ANEE

CENTRE OF EXCELLENCE IN ANCIENT NEAR EASTERN EMPIRES
UNIVERSITY OF HELSINKI www.helsinki.fi/anee

BABYLEMMATIZER

Aleksi Sahala

(Co-authors: Tero Alstola, Jonathan Valk, Krister Lindén)

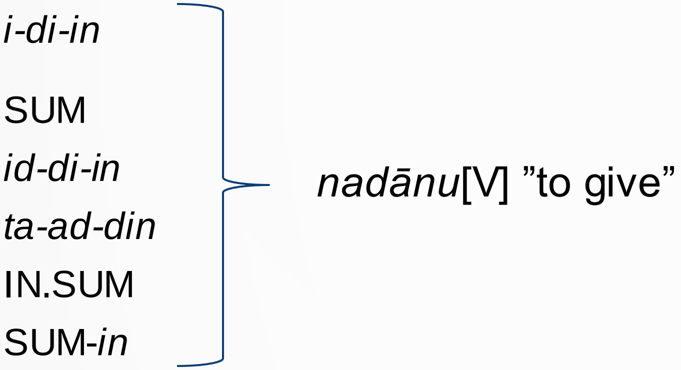
University of Helsinki

AKKADIAN LANGUAGE

- East-Semitic language
- Best known as the language of the Old Akkadian Empire, Babylonia and Assyria
- Documented ca. 2350 BCE – 100 CE
- Important works: The Epic of Gilgameš, Law code of Hammurabi
- Important resources: Open Richly Annotated Cuneiform Corpus (**Oracc**)
- Relevant data set to this publication: Achemenet, especially Neo-Babylonian administrative and legal documents from the late first millennium BCE

WHY LEMMATIZE?

- Effective way to normalize variation in form and spelling
  - Enables searching, data analysis etc.



i-di-in

SUM

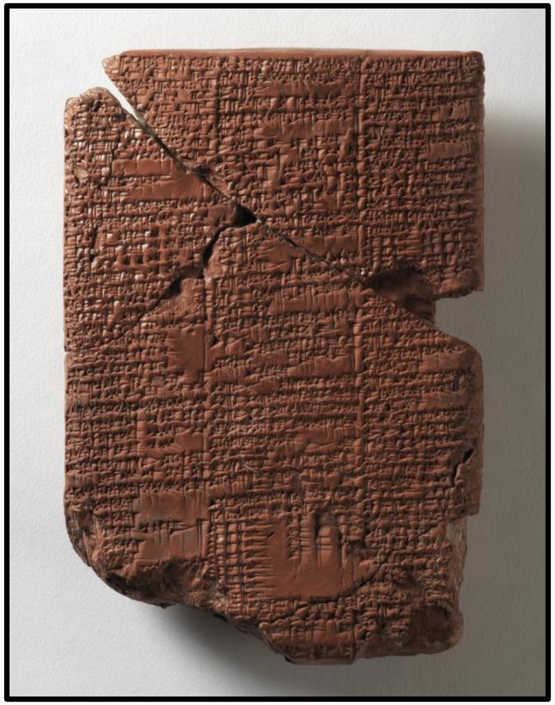
id-di-in

ta-ad-din

IN.SUM

SUM-*in*

nadānu[V] ”to give”



(Image source: Metropolitan Museum of Art)

HOW BABYLEMMATIZER WORKS? INITIAL STEP

- Neural networks for POS-tagging and initial lemmatization
  - TurkuNLP's Lemmatizer (Kanerva et al. 2018)
  - TurkuNLP's POS-tagger (Dozat et al. 2017)

1. POS-tag the input text (acc. ~ 97%)
2. Give raw predictive lemmatization for the text (acc. ~85%)

HOW BABYLEMMATIZER WORKS?
OVERRIDING WITH DICTIONARY

- Dictionary-based post-correction to re-lemmatize all in-vocab words as follows:

1. Calculate probabilities for all lemmata for each wordform in the training data
2. If probability of any lemma is >60%, consider it lowly ambiguous
3. Replace predicted lemma with this

HOW BABYLEMMATIZER WORKS?
DISAMBIGUATION STEP

- Disambiguation of ambiguous wordforms
- Relies on TurkuNLP's high POS-tagging accuracy

1. Calculate co-occurrence probabilities for all Lemma[POS] with their preceding and following POS-tags (for the given wordform)
2. Use this context information to re-lemmatize all ambiguous lemmata.
3. Especially useful for logograms: e.g. if IGI was always *pānu* before, now it may become *amāru*, *īnu*, *šību* etc.

HOW BABYLEMMATIZER WORKS? CLEANING

- Flag impossible predictions: *gītu*[V]
- Remove lemmatizations for too broken words
  - x-x-x-tu aššatu[N] → x-x-x-tu _[u]
- Remove lemmatization of numbers
  - 1 išten[NU] → 1 _[n]

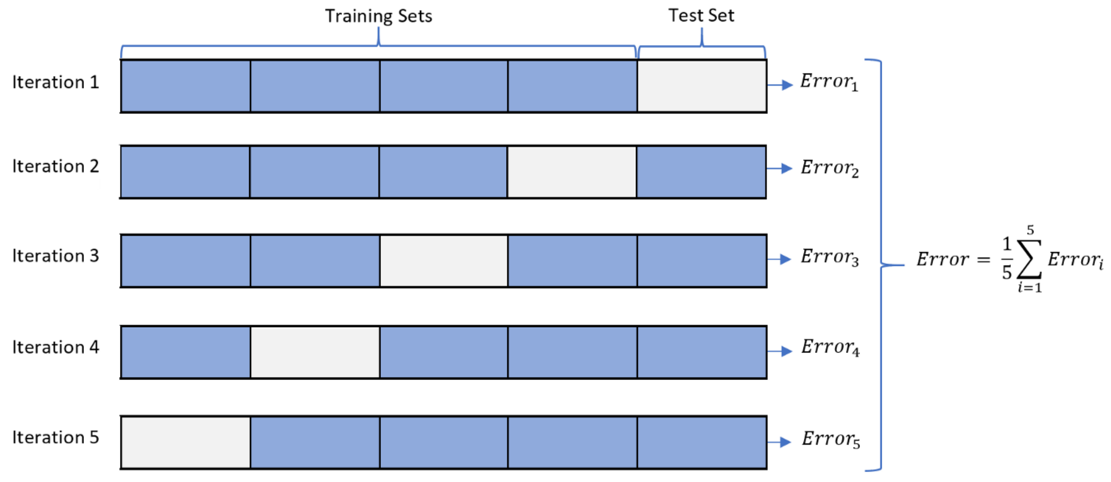
CONFIDENCE SCORING

- Help finding most likely incorrect lemmatizations

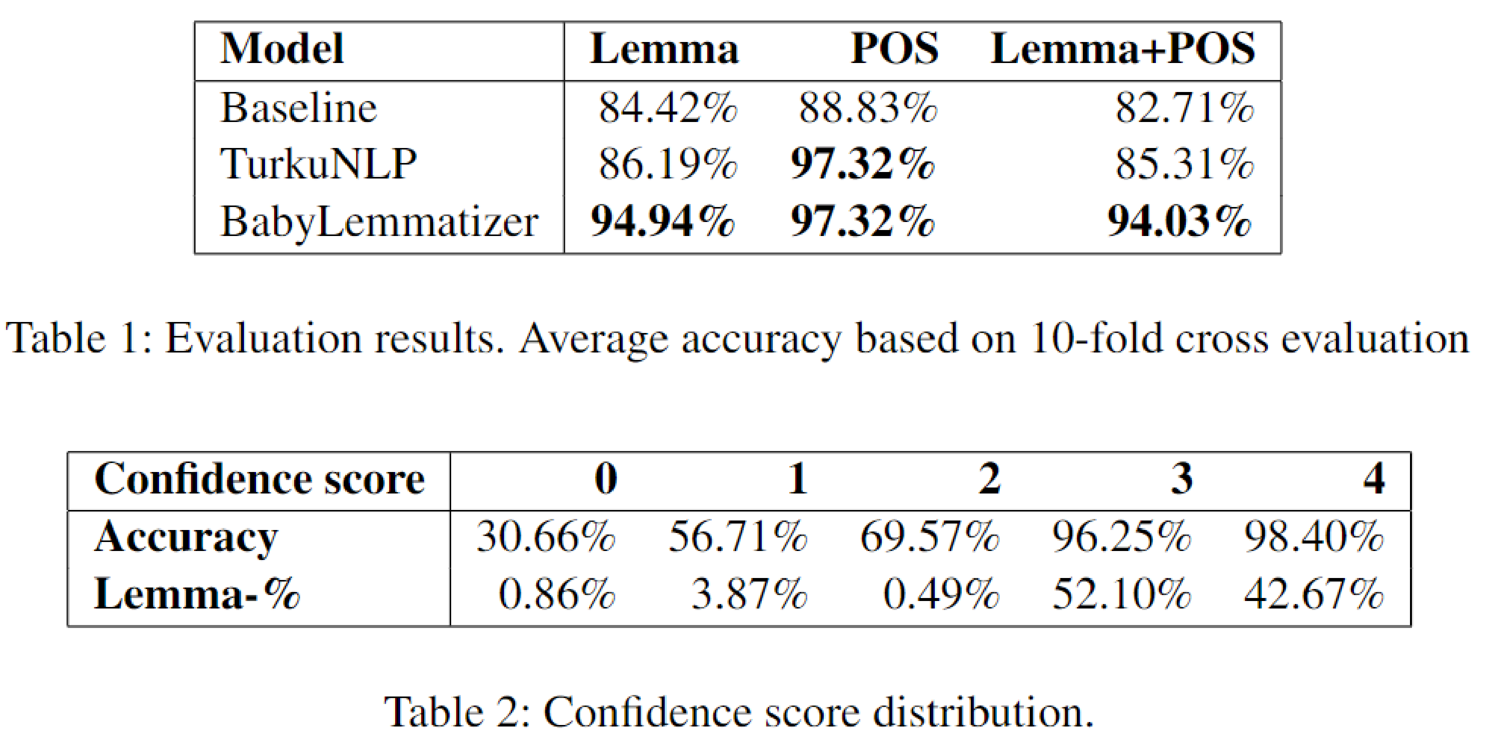
0	out-of-vocab logograms
1	out-of-vocab syllabic spellings
2	highly ambigous unresolved in-vocab words
3	low ambiguity in-vocab words
4	in-vocab words in known POS contexts

EVALUATION SETTING

- Train BabyLemmatizer with 500,000 Akkadian words (first millennium) from Oracc.
- Use 80/10/10 train/dev/test split
- Use 10-fold cross validation



RESULTS



Model	Lemma	POS	Lemma+POS
Baseline	84.42%	88.83%	82.71%
TurkuNLP	86.19%	**97.32%**	85.31%
BabyLemmatizer	**94.94%**	**97.32%**	**94.03%**

Table 1: Evaluation results. Average accuracy based on 10-fold cross evaluation

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.

TEST CASE: STRASSMEIER CORPUS OF ACHEMENET

- Neo-Babylonian legal documents: out-of-domain to our training data
- Manual validation of ca. 1000 lemmata
- Measure lemma+POS accuracy
- Initial accuracy: **90.2%**
- After fixing words with frequency of >5 belonging to confidence classes 0 and 1 and retraining the model: **94.5%**

FUTURE

- Using BabyFST to confirm OOV lemmata and provide morphological analyses
- Use BabyLemmatizer to disambiguate BabyFST's morphological analyses
- Aim to find more sophisticated disambiguation

Thank you!