Universal Language Model Fine-tuning for Patent Classification

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

MACQUARIE University

ALTA Shared Task 2018

The ALTA Shared Task for 2018 is to automatically classify Australian patents into one of the principal International Patent Classification sections. Data files, submissions and results were managed using Kaggle in Class.

IPC Symbols

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

ALTA Data set

The data provided contains 4972 Australian patents.

- 3972 in Training set
- 1000 in Test set
- 8 unbalanced classes

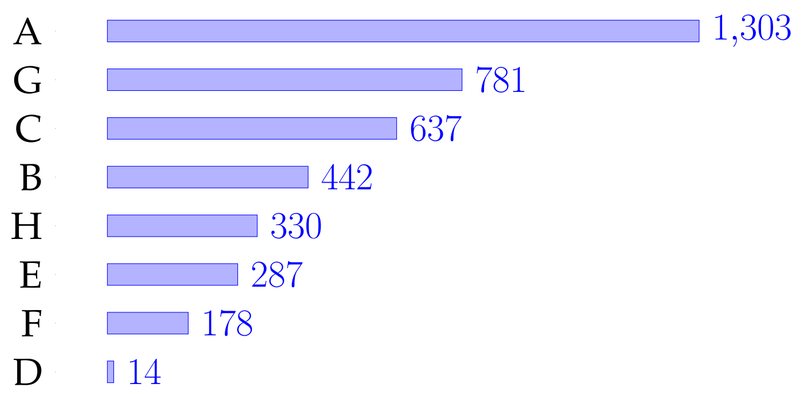

Figure 1: ALTA training set counts by IPC Section

WIPO-alpha Data set

WIPO-alpha is a collection of patent applications form the World Intellectual Property Organization used to increase the training data.

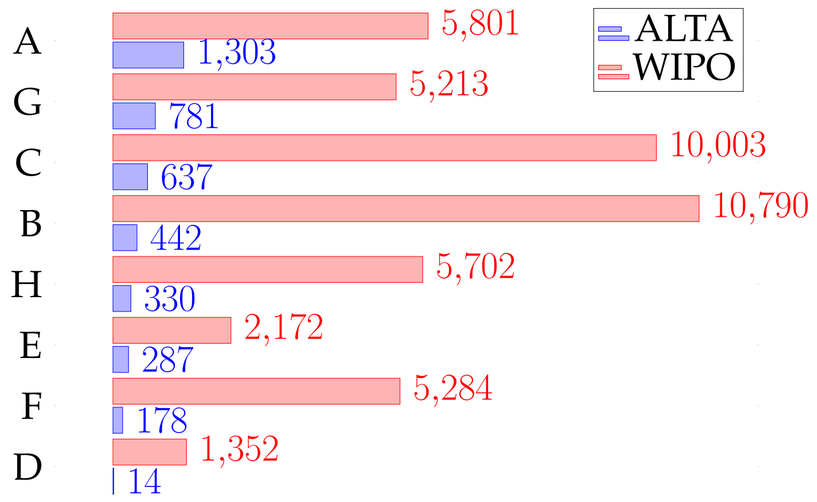

Figure 2: WIPO-Alpha training set counts by IPC Section

Methods

SVM

- First 3500 characters
- Tf-Idf
- Unigrams and Bigrams
- Linear kernel

ULMFiT

Universal Language Model Fine-tuning (ULMFiT) is a transfer learning technique introduced by Howard and Ruder (2018).

1 General-domain language model pretrainig:

- State of the art language model - AWD LSTM
- Large dataset - Wikitext-103

2 Target task language model fine-tuning:

- ALTA training and test sets
- WIPO-Alpha training set

3 Target task classifier fine-tuning:
Concatenate last hidden layer with max and mean pooling of hidden layers.

G/H Decider

We use two additional SVM classifiers in an attempt to reduce the errors between section `G` and `H`.

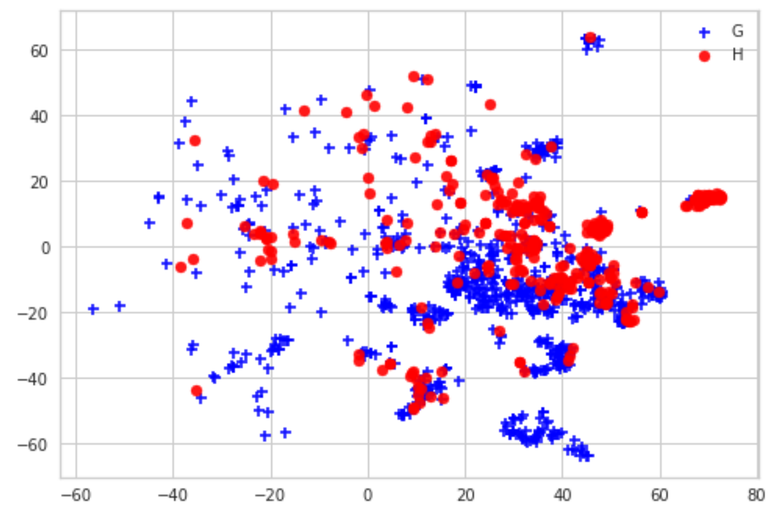

Figure 3: t-SNE of Tf-idf

Results

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 1: F1 micro scores

Conclusions

Patent classification for the 2018 ALTA Shared Task has proven to be a good representation of the challenges of Language Technology. We show that with enough data ULMFiT outperforms SVM for patent classification.

Acknowledgments

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