PatentPal: Summarising and Retrieving Patents

Deepi Garg
IIIT-Delhi
India
2018389

Himanshi Mathur
IIIT-Delhi
India
2018037

Vanshika Goel
IIIT-Delhi
India
2018202

Vibhu Agrawal
IIIT-Delhi
India
2018116

ABSTRACT

With the advent of technology, inventions and innovations are on a rise. To protect their interests and gain recognition, inventors file patents. However, before filing patent, it is essential to have a brief summary of all the patents filed in the domain so far, in order to ensure that there is no overlap of patent filed. With the great influx of patents every year, it is essential to have a fast retrieval and summarising system for the same. Hence, we propose, PatentPal - which can summarise the patents filed, create a summary for a new patent and find similar patents, given another patent or appropriate keywords. The codebase can be found on `https://github.com/vibhu18116/IR-Project-Group23-PatentPal`

1. MOTIVATION

Inventorship is an art, proving you smart
But so many patents, where to start?

In the present world, the information is available at a click away. The resources thus provided and otherwise allow exploration at a large scale, resulting into innovations and inventions. The inventions need to be patented to claim the work, get recognition and obtain some materialistic rewards. However, it is essential to maintain uniqueness of the work done and carry forward the previous literature to ensure that the efforts made are in the best of "human interests" and to prevent "reinvention of wheel".

However, the process of filing a patent is a long and tedious one. One has to lookup if a similar work already exists, then draft a patent proposal and attach a summary to it. However, due to plethora of patents filed per year and the patents in force as illustrated in Figure 2, it is very difficult to look at every patent, and it is better to have a comprehensive summary rather than having to read the whole of it. The ratio of design patent applications filed, and accepted through various years is shown in Figure 3. It demonstrates a lot of patents are not accepted, where one of the possible reasons could be that the same or very similar work already exists. Similar statistics exist across other domains as well. Hence, it is critical to have an efficient retrieval system for existing patents.

Keeping such a scenario in mind, we intend to develop a system which can provide comprehensive summaries of the existing patents and to summarise any new patent based on the learning of present corpus. The system would also find similar patents already filed, given another patent, query keywords or a general query.

Figure 1: Depicting problems faced by inventors and researchers while filing patents

2. LITERATURE REVIEW

2.1 Summarisation of Patent Documents

Summarisation of documents differ from one domain to another domain. It is very difficult to have a general summariser that summarises legal as well as research papers well. Summarisation either works on word level or sentence level by finding their similarity and importance score in the document and corpus. Most of the work done so far focusses on word level. [1]

Trappey et al presented a 2 step approach to summarise patent documents. The first step extracts key words and phrases using two extraction algorithms. The first algorithm uses the TF-IDF concept clustering approach and the second one uses a specific ontology embedded in the system to calculate the frequency of mapping words. The second step involves summary generation which includes key phrase extraction, the creation of the paragraph and key word frequency matrix, the computation of paragraph importance, document concept clustering, and the building of the visualization summary tree. [2]

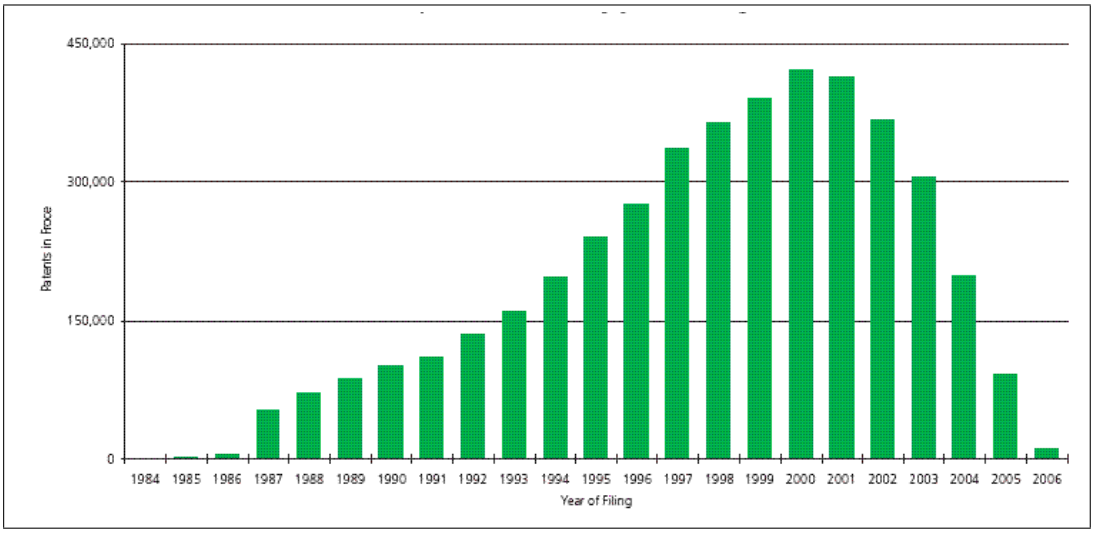

Figure 2: Number of patents in force by year of filing, 2006. The graph does not contain data for the Japan Patent office and the State Intellectual Property Office of China. Source: WIPO Statistics Database (`https://www.wipo.int/ipstats/en/statistics/patents/wipo_pub_931.html`)

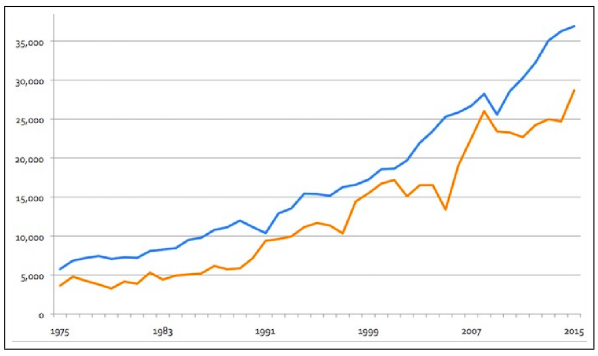

Figure 3: Number of design patent applications filed vs number of patents issued per year. Blue line indicates Applications Filed and orange line indicates Patents Issued (`https://www.ipwatchdog.com/2016/09/10/design-patents/id=72714/`)

Hu et al proposed "patent keyword extraction algorithm (PKEA)" to extract keywords from patent documents for summary generation. The method applies a skip-gram model to the pre-trained word embeddings to obtain a table, which is then used to generate the centroid vector using k-means algorithm. The cosine similarity between the centroids and candidate keywords is calculated to extract the top n important keywords. [3]

A graph-based sentence ranking method was put forth by Yeh to produce document summaries. The document is modelled into a network of sentences connected on the basis their lexical overlap. A feature profile is created using three surface level features, namely centroid, position and first sentence overlap. The sentence similarity network and feature profile are used to rank the sentences by applying the spreading activation theory (Quillian, 1968). The summary is generated by adding the top ranking non-redundant sentences, which are then ordered chronologically. [4]

2.2 Retrieval of Related Patents

The quick retrieval of the documents is also paramount to the success of project. Sharma et al applied a semantic expansion technique to enhance the patent search process. This technique incorporates external sources for expansion. The method first filters the abstracts on the basis of IPC using TF-IDF and generates a keyword vector from the abstracts. It then expands the queries and constructs the vector space using external sources like WordNet and Wikipedia. It finally calculates the document similarity using cosine similarity and extended Jaccard Coefficient. [5]

Helmers et al used features like high dimensional sparse bag-of-words (BOW) vectors with tf-idf, neural network language models (NNLM), word2vec (combined with BOW vectors) or doc2vec instead of the keyword approach to find the cosine similarity between two patent documents. [6]

2.3 Evaluation Metrics

The summaries can be generated using extractive or abstractive method. Extractive method tries to figure out most important sentences and use them as it is in the summary. On the other hand, abstractive method tries to summarise the content by paraphrasing and retaining the original context. Here we intend to use abstractive method for summarisation as the dataset has well written human summaries associated with each patent which would act like gold-standard summaries. To evaluate the summaries produced by the model, we would use Rouge [7], which stands for Recall-Oriented Understudy for Gisting Evaluation. Its goal is to provide a measure of quality of an automatically generated summaries in comparison against a reference summary produced by humans.

To evaluate ranking and viability of the retrieved documents, similarity measure with thresholds is used. Precision, Recall and F1 score can be used to check the correctness and accuracy of the retrieved documents.

2.4 Novelty of Work

Most models which rely on a keyword-based search for summarisation as well as retrieval often produce suboptimal results. Bag of Words can fetch out patents with exact similar words without any regard of their meaning in the required context. We aim to club the approaches to obtain a context sensitive summary for the documents which would also retain words important to the flow and meaning of legal sense. Compiling it, we will embed them into a web-platform capable of summarising the patent documents well.

To the best of our knowledge, there is no existing platform to aid with writing summaries for the patents. Most existing platforms retrieve similar documents and present the human written summary from the same.

3. DATASET

Summarisation is a not so straightforward task as evaluation of summaries require comparison from well written human summaries. BigPatent [8] consists of 1.3 million records of U.S. patent documents along with human written abstractive summaries. The data has been collected from Google Patents Public Datasets using BigQuery. It contains patents filed after 1971 across nine different technological areas. We shall use each patent's abstract as the goldstandard summary and its description as the input. The dataset is publicly available on `https://evasharma.github.io/bigpatent/`.

Compared to existing summarization datasets, BigPatent has the following properties:

- Summaries contain richer discourse structure with more recurring entities
- Salient content is evenly distributed in the input
- Lesser and shorter extractive fragments present in the summaries

4. DATA EXPLORATION

The first step after obtaining the dataset was to understand it before proceeding to adopt summarization techniques and indexing. We plotted Word Clouds of summary and description for 1000 documents and found results as shown in Figure 4 and Fgure 5. Figure 4 has word cloud for abstract and 5 for the description of the chosen patents. There are a lot of common words between the two plots like "device", "portion", "include", "invention", "system", "body", "embodiment", etc. Figure 6 shows the distribution of the number of common words between abstract and description, which accounts for almost 30-60 words.

To see the compression ratio achieved by the summaries in the domain of patents, we plotted the number of words in

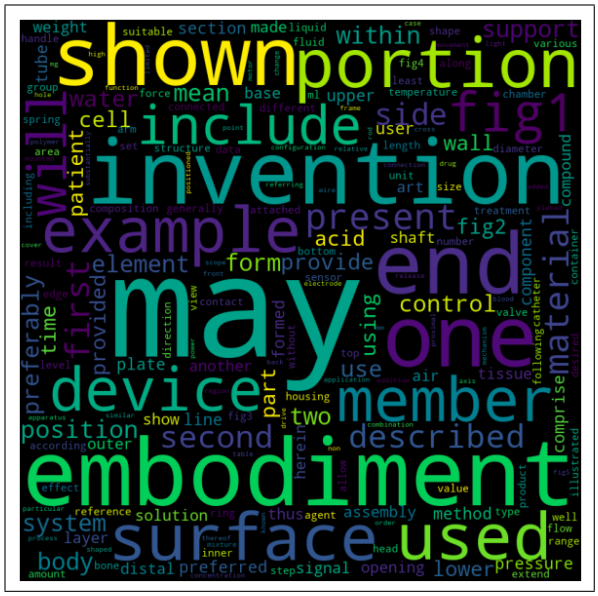

Figure 4: Wordcloud for abstract

the abstract to the number of words in the description ratio. We also did the same analysis on the sentence level. Most abstracts were compressed to a ratio of 1-5% in terms of the number of sentences and 2-6% in terms of the number of words. Figure 7 shows the ratio plot for sentences and Figure 8 shows the ratio plot for words.

To see the correlation between the article and summary length, we again plotted the regression curve between abstract and description as shown in Figure 9. We see the slope of the regression line is positive but the r-squared value is meagre 0.02% indicating that the relationship between the sentence length in abstract and description is very low.

The majority of descriptions have a length of about 200 sentences and abstracts have a length of about 2-3 sentences which is in accordance with the heuristic ratio found above. On average, an English sentence has 15-20 words including all the stopwords and supporting words. This makes about 35 words out of a 60-word abstract come straightforwardly from the description, showing scope for extractive summarization models to maybe perform well.

5. BASELINE MODELS

We constructed baseline models for summarization and retrieval of patent documents. Summarisation can be performed in an extractive or abstractive sense. For retrieval, we need to index the documents, preprocess the documents and queries and then build a logic for reteiving relevant documents.

5.1 Extractive Summary

Extractive summary figures out the key phrases and tokens and constructs a summary using them. The summary constructed has words and phrases from the description in ver-

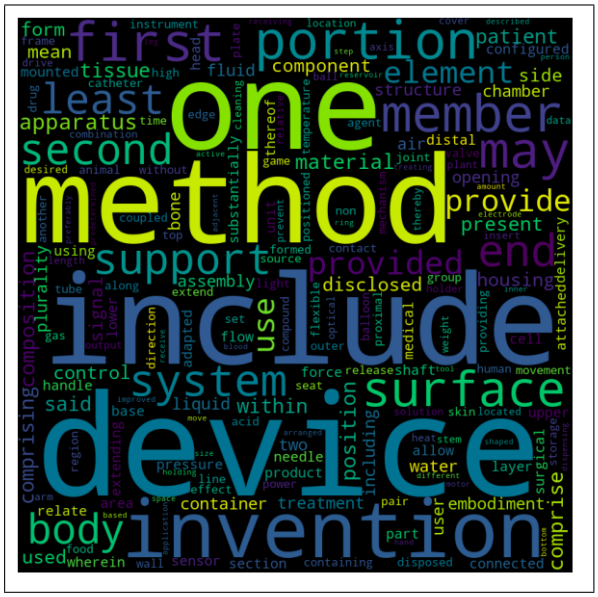

Figure 5: Wordcloud for description

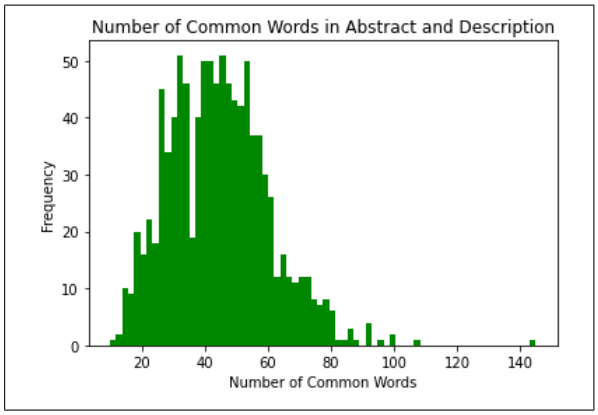

Figure 6: Number of common words between abstract and description

batim. The following are the baseline models used for extractive summaries

5.1.1 *LSTM (Long Short Term Memory)*

LSTM is an artificial recurrent neural network (RNN) architecture that has feedback connections [9]. A unidirectional LSTM model with 25 neurons and a bidirectional LSTM model with 25 neurons were trained. The embeddings for sentences were obtained using BERT provided by Google.

5.1.2 *TextRank*

TextRank is a graph-based ranking model for text processing that can be used in order to find most relevant sentences in text and to find keywords [10]. The found sentences are then appended to form the final summary. A limit of maximum of 15 phrases and 5 sentences was imposed on summary.

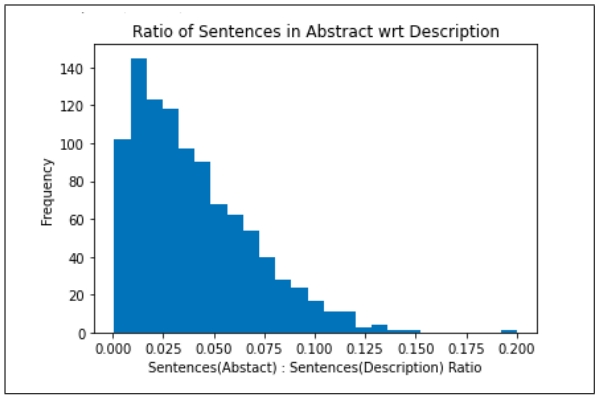

Figure 7: Ratio of abstract length and description length in terms of number of sentences

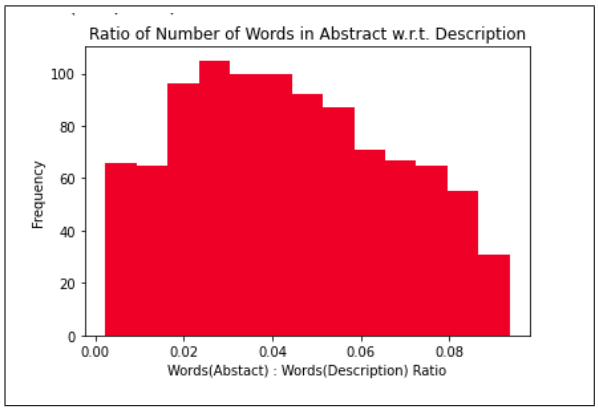

Figure 8: Ratio of abstract length and description length in terms of number of words

5.1.3 *Gensim*

Gensim is an NLP processing library that provides tools for summarization using the PageRank algorithm [11].

5.1.4 *Using Cosine Similarity*

The cosine similarity metric measures the similarity between two non-zero vectors of an inner product space by measuring the cosine of the angle between them. The similar vectors have the product closer to 0 and for different vectors, the product increases towards 1 [12]. BetterNLP pre-trained model for summarization was used along with cosine similarity for this task. It ranks sentences on the basis of cosine similarity and then picks the top ranking sentences from there to form the summary.

5.1.5 *Using TF-IDF Ranking Method*

TF-IDF (Term Frequency Inverse Document Frequency) finds out the rank of the sentences using the frequency with which the terms occur in it and occur in general in the corpus [13]. BetterNLP pre-trained model for summarization was used along with TF-IDF similarity for this task. It ranks sentences on the basis of TF-IDF similarity and then picks the

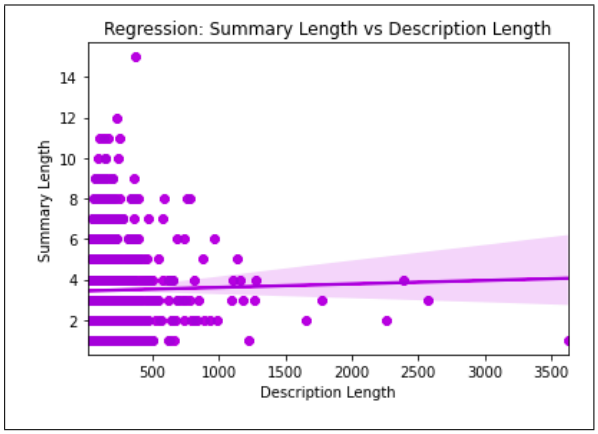

Figure 9: Regression between number of sentences in description and abstract

top ranking sentences from there to form the summary.

5.2 Abstractive summary

The abstractive summary does not pick words and phrases directly from the text but rather paraphrases the text. It uses a different set of vocabulary to construct the final summary. The following model was used for the abstractive summary.

5.2.1 PEGASUS

PEGASUS is a pre-trained large Transformer-based encoder-decoder model. In PEGASUS, important sentences are removed/masked from an input document and are generated together as one output sequence from the remaining sentences. It has been trained on different datasets. In specific, the model used here was trained on the big-patent dataset.

5.3 Summarisation - Evaluation

The summaries generated were scored using ROUGE score. ROUGE-1, ROUGE-2 and ROUGE-L were used to check correctness of the generated summaries against the human written GOLD summaries available as abstract along with descriptions.

ROUGE-1 matches the number of unigrams between the two summaries. ROUGE-2 similarly matches the number of bigrams. ROUGE-L finds the ratio of Longest Common Subsequence Length between the two summaries generated. The results obtained for different models are summarised in Table 1.

5.4 Summarisation - Interpretation

PEGASUS - the abstractive model performs the best. It has the highest ROUGE scores. It probably happens because the pre-trained model used has been trained on Big-PATENT dataset itself. Also, the gold summaries that are being compared against have been written by humans who use the abstractive method by paraphrasing the main content. Henceforth, PEGASUS will be used as the baseline model for summarisation.

5.5 Patent Retrieval

In order to retrieve relevant patents based on a query input by the user, the following steps were followed:

1. All the patent documents were traversed and their data was preprocessed. The preprocessing steps involved removing accented words, punctuations and stopwords followed by stemming the text.

2. In order to create the inverted index, we built the vocabulary from the documents and for each term in the vocabulary, we stored a list of IDs of documents containing the term in sorted order. The index was sorted alphabetically on the basis of terms.

3. In order to create the positional index, for each term in the vocabulary, we store the document ids containing that term along with a list of the positions at which the term occurs.

4. For the inverted index model, the documents containing all words from the query (i.e. boolean AND type query) were retrieved. For the positional index model, the documents containing all positional bigrams from the query were retrieved.

5.6 Patent Retrieval - Interpretation

The postional indexing method is performing faster but retrieves very few and specific documents which leads to document famine. We would further need to build upon this retrieval model to optimise time of query and ensure enough and relevant documents are fetched.

6. PROPOSED METHOD

To create a web platform for summarising the patents and retrieving the related patents, we need to finalise the summary model. Since, the training set has more than 10k records, currently all experiments have been carried on a smaller subset of documents. Once the model would be finalised, it would be extended for the remaining model.

For retrieval, after preprocessing, currently inverted index and positional indexing have been used along with boolean retrieval methods. We would further explore topic clustering, top-K-rank retrieval, etc.for query optimisation.

Finally we would integrate both in a web platform and deploy it on heroku. We would collect user feedback to further improve the performance of the proposed methodology.

6.1 Primary Goals

Model Name	ROUGE-1		ROUGE-2		ROUGE-L	
	Recall	F1	Recall	F1	Recall	F1
Extractive Summrisation						
LSTM	36.11	33.64	9.70	8.95	21.2	19.68
Bi-LSTM	36.37	33.58	9.90	9.12	21.53	19.76
TextRank	24.57	31.66	6.77	8.83	13.85	17.97
Gensim	20.96	25.47	4.73	5.83	12.93	15.67
Cos score	24.54	30.47	7.37	9.16	14.48	18.00
TF-IDF	25.66	30.39	6.30	7.56	14.14	16.79
Abstractive Summrisation						
PEGASUS	57.72	36.41	25.43	15.73	41.32	25.73

Table 1: ROUGE scores for diffeent summarisation models

Query	No. of docs in inverted indexing	Time taken by inverted indexing	No. of docs in positional indexing	Time taken by positional indexing
fan cold cooling rotate metal steel cover heating	4	0.107	0	0.062
dentist typically supplies the technician with a full face photograph	1	0.081	1	0.049
may be straight - chained or branched . an optionally substituted alkyl	67	0.098	1	0.062
interchangeable retail display in accordance with the invention	3	0.048	1	0.041
numerous specific details are set forth in order to provide a thorough	94	0.126	2	0.085
effects of garlic extracts containing allicin for prostate tumor treatment	1	0.097	1	0.050
cooking device 10 will now be described with respect to the figures	17	0.102	1	0.053
figures are not drawn to scale and they are provided merely	23	0.070	1	0.035
accordance with the present invention will be described	4671	0.078	207	0.057
drawings wherein like numerals refer to like matter throughout	38	0.125	1	0.069

Table 2: Performances of inverted indexing and positional indexing models

6.1.1 Data Analysis phase

1. Understand the dataset corpus.
2. Pre-process the dataset to remove any redundant information.
3. Extract useful information from the dataset in a proper format using Stanford NLP Parser and OpenNLP toolkit.
4. Plots graphs, wordclouds, etc. to see the different linguistic features of the dataset.

6.1.2 Feature Extraction phase

1. Infer appropriate NLP vectorisers to use from statistics obtained from previous step.
2. Select models to perform summarisation task and train it.

6.1.3 Model evaluation phase

1. Analyse the performance of selected model.
2. Tune it better, in case the performance is not satisfactory.
3. Test the accuracy over the unseen test set.
4. Plot the graphs to see the performance of the trained model.

6.1.4 Web Platform

1. Build the web platform for accepting new patents.
2. Return summary of the input patent.
3. Deploy the platform on Heroku/ DigitalOcean.

6.2 Secondary Goals

1. Read about Lucene and other indexing and searching toolkits to analyse usability and feasibility for the proposed task.
2. Build a system to return similar documents given a input document.
3. Find similar patents from the keyword.
4. Customise the web platform to act as search engine for the patents in the dataset.

6.3 Stretch Goals

1. Allow online learning for the summariser and retrieval system based on any new user input by including user feedback.

7. REFERENCES

[1] Pirmin Lemberger. Deep learning models for automatic summarization. *arXiv preprint arXiv:2005.11988*, 2020.

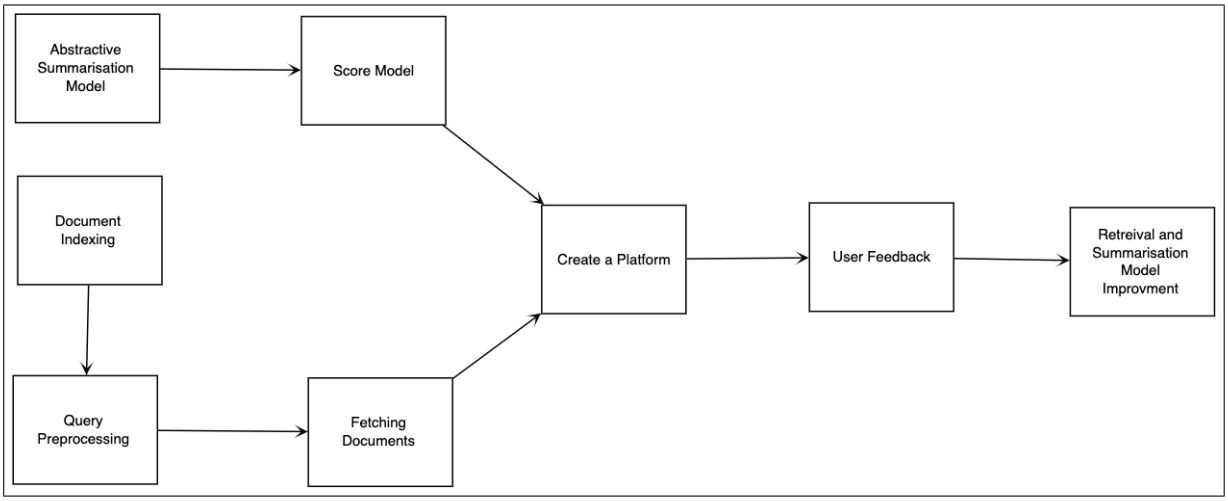

Figure 10: Proposed Methodology

[2] Amy Trappey, Charles Trappey, and Chun-Yi Wu. Automatic patent document summarization for collaborative knowledge systems and services. *Journal of Systems Science and Systems Engineering*, 18:71–94, 03 2009.

[3] Jie Hu, Shaobo Li, Yong Yao, Liya Yu, Yang Guanci, and Jianjun Hu. Patent keyword extraction algorithm based on distributed representation for patent classification. *Entropy*, 20:104, 02 2018.

[4] Jen-Yuan Yeh, Hao-Ren Ke, and Wei-Pang Yang. ispreadrank: Ranking sentences for extraction-based summarization using feature weight propagation in the sentence similarity network. *Expert Systems with Applications*, 35:1451–1462, 10 2008.

[5] Pawan Sharma, Rashmi Tripathi, and R.C. Tripathi. Finding similar patents through semantic query expansion. *Procedia Computer Science*, 54:390–395, 12 2015.

[6] Lea Helmers, Franziska Horn, Franziska Biegler, Tim Oppermann, and Klaus-Robert Müller. Automating the search for a patent's prior art with a full text similarity search. *PLOS ONE*, 14:e0212103, 03 2019.

[7] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. page 10, 01 2004.

[8] Eva Sharma, Chen Li, and Lu Wang. Bigpatent: A large-scale dataset for abstractive and coherent summarization. pages 2204–2213, 01 2019.

[9] Hasim Sak, Andrew W Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. 2014.

[10] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[11] Keyvan Khosrovian, Dietmar Pfahl, and Vahid Garousi. Gensim 2.0: A customizable process simulation model for software process evaluation. In *International conference on software process*, pages 294–306. Springer, 2008.

[12] Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. Semantic cosine similarity. In *The 7th International Student Conference on Advanced Science and Technology ICAST*, volume 4, page 1, 2012.

[13] Cheng-Hui Huang, Jian Yin, and Fang Hou. A text similarity measurement combining word semantic information with tf-idf method. *Jisuanji Xuebao(Chinese Journal of Computers)*, 34(5):856–864, 2011.

8. APPENDIX

Summary Examples

Golden Summary
a fitting for inserting a needle or catheter into a blood vessel and for positionally stabilizing the inserted needle or catheter in situ includes separate needle holder and stabilizer elements . a retainer member is secured at its proximal end to the holder and extends in spaced relation along the holder to a distal end . the stabilizer has a bi - wing shape with a slot contoured to receive the retainer member in longitudinally slidable relation at a location below a holder supporting region on the top surface of the stabilizer . the wings can be flexed upward and toward one another to positively grip the holder for insertion of the needle or catheter . when unflexed , the stabilizer slot constrains the retainer element which in turn holds the stabilizer in the space between the retainer and the holder . removal of the needle is achieved by withdrawing the retainer longitudinally from the stabilizer slot .

PEGASUS Summary
A needle support includes a holder member for supporting a needle and a retainer member extending between the holder member and the rearward end of the needle. The retainer member is resiliently urged against the holder member so as to preclude mutual axial displacement of the needle and holder member during insertion of the needle into a blood vessel.

LSTM Summary
holder member 11 may be molded about the contained rearward end of needle 12 or may be otherwise secured thereto in any conventional manner so as to preclude mutual axial displacement of needle 12 and holder 11 during insertion of the needle into a blood vessel in the manner described herein . portions of bottom surface 22 are provided with a suitable adhesive coating , as is conventional for such applications , so that the bottom surface can adhere to a patient &#

39 ; s skin for initial positional stabilization after insertion of needle 12 into a blood vessel . in order to withdraw the needle from the patient &# 39 ; s blood vessel , the surgical tape is first removed .', 'a fitting for inserting a needle or catheter into a blood vessel and for positionally stabilizing the inserted needle or catheter in situ includes separate needle holder and stabilizer elements . a retainer member is secured at its proximal end to the holder and extends in spaced relation along the holder to a distal end . the stabilizer has a bi - wing shape with a slot contoured to receive the retainer member in longitudinally slidable relation at a location below a holder supporting region on the top surface of the stabilizer . the wings can be flexed upward and toward one another to positively grip the holder for insertion of the needle or catheter . when unflexed , the stabilizer slot constrains the retainer element which in turn holds the stabilizer in the space between the retainer and the holder . removal of the needle is achieved by withdrawing the retainer longitudinally from the stabilizer slot .