Evaluating Code-Switching Translation with LLMs

LING 351 – Language Technology and LLMs

Paper Presentation by Sam Moll

Table of Contents

Paper Overview, Background, Research Question

INTRODUCTION

Paper Overview

TITLE

Evaluating Code-Switching Translation with Large Language Models

AUTHORS

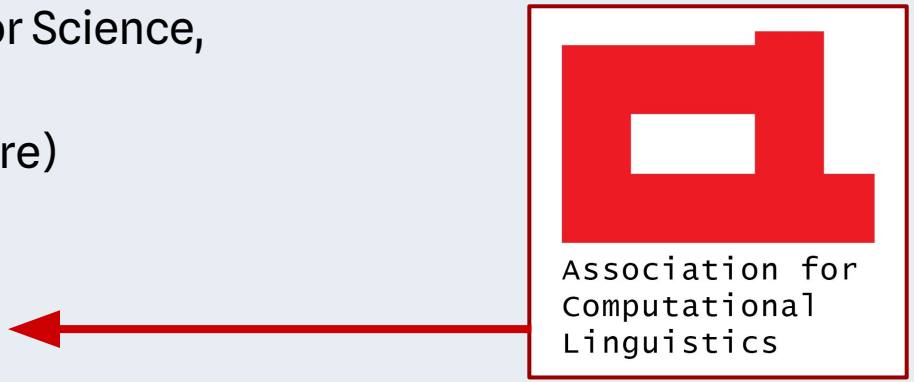
Muhammad Huzaifah, Weihua Zheng, Nattapol Chanpaisit, Kui Wu

FROM

Institute for Infocomm Research, Agency for Science, Technology and Research (Singapore),
Nanyang Technological University (Singapore)

PUBLICATION

Published in ACL Anthology in **May 2024**



What is **Code-Switching?**

Code-switching is the alternation of multiple of multiple languages in a single utterance.

- Utterance → Unit of speech
  - Stretch of spoken language that is preceded by silence and followed by silence or a change of speaker

- Common phenomenon in multilingual communication
- Prevalent online (social media)

Why is **Evaluating Code-Switching Translation** with LLMs Important?

Properly translating code-switching means preserving the original **meaning, nuance, and tone** of a message.

- Code-switching remains a challenge for neural machine translation (NMT) models
- Why might large language models (LLMs) be better suited for code-switching translation?

NMT Models v.s. LLMs - Parallel Data

NMT Models

- Neural Machine Translation
- Designed for monolingual text
- Reliant on huge amounts of parallel data

LLMs

- Large Language Models
- Trained for language modelling in which parallel data unnecessary

Parallel Data consists of data sets of translation pairs with both sentences and their translations.

- Scarce for code-switched text

Previous Research on LLM Translation

LLMs shown to

- Improve performance across variety of NLP problems (Gao et al., 2021)

Gao, Leo et al. (2025). A framework for few-shot language model evaluation. *Zenodo*. https://doi.org/10.5281/zenodo.16737642

- Provides a framework to test generative language models on a large number of different evaluation tasks

Previous Research on LLM Translation

LLMs shown to

- Improve performance across variety of NLP problems (Gao et al., 2021)
- Improve quickly with new iterations & releases (Brown et al., 2020)

Brown, Tom et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–190)

- Scaling up language models greatly improve performance

Mixed results on LLM use for general translation (Zhu et al., 2023)

- Competitive with NMT models on high-resource languages

High-resource languages → Languages with vast amount of available and accessible corpora and training data

Zhu, Wenhao et al. (2023). Multilingual machine translation with large language models: Empirical results and analysis. a*rXiv preprint arXiv:2304.04675.*

- Investigate the advantages and challenges of LLMs for multilingual machine translation (MMT)

Mixed results on LLM use for general translation (Zhu et al., 2023)

- Competitive with NMT models on high-resource languages
- Behind other NMT models on low-resource languages

Low-resource languages → Languages with sparse amount of available and accessible corpora and training data

Zhu, Wenhao et al. (2023). Multilingual machine translation with large language models: Empirical results and analysis. a*rXiv preprint arXiv:2304.04675.*

- Investigate the advantages and challenges of LLMs for multilingual machine translation (MMT)

Mixed results on LLM use for general translation (Zhu et al., 2023)

- Competitive with NMT models on high-resource languages
- Behind other NMT models on low-resource languages
- Can acquire translation ability in a resource efficient way
    - Indicates promising future in multilingual machine translation

Zhu, Wenhao et al. (2023). Multilingual machine translation with large language models: Empirical results and analysis. a*rXiv preprint arXiv:2304.04675.*

- Investigate the advantages and challenges of LLMs for multilingual machine translation (MMT)

Research Question

How effectively can large language models (LLMs) translate code-switched text compared to neural machine translation (NMT) models and commercial engines?

04

LLMs, NMT Models, Datasets, Evaluation, Prompting

METHODOLOGY

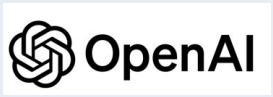
Setup: Large Language Models

Definition of LLM

- Language model trained on vast amounts of data in order to understand, generate, and manipulate language human language

Chose 6 well-benchmarked LLMs to give comprehensive representation of current* state-of-the-art. (*June-Sep 2023)

- Evaluated only biggest available version for each



GPT-4

- Latest LLM offered by OpenAI *(at the time of the paper)*
- Remarkable capabilities in general translation
- Directly accessed through ChatGPT interface

GPT-3.5

- OpenAI model that previously powered ChatGPT
- Responsible for explosion of interest in general public
- Previously leading LLM in many NLP benchmarks

Bard

- Conversational chatbot released by Google
- Based on PaLM-2 model
  - Trained on large corpora of multilingual text
  - Claimed to excel at multilingual tasks (translation)

LLaMA-2

- Family of LLMs released by Meta
- Further improved popular predecessor (LLaMA)
  - One of first open-source LLMs trained at scale
  - Well adopted by research community

Falcon

- Developed by the Technology Innovation Institute (TII)
- 40B variant
- Best performing LLM on Open LLM leaderboard at release

Phoenix

- Developed by Arise AI
- Focuses on multilingual performance
  - Chinese & non-Latin languages
- Uses BLOOMZ as base model
  - Fine-tuned on multilingual conversation data

Setup: Neural Machine Translation Models

Definition of NMT Model

- Language model that uses a large, single neural network to translate sentences from a source language to a target language
- Encodes meaning of input into word embeddings (vectors)
- Dominant approach to machine translation

Compared chosen LLMs against 3 NMT systems commonly used for translation.

- Since not built for code-switched translation, have to specify a matrix (dominant) language (minor languages called embedded language)

Google Translate

- Leading commercial machine translation engine developed by Google
- Supports 243 languages

DeepL

- Leading commercial machine translation engine
- Supports 36 languages

NLLB

- “No Language Left Behind”
- Massive multilingual translation model developed by Meta AI
- Trained on data using data mining techniques tailored for low-resource languages
- Supports 200 languages (75 minority languages)

Setup: Datasets

- Lack of high-quality parallel code-switched data
  - Corpora containing both code-switching & translations
- Prior datasets using speech and social media data
  - Where code-switching most common
- Supplement with synthetic data generated from Flores-200
- Only consider code-switching sources made up of two languages (one dominant, one minor) to a single target language

Speech & Social Media Corpora

3 open-source datasets used

Hindi-English (HI-EN)

- Translated to English
- From LinCE (Linguistic Code-switching Evaluation) dataset
- Spoken data
- 892 lines

Spanish-English (SP-EN)

- Translated to English
- From Bangor Miami speech dataset
- Spoken data
- 3204 lines

Indonesian-English (ID-EN)

- Translated to Indonesian
- Derived from Twitter/X posts
- Dominant language overlaps
- 815 lines

Synthetic Data: Flores-200

- Constructed pseudo-code-switching data from Flores-200 dataset

Flores-200

- "Facebook Low Resource Languages Evaluation Sets"
- Multilingual translation dataset
- Supports 204 languages
- Designed to support research for low-resource languages

Algorithm 1: Synthetic code-switching data generation

Data: bwa, ms, es, pt, pos, ner
Result: Code-switched sentence

```
begin
    for each Name Entity e in ner do
        if translation of e exists in es then
            Replace e in ms with its
              translation from es;
    for each node n in pt do
        if node n is switchable according to
          Matrix Language Theory then
            Set switch_label(n) to True;
        else
            Set switch_label(n) to False;
    for each node n with switch_label(n) as
      True do
        if lexicality of n is not in {noun,
          adjective, verb} based on pos then
            Set switch_label(n) to False;
    for each node n with switch_label(n) as
      True do
        if translation of node's word exists in
          bwa and is in es then
            Replace word of node n in ms
              with its translation from es;
        else
            Continue without replacement;
    return Modified ms as code-switched
      sentences;
```

Synthetic Data: Flores-200

- English-Chinese (EN-ZH) to Chinese (ZH)
- German-Turkish (DE-TR) to English (EN)
- French-Italian (FR-IT) to Japanese (JA)
- Tamil-English (TA-EN) to Czech (CS)

Algorithm 1: Synthetic code-switching data generation

Data: bwa, ms, es, pt, pos, ner
Result: Code-switched sentence

```
begin
    for each Name Entity e in ner do
        if translation of e exists in es then
            Replace e in ms with its translation from es;
    for each node n in pt do
        if node n is switchable according to Matrix Language Theory then
            Set switch_label(n) to True;
        else
            Set switch_label(n) to False;
    for each node n with switch_label(n) as True do
        if lexicality of n is not in {noun, adjective, verb} based on pos then
            Set switch_label(n) to False;
    for each node n with switch_label(n) as True do
        if translation of node's word exists in bwa and is in es then
            Replace word of node n in ms with its translation from es;
        else
            Continue without replacement;
    return Modified ms as code-switched sentences;
```

Setup: Evaluation Metrics

Primary metric

BLEU (Bilingual Evaluation Understudy metric)
Measures similarity between generated text and reference text by looking at overlap of word sequences (n-grams)

Complementary metrics

ChrF++ (Character n-gram F-score)
Measures similarity between generated text and reference text by looking at character sequences (n-grams)

TER (Translation Error Rate)
Calculates minimum number of edits needed to change generated text into reference text

Prompts	BLEU	ChrF++	TER
P1	**37.70**	56.18	52.53
P2	37.50	**56.22**	**51.86**
P3	36.98	55.61	54.98

Prompting Strategy

- Modified prompts used by previous study for monolingual translation
  - “Provide ten concise prompts or templates that can make you translate code-switched sentences”
- Evaluated performance with subset of 100 random lines

	Prompts
P1	`Translate the following code-switched [SRC] sentences to pure [TGT] line by line. Do not output any additional text other than the translations: \n [SRC1] \n [SRC2] ...`
P2	`Translate the following [SRC] sentences to pure [TGT] line by line. Do not output any additional text other than the translations: \n [SRC1] \n [SRC2] ...`
P3	`Please provide the [TGT] translation for these sentences line by line. Do not output any additional text other than the translations: \n [SRC1] \n [SRC2] ...`

Table 2: Candidate prompts. `\n` denotes a newline while `[SRC1]` and `[SRC2]` are source sentences.

05

Results and Discussion

FINDINGS

Relative LLM Performance

GPT-4 outperforms all other LLMS, followed by GPT-3.5.

- GPT-4 Overall more accurate
- GPT-3.5 More natural sentence structure at times
- Bard Significant variation in performance
    - Higher tendency for mistranslations
- LLaMA-2 Frequent untranslated words
- Falcon High failure rate for several languages
- Phoenix Unnatural sound

Comparative Performance

- **GPT-4** excelled in 4/7 datasets
  - HI-EN→EN SP-EN→EN ID-EN→ID DE-TR→EN
  - Better for:
    - High-resource languages
    - Translating into English

- **Google Translate** outperformed for other 3/7 datasets
  - EN-ZH→ZH FR-IN→JA TA-EN→CS

- NLLB-54B performed comparably to GPT-4 for low-resource TA-EN→CS

Model	HI-EN→EN		SP-EN→EN		ID-EN→ID		*EN-ZH→ZH*		*DE-TR→EN*		*FR-IT→JA*		*TA-EN→CS*	
	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++
GPT-4	37.8	60.4	53.9	71.2	57.3	74.1	44.9	30.2	45.4	67.6	25.9	26.1	19.1	45.2
GPT-3.5	30.5	54.9	48.6	69.1	48.7	66.6	41.1	27.1	43.1	66.5	24.4	25.0	6.8	29.5
Bard-PaLM2	23.9	42.1	45.6	61.6	28.9	49.8	44.0	30.3	43.1	66.6	24.9	24.6	15.8	38.5
LLaMA-2-70B	25.7	49.4	40.2	61.2	34.3	50.4	34.4	23.6	37.2	60.8	19.9	20.5	0.9	16.6
Falcon-40B	5.9	25.0	15.4	34.4	N/A	N/A	20.8	15.3	25.6	52.1	1.3	4.7	N/A	N/A
Phoenix-7B	7.0	30.2	31.9	51.9	28.2	40.2	39.4	27.3	17.8	40.6	6.1	9.2	1.5	16.2
Google T	28.5	51.6	49.1	69.4	54.6	70.0	47.5	35.0	27.7	50.3	26.5	25.5	22.4	48.0
DeepL T	N/A	N/A	47.6	68.1	52.7	69.1	46.4	34.6	28.0	50.6	25.4	26.2	N/A	N/A
NLLB-1.3B	8.0	30.6	46.7	67.0	53.5	69.4	28.2	19.7	32.8	55.6	15.8	19.6	15.6	40.5
NLLB-54B	10.4	29.9	47.1	66.7	54.3	68.4	28.7	20.8	34.9	57.2	16.6	19.9	18.8	43.9
Copy	5.1	28.8	27.6	42.1	49.5	65.0	12.9	10.6	2.3	20.4	0.2	1.4	0.7	4.9

Table 4: BLEU and ChrF++ across various code-switching datasets for a collection of LLMs. They are evaluated against baselines containing commercial MT engines (Google and DeepL translate) and massive multilingual MT models (NLLB). "Copy" baseline are scores between untranslated source and reference target. Synthetic datasets are italicized.

Summary of Findings

Discussion of Results
High vs Low Resource Languages

NMT models performed better on **low-resource** language pairs.

LLMs were more accurate and reliable for **high-resource** language pairs.

A language's coverage in an LLM's training data directly correlates with its overall performance in translating that language.

Discussion of Results
Degree of Code-Switching

Performance of both LLM and NMT models **greatly deteriorates as degree of code-switching increases**.

- Deterioration more extreme for NMT models
- Google Translate outperformed GPT-4 for all monolingual baselines (EN→ZH / EN→DE / DE—> EN)
- However performance significantly drops as code-switching introduced

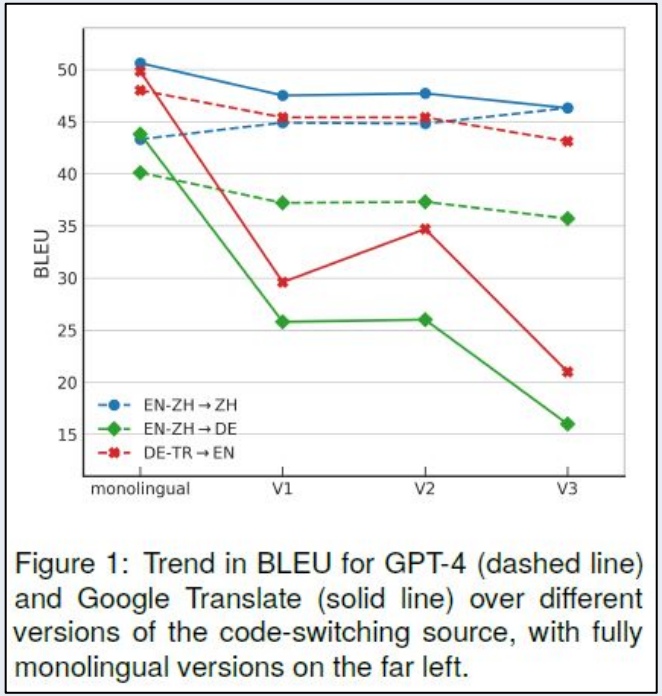


Figure 1: Trend in BLEU for GPT-4 (dashed line) and Google Translate (solid line) over different versions of the code-switching source, with fully monolingual versions on the far left.

Final Conclusion of Findings

- **GPT-4** exhibited superior performance across both high and low resource languages.
  - Other LLMs showed varying ability

- **Google Translate** and **DeepL** (NMT commercial engines) performed well on select datasets when low code-switching amount.

- **GPT-3.5** followed closely behind GPT-4 in high-resource languages, but behind NMT models for low-resource languages.

Further Improvements
Prompt Engineering

In-context learning

- Augments prompts to include demonstrations of desired task
- Shown to boost performance instead of fine-tuning the model
- Also shown to benefit monolingual translation
- Investigated different strategies

Task-related examples

- Code-switched sentences & translations in distinct languages

In-domain examples

- Sources from same type of data as test set
- Example pairs from dominant language to target language > code-switched examples

Further Improvements
Prompt Engineering

Pivot Translation

1. Translate in pivot language
   - Dominant (matrix) language for monolingual counterpart
   - High-resource language (EN)
2. Translate into target language

- Beneficial for low-resource languages & distant translation pairs
- Not as effective when target high-resource

Direction	Pivot Matrix	Pivot EN	Result BLEU	Result ChrF++	Result TER
DE-TR→EN	(direct)		45.1	67.6	**36.5**
	✓		**45.2**	**67.9**	36.7
FR-IT→JA	(direct)		25.1	27.7	62.9
	✓		26.2	27.8	63.0
		✓	**28.5**	26.2	**60.2**
	✓	✓	27.4	**29.0**	61.3
TA-EN→CS	(direct)		16.3	41.0	69.5
	✓		15.6	41.2	69.4
		✓	**17.5**	**43.2**	**66.4**
	✓	✓	16.7	41.6	71.7
EN-ZH→ZH	(direct)		44.4	28.7	42.3
	✓	✓	**45.1**	**29.0**	**41.0**

Table 6: Results for matrix and English language pivot translation strategies. Note that for EN-ZH→ZH the two strategies are equivalent.

06

Evaluation, insights, critiques

COMMENTARY

Evaluation of the Study

POSITIVES

- Compared a wide range of LLMs and NMT models
- Compared across diverse language pairs and resource levels
- Used speech and social media datasets for different code-switching conditions
- Used multiple evaluation metrics (BLEU, ChrF++, TER)
- Clearly explained how the base prompt was chosen
- Explored how prompt engineering could further improve LLM performance
- Built on previous studies to show LLMs as a promising alternative to NMT models

Evaluation of the Study

CRITIQUES

- No clear metric for determining whether a language pair is high-resource or low-resource included
- Caution with synthetic data, may not fully reflect natural code-switching
  - Understandable due to lack of code-switching parallel data

GENERAL CONSENSUS

Comprehensive study that demonstrates promising ability for GPT-4 and other LLMs to rival/surpass NMT models for code-switching translation

07

QUIZ!

Which is a result found when comparing LLMs to NMT models for code-switching translation?

A LLMs outperformed NMT models on code-switching translation for low-resource languages

B NMT models showed little to no deterioration in performance as the amount of code-switching in the text increased

C A language's coverage in an LLM's training data directly correlates with its overall performance in translating that language

D NMT models can improve on code-switching translation through prompt engineering and in-context learning

What is NOT a way in which LLM code-switching translation performance can be improved?

A Providing the LLM with examples of the code-switching translation task directly inside the prompt

B Reducing the amount of context given in the prompt by removing example translations from the prompt to avoid overwhelming the LLM

C Prompting the LLM to translate through a 3rd high-resource language before producing the final target translation

D Converting the code-switching input to its monolingual counterpart by prompting the LLM to first pivot to the dominant language

Which is a valid argument for deciding whether NMT models or LLMs are better at handling code-switching translation?

A NMT models because they're designed for translating monolingual text

B NMT models because they're adaptable to multiple languages

C LLMs because they don't require large amounts of parallel data

D LLMs because they are trained on large amounts of code-switched text

What is the difference between low-resource and high-resource languages?

A Low resource languages require less computing power to train a model on, high resource languages require more computing power

B Low resource languages have more accurate results with language models, high resource languages have less accurate results

C Low resource languages are spoken by few people in the world, high resource languages are spoken by a majority

D Low resource languages have limited data available for training language models, high resource languages have a lot of available data for training

Questions?

THANK YOU!

CREDITS: This presentation template was created by **Slidesgo**, and includes icons, infographics & images by **Freepik**

Please keep this slide for attribution