Universal Language Model Fine-tuning for Patent Classification

Jason Hepburn
Macquarie University
Sydney, Australia
`jason.hepburn@students.mq.edu.au`

Abstract

This paper describes the methods used for the 2018 ALTA Shared Task. The task this year was to automatically classify Australian patents into their main International Patent Classification section. Our final submission used a Support Vector Machine (SVM) and Universal Language Model with Fine-tuning (ULMFiT). Our system achieved the best results in the student category.

1 Introduction

For the last nine years the Australasian Language Technology Association (ALTA) has run a shared task competition for students. This year the shared task is to classify patent applications into their primary section code (Mollá and Seneviratne, 2018).

Patent applications are classified and compared to previous inventions in the field. Accurate classification of patents is crucial to patent officers, potential inventors, and industry. The patent classification process is dependant on human labour and with the rate of submissions increasing there is an ever greater need for an Automated Patent Classification (APC) system (Fall et al., 2003).

The International Patent Classification (IPC) has a tree structured class hierarchy (Silla and Freitas, 2011). At the highest level of this hierarchy is the IPC Section designated by the capital letters A to H (Table 1). Following the tree structure from Sections are Classes, Sub-classes, and Groups. There are approximately 69,000 different categories at the group level. The classification taxonomy is revised annually and previous patents can be reclassified (D'hondt et al., 2013).

Most patents have a main code in addition to a set of secondary codes. These secondary codes can be very distant to each other. For some codes it is obligatory to also assign other codes (eg. All `C12N` are also classed `A61P`). Codes can have *placement* rules defining a preference for one code when two may apply.

A	Human necessities
B	Performing operations, transporting
C	Chemistry, metallurgy
D	Textiles, paper
E	Fixed constructions
F	Mechanical engineering, lighting, heating, weapons, blasting
G	Physics
H	Electricity

Table 1: IPC Sections

At the semantic level all patents are different as they must describe a new idea or invention. Some terms, phrases, or acronyms can have very different meaning in different fields. Applicants try to avoid narrowing the scope of the invention and as such can use vague or general terms. As an example, pharmaceutical companies tend to describe every possible therapeutic use for an application. This can make it difficult to classify these patents.

We structure this paper as follows: Section 2 introduces related research for APC; Section 3 describes the data set provided for the competition; Section 4 describes the methods used; Section 5 presents and discusses the results; Section 6 concludes this paper.

2 Related works

With the need for reliable and efficient APC systems considerable research has been conducted in this area.

Fall et al. (2003) introduce the publicly available WIPO-alpha data set for patent classification (See section 3.2). They give a comprehensive description of the problem and much of its complex-

Jason Hepburn. 2018. Universal Language Model Fine-tuning for Patent Classification. In *Proceedings of Australasian Language Technology Association Workshop*, pages 93−96.

ities. One such complexity is the similarities of section `G` and `H` which are ”Physics” and ”Electricity” respectively. The authors give a detailed analysis of the classification errors between these two sections.

Various classification models are tested and compared including Naïve Bayes, K-Nearest Neighbours, and SVM. Fall et al. (2003) show that the best performing model is a SVM with a linear kernel using only the first 300 words of the document.

Benzineb and Guyot (2011) describe in great detail the task and challenges of APC. APC can be used to classify new applications as well as help with searches for similar prior art. Interestingly they noted that SVMs are more accurate than Neural Network approaches.

D'hondt et al. (2013) assess the use of statistical and linguistic phrases for patent applications. Adding phrases, particularly bigrams, to unigrams significantly improves classification.

Seneviratne et al. (2015) build on Falls work with a focus on improving the efficiency of classification. Dimensionality reduction is used in the form of a signature approach to reduce computation and enable a larger vocabulary. For top predictions a marginal improvement is made.

3 Data sets

In this section we describe the two data sets used by our system. The first data set is provided for the ALTA Shared task [1]. The second is the WIPO-alpha data set introduced by Fall et al. (2003).

3.1 ALTA

The data provided contains 4972 Australian patent applications. 3972 of them are part of the training set labelled with the main IPC section. The other 1000 applications in the test set are unlabelled.

The section counts are significantly unbalanced with the largest, section A, having 1303 compared to section D having 14 (see Figure 1).

3.2 WIPO-alpha

WIPO-alpha is a collection of patent applications form the World Intellectual Property Organization. The documents are all in English and published between 1998 and 2002. Each patent is a structured XML document. This allows for analysis of separate parts of the documents such as the title or abstract. Documents include the full IPC main classification as well as secondary classifications.

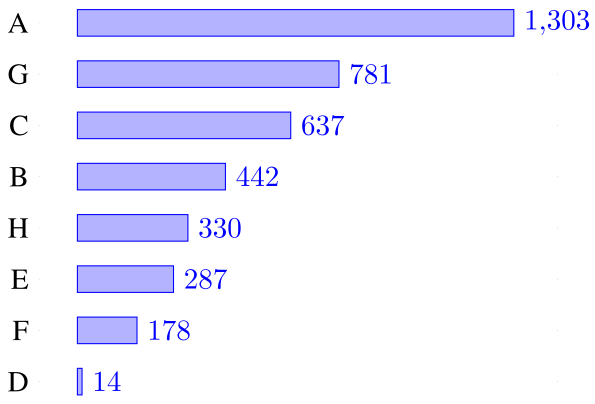

Figure 1: ALTA training set counts by IPC Section

There are 75,250 documents in the data set split into approximately 60% train and 40% test. The splitting of the train and test sets has tried to maintain an equal distribution IPC main group level.

4 Methodology

We used several statistical classifiers to complete this task. In this section we describe in detail the methods used and the steps they involved. Section 4.1 describes the pre-processing of the ALTA and WIPO-alpha data sets. Section 4.2 describes the SVM classifier motivated by Fall et al. (2003). Section 4.3 describes ULMFiT from Howard and Ruder (2018) and how it is adapted to this task. Section 4.5 describes the system used to deal with the classification errors between Section `G` and `H`.

4.1 Pre-processing

During the exploration of the data it was found that there is a large variation of document length. There are 48 documents in the ALTA training data set which contained only ”`NA parse failure`”. These documents were excluded from the training set and when found in the test set automatically classified as Section `A` which is the majority class. Looking closer at the large documents some contain long strings of DNA and amino acid sequences. The largest document appears to contain a large number of incorrectly encoded characters. Motivated by Fall et al. (2003), this and other noisy data is avoided by only using a small portion of the begining of the document.

[1] www.alta.asn.au/events/sharedtask2018

Patent documents from the WIPO data set are in XML format. These documents were converted into plain text to best replicate the format of the target ALTA documents. This was achieved by concatenating the document Title, Abstracts, and Claim.

4.2 SVM

For the SVM classifier we use the Python Scikit-learn (Pedregosa et al., 2011) library. Documents are indexed using term frequency-inverse document frequency (tf-idf) and only using the first 3500[2] characters. Motivated by D'hondt et al. (2013) we use unigrams and bigrams. As with the work of Fall et al. (2003) linear kernels for the SVM were found to perform best.

4.3 ULMFiT

Universal Language Model Fine-tuning (ULMFiT) is a transfer learning technique introduced by Howard and Ruder (2018). This technique uses the following three steps: a) General-domain language model pretrainig (4.3.1); b) Target task language model fine-tuning(4.3.2); and c) Target task classifier fine-tuning (4.3.3).

4.3.1 General-domain language model pretraining

The first step is to carry out unsupervised training of a language model on a large corpus to create a general-domain language model. As this step is not domain specific here we have used the pretrained model[3] from Howard and Ruder (2018). This model uses the state of the art language model AWD LSTM trained on Wikitext-103 (Merity et al., 2017)

4.3.2 Target task language model fine-tuning

The general-domain language model is then fine-tuned on data from the target task. The pretraining allows this stage to converge faster and results in a robust language model even for small datasets. A key advantage here is that words that are uncommon in the target training set retain robust representations from the pretraining. As this fine-tuning is also unsupervised here we use both the ALTA training and test sets as well as the WIPO-alpha training set [4].

Data	Model	Private	Public	Mean
ALTA	SVM	0.714	0.722	0.718
	ULMFiT	0.662	0.712	0.687
WIPO	SVM	0.684	0.728	0.706
	ULMFiT	0.738	0.730	0.734
Both	SVM	0.748	0.754	0.751
	ULMFiT	**0.770**	0.760	0.765
Ensemble		0.764	0.772	0.768
Ensemble + G/H		0.752	**0.784**	0.768

Table 2: F1 scores

4.3.3 Target task classifier fine-tuning

The final step adds two additional linear blocks to the pretrained language model. The first linear layer takes as the input the pooled last hidden layers of the language model and applies a ReLU activation. The last layer is fully connected with a softmax activation to output the probability over the target classes.

4.4 Ensemble

The ensemble stage is combined using hard voting. The four systems that had the highest results on the public set were used. Specifically this includes SVM and ULMFiT trained only with WIPO-alpha and the same models trained with the combined ALTA and WIPO-alpha data. Ties were broken by defaulting to the best performing system which was ULMFiT trained on the combined ALTA and WIPO-alpha data.

4.5 G/H decider

To reduce many of the errors that occur between section `G` and `H` we use two more SVM classifiers trained only on the ALTA training set. The first is a binary classifier to separate the `G/H` from `Not G/H`. The second classifier is trained to separate section `G` from `H`. These classifiers were applied at the ensemble stage such that if the first model classified the document as `G/H` then the ensemble label was overridden by the `G` or `H` label of the second model.

5 Results

Results for this task were evaluated by micro-averaged F1-Score and shown in Table 2.

When only using the smaller ALTA data set SVM outperformed ULMFiT. Training with the larger WIPO-alpha data significantly improved the performance of ULMFiT. This validated the use of

[2] Testing of different lengths found that 3500 characters performed best.

[3] `http://files.fast.ai/models/wt103/`

[4] Fine-tuining on only the ALTA data set performed poorly compared to SVM

the WIPO-alpha data set as it performed better on the ALTA test set despite not using the ALTA data for training.

Training with both data sets together improved both models further.

The performance of some models turned out to be quite different on the private and public splits of the test set. The model that performed best on the public set was third on the private set and the best performance on the private set was third on the public set. The final results on the Kaggle [5] leaderboard also showed similar changes in results for other teams.

Kaggle's default is to take the two best performing submissions from the public scores as the final submission to the competition. From these two the best private score is used as the final result. This mean that our best performing private score was not available for the final result.

When viewing only the public results it appeared that the Ensemble with G/H decider (section 4.5) performed best. The mean of the public and private scores show that both ensembles performed the same with a score of 0.768. The best private score was achieved with ULMFiT trained on both the ALTA and WIPO-apha data.

6 Conclusion

Patent classification for the 2018 ALTA Shared Task has proven to be a good representation of the challenges of Language Technology. In this paper we describe some of the challenges of patent classification. We show that ULMFiT outperforms SVM for patent classification.

Acknowledgments

We would like to thank Dr. Diego Mollá Aliod for his time and support with this task and paper.

References

Karim Benzineb and Jacques Guyot. 2011. *Automated Patent Classification*, Springer Berlin Heidelberg, Berlin, Heidelberg, pages 239–261. https://doi.org/10.1007/978-3-642-19231-9$_1$2.

Eva D'hondt, Suzan Verberne, Cornelis Koster, and Lou Boves. 2013. Text representations for patent classification. *Computational Linguistics* 39(3):755–775. https://doi.org/10.1162/COLI_a_00149.

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

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 328–339. http://aclweb.org/anthology/P18-1031.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and Optimizing LSTM Language Models. *arXiv preprint arXiv:1708.02182* .

Diego Mollá and Dilesha Seneviratne. 2018. Overview of the 2018 alta shared task: Classifying patent applications. In *Proceedings 2018 Australasian Language Technology Workshop ALTA 2018.*

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Dilesha Seneviratne, Shlomo Geva, Guido Zuccon, Gabriela Ferraro, Timothy Chappell, and Magali Meireles. 2015. A signature approach to patent classification. In Guido Zuccon, Shlomo Geva, Hideo Joho, Falk Scholer, Aixin Sun, and Peng Zhang, editors, *Information Retrieval Technology*. Springer International Publishing, Cham, pages 413–419.

Carlos N. Silla and Alex A. Freitas. 2011. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery* 22(1):31–72. https://doi.org/10.1007/s10618-010-0175-9.

[5] www.kaggle.com/c/alta-2018-challenge