Exploring Language Family Relationships by Fine-Tuning a Pre-Trained Text-to-Text Transformer

Mehmed Can Olgac
Yale University
24 Hillhouse Ave
New Haven, CT 06511
mehmed.olgac@yale.edu

Megan Zhang
Yale University
24 Hillhouse Ave
New Haven, CT 06511
megan.zhang@yale.edu

Abstract

Languages are often grouped into families on the basis of similar features, such as grammatical structures. We aim to investigate one such grouping, the proposed Altaic family, which includes not only Turkic and Mongolic languages but also Koreanic and Japanese languages on the basis of shared grammatical structures. Although such proposals are not widely supported in the linguistic community, we seek to explore similarities between languages through fine-tuning a text-to-text transformer in translation tasks of various languages. We compare the accuracy of English-Turkish translations resulting from having fine-tuned a model in translations of other languages to English: Japanese-English and French-English. We find that models fine-tuned in either language are able to translate some simple English sentences, both reversible and irreversible, with high accuracy. However, both fine-tuned models were unable to translate more complex sentences with accuracy. We also did not find any significant difference in the translation accuracy of the model fine-tuned in Japanese as opposed to French. Further experimentation is needed to be able to observe differences in translation.

1 Introduction

Languages can be classified based on several different features. Some are said to be "genetically" related, sharing common ancestry (a proto-language) with others in their language family. Languages can also be classified according to their sentence structures (e.g. subject, object, and verb placement) – this is known as typological classification. These classifications are still being studied, and the exact classification of languages into families is still a topic of debate.

This project is in part motivated by the Altaic language family proposal, also known as the Ural-Altaic language family proposal. The Turkic languages are a language family of at least 35 languages, including Turkish, Azerbaijani, and Uzbek; these languages are thought to have originated in a region of East Asia spanning from Mongolia to Northwest China. The Altaic language family is a proposed language family which would include Turkic, Mongolic, and Tungusic language families; in addition, Japonic and Koreanic languages have been proposed by some scholars (Starostin, 2016). Figure 1 shows the proposed geographic distribution of this language family (Kassian et al., 2021). However, many comparative linguists have rejected the proposal of the Altaic family, and attempts to include Korean and Japanese into the proposed family have also been unsuccessful.

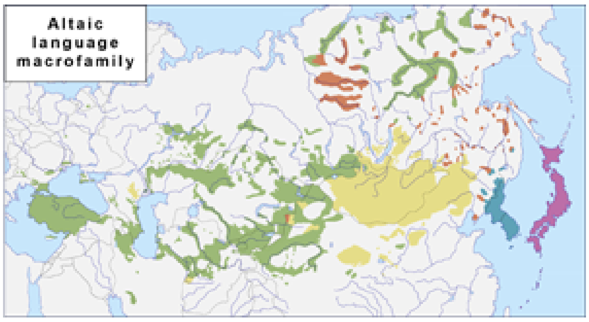

Figure 1: Geographical extent of the hypothesized Altaic language family (Kassian et al., 2021).

Nonetheless, the Altaic family theory is still supported by a small minority of scholars, and holds some cultural (and possibly political) significance. It is often argued in the Turkish national educational system that Turkish belongs to the Ural-Altaic language family and that Turkish and Japanese are grammatically very similar; for example, both are SOV (Subject, Object, Verb) languages and are also considered to be agglutinative. On the other hand, as a result of centuries of coexistence, Turkish acquired many words and phrases from both Arabic and Persian, and even after the simplification of Ottoman Turkish to Modern Turkish in the 1920s and 1930s, the influence of both of these languages is still noticeable. However, grammar structures differ more between Turkish and languages like Arabic and Persian.

Our goal is to explore the proposed relationship between Turkish and other languages, including Japanese. We do this through the lens of machine translation; we seek to evaluate and compare the efficacy of using different languages of varying similarity to Turkish to fine-tune a translation model. We aim to observe whether a model which has learned translation in a particular language will achieve better performance on given translation tasks in a separate target language. In this case, we evaluate the effect of fine-tuning a model on Japanese-to-English translation on the model's accuracy in Turkish-to-English translation. The accuracy will be compared to that of fine-tuning on other language-to-English translations.

2 Background

For this project we make use of transfer learning by fine-tuning an mT5 model, a multilingual pre-trained text-to-text transformer. The mT5 model is pre-trained on 101 different languages. For fine-tuning, we use a dataset of sentence pairs between English-Japanese and English-French. along with English-Turkish sentence pairs.

As previously mentioned, we take interest in Japanese due to the grammatical similarities and proposed relationship with Turkish. French serves as a control as it is considered to be a Romance language of the Indo-European family and generally shares less grammatical similarity with Turkish. Other datasets containing sentence pairs between English and other languages, such as Arabic and Persian, are also of interest to future work on this topic.

Transfer learning has been relevant in neural machine translation (NMT), and particularly so in regard to low-resource languages. Zoph et al. (2016) apply transfer learning to an encode-decoder framework for NMT by first training a high-resource language pair, known as a parent model, and then transferring some of the learned parameters to a low-resource pair (or child model) to initialize training. This method results in improved BLEU scores on low-resource language pairs. Our approach is partially motivated by this method; to better observe and understand how translation to Turkish can be improved, our fine-tuning data has been artificially limited (additionally, time and resource constraints lead us to take a very limited dataset).

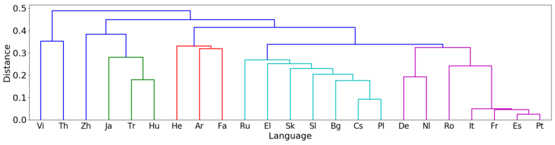

Figure 2: Hierarchical clustering based on language embeddings (Tan et al., 2019). These previous results indicate a closer relationship between Turkish and Japanese.

3 Related Work

Previous work in computational linguistics has also explored the relationship between languages and language families. (Tan et al., 2019) discuss the process of training for multilingual NMT and explore the idea of clustering language pairs together and training one model for each cluster, giving consideration to the fact that similar language pairs are likely to boost each other in model training while language pairs which differ greatly may negatively affect the training process if they are clustered together.

Tan et al. find that clustering languages by a representative embedding is able to improve translation accuracy compared to clustering languages by prior knowledge of language families. Interestingly, the hierarchical clustering method Xu et al. use to group languages places Japanese and Turkish within the same cluster, reflecting knowledge of morphological typology (Figure 2). As expected, French is placed in a separate cluster containing roughly Romance languages, and located a much further distance in the embedding space. We take a similar training approach in this project, but only compare results from fine-tuning on one language pair (as well as a reduced set of Turkish translations) at a time.

Other studies, such as Dabre et al. (2017), examine multilingual NMT in the context of language pairs by applying Zoph et al.'s transfer learning approach. Dabre et al. explore how different choices of parent models affect the performance of child models using multiple language pairs. In this case, using a parent model with a linguistically similar language was shown to achieve better results (measured by BLEU score) with the child language model.

If Japanese is closely related to the Turkish language, we would expect similar results to Dabre et al.; models pre-trained in Japanese translation should perform better compared to models pre-trained in other languages. However, it is possible that better performance may be an indicator of grammatical structures that are common among many different languages, rather than a language family relationship.

4 Methods

4.1 Model

multilingual-T5 is, as its name implies, the multilingual version of Google's Text-To-Text Transfer Transformer – T5 for short. T5 leverages the powerful idea of transfer learning, where a model is first pre-trained on a data rich task before being fine-tuned for a particular task (Raffel et al., 2020). The pre-training for T5 used data from the Common Crawl project, a project that extracts close to 20 TB of data from the world wide web on a monthly basis. This data was then cleaned (as it included explicit and obscene language, as well as incomplete sentences, Javascript, and many other unwanted text data). This cleaning resulted in what the creators of T5 called The *Colossal Clean Crawled Corpus*, or C4.

A general transformer architecture model architecture is shown in Figure 3 (Vaswani et al., 2017). The transformer model is pretrained on this unlabeled data via unsupervised learning procedures. The only relevant difference between T5 and mT5 is that the latter was pre-trained on a new Common Crawl-based dataset covering 101 languages. The details of the T5 model are outlined in Raffel et al. (2020). The details of the mT5 model are outlined in Xue et al. (2021), published around eight months after the initial paper in March 2021.

It should be noted that Google released smaller versions of mT5 as fine-tuning the full version would require hardware that not many individuals or researchers possess around the world. Hence, in our study, we used the smallest version of the available mT5 models available on `https://huggingface.co/`, called `mt-small` (Google, 2021).

4.2 Data

We used the Tatoeba Multi-Language Dictionary dataset, also available on huggingface.co , to acquire datasets of English to Japanese, English to French, and English to Turkish translations (Tatoeba.org, 2021). There were 186,511 English to Japanese, 306,133 English to French, and 689,302 English to Turkish translations available to us through tatoeba. We randomly sampled around 20,000 translations from all three of these categories for our train-sets and around 2,000 translations for our test sets. We will further discuss both the reasons and the consequences of these numbers in the Results and Discussion section of the paper.

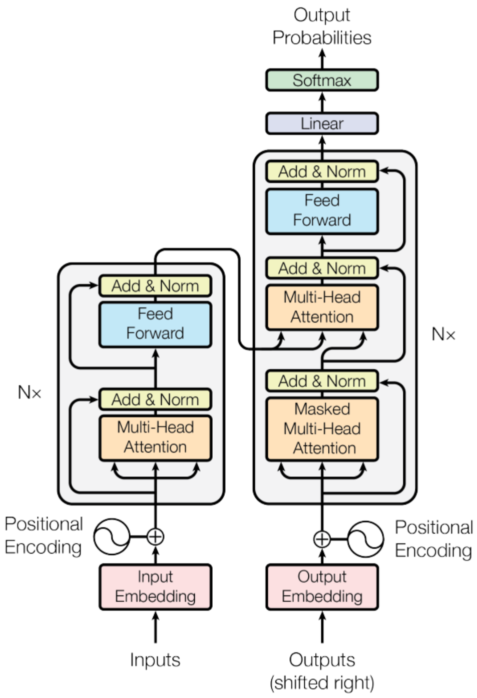

Figure 3: Model Architecture of a transformer (Vaswani et al. 2017).

4.3 Procedure

We took two pre-trained small mT5 models and fine-tuned each using two translation tasks. The first mT5 model was fine-tuned on English to Japanese translations for 8 epochs and with a batch size of 16 and an initial learning rate of 5e-4. Then, it was further fine-tuned on English to Turkish translations with exactly the same hyper-parameters. We followed the exact same procedures in fine tuning the second model, but instead of English to Japanese translations in the first round of fine-tuning, we used English to French translations. To fine-tune the models, we modified instructions provided at `https://giters.com/ejmejm/multilingual-nmt-mt5` (Meyer, 2021).

4.4 Evaluation

We came up with a short list of simple and complex English sentences that we fed into both models. The set of evaluation sentences were constructed so that a perfect translation would utilize case marking, where Turkish is more similar to Japanese than it is to French.

We included English sentences which were both reversible and non-reversible to examine differences in their translation. Semantically reversible sentences have the property where when the subject and object of the sentence are switched (e.g. ”Mary likes John”), the sentence remains meaningful; on the other hand, non-reversible sentences, such as ”I ate an apple”, are not meaningful when the subject and object are switched (Richardson, 2010). Reversible sentences tend to be more easily misinterpreted; however, in languages with case marking, certain modifiers specifically denote the subject and object. We aimed to evaluate differences in translation of reversible and non-reversible sentences by models which had and had not been fine-tuned in languages with case marking.

The translations of both models were then evaluated by individuals who are proficient in both English and Turkish and their evaluations were recorded. Our original intention was to use BLEU scores in order to evaluate and compare the accuracy of translations by each model. However, as we will see in Section 5, we found that using BLEU scores would not create a meaningful metric of comparison for our models.

5 Results and Discussion

5.1 First Round of Training

Fine tuning both models took around 3 hours using a GPU via Google's Collaboratory. Each training module–English to Japanese, English to French, English to Turkish–took around 90 minutes to complete all eight epochs.

In all three training modules, the model was trained to overfit. For example, while the average loss for the English to Japanese translation module was as low as 0.071 on the eighth epoch, the test loss of the model kept going up after the second epoch and went as high as 3.28. While the discrepancy was less severe in the other two modules–English to French and English to Turkish– a similar trend was observed and models were trained to overfit.

Partially as a result of this overfitting, both models outputted rather meaningless translations for the complex sentences even though they both successfully translated simple sentences such as but not limited to:

- `Mary likes John.`
- `John likes Mary.`
- `I love you.`
- `I read books.`
- `I had dinner.`

5.2 Discussion of Preliminary Results

Unfortunately, the preliminary results were not very useful in determining the effect of finetuning on English to Japanese or English to French translations on the task of translating from English to Turkish as the models performed very similarly. Yet some interesting observations can still be made even if we cannot definitively argue that they were a result of the difference in the fine-tuning processes of the two models.

One causal inference we can make from these results, albeit a weak one, is that given that the second model, which was fine-tuned on English to French translations first, performed decently with sentences like “Mary likes John” and “John likes Mary”, where the Turkish language uses case marking, we cannot clearly see a benefit in finetuning with English to Japanese translations over English to French translations.

On the other hand, and more interestingly, the first model, fine-tuned on English to Japanese translations, always outputted translations to the complex sentences where the sentences followed the SOV word order, whereas the second model outputted some translations where the sentences did not end with verbs. Of course, the number of complex sentences fed into the two models were very small, but this observation might hint at some transfer of learning from English to Japanese translations to English to Turkish translations.

5.3 Second Round of Training

Given that the preliminary results were inconclusive on the subject of study of this paper: primarily the causal effect of training a model on English to Japanese translations as opposed to a grammatically unrelated language to Turkish such as French on the performance of the model in English to Turkish translations, we have decided to retrain both models by changing some of the hyperparameters in the hope that we could capture the causal effect, if any, better.

In this second training, we increased the number of examples in the primary fine-tuning modules and decreased the examples in the secondary fine-tuning module. In particular, we have trained the new models on 100,000 test translations from English to Japanese and French respectively while we decreased the number of test translations from English to Turkish to 10,000 in both instances. Furthermore, we decreased the number of training epochs to 4 in order to prevent overfitting by training the model too intensely on the training examples. We anticipated that such modifications in parameters would allow us to see more clearly the effect of the primary fine tuning on Japanese or French on Turkish better than the previous set up.

5.4 Discussion of Results of Second Training

The first module of fine-tuning took around 90 minutes in the second training as well since we trained the model only for two epochs instead of eight epochs. Both the test loss and the average loss were around 1.000 for the first module of fine-tuning of the first model–English to Japanese then English to Turkish. Hence, this fixed the problem of overfitting.

The second module of fine-tuning of the first model took only around 10 minutes. Both the test loss and the average loss were much lower at 0.45 and 0.36 for the first module of fine-tuning of the second model–English to French then English to Turkish. We have seen a similar phenomenon in parts 1 and 2 of this section where the loss of the English-Japanese model was much higher even in the first training then the loss of the English-French model.

We were not sure as to why this might have been the case. One plausible explanation is that the mC4 dataset might have had much more or much better, or both, data in French. This is a plausible assumption to make given that French is one of the most widely spoken languages all around the world as well as online–certainly more than Japanese according to our assumptions. Furthermore, if this is indeed the case, our entire causal claim may be put into jeopardy due to this factor acting as a confounding variable. One way around this problem might be to find another language other than French, which the mT5 model have had around the same level of exposure during pre-training given of course that this new language, like French, is grammatically not closely related to Turkish.

In any case, bracketing such concerns until the conclusion section (Section 6) of this paper, we fed in our example sentences into this model and evaluated the translations in context of the other translations from the other three models. While the translations of the model that was fine-tuned on English to Japanese translations were preferable, in most cases, the shortcomings of the translations of the latter model did not stem from their grammatical structures but rather from word choice. In other words, we cannot, by looking at the translations of our example sentences, conclude that the model that was fine-tuned on English to Japanese translations did significantly better than the mT5 model that was fine-tuned on English to French translations with respect to grammatical accuracy.

The following are a selection of sentences translated by the models and the respective translations for the reader's reference:

Models from the first round of fine-tuning (20,000 examples in the train set for both):

- a - Japanese 1
- b - French 1

Models from the second round of fine-tuning (100,000 examples for Japanese/French and 10,000 for Turkish in the train set):

- c - Japanese 2
- d - French 2

Examples:

1. Mary likes John.
   (a) Mary John'u seviyor.
   (b) Mary John'u seviyor.
   (c) Mary, John'u seviyor.
   (d) Mary Mary John'u seviyor.
2. John likes Mary.
   (a) John Mary'yi sever.
   (b) John Mary'yi seviyor.
   (c) John, Mary'yi seviyor.
   (d) John Mary'yi seviyor.
3. I read books.
   (a) Kitaplar okurum.

   (b) Kitap okudum.
   (c) Ben kitaplar okudum.
   (d) Bazı kitapları okudum.

4. I had dinner.
   (a) Akşam yemeği yedim.
   (b) Akşam yemeği yedim.
   (c) Yemek yedim.
   (d) Bir akşam yemeği yedim.

5. It was a magnificent day, rendered mild by the autumn sun.
   (a) O, yaz aylarının ortasına çıkan harika bir gündü.
   (b) Güzücü bir gündü, bahar gününden batar.
   (c) Güzel bir gündü, yaz ışığından dolayı ışıltılı görünüyor.
   (d) Bu harika bir gün, bahar şemsiyesine göre sıcaktır.

6. I often wonder what would have happened to me if I hadn't made that decision.
   (a) Ben sık sık bana ne olmamış olacağımı merak ediyorum.
   (b) Bir zamanlar bana ne olacağını zaman bana ne olmuş olacağını sık sık merak
   (c) Çoğunlukla bana ne olacağını merak ediyorum.
   (d) Ben sık sık bana ne olur olacağını merak ediyorum.

6 Conclusions

The Ural-Altaic language family hypothesis has been discredited within linguistic circles and people have mostly moved away from it. However, in the Republic of Turkey, even today, people have the misconception that their language is grammatically closer to Japanese than any of the neighboring languages. Even though the two languages do not belong to the same language family they do share some grammatical similarities, among which the SOV word order is the most important.

In this experimental study, we strived to create a controlled experiment where we explored the possibility of assessing this conception/misconception using Natural Language Processing (NLP) models and the powerful idea of transfer learning from the same field. In particular, we have fine-tuned two small mT5 models on English to Japanese and English to French tasks respectively and tried to see how they would perform on the task of translating English sentences into Turkish.

We fed the models a selection of simple and complex English sentences and evaluated the quality of translations outputted by each. While both models performed decently on simple sentences and rather poorly on the more complex ones, we did not identify any ways in which the model fine-tuned on English to Japanese translations outperformed the model fine-tuned on English to French translations.

There are likely several reasons for the failure of our model to translate more complex sentences accurately. The first is the lack of scale; this includes a lack of training data. We reduced the size of the training data per language in order to fit our resource constraints, which took three hours per model to train. However, much more time and training data is likely needed in order to produce more accurate translations, which is why our resulting translations were insufficient.

Another reason we may have failed to notice a difference between Japanese-fine-tuned model translation and French-fine-tuned translation is that the models were fine-tuned with sub-optimal ratios of each dataset. The idea of transfer learning to improve machine translation has been mostly applied in the context of low-resource language pairs, where a model would first be trained on a much larger dataset and then trained on a smaller dataset in a different language. However, in our experiment we included a equal number of sentences between Japanese, French, and Turkish in the first training. Additionally, since this number was small, and both models were fine-tuned on the Turkish as well as the original language, it seems possible that the model's translation would not have seen a large difference when fine-tuning on different languages. We tried to adjust this problem by changing the ratio between Japanese/Turkish and French/Turkish datasets in the second training, but did not achieve better results. Further adjustments of the training dataset sizes are likely needed.

7 Future Directions

The first step towards improving upon this project is to improve the overall accuracy of the fine-tuned mT5's translations. This will require increasing the scale of the training; i.e. increasing the size of the translations datasets, which will result in training the model for a longer period of time. After higher translation accuracy is achieved, and the translations can be measured meaningfully through some metric such as BLEU score, we will then be able to compare models fine-tuned in different languages. Another avenue we can explore is using different language and translation models which might achieve better translation accuracy. Given more resources, a bigger model (such as the base version of mt5) may be able to achieve higher accuracy.

After the general translation accuracy is improved, we would like to further explore this topic by fine-tuning the model in languages other than Japanese and French. We chose the two languages based on availability of datasets, but aim to find datasets with varying degrees of relation to Turkish. For example, Arabic and Persian, mentioned previously, have similarities in vocabulary but not in grammar structures. Korean is also proposed to be part of the Altaic family and would make an interesting case study as well.

Additionally, experimenting with other languages which have similar grammar structures (e.g. other Subject-Object-Verb languages) but have no proposed family relation to Turkish could shed insight into the proposed family relationship. As discussed previously, similar grammatical structures do not necessarily imply a ”common ancestor” relationship. Although this problem cannot be solved simply through examination of machine translation results, it is still an interesting proposition worth exploring. This study could also be extended to target languages other than Turkish.

Acknowledgments

We would like to thank Professor Bob Frank and Sophie Hao for their guidance and support throughout this project and course. This project would not have been possible without their time and expertise. We would also like to thank Edan Meyer for guiding us through the process of fine-tuning mT5 through his online resources.

References

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 282–286. The National University (Phillippines).

Google. 2021. Google/mt5-small. Accessed: December 2021.

Alexei S. Kassian, George Starostin, Ilya M. Egorov, Ekaterina S. Logunova, and Anna V. Dybo. 2021. Permutation test applied to lexical reconstructions partially supports the altaic linguistic macrofamily. *Evolutionary Human Sciences*, 3:e32.

Edan Meyer. 2021. Multilingual nmt mt5.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Thomas M. S. Price C. J. Richardson, F. M. 2010. Neuronal activation for semantically reversible sentences. *Journal of Cognitive Neuroscience*, 22(6):1283–1298.

George Starostin. 2016. Altaic languages.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering.

Tatoeba.org. 2021. Tatoeba dataset. Accessed: December 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.