Deep Past Challenge - Translate Akkadian to English

Overview

The Deep Past Challenge asks a bold question: Can AI decode 4,000-year-old business records?

In this competition, you will help decode the everyday business records of ancient Assyrian merchants. Using data from 8,000 cuneiform texts, your goal is to build a translation system for Old Assyrian. Thousands more like them lie unread in museum drawers around the world. Your work can help bring their voices back into the story of humanity.

Description

Four thousand years ago, Assyrian merchants left behind one of the world's richest archives of everyday and commercial life. Tens of thousands of clay tablets record debts settled, caravans dispatched, and discuss day-to-day family matters. Today, half of these tablets remain silent, not because they're damaged, but because so few people can read the language pressed into their clay. Many have sat untranslated in museum drawers for more than a century.

The Deep Past Challenge turns this ancient mystery into a modern machine-learning problem by inviting competitors to help unlock the largest untranslated archive of the ancient world. We invite you to build translation models for Old Assyrian cuneiform tablets: Bronze Age texts that have sat unread in museum collections for over a century. Old Assyrian—the dialect used on these tablets—is an early form of Akkadian, the oldest documented Semitic language.

Nearly twenty-three thousand tablets survive documenting the Old Assyrian trading networks that connected Mesopotamia to Anatolia. Only half have been translated, and less than a dozen scholars in the world are specialized to read the rest.

These aren't the polished classics of Greece and Rome, curated and copied by scribes who chose whose voices survived. These are unfiltered, straight from the people who wrote them: letters, invoices and contracts written on clay by ancient merchants and their families. They're the Instagram stories of the Bronze Age: mundane, immediate, and breathtakingly real.

Your task is to build neural machine-translation models that convert transliterated Akkadian into English. The challenge: Akkadian is a low-resource, morphologically complex language where a single word can encode what takes multiple words in English. Standard architectures built for modern, data-rich languages fail here. Crack this problem and you'll give voice to 10,000+ untranslated tablets. And you'll do more than revive the past: you'll help pioneer a blueprint for translating the thousands of endangered and overlooked languages—ancient and modern—that the AI age has yet to reach.

Evaluation

linkkeyboard_arrow_up

Submissions are evaluated by the Geometric Mean of the BLEU and the chrF++ scores, with each score's sufficient statistics being aggregated across the entire corpus (that is, each score is a micro-average).

You may refer to the SacreBLEU library for implementation details. A notebook implementing the metric on Kaggle may be found here: Geometric Mean of BLEU and chrF++.

Submission File

For each id in the test set, you must predict an English translation of the associated Akkadian transliteration. Each translation should comprise a single sentence. The file should contain a header and have the following format:

```
id,translation
0,Thus Kanesh, say to the -payers, our messenger, every single colony, and the...
1,In the letter of the City (it is written): From this day on, whoever buys meteoric...
2,As soon as you have heard our letter, who(ever) over there has either sold it to...
3,Send a copy of (this) letter of ours to every single colony and to all the trading...
...
```

Code Requirements

linkkeyboard_arrow_up

￼

Submissions to this competition must be made through Notebooks. In order for the "Submit" button to be active after a commit, the following conditions must be met:

- CPU Notebook <= 9 hours run-time
- GPU Notebook <= 9 hours run-time
- Internet access disabled
- Freely & publicly available external data is allowed, including pre-trained models
- Submission file must be named submission.csv

Dataset Instructions

linkkeyboard_arrow_up

By far the biggest challenge in working with Akkadian / Old Assyrian texts is dealing with the formatting issues. As they say, “garbage in, garbage out” and unfortunately, the format of text in transliteration poses challenges at each step of the ML workflow, from tokenization to the transformation and embedding process.

To mitigate these issues, we provide the following information and suggestions in handling the different formatting challenges in both the transliterated and translated texts.

Texts in Transliteration

- Main formatting challenges: in addition to the standard transliteration format, with hyphenated syllables, additional scribal additions have encumbered the text with superscripts, subscripts, and punctuations only meaningful to specialists in Assyriology (Complete Transliteration Conversion Guide).
- Capitalization is also a challenge, as it encodes meaning in two different ways. When the first letter of a word is capitalized it implies the word is a personal name or a place name (i.e. proper noun). When the word is in ALL CAPS, that implies it is a Sumerian logogram and was written in place of the Akkadian syllabic spelling for scribal simplicity.
- Determinatives are used in Akkadian as a type of classifier for nouns and proper nouns. These signs are usually printed in superscript format adjacent to the nouns they classify.

To avoid the potential confusion of reading a determinative sign as part of a work, we have followed the standard transliteration guide and retained curly brackets around these. While this may pose challenges in ML, we note that this is the only use of curly brackets in the transliteration (e.g. a-lim{ki}, A-mur-{d}UTU).

- Broken text on the tablet: as these are ancient texts, they include a number of breaks and lacunae. In order to standardize these breaks, we suggest using only two markers, one for a small break of a single sign <gap> and the other for more than one sign up to large breaks <big_gap>.
- For the purpose of this challenge, we include suggestions of how best to handle these formatting issues below.

Texts in Translation

There is currently no complete or extensive database for translations of ancient cuneiform documents, and this is especially true for the Old Assyrian texts. For this reason, we gathered together the books and articles with the translations and commentaries of the Old Assyrian texts and we digitized them with an OCR and LLM for corrections. Even after all that work, there are still a number of formatting issues with these translations, which makes this a central component of the challenge for successful machine translation development.

Translations usually retain the same proper noun capitalization, and these proper nouns in general are where most ML tasks underperform. To account for these issues, we have included a lexicon in the dataset which includes all the proper nouns as specialists have normalized them for print publications.

Modern Scribal Notations

Lastly, it is important to note that there are modern scribal notations that accompany the text in transliteration and translation. The first of these include line numbers. These are typically numbered 1, 5, 10, 15, etc. However, if there are any broken lines, the line numbers will have an apostrophe immediately following (‘), and if there is a second set of broken lines, the line numbers will have two trailing apostrophes (‘’). These are not quotation marks, but a scribal convention editors sometimes use in publication.

Additional scribal notations include:

- Exclamation marks when a scholar is certain about a difficult reading of a sign !
- Question mark when a scholar is uncertain about a difficult reading of a sign ?
- Forward slash for when the signs belonging to a line are found below the line /
- Colon for the Old Assyrian word divider sign :
- Comments for breaks and erasures in parentheses ( )
- Scribal insertions when a correction is made in pointy brackets < >
- The demarcation of errant or erroneous signs in double pointy brackets << >>
- Half brackets for partially broken signs ⸢ ⸣
- Square brackets for clearly broken signs and lines [ ]
- Curly brackets for determinatives (see below) { }

Formatting Suggestions for Transliterations and Translations:

Remove (modern scribal notations):

- ! (certain reading)

- ? (questionable reading)
- / (line divider)
- : OR . (word divider)
- < > (scribal insertions, but keep the text in translit / translations)
- ⌜ ⌝ (partially broken signs, to be removed from transliteration)
- [ ] (remove from document level transliteration. e.g. [KÙ.BABBAR] → KÙ.BABBAR)

Replace (breaks, gaps, superscripts, subscripts):

- [x] <gap>
- … <big_gap>
- [… …] <big_gap>
- ki {ki} (see full list below)
- il5 il5 (same for any subscripted number)