

DeepKEA: Employing Deep Learning Models for Keyword Extraction from Patent Documents

Rima Dessi¹, Hidir Aras¹ and Lei Zhang¹

¹FIZ Karlsruhe - Leibniz Institute for Information Infrastructure, Germany

Abstract

Patents are an important source for technological innovation, utilized by companies to disclose their inventions as well as protect intellectual properties legally. Due to the exponential growth of available patent data, keyword extraction has become an important task for the efficient analysis and organization of patent documents. Keywords extracted from the text of a patent comprise highly relevant terms or phrases that represent the content of the patent document. Further, such terms are exploited by various patent applications such as freedom-to-operate analysis, prior-art-search, etc. Most of the existing methods are not able to extract useful terms that help to understand the core of the patent, i.e., the invention. Moreover, these approaches focus on either supervised settings, which require large amounts of training data, or unsupervised settings that cannot extract semantically meaningful keywords. To fill these gaps, this paper proposes a weakly-supervised deep neural network model (DeepKEA) which is designed to extract terms that are closely related to the topic of a given patent document and the invention it describes. It consists of two main modules: (1) a training data generation module, (2) a deep neural network module. The experiments show that our model yields better performance than existing baselines.

Keywords

information retrieval, deep learning, keyword extraction, patent analysis


1. Introduction


Intellectual property (IP) rights play an important role in the creation, dissemination, and use of new knowledge for further technological innovation. Patent documents which are complex, heterogeneous, and lengthy in nature, contain scientific, technical, legal, and business-relevant information. The so-called full text of a single patent document consists of a title, an abstract, claims, and a detailed description. While the description part describes the embodiment of the invention, its use, and the benefits it offers for target applications, the claims give a clear definition of what the patent legally protects, i.e. they define the scope and boundaries of an invention for the purpose of legal protection. In order to deal with steadily growing patent data, researchers started to employ AI-based approaches to support experts in patent retrieval and analysis processes. One crucial task for patent searchers is to find the right information that can be used to support business-critical decisions. To seek such crucial information and explore patent data fast and efficiently, various automatic keyword extraction methods have been developed [1, 2, 3].

LIRAI'23: 1st Legal Information Retrieval meets Artificial Intelligence Workshop co-located with the 34th ACM Hypertext Conference, September 4, 2023, Rome, Italy

✉ rima.dessi@fiz-karlsruhe.de (R. Dessi); hidir.aras@fiz-karlsruhe.de (H. Aras); lei.zhang@fiz-karlsruhe.de (L. Zhang)

ORCID 0000-0001-8332-7241 (R. Dessi); 0000-0002-3117-4885 (H. Aras); 0000-0001-8184-9353 (L. Zhang)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Recently, several supervised approaches have been proposed [4, 5], however, they require a large amount of labeled data. Obtaining labeled data is an expensive and time-consuming task. On the other hand, the most prominent unsupervised approaches employ bag-of-words (BoW), graph-based, and topic-modeling techniques to perform the keyword extraction task. The BoW methods rely on measures such as Term Frequency-Inverse Document Frequency (TF-IDF) [6], and they do not require any training data. However, these methods cannot capture the semantic meaning of the text and the extracted keywords. The graph-based approaches [7] model the text into graphs and select the most scored nodes as keywords; however their performance drop with an increasing number of extracted keywords. Finally, Latent Dirichlet Allocation (LDA) is a popular topic modeling technique which does not require any training data, yet the extracted keywords with this technique are too general to convey the meaning of the text [8]. Therefore, the keywords extracted via statistical or hybrid methods [7] cannot fulfill the requirement of the patent experts.

In this study, we present our preliminary work toward a novel patent keyword extraction model (DeepKEA) based on deep neural networks. DeepKEA starts by extracting noun phrases from the abstracts and claims of patent documents. These extracted noun phrases serve as initial candidates for relevant terms. To refine these phrases expert validation is employed. In other words, each extracted noun phrase is reviewed and validated by a patent expert internally. It is important to note that this study focuses on the abstracts and claims, and the exploration of noun phrases within patent descriptions will be addressed in future work. In the second step, DeepKEA uses the extracted keywords and their corresponding original patent text to train a deep neural network. Finally, the trained model allows to extract a list of keywords for a given arbitrary patent document.

Overall, the main contributions of the paper are:

- A pipeline to generate training data for the patent keyword extraction task,
- A neural network architecture for generating embeddings of patent documents and keywords,
- The adoption of an approximate nearest neighbor search for efficient keyword extraction based on dense vectors.

2. Keyword Extraction from Patents (DeepKEA)

Problem Formulation. Given an input patent document d , which contains a set of noun phrases $K_d = \{k_1, k_2, \dots, k_m\}$, the goal is to output the most relevant top- N noun phrases as keywords $K'_d = \{k'_1, k'_2, \dots, k'_N\}$, where $K'_d \subseteq K_d$.

Overview. The general workflow of DeepKEA is shown in Figure 1. There are two main modules of the proposed workflow: (1) the *Training Data Generation Module*, and (2) the *Deep Neural Network Module*. Section 2.1 and Section 2.2 provide a detailed description of each module and the feature sets that have been utilized by each module.

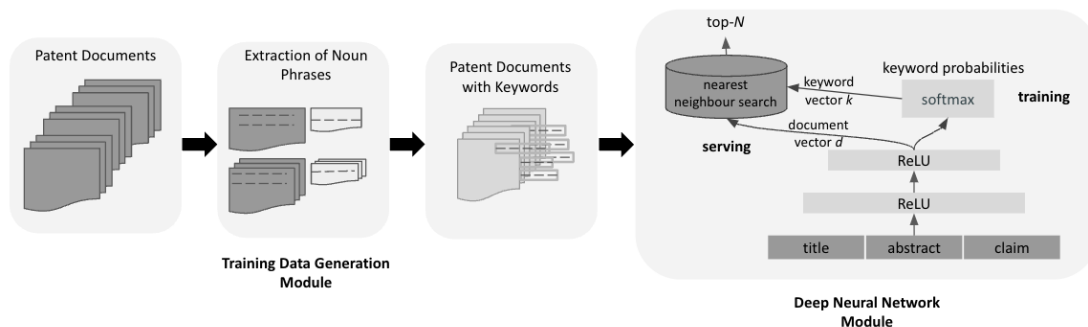


Figure 1: The workflow of DeepKEA

2.1. Training Data Generation Module

The *training data generation module* is responsible for assigning meaningful noun phrases present in the abstracts and claims as keywords to the respective patent documents. The noun phrases are identified by employing an NLP pipeline using Spacy¹. These extracted noun phrases are validated and refined internally by patent experts. This examination involves the reviewing of keywords based on the content of the corresponding patent document. This curation process ensures that the extracted noun phrases align with the content of the patent document. Further, they serve as a set of keywords specifically assigned to the corresponding patent documents for the training phase.

2.2. Deep Neural Network Module

The second module of the workflow involves the training and serving of a *deep neural network model*. We have adapted an approach which uses deep learning for recommendation systems [9] to our keyword extraction task, where we consider keyword extraction as extreme multi-class classification. The prediction problem is to classify a specific keyword (as a class) among all keyword classes based on the features of a patent document. For this study, we have used the title, abstract and claim parts of the patent text. Three different embedding vectors, i.e., *Title Embedding*, *Abstract Embedding* and *Claim Embedding* are utilized as the input to the model (see Fig. 1). Each input embedding is obtained by exploiting the Sentence Transformers with BERT for Patents² which has been trained by Google on over 100M patents. Given the concatenation of these input embeddings, two fully connected hidden layers are added on top, the output of which can be thought of as the embeddings of patent documents. Based on that, the softmax layer outputs a probability distribution over all keyword classes, each of which can be thought of as a separate keyword embedding. In training, a cross-entropy loss is minimized with gradient descent on the output of the softmax. At the end of the training stage, the model produces two separate sets of output embeddings: one for patent documents and the other for keywords. Those can be thought of as semantic representations of the patent documents and keywords, respectively.

¹<https://spacy.io/>

²<https://huggingface.co/anferico/bert-for-patents>

At serving time, we need to compute the most likely N classes (keywords) for each patent document based on the generated patent and keyword embeddings. However, scoring a large amount of keywords is expensive. Therefore, to address the challenge of computational expense, an approximate nearest neighbor lookup based on inner-product is performed to efficiently generate the top- N keywords, for which Faiss³, a library for efficient similarity search of dense vectors, is used.

Finally, when confronted with a new patent document in the serving stage, the workflow involves creating embeddings for the document’s title, abstract, and claim sections using the sentence transformer with BERT for patents, as previously described. Subsequently, the trained deep neural network, DeepKEA takes these embeddings as input and generates the embedding of the given patent document. Utilizing inner-product similarity top- N keywords are identified by referencing precomputed keyword embeddings.

3. Experimental Results

This section provides a description of the dataset and the baselines, followed by the experimental results and a comparison to the baseline approaches.

3.1. Dataset and Baselines

For the evaluation of the proposed model, we gathered internally a dataset consisting of 9,664 unique patents. The patent documents are utilized as input to the proposed workflow. Each patent document is paired with its keywords that are extracted from the abstracts and claims by applying the first step of the workflow, i.e., training data generation. The extracted noun phrases are then validated by experts. Further, the keywords that appeared less than 100 times in the entire dataset are filtered out. After applying the first step the average number of extracted keywords per patent is 13.71, resulting in a total of 14,957 unique keywords. The dataset⁴ is organized into pairs consisting of keywords and their corresponding patent documents. Finally, the data is split into 13,586 and 1,371 pairs as train and test data, respectively. The test data consists of 100 unique documents and their corresponding keywords.

To evaluate the performance of DeepKEA, two different baselines are selected:

- **TF-IDF** is a standard baseline due to its simplicity and effectiveness. It is applied on noun phrases of the patent documents and based on the tf-idf score top- N keywords are assigned.
- **BERT for Patents** exploits the vector similarity between documents and keywords. Each document and its keywords (noun phrases present in the document) are firstly converted into their vector representations with the help of sentence transformers with BERT for Patents. Based on the vector similarity between the document and keywords, top- N keywords are assigned.

³<https://faiss.ai/>

⁴<https://github.com/rima-turker/Patent-Keyword-Extraction.git>

Table 1

Performance of DeepKEA on the test collection, compared against two baseline methods

Model	P@10	R@10	F1@10	P@15	R@15	F1@15	P@20	R@20	F1@20
DeepKEA	0.678	0.511	0.583	0.626	0.703	0.662	0.604	0.849	0.705
BERT for Patents	0.572	0.432	0.492	0.564	0.634	0.597	0.551	0.770	0.643
TF-IDF	0.313	0.240	0.272	0.351	0.399	0.374	0.343	0.510	0.410

3.2. Evaluation

Table 3.2 illustrates the performance of DeepKEA in comparison to the baselines. We have utilized standard metrics, namely precision, recall and f1-score at different N . The proposed model outperforms the baselines on each N value. Although TF-IDF is a standard approach for the keyword extraction task, it could not perform well with the noun phrases due to the sparsity issue. It should be noted that we have also tried to use words instead of phrases. However, single words as keywords could not represent the fundamental idea of patent documents compared to phrases. Moreover, BERT outperforms the TF-IDF on each experiment. The reason is BERT captures the semantic meaning of noun phrases and documents for assigning the keywords, while TF-IDF assigns keywords based on their occurrences such that it does not consider any semantics. On the other hand, BERT simply relies on vector similarities based on pre-trained embedding models. Whereas, DeepKEA first generates training data and then trains a deep neural network to perform keyword extraction. Hence, the proposed model utilizes more task-oriented semantics than the baselines. Therefore, the obtained results with DeepKEA are much more promising.

4. Conclusion and Future Work

In this preliminary study, we present DeepKEA, a deep neural network model for extracting highly relevant keywords (noun phrases) from patent documents. First, the training data is generated by leveraging the abstracts and claims of patent documents. This initial set of noun phrases is then subjected to validation and filtering by domain experts. Second, the training data is used to train a deep neural network for obtaining the embeddings of documents and keywords. Finally, the model assigns keywords to individual (unseen) documents by applying an approximate nearest-neighbor search based on dense vectors. The experimental results showed that DeepKEA outperforms the baselines. As for future work, we aim to (1) improve the training data generation module by exploiting additional semantic information as well as description part, (2) improve the deep neural network module by including more structured features of patent documents, such as CPC (Cooperative Patent Classification) codes, citations, inventors and applicants.

References

- [1] S. Suzuki, H. Takatsuka, Extraction of keywords of novelties from patent claims, in: COLING, ACL, 2016.
- [2] R. Alzaidy, C. Caragea, C. L. Giles, Bi-lstm-crf sequence labeling for keyphrase extraction from scholarly documents, in: WWW, 2019.
- [3] Q. Zhang, Y. Wang, Y. Gong, X. Huang, Keyphrase extraction using deep recurrent neural networks on twitter, in: EMNLP, 2016.
- [4] R. Alzaidy, C. Caragea, C. L. Giles, Bi-lstm-crf sequence labeling for keyphrase extraction from scholarly documents, in: The world wide web conference, 2019.
- [5] Y. Zhang, M. Tuo, Q. Yin, L. Qi, X. Wang, T. Liu, Keywords extraction with deep neural network model, *Neurocomputing* (2020).
- [6] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, *Information processing & management* 24 (1988) 513–523.
- [7] Z. Huang, Z. Xie, A patent keywords extraction method using textrank model with prior public knowledge, *Complex & Intelligent Systems* 8 (2022) 1–12.
- [8] J. Hu, S. Li, Y. Yao, L. Yu, G. Yang, J. Hu, Patent keyword extraction algorithm based on distributed representation for patent classification, *Entropy* (2018).
- [9] P. Covington, J. Adams, E. Sargin, Deep neural networks for youtube recommendations, in: *Proceedings of the 10th ACM Conference on Recommender Systems*, ACM, 2016.