

Moreover, Proximity Information was explored when questions were expanded by selecting expansion terms based on a given query topic (Khode et al., 2017; Shen, He, Gao, Deng, & Mesnil, 2014). Weights were given to expansion terms based on the assumption that the closest query terms are more likely to be related to a question and were therefore given higher weights. While this approach performed well, it is only limited to the number of words defined by the system.

Furthermore, the text of a patent document is structured in a standardized way (Khode et al., 2017). One study exploited the patent

document structure by using XML Query Language to find the similarity between two patent application structures (Golestan Far, Sanner, Bouadjenek, Ferraro, & Hawking, 2015), where user-specific questions were automatically rewritten into query language, which enabled search to be done using special XML query functions. Moreover, a method was proposed which employs Boolean Search and Decision Trees, where a Boolean Query is defined as a sequence of terms in a decision tree (Kim & Croft, 2015). However, these methods do not take the meaning of words, word order, nor context into consideration.

Other studies show how Pseudo-Relevance Feedback was implemented to enhance patent search. An ocular relevance feedback system was used to extract terms from judged relevant documents, then incorporated into proceeding query searches (Hristidis et al., 2010; Golestan Far et al., 2015). Moreover, there are existing studies which are focused on Citation Analysis, where citations, along with text content were used in a patent search. It was assumed that if a patent is cited by a large number of other patents, it is deemed more important in answering a question than one that is not cited as much (Fujii, 2007).

Additionally, some studies mention that query expansion is not sufficient, and the lexical relations should be considered (Andersson et al., 2017). With the use of two filters, after patent documents were represented as vectors, the pointwise mutual information and cosine similarities were calculated as syntagmatic and pragmatic filters respectively. Consequently, not many approaches take semantics into consideration when looking for answers to questions about patents. In Bhavnani et al. (2008), domain semantics are explored where Latent Semantic Analysis is introduced as one way to look at semantic information in search questions. Latent Semantic Analysis was proposed in Helmers et al. (2019) to identify similar overarching topics between a question and a document containing the answer and returns it to the user. Similarly, Latent Dirichlet Allocation was used to cluster questions and analyze trends in patent documents and to return relevant answers with topics to a user (Sales, Freitas, Handschuh, & Davis, 2015). Another method employs search diversification to improve answers to queries by identifying topic phrases that would represent underlying topics (Kim & Croft, 2015). Other studies consist of a Bag-o-Words approach, where the similarity between two documents was computed by representing both the question and patent documents as vectors and then using cosine similarity to measure how close the two are. Such methods disregard semantics and word order, which may not be an ideal model in patent question and answering.

Moreover, some studies implemented the use of word embeddings to garner context from a given search query (Kim & Croft, 2015; Helmers et al., 2019; Hofstätter, Rekabsaz, Lupu, Eickhoff, & Hanbury, 2019; Loginova, Varanasi, & Neumann, 2020; Bandyopadhyay, Ganguly, Mitra, Saha, & Jones, 2018). In Hofstätter et al. (2019), word2vec was considered, along with doc2vec to enhance semantic search in patent retrieval. However, these embeddings look at words from a particular window size, and in one direction, which only caters to the local context of documents. Furthermore, another study proposed that the global context of words to be taken into consideration as well, by using Latent Semantic Indexing to gather major topics in documents and return them as answers to patent searches. This method deals with homonyms well which might be important in a patent retrieval scenario as different inventions can be described using the same words.

2.2. Question answering systems

A Question Answering (QA) System is a specialized area of the Information Retrieval System that focuses on automatically finding answers to questions asked in human languages, to aid in the search for information. It is different from Information Retrieval in the sense that instead of returning documents, it returns snippets of relevant documents containing the answer to a question (Zhang et al., 2003). The Question Answering domain consists of two paradigms: Information-Retrieval-Based and Knowledge-Based paradigms. The former depends on extensive quantities of textual information to retrieve relevant documents to a given question, then returns a pertinent span of text as the answer. The latter paradigm describes a system that builds a logical semantic representation of a query, then mapping that query to a structured database (Jurafsky, 2000). Subsequently, the most common type of QA system concerns *Factoid Questions*, where the goal is to answer a question with a small text segment such as "How many countries are in South America?". However, recent research has shown the significance of *N* on-factoid Based Questions, which consists of open-ended questions (Kim et al., 2011; Qu, Yang, Croft, Scholer, & Zhang, 2019; Zhang et al., 2003), such as "How to fix a broken iPhone Screen?"

Question Answering for Patent Retrieval. In conducting prior art search, the initial stages consist of Sensemaking, where patent agents are rapidly trying to make sense of the novelty of an invention by constructing representations to organize existing patents and to learn the details about them (Kim et al., 2011). According to Russell and Stefik, in Sensemaking Theory, when users perform a search for information, they attempt to make sense of it by perpetually modifying, developing and refining queries (Jurafsky, 2000).

The question answering Process mainly consists of three components – question processing, document processing, and Answer processing. There are several studies which endeavor to improve with either one of the three categories or a combination of one or more. In Harabagiu et al. (2000), the authors used *WordNet* to boost the answer processing component, which did show improvement, while another treated a query as a question and returned a set of ranked documents. The semantic representations for the query were compared to that of the documents to discover the answer. This approach showed up to 53.33% in precision. However, the system was described as open-domain but only consisted of a small number of business-driven data. Moreover, these have shown probabilistic methods can be used in QA systems. In Xu (2003), the authors explored a hybrid approach by using probabilistic methods and other extraction methods such as name finding and relation extraction. All in all, none of the aforementioned approaches take semantics into context, neither word order despite the fact that in any Information Retrieval task, this has shown to be important (Bhavnani et al., 2008).

Topic Modeling in Answer Processing. Topic modeling has been used in search diversification in the IR Domain. Carterette and Chandar (2009) show that traditional models of retrieval assume documents are independently relevant. However, when the goal is to retrieve diverse topics, such methods are not sufficient. Therefore, several studies used LDA to propose a model with the goal of finding a set of documents that cover the different aspects of an information need (Tang, Li, Zhang, & and Mei, 2016; Carterette & Chandar, 2009). Similarly, Vikraman, Croft, and O'Connor (2018) discussed how retrieving precise answers is an important task. They evaluated the impact of applying existing document diversification frameworks to the problem of answer diversification and used the LDA in their framework. The results showed that LDA outperformed the other two methods in two of the three test datasets. Moreover, Yu et al. (Yu, Mohan, Putthividhya, & Wong, 2014) explored the use of LDA in diversified search results in e-commerce sites. The LDA model can discover meaningful user intents and the LDA-based approach outperforms the baseline production ranker and three other diversified retrieval approaches. Therefore, LDA is shown to be a valid method to gain diversified answers that consider the thematic nature of natural language.

Table 1 shows an overview of the recent methods. This overview provides an insight into the techniques and approaches used in the patent retrieval, as well as the Question Answering domains. Our area of focus would mainly encompass a framework that would enable patent agents to ask questions about new inventions to help determine validity for a new patent application and to aid in sensemaking. To enable this system to understand the context in natural language, we will be using bidirectional neural embeddings alongside other state-of-the-art techniques, such as Topic Modeling. As a result, rather than requiring patent

Table 1
Summary of recent literature work in depression detection.

Paper	Research Objective	Research Methodology	Key Findings
Helmers et al. (2019)	- Patents guarantee their creators protection against infringement. - Current search methods are time-consuming and prone to errors.	Transform patent texts into numerical vectors using different methods BoW, LSA, word2vec, doc2vec, and Bow + word2vec	- BoW feature representations outperform all others on the entire document. - The next performing representation is doc2vec on the claims and abstract sections.
Golestan Far et al. (2015)	Investigate the influence of term selection on retrieval performance by using the description section of a patent query on a language model, and scoring functions.	- Indexed each section of the patent application in a separate field - Developed an ocular query by defining a relevance feedback system that extracts terms from judged relevant documents	- Query reduction should be enough for effective prior art patent retrieval - There are ways to eliminate poor query terms, such as negative words which when done, improves query retrieval performance
Kim et al. (2011)	- Novel Boolean query suggestion technique - Generate Boolean queries by decision trees learned from pseudo-labeled document	- convert the path of the tree to a Boolean query - a learned decision tree could imply a Boolean query representing a set of relevant documents	- The system can not only generate many effective Boolean queries but also select highly effective queries for the suggestion - Effective queries can be identified by searchers in real environments.
Kim and Croft (2015)	Since a patent document generally contains long and complex descriptions, generating effective search queries can be complex and difficult	-identify topic phrases is generating a list of effective phrases for diversification - used as query patents - developed a 'diversification' algorithm that gives weight to each topic term in a query - used unigram terms instead of phrases to express patent topics because patent documents frequently contain longer technical terms	- Given an initial retrieval result of each query patent, can identify topic phrases to represent underlying query topics and diversify based on the identified phrases - Diversification can increase the ranks of relevant documents related to diverse topics, and enabling the user to recognize the diverse aspects of query patents
Fujii (2007)	Propose a method that uses text content and citations to enhance search	- For the text-based retrieval, the claim(s) in each document was used to perform word-based indexing - perform the text-based retrieval and obtain top N documents. They then compute the citation-based score for each of the N documents	A combination of the text-based and citation-based methods improved the text-based method
Marrara and Pasi (2015)	- to classify patents by exploiting their XML structure - uses XML formatted documents - propose an approach - that relies on the recent outcomes of research in XML Retrieval.	- to classify patents by exploiting their XML structure - uses XML formatted documents - propose an approach - that relies on the recent outcomes of research in XML Retrieval.	- Showed that flexible constraint on tag names similar - Showed how different flexible constraints can be combined to compute an overall relevance degree of a document fragment
Samarinas and Tsoumakas (2018)	- propose a new ranking method for factoid answers that take into account the semantic similarity of the context - Describe methods used for non-factoid answer extraction and ranking	- GloVE pre trained word embeddings on Wikipedia - Extracted factoid answers after the removal of Boilerplate content from documents, then ranking of factoid answers	- Words that are closer to the factoid in the context window contribute more to the context score - the number category has the lowest MRR
Jia et al. (2018)	- propose a method for temporal QA that can run on top of any KB-QA engine. - Decompose temporal questions and rewrite the resulting sub-questions so that it can be separately evaluated by a KB	Given a query, it works in four stages: 1. Detect that a question is temporal by using 'temporal signals.' 2. Decompose and rewrite a question 3. Obtain an answer 4. Apply constraint-based to produce a final answer	- TEQUILA enables KB-QA systems to answer composite questions with temporal conditions - outperforms state-of-the-art baselines in F1 scores
Hofstätter et al. (2019)	- create a model that considers both the local and global contexts of words - enhance the skip-gram model using global retrofitting - Filter global similarities using global context	- Propose different models of skip-gram, LSI and retrofitting, or a combination of them - Conducted retrofitting on skip-gram embedding by using an external resource - Added a post-filter to remove words appearing in the skip-gram that does not occur in the external resource	Observed some improvements over baselines on the CLEF-IP 2013 task
Andersson et al. (2017)	To propose an automatic query expansion method to identify domain-specific lexical-semantic relations	- compared several different features for QF such as phrases, words, bigrams - word2vec model was trained on patent data on 300 dimensions	- The AQE method underperforms in comparison to the pure NLP method - There is an improvement when using 5 terms for expansion, but it is not statistically significant enough

agents to read through long and mundane documents, our approach would enable them to simply ask questions about a certain topic, invention or domain. We would like to introduce a framework which uses Bidirectional Encoder Representation from Transformers (BERT), which is a neural pre-trained embedding that takes semantics into context by looking at text sequences in both the left and right direction of a given question (Saraswat et al., 2019). As a result, it is expected that questions should be understood in a global context and answers should be given based on such. For this study, the full text of patent applications will be explored to eliminate any initial brainstorming to come up with keywords.

3. Conceptual framework

Our goal is to devise a new question and answering system which would aid in sensemaking for patent prior art search. The main basis of the study is twofold; the first is to propose a new representation to explore the effects of full-text search vs. keyword search by use of bidirectional neural embeddings and show how the semantics and syntax of natural language can affect patent retrieval. The second aspect is to come up with a new representation for the search results returned to the patent agent in which the results cover all topics across various domains, by the use of topic modeling. The proposed framework in this study is beneficial to lawyers and agencies to have a more convenient and efficient way of doing prior art search. In this section, we first present a basic version called *NFQA*, Non-Factoid Question Answering system. Then, we introduce two improved versions of NFQA called, NFQA+ and NFQA++ by proposing effective components to enhance the framework.

3.1. NFQA: The baseline framework

Fig. 1 shows our basic framework for prior art search. Our framework begins with data collection and then modeling different aspects of the prior art search. There are two main components of *forming the question based on user input query* (e.g., a patent) and *f* inding the answer based on the content modeling and representations of the input database. In the first component, an effective representation of the input patent is built. Given all the patents in the database, NFQA builds and vectorizes an effective representation of the database content. Given the representation of the query and the patents in the database, the framework builds a document as an answer based on the retrieved contents. Below, we present the details of each component.

3.1.1. Data collection

The data that make up the answer component for NFQA can be collected from different data sources. In this study, we consider two different sources: (1) the Google Patents Public Data[3], which consists of different datasets regarding patents and publications, (2) CLEF-IP 2011 benchmark dataset (Piroi, Lupu, Hanbury, & and Zenz, 2011) for prior art search[4]. Below, we explain the details of the data collection for each dataset:

- *Google Patent dataset:* Full-text patent applications from over seventeen (17) different countries worldwide, provided by the ISI Claims Patent Services are contained in this particular database. The data is stored and accessed through Google's BigQuery application, and SQL queries are written to select features used in this paper. The titles and abstracts from 50,000 patent publications from different categories of applications including utility, health, and technology are in the dataset. Overall, there are a total of 170,830 sentences, and over 1067,650 words, with approximately 58,000 unique words across more than 15 different Part of Speech tags. The average number of words in a sentence is 31.8. Moreover, on average, the Part-of-Speech with the highest count is the most *Noun Phrases*, followed by *Adjectives*, which is approximately 50% less. This is representative of natural language since these Part-of-Speeches represents the descriptive and lexical nature of language. The other dominant Part-of-Speech categories in the dataset are Prepositions, followed by Verb Phrases.

- *CLEF-IP 2011:* The dataset is created for different tasks including prior art search. The topics and corpus are similar to the CLEF-IP 2010 dataset [5]. NFQA is designed to use a patent application's title and abstract as input. The dataset is downloaded from the CLEF-IP 2011 corpus [6], which consists of over 500,000 unique patent application documents, produced in German, French and English. We are particularly interested in the A1 type of documents, which are the publication of application with search report because these consist of the title and abstract of patent applications. Furthermore, our focus is on patents in English. The A1 files are in XML format whose filename contains 'A1'. A python script is implemented to recursively parse all files and extract the abstract and the title tags, where the language attribute is set to English. The transformed dataset contained approximately 310,000 unique English patent abstracts.

3.1.2. Modeling and representations

The main component of any prior art search framework is how to build an effective representation of the patents in the database and the input query such that the output is effectively retrieved based on what a patent agent is looking for. In this paper, we leverage BERT (Devlin, Chang, Lee, & and Toutanova, 2018) as the main component to model both input query (e.g., the patent) and the patents in the database.

BERT Pre-trained embeddings are neural embeddings that enable natural language to be represented in a vector space. BERT embeddings were trained using a deep learning, multi-layer bidirectional transformer encoder on Wikipedia and Books corpus by conditioning tokens from both left and right contexts. Thus, it enables a bidirectional model that is based on the transformer architecture and utilizes a 'Masked Language Model' by replacing the sequential nature of RNN with a fast Attention-based approach (Devlin et al., 2018). In summary, BERT takes a sequence of words as input, continuously flowing up a stack of encoders, while each layer applies self-attention. Self-attention looks at other positions in the input sequence for clues that can lead to a better encoding for the word. Then, it passes the results through a feed-forward neural network, which ultimately sends it to the following encoder. BERT's language modeling task masks 15% of words in the input and asks the model to predict the missing word. As a result of this, the BERT pre-trained embedding is birthed.

BERT's word vectors encode and represent the complex and rich linguistic nature of natural language such as syntactic and semantic features. In this paper, BERT embedding is selected as the preferred method of pre-trained embedding because of its deep bidirectional and contextual abilities. Moreover, BERT can fine-tune the pre-trained model on other tasks. Given a patent p, to fine-tune the model for the prior art search and represent the patents effectively, we build a pre-tune dataset as follows. Each patent in a our dataset is represented as a set of sentences s_i. Given an input patent p, we build pairs of $\langle p, s_i \rangle$ where s_i is a sentence representing a content related to the input patent p. Then, p and s_i are separated with $[SEP]$ token. Given each token in the text, its final embedding is fed into a classifier. The classifier considers a single set of weights for every word. To come up with a probability distribution over all of the words, we use the softmax activation. This process will fine tune the pre-trained BERT to give a more representative and meaningful distribution vector to each sentence. Fig. 2 shows how we build the representation of the sentences using the pre-trained BERT. Note that, we use the same pre-trained embedding model to represent the input query (e.g., a patent).

3.1.3. Question and answer processing

Given the pre-trained fine-tuned representation, to find the answer for a given question, NFQA builds the representation of both input question and the possible answers. Given the patent application dataset, the abstracts are tokenized into sentences and then represented in a vector space using the modeling discussed in previous section. This results in multiple vectors, one per sentence. This approach ensures that each sentence is captured in a vector space based on its surrounding context, semantics, and syntax. These sentences and vectors would ultimately represent the answers that will be given to a patent agent based on the questions he inquires. Additionally, to process the question component of this model, a similar procedure is followed. That is, the input patent is considered and processed as a sentence and is vectorized using the pre-trained BERT model. One of the disadvantages of BERT is that independent sentence embeddings are not calculated. Therefore, it will be more challenging to extract sentence embeddings from BERT. To resolve this problem, we follow (May, Wang, Bordia, Bowman, & and Rudinger, 2019) to derive a fixed sized vector by calculating the average of the outputs. Followed by these steps, to build the answer to an input query, the similarity of the input patent and each sentence in the database is computed. To do this, we use a distance measure, *soft cosine similarity*, which computes the angular distance between the input query and each sentence in the database. Assume that S_p and S_{db} are embedded vectors representing the input patent and a sentence in the patent database, respectively. The cosine similarity is computed through the matrix M to indicate the similarity between the two embedded vectors. Matrix M is based on *L*evenshtein distance (Navarro, 2001); then the soft cosine function is mathematically modeled as follows (Sidorov, Gelbukh, Gómez-Adorno, & Pinto, 2014):

[3] https://www.google.com/?tbm=pts
[4] Other repository can also be used in our framework
[5] http://www.ifs.tuwien.ac.at/clef-ip/download/2010/index.shtml
[6] http://www.ifs.tuwien.ac.at/clef-ip/download/2011/index.shtml

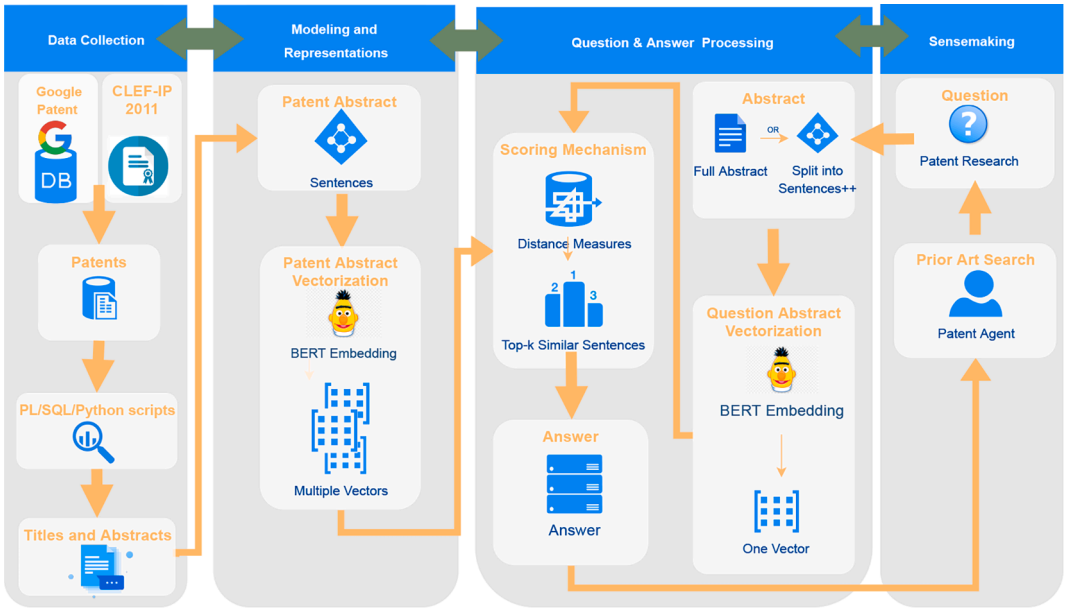

Fig. 1. NFQA: The baseline framework.

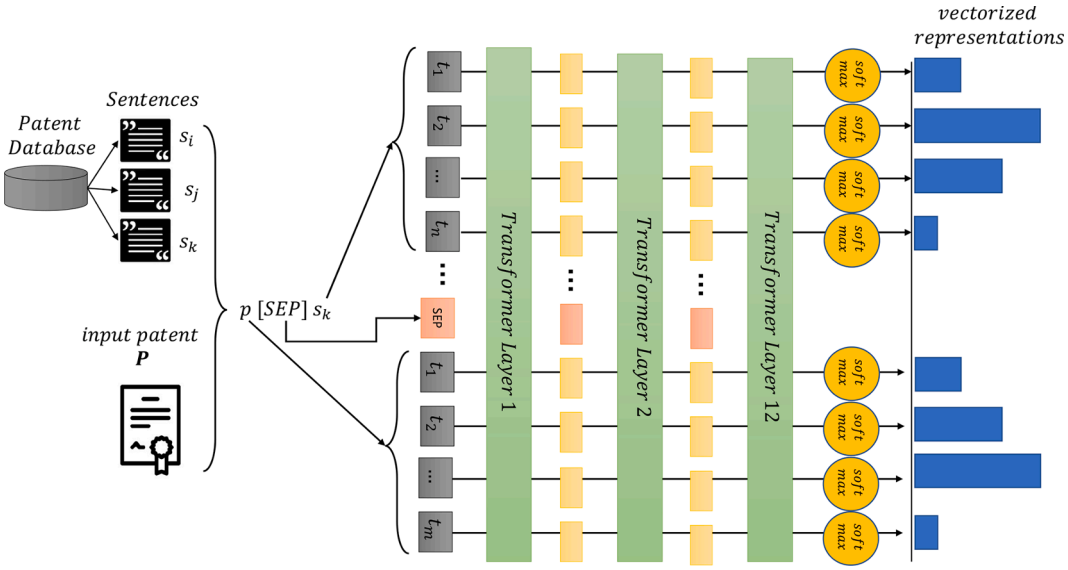

Fig. 2. Pre-trained BERT fine tuning for prior art search.

$$Soft - Cosine(S_p, S_{db}) = \frac{\sum_{i,j=1}^{N} M_{ij} S_{p_i} S_{db_j}}{\sqrt{\sum_{i,j=1}^{N} M_{ij} S_{p_i} S_{db_j}} \cdot \sqrt{\sum_{i,j=1}^{N} M_{ij} S_{db_i} S_{db_j}}}$$

3.1.4. Sensemaking

The patent agent plays a major role in this stage of our framework. In this vertical, the patent agent inquires about a given field to make sense of the information or to conduct prior art search. The question asked by the patent agent, which is in the format of a sentence or paragraph, ideally an abstract, goes through the same modeling process as discussed before, where tokenization occurs, followed by vectorization, then a similarity analysis and scoring mechanism. Note that, the output result is a document representing the most similar contents (e.g., sentence, paragraphs) from patents stored in the database with respect to the input patent. In our experiment, we will show an example of the working system.

3.2. *NFQA+: Bringing topics to the modeling*

NFQA focuses on the sentence-based analysis to find the best answer. This approach lacks considering the structure of the document as well as the topics represented by the document. As results, the output answer will be more focused on the similar sentences with the similar meaning to the input query and might not cover all the topics presented in the input query, thus missing important information. To address this, we argue that the modeling should also consider the topics and their similarities into account. As the first improvement, i.e., NFQA+, we discover the topics represented by the top-m (m is much larger than k sentences for the final answer) similar sentences retrieved by NFQA. *Topic modeling* facilitates the discovery of hidden topics found in a large corpus of text, by allowing the underlying semantic features of natural language to be revealed through probability distributions. In this paper, we apply the

popular topic modeling method called LDA (Blei, Ng, & and Jordan, 2003). Note that, our goal is generate the a probabilistic topic model, therefore any other topic modeling technique that represent a document as a distribution over topics can be plugged in here.

In LDA, it is assumed that a chunk of text has a probability distribution of topics, while a topic has a probability distribution of words. A corpus D is a group of texts, which in this case are all the closest answers to the given question. The words in the corpus of these answers make up the vocabulary. Each answer is considered as one document which is a sequence of N words, while a word is a single unit of meaning in the data – an element in the vocabulary. Meanwhile, a topic z is an abstract notion of a probability distribution over the vocabulary (Blei et al., 2003). LDA assumes that words carry strong semantic information and similar documents carry the same words. For the generation of texts, for each w in D, we assume:

1. Select θ, a topic distribution in the patent applications
2. Select length N
3. For each word n in N in each abstract w,
4. Select a topic z_n from step 1
5. Select a word for the corresponding topic z_n from distribution β, which is the distribution over words in the vocabulary.

The equation below expresses the process over a joint distribution, where K is the number of topics.

$$p(\beta, \theta, z|W_D) = \Pi_{i=1}^{K} p(\beta_i)\Pi_d^D p(\theta_d)(\Pi_{n=1}^{N})p(Z_{d,n})p(p(W_{D,n}|\beta, Z_{d,n})) \quad (1)$$

Additionally, given the above, we use the following equation to make inferences to measure the distribution of the latent variables (topics) given the observed corpus.

$$p(\beta, \theta, x, W_D) = \frac{p(\beta, \theta, z, W_D)}{p(W_D)} \quad (2)$$

We use topic modeling in an unconventional method to devise overall themes from the *top-m* closest sentences to a given question. As such, LDA is executed on the sentences with the highest m similarity scores. Then a topic-based filtering is employed. LDA is an unsupervised learning algorithm and we do not need to have the ground truth topics of the input contents. However, one of the most important parameters in finding the best topic model is the number of topics. To find the right number of topics, we use two common performance measures *perplexity* and *coherence*. The model with the highest perplexity and coherence is selected as the best model and its number of topic is picked as the best number of topics. In this study, Gensim package (Srinivasa-Desikan, 2018) is used to run LDA and measure the coherence and perplexity for all the topic models built using our framework. Algorithm 1 shows the details of the proposed method to discover a more diversified answer using the topic modeling. Given *t*op-m sentences from BERT answer processing, we first run LDA over all the text in our database to build the topic model and find topic distribution vectors (i.e., V_t). Then, using a function called *s*how_topics we build vectors representing *t*op-m sentences (i.e., T_{ans}) as well as the input query patent (i.e., T_p). In the last step, we first discover the most representative sentence per topic and then calculate the similarity of each vector in T_{ans} and S_{T_p} and return *t*op-k most similar sentences as the answer. In our experiments, we show that the topic-based filtering step will improve the quality of the results by considering more diversified sentences as the answer compared to those returned by basic NFQA.

Algorithm 1: Topic-based Answer Discovery

Input: patent database D, top-m sentences S_{top}, query p
Output: Answer document
$V_t \leftarrow$ Run LDA over D
$T_{ans} \leftarrow \emptyset$
for each sentence $s_i \in S_{top}$ **do**
$v = show\_topics(s_i, V_t)$

(continued on next column)

(continued)

Algorithm 1: Topic-based Answer Discovery

Add v to T_{ans}
end for
$T_p \leftarrow show\_topics(p, V_t)$
Calculate cosine similarity between the most representative sentence of each topic S_{T_p} and T_{ans}
Return Top-k similar vectors in T_{ans} to S_{T_p}

3.3. *NFQA++: Query vectorization*

The question comes from the patent agent in the form of one or more sentences (e.g., patent abstract). In NFQA and NFQA+, the entire paragraph of words, which in this case is the abstract, is treated as a single vector and embedded in vector space using BERT embeddings and then topic modeling. We argue that it might be helpful to consider each sentence of the input query separately. As another improvement to the basic framework, to represent more diverse semantic and syntactic content, we modify the framework such that the input query is processed as sentence in which each sentence in the question abstract will be a separate vector. Thus, this model's question abstract consists of multiple vectors.

Example 1. Take the following paragraph of text into consideration:

"*A patent gives the owner of an invention the exclusive rights to make, use and sell their invention. Before a new patent application is filed, patent lawyers are required to engage in Prior Art Research. To determine the likelihood that an invention is novel, valid or to make sense of the domain prior art search is used. To perform this search, existing platforms utilize keywords and Boolean Logic. These disregard the syntax and semantics of Natural Language and thus, making the search extremely difficult.*"

This is a body of text which will be considered as a question in both NFQA and NFQA+. The paragraph contains five (5) sentences. In the NFQA and NFQA+, the entire question abstract is preserved as one and represented in vector space via BERT embedding. Thus, the question would be considered as one vector. In our NFQA++ model, we split the paragraph by sentence, keeping the word order and meaning, then we represent it in a vector space via BERT embeddings. Thus, we will eventually have one vector per sentence in both the question and the answer, and compute pairwise vector comparisons via cosine similarity. Similarly, each question from each abstract in our dataset is vectorized in this manner.

Fig. 3 illustrates the complete version of the framework, NFQA++. In our experiments we evaluate all three versions and discuss the details of their capabilities.

4. Experiments

In this section, we evaluate the performance of the proposed framework and compare it to the state-of-the-art techniques as baselines. Below, we first present the experimental setups, then the evaluations will be done on different types of question abstracts which are created to explore how the framework performs on different aspects of natural language.

4.1. *Experimental design and datasets*

To test the robustness of our framework, as well as to explore the different aspects of language, we devise a way to test four different categories of question abstracts, which are selected at random from the collected dataset. The four categories are as follows:

- Long Abstracts
- Short Abstracts
- Abstracts with Specialized Jargon

- Joint Abstracts, which were a combination of sentences from a random selection of Abstracts, were used to test each of the four Models.

Each category consists of one hundred test queries, which are used as the test cases to test across all of the models. Given each category of the abstracts, each experiment is conducted 20 times, and the average values of the performance measures are reported. Examples of each type of question abstract can be referenced from the Appendix of this paper. In Google dataset, once the patents are collected, we first pick 100 patents for each category. We review the patent carefully to make sure that they belong to the correct category. The similar procedure is applied to find the right patents for each category in CLEF-IP 2011 dataset. CLEF-IP 2011 provides a input query file that contains the patent ID as the query and its related and unrelated patents in the corpus. We review the patents in the corpus and choose our patent queries by finding the right category for each. Then, we parse the file given by CLEF-IP 2011 to find the related and unrelated patent with respect to each query.

4.2. Baselines

To the best of our knowledge, our framework is the first non-factoid question answering platform for prior art search. As baselines, two different methods are implemented to compare and benchmark metric scores. The first one is a modified version of the method proposed in (Helmers et al., 2019) which is based on *Doc2Vec* (Cer et al., 2018). This method is called *D*oc2VecQA in the experiments. As the second baseline, we develop another method, called *USEQA* exploiting *U* niversal Sentence Encoder (USE), which is a pre-trained embeddings (Le & Mikolov, 2014). Both methods are used in a similar manner as our basic framework (i.e., NFQA). That is, after vectorization of both the question abstract and the abstract dataset, the sentences are represented in a vector space, then the cosine similarity is used to find the top-k most similar sentences to the input abstract.

We consider all three different versions of our proposed framework. For reference, Table 2 is provided below which simplifies and highlights the differences in the three proposed versions.

4.3. Datasets and query formation

In this study, we use two real datasets in our experiments. As discuss.

4.4. Evaluation metrics

Modified versions of the *ROUGE-L* and the *METEOR* metrics are used to compare the performance for the baselines to the proposed models.

(1) Recall-Oriented Understudy for Gisting Evaluation-Longest Common Sequence (ROUGE-L): Traditionally, the ROUGE metric, first introduced by Lin (2004), is used to measure the performance of an automatic text summarization model, by comparing overlapping n-grams in a machine produced summary verses a human-produced summary. However, in this case, we use patent abstracts from the dataset, which consists of the answers, and compared it to the question from the patent agent. The ROUGE-L method can be implemented using the following formula,

$$ROUGE-L = \frac{\sum_{Q \in Ans} \sum_{lcs \in Q} Cm(lcs)}{\sum_{Q \in Ans} \sum_{lcs \in Q} Cnt(lcs)} \quad (3)$$

where: *lcs* represents the longest common sequence between the question and the answer, *Q* represents the Question, *Ans* represents the Answer, *Cnt* count of longest common sequences, and *Cm* is the count-match of longest common sequences. Note that the longest common

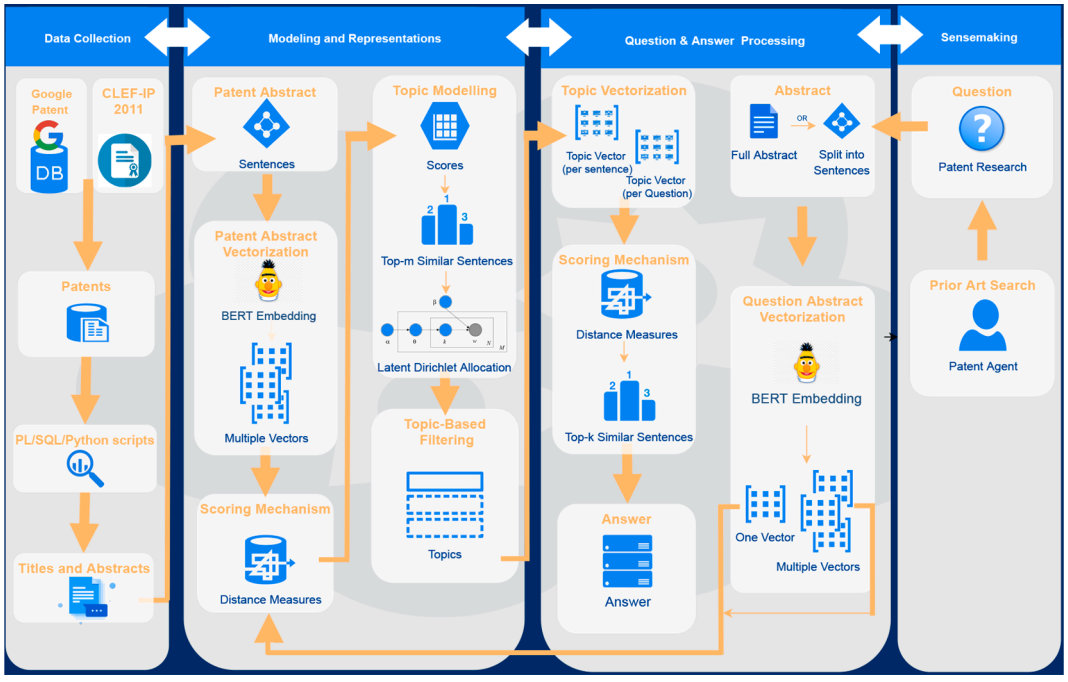

Fig. 3. NFQA++ framework.

Table 2
Summary of the proposed Models in our Framework.

Model	Input	Modeling and Representations	Output
NFQA	Abstract as Question - One vector	• Tokenize Dataset on sentences • Apply fine-tuned BERT on each sentence in the dataset and the input query • Calculate soft cosine distance between the input query and each sentence	Top-k similar sentences to input as Answer
NFQA+	Abstract as Question - One vector	• Tokenize Dataset by sentences • Apply fine-tuned BERT to each sentence Dataset and to the input query • Calculate soft cosine distance between the input query and each sentence in the dataset • Get top-m most similar sentences to the input query • Apply LDA on each of the m sentences • Apply LDA on the input query • Calculate cosine distance between the topic vector in the Question Abstract and each topic vector in top-m similar sentences	Top-k similar sentences from where each topic was represented as Answer
NFQA++	Abstract tokenized as sentences as Question- Multiple vectors	• Tokenize Dataset by sentence • Apply fine-tuned BERT to each sentence in the input query and the dataset • Calculate soft cosine distance between each sentence in the input query and the dataset • Get top-m most similar sentences to the input query sentences • Apply LDA on each of the m sentences • Apply LDA on each sentence of the input query • Calculate cosine distance between the topic vector in the Question Abstract and each topic vector in top-m similar sentences	Top-k similar sentences from where each topic was represented as Answer

sequence which reflects sentence-level word order, thus making it an ideal metric to measure the use of syntax of natural language (Lin, 2004).

(2) Metric for Evaluation of Translation with Explicit ORdering (METEOR): The Meteor Metric approximates the similarity between two groups of texts by matching the first text to the second group of text. Like the above, in this case, the METEOR metric will be used to look at the similarity between the question and the retrieved answers in our system. The semantic equivalence of two bodies of words is done by looking at synonyms in the texts, done by the use of WordNet, as well as inflections was considered, done via stemming. Alignments of words are based on exact, stem, synonyms, and paraphrase matches between words and phrases (Banerjee & Lavie, 2005).

METEOR evaluates the answers to a question by computing a score based on explicit word-to-word matches between the question and an answer. To do this, it computes the weighted F-score, as in the following formula:

$$F = \frac{PR}{\alpha P + ((1-\alpha)R)} \tag{4}$$

where, F is the F-score, which is the weighted harmonic mean, P is the precision of words, R is the recall, α is represented as the weight.

The above formula in (4) only looks at unigram counts. Thus, to account for the word order in the question and answers for longer segments of words, a penalty function is introduced, where gamma determines the maximum penalty (Banerjee & Lavie, 2005).

$$Penalty = \gamma(\frac{c}{m})^{\beta}, where 0 \leqslant \gamma \leqslant 1 \tag{5}$$

Additionally, c is the number of matching groups of words, and $m$is the total number of matches. Thus, if $m$is adjacent, the number of word groups is lower and the penalty decreases. Thus, the METEOR score is computed as follows:

$$METEOR = (1 - Penalty) \times F \tag{6}$$

4.5. Evaluations on long abstracts

Fig. 4 presents the performance of different methods on long abstracts in Google patent dataset. Long abstracts are used as questions to test whether the models are sensitive to the length of sentences and the number of words used when represented in an embedded vector space. NFQA slightly surpasses NFQA+ in the METEOR score in this category. Overall, for the length of the question did not seem to affect the models we proposed in this paper. The performance superiority of NFQA++ implies that the most important component that our framework works better than the baseline on the long abstract is the combination of topic modeling and sentence-based similarity analysis component. The topics present more diversified results compared to the time that only embedding being used to find the answers.

Fig. 5 shows the results of different methods on CLEF-IP 2011 dataset. On average, the performance of NFQA++ surpasses NFQA+ in both the METEOR and ROUGE-L on this dataset. In terms of ROUGE-L, this category of abstracts shows that vectorizing question abstracts as single sentences work best. Overall, when taking both metrics into consideration, the performance superiority of NFQA++ implies that the most important component that our framework works better than the baseline on the long abstract is the combination of topic modeling and sentence-based similarity analysis component. The topics present more

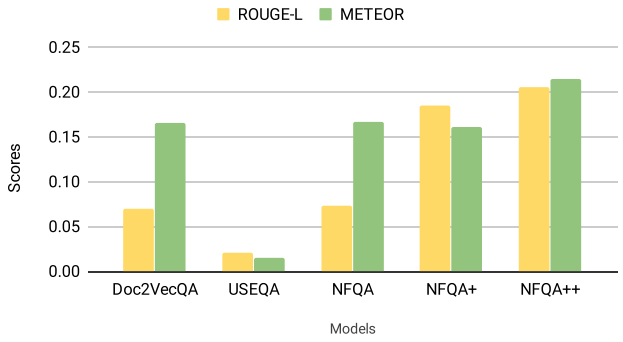

Fig. 4. Google dataset – Average METEOR and ROUGE-L scores by models on long abstracts.

diversified results compared to the time that only embedding is being used. As shown, Doc2VecQA works significantly better than USEQA where in most of the cases could not return any of related patents to the query.

4.6. Evaluations on short abstracts

The short abstracts are another common query type. This type is more similar to the input query of factoid question answering systems. Fig. 6, shows the results for the proposed methods as well as the baselines on Google dataset. NFQA+ model has a slightly higher score than the NFQA++ model in ROUGE-L. This could be attributed to the multiple vectorizations of question sentences, added to the topic modeling based filtering. Since the larger portions of words in NFQA+ used in topic modeling could garner more coherent topics, this results in a slightly better performance of NFQ + in terms of ROUGE-L. The other observation in this experiment is the big gap between NFQA and the other two versions. This also confirms the fact that the topic modeling played an important role to retrieve the best answer to a given question particularly when the length of the input query is short.

Fig. 7 shows the performance of different models on CLEF-IP 2011. In this category, NFQA outperforms other models in terms of METEOR. This is due to the fact that the input queries in this category of the dataset are very short, thus vectorizing the input query as different sentences might not help improving the quality of the results in terms of METEOR. However, NFQA+ and NFQA++ outperform NFQA significantly in terms of ROUGE-L. This means that the similarity of the majority contents between the input query and the retrieved patents is much higher when we use topic modeling. Regardless, considering both metrics, our frameworks NFQA, NFQA+ and NFQA++ surpass the baselines in both metrics, where NFQA achieves the highest METEOR score, and NFQA++ achieves the best ROUGE-L score. Doc2Vec has a lower score than the our models in both METEOR and ROUGE-L. The superiority of the proposed models verifies the fact that they are resilient against specialized words and return meaningful answers for a given query with short queries.

4.7. Evaluations on abstracts with specialized jargon

One of the most common type of contents in legal domain is the ones with specialized words. Therefore, a question answering system in this domain should be able to handle texts with highly specific and specialized words effectively. In this section, we investigate the performance of the proposed frameworks on such type of patents. To do so, we consider abstracts with specialized jargon as the input query. This category of abstract is used to test how well the model can resist special jargons used in patent applications. Fig. 8 shows the results of the

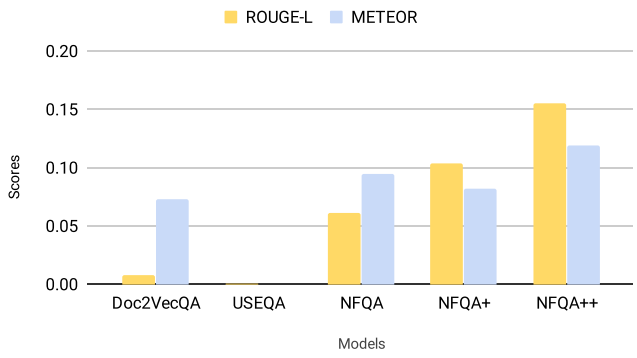

Fig. 5. CLEF-IP 2011 dataset – Average METEOR and ROUGE-L scores by models on long abstracts.

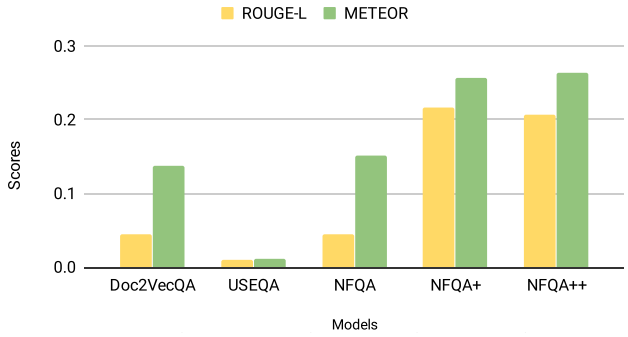

Fig. 6. Google dataset – Average METEOR and ROUGE-L scores by models on short abstracts.

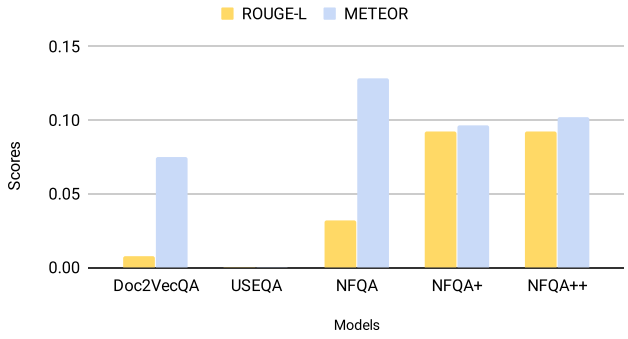

Fig. 7. CLEF-IP 2011 dataset – Average METEOR and ROUGE-L scores by models on short abstracts.

methods on abstracts with specialized jargon. Interestingly, the average performance of all the proposed frameworks (i.e., NFQA, NFQA+ and NFQA++) is higher in terms of ROUGE-L and METEOR compared to their performance on the other two types of abstract categories evaluated in the previous sections. This verifies the fact that the proposed methods are resilient against specialized words and return meaningful answers for a given query with special words. It can be seen that once again, the models proposed in this paper performed well in this regard, with NFQA++ being the top-performing model. Note that, the performance of the baselines are significantly lower than the proposed methods, while Doc2VecQA works better than USEQA in terms of both performance measures.

As Fig. 9 shows, NFQA+ performs best in terms of ROUGE-L on CLEF-IP 2011, while NFQA++ surpasses the other models in METEOR. The proposed frameworks performed well in this regard, with NFQA++ being the top-performing model when looking at the coherence of both performance metrics. This suggests that our NFQA++ model, with multiple question embeddings and topic-based filtering, takes specialized jargons into consideration. The performance of the baselines are significantly lower than the proposed methods in terms of ROUGE-L, while Doc2VecQA works better than USEQA in terms of both performance measures and its METEOR score is slightly lower than the proposed frameworks.

4.8. Evaluations on joint abstracts

The other common type of contents in legal domain is joint abstracts. These types of abstracts mainly represent the contents with non-coherence in the language. This is common type of patents in business when a technology is adapted in a particular domain (e.g., Blockchain in

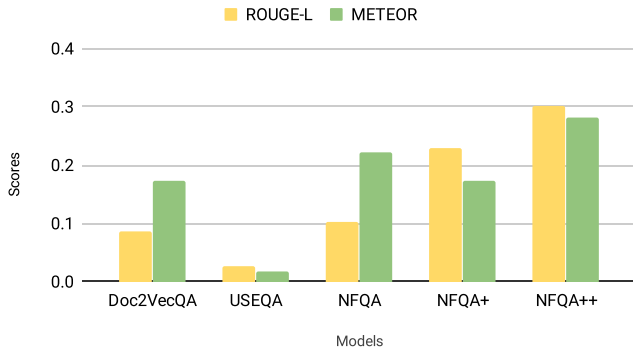

Fig. 8. Google dataset – Average METEOR and ROUGE-L scores by the models on specialized jargon abstracts.

retail). Such abstracts do not have consistency in their content and they might represent the information from different domains. In this section, we present the performance of the proposed frameworks on this type of abstracts. Fig. 10 shows the results in terms of ROUGE-L and METEOR. NFQA++ outperforms other methods significantly in terms of both measure. Although topic modeling seems to be a reasonable addition to improve the performance of the QA system, the results show that having topic modeling without detailed analysis of content would not be that effective. As shown NFQA++ achieves high scores due to the sentence-by-sentence analysis of the input query against the database. The three proposed models seem to address this component of the test well, with NFQA++ surpasses the ROUGE-L and METEOR scores for the other models by almost 50%.

As Fig. 11 illustrates, NFQA+ performs the best in METEOR scores in this category on CLEF-IP 2011, indicating that better results are retrieved when looking at a question as a whole in this category. In terms of both measures, NFQA+ and NFQA++ perform well coherently. In terms of ROUGE-L, the performance of the models increases with vectorizing queries and applying topic-modeling. This indicates that in abstracts that cover multi-domain language, our NFQA++ model performs relatively well.

4.9. Overall performance and discussions

This section discusses the results of the experiments performed to evaluate the models regardless of the type of the abstracts. We show that the three proposed models surpass the baseline models, Doc2VecQA and USEQA in both the ROUGE-L and METEOR scores. Our NFQA++ model performs the best amongst our three proposed models, with the highest scores in both the METEOR and ROUGE-L metrics. This is followed by

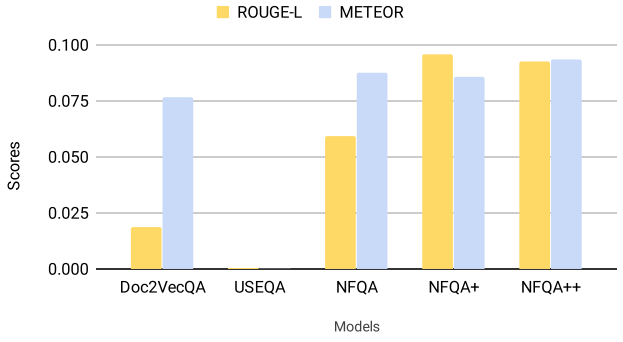

Fig. 9. CLEF-IP 2011 dataset – Average METEOR and ROUGE-L scores by models on specialized jargon abstracts.

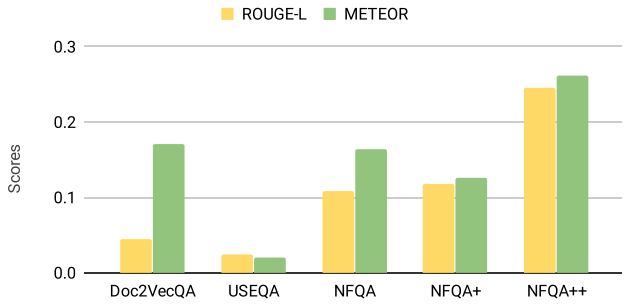

Fig. 10. Average METEOR and ROUGE-L scores by models on joint abstracts in Google dataset.

the NFQA+ model, which performs the second best, and the NFQA followed closely. For the METEOR metric, NFQA+ surpasses NFQA's score, which indicates that the former model performed better in regard to the semantic nature of language, whereas NFQA+ and NFQA++ performed similarly in the METEOR metric. On the other hand, the USEQA model significantly underperformed in both ROUGE-L and METEOR, when compared to the other models used in this paper.

Fig. 12 shows the overall performance of the methods in terms of ROUGE-L. The ROUGE-L metric is mainly used to test how powerful and robust the models are in terms of the syntactic structure of natural language since it considers n-gram and longest common sequences, which essentially deals with word order in language. Our NFQA++ model proved to be sufficiently robust to take syntax into consideration, with having the highest score in three out of the four categories, with NFQA slightly surpassing NFQA+ by a score of 0.0096 in the Short Abstract Category.

Moreover, Fig. 13 shows the overall performance of the methods in terms of METEOR. The METEOR metric is chosen to measure the performance of the semantic aspect of our models, since it considers synonymy and root words in natural language. Across all four categories, NFQA++ surpasses the other models on average, indicating that the model addresses the synonymy in language better than not only our other two proposed models, as well as the baseline models.

Irrespective of the type of the abstracts, on average, our three proposed frameworks perform well. Fig. 14 and Fig. 15 show the results of different methods on CLEF-IP 2011. In one of the two metrics, ROUGE-L, the three frameworks significantly surpass the Doc2VecQA as well as USEQA models. The NFQA++ model performs the best amongst our three proposed models, with the highest scores in ROUGE-L metrics. This is followed by the NFQA+ model, which performs the second best, and the NFQA follows closely.

For the METEOR metric, NFQA++ surpasses the scores of NFQA and NFQA+, which indicates that the former model performs better in

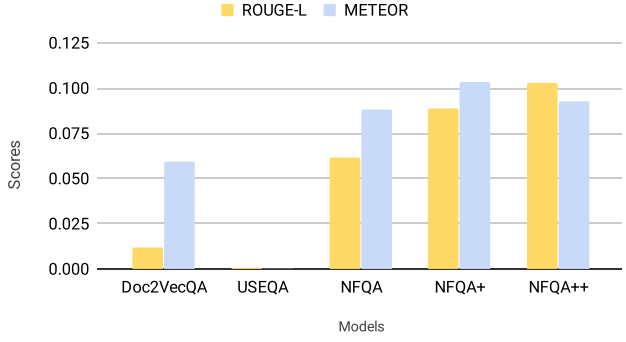

Fig. 11. Average METEOR and ROUGE-L scores by models on joint abstracts in CLEF-IP 2011.

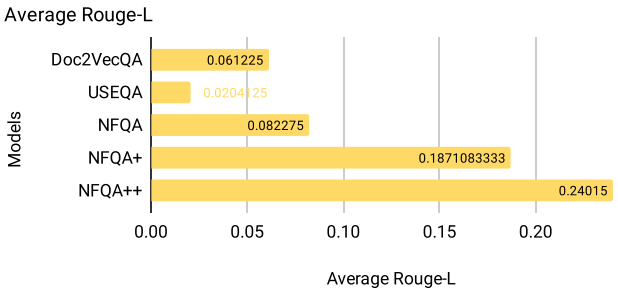

Fig. 12. Average ROUGE-L scores for different methods on Google dataset.

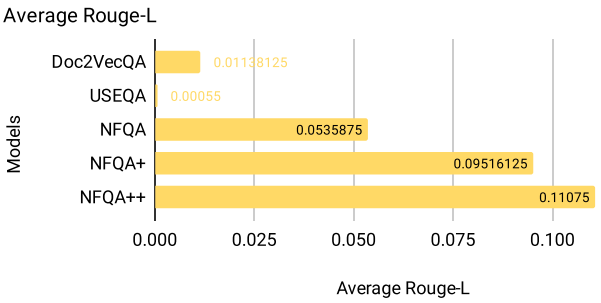

Fig. 14. Average ROUGE-L scores for different methods on CLEF-IP 2011.

regard to the semantic nature of language, whereas NFQA+ and NFQA performed similarly in the METEOR metric. On the other hand, the USEQA model significantly underperformed in both ROUGE-L and METEOR, when compared to the other models used in this paper. The ROUGE-L metric is mainly used to test how powerful and robust the models are in terms of the syntactic structure of natural language since it considers n-gram and longest common sequences, which essentially deals with word order in language. Our NFQA++ model proved to be sufficiently robust to take syntax into consideration, with having the highest score in three out of the four categories, with NFQA+ slightly surpasses NFQA++ in the Joint Abstract Category in terms of METEOR.

The METEOR metric is chosen to measure the performance of the semantic aspect of our models, since it considers synonymy and root words in natural language. Across all four categories, NFQA++ surpasses the other models on average, indicating that the model addresses the synonymy in language better than not only our other two proposed models, as well as the baseline models.

4.10. Sensemaking case study

Fig. 16 shows an example of the input abstract, and the generated answer by NFQA. Given the sentences in the answer, the average similarity of each article is also reported in this figure. In this example, the number of retrieved sentences for topic modeling is set to 40 and the number of sentences in the final answer is set to 4. Note that the answer is a combination of the most relevant sentences to the input query while covering different topics represented in the input query.

To evaluate the effectiveness of the generated answers from a user point-of-view, we conduct a user study. The generated answers of five abstracts are given to ten experts. We ask each to evaluate the quality of

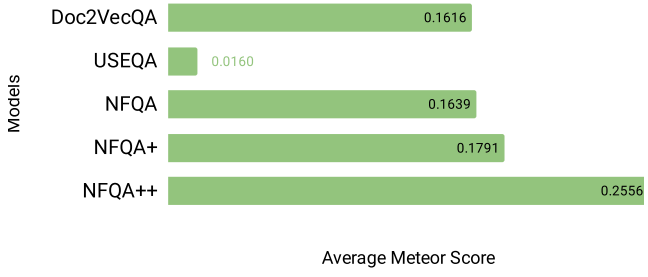

Fig. 13. Average METEOR scores for different methods on Google dataset.

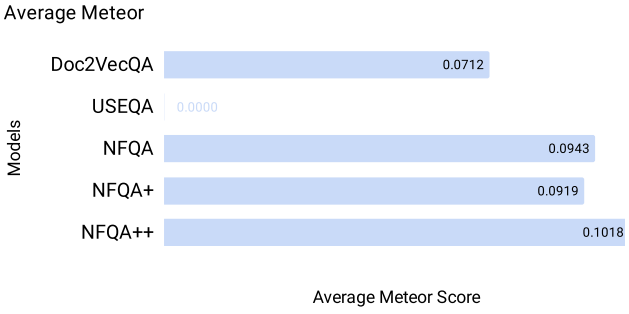

Fig. 15. Average METEOR scores for different methods on CLEF-IP 2011.

the generated answers by searching the keywords in each abstract in Google patent repository and then give a score between zero to one to rate how effective the generated answer is with respect to the given patent. Table 3 shows the average results in percentage. The results verify that the answers generated by NFQA++ are well-accepted by the experts in the user study.

5. Conclusion, limitations and future works

In this study, we proposed a question answering-like System for patent retrieval to help lawyers better engage in prior art search and aid in Sensemaking. With the use of Bidirectional Pre-Trained Embedding and topic modeling, we facilitated an effective way for word order, semantics and the thematic nature of Natural Language to be considered in Patent Research, as current methods disregard these areas. Our three proposed models, NFQA, NFQA+, and NFQA++ showed to surpass the baseline models significantly in terms of different performance measures. Ultimately, our most robust model is the NFQA++ model, which has the highest ROUGE-L and METEOR scores amongst all the models. This is an indication that we have dealt with the semantics and syntax of language in this framework as this model performed well across the different types of long, short, joint and jargon-filled question abstracts.

We acknowledge that our framework has some limitations. The topic of prior art search is a very broad topic and there are several directions that can be considered as the next steps of this study. User feedback plays an important role in prior art search. Each patent analyst has a unique opinion about the related patents to a given query. Therefore, we will expand the framework to take user feedback into account during the modeling and answer discovery. This results in a more interactive

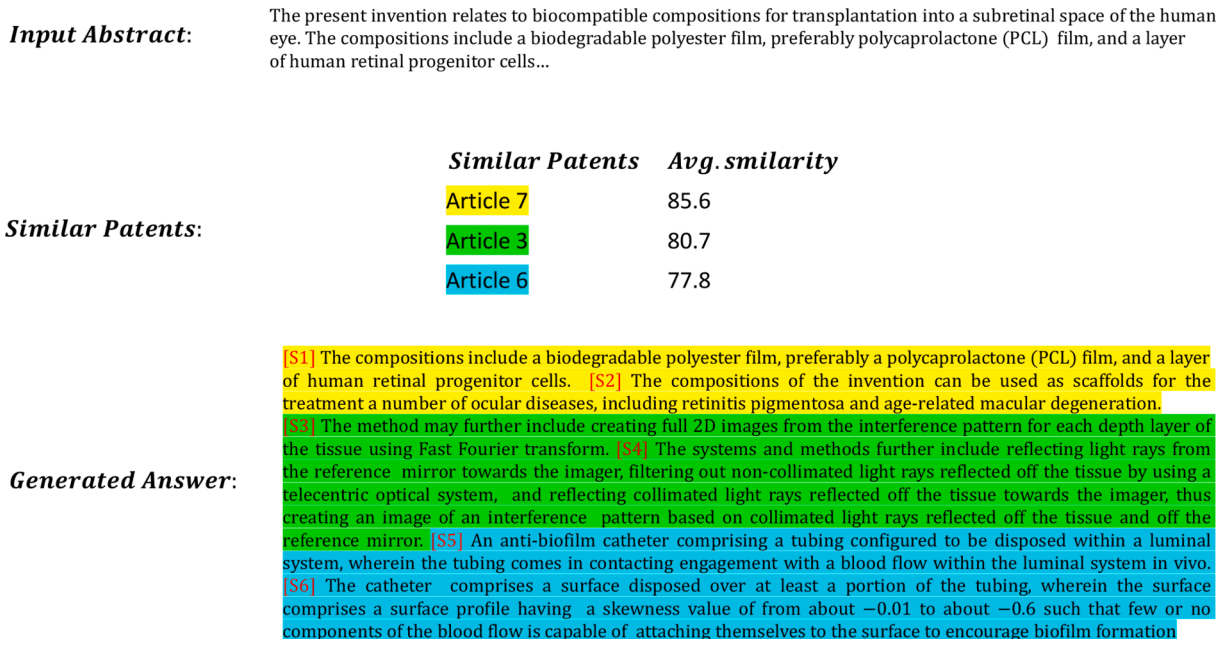

Fig. 16. An example of NFQA++.

Table 3
The average score of five generated answers in the user study.

Abstract	Possible Response Range	Actual Response Range	User Judge Mean (SD)	Avg Similarity (%)
A method of screening for breast cancer, including determining at least one first electrical …	0–100	45–100	68 (10)	82
The present inventors found that the natural flavor intrinsic to soybean is retained more than …	0–100	65–100	73 (9)	78
In various embodiments, lighting systems include an electrically insulating carrier having a plurality…	0–100	40–100	63 (18)	74
Systems and methods for imaging within depth layers of a tissue include illuminating light rays …	0–100	60–100	76 (12)	85
cA sending/receiving system includes first and second sending/receiving apparatuses…	0–100	55–100	71 (14)	79

system that is able to adapt to the factors presented by the user in his/her feedback. Active learning approach is one of the potential directions to improve the quality of answers based on users' feedback.

Furthermore, in this study, our focus was only on patents in English. One of the important future directions will be to investigate the effectiveness of the proposed framework in multilingual prior art search. We will also consider other neural embeddings, such as XLNet and GPT-2 to add to our framework. This will lead to a more comprehensive framework in terms of language understanding. In addition, testing will also be done to more categories of question abstracts to see how the models address other aspects of natural languages, such as polysemy and ambiguity. Last but not least, we will investigate the role of external resources such as knowledge graphs to represent patent contents more effectively for the task of prior art search.

CRediT authorship contribution statement

Morteza Zihayat: Supervision, Conceptualization, Methodology, Software. **Rochelle Etwaroo:** Writing - original draft, Investigation, Validation, Visualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported in part by Natural Sciences and Engineering Research Council of Canada (NSERC) under Discovery Grant (RGPIN-2018–05041, PI: Morteza Zihayat). The authors would like to acknowledge the efforts of Dr. Mehdi Kargar for his support and inputs during this study.

Appendix A. Appendix

The following subsections represent test case question examples (Abstracts) used in our experiments.

A.1. Case 1: Long Abstract

1. A document box includes a top container portion and a bottom container portion. The top container portion has a rear wall with a hollow rib integrally disposed thereon. The bottom container portion is rotatably coupled to the top container portion about an axis of rotation, and the top container portion is positionable between a closed position and an open position on the bottom container portion. The bottom container portion having a rear wall and a bottom wall. A raised member is positioned on the bottom wall adjacent to the rear wall. The top container portion where rotated to the open position releasably positions the raised member within the hollow recessed rib, thereby defining a hold open feature for the top container portion.
2. The present invention discloses a contact structure of a low-voltage electrical apparatus. The contact structure is in a dual-breakpoint form, and comprises: two U-shaped static contacts, the U-shaped static contact enabling the current direction in the static contact to be opposite to the current direction in a movable contact; a contact bridge; two movable contacts, disposed on the contact bridge, and respectively corresponding to the two static contacts; a contact support member, disposed on the movable contacts and connected to the movable contacts; two main contact springs, symmetrically disposed under the movable contacts and forming an angle with the contact bridge; and a spring support member, disposed under the two movable contacts and connected to the two main contact springs. At a contact position of the static contact and the movable contact and at a repulsed open position of the static contact and the movable contact, the angle between the main contact spring and the contact bridge is between $-\beta$ and $+\alpha$.
3. Alveolar macrophages contribute to host defenses against influenza. Enhancing their function contributed to protection against influenza and other acute lethal pulmonary infections. Wild-type mice and Tg mice expressing GM-CSF in the lung were infected with influenza virus, and lung pathology, weight loss and mortality were measured. GM-CSF was also administered to lungs of wild-type mice that were infected with influenza virus. All Tg mice expressing GM-CSF in the lungs survived with greatly reduced weight loss and lung injury and histologic evidence of a rapid host inflammatory response that controlled infection vs. wild-type mice not expressing GM-CSF in the lungs. This resistance to influenza was abrogated by elimination of alveolar phagocytes, but not by depletion of T cells, B cells or neutrophils. Tg mice had far more alveolar macrophages than wild-type mice and were more resistant to influenza-induced apoptosis. Delivery of intranasal GM-CSF to wild-type mice also conferred influenza resistance. Therefore, GM-CSF confers resistance to influenza by enhancing innate immune mechanisms that depend on alveolar macrophages. Pulmonary delivery of GM-CSF is therefore useful for reducing the significant morbidity and mortality due to influenza virus and is similarly useful in pulmonary infection caused by other infectious viral and bacterial agents.

A.2. Case 2: Short abstract

1. The present invention provides a drought tolerance associated protein, DT1, a nucleic acid molecule encoding the DT1 protein and application thereof.
2. A night light supported by a toilet seat assembly for directing illumination toward a toilet bowl.
3. The subject invention includes methods and plants for controlling European corn borer, said plants comprising a Cry1Ab insecticidal protein and a DIG-3 insecticidal protein to delay or prevent development of resistance by the insect.

A.3. Case 3: Combination of different abstracts joined together

1. A downhole material and a soluble glass dispersed within the material. A method for operating in a borehole. The improved cable system uses a preferred aluminum braided cable to provide low resistance combined with resistance to sour fluids. A new spring assembly increased tension and extends cable life by minimizing slack in the cable. A raised member is positioned on the bottom wall adjacent to the rear wall. The registration includes associating a subscription identification of a cellular services plan with an identification of the electronic device, from which a registration server may create an account associated with the cellular-based communications. The stream of beacons is broadcast over a wireless communication channel to mobile devices within range. A list of broadcasted beacons is stored in a table along with a time and location of broadcast. Subsequent to broadcasting, a stream of beacons is detected.
2. Embodiments of a mattress ventilation foundation and sleep system are disclosed. Foundation embodiments typically may be configured to provide ventilation to a supported mattress, for example through an upper surface of the foundation allowing airflow therethrough. Embodiments are provided for registering an electronic device of a subscriber for cellular-based communications. According to certain aspects, the cellular-based communications may be facilitated by a data center while the electronic device is not easily within range of a cellular network, such as when the electronic device is traveling on an aircraft.The tape cartridge includes a tab portion which is gripped at least at the time of unloading. The tab portion is arranged at or near a position where pull-out resistance due to unloading is balanced within a plane intersecting with an unloading direction.
3. Techniques involving gestures and other functionality are described. In one or more implementations, the techniques describe gestures that are usable to provide inputs to a computing device. A variety of different gestures are contemplated, including bimodal gestures (e.g., using more than one type of input) and single modal gestures. A tape cartridge or the like which can be pulled out of a cartridge loading section while maintaining a loading posture is to be provided. A tape cartridge is configured to be able to be loaded in and unloaded from a tape printing device. Additionally, the gesture techniques may be configured to leverage these different input types to increase the amount of gestures that are made available to initiate operations of a computing device.

A.4. Case 4: Abstracts containing numbers and special characters

1. A multi-part kit system comprising (i) a solid part A which comprises 10 to 80 wt.% of peroxy compound selected from the group consisting of KHSO 5, K 2 S 2 O 8, Na 2 S 2 O 8, magnesium monoperoxyphthalate hexahydrate, sodium percarbonate and sodium perborate, 0.1 to 10 wt.% of LiCl, NaCl and/or KCl and 1 to 20 wt.% of H 2 N(CH 2) n SO 3 H with n = 0, 1, 2 or 3, and (ii) a liquid part B in the form of an aqueous solution

which comprises 0 to 20 wt.% of nonionic surfactant, 3.6 to 20 wt.% of amphoteric surfactant and 0.5 to 20 wt.% of at least one compound comprising substituted ammonium selected from the group consisting of dihydrocarbyl dimethylammonium chlorides or bromides, didecyl methyl-poly(oxyethyl) ammonium propionate, chlorhexidine gluconate, cetylpyridinium chloride or bromide, and polyhexamethylene biguanide hydrochloride, wherein at least one of the two hydrocarbyl residues comprises 8 to 18 carbon atoms.

2. Two polymorphic forms ofbis[(E)-7-[4-(4-fluorophenyl)-6-isopropyl-2-[methyl(methylsulfonyl) amino]pyrimidin-5-yl](3R,5S)-3,5-dihydroxyhept-6-enoic acid] calcium salt, processes for making them and their use as HMG Co-A reductase inhibitors are described.
3. The aim of the invention is to improve the detergent power of solid textile detergents having a pH value ranging from 4 to under 10 in a solution containing 1 wt.% in demineralized water at 20 °C. Said aim is achieved by using a protease which comprises an amino acid sequence that is at least 80% identical to the amino acid sequence set forth in SEQ ID NO. 1 and contains, at position 99 of SEQ ID NO. 1, the amino acid glutamic acid (E) or aspartic acid (D), or the amino acid asparagine (N) or glutamine (Q), or the amino acid alanine (A) or glycine (G) or serine (S).

References

Andersson, L., Rekabsaz, N. and Hanbury, A. (2017) Automatic query expansion for patent passage retrieval using paradigmatic and syntagmatic information. In The first WiNLP workshop will be co-located with ACL.

Bandyopadhyay, A., Ganguly, D., Mitra, M., Saha, S. K., & Jones, G. J. (2018). An embedding based ir model for disaster situations. *Information Systems Frontiers, 20* (5), 925–932.

Banerjee, S. and Lavie, A. (2005) Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization (pp. 65–72).

Bhavnani, S., Clarkson, G., and Scholl, M. (2008) Collaborative search and sensemaking of patents. In CHI'08 Extended Abstracts on Human Factors in Computing Systems (pp. 2799–2804).

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003) Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022.

Carterette, B., & Chandar, P. (2009). Probabilistic models of ranking novel documents for faceted topic retrieval. In *In Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 1287–1296).

Cer, D., Yang, Y., Kong, S.-Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. (2018) Universal sentence encoder. arXiv preprint arXiv:1803.11175, 2018.

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

Fujii, A. (2007). Enhancing patent retrieval by citation analysis. In *In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 793–794).

Goldberg, Y. and Levy, O. (2014) word2vec explained: deriving mikolov et al'.s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.

Golestan Far, M., Sanner, S., Bouadjenek, M. R., Ferraro, G., & Hawking, D. (2015). On term selection techniques for patent prior art search. In *In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 803–806).

Harabagiu, S., Moldovan, D., Pasca, M., Mihalcea, R., Surdeanu, M., Bunescu, R., Girju, R., Rus, V., & Morarescu, P. (2000). Falcon: Boosting knowledge for answer engines. *TREC, 9*, 479–488.

Helmers, L., Horn, F., Biegler, F., Oppermann, T., & Müller, K.-R. (2019). Automating the search for a patent's prior art with a full text similarity search. *PloS one, 14*(3).

Hofstätter, S., Rekabsaz, N., Lupu, M., Eickhoff, C., & Hanbury, A. (2019). In *Enriching word embeddings for patent retrieval with global context. In European Conference on Information Retrieval* (pp. 810–818). Springer.

Hristidis, V., Ruiz, E., Hernández, A., Farfán, F., & Varadarajan, R. (2010). Patentssearcher: a novel portal to search and explore patents. In *In Proceedings of the 3rd international workshop on Patent information retrieval* (pp. 33–38).

Jia, Z., Abujabal, A., Saha Roy, R., Strötgen, J., & Weikum, G. (2018). Tequila: Temporal question answering over knowledge bases. In *In Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (pp. 1807–1810).

Jurafsky, D. (2000). *Speech & language processing*. Pearson Education India.

Jürgens, J. J., & Womser-Hacker, C. (2014). *Limitations of automatic patent ir. Datenbank-Spektrum, 14*(1), 5–17.

Khode, A., & Jambhorkar, S. (2017). A literature review on patent information retrieval techniques. *Indian Journal of Science and Technology, 10*(36), 1–13.

Kim, Y., & Croft, W. B. (2015). Improving patent search by search result diversification. In *In Proceedings of the 2015 International Conference on The Theory of Information Retrieval* (pp. 201–210).

Kim, Y., Seo, J., & Croft, W. B. (2011). Automatic boolean query suggestion for professional search. In *In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (pp. 825–834).

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *In International conference on machine learning* (pp. 1188–1196).

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *ProceedingsofWorkshop on Text Summarization Branches Out, Post2Conference Workshop of ACL*.

Loginova, E., Varanasi, S., & Neumann, G. (2020). Towards end-to-end multilingual question answering. *Information Systems Frontiers*, 1–15.

Marrara, S. and Pasi, G. (2015) Flexibility in patent search. In 2015 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology (IFSA-EUSFLAT-15). Atlantis Press.

May, C., Wang, A., Bordia, S., Bowman, S. R., and Rudinger, R. (2019) On measuring social biases in sentence encoders. arXiv preprint arXiv:1903.10561.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR), 33*(1), 31–88.

Piroi, F., Lupu, M., Hanbury, A., and Zenz, V. (2011) Clef-ip 2011: Retrieval in the intellectual property domain. In CLEF (notebook papers/labs/workshop).

Qu, C., Yang, L., Croft, W. B., Scholer, F., & Zhang, Y. (2019). Answer interaction in non-factoid question answering systems. In *In Proceedings of the 2019 Conference on Human Information Interaction and Retrieval* (pp. 249–253).

Sales, J. E., Freitas, A., Handschuh, S., & Davis, B. (2015). Linse: A distributional semantics entity search engine. In *In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1045–1046).

Samarinas, C., & Tsoumakas, G. (2018). Wamby: An information retrieval approach to web-based question answering. In *In Proceedings of the 10th Hellenic Conference on Artificial Intelligence* (pp. 1–8).

Saraswat, N., Verma, I., & Gupta, V. (2019). Catch-phrase based document representation for improved prior art search. In *In Proceedings of the ACM India Joint International Conference on Data Science and Management of Data* (pp. 210–216).

Shalaby, W., & Zadrozny, W. (2019). *Patent retrieval: a literature review. Knowledge and Information Systems* (pp. 1–30).

Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G. (2014). Learning semantic representations using convolutional neural networks for web search. In *In Proceedings of the 23rd International Conference on World Wide Web* (pp. 373–374).

Sidorov, G., Gelbukh, A., Gómez-Adorno, H., & Pinto, D. (2014). Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas, 18*(3), 491–504.

Srinivasa-Desikan, B. (2018). *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Birmingham, UK: Packt Publishing Ltd.

Tang, J., Li, C., Zhang, M., and Mei, Q. (2016) Less is more: Learning prominent and diverse topics for data summarization. arXiv preprint arXiv:1611.09921.

Vikraman, L., Croft, W. B., & O'Connor, B. (2018). Exploring diversification in non-factoid question answering. In *In Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval* (pp. 223–226).

Xu, J., Licuanan, A. and Weischedel, R. M. (2003) Trec 2003 qa at bbn: Answering definitional questions. In TREC (pp. 98–106).

Xue, X., & Croft, W. B. (2009). Automatic query generation for patent search. In *In Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 2037–2040).

Yu, J., Mohan, S., Putthividhya, D., & Wong, W.-K. (2014). Latent dirichlet allocation based diversified retrieval for e-commerce search. In *In Proceedings of the 7th ACM international conference on Web search and data mining* (pp. 463–472).

Zhang, D., & Lee, W. S. (2003). A web-based question answering system. *Massachusetts Institute of Technology (DSpace@MIT)*.