Temporal Text Ranking and Automatic Dating of Texts

Vlad Niculae[1], Marcos Zampieri[2], Liviu P. Dinu[3], Alina Maria Ciobanu[3]
Max Planck Institute for Software Systems, Germany[1]
Saarland University, Germany[2]
Center for Computational Linguistics, University of Bucharest, Romania[3]
vniculae@mpi-sws.org, marcos.zampieri@uni-saarland.de,
ldinu@fmi.unibuc.ro, alina.ciobanu@my.fmi.unibuc.ro

Abstract

This paper presents a novel approach to the task of temporal text classification combining text ranking and probability for the automatic dating of historical texts. The method was applied to three historical corpora: an English, a Portuguese and a Romanian corpus. It obtained performance ranging from 83% to 93% accuracy, using a fully automated approach with very basic features.

1 Introduction

Temporal text classification is an underexplored problem in NLP, which has been tackled as a multi-class problem, with classes defined as time intervals such as months, years, decades or centuries. This approach has the drawback of having to arbitrarily delimit the intervals, and often leads to a model that is not informative for texts written within such a window. If the predefined window is too large, the output is not useful for most systems; if the window is too small, learning is impractical because of the large number of classes. Particularly for the problem of historical datasets (as the one we propose here), learning a year-level classifier would not work, because each class would be represented by a single document.

Our paper explores a solution to this drawback by using a *ranking* approach. Ranking amounts to ordering a set of inputs with respect to some measure. For example, a search engine ranks returned documents by relevance. We use a formalization of ranking that comes from *ordinal regression*, the class of problems where samples belong to inherently ordered classes.

This study is of interest to scholars who deal with text classification and NLP in general; historical linguists and philologists who investigate language change; and finally scholars in the digital humanities who often deal with historical manuscripts and might take advantage of temporal text classification applications in their research.

2 Related Work

Modelling temporal information in text is a relevant task for a number of NLP tasks. For example, in Information Retrieval (IR) research has been concentrated on investigating time-sensitivity document ranking (Dakka and Gravana, 2010). Even so, as stated before, temporal text classification methods were not substantially explored as other text classification tasks.

One of the first studies to model temporal information for the automatic dating of documents is the work of de Jong et al. (2005). In these experiments, authors used unigram language models to classify Dutch texts spanning from January 1999 to February 2005 using normalised log-likelihood ratio (NLLR) (Kraaij, 2004). As to the features used, a number of approaches proposed to automatic date take into account lexical features (Dalli and Wilks, 2006; Abe and Tsumoto, 2010; Kumar et al., 2011) and a few use external linguistic knowledge (Kanhabua and Nørvåg, 2009).

A couple of approaches try to classify texts not only regarding the time span in which the texts were written, but also their geographical location such as (Mokhov, 2010) for French and, more recently, (Trieschnigg et al., 2012) for Dutch. At the word level, two studies aim to model and understand how word usage and meaning change over time (Wijaya and Yeniterzi, 2011), (Mihalcea and Nastase, 2012).

The most recent studies in temporal text classification to our knowledge are (Ciobanu et al., 2013) for Romanian using lexical features and (Štajner and Zampieri, 2013) for Portuguese using stylistic and readability features.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 17–21,
Gothenburg, Sweden, April 26-30 2014.

3 Methods

3.1 Corpora

To evaluate the method proposed here we used three historical corpora. An English historical corpus entitled Corpus of Late Modern English Texts (CLMET)[1] (de Smet, 2005), a Portuguese historical corpus entitled Colonia[2] (Zampieri and Becker, 2013) and a Romanian historical corpus (Ciobanu et al., 2013).

CLMET is a collection of English texts derived from the Project Gutenberg and from the Oxford Text Archive. It contains around 10 million tokens, divided over three sub-periods of 70 years. The corpus is available for download as raw text or annotated with POS annotation.

For Portuguese, the aforementioned Colonia (Zampieri and Becker, 2013) is a diachronic collection containing a total of 5.1 million tokens and 100 texts ranging from the 16th to the early 20th century. The texts in Colonia are balanced between European and Brazilian Portuguese (it contains 52 Brazilian texts and 48 European texts) and the corpus is annotated with lemma and POS information. According to the authors, some texts presented edited orthography prior to their compilation but systematic spelling normalisation was not carried out.

The Romanian corpus was compiled to portrait different stages in the evolution of the Romanian language, from the 16th to the 20th century in a total of 26 complete texts. The methodology behind corpus compilation and the date assignment are described in (Ciobanu et al., 2013).

3.2 Temporal classification as ranking

We propose a temporal model that learns a linear function $g(x) = w \cdot x$ to preserve the temporal ordering of the texts, i.e. if document[3] x_i predates document x_j, which we will henceforth denote as $x_i \prec x_j$, then $g(x_i) < g(x_j)$. Such a problem is often called *ranking* or *learning to rank*. When the goal is to recover contiguous intervals that correspond to ordered classes, the problem is known as *ordinal regression*.

We use a pairwise approach to ranking that reduces the problem to binary classification using a linear model. The method is to convert a dataset of the form $\mathcal{D} = \{(x, y) : x \in \mathbb{R}^d, y \in \mathcal{Y}\}$ into a pairwise dataset:

$$\mathcal{D}_p = \{((x_i, x_j), \mathbf{I}[y_i < y_j]) : (x_i, y_i), (x_j, y_j) \in \mathcal{D}\}$$

Since the ordinal classes only induce a partial ordering, as elements from the same class are not comparable, $\mathcal{D}_p$ will only consist of the comparable pairs.

The problem can be turned into a linear classification problem by noting that:

$$w \cdot x_i < w \cdot x_j \iff w \cdot (x_i - x_j) < 0$$

In order to obtain probability values for the ordering, we use logistic regression as the linear model. It therefore holds that:

$$\mathbf{P}(x_i \prec x_j; w) = \frac{1}{1 + \exp(-w \cdot (x_i - x_j))}$$

While logistic regression usually fits an intercept term, in our case, because the samples consist of differences of points, the model operates in an affine space and therefore gains an extra effective degree of freedom. The intercept is therefore not needed.

The relationship between pairwise ranking and predicting the class from an ordered set $\{r_1, ...r_k\}$ is given by assigning to a document x the class r_i such that

$$\theta(r_{i-1}) \leq g(x) < \theta(r_i) \quad (1)$$

where θ is an increasing function that does not need to be linear. (Pedregosa et al., 2012), who used the pairwise approach to ordinal regression on neuroimaging prediction tasks, showed using artificial data that θ can be accurately recovered using non-parametric regression. In this work, we use a parametric estimation of θ that can be used in a probabilistic interpretation to identify the most likely period when a text was written, as described in section 3.3.

3.3 Probabilistic dating of uncertain texts

The ranking model described in the previous section learns a direction along which the temporal order of texts is preserved as much as possible. This direction is connected to the chronological axis through the θ function. For the years t for

[1] https://perswww.kuleuven.be/~u0044428/clmet

[2] http://corporavm.uni-koeln.de/colonia/

[3] For brevity, we use x_i to denote both the document itself and its representation as a feature vector.

which we have an unique attested document x_t, we have that

$$x \prec x_t \iff g(x) < g(x_t) < \theta(t)$$

This can be explained by seeing that equation 2 gives $\theta(t)$ as an upper bound for the projections of all texts written in year t, and by transitivity for all previous texts as well.

Assuming we can estimate the function θ with another function $\hat{\theta}$, the cumulative densitiy function of the distribution of the time when an unseen document was written can be expressed.

$$P(x \prec t) \approx \frac{1}{1 + \exp(w \cdot x - \hat{\theta}(t))} \qquad (2)$$

Setting the probability to $\frac{1}{2}$ provides a point estimate of the time when x was written, and confidence intervals can be found by setting it to p and $1 - p$.

3.4 Features

Our ranking and estimation model can work with any kind of numerical features. For simplicity we used lexical and naive morphological features, pruned using χ^2 feature selection with tunable granularity.

The lexical features are occurrence counts of all words that appear in at least p_{lex} documents. The morphological features are counts of character n-grams of length up to w_{mph} in final positions of words, filtered to occur in at least n_{mph} documents.

Subsequently, a non-linear transformation ϕ is optionally applied to the numerical features. This is one of $\phi_{\text{sqrt}}(z) = \sqrt{z}$, $\phi_{\log}(z) = \log(z)$ or $\phi_{\text{id}}(z) = z$ (no transformation).

The feature selection step is applied before generating the pairs for classification, in order for the χ^2 scoring to be applicable. The raw target values used are year labels, but to avoid separating almost every document in its own class, we introduce a *granularity* level that transforms the labels into groups of n_{gran} years. For example, if $n_{\text{gran}} = 10$ then the features will be scored according to how well they predict the decade a document was written in. The features in the top p_{fsel} percentile are kept. Finally, C is the regularization parameter of the logistic regression classifier, as defined in *liblinear* (Fan et al., 2008).

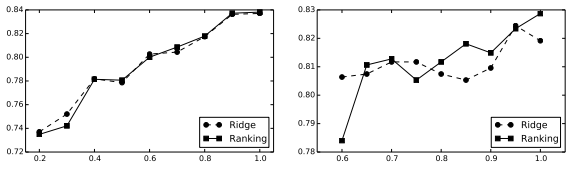

Figure 1: Learning curves for English (top) and Portuguese (bottom). Proportion of training set used versus score.

4 Results

Each corpus is split randomly into training and test sets with equal number of documents. The best feature set is chosen by 3-fold cross-validated random search over a large grid of possible configurations. We use random search to allow for a more efficient exploration of the parameter space, given that some parameters have much less impact to the final score than others.

The evaluation metric we used is the percentage of non-inverted (correctly ordered) pairs, following (Pedregosa et al., 2012).

We compare the pairwise logistic approach to a ridge regression on the same feature set, and two multiclass SVMs, at century and decade level. While the results are comparable with a slight advantage in favour of ranking, the pairwise ranking system has several advantages. On the one hand, it provides the probabilistic interpretation described in section 3.3. On the other hand, the model can naturally handle noisy, uncertain or wide-range labels, because annotating whether a text was written before another can be done even when the texts do not correspond to punctual moments in time. While we do not exploit this advantage, it can lead to more robust models of temporal evolution. The learning curves in Figure 1 further show that the pairwise approach can better exploit more data and nonlinearity.

The implementation is based on the *scikit-learn* machine learning library for Python (Pedregosa et al., 2011) with logistic regression solver from (Fan et al., 2008). The source code will be available.

4.1 Uncertain texts

We present an example of using the method from Section 3.3 to estimate the date of uncertain, held-out texts of historical interest. Figure 2 shows the process used for estimating θ as a linear, and in the case of Portuguese, quadratic function. The

	size	p_{lex}	n_{mph}	w_{mph}	ϕ	n_{gran}	p_{fsel}	C	score	ridge	century	decade	MAE
en	293	0.9	0	3	$\phi_{\log}$	100	0.15	2^{9}	0.838	0.837	0.751	0.813	22.8
pt	87	0.9	25	4	ϕ_{sqrt}	5	0.25	2^{-5}	0.829	0.819	0.712	0.620	58.7
ro	42	0.8	0	4	$\phi_{\log}$	5	0.10	2^{28}	0.929	0.924	0.855	0.792	28.8

Table 1: Test results of the system on the three datasets. The score is the proportion of pairs of documents ranked correctly. The column *ridge* is a linear regression model used for ranking, while *century* and *decade* are linear SVMs used to predict the century and the decade of each text, but scored as pairwise ranking, for comparability. Chance level is 0.5. MAE is the mean absolute error in years. The hyperparameters are described in section 3.4.

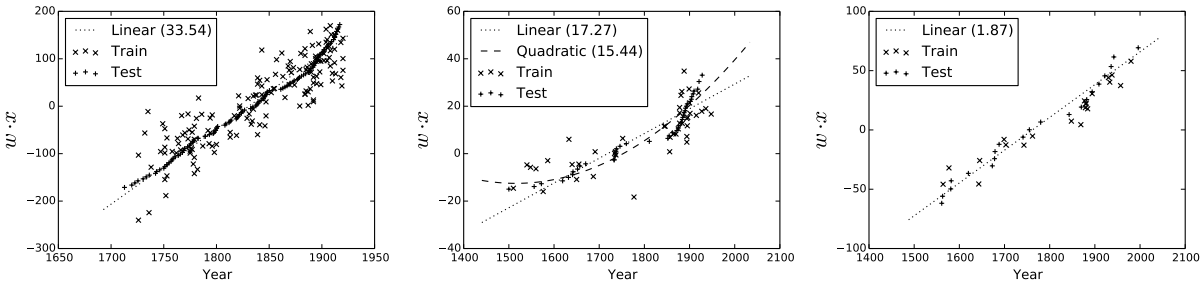

Figure 2: Estimating the function θ that defines the relationship between years and projections of documents to the direction of the model, for English, Portuguese and Romanian (left to right). In parantheses, the normalized residual of the least squares fit is reported on the test set.

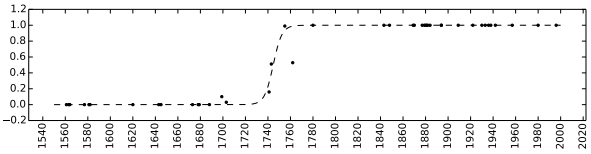

Figure 3: Visualisation of the probability estimation for the dating of C. Cantacuzino's *Istoria Țării Rumânești*. The horizontal axis is the time, the points are known texts with a height equal to the probability predicted by the classifier. The dashed line is the estimated probability from Equation 2.

estimation is refit on all certain documents prior to plugging into the probability estimation.

The document we use to demonstrate the process is Romanian nobleman and historian Constantin Cantacuzino's *Istoria Țării Rumânești*. The work is believed to be written in 1716, the year of the author's death, and published in several editions over a century later (Stahl, 2001). This is an example of the system being reasonably close to the hypothesis, thus providing linguistic support to it. Our system gives an estimated dating of 1744.7 with a 90% confidence interval of $1736.2 - 1753.2$. As publications were significantly later, the lexical pull towards the end of 18th century that can be observed in Figure 3 could be driven by possible editing of the original text.

5 Conclusion

We propose a ranking approach to temporal modelling of historical texts. We show how the model can be used to produce reasonable probabilistic estimates of the linguistic age of a text, using a very basic, fully-automatic feature extraction step and no linguistic or historical knowledge injected, apart from the labels, which are possibly noisy.

Label noise can be atenuated by replacing uncertain dates with intervals that are more certain, and only generating training pairs out of non-overlapping intervals. This can lead to a more robust model and can use more data than would be possible with a regression or classification approach. The problem of potential edits that a text has suffered still remains open.

Finally, better engineered and linguistically-motivated features, such as syntactic, morphological or phonetic patterns that are known or believed to mark epochs in the evolution of a language, can be plugged in with no change to the fundamental method.

References

H. Abe and S. Tsumoto. 2010. Text categorization with considering temporal patterns of term usages. In *Proceedings of ICDM Workshops*, pages 800–807. IEEE.

A. Ciobanu, A. Dinu, L. Dinu, V. Niculae, and O. Sulea. 2013. Temporal text classification for romanian novels set in the past. In *Proceedings of RANLP2013*, Hissar, Bulgaria.

W. Dakka and C. Gravana. 2010. Answering general time-sensitive queries. *IEEE Transactions on Knowledge and Data Engineering*.

A. Dalli and Y. Wilks. 2006. Automatic dating of documents and temporal text classification. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 17–22, Sidney, Australia.

F. de Jong, H. Rode, and D. Hiemstra. 2005. Temporal language models for the disclosure of historical text. In *Proceedings of AHC 2005 (History and Computing)*.

H. de Smet. 2005. A corpus of late modern english. *ICAME-Journal*.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

N. Kanhabua and P. Nørvåg. 2009. Using temporal language models for document dating. In *ECML/PKDD*, pages 738–741.

W. Kraaij. 2004. *Variations on language modeling for information retrieval*. Ph.D. thesis, University of Twente.

A. Kumar, M. Lease, and J. Baldridge. 2011. Supervised language modelling for temporal resolution of texts. In *Proceedings of CIKM11 of the 20th ACM international conference on Information and knowledge management*, pages 2069–2072.

R. Mihalcea and V. Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *Proceedings of ACL*, pages 259–263. Association for Computational Linguistics.

S. Mokhov. 2010. A marf approach to deft2010. In *Proceedings of TALN2010*, Montreal, Canada.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Fabian Pedregosa, Alexandre Gramfort, Gaël Varoquaux, Elodie Cauvet, Christophe Pallier, and Bertrand Thirion. 2012. Learning to rank from medical imaging data. *CoRR*, abs/1207.3598.

H.H. Stahl. 2001. *Gânditori şi curente de istorie socială românească*. Biblioteca Institutului Social Român. Ed. Univ. din Bucureşti.

S. Štajner and M. Zampieri. 2013. Stylistic changes for temporal text classification. In *Proceedings of the 16th International Conference on Text Speech and Dialogue (TSD2013), Lecture Notes in Artificial Intelligence (LNAI)*, pages 519–526, Pilsen, Czech Republic. Springer.

D. Trieschnigg, D. Hiemstra, M. Theune, F. de Jong, and T. Meder. 2012. An exploration of language identification techniques for the dutch folktale database. In *Proceedings of LREC2012*.

D. Wijaya and R. Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proc. of the Workshop on Detecting and Exploiting Cultural Diversity on the Social Web (DETECT)*.

M. Zampieri and M. Becker. 2013. Colonia: Corpus of historical portuguese. *ZSM Studien, Special Volume on Non-Standard Data Sources in Corpus-Based Research*, 5.