Vibha Hodachalli
Vaageesha Das
Raquel Buege

Project Proposal

1. *What is the problem or task you are focusing on?*

   The task that we are focusing on is machine translation from Quechua to Spanish.

2. *What is the format of the input and output of this task? For example, each input could be a sentence of text and the output could be a label from a discrete set of possible labels. Provide at least one example of input and output from your data.*

   The input of the task would be a sentence in Quechua and the output would be the translated sentence in Spanish.

   Example:

   Input: Ñoqa qorpa rishani
   Output: Yo voy al hotel.

3. *What data are you using? Please explain where these datasets are from and how they were constructed. Provide links to any URLs if the data is hosted online or links to papers if the dataset is published somewhere. If the data has labels or "gold" text that you are predicting or generating, where do those labels come from?*

   We will be using a corpus of parallel Quechua and Spanish translations. The data that we are using is provided by the Hugging Face Platform. We will use the reference translations in this dataset as gold labels for tuning and evaluating the MT system. The data was retrieved from a series of blogs and other apps. According to the platform, this is specifically where the data was taken from (as exactly listed on the website):

   1. "Mundo Quechua" by "Ivan Acuña" - available here
   2. "Kuyakuykim (Te quiero): Apps con las que podrías aprender quechua" by "El comercio" - available here
   3. "Piropos y frases de amor en quechua" by "Soy Quechua" - available here
   4. "Corazón en quechua" by "Soy Quechua" - available here
   5. "Oraciones en Español traducidas a Quechua" by "Tatoeba" - available here

6. "AmericasNLP 2021 Shared Task on Open Machine Translation" by "americasnlp2021" - available here

URL: https://huggingface.co/datasets/somosnlp-hackathon-2022/spanish-to-quechua

4. *What approach are you taking to building a NLP system to handle this task? What software packages are you planning to use to build this system? Except in some cases, the approach should draw on statistical approaches we've covered in class so far, such as n-gram representations of text. Talk to the instructor if you are not sure about this.*

   We are planning to train a *statistical machine translation* model on the Spanish-Quechua bitext using Moses. Since our bitext consists of phrases, we will use Moses' phrase-based decoder. Moses works similarly to n-gramming as we've learned in class where it produces the most likely set of phrases based on learned probabilities.

5. *How are you evaluating your approach? What performance metrics are you going to use?*

   One way we are planning on evaluating our approach is through the BLEU score (Bilingual Evaluation Understudy) which compares machine translation output to a professional reference translation. Moses has a perl script (multi-bleu.perl) that we can use as our BLEU scoring tool. Another evaluation metric we are planning to use is chrF (CHaRacter-level F-score) which compares MT output and reference translation using character n-grams; it can be done using the evaluate module and its chrf metric from Hugging Face. The evaluate module also has a bleu metric that we may use. We are also considering implementing an error analysis similar to what we had done for Homework 2.

6. *What kinds of ethical issues may be raised by your model or data?*

   Ethical issues that may be raised by our model are biases showing in our translations from Quechua to Spanish. This could be due to the fact that our data could be biased, and therefore since we are using this to make our models, then our models would reflect the same problems. Another ethical issue could be the loss of meaning in translation. There is potential for some niche aspects of Quechua to be lost when phrases are automatically translated to Spanish. Additionally, there is a potential for the model to replace actual folks who translate Quechua to Spanish.

7. *What are the proposed steps needed for completion of the project?*

   The first step we will need to take in using the phrase-based decoder is training the translation model. We will have to create a phrase translation table which will be the source of knowledge for the machine translation decoder. We are able to control the decoder and specify parameters using a configuration file. We will have to look into what parameters we want to adjust or control when running the decoder. After we have built the model, we will run the decoder and evaluate the model using BLEU and chrF and fine tune our model based on the results of the evaluations.

8. *What are roles and tasks of each person in the group? Though group members will contribute in various capacities, it is best if each person is responsible for at least one aspect of the project.*

   All three individuals will collaborate on all parts of the project but the specific roles each person will take on is: Vibha will engage in research and take care of the linguistic verification, Raquel will do quality assurance on the language translation as a human translator, and Vaageesha will oversee the implementation. With the model and two evaluations, Vibha will take care of the model, Raquel will take care of BLEU, and Vaageesha will take care of chrF evaluation.

Sources:
https://web.stanford.edu/~jurafsky/slp3/13.pdf
https://www.scribd.com/document/525044640/ORACIONES-EN-QUECHUA
http://www2.statmt.org/moses/?n=Moses.Tutorial (Moses phrase-based tutorial)
http://www2.statmt.org/moses/?n=Moses.SupportTools#ntoc5 (Moses scoring translations with BLEU)
https://machinetranslate.org/chrF
https://machinetranslate.org/bleu
https://huggingface.co/spaces/evaluate-metric/chrf (chrF evaluation)