I made some big mistakes while trying to translate Cairo Geniza Documents.

Automated translation would be very useful to the Geniza Research Community. Right now, there are roughly 6000 documents with transcriptions in the Princeton Geniza Archive. However, there are only roughly 600 documents with translations. Of these, half of them are in Judeo-Arabic.

My first attempt to translate documents fom My first big mistake was being too early.

Right now, I am trying to figure out how to pick the best documents to try to translate. I am pretty certain that I am prompting about as well as I will be able to but I realized that my dataset isnt as good as I had hoped.

I think the problems are

1. Document transcriptions and translations don't always overlap so my AI translator might be translating text it cant compare itself against.

2. The documents are often unclean with a bunch of empty text because many of the documents are ripped or torn or otherwise missing stuff

So right now I am trying to fix these problems.

I fixed #1 by doing a new scrape and actually lining this stuff up

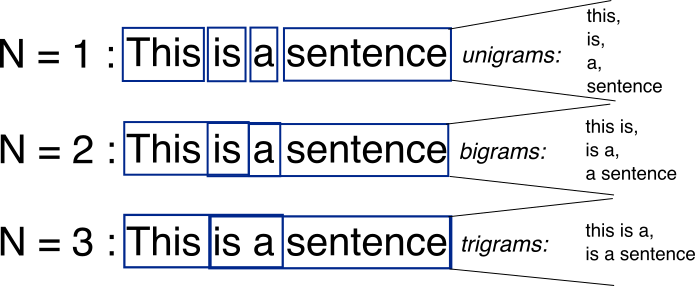
I am currently working on #2. I am thinking of making a little ngram model of clean text and comparing actual text against it to find its perplexity. It might just be easier to take the # of "...?[](){}" characters and divide it by the total # of characters. I imagine id get roughly the same result

Title: "Challenges and Progress in AI-Assisted Translation of Cairo Geniza Documents"

1. Preface

I am hoping that this might be of interest to other people either working on translating low resource language documents or people working on the Cairo Geniza. I don’t want to explain terms that will be easily understood by each group to each group, as they have each been discussed in detail elsewhere. I am going to use some terms familiar to people in each community without doing the proper work to introduce them. So I will describe them here and will link to a smarter person’s full explanation of them as well.

**Machine Translation Terms**:

What is a BLEU score? The primary programming task for a BLEU implementor is to compare n-grams of the candidate with the n-grams of the reference translation and count the number of matches. These matches are position-independent. The more the matches, the better the candidate translation is. The standard BLEU score (which I use) finds all 1-grams, 2-grams, 3-grams, and 4-grams in each text and compares how many are the same in the two texts.[[1]](#footnote-1) I think the best way to understand a n-gram is with an example.  (source: deepAI)

**Cairo Geniza Terms:**

What is Judeo-Arabic? Judeo-Arabic is a language still spoken today by roughly 250,000 people. It is (and was) the vernacular language of Jews living in the Muslim-majority Middle East. It’s often been compared to Yiddish, which was the vernacular language of Northern European Jews. Like Yiddish, it is written in the Hebrew Script and someone unfamiliar with it would recognize it to be Hebrew. However, also like Yiddish, many of the word meanings are actually in the language of the surrounding majority I am actually not super familiar with it (I do not read or write Judeo Arabic, Hebrew, or Arabic) but it has been explained to me to effectively be “Hebrew Letters with Arabic Meaning,” which is to say that

1. Introduction

Currently there are roughly six thousand documents with a transcription in the Cairo Geniza. Of them:

* Roughly 3000 are transcribed Judeo Arabic Documents
* 633 have a translation into English or Hebrew from their original language
* 272 are transcribed Judeo Arabic Texts with an English Translation.

Ever since I first learned about the Cairo Geniza, I knew there was cool information that would require me to either learn Judeo-Arabic or figure out a way to mass translate the documents into English. Years ago, I had briefly tried to do a translation with a huggingface pipeline. Unfortunately, I did not know what I was doing and got nonsense as a result.

More recently, with the proliferation of large language models, I decided that it was time to try it again.

1. Initial Approach and Challenges

The first attempt with a LLM was very simple. I copied the raw text from five different transcriptions from the Princeton Geniza Project website, pasted them into the GPT-4 window with the instruction to translate them into English, copied the output text to a txt file, copied the human translation from the Princeton Geniza Project website into a separate txt file, and compared them using a BLEU score.

* + Explain the two main problems you encountered: a. Misalignment of transcriptions and translations b. "Unclean" documents with missing text

1. Data Preparation Improvements
   * How you addressed problem #1 with a new scrape to align transcriptions and translations
   * Your current work on problem #2:
     + Proposed ngram model for clean text comparison
     + Alternative method: ratio of special characters to total characters
2. Translation Process and Evaluation
   * Describe your translation attempts with and without summaries
   * BLEU score results:
     + Without description: up to 0.2
     + With description: up to 0.27
   * Discussion on the potential trade-offs of using summaries
3. Ongoing Challenges and Considerations
   * The dilemma of using summaries: potential for improved accuracy vs. risk of "made-up" translations
   * The importance of ground truth comparisons
   * Your focus on ~300 documents with both transcriptions and English translations
4. Future Steps
   * Your plans for improving the dataset and translation process
   * Potential for increasing BLEU scores with refined data
5. Conclusion
   * Recap of progress made and challenges overcome
   * The potential impact of this work on Geniza research
   * Call for collaboration or feedback from the research community

1. Read the original paper here: https://aclanthology.org/P02-1040.pdf [↑](#footnote-ref-1)