Get Your Hands Dirty: Evaluating Word2Vec Models for Patent Data

Hidir Aras[1], Rima Türker[1,2], Dieter Geiss[1], Max Milbradt[1], and Harald Sack[1,2]

[1] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
[2] Karlsruhe Institute of Technology, Institute AIFB, Germany
{firstname.lastname}@fiz-karlsruhe.de
{firstname.lastname}@kit.edu

Abstract. Patent search systems allow complex queries to be formulated by combining different search terms using boolean and other operators such as proximity, wildcards, etc. in order to find relevant patents. This widely adopted approach is based on exact match, making it difficult to efficiently identify and analyze relevant patents, as the search terms often do not match the terminology used by the inventors. Another problem concerns the large number of relevant hits due to weekly and monthly updates of patent applications and grants. Although some semantic search systems for patents based on latent semantic analysis have been implemented as black-box systems in the past, word embeddings that have been successfully applied to generate semantic representations of text have rarely been employed and evaluated for a (large) patent corpus. The work described here aims to evaluate semantic representations for patent data via a pre-trained general model in comparison to an adapted word embedding model created from a patent corpus in order to contribute to a multitude of semantic analysis tasks for patents such as similarity search, content analysis, entity linking etc..

1 Introduction

Patents are regarded as an important source allowing companies to define new business strategies and support high-level decision making processes. With increasing complexity and volumes of patent data enhanced search systems and novel methods for analyzing patents [1] which aid the time-consuming patent reviewing process are required. Herewith, experts can gather valuable insights for detecting novel inventions, analyzing patent trends, identifying technological hotspots, etc. The expectations towards new types of search systems for patents are high [2], as information professionals not only wish to find more accurate results but also detect frequently hits which could not be found using traditional search. Although patents provide valuable scientific information which can be gathered via text mining [3] and are able to indicate to novel scientific relationships in earlier literature, they clearly focus on commercial applications, e.g. use of drugs for medical purposes. Besides that, patents also entail considerable difficulties due to their broad claims, non-relevant references embedded in patent

text which may lead to wrong relations, and the heavy use of acronyms leading to more false positives. In this paper, we evaluate word embeddings models created from different corpora for calculating the semantic similiarity of patent documents, a task which is crucial in several patent analysis use cases such as prior art, freedom to operate and infringement.

2 Related Work

In [4] latent semantic indexing (LSI) was applied for automatic indexing in information retrieval. However, the results only showed little improvement over the vector space model. Alternatively, commercial systems like *TotalPatent*[1] or *OctiMine*[2] are accessible only as black-box systems for search and retrieval based on the (semantic) similarity of patent documents to determine their relevance for a given query. Other systems like *PatBase*[3] also enable semantic search based on the semantic analysis of citation networks. In [5] the peculiarities of patent search systems such as semantic similarity and semantic search are described in more detail. The work in [6] and [7] describe semantic representations of paragraphs and short text, while the research needed to study the semantic similarity complex document types such as patents is still lacking.

3 Approach: Using Word Embeddings for Patent Data

Dataset. In order to create a domain-specific word embedding model, a subset of patent documents from the WIPO (World Intelectual Property Organization) and the EPO (European Patent Office) patent databases has been sampled by filtering for specific patent classification codes as shown in Table 1. In total

Search Fields	Used IPC, CPC Codes
IPC, CPC	G06* AND NOT G06M* OR H04L0009* OR H04W0012-00 OR H04H0060-23 OR G11*

Table 1. Patent query based on IPC/CPC classification codes.

410.607 patent documents have been retrieved and the English patent description texts were used to generate the patent embedding model, furtheron referred to as IT corpus.

Model Creation. To create a word2vec model, the Gensim library was applied using default parameters. The model creation and document vector representation were based on the description text of the patent documents. For comparison the pre-trained Google word2vec model has been used as a baseline.

[1] https://www.lexisnexis.com/totalpatent
[2] https://www.octimine.com
[3] https://www.patbase.com

4 Evaluation

4.1 Setup

The evaluation[3] of both models for the task of semantic similarity of patent documents was performed based on a randomly selected set of 25 documents (14 WIPO and 11 EPO) from the IT corpus. For each document the 10 most similar patent documents were determined based on the vector representation of the description text. To obtain document vectors, word vectors for the patent description text were averaged. The cosine similarity was used to compute the similarity among document vectors. As to our best knowledge there exists no ground-truth for this task, a qualitative evaluation was carried out with patent experts. In the evaluation they assesed the relevance of the top 10 similar patents according to the following 3-level scale: 0: irrelevant patent, 1: related patent, 2: similar patent.

4.2 Results

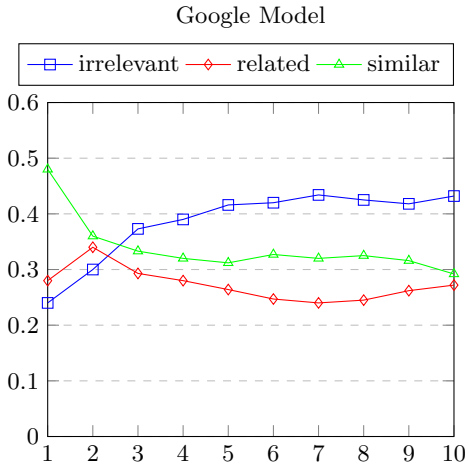

Fig. 1. Performance of the Google Model, x-axis corresponds to rank and the y-axis corresponds to percentage(%).

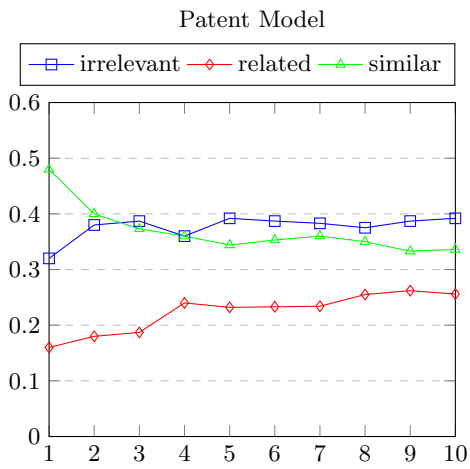

Fig. 2. Performance of the Patent Model, x-axis corresponds to rank and the y-axis corresponds to percentage(%).

In order to compare the results of both models we analyzed the cumulated and normalized scores (Fig. 1 and Fig. 2) in dependence of the rank for the top 10 documents. Looking at the similarity, we see better average score for the Patent Model with increasing rank, while the relatedness shows a different picture. Here, the higher ranks of the Google Model[4] show better results, while with increasing rank the relatedness score meet at the same level. It can also be

[3] https://github.com/dlatfiz/w2v4pat
[4] https://code.google.com/archive/p/word2vec/

observed that the average score for the irrelevant documents rise with increasing rank in the Google Model, while in the Patent Model the score stabilize around 40%. We assume that for the similarity the representation of domain-specific words in the customized model is more sophisticated compared to the Google model, which was trained on general (news) data. In contrast, the patent model was trained only with a much smaller number of domain-specific patent data not being able to cover all aspects of relatedness appropriately.

5 Conclusion and Future Work

In this paper, we have evaluated semantic representations of patent texts via word embeddings created from distinct corpora. We compared the results for the two regarded distinct word2vec models for the tasks of semantic similarity and relatedness. The achieved results showed that the mean average scores for the domain specific word embedding model are much higher in comparison to the general Google word2vec model, while the relatedness aspect must be evaluated in additional experiments employing a more fine-grained scoring scheme for the experts. In future work, we aim to enhance our approach by considering a more sophisticated pre-processing also taking into account the inherent structure of a patent document. In this work, the evaluation was performed based on the patent description text only, while in future we also want to analyze how abstracts and claims of the patent text can be exploited for different patent analysis use cases in different selected domains such as life science, engineering, etc.

References

1. Assad Abbas, Limin Zhang, Samee U. Khan, A literature review on the state-of-the-art in patent analysis, World Patent Information, Volume 37, 2014.
2. Mihai Lupu, Katja Mayer, Noriko Kando, and Anthony J. Trippe. Current Challenges in Patent Information Retrieval (2nd ed.). Springer Publishing Company, Incorporated, 2017.
3. Aras, Hidir, René Hackl-Sommer, Michael Schwantner and Mustafa Sofean. Applications and Challenges of Text Mining with Patents. IPaMin@KONVENS, 2014.
4. A. Moldovan, R. I. Boţ, G. Wanka. Latent semantic indexing for patent documents, International Journal of Applied Mathematics and Computer Science 15(4), 551-560, 2005.
5. Björn Jürgens, Nigel Clarke, Study and comparison of the unique selling propositions (USPs) of free-to-use multinational patent search systems, World Patent Information, 2014.
6. Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 (ICML'14), Eric P. Xing and Tony Jebara (Eds.), Vol. 32. JMLR.org II-1188-II-1196.
7. Tom Kenter and Maarten de Rijke. 2015. Short Text Similarity with Word Embeddings. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15). ACM, New York, NY, USA, 1411-1420.