Unsupervised Sumerian Personal Name Recognition

Liang Luo, Yudong Liu, James Hearne and **Clinton Burkhart**
Computer Science Department
Western Washington University
Bellingham, Washington 98226
{luol@students.-yudong.liu@-james.hearne@-burkhac@students.}wwu.edu

Abstract

This paper describes an unsupervised named-entity recognition (NER) system to identify personal names in Sumerian cuneiform documents from the Ur III period. We are motivated by the needs of social and economic historians of that period to identify specific persons of importance and such historically relevant facts as can be discerned by the surviving texts. The work was confronted by the challenges posed by the fact that Sumerian is not a well understood language and the texts come down to us in damaged condition. We based our recognizer on the Decision List CoTrain algorithm, subjecting it experimentally to modifications to accommodate the nature of the data and narrower task it was originally devised for. We achieved 92.5% recall and 56.0% precision, results that are usable by the economic and social historian. We described the results of our work and suggest further applications of the techniques we have devised, also in the analysis of ancient Sumerian texts.

Introduction

The vast majority of clay tablets from the Third Dynasty of Ur, also known as the Neo-Sumerian Empire or the Ur III Empire, record financial transactions, such as records of cattle deliveries, receipt of metals, repayment of loans, and so forth. Importantly, in addition to the provider and recipient of transference, tablets consistently enumerate lists of witnesses. This fact makes the tablet an invaluable resource for the social history of the time since they record, implicitly, on each tablet, lists of persons who knew one another and who enjoyed professional relations with one another (Widell 2008). The recovery of personal names on the tablets suggests the possibility of reconstructing social networks of actors in the mercantile class and also, given the overlap, their social network connections to royalty.

Use of the tablets for such purposes is impeded in two ways. First, they are by and large written in Sumerian, a language already in its twilight at the time the tables were being written and which, even after nearly two centuries of scholarly investigation, is ill-understood. A causal 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, greatly impeding decipherment. The problem is further compounded by the fact that the tablets come down to us in many cases damaged by time and the circumstances of their recovery which was, in many cases, looting. Second, although there are now scholars well enough able to make sense tablets relating to merchant records (Garfinkle 2012), 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 (Molina 2008); further, new tablets are still being uncovered, even though the region they are found in, modern Iraq, has been troubled by warfare for well over a decade.

Our approach to this problem, described below, has been to apply methods from the natural language processing (NLP) and machine learning traditions (Collins and Singer 1999) to the problem of identifying the personal names in each legible tablet. Needless to say, this effort presupposes the cooperation of a domain expert, cooperation generously provided by Steven Garfinkle, a Sumerologist in our home university.

Background

Previous Work

Research on the use of techniques developed in the natural language processing and machine learning traditions to ancient languages is not abundant. 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. (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 both into the Sumerian language, officials participating in the activities the tablets record and in document classification.

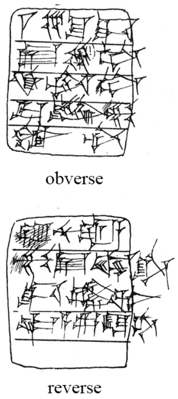
Cuneiform Tablet	Transliteration	English Translation
	`&P100079 = AAS 092`	
	`@tablet`	
	`@obverse`	
	`1. 1(barig) sze zi3-da`	60 liters barley flour
	`2. e2-kikken-ta`	mill
	`3. ki ARAD2-ta`	from ARAD2
	`4. ur-lugal muhaldim`	(to) Ur-lugal the cook
	`5. szagina`	(of the) general
	`@reverse`	
	`1. kiszib3 nu-ra-a`	sealed by Nuraya
	`@date`	
	`2. mu e2 |PU3.SZA|-da-gan ba-du3`	year: the temple of Szulgi was built

Figure 1: An example of a transliterated tablet with an ID of P100079.

Available Data

There are two main Sumerian tablet databases: the cuneiform Digital Library Initiative (CDLI, http://cdli.ucla.edu/) and the Database of Neo-Sumerian Texts (BDTNS, http://bdtns.filol.csic.ed/). CDLI is a project that concentrates on electronic documentation of ancient cuneiform, consisting of cuneiform texts, images, transliterations and glossaries of 3500 years of human history. It is managed by UCLA, USA and the Max Planck Institution for the History of Science, Berlin. BDTNS is a database that manages more than 95,300 administrative Sumerian cuneiform tablets during the Neo-Sumerian period. In this project, we make use of the CDLI repository because its restriction to the ASCII character set is more convenient than the UTF-8 encoding used by BDTNS. Of the CDLI repository we use the 53,146 tablets having lemmata.

Data

Words and Signs

Although the texts we are investigating were originally written in cuneiform script, scholars have traditionally worked with transliterations using the English alphabet. Homophony, which is very common in cuneiform, is handled by numerical subscripts, i.e., "gu," "gu_2" and "gu_3" refer to distinct signs with the same sound value. Sumeriologists have developed a system of in-text annotations important to the application of NLP techniques. Figure 1 shows the tablet with an id of P100079 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.

Sumerian written in cuneiform script does not have the concept of upper- or lowercase, and as a result, in monolingual contexts, the typical step of case normalization is not necessary as all text is rendered in lowercase in the transliteration format used by CDLI. However, signs rendered in uppercase occur frequently and denote a number of special circumstances, most commonly, that the phonetic equivalent of the sign is unknown. With respect to our system, the presence of uppercase signs is rarely problematic as long as the spelling and casing is consistently attested in the corpus. This consistency is provided by the lemmatization required of Sumerologists in order to submit initial transliterations (Foxvog 2014).

Royal epithets notwithstanding, Sumerian personal names are exclusively comprised of a single word, almost always consisting of at least two signs. In cases where disambiguation is required, a patronymic may added (for example, szu-esz4-tar2 dumu zu-zu, "Su-Estar, son of Zuzu"). This disambiguation is frequent in practice due to the relatively shallow pool of personal names used (Limet 1960).

Lemmata

62.1% of the tables in the CDLI corpus are accompanied by annotations, dubbed "lemmata" which provide, stemification, translation and categorization of linguistic elements within a line. Thus, the example given in Figure 2 gives a line of text followed by its lemmatization, indicated with a line-initial "#." In this example, the lemmatization indicates that the word "GAN2" derives from the Sumerian stem "iku" and is understood to mean "unit." Similarly, "ur-{gesz}-gigir" is a personal name; "nu-banda3" with the Sumerian stem "nubanda" means "overseer", a common profession in Sumerian society; the word "gu4" with the Sumerian stem "gud" means "ox." Sometimes, when the function of a word is uncertain, the lemmatization offers more than one cate-

```
4. GAN2 ur-{gesz}gigir nu-banda3 gu4
#lem: iku[unit]; PN; nubanda[overseer]; gud[ox]
```

Figure 2: A lemmatized transliteration.

gory. For example, lemma “GN|FN” indicates that the corresponding can be either a geographical name or the name of a field (analogous to names such as “Potter’s Field” in English).

Damaged Tablets

In CDLI corpus, some transliterations have “[” and “]” attaching to a sign. It indicates the sign is damaged. More specifically, “[” indicates that the following sign is damaged on the left edge, whereas “]” indicates the following sign is damaged on the right edge (Sahala 2012). “[x]” indicates that there is a sign that cannot be read due to the damage on both edges and “[...]” indicates that there are several unreadable signs. This will be discussed in more detail in the Annotation section.

Method

The NER system has three components: the pre-processing component, the Decision List Co-Train (DL-CoTrain) (Collins and Singer 1999) component and the post-processing component. We will discuss each component in detail in the following.

Pre-processing

The transliteration standard used to represent the Sumerian corpus contains metacharacters specific to the many different types of damage and modification that the original cuneiform tablets are subject to. Since these characters arouse unnecessary complications, we removed them as noise. After removing these metacharacters, we also pre-tag the corpus with a limited amount of pre-knowledge provided by our language expert.

Noise Removal When the Sumerologists transliterate the tablets, they use metacharacters such as “[...]” and “#” to indicate damage to the text, and “!”, “?”, “*”, and “<...>” to represent correction, querying or collation (Tinney and Robson 2014). For “[...]” and “<...>” cases, the Sumerologists put their “best guess” within the brackets. For example, in the word “[nu]-su”, the first sign was originally damaged but restored by the Sumerologists as the “best guess”. Our system removes the metacharacters as noise, and treats the resulting text as if it were otherwise unannotated.

Annotation To utilize the pre-knowledge from the language experts and (Weibull 2004), we apply a tag set of 13 tags to pre-annotate the corpus. The 13 tags in the tag set {“GN”, “FN”, “TN”, “WN”, “MN”, “n”, “TITLE”, “UNIT”, “GOODS”, “OCCUPATION”, “YEAR”, “MONTH”, “DAY”} represent geographical names, field names, temple names, watercourse names, month names, numbers, title names, unit names, trade goods names, occupation names and indicators for year, month and day, respectively.

Decision List CoTrain (DL-CoTrain)

Our NER system is built over DL-CoTrain model (Collins and Singer 1999), utilizing contextual and spelling rules to create a decision list.

Contextual Rules and Spelling Rules A contextual rule specifies the context for a named-entity with the window size of 1 or -1 (the right word or the left word). For example, according to the contextual rule “right_context=TITLE → Person”, “nam-zi” is recognized as a personal name in “nam-zi simug” given that “simug” is pre-tagged as “TITLE” (Smith) in the pre-processing phase.

A spelling rule specifies the spelling of a named-entity. It is a sign sequence that can be either the full string of an entity or is contained as a substring of the entity. For example, “contains(e2-kikken) → Person” is a spelling rule. By applying the rule, the word “e2-kikken-ta” is recognized as a personal name. With the spelling rule “full-string={d}en-lil2 → Person”, the word “{d}en-lil2” is recognized as a personal name.

Seed Rules As suggested by the language expert, the following three contextual rules are used as the seed rules for our NER system:

left_context=giri3 → Person

left_context=kiszib3 → Person

left_context=mu-DU → Person

The first rule indicates that a person is acting as an intermediary in the transaction. The second rule indicates that the tablet was sealed by the named individual, and usually appears in administrative records. The last rule indicates that a delivery was made to the named individual. Since these seed rules have a high specificity to personal names, each of them is given a strength of 0.99999.

Algorithm The major task of the system is to learn a decision list to classify a word as a personal name. Initialized with the 3 contextual seed rules, the decision list is applied to label the training data to get spelling rules. In the next iteration, the newly obtained spelling rules are applied to label the training data to get new contextual rules. In this alternating process, each iteration produces a new set of rules which are ranked by their strength. In our system, the top 20 rules with the highest strength are added to the decision list.

In (Collins and Singer 1999), the strength of a rule ($x \rightarrow y$ where x is either a contextual feature or a spelling feature, and y is the label that belongs to one of the categories {Person, Organization, Location}) is defined as follows:

$$strength = \frac{Count(x, y) + \alpha}{Count(x) + k\alpha} \qquad (1)$$

where $Count(x, y)$ denotes the number of times a feature x is seen with label y and $Count(x)$ is the total occurrence of feature x. α is a smoothing parameter and k is the number of possible labels.

In our NER system, we first adopted the same formula, restricting the lable y to PN and hence $Count(x, PN)$ denotes the number of times that a feature x and a personal name (PN) appear together in training data, and $Count(x)$ is the total occurrence of feature x in training data. The value of α is set to 0.1 and k is set to 2 experimentally. This practice responded to the problem of finding rules which correctly identify a single name. However, in subsequent iterations it yielded no new names because it identified a context that occurred only once. We therefore experimentally settled on a ranking criterion that made use of frequency of some feature x, reverting to the original formula in the above Equation (2) in the case of ties.

This algorithm, if iterated 150 cycles, produces a decision list of over 2000 rules and approximately 17,000 personal names in these Sumerian texts. More results are rehearsed below.

Post-Processing

Our domain expert suggested that two post-processing rules could be applied to eliminate false positives. The application of the following rules improved the performance by 0.5%.

- A word that starts with a number should not be a name.
- A word following the word "iti" (month indicator) should not be a name.

Experiments

We trained our NER system by using the algorithm described above and applied the rules from the decision list on the testing data set. The result was compared with the result from a baseline system on the same testing data set.

Experimental Setup

We used a 5-fold cross-validation model to test our NER system. In each fold, we randomly picked 85% of the tablets from the corpus as our training set and the remaining 15% of the tablets as the testing set. Our NER system spent 150 iterations to be trained. During each iteration, 20 new rules were generated and added to the decision list. We use the lemmatization as the gold standard data set in this experiment.

Experimental Results

We ran a baseline system on the same testing set from each fold to compare with the results from our NER system.

Baseline System The baseline algorithm employs a naive pattern-matching algorithm to attempt to identify personal names based solely on sequences of words that appear commonly in the corpus (Brewer et al. 2014). Those 12 patterns are: "dumu PN TITLE", "PN szu ba-ti", "ki PN", "1(disz) PN", "n UNIT PN", "kiszib3 PN", "PN TITLE", "PN i3-dab5", "n UNIT PN", "giri3 PN", "n PN" and "mu-DU PN", where "PN" indicates a personal name and the meanings of other tags can be found in the Annotation Section.

The baseline system takes the transliterated text of each tablet and populates a queue with every word in that tablet. Each pattern in the collection is then tested against the words at the beginning of the queue in decreasing order of pattern length, since it is assumed that the longer patterns are most specific. If, at any time, each non-personal-name word in the pattern is at least a partial match with the corresponding word at the beginning of the queue, the baseline marks the word in the queue corresponding to the personal name in the pattern as a personal name and no further patterns are tested against the queue. This process is repeated until the queue is emptied of words.

Results Table 1 shows the results of our DL-CoTrain system from 5 folds at the 90^{th} iteration with the baseline system.

System		Precision	Recall	F-1
Baseline	Fold 1	64.1%	30.1%	41.0%
	Fold 2	64.2%	29.6%	40.5%
	Fold 3	64.6%	29.5%	40.5%
	Fold 4	65.4%	30.1%	41.2%
	Fold 5	65.1%	29.7%	40.8%
	Average	**64.6%**	**29.8%**	**40.8%**
DL-CoTrain	Fold 1	56.2%	92.9%	70.0%
	Fold 2	55.8%	92.1%	69.5%
	Fold 3	56.3%	92.8%	70.1%
	Fold 4	56.3%	92.4%	70.0%
	Fold 5	55.3%	92.5%	69.2%
	Average	**56.0%**	**92.5%**	**69.8%**

Table 1: The Precision, Recall and F-1 Measure of the baseline system and our DL-CoTrain system.

Table 1 shows that our NER system has an average Precision score of 56.0% and an average Recall score of 92.5%. On the other hand, the baseline system has an average Precision score of 64.6% and an average Recall score of 29.8%. The result shows that our NER system can recover 92.5% of the personal names annotated in the lemmatization. The 56.0% Precision indicates that around 44% of the names found by our NER system are not labeled as a personal name in the lemmatization. However, preliminary investigation into these names by our language experts has suggested that some of the false positives are, in fact names, are arise from the very conservative policy followed by the authors of the lemmata when working with damaged tablets. Our response to this fact can be found in (Liu et al. 2015).

More Experiments and Some Analysis

In order to see the impact of the number of iterations, and the window size of the context to the result, we also ran the following experiments.

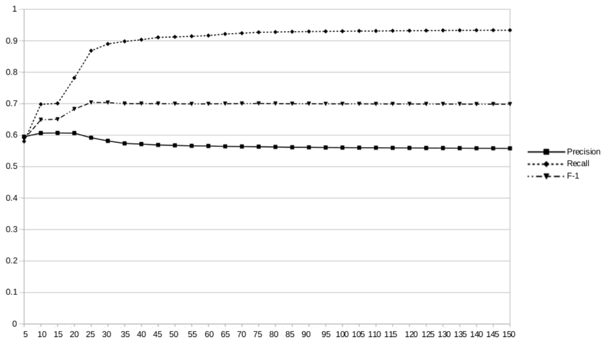


Figure 3: The changes on Precision, Recall and F-1 Measure for 150 iterations.

Iterations The tendency of Precision, Recall and F-1 Measures with iterations is shown in Figure 3. Based on the figure, both Precision and Recall become solid after the 90^{th} iteration. For this reason, we believe that the 90^{th} iteration is a good place to settle upon.

Tri-Gram and 4-Gram As described in Method Section, when we induce contextual rule, we considered about bi-grams case, whereas in the baseline system, we used some tri-gram contextual patterns for matching. We, then ran two more experiments in which the system induced only tri-gram and 4-gram patterns.

System	Precision	Recall	F-1
Bi-Gram (90^{th})	55.8%	92.1%	69.5%
Tri-Gram (90^{th})	77.7%	26.7%	39.7%
4-Gram (90^{th})	77.2%	26.5%	39.5%
Bi-Gram (150^{th})	55.4%	92.1%	69.2%
Tri-Gram (150^{th})	72.6%	31.0%	43.4%
4-Gram (150^{th})	77.2%	28.5%	41.6%

Table 2: Precision, Recall and F-1 Measure for the bi-gram, tri-gram and 4-gram models at the 90^{th} and the 150^{th} iteration.

As shown in Table 2, at the 90^{th} iteration, the tri-gram model and 4-gram model still maintained a high Precision but failed to find as many names as the bi-gram model did. At the 150^{th} iteration, the Precision of the tri-gram model decreased by 5.1% but the Recall increased by 4.3%; the Precision of the 4-gram model maintained and the Recall increased by 2.0%. This result shows that, not unexpected, it takes more iterations (i.e. longer time) to train a tri-gram or 4-gram model. A method that combines these models may improve both Precision and Recall. However, we do not know when it is optimal to switch to tri-gram or 4-gram rule induction.

Conclusions and Future Work

The immediate goal of this research is to develop an automated system to recognize Sumerian personal names from the transliterations of the tablets. The recognition of personal names can greatly facilitate the recognition of dates and events in Sumerian texts. Such information can further help the historians to develop an environment for investigating Sumerian sociology. However, due to the availability of more and more excavated tablets, and the lack of knowledge of this language, extracting such information from Sumerian texts is becoming a serious challenge faced by both the historians and the computer scientists.

In this paper, we adopted the unsupervised DL-CoTrain algorithm for modern English processing to recognize Sumerian personal names. Our experimental results showed that the NER system built over the DL-CoTrain method, with the minimal amount of prior knowledge, performs very well in recovering names from Sumerian texts.

From the previous analysis, we would like to try a method that also takes a larger contextual window into account, to further improve the precision of the system. With the current annotation, we would hope that a supervised learning method can better the system performance. In fact, we have been informed by the historian in our home university that from his perspective, an NER system with a higher recall and a lower precision is considered better than a system with a lower recall, and a higher precision, because he will have more room to do the verification on the result for the former system, especially when the lemmatization was annotated by a more critical and conservative approach. If given the opportunity, we also would like to see how the system works when adopted to a different ancient language domain or to Sumerian texts of different genres.

Acknowledgments

The authors wishes to thank Steven Garfinkle from Department of History of Western Washington University for his assistance and valuable input as a language expert in this work.

References

Brewer, F.; Burkhart, C.; Houng, J.; Luo, L.; Riley, D.; Toner, Brandon. Liu, Y.; and Hearne, J. 2014. A preliminary study into named entity recognition in cuneiform tablets. In *The third Pacific Northwest Regional Natural Language Processing Workshop*, 1–3.

Collins, M., and Singer, Y. 1999. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 100–110.

Foxvog, D. 2014. An introduction to sumerian grammar.

Garfinkle, S. 2012. *Entrepreneurs and Enterprise in Early Mesopotamia: A Study of Three Archives from the Third Dynasty of Ur.* Ithaca, NY USA: Cornell University Studies in Assyriology and Sumerology (CUSAS).

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.

Limet, H. 1960. *L'Anthroponymie sumerienne dans les documents de la 3e dynastie d'Ur.* Paris: Socit d'dition Les Belles Lettres.

Liu, Y.; Burkhart, C.; Hearne, J.; and Luo, L. 2015. Enhancing sumerian lemmatization by unsupervised named-entity recognition. In *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015) May 31-June 5, 2015, Denver, Colorado, United States.* The Association for Computational Linguistics.

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 Cientficas. 19–54.

Sahala, A. 2012. Notation in sumerian transliteration. Technical report, University of Helsinki.

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.

Tinney, S., and Robson, E. 2014. Oracc: The open richly annotated cuneiform corpus.

Weibull, N. 2004. A historical survey of number systems. Technical report, Chalmers University of Technology.

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 Investigationes Cientficas. 207–223.