Don't Go Far Off: An Empirical Study on Neural Poetry Translation



Tuhin Chakrabarty
tuhin.chakr@cs.columbia.edu

Arkadiy Saakyan
a.saakyan@columbia.edu

Smaranda Muresan
smara@cs.columbia.edu

Natural Language Processing
Columbia University

https://github.com/tuhinjubcse/PoetryTranslationEMNLP2021

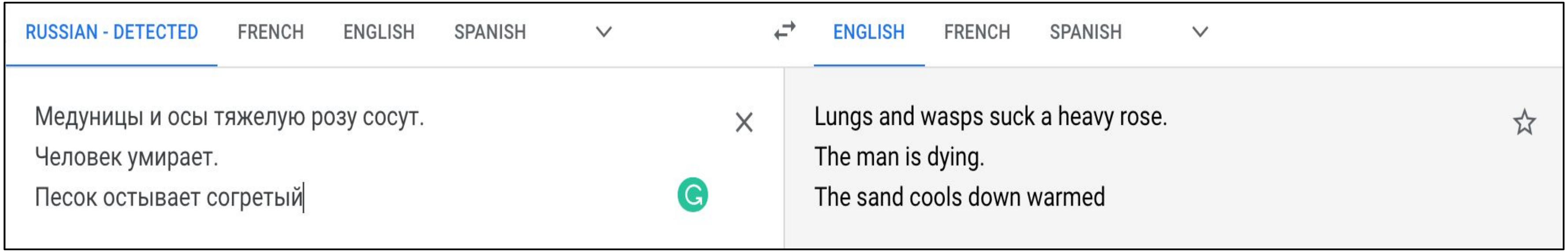
Introduction

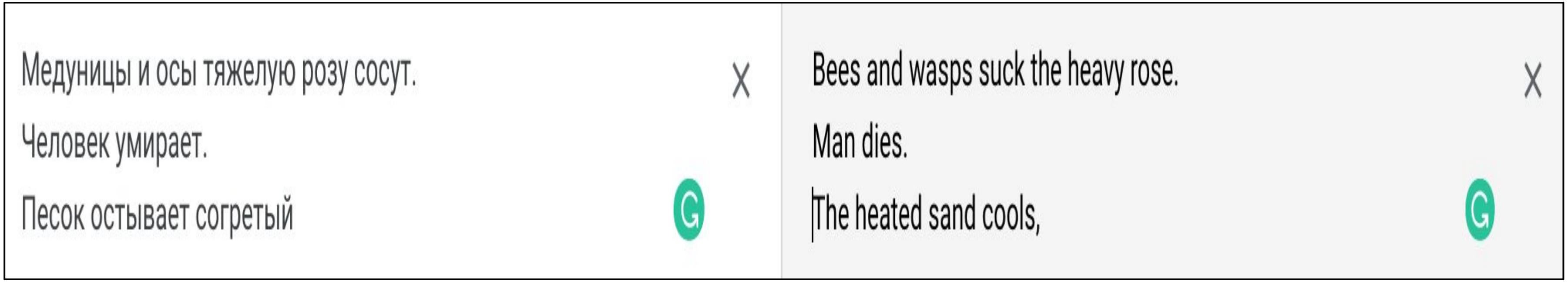
Poetry is *“that which gets lost out of both prose and verse in translation“* (Frost, 1961)

Challenges with Poetry Translation

- The lack of open-sourced multilingual parallel poetic corpora
- The intrinsic complexities involved in preserving the semantics, style and figurative nature of poetry

Translation from MBART along with Gold Translation





Efforts in Poetic Translation

Genzel et al. (2010) produce poetry translations with meter and rhyme using phrase-based statistical MT approaches

Ghazvininejad et al. (2018) present a neural **French to English** poetry translation system that focuses on rhyme and fixed rhythm patterns rather than meaning

Contributions

- Release parallel poetic corpora enabling translation from Russian, Spanish, Italian, Dutch, German, and Portuguese to English

- Release test sets for poetry translation from Romanian, Ukrainian and Swedish to evaluate the zero-shot performance of our models

- Show Multilingual fine-tuning of pre-trained models on poetic text significantly outperforms multilingual fine-tuning on non-poetic text that is 35X larger in size (177K vs 6M), both in terms of automatic and human evaluation metrics such as faithfulness

- We also show that multilingual fine-tuning on languages belonging to the same language family sometimes leads to improvement over fine-tuning on all languages

New Parallel Corpus

NEW

Language Pair	Source	Train	Valid	Test
Spanish-English	https://www.poesi.as/ https://lyricstranslate.com/	37,746	2059	536
Russian-English	https://ruverses.com/	50,001	4186	548
Portugese-English	http://www.poemsfromtheportuguese.org/ https://www.poetryinternational.org/ https://lyricstranslate.com/	15,199	699	140
German-English	http://www.poemswithoutfrontiers.com/ https://www.poetryinternational.org/ https://lyricstranslate.com/	17,000	1,050	1295
Italian-English	https://digitaldante.columbia.edu/ https://www.poetryinternational.org/ https://lyricstranslate.com/	34,534	1,997	528
Dutch-English	https://www.poetryinternational.org/ https://lyricstranslate.com/	23,403	1,000	159

Models

Non-Poetic Bi (OPUS)	Non-Poetic Multi (ML50)	Non-Poetic Multi (OPUS):
mBART50 on Non-Poetic data from OPUS100 for respective languages bilingually.	**mBART-large-50-many-to-one** model multilingually fine-tuned on 50 languages from the ML50 data	mBART-large-50-many-to-one model multilingually fine-tuned on on Non-Poetic data for 6 languages from OPUS100 (6M parallel sentences).

ML50 is 4 times larger than OPUS and created using all of the data that is publicly available (e.g., WMT, IWSLT, WAT, TED).

Models

Poetic	Poetic Lang Family	Poetic All
mBART50 fine-tuned bilingually (e.g., Ru-En, Es-En, It-En) on poetic data	**mBART-large-50-many-to-one** multilingually fine-tuned on poetic data for all languages belonging to the same language family	**mBART-large-50-many-to-one** multilingually fine-tuned on all poetic data combined.

Original: Люблю я пышное природы увяданье,

Gold: I love the lavish withering of nature,

mBART: I love the splendor of nature's devotion,

Poetic: I love the luxuriant decay of nature,

Evaluation

- **BLEU**
- **BERTScore** It computes a similarity score using contextual embeddings for each token in the system output with each token in the reference. We report F1-Score of *BERTScore*. We use the latest implementation to date which replaces BERT with *deberta-large-mnli*
- **COMET**
  - COMET models were trained to simulate the ratings of the WMT Metrics shared task, which are normalized per annotator
  - It leverages recent breakthroughs in cross lingual pre-trained language modeling resulting in highly multilingual and adaptable MT evaluation models that exploit information from both the source input and a target-language reference translation in order to more accurately predict MT quality.
  - We rely on the recommended model wmt-large-da-estimator-1719

Automatic Evaluation Results

Model	BLEU	BERTScore	COMET
Non-Poetic Bi(OPUS) Ru-En	12.4	65.4	-47.83
Non-Poetic Multi(ML50) Ru-En	13.0	67.5	-37.55
Non-Poetic Multi(OPUS) Ru-En	12.8	67.2	-39.5
Poetic Ru-En	11.9	64.3	-55.14
Poetic LangFamily	-	-	-
Poetic All	**17.0**	**70.2**	**-25.71**
Non-Poetic Bi(OPUS) Es-En	26.9	74.6	1.43
Non-Poetic Multi(ML50) Es-En	5.1	58.9	-60.98
Non-Poetic Multi(OPUS) Es-En	28.0	75.6	4.84
Poetic Es-En	26.8	74.3	-3.09
Poetic LangFamily	30.9	**77.2**	**12.14**
Poetic All	**31.2**	76.6	10.10
Non-Poetic Bi(OPUS) Pt-En	9.5	63.3	-47.27
Non-Poetic Multi(ML50) Pt-En	7.3	62.7	-53.48
Non-Poetic Multi(OPUS) Pt-En	9.2	64.0	-42.86
Poetic Pt-En	9.6	63.4	-50.93
Poetic LangFamily	**12.5**	66.4	-39.36
Poetic All	12.2	**66.6**	**-35.89**
Non-Poetic Bi(OPUS) It-En	22.2	70.3	-14.85
Non-Poetic Multi(ML50) It-En	17.0	68.7	-24.53
Non-Poetic Multi(OPUS) It-En	22.9	71.1	-8.87
Poetic It-En	18.8	69.3	-24.21
Poetic LangFamily	**25.4**	**72.2**	**-7.35**
Poetic All	24.6	71.6	-8.87
Non-Poetic Bi(OPUS) De-En	15.2	68.6	-27.95
Non-Poetic Multi(ML50) De-En	20.1	73.4	-5.88
Non-Poetic Multi(OPUS) De-En	17.8	70.9	-16.77
Poetic De-En	16.8	70.2	-23.07
Poetic LangFamily	20.5	73.6	-4.22
Poetic All	**22.7**	**74.6**	**-0.52**
Non-Poetic Bi(OPUS) Nl-En	24.5	72.5	-4.83
Non-Poetic Multi(ML50) Nl-En	23.8	72.2	-6.73
Non-Poetic Multi(OPUS) Nl-En	26.1	72.9	-4.83
Poetic Nl-En	26.5	71.6	-12.73
Poetic LangFamily	**32.1**	74.3	-3.74
Poetic All	30.7	**74.5**	**-1.90**

Human Evaluation Results

Human judges were asked to evaluate on a binary scale whether:

- the model introduces hallucinations or translates the input into something arbitrary, i.e. and at the same time
- the syntactic structure is poetic and the translations are rich in poetic figures of speech (e.g., metaphors, similes, personification).

	NonPoetic Best	Poetic Best
Ru-En	20%	80%
Es-En	0%	100%
Pt-En	40%	60%
De-En	28%	72%
It-En	28%	72%
Nl-En	0%	100%

Performance on Unseen Languages

Our multilingually fine-tuned poetic model outperforms the other two multilingual models fine-tuned on Non-Poetic data, even though the languages were not contained in the fine-tuning data.
M1=Non- Poetic Multi(ML50); **M2**=Non-Poetic Multi(OPUS); **M3**=Poetic All.

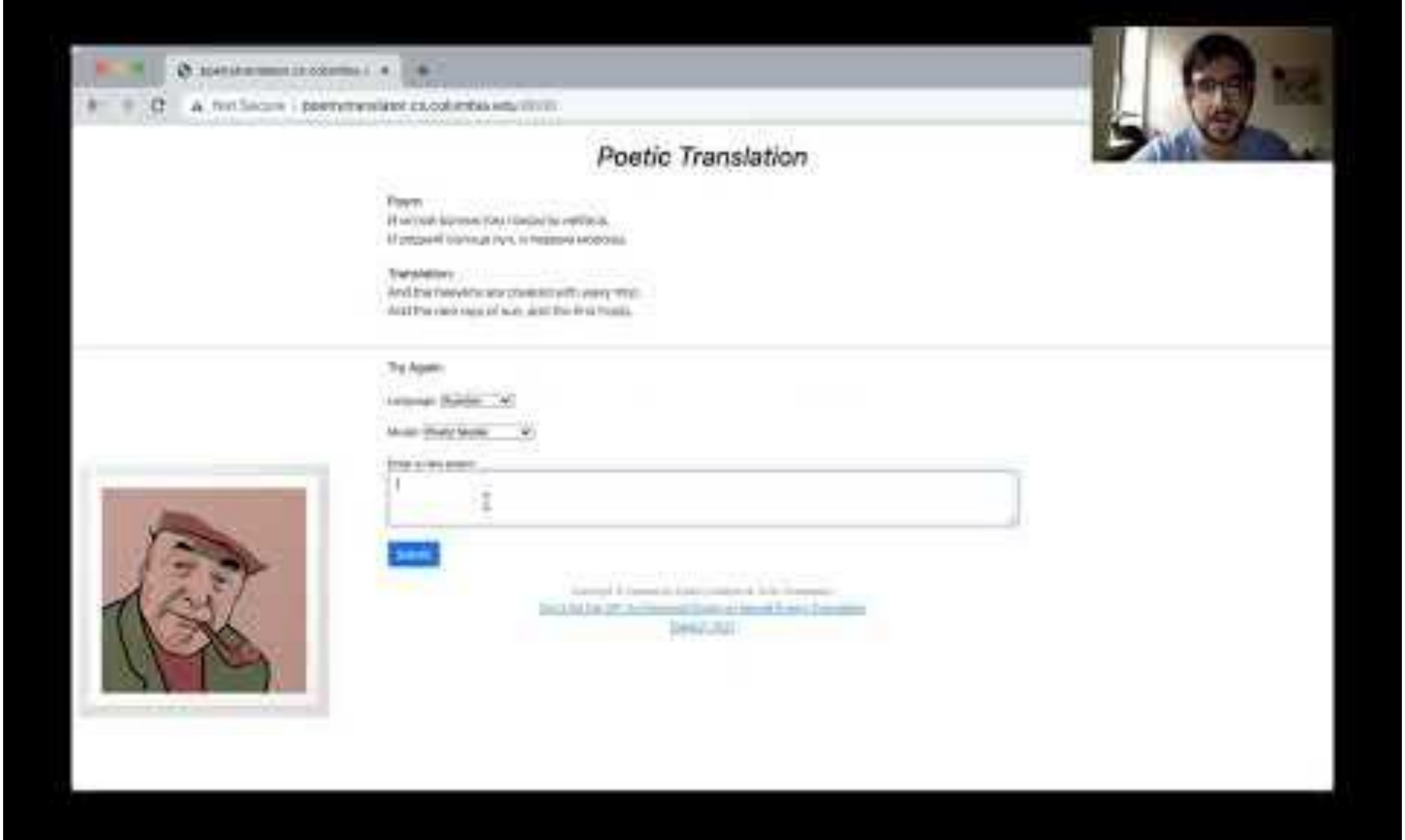
		BLEU	BERTScore	COMET
Ukranian	M1	9.2	64.2	-39.61
	M2	9.1	65.0	-40.46
	M3	**15.1**	**67.3**	**-32.10**
Romanian	M1	**30.1**	74.7	13.71
	M2	24.4	73.6	9.43
	M3	29.9	**76.1**	**18.15**
Swedish	M1	14.3	68.0	-24.21
	M2	16.6	66.4	-30.47
	M3	**19.5**	**71.3**	**-14.97**

Conclusions

- We release poetic parallel corpora for 6 language pairs
- Our work shows the clear benefit of domain adaptation for poetry translation
- It further shows that improvements can be achieved by leveraging multilingual fine-tuning, and that the improvements transfer to unseen languages
- Future directions include addition of new languages and larger corpora, adapting low-resource machine translation techniques for poetry translation, translating to languages that are morphologically richer than English, as well as working on better evaluation metrics to detect hallucinations

И мглой волнистою покрыты небеса,
И редкий солнца луч, и первые морозы

The sky engulfed by tides of rippled gloom,
The sun's scarce rays, approaching frosts,


Poetic Translation

Thank you!

tuhin.chakr@cs.columbia.edu
a.saakyan@columbia.edu
smara@cs.columbia.edu

QUESTIONS ?