UNSPSC Classification on Demand

Author: Mitchell Gray
Date: 2022-08-22

Background

Personal Motivation

Nice Classification System

My work involves the processing of data about Intellectual Property rights for the Australian IP office, IP Australia. In this capacity, I deal with trade marks which have an associated internationally recognised classification system called the Nice Classification system[1]. Although Nice is standard in the trade mark[2] domain, in practice when performing analysis of filing activities either at a jurisdictional level or a firm level, this system is very difficult to interpret. Nice provides just 45 classification categories to identify the economic activity associated with a trade mark. This number of categories is reasonable for a coarse overview of the activities associated, however, the classification system is extremely difficult to interpret for economists and other interested policy stakeholders. Consider for example, Nice Class 19:

> Materials, not of metal, for building and construction; rigid pipes, not of metal, for building; asphalt, pitch, tar and bitumen; transportable buildings, not of metal; monuments, not of metal.[3]

Although this provides some indication of what goods marks associated with this classification are, it is also not interpretable. This classification code includes Non-metallic building materials, but then including some pipes and ornamental products including monuments. While this is one specific example, providing interpretable analysis from Nice class information which are all at least as complex as the above is very difficult. It is necessary to have a better classification system to allow both more granular reporting.

Goods & Services Item Description

In addition to a requirement to declare a Nice Classification area or areas in which their trade mark will operate, applicants for trade marks must also submit descriptions of good and services on which their mark will appear. As a result, the descriptive information held in plain text is much richer than the Nice Classification code. It is ideal, therefore, for IP Australia to employ automated goods and services description categorisation activities using this natural language channel of information.

United Nations Standard Product & Service Classification (UNSPSC)

The United Nations promulgates a standard called the United Nations Standard Product & Service Classification (UNSPSC). This classification system provides a classification hierarchy of four levels:

1. Segment
    1. Comprising 57 high level categories of goods and services
2. Family
    1. Lower level classification nested within the segments. In total, across the 57 segments, there are 465 product and service families
3. Class
    1. Again, nested within the families, there are in total 5313 product and service classes within the 465 product and service families.
4. Commodity
    1. In total 71502 individual commodities are identified, again nested within the higher level classes.

Information in the above obtained through exploration of the classification code data file available from the US data.gov portal[4]. Unfortunately, classes are not balanced, so that some Segments have fewer Families than others. At the highest level of granularity, the largest

Segment contains over 31,000 Commodities and the smallest segment contains 38. Despite this, the granularity of the standard, and the fact that it is nested allows us to select the specificity of our classification information for a specific reporting audience. In practice, granularity at the level of the specific commodity is likely to be both unattainable with any accuracy and probably undesirable as this level of granularity provides is too much for a user to make sense of.

In addition, the product and services descriptions provided by the UNSPSC are much more interpretable for normal users than the descriptions provided by Nice. Consider, UNSPSC Class code 10151500:

> Vegetable seeds and seedlings

This classification and the majority of the remainder of the UNSPSC classification symbols are both specific and intepretable. Even at the high level, these classification systems remain interpretable. Consider Family Code 93170000:

> Trade policy and regulation

This combination of scalable granularity and interpretability makes this classification system an ideal candidate for an additional analysis focused classification system for goods and services item descriptions.

Proposed Context

IP Australia may develop a requirement to train a classifier of goods and services item descriptions into an associated UNSPSC class label. This classifier would need to classify these descriptions into at least the UNSPSC segment with a sufficient degree of accuracy. Additionally, it is envisaged that this exercise may need to be done at the time the application is processed. Therefore, the model, once trained, will need to be deployed to a number of endpoints sufficient to service the Trade Mark division of IP Australia (approximately 200 users).

While the training data proposed to be used is not trade mark related, the purpose of the descriptions - as an identifier of the goods and services - should mean that a classifier trained on government purchasing data will have acceptable performance when transferred to the trade mark context.

At first instance this system will trained to provide the UNSPSC market segment only. However, given the volume of training data available, extension to UNSPSC Family classification may be attempted.

Availability of Labelled Data

While I have access to significant holdings of internal data regarding trade mark goods and services descriptions, this information is internally held government data and is not suitable for disclosure for academic purposes. In addition, disclosure of this data source, while it is the most true to life for the intended use case, will not assist; IP Australia does not have labelled data using this classification system.

Government Sources

The UNSPSC system has been employed to categorise the goods and services that various governments have procured across a range of jurisdictions. As the legislative frameworks of many of these jurisdictions require purchasing transparency in the expenditure of public funds, many of these governments provide large volumes of unit record data which provides a plain description of the goods and services procured along with the assigned UNSPSC classification code assigned to the procurement. This information is critical in providing high quality labelled data for a classification system. In practice, there are large samples available of labelled data that can be used to train a classifier.

1. Australian Government Historical Contract Notice dataset [5]
    1. Data provided across multiple files which captures both plain text description and the UNSPSC label assigned.
    2. Approx 80,000 records per year for each of 5 years. Approximately 400,000 tagged records

2. Californian Historical Purchase Order Data 2012-2015 [6]
    1. This dataset provides information across the years of interest regarding products and services procured by the government of California and what UNSPSC code was assigned.
    2. Approximately 300,000 tagged records.
3. Canadian Government data
    1. NIBS-GSIN concordance
        1. Canada is undertaking a process to replace its Goods And Services: Identification Number (GSIN) procurement categorisation system with the UNSPSC. As part of this, Canada has provided a concordance that maps old GSIN codes to the most relevant UNSPSC code. This provides an additional language sample in the form of the GSIN descriptions to exploit as training data[7]
    2. Tender Request data:
        1. Canada provides historical tender information in the form of a spreadsheet detailing the product and service to be procured along with a mix of legacy GSIN and UNSPSC classification codes. The concordance above can be used to convert the GSIN tagged records to UNSPSC codes providing an additional source of data. [8]
    3. It should be noted that the coarse nature of the GSIN classification system may make this channel of data of limited utility unless only high levels of UNSPSC classification from the language model are desirable.
    4. Approximately 200,000 tagged records.

Exploitation of the UNSPSC Hierarchy

Depending at which level (Segment, Family, Class or Commodity) one wished to train a classifier, it would be possible to exploit the hierarchical nature of the UNSPSC system to provide a particularly clean source of labelled data in the form of lower class code descriptions. That is, if the classifier is only to be trained to classify records to the level of UNSPSC segment, any descriptions present in the data for Family, Class and Commodity can be regarded as labelled training data for this classifier.

Since this source of data is particularly clean, it will be used alongside other, more realistic data sources from the government data sources above.

Data Wrangling

Taken together, the sources of labelled data above represent tagged samples approaching 900,000 records. This will require a significant amount of data wrangling to ensure data is preprocessed consistently and is sufficiently cleaned to enable the classification algorithms to work correctly.

Can the Canadian Data be Used?

The Canadian tender notices data uses the GSIN which is proposed to be converted to the UNSPSC using the Canadian Government provided concordance. Given that the initial goal of this project is to deploy endpoint classifiers to classify the UNSPSC segment of the text only, this may be possible. However, inspection of the data indicates that it is likely that there will be a loss of information through the conversion as the levels of granularity between the standards for specific goods and services are not consistent. As a result, depending upon whether the project is extended to the UNSPSC Family, this dataset may need to be abandoned.

Sequence Length

Inspection of all of the government data sources reveals a range of sequence lengths for goods and services item descriptions. This presents some difficulty when training the classifier in two different ways:

1. Long sequences contain a large amount of semantic content, but often need to be split into sentences for transformation.
2. Sequences which are very short which list brand names may be of little value for training. Sentence transformer models rely on the information contained in the sequence. Where there is only a few tokens in a description, this may not be ideal for training.

Sequence Content Curation

Some of the goods and services item descriptions may contain brand names and other content which, depending upon vocabulary size, may not be useful for training. Preliminary exploration of the dataset has identified a number of inclusions in the description text fields such as model numbers and purchase order numbers which are likely to result in limited training value if used. Finding a robust way to curate the training and test sets to remain true to life while minimising the inclusion of unnecessary noise will be critical to later training model efficacy.

Unbalanced Classes in the Training Data

In high level explorations completed to date, it has been noted that the classes for this classification are unbalanced, both in terms of the number of specific items (commodities) included under higher classification standards but also in terms of the prevalence with which they appear in the government data sources identified. Some products and services are not often acquired by governments. As a result, some classes may have relatively few members. Care will need to be taken in wrangling and preparing the data for use in training that the class imbalance issue is addressed.

Natural Language Processing Approach

Baseline Model

Until the emergence of transformer language models, a common approach to language classification was the use of Recurrent Neural Networks (RNN) and Long Short-Term Memory Networks (LSTM). My team within IP Australia has no history applying neural network methodology to text classification, however, the application of one of these earlier methods would be desirable in our context for a number of reasons:

1. Government security frameworks are suspicious of utilising pre-trained models[9]. Training a RNN may be preferable even with

diminished performance as this system does not require the use of external artefacts

2. RNNs are simpler than transformer models and allow training for a specific context. It is possible that the way that an RNN works will be favourable for the often quite short goods and services item descriptions used on products.

I have chosen to use a simple RNN model, guided by some online provided tutorials as the baseline for this project [10]. This model will be trained using the same data and pre-processing (up until transformation).

Proposed Model

Since the arrival of transformer models, previous technologies have been largely abandoned in favour of these transformers. These transformers utilise the mechanism of self-attention to provide additional contextual information for sequence embeddings than their predecessors[11]. These models use sequence embeddings to create a trainable vector space in which words which are associated have similar vectors. The seminal work of Vaswani et al (2017) from Google demonstrated that attention could replace the requirement for other forms of sequence embedding including LSTM and RNN models, and replacing them with attention head networks only [12].

Since the advent of this approach, pretrained variants of these networks have appeared regularly which provide better performance or extend functionality in some way. See for example Devlin et al (2018) introducing BERT or Radford et al (2019) introducing GPT-2[13]. Sanh et al (2020) note the rising number model complexity of transformer models and seek to identify ways to reduce the complexity of these models to speed up training processes while retaining the majority of the performance of larger models[14].

I propose to use a transformer model which is pretrained from the huggingface library[15]. Huggingface provides a repository of pretrained language models with detailed instructions for how to fine tune these for specific use cases [16]. My proposed solution is to use DistilBERT. While the purpose of this model is to reduce the training time by requiring fewer parameters (~40% fewer) when constructing the model, I believe that the fact that this model may also be faster to fine tune and reduce AWS employment costs by allowing training to fit within smaller instances sizes for my use case.

Proposed Metrics

The proposed metric of performance on which the models will be compared is the classification accuracy for the model at the chosen level of granularity of classification to UNSPSC from a training set.

Proposed Solution

The proposed solution process is as follows:

1. Download and storage of sources identified above in an s3 bucket for storage.
2. Exploration and pre-processing of this data to produce a consolidated dataset of UNSPSC classification codes along with known goods and services descriptions.
3. Storage of the cleaned dataset in s3 for later access by training routines.
4. Training and deployment of a Recurrent Neural Network Classifier baseline using PyTorch in AWS Sagemaker.
5. Fine tuning of a pre-trained DistilBERT classifier using AWS Sagemaker hyperparameter optimisation routines and including profiling reporting to ensure model training is appropriately utilising resources
6. Deployment of the optimal model to a AWS Sagemaker endpoint. Configuration of this endpoint for autoscaling to enable a workforce of around 200 to make API calls for inference on demand without significant slowdown. The workforce is expected to make around 4 calls to the API per person per day.
7. Configuration of an AWS Lambda function to serve as the API communication point for the Sagemaker endpoint. Concurrency will

be set up to enable this function to handle multiple concurrent requests.

Conclusion

UNSPSC tagging of goods and services descriptions allows a high level understanding of products and services without the need to reference the descriptions in detail. UNSPSC provides a superior classification system both in terms of granularity and interpretability. The adoption of this classification system for analytics work would enable IP Australia to develop a good understanding of the economic activity associated with trade makr filings that it receives.

The use of tagged datasets in a similar domain allows the creation of a model capable of doing this without the cost and difficulty of manually tagging and sharing internal data. Using an AWS environment it will be possible to fine tune and deploy a model capable of completing this inference in real time. The proposed solution of deploying endpoints based upon fine-tuned state of the art language model to provide maximum accuracy through and AWS environment provides a clear and cost-effective path to the goal of real-time tagging of filing activities to understand changes in the trade mark system.

---

1. See https://www.wipo.int/classifications/nice/en/↩
2. In the Australian context, the "Trade Mark" is spelled as two words, not one. See for example: https://www.ipaustralia.gov.au/trade-marks/understanding-trade-marks/trade-mark-basics.↩
3. Refer to https://www.wipo.int/classifications/nice/nclpub/en/fr/?basic_numbers=show&class_number=19&explanatory_notes=show&lang=en&menulang=en&mode=flat¬ion=&pagination=no&version=20220101↩
4. Refer to https://catalog.data.gov/dataset/unspsc-codes and .csv file attached to this page. Exploration completed using Pandas↩
5. Refer to https://data.gov.au/dataset/ds-dga-5c7fa69b-b0e9-4553-b8df-2a022dd2e982/details?q=UNSPSC. Files are provided on an

annual basis in line with the Australian Financial Year which runs from July to June. ↩

6. Refer to https://catalog.data.gov/dataset/purchase-order-data↩
7. Refer to https://open.canada.ca/data/en/dataset/588eab5b-7b16-4a26-b996-23b955965ffa↩
8. Refer to https://open.canada.ca/data/en/dataset/ffd38960-1853-4c19-ba26-e50bea2cb2d5↩
9. Production environments in my workplace are sealed from access to the internet. Acquisition of pre-trained network weights *cannot* be completed programatically and must be subject to approvals.↩
10. See for example https://coderzcolumn.com/tutorials/artificial-intelligence/pytorch-rnn-for-text-classification-tasks ↩
11. https://towardsdatascience.com/an-intuitive-explanation-of-self-attention-4f72709638e1↩
12. Vaswani et al (2017) Attention is All You Need. ArXiv preprint from: https://arxiv.org/pdf/1706.03762.pdf↩
13. Devlin et al (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv preprint from: https://arxiv.org/pdf/1810.04805.pdf, Radford et al (2019) Language Models are Unsupervised Multitask Learners. Available from: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf ↩
14. Sanh et al (2020) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv preprint from: https://arxiv.org/pdf/1910.01108.pdf↩
15. Refer to https://huggingface.co/models?sort=downloads↩
16. https://huggingface.co/docs/transformers/notebooks contains many useful examples.↩