Benchmarking Google Translate

Philosophy of Computation Lab IV

Henry Blanchette

April 5, 2019

Abstract

TODO

1 Introduction

2 Transcript Setup

I selected transcript sections that reflect a variety of writing styles, including modern English, English, technical writing, storytelling, and English translated from other languages.

Transcripts:

T1. The Bible, Genesis

T2. Melville's Moby Dick, Chapter 1

T3. Mariam-Webster English Dictionary, definition of Abdicate

T4. Bedau's patentsample.txt

T5. Shakespeare's Henry IV, Part 1

3 Translation task

Start with a sequence of languages $L_0, \ldots, L_n$ and a transcript in L_0, called the *original transcript.* GT translates the L_0-transcript to L_1, and then translates the resulting L_1-transcript to L_2, and so on until the transcript has been translated to L_n. At the end, there is left a L_n-transcript. Then, GT translates this L_n-transcript back to L_0 - the result is the *processes transcript.* The differences between the original and processed transcripts are measured to rate GT's success at this task. The goal of this scoring is to rate GT according to how well it preserve the meaning and grammatic structure of the original transcript.

4 Translation Success Measure

I rated GT's success at the task by how close the processed transcript was to the original transcript in terms of meaning and grammar. For each of these dimensions, I categorized an ranking of ordered performance classes.

Grammar is how properly-constructed the processed transcript is according to the rules of L_0 and the grammatical structure of the original transcript. The following are the classes of grammar performance I partitioned in order of increasing success. They are meant to be "evenly spaced" in the space of deviations from the exact grammar and structure of the original transcript.

Grammar Class	Description
1	completely confused
2	mostly confused
3	often confused
4	sparsely confused
5	effectively perfect

Meaning is how close the processed transcript is to the original transcript in meaning. The following are the classes of meaning performance I partitioned in order of increasing success. They are meant to be "evenly spaced" in the space of deviations from reflecting the exact meaning of the original transcript.

Meaning Class	Description
1	irrelevant
2	sparsely relevant
3	often relevant
4	mostly accurate
5	effectively perfect

5 Predictions

In regards to language speaker counts, I hypothesized that GT will perform better on these languages than the languages in Experiment 2 because I assume that (1) having more native speakers correlates with having more training data for GT's algorithms, and (2) having more data to train on yields a more accurate GT in the ways measured by this experiment. Assumption 1 seems very likely in general, but assumption 2 maybe depends heavily on how exactly GT works and whether it actually works well at all.

In regards to transcript style, I hypothesized that GT will perform better the more technical and less artsy the transcript. For example, I think that it will not well-maintain neither the grammatical structure nor meaning of the Shakespeare prose because it contains analogies and uncommon grammatic structures.

6 Experiment 1: Translation Ring with Well-Documented Languages

6.1 Language Setup

I selected from the top 5 languages (without English) by native speaker count. The following are the languages used in this experiment in order of decreasing native speakers count: Chinese (simplified), Spanish, Hindi, Arabic, Portuguese.

6.2 Experimental Design

I ran each transcript through the following trials, where the selected languages and their order was chose randomly:

Trial 1: Chinese → Arabic → Spanish → Portuguese → Hindi

Trial 2: Hindi → Chinese → Portuguese → Arabic → Spanish

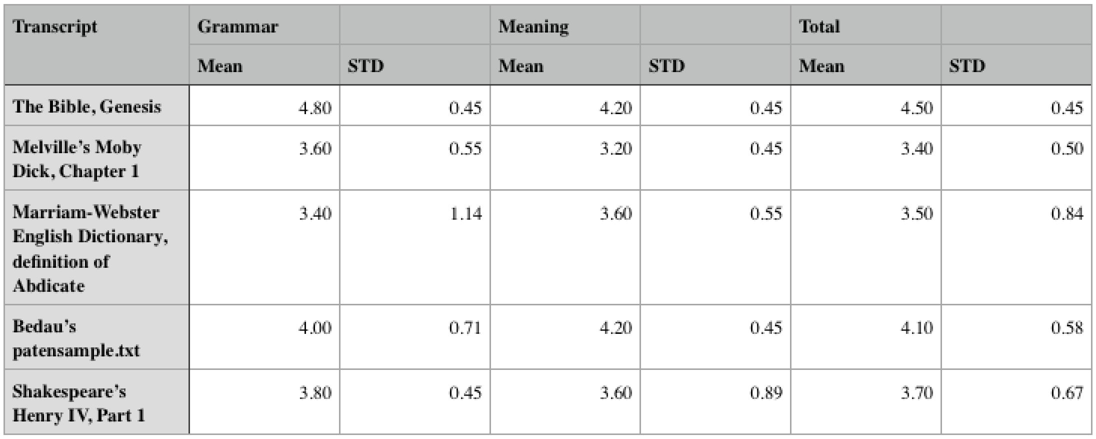
Trial 3: Hindi → Spanish → Arabic → Chinese → Portuguese

Trial 4: Chinese → Arabic → Spanish → Portuguese → Hindi

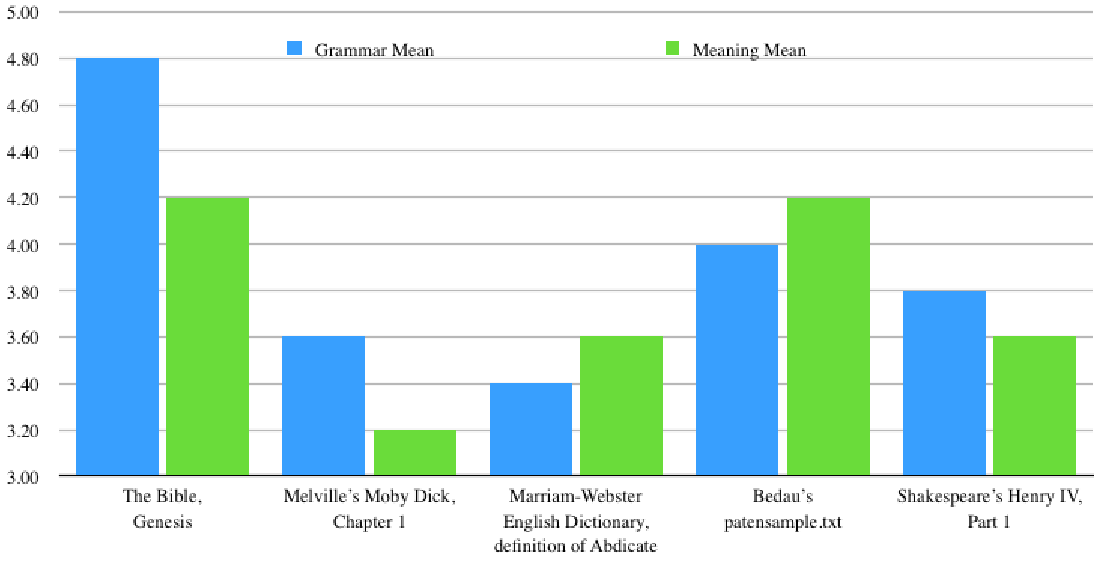
Trial 5: Arabic → Chinese → Portuguese → Spanish → Hindi

6.3 Results

Total Scores



Transcript	Grammar		Meaning		Total	
	Mean	STD	Mean	STD	Mean	STD
The Bible, Genesis	4.80	0.45	4.20	0.45	4.50	0.45
Melville's Moby Dick, Chapter 1	3.60	0.55	3.20	0.45	3.40	0.50
Marriam-Webster English Dictionary, definition of Abdicate	3.40	1.14	3.60	0.55	3.50	0.84
Bedau's patensample.txt	4.00	0.71	4.20	0.45	4.10	0.58
Shakespeare's Henry IV, Part 1	3.80	0.45	3.60	0.89	3.70	0.67



7 Experiment 2: Under-Documented Language Translation Ring

7.1 Language Setup

I selected from the bottom 5 languages by native speaker count that Google Translate supports . The following are the languages used in this experiment in order of decreasing native speaker count: Nepali, Sinhala, Greek, Hungarian, Zulu.

7.2 Experimental Design

I ran each transcript through the following trials, where the selected languages and their order was chose randomly:

Trial 1: Zulu → Hungarian → Nepali → Sinhala → Greek

Trial 2: Nepali → Hungarian → Greek → Sinhala → Zulu

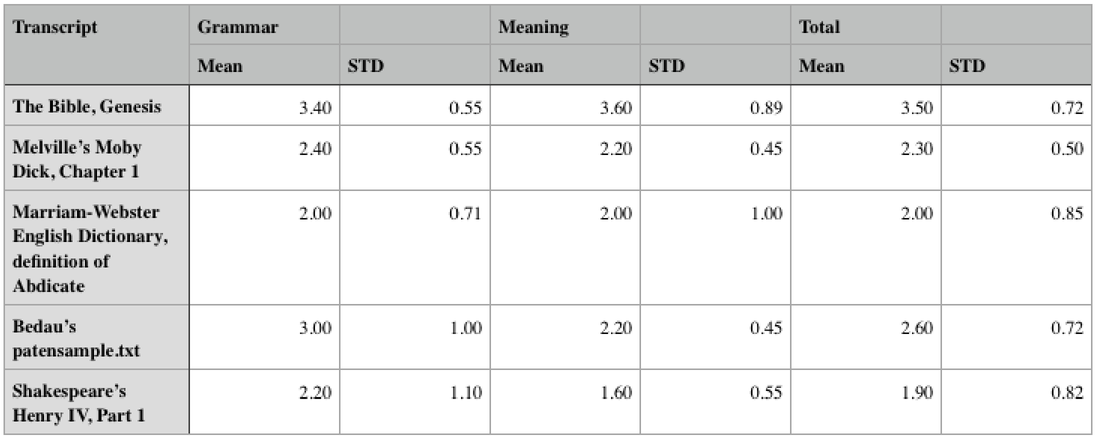
Trial 3: Nepali → Sinhala → Zulu → Greek → Hungarian

Trial 4: Sinhala → Greek → Nepali → Zulu → Hungarian

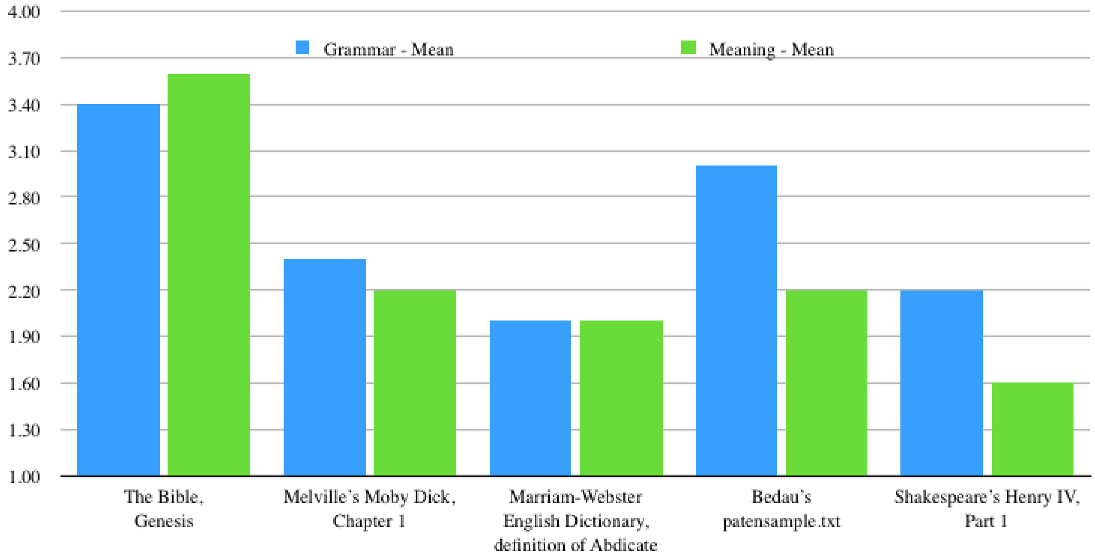
Trial 5: Hungarian → Greek → Sinhala → Zulu → Nepali

7.3 Results

Total Scores



Transcript	Grammar		Meaning		Total	
	Mean	STD	Mean	STD	Mean	STD
The Bible, Genesis	3.40	0.55	3.60	0.89	3.50	0.72
Melville's Moby Dick, Chapter 1	2.40	0.55	2.20	0.45	2.30	0.50
Marriam-Webster English Dictionary, definition of Abdicate	2.00	0.71	2.00	1.00	2.00	0.85
Bedau's patensample.txt	3.00	1.00	2.20	0.45	2.60	0.72
Shakespeare's Henry IV, Part 1	2.20	1.10	1.60	0.55	1.90	0.82



8 Analysis

9 Conclusion

Bibliography