

Dating Tablets in the Garshana Corpus

**James Hearne, Connor Anderson, Yudong Liu
Dannon Dixon, Dario Fenstermacher Ritchie**

Computer Science Department
Western Washington University
Bellingham, Washington 98226

Abstract

This paper reports on an effort to assign dates to undated tablets in the Garshana corpus, a fully curated and annotated collection of Sumerian tables from the Ur III era. Of the 1488 tablets, 92% are dated, giving a strong training set for dating those undated. Two approaches were pursued: (1) a naive one which determined the date of a tablet by simply counting the name overlap with tablets of known year and (2) invoking a collection of machine learning algorithms of established robustness. The naive method reports an accuracy of 45-55% and the machine learning algorithms achieve 78.8-84.15% accuracy.

Introduction

The Third Dynasty of Ur, a 21st to 20th century BC Sumerian was a period of accelerated economic activity that gave rise to advances in record-keeping. As a consequence, we have in the archaeological record a vast residue documenting commercial and social relations. The vast majority of these tables recording a financial transaction include (1) its agent, (2) its patient, (3) a list of witnesses, (4) a named scribe and (5) a date. Fig. 1 shows the tablet with an id of P105955 from CDLI repository (<http://cdli.ucla.edu/>). The original cuneiform script is on the left; the transliteration is in the middle and the modern English translation is on the right.

Such information supports a variety of historical studies, such as the investigation of specific historical actors, or, in the spirit of the data mining tradition, the development of a complete social network of ancient Sumerian merchant society. Such a social network is one of the larger purposes of the project of which the current study is a part. The development of a social network can serve at least two valuable purposes. The obvious one is that such a network would support a study of merchant society that included the identification of significant actors, ranges of movement and perhaps average lifespans.

It also serves a meta-historical purpose. Most of the tables now available to scholars were, in effect, looted, with consequence that their original provenance is uncertain. This fact renders difficult or impossible the ability to reconstruct the

original state-owned archives from which they came, depriving scholars of the ability to reconstruct local economic activity. Tablets originally stored in the same local archive are now scattered throughout the world in museums and universities. The ability to identify dense regions of a larger social network, quasi-cliques, would support the assignment of a tablet to a specific but now scattered archive. Since names were reused in Sumerian society, as they are in most, the identification of tablet dates is essential for such an assignment.

This overarching goal requires the identification of two sorts of information from the original tablets, dates and personal names. It also requires, as this study reveals, the identification of geographical names.

Impediments

To mine such information requires overcoming two problems. The vast majority of tablets are written in Sumerian, an isolate language ill-understood even after over 150 years of scholarly investigation. A casual exploration conducted by us showed that roughly half of the lexical items found in the UR III Corpus are *hapax-legomena*, words having but a single attestation, a fact that greatly impedes decipherment. In addition, although there are now scholars able to make sense of tablets relating to merchant records, the corpus is of a size too large for even a community of scholars to master. One source estimates over 90,000 published tablets and tens of thousands yet unpublished (Widell 2008). The sheer size of the corpus necessitates seeking the automatic extraction of personal names, geographical names and dates. Fortunately, most tablets are dated and, where dates are identifiable, with a few exceptions, can be automatically recognized.

The Garshana Corpus

The Garshana corpus is a fully curated and annotated corpus of tables from the Ur III period provided to us by Manuel Molina of CSIC¹. It is fully curated in the sense that all place names, personal names and dates are identified and normalized, making machine processing possible. Since 92% of the 1,488 tablets are properly dated, it provides an ideal machine learning environment for testing dating techniques. The corpus is temporally circumscribed, covering a span of

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹<http://cchs.csic.es/es/personal/manuel.molina>



Cuneiform Tablet	Transliteration	English Translation
	&P105955 = BIN 03, 149	(unique tablet identification)
obverse	@tablet	
	@obverse	(front of tablet)
	1. 1(disz) sila4	1 lamb
	2. ki ab-ba-sa6-ga-ta	from Ab-ba-sa-ga (the seller)
	3. ur-{d}szul-pa-e3	Ur-Šul-pa-e (the buyer)
	@reverse	(back of tablet)
reverse	1. i3-dab5	received
	2. iti u5-bi2-gu7	month: 3
	3. mu hu-uh2-nu-ri{ki} ba-hul	year: Huhnuri was destroyed
	@seal	(tablet sealed by)
	1. ur-{d}szul-pa-e3	Ur-Šul-pa-e
	2. dub-sar	the scribe
	3. dumu da-a-da kuruszda	son of Da-a-da the fattener

Figure 1: Tablet with ID of P105955 from CDLI.

40 years; the dates in the corpus are sparse and only twelve unique years are present. In the corpus, years are often named after events, instead of chronologically. Therefore, inconsistencies occur when similar events happen more than once. For example, there is an ambiguity between *Shulgi Year 42* and *Amar-Suen Year 6*. For *Shulgi 42*, it's representation was "Year: The king destroyed Shashrum". For *Amar Suen 6*, it's "Year: Amar-Suen, the king, destroyed Shashrum for the second time...". Both were abbreviated to "Year: Shashrum was destroyed" in the tablets. Therefore, an ambiguity occurred. Such ambiguity can be potentially solved if a tablet can be properly dated.

Previous Work

Research on the uses of techniques developed in the natural language processing and machine learning traditions to ancient languages is not voluminous. By and large, the application of computer technology has been limited to electronic publishing and string searching. Such efforts applied to Sumerian are even more sparse. Tablan et al. (Tablan et al. 2006) described the creation of a tool for Sumerian linguistic analysis and corpus search applied to Sumerian literature. This work focused on the application of a morphological model for noun and verb recognition, developed by Sumerologists. However, it did not extend beyond simple word and morpheme recognition. Other work developed an approach to extraction of information from the same economic documents of concern to us by beginning with an ontology of the world presupposed by this corpus and, in conjunction with syntax and semantic analysis (Jaworski 2008). The claim of this work is that it supports research into the Sumerian language, officials participating in the activities the tablets record, and into document classification. In addition, the authors previous work apply NLP and machine learning techniques to the problem of proper name recognition (Luo et al. 2015) (Liu, Hearne, and Conard 2016).

Methods

Naive Approach

Under this heading two sorts of experiments were conducted. In the first one, dates were determined by the raw, un-weighted overlap of names in tablets of known date with those whose date is unknown. For the second one, in the spirit of social networks, we hypothesized that some ancient actors were of greater importance than others and developed a weighted formula for determine name overlap, using this formula:

For every unique personal name PN_i in a tablet, let (1) $C(PN_i, YR_j)$ denotes the number of personal name PN_i observed in (tablets of) a given year YR_j , and (2) $C(YR_i)$ denotes the total number of personal names observed in (tablets of) YR_j :

$$\text{weighted sum} += C(PN_i, YR_j) / C(YR_i)$$

Fig. 2 shows the results of the two naive methods by adopting different "window size" (i.e., the number of years under consideration).

We applied this method to increasing year windows, from an exact year match to windows of five years. Significantly, the performance of the weighted criterion is an improvement over the unweighted measure overlap, but the two measures give progressively similar precision as the window of tolerance broadens. With a window size of 1, using only proper names, we achieved an accuracy of 45-55%.

Machine Learning Methods

To compensate for the limitations of either naive approach, we both (1) invoked established machine learning algorithms and (2) augmented the training data set by adding toponyms or geographic names (GN). Five machine learning algorithms were chosen for this investigation, all of them implemented by sci-kit learn (Pedregosa et al. 2011). Each im-

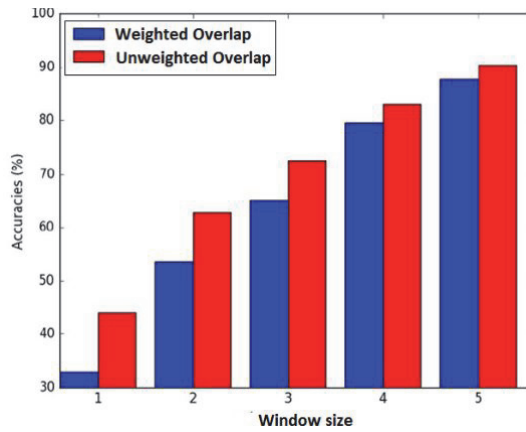


Figure 2: results of naive methods with window sizes of 1 to 5 where size of 1 indicates an exact match).

plementation had a number of options for customizing its performance, such as learning rate and number of training iterations. The values for these parameters were chosen by first selecting a list of reasonable values for each option, then for each combination of values a model was trained, and the model giving the best accuracy was kept. The algorithms are: (1) Multinomial Naive Bayes (2) Decision Trees, (3) Stochastic Gradient Descent, (4) Perceptrons and (5) Support Vector Machine. In sci-kit learn, SVM uses “one-against-one” for multi-classification.

Each model is trained and tested following 5-Fold Cross Validation methods, with an 80-20 split train/test set.

Results

Performance of the computational experiments give below in Table 1. We calculated estimates of undated tablets by consulting overlapping features of tables of known date. The feature set was necessarily meager and meaningful features were restricted to (1) personal names or (2) personal names together with geographic names. Attention to the table of results shows that the inclusion of geographic names (GNs) improved results considerably, as much as 25%, in almost all of the methods.

Method	Accuracy with no GNs	Accuracy with GNs
Decision Tree	56.76	78.80
Perceptron	51.49	76.76
Naive Bayes	54.69	72.65
Gradient Descent (SGD)	58.51	82.69
SVM	61.05	84.15

Table 1: Accuracy (%) without and with Geographic Names (GNs)

Conclusion and Future Work

In this paper we reported our work on dating tablets of Garshana corpus. Due to that the tablets were dated with events

but not chronologically, and similar event may occur more than once, tablets may be dated ambiguously. The ultimate goal of dating tablets is to solve such ambiguity. Our work is an exploration of such a possibility.

Based on the overlap of the important information, such as personal names and geographical names, on the tablets, we adopt a naive method and a set of machine learning methods to date the tablets. Results show that the machine learning methods outperform the naive method, and integration of the geographic names in the machine learning methods have greatly improved the performance.

For future work, we would like to integrate more features such as relations between persons, and profession names. A graphical model that captures the relations between person names across tablets could also potentially improve the performance. We also intend to extend this method to uncurated corpora of Sumerian texts toward the ultimate goals of exploring ancient social networks. This work will also ultimately be brought to bear in the task of reassembling tablets which were in the same ancient archive and which are now distributed in museums and universities throughout the world (Molina 2008).

Acknowledgments

The authors wish to thank Mr. Manuel Molina for making available the data set.

References

- Jaworski, W. 2008. Contents modeling of neo-sumerian ur iii economic text corpus. In *Proceedings of the 22nd International Conference on Computational Linguistics*, 369–376.
- Liu, Y.; Hearne, J.; and Conard, B. 2016. Recognizing proper names in ur iii texts through supervised learning. In *Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2016, Key Largo, Florida, May 16-18, 2016.*, 535–540.
- Luo, L.; Liu, Y.; Hearne, J.; and Burkhart, C. 2015. Unsupervised sumerian personal name recognition. In *Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2015, Hollywood, Florida, May 18-20, 2015.*, 193–198.
- Molina, M. 2008. The corpus of neo-sumerian tablets: An overview. In Garfinkle, S., and Cale Johnson, J., eds., *The Growth of an Early State in Mesopotamia: Studies in Ur III Administration*. Madrid: Consejo Superior de Investigaciones Científicas. 19–54.
- Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.
- Tablan, V.; Peters, W.; Maynard, D.; and Cunningham, H. 2006. Creating tools for morphological analysis of sumerian. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 1762–1765.
- Widell, M. 2008. The ur iii metal loans from ur. In Garfinkle, S., and Cale Johnson, J., eds., *The Growth of an Early State in Mesopotamia: Studies in Ur III Administration*. Madrid: Consejo Superior de Investigaciones Científicas. 207–223.