Computing Science

Do smaller natural language models offer satisfactory translation performance compared to bigger models?

Stanislav Stoyanov

No portion of the work contained in this document has been submitted in support of an application for a degree or qualification of this or any other university or other institution of learning. All verbatim extracts have been distinguished by quotation marks, and all sources of information have been specifically acknowledged.

Signed:

Date: December 2, 2022

CS4040 Report

Department of Computing Science
University of Aberdeen
King's College
Aberdeen AB24 3UE

December 2022

Do smaller natural language models offer satisfactory translation performance compared to bigger models?

Stanislav Stoyanov

1 Introduction

Machine translation of natural language has significantly improved in the past few years. Neural models have completely replaced statistical models and are growing bigger and more complex. A great example of that phenomenon would be Google Translate. In the past 8 years, it has evolved into a monster of a neural machine with impressive multilingual capabilities in text-to-text, speech-to-text, and text-to-speech tasks.[3] Translation models have come a long way and at this point are able to exhibit the human-specific complex language abilities that have been the missing piece in machine translation.[15] Natural language model evaluation is an interesting task because there is no right or wrong way of doing it, and there is not a single established correct way of doing it quite a lot of people are working on ways of developing various metrics which often are case specific. The metrics that I am interested in are mainly efficiency and correctness. Sure I can just pick the biggest model I can find, which will certainly give me the most proficient output. But at what point can we achieve satisfactory performance and stop wasting time and energy with bigger and bigger models? Are bigger models always more precise and is their environmental footprint always worth it? When using bigger models, users often come across different hurdles as running time and disk space are limited. Can we mitigate our carbon footprint and conserve resources and time by using smaller language models? In this experiment, I will try to evaluate to what extent a bigger model is more accurate than a smaller one specifically in a translation task. To achieve this task, I need to develop a metric to evaluate translations. I am only interested in comparing a big model to a smaller one in a real-world situation. By real-world situation, I mean a setting in which the model is met with most certainly a completely new subject. Everyday translations, such as common phrases and beginner conversations are of no interest as such translations are already widely available and easily accessible. I am interested in the ability of models to work with text that they haven't been met with as this is a more complex task which is worth exploring. If the performance of the big model can be shown to be only marginally better than the performance of the small then we can assume that the smaller version of the model is viable for 'normal' use. This research does not suggest that bigger language models are not necessary or not relevant in any way. They are crucially important for AI and I personally believe that developing a completely human-like model is a reasonable task to pursue.

2 Background and Related Work

It took reading quite a lot of papers in order to become fascinated with the evaluation process. There is a lot of work that is concerned with the issue of benchmarks and their inability to offer an evaluation

of real-world performance, which is different due to being noisy, unpredictable and generally more problematic.[2] This means that we need more concrete evaluation metrics for specific cases. A crucial metric for translation language models is human evaluation. Due to the circumstances of this course, I have not been able to incorporate this in any way. This is however a very important thing to study because naturalness is an invaluable feature of translation models. This adds a human element to the evaluation process and comes with a fair amount of drawbacks and difficulties.[17] An approach to developing a good metric is to combine simple metrics. This can be done in order to mitigate the weaknesses of each metric and combine their strengths.[21] A convincing study uses suitable metrics which enhance the credibility of the experiment.

3 Research Question

The main question that I will try to investigate is whether or not and/or to what extent a smaller model would be of sufficient quality for a 'normal' work compared to a bigger one. Today, many huge language models are available for a lot of different tasks. With their great capabilities come a number of drawbacks. They are slow to download, slow to run, slow to analyze and evaluate and require a lot of resources to manage and run like disk space and electricity. Personal computers have come a long way but not everyone has access to the latest technology. A project that is not focused on Natural Language Processing may want to use a language model for some task and in such work, speed and lightness might be priorities. A great concern would also be the environmental impact of using big and slow models. I am quite sure that I myself have used quite a significant amount of electricity compared to the usual PC user just by running my slow codes for hours and exploring and abusing various language models. Training and using models with billions of parameters is practically impossible without dedicated equipment such as a few graphics cards and days of training time. This is often not feasible so to incorporate a natural language model in a project there might be limitations on how big it is. In this mini-project, I will try to compare the performance of translation models and try to prove the hypothesis that small models usually offer sufficient performance. Evaluating the performance of translation models is quite a daunting task and my experiment will probably leave a lot to be desired but bad research is still research and conclusions can still be drawn.

4 Experimental Design

In this experiment, I used two versions of language translation models, which are trained on different amounts of training data. The experiment consists of using two sets of corresponding models (in this case English to French, French to English, English to Hungarian, and Hungarian to English) to translate entries to the non-English language and then back to English. This method removes the aspect of comparing multi-lingual data, which simplifies the task from virtually impossible to very hard. Doing two translations also highlights any weaknesses that the models exhibit - inaccurate, missing, wrong, or unsuitable information will be twice as likely. The main weaknesses that I am concerned with are 2 - first of all, is general accuracy and correctness of translation and second will be the retention of original words. As to whether such back-and-forth translations should explicitly

retain the exact same words is a question for a different study. Here I assume that retaining the same words is a desirable trait. The first one is quite explanatory feature of interest in translation models - we want target sentences to have the same meaning. This I tested with two models for sentence comparison. They use a database of word vectors in which similar words have similarly pointed and weighted vectors. This way they are able to produce an output score from 0 to 1 where 0 is a completely different sentence and 1 is exactly the same sentence. The second metric I am interested in is exactness. For this metric, I wrote a very simple python script which counts the words that are seen in both the original and target sentences and divides their number by the total amount of words, which algorithm I call the 'dummy' comparison model. This way I get a score, which is also from 0 to 1, which tells me how similar the sentences are but in a more literal sense. I have noticed that a few of the sentences get scores of 1.0 which means that the models do manage to reach the performance I was looking for, which suggests that the metric is not unreasonable. I have done a few steps to mitigate external factors. I have used two languages - one close to English - French and one not at all close to English - Hungarian. Although unrelated to my main focus of the experiment, the data clearly shows that either (or probably actually both) the models are better trained for English-French-English translation or that translation to a similar to the English language is simply a fundamentally easier task. The data I have collected is also not random. This may sound like a liability but in the circumstances of this experiment, I could not use huge amounts of randomized data. I used 20 sentences from a Wikipedia page which contains quite a lot of subject-specific terms and technical language.[20] I also used 20 sentences from a news article [1] which serves as an example of something from our everyday lives and 20 sentences from recipes [22], which I chose because of how necessary literal accuracy is in such translations. A lot of evaluation can be done on just these 60 sentences, but due to the limitations and the nature of this project, I had to focus on the basics. I have tested all 60 sentences with the 2 sets of translation models (total of 240 translations) and evaluated each of the 120 pairs of source and target English text with each of the 3 metrics. This way I have generated scores for every sentence and metric, for every sentence overall combined and for every sentence without considering my 'dummy' comparison model. I have included a statistic without considering the 'dummy' model as it can objectively be considered an unfair way of comparing the performance of translation models and is inherently a bit flawed. To stay on track with my hypothesis, I have focused on only comparing same-metric scores for each model - big and small. I believe that quite a lot can be done to improve the metrics, test data set, models and a few other aspects but I have used these as they make the most sense in the given time frame and workforce.

Thanks to "Huggingface.co"[14], "Helsinki-NLP"[4], "MarianMT" model[16], "Tatoeba" collection of sentence translations[19], "SentenceTransformer" python library[13], "spaCy" python library[18] and everyone working on these projects I was able to complete my little experiment. The models that I have used are "opus-mt-tc-big-fr-en"[9], "opus-mt-tc-big-en-fr"[5], "opus-mt-en-fr"[6], "opus-mt-fr-en"[10],"opus-mt-tc-big-en-hu"[7], "opus-mt-tc-big-hu-en"[11], "opus-mt-en-hu"[8], "opus-mt-hu-en"[12] and the two comparison models offered by SentenceTransformers[13]

```
C:\Users\Lenovo\Desktop\researchminiproject>py results.py
Language:fr
big wins ST with 0.9373 against 0.9124 with 36.5 rounds against 23.5
big wins spacy with 0.9711 against 0.9673 with 38.5 rounds against 21.5
big wins dummy with 0.8267 against 0.7872 with 41.0 rounds against 19.0
overall score of big model 54.702
overall score of small model 53.3366
overall score of big model when not considering dummy model is 57.2524
overall score of small model when not considering dummy model is 56.3902
big model is 2.56% better
```

Figure 1: Statistical results of the EN-FR-EN models

```
C:\Users\Lenovo\Desktop\researchminiproject>py results.py
Language:hu
big wins ST with 0.9146 against 0.87 with 39 rounds against 21
big wins spacy with 0.9529 against 0.9455 with 38 rounds against 22
big wins dummy with 0.7327 against 0.682 with 37.5 rounds against 22.5
overall score of big model 52.0042
overall score of small model 49.9497
overall score of big model when not considering dummy model is 56.0252
overall score of small model when not considering dummy model is 54.465
big model is 4.109999999999999% better
```

Figure 2: Statistical results of the EN-HU-EN models

and spaCy[18].

5 Results

The amount of dimensions and size of the data allows us to present it in a few different ways each of which suggests a different hypothesis. The main thing I was interested in was performance based on the compound metric. This would be evaluated by comparing summations of the scores for each metric for each sentence and generating some sort of overall score for each model. I have classified each sentence as a ”round” in which the models get scores and compete. I have generated scores for each of the two tuples of translation models for each of the three metrics. Then I calculated overall scores for each model and a bonus score, based only on the ”spaCy”[18] and ”SentenceTransformer”[13] similarity models as the usage of the dummy model is a bit questionable as I previously mentioned. The values are shown in figures [1] and [2].

We can see from the number of rounds that the big models win compared to the small ones and the overall scores that the big model performs better. However, the difference in score is not very significant and in the scores that do not consider the dummy model it is even less. The differences in scores that the ’dummy’ model gives are broader everywhere which shows that my hypothesis of the ’dummy’ model being too rough for use in this experiment is correct. Despite this, the ’dummy’ model does give some valuable insight into a quality that the other 2 models fail to show so clearly. Despite losing overall every time, the small models did win a significant amount of rounds with a ratio of around 1 to 2. This shows that the smaller model is not only viable overall, but in some sentences even better than the bigger one. The results are the same within both tuples of translation models with the only difference being that the scores in English-French-English are a bit bigger. This would prove

either of two things - that English-French-English models are more advanced, better trained or just better overall or that translation to a language that is closer to the source is a fundamentally easier task. I would speculate that both are true but that's a question for a different study. I used two ways to plot the generated data. Firstly I used a bar chart to chart the overall scores for each sentence with two scores per sentence - one for the big model and one for the small one. I also coloured the bars with the following colour scheme: 'Green' for scores above 0.9, "Yellow" for scores between 0.9 and 0.75 and "Red" for scores below 0.75. This gives us an insight into how good the models did. The clearest inference we can do from that colouring is that the English-French-English models did a lot better. (Figures [3] and [4]) I have also included 2 bonus bar charts for the scores that are calculated only on the "SentenceTransformer" and "spaCy" models on both tuples. (Figures [5] and [6])

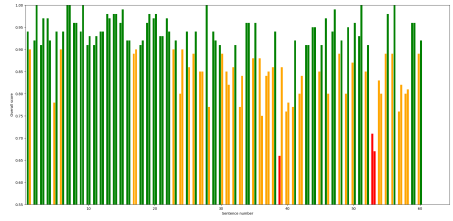

Figure 3: Overall performance of EN-FR-EN models.

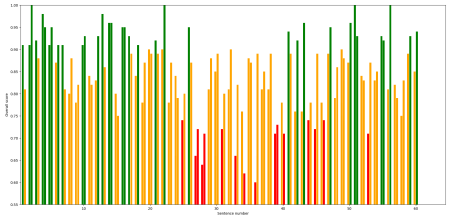

Figure 4: Overall performance of EN-HU-EN models.

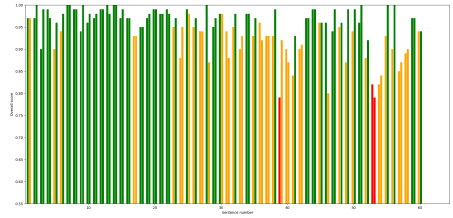

Figure 5: Overall performance of EN-FR-EN models, when not accounting for 'dummy' model scores.

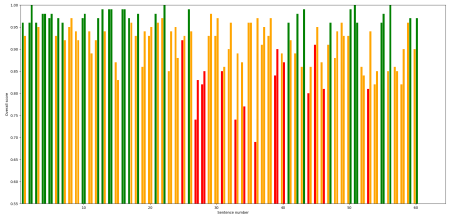

Figure 6: Overall performance of EN-HU-EN models, when not accounting for 'dummy' model scores.

Secondly, I used a similar bar chart but in this case, I coloured the bars based on which model had a higher score. "Green" when the bigger model has a better score, "Yellow" when they are equal and "Red" when the smaller model performed better. This way we can still clearly see that the En-Fr-En models did better but we can also clearly see that the bigger models perform better overall.(Figures [7] and [8]) Here I have also included 2 bonus charts for the scores that are calculated only on the 'non-dummy' models. (Figures [9] and [10])

I have also calculated how much the big models are better than the small ones which evaluated to around 2.5% for the English-French-English models and around 4% for the English-Hungarian-English models.

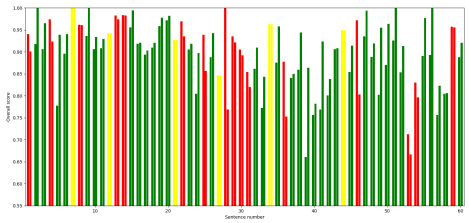

Figure 7: Performance comparison of EN-FR-EN models.

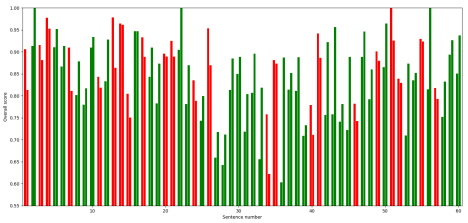

Figure 8: Performance comparison of EN-HU-EN models.

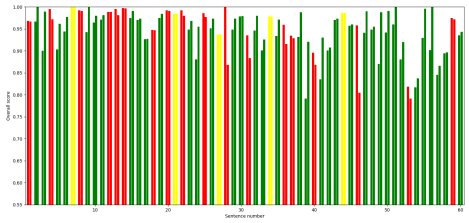

Figure 9: Performance comparison of EN-FR-EN models, when not accounting for 'dummy' model scores.

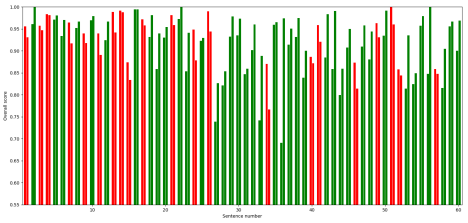

Figure 10: Performance comparison of EN-HU-EN models, when not accounting for 'dummy' model scores.

6 Discussion

The only conclusion I can derive from this experiment with 100% certainty is that model evaluation is a very complicated task. This experiment may be far from decent but it does give us valuable insight. It is a step towards a successful experiment. Despite not giving clear factual evidence and sufficient and precise data, it clearly reveals quite a few of the hurdles that need to be overcome in future work. Causation is generally hard to prove, but by mitigating external factors as much as possible, this study makes a statement on the main hypothesis. I have also found that the data can be used to suggest and even try to prove other hypotheses. This is valuable because it raises different side questions. When analysing my own experiment, I managed to ask a lot of questions, each of which is relevant enough for a future study. The amount of models that can be tested, types of test data and languages is enormous and with the data that is publicly available today, there are limitless possibilities of experiments that can be done with an equally big amount of hypotheses that can be raised. The type of data to be translated, the models used, their size, and the languages they are based multiplied by the number of metrics that can be of interest when comparing and/or evaluating models is what makes this task interesting. There are comparison metrics that I have not even considered incorporating in this study such as model disk size, time to compile, time to download, naturalness and many more. A lot can be done on the matter if the main focus is mitigating the waste of computing hours and electricity use and using the most efficient and as small as practically reasonable models. Another thing that I have not considered at all is training the models on specific data. This is completely irrelevant in

everyday cases such as the one I am investigating here but is invaluable for specific tasks. If sufficient training data is available, the models can be used to accurately translate specific texts such as my version of a technical text and recipes. Specific tasks require specific translation behaviour, which will probably also create the need for specific metrics and methods of evaluation and comparisons.

7 Conclusion & Future Work

It would be a long shot to try to make conclusions on the data that this experiment has generated. More data and better metrics are needed to start developing conclusions. Cross-analysis of scores will also be needed in order to show that the scores are similarly distributed and the only difference is that they are higher for some models. Future work can be done in a lot of directions. The one I am interested in would be gathering data about the exact size of the models, the size of their training sets, the amount of time they take to download and run and most importantly the difference between these values. A very interesting project to develop would be a comparison between size-related performance. It is not unreasonable to try to show a performance difference of a few per cent with a data set size difference of a few magnitudes. This would clearly show that smaller models are capable and we often do not need to waste valuable time and resources with unnecessarily big models.

8 Reflective Analysis

The main problem that I encountered while conducting my experiment was ironically the size and speed of the models that I used. In the last stages, my code was taking almost an hour to compile. A few minutes were lost in loading the 6 models that I was using. This was caused by a lot of factors such as the hardware, the not-optimized code, and the number of sentences I used. The main issue I was faced with was finding small mistakes in the code and formatting of sentences which would compound into bigger issues later on in the data and the correct way of plotting my findings. A weakness of my project that I can easily point out is that I have not based my reasoning and research on real papers enough. I should have spent more time linking my work to other related studies. The only thing I would do differently is to spend more time on this experiment as it has been one of the most fun things I have done during my studies.

References

[1] Capybara wikipedia page and ”giant rodents ’invaded’ a wealthy gated community.” article on time.com.

[2] Evaluating the robustness of neural language models to input perturbations.

[3] Google translate wikipedia page.

[4] Helsinki NLP project.

[5] Helsinki-NLP EN-FR big model.

[6] Helsinki-NLP EN-FR small model.

[7] Helsinki-NLP EN-HU big model.

[8] Helsinki-NLP EN-HU small model.

[9] Helsinki-NLP FR-EN big model.

[10] Helsinki-NLP FR-EN small model.

[11] Helsinki-NLP HU-EN big model.

[12] Helsinki-NLP HU-EN small model.

[13] Huggingface.co sentence similarity tutorial.

[14] Huggingface.co website.

[15] Language models show human-like content effects on reasoning.

[16] Marian MT model.

[17] Rankme: Reliable human ratings for natural language generation.

[18] SpaCy python library.

[19] Tatoeba dataset.

[20] Toyota A engine Wikipedia page.

[21] Unifying human and statistical evaluation for natural language generation.

[22] BBC goodfood recipes.