GSoC 2021 Project Proposal for Apertium

Contact:

Name: Gourab Chakraborty

E-mail address: 19bcs118@iiitdwd.ac.in

Whatsapp Contact: 9432200935

IRC: gourab337

Why is it that you are interested in Apertium?

I have worked in various NLP projects (including creating the multi-language support for Hindi/English/Kannada for our University website and one internship) and want to help create a new language pair (Hindi to Bengali) for Apertium (Bengali is my native language and a widely used language in India).

Hindi is the 4th most spoken language and Bengali is the 5th most spoken language in the world and also there are a lot of amazing works in Bengali (including Nobel Prize winning literatures) and hence I believe we should have a translation language pair for Hindi-Bengali so that one language speaker can get access to literatures of the other language via effective translation.

I have worked in previous projects in the field of language processing and am confident that I can positively contribute to this amazing initiative by Apertium.

Which of the published tasks are you interested in? What do you plan to do?

Creation of a new language pair that is ready for publication.

There is a lot of research and literature works in Bengali (in religion, philosophy, music, songs, poems and scripts) that India is currently missing out on but also the same is true

the other way round. Like there are many literatures in Hindi that Bengali speakers are missing out on. Many aren't even properly monetised (since the works are read in a limited region). Creating an efficient translator would open up these works to a much wider audience and hence benefit scholars, casual readers, Indians who find reading their native language more comfortable than the source language. Also Hindi-Bengali pair is an essential which is missing in Apertium's vast library of language pairs.

For this reason, I believe that Google and Apertium should sponsor this project for GSoC 2021. And I would ensure that I give my 100% in creating apertium-hin-ben that would be ready for publication.

Deliverables :

Regarding the said task I'm interested in, I took a look at the Apertium wiki (https://wiki.apertium.org/wiki/Bengali_and_English) and http://rua.ua.es/dspace/handle/10045/12029 for reference.

My main focus will be on creating the Hindi-Bengali language pair during the duration of GSoC. I plan to attain the said task by:

1. Finishing the morphological analysis of Bengali.
2. Creating/expanding the transfer rules.
3. Creating the lexical selection rules.
4. Adding several thousand words in the bidix.
5. Testing on real texts to fine-tune the translator and presenting a finished translator with a WER of less than 25%, ready for publication, at the end of the project.
6. Proper documentation for the above mentioned deliverables.

Proposal timeline / Work Plan:

This week-by-week timeline provides a rough guideline of how the project will be done.

Before May 20:

- Familiarize with the code and the community, the version control system, the documentation and test system used, and the Apertium engine.
- Learn how to make an XML dictionary in order to make language pairs so that I can edit the Apertium dictionary and code.

May 20 - June 7 (Before the official coding time):

- To do self coding with XML to improve my further understanding and ease of use with Apertium.
- During this period I will remain in constant touch with my mentor and the Apertium community. I will remain active on IRC and mailing lists to discuss and finalize on the modifications.
- Thus with the help of my mentor I will become absolutely clear about my future goals, the final implementations that need to be done as well as the approach that I will follow to create the apertium-hin-ben language pair.

June 7 - June 25 (Official coding period start):

- Gather a large collection of Bengali literature with different types of inflection systems (for both Bangladeshi Bengali and Indian Bengali).
- Analyze morphologically to cover all the possible phonological processes in the language and also take care of exception cases.
- Improve the morphological analyser as part of the Apertium machine translation system.

June 26 - July 15 :

- Improve the morphological generator as part of the Apertium machine translation system.
- Create / expand the transfer rules.
- Create the lexical selection rules.

July 16 Evaluation Phase 1

June 16 - June 25 :

- Making further changes in the code to improve the functionality, exception handling, bug removal.
- Keep adding words in bidix (up to several thousand words).

June 25 - August 5 :

- To be in constant touch with the Apertium's developers and to let them know about our progress.
- Most of the time will be consumed for rigorous testing and bug fixes.
- Testing on real texts (gathered earlier) to fine-tune the translator to ensure a word error rate of less than 25%, ready for publication, at the end of the project.

August 6 - August 13 :

- For Documentation (GitHub pages, Apertium Wiki and/or Medium blog).

A buffer period of 10 days has been kept for any unpredictable delay.

About Me:

I'm a competitive programmer with a passion for development. I like to see my code turn into meaningful projects that impact our society. I am pursuing **Computer Science and Engineering** from **Indian Institute of Information Technology, Dharwad** and am currently in my 2nd year of my undergraduate course. My technical skills include C++(Strong), Python(Strong), Javascript, HTML, CSS, XML, bash(strong), SQL, APIs.

Past year I created COVID-19 Tracker, Edubile and was involved in developing a website for our institute. For our Institute web dev, I was responsible for the frontend part (HTML / CSS / JavaScript) and the multi-language support part

(Majority of the website gets translated using an API connected NLP language translator, which was built from ground up to save on the Google Translate API charges), the rest part of the website had to be hard coded for the 3 different languages to take care of Institute branding issues. In the same web dev project, I had the complete responsibility of creating and deploying an NLP (Pytorch based) Chatbot on the Django site. The website codebase is in our Institute's private repo, but I have the Chatbot repo public with complete demo and documentation in my GitHub account. I believe my frontend/backend/api skills might add value to the team if necessary.

I also did an internship last year around October/November where I had to create a Chatbot browser extension based on DialogueFlow. It takes queries related to Dev/ComputerScience and returns the most relevant Text article and Youtube Video. During my time at the internship, I worked directly under the guidance of the CEO, Mr. Yasin Shah.

GitHub: https://github.com/gourab337 , Devfolio: https://devfolio.co/@gourab337 , LinkedIn: https://www.linkedin.com/in/gourab-chakraborty-71a38a182/

This year I'm working on Project Gateway, which is a cross-platform campus management app made on Flutter. I have previously worked on open-source projects with both large and small teams and am confident that I would be able to deliver on the responsibilities that would be assigned to me.

I don't have any internships/jobs this summer and I will be able to give 30 hours and more, every week for the duration of 10 weeks during which GSoC would last as I am confident about my time management skills (had similar working schedule during my internship last year) and I give my hundred percent focus to what I do.

I'm very much interested in the field of language processing and want to explore this field of machine translation and morphological analysis. I believe that I would be able to add positive value to the Apertium team by creating the Hindi-Bengali

language pair. I would be really grateful if I am given the opportunity to work on this project. A large scale open source project like Apertium, would really help in improving my skills in the long run.