BabyLemmatizer : A Lemmatizer and POS-tagger for Akkadian

Sahala, Aleksi

2022

Sahala , A , Alstola , T , Valk , J & Linden , K 2022 , BabyLemmatizer : A Lemmatizer and POS-tagger for Akkadian . in T Erjavec & M Eskevich (eds) , CLARIN Annual Conference Proceedings, 2022 . CLARIN Annual Conference Proceedings , CLARIN ERIC , Utrecht , pp. 14–18 , CLARIN Annual Conference , Prague , Czech Republic , 10/10/2022 .

http://hdl.handle.net/10138/348412

unspecified
publishedVersion

BabyLemmatizer: A Lemmatizer and POS-tagger for Akkadian

Aleksi Sahala
University of Helsinki, Finland
aleksi.sahala@helsinki.fi

Tero Alstola
University of Helsinki, Finland
tero.alstola@helsinki.fi

Jonathan Valk
University of Helsinki, Finland
jonathan.valk@helsinki.fi

Krister Lindén
University of Helsinki, Finland
krister.linden@helsinki.fi

Abstract

We present a hybrid lemmatizer and POS-tagger for Akkadian, the language of the ancient Assyrians and Babylonians, documented from 2350 BCE to 100 CE. In our approach the text is first POS-tagged and lemmatized with TurkuNLP trained with human-verified labels, and then post-corrected with dictionary-based methods to improve the lemmatization quality. The post-correction also assigns labels with confidence scores to flag the most suspicious lemmatizations for manual validation. We demonstrate that the presented tool achieves a Lemma+POS labeling accuracy of 94%, and a lemmatization accuracy of 95% in a held-out test set.

1 Introduction

Application of computational methods to historical text corpora provides interesting opportunities for studying large-scale phenomena that are difficult to perceive through close reading of texts. This often requires careful normalization of the language, because in many past societies spelling conventions were not fully standardized, and the corpora can contain documents written in several synchronic and diachronic variants of the language. The languages can also be morphologically complex, which further complicates even such fundamental tasks as searching for all attestations of a certain word in the corpus.

One way to normalize historical languages is lemmatization, which labels words with their dictionary forms regardless of their morphology and spelling. In this paper, we present a lemmatizer for Akkadian, an extinct language that was widely used in ancient Mesopotamia.

The motivation for this tool emerges from close co-operation between the FIN-CLARIN coordinated Language Bank of Finland and the Centre of Excellence in Near Eastern Empires, a University of Helsinki-based research project focusing on the study of the Near East in the first millennium BCE. As part of this co-operation, the Language Bank of Finland collects corpora of ancient Mesopotamian texts written in the Akkadian language in the Korp concordance service.[1] Korp offers several useful functionalities for historians from flexible search options to generating statistics from text metadata.(Borin et al., 2012)

At present, Korp hosts a version of the Open Richly Annotated Cuneiform Corpus (Oracc),[2] which comes with human-verified lemmatization. The next Akkadian corpus to be included in Korp is Achemenet,[3] which has not been lemmatized. The only Akkadian lemmatizer currently available (Tinney, 2019) requires extensive human supervision. To minimize the need for human intervention, our aim is to lemmatize the Achemenet corpus by first training the TurkuNLP's lemmatizer using the available Oracc data, and then applying simple dictionary-based post-correction scripts.

2 The Akkadian Language

Akkadian is an extinct East Semitic language documented in hundreds of thousands of cuneiform tablets excavated across the Near East. The earliest written exemplars of Akkadian date back to the Sargonic Pe-

[1] http://korp.csc.fi

[2] http://oracc.org/

[3] http://www.achemenet.com/

riod (2350-2170 BCE), after which the language is mostly documented in its two main dialects: Assyrian (1950-600 BCE) and Babylonian (2100 BCE - 100 CE) (von Soden, 1995).

Like other Semitic languages, Akkadian morphology employs nonconcatenative root-pattern morphotactics in stem formation and concatenative morphotactics in the attachment of various grammatical affixes to the stems. For example, the verbal form *ludlul* ”let me praise (it)!” consists of the first person singular precative suffix {lu} attached to the preterite stem {dlul}, which is formed from the root *dll* of the verb *dalālu* ”to praise”. Although the morpheme boundaries are transparent in this example, various morphophonological processes often obscure the underlying structure of the word, complicating recognition of the root radicals (von Soden, 1995).

Another layer of complexity emerges from the cuneiform script that developed to represent the linguistically unrelated Sumerian language before being adopted to represent Akkadian in the 24th century BCE. Although the Akkadian language was generally written syllabically, scribes sometimes favoured the use of Sumerian logograms, especially in certain genres of text. The Akkadian verbal form *iddin* ”(s)he gave it”, can for instance, be spelled syllabically as *id-din, i-din, id-di-in* or *i-di-in*, but logographic or logo-syllabic spellings like SUM and SUM-*in* are also attested.

In Akkadian transliteration, logograms are represented in capital letters and named after their base reading values in Sumerian rather than Akkadian. For this reason, the character level relationship between the graphemic and phonemic forms of logographic spellings is typically suppletive. Many logograms are also ambiguous and can have different readings in different contexts. For example, the Sumerian logogram IGI (depicting an eye) can indicate any form of the words *īnu* ”eye”, *pānu* ”front”, *mahru* ”before” and *amāru* ”to see”.

2.1 Digital Resources

For an extinct language, Akkadian is fairly well resourced, and texts comprising about 3-4 million tokens in total have been digitized.[4] Some larger text corpora are Oracc (the Open Richly Annotated Cuneiform Corpus) with 19,000 Akkadian texts, Achemenet with 3,000 texts and Archibab with 10,000 texts. A complete survey of Akkadian digital resources is given in Charpin (2014).

3 Previous Work

Due to the previously discussed complexity of the Akkadian morphology and script, lemmatization is considered a mandatory step into making any digital corpus of Akkadian searchable or suitable for computational analysis (Maiocchi, 2019). To date, however, only Oracc provides extensive lemmatization for Akkadian texts, totalling about 1.5 million lemmatized words. Oracc is lemmatized using a dictionary-based tool known as L2 (Tinney, 2019), which populates new texts with lemmata and POS-tags based on a labelled glossary extracted from previously lemmatized texts. Texts are then checked manually word-by-word, filling in lemmata for out-of-vocabulary words and resolving possible ambiguities.

4 Description of BabyLemmatizer

BabyLemmatizer combines the use of neural networks and dictionary-based lemmatization. The backbone of our tool is Turku Neural Parser Pipeline (TurkuNLP) (Kanerva et al., 2018), a state-of-the-art neural lemmatizer and POS-tagger, for which we train a model using Oracc data. In the lemmatization process, we first provide the text with POS-tagging and raw lemmatization with TurkuNLP and then apply post-corrections to the result to improve lemmatization accuracy. Our post-correction involves three steps:

The first step overrides all predictions for in-vocabulary words. This minimizes the effect of mislearned character level relationships between spellings and their lemmata. We calculate the degree of ambiguity for all lemmatizations in the training data and create a *master glossary* of word forms that have a low degree of ambiguity. We then override all lemmatizations of in-vocabulary words using this data. The degree of ambiguity for a word form is considered to be low, if any lemma+POS label constitutes over N percent of all the labels assigned to it in the training data. Based on our experiments, an N-value

[4] This is our crude estimate based on the number of texts listed in various text corpora.

of 60% seems to produce consistently good results. We leave highly ambiguous lemmata as they are. The second step aims to assign correct lemmata to ambiguous word forms, especially logograms. Here we calculate co-occurrence probabilities for lemmata and their adjacent POS-tags in the training data, and then assign the most likely lemmata for all word forms in the text. We rely on POS-tags instead of surrounding lemmata due to the very reliable POS-tagging accuracy of TurkuNLP. This step allows us to reconfirm that our close-to-unambiguous lemmata are likely correct, and that the ambiguous word forms are lemmatized with the most likely option. The minimum probability threshold is adjustable, but in our experiments we always accept the most likely lemma in the given context.

Finally, we apply various other post-corrections to the data, such as removing the lemmatization from numbers and words that occur in badly damaged sections of the tablet (unreadable signs are indicated in transliteration with x, as in *x-x-in-nu*, which makes them easy to find). This is done to make the lemmatizations more consistent with Oracc conventions and to prevent TurkuNLP from attempting to predict reconstructions that are beyond human comprehension. We also heuristically detect some obvious lemmatization errors, such as verbs that show impossible or very unlikely dictionary form patterns. Nonetheless, these can only be flagged, but not fixed automatically.

4.1 Confidence scoring

Post-processing also assigns lemmatizations with confidence scoring that helps Assyriologists identify the most likely incorrect lemmata. The lowest scores of 0 and 1 are assigned to OOV words containing logograms and to syllabic spellings. The score of 2 is assigned to highly ambiguous in-vocabulary words in previously unseen POS contexts. The second highest score of 3 is assigned to in-vocabulary words that show low ambiguity, and the highest score of 4 to lemmata that exist in previously seen POS contexts.

5 Evaluation

For evaluation, we train ten models for the first millennium BCE Babylonian texts from Oracc comprising ca. 500,000 words in total. We use a 90/10/10 train/dev/test split and estimate the model's accuracy against two baseline models by using 10-fold cross-validation.[5] Our first **baseline** model is a dictionary-based lemmatizer and POS-tagger that labels the word forms in our test set with their most common lemmata and POS-tags seen in the training data. To measure the effect of our post-corrections, we use **TurkuNLP** without any post-correction scripts as the second baseline model. The results are presented in Table 1.

Model	**Lemma**	**POS**	**Lemma+POS**
Baseline	84.42 ±0.33	88.83 ±0.31	82.71 ±0.34
TurkuNLP	86.19 ±1.32	**97.32** ±0.10	85.31 ±1.31
BabyLemmatizer	**94.94** ±0.17	**97.32** ±0.10	**94.03** ±0.35

Table 1: Average accuracy (%) based on 10-fold cross-validation

In Table 2, we measure lemmatization accuracies in different confidence classes, as well as the proportion of lemmata that are assigned to each confidence class in our evaluation setting.

Confidence score	**0**	**1**	**2**	**3**	**4**
Accuracy	30.66%	56.71%	69.57%	96.25%	98.40%
Lemma-%	0.86%	3.87%	0.49%	52.10%	42.67%

Table 2: Confidence score distribution.

[5] In this experiment we use the default network architectures for training TurkuNLP's lemmatizer and tagger

5.1 Manual Evaluation

To test our lemmatizer, we apply it to a sub-corpus of Achemenet comprising 107,778 words. This is an Akkadian corpus with a different genre and time period distribution than our previous test sets. We use a model trained with the same Oracc data and train/dev/test split as in our evaluation setting described above, with an added glossary of Akkadian personal names from Prosobab (Waerzeggers and Groß et al., 2019). We then generate glossaries of the most common words that were assigned with the two lowest confidence classes and manually correct lemmata and POS-tags for word forms in the glossary file that have a frequency of >3 (for class 0) and >5 (for class 1) in the data. There were 315 unique corrected word forms, comprising 3.87% of the unique word forms covering 4.77% (5,037) of the 107,778 words in the sub-corpus.

To measure the accuracy of the lemmatizer and the effect of our manual corrections, we randomly select texts from our lemmatization results amounting to ca. 1,000 tokens for manual evaluation. We first evaluate the initial lemmatization without any manual corrections to the glossaries as a baseline. Then we apply our corrections to the lemmatization results in two ways: first, as a part of our *master glossary* of unambiguous lemmata (used in step 1 of post-processing), and second, by adding our manual corrections to the training data for TurkuNLP to see how much the system can learn from the corrections. The training data is added by first lemmatizing the text with a corrected master glossary and then replacing all words with the lowest two confidence scores with underscores to prevent the neural network from learning likely erroneous lemmata. Results are shown in Table 3.

	Lemma	**POS**	**Lemma+POS**
Baseline	93.0%	94.6%	90.2%
Glossary Override	96.2%	96.0%	93.8%
Retrained NN	**96.6%**	**96.1%**	**94.5%**

Table 3: Improvement in accuracy after corrections.

As can be seen from Table 3, our Lemma+POS labeling accuracy improves 4.3% when manually correcting only 3.87% of the unique word forms. The final results can be considered satisfactory for our current needs, which is to make the corpus searchable in Korp.

6 Conclusions

We presented a hybrid lemmatizer and POS-tagger for Akkadian, and demonstrated an increase of ca. 10% in Lemma+POS labeling accuracy compared with our baseline models. We also tested the lemmatizer on a previously unlemmatized Akkadian corpus with a different chronological and genre distribution than our training data. This test demonstrated that the system can reach a Lemma+POS labeling accuracy close to 95% after minor manual corrections.

References

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, pages 474–478, Istanbul. ELRA.

Dominique Charpin. 2014. Ressources assyriologiques sur internet. In *Bibliotheca Orientalis, 71(3-4).*, October.

Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 133–142, Brussels, Belgium, October. Association for Computational Linguistics.

Massimo Maiocchi. 2019. Thoughts on ancient textual sources in their current digital embodiments. In S. Valentini and G. Guarducci, editors, *Between Syria and the Highlands: Studies in Honor of Giorgio Buccellati and Marilyn Kelly-Buccellati*, pages 262–268. CAMNES.

Steve Tinney. 2019. *L2: How it Works, http://oracc.org/doc/help/lemmatising/howl2works.*

Wolfram von Soden. 1995. *Grundriss der akkadischen Grammatik (3rd edition)*. Pontifical Biblical Institute, Rome.

Caroline Waerzeggers and Melanie Groß et al. 2019. *Prosobab: Prosopography of Babylonia (c. 620-330 BCE), https://prosobab.leidenuniv.nl*.