Responding to the “Arrival”: Essential Background Information and Strategies for Language Instructors in the Age of Human-Like Language Technologies (Machine Translation and Large Language Models)

Dr. Joel A. Walsh
October 12, 2023
Introductions

- Name
- Primary Language, Languages taught
- Favorite Bay area restaurant
Objectives

1. To be able to define what a language model is/does, and how machine translation works
2. To understand weaknesses and strengths of these technologies
3. To walk away from today and Friday’s workshop with concrete tools and next steps for working with these models in the future
- Machine Learning: Models that learn how to do tasks from data.

“…for some tasks, however, we do not have an algorithm, despite decades of research. Some of these are tasks we as human beings can do, and do effortlessly, without even being aware of how we do them. We can recognize a person from a photograph; we can move in a crowded room without hitting objects or people; we can play chess, drive a car, and hold conversations in a foreign language.”

Some definitions

- Machine Translation: Using rules or probability models to translate from one language to another

- Large Language Models (LLMs): Language Models predict the next token in a sequence, given the tokens that come before it. “Large” generally means trained on internet-scale corpora, parameter counts in the billions

- Natural Language Processing: Natural Language Processing (NLP) is concerned with, “…the set of methods for making human language accessible to computers”. *Eisenstein, J. Introduction to Natural Language Processing. (2018)*
What are some competencies and skills that students should gain by the end of a language course?

- Take 2-3 minutes to write down as many as you can in the padlet. Feel free to discuss in the chat, or in person once you are done.
- Be general (habits of mind, conceptual knowledge)! Be specific (tasks, interactions)! Be ungovernable (?)!
- At the end of 3 minutes, we will share out as a group, with a plot twist.
How many of your skills and competencies involved students creating...

How many of your skills and competencies involved other aspects of...
Agenda

(I. History of language technologies
II. Anatomy of modern Large Language Models and Transformers-based MT technologies
III. Implications of /strategies for use of language technologies in language instruction
(With some diversions along the way)
Early foundational machine translation technology

Predating Georgetown-IBM experiment

• 9th century- Al-Kindi develops many of the mathematical tools used in later systems - ie frequency analysis, probability and statistics applications

• 1940s - Claude Shannon and information theory, advances in cryptography from WW1
I. History of language technologies

- Early attempts to translate Russian/ DARPA involvement
- ELIZA -Joseph Weizenbaum @MIT
- The rise of statistical/empirical methods for natural language processing
- Winter
- Imagenet/ GPU revolution
- Attention and Transformers
- BERT -> ChatGPT
Early foundational machine translation technology

The Noisy Channel Model

\[ \hat{e} = \arg \max_e p(e|f) \]

Source:
A Mathematical Theory of Communication
By C. E. SHANNON

The math:
Here
With partners: Given this scenario, identify the following

\[ \hat{e} = \arg \max_e p(e|f) \]
Early foundational machine translation technology

Predating Georgetown-IBM experiment

- 1949 - Warren Weaver’s proposals in “Translation memorandum” - recasting machine translation as a cryptography problem, urges researchers to use computers. Word to word translations will not do, future systems would need to use context.
Georgetown IBM experiment -1954

Epic fail

- Russian to English demonstration
- Contained 6 grammatical rules, 250 lexical items
- Used carefully curated passages
- Well publicized but ultimately brittle
- By 1966 machine translation was essentially dead, funding had dried up
I. History of language technologies

Eliza is a mock Rogerian psychotherapist. The original program was described by Joseph Weizenbaum in 1966. This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you?
YOU: Men are all alike.
ELIZA: What is the connection, do you suppose?
YOU: They're always bugging us about something or other.
ELIZA: Can you think of a specific example?
YOU: Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here?
YOU: He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU: It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy?
YOU:
I. History of language technologies

• ELIZA - 1964 - MIT, created by Joseph Weizenbaum
  • First known chatbot
  • Used pattern matching and substitution
  • Weizenbaum was surprised to see how much people anthropomorphized ELIZA
  • Original source code was found in MI archives, many online versions exist:
    https://psych.fullerton.edu/mbirnbaum/psych101/eliza.htm
  • Let’s try! : https://shorturl.at/fLOW3
The rise of empirical natural language processing (80s - 90s)

- “These methods employ learning techniques to automatically extract linguistic knowledge from natural language corpora rather than require the system developer to manually encode the requisite knowledge”

- An Overview of Empirical Natural Language Processing Eric Brill and Raymond J. Mooney. 1999
The rise of empirical natural language processing (80s -90s)

- Key contributions
  - Language modeling- assigning probabilities to sentences
  - Applying machine learning via supervised learning to form language models
  - Modeling sentences as sequences (Hidden Markov Models)
  - Deriving “features” (n-grams, parts of speech, etc.) to translate, predict text, etc.
2000s - early 2010s

The neural network revolution

- Letting the models learn the features from data, only now we need lots of data

- Imagine - 2012 - Exploiting advances in hardware, ie GPUs

- Seq2seq and encoder-decoder models

- 2017 - A mathematical language for context in translation and token prediction (Attention) rises, and eventually becomes transformers
The rise of Transformers and LLMs - 2018-now

- LLMs begin to show nearly state-of-the-art capabilities in translation
- Transformers and next symbol prediction as a flexible and powerful general purpose computing paradigm
II. Anatomy of modern Large Language Models and Transformers-based MT technologies

- Older models: rule-based
- Empirical, statistical - some rules, also learn from data
- Neural - learn solely from data, rules (ie syntax) emerge with enough data - but what does this “data” look like?
II. Anatomy of modern Large Language Models and Transformers-based MT technologies

Where does the data come from?

Gold standard example: Parallel Corpus made by experts

United Nations Parallel Corpus

Introduction

The United Nations Parallel Corpus v1.0 is composed of official records and other parliamentary documents of the United Nations that are in the public domain. These documents are mostly available in the six official languages of the United Nations. The current version of the corpus contains content that was produced and manually translated between 1990 and 2014, including sentence-level alignments.
II. Anatomy of modern Large Language Models and Transformers-based MT technologies

### Corpus statistics

Statistics for pair-wise aligned documents:

<table>
<thead>
<tr>
<th></th>
<th>ar</th>
<th>en</th>
<th>es</th>
<th>fr</th>
<th>ru</th>
<th>zh</th>
</tr>
</thead>
<tbody>
<tr>
<td>ar</td>
<td>–</td>
<td>111,241</td>
<td>113,065</td>
<td>112,605</td>
<td>111,896</td>
<td>91,345</td>
</tr>
<tr>
<td></td>
<td></td>
<td>18,539,207</td>
<td>18,578,118</td>
<td>18,281,635</td>
<td>18,863,363</td>
<td>15,595,948</td>
</tr>
<tr>
<td>en</td>
<td>456,552,223</td>
<td>–</td>
<td>123,844</td>
<td>149,741</td>
<td>133,089</td>
<td>91,028</td>
</tr>
<tr>
<td></td>
<td>512,087,009</td>
<td></td>
<td>21,911,121</td>
<td>25,805,088</td>
<td>23,239,280</td>
<td>15,866,041</td>
</tr>
<tr>
<td>es</td>
<td>459,363,823</td>
<td>590,672,799</td>
<td>–</td>
<td>125,098</td>
<td>115,921</td>
<td>91,704</td>
</tr>
<tr>
<td></td>
<td>593,671,507</td>
<td>678,778,068</td>
<td></td>
<td>21,915,504</td>
<td>19,993,922</td>
<td>15,428,381</td>
</tr>
<tr>
<td>fr</td>
<td>452,833,187</td>
<td>668,518,779</td>
<td>674,477,239</td>
<td>–</td>
<td>133,510</td>
<td>91,613</td>
</tr>
<tr>
<td></td>
<td>597,651,233</td>
<td>782,912,487</td>
<td>688,418,806</td>
<td></td>
<td>22,381,416</td>
<td>15,206,689</td>
</tr>
<tr>
<td>ru</td>
<td>462,021,954</td>
<td>601,002,317</td>
<td>623,230,646</td>
<td>691,062,370</td>
<td>–</td>
<td>92,337</td>
</tr>
<tr>
<td></td>
<td>491,166,055</td>
<td>569,888,234</td>
<td>513,100,827</td>
<td>557,143,420</td>
<td></td>
<td>16,038,721</td>
</tr>
<tr>
<td></td>
<td>387,931,939</td>
<td>381,371,583</td>
<td>382,052,741</td>
<td>377,884,885</td>
<td>392,372,764</td>
<td></td>
</tr>
</tbody>
</table>
II. Anatomy of modern Large Language Models and Transformers-based MT technologies

Not as good but still pretty decent: exploiting crowdsourced data - mBERT and Wikipedia

- Top 104 languages with the largest Wikipedia usage, using a masked language modeling (MLM) objective
- Some semblance of curation and parallel semantics across articles

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, 2018
What does this mean for machine translation?

## Tiers of resources multilingual models

<table>
<thead>
<tr>
<th>WikiSize</th>
<th>Languages</th>
<th># Languages</th>
<th>Size Range (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>io, pms, scn, yo</td>
<td>4</td>
<td>[0.006, 0.011]</td>
</tr>
<tr>
<td>4</td>
<td>cv, lmo, mg, min, su, vo</td>
<td>6</td>
<td>[0.011, 0.022]</td>
</tr>
<tr>
<td>5</td>
<td>an, bar, br, ce, fy, ga, gu, is, jv, ky, lb, mm, my, nds, ne, pa, pnb, sw, tg</td>
<td>19</td>
<td>[0.022, 0.044]</td>
</tr>
<tr>
<td>6</td>
<td>af, ba, cy, kn, la, mr, oc, sco, sq, tl, tt, uz</td>
<td>12</td>
<td>[0.044, 0.088]</td>
</tr>
<tr>
<td>7</td>
<td>az, bn, bs, eu, hi, ka, kk, lt, lv, mk, ml, nn, ta, te, ur</td>
<td>15</td>
<td>[0.088, 0.177]</td>
</tr>
<tr>
<td>8</td>
<td>ast, be, bg, da, el, et, gl, hr, hy, ms, sh, sk, sl, th, war</td>
<td>15</td>
<td>[0.177, 0.354]</td>
</tr>
<tr>
<td>9</td>
<td>fa, fi, he, id, ko, no, ro, sr, tr, vi</td>
<td>10</td>
<td>[0.354, 0.707]</td>
</tr>
<tr>
<td>10</td>
<td>ar, ca, cs, hu, nl, sv, uk</td>
<td>7</td>
<td>[0.707, 1.414]</td>
</tr>
<tr>
<td>11</td>
<td>ceb, it, ja, pl, pt, zh</td>
<td>6</td>
<td>[1.414, 2.828]</td>
</tr>
<tr>
<td>12</td>
<td>de, es, fr, ru</td>
<td>4</td>
<td>[2.828, 5.657]</td>
</tr>
<tr>
<td>14</td>
<td>en</td>
<td>1</td>
<td>[11.314, 22.627]</td>
</tr>
</tbody>
</table>

Table 1: List of 99 languages we consider in mBERT and its pretraining corpus size. Languages in **bold** are the languages we consider in §5.
II. Anatomy of modern Large Language Models and Transformers-based MT technologies

Dear God what have we done: LLMs trained using masked language modeling (MLM) over internet level corpora, translation learned along the way from all manner of cursed sources
What differences between source and target languages effects MT quality?

- Linguistic typology
- Lexical divergence
- Morphological typology
- Referential density - ex: pro-drop languages
Extreme *she* occupations

1. homemaker  
2. nurse  
3. receptionist  
4. librarian  
5. socialite  
6. hairdresser  
7. nanny  
8. bookkeeper  
9. stylist  
10. housekeeper  
11. interior designer  
12. guidance counselor

Extreme *he* occupations

1. maestro  
2. skipper  
3. protege  
4. philosopher  
5. captain  
6. architect  
7. financier  
8. warrior  
9. broadcaster  
10. magician  
11. fighter pilot  
12. boss

Figure 1: The most extreme occupations as projected on to the *she–he* gender direction on g2vNEWS. Occupations such as *businesswoman*, where gender is suggested by the orthography, were excluded.

Attention is all you need

Foundational paper behind machine translation and LLMs like chatGPT - utilizing the MLM objective

---

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* †
illia.polosukhin@gmail.com
What is “attention”

A bunch of matrix algebra, but it’s more intuitive than that

\[
\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O
\]

where \( \text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \)

\[
\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V
\]

Where the projections are parameter matrices \( W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}, W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}, W_i^V \in \mathbb{R}^{d_{\text{model}} \times d} \) and \( W^O \in \mathbb{R}^{h\times d_{\text{model}}} \).
Take two minutes and fill in the blanks with the first word that comes to mind. Put an asterisk(s) next to the part or parts of the sentence that most influenced your prediction.

1. I travelled to the beach by ________.

2. I ____ to the store.

3. We _______ to the national park.

4. Did you try the ________ of the day at Cheese Board?

If you are attending remotely, write out your answers on a piece of paper. When prompted, put your answer in the chat!
A. Take a guess at the probability \([0,1]\) of your prediction occurring on the internet.
B. What other words in the sentence made you make the predictions that you did?

1. I travelled to the beach by ________.

2. I ____ to the grocery store.

3. We ________ to the national park.

4. Did you try the ________ of the day at Cheese Board?
With your elbow partners: fill in the blanks for the translation
Please use Google translate if you aren’t conversant in Spanish!

English: They went to the house to feed themselves.

Spanish: ________ a la casa para ________.
"Reducing gender bias in Google Translate". Blog post. James Kuczmarski

1. I travelled to the beach by __________.

2. I ____ to the grocery store.

3. We _______ to the national park.

4. Did you try the ________ of the day at Cheese Board?
How does the model get a sense of what to predict?

“Masked” training

Sentence A:
[MASK] you try the pizza of the day?
Did [MASK] try the pizza of the day?
Did you [MASK] the pizza of the day?

Sentence B initial predictions:
[Did] you try the pizza of the day?
Did [they] try the pizza of the day?
Did you [eat] the pizza of the day?

The “teacher” answer:
Did you try the pizza of the day?
Prediction architecture: Transformers in machine translation

Figure 10.5 The encoder-decoder transformer architecture for machine translation. The encoder uses the transformer blocks we saw in Chapter 9, while the decoder uses a more powerful block with an extra cross-attention layer that can attend to all the encoder words. We’ll see this in more detail in the next section.
¿Todavía están en casa?

Lightly trained neural network English translation:
Are you still at home?
Transformers/Attention: Not just for symbols

Which part of the picture leads to the prediction?
Can you explain to me how RL with human feedback works for language models like I am five?

Sure! So, imagine you have a robot named Rufus who wants to learn how to talk like a human. Rufus has a language model that helps him understand words and sentences.

First, Rufus will say something using his language model. For example, he might say "I am a robot."

Then, a human will listen to what Rufus said and give him feedback on whether it sounded like a natural sentence a human would say. The human might say, "That's not quite right, Rufus. Humans don't usually say 'I am a robot.' They might say 'I'm a robot' or 'I am a machine.'"

Rufus will take this feedback and use it to update his language model. He will try to say the sentence again, using the new information he received from the human. This time, he might say "I'm a robot."

The human will listen again and give Rufus more feedback. This process will continue until Rufus can say sentences that sound natural to a human.

Over time, Rufus will learn how to talk like a human thanks to the feedback he receives from humans. This is how language models can be improved using RL with human feedback.
What is Reinforcement Learning with Human Feedback? RLHF

Or, how ChatGPT made a BIG usability and performance leap

Source: Hugging face blog
Who are the humans in Human Feedback?

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

Published on January 6, 2022. In Endless Origins

Is AI fast becoming a technology built on worker exploitation from Global South?

While working with crowd work platforms for datasets, it is essential to consider annotator subjectivity as it has the capability to make the data set of extremely high or low quality, which in turn affects the whole ML model.

Source: Analytics India magazine
What happens when AI trains AI?

- Google Translate - digital watermark
- OpenAI - not so much.
III. Implications of strategies for use of language technologies in
THE PUBLIC

CHATGPT

IS THIS... ARTIFICIAL GENERAL INTELLIGENCE?
What is “emergence”? Is This AGI?

- AGI : Artificial General Intelligence
  - No, what does that even mean?
    - Physical symbol systems are powerful systems for representing the world and predicting symbols is all LLMs can do

- Emergence: Train a model for one thing, and it learns something else on the way. Is this AGI?
François Chollet @fchollet

Every self-supervised system ever developed has displayed emergent properties of some kind.

When Mikolov trained word vectors in 2013, he noticed that some vectors in the resulting space encoded useful semantic transformations, like "plural(x)" or "gender(x)". That's emergence.

François Chollet @fchollet · Aug 10

He had not designed or trained the system to do this. It was picked up as a by-product of learning word co-occurrences. Learn to organize words in a vector space, and your space will turn out to have interesting properties you did not anticipate.

François Chollet @fchollet · Aug 10

Emergent learning happens because information space is not random. It's highly organized. So that if you learn one aspect of its organization, you will pick up other aspects as a by-product.

It's especially prevalent with self-supervised learning.
Resource levels in CommonCrawl, and thus most LLMs

<table>
<thead>
<tr>
<th>Language</th>
<th>Code</th>
<th>Pop. (M)</th>
<th>CC Size (%)</th>
<th>Cat.</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>en</td>
<td>1,452</td>
<td>45.8786</td>
<td>H</td>
</tr>
<tr>
<td>Russian</td>
<td>ru</td>
<td>258</td>
<td>5.9692</td>
<td>H</td>
</tr>
<tr>
<td>German</td>
<td>de</td>
<td>134</td>
<td>5.8811</td>
<td>H</td>
</tr>
<tr>
<td>Chinese</td>
<td>zh</td>
<td>1,118</td>
<td>4.8747</td>
<td>H</td>
</tr>
<tr>
<td>Japanese</td>
<td>jp</td>
<td>125</td>
<td>4.7884</td>
<td>H</td>
</tr>
<tr>
<td>French</td>
<td>fr</td>
<td>274</td>
<td>4.7254</td>
<td>H</td>
</tr>
<tr>
<td>Spanish</td>
<td>es</td>
<td>548</td>
<td>4.4690</td>
<td>H</td>
</tr>
<tr>
<td>Italian</td>
<td>it</td>
<td>68</td>
<td>2.5712</td>
<td>H</td>
</tr>
<tr>
<td>Dutch</td>
<td>nl</td>
<td>30</td>
<td>2.0585</td>
<td>H</td>
</tr>
<tr>
<td>Polish</td>
<td>pl</td>
<td>45</td>
<td>1.6636</td>
<td>H</td>
</tr>
<tr>
<td>Portuguese</td>
<td>pt</td>
<td>257</td>
<td>1.1505</td>
<td>H</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>vi</td>
<td>85</td>
<td>1.0299</td>
<td>H</td>
</tr>
<tr>
<td>Turkish</td>
<td>tr</td>
<td>88</td>
<td>0.8439</td>
<td>M</td>
</tr>
<tr>
<td>Indonesian</td>
<td>id</td>
<td>199</td>
<td>0.7991</td>
<td>M</td>
</tr>
<tr>
<td>Swedish</td>
<td>sv</td>
<td>13</td>
<td>0.6969</td>
<td>M</td>
</tr>
<tr>
<td>Arabic</td>
<td>ar</td>
<td>274</td>
<td>0.6658</td>
<td>M</td>
</tr>
<tr>
<td>Persian</td>
<td>fa</td>
<td>130</td>
<td>0.6582</td>
<td>M</td>
</tr>
<tr>
<td>Korean</td>
<td>ko</td>
<td>81</td>
<td>0.6498</td>
<td>M</td>
</tr>
<tr>
<td>Greek</td>
<td>el</td>
<td>13</td>
<td>0.5870</td>
<td>M</td>
</tr>
<tr>
<td>Thai</td>
<td>th</td>
<td>60</td>
<td>0.4143</td>
<td>M</td>
</tr>
<tr>
<td>Ukrainian</td>
<td>uk</td>
<td>33</td>
<td>0.3304</td>
<td>M</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>bg</td>
<td>8</td>
<td>0.2900</td>
<td>M</td>
</tr>
<tr>
<td>Hindi</td>
<td>hi</td>
<td>602</td>
<td>0.1588</td>
<td>M</td>
</tr>
<tr>
<td>Bengali</td>
<td>bn</td>
<td>272</td>
<td>0.0930</td>
<td>L</td>
</tr>
<tr>
<td>Tamil</td>
<td>ta</td>
<td>86</td>
<td>0.0446</td>
<td>L</td>
</tr>
<tr>
<td>Urdu</td>
<td>ur</td>
<td>231</td>
<td>0.0274</td>
<td>L</td>
</tr>
<tr>
<td>Malayalam</td>
<td>ml</td>
<td>36</td>
<td>0.0222</td>
<td>L</td>
</tr>
<tr>
<td>Marathi</td>
<td>mr</td>
<td>99</td>
<td>0.0213</td>
<td>L</td>
</tr>
<tr>
<td>Telugu</td>
<td>te</td>
<td>95</td>
<td>0.0183</td>
<td>L</td>
</tr>
<tr>
<td>Gujarati</td>
<td>gu</td>
<td>62</td>
<td>0.0126</td>
<td>L</td>
</tr>
<tr>
<td>Burmese</td>
<td>my</td>
<td>33</td>
<td>0.0126</td>
<td>L</td>
</tr>
<tr>
<td>Kannada</td>
<td>kn</td>
<td>64</td>
<td>0.0122</td>
<td>L</td>
</tr>
<tr>
<td>Swahili</td>
<td>sw</td>
<td>71</td>
<td>0.0077</td>
<td>X</td>
</tr>
<tr>
<td>Punjabi</td>
<td>pa</td>
<td>113</td>
<td>0.0061</td>
<td>X</td>
</tr>
<tr>
<td>Kyrgyz</td>
<td>ky</td>
<td>5</td>
<td>0.0049</td>
<td>X</td>
</tr>
<tr>
<td>Odia</td>
<td>or</td>
<td>39</td>
<td>0.0044</td>
<td>X</td>
</tr>
<tr>
<td>Assamese</td>
<td>as</td>
<td>15</td>
<td>0.0025</td>
<td>X</td>
</tr>
</tbody>
</table>

Table 1: List of languages, language codes, numbers of first and second speakers, data ratios in the CommonCrawl corpus, and language categories. The languages are grouped into categories based on their data ratios in the CommonCrawl corpus: High Resource (H, > 1%), Medium Resource (M, > 0.1%), and Low Resource (L, > 0.01%), and Extremely-Low Resource (X, < 0.01%).

For more info on linguistic diversity in NLP

The State and Fate of Linguistic Diversity and Inclusion in the NLP World

Pratik Joshi*  Sebastín Santy*  Amar Budhiraja*
Kalika Bali  Monojit Choudhury
Microsoft Research, India
{t-prjos, t-sesan, amar.budhiraja, kalikab, monojitc}@microsoft.com
High resource v low resource, alphabet similarity

- GPT-4, the latest LLM from OpenAI, excels at English, Spanish, Italian, Indonesian and other Latin alphabet-based languages, but it struggles with Thai, Punjabi and other languages based on different alphabets. Baidu's Ernie Bot is best with Chinese, which it was trained on.

Zero shot:

Translate the following from English to French: donkey

The word "donkey" is translated to French as "âne."

Few shot:

Translate the following from English to French: donkey -> âne. chicken ->

The word "chicken" is translated to French as "poulet."

* Only trained to predict the next tokens given the prompt!
Language models have an element of randomness, are constantly updated

- LMs are not generally not deterministic, but some widely available ones let you toggle this! But making it deterministic takes some of the power away.
- Most sample from a distribution $p(\text{next_token} | \text{previous_tokens})$
- What works in one session might not work in another, or at a later date
Large Language Models Are State-of-the-Art Evaluators of Translation Quality

Tom Kocmi and Christian Federmann
Microsoft, One Microsoft Way, Redmond, WA-98052, USA
{tomkocmi, chrife}@microsoft.com

Abstract

We describe GEMBA, a GPT-based metric for assessment of translation quality, which works both with a reference translation and without. In our evaluation, we focus on zero-shot prompting, comparing four prompt variants in two modes, based on the availability of the reference. We investigate nine versions of GPT models, including ChatGPT and GPT-4. We show that our method for translation quality assessment only works with GPT 3.5 and larger models. Comparing to results from WMT22’s Metrics shared task, our method achieves state-of-the-art accuracy in both modes when compared to MQM-based human labels. Our results are valid on the system level for all three WMT22 Metrics shared task language pairs, namely English into German, English into Russian, and Chinese into English. This provides a first glimpse into the usefulness of pre-trained, generative large language models for quality assessment of translations. We publicly release all our code and prompt templates used for the experiments described in this work, as well as all corresponding scoring results, to allow for external validation and reproducibility.1
Implications?

Sarah Chieng
@SarahChieng

At MIT, our teachers required that we use chatGPT.

ex: our Chinese teacher told us to use chatGPT to correct our grammar before posting our essays for other students to read. She didn't want any incorrect grammar to confuse other students.

Why are schools still banning chatGPT?!?

Last edited 10:19 AM · Sep 13, 2023 · 76.8K Views

While preliminary results indicate that the GEMBA metric performs very well when compared to other automated metrics evaluated as part of the WMT22 Metrics shared task, it is important to note that these results are based on human labels for only three language pairs. We expect that the metrics performance may suffer for other language pairs, mainly under-resourced languages similar to Hendy et al. (2023) who show low translation quality for such languages.
### Build your own attention maps

**Instructions:**
Shade in the parts of the matrix where the source language influences the target language. Justify each row with a short phrase explaining why.

<table>
<thead>
<tr>
<th>The</th>
<th>agreement</th>
<th>on</th>
<th>the</th>
<th>European</th>
<th>Economic</th>
<th>Area</th>
<th>was</th>
<th>signed</th>
<th>in</th>
<th>August</th>
<th>1992</th>
<th>&lt;end&gt;</th>
</tr>
</thead>
</table>
Q: When the output is tasked with rewriting in hiragana what are some big (sentence-level, syntactical, etc.) and small (word-level) errors that you see? (Hint: Ramen is not from Japan!!)

Critiquing output - plenty of room for critiques!

Translate this into Japanese (romanized English). "While the noodles at Ramen Shop are fine, I dislike the long waits and feel that the traditional ramen styles like tonkatsu or shio are done much better elsewhere"

The sentence "While the noodles at Ramen Shop are fine, I dislike the long waits and feel that the traditional ramen styles like tonkatsu or shio are done much better elsewhere" can be translated into romanized Japanese as:

"Ramenya no men wa maamaa desu ga, nagai matsuri ga kirai de, tonkotsu ya shio no dentō-teki na ramen sutaizu wa hoka no tokoro de yoku dekiteiru to omoimasu."

Can you do this in hiragana

Certainly, here is the sentence in hiragana:

"らーめんやの麺はまあまあですが、長い待ち時間が嫌いで、とんこつや塩の伝統的なラーメンスタイルは他のところでよくできていると思います。"
First draft -> ChatGPT evaluation. Revise, resubmit.

• Assignment: Write a short paragraph detailing your plans for the weekend. Submit your paragraph to a language model, and prompt the language model to identify (but not fix!) any errors. Rewrite and submit the rough draft, comments, and revised version.
Workshop tomorrow

10 AM - 12 PM: B-21 Dwinelle

• Exploring 2,048 different translation models in Huggingface using Google Colab
• Prompting best practices
• If-then collective composition assignments - adapting for language classes
Workshop

Agenda

- Two models: Bard and ChatGPT
  - Prompting best practices
    - Chain of Thought, revision
    - API Playground and temperature
    - Notes on context windows
- Multimodal capabilities
- Advanced data analysis plugin
- Hugging face transformers in Colab
- Executing Python code in Colab
Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis

Wenhan Zhu$^{1,2}$, Hongyi Liu$^3$, Qingxiu Dong$^4$, Jingjing Xu$^2$
Shujian Huang$^1$, Lingpeng Kong$^{2,5}$, Jiajun Chen$^1$, Lei Li$^6$

$^1$ National Key Laboratory for Novel Software Technology, Nanjing University
$^2$ Shanghai AI Lab
$^3$ Shanghai Jiao Tong University
$^4$ Peking University
$^5$ The University of Hong Kong
$^6$ University of California, Santa Barbara

zhuwh@smail.nju.edu.cn, liu.hong.yi@sjtu.edu.cn, dqx@stu.pku.edu.cn, jingjingxu@pku.edu.cn
huangsj@nju.edu.cn, lpk@cs.hku.hk, chenjj@nju.edu.cn, lillei@cs.ucsb.edu

https://huggingface.co/tasks/translation
Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei  Xuezhi Wang  Dale Schuurmans  Maarten Bosma
Brian Ichter  Fei Xia  Ed H. Chi  Quoc V. Le  Denny Zhou

Google Research, Brain Team
{JasonWei, Dennyzhou}@google.com

**Standard Prompting**

**Model Input**

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The answer is 27. ✗

**Chain-of-Thought Prompting**

**Model Input**

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls, 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓