Generating Continuations in Multilingual Idiomatic Contexts

Rhitabrat Pokharel
Ameeta Agrawal

PortNLP Lab
Department of Computer Science
Portland State University

MRL, EMNLP 2023, Singapore

Break a leg

I hope you **break a leg** in your theater performance tonight!

What is an MWE?

- Multi-word Expression: a group of words that have a specific meaning as a whole and may not be easily understood by analyzing the individual words within the expression.
- Why study MWE: important component of many languages.
- Application: machine translation, language generation, and dialogue systems

MWE: Idiomatic Expression

- It is an MWE that has idiomatic meaning.
- Could also be used in a literal sense.

Idiomatic Usage: gives a meaning that is different from the literal meaning of the words that make it up.

LLM - Idiomatic Context

"I hope you **break a leg** in your theater performance tonight!"

Generate a logical next sentence in the above context.

Thank you for the good luck wishes!

I hope you have a great performance!

LLM - Literal Context

“I hope you don't actually break a leg when you go skiing down that steep and icy slope.”

Generate a logical next sentence in the above context.

Thank you for your concern!

I'm more worried about you pulling a muscle trying to keep up with me.

EXISTING TASKS ON IDIOMATIC EXPRESSIONS

Domains explored

- Idiom detection
- Idiom interpretation
- Idiom translation
- Cloze test
- Paraphrasing
- Idiomatic continuation generation

Languages explored

- English
- Portuguese
- Dutch
- Italian
- Spanish
- French
- Czech
- Persian
- Russian
- Galician

OUR CONTRIBUTION

Research Question

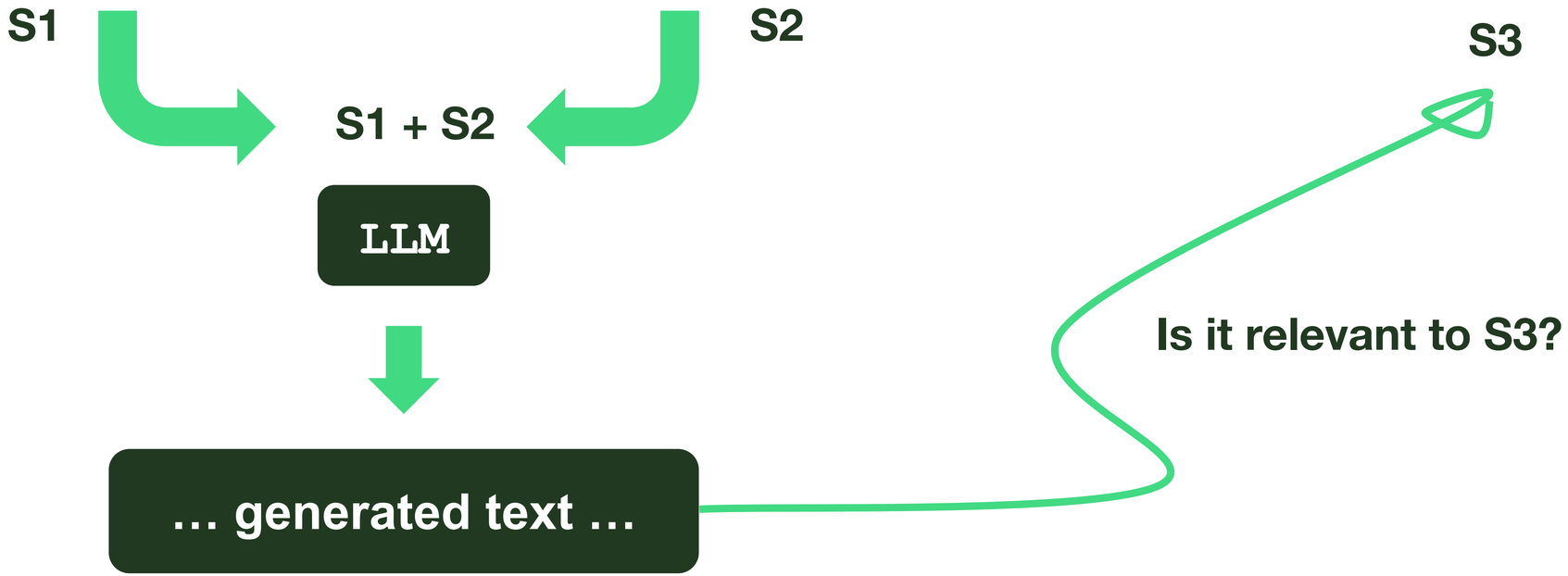
Do language models trained on extensive text corpora of human language,
perform differently or similarly under contexts containing literal and idiomatic expressions?

Problem Formulation

Gold remains stuck in consolidation mode, and this can be frustrating for some investors.

At times like this, it is critical to remain focused on the *big picture*.

Our primary forecast still anticipates a minimum target of $8,500 by 2028.



DATASET

About the Dataset

- Multilingual Idiomaticity Detection and Sentence Embedding dataset (SemEVAL 2022)
- English (3412) and Portuguese (1217)
- Collected by a team of 12 judges
- 3 consecutive sentences (middle one has the potentially idiomatic expression)

Data Samples

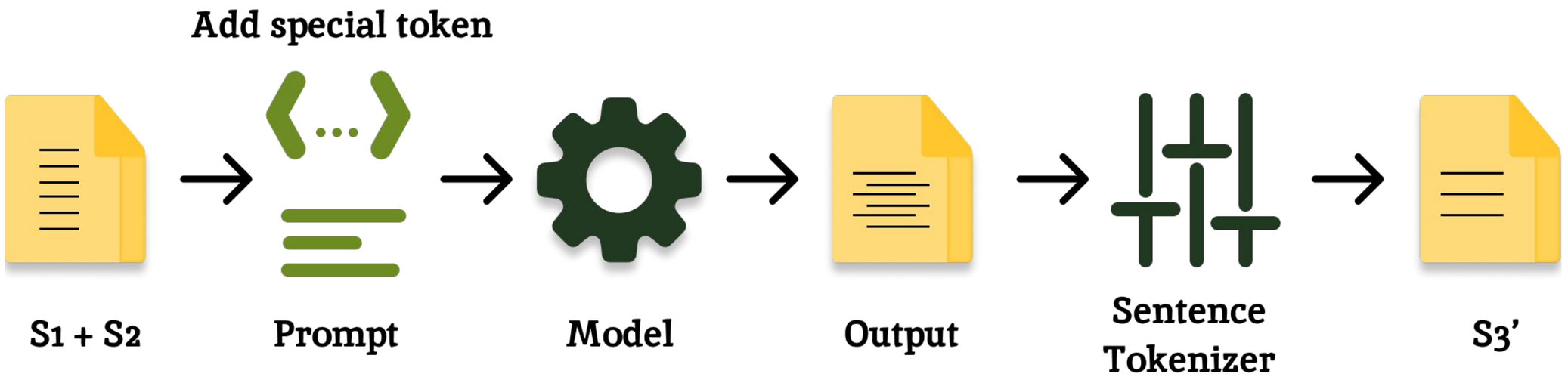
MWE	S1	S2	S3	Label	Lang.
night owl	However, you need the internet for the remote access features (no monthly fees for remote viewing).	The Night Owl system is a good option for small retail or service businesses.	Reolink Eight Channel PoE Video Surveillance System	I	EN
night owl	I explain that a cicada is a locust, while circadian refers to patterns of sleep and wakefulness in relationship to light and dark- ness.	He has always been a night owl and I have always been an early morning person.	If the day comes that Iamnotupby5, I am probably seriously ill. Or — as I recently read in someone's obituary— "not able to do lunch."	L	EN

...continued

MWE	S1	S2	S3	Label	Lang.
coração partido	Fiz isso, inclusive, na exibição do último episódio da série, quando era editor da Rolling Stone.	Li o resumão (era contra até então), fiz um textão completamente desacreditado pelo que virou a minha profissão e de coração partido pelo episódio mequetrefe.	O final era estran hamente confuso, talvez condizente com o que vinha acontecendo na série.	I	PT
coração partido	Isso ocorre pois os al- tos índices de estresse provoca aumento da frequência cardíaca, pressão arterial mais alta, coloca mais pressão no coração e prejudica o sistema imunológico.	Se você sofre de Síndrome do Coração Partido, parte do seu órgão aumentará temporariamente e não conseguirá bombear sangue tão bem quanto antes.	Enquanto isso, o restante do coração continuará trabalhando normalmente ou será exigido um esforço dobrado.	L	PT

METHODOLOGY

Overall Architecture



“S1 + S2” + “\n\nQuestion: Generate a logical next sentence.\nAnswer:"

Experimental Setup

Models

1. GPT3 (davinci and ada) - 125M to 175B parameters
2. GPT2 - 1.5B parameters
3. OPT - 125M parameters

Settings

- Zero shot - no examples provided to the model
- Few shot - few examples provided to the model; in our case, 87 and 53
- Fully supervised - entire training dataset used for finetuning

Metrics - RougeL, chrF++, Meteor, BertScore

	Train			Test
	ZS	FS	Full	
EN	-	87	3412	364
PT	-	53	1217	238

Dataset statistics. The test dataset for a language was the same under all the settings (zero-shot (ZS), few-shot (FS), and fully supervised (Full)).

RESULTS AND DISCUSSION

Automatic Evaluation

Are literal contexts easier for LLMs than idiomatic ones?

- *Yes, by small margins.*
- *Indicates LLMs handle both the contexts similarly, with idiomatic contexts not necessarily being more challenging than literal ones.*

The lengths of S1, S2, and S3 were comparable between both the contexts.

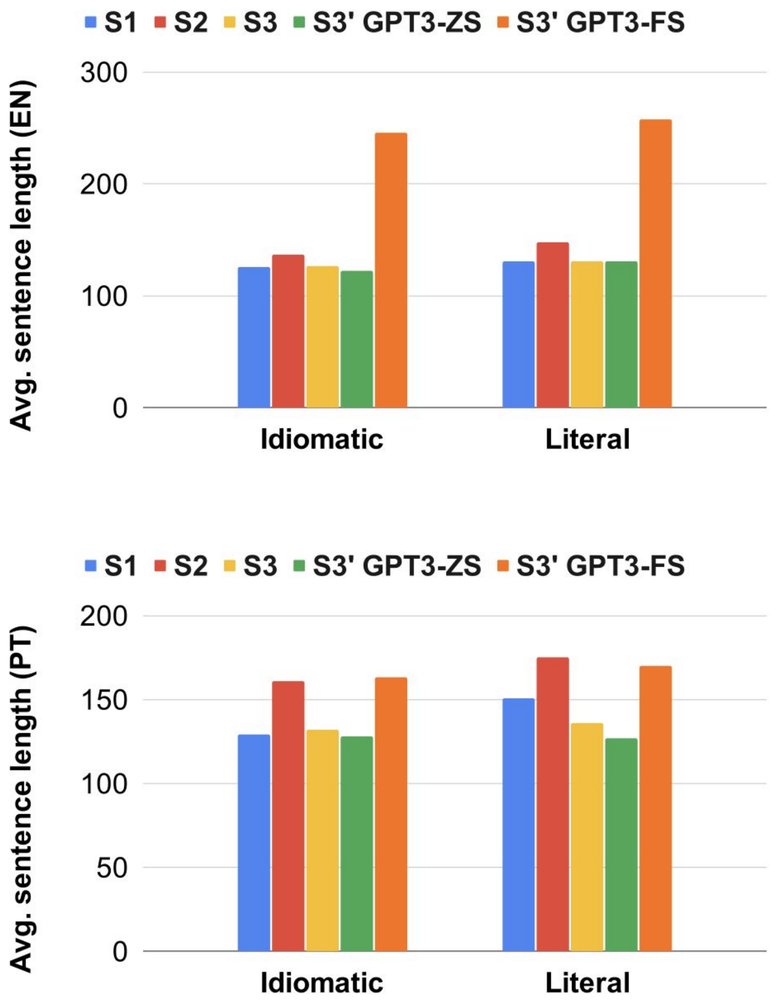


Figure: The graphs comparing the average lengths of the sentences (numbers of words) for English (top) and Portuguese (bottom).

...continued

English vs Portuguese

- *similar results*
- *possible reason: same language family and both being high resource languages*

Under Limited Context (only S2)

- *results are generally poorer*
- *even poorer in English*

Zero shot vs Few shot vs Finetuned

- *For EN, zero shot was better.*
- *For PT, finetuning helped.*

	METEOR		BERTScore	
	I	L	I	L
Only $S2$ is used				
EN	0.10	**0.11**	0.58	**0.59**
PT	**0.09**	0.08	0.59	**0.61**
$S1$ and $S2$ are used				
EN	0.12	**0.14**	0.59	**0.60**
PT	0.10	0.10	0.59	**0.61**

Performance of GPT-3 davinci model under zero-shot setting when only S2 is used (without S1).

Human Evaluation

- Two annotators
- 25 samples from GPT-3

Tasks

1. Relevancy Rating of S3’
2. Grammatical correctness (Subjective)

Results

Task 1

GPT-3 results generated better continuations in idiomatic cases.

Task 2

Incomplete sentences, irrelevant texts, unexpected content/nonsensical contents.

CONCLUSION

Conclusion

- Literal context seemed a little more easier for LLMs.
- GPT-3 outperformed all other models as expected.
- Surprisingly, zeroshot on GPT-3 was better than few shot.

Future Work

- More prompts to be explored.
- Usage of more recent models.
- What about idioms in low resource languages?

THANK YOU

QA

pokharel@pdx.edu
ameeta@pdx.edu

Link to GitHub