Final Report:
“Collecting & Preparing Multilingual Language Data”
Date Report Prepared: April 29, 2024

Executive Summary

Introduction

This report discusses the challenges and strategies involved in collecting and preparing multilingual language data, particularly focusing on low-resource languages for AI applications such as machine translation. Emphasis is placed on the necessary conditions for dictionary and corpus readiness to ensure successful translation tasks.

Dictionary Readiness

Key to machine translation success is identifying atomic units (lemmas) and their multiple senses across languages. For low-resource languages, establishing a robust word to lemma and lemma to sense mapping is crucial. Dictionary readiness in AI contexts does not prioritize human readability but focuses on effective translation by leveraging parallel sentence data. Proper nouns are treated differently since they carry inherent meanings and are easier to translate without deep lexical analysis.

Corpus Readiness

Corpus readiness involves creating parallel translations that collectively cover all required linguistic elements for a translation task. The efficiency of this process is enhanced by large language models which identify and define new word senses and grammatical features within the corpus. For example, studies on biblical texts show that strategic sampling of verses can yield maximal coverage with minimal data, aiming to establish a minimal threshold of linguistic elements that a machine learning model needs to function effectively.

Research Questions

The report raises several research questions about optimal data collection strategies, the level of incentives necessary to gather sufficient data, and the ideal amount of data needed before hitting diminishing returns. These questions guide the ongoing efforts to refine data collection methodologies, particularly in the context of low-resource languages.

Multilingual Model Building

Recent advancements in neural machine translation (NMT) have favored multilingual models over language-pair-specific models due to their scalability and ability to generalize across languages. This report presents a case study on the Alas language,

showing that multilingual training incorporating closely related languages can enhance translation accuracy, as evidenced by improved BLEU and ChrF scores.

Conclusion

The findings suggest that multilingual training holds significant promise for improving machine translation, particularly for languages closely related to the training set. However, the effectiveness varies, and in some cases, like with the Toba language, it does not show noticeable improvement. Further research and experimental trials are necessary to understand the conditions under which multilingual training will be most beneficial.

This summary provides an overview of the report's findings and the ongoing questions in the field of AI-driven language translation, particularly focusing on the challenges faced by low-resource languages. The insights gained from this research are crucial for developing more effective and efficient translation technologies.

Full Report

“Collecting & Preparing Multilingual Language Data”

Introduction

For AI tools such as machine translation to succeed within the context of low resource languages, requirements for dictionary and corpus readiness must be defined.

Dictionary Readiness

At a fundamental level, for machine translation tasks to succeed, one needs to find atomic (i.e. minimal) units from one language to another. In the context of low resource languages, XRI’s strategy is to find these correspondences between the target language, and the language of wider communication (LWC), which is most relevant in that target language’s context.

While words frequently act as a stand-in for atomic units in a language, this introduces unreliability on two key fronts. The first is that words generally accept grammatical morphology, which means that the same underlying word meaning might have a multiplicity of surface forms. Since surface forms in language are ubiquitously generated as a result of underlying forms (sometimes referred to as lemmas) in combination with regular patterns, it is a common assertion that not all surface forms need to be captured. Rather, a multilingual dictionary can be made up of lemma mappings between languages, while the capturing of these regular patterns can be addressed in corpus readiness.

The second way in which words can be false atomic units is with regard to word sense. Whereas surface forms of words can be seen as the decompressed version of an underlying form, words themselves can be seen as a compressed form of word senses. This is to say that the various meanings of a word cannot be seen as collected, unless explicitly mentioned by sense definition. To exemplify this, consider two senses of the word ‘fast’: ‘with great speed’ & ‘to forego food for a time’. Here the word itself obscures the fact that there are (at least) two meanings stacked on top of a single encoding. In terms of data collection, we should expect to consistently see multiple meanings per word. Additionally, it would be surprising if word to sense mappings were consistent across languages, which means that it is insufficient to collect a parallel translation for a word and consider all instances of that word properly defined.

Thus, dictionary readiness needs two key components: first, a word to lemma mapping, where multiple surface forms are correlated to a single lemma form, and secondly a lemma to sense mapping, where a lemma has multiple senses associated with it.

A notable exception to this is proper nouns. While many names have meaning encoded within their morphology, this meaning is considered superfluous in most contexts, and the surface form takes precedence. For example, the place name Bethlehem is the Anglicized version of a combination of Hebrew roots for ‘house’ and ‘bread’, giving the meaning house of bread. However, in order to translate a sentence including the word Bethlehem, one does not need to know the word house or bread. This is to the translator’s advantage. For example, in an analysis of Gondi-Hindi data, when accounting for lemmas, compared against word coverage for the Bible, initial coverage is 36%; however, when considering practical coverage, that number will in fact be higher, as proper nouns effectively self-define themselves within context, and do not require explicit collection.

As a final note, while machine translation systems do make use of a dictionary, these dictionaries are not intended for human readability, but rather for effective translation. This means that the dictionary is built as a part of the data processing based on the parallel sentences provided. Thus, in terms of machine learning, the atomic units are sub-words most closely related to morphemes. In practical terms this puts the onus of dictionary readiness entirely in the context of the parallel sentences. Dictionary readiness will be achieved when there is at least one (but quite possibly more than one) instance of each word sense needed for translation across the entire training set of data. The results we achieved when approaching data collection with a view toward dictionary readiness demonstrated that this concept applied greatly increases the performance of the AI models.

Corpus Readiness

Given the exploration of dictionary readiness, corpus readiness can be defined in terms of a series of parallel translations, in which the sum of all elements within the total sentences contains (at least) one instance of every element known to be needed for the translation task. To ground this in a practical example: any given text, such as the book of Mark, has a finite amount of word senses, morphological features, and grammatical constructions. If each of these elements in Mark (or any given corpus) are captured and represented in parallel, then in theory we have everything we need to produce a translation of Mark, and the generalization of this process can be applied in novel contexts.

While there have been many cases of monumental human effort to identify the deepest complexities of a single corpus, XRI is developing tools that can be tasked to any given corpus, yielding all elements identified within the text by leveraging the power of large language models, which are able to identify new word senses, along with their definitions, as well as morphological features and grammatical rules.

Initial analyses were focused on economy of coverage: essentially how few sentences do we need in order to receive maximal coverage. This suggests that not all sentences are equal with regard to impact towards corpus coverage. This aligns with the fact that all languages have high frequency grammatical words, which when collected would account for the majority of the text of the corpus. However, ultimately focus shifts toward elements of minimal impact. While a maximally impactful dataset could be generated, the result is unnecessarily small, representing a maximally compressed analysis. For example, when analyzing coverage of the following books of the Bible: Matthew, Mark, Luke, John, Genesis. Maximal economy of coverage of elements can be achieved by targeted sampling of 2000 verses (out of approximate 5000). This is to say with the components of 2000, we should be able to reconstruct the following 3000.

This assumes perfect implementation, which is highly unlikely. In terms of machine learning, quality of data has proved to be integral time and again; however, there is a minimum threshold. We are interested in finding that minimum threshold against multiple elements (at least word senses, morphological features, and grammatical rules). There are also many different approaches to grammatical analysis. Our current approach for grammar is centered around Construction Grammar, as it is verbally expressed, which is conducive to working with large language models.

To investigate this minimum threshold for elements, we then took our 2000 verses and their identified components, and synthesized 8000 sentences, where every element appears a minimum of N times where N = 4. If we wanted to increase N, we could take that same original 2000 verses, and synthesize a greater number of sentences, until we have the desired holistic minimal coverage.

With this dataset, recently synthesized, we can begin to perform experiments relating to Machine learning’s sensitivity to particular elements, by selective exclusion. These simulations will be insightful as to the minimum size of N for various elements. Initial hypotheses posit that N will be different for different elements when relating to machine learning models. It might be possible that the model only needs to see the word sense ‘fast.1: with great speed’ two times but needs to see five examples of existential sentence constructions. While these experiments still need to be run, we can confirm that approaching data collection with a view toward achieving corpus readiness leads to higher performing AI models.

Key Questions

XRI continues to posit that intelligent data collection will result in high quality datasets and that these datasets will be able to yield corpus specific readiness with greater efficiency and effectiveness. However, to prove this we endeavored to answer the following research questions.

Research Questions

Question 1: What is the ideal set of prompts for collecting this data?

Regarding ideal prompts for data collection, earlier analyses indicated a necessary awareness of not only all vocabulary items (as seen at the word sense level), but also invisible grammatical and syntactic elements. This remains true, and we have continued to see positive results using our collected dataset in the contexts of model training for translation. However, continued investigation has also revealed that in the context of collecting language data in low-resource languages for the purposes of machine translation, precise care also needs to be given for a trilingually aware dataset. There is the target language itself, which may have unique features with significant impact on successful machine translation. There is the language of wider communication, which is the medium by which the target language will be prompted, and finally there is the language of grammatical analysis, presumably English, which is necessary on several accounts, most notably English is the language most integrated with cutting-edge AI, which allows for detailed and nuanced analyses of a target domain or corpus.

Question 2: What level of incentives are needed to obtain the minimum target data set?

This question should perhaps be rephrased to the following: 'What model of incentives is necessary to obtain the target dataset? While either question certainly merits more research, at this point XRI has a proven procedure that has worked in multiple languages. Namely a small selection of contracted target language experts, managed by a local entity that liaises with a technical expert. This model has proven to be more effective than both paid crowdsourcing and volunteer crowdsourcing. Our informal analysis for why this would be includes the following:

- For every contributor there is an implementation period in which a contributor familiarizes himself or herself with the nature of the task, the medium of communication, and any other expectations. If a crowdsourced volunteer only contributes a few sentences, that individual will always be in this implementation period.
- A local entity is more likely to be able to navigate cultural expectations relating to work timelines, quality of data, technological confusion, or important cultural norms.

- A contracted incentive, unlike a per sentence crowdsourced paying structure, will encourage speakers to complete all tasks, both by quantifying the work from potentially infinite to a stated 800 parallel translations tasks, 800 vocabulary & validation tasks, as well as by giving a stated incentive related to the completion.

Question 3: What is the right amount of data before diminishing marginal returns sets in?

At this point, using the data collected, we can build a model that is evaluated by a separate set of speakers as either accurate, natural, or in some cases both. This is not to guarantee that such a model is up to production quality, but these are clearly positive indicators that we are getting the right kind of data, and we are getting enough data to be at least partially effective. This model is trained on the 8000 sentences dataset, which is domain specified.

Domain specification is an important caveat, as it is directly tied to the idea of diminishing marginal returns. Logically if the domain of interest is small, then the proportion of necessary data will be smaller. Conversely a large domain or multi-domain target for machine translation is anticipated to have larger data needs.

At this point it is impossible to empirically explore the impact of more data (we can for example divide our dataset and try on smaller portions of data (e.g. 4000)). We cannot, however, answer the question: 'How much better would our model be if we had 12,000 sentences.

To posit an informal prediction: by 8000 sentences, given the specified domain, diminishing marginal returns is already having an impact, and every new sentence is less valuable than the previous. However, that impact is not significant enough that 8000 sentences is to be considered the ideal number for collection. These questions, however, need to be answered not only in a theoretical sense, but also within the pragmatic context of language collection.

Multilingual Model Building

In recent years, neural machine translation (NMT) engines have become highly accurate, even achieving in some cases superhuman levels of accuracy. This improvement in quality has been driven by better neural architectures, more compute resources, and vast amounts of data. One of the significant areas of architectural improvement has been the shift away from creating models to translate between specific language pairs, and instead creating models which know a set of languages and can translate from any language in the set to any other language in the set. These models are called multilingual models.

Multilingual models have several advantages over systems trained for a dedicated pair of languages. One obvious benefit is scale. Meta's No Language Left Behind (NLLB) speaks 200 languages, but is only a single multilingual model. If one were to create a dedicated model for each language pair, they would have to train 19,900 separate models.

A more important advantage for our purpose is that a multilingual model can generalize from languages it knows well to languages unknown to it. In fine-tuning it to learn a new language, it will do better if that language shares grammar and/or vocabulary with other languages it already knows. Along these lines, if it is taught several similar languages at once rather than only being taught one of them, then it can use what it learns about one language to help speak another one better.

Batak Alas

Alas (also known as Batak Alas or Alas-Kluet) is an Austronesian language for which we have completed a data collection of 8,000 sentences. From this collection, we have fine-tuned NLLB-200-distilled-1.3M to achieve an accuracy of about 25% BLEU. We have demonstrated the utility of this model to Alas translators, but there is still room for improvement. One possible improvement arises because we have not just collected Alas data, but also Toba and Gayo. All three of these languages are closely related to Indonesian (the source language in these collections), suggesting that fine-tuning NLLB on Gayo and Toba as well as Alas may improve recognition scores for Alas.

To test this, we trained two Alas models on purely Alas data. For the first model, we used the first 7600 sentences as training data, the next 200 as dev, and the final 200 as test. For the second model, we used the first 200 as dev, the next 200 as test, and the remaining 7600 as training data. For each model, we measured the BLEU and ChrF scores on the dev and test data every 5 epochs. For each training setup, we consider models from 4 epochs: the epoch with the best dev BLEU, best test BLEU, best dev ChrF, and best test ChrF. For each of these epochs, we used the three scores that were not used to choose the epoch, resulting in 6 BLEU scores and 6 ChrF scores per training setup, or 12 of each across the two setups. From these, we calculated a mean and standard deviation BLEU and ChrF score for Alas models trained purely on Alas data. These 12 numbers are not fully independent, and indeed there were a few cases where the different methods of choosing the epoch resulted in the same choice of epoch. We expect this dependence results in a small underestimate of the error bars, which we am ignoring for now.

For the multilingual training, we repeated the experiment including Gayo and Toba data in the training sets. Since the Indonesian data in the parallel sentences is common across all three languages, we excluded the Gayo and Toba data whose Indonesian source sentences were in the dev or test sets. Following the same setup as before, we trained two models using different data splits, and found the mean and standard deviation of the BLEU and ChrF of the 12 samples. The results of these experiments are shown in Table 1.

	BLEU	ChrF
Monolingual	25.03 +/- 0.78	64.67 +/- 0.32
Multilingual	26.69 +/- 0.72	65.98 +/- 0.85

Table 1: Comparison of monolingual and multilingual models for Alas

This data is rough; it would be better to create more validation splits and get tighter and more stable error bars to confirm the statistical significance of these measures, but since the BLEU score has improved by over two standard deviations, it appears that the multilingual training has made a measurable improvement.

Batak Toba

While significant, the improvement to for Alas is not large. One possibility for this is that Alas is already quite similar to Indonesian, and adding Toba and Gayo to the training set doesn't add much that Indonesian doesn't already include. A simple quantitative measure of this is to compute the BLEU and ChrF scores between the source sentence and their target translations. If these measures are high, this at least crudely indicates similarity between the source and target languages. Table 2 shows these similarity measures for the three languages we have collected:

	src/target BLEU	src/target ChrF
Alas	4.3	37.8
Gayo	2.7	31.6
Toba	0.9	23.3

Table 2: Measures of similarity between Indonesian and the collected languages

This indicates Alas is closest to Indonesian, followed by Gayo, and finally Toba. If indeed the similarity between Alas and Indonesian is why Gayo and Toba didn't help the Alas NMT much, then we would expect Toba to be much improved by the addition of Alas and Gayo, as Toba is the least similar to Indonesian.

To test this hypothesis, we repeated the Alas experiments on Toba, with the results shown in Table 3.

	BLEU	ChrF
Monolingual	17.50 +/- 1.01	51.54 +/- 0.57
Multilingual	17.29 +/- 1.37	51.27 +/- 1.03

Table 3: Comparison of monolingual and multilingual models for Toba

In contrast to Alas, the monolingual and multilingual Toba NMTs are statistically indistinguishable. While it would be good to verify this with more validation splits to get more statistically significant results, it does not bode well for the conjecture above.

Multi Model Building Conclusions

Multilingual training shows promise, and is helpful for improving the accuracy of an Alas NMT. It has not shown any improvement on Toba, however. This result is puzzling and deserves further study. At present, we would characterize multilingual training as a useful tool to have in the toolbox, but we are unable to predict the cases in which it is likely to be helpful without simply trying the experiments.

Final Conclusions

Our research has demonstrated that collecting data with robust attention to both dictionary and corpus readiness as defined in this report leads to impressive improvements in training AI translation models. Our method contrasts with the traditional method of scraping existing data from the internet, cleaning it, and then using it to train translation models. By collecting and preparing multilingual language data with the methods applied above, we can reduce both the time and cost of building translation models for low-resource languages. Building translation models using language data from multiple related languages shows promise in some but not all circumstances.