Text Retrieval and NLP in Intellectual Property (IP)

Matthew Mechtly

Introduction

The rate at which patent applications are being submitted is increasing linearly[1]. Consequently, the rate at which applications are accumulating is accelerating. This means that the potential overlap between existing documents and new patents is $O(n^2)$, where 'n' is the cumulative number of patents. In other words, the potential conflict in exclusive claims between patents are becoming increasingly more difficult to manage using human judgment exclusively. Already, this computational complexity has decreased intellectual property protection and made enforcement more difficult[2]. Therefore, computer scientists, linguists, and lawyers need to create more automated approaches to managing both existing patents and new patent applications.

Within the specialty of intellectual property, many text-centered tasks related to patent processing must be undertaken. For patent offices and patent holders alike, the ability to correctly categorize patents and patent applications by the correct classification code is a necessity[3]. While doing so is a matter of completing the required workflow for patent offices, it is a matter of improving recall for business-related applications. From finding patents that already exist prior to investing capital in R&D, to the discovery of infringing patent applications, the financial implications are immense. Additionally, the ability to identify markets with significant growth opportunity by data mining patents enables entrepreneurs to better compete, thereby benefiting consumers. Lastly, the ability the take the unstructured text data within patents and annotate those patents with structured metadata facilitates querying with a language like SQL[4]. These are some of the many possible patent-related applications which have the potential to be automated.

Unfortunately, patent-specific datasets do not permit task-automation to the extent that do many other sectors, such as the biomedical and academic fields[5]. Specifically for querying, text retrieval techniques like BM25 often do not perform better than a coin-flip, depending on the dataset in question[6]. Additionally, machine learning algorithms that work well elsewhere tend to not perform as well in the intellectual property space either. While SVMs don't work well because of the significant overlap in data, KNN implementations fall short because of the many dimensional spaces created by extensive jargon—much of which is purposely included to obscure patent applications from those who would challenge the claims therein. For tree-based algorithms, there is a tendency to over-fit models when extensive training data is not available. This is almost always the case during patent-related endeavors. Even algorithms like Naive Bayes—which should address most of the issues associated with patent-related data—only achieve F1 scores around 0.4 for many classification tasks[5]. In short, intellectual property data is difficult to work with along multiple dimensions. Consequently, the intellectual property space has been less studied than most other domains, much to the detriment of the resulting tools.

Body

With regard to text-based analytics—whether text retrieval, text mining, or deep-learning-based NLP—the intellectual property field lags behind many other industries[7]. As noted previously, this is at least partly due to the complexity and sparsity of the underlying data itself. Much of what work has been

done has employed the CLEF-IP dataset, the standard dataset to use for classification tasks. These classification tasks are especially useful for the discovery of potentially conflicting patents[8].

Fortunately, this trend has begun to reverse in the past decade as more research has addressed intellectual property applications specifically. Starting in 2011, Trappey et al. employed a supervised learning approach to quantify patent quality, since patent quality is inextricably linked with market competitiveness[9]. By first extracting the relevant text-related characteristics—including but not limited to patent references cited, links to other publications, and claims—the researchers were able to extract several pieces of data from the unstructured patent text. Then, after projecting them to a lower dimension using principal component analysis, the researchers were ultimately able to achieve a classification accuracy of 85% using only 399 patents for the training set. At the time, this 85% figure was an impressive rate for patent sale or licensing classification. Patent quality in this particular situation was assumed to be equivalent with sale and/or licensing of the patent in question.

One year later, Hido et al. approximated "patent quality" using a more direct measure: whether a particular patent application was actually granted[10]. Instead of simply measuring the frequency of some of the more obvious text features like Trappey et al., these researchers also extracted features related to the syntax of sentences (e.g. maximum depth of dependencies in a sentence), the age of a word (how many years since a particular keyword first appeared in a patent, since novelty is an important aspect of patent granting), as well as a TF-IDF implementation for terms. Using a simple logistic regression model, the researchers achieved an AUC of 0.6. While this result cannot be compared directly to the previous paper since the datasets and evaluation criteria were different, Hido et al. clearly implemented more complex feature engineering. More specifically, Hido et al. integrated the more advanced NLP techniques of syntactic analysis.

Pivoting away from patent classification into patent search, Vrochidis et al. successfully combined both text and image data from patents to improve patent-specific search[11]. Due to the lack of color and the relative lack of features from which to learn in the images themselves, learning patent images has been more difficult than learning images in many other domains[12]. Still, Vrochidis et al. were able to improve upon using text analysis alone. Specifically, the researchers accomplished this by extracting low level image features, integrating these with the text-based features more typically used for patent-related search, and then learning from all those features. Using an SVM, the researchers achieved an overall precision of 91% with an F-score of 79%[11]. While this paper didn't introduce any ground-breaking alterations to text retrieval specifically, this was the first successful NLP and computer vision integration in the intellectual property space.

Three years later in 2015, a pair of researches implemented a more advanced topic mining technique to further improve patent-related tasks. Venugopalan and Varun created a probabilistic topic model based on a Latent Dirichlet Allocation (LDA) model[14]. This led to a 98% accuracy in relevance classification and an 87% accuracy in category classification. This began a period in which several papers employed LDA to improve task performance. Suominen and Seppanen used LDA on the full-texts of patents to quantify the "knowledge profile" of various companies[15]. In this situation, a "knowledge profile" is simply a company's area of expertise, as evinced by the published of successful patents and characterized by the latent topics associated with each patent. Govindarajan et al. used LDA to perform topic modeling, thereby allowing them to create a domain-specific ontology[16]. In other words, they were able to create a hierarchical relation of topics using LDA. Finally, in 2019, Clerq et al. extracted topics from 17,500 electric vehicle patents using LDA, resulting in an F1 score of 74% for a nine-class categorization[17].

From relevance rating to categorical classification to using LDAs to create topic hierarchies, various models were being developed to answer an array of questions; however, during this time period deep learning began to outperform other models in the intellectual property space. This was primarily due to the availability of large corpora of patent data as well as access to cheaper computation capabilities. In several studies, these deep learning NLP techniques even outperformed human judgment[12]. The CNN DeepPatent—using word embeddings—was the best classifier for large datasets until recently. Then, in 2020, Lee and Hsiang used a Bidirectional Encoder Representation from Transformers (BERT) model to develop the most state of the art patent classification model yet[18]. Given that BERT is the current state of the art model for NLP-related tasks in the intellectual property space, extra detail will be devoted to explaining its intricacies.

BERT—at its most basic construction—is a series of encoders[19]. The architecture of each encoder is roughly equivalent to that outlined in the landmark paper on the attention capabilities of transformers[20]. With respect to model training, BERT's training can be divided into two different distinct parts: pretraining and fine-tuning. In pretraining, the encoders are trained by learning on two tasks simultaneously: Next Sentence Prediction (NSP) and Masked Language Modeling (MLM). Simply put, NSP calculates the probability that sentence X is followed by Sentence Y, while MLM is a technique that determines the probability distribution of words for a fill-in-the-blank task[19]. These encoders produce word embedding vectors that roughly correspond to the contextual meaning of words.

Because of the tremendous computational cost of pretraining the series of encoders in BERT (estimates have put the training cost of a the released pre-trained BERT model at $6,912, due of the computational requirements[21]), it is typically more efficient for researchers to start with the pre-trained model and do the fine-tuning required in order to tackle their domain-specific problem. In the fine-tuning stage of BERT, only the fully connected output layer needs to be retrained for the specific task at hand (whether Q&A, summarizing text, sentiment analysis, classification, etc.). During this fine-tuning phase, the internal parameters of the encoders are only slightly tweaked. This fine-tuning process is exactly what Lee et al. undertook to arrive at their final model[19].

In order to ensure comparability to the previous patent classification literature, Lee et al. selected to classify based on the 639 sub-classes of the CLEF-IP dataset. In all, over 3-million different patents were used. After fine-tuning had been completed, Lee et al. had successfully created "PatentBERT". Previously, the precision score for DeepPatent in this multi-level classification task was 74%. Using the same text features from the patents, PatentBERT was able to achieve a precision of 82%: an increase in precision of 8%. Moreover, while "claims" were previously considered to be inconsequential to the resulting classification ability of a model, adding them to the model allowed PatentBERT to increase that precision to 85%. This indicates that the long form text data that was previously uninterpretable by other models was understood and was therefore a useful feature for classification with PatentBERT. Precision aside, the highest F1 score achieved by DeepPatent was less than 43%, while PatentBERT achieve an F1 score above 66%[19].

Conclusion

Despite the difficulties associated with the datasets in the field, substantial progress has been made in automating intellectual property related tasks over the past decade. The need for this automation continues to increase rapidly due to accruing patents. Looking toward the future, there exist several research avenues through which intellectual property analytics could be improved. Any technology that is able to better perform image classification and evaluation will aid in patent-related tasks.

Unfortunately, due to the colorlessness and atypical nature of the images found in patents, the ceiling on this activity is likely low relative to other computer vision tasks. On the other hand, patents contain rich text data. Yes, many of the words in patents are uncommon jargon. However, the ability to pre-train word embeddings in other domains and transfer these embeddings for use in IP-specific applications means that rich contextual understanding of patents may be achievable using millions of data points and creative neural network architectures that incorporate attention. Ultimately, this means that more tasks will be automated and done more thoroughly, resulting in better decision-making in patent offices and board rooms alike.

References

[1] Bergquist, K., Khan, M., Lamb, R., Le Feuvre, B., Letnikava, A. and Zhou, H., 2020. World Intellectual Property Indicators 2020. [ebook] Geneva: World Intellectual Property Organization, p.27. Available at: <https://www.wipo.int/edocs/pubdocs/en/wipo_pub_941_2020.pdf> [Accessed 24 October 2021].

[2] Trappey, Amy & Lupu, Mihai & Stjepandic, Josip. (2020). Embrace artificial intelligence technologies for advanced analytics and management of intellectual properties. World Patent Information. 61. 101970. 10.1016/j.wpi.2020.101970.

[3] C. J. Fall, A. Törcsvári, K. Benzineb, and G. Karetka. 2003. Automated categorization in the international patent classification. SIGIR Forum 37, 1 (Spring 2003), 10–25. DOI:https://doi.org/10.1145/945546.945547

[4] Milan Agatonovic, Niraj Aswani, Kalina Bontcheva, Hamish Cunningham, Thomas Heitz, Yaoyong Li, Ian Roberts, and Valentin Tablan. 2008. Large-scale, parallel automatic patent annotation. In Proceedings of the 1st ACM workshop on Patent information retrieval (PaIR '08). Association for Computing Machinery, New York, NY, USA, 1–8. DOI:https://doi.org/10.1145/1458572.1458574

[5] Cassidy, Caitlin. (2020). Parameter tuning Naïve Bayes for automatic patent classification. World Patent Information. 61. 101968. 10.1016/j.wpi.2020.101968.

[6] Murata, Masaki & Kanamaru, Toshiyuki & Shirado, Tamotsu & Isahara, Hitoshi. (2005). Using the k nearest neighbor method and bm25 in the patent document categorization subtask at ntcir-5.

[7] Aristodemou, Leonidas & Tietze, Frank. (2018). The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data. World Patent Information. 55. 37-51. 10.1016/j.wpi.2018.07.002.

[8] Harris, Steve & Trippe, Anthony & Challis, David & Swycher, Nigel. (2020). Construction and evaluation of gold standards for patent classification—A case study on quantum computing. World Patent Information. 61. 101961. 10.1016/j.wpi.2020.101961.

[9] Trappey, Amy & Trappey, Charles & Wu, Chun-Yi & Lin, Chi-Wei. (2012). A patent quality analysis for innovative technology and product development. Advanced Engineering Informatics. 26. 26-34. 10.1016/j.aei.2011.06.005.

[10] Hido, Shohei & Suzuki, Shoko & Nishiyama, Risa & Imamichi, Takashi & Takahashi, Rikiya & Nasukawa, Tetsuya & Id^|^eacute, Tsuyoshi & Kanehira, Yusuke & Yohda, Rinju & Ueno, Takeshi & Tajima, Akira & Watanabe, Toshiya. (2012). Modeling Patent Quality: A System for Large-scale Patentability Analysis using Text Mining. Journal of Information Processing. 20. 655-666. 10.2197/ipsjjip.20.655.

[11] Vrochidis, Stefanos & Moumtzidou, Anastasia & Kompatsiaris, Ioannis. (2012). Concept-based patent image retrieval. World Patent Information. 34. 292-303. 10.1016/j.wpi.2012.07.002.

[12] Lupu, Mihai. (2017). Information retrieval, machine learning, and Natural Language Processing for intellectual property information. World Patent Information. 49. A1-A3. 10.1016/j.wpi.2017.06.002.

[13] Hido, Shohei & Suzuki, Shoko & Nishiyama, Risa & Imamichi, Takashi & Takahashi, Rikiya & Nasukawa, Tetsuya & Ideacute, Tsuyoshi & Kanehira, Yusuke & Yohda, Rinju & Ueno, Takeshi & Tajima, Akira & Watanabe, Toshiya. (2012). Modeling Patent Quality: A System for Large-scale Patentability Analysis using Text Mining. Journal of Information Processing. 20. 655-666. 10.2197/ipsjjip.20.655.

[14] Venugopalan, Subhashini & Rai, Varun. (2014). Topic based classification and pattern identification in patents. Technological Forecasting and Social Change. 94. 10.1016/j.techfore.2014.10.006.

[15] Suominen, A., Toivanen, H., & Seppänen, M. (2017). Firms' knowledge profiles: Mapping patent data with unsupervised learning. Technological Forecasting and Social Change, 115, 131-142.

[16] Govindarajan, Usharani Hareesh & Trappey, Amy & Trappey, Charles. (2018). Immersive Technology for Human-Centric Cyberphysical Systems in Complex Manufacturing Processes: A Comprehensive Overview of the Global Patent Profile Using Collective Intelligence. Complexity. 2018. 10.1155/2018/4283634.

[17] De Clercq, Djavan & Diop, Ndeye-Fatou & Jain, Devina & Tan, Benjamin. (2019). Multi-label classification and interactive NLP-based visualization of electric vehicle patent data. World Patent Information. 58. 101903. 10.1016/j.wpi.2019.101903.

[18] Lee, Jieh-Sheng & Hsiang, Jieh. (2020). Patent classification by fine-tuning BERT language model. World Patent Information. 61. 101965. 10.1016/j.wpi.2020.101965.

[19] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL*.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[21] Wiggers, K. (2021, February 10). *Neureality emerges from stealth to accelerate AI workloads at scale*. VentureBeat. Retrieved October 27, 2021, from https://venturebeat.com/2021/02/10/neureality-emerges-from-stealth-to-accelerate-ai-workloads-at-scale/.