Location and Temporal Classification with a Multi-Task Recurrent Neural Network

Eve Fleisig (@efleisig) and Chris Pan (@chrispan)
COS 484
December 7, 2019

Abstract

Archiving tasks often require researchers to classify documents based on multiple related attributes. The Princeton Prosody Archive is a database of digitized books published between 1570 and 1923 that has undergone little NLP analysis; however, such analysis could be essential to classifying documents when information such as the year and location of publication is sometimes missing. Thus, we trained a multi-task recurrent neural network to automatically classify documents by time and location of publication based on their contents. We found that our multi-task models were able to outperform our optimal single-task model on time classification and significantly outperform our optimal single-task model on region classification. The optimal single-task model for both region and time classification is a bidirectional two-layer recurrent neural network, as is the optimal multi-task model for time classification; the optimal multi-task model for region classification uses a unidirectional model with a single-layer RNN. Since multi-task RNNs have not yet been applied to historical dating problems, the success of our work could indicate that multi-task RNNs have larger applicability to archiving tasks where documents may need to be classified based on multiple related attributes.

1 Introduction

The Princeton Prosody Archive is a database of digitized books published between 1570 and 1923, including historical documents on the history of poetry, grammar, literature, phonetics, and their intersections. These unclassified documents have undergone little analysis using NLP techniques; furthermore, relevant metadata such as the year and location of publication is sometimes missing. Thus, we train a multi-task recurrent neural network to automatically classify documents by their time and location of publication based on their contents.

Because changes in language usage spread gradually from one location to another over time, information about the time period in which a work was written can help to pinpoint its location, and vice versa (Baxter et al., 2009). Thus, we hypothesized that allowing both classification tasks to share information would improve their performance.

This work is relevant not only to historical researchers who need to automatically classify large amounts of data, but also to the NLP research community, as multi-task RNNs have not been applied to historical dating problems. The success of our work could indicate that multi-task RNNs have broader applicability to archiving tasks where documents may need to be classified based on multiple related attributes.

2 Related Work

NLP methods have been used since the early 2000s to attempt to classify documents by time period (Graliński and Wierzchoń, 2018), but deep learning techniques have only recently been applied to this problem. Research on classifying English documents as British or American includes substantial work using support vector machines (Lui and Cook, 2013), while broader work involving dialect-based classification of documents by region of origin includes Khurana et al. (2017)'s work on classification of Arabic dialects using convolutional neural networks.

Liebeskind and Liebeskind (2019) found that recurrent neural networks (RNNs) outperform a number of other supervised learning models for temporal document classification, although multi-task learning has not yet been applied to temporal classification. Studies such as Enguehard et al. (2017) have found that multi-task learning lowers RNNs' error rates. Enguehard et al. noted that single-task RNNs were often unable to handle complex cases of the subject-verb agreement task, which serves as a good diagnostic for a network's sensitivity to sentence structure. By contrast, multi-task RNNs were better able to handle complex subject-verb agreement cases. These results motivated us to investigate whether multi-task RNNs would have similar advantages when applied to historical classification tasks. In addition, Liu et al. (2016) compared the performance of multi-task RNNs with different architectures on classification tasks with comparatively small quantities data. They found that uniform-layer architecture for multi-task RNNs performed well, particularly compared to coupled-layer architecture. We therefore constructed a uniform-layer multi-task RNN for this project.

3 Implementation

3.1 Dataset

Our dataset consists of raw text from the books in the Princeton Prosody Archive, as well as metadata including works' titles, subtitles, and date and location of publication. We first categorized the documents by country of origin. Because documents are labelled with their city of publication, we categorized the cities as American, British, continental European, or other/ambiguous[1]. This dataset consisted of 3146 American documents, 1808 British documents, 82 continental European documents, and 218 other documents. To train the classifier to distinguish only between American and British documents, we removed the extraneous European and ambiguous data. We then sampled 1808 American works in order to train the model on equal amounts of American and British data.

In keeping with the methodology used in related studies such as Mitra et al. (2015), we divided the documents into four time periods (1700-1842, 1843-1874, 1875-1901, and 1902-1923), such that all time periods contain a roughly equal amount of data, and each time period contains balanced amounts of British and American data.

We then used an 80/20 split to randomly partition the data (1808 British documents and 1808 American documents, 3616 in total) into a training set, consisting of 2688 documents, and a test set, consisting of 672 documents.

After initially training models based on the documents' titles and subtitles (capped at 25 words), we experimented with training models on the first paragraph of the full text of the documents. However, the models performed better on the titles and subtitles alone than on the longer documents (analyzed in Section 5). Given that we were unable to gain meaningful improvements from using the full data, we ultimately chose to use documents' titles and subtitles alone. The title and subtitle information for each document was fed into our models token by token, and the predicted classes were compared against the location and time period of origin extracted from the metadata.

3.2 Models

Our multi-task model uses shared hidden layers and RNN features for both tasks (time and region classification). The last output of the shared RNN then branches off into two separate paths–one for time period classification and one for region classification–each consisting of a linear layer, a sigmoid layer, and a second separate linear layer to project output into dimensions equal to number of classes for both tasks. We then obtain probabilities for the maximum likelihood estimate using a softmax layer. This architecture is illustrated in Figure 1. The models use word embeddings trained on the dataset, as opposed to pretrained word2vec embeddings, since these may be inaccurate for older documents.

As a baseline for performance comparison, we built two separate single-task RNNs, one for time period classification and one for region classification.

3.3 Training

We tested the multi-task model and baseline while varying several parameters, including the learning rate, training with mini-batch gradient descent or the full training set, the hidden layer size, and the model's directionality.

Upon repeating experiments with different random initializations of RNN features as well as the first hidden and context states, we found that our model encountered a significant number of local minima. These could be unavoidable consequences of our noisy dataset, or due to the fact that a more complex model is necessary to model the dataset; we tentatively hypothesize that the level of local minima could be due to a combination of both. To account for the existence of these frequent local

[1] Cities such as Hartford exist in both the US and UK, impeding classification of some documents by country of origin. We thus removed these documents from the dataset.

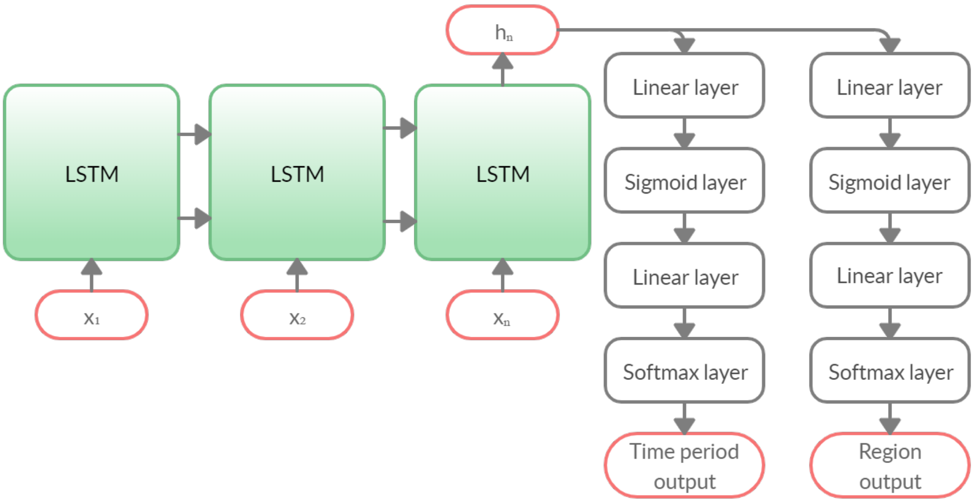

Figure 1: Structure of the optimal multi-task model (unidirectional single-layer LSTM)

minima, we chose to train the model using gradient descent on the whole dataset rather than mini-batch gradient descent, since mini-batch gradient descent is often less consistent than processing the entire dataset at once.

Additionally, these local minima caused the best multi-task model for one of the tasks, region classification, to be comparatively simple: a unidirectional model with a single hidden layer. We hypothesize that given the right training parameters and initialization, the bidirectional two-layer multi-task RNN could perform as well on region classification as the single-layer multi-task RNN did, a prospect that we are interested in exploring in future work (Section 5). For the present research, we found that because the optimization goal was far from convex, the multi-task model performed better on region classification with a single layer.

After experimenting with hyperparameters and training parameters, we found that the optimal multi-task model for region classification was a single-layer unidirectional long-short term memory model (LSTM) with a hidden size of 64 and an embedding size of 64, while the optimal multi-task model for time period classification was a two-layer bidirectional long-short term memory model (LSTM) with a hidden size of 32 and an embedding size of 64. The optimal single-task baseline (for both classification tasks) consisted of two single-task RNNs, each using two-layer bidirectional LSTMs with a hidden size of 32.

Both the optimal multi-task model and the optimal single-task baseline performed best with gradient descent on the full train set instead of mini-batch gradient descent. As we had observed, mini-batch gradient descent may go down a gradient that is suboptimal as it converges, since it does not reflect the gradient of the full cost function. Instead, stepping based on the entire dataset at once produced more consistent results; moreover, it sped up the convergence of our algorithm. The optimal multi-task model and the optimal single-task baseline were both trained for 50 epochs with a learning rate of 0.01 that did not decay. The gradient was optimized using the Adam algorithm.

Code is available at github.com/chrispan68/COS-484-Final-Project.

4 Evaluation

After training the models with varying parameters, we analyzed and compared the performance of the single-task baseline model and the multi-task model results on different parameters. All optimal results for different tasks and models were found to occur when using bidirectionality and two layers or unidirectionality and a single layer.

With one layer, the multi-task model conspicuously outperforms the single-task model in both region classification (average single-task F1=0.741; average multi-task F1=0.773) and temporal classification (average single-task F1=0.526; average multi-task F1=0.552). With two layers, the single-task model outperforms the multi-task model in region classification (average single-task F1=0.748; average multi-task F1=0.739), but the multi-task model outperforms the single-task model in tem-

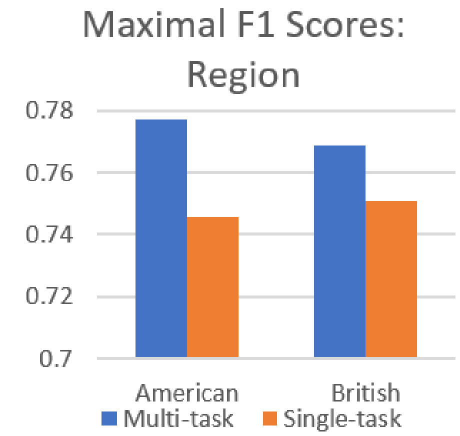

Figure 2: Maximal F1 scores for region classification. The multi-task model is a unidirectional single-layer model; the single-task model is a bidirectional two-layer model.

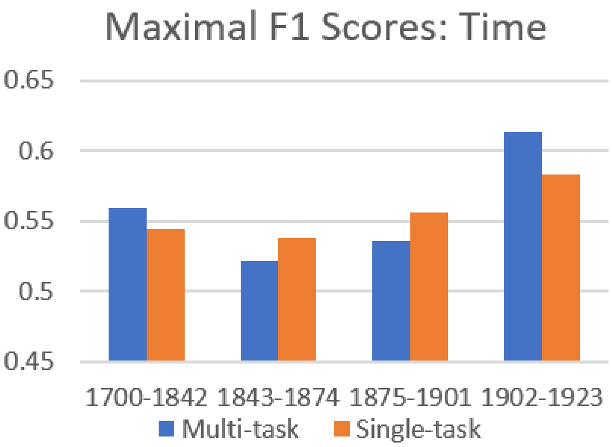

Figure 3: Maximal F1 scores for region classification. Both the single- and multi-task models are bidirectional two-layer models.

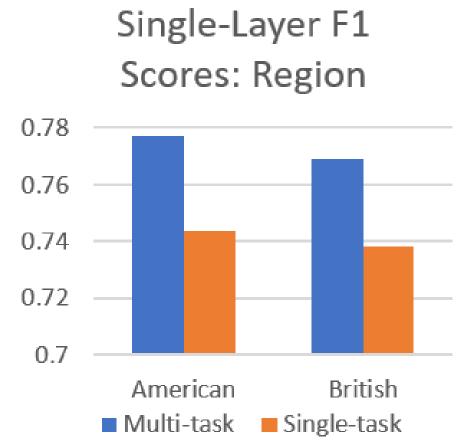

Figure 4: F1 scores for region classification using a single layer and unidirectionality (which produced optimal scores for the multi-task model).

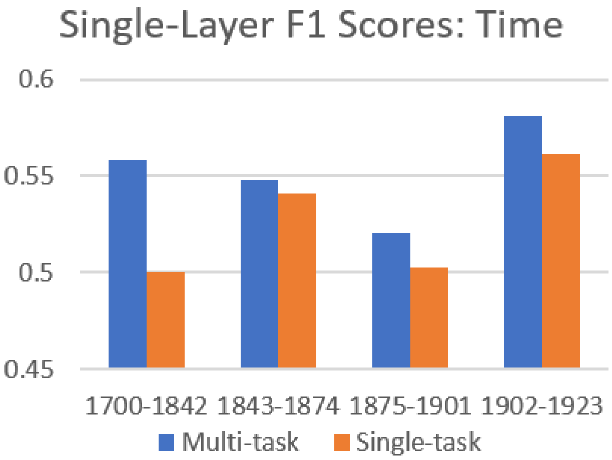

Figure 5: F1 scores for time period classification using a single layer and unidirectionality.

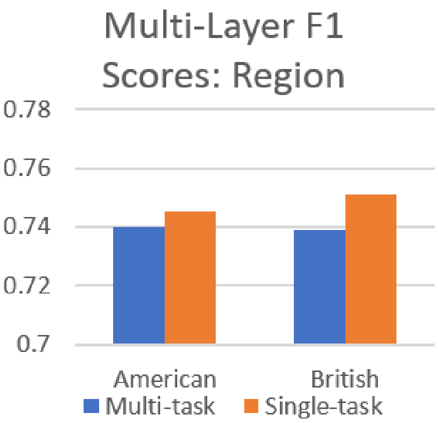

Figure 6: F1 scores for region classification using two layers and bidirectionality. (Note: F1 scores for time classification on these parameters were optimal for both single- and multi-task models; see Figure 3.)

Model Type	Layers	Time Period				
		1700-1842	1843-1874	1875-1901	1902-1923	Average
Multi-task	1	0.5584	**0.5483**	0.5206	0.5810	0.5520
Multi-task	2	**0.5595**	0.5214	**0.5356**	**0.6136**	**0.5575**
Single-task	1	0.5	0.5409	0.5026	0.5615	0.5263
Single-task	2	0.5442	0.5372	0.5564	0.5824	0.5550

Table 1: F1 scores for temporal classification. The two-layer multi-task model performed best. Using one layer or two layers, the multi-task models outperformed the single-task models on the same parameters.

Model Type	Layers	Region		
		American	British	Average
Multi-task	1	**0.777**	**0.769**	**0.773**
Multi-task	2	0.7398	0.7389	0.73935
Single-task	1	0.7439	0.7384	0.74115
Single-task	2	0.7455	0.7509	0.7482

Table 2: F1 scores for region classification. The optimal multi-task model (with one layer) significantly outperformed all single-task models.

poral classification (average single-task F1=0.555; average multi-task F1=0.558).

The single-layer multi-task model performed much better than all single-task baseline models at region classification (Table 2), and the two-layer multi-task model outperformed all single-task baseline models at time classification (Table 1). Thus, multi-task models outperformed the single-task models, although the optimal multi-task models for each task had different parameters. Whereas with two layers, the multi-task model performed best at time period classification (though it performed slightly below the equivalent single-task model on region classification), the single-layer multi-task model performed best at region classification (and, moreover, performed slightly better than the equivalent single-task model on time period classification).

5 Conclusions and Future Work

We conclude that multi-task models do improve performance on related historical classification tasks. In particular, when optimizing for region classification using a single-layer model, performance on the time-period classification task is comparable to that of the optimal single task model, but region classification saw large performance increases.

The finding that using a multi-layer model increases performance on time period classification beyond all single-task baselines suggests that further research could result in multi-layer multi-task models that also outperform multi-layer single-task models on region classification, paralleling the benefits of multi-task models on both tasks using a single layer.

We hypothesize that these performance benefits could stem from two sources: (1) Converting two single task models into a single multitask model with shared RNN cells and hidden layers would serve as a way to regulate the model against overfitting. By forcing the hidden layers to store information about both time period classification and region classification, the network is discouraged from learning trends about one task that are very specific to the training set and do not generalize to the test set. (2) The two tasks appear to be related enough that having temporal information encoded into the hidden layers could substantially benefit our model's performance on the region classification task, and vice versa.

5.1 Future Work

The model's convergence to multiple local minima (as observed from variation between randmoly initialized runs) suggests that more complex models may better capture all relevant characteristics of the tasks. The occurrence of these frequent local minima may be a byproduct of the noise inherent in our data set, but we hypothesize that their existence is in part induced by the relative simplicity of this model. These local minima may be responsible for the fact that the best multi-task model for region classification was a unidirectional model with a single hidden layer, even though the best multi-task model for time classification was a bidirectional two-layer model. In future work, we hypothesize that by increasing model complexity, the bidirectional two-layer multi-task RNN could arrive at similar results on region classification.

Additionally, the fact that using more text from the documents did not improve performance also suggests that increased complexity could boost performance. We speculate that either our model was not complex enough to leverage the information present in the full data set, or our data set was still not large enough to properly extract relevant information from the first paragraphs of the full text alone.

Our findings on methods that improved the models discussed in this paper suggest how the model's complexity could be increased in order to better capture relevant information. We found that increasing the complexity of the differentiation stage of the multi-task model by adding a linear and a sigmoid layer boosted performance, so we could investigate the use of more complex methods to decode the encoding generated by the RNN. It is quite plausible that the complexity of our multi-task model is still bottlenecked by the last three layers, since increasing the layer count and directionality of our multi-task model did not result in the performance improvements on region classification seen on the single-task models.

More broadly, different model architectures could also be explored, particularly multi-task attention-based transformer models. Such models may be more effective at extracting high-level information from the documents.

Furthermore, extensions of this work could examine the application of multi-task RNNs to other archival classification tasks, such as identifying the subject matter or author of a document. Larger multi-task networks classifying documents in more than two ways may increase the scope and applicability of this work.

References

Gareth Baxter, Richard Blythe, William Croft, and Alan McKane. 2009. Modeling language change: An evaluation of trudgill's theory of the emergence of new zealand english. *Language Variation and Change*, 21:257 – 296.

Émile Enguehard, Yoav Goldberg, and Tal Linzen. 2017. Exploring the syntactic abilities of RNNs with multi-task learning. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 3–14, Vancouver, Canada. Association for Computational Linguistics.

Filip Graliński and Piotr Wierzchoń. 2018. Retroc – a corpus for evaluating temporal classifiers. pages 101–111.

Sameer Khurana, Maryam Najafian, Ahmed Ali, Tuka Al Hanai, Yonatan Belinkov, and James Glass. 2017. Qmdis: Qcri-mit advanced dialect identification system. In *Proc. Interspeech 2017*, pages 2591–2595.

Chaya Liebeskind and Shmuel Liebeskind. 2019. Deep Learning for Period Classification of Historical Texts. Working paper or preprint.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 2873–2879. AAAI Press.

Marco Lui and Paul Cook. 2013. Classifying English documents by national dialect. In *Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013)*, pages 5–15, Brisbane, Australia.

Sunny Mitra, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2015. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, 21(5):773–798.

A Contributions

Eve scraped and collected the dataset, then cleaned and preprocessed the OCR text. Chris trained the initial models. We debugged, analyzed, wrote, and presented the work together.

B Honor Code

This paper represents our own work in accordance with University regulations.