Semantics of Empire: A Neural Machine Translation Approach for Ottoman Turkish Texts

Stanford CS224N Custom Project

Merve Tekgürler
Department of History and the Symbolic Systems Program
Stanford University
mtekgurl@stanford.edu

Abstract

Semantics of Empire presents the preliminary results of developing a neural machine translation (NMT) model to translate Ottoman Turkish (OT) into English. Our research is motivated by the need to enhance the accessibility of primary or historical sources in higher education. The proposed NMT model aims to provide a first-pass translation tool for scholars, facilitating the integration of non-English historical texts into teaching and research, thereby democratizing access to diverse historical accounts. The project also investigates the potential of multilingual NMT for languages with limited resources, using OT as a case study and leveraging its relation to the more resourced modern Turkish. Similarly, OT's status as an extinct language with no possibility of generating new texts make it an ideal candidate for testing sentence alignment techniques for utilizing existing archival and translated materials.

Despite the lack of a dedicated MT system for OT-EN translation, and the considerable linguistic differences between OT and modern Turkish, including vocabulary and syntactic structures, there is a growing interest in leveraging large language models (LLMs) like OpenAI's ChatGPT for initial translation efforts. This paper presents the first structured analysis of the 'emergent capability' of LLMs to translate OT, assessing their effectiveness and reliability. The findings highlight the potential of generative models in translating extinct languages, while also pointing out the limitations and challenges in ensuring accurate and interpretable translations.

1 Key Information to include

- Mentor: Nelson Liu
- External Collaborators (if you have any): N/A
- Sharing project: N/A

2 Introduction

This paper tackles the open problem of engineering a neural machine translation model for translating Ottoman Turkish into English. There are two underlying motivations. One, we aim to increase the availability of primary sources in History education. In the higher education institutions in the US, history is taught by and large only using English source materials. This skews the perception of history and more importantly whose histories we as historians are presenting as histories that everybody should know about. Lack of reliable translations pose a barrier to scholars who desire to teach non-English materials. Even those scholars who read the languages and work with them in their research, do not have the time to create translations for instruction. A first-pass machine translator could lower these barriers by providing scholars with a rough translation that they can edit in a much

short time frame. They can incorporate this into their teaching or use it as a starting point to create authoritative editions.

Secondly, we see the case of Ottoman Turkish as an interesting challenge that has broader implications for other languages. OT is a low-resourced language with a higher-resourced related language, Turkish. As such, it is a great case to test Multilingual Neural Machine Translation Saleh et al. (2021); **?** and transfer learning approaches Zoph et al. (2016); Li et al. (2022); Liu et al. (2023) in related languages. Moreover OT is an historical, extinct language. We cannot produce more new data in OT. The only feasible option is to turn archival documents, manuscripts, novels, and other works along with their translations into data. Thus, researchers are limited to utilizing the existing sources in creative ways. Hence, this project becomes highly relevant for testing the limits of sentence alignment for mining bi-text data.

To the best of our knowledge, there is no MT system specifically designed for OT-EN translation. Current tools for Turkish-English translation are not directly adaptable for this task. The differences between the two languages are not insignificant. OT vocabulary contains high numbers of Persian and Arabic words and uses syntactic forms such as *ezafe*[1] that are no longer common or even present in Turkish. These differences render NMT systems unreliable if not entirely unusable.

Recently, Ottomanists turned to large language models, particularly to OpenAI's ChatGPT for support. We know anecdotally that ChatGPT is considered an excellent first-pass translator. Yet, there is no quantitative analysis of this empirical insight. As this paper shows, these generative models display impressive results, without any specific training in OT. However, they are not always reliable or easily interpretable. Our findings on the 'emergent capability' of LLMs to translate OT offers the first structured critique in this use case.

3 Related Work

This research project is at the intersection of historical NLP, Digital History, neural machine translation, and NLP research on low-resourced languages. By historical NLP, we are referring to works like those on Coptic (Enis and Megalaa) or Latin (Martínez Garcia and García Tejedor, 2020) that study these historical languages within the field of NLP. The use of NLP methods in History research has increased in the recent years Jo (2020); de Bolla (2023); Guldi (2023). Our work recognizes the value that computational approaches add to History scholarship. At the same time, we argue that Digital History, much like NLP has a bias towards English. Non-English languages are extremely underrepresented in this field. Thus, we see similarities between our work and those of NLP researcher studying other non-English languages Doumbouya et al. (2023). Specifically, this paper deals with two distinct yet intimately related tasks: sentence alignment and machine translation. Statistical MT research developed concurrently with sentence alignment efforts. The first sentence alignment work by Gale and Church (1991) used the same corpus, Canadian Hansards[2], as the seminal SMT paper by Brown et al. (1990). As such, it is intuitive to envision these two tasks within the scope of the same project.

3.1 Sentence Alignment

Sentence alignment is the is the task of finding matching sentences in two parallel documents (Steingrimsson et al., 2023). A sentence alignment algorithm parses through parallel texts calculating the similarity of sentences in the source text with the those in the target set to determine which sentence or sentences correspond to one another. Often used in the context of MT, sentence alignment could be a one-to-one match, meaning one sentence in the source text is translated as exactly one sentence in the target text or one-to-many, many-to-one, many-to-many or a sentence might be omitted in the translation or a new sentence might be inserted by the translator. This complexity makes sentence alignment a challenging task in NLP since the statistical MT era.

Gale and Church (1991) introduced a method to align sentences based on a correlation of character lengths between a paragraph and its translation. Hunalign (Varga et al., 2005) expands upon this work by incorporating a lexicon and a token based approach searching for shared words in parallel

[1]This is a grammatical particle that links two words, most commonly for the possessive case. For further information, please refer to `https://en.wikipedia.org/wiki/EzÄĄfe`.

[2]Canadian Hansards are the bilingual French-English records of the Canadian parliamentary proceedings.

sentences. Bleualign (Sennrich and Volk, 2011) reimagined sentence alignment by incorporating a translation step and calculating the similarity between the target text and the translation of the source text, now in the same language as the target text. Xu et al. (2015) stand out as one of the earliest works to identify the potential of sentence alignment for creating MT datasets from literary works.

Entering the era of neural NLP, Thompson and Koehn (2019) developed VecAlign, using Language-Agnostic SEntence Representations (LASER) (Artetxe and Schwenk, 2019) for calculating the semantic similarity of sentences. The algorithm operates recursively and in linear time. LASER embeddings are a bidirectional long-short-term memory (BiLSTM) based word embedding method. VecAlign architecturally does not depend on LASER and can be used with any float32 embedding. The neural alignment trend is followed by Bertalign (Liu and Zhu, 2022), which uses Language-agnostic BERT Sentence Embedding or LaBSE Feng et al. (2022). The goal of the authors is to create parallel data of Chinese-English literary texts. They develop a two part algorithm that first identifies one-to-one matches and then aligns the rest of the text using them as anchors. The other system that uses LaBSE is SentAlign Steingrimsson et al. (2023). SentAlign uses Dijkstra's algorithm for optimal path finding. In context of sentence embeddings, language-agnostic means that these embeddings cluster sentences in multiple languages in the same vector space based on semantic similarity of content. Regular multilingual embeddings, like XLM-RoBERTa often cluster texts roughly by language regardless of the meaning of the sentences. While there are strong and accurate criticism to this claim of language agnosticity (Chen and Avgustinova, 2021), LaBSE performes very successfully in tasks related to cross-lingual semantic similarity (Chimoto and Bassett, 2022).

3.2 Neural Machine Translation

Neural machine translation is a sequence-to-sequence task that encodes a sentence in the source language and decodes its translation in the target language. NMT is a wide and diverse field of study. Under the umbrella of NMT, there are three tangents that are related to our project. The first one is the question of text domain in translation, second transfer learning, and third multilingual NMT.

Text domain presents a challenge for the deployment of NMT systems. A model trained on parliamentary record, news, Internet content and other contemporary texts do not perform well on historical or literary data. More importantly, increasing the out-of-domain data does not help increase the model performance on in-domain data (Luong and Manning, 2015). Wang et al. (2017) propose an adaptation through data selection that scores the similarity between out-of-domain and in-domain data to identify which sections of the dataset can be used effectively.

Transfer learning is a technique in machine learning where a model developed for a specific task is reused as the starting point for a model on a second task, leveraging the knowledge gained from the first task. Zoph et al. (2016) is the first paper that applies transfer learning in NMT. Low-resourced languages suffer from a lack of existing data to train NMT models, which the authors ameliorate with transfer learning using parent-child models. They a parent model with a high-resourced language pair and use the parameters of that model in the initialization of the child model. More recent works utilize the power of transfer learning at every stage of the child model's training. ConsistTL (Li et al., 2022) maintains a prediction consistency between the parent and child models by constructing semantically equivalent instances for the parent model during the child's training. kNN-TL (Liu et al., 2023) utilizes k-nearest-neighbor (kNN) techniques to leverage parent model knowledge throughout the entire development process of the child model.

Multilingual NMT refers to a single NMT system that is capable of translating between multiple languages. Bala Das et al. (2023) develops a MNMT system for multiple Indic languages with a shared encoder-decoder architecture. The model handles multiple language pairs simultaneously, facilitating efficient knowledge transfer and resource sharing across the languages. Saleh et al. (2021) points out that knowledge from unrelated languages can degrade translation performance (negative transfer) and proposes a hierarchical structure where languages are clustered based on linguistic typology and phylogeny.

4 Approach

4.1 Sentence Alignment

For this task, we used SentAlign based on its superior performance compared to other systems discussed in the Experiment section. We used LaBSE and retained the suggested settings including setting 0.4 minimum for acceptable similarity score in an aligned pair of sentences. We did not change anything with this codebase.

4.2 Neural Machine Translation

NMT training predominantly focused on fine-tuning an existing NMT model for Turkish-English translation. We chose the Opus-MT model developed by Helsinki NLP group, which can be accessed on Hugging Face. This model is licensed under Creative Commons which allows the free use and further development of this model for educational purposes. Helsinki NLP developed this model within the framework of the Tatoeba Challenge (Tiedemann, 2020) in 2021. We fine-tuned the March 2022 updated version of the model. Further details about the Helsinki model can be found on their GitHub repository.

During the fine-tuning steps, we used Hugging Face Seq2SeqTrainer. The Trainer class allows for a feature-complete training of Hugging Face models in PyTorch without having to manually code a training loop. Thus, the training process is optimized and streamlined, which preserves time and resources. Moreover, we are easily able to set important model configurations such as batch-size and gradient accumulation.

We experimented with 3 fine-tuning approaches. The first two were trained on the dataset detailed below. The third approach followed the conceptual premise of multilingual training. We merged the two test sets with our training data and fine-tuned the same model. We left out our the Osman Aga manuscript as the test case. Our baselines are the original Helsinki NLP model, GPT-3.5 Turbo, GPT-4 Turbo, Gemini and Cohere Aya. We ran two tests with Gemini, one with the standard safety settings and one with safety settings entirely disabled.

5 Experiments

5.1 Data

We hand-curated a dataset for this project. Our goal is to build a neural machine translation (NMT) system for Ottoman Turkish (OT) that can be used as a first-pass translator. OT is an historical, low-resourced language with a higher-resourced related language, Turkish. The main challenge with the existing NMT Turkish-English datasets is vocabulary. Compiled from news sources, European Union publications, and other contemporary materials (Tiedemann, 2020), these datasets are limited in their overlapping vocabulary with OT. Thus, we turned to novels in Ottoman and Modern Turkish with English translations to create more parallel data with ground truth translations.

We identified 13 novels (see Appendix) with translations and acquired text files using ABBYY FineReader 15 for PDFs and ebooklib for epubs. NMT datasets contain sentence pairs where the target sentence in this case English, is the translation of the source sentence, Turkish. We split novels into sentences and align them at sentence level using the SentAlign algorithm (Steingrimsson et al., 2023). After the alignment, we developed a heuristic to determine which sentences pairs will be in the final dataset. While the alignment removed majority of the sentences with OCR errors and content like commentaries in one language not found in the other, there were still some erroneous sentences. We cleaned all sentences that do not contain at least 2 letters in either language. SentAlign lists similarity scores for each sentence pair. Our initial alignment process only accepted those sentence pairs with 0.4 or higher similarity scores. We decided on further cleaning more upon a closer investigation of random samples from the dataset. For each novel, we clustered the sentences into 10 clusters using k-means clustering and removed those sentence pairs that have a lower similarity score than that of the the 3rd lowest ranked centroid. We applied this approach dynamically to all novels. Finally, we merged all the novels together and shuffled all the sentences and created train and validation sets with 80/20 divide using scikit-learn.

For testing purposes we set aside one Ottoman novel from 1875, which contains language more similar to the training data and one Ottoman manuscript from 1781, which is an example of what we aim for our final model to specialize on. The novel is like a bridge between Ottoman and Modern Turkish. We also identified one more Ottoman manuscript called Osman Ağa, which we processed in the same way and set aside for testing the multilingual model, which we trained on the novel and manuscript test sets merged with the training set. Osman Ağa was written in 1724. Having read all 3 test sets, we believe that the language of the novel is closest to Turkish, with the manuscript from 1781 in the middle and Osman Ağa the least similar.

Table 1: Dataset Statistics

Dataset Name	Number of Sentence Pairs
Train Set	41,782
Dev Set	10,447
Test Set 1: Novel	2,694
Test Set 2: Manuscript	425
Ottoman Train and Dev Set	3,323

5.2 Evaluation method

5.2.1 Sentence Alignment

We tested the performance of three sentence alignment approaches, SentAlign, VecAlign, and Negar92 on one book chapter. We created the ground truth alignment data manually and then extracted the match indexes. We evaluated the performance of the system based on reading the output and by running an evaluation script that was a part of the VecAlign repository. This evaluation script reports precision, accuracy, and F1 scores.

5.2.2 Neural Machine Translation

We used Bilingual Evaluation Understudy or BLEU scores (Papineni et al., 2002) and character n-gram F-score or chr-F (Popović, 2015). The differences in translation practices with regards to person and placenames could be causing lower BLEU scores, when even if the model translation is not wrong. For example, Ottoman Temeşvar in modern-day Romania can be translated as Timișoara using its current name or as Temesvár using its historical Hungarian name or just retained as Temeşvar. We should account for these complexities when evaluating historical translation with chr-F scores. We also read through a sample of the model translations ourselves as the expert.

5.3 Experimental details

5.3.1 Sentence Alignment

We extracted one chapter of one novel in Turkish and English translation and divided it up into sentences. We aligned the chapter first with Negar92 with the notebook on their GitHub repository, then with SentAlign using the scripts on their GitHub, and finally with VecAlign using their GitHub as well as Google Colab GPU runtime for embedding the sentences. Since the first two systems use LaBSE, we tested VecAlign both with LASER and LaBSE.

5.3.2 Neural Machine Translation

The first fine-tuning experiment used a batch size of 4 and trained for 3 epochs. We evaluated the model once per epoch based on loss. The training took 2 hours, 31 minutes, and 50 seconds on one Nvidia T4 GPU. By the end of the training evaluation loss was stable around 2.16 and not increasing despite the constant decrease of loss throughout the training, reaching as low as 1.63. Hence, we adjusted the hyperparameters in the second fine-tuning round. We set an early stopping condition based on the BLEU scores evaluated every 1000 steps and increased the epoch size to 6. BLEU scores are computationally expensive to calculate and this new approach resulted in out-of-memory errors. We reduced the batch size to 2 and incorporated gradient accumulation of 2, to re-approximate the batch size of 4. We also adjusted our code so that the BLEU score calculation uses CPU instead

of the GPU memory. The training took 3 hours, 20 minutes, and 33 seconds and stopped early at 2.3 epochs. Controlling for the BLEU score revealed that although the evaluation loss was not increasing, BLEU scores started decreasing around 2.2 epochs. For the third fine-tuning experiment, we decided to explore multilingual training in more detail. We merged the test sets novel and manuscript with our training data and thereby created a mixed Ottoman-Modern Turkish dataset. We trained the model with this data using the same configurations as the first fine-tuning experiment (batch size 4 and 3 epochs). There was an issue with the virtual machine that unfortunately resulted in the loss of the training logs. However we evaluated this model on the Osman Aga manuscript, which we report below.

5.4 Results

5.4.1 Sentence Alignment

Based on the results listed below and a qualitative evaluation of the alignment output[3], we decided to use SentAlign to create our novel corpus. We were particularly surprised by how inefficient the Negar92 algorithm was. It took longer to align of chapter with 72 source and 80 target sentences with Negar92 than SentAlign to align the entire book with 8881 source and 9830 target sentences. Since both approaches use LaBSE, the differences in search algorithm implementation must have caused these results. We ran two comparisons with VecAlign (Thompson and Koehn, 2019), using LASER and LaBSE. LASER yielded worse results than LaBSE which is consistent with our review. LaBSE is considered generally as the state-of-the-art 'language-agnostic' embedding option. Moreover Vecalign was faster than the other two systems. After the embeddings are acquired, it took less than seconds to align the chapter. However, VecAlign was very complicated to set up and it requires that embeddings for overlapping sentences are acquired externally. We created the overlap text files using Vecalign, transfered those files to GPU runtime on Google Colab, calculated embeddings and retransfered them to Vecalign to run the alignment step. Also, VecAlign does not return a text file of aligned sentence pairs, only their indexes. This could be improved upon with the use of other repositories such as the one by (Forgac and Kelebercova1, 2023). Ultimately, we found it simpler to use SentAlign which directly produces the aligned output from text files.

Table 2: Alignment Performance Metrics

Model	Precision	Recall	F1 Score	Runtime on CPU
Negar92	0.620	0.508	0.559	4h 20min 14s
SentAlign	0.902	0.902	0.902	1min 42s
VecAlign (with LASER)	0.708	0.754	0.730	ca. 1min*
VecAlign (with LaBSE)	0.887	0.902	0.894	ca. 20s*

* This includes the time to run the embedding model on both files.

5.4.2 Neural Machine Translation

The NMT training showed promising results. Table 3 shows the BLEU and chr-F scores for each fine-tuned model and the baseline models tested on the 3 test sets. Manuscript, Novel, and Osman Ağa were used for all models. Since we used the Manuscript and Novel for fine-tuning the multilingual model, we only reported its performance on the Osman Ağa. We observe a steady increase in the chr-F scores for Osman Ağa from the Helsinki NLP baseline through the fine-tuning. More importantly, the 1.04 increase in the BLEU score from the Helsinki NLP baseline (2.83) to the multilingual fine-tuning (3.87) is encouraging. Even with an off-the-shelf fine-tuning approach, we were able to increase the model's performance. These scores are bold in Table 3.

Moreover, by fine-tuning the model on a corpus of novels, we reached BLEU scores higher than the Helsinki NLP and Cohere Aya baselines and comparable to Gemini. The scores are italicized in Table 3. There were some challenges with acquiring a baseline from the Gemini model. We used the latest model through API calls and discovered that the model refuses to translate some sentences due to safety concerns. Instead of translating the sentence, the model returned a JSON that evaluated the safety based on 4 factors: harassment, hate speech, sexually explicit, and dangerous content. 6.54% of Novel, 9.67% of Manuscript and 13.69% of Osman Ağa were not translated. Table 4 in the

[3]Alignment results with full sentences can be found here.

Table 3: BLEU and chrF Scores

Model	Novel		Manuscript		Osman Ağa	
Evaluation Metrics	BLEU	chrF	BLEU	chrF	BLEU	chrF
GPT-4	11.68	39.47	9.75	41.41	7.97	37.71
GPT-3.5	11.14	38.09	8.23	38.58	7.11	35.84
Gemini (no safety)	*11.11*	37.38	9.04	39.04	7.84	36.60
Gemini*	*10.97*	37.25	9.06	39.55	7.85	36.61
Fine-tune (v1)	*10.94*	33.52	3.29	23.24	**2.78**	**20.07**
Fine-tune (v2)	*10.62*	33.07	3.31	23.47	**2.85**	**20.16**
Cohere Aya	*10.29*	33.92	5.46	29.52	5.74	28.91
Helsinki NLP	*9.74*	33.25	3.44	22.21	**2.83**	**19.39**
Multilingual Fine-tune †	–	–	–	–	**3.87**	**24.23**

* We omitted the empty translations from the calculations.
† No scores reported because we trained this model on Novel and Manuscript.

Appendix shows this in more detail. Comparing this with the version of the model with no safety settings, we see that there are 5 sentence in Novel, 1 sentence in Manuscript and 3 in Osman Ağa. These sentences did not trigger the safety settings of the model and we cannot interpret why they might not have been translated based on reading them alone.

Finally, we identified an unusual trend regarding the differences between chr-F and BLEU scores. Figure 2 in Appendix contains two plots that show the BLEU and chr-F scores for each model, same as in Table 3. By plotting these scores, we demonstrate that GPT-4, GPT-3.5 and Gemini behave differently from the Helsinki NLP model and the fine-tuned translation models. As expected all BLEU scores decrease with increasing difference of the test texts from Modern Turkish. However for the 3 LMMs, their chr-F score for manuscript is higher than their chr-F for novel. Typically, we expect both the BLEU scores and chr-F to decrease on the same test because this indicates overall poorer performance. We do not have a straightforward explanation for this. Relatedly, Cohere Aya was the only model to score higher on BLEU score for Osman Ağa than for Manuscript.

6 Analysis

In this section, we will only reflect on the performance of our NMT training evaluated by the author as the expert in Ottoman Turkish reading the model outputs. Our NMT system faced some challenges with the dataset. Sentence alignment included many-to-many, one-to-many, and many-to-one matches. This means that some of the training data, although aligned at sentence level, was not sentences in the sense that they did not end in punctuation. The model was originally trained on data that contained shorter sentences and upon inspection, we did not see any longer source-target pairs. This means that the model learnt strongly to predict an end of sentence token after seeing a punctuation. There were multiple predictions where the translation was cut short. The model only translated the first sentence in italic below.

Novel: *İş bu iki genç tayfa kürek oturaklarına oturdukları zaman Ziklas dümen yekesini ele alır ve Râkım ile Misters Ziklas dahi karşı karşıya oturur idi.* Yelken fora etmek, yelken sarmak iki nefer genç tayfaların hizmeti olup şâyet kürek çekmek icâb eder ise o zaman dümen yekesini Misters Ziklas eline alıp kocası ile Râkım ve iki nefer güzel tayfalar dahi küreğe geçerler idi.

Fine-tune V2: When the two young men sat on the rowing couch, Ziklas took charge of the rakı and even Rakım and Misters Ziklas sat opposite each other.

Sometimes the model was able to translate only a part of the text accurately and then it diverged in topic. Below, the italicized part is accurately translated but the rest of the source sentence is missing and the model prediction is entirely inaccurate. However, this is curious because the source sentence contains the word Davudpaşa, which is a place name but Davud is the Ottoman version of the English name David, today used predominantly in reference to the story of the King David. Thus, there is some relationship between the prediction and the source text.

Manuscript: *İstanbul'a geldiğinin onuncı güni* sefer-i mezkûr maslahatı içün huzûr-u hümâyûnda müşâvereye şürû' olunmağın hiç hastelikden haberi olmamak sûretinde pâdişâh karşusunda iki sâ'at koçan gibi oturub mizâc-ı hümâyûna muvâfık sadedlerle izhâr-ı şecâ'at ve düşmam gözine kesdirüb bir gün evvel Davudpaşa sahrasına çıkmağa 'azimet ve Cemaziyelevvel'in yirmi beşinde yine huzûr-u şehriyârîde sefere fâtiha okudtırub Moskov balyozı Abreşkof Yedikule'ye gönderilmekle 'ilân-ı cenk ve cidâla himmet eyledi.

Fine-tune V2: *On the tenth day of his visit to Istanbul,* he was sent to the mosque to read the story of David, and on the second day he said, "I mustn't forget about it.

Some errors stem from not having seen enough ezafe constructs. The model is not able to accurately tokenize the - and decode it as the possessive syntactical marker.

Manuscript: *Bervech-**i** muharrer Tatar taifesi simât-**ı** şehriyârî* ile muğtenim-i safâya ta'yîn olunub ve reis efendi balyoz ile muhal kabilinden olan timûn mûm itmeğe ikdam idüb himmet ve gayretle bitmez iş olmaz laklakasıyla minâreye dışaradan çıkmağa çalışub iş bitmezse de makamında pâydâr olmağı ganimet 'add idüb riyaseti gerçi üc seneye yaklaşdırdı.

Fine-tune V2: The Tatars of Bervech-i-Mutari-i-de-Tayar-i-i-Tayar-i-Tayar-i-

There are also examples where the model was to some extend already doomed to fail. In this example below, it is not immediately clear who is the subject of the verb to see. Please note the nickname Old Osman in gold translation, which was the translators decision. Since this is a literary work, these nuances are difficult to represent at sentence level. All models including GPT-4 fail at translating this sentence accurately and clearly. GPT-3.5 came close to it but only by leaving the object of seeing ambiguous in English. Did Osman see a gray hair in his beard or seeing that he has a gray beard, the other conversation partner is thinking that he is old? GPT-4 captures this nuance but it misses the mark on the object of the verb to think. Uncle Osman is referring to himself in the third person, which is lost in GPT-4 translation.

Novel: Osman amca - Sen Osman amcanın sakalında kır görüp de ihtiyar olmuş mu zannediyorsun?

Ground Truth: old osman: Just because you see some grey in his beard, do you think that Old Osman has become an old man?

Helsinki Baseline: Uncle Osman - Do you think Uncle Osman saw a break in his beard and became an old man?
Fine-tune V1: Uncle Osman-Do you think Uncle Osman is old when he sees gray in his beard?
par **Fine-tune V2:** Uncle Osman-Do you think he's grown old when he sees gray on Uncle Osman's beard?
Cohere Aya: Osman uncle - You think Osman uncle has gray hair and is old?
Gemini: Uncle Osman - Do you think you have seen white color in Uncle Osman's beard and he has become too old?
GPT-3.5: Uncle Osman - Do you think that Uncle Osman has turned old seeing a gray hair in his beard?

GPT-4: Uncle Osman - Do you think he has become old just because you see gray in Uncle Osman's beard?

However on the following examples the model performed really nicely. The first example contains an unusual spelling of the verb to say, disek instead of desek, and they the model was able to translate it. Similarly in the second one deyü is used instead of diye but it did not cause any translation errors.

Manuscript: Moskov üzerine seferimiz var *disek* evvelki gibi olur.

Osman Ağa: Beni öldürdünüz *deyü* taşra çıkmağa korkarlar.

Fine-tune V1: If we said we had a campaign on Moskov, it would be just like before.

Multilingual Fine-tune: They're afraid of going out into the countryside because you killed me.

7 Conclusion

The Semantics of Empire paper discussed the preliminary process of creating a dataset and training an neural machine translation (NMT) model for Ottoman Turkish. Our results demonstrate that it is indeed possible to train a model that competes with existing options. The relative success of our approach stems from the fact that we have curated a dataset that brought an out-of-domain Turkish-English translation model closer to our in-domain texts. Moreover, we were able to leverage powerful fine-tuning pipelines on Hugging Face and use an existing model trained on more parameters than we would have been able to with our limited resources. We believe that training an NMT model is an iterative task. Our aim is to expand our corpus in collaboration with other scholars in the field

and experiment with more fine-tuning configurations, as to eventually implement a transfer learning based model.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Sudhansu Bala Das, Atharv Biradar, Tapas Kumar Mishra, and Bidyut Kr. Patra. 2023. Improving multilingual neural machine translation system for indic languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(6).

Peter de Bolla. 2023. *Explorations in the Digital History of Ideas: New Methods and Computational Approaches*. Cambridge University Press, Cambridge.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Yu Chen and Tania Avgustinova. 2021. Are language-agnostic sentence representations actually language-agnostic? In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 274–280, Held Online. INCOMA Ltd.

Everlyn Chimoto and Bruce Bassett. 2022. Very low resource sentence alignment: Luhya and Swahili. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 1–8, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Moussa Koulako Bala Doumbouya, Baba Mamadi Diané, Solo Farabado Cissé, Djibrila Diané, Abdoulaye Sow, Séré Moussa Doumbouya, Daouda Bangoura, Fodé Moriba Bayo, Ibrahima Sory 2. Condé, Kalo Mory Diané, Chris Piech, and Christopher Manning. 2023. Machine translation for nko: Tools, corpora and baseline results.

Maxim Enis and Andrew Megalaa. Ancient voices, modern technology: Low-resource neural machine translation for coptic texts. In *Coptic Translator*, pages 1–15.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Munkova Dasa Munk Michal Forgac, Frantisek and Livia Kelebercova1. 2023. Evaluating automatic sentence alignment approaches on English-Slovak sentences. *Scientific Reports*, 13(1).

William A. Gale and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, ACL '91, page 177–184, USA. Association for Computational Linguistics.

Jo Guldi. 2023. *The Dangerous Art of Text Mining: A Methodology for Digital History*. Cambridge University Press, Cambridge.

Eun S. Jo. 2020. *Foreign Relations of the United States Series, 1860-1980: A Study in New Archival History*. Ph.D. thesis, ProQuest Dissertations and Theses. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-06-21.

Zhaocong Li, Xuebo Liu, Derek F. Wong, Lidia S. Chao, and Min Zhang. 2022. ConsistTL: Modeling consistency in transfer learning for low-resource neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8383–8394, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lei Liu and Min Zhu. 2022. Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634.

Shudong Liu, Xuebo Liu, Derek F. Wong, Zhaocong Li, Wenxiang Jiao, Lidia S. Chao, and Min Zhang. 2023. kNN-TL: k-nearest-neighbor transfer learning for low-resource neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1891, Toronto, Canada. Association for Computational Linguistics.

Minh-Thang Luong and Christopher Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam.

Eva Martínez Garcia and Álvaro García Tejedor. 2020. Latin-Spanish neural machine translation: from the Bible to saint augustine. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 94–99, Marseille, France. European Language Resources Association (ELRA).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Fahimeh Saleh, Wray Buntine, Gholamreza Haffari, and Lan Du. 2021. Multilingual neural machine translation: Can linguistic hierarchies help? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1313–1330, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rico Sennrich and Martin Volk. 2011. Iterative, MT-based sentence alignment of parallel texts. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 175–182, Riga, Latvia. Northern European Association for Language Technology (NEALT).

Steinthor Steingrimsson, Hrafn Loftsson, and Andy Way. 2023. SentAlign: Accurate and scalable sentence alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 256–263, Singapore. Association for Computational Linguistics.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Nagy, László Németh, and Viktor Tron. 2005. Parallel corpora for medium density languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2005)*, pages 590–596, Borovets, Bulgaria. INCOMA Ltd.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 560–566, Vancouver, Canada. Association for Computational Linguistics.

Yong Xu, Aurélien Max, and François Yvon. 2015. Sentence alignment for literary texts: The state-of-the-art and beyond. *Linguistic Issues in Language Technology*, 12.

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.

A Appendix

A.1 Figures and Tables

Table 4: Gemini: Translation vs Safety

Test Set	Total Sentences	Not Translated	Percentage Not Translated
Manuscript	424	41	9.67%
Novel	2,693	176	6.54%
Osman Ağa	628	86	13.69%

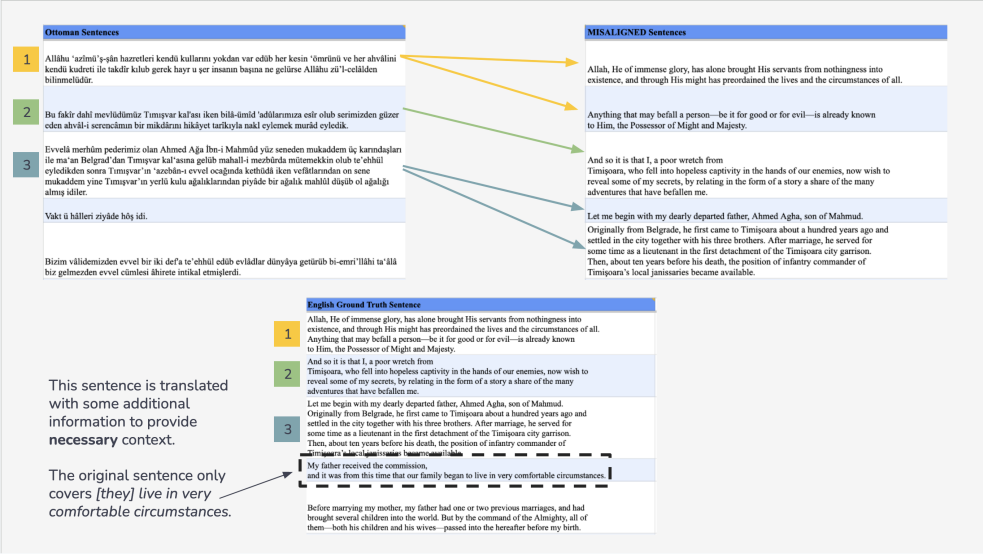

Figure 1: Problem with Sentence Alignment

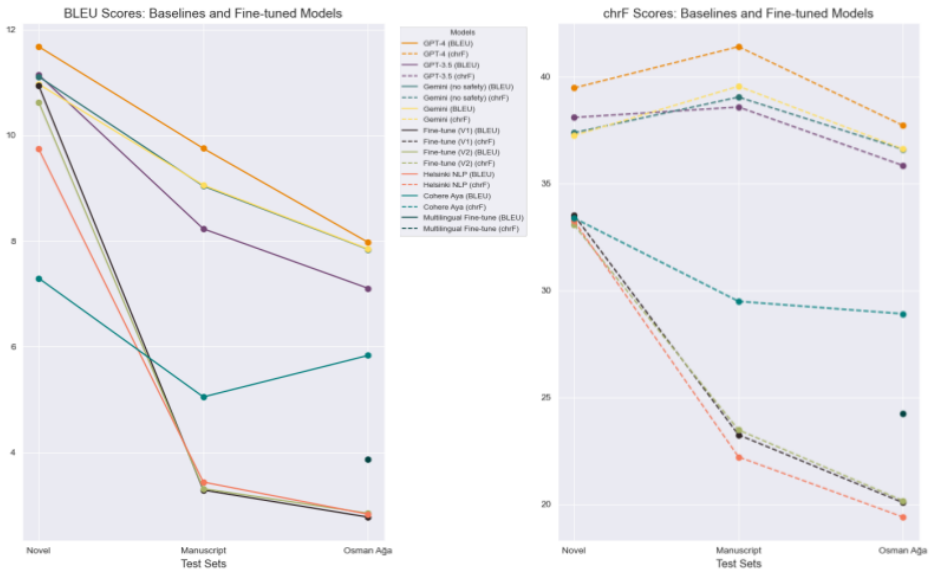

Figure 2: BLEU and chr-F Scores for Test Set per Model

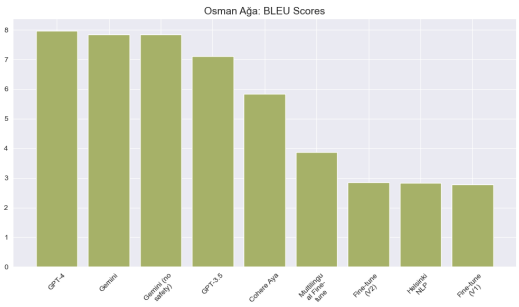

Figure 3: Osman Ağa: BLEU Scores

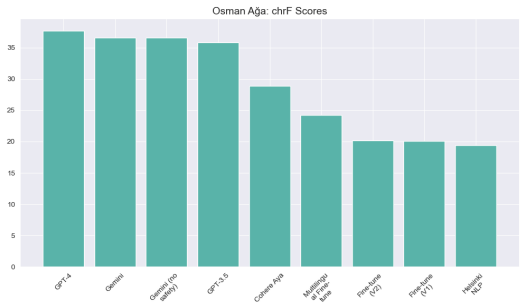

Figure 4: Osman Ağa: chr-F Scores

A.2 Novel Corpus

The 13 novels and 2 manuscripts (first two items) listed below were selected based on a combination of factors: personal domain knowledge, support of Eyüp Eren Yürek, friend of the author and a scholar in Comparative Literature, and this bibliography. We believe that data acquisition is an iterative process and intend to increase the number of novels in our pipeline reflecting on the results of the model training.

Since most of these novels and almost all the translations are under copyright, we cannot not publish the dataset in its entirety. Instead, we published a small (10 percent of the dataset) random sample of sentences as permitted by fair use laws for publication.

1. Osman Ağa. Memoirs (1724). tr. Prisoner of Infidels (2020)
2. Ahmed Resmi. Hulâsatül-itibâr (1781). tr. A Summary of Admonitions (2011)
3. Namık Kemal. Intibah (1876). tr. The Awakening (2018)
4. Ahmet Mithat Efendi. Felatun Bey ile Rakım Efendi (1875). tr. Felatun Bey and Rakim Efendi (2016)
5. Halide Edib Adıvar. Sinekli Bakkal (1936). tr. Clown and His Daughter (1935)
6. Ahmet Hamdi Tanpınar. Huzur (1946). tr. A Mind at Peace (2011)
7. Yaşar Kemal. İnce Memed (1955). tr. Memed, My Hawk (2016)
8. Ahmet Hamdi Tanpınar. Saatleri Ayarlama Enstitusu (1961). tr. The Time Regulation Institute (2013)
9. Halide Edib Adıvar. Türk'ün Ateşle İmtihanı (1962). The Turkish Ordeal (1928)
10. Orhan Pamuk. Benim Adım Kırmızı (1998). tr. My Name Is Red (2010)
11. Mario Levi. Istanbul Bir Masaldı (1999). tr. Istanbul Was a Fairy Tale (2012)
12. Orhan Pamuk. Beyaz Kale (1985). tr. The White Castle (1998)
13. Ayşe Kulin. Nefes Nefese (2002). tr. Last Train to Istanbul (2006)
14. Ayşe Kulin. Umut (2008). tr. Love in Exile (2016)
15. Elif Şafak. Baba ve Piç (2016). tr. The Bastard of Istanbul (2018)