Adaptive Machine Translation with Large Language Models

Ali Salem

Agenda

1. Introduction
2. Experimental Setup
3. Adaptive MT with Fuzzy Matches
4. GPT-3 vs Encoder-Decoder MT Models
5. Incorporating Encoder-Decoder MT
6. Terminology-Constrained MT
7. ChatGPT
8. BLOOM and BLOOMZ
9. Conclusion

Introduction

- Adaptive Machine translation (MT) is a type of machine translation that utilizes feedback from users to improve the quality of the translations over time.
- There are still several challenges to effectively incorporate user feedback into the translation process, especially at inference time.
- This work aims to investigate how we can utilize in-context learning to improve real-time adaptive MT.
- By feeding an LLM at inference time with a prompt that consists of a list of translation pairs, it can then simulate the domain and style characteristics.

Adaptive MT with Fuzzy Matches

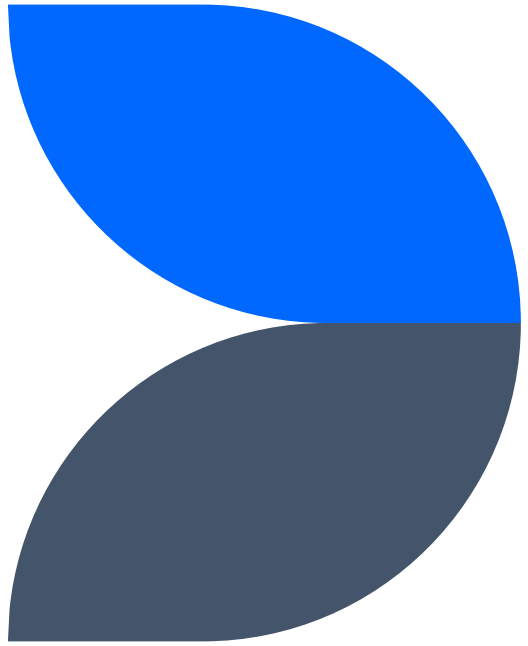
In translation environments, similar approved translated segments are usually referred to as “fuzzy matches”, and are stored in parallel datasets, known as translation memories (TMs).

In particular, the aim of this work is understanding the quality with which such models can perform the following tasks:

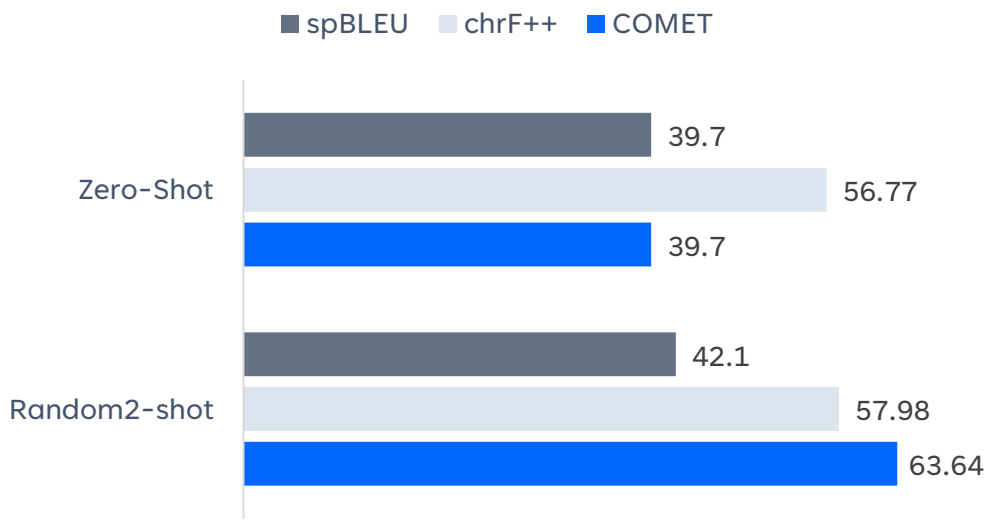
- Adapting new translations to match the terminology and style of previously approved TM fuzzy matches, at inference time;
- Matching or outperforming the quality of translations generated by encoder-decoder MT models across a number of languages;
- Fixing translations from stronger encoder MT systems using fuzzy matches, which is especially useful for low-resource languages; and
- Terminology-constrained MT, by first defining terminology in the relevant sentences or dataset, and then forcing new translations to use these terms.

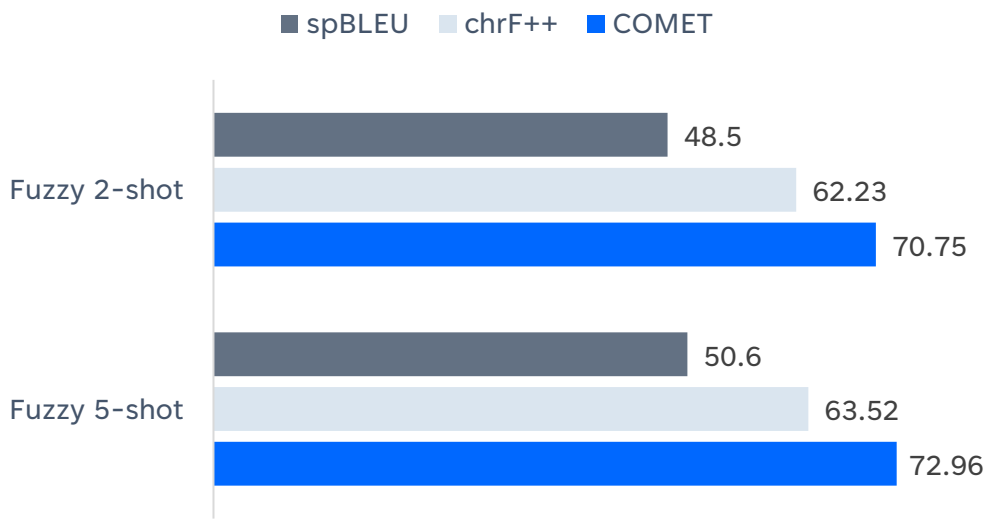
Experimental Setup

- Using GPT-3.5 text davinci- 003 model via its official API
- Using the domain specific dataset, TICO-19 which includes 3070 unique segments.
- Focusing on a range of languages with diverse scripts and amounts of resources, namely English as the source language, and Arabic, Chinese, French, Kinyarwanda, and Spanish as the target languages.



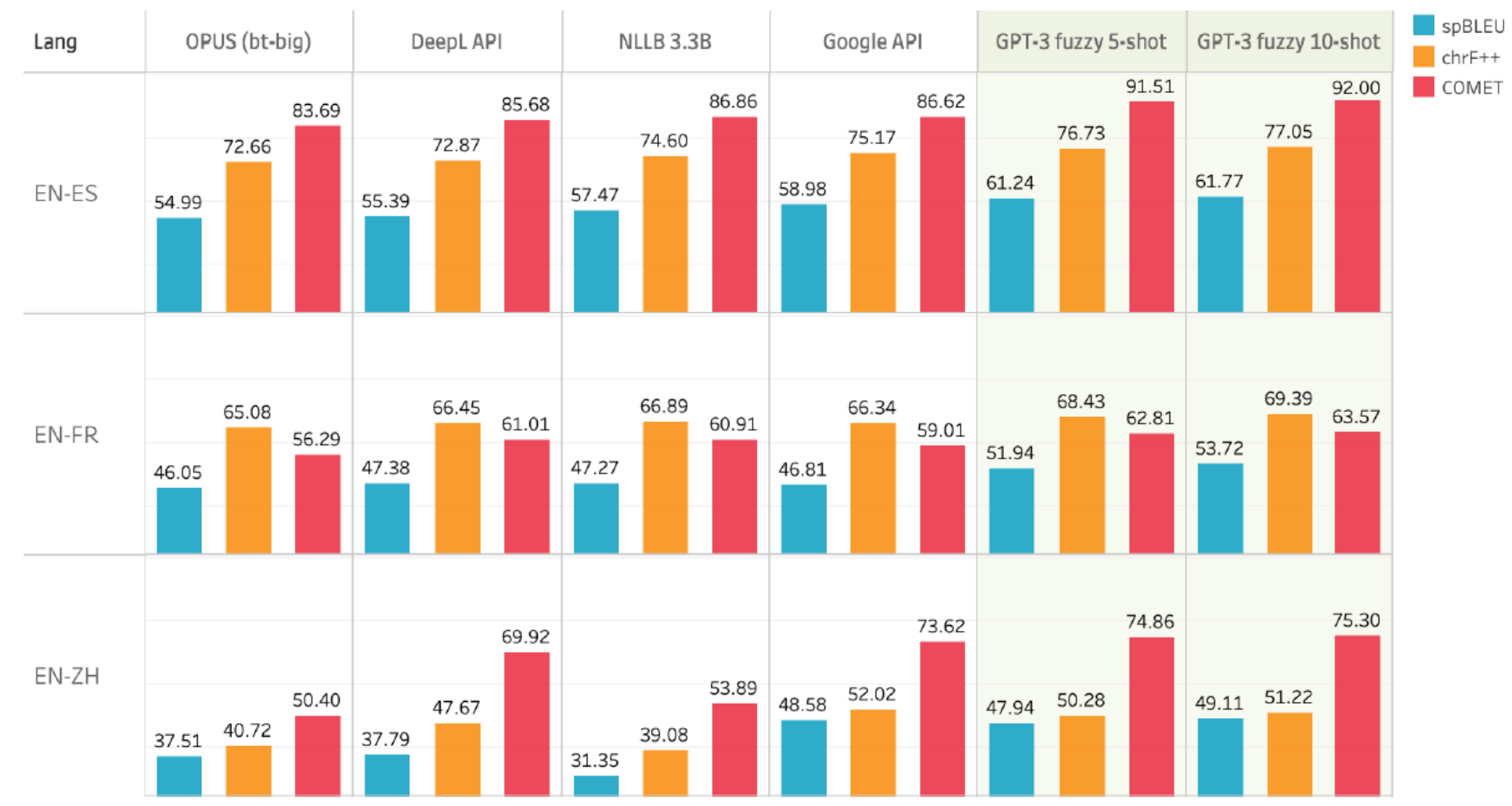
Evaluation results for GPT-3.5





Autoregressive LLMs such as GPT-3 can perform well on diverse tasks, through zero-shot, one-shot, and few-shot in-context learning without weight updates.

GPT-3 vs Encoder-Decoder MT Models



Incorporating Encoder-Decoder MT

Fuzzy matches + new segment MT

Incorporating a translation from an encoder-decoder MT model with fuzzy matches, we could achieve substantial improvements over the baseline MT performance.

Bilingual Terminology Extraction

Terminology extraction is the task of automatically defining domain-specific terms in a dataset.

Adding MT of the new segment from an encoder-decoder model to fuzzy matches, which enhanced GPT-3.5 in-context learning.

Fuzzy matches + all segments MT

Terminology-Constrained MT

We investigate three scenarios:

1. Few-shot translation with 2 fuzzy matches
2. Integrating terms from a glossary including all terms from the dataset
3. Zero-shot translation, i.e. without any fuzzy matches.

Lang	GPT-3 Context	Human Eval. ↑	Terms ↑
EN-AR	Zero-shot	2.80	0.67
	Zero-shot + glossary terms	**3.19**	**0.94**
	Fuzzy two-shot	2.89	0.80
	Fuzzy two-shot + glossary terms	**3.03**	**0.94**
EN-ES	Zero-shot	3.76	0.87
	Zero-shot + glossary terms	**3.93**	**0.96**
	Fuzzy two-shot	3.77	0.89
	Fuzzy two-shot + glossary terms	**3.84**	**0.97**
EN-FR	Zero-shot	3.55	0.89
	Zero-shot + glossary terms	**3.64**	**0.97**
	Fuzzy two-shot	3.50	0.91
	Fuzzy two-shot + glossary terms	**3.55**	**0.92**

ChatGPT

OpenAI has released new conversational models, publicly

referred to as ChatGPT. This range of models includes:

GPT-3.5 Turbo and GPT-4. While gpt- 3.5-turbo is more efficient than text-davinci-003, it shows comparable quality for both zero-shot and few-shot translation (with fuzzy matches).

The newest model gpt-4 provides better zero-shot translation quality, while the quality of few-shot translation is relatively similar to that of the two other models.

Lang	Model	Context	spBLEU ↑	chrF++ ↑	TER ↓	COMET ↑
EN-AR	GPT-3.5 Davinci	0-shot	27.6	48.36	70.6	41.28
	GPT-3.5 Turbo		38.06	56.35	61.34	62.68
	GPT-4		**40.29**	**57.86**	**59.55**	**64.25**
	GPT-3.5 Davinci	2-shot	38.41	56.57	62.31	57.36
	GPT-3.5 Turbo		46.04	62.18	55.03	73.35
	GPT-4		**47.52**	**63.28**	**53.04**	**73.7**
EN-ES	GPT-3.5 Davinci	0-shot	53.91	72.61	36.86	84.0
	GPT-3.5 Turbo		52.91	70.87	38.86	82.28
	GPT-4		**56.93**	**74.41**	**34.35**	**87.89**
	GPT-3.5 Davinci	2-shot	59.64	75.83	32.56	90.37
	GPT-3.5 Turbo		**60.35**	**76.51**	32.05	91.57
	GPT-4		60.16	**76.51**	**31.77**	**91.86**
EN-FR	GPT-3.5 Davinci	0-shot	44.87	65.29	50.34	58.67
	GPT-3.5 Turbo		46.85	66.75	48.31	61.34
	GPT-4		**47.39**	**67.14**	**48.03**	**61.93**
	GPT-3.5 Davinci	2-shot	49.79	67.41	46.79	61.38
	GPT-3.5 Turbo		**49.88**	68.33	46.27	63.62
	GPT-4		49.75	**68.38**	**45.97**	**64.04**
EN-RW	GPT-3.5 Davinci	0-shot	2.82	22.53	143.12	N/A
	GPT-3.5 Turbo		5.31	29.77	114.34	N/A
	GPT-4		**8.95**	**35.28**	**93.15**	N/A
	GPT-3.5 Davinci	2-shot	12.23	36.66	105.54	N/A
	GPT-3.5 Turbo		12.49	39.37	105.51	N/A
	GPT-4		**16.78**	**44.21**	**83.31**	N/A
EN-ZH	GPT-3.5 Davinci	0-shot	32.41	40.82	99.45	59.87
	GPT-3.5 Turbo		36.83	45.77	99.83	69.13
	GPT-4		**37.65**	**47.02**	**99.37**	**70.75**
	GPT-3.5 Davinci	2-shot	46.18	49.12	69.0	73.9
	GPT-3.5 Turbo		**45.95**	49.79	74.53	74.63
	GPT-4		45.37	**50.26**	79.29	**74.9**

BLOOM and BLOOMZ

BLOOM is a general-purpose LLM, BLOOMZ belongs to a family of models capable of following human instructions in a zero-shot manner. We found that “greedy search” achieves better results for BLOOM. When providing each system with two fuzzy matches, generally GPT-3.5 outperforms both BLOOM and BLOOMZ for most language pairs, except English-to-Arabic translation. The English to- French translation quality of BLOOM and GPT-3.5 is comparable.

Lang	System	spBLEU ↑	chrF++ ↑	TER ↓	COMET ↑
EN-AR	BLOOM fuzzy 2-shot	**43.19**	**59.48**	**57.58**	**67.36**
	BLOOMZ fuzzy 2-shot	36.29	53.33	66.86	58.4
	GPT-3 fuzzy 2-shot	38.41	56.57	62.31	57.36
EN-ES	BLOOM fuzzy 2-shot	57.67	74.25	34.86	86.48
	BLOOMZ fuzzy 2-shot	53.07	70.44	40.45	81.38
	GPT-3 fuzzy 2-shot	**59.64**	**75.83**	**32.56**	**90.37**
EN-FR	BLOOM fuzzy 2-shot	**50.52**	66.81	**46.45**	55.74
	BLOOMZ fuzzy 2-shot	45.1	62.73	51.69	47.49
	GPT-3 fuzzy 2-shot	49.79	**67.41**	46.79	**61.38**
EN-RW	BLOOM fuzzy 2-shot	10.95	31.87	91.07	N/A
	BLOOMZ fuzzy 2-shot	**12.26**	35.44	**88.36**	N/A
	GPT-3 fuzzy 2-shot	12.23	**36.66**	105.54	N/A
EN-ZH	BLOOM fuzzy 2-shot	40.62	40.62	75.24	66.23
	BLOOMZ fuzzy 2-shot	34.82	38.23	80.03	59.92
	GPT-3 fuzzy 2-shot	**46.18**	**49.12**	**69.0**	**73.9**

Conclusion

1

Conducting several experiments to assess the performance of GPT-3.5 across multiple translation tasks, namely adaptive MT using fuzzy matches, MT post-editing, terminology extraction and terminology constrained MT

2

compared its translation quality with strong encoder-decoder MT systems

3

While some high-resource languages such as English-to-French, English-to Spanish and even English-to-Chinese show excellent results, other languages have lower support

4

This means that different pipelines can be adopted in production for different language pairs, based on the level of support of these languages by an LLM.

5

In the future, we might consider starting with fine-tuning the model, and then conducting similar experiments. This can be especially beneficial for low-resource languages and rare domains efficiency.