NLP 245: Course Project Report

Malini Kar and Abigail Kufeldt

June 12, 2023

1 Research Problem

1.1 Problem Definition

In this project, we're interested in building a demo of a social bot intended to help language learners practice their grammar and conversational skills. We will be focusing on the English-French language pair specifically for this demo.

We intend to produce an agent that is useful to second language learners of French (with fluency in English). The goal of this tool is to help language learners with immersion, ie. conversing only in French. When the user makes mistakes, the agent will suggest corrections for those mistakes before proceeding with the conversation.

1.2 Problem Significance

This problem is significant because language acquisition relies heavily on consistent practice in the target language. However, many language learners, especially beginners, may be too embarrassed with their conversational ability to want to practice their skills with a real human. Having a conversational agent capable of encouraging the user to speak in their target language will allow the learner to more quickly acquire the language, and without shame or embarrassment. It is also important that beginner language learners have the freedom to make mistakes and get accurate corrections for them in real time.

This is a difficult problem to solve as it requires a bot to behave in the same way a human interlocutor would, which is not necessarily possible at this point in time. Thus, as the system designers, we must make some decisions about what capabilities we want to prioritize in this bot. We decided to prioritize the following: (1) immersion, ie. the bot aims to keep the user speaking in the target language, and (2) grammar error correction, ie. the bot will correct any ungrammatical French utterances from the user by informing them the correct way of stating that phrase.

There are a multitude of other potential capabilities for this type of bot, such as code-switching within a dialog turn, or answering more complex grammatical questions from the user, but we concluded those types of features were out of scope for this project.

2 Prior Work

Memrise, a prominent online language learning platform, has recognized the potential of chatbot technology in facilitating language acquisition. As part of their approach, Memrise developed MemBot[1], a chatbot designed to augment the language learning experience. They are not open source, however, and the only information we have about the agent is that it is powered by GPT-3.

The study “Learn to Speak Like A Native: AI-powered Chatbot - Simulating Natural Conversation for Language Tutoring“[2] proposes a language learning chatbot that creates a simulated conversational environment for learners to practice their target language. The chatbot is built upon a deep neural network trained on MultiWOZ and personaChat datasets, incorporating a novel chatbot structure to optimize performance. The evaluation results indicate that the chatbot achieves higher accuracy in generating appropriate responses, covering various domains of daily life conversations with a wide vocabulary base and diverse sentence structures. This paper's findings suggest that the proposed chatbot can offer valuable support to language learners, facilitating their acquisition of conversational skills and providing a more immersive and interactive learning experience.

The study “BookBuddy: Turning Digital Materials Into Interactive Foreign Language Lessons Through a Voice Chatbot“[3] addresses the limitation of digital materials in facilitating conversational practice, a crucial aspect of foreign language learning. Handling advancements in chatbot technologies, the researchers developed BookBuddy, an innovative virtual reading companion that transforms any digital reading material into an interactive English lesson. The pilot study involved five 6-year-old native Chinese-speaking children learning English, and the preliminary results indicate that the children enjoyed interacting with the virtual tutor and exhibited high engagement during the English-speaking sessions. This paper's findings demonstrate the potential of using voice chatbot technology to enhance language learning by providing interactive and engaging conversational practice opportunities using digital materials.

3 Design

3.1 Chatbot

We use RASA as our dialog system platform for this project. We train the RASA NLU to recognize when the user wants to initiate learning, as well as understand simple greetings and goodbyes. Within the Rasa backend, intents are a fundamental concept, representing the goals or purposes of the user's messages. By defining and training the chatbot with different intents relevant to language tutoring, in particular "greet," "vocabulary_question," and "conversation_practice," the chatbot can understand the user's intentions and respond accordingly.

Custom actions are another crucial component of our Rasa backend. These actions allow us to implement specific behaviors and functionalities for the chatbot. In the context of the language tutor chatbot, custom actions are used for the tasks of providing vocabulary definitions, simulating conversations, and performing grammar correction. Each custom action is implemented as a Python class and contains the necessary logic to execute the desired functionality.

To integrate the Rasa backend with our language tutor chatbot, we define the custom actions required for your application, the action_answer_qa and action_grammar_correction mentioned earlier. These actions can be implemented within separate Python classes, each responsible for executing a specific task. We connect to GPT3.5 via these custom action classes.

Within the Rasa configuration files, we specify the intents, training data, and action mappings for the chatbot. The training data includes example conversations, intents, and entities specific to language tutoring scenarios. This data is used to train the Rasa model, enabling the chatbot to understand user input and generate appropriate responses.

During runtime, when a user sends a message, the Rasa backend processes the message, predicts the user's intent, and determines the appropriate action to take. If the action is a custom action, such as performing grammar correction, the backend executes the corresponding action class and generates a response.

3.2 Grammar Error Correction (GEC)

We also need to have a grammar error correction (GEC) model that we can integrate into the RASA framework. To do so, we plan to finetune a pretrained French LM model on the Multilingual GEC dataset[6]. We plan to use mT5[5] to do so, a massively multilingual T5 model pretrained on 101 languages, including French. Our baseline for this set of experiments would be Barthez without finetuning, just zero-shot evaluated on our data.

We are further interested in comparing the performance of mT5 finetuned for GEC against GPT-3 zero-shot evaluated on this task. We use the `gpt-3.5-turbo` model from OpenAI in conjunction with the LangChain API to prompt GPT-3. After analyzing the performance of both approaches, the better of the two will be what we use in the backend of our RASA chatbot.

3.3 GEC Dataset

Multilingual GEC is a dataset of sentences with errors and their corrected counterparts in English, French, German, and Spanish. There are 67,157 French sentence pairs. We finetune the grammar correction model on just the FR data points from the Multilingual GEC data.

Note that the original data as seen on HuggingFace had a 99/1 train-test split, so we decided to take the train split only and divide that into 90/5/5 train-dev-test datasets. The table below reflects the size of each set after splitting in this fashion.

Split	**# Datapoints**
Train	59,850
Dev	3,325
Test	3,325

Table 1: Multilingual GEC Dataset Splits

4 Evaluation Plan

Since this project is focused on producing a demo of the intended chatbot, there's no clear evaluation metric we can apply to our demo to know when we've accomplished our task.

That said, we can measure the performance of our grammar correction model using traditional ML metrics for this task. Drawing from Choshen and Abend's 2018 paper[3] comparing different metrics in the GEC task, we originally evaluate our model's performance using BLEU and GLEU scores. GLEU score is a GEC-specific metric, similar to BLEU although it penalizes unchanged n-grams in the output that are changed in the reference text, thus encouraging the model to actually alter the inputted text.

However, in the process of working on our project, we discovered that the NLTK implementation of GLEU score that we were planning on using doesn't seem to be the same metric as described in Choshen and Abend 2018, as it doesn't utilize source sentences at all. Due to time constraints we were unable to implement this metric ourselves and reluctantly decided to rely on BLEU score alone to get a gauge of how well our models perform on the GEC task.

We also compare the performance of zero-shot and few-shot GPT-3.5 on the dev and test sets of Multilingual GEC against our finetuned mT5 model.

5 Experimental Results

5.1 GEC Model: Baseline

The zero-shot mT5 model was unable to complete this task to any degree of accuracy as it had no knowledge of the task at hand. Thus the model outputs were virtually nonsense, and, although we computed the evaluation metrics for consistency, the results were not very useful.

In the table below, we report the dev and test set zero-shot evaluation results for our baseline model, mT5.

Split	**BLEU**
Dev	0.0044
Test	0.0043

Table 2: Eval Results for Baseline Model

5.2 GEC Model: Finetuned mT5

We finetune mT5 on the Multilingual GEC dataset, training for 2 epochs with a learning rate of 0.001 using the Trainer API from HuggingFace. The figure below illustrates training and validation loss during training; note that the model converged quickly, only needing 1.5 epochs despite the size of the training data.

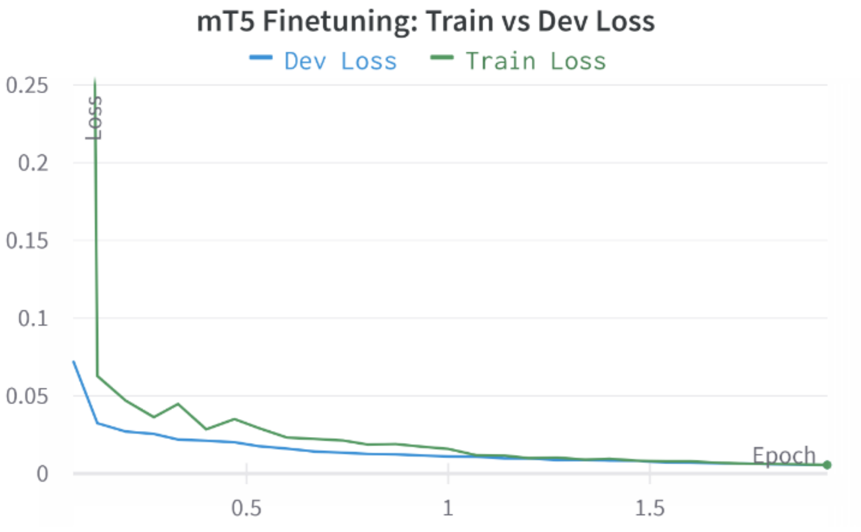

Table 2 below illustrates the dev and test evaluation results for this finetuned model.

Split	**BLEU**
Dev	0.9455
Test	0.9446

Table 3: Eval Results for Finetuned mT5

Although this is a startlingly high BLEU score, the dataset is not a particularly difficult task for a model like mT5 that has already been pretrained on the French language. Thus, this high BLEU score indicates thatthe model has learned to perform the types of grammatical corrections represented in the Multilingual GEC dataset very well.

5.3 GEC Model: Prompt GPT-3.5

In this part of our project, we prompt the `gpt-3.5-turbo` model from OpenAI to perform GEC on the Multilingual GEC test set. After experimenting with a few different prompt variations, we use the following as our prompt:

```
Correct the errors in the following French text without telling
me what you did:
```

The table below reports our evaluation results for zero-shot and few-shot prompting GPT on the test set:

Approach	**BLEU**
Zero-shot	0.7833
Few-shot	0.8185

Table 4: Eval Results for GPT-3.5 Prompting

It is unsurprising that GPT fails to surpass the performance of mT5 in terms of BLEU score considering that mT5 was finetuned for this task and, ultimately, BLEU is simply measuring how well a given model performs the task set by the evaluation data. Thus, although it may be the case that mT5 is better at making the grammatical corrections that are represented in the data, it's not clear that mT5 would be a better backend model for our bot. Because of this, it would be prudent to perform some error analysis to get insight into how each model behaves.

6 Error Analysis

We compared the performance of finetuned mT5 and the better of the two GPT-3.5 approaches, the few-shot version, on the following input:

```
bonjour monisserie
```

We chose this input because *monisserie* is not a valid French word, but could be a mistaken input from a very confused early language learner. We wanted to use an extreme example that would be more reflective of a human user than the Multilingual GEC dataset, to gauge how each model responds.

Given this input, mT5 predicts the following correction:

```
Bonjour, mon-série.
```

where *mon série* translates to something like *my series* or *my list*, depending on context. Obviously, neither reading is semantically valid in this case, but one could interpret mT5 as choosing to make the least number of changes to the incorrect input necessary to produce a grammatical French utterance, which it achieved.

In stark contrast, few-shot prompting GPT-3.5 produces the following:

```
Bonjour pâtisserie. (Assuming you meant to say "Hello bakery"
instead of "Hello my miseries")
```

Here, although still not particularly semantically sound, GPT appears to choose the most logical French word that is morphologically similar to the incorrect word. Interestingly, it acknowledges the possibility that its prediction does not match the user's intent, and offers another possible phrase that the user may have been trying to express—*mes misères* meaning *my miseries.*

Note that a future iteration of this chatbot would benefit from having an option to customize whether the bot converses with the user entirely in French (ie. ”immersion mode”) as opposed to producing outputs such as this, whether the corrected grammar is in French but the bot adds additional comments to the user in English (ie. ”beginner mode”).

From this example alone the difference between the models is clear, and also reflects our intuition given our knowledge of the GPT series in general. Prompting GPT-3.5 as opposed to using a finetuned model like mT5 will result in more linguistically-intuitive grammar corrections, as well as a more interactive experience for the user. Because of this, we decided to use the few-shot prompting approach with GPT-3.5 as the backend of our bot.

7 References

[1] Memrise. (n.d.). https://www.memrise.com/blog/introducing-membot.

[2] Tu, J. (2020). Learn to speak like a native: AI-powered chatbot simulating natural conversation for language tutoring. Journal of Physics: Conference Series, 1693(1), 012216. https://doi.org/10.1088/1742-6596/1693/1/012216.

[3] Ruan, S., Willis, A., Xu, Q., Davis, G. M., Jiang, L., Brunskill, E., Landay, J. A. (2019). BookBuddy: Turning Digital Materials Into Interactive Foreign Language Lessons Through a Voice Chatbot. Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale. https://doi.org/10.1145/3330430.3333643.

[4] Choshen, L., Abend, O. (2018). Automatic metric validation for grammatical error correction. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).
https://doi.org/10.18653/v1/p18-1127.

[5] mT5. (n.d.). https://github.com/google-research/multilingual-t5.

[6] Multilingual GEC Dataset. (n.d.).
https://huggingface.co/datasets/juancavallotti/multilingual-gec.