Probabilistic Modeling of Chronological Dates to Serve Machines and Scholars

Habring, Andreas
andreas.habring@uni-graz.at
University of Graz, Austria

Nicolaou, Anguelos
anguelos.nicolaou@gmail.com
University of Graz, Austria

Luger, Daniel
daniel.luger@uni-graz.at
University of Graz, Austria

Atzenhofer-Baumgartner, Florian
florian.atzenhofer-baumgartner@uni-graz.at
University of Graz, Austria

Lamminger, Florian
florian.lamminger@uni-graz.at
University of Graz, Austria

Decker, Franziska
franziska.decker@uni-graz.at
University of Graz, Austria

Aoun, Sandy
sandy.aoun@uni-graz.at
University of Graz, Austria

Kovács, Tamás
tamas.kovacs@uni-graz.at
University of Graz, Austria

Vogeler, Georg
georg.vogeler@uni-graz.at
University of Graz, Austria

Holler, Martin
martin.holler@uni-graz.at
University of Graz, Austria

Probabilistic Modeling of Chronological Dates to Serve Machines and Scholars

We present an approach for modeling dates of creation of documents which allows scholars to express uncertainty when annotating data, acts as a differentiable loss function for training models and enables unbiased interpretable performance evaluation under uncertainty.

Finding an appropriate date to a historical document is a task of digital diplomatics (Gervers, 2000; Gervers and Margolin, 2003; Feuerverger et al., 2005; Tilahun et al., 2012; Tilahun et al., 2014). Specifically Stokes, 2015 provides an analysis of the intricacies in modeling dates in the case of medieval documents. In this context, the time of creation and its uncertainty can be suggested by scholars as well as machine learning models. A framework that allows to compare two different estimates over timing a document in a way that agreement between two opinions can be meaningfully quantified is needed in order to train neural networks that estimate dates, evaluate the performance of such >neural networks, or even analyse quantitatively the consensus among scholars. On top of that, from a machine learning perspective, timing documents can be approached in very different ways: choosing from a set of time periods (classification) (Cloppet et al., 2016), predicting a time span (regression) (Seuret et al., 2021), etc. These approaches produce incomparable outputs and in order to decide which one performs best, a method of quantitatively comparing them is necessary. The core problem is how uncertainty in >a prediction is expressed, whether it is by humans or algorithms. The most straight forward way for an expert to describe an ambiguous date is a pair of "not before" and "not after" limits. While this also allows to express exact time-points by setting "not before" and "not after" to be identical, it is not clear whether this format is optimal for machine learning as well. Furthermore, defining an objective similarity measurement between pairs of "not before" and "not after" is not trivial especially if the ranges are interpreted as plausibility. Moreover, even more delicate formats may occur, such as knowing only a day without the corresponding year (Thaller, 2020).

We propose to model all forms of expressing an opinion on a documents date as Probability Density Functions (PDF) over time. Figuratively speaking, each point in time is

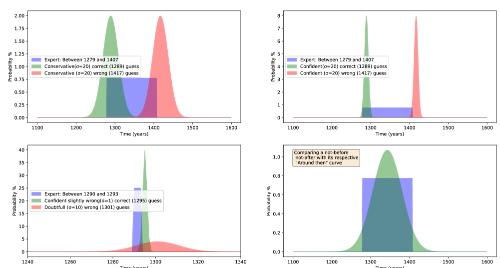

Figure 1. Intersection of closed form PDF’s as a quantification of agreement

assigned a probability of being the time of writing of the given document. This probability can be 100% on a specific day and 0% at any other day, in the case of knowing the exact day of writing, it can be evenly spread across an interval, in the case of knowing only a time span, or it can be any other distribution, e.g., normally distributed around a certain time point. We also propose to quantify agreement between two options by a suitable distance measure between the PDF's. One possibility to achieve this is the Wasserstein distance (Frogner et al., 2015). As an alternative, one can also measure the surface where the PDF's overlap. While both possibilities can accommodate any kind of PDF over time, the two major ways of expressing ambiguity is "not before, not after" which we model as a uniform distribution and "around then" which we model as a Gaussian distribution. In Figure 1 these two ways of expressing ambiguity are illustrated. Both can be represented by a pair of numbers, start and end point in the case of a time span and mean and standard deviation in the Gaussian case, and both can also accommodate a fixed point in time as long as there is a minimum meaningful time-interval (dt) defined. In order to obtain a differentiable loss function for training a machine learning method, any form of PDF can be convolved with a Gaussian kernel while introducing negligible distortions from an information theory perspective. In the context of performance evaluation for point predictions, it makes more sense to define any moment in the interval defined by scholars as 100% plausible. One way to achieve this (which is also applicable for "around then" intervals from scholars) is to integrate a delta peak (modeling the point prediction), or any other PDF modeling a more nuanced prediction, against a density function renormalized to have at least one maximum of 100%.

Bibliography

Cloppet, F., Eglin, V., Stutzmann, D., Vincent, N., et al. (2016). Icfhr2016 competition on
the classification of medieval handwritings in latin script. 2016 15th International
Conference on Frontiers in Handwriting Recognition (ICFHR), 590–595.

Feuerverger, A., Hall, P., Tilahun, G., & Gervers, M. (2005). Distance measures and
smoothing methodology for imputing features of documents. Journal of Computational
and Graphical Statistics 14, 14, 255–262.

Frogner, C., Zhang, C., Mobahi, H., Araya, M., & Poggio, T. A. (2015). Learning with a
Wasserstein loss. Advances in neural information processing systems, 28.

Gervers, M. (Ed.). (2000). Dating undated Medieval Charters. Boydell Pr.

Gervers, M., & Margolin, M. (2003). Application of Computerized Analyses in Dating
Procedures for Medieval Charters. Le Médiéviste et l'ordinateur, 42 http://lemo.irht.cnrs.fr/42/mo42_01.htm

**Seuret, M., Nicolaou, A., Rodriguez-Salas, D., Weichselbaumer, N., Stutzmann, D.,
Mayr, M., Maier, A., & Christlein, V.** (2021). Icdar 2021 competition on historical
document classification. International Conference on Document Analysis and Recognition,
618–634.

Stokes, P. (2015). The problem of digital dating: A model for uncertainty in medieval
documents. Digital Humanities 2015 Book of Abstracts. http://peterstokes.org/pubs/Stokes_digital_dating.pdf

Thaller, M. (2020). On vagueness and uncertainty in historical data. Retrieved October
26, 2022, from https://ivorytower.hypotheses.org/88

Tilahun, G., Feuerverger, A., & Gervers, M. (2012). Dating medieval English charters.
The Annals of Applied Statistics, 6(4), 1615–1640 https://doi.org/10.1214/12-%20AOAS566

Tilahun, G., Gervers, M., & Feuerverger, A. (2014). Statistical methods for applying
chronology to undated English medieval documents. In A. Ambrosio, S. Barret, & G. Vogeler (Eds.), Digital diplomatics (pp. 211–224). Böhlau Verlag. https://doi.org/10.7788/boehlau.9783412217020.211