Contact Information:

Name: Gourab Chakraborty
Location: Kolkata, India
Email: 19bcs118@iiitdwd.ac.in
IRC: gourab337
GitHub: gourab337
Whatsapp: +91 9432200935

GSoC 2021 Proposal: hin-ben

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).

Apertium is free/open-source software and there is a lot of good work done and being done in it. It is not only machine translation, but also free resources that can be used for other purposes: e.g. dictionaries, morphological analysers, spell checkers, etc. Hindi/Bengali is currently missing out on a lot of these and hence I feel it is essential to get it done.

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?

1. I'd like to create the pair Hindi-Bengali to publication.
2. It is one of the translation pairs that is incomplete in Apertium and I wish to complete it.

My proposal

Title

Adopting the Hindi-Bengali language pair (unreleased language pair).

Major goals

- Creating an hin-ben repository in Apertium. This should include:
  - Creating/expanding the transfer rules.
  - Creating the lexical selection rules.
  - Adding several thousand (in the range of 15,000) words in the hin-ben bidix.
- 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.
- Proper documentation for the above mentioned deliverables.

Reasons why Google and Apertium should sponsor it

- Apertium is free and open source. The resources generated by this project will be generally available, not just for use as a translator.
- There has already been some work on this language pair, but it is unfinished. This project will generate a deliverable that will be a functional machine translator.
- This will not only make Apertium a functional Bengali morphological analyser/generator, which can be used for other language pairs, but will also give visibility to this fact by creating a functional translator based on Bengali.
- It would be possible to make a proposal based on, for example, the Bengali-Assamese pair. The problem is that it would involve working on two languages that have very limited resources in Apertium. The time constraints of the project make it reasonable to concentrate on a pair where one of the languages is already "mature" in Apertium. A pair such as Bengali-Assamese is an excellent choice for a GSoC project once one of the two languages is functional in Apertium.

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.

Resources

I am a native speaker of both Bengali and Hindi. There are several sources like glosbe, wiktionary that can be used to fill up the bidix fast. I have a friend who is a student of Bengali

Literature and I am in contact with him. She has agreed to extend her help in case I need it. This will definitely help me in translating the words faster. I can build up from the pre-existing repository https://github.com/srj31/apertium-ben-hin. This repository has 411 entries in it's bdix (20 Adj, 20 Adverbs, 6 Conjunctions, 5 Determiners, 63 Nouns, 9 Postpositions, 23 Pronouns, 265 Verbs). This is very less as compared to the task of putting in ~18,000 entries that I have set. But this is definitely better than any work done in the same from other developers. In my work plan, I have considered adding words from 0, so this is definitely a headstart for me.

The Apertium Bengali Morphological Analyser (https://github.com/apertium/apertium-ben) has a Naive Coverage of 80.35% (Prothom-Alo) and 68.21% (Wikipedia). Precision of the analyser on the Prothom-Alo corpus (Random 1000 words) is 99.8% and Recall of 88.29%. Ref: Abu Zaher Md. Faridee and Francis M. Tyers (2009) "Development of a morphological analyser for Bengali". Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation. pp. 43--50

Collection of Resources and Corpus that I can build upon:
https://github.com/banglakit/awesome-bangla

Work done on Hindi Morphological Disambiguator: https://github.com/apertium/apertium-hin

The Hindi morphological analyser at (https://github.com/apertium/apertium-hin) seem to be used in several apertium language pairs like Hindi-English, Hindi-Punjabi, Hindi-Gujarati, Hindi-Marathi, Hindi-Assamese. So I am planning to use this morphological analyser for improving Hindi-Bengali due to limited window of GSoC this year.

Workplan

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

Before 17 May:

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

Week	Dates	Goals	Bdix	WER	Coverage
Post	17 May - 7	- Find language resources for			

Application Period	June	Bengali, Hindi. ● Create an Bengali corpus ● Create a Hindi corpus			
1	7 June - 14 June	● Closed categories: determinants, prepositions, conjunctions, relative pronouns, other pronouns			
2	15 June - 22 June	● apertium-ben: verbal paradigms ● Expand bilingual dictionary ● Lexical selection rules	~1,500		
3	23 June - 28 June	● Expand bilingual dictionary ● Lexical selection rules	~3,500		
4	29 June - 6 July	● Expand bilingual dictionary ● Lexical selection rules	~6,000		
5	7 July - 14 July	● Expand bilingual dictionary ● Lexical selection rules ● Transfer rules hin-ben **Student Evaluation Phase 1**	~8,500	<40% (hin-ben)	
6	15 July - 22 July	● Expand bilingual dictionary ● Lexical selection rules ● Transfer rules hin-ben	~11,500		
7	23 July - 30 July	● Expand bilingual dictionary ● Lexical selection rules ● Transfer rules hin-ben ● Manual disambiguation of hin texts ● Disambiguation rules hin ● Building morphological disambiguator for hindi.	~13,500		
8	31 July - 7 Aug	● Expand bilingual dictionary ● Lexical selection rules ● Transfer rules ben-hin ● Manual disambiguation of hin and ben texts. ● Finish building morphological disambiguator for hindi. ● Build morphological disambiguator for bengali. ● Disambiguation rules hin.	~14,000		

9	8 Aug - 15 Aug	● Expand bilingual dictionary. ● Lexical selection rules. ● Finish building morphological disambiguator for Bengali. ● Transfer rules ben-hin. ● Disambiguation rules ben. ● Testvoc ben-hin, hin-ben: closed categories, vblex.	~14,500		
10	16 Aug - 23 Aug	● Manual disambiguation of ben. ● Testing the ben-hin translator. ● Add words, rules. ● Testvoc ben-hin, hin-ben: adj, adv. ● Testvoc ben-hin, hin-ben: n.	~15,000	<40% (ben-hin) <30% (hin-ben)	
		● Documentation (GitHub / Apertium Wiki / Medium Blog)			

List your skills and give evidence of your qualifications

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.

Coding Challenge:

2021 Coding Challenge for adopting an unreleased language pair to publication was not available in the Apertium Wiki so I proceeded with this.

https://github.com/gourab337/apertium-hin-ben

I have been studying the Apertium Wiki and working on apertium-ben-asm too at the same time. My work files for the apertium-hin-ben are in this repo:

https://github.com/gourab337/apertium-ben

https://github.com/gourab337/apertium-hin

Here I have been trying to build on top of the pre-existing hin-ben repo for apertium.

As the timeline for GSoC is short this year, I will continue expanding this repo so that I'm ahead of the above timeline.

List any non-Summer-of-Code plans you have for the Summer

This year my GSoC schedule will be in alignment with my summer vacation in college and due to the Coronavirus, I don’t have any travel plans during the summer. Just to clarify, I’ll be available for 30+ hours a week. Also since this is a topic of interest to me I’ll be able to put in more.