Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages, pages 111–113
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Data-driven Choices in Neural Part-of-Speech Tagging for Latin

Geoff Bacon
Department of Linguistics
University of California, Berkeley
bacon@berkeley.edu

Abstract

Textual data in ancient and historical languages such as Latin is increasingly available in machine readable forms, yet computational tools to analyze and process this data are still lacking. We describe our system for part-of-speech tagging in Latin, an entry in the EvaLatin 2020 shared task. Based on a detailed analysis of the training data, we make targeted preprocessing decisions and design our model. We leverage existing large unlabelled resources to pre-train representations at both the grapheme and word level, which serve as the inputs to our LSTM-based models. We perform an extensive cross-validated hyperparameter search, achieving an accuracy score of up to 93 on in-domain texts. We publicly release all our code and trained models in the hope that our system will be of use to social scientists and digital humanists alike. The insights we draw from our inital analysis can also inform future NLP work modeling syntactic information in Latin.

Keywords: Part-of-speech tagging, Latin, LSTM, grapheme tokenization

1. Introduction

Textual data in historical and ancient languages (such as Latin and Ancient Greek) is increasingly available in digital form. As such, computational tools for analyzing and processing this data are highly useful among social scientists and digital humanists. In order to promote the development of resources and language technologies for Latin, the CIRCSE research centre[1] organized EvaLatin: a shared competition on part-of-speech tagging and lemmatization in Latin. This paper describes our system that participated in the part-of-speech tagging task of EvaLatin (Sprugnoli et al., 2020).

Our system was heavily informed by a detailed exploratory analysis of the training data. This analysis guided both our preprocessing decisions as well as the structure of the model. We assembled a large unlabelled corpus of Latin to train embeddings at both the grapheme and word level. Our system combines these pre-trained embeddings in LSTMs to predict part-of-speech tags. In this way we are able to leverage the wealth of unlabelled but machine-readable text in Latin available, as well as recent progress in neural network models of language. To fine-tune our system, we perform an extensive cross-validated hyperparameter search.

The remainder of the paper is structured as follows. In the next section, we outline the main findings of our exploratory data analysis that guided our approach. We then discuss the preprocessing decisions that were informed by this analysis in section 3. Section 4 describes our system, including our cross-validated hyperparameter optimization. In section 5 we present our results. Finally, section 6 highlights our plans for improving our method as well as the open and reproducible nature of this research.

2. Exploratory data analysis

Prior to making any modeling decisions, we performed a detailed exploratory analysis of the EvaLatin dataset. The goal was to find insights in the data that could be leveraged during the modeling stage. To do this, we analyzed the training data from three viewpoints, each focusing on a different level of the data: dataset-wide, orthographic forms and part-of-speech labels. In this section, we highlight the main findings from our analysis that guided the development of our system.

The training dataset contains 14,399 sentences with a total of 259,645 words. This is sizeable yet still significantly smaller than part-of-speech datasets in many other languages. The moderate size of labelled data available motivated us to investigate external unlabelled data (described in Section 4.1). Most (75%) sentences have under 24 tokens, with the average having 18. The vast majority (95%) of sentences have at most 40 tokens. A common concern in sequence-based neural networks is their recency bias which is a shortcoming when the data displays long-distance dependencies. However, with sentences of such moderate length, this concern is not pressing.

At the level of the orthographic form, we found numerous insights that guided our modeling. There are 43,767 unique forms in the training data, of which more than half (24,376) only appear once. The vast majority (90%) of forms appear at most 7 times in the training data. The large number of forms, and especially the large number of hapax legomena, suggest the need to include sub-word information, e.g. character-based models. There are 126 unique characters in the training data, a number which we could massively reduce by focusing on Latin characters (47 unique). Within the Latin characters, we noted that over 98% of instances are lower case. We further noted that capitalization is used in one of four ways: i) as the first character of a sentence, ii) as the first character of a proper noun (abbreviated or not), iii) in Roman numerals, or iv) in the token "HS". Although capital letters are an important signal for proper nouns, case folding would again halve the size of the character vocabulary. Full stops were also used in one of four ways: i) in abbreviations of proper nouns, ii) in lacunae, iii) for the noun "salus", almost always preceded by "suus", or iv) other abbreviations, whose full form is not found elsewhere in the sentence. As all Greek forms have the part-of-speech X, we can effectively represent any Greek word with a single

[1] https://github.com/CIRCSE

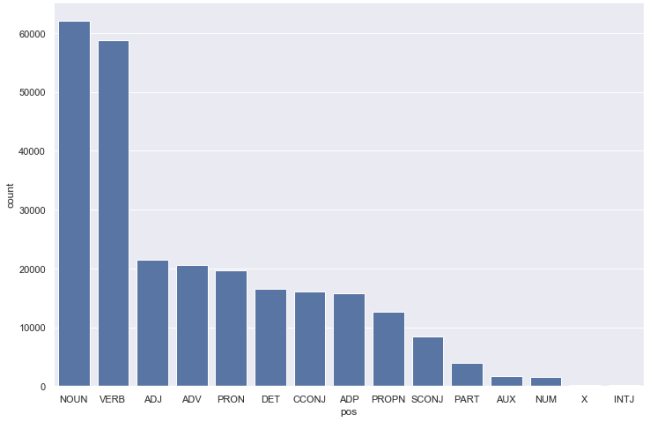

Figure 1: The frequency distribution over part-of-speech tags in the training data. Nouns and verbs are by far the most frequent tags, while AUX, NUM, X and INTJ are extremly rare.

form. Taken together, these insights suggest heavy preprocessing to reduce the character vocabulary, which we describe in Section 3.

Although there are a total of 15 part-of-speech tags in the dataset, the tags are clearly separated into three groups by frequency. The distribution over tags is illustrated in 1. Nouns and verbs are by far the most frequent tags (each accounting for around 23% of all tokens, totalling over 45% together). The next group consists of ADJ, ADV, PRON, DET, CCONJ, ADP, PROPN, SCONJ and PART tags, and each account for 1-8% of tags. The last group consists of AUX, NUM, X and INTJ tags, which each account for less than 1% of tokens. As a baseline, predicting NOUN for all words would have an accuracy of 23% in the training data. The NOUN, VERB, ADJ tags and PRON are identified by lexical root, morphology and syntactic context. Thus, it is important to explicitly include these information sources in the model, for example, with a contextual model. The ADV, DET, ADP and CCONJ tags are often tied to a particular orthographic form, which suggests that word-type representations would be effective in identifying them. Identifying tags which rely on inflectional morphology could be handled by character-based models and sub-word representations. As Latin's inflectional morphology is entirely suffixing, models would benefit from explicit end of word information.

3. Preprocessing

Based on our initial data analysis, our preprocessing was designed to remove as much noise from the data as possible that is not relevant to the task of part-of-speech tagging. To that end, we made significant preprocessing decisions. We replaced the following classes of word forms with placeholder characters as their specific forms do not matter for part-of-speech tagging: i) Greek words, ii) proper noun abbreviations and iii) lacunae. All remaining forms were lowercased. We also added start and end characters for word boundaries to assist modeling inflectional morphology. Furthermore, we tokenized orthographic forms into graphemes rather than characters (Moran and Cysouw, 2018). Thus, character bigrams such as ⟨qu⟩ and ⟨ph⟩ are represented as a single grapheme in our models, rather than two.

4. System

Our system is broadly composed of three sections: i) pretrained domain-specific grapheme and word embeddings, ii) grapheme-level LSTMs, and iii) word-level bidirectional LSTMs. In this section, we first describe the unlabelled corpus of Latin text we curated to pre-train embeddings. We then describe the training procedure of the embeddings, followed by the structure of our model. Finally, we describe our extensive hyperparameter search to fine-tune our system.

4.1. Unlabelled corpus

Given the moderate size of the labelled training data discussed in Section 2, we opted to leverage unlabelled data to improve performance. Concretely, we curated an unlabelled corpus of Latin texts in order to learn non-contextual grapheme and word embeddings. We sourced this corpus from the Perseus Project, the Latin Library and the Tesserae Project through the CLTK library (Johnson, 2014 2020). The resulting corpus totalled over 23 million words.

4.2. Embeddings

We trained grapheme and word embeddings on this unlabelled corpus. In order to capture as much inflectional morphology as possible in the word embeddings, we used fastText (Bojanowski et al., 2017) which benefits from sub-word information. For grapheme embeddings, where subsymbolic information is not available we used the closely related word2vec (Mikolov et al., 2013). We trained grapheme embeddings of dimension $d_g \in \{5, 10, 20\}$ and word embeddings of dimension $d_w \in \{10, 25, 50, 100, 200, 300\}$ with n-gram lengths from 2 to 4. As part-of-speech tagging is a syntactic task, we fixed a low window size (3) for both sets of embeddings and trained for 10 epochs.

4.3. Model

Our part-of-speech tagging model is structured as follows. A unidirectional LSTM reads words as the preprocessed sequence of graphemes, representing them with their pretrained embeddings. The final hidden state of that model is concatenated with the pre-trained word embedding. This concatenation (of size $d_g + d_w$) represents the input to a bidirectional LSTM at a single time step. At each time step, the output of the bidirectional LSTM is passed through a linear layer to produce probabilities over part-of-speech tags. All parameters within the model, including the pretrained embeddings, are trainable.

4.4. Hyperparameter optimization

We ran extensive hyperparameter optimization to fine-tune our model. In particular, we performed a grid search over the following hyperparameters: grapheme embeddings ($d_g \in \{5, 10, 20\}$), word embeddings ($d_w \in \{10, 25, 50, 100, 200, 300\}$), hidden size of bidirectional

Subtask	Text	Accuracy
Classical	Bellum Civile	93.08
	In Catilinam	93.02
	De Providentia	90.63
	De Vita Beata	90.72
	Agricola	89.71
	Germania	87.38
	Epistulae	90.02
Cross-Genre	Carmina	73.47
Cross-Time	Summa Contra Gentiles	76.62

Table 1: The official evaluation results of our system on the EvaLatin shared task. Our system performed well on other Classical texts but saw significant performance drops on out-of-domain texts.

LSTM ($d_h \in \{50, 100, 200, 300\}$) and batch size ($b \in \{8, 16\}$). To evaluate each hyperparameter setting, we used 5-fold cross-validation of the training data. We trained for up to 10 epochs, with early stopping. In total, we trained 1,440 models on a single GPU.

5. Results

In this section, we analyze the results of our hyperparameter search and the errors our system makes, as well as report on the official evaluation.

Averaging over the five cross-validation folds, our best performing model achieved 95.3% accuracy on the training set. We observed a strong positive correlation between the dimensionality of the word embeddings and performance (Pearson's correlation $\rho = 0.725$) and a moderate positive correlation between the dimensionality of the hidden state of the bidirectional LSTM and performance ($\rho = 0.253$). The dimensionality of the grapheme embeddings and performance were weakly correlated ($\rho = 0.042$). All of the 1,440 models we trained achieved above 99% top 3 accuracy. The most common errors we observed were incorrectly tagging adjectives as nouns (12 % of errors) or nouns as adjectives (11%).

The official evaluation metric used in the EvaLatin evaluation was accuracy. The scores of our model on individual texts across the three subtasks are illustrated in Table 1. Our system performed well on in-domain texts (the Classical subtask) but saw significant drops in performance in out-of-domain texts spanning different genres and time periods of the language.

6. Discussion

Our approach was one heavily informed by an initial exploratory data analysis of the training dataset. We relied on significant preprocessing to remove noise from the data and leveraged a large unlabelled corpus of Latin texts. Our extensive hyperparameter search fine-tuned our system. Although our system performed well on in-domain texts, this high performance did not carry well across to other domains and time periods. Future work could investigate the use of external labelled resources to improve performance out of domain.

In order to facilitate engagement with our work, we make all our code and trained models publicly available at `https://github.com/geoffbacon/verrius`. In future work, we plan to make our models freely available through an API for research purposes. With the increased availability of digitized documents in ancient languages like Latin, computational tools for processing linguistic data grow in usage. We hope that our system will be of use to social scientists and digital humanists alike.

7. References

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Johnson, K. (2014–2020). CLTK: The classical language toolkit. `https://github.com/cltk/cltk`.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Moran, S. and Cysouw, M. (2018). *The unicode cookbook for linguists: managing writing systems using orthography profiles*.

Sprugnoli, R., Passarotti, M., Cecchini, F. M., and Pellegrini, M. (2020). Overview of the evalatin 2020 evaluation campaign. In Rachele Sprugnoli et al., editors, *Proceedings of the LT4HALA 2020 Workshop - 1st Workshop on Language Technologies for Historical and Ancient Languages, satellite event to the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Paris, France, May. European Language Resources Association (ELRA).