Learning Patent Speak: Investigating Domain-Specific Word Embeddings

Julian Risch
Hasso Plattner Institute
University of Potsdam
Potsdam, Germany
julian.risch@hpi.de

Ralf Krestel
Hasso Plattner Institute
University of Potsdam
Potsdam, Germany
ralf.krestel@hpi.de

***Abstract*—A patent examiner needs domain-specific knowledge to classify a patent application according to its field of invention. Standardized classification schemes help to compare a patent application to previously granted patents and thereby check its novelty. Due to the large volume of patents, automatic patent classification would be highly beneficial to patent offices and other stakeholders in the patent domain. However, a challenge for the automation of this costly manual task is the patent-specific language use. To facilitate this task, we present domain-specific pre-trained word embeddings for the patent domain. We trained our model on a very large dataset of more than 5 million patents to learn the language use in this domain. We evaluated the quality of the resulting embeddings in the context of patent classification. To this end, we propose a deep learning approach based on gated recurrent units for automatic patent classification built on the trained word embeddings. Experiments on a standardized evaluation dataset show that our approach increases average precision for patent classification by 17 percent compared to state-of-the-art approaches.**

***Index Terms*—Document Classification, Deep Learning, Word Embedding, Patents**

I. Introduction

In 2017, a record number of 320,003 U.S. patents has been granted by the U.S. Patent and Trademark Office[1]. All granted U.S. patents since 1976 are publicly available as full text[2]. These large text collections represent an extensive amount of human knowledge in an almost unstructured form. This makes mining information from them challenging and automatic classification and retrieval a hard problem.

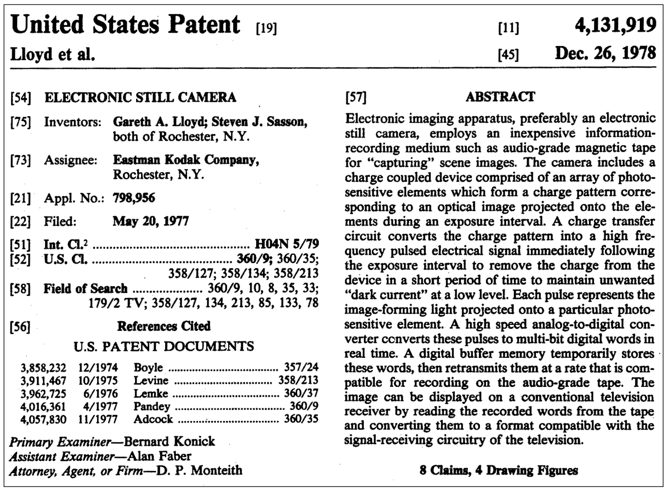
Not only the number of documents but also the patent-specific vocabulary make the tasks more difficult. Because of the underlying legal purpose of patent documents, they follow a specific writing style. Patent applications need to define the scope of an invention and need to delimit an invention from others whilst covering as much variation of the invention as possible. As a consequence, descriptions of an invention use vague language. For example, a patent calls an invention “electronic still camera” and “electronic imaging apparatus”, whereas such a device is called “digital camera” in colloquial speech (Fig. 1). A patent’s claims are a controversial subject, because a patent grants rights and also limits the rights of others. Patents grant a monopoly for a limited time in exchange for the disclosure of the invention so that others can license it.

Unstructured text sections, such as abstracts, descriptions, and claims, make up the largest part of a patent. The claims section is essential for defining the scope of an invention. It describes the extent of the monopoly rights granted by the patent. Court decisions of the past precisely define the meaning of “patent speak”. An example are the slight differences of “consist of” and “comprise”[3]: “consist of” implies an exhaustive enumeration, whereas “comprise” commences an enumeration that is not necessarily exhaustive. Classifying patents is challenging because of patent-specific language use — even for domain experts.

The International Patent Classification (IPC) is a hierarchical classification system for patents. It has been periodically revised and adapted to the upcoming of new fields of invention. The system considers 4 levels of hierarchy: sections, classes, subclasses, and group. For example, the U.S. patent no. 4131919 with the IPC code H04N 1/21 is in group H04N 1/21, which is in the subclass H04N, the class H04, and section H, The subparts of this code correspond to the section “electricity”, class “electric communication technique”, subclass “pictorial communication, e.g. television”, and group “Intermediate information storage”. An excerpt of this patent is depicted in Fig. 1 with the depricated IPC code H04N 005/79.

This complicated classification system is applied at several different steps in the patenting process. On the one hand, patent applicants need to search for prior art, if they file a patent. They need to retrieve patents about similar inventions although they might use different words for description. On the other hand, patent examiners in a patent office need to check a patent application for its “inventive step or non-obviousness” and its “novelty”. A patent examiner specialized in the field of the invention needs to be matched to the patent application. Finally, patent courts and patent attorneys deal with the infringement and validity of granted patents. All three scenarios involve an information retrieval task, where patents similar to a given patent need to be found. Based on their

[1] https://www.ificlaims.com/rankings.htm
[2] https://bulkdata.uspto.gov/
[3] https://www.epo.org/law-practice/legal-texts/html/guidelines/e/f_iv_4_21.htm



United States Patent [19] [11] **4,131,919**

Lloyd et al. [45] **Dec. 26, 1978**

[54] **ELECTRONIC STILL CAMERA**

[75] Inventors: **Gareth A. Lloyd; Steven J. Sasson,** both of Rochester, N.Y.

[73] Assignee: **Eastman Kodak Company,** Rochester, N.Y.

[21] Appl. No.: **798,956**

[22] Filed: **May 20, 1977**

[51] **Int. Cl.**^{2} **H04N 5/79**

[52] **U.S. Cl.** **360/9;** 360/35; 358/127; 358/134; 358/213

[58] **Field of Search** .... 360/9, 10, 8, 35, 33; 179/2 TV; 358/127, 134, 213, 85, 133, 78

[56] **References Cited**

U.S. PATENT DOCUMENTS

3,858,232	12/1974	Boyle	357/24
3,911,467	10/1975	Levine	358/213
3,962,725	6/1976	Lemke	360/37
4,016,361	4/1977	Pandey	360/9
4,057,830	11/1977	Adcock	360/35

Primary Examiner—Bernard Konick
Assistant Examiner—Alan Faber
Attorney, Agent, or Firm—D. P. Monteith

[57] **ABSTRACT**

Electronic imaging apparatus, preferably an electronic still camera, employs an inexpensive information-recording medium such as audio-grade magnetic tape for "capturing" scene images. The camera includes a charge coupled device comprised of an array of photosensitive elements which form a charge pattern corresponding to an optical image projected onto the elements during an exposure interval. A charge transfer circuit converts the charge pattern into a high frequency pulsed electrical signal immediately following the exposure interval to remove the charge from the device in a short period of time to maintain unwanted "dark current" at a low level. Each pulse represents the image-forming light projected onto a particular photosensitive element. A high speed analog-to-digital converter ccnverts these pulses to multi-bit digital words in real time. A digital buffer memory temporarily stores these words, then retransmits them at a rate that is compatible for recording on the audio-grade tape. The image can be displayed on a conventional television receiver by reading the recorded words from the tape and converting them to a format compatible with the signal-receiving circuitry of the television.

8 Claims, 4 Drawing Figures

Fig. 1. Patent documents follow a standardized structure and consist of several fields, such as title, abstract, and claims, but also references. This example is an excerpt of U.S. patent no. 4131919.

similarity, similar patents mutually limit their scopes.

The IPC systematically classifies patents into topical subclasses. Thereby the retrieval of similar patents can be performed by looking up patents in the same subclass. However, manually classifying patents into such subclasses is costly in terms of working power and needs domain-specific knowledge due to the complexity of the IPC. The goal of automated patent classification is to save these costs and associate a given patent document with its correct subclasses automatically. Smith summarizes the applications of automated patent classification as (1) matching patent applications with a patent examiner who is a domain expert for the field of invention, (2) classification of external documents so that they can easily be retrieved during the patent examination process, and reclassification of older patents labeled with outdated classification schemes [1]. In practice, patents can be associated with multiple subclasses. Therefore, patent classification is not a multi-class but a multi-label classification task. In fact, our example patent in Fig. 1 is associated with 2 IPC subclasses. In total, the IPC knows 637 subclasses.

In this paper, we propose to improve automatic patent classification by leveraging recent deep learning techniques. In particular, we train fastText word embeddings on a large dataset of more than 5 million patents. We use these embeddings together with bi-directional Gated Recurrent Units (GRUs) to classify patents. Experiments show that our approach is superior to state-of-the-art approaches in terms of three evaluation measures. For example, we increase micro-average precision at predicting a patent's subclass by 17 percent. Further, we find that domain-specific word embeddings trained on patent documents outperform standard word embeddings trained on Wikipedia pages by 9 percent when combined with a GRU-based neural network.

Our contribution is thus twofold:

1) Computation of word embeddings on the second largest corpus ever used for training and providing these word embeddings for download[4].
2) Proposing a deep neural network architecture based on bi-directional Gated Recurrent Units (GRUs) for patent classification.

Section II summarizes related work in the field of automatic patent classification and gives an overview of different word embedding approaches. The three datasets used in this paper are described in Section III and Section IV describes our approach to capture semantics in patent language by domain-specific word embeddings and automatically classify patents. We evaluate our approach with three experiments in Section V and conclude in Section VI.

II. Related Work

Fall et al. established a collection of around 75,000 excerpts of English-language patent applications as a de-facto standard dataset for the evaluation of automatic patent classification [2]. The dataset is called WIPO-alpha[5] and is provided by the World Intellectual Property Office (WIPO). Fall et al. further propose three evaluation measures that are tailored to the patent classification task, where a patent is typically associated with a main subclass, but also with several incidental subclasses. We apply the three measures in our experiments and describe them in detail in Section V. In general, the micro-precision of assigning the correct class to a given patent is evaluated.

Seneviratne et al. propose to generate signatures from patents instead of using the full vocabulary as features [3]. They evaluate their patent classification approach on IPC class level (114 classes) and subclass level (451 subclasses) on the WIPO-alpha dataset. While they improve classification performance in comparison to Fall et al., they optimize also the time required to index and search a patent collection. Other results on the WIPO-alpha dataset have been published by Nguyen (macro-f1: 0.452, micro-f1: 0.755) [4], Rousu et al. (micro-f1: 0.767) [5], and Qiu et al. (macro-f1: 0.418) [6].

Several researchers conducted their experiments on other datasets, which makes a direct comparison with their results impossible. An ensemble of different classifiers slightly improves micro-F1 score on a refined version of the WIPO-alpha dataset according to Mathiassen and Ortiz-Arroyo [7]. They report a micro-f1 of 0.867. Instead of IPC, Tran and Kavuluru use the Cooperative Patent Classification (CPC) system, which replaces the earlier the U.S. Patent Classification (USPC) system [8]. They report a micro-f1 of 0.700 on a dataset of patents with 654 subclasses. Dhondt et al. report a micro-f1 0.751 and a micro-precision of 0.800 on a subset of 532,264 English abstracts from the so called CLEF-IP 2010 corpus [9].

The CLEF-IP 2010 corpus from the Conference and Labs of the Evaluation Forum's track for retrieval experiments in the intellectual property domain (CLEF-IP) considered two

[4] https://hpi.de/naumann/projects/repeatability/text%2Dmining.html

[5] WIPO-en-alpha dataset, World Intellectual Property Office, Geneva, Switzerland, 2002

tasks: (1) recommending patents as prior art for another patent and (2) patent classification according to the International Patent Classification system (IPC). Verberne and Dhondt find that using not only abstracts but also full description texts improves classification performance [10]. With regard to the usefulness of metadata, such as applicants, inventors, and address, they conclude that it does not improve classification. This result contradicts Beney, who finds that applicant and address improves classification [11]. They argue that names and addresses identify companies, which work in restricted domains. Derieux et al., results are language specific, classification on English patents is at least 10% better than on German and French patents [12]. The observation that language has a strong influence on the classification motivates further investigation of patent-specific language use. In this paper, we consider to model this language use with patent-specific word embeddings. To this end, we summarize work in the field of word embeddings.

The upcoming of word embeddings or, more general speaking, dense vector representations to capture the semantic meaning of words influences many natural language processing tasks. With Word2Vec, Mikolov et al. propose an efficient way to train word embeddings [13]. As a consequence, they are able to train embeddings on large datasets with billions of words. A similar approach, termed global vectors (GloVe), trains word embeddings on global word-word co-occurrence counts rather than on context windows of limited size [14]. A disadvantage of both Word2Vec and GloVe is the inherent out-of-vocabulary problem: a word that occurs only in the test data but not in the training data has no vector representation in the word embedding space. To overcome this problem, Bojanowski et al. introduce another context-window-based approach, which they call fastText [15]. fastText word embeddings incorporate information about character n-grams as subparts of a word. As a consequence, they overcome the out-of-vocabulary problem of other word embedding approaches by falling back to embeddings of character n-grams if a word is unknown.

Recently, deep learning approaches for patent classification have been proposed. Xia et al. outline a general deep learning approach for patent classification based on sparse auto-encoders and deep belief networks [16]. However, their proposal is limited to a theoretical approach and lacks practical experiments. Grawe et al. automatically classify patents based on word embeddings and long-short term memory units (LSTMs) in a neural network [17]. Their approach is similar to ours but has several limitations: (1) it considers only 50 different classes, (2) it achieves only 63% accuracy, and (3) as opposed to our approach it suffers from out-of-vocabulary problems, which is inherent to the applied Word2Vec model.

Instead of content-based approaches, which consider only a patent's text sections, Li et al. propose a citation-based approach [18]. They exploit co-citation relations among patents. Further, they leverage the fact that patents reference other patents in the same field to explain the novelty of their ideas. These references are not limited to patents, but also include scientific papers. Cross-collection topic models can be used to recommend references across these different document collections [19]. While these approaches can help to retrieve similar patents and can therefore be of help in the patenting process, we solely focus on content-based classification of patents in this paper. Similar to the IPC system in the patent domain is the Medical Subject Headings (MeSH) ontology in the medical domain. Eisinger et al. compare automatic document classification for the two classification schemes [20]. They leverage class co-occurrence frequencies to enrich labeled classes and propose a guided search as an application.

TABLE I
A Comparison of the Three Patent Datasets

Dataset	# Documents	# Tokens
WIPO-alpha	75,250	561 million
USPTO-2M	2 million	235 million
USPTO-5M	5 million	38 billion

III. Dataset

In this paper, we consider three different datasets of patent documents. Tab. I gives an overview of the datasets, their number of documents, and their number of tokens. The first dataset is the WIPO-alpha dataset established by Fall et al., which is a de-facto standard for the evaluation of automated patent classification and has been widely used [2]–[6]. The dataset contains more than 75,000 patents with title, abstract, claims, and full description. Further, each patent is associated with a main subclass and incidental subclasses.

The second dataset is much larger and contains 5.4 million patents granted by the United States Patent and Trademark Office (USPTO). The USPTO keeps records of all U.S. patent activity since 1790. On their website[6], they provide free bulk downloads of full text patent publications from 1976 to 2016. We use this full dataset and refer to it as USPTO-5M. Each patent contains bibliographic data, such as title, inventor, owner, filing date, and granting date. Furthermore, author information, patent type classification, claims, abstract, links to other patents or papers, and a detailed description of the invention are provided. For our experiments, we focus on textual data and leave out figures and their captions. In comparison to WIPO-alpha, USPTO-5M is 70 times larger in terms of number of documents and also number of tokens.

The third dataset is called USPTO-2M and contains 2 million patents. It is publicly available online[7] in a preprocessed JSON format so that other researchers can use it easily. The dataset is split into a training set with 1.95 million documents and a test set with the remaining 50,000 documents. Further, it is limited to titles, abstracts, document identifiers, and subclasses. In total, there are 637 subclasses. In contrast to WIPO-alpha, USPTO-2M does not distinguish between main subclass and incidental subclasses.

[6] https://www.uspto.gov/learning%2Dand%2Dresources/electronic%2Dbulk%2Ddata%2Dproducts

[7] http://mleg.cse.sc.edu/DeepPatent/

IV. Deep Learning for Patent Classification

Our goal is to automatically classify patents into their assigned subclasses. The large amount of available patents and their full text plus the recent success of deep learning for natural language processing motivate to investigate deep learning for patent classification. To this end, we propose to use word embeddings to capture the semantics of the specific language that is used in patents. Further, we propose a neural network architecture to automatically classify patents based on the inferred word embeddings.

A. *Domain-Specific Word Embeddings*

Word embeddings are a basic ingredient for a variety of tasks in natural language processing. They represent words as dense vectors in a vector space. Pre-trained on a large number of tokens, relations of these representations in a vector space can mirror semantic relations of words [13].

We propose to train fastText word embeddings based on the method by Bojanowski et al. [15] with 100, 200, and 300 dimensions. We transform all characters to lowercase and discard all words that occur less than ten times. The used context window size is 5.

We train the embeddings on our dataset USPTO-5M, which contains 38 billion tokens and publish the resulting word embeddings online[8]. To the best of our knowledge, this is the second largest number of tokens ever used to train word embeddings. It contains more than twice the number of tokens of the English Wikipedia (16 billion) and is only exceeded by the Common Crawl dataset, which consists of 600 billion tokens. We assume that the embeddings are helpful not only for patent classification but also for other tasks in the patent domain and hope that other researchers can build on our results.

B. *Neural Network Architecture*

Given a patent document, our goal is to infer its main subclass and also potential incidental subclasses. We investigate how domain-specific word embeddings can help to solve this classification problem. Therefore, we extract a patent document's title and abstract and consider only the sequence of the first 300 words. We choose this limitation to be comparable to related work in our evaluation [2], [3]. Longer sequences linearly increase runtime and memory need.

Fig. 2 visualizes the network architecture. For each word in the input sequence, we calculate its word embedding based on our pre-trained, domain-specific fastText model. This sequence of word embeddings is processed by a spatial dropout, which randomly masks 10% of the input words to make the neural network more robust. The remaining 90% of the sequence serve as input to the next layer in the neural network. In particular, we propose a deep neural network architecture based on gated recurrent units (GRUs). We use bi-directional GRUs so that the input sequence is processed in two directions: correct order and reverse order of the words. The outputs of these two processing steps are averaged and followed by a dropout of 10%, again to make the network more robust.

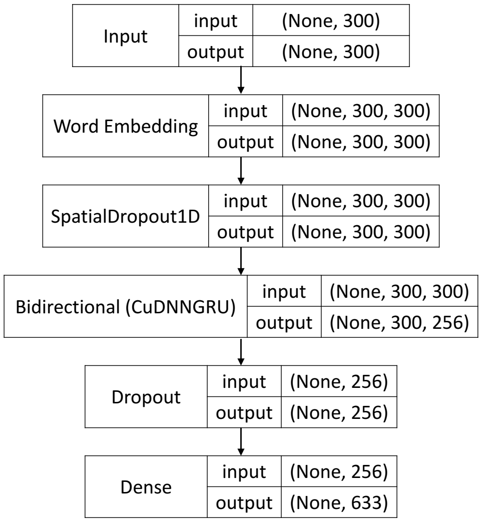


Fig. 2. The neural network uses pre-trained word embeddings, spatial dropout, GRUs, dropout, and a dense layer with softmax activation.

In Section I, we pointed out that patent classification is not a multi-class but a multi-label classification task. A typical final layer of our neural network would therefore be a dense layer with as many units as subclasses and a sigmoid activation. Instead, we use a dense layer with as many units as subclasses and a softmax activation. Thereby, we train the model for the multi-class classification task only and aim to predict a patent's main subclass. For training the neural network, the softmax activation together with a categorical loss function considers only a single subclass as correct. If our model predicts any other subclass, such as any incidental subclasses, the prediction is considered wrong during training.

However, during testing, we consider the probabilities output by the softmax activation for all subclasses. We consider the top three subclasses with the highest probabilities as our final prediction. Although the neural network is trained to predict only the main subclass and not the incidental subclasses, our experiments in Section V show that the model achieves competitive results for both tasks.

Training of the neural network until conversion takes 13 epochs with a batch size of 256. With a larger batch size, more subclasses are covered in a particular epoch. The more diverse set of subclasses potentially prevents the model from optimizing for a small subset of all subclasses per epoch only. However, we find no significant difference in classification performance if we train the model with a batch size of 32 until convergence for 5 epochs. We assume that smaller batches, which cover less subclasses, have no negative effect on classification performance at our task, but we did not conduct experiments to further investigate this matter.

V. Experiments

For our experiments, we use three evaluation measures as proposed by Fall et al. [2]. Fig. 3 visualizes the three

[8] https://hpi.de/naumann/projects/repeatability/text%2Dmining.html

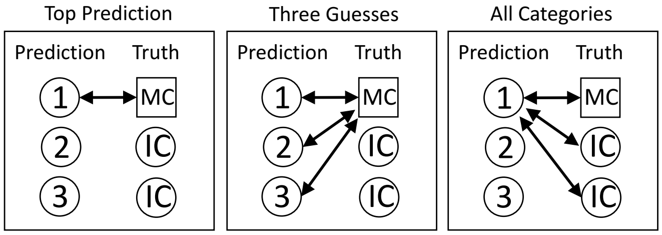


Fig. 3. Three evaluation measures for the task of patent classification. The predicted, ranked subclasses are compared to the ground truth main subclass (MC) and the incidental subclasses (IC). (adapted from Fall et al. [2])

TABLE II
A Comparison of Micro-Average Precision for Different Numbers of Word Embedding Dimensions on the WIPO-alpha dataset.

Evaluation Measure	Word Embedding Dimensions			
	100-patent	**200-patent**	**300-patent**	**300-wiki**
Top-Prediction	45%	48%	**49%**	42%
Three-Guesses	70%	**72%**	**72%**	67%
All-Categories	54%	56%	**57%**	50%

evaluation measures and how they differ in comparing the ranked, top three predicted subclasses with the ground truth main subclass (MC) and incidental subclasses (IC). These measures are tailored to the practical application of patent classification. The measure "top prediction" compares only the top-ranked prediction to the main subclass. The measure "three guesses" compares not only the top-ranked but the three top-ranked ranked predictions to the main subclass. The prediction is successful if one of the top three predictions matches the ground truth main subclass. Both measures, "top prediction" and "three guesses" evaluate only a multi-class classification task. In contrast, the measure "all categories" considers also the incidental subclasses as ground truth information and thus evaluates based on a multi-label ground truth. The measure checks whether the top prediction is included in the set of the main subclass and all incidental subclasses. In theory, this set could contain more than three subclasses. However, in practice the set contains less than two subclasses on average.

We run three experiments to show that our domain-specific word embeddings are beneficial for the task of patent classification. In the first experiment, we compare the classification performance of four different approaches. Three of them use our pre-trained, patent-specific word embeddings and differ only in the number of word embedding dimensions (either 100, 200, or 300). The fourth approach uses generic 300-dimensional word embeddings trained on Wikipedia pages. All four approaches have the same neural network architecture as described in Section IV-B. We use the WIPO-alpha dataset and apply the three evaluation measures: "top prediction", "three guesses", and "all categories".

Tab. II lists the results of our first experiment. The patent-specific word embeddings, which we trained on 38 billion tokens, outperform word embeddings trained on English Wikipedia pages. This superiority holds if we use 300-dimensional word embeddings for both approaches. However, if we train domain-specific word embeddings with only 100-dimensional vectors, 300-dimensional word embeddings trained on Wikipedia are almost as good as domain-specific word embeddings. The performance of 200- and 300-dimensional domain-specific word embeddings differ only slightly.

The second experiment compares our best model to state-of-the-art approaches for patent classification to show that our approach achieves competitive results. To this end, we use the same experiment setup as Fall et al., again on the WIPO-alpha dataset and are thereby able to compare with results reported in related work [2], [3]. Tab. III lists the results of our second experiment. Our best model with domain-specific word embeddings outperforms the best other approach by up to 17 percent (42 percent compared to 49 percent precision).

The third experiment evaluates our approach on a more recent and larger dataset than WIPO-alpha, called USPTO-2M. Unfortunately, this dataset does not distinguish between main subclasses and incidental subclasses. For training our approach, we consider the first listed subclass of each patent as its main subclass. For the majority of patents only one subclass is listed anyways.

The measure "all categories" is not influenced by the fact that the dataset does not explicitly list main subclasses. Both other measures, "top prediction" and "three guesses", can only be approximated, because we can only guess the ground truth main subclass out of the set of all listed subclasses. Another limitation of the dataset is that it does not contain patents' full texts but only their abstracts and titles. However, the WIPO-alpha and the USPTO-2M dataset are still quite similar and we assume that the task of patent classification is equally difficult on both datasets. We use the patents of the years 2006 to 2013 as training data and the patents of the year 2014 as test data.

Because of the size of the dataset and memory constraints during training, we can only process the first 30 words of each patent (instead of the first 300 words as in our other experiments). For the same reason, we can use only 100-dimensional and no 300-dimensional word embeddings. Tab. IV lists the results of our third experiment. Surprisingly, the classification results are even better on the USPTO-2M dataset with the limited approach than on the WIPO-alpha dataset with our more complex approach. The USPTO-2M dataset contains 25 times more training samples than the WIPO-alpha dataset. We assume that the larger number of training samples is the main reason for the model's strong performance.

Together, the three experiments show that domain-specific word embeddings together and a GRU-based neural network achieve competitive results at the task of patent classification. In particular, patent-specific word embeddings outperform generic word embeddings trained on Wikipedia pages. However, memory constraints during training limit our approach for the USPTO-2M dataset.

TABLE III
A COMPARISON OF MICRO-AVERAGE PRECISION FOR STATE-OF-THE-ART APPROACHES [2], [3] AND OUR NEURAL NETWORK WITH WIKIPEDIA WORD EMBEDDINGS (RNN-WIKI) AND PATENT WORD EMBEDDINGS (RNN-PATENT) ON THE WIPO-ALPHA DATASET.

Evaluation Measure	Naive Bayes [2]	k-NN [2]	SVM [2]	SNoW [2]	k-NN [3]	RNN-wiki	RNN-patent
Top-Prediction	33%	39%	41%	36%	42%	45%	**49%**
Three-Guesses	53%	62%	59%	56%	67%	69%	**72%**
All-Categories	41%	46%	48%	43%	50%	53%	**57%**

TABLE IV
MICRO-AVERAGE PRECISION FOR OUR NEURAL NETWORK WITH PATENT WORD EMBEDDINGS (RNN-PATENT) WITH 100 DIMENSIONS (LIMITED TO THE FIRST 30 WORDS OF EACH PATENT) ON THE USPTO-2M DATASET.

Evaluation Measure	RNN-patent
Top-Prediction	53%
Three-Guesses	75%
All-Categories	64%

VI. CONCLUSIONS

In this paper, we studied the task of automatic patent classification. We proposed to apply domain-specific fastText word embeddings, which we trained on a large dataset of full texts of more than 5 million patents. Based on these word embeddings that capture the special characteristics of patent speak, we trained a deep neural network with GRUs. Our model is trained with a softmax activation for the task of multi-class classification but is applicable also for multi-label classification. We evaluate our approach with three standard measures in three experiments and improve micro-average precision by 17 percent compared to the state-of-the-art. Further, we find that domain-specific word embeddings, trained specifically on patent documents, outperform generic word embeddings trained in Wikipedia pages. We publish our trained word embeddings and hope that other researchers can profit from the improved semantic representation of patent language. A path for future work is the application of deep learning approaches to other tasks that involve natural language processing in the patent domain, such as classic patent retrieval or reference recommendation. These approaches can surely benefit from pre-trained, domain-specific word embeddings that capture patent speak. Further, an investigation of new neural network architectures tailored to the needs of the patent domain and its hierarchical classification system is promising.

REFERENCES

[1] H. Smith, “Automation of patent classification,” *World Patent Information*, vol. 24, no. 4, pp. 269–271, 2002.

[2] C. J. Fall, A. Törcsvári, K. Benzineb, and G. Karetka, “Automated categorization in the international patent classification,” in *Acm Sigir Forum*, vol. 37, no. 1. ACM, 2003, pp. 10–25.

[3] D. Seneviratne, S. Geva, G. Zuccon, G. Ferraro, T. Chappell, and M. Meireles, “A signature approach to patent classification,” in *Asia Information Retrieval Symposium*. Springer, 2015, pp. 413–419.

[4] N. Nguyen, “Improving hierarchical classification with partial labels.” in *ECAI*, 2010, pp. 315–320.

[5] J. Rousu, C. Saunders, S. Szedmak, and J. Shawe-Taylor, “Kernel-based learning of hierarchical multilabel classification models,” *Journal of Machine Learning Research*, vol. 7, no. Jul, pp. 1601–1626, 2006.

[6] X. Qiu, X. Huang, Z. Liu, and J. Zhou, “Hierarchical text classification with latent concepts,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, 2011, pp. 598–602.

[7] H. Mathiassen and D. Ortiz-Arroyo, “Automatic categorization of patent applications using classifier combinations,” in *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, 2006, pp. 1039–1047.

[8] T. Tran and R. Kavuluru, “Supervised approaches to assign cooperative patent classification (cpc) codes to patents,” in *International Conference on Mining Intelligence and Knowledge Exploration*. Springer, 2017, pp. 22–34.

[9] E. D’hondt, S. Verberne, C. Koster, and L. Boves, “Text representations for patent classification,” *Computational Linguistics*, vol. 39, no. 3, pp. 755–775, 2013.

[10] S. Verberne and E. D’hondt, “Patent classification experiments with the linguistic classification system lcs,” in *CLEF (Notebook Papers/Labs/Workshop)*, 2010.

[11] J. Beney, “Lci-insa linguistic experiment for clef-ip classification track,” in *CLEF (Notebook Papers/Labs/Workshops)*, 2010.

[12] F. Derieux, M. Bobeica, D. Pois, and J.-P. Raysz, “Combining semantics and statistics for patent classification,” in *CLEF*, 2010.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.

[14] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[15] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[16] B. Xia, L. Baoan, and X. Lv, “Research on patent document classification based on deep learning,” in *Proceedings of the International Conference on Artificial Intelligence and Industrial Engineering (AIIE)*, 2016.

[17] M. F. Grawe, C. A. Martins, and A. G. Bonfante, “Automated patent classification using word embedding,” in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec 2017, pp. 408–411.

[18] X. Li, H. Chen, Z. Zhang, J. Li, and J. F. Nunamaker, “Managing knowledge in light of its evolution process: An empirical study on citation network-based patent classification,” *Journal of Management Information Systems*, vol. 26, no. 1, pp. 129–154, 2009.

[19] J. Risch and R. Krestel, “My approach = your apparatus?” in *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, ser. JCDL ’18. New York, NY, USA: ACM, 2018, pp. 283–292.

[20] D. Eisinger, G. Tsatsaronis, M. Bundschus, U. Wieneke, and M. Schroeder, “Automated patent categorization and guided patent search using ipc as inspired by mesh and pubmed,” in *Journal of biomedical semantics*, vol. 4, no. 1. BioMed Central, 2013, p. S3.