Morphological parser for Sumerian

Mohammad Sameer
Aligarh Muslim University
+91 999 076 5607
mhdsmr2@gmail.com
gitlab.com/m-sameer
github.com/m-sameer

Project brief

The project aims to have a rule-based parser that performs morphological annotations on Sumerian using SFST. This project would be a fine tool to analyse the CDLI dataset of Sumerian as other look-up or statistics based parsers have an absurd way to interpret text and would reliably make mistakes in much of the data just like doing Google translate on Latin.

Project detail

As for now, all the translation that can be done of Sumerian is that by Assyriologists, which are not that many, or by non-rule-based parsers who don't do a good job at translating the sentences. This project aims to do the first step in the right direction to develop a rule-based morphological parser to ease the growth of further development in the study of the language of Sumer and Akkadian civilizations.

SFST is a finite-state transducer that will be used to write the morphologies for Sumerian. The analysis would be on three-level:

i. Morphological, turning the grammatical features into morphemes with morpheme boundaries and zero morphemes.

ii. Morphophonological, transcribing the morphemes to the conventional transcription (by removing morpheme boundaries).

iii. Transliteration, changing to make it orthographically accurate by the conventions of CDLI orthography.

SuMor is a toy grammar developed by Prof. Christian Chiarcos as a kickstart to this project. As it's a prototype for illustrative purposes. I would directly work on a fork of this repo and merge it after coordinating with the mentors.

Prime Difficulties

Prime difficulties I may encounter during the project might be the following:

- The unusual language of Sumerian, being a language isolate itself and having no sister or daughter language of its own, I could get a feel for it only through descriptive grammar and taking the abstract idea and constructing a language in my head.

- Having to conform to CDLI transliteration conventions as it's uncommon for morphological parsers to have transducers for orthographical purposes.

- Some difficult concepts of linguistics phenomenon not present in languages I'm fluent in which would add a bit more friction for me towards grasping Sumerian.

Timeline

May 20 - June 12 (Community Bonding Period)

Getting familiar with tools that are going to be used in building the parser:

- Read the Introduction to Sumerian Grammar by Daniel A. Foxvog thoroughly to get familiar and start getting some plans on how to map out the grammatical structures in SFST beforehand.
- Analyse some examples from the morphologically annotated CDLI corpora, esp., github.com/cdli-gh/mtaac_gold_corpus/tree/workflow/morph/to_dict and github.com/cdli-gh/mtaac_gold_corpus/tree/workflow/morph/external, by hand.
- Identify and try to resolve how annotation terminology in grammar and the CDLI examples relate to each other.
- Discuss with the mentor about working and scheduling meetings that'll take place further down the line.

June 13 - July 25 (Coding, or rather transducing, Phase I)

Following are most of the things I'll be completing in those 6 weeks:

- Read up on CDLI conventions to write Sumerian in Latin script
- Build a lexicon from CDLI text files to build a Sumerian lexicon
    - Write a script (probably in awk or python, I think python would be preferred because of its ease for new contributors) that builds a TSV lexicon file by reading the Gold Corpus' conll.
    - Testing would be done by checking against the MTAAC's Annotated Files for Morphology the files have the forms, segmented forms, meanings, and POSs. For coverage see: github.com/smc/mlmorph/blob/master/tests/coverage-test.py
    - Compile training, dev and test files over which we'll test the accuracy of our morphology.

- Learn more about how morphological parsers are made by dwelling into the code of, primarily, SMOR (German morphology) and mlmorph (Malayalam morphology)
- Add tests, wiring MTAAC corpus to SuMor
- Get my whitepaper draft reviewed.

July 25 - July 29 (Phase I Evaluation)

- Document the work done, for the way of documenting see: github.com/cdli-gh/annodoc
- Quantitative evaluation: precision, recall and coverage against dev corpus.
- Submit everything to the mentor

July 29 - August 19 (Transducing, Phase II)

- Add transducers for irregular verbal and nominal forms obtained from tests from the MTAAC dev corpus.
- From now on it would be appropriate to focus only on transliteration transducers.
- Weekly evaluation report of precision, recall and coverage over the dev corpus.

August 20 - Sept 4 SPLIT OFF FINAL (Phase II Two weeks on eval)

- Evaluation precision, recall and coverage overtraining, test and dev corpus.
- Compile a progress diagram from these numbers and those of earlier reports
- Draft a whitepaper, analyse and document achievements and challenges

Sept 5 - Sept 12 (Sadly the final week of GSoC)

- Wrapup: I'll check if everything is working like it's supposed to, that my work can be replicated and my code can be deployed.

Time Availability

From the 5th of August, I have my semester exams that month so I will not be putting so many hours that I would be capable of in Phase I so I would like to have my Phase I have 35 hour weeks, 4 hours for weekday and 6 for weekends, and my Phase II have 25 hour weeks, 3 hour weekdays and 5 hour weekends.

I have no other commitments in my summers. I'll continue to maintain the project as long as I have my cognitive health.

About me

I'm a freshman, majoring in Psychology at Aligarh Muslim University, India. Have been doing computer programming for a few years primarily fixing bugs in mobile and web applications, correcting UIs in open source, and doing competitive programming on CodeChef and CodeForces.

Why CDLI?

I would be honoured to work with an organisation working towards preserving a language, and not just a language but a culture frozen pure in its expression of the Collective Unconscious. From the stories of the king of Ur to the majestic Utu, we have so much to explore in the realm of mythical stories as we're having an epistemological crisis over if human nature is a mere social construction? After skimming through some of the Jungian works, one is tempted to think otherwise.

And apart from all that, I would work on morphology, a subfield of linguistics and linguistics intrigues me a lot, in which I plan to do my master's degree.

Why me?

I would like to intensify initial experiences with previous open source projects. Became a Google Code-In finalist 2018 from Digital Impact Alliance. And from that I gained experience on how to work with mentors in different time-zones and read the manual first before asking stupid questions (the first rule in OSS repositories :P).

To count some of my previous learning voyage : Made patches to Apache Fineract: github.com/apache/fineract/pull/493, Added translation for Urdu (my native language) in Mifos Mobile: github.com/openMF/mifos-mobile/pull/965, Changes to UI to make it more beautiful: github.com/openMF/mobile-wallet/pull/125, Little contributions to web apps here and there: github.com/openMF/community-app/pull/3059.

I forked cdli-gh/sumor and did some changes to extend it for another example in Jagersma C5.1: github.com/m-sameer/sumor.

And secondly I want to learn linguistics as it's been my interest since childhood. When I was about 10, I figured out, by myself, that there's a thing which I now know as place of articulation or مخرج الحرف (in Arabic) by changing the sound (by moving tongue or opening vocal tract) of every phoneme a little till it became another phoneme (which gave me an unnamed idea in my head that there can be difference between the sounds of phonemes but there's a limit till the phoneme changes, which I now know as allophones and phonemes) and writing down from where (the position of the tongue) till where it could be considered the same sound Getting the range of allophones that can be considered a particular phoneme in a range.