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. Bedau's patentsample.txt

T3. Shakespeare's Henry IV, Part 1

T4. Melville's Moby Dick, Chapter 1

T5. Mariam-Webster English Dictionary, definition of Abdicate

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 rate GT's success at the task by how close the processed transcript is the the original transcript in terms of meaning and grammar. For each of these dimensions, I categorized an ordered ranking system.

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 grammar classes I used in order of increasing success.

G1. completely confused

G2. mostly confused

G3. often confused

G4. sparsely confused

G5. passing

G5. perfect

Meaning is how close the processed transcript is to the original transcript in meaning. The following are the grammar classes I used in order of increasing success.

M1. no meaning

M2. irrelevant

M3. sparsely relevant

M4. often relevant

M5. mostly accurate

M6. perfect

5 Experiment 1: Well-Documented Language Translation Ring

5.1 Language Setup

I selected from the top 5 languages (without English) by native speaker count . I hypothesized that this would correlate with the amount of effort that Google has put into training translations to and from these languages, which should yield more coherent and thus easier-to-score processed texts from this task.

The following are the languages used in this experiment in order of decreasing native speakers count:

L1. Chinese (simplified)

L2. Spanish

L3. Hindi

L4. Arabic

L5. Portuguese

5.2 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

5.3 Predictions

5.4 Results

5.5 Analysis

6 Experiment 2: Under-Documented Language Translation Ring

6.1 Language Setup

I selected from the bottom 5 languages by native speakers that Google Translate supports .

Languages:

L1. Nepali

L2. Sinhala

L3. Greek

L4. Hungarian

L5. Zulu

6.2 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