Patent Application Classification

Fernando Benites[1], Shervin Malmasi[2,3], Marcos Zampieri[4]
benf@zhaw.ch, shervin.malmasi@mq.edu.au, m.zampieri@wlv.ac.uk

[1]Zurich University of Applied Sciences, Switzerland [2]Harvard Medical School, United States
[3] Macquarie University, Australia [4] University of Wolverhampton, United Kingdom

	Training	Public (Validation)	Private (Test)
(1) Baseline 20k feats.	0.709	0.710	0.692
(2) Baseline 40k feats.	0.715	-	-
(3) Baseline w/ WIPO-alpha	0.775	0.758	0.744
(4) Semi-supervised	0.734	0.728	0.704
(5) Ensemble w/ WIPO-alpha + gamma	0.787	0.776	0.778

Table: F1-micro performance of the systems in training (10-fold CV), in the validation and in the test sets (train, public and private leaderboard).

Abstract

We present methods for the automatic classification of patent applications using an annotated dataset provided by the organizers of the ALTA 2018 shared task - Classifying Patent Applications. The goal of the task is to use computational methods to categorize patent applications according to a coarse-grained taxonomy of eight classes based on the International Patent Classification (IPC). We tested a variety of approaches for this task and the best results, 0.778 micro-averaged F1-Score, were achieved by SVM ensembles using a combination of words and characters as features. Our team, BMZ, was ranked first among 14 teams in the competition.

Introduction

According to statistics of the World Intellectual Property Organization (**WIPO**), the number of patent applications filled across the world keeps growing every year. To cope with the large volume of applications, companies and organizations have been investing in the development of software to process, store, and categorize patent applications with minimum human intervention.
We present a system to automatically categorize patent applications from Australia according to the top sections of the IPC taxonomy using a dataset provided by the organizers of the **ALTA 2018** shared task on Classifying Patent Applications (Molla and Seneviratne, 2018)

Data

ALTA 2018 shared task data:

- 5,000 documents Training
- 1,000 documents Validation
- 1,000 documents Test
- 8 Classes
- OCR documents (including OCR artifacts such as 'NA\\nparse failure" and page numbers)

Additional data from WIPO was used:

- alpha: additional 75243 documents
- en: additional 100000 documents randomly selected from 1.1 million

Methods

For feature extraction (Malmasi and Zampieri(2017):

- Term Frequency (TF) of n-grams: 3-6 chars, 1-2 words
- Term Frequency (TF-IDF) of n-grams: 3-6 chars, 1-2 words

Classifier:

- Base linear Support Vector Machines (SVMs)
- Ensemble with Meta

SVMs have proven to deliver very good performance in a number of text classification problems. It was previously used for complex word identification (Malmasi et al., 2016a), triage of forum posts (Malmasi et al., 2016b), dialect identification (Malmasi and Zampieri, 2017), hate speech detection (Malmasi and Zampieri, 2018), and court ruling prediction (Sulea et al., 2017a).

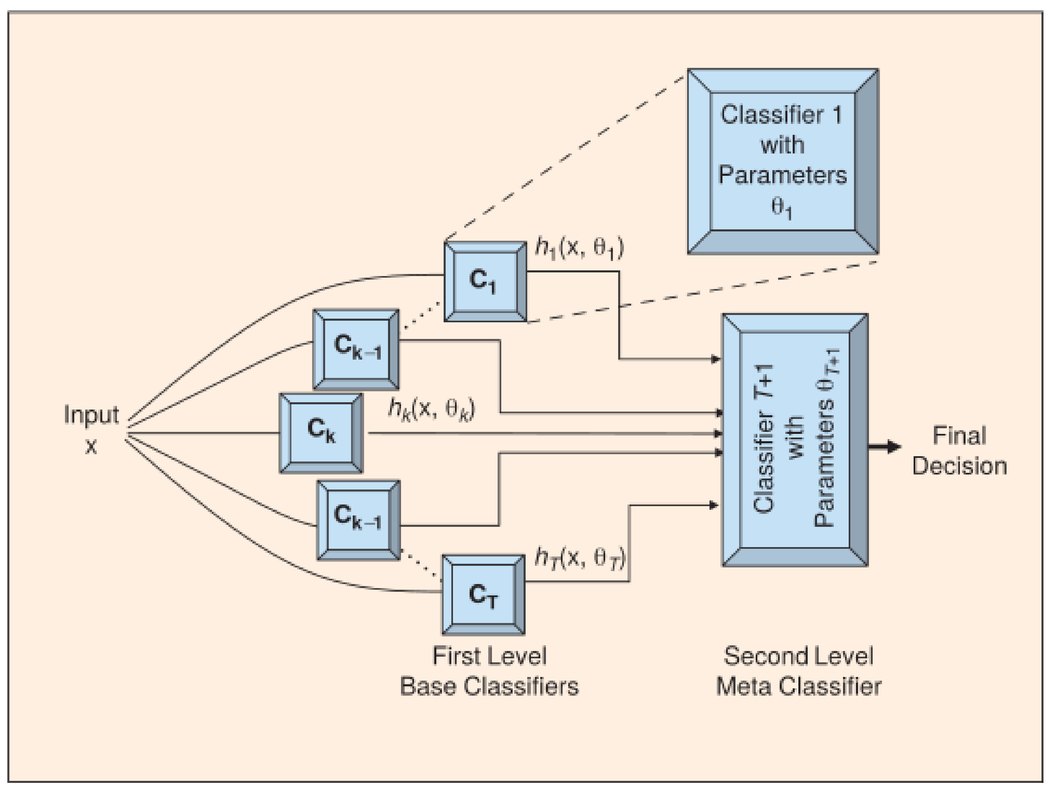

Figure: Meta Classifier (from Polikar (2006))

Results

We present the results obtained in the training stage, the public leaderboard, and the private leaderboard in Table 1. The shared task was organized using Kaggle[a], a data science platform, in which the terms Public Leaderboard and Private Leaderboard are used referring to what is commonly understood as development or validation phase and test phase. This is important in the system development stage as it helps preventing systems from overfitting. We used 10-fold cross validation in the training setup.
As can be seen in Table 1, the ensemble system with additional data achieved the best performance. This can be attributed to the use of large amounts of additional training data, a semi-supervised approach, and an ensemble model with many features.

References

- Robi Polikar. 2006. Ensemble based systems in decision making. Circuits and Systems Magazine, IEEE, 6(3):21–45.
- Diego Molla and Dilesha Seneviratne. 2018. Overview of the 2018 ALTA Shared Task: Classifying Patent Applications. In Proceedings of ALTA 2018.
- Shervin Malmasi, Mark Dras, and Marcos Zampieri.2016a. LTG at SemEval-2016 task 11: Complex Word Identification with Classifier Ensembles. In Proceedings of SemEval.
- Shervin Malmasi, Marcos Zampieri, and Mark Dras. 2016b. Predicting Post Severity in Mental Health Forums. In Proceedings of CLPsych Workshop.
- Shervin Malmasi and Marcos Zampieri. 2017. German dialect identification in interview transcriptions. In Proceedings of the VarDial Workshop.
- Shervin Malmasi and Marcos Zampieri. 2018. Challenges in Discriminating Profanity from Hate Speech. Journal of Experimental Theoretical Artificial Intelligence 30(2):187–202.
- Octavia-Maria Sulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2017b. Predicting the Law Area and Decisions of French Supreme Court Cases. In Proceedings of RANLP.

[a]https://www.kaggle.com/