A Preliminary Study into Named Entity Recognition in Cuneiform Tablets

Felicity Brewer, Clinton Burkhart, Derek Riley, Brandon Toner, Joe Houng, Liang Luo, Yudong Liu and **James W. Hearne**
Computer Science Department
Western Washington University
{brewerf, burkhac, rileyd3, tonerb2, houngj, luol}@students.wwu.edu
{yudong.liu, james.hearne}@wwu.edu

1 Introduction

Cuneiform is a writing system first developed by the ancient Sumerians of Mesopotamia c. 3500-3000 B.C.E., and was used by people throughout the ancient Near East to write several different languages (including Akkadian, Babylonian, and Sumerian). By the middle of the Third Millennium B.C., Cuneiform primarily written on clay tablets recorded daily events, astronomy, and literature. So far, a lot of effort has been put to build online Digital Corpus of Cuneiform Texts (BDTNS, 2014; CDLI, 2014), aiming at providing resources compared to modern dictionaries in the education of scribes and in the study of social development in the period.

Most of the contemporary research on these Cuneiform tablets involves an Assyriologist manually collecting and processing information through static web pages or books where the tablet transliterations are displayed, which is a very time-consuming process. Our goal is to create tools that automate the information extraction process, and facilitate Assyriologists to reconstruct a social network of ancient Sumer during the Ur III period.

We use transliterations from the Cuneiform Digital Library Initiative as source material (CDLI, 2014). The Ur III period ranges from 2047 to 1940 B.C.E., and the kingdom of Sumer spanned, approximately, the southern half of modern-day Iraq. Almost all of the texts are financial records, having a year and month, witness names, and names of other parties participating in the transaction. The most relevant data items are personal, year, and month names. This information indicates social and business relationships between the persons mentioned in the tablet. These relational networks can be expanded by combining information from multiple tablets. We can identify an individual based on their appearance in multiple tablets that are dated to around the same time period. To accomplish this analysis, the two most important data items we extract from the tablets are personal names and dates.

2 Data

From various resources (Garfinkle, 2012) a list of 173 year names, a list of 48 month names, and another list of about 7,500 personal names from Ur III have been collected, but we have no way of verifying their accuracy or completeness.

Generally, in Sumerian, variant spellings of all proper names, including personal names, are very common in the corpus. See more discussion in 2.1. Also, damaged text and transliterational noise is another challenge we have to tackle in this task.

In our current very preliminary implementation, a personal name is first identified by looking up the given list. Approximately 3,000 personal names were found in the current Ur III corpus (58,634 tablets). We then applied a simple heuristic rule where all the occurrences containing the determinative {d} (an indicator of a name having a divine origin) have been collected. This gives us another over 6,000 new names. To handle damage, particularly to the beginning and ends of texts, and minor textual differences due to scribal variety, the Levenshtein distance was used preliminarily due to its simplicity. We use the distance to create a confidence metric which indicates how similar the word is to a specific known entity for recognizing dates.

2.1 Identifying and Distinguishing Personal Names

A personal name, PN, in the text may contain modifiers that indicate that person's role in the transaction, type of name (divine, etc), their official title, or their relationship to another PN. Most occurrences of PNs are completely undecorated by any of these modifiers, but our methods need to account for a modified PN. A PN followed by "-sze3" (a transliterated Sumerian terminative case ending) indicates that the PN brought something

to the transaction and "-ke4" after a PN means that the person received something. These "case endings" may also appear anywhere inside a PN or other word as part of the word itself, instead of as modifiers. The case endings are included in approximately half of all situations they are applicable. Determinatives are symbols that indicate a connected word belongs to a specific category of noun. Usually, the determinative will succeed the word it modifies but sometimes the determinative precedes the word or is contained within the word being modified. Determinatives are very common and are always enclosed in curly braces. For example, the transliteration "ur-{d}nusku" refers to a person named "Ur-Nusku", whose name has a divine origin. Official titles always succeed the PN they refer to. For example, the transliteration "szu-er3-ra **simug**" translates to "Szu-Erra **the smith**". Like titles, a PN may be followed by a word indicating their relationship to some other PN. For example, "ur-nusku **dumu** ka-ka" translates to "Ur-Nusku **son of** Kaka" (Garfinkle, 2012). All together, there is a general pattern to the appearance of personal names: PN [T] [R PN [T]], where PN represents "personal name", T represents "title", and R represents "relationship to". We have encountered only one case where that pattern was broken; the transliterated text matched the pattern R PN T.

The recognition process must also accomodate scribal variations. For example, the name of the city of Urbilum is rendered in the texts in many ways, including, but not limited to: ur-bi-lumki, ur-bi2-lumki, ur-bi2-i3-lumki, ur-bi2-il6-lumki, ur-bi2:i3-lumki, ar-bi2-lumki and ar-bi2-li-lumki.

2.2 Identifying and Ordering Dates

Dates were included on every tablet at the time of its creation, but the specification of dates exhibit enormous syntactic variation. In addition, some of the tablets come to us highly damaged and lack fully legible dates. In addition, each province of Sumer had a unique set of 13 month names. Year names were standardized but each year was named after a significant event that occurred during that year. Before a year was named, it was simply referred to as "the year after LastYearName". There is no consistent connection between the progression of time and the names of the years, complicating the parsing.

The full date was written as a month and a year, usually on the last few lines of a tablet. Dates may also appear inside the tablet when it documents a loan, since the terms of the loan indicate a time period until repayment. All month names are, unless the tablet was damaged, preceded by "iti" and years are preceded by "mu". "Iti" is a unique symbol and does not appear in any other context besides indicating that the following text is a month name or damaged beyond recovering. "mu", however, is followed by a year name about 80% of the time. The other occurrences are followed by damaged text that used to be a year name or "mu" has another meaning in that context. For example, "mu lugal-bi al-pa" translates to "In the name of the king", and "mu" does not relate to a year.

3 Discussion and Conclusion

In this paper, we reported our very preliminary work on Personal names and Dates recognition from the Cuneiform tablets. Since the vast majority of texts from the Ur III corpus are economic in nature and have a well-established, formulaic structure, identifying references to names, months, and years in the corpus is relatively straightforward. More problematic, however, is the resolution of these month and year references to an "objective" calendar, a requirement for chronological ordering, grouping the texts, or constructing a timeline of activity for individuals named in those texts. Another technological challenge is the non-standard representation of personal names. The primary reasons for these complications involve physical damage to the text, poor differentiation between the canonical names of different years, distinct canonical names for the same year, lack of standardization between the local calendar months to which the dates refer, and inconsistent use of personal name modifiers. Furthermore, physical damage to a text may prevent an unambiguous resolution; as a result, a confidence level must be associated with all resolutions.

The current method is unable to fulfill the NE task with a high recall. For example, occurrences such as "dumu ba-zi-ta" as a person are yet to be recognized. The Levenshtein distance algorithm did not perform acceptably as a result of loss of text within a string, especially when the number of signs lost was unknown. We hope that more advanced named entity recognition method will augment our current techniques, resolving some of the technical challenges our naive approach has been unable to thus far.

References

Richard Firth. 2013. *Notes on Year Names of the Early Ur III Period: Szulgi 20-30*. Cuneiform Digital Library Journal, 2013:1.

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

Steven J. Garfinkle. 2012. *Entrepreneurs and Enterprise in Early Mesopotamia: A Study of Three Archives from the Third Dynasty of Ur*. CDL Press, Bethesda, Maryland.

Wojciech Jaworski. 2008. *Contents Modelling of Neo-Sumerian Ur III Economic Text Corpus*. Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008).

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