Query Oriented Extractive-Abstractive Summarization System (QEASS)

K.Girthana
Research Scholar, Dept. of Information Science &
Technology, Anna University, CEG Campus
Chennai, TamilNadu, India
keerthi3110@gmail.com

S. Swamynathan
Professor, Dept. of Information Science & Technology,
Anna University, CEG Campus
Chennai, TamilNadu, India
swamyns@annauniv.edu

ABSTRACT

This work proposes a query oriented extractive-abstractive summarization system where the query is synthesized and expanded from the novel details provided by the patent analyst and the domain ontology. Since the search and patent document retrieval using the formulated semantic query alone will not satisfy the user requirements, this work filters and summarizes the retrieved document set both extractively and abstractively. Summarization makes use of deep learning techniques as their structure mimics the human brain. The proposed work was evaluated using Google patent dataset. The retrieval results of semantic query expansion using domain ontology are compared with Google Prior-art search query results and WordNet based semantic query expansion retrieval sets. The summarization results of the retrieved document sets are compared with the human summaries.

CCS CONCEPTS

• **Information systems → Ontologies**; **Query reformulation**; **Summarization**;

KEYWORDS

Domain ontology, Query Expansion, Document ranking, Summarization

ACM Reference Format:
K.Girthana and S. Swamynathan. 2019. Query Oriented Extractive-Abstractive Summarization System (QEASS). In *6th ACM IKDD CoDS and 24th COMAD (CoDS-COMAD '19), January 3–5, 2019, Kolkata, India.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3297001.3297046

1 INTRODUCTION

Innovation is the key to sustainable growth of a country and Intellectual Property Rights (IPR) mainly patents plays a great role in the economic development and securing innovations in a country. Due to increase in number of patent filings and non-patent literature, prior-art search on patents has a pivotal role in patent grant process. Earlier works are either keyword-based or uses domain independent knowledge base (Wikipedia, WordNet, Wikitionary) or combination of both along with other patent metadata [1, 6]. The relevant document retrieval was marginal and the problem with these approaches is some of the relevant documents are missed or more number of irrelevant documents retrieved. This Query Oriented Extractive-Abstractive Summarization System (QEASS) improves the retrieval performance through expansion of prior-art search query with domain ontology (domain dependent knowledge base). Earlier works also uses International Patent Classification (IPC) definitions as domain knowledge base and performed expansion [7]. Since, the definitions were not well defined for all domains; this method is restricted to few domains. Also the relevant document listing alone will not be useful to the patent analyst. It is because of the patent document length (around 28-32 pages). Moreover, it will be difficult to find the relevant parts of each document about the search topic. Therefore, representative sentences can be generated and condensed into a summary that best represents the search topic in the document. Most of the works on document summarization uses text mining approaches along with statistical and linguistic analysis and are extractive in nature [3–5]. Redundancy and coherence are common problems that exist with multi-document extractive summarization. But, abstractive summarization though difficult, overcomes extractive summary problems through interpretation and mapping concepts to key phrases using Natural language processing techniques. Recent works on text summarization suggest that Deep Learning [2, 10] achieves impressive results in generating headlines for news articles (Text generation), speech recognition, Machine Translation, and Document Summarization. Very few works on summarization focus on patent document genre. Some have analyzed the patent structure and used template based document summarization method [8, 9]. To overcome the above said problems, the QEASS constructs a domain ontology that better describes the domain concepts and their relation and uses this domain ontology for efficient retrieval and processing of patent documents. Also summarizes the retrieved patent documents through Deep learning approaches mainly Restricted Boltzmann Machine (RBM). RBM is a probabilistic approach with a visible layer and a hidden layer. The rest of the paper is organized with the detailed methodology of QEASS in section 2, and section 3 evaluates the system and compares the efficiency of proposed semantic query expansion and document processing with Google prior-art search system and WordNet query expansion system. Finally section 4, concludes the system with possible future enhancements.

CoDS-COMAD '19, January 3–5, 2019, Kolkata, India

ACM ISBN 978-1-4503-6207-8/19/01.
https://doi.org/10.1145/3297001.3297046

2 QUERY ORIENTED EXTRACTIVE - ABSTRACTIVE SUMMARIZATION SYSTEM

Figure 1 illustrates the flow of the Query oriented Extractive - Abstractive Summarization System (QEASS). It consists of mainly five components and are detailed in the following sub-sections.

2.1 Domain Ontology Construction

The domain ontology construction follows a knowledge engineering methodology which involves both domain experts and users. It acts as the knowledge base for efficient patent retrieval and processing. The ontology is conceptualized by considering smartdevice (smartphone, smartwatch and smarthome) related patent documents and technical specification documents. A glossary of terms identified through competency questions and domain documents are enhanced with their abbreviations, synonyms, types, components, and other related terms. The domain experts' analyses, filters, groups, and abstract them as concepts. The smartdevice domain ontology provides a specification of various hardware components such as processors, sensors and memory, software components (operating system, Middleware) and communication and display technologies. The relations are built within and between the concepts. The expressivity of the constructed smart device domain ontology is SROIQ (D) and is attained through the object and data properties, cardinality and a set of axioms. The smart device domain ontology is available at "github.com/Girthana/SD-ontology".

2.2 Initial Query Builder

The initial query builder extracts the candidate key-phrases and builds the initial query from the novel patent document or the innovative details provided by the patent analyst. The candidate key-phrases are extracted using Apache OpenNLP and TextRank. The system extracts the phrases of the form {(<JJ>* <NN.*>+ <IN>)? <JJ>* <NN.*>+} from the abstract , technical field and description of a patent document. The tagged components are Adjective (JJ), noun (NN) and a preposition (IN). The extracted phrases are scored based on the TF-IFF score. The TF-IFF is calculated as in equation 1.

$$tf - iff_{t,f,F} = \frac{n_{t,f}}{\sum_{n=1}^{k} n_{k,f}} \cdot \log \frac{|F|}{|1 + [f \epsilon F : t \epsilon f]|} \tag{1}$$

where $n_{t,f}$ is the number of occurrences of the term t in the field f; $\sum_{n=1}^{k} n_{k,f}$ is the size of the field f; $|F|$ is the cardinality of fields in the document and $[f \epsilon F : t \epsilon f]$ represents the field frequency.

The candidate key-phrases that occur in both approaches are given priority, and the key-phrases obtained are compared with the Google prior-art search query and Table 1 depicts the same for the sample application document "US5515043".

Table 1: Initial Query Builder Vs. Google Prior-art Search Query

Patent Document No.	Initial Query Key Terms/Phrases (Noun-phrase + PageRank)	Google Patent Prior-Art Search Query Key Terms
US5515043	vehicle, cellular phone, gps, alarm sensor, gps locating assembly, satellite-based, gps receiver ...	vehicle, alarm, apparatus, fig, power

2.3 Semantic Query Expander

This query expander expands the initial query formulated using domain-dependent knowledge base (smart device domain ontology) and domain independent knowledge base (WordNet).

Domain Ontology Query Expander. The initial query is composed of ten terms. For each term in the initial query, a set of concept terms are chosen based on the similarity with the concepts annotation properties and added to the candidate list for expansion. Then each candidate concept terms are further expanded in two levels. The first level expansion includes expansion with annotation properties and the data properties associated with the concept. While the second level involves retrieving the linked concepts for a candidate concept based on the path distance and the context of the initial query. The final list of expansion occurs based on the concept occurrence count in the patent document and for each term, the expansion list is restricted to three concepts. Table 2 explains the domain ontology query expansion at both levels for a sample query. The initial query is tokenized and the corresponding concepts are retrieved for the token from the domain ontology. As seen from the table, the concept retrieved for cellphone is *Smartphone* and it doesn't have any other annotation properties. But in case of Bluetooth, first level expansion with Abbrev annotation property yields *BT*, and *Radio Frequency* and *Frequency hopping spread spectrum* defined through its data properties. In the second level, the other concepts that are directly linked or at two hop distance with the concept are chosen. This includes ancestors, siblings, children and other concepts associated through object properties.

Table 2: Domain ontology query expansion example

Initial Query : Cellphone bluetooth querying device radio communication		
Query Terms / Phrases	**First level Expansion concepts**	**Second level Expansion concepts**
Cellphone	Smartphone	-
Bluetooth	BT , Radio Frequency, Frequency Hopping spread spectrum	Connectivity, Wireless communication, Bluetooth Antenna, Bluetooth transceiver, Wi-Fi, GPS, and 5 more concepts
Querying device	Bluetooth	Link Management Protocol, LMP, MLME, and 4 more concepts
Expanded Query: Cellphone or Smartphone Bluetooth or BT or Radio Frequency or Connectivity or Bluetooth Antenna or Bluetooth transceiver Querying device or Bluetooth and more concepts		

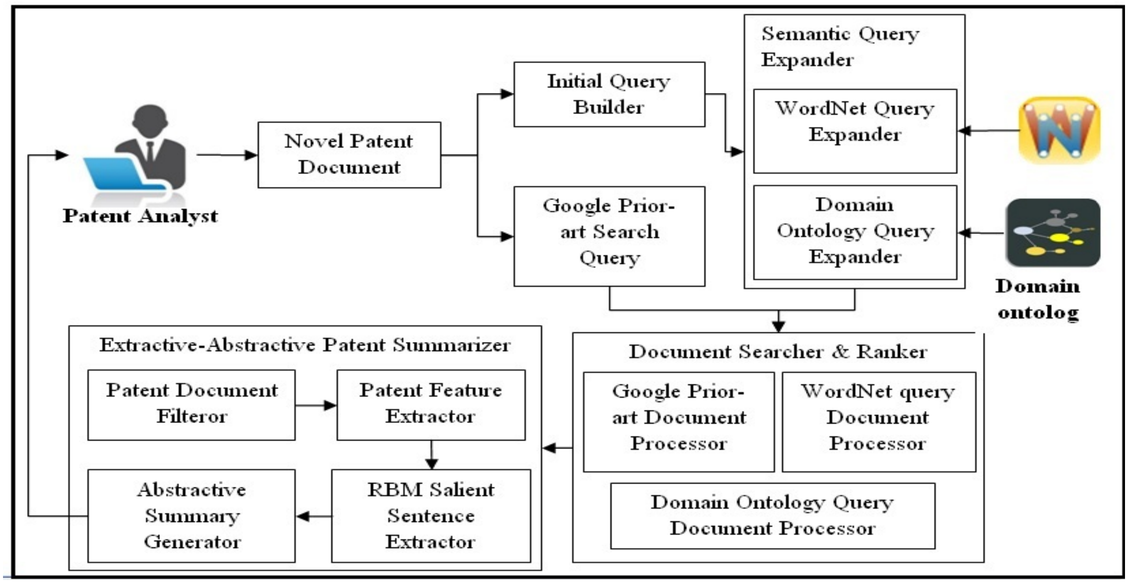
Patent Analyst
Novel Patent Document
Initial Query Builder
Google Prior-art Search Query
Semantic Query Expander
WordNet Query Expander
Domain Ontology Query Expander
Domain ontolog
Extractive-Abstractive Patent Summarizer
Patent Document Filteror
Patent Feature Extractor
Abstractive Summary Generator
RBM Salient Sentence Extractor
Document Searcher & Ranker
Google Prior-art Document Processor
WordNet query Document Processor
Domain Ontology Query Document Processor

Figure 1: QEASS Flow

Table 3: Comparison of prior-art search queries of different sources

Patent Title	**Google Patent Search**		**WordNet-based query expansion**		**Domain Ontology-based query expansion**	
	Prior-art Search query	**Total documents**	**Prior-art Search query**	**Total documents**	**Prior-art Search query**	**Total documents**
Antenna apparatus for smart phone	antenna body end part upper	27282	**Initial Query:** stretchable antenna foldable antenna smartphone			
			stretchable **stretchy elastic** antenna **aerial transmitting-aerial** foldable **foldaway folding collapsible** smart phone	533	stretchable antenna **cellular_antenna functional_antenna top or upper_end transmit receive** foldable antenna smart phone	312

WordNet Query Expander. The WordNet query expander reformulates the original query by adding new terms from the lexical knowledge base, WordNet. Since retrieving all the synsets associated with each term results in ambiguity, this work restricts to top three synsets of the same sense for semantic query generation to improve the precision and query response time.

2.4 Patent Document Retriever and Ranker

The patent document retriever and ranker compose three different processors according to the search query component. The Google prior-art document processor uses Google prior-art search query for search while the other two uses their corresponding semantically expanded query. The retrieved patent documents are re-ranked based on the cosine similarity with the source patent document and passed to the next module for further processing.

2.5 Extractive-Abstractive Document Summarizer

This summarizer analyses the retrieved patent document set and provides a gist of each document about the search topic.

Patent document filteror. The retrieved patent document set has multiple thousands of documents, and all the documents in this set need not be relevant. So, the document set is filtered with the International Patent Classification (IPC) code till sub-class or group level. IPC is a hierarchical classification system with multiple levels.

Patent Feature Extractor. It involves extracting features such as title and search query similarity (T_SQ_i) (Equation 2), sentence field position (Equation 3), and Term Frequency-Inverse Field Frequency (TF_IFF) (Equation 1). T_SQ_i is based on the number of common occurrences between the title (T) - sentence (sen_i) and search query (SQ) - sentence (sen_i). The notations T and SQ in Equation 2 represents the set of words in Title and Search Query, sen_i represents the

Table 4: Extractive Summarizer Metrics

Patent Title	Total No. of Sentences		Comp. Rate	RBM Salient Sentence extractor		TextRank	
	before	after		Precision	Recall	Precision	Recall
CN101916462A	101	78	23%	0.522	0.888	0.4484	0.299
CN103793833B	53	38	28%	0.518	0.633	0.668	0.547
CN104392501A	92	64	30%	0.318	0.551	0.076	0.012
CN105867321B	90	67	26%	0.386	0.801	0.594	0.529
CN106023332A	97	72	26%	0.531	0.512	0.569	0.235

set of words in an i^{th} sentence of a document and |T| and |SQ| represents the length of the title and search query respectively. Similarly, in Equation 3, i is the currently processed sentence index, and m is the total number of sentences considered for a document. These extracted features are mapped to form sentence-feature matrix.

$$T - SQ_i = \frac{|T \cap Sen_i| + |SQ \cap Sen_i|}{\log|T| + \log|SQ|} \tag{2}$$

$$SF_i = \begin{cases} 0.25, & \text{if } sen_i \text{ is at the beginning} \\ 0.75, & \text{if } sen_i \text{ is at the end} \\ max[i^{-1}, (m-i+1)^{-1}], & \text{otherwise} \end{cases} \tag{3}$$

RBM Salient Sentence Extractor. RBM is a generative model, best extracts the features (salient sentences) using the sentence feature matrix. The RBM visible layer encompasses three perceptrons for the features, and the hidden layer composes two perceptrons that describe whether the sentence is essential or not. The system is trained in an unsupervised way using contrastive divergence with a learning rate of 0.1 and with the inputs. Each sentence feature vector is passed to the hidden layer along with the learned weights and biases. Once learning is completed, it outputs binary values indicating whether the sentence is important or not. For each document, the sentences whose RBM output is one is included in the salient sentences set.

3 EXPERIMENTAL RESULTS

The experiments are conducted with the Smart device (smartphone, smarthome and smartwatch) patent document set collected from the Google patent search engine. It includes patent documents from various patent databases across the world. Around 900 patent documents are retrieved for each query submitted to the search engine through Google patent search API. Since the patent document is lengthy and only textual fields are considered for processing.

3.1 Query expansion using domain ontology

The performance of semantic query expansion using smart device domain ontology is validated by comparing with WordNet-based query expansion and Google prior-art search query. Table **??** depicts Google prior-art search query and semantically expanded query of WordNet and Domain ontology. The Google prior-art search query has generic terms that may increases retrieval of more irrelevant documents. Further expansion with the generic query will not yield better results. So the initial query is formulated, and used for semantic expansion. The document processor results of the three approaches are compared and evaluated in terms of Mean Average Precision (MAP) and recall by considering the top 100 documents. MAP and recall of document retrieval using Google prior-art search query , WordNet-based query expansion and Domain ontology based expansion are 0.142 and 0.234, 0.173 and 0.421, and 0.332 and 0.734 respectively. The improvement of domain ontology based query expansion was due to the number of search query terms and query term quality in both the system.

3.2 Extractive Document Summarizer

The sentence feature vector, when fed to the RBM, outputs the binary value for the sentences to be included in the summary. For this sample query, Table **??** depicts the summary statistics along with metrics for top 5 retrieved patent documents after filtering. It can be observed that the average compression rate of the summaries obtained using the proposed approach is around 30%.

4 CONCLUSION

The QEASS system summarizes the prior-art search patent documents using deep learning techniques in both extractive and abstractive manner. In order to improve the prior-art search results (patent documents), this system formulates an initial prior-art search query from the novel patent document and expands them using the smart device domain ontology. The proposed prior-art search query and its expansion were compared with Google prior-art search query and WordNet-based query expansion approaches. Better performance was achieved for domain ontology based query expansion than other methods in terms of recall. Also, the extractive summaries generated for each patent document has an compression rate of around 30% and when compared with that of human generated summaries had an average precision and recall of around 69% and 75% respectively. As a future work, this system plan to abstractively summarize the output of extractively summarizer using deep learning techniques.

ACKNOWLEDGMENTS

The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions. The work is supported by the Visvesvaraya PhD Scheme for Electronics and ITnder Proceedings No. : 3408/PD6/DeitY/2015.

REFERENCES

[1] Bashar Al-Shboul and Sung-Hyon Myaeng. 2014. Wikipedia-based query phrase expansion in patent class search. *Information retrieval* 17, 5-6 (2014), 430–451.

[2] Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015. Ranking with Recursive Neural Networks and Its Application to Multi-Document Summarization.. In *AAAI*. 2153–2159.

[3] Wesley T Chuang and Jihoon Yang. 2000. Extracting sentence segments for text summarization: a machine learning approach. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 152–159.

[4] Harold P Edmundson. 1969. New methods in automatic extracting. *Journal of the ACM (JACM)* 16, 2 (1969), 264–285.

[5] Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development* 2, 2 (1958), 159–165.

[6] Walid Magdy, Patrice Lopez, and Gareth JF Jones. 2011. Simple vs. sophisticated approaches for patent prior-art search. In *European Conference on Information Retrieval*. Springer, 725–728.

[7] Parvaz Mahdabi and Fabio Crestani. 2014. Patent query formulation by synthesizing multiple sources of relevance evidence. *ACM Transactions on Information Systems (TOIS)* 32, 4 (2014), 16.

[8] Amy JC Trappey and Charles V Trappey. 2008. An R&D knowledge management method for patent document summarization. *Industrial Management & Data Systems* 108, 2 (2008), 245–257.

[9] Amy JC Trappey, Charles V Trappey, and Chun-Yi Wu. 2009. Automatic patent document summarization for collaborative knowledge systems and services. *Journal of Systems Science and Systems Engineering* 18, 1 (2009), 71–94.

[10] Yong Zhang, Meng Joo Er, Rui Zhao, and Mahardhika Pratama. 2017. Multiview convolutional neural networks for multidocument extractive summarization. *IEEE transactions on cybernetics* 47, 10 (2017), 3230–3242.