Research report template: google doc here

TIMELINE

02.04	Deadline for submitting Report version 1
04.04	Deadline for Feedback 1
25.04	Finalize the data collection and inspection
30.04	Deadline for submitting Report version 2
02.05	Deadline for Feedback 2
10.05	Finalize literature review and background research, settle down on the methodology and focus
17.05	Gather all the results and proceed with evaluation and statistical analysis
30.05	Final Report deadline (tentative)

1. Introduction

- definition of the phenomenon
- research question: what do we want to find out about this phenomenon
- motivation: why does the world (or at least your research community) need to know about your study

The submissions to the AmericasNLP 2021 Shared Task on Open Machine Translation (OMT) offer vast possibilities.[1] Not only do they offer state-of-the-art solutions to the problem of machine translation into low-resource languages, but also a comparative analysis of the approaches adopted by different teams. In this study, we take a closer look at the datasets offered to all the teams and collected by each team.

While the open design of the shared task makes it impossible to compare the *models* submitted by the teams, we can take a lexicographic approach and examine the *datasets/corpora* used by the teams. This is well in line with the resurgent data-centric paradigm in the research

[1] Findings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas ---> https://aclanthology.org/2021.americasnlp-1.23/

community, as it (a) offers an alternative to costly and often inaccessible resources needed for training of large models and (b) proves to be invaluable in low-resource settings.

We adopt an exploratory approach, and will zoom in on the train, development (?) and test sets in search of the patterns which could have a significant impact on the results. In particular, we are going to focus on what we a priori know to be challenging/impossible for the models: out-of-vocabulary tokens (OOV) to see how their distributions across different datasets correlate with drops and gains in performance.[2]

2. State of the art / Background

- what is already known about the phenomenon
- theories
- Concepts
- the gap: what the others got wrong or missed (and your study will fix)

Translation into and from low-resource languages is a topic of growing interest in the research community, including the recent spark of interest into multimodal NLP models and their application to low-resource languages. While the state-of-the-art models show steady improvements against benchmarks, these improvements are slowing down.

Hence, there is a noticeable focus shift - both in the research community and in the industry - towards data-centric approaches. This involves carefully curated dataset construction and annotation as key factors influencing model performance.[3] It has been demonstrated that a more carefully curated pre-training dataset can contribute to remarkable improvements in language modeling.[4] The choice of an appropriate tokenization approach alone can yield significant improvement in model performance, or sabotage the improvement efforts.[5] [6]

If anything has been shown with clarity even by the most superficial analysis of the submissions, it is that careful curation of the corpora/datasets is a key factor contributing to the MT model success. The research gap we thus hope to bridge is the influence of dataset size, richness and diversity on system output quality.

[2] Translation of Unknown Words in Low Resource Languages → https://www.cs.jhu.edu/~phi/publications/translating-unknown-words.pdf

[3] Andrew Ng. 2021. MLOps: From Model-centric to Data-centric AI. Virtual Event. (2021). https://www.deeplearning.ai/wp-content/uploads/2021/06/MLOps-From-Model-centric-to-Data-centric-AI.pdf.

[4] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. (2020). arXiv:cs.CL/2101.00027

[5] Should we find another model?: Improving Neural Machine Translation Performance with ONE-Piece Tokenization Method without Model Modification → https://aclanthology.org/2021.naacl-industry.13.pdf

[6] Impact of Tokenization on Language Models: An Analysis for Turkish → https://arxiv.org/pdf/2204.08832.pdf

3. Your approach

- your research question or a concrete problem that you are trying to solve
- expected or possible answers (hypotheses) or your proposed solution
- the intuition behind your hypotheses or proposed solution

The lack of knowledge of the target languages poses a series of challenges. To begin with, the human evaluation of system outputs (fluency and adequacy scores) was extremely narrow in scope and only covered two language pairs: Spanish to Shipibo-Konibo and Spanish to Otomí. Knowing the large disagreements between human annotators on such tasks, this raises concerns as to the validity of human translation quality ratings. The resulting "inter-annotator agreement" of the sole annotator is equal to 1, which cannot be considered to be scientifically trustworthy. Secondly, the inaccessibility of the target languages we have no understanding of makes it impossible to inspect the corpus and collect statistics on the vocabulary richness, POS distributions, etc.

It is thus beneficial to adopt an “alien view” on the datasets, as if we had no knowledge about any of the languages in the project, and examine the data for patterns and overlaps. We hypothesize that a careful lexicographic analysis of the corpora is able to point to the variables impacting model performance that can be viewed as performance predictors.

Neural machine translation systems, despite a vast superiority over statistical NMT systems in dealing with rare and OOV words, still show weakness in translating low-frequency words.[7] Although sub-word tokenization techniques can be employed to overcome this issue, the problem remains.[8] The first hypothesis addresses this problem and can be formulated as follows: There is a negative correlation between the share of OOV (out-of-vocabulary units) in the target language dataset (#TODO train? Dev? test?)

Further, we are planning to inspect the split into two tracks to see the impact of integrating the development set during training the successfulness of the submission. Clearly, all the other adapted parameters specific for a given submission must be taken into account as well. Knowing that (a) the development and test sets came from similar/same domains, and that (b) the inclusion of the dev set (track 1) resulted in a significant performance improvement, we hypothesize that the impact can be tracked down to the overlaps in vocabulary between the dev and test sets and, as a consequence, a drop in OOV rate.

#TODO: integrate the 2 tracks: → since the difference is in the use of the dev set during training, we could control for its impact if all the other parameters remain unchanged.

[7] Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation, pages 28–39, Vancouver. Association for Computational Linguistics.

[8] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

#TODO see if smth else was changed in each submission.

- Track 1: the development set is part of the training data (common practice in the machine translation community)

- Track 2: the development set is not part of the training data (mimicking a more realistic low-resource setting)

4. Data and methods

- what information (=data) you collect
- how you collect it
- how you organize the collected data
- what analyzes you apply what tools you use to perform the analyses

4.1. Data[9]

Data table here: RBCNLP_project_table

Unfortunately, the organizers and most participants of the AmericasNLP 2021 Shared Task do not provide output translations produced by the submitted models. This precludes us from inspecting the system outputs and verifying the evaluations. While it would theoretically be possible to replicate the results by running the models on the test sets, the ends would not justify the means: this would be an enormous waste of computation power with no clear benefit.

Luckily, we do have access to all the datasets (train, dev, test) for all languages: both those provided by the shared task organizers as well as those submitted by the participants.[10]

As we know from the project documentation, the domains of the training ("train") data differ from that of the development ("dev") and test sets for all language pairs, but dev and test sets are taken from the same domain and use the same orthography. We also know that none of the dev and test sets are tokenized, while some of the training sets are.[11] Therefore, we must select a tokenization technique for the original datasets provided by the Shared Task organizers Knowing how large a potential impact of tokenization on both performance and performance evaluation can be, we must be careful when comparing the train and test sets.

4.2. Method

We begin the inspection of the datasets with a comparison of dataset sizes and their actual vocabulary richness (word frequency dictionaries & TTR rates). This step is based on the assumption that TTR (type-to-token ratio) might be a good complement to the simple word count, since it offers a quantitative measure of the vocabulary size of the model. Although modern sub-word approaches to tokenization and translation attempt to overcome the issue of

[9] NB: none of the dev and test sets are tokenized
[10] https://github.com/AmericasNLP/americasnlp2021/tree/main/test_data
[11] https://github.com/AmericasNLP/americasnlp2021/blob/main/data/information_datasets.pdf

unknown words and neologisms, neural machine translation systems still show weakness in translating low-frequency words.

Test set:

In the first step, we need to extract raw statistical information on the vocabulary of the test sets in all languages. To this end, we tokenize the test sets and create frequency dictionaries for each of the test sets per language.[12] We can now compute the vocabulary size of each test set and save the output to one file.[13] We then compute the TTR (type-to-token ratio) for each document and add the results to the database.[14]

file	**vocabulary_size**	**TTR**	**file_es**	**vocabulary_size _es**	**TTR_es**
test_dict.test.bzd.txt	1909	0.163	test_dict.test.es.txt	2465	0.245
test_dict.test.hch.txt	3089	0.313	test_dict.test.es.txt	2465	0.245
test_dict.test.aym.txt	3700	0.550	test_dict.test.es.txt	2465	0.245
test_dict.test.gn.txt	2684	0.414	test_dict.test.es.txt	2465	0.245
test_dict.test.cni.txt	3082	0.480	test_dict.test.es.txt	2465	0.245
test_dict.test.shp.txt	2921	0.331	test_dict.test.es.txt	2465	0.245
test_dict.test.oto.txt	2235	0.215	test_dict.test.es.txt	2465	0.245
test_dict.test.nah.txt	2531	0.386	test_dict.test.es.txt	2465	0.245
test_dict.test.quy.txt	3462	0.513	test_dict.test.es.txt	2465	0.245

Dev set:

The procedure is repeated for the development set to compute the respective values.[15] This time, however, each dev set per language has its unique counterpart in Spanish, since the development sets are different for each language pair.

file	**vocabulary_size**	**TTR**	**file_es**	**vocabulary_size_es**	**TTR_es**
dev_dict.dev.cni.txt	2863	0.473	dev_dict.dev.es.txt	2577	0.269

[12] See test_dict_create.py
[13] See test_dict_count.py → test_dict_counts.txt
[14] TTR.py → TTR_counts.txt
[15] See dev_dict_create.py, dev_dict_count.py, TTR.py

dev_dict.dev.aym.txt	3920	0.555	dev_dict.dev.es.txt	2855	0.257
dev_dict.dev.bzd.txt	2195	0.169	dev_dict.dev.es.txt	2855	0.257
dev_dict.dev.gn.txt	2943	0.410	dev_dict.dev.es.txt	2855	0.257
dev_dict.dev.oto.txt	1483	0.293	dev_dict.dev.es.txt	1614	0.316
dev_dict.dev.nah.txt	1729	0.403	dev_dict.dev.es.txt	1828	0.289
dev_dict.dev.quy.txt	3839	0.521	-	-	-
dev_dict.dev.tar.txt	2589	0.250	dev_dict.dev.es.txt	2855	0.257
dev_dict.dev.shp.txt	3056	0.336	dev_dict.dev.es.txt	2855	0.257
dev_dict.dev.hch.txt	3358	0.329	dev_dict.dev.es.txt	2855	0.257

#TODO Train set

#TODO the above analysis for individual submissions including enriched datasets

#TODO OOV analysis

5. Findings

- counts, percentages, outcomes of statistical tests, often given in tables or graphs
- descriptions of these tables and graphs
- and/or clear statements containing new facts established in your study

Data table here: RBCNLP_project_table

We expect a significant impact of larger vocabulary overlaps between source and target test sets on the model evaluation results. As an indication that this might be a correct assumption, we have seen in the AmericasNLP 2021 that it is (a) careful curation of the dataset and (b) its enrichment beyond what was offered by the task organizers is what makes a huge difference.

6. Interpretation

- relate the findings to the starting expectations/hypotheses

- explain how your findings improve the knowledge about the phenomenon you studied

7. Discussion

- obstacles
- limitations
- alternative explanations
- Speculations
- broader relevance

7.1. Discussion

limitations:

- “sloppy” tokenization principles when calculating TTR – more fine-grained normalization (orthography) not taken into account

- this is a case study --> generalization of the findings on other studies and real-life scenarios must follow if we are to validate the findings;

broader relevance:

- relevance for "classical" neural MT systems as well as multimodal models;

alternative explanations:

- Since we are analyzing *approaches*, not *models* per se, we lack transparency as to which factor contributes to which degree to the changes in model performance. It is hard to distinguish between the influence of dataset properties and neural model architecture, hyperparameters, training regimes, etc.

7.2. Future work

- ablation studies to compare model performance instead of general approach effectiveness

- character noise infusion --> exploit close relation of some of the languages

8. Conclusion

- synthesis, the main message of your study