Enhancing Sumerian Lemmatization by Unsupervised Named-Entity Recognition

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

Abstract

Lemmatization for the Sumerian language, compared to the modern languages, is much more challenging due to that it is a long dead language, highly skilled language experts are extremely scarce and more and more Sumerian texts are coming out. This paper describes how our unsupervised Sumerian named-entity recognition (NER) system helps to improve the lemmatization of the Cuneiform Digital Library Initiative (CDLI), a specialist database of cuneiform texts, from the Ur III period. Experiments show that a promising improvement in personal name annotation in such texts and a substantial reduction in expert annotation effort can be achieved by leveraging our system with minimal seed annotation.

1 Introduction

Because the Sumerian cuneiform writing system is historically the earliest, Sumerian culture is the earliest recorded civilization. The large number of clay tablets that have been recovered from Mesopotamia reveal "an almost obsessive concern for the preservation of daily events of the time: the digging of ditches, the care of livestock, the storage of grain, and so on. Their survival allows insight into the lives of the city dwellers of remote antiquity" (Garfinkle, 2012). Today, most cuneiform texts are held in public institutions, but the texts are widely separated both from each other and often from scholars by great distances and expensive journeys. Current projects like the Digital Library Initiative (CDLI, 2014) and the Database of Neo-Sumerian Texts (BDTNS, 2014) aim to provide scholars immediate access to virtual collections of tens of thousands of texts.

The Ur III period (2112-2004 BCE) is particularly abundant in surviving texts. Because this era was the specialty of our principle informant, an Assyriologist at our home university, we focus on the tablets that are from this era. The vast majority of these tablets record financial transactions, such as records of cattle deliveries, receipt of metals, repayment of loans, and so forth.

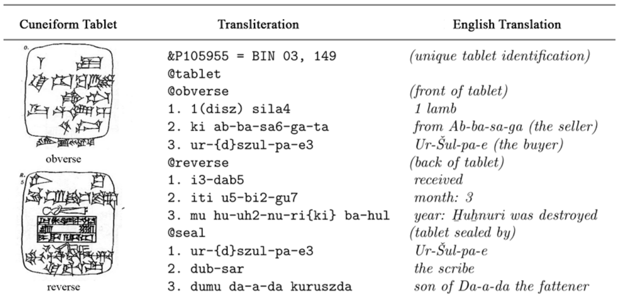
Figure 1 shows a tablet from the CDLI repository. For expository purposes, we arranged the original cuneiform drawings on the left (which are not input to our computations), with its transliteration (a one-to-one transcription of signs in a cuneiform text to computer readable text) in the middle, and the modern English translation on the right. The original CDLI data includes transliterations in ASCII format and inline lemmatization markup. More detail about CDLI data will be introduced in Section 2.

As we can see in Figure 1, in addition to the provider and recipient of transference, tablets consistently enumerate lists of witnesses ("sealed by"). 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 enjoyed professional relations (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.

Motivated by this perspective, we built an unsupervised Sumerian named-entity recognition (NER) system, also to accommodate the facts of 1) Sumerian is a dead language; 2) the corpus is of a size too large for even a community of scholars to master; 3) the tablets come in many cases damaged by time and

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1446–1451,
Denver, Colorado, May 31 – June 5, 2015. ©2015 Association for Computational Linguistics

the circumstances of their recovery which was, in many cases, looting; 4) new tablets are still being uncovered. More detail on our Decision List Co-Train method (Collins and Singer, 1999) can be found in Section 3. In the process of evaluating our NER system, we noticed that a major inconsistency between our result and the lemmata in CDLI lies in the annotation of personal names with missing signs in damaged tablets. For example, "szu-[x]-lum" is not labeled as a name in the lemmata, but our system does so with a high confidence score. As shown this word contains a damaged sign (indicated by "[x]"). Inconsistencies of this kind account for around 50% of the total false positives in our result. With the help of the Sumerologist at our home university, around 40% of such damaged occurrences have been easily verified as personal names. This suggests that the original lemmatization is performed by a more critical and conservative approach. Our work offers a promising automation tool for the annotation task on this corpus by making good recommendations on name candidates to the annotators.


Cuneiform Tablet
obverse
reverse

Transliteration	English Translation
&P105955 = BIN 03, 149	*(unique tablet identification)*
@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)*
1. i3-dab5	*received*
2. iti u5-bi2-gu7	*month: 3*
3. mu hu-uh2-nu-ri{ki} ba-hul	*year: H̲uh̲nuri 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.

2 CDLI and the Annotations

The CDLI is a collaborative project with cuneiform text capturing and processing efforts underway in North America, Europe and the Middle East. It aims to provide an open access to electronic documentation of ancient cuneiform, consisting of texts, images, transliterations and glossaries of 3500 years of human history. Adhering to the open-source policy, any contribution to the collection by providing electronic catalogues, transliterations, or images of cuneiform artifacts is welcomed (CDLIwiki, 2014).

When represented in Roman script in transliterations, the syllable signs that make up a Sumerian word are written joined together with dashes. As there is no concept of upper- or lowercase in cuneiform writing, signs in transliterations typically occur in lowercase. However, signs rendered in uppercase do occur when the phonetic equivalent of the sign is unclear, tentative or fairly new (Sahala, 2012). One important property of Cuneiform is a high degree of homophony (referred to in the literature on Cuneiform as 'polyvalence'). This phenomenon is conventionally handled by numerical subscripts. For example, "du" means "to go", "du$_3$" means "to build" (Tablan et al., 2006).

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).

In the lemmatization information exposed by CDLI that we make use of in our NER task, when the word is a noun or verbal form, the two types of information included in the lemmata are 1) the *citation form*, rendered as the Sumerian stem; 2) the *guide word*, which functions as a disambiguator and is generally the English translation of the stem; otherwise, the lemma contains only the part of speech, as is the case with proper names and numbers. For example, in the following excerpt (CDLI No: P100032), wherein text is presented with interlinear lemmata (English translation: *Egi-zi-mah received 2 oxen from runner.*), we see both types of lemmatization.

```
1. 2(disz) gu4
#lem: n; gud[ox]
2. ki kas4-ta
#lem: ki[place]; kasz[runner]
3. egi-zi-mah i3-dab5
#lem: PN; dab[seize]
```

On line 3, the verbal form i3-dab5, indicating the receipt of an animate object, is lemmatized with the citation form dab, which is the Sumerian root for this form, and guide word "seize", the best English translation of the citation form. On lines 1 and 3, we have a number lemmatized with the part of speech n and the personal name egi-zi-mah with the part of speech PN, respectively. These annotated

PNs are used as gold standard labels to evaluate our NER system.

In the study of the Ur III corpus, the most exhaustive infrastructure and documentation for lemmatization is that provided for “the Open Richly Annotated Cuneiform Corpus (Oracc)” (ORACC, 2014). The lemmatizer for the Oracc system is accessed via an Emacs interface designed to encourage simultaneous transliteration and lemmatization by a human expert. The process begins with the human expert submitting an unlemmatized transliteration in a format called ATF (ASCII Transliteration Format). This format is the standard interchange format for transliteration across many projects dealing in and exchanging Assyriological textual representations (such as CDLI, BDTNS, the Pennsylvania Sumerian Dictionary (PSD, 2006), and Digital Corpus of Cuneiform Lexical Texts (DCCLT, 2014)). Via the Emacs interface, the transliteration is submitted to the linguistic annotatation system, which identifies an existing project-specific glossary based on directives provided by the human expert in the transliteration, and returns a preliminary lemmatization whose completeness and content depends on the referenced project glossary. The transliterator may then modify any automatically-generated lemmata, or, in the case of new words or new senses in which existing words used, manually lemmatize the word to allow the lemmatizer to “harvest” the new lemma and add it to the glossary. Oracc’s lemmatizer also performs normalization and morphological analysis in order to automatically and consistently identify words in the text. The lemmatizer is not designed to “learn” new insights or induce new rules regarding Sumerian morphology on the basis of new lemmata harvested from submissions, but rather serves as a mechanism to consistently apply rules that have been harvested.

Based on our statistics, 53,146 tablets (about 60%) of the CDLI repository are accompanied by the in-line annotations described above. That is the amount of the tablets we used for the NER System.

3 Sumerian Personal Name Recognition

3.1 Related Work

To our knowledge, no previous empirical research exists directly addressing the question of how to recognize named entities from the Sumerian text. Our very preliminary work on this task (Brewer et al., 2014) uses an existent name list to recognize existing names, and applies simple heuristics and a similarity measure to recognize unseen personal names and dates. And at the time, no comprehensive evaluation and analysis could be done due to the unavailability of the language expert.

The investigation most closely related to ours is found in (Jaworski, 2008), which describes a system for processing Sumerian economic documents. Even though we borrowed 3 rules from their work as our seed rules (more details can be found in Section 3.2), and we are dealing with the same language in the same domain, there are a few important differences between our work and theirs. 1) Their goal is to model the content of the text by using an ontology driven method, whereas our goal is to extract named entities from the text by using some statistical method. 2) Their data set is strictly smaller than ours. The corpus used in their work was restricted to ∼12,000 tablets containing transactions involving animals, with the contents of these transactions being extracted via an a priori ontology. Our work is addressed to almost the entire corpus where the lemmatization is available, ∼53,000 tablets. 3) Their work involved no learning but rather the application of pre-defined Finite State Methods for entity recognition.

Supervised named entity recognition has achieved excellent performance (Bikel et al., 2002) (Zhou and Su, 2002) (McNamee and Mayfield, 2002) (MaCallum and Li, 2003) (Oliver et al., 2003). Semi-supervised approaches and unsupervised approaches have also achieved notable success on this task. Although our research also has a fairly large amount of data, unlike the previous unsupervised methods (Etzioni et al., 2005) (Nadeau et al., 2006) (Li et al., 2012), we do not have extremely large external corpora such as Wikipedia to retrieve very precise, but sparse features. Our work adopted the DL-Cotrain method proposed in (Collins and Singer, 1999). However, all their features are at the word sequence level, instead of at the token level. As noted in Section 2, there is no concept of upper- or lowercase in cuneiform writing, features on capitalization are not relevant here. Another important observation is that Sumerian personal names are exclusively comprised of a single word, thus our spelling features are on the token level. In addition, unlike their work where POS and parsing information is used for named entity candidate selection, we do not have the candidate selection component given that no Sumerian POS tagger or parser available. In fact, further complicating factors in determining syntactic features include the lack of standardization in spelling and inconsistent scribal quality.

3.2 Our System

Our NER system has three components: the preprocessing component, the Decision List Co-Train (DL-CoTrain) (Collins and Singer, 1999) component and the post-processing component.

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.

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.

After the above pre-processing step, we applied the DL-CoTrain method by utilizing contextual and spelling rules to create a decision list.

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(ab-ba) → Person" is a spelling rule. By applying the rule, the word "ab-ba-sab-ga-ta" is recognized as a personal name. With the spelling rule "full-string = ur-{d}szul-pa-e3 → Person", the word "ur-{d}szul-pa-e3" is recognized as a personal name.

We use the following three contextual rules (Jaworski, 2008) as the seed rules for the system 1) left_context = giri3 → Person 2) left_context = kiszib3 → Person 3) 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.

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 NER system, we experimentally settled on a ranking criterion that made use of frequency of some feature x, instead of (smoothed) relative frequency as used in (Collins and Singer, 1999), in order to avoid the problem of some context feature occurs once only as the cue of a personal name, and reverting to the relative frequency formula in the case of ties.

The two post-processing rules are applied to eliminate false positives 1) A word that starts with a number should not be a name; 2) A word following the word "iti" (month indicator) should not be a name. The application of these 2 rules improved the performance by 0.5%.

4 Experiments and Evaluation

We used a 5-fold cross-validation model to train and test our NER system. In each fold, we randomly picked 85% of the tablets from the corpus for training and the remaining 15% of the tablets for testing. With the top 20 new rules from each iteration being added to the decision list, the system produces a decision list of over 2000 rules and approximately 17,000 personal names in these Sumerian texts, after 150 iterations.

When the lemmata is used as the gold standard data set in this experiment, the system achieved 91.4% recall and 39.6% precision score on average from the 5-fold cross-validation. The low precision motivated us to take a closer look at the cause of the false positives from our system.

Using fold-2 as an example, the system reported 16,657 personal names, and there are 7,406 annotated names in the lemmata. Among all these 7,406 names, 91.4% has been correctly identified by the system. However, 60.6% of the names reported by our system are not labeled as names in the lemmata. Through error analysis, we found that nearly 50% of these false positive names contain“missing” or “damaged” signs in the transliteration (i.e., annotated as [x] or [u] in the lemmata). They were therefore not annotated at all in the lemmata, even though their linguistic context clearly shows that they are personal names. For example, “`szu-x-lum`” in “`giri3 szu-x-lum`” is a word in the testing data labeled as a name by our system after applying one of the seed rules. However, owing to physical damage to the word in the original tablet (flagged by “`x`” in the lemmata), it is unannotated. As a result, it's reported as a false positive in the evaluation.

Based on this observation, we asked the Sumeriologist at our home university to verify the “false positives” that contain “missing” or “damaged” signs (marked in the lemmata as either “unknown” (part of speech `X`) or “unlemmatizable” (part of speech `u`)), restricting our concern to damaged signs to limit the imposition on his time. It turns out that over 40% of such names should have been labeled as a name in the first place. This elevates the precision to 55.8% from 39.4% without sacrificing the recall, for fold-2 testing data. Similar performance gain is obtained for other folds.

Due to the large number of “false positives” and time constraints, we cannot impose on our Assyriologist informant the task of verifying all of the system reported names for us at the moment. However, the current evaluation result reveals that the systematic lemmatization on CDLI, as discussed in Section 2, follows an extremely conservative approach. We suspect that the reason for this is to avoid labeling damaged personal names as such is to prevent partial or potentially incorrect sign information from being reused by the morphological analyzer in future runs of the lemmatizer. Our result suggests that the existing lemmata has its own limitation and should not be fully relied on for evaluation for our NER task. It also suggests that our NER system can be used for automatic annotation task given that it performs well in recovering names based on the context and spelling features, even with the minimal prior knowledge. More details of the algorithms and result can be found in (Liu et al., 2015).

5 Conclusions and Future Work

We have shown that a DL-CoTrain based name tagger, with only three initial seed rules and unlabeled data, performs well in recovering personal names from Sumerian texts. This work can potentially make the annotation job much less costly, especially when the expert resource is extremely scarce.

Our results show that the existing lemmatization on CDLI corpus was generated by a, perhaps, excessively conservative policy, especially when one or more signs in the name have sustained damage. As a result, we consider that the existing lemmata cannot be fully relied on, especially for damaged names, for our NER evaluation. Our system is able to make good guesses on such damaged occurrences, based on the context and the spelling features. Confirmed by the language expert, such a high-recall, not-so-high-precision system can be particularly useful for the corpus annotators because they can simply focus on and verify the system's recommended names. Furthermore, we would expect that by applying supervised learning or combining with gazetteer-based method, and by extending the current method to recognizing other types of names in the texts, our system can work even better as an automation tool for such an annotation task.

References

Steven Garfinkle. 2012. *Entrepreneurs and Enterprise in Early Mesopotamia: A Study of Three Archives from the Third Dynasty of Ur*, 36–136 Cornell University Studies in Assyriology and Sumerology (CUSAS), Ithaca, NY USA.

Michael Collins and Yora Singer. 1999. *Unsupervised Models for Named Entity Classification*, 100–110. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora

Wojciech Jaworski. 2008. *Contents Modeling of Neo-Sumerian Ur III Economic Text Corpus*, 369–376. Proceedings of the 22nd International Conference on Computational Linguistics

Aleksi Sahala. 2012. *Notation in Sumerian Transliteration*.

Nikolai Weibull. 2004. *A Historical Survey of Number Systems*, 1–13.

Valentin Tablan, Wim Peters, Diana Maynard, and Hamish Cunningham. 2006. *Creating Tools for Morphological Analysis of Sumerian*, 1762–1765. Proceedings of the Fifth International Conference on Language Resources and Evaluation

Magnus Widell. 2008. *The Ur III Metal Loans from Ur*, 207–223. Consejo Superior de Investigationes Cientficas, Madrid.

Database of Neo-Sumerian Texts. 2014. *http://bdtns.filol.csic.es*

Cuneiform Digital Library Initative. 2014. *http://cdli.ucla.edu*

Cuneiform Digital Library Initative wiki. 2014. *http://cdli.ox.ac.uk/wiki/*

Oracc: The Open Richly Annotated Cuneiform Corpus. 2014. *http://oracc.museum.upenn.edu/*

PSD: The Pennsylvania Sumerian Dictionary. 2006. *http://psd.museum.upenn.edu/*

DCCLT - Digital Corpus of Cuneiform Lexical Texts. 2014. *http://oracc.museum.upenn.edu/dcclt/*

Felicity Brewer, Clinton Burkhart, Joe Houng, Liang Luo, Derek Riley, Brandon Toner, Yudong Liu, and James Hearne. 2014. *A Preliminary Study into Named Entity Recognition in Cuneiform Tablets*, 1–3. The third Pacific Northwest Regional Natural Language Processing Workshop

Daniel Foxvog. 2014. *An Introduction to Sumerian Grammar*.

Henri Limet. 1960. *L'Anthroponymie sumerienne dans les documents de la 3e dynastie d'Ur*. Société d'Édition Les Belles Lettres, Paris.

Manuel Molina. 2008. *The Corpus of Neo-Sumerian Tablets: An Overview*, 19–54. Consejo Superior de Investigationes Cientficas, Madrid.

Steve Tinney and Eleanor Robson. 2014. *Oracc: The Open Richly Annotated Cuneiform Corpus*. http://oracc.museum.upenn.edu/doc/about/aboutoracc/index.html

Roger Woodard. 2008. *The Ancient Languages of Mesopotamia, Egypt and Aksum*. Cambridge University Press, Cambridge, UK.

Steve Tinney and Eleanor Robson. 2014. *Oracc: Linguistic Annotation*. http://build.oracc.org/doc/builder/linganno/

Yudong Liu, Clinton Burkhart, James Hearne, and Liang Luo. 2015. *Unsupervised Sumerian Personal Name Recognition*. Proceedings of the Twenty-eighth International Florida Artificial Intelligence Research Society Conference, May 18-20, 2015, Hollywood, Florida, USA. AAAI Press, 2015.

Andrew McCallum and Wei Li. 2003. *Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons*. Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003.

Oliver Bender, Franz Josef Och, and Hermann Ney. 2003. *Maximum Entropy Models for Named Entity Recognition*. Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003.

Paul McNamee and James Mayfield. 2002. *Entity Extraction Without Language-specific Resources*. Proceedings of the 6th Conference on Natural Language Learning - Volume 20.

Guodong Zhou and Jian Su. 2002. *Named Entity Recognition using an HMM-based Chunk Tagger*. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.

Daniel M. Bikel and Richard Schwartz and Ralph M. Weischedel. 1999. *An Algorithm That Learns What's in a Name*. Machine Learning.

Etzioni, Oren and Cafarella, Michael and Downey, Doug and Popescu, Ana-Maria and Shaked, Tal and Soderland, Stephen and Weld, Daniel S and Yates, Alexander. 2005. *Unsupervised named-entity extraction from the web: An experimental study*. Artificial intelligence - Volume 165.

Nadeau, David and Turney, Peter and Matwin, Stan. 2006. *Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity*. Advanced in Artificial Intelligence - Lecture Notes in Computer Science - Volume 14013.

Li, Chenliang, Weng, Jianshu and He, Qi and Yao, Yuxia and Datta, Anwitaman and Sun, Aixin and Lee, Bu-Sun. 2012. *Twiner: named entity recognition in targeted twitter stream*. Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieva.