Computational Linguistics in the Netherlands Journal 14 (2025) 241-254 Submitted 12/2024; Published 07/2025

Location-focused translation of flooding events in news articles

Suzan Lejeune* SUZAN.LEJEUNE@RU.NL
Iris Hendrickx* IRIS.HENDRICKX@RU.NL

* *Centre for Language Studies, Centre for Language and Speech Technology, Radboud University, Nijmegen, Netherlands*

Abstract

We are interested in the automatic extraction of information on flooding events in the Philippines from local news papers. Given that the majority of existing information extraction tools have been developed for English, this study aims to investigate the feasibility of using open-source machine translation (MT) tools to translate Tagalog news items to English. Extra care should be taken when translating location names, as precise location information is indispensable for effective disaster management. We fine-tuned an open-source multi-lingual MT model for disaster news in Tagalog. We investigated several methods to enhance the model performance on location translation and evaluated the different versions to compare the translation quality of locations using a custom location-focused evaluation metric. To this end, two new Tagalog-English datasets specific to the domain were introduced for the purposes of fine-tuning and evaluation. We tested out fine-tuning on domain specific data and two masking techniques using either general masks or database-look-up of names. Contrary to our expectations, our findings show that the base open-source multi-lingual MT model was already proficient in location translation. Our analysis indicates that fine-tuning on domain-specific data improves overall machine translation quality. Our manual analysis provides insight into specific errors of location translation and the unique effects of the fine-tuning techniques.

1. Introduction

The Philippines is a country regularly affected by hurricanes. This leads to the flooding of larger cities and local towns alike, which threatens people's lives, impacts their health and livelihoods, and causes damage to infrastructure (Alcantara 2019). As certain areas are more prone to be affected by hurricanes than others, disaster managers want to know where these risk areas are located for effective future aid in impacted areas. One potential source for collecting fast and up-to-date information about flooding sites is the media. As the Philippines has two official languages, English and Tagalog, part of the news media coverage is only available in Tagalog. Current automatic tools for flooding event and location extraction have mainly been developed for English (Nasar et al. 2021). Retraining these tools for another language is often costly. A more sustainable solution is to automatically translate Tagalog news coverage of flooding events into English.

Machine translation (MT) research is a well-established field with a rich history (Hutchins 2023). The translation of named entities such as locations is of utmost importance as it impacts the quality of the translation greatly (Hassan et al. 2018). Consider the Filipino location *'Brgy. Maliwanag'*. If this is erroneously translated to *'Brgy. Clearly'*, any sense of location will be lost, thus clearly impacting our understanding of the meaning.

Tagalog is an Austronesian language that uses the Latin script. The syntactic sentence structure typically follows a Verb–Subject–Object (VSO) word order, though with some flexibility. It has a complex focus system that alters verb forms depending on whether the emphasis is on the actor, object, or location (Brown and Ogilvie 2010). Key translation challenges include word order differences, inflections, capitalization and lack of direct equivalents for cultural terms, making precise

translation to English difficult (Langga and Alico 2020). Tagalog additionally has the disadvantage of being a low-resource language. As such, the classic pretrain-fine-tune MT approach that works for high-resource languages will not work as effectively for Tagalog as a result of the lack of data (Park et al. 2020). Instead, multilingual MT models can be used to leverage knowledge from different languages to translate to low-resource languages (Conneau and Lample 2019). By fine-tuning on parallel English–Tagalog data, such models can acquire more knowledge from the process compared to their bilingual counterparts.

Since our goal is to translate Tagalog news coverage of flood events into English — so that we can eventually apply existing event extraction tools — our research question in this study is as follows:

> *"How does fine-tuning a multilingual MT model influence the translation quality of locations for Tagalog?"*

We hypothesize that fine-tuning on domain-specific flood-related data will increase the translation quality of locations when translating flood-related news. In addition, we expect that data augmentation will improve the translation quality of location names in flood-related news. Since many location names are rarely used in news, the MT model should not translate them into the target language but instead preserve them as written in the source language. We aim to train the MT model to retain these names by augmenting the data with masking strategies. We test this hypothesis by comparing MT models trained on varying amounts of data, as well as data augmented using various techniques.

2. Related Work

The pretrain-fine-tuning approach has been widely used for improvement in various NLP techniques (Devlin et al. 2019, Yang et al. 2019, Raffel et al. 2020). However, large datasets are necessary for this approach to work effectively, which are not readily available for low-resource languages. To amend this problem, multilingual MT models can be used. The goal of multilingual MT is to translate between any language pair. As these models are trained on multiple languages, information between similar languages can be shared, thus benefiting the translation of low-resource languages (Aharoni et al. 2019). However, training on multiple languages means that model capacity must be divided over all languages. For this to work well, the model capacity must in turn be larger compared to bilingual models.

Previous work has shown that such multilingual MT models can be used effectively for translating to low-resource languages by applying optimization techniques such as Sparsely Gated Mixture of Experts (MoE) (Almahairi et al. 2016, Bengio et al. 2013), self-supervised learning (Bapna et al. 2022, Chi et al. 2021, Ma et al. 2021), or backtranslation (Edunov et al. 2018, Sennrich et al. 2016). In this study we use a recent multilingual MT model that uses MoE as optimization technique and has state-of-the-art performance: No Language Left Behind (NLLB) (Costa jussà et al. 2022).

Entity translation

Named entities such as persons, locations, and organizations carry important meaning in a text. Automatic translation of these named entities poses a significant challenge for machine translation systems (Xie et al. 2022). This is due to the fact that many names are infrequent occurrences connected to particular topics or specific to a certain time or location. In our case, location names related to flooding events serve as a good example of such low-frequent time- and space-specific entities.

MT systems in general either treat named entities the same as any other word in the text to be translated, or they apply an additional step in the translation process to first tag entities in the text (Xie et al. 2022). In this additional first step, named entity recognition (NER) is used to extract

the entities, which are then tagged by either replacing them with placeholders (Wang et al. 2017), indicating the boundaries of the entities (Li et al. 2018), or labeling the entities with an entity embedding (Ugawa et al. 2018). Note that this extra step of adding entity tags is applied in both the training and testing phase. In the testing phase, the input text is first processed with NER tools to extract and tag the entities before translation of the sentence, and afterwards a post-processing step is needed to remove the added tags and get the final translation result (Xie et al. 2022). However, since this is a two-step approach and the NER tool might not be completely accurate, this can lead to cascading errors (Huang et al. 2003, Huang et al. 2004, Lambert et al. 2011).

In our study, on the other hand, we follow a data augmentation approach that will only add tags to the training data, and thus keep the classic pretrain-fine-tune approach. We implemented two such data augmentation approaches that have been developed for MT of named entities. We followed the approach of Post et al. (2019) who used a masking method on the training data so that their model can be trained to learn to demask them, thus aiming to teach the decoder to reliably translate these entities.

As a second augmentation approach we implemented DEnoising Entity Pre-training (DEEP) (Hu et al. 2022), which is quite similar to the masking method. Here, entities in the text are replaced with their translations for pre-training. In contrast with general masking, these entities translations are taken from a knowledge base (KB), Wikidata (Vrandečić and Krötzsch 2014). This teaches the model the manner in which these entities should be translated and how the translations fit into the context.

The model is then pre-trained on this augmented data, before it is finally fine-tuned on non-augmented parallel data. By making the model denoise the entity augmented data, no change to the architecture is needed. Additionally, the model can exploit the sentence context to improve translation quality. Since the model learns to denoise using the sentence context, even yet unseen entities can be translated (Hu et al. 2022).

3. Methodology

We aim to investigate the effect of various fine-tuning methods for a multilingual machine translation model on the automatic translation of Tagalog into English, with a particular focus on the translation quality of location names. . In section 3.1, we will discuss the multilingual MT model that we used. Then, subsequent section 3.2 will describe the evaluation of the experiments conducted, including the dataset that was used, and the metrics employed. The translation quality of different types of locations was evaluated using both automatic evaluation metrics and a manual evaluation and error analysis. Finally, the section 3.3 concludes with a detailed account of the fine-tuning of the model.

3.1 MT Model

In this paper we use No Language Left Behind (NLLB) (Costa jussà et al. 2022), an open source sequence-to-sequence multilingual Transformer, as our base MT model that we fine-tune ourselves. NLLB includes 200+ languages, including many low-resource languages, and uses several techniques to minimize interference between unrelated languages, and training on high-quality monolingual data. In this way, the model balances the high- and low-resource languages to perform well for both. We use the tools provided on Hugging Face to help us fine-tune the model[1]. All code that we used in our experiments, as well as our datasets, can be found at `https://github.com/slejeune/tagalog-MT`.

In this study, we compare the performance of the out-of-the-box NLLB model with several different custom fine-tuned versions of NLLB. The versions we use are as follows.

- **NLLB**: The base NLLB model, without additional fine-tuning.

1. https://huggingface.co/facebook/nllb-200-distilled-600M

- **NLLB-full**: The first fine-tuned version of NLLB is fine-tuned on the full custom fine-tuning dataset we created (detailed in section 3.3). This model has the advantage of a large amount of data, but no further augmentation techniques are used.
- **NLLB-DEEP**: This fine-tuned model uses the larger entity augmented fine-tuning dataset to further pre-train the NLLB model. Then, it uses the smaller unprocessed parallel fine-tuning subset to further fine-tune on the flooding domain data.
- **NLLB-subset**: We include a fine-tuned version of NLLB that is only fine-tuned on the smaller unprocessed subset of the custom parallel fine-tuning dataset. Since the entity augmented and masked versions first pre-train on the entity augmented or masked data and then further fine-tune on the smaller subset, there is a possibility that potential change in performance can also be attributed to the fine-tuning on the smaller subset alone. This version is thus included to compare with the entity augmented and masked versions to fully investigate the effect of the entity augmentation and masking.
- **NLLB-masked**: This fine-tuned model uses the same approach as NLLB-DEEP, except it uses the generally masked data for the pre-training instead of the entity augmented data.

3.2 Evaluation

We collected a set of Tagalog news articles reporting on flooding events in the Philippines for the evaluation of the translation quality of locations. As we wanted to know whether certain types of location names are be easier or more difficult to translate, we also manually labeled each location with its type (city, street, river, etc.). We also noted that not every location name mentioned in an article is directly related to the flooding event. Therefore, we performed an extra analysis on the dataset where we split the location mentions into relevant flooding locations and all other locations not relevant to the flooding event.

3.2.1 Evaluation Dataset

The evaluation dataset is comprised of 47 articles written in Tagalog from Filipino news websites reporting on flooding disasters in the Philippines. We translated these articles, with a total of 14,539 words in Tagalog, to English to create a parallel dataset. The articles are first automatically translated by Google Translate (the industry standard) to English, before being manually corrected by a native speaker of Tagalog. When we compare the post-edited version against the original translated version by Google we observe a BLEU score of 97.5, indicating high overlap between both versions. This shows that manual post-editing was minimal.
The location categories were assigned manually according to the categories to which each named location belonged. The following categories of location were distinguished:

- **Streets:** All streets, roads and highways that are directly referred to by name.
- **Barangays:** Barangays are the smallest administrative subdivision in cities or municipalities. All barangays that are referred to by name, either written fully or abbreviated with 'Brgy.'. Town neighborhoods are also included in this category.
- **Cities:** All cities and towns that are referred to by name.
- **Municipalities:** All municipalities and areas bigger than cities but smaller than provinces that are referred to by name.
- **Provinces:** All provinces or larger parts of the Philippines that are referred to by name.
- **Rivers:** All rivers that are referred to by name. If 'river' is next to the name, the word is *not* counted as part of the name.
- **Bridges:** All bridges that are referred to by name.
- **Buildings:** Buildings of all types, plazas, and other very specific locations that are referred to by name.

- **Descriptive:** Locations that are very specific but not directly referred to by name. Instead, these locations are described.
- **Relevance:** A secondary category that can additionally be true together with one of the above categories. This marks whether the location is relevant to the flooding event described in the article or not. An example of a non-relevant location would be the publishing location of the article.

Please note that we cannot make the evaluation dataset easily available as the data consists of copyrighted material. We do provide the URLs and titles of the news articles, and we can provide the translated sentences containing location names on request.

3.2.2 Evaluation metrics

We use the well-established MT evaluation metrics BLEU (Papineni et al. 2002) and COMET (Rei et al. 2020) to evaluate the overall translation quality of the various models. However, neither BLEU nor COMET give us insight into whether the locations are translated correctly. Therefore, we make use of a location specific evaluation metric.
In our use-case of disaster relief, we prioritize the inclusion of all locations. We used a manually tagged list of locations and verified whether the translated sentence has the correct location included or not. Then, we calculate the F-score over these cases, which we will refer to as the 'location F-score'. There may be translations of locations that may not be equal to the exact location string of the reference location name, but can still be considered as an alternative correct translation. For this reason, we did an additional qualitative manual evaluation of the location names to gain insight into the types of errors that are made.

3.3 Fine-tuning

We hypothesize that fine-tuning on domain-specific flood-related news articles will increase the translation quality of locations when translating flood-related news. It is evident that the acquisition of such a domain-specific dataset would be required. However, our endeavors to scrape articles from the internet revealed the substantial time investment is required to extract hundreds of domain-specific Tagalog news items. Instead of collecting Tagalog news articles, we chose to scrape the web for English news articles about flooding events, as these were much more accessible. It should be noted that these articles describe general flooding events not specific to the Philippines. We then automatically translated these to Tagalog using Google Translate. Given our interest in evaluating the quality of the translation from Tagalog to English, our selection of English articles for scraping also ensures that the English component of the resulting parallel dataset is of a high (native) quality.

Preprocessing We chose to limit the length of articles to between 100 and 4000 characters. This limit filters out short articles that do not contain any useful content and lengthy articles that are likely to not contain any useful information about flooding events but are instead opinion pieces. The remaining articles were manually reviewed to ensure the content is about a flooding event.

If articles have 50 successive characters that overlap with another article, the article that is published later is marked as a duplicate and filtered out. Finally, if an article contains less than three sentences or contains characters that imply that JavaScript, or other code artifacts are included in the body of the article, it gets filtered out as well to improve the quality of our fine-tuning dataset. The English text is then automatically translated into Tagalog using Google Translate to make this set of articles into a parallel dataset. The resulting fine-tuning dataset is comprised of 816 articles, with a total of 359,688 words in Tagalog.

Location augmenting We created several versions of the fine-tuning dataset by augmenting the location entities in two ways in order to compare the effects of the data enhancement techniques on location translation. Specifically, we applied a general masking method for augmentation and the

DEEP approach which uses a KB (Wikidata) lookup to replace English location names with their Tagalog counterparts.

First, we extract locations from the English data by using the NER tool from spaCy[2]. When implementing the DEEP approach, we observed that the KB lookup for the entity augmentation in Tagalog had a succession rate of 8.3%. This means that only 0.7% of the fine-tuning data was replaced with Tagalog entity information. To remedy this, we decided to do additional general masking on articles that were masked less than 1% after entity augmentation. The additional words that are masked are randomly selected content words. After the additionally masking 3.4% of all tokens are augmented. For the general masking approach, we masked the same tokens as those that were selected in the DEEP approach. We show an example sentence from the fine-tuning set with the DEEP and masking approach in example 1.

(1) **Original** In the summer of 2021, **Germany** was hit by deadly floods that killed more than 230 people.

Masked In the summer of 2021, <**MASK**> was hit by deadly floods that killed more than 230 people.

DEEP In the summer of 2021, **Alemanya** was hit by deadly floods that killed more than 230 people.

We also investigate the effect of the size of the fine-tuning set. We randomly sampled 10% of the original parallel data to create a small (not augmented) fine-tuning set that was used in the fine-tuned version NLLB-subset. The remaining 90% of the articles for the larger DEEP and masked datasets is taken from either the entity augmented data or the generally masked data and is used for the additional pre-training.

4. Results

4.1 Sentence quality

We first report the overall Tagalog translation quality on our evaluation set of 47 news articles using BLEU and COMET in Table 1. All fine-tuned versions of the NLLB model show improvement over the base NLLB model. Of all fine-tuned NLLB versions, NLLB-full scores highest.

	NLLB	NLLB-full	NLLB-DEEP	NLLB-subset	NLLB-masked
BLEU	50.82	**53.70**	51.09	51.634	52.32
COMET	0.606	**0.649**	0.627	0.629	0.635

Table 1: An overview of all BLEU and COMET scores for the base NLLB model and all fine-tuned versions of the NLLB model. The highest scores for the NLLB models are in bold.

4.2 Location evaluation

We present the results for the location F-scores calculated over all locations and individual location types we tagged in the evaluation set. The results for all models on the different location types are in shown in Table 2. We see that for most location types NLLB-full performs best. Only for the 'rivers' category does NLLB-masked perform better, and for the 'buildings' category all models perform similarly. This implies that the traditional method of fine-tuning on large sets of data is the most effective, even for accurately translating entities such as locations. We observe that the location descriptions were the most difficult location to translate as can be expected due to the free form that these names can have. Another category location names that were more difficult

2. https://spacy.io/

than other names were the barangays. Many of the errors for this category are due to abbreviation punctuation marks that are mistaken for end-of-sentence punctuation as we will illustrate in section 4.4.

	NLLB	NLLB-full	NLLB-DEEP	NLLB-subset	NLLB-masked
Streets (n=81)	93.42%	**96.15%**	94.81%	**96.15%**	**96.15%**
Barangays (n=149)	85.16%	**88.64%**	81.45%	85.60%	86.49%
Cities (n=140)	96.27%	**97.42%**	94.30%	94.70%	95.49%
Municipalities (n=220)	96.96%	**98.15%**	95.73%	96.96%	96.71%
Provinces (n=374)	98.77%	**99.46%**	97.93%	98.90%	99.04%
Rivers (n=31)	96.43%	96.43%	96.43%	96.43%	**98.24%**
Bridges (n=4)	**100%**	**100%**	85.71%	**100%**	**100%**
Buildings (n=29)	94.55%	94.55%	94.55%	94.55%	94.55%
Descriptive (n=5)	**75.00%**	33.34%	**75.00%**	**75.00%**	**75.00%**
All (n=1028)	95.58%	**96.78%**	94.35%	95.68%	95.95%

Table 2: An overview of all location F-scores for the base NLLB model and all fine-tuned versions of the NLLB model. The scores are separated by location type, and the number of samples per location type is included. The highest scores for the NLLB models are in bold.

4.3 Relevance

The location F-scores split by relevance are shown in Table 3. A 'relevant' location in this context means any location that is directly referring to a flooding location. Since these F-scores are calculated over the same translation as the scores in Table 2, the best performing model version is the same as when the locations are split by location type. More interestingly, there seems to be very little difference in the location F-scores between locations that are relevant to flooding events and locations that are not relevant, and neither category consistently scores better or worse.

	NLLB	NLLB-full	NLLB-DEEP	NLLB-subset	NLLB-masked
Relevant (n=737)	95.56%	**96.82%**	94.65%	95.48%	95.93%
Non-relevant (n=291)	95.64%	**96.72%**	93.59%	96.18%	96.18%

Table 3: An overview of the location F-scores for the base NLLB model and all fine-tuned versions of the NLLB model. The scores are separated by whether a location is classified as relevant or not for a flooding event, and the number of samples is included. The highest scores for the NLLB models are in bold.

4.4 Qualitative evaluation

The scores given in Table 1, Table 2 and Table 3 do not reflect the full evaluation results as they give no insight into the types of translation errors. To acquire a more complete view on the translations of locations, we conducted a manual analysis of all location translation errors. These errors were subsequently categorized to identify common patterns of mistakes. The errors were classified into four

main types: anglicization, rare names, compounds, and abbreviations. Next, we provide examples to illustrate each of these categories.

4.4.1 Anglicization

The translation of location names tends to be incorrectly anglicized, especially when a direction is present in the place name, such as 'south'. One such example is shown in Table 4, this example shows multiple ways in which a location can be incorrectly anglicized. NLLB-DEEP consistently anglicizes location names more often than the other version. In one extreme example, it translated 'Barangay San Pedro' into 'St Peter's Village', while all other versions of the model kept 'San Pedro' in the translation. It should be noted that this behavior shows up in all versions, although NLLB-DEEP tends to exaggerate this effect.

Tagalog	Sa **Camarines Sur** naman, sari-saring pinsala ang tumambad sa mga residente matapos ang nasa walong oras na pagbayo ni Rolly.
Gold standard	In **Camarines Sur**, residents suffered various injuries after Rolly's eight-hour storm.
NLLB	In the **southern Camarines**, residents were hit by a series of injuries after Rolly's eight-hour ride.
NLLB-full	In the **Camarines Sur**, however, self-inflicted injuries hit residents after Rolly's eight-hour ride.
NLLB-DEEP	In the **South Camarines**, however, residents were hit by a series of injuries after Rolly's eight-hour ride.
NLLB-subset	In the **southern Camarines**, residents were hit by a series of injuries after Rolly's eight-hour ride.
NLLB-masked	In the **southern Camarines**, residents were hit with a series of injuries after Rolly's eight-hour ride.

Table 4: An overview of an example sentence and the translation given by all different versions of the NLLB model. This example shows over-anglicization of the gold standard location name.

4.4.2 Rare and specific names

As can be seen in Table 2, larger locations such as big cities or provinces occur more frequently in the news and might be learned from the training or fine-tuning data. Contrary to this, small-scale or very local locations such as streets are rarely mentioned in the news, and are therefore harder to learn to translate correctly. An example of this is shown in Table 5.

4.4.3 Compound named entities

Location names that include a larger location as part of their name, such as 'España Avenue' shown in Table 5, tend to get translated as if they are separate names. In this particular example, 'España' is not seen as part of the street name, but is instead incorrectly translated to 'Spanish' or 'Spain'. This example also shows that the correct translation of the Tagalog term 'burmaha', indicating 'flooding', is not translated correctly by NLLB.

Tagalog	Samantala, bumaha rin sa **España Avenue** sa Maynila.
Gold standard	Meanwhile, **España Avenue** in Manila was also flooded.
NLLB	Meanwhile, **the Spanish Avenue** in Manila also opened.
NLLB-full	Meanwhile, it also slumped on E**spaña Avenue** in Manila.
NLLB-DEEP	Meanwhile, **Spain Avenue** in Manila also fell.
NLLB-subset	Meanwhile, it also rises on **España Avenue** in Manila.
NLLB-masked	Meanwhile, it also takes up **España Avenue** in Manila.

Table 5: An overview of an example sentence and the translation given by all different versions of the NLLB model. This example illustrates the translation of a specific location name, as well as a compound named entity.

4.4.4 Abbreviation

Errors of punctuation and abbreviation are frequently observed. Table 6 shows how the abbreviation of 'Barangay' and its punctuation mark lead to many different types of mistakes by the various models. Very often, the location name after the abbreviation is left out, or in more extreme cases (here shown for the NLLB models) the remainder of the sentence after the abbreviation is completely ignored in the translation. In this particular instance, the original Google Translate (GT) version is also reported. This version contains an error due to the punctuation mark, which results in a complete alteration of the sentence's meaning. Consequently, it is no longer possible to extract the correct location from the sentence.

Tagalog	Sa Mansalay, Oriental Mindoro, kapwa nalunod naman sina Henry Pastor, 20, at Eddie Gado, 40, ng Sitio Maalin, **Brgy. Maliwanag**, aniya.
Gold standard	In Mansalay, Oriental Mindoro, both Henry Pastor, 20, and Eddie Gado, 40, of Sitio Maalin, **Brgy. Maliwanag**, he said.
NLLB	In Mansalay, Oriental Mindoro, both Henry Pastor, 20, and Eddie Gado, 40, of Sitio Maalin, **Brgy**, drowned.
NLLB-full	In Mansalay, Oriental Mindoro, both drowned Henry Pastor, 20, and Eddie Gado, 40, of Sitio Maalin, **Brgy**.
NLLB-DEEP	In Mansalay, Oriental Mindoro, both Henry Pastor, 20, and Eddie Gado, 40, of Sitio Maalin, **Brgy** drowned.
NLLB-subset	In Mansalay, Oriental Mindoro, both Henry Pastor, 20, and Eddie Gado, 40, of Sitio Maalin, **Brgy** drowned.
NLLB-masked	In Mansalay, Oriental Mindoro, both Henry Pastor, 20, and Eddie Gado, 40, of Sitio Maalin, **Brgy** drowned.
GT	In Mansalay, Oriental Mindoro, both Henry Pastor, 20, and Eddie Gado, 40, of Sitio Maalin, **Brgy. Clearly**, he said.

Table 6: An overview of an example sentence and the translation given by all different versions of the NLLB model and Google Translate. This example shows data loss when unexpected punctuation is present.

5. Discussion

In this study, we investigated several methods to enhance the model performance on location translation. We hypothesized that fine-tuning on domain-specific flood-related data would increase the translation quality of locations when translating flood-related news. We also expected that data augmentation would improve the translation quality of location names in flood-related news. Our results demonstrated that the baseline multilingual MT model already performs well in translating

location names, even without fine-tuning. However, fine-tuning with domain- and language-specific data did lead to further improvements, not only in overall translation quality but also in the accuracy of location name translation.

However, it is important to note that these results should not be interpreted without careful consideration. While some locations were translated correctly, the overall translation quality of a sentence can simultaneously be low. This results in a scenario where we cannot discern whether the correctly translated location is relevant for the flooding event or not. Location translation quality should therefore not come at the cost of general translation quality.

Contrary to our previous work (Hu et al. 2022), we did not observe much effect in our experiment with the augmentation with DEEP. The entity augmentation approach using DEEP showed much promise in that the results it generated were very different from the traditionally fine-tuned models. However, only a small percentage (8%) of the Tagalog locations were present in the knowledge base. If the proportion of entities covered by the knowledge base were higher the influence of entity augmentation techniques would likely increase, thus hopefully leading to further improvements in location translation quality. Future research could explore this hypothesis by applying similar methods to another low-resource language that has a larger knowledge base, to study whether the size of the knowledge base influences the results.

A noteworthy phenomenon in our case study is the Filipino location names that are native in English as the country is bilingual. Such location names are easily translated since nothing needs to be changed. The inclusion of such names may cause the F-score to be higher than expected.

The choice to use automatic Tagalog translations of English news articles for the fine-tuning of models caused many locations to become over-anglicized. An interesting avenue for further research would be investigating what the results of fine-tuning are when the fine-tuning data consists of original Tagalog texts about Filipino floodings.

5.1 Limitations

Our study had several limitations. We only investigated the translation quality of one state-of-the-art open-source multi-lingual MT system, NLLB, but in future work it would be interesting to also take into account other commercial multi-lingual MT systems such as DEEPL, ModernMT or ChatGPT (Gao et al. 2024).

This study focuses specifically on Tagalog-to-English translation of location names within the domain of disaster-related news. As such, the findings may not generalize to other language pairs, subject areas or types of named entities. Additionally, the fine-tuning dataset was collected during a narrow time-frame, from December 2023 to January 2024. This may have led to the over-representation of certain themes, such as flooding events that occurred around New Year, and may not fully capture the variability of news reporting on flooding events.

Finally, the study did not evaluate the end-to-end performance of an actual information extraction pipeline using the translated texts. While the base multilingual MT model performed well, our approach still relies on automatic translation. This may introduce errors in translation that could propagate into downstream information extraction tasks. Thus, the practical impact on an actual information extraction pipeline for disaster response systems remains to be validated.

6. Conclusion

We investigated the extent to which an open-source multilingual MT model can be used to translate Tagalog news about flooding disasters into English, with the focus on accurately translating location names, as these are crucial for effective disaster management in the Philippines. We experimented with several methods for fine-tuning the MT model for this specific domain and language pair.
We created two new parallel Tagalog-English datasets about flooding events in the Philippines. We collected a small carefully curated dataset of Tagalog news articles about floods in the Philippines

that was translated to English and manually annotated with location types. We also collected a larger parallel sample of English flood-related news articles that were automatically translated to Tagalog for fine-tuning. We experimented with two methods for data augmentations, masking the entities with either their translation from a KB or with a general mask. We then fine-tuned our model on either the full fine-tuning data, the entity augmented data, the masked data or a subset of the fine-tuning data. This resulted in four different versions that we compared in both a quantitative and qualitative manner.

Our results indicate that the base MT model performs well in translating location names. However, fine-tuning on domain- and language-specific data does help to improve both the overall BLEU and COMET scores, and the translation of location names in particular. To obtain a more comprehensive understanding of the errors in location translation, we performed a manual analysis of these errors. We identified four common types of mistakes in location translation: anglicization, rare names, compounds, and abbreviations.

Overall, our study demonstrates that automatic translation of Tagalog news into English is a viable approach resulting in only minimal errors in the translation of location names. With this method in place, existing information extraction tools developed for English can be effectively applied to retrieve key information related to flooding events in the Philippines.

7. Acknowledgments

This publication is part of the project 'Indeep: Interpreting Deep Learning Models for Text and Sound' with project number NWA.1292.19.399, which is partly financed by the Dutch Research Council (NWO). We want to thank Jane Arleth dela Cruz for verifying the test set data. We would also like to thank the reviewers for their valuable feedback and constructive suggestions, which helped improve the quality and clarity of this work.

References

Aharoni, Roee, Melvin Johnson, and Orhan Firat (2019), Massively multilingual neural machine translation, *in* Burstein, Jill, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 3874–3884. https://aclanthology.org/N19-1388.

Alcantara, Jonathan (2019), Overview of the Societal Impacts of Floods in the Philippines, *Parliament Institute of Cambodia*.

Almahairi, Amjad, Nicolas Ballas, Tim Cooijmans, Yin Zheng, Hugo Larochelle, and Aaron Courville (2016), Dynamic capacity networks, *Proceedings of the 33rd International Conference on Machine Learning*, ICML'16, p. 2549–2558.

Bapna, Ankur, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes (2022), Building machine translation systems for the next thousand languages. https://arxiv.org/abs/2205.03983.

Bengio, Yoshua, Nicholas Léonard, and Aaron Courville (2013), Estimating or propagating gradients through stochastic neurons for conditional computation. https://arxiv.org/abs/1308.3432.

Brown, Keith and Sarah Ogilvie (2010), *Concise encyclopedia of languages of the world*, Elsevier.

Chi, Zewen, Li Dong, Shuming Ma, Shaohan Huang, Saksham Singhal, Xian-Ling Mao, Heyan Huang, Xia Song, and Furu Wei (2021), mT6: Multilingual pretrained text-to-text transformer with translation pairs, *in* Moens, Marie-Francine, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 1671–1683. https://aclanthology.org/2021.emnlp-main.125.

Conneau, Alexis and Guillaume Lample (2019), Cross-lingual language model pretraining, *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA.

Costa jussà, Marta R., James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang (2022), No language left behind: Scaling human-centered machine translation., *CoRR*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019), BERT: Pre-training of deep bidirectional transformers for language understanding, *in* Burstein, Jill, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. https://aclanthology.org/N19-1423.

Edunov, Sergey, Myle Ott, Michael Auli, and David Grangier (2018), Understanding back-translation at scale, *in* Riloff, Ellen, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Brussels, Belgium, pp. 489–500. https://aclanthology.org/D18-1045.

Gao, Ruiyao, Yumeng Lin, Nan Zhao, and Zhenguang G Cai (2024), Machine translation of chinese classical poetry: a comparison among chatgpt, google translate, and deepl translator, *Humanities and Social Sciences Communications* **11** (1), pp. 1–10, Palgrave.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, and Ming Zhou (2018), Achieving human parity on automatic chinese to english news translation.

Hu, Junjie, Hiroaki Hayashi, Kyunghyun Cho, and Graham Neubig (2022), DEEP: DEnoising entity pre-training for neural machine translation, *in* Muresan, Smaranda, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Dublin, Ireland, pp. 1753–1766. https://aclanthology.org/2022.acl-long.123.

Huang, Fei, Stephan Vogel, and Alex Waibel (2003), Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization, *Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition*, pp. 9–16.

Huang, Fei, Stephan Vogel, and Alex Waibel (2004), Improving named entity translation combining phonetic and semantic similarities, *Proceedings of the Human Language Technology Conference*

of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, pp. 281–288.

Hutchins, W John (2023), Machine translation: History of research and applications, *Routledge encyclopedia of translation technology*, Routledge, pp. 128–144.

Lambert, Patrik, Holger Schwenk, Christophe Servan, and Sadaf Abdul-Rauf (2011), Investigations on translation model adaptation using monolingual data, *Sixth Workshop on Statistical Machine Translation*, pp. 284–293.

Langga, Prences Mae M and Jerryk C Alico (2020), Students' proficiency and challenges in filipino-to-english translation: The case of filipino senior high school students in a private institutio, *International Journal of Linguistics, Literature and Translation* **3** (4), pp. 51–62, Al-Kindi Center for Research and Development.

Li, Zhongwei, Xuancong Wang, AiTi Aw, Eng Siong Chng, and Haizhou Li (2018), Named-Entity Tagging and Domain adaptation for Better Customized Translation, *Proceedings of the Seventh Named Entities Workshop*, Association for Computational Linguistics, pp. 41–46.

Ma, Shuming, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei (2021), Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. https://arxiv.org/abs/2106.13736.

Nasar, Zara, Syed Waqar Jaffry, and Muhammad Kamran Malik (2021), Named entity recognition and relation extraction: State-of-the-art, *ACM Computing Surveys (CSUR)* **54** (1), pp. 1–39, ACM New York, NY, USA.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002), Bleu: a method for automatic evaluation of machine translation, *in* Isabelle, Pierre, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318.

Park, Chanjun, Yeongwook Yang, Kinam Park, and Heuiseok Lim (2020), Decoding strategies for improving low-resource machine translation, *Electronics*. https://www.mdpi.com/2079-9292/9/10/1562.

Post, Matt, Shuoyang Ding, Marianna Martindale, and Winston Wu (2019), An exploration of placeholding in neural machine translation, *Proceedings of Machine Translation Summit XVII: Research Track*, pp. 182–192.

Raffel, Colin, 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, *Journal of Machine Learning Research* **21** (140), pp. 1–67. http://jmlr.org/papers/v21/20-074.html.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie (2020), COMET: A neural framework for MT evaluation, *in* Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, pp. 2685–2702. https://aclanthology.org/2020.emnlp-main.213.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016), Improving neural machine translation models with monolingual data, *in* Erk, Katrin and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), Association for Computational Linguistics, Berlin, Germany, pp. 86–96. https://aclanthology.org/P16-1009.

Ugawa, Arata, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura (2018), Neural machine translation incorporating named entity, *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3240–3250.

Vrandečić, Denny and Markus Krötzsch (2014), Wikidata: a free collaborative knowledgebase, *Commun. ACM* **57** (10), pp. 78–85, Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2629489.

Wang, Yuguang, Shanbo Cheng, Liyang Jiang, Jiajun Yang, Wei Chen, Muze Li, Lin Shi, Yanfeng Wang, and Hongtao Yang (2017), Sogou neural machine translation systems for WMT17, *Proceedings of the Second Conference on Machine Translation*, Association for Computational Linguistics, Copenhagen, Denmark, pp. 410–415.

Xie, Shufang, Yingce Xia, Lijun Wu, Yiqing Huang, Yang Fan, and Tao Qin (2022), End-to-end entity-aware neural machine translation, *Mach. Learn.* **111** (3), pp. 1181–1203, Kluwer Academic Publishers, USA. https://doi.org/10.1007/s10994-021-06073-9.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le (2019), XLNet: Generalized Autoregressive Pretraining for Language Understanding, *Advances in Neural Information Processing Systems*, Vol. 32, Curran Associates, Inc.