Toward Automatically Assembling Hittite-Language Cuneiform Tablet Fragments into Larger Texts

Stephen Tyndall
University of Michigan
styndall@umich.edu

Abstract

This paper presents the problem within Hittite and Ancient Near Eastern studies of fragmented and damaged cuneiform texts, and proposes to use well-known text classification metrics, in combination with some facts about the structure of Hittite-language cuneiform texts, to help classify a number of fragments of clay cuneiform-script tablets into more complete texts. In particular, I propose using Sumerian and Akkadian ideogrammatic signs within Hittite texts to improve the performance of Naive Bayes and Maximum Entropy classifiers. The performance in some cases is improved, and in some cases very much not, suggesting that the variable frequency of occurrence of these ideograms in individual fragments makes considerable difference in the ideal choice for a classification method. Further, complexities of the writing system and the digital availability of Hittite texts complicate the problem.

1 Introduction

The Hittite empire, in existence for about 600 years between 1800 and 1200 BCE, left numerous historical, political, and literary documents behind, written in cuneiform in clay tablets. There are a number of common problems that confront Hittite scholars interested in any subdiscipline of Hittitology, be it history, philology, or linguistics. Horst Klengel summarizes the issue most crucial to this paper:

> Some general problems, affecting both philologists and historians, are caused by the Hittite textual tradition itself. First, the bulk of the cuneiform material is fragmentary. The tablets, discovered in various depots in the Hittite capital and in some provincial centers, normally were of a larger size. When the archives were destroyed, the tablets for the most part broke into many pieces. Therefore, the joining of fragments became an important prerequisite for interpretation(Klengel, 2002).

Most Hittite texts are broken, but a number exist in more than one fragmentary copy.

Figure 1 shows a photograph, taken from the University of Meinz *Konkordanz der hethitischen Texte*[1], of a typical Hittite cuneiform fragment.

Complete or partially-complete texts are assembled from collections of fragments based on shape, writing size and style, and sentence similarity. Joins between fragments are not made systematically, but are usually discovered by scholars assembling large numbers of fragments that reference a specific subject, like some joins recently made in Hittite treaty documents in (Beckman, 1997).

Joins are thus fairly rare compared to the frequency of new publishing of fragments. Such joins and the larger texts created therewith are catalogued according to a CTH (*Catalogue des Textes Hittites*[2]) number. Each individual text is composed of one or more cuneiform fragments belonging to one or more copies of a single original work.

[1] available at http://www.hethport.uni-wuerzburg.de/HPM/hethportlinks.html

[2] available at http://www.hethport.uni-wuerzburg.de/CTH/

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 243–247,
Jeju, Republic of Korea, 8-14 July 2012.

Figure 2 shows a published join in hand-copied cuneiform fragments. In this case, the fragments are not contiguous, and only the text on the two fragments was used to make the join.

The task then, for the purposes of this paper, is to connect unknown fragments of Hittite cuneiform tablets with larger texts. I'm viewing this as a text classification task, where larger, CTH-numbered texts are the categories, and small fragments are the bits of text to be assigned to these categories.

2 The Corpus of Hittite

Hittite cuneiform consists of a mix of syllabic writing for Hittite words and logographic writing, typically Sumerian ideograms, standing in for Hittite words. Most words are written out phonologically using syllabic signs, in structure mostly CV and VC, and a few CVC. Some common words are written with logograms from other Ancient Near Eastern languages, e.g. Hittite *antuhša-* 'man' is commonly written with the Sumerian-language logogram transcribed LÚ. Such writings are called Sumerograms or Akkadograms, depending on the language from which the ideogram is taken.

The extant corpus of Hittite consists of more than 30,000 clay tablets and fragments excavated at sites in Turkey, Syria, and Egypt (Hoffner and Melchert, 2008, 2-3). Many of these fragments are assigned to one of the 835 texts catalogued in the CTH.

3 Prior Work

A large number of prior studies on text classification have informed the progress of this study. Categorization of texts into genres is very well studied (Dewdney et al., 2001). Other related text classification studies have looked at classifying text by source, in contexts of speech, as in an attempt to classify some segments of speech into native and non-native speaker categories (Tomokiyo and Jones, 2001), and writing and authorship, as in the famous Federalist Papers study(Mosteller and Wallace, 1984), and context, as in a categorization of a set of articles according to which newspaper they appeared in (Argamon-Engelson et al., 1998).

Measures of similarity among sections of a single document bear a closer relation to this project than the works above. Previous studies have examined in-

Figure 1: Photograph of a Hittite Tablet Fragment

Figure 2: Published Fragment Join

ternal document similarity, using some vector-based metrics to judge whether documents maintain the same subject throughout (Nicholson, 2009).

Very little computational work on cuneiform languages or texts exists. The most notable example is a study that examined grapheme distribution as a way to understand Hurrian substratal interference in the orthography of Akkadian-language cuneiform texts written in the Hurrian-speaking town of Nuzi (Smith, 2007). Smith's work, though using different classifying methods and and an enormously different corpus on a language with different characteristics, is the most similar to this study, since both are attempts to classify cuneiform fragments into categories - in Smith's case, into Hurrian-influenced Nuzi Akkadian and non-Nuzi standard Akkadian.

4 The Project Corpus

For this project, I use a corpus of neo-Hittite fragment transcriptions available from H. Craig Melchert (Melchert,). The corpus is one large text file, divided into CTH numbered sections, which themselves are divided into fragments labeled by their publication numbers - mostly KUB, which stands for *Keilschrifturkunden aus Boghazköi* or KBo, *Keilschrifttexte aus Boghazköi*, the two major publications for Hittite text fragments.

I restricted the fragments used in this project to fragments belonging to texts known to exist in at least two copies, a choice that produces a larger number of fragments per text without requiring a judgment about what number of fragments in a text constitutes "fragmented enough" for a legitimate test of this task. This leaves 36 total CTH-numbered texts, consisting of 389 total fragments.

The fragments themselves are included as plain text, with restorations by the transcribers left intact and set off by brackets, in the manner typical of cuneiform transcription. In transcription, signs with phonemic value are written in lower case characters, while ideograms are represented in all caps. Sign boundaries are represented by a hyphen, indicating the next sign is part of the current word, by an equals sign, indicating the next sign is a clitic, or a space, indicating that the next sign is part of a new word.

```
{KUB XXXI 25; DS 29}
x
[              ]A-NA KUR URUHa[t-ti?
[            i]s-tar-ni=sum-m[i
[             ]x nu=kn ki-x[
[             ] KUR URUMi-iz-ri=y[a
[is-tar-ni]=sum-mi e-es-du [

[    ] nu=kn A-NA KUR URUMi-iz-ri[
[A-NA EGI]R UDmi is-tar-ni=su[m-mi
```

This fragment, KUB XXI25, is very small and broken on both sides. The areas between brackets are sections of the text broken off or effaced by erosion of tablet surface material. Any text present between brackets has been inferred from context and transcriber experience with usual phrasing in Hittite. In the last line, the sign *EGIR*, a Sumerian ideogram, which is split by a bracket, was partially effaced but still recognizable to the transcriber, and so is split by a bracket.

5 Methods

For this project, I used both Naive Bayes and Maximum Entropy classifiers as implemented by the MAchine Learning for LanguagE Toolkit, MALLET(McCallum, 2002).

Two copies of the corpus were prepared. In one, anything in brackets or partially remaining after brackets was removed, leaving only characters actually preserved on the fragment. This copy is called *Plain Cuneiform* in the results section. The other has all bracket characters removed, leaving all actual characters and all characters suggested by the transcribers. This corpus is called *Brackets Removed* in the results section. By removing the brackets but leaving the suggested characters, I hoped to use the transcribers' intuitions about Hittite texts to further improve the performance of both classifiers.

The corpora were tokenized in two ways:

1. The tokens were defined only by spaces, capturing all words in the corpus.

2. The tokens were defined as a series of capital letters and punctuation marks, capturing only the Sumerian and Akkadian ideograms in the text, i.e. the very common Sumerian ideogram *DINGER.MEŠ*, 'the gods'.

The training and tests were all performed using MALLET's standard algorithms, cross-validated,

Table 1: Results for Plain Corpus

Tokenization	Naive Bayes	Max Ent
All Tokens	.55	.61
Ideograms Only	.44	.51

Table 2: Results for Tests on Corpus with Brackets Removed

Tokenization	Naive Bayes	Max Ent
All Tokens	.64	.67
Ideograms Only	.49	.54

splitting the data randomly into ten parts, and using 9 parts of the data as a training set and 1 part of the data as a test set. This means that each set was tested ten times, with all of the data eventually being used as part of the testing phase.

6 Results and Discussion

Accuracy values from the classifiers using the Plain corpus, and from the corpus with the Brackets Removed, are presented in Tables 1 and 2, respectively. The measures are raw accuracy, the fraction of the test fragments that the methods categorized correctly.

The results for the Plain Corpus show that the Naive Bayes classifier was 55% accurate with all tokens, and 44% accurate with ideograms alone. The Maximum Entropy classifier was 61% accurate with all tokens, and 51% accurate with ideograms only.

Both classifiers performed better with the Brackets Removed corpus. The Naive Bayes classifier was accurate 64% of the time with all tokens and 49% of the time with ideograms only. The Maximum Entropy classifier was 67% accurate with all tokens, and 54% accurate with ideograms only.

The predicted increase in accuracy using ideograms was not upheld by the above tests. It may be the case that Sumerograms and Akkadograms are insufficiently frequent, particularly in smaller fragments, to allow for correct categorization. Some early tests suggested occasional excellent results for this tokenization scheme, including a single random 90-10 training/test run that showed a test accuracy of .86, much higher than any larger cross-validated test included above. This suggests, perhaps unsurprisingly, that the accuracy of classification using Sumerograms and Akkadograms is heavily dependent on the structure of the fragments in question.

Maximum Entropy classification proved to be slightly better, in every instance, than Naive Bayes classification, a fact that will prove useful in future tests and applications.

The fact that removing the brackets and including the transcribers' additions improved the performance of all classifiers will likewise prove useful, since transcriptions of fragments are typically published with such bracketed additions. It also seems to demonstrate the quality of these additions made by transcribers.

Overall, these tests suggest that in general, the 'use-everything' approach is better for accurate classification of Hittite tablet fragments with larger CTH texts. However, in some cases, when the fragments in question have a large number of Sumerograms and Akkadograms, using them exclusively may be the right choice.

7 Implications and Further Work

In the future, I hope to continue with a number of other approaches to this problem, including lemmatizing the various Hittite noun and verb paradigms. Additionally, viewing the problem in other ways, e.g. regarding tablet fragments as elements for connection by clustering algorithms, might work well.

Given the large number of small fragments now coming to light, this method could speed the process of text assembly considerably. A new set of archives, recently discovered in the Hittite city of Šapinuwa, are only now beginning to see publication. This site contains more than 3000 new Hittite tablet fragments, with excavations ongoing(Süel, 2002). The jumbled nature of the dig site means that the process of assembling new texts from this site will be one of the major tasks in for Hittite scholars in the near future. This attempt at speeding the task is only the beginning of what I hope will be a considerable body of work to help build more complete texts, and therefore more complete literatures and histories, of not only Hittite, but other cuneiform languages like Akkadian and Sumerian, some of the world's earliest written languages.

References

S. Argamon-Engelson, M. Koppel, and G. Avneri. 1998. Style-based text categorization: What newspaper am i reading. In *Proc. of the AAAI Workshop on Text Categorization*, pages 1–4.

G. Beckman. 1997. New Joins to Hittite Treaties. *Zeitschrift für Assyriologie und Vorderasiatische Archäologie*, 87(1):96–100.

N. Dewdney, C. VanEss-Dykema, and R. MacMillan. 2001. The form is the substance: Classification of genres in text. In *Proceedings of the workshop on Human Language Technology and Knowledge Management-Volume 2001*, pages 1–8. Association for Computational Linguistics.

H.A. Hoffner and H.C. Melchert. 2008. *A grammar of the Hittite language*. Eisenbrauns.

Horst Klengel. 2002. Problems in hittite history, solved and unsolved. In Simrit Dhesi K. Aslihan Yener, Harry A. Hoffner Jr., editor, *Recent developments in Hittite archaeology and history: papers in memory of Hans G. Güterbock*, pages 101–109. Eisenbrauns.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

H. Craig Melchert. Anatolian databases. `http://www.linguistics.ucla.edu/people/Melchert/webpage/AnatolianDatabases.htm`.

F. Mosteller and D.L. Wallace. 1984. Applied bayesian and classical inference: The case of the federalist papers.

C. Nicholson. 2009. Judging whether a document changes in subject. In *Southeastcon, 2009. SOUTHEASTCON'09. IEEE*, pages 189–194. IEEE.

S.P. Smith. 2007. *Hurrian Orthographic Interference in Nuzi Akkadian: A Computational Comparative Graphemic Analysis*. Ph.D. thesis, Harvard University Cambridge, Massachusetts.

A. Süel. 2002. Ortaköy-sapinuwa. In Simrit Dhesi K. Aslihan Yener, Harry A. Hoffner Jr., editor, *Recent developments in Hittite archaeology and history: papers in memory of Hans G. Güterbock*, pages 157–165. Eisenbrauns.

L.M. Tomokiyo and R. Jones. 2001. You're not from 'round here, are you?: naive bayes detection of non-native utterance text. In *Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, pages 1–8. Association for Computational Linguistics.