Exploring Unsupervised Machine Translation for Highly Under resourced Languages(Hausa)

Stanford CS224N Custom Project

Zouberou Sayibou & Sajid Farook
Department of Computer Science
Stanford University
zouberou@stanford.edu & sajidof@stanford.edu
Project Mentor: Soumya Chatterjee (soumyac@stanford.edu)

Abstract

We explore various approaches to address the under-explored challenge of machine translation (MT) for resource-poor languages, namely Hausa, without parallel data. We first develop a new embedding model outperforming the current FastText baseline. Then, we train 3 supervised and semi-supervised translation models on Hausa-English and French-English translation respectively. Our fully-supervised model struggles uniquely on translating more general, representative Hausa sentences, highlighting major issues with under-resourced languages' reliance on religious text translations for parallel data for supervised models. Next, we develop a method of manually generating a bilingual dictionary using Google Translate to either replace or validate established unsupervised embedding alignment algorithms. We show that either approach can be more effective depending on the language and dictionary, but that the method produces substantial performance improvements compared to a fully unsupervised approach that foregoes a dictionary entirely. Overall, our work is a valuable contribution to the literature which is specifically lacking in experimentation on semi-supervised methods to improve MT models for languages with limited resources in particular.

1 Introduction

Machine translation (MT) models from developing countries and under-resourced communities tend to perform poorly because of the lack of parallel training data and other resources. Moreover, machine translation literature seldom studies MT for such languages. Hausa, for instance, has over 54 million speakers, but default FastText word embedding include just over 4,000 embeddings, while reliable parallel datasets are limited primarily to the JW300 dataset of just over 230,000 parallel sentences from Biblical translation. In this paper, we investigate ways to bridge this gap and improve MT performance on low-resource languages, with a focus on Hausa to English translation. Specifically, we investigate the potential for avoiding the need for parallel data by comparing a supervised model with two semi-supervised models to determine when maintaining supervision is more important for MT performance than increasing dataset quality and size. In addition to studying this trade-off, our work offers some findings about the importance of linguistic similarity, training dataset representatives, and initial word embeddings, in MT performance for under-resourced languages like Hausa.

2 Related Work

The possibility for unsupervised MT methods was unlocked when Conneau et al. (2018) showed that high-quality cross-lingual word embeddings could be generated without the need for parallel data. This was done using adversarial training to find the rotation matrix that, when applied to each word embedding in the target language, aligns corresponding words in either language in the same latent

space, effectively yielding a parallel dictionary. Sennrich et al. (2016) had also shown that iterative back-translation can be used to synthetically generate parallel data from a high-resource language to a lower-resource language as ground-truth parallel data for training a MT model in a supervised way, partially mitigating the aforementioned parallel data bottleneck.

Later, Lample et al. (2018a) combined these ideas to develop a fully unsupervised model that begins with inferring a bilingual dictionary using the aforementioned embedding-alignment method from Conneau et al. (2018) and then using iterative backtranslation through bidirectional LSTMs to smooth naive word-by-word translations generated by the dictionary. However, this paper only tests their method on French-English translation even though such highly-resourced languages do not suffer from the data shortage that make unsupervised approaches necessary. Later, Lample et al. (2018b) showed that augmenting this approach with a phrase-based statistical method (PBSMT) to generate a phrase-table to inform phrase-by-phrase translations perform better on low-resource languages in particular such as Urdru and Romanian, compared to purely neural methods.

The limitation of these papers is that they exclusively explore fully unsupervised models without considering the potential for augmenting these approaches with supervised methods using the little parallel data that does exist or can be generated for under-resourced languages. Perhaps the most notable work in semi-supervised MT literature is by Cheng et al. (2016) which uses parallel data to train models in both translation directions, which are then fine-tuned to reconstruct a larger monolingual corpus by translating it to and from another language. However, the supervised models they start with were trained on over 2.5 million parallel sentences, meaning that the application to resource-poor languages is once again limited.

Overall, we identify a distinct gap in the literature concerning semi-supervised approaches to low-resource languages in particular. We take particular interest in the generation of a bilingual dictionary. While the adversarial training method from Conneau et al. (2018) used by most unsupervised MT papers is effective for aligning linguistically similar languages, the potential for initializing and improving aligned embeddings through other methods such as a Google-translate-generated dictionary are under-explored. This paper tests different approaches, both in Hausa-English and French-English translation, to better understand how semi-supervised approaches may support MT for under-resourced languages.

3 Approach

Our main experiment compares three approaches for translation from Hausa to English and French to English.

Firstly, a *fully supervised approach*. We use a bidirectional LSTM encoder and a unidirectional LSTM decoder with multiplicative attention, as is used in CS224N assignment 4, which we borrow the code set-up from. The goal here is not to maximize performance, but rather to serve as a baseline model using a traditional supervised method to gauge the extent to which limited parallel corpora inhibits model performance for Hausa. Because no reliable open-source Hausa MT models are available, this is used as our baseline and benchmark.

Secondly, *unsupervised alignment*. This is a mostly an unsupervised approach that utilizes Lample et al. (2018b) PBSMT method to count individual n-grams of up to 4 words in the source language (Hausa), matches each with the top 100 most similar word embeddings from the target language using cosine similarity, and augments this with a language model (KenLM, for speed) to select the best phrase-by-phrase translation. To determine the most similar words, we begin by generating cross-lingual word-embeddings through using adversarial training to learn the top 5 rotations to align embeddings in either language in latent space and chose the best rotation by computing whether k-nearest neighbors includes the corresponding translation based on a bilingual word dictionary (English-French was used in the paper) (Conneau et al., 2018). Given that there are no available Hausa-English dictionary datasets, we decided to generate a dictionary by passing every English word from English-French dictionary from Conneau et al. (2018) through Google Translate API.

Lastly, *supervised alignment*. This is a semi-supervised approach that follows the same method as unsupervised alignment. However, it forgoes adversarial training to rotate and align embeddings and instead relies solely on the bilingual dictionary to generate word-for-word translations, which is used to assign new embeddings to target language words instead. Hence, while unsupervised alignment

finds cross-lingual embeddings by learning a rotation and validating it against a dictionary, supervised alignment finds embeddings using the dictionary itself.

In addition to these three approaches, we also trained a final model - identical characters alignment - to serve as a lower-bound baseline. This follows the same procedure as supervised alignment, except the dictionary used to initialize cross-lingual embeddings only included mappings all words that are seen in both languages' monolingual corpora (and are therefore presumed to mean the same thing, intuitively containing mostly proper nouns). Because no adversarial training is being used to align the language and the input dictionary contains little information, we expect our results to be almost random, which serves as a lower-bound baseline for our other approaches.

In addition to our three main models, we ran a few additional experiments. The only pre-trained Hausa word embeddings are fastText embeddings that contain only 4,347 words trained on Wikipedia Bojanowski et al. (2017). We also trained our own Hausa word embeddings using Word2Vec's Continuous Bag of Words (CBOW) model using the Gensim library on our monolingual corpus of over 3,000,000 sentences, producing around 200,000 embeddings. Performance of each embedding model was compared on our supervised alignment task.

Lastly, to test the potential for overfitting to the smaller, niche datasets that under-resourced languages like Hausa are relegated to using, we evaluate our fully supervised Hausa model on both 10% of the JW300 dataset it was trained on versus a distinct test test from FLORES.

We primarily use code provided by Lample et al. (2018b), but write our own code to download and process Hausa data, generate the Hausa dictionary, integrate it in the base code, and generate our Word2Vec embeddings.

4 Experiments

	Training Set Source			**Training Set Size**			**Test Set Source**			**Test set size**		
Model	**Ha**	**Fr**	**En**	**Ha**	**Fr**	**En**	**Ha**	**Fr**	**En**	**Ha**	**Fr**	**En**
Fully Supervised	JW300	CommonCrawl	JW300/ CommoCrawl	188,613	2,612,877	188,613/ 2,612,877	JW300 / FLORES+	CommonCrawl	JW300 /CommonCrawl FLORES+	53,350 / 900	2,212	53,350/ 2,212/ 900
Supervised alignment/ Unsupervised alignment/ Identical characters alignment	NNLB+ ParaCrawl + Wikimedia+ CCMatrix	WMT14	NNLB+ ParaCrawl + Wikimedia+ CCMatrix/ WMT14	10,000,000	10,000,000	10,000,000	FLORES+/	WMT17	FLORES+	900	2051	900

Table 1: Datasets used for different experiments

Note that one experiment involved running the same model (supervised alignment) against two evaluation sets. Dictionaries are excluded from this table as they are not models, but English-French dictionary was sourced from Conneau et. al 2017's open-source data, while the English-Hausa dictionary was generated using the the same English words included in the French dictionary (for consistency) using Google Rranslate. Similarly, our Hausa Word2Vec embedding model is excluded from this table, but it was trained on the ParaCrawl (3.5M sentences) monolingual data Bañón et al. (2020).

To elaborate on our datasets, JW300 is a translation of Biblical texts, while CommonCrawl is a dataset comprised of raw web page data, metadata extracts, and text extracts from the web. FLORES is sourced from a range of Wikipedia articles.

Our fully supervised model (LSTMs with attention) was trained over around 2 hours with a learning rate of .0005, batch size of 64, with dropout rate of 0.3. Both unsupervised models were trained over 4-5 hours with 5 iterations of adversarial training alignment and a maximum vocabulary size of 200,000 words.

4.1 Results

Our results show that our embedding model significantly outperformed the default FastText model by 2.85 BLEU on the selected task. This highlights the weakness of existing pre-trained embedding

models for under-resourced language like Hausa, as well as the importance of starting with high-quality embeddings in MT tasks. We suspect that the lack of embeddings (just over 4347, compared to our model's 200,000) was the main reason for this poor performance.

FastText	Word2Vec CBOW
0.80	3.65

Table 2: Comparing FastText (Bojanowski et al., 2017) and Word2Vec CBOW (our embeddings model) used on Hausa

Model	Hausa-English	French-English
Fully Supervised	14.94 / 1.69	21.77
Identical characters alignment	0.93	5.83
Unsupervised alignment	0.80 / 3.65	9.02
Supervised alignment	1.58	11.10

Table 3: The BLEU scores for different models are presented. For the Identical Characters Alignment, the model without a dictionary is used. The Unsupervised Alignment scores are shown for two different embedding methods: fastText and Word2vec. Note that the Supervised Alignment was evaluated on two datasets, and the scores reflect the performance of the Word2vec model.

The above findings show different results for Hausa and French translation tasks. For Hausa, we find firstly that the supervised model is the best when evaluated against a subset of data from the same dataset (JW300 Biblical translations) with BLEU 14.94, but substantially worse when evaluated against a different dataset (FLORES+) 1.69. This is likely because the JW300 dataset is not representative of most Hausa text, being sourced from the Bible and covering vocabulary relating to religion, morality, etc. FLORES, on the other hand, is sourced from Wikipedia, covering a wide range of often technical concepts and words that were absent from JW300. In addition to the amount of vocabulary that was not learned during training, the rhetorical style from JW300 is drastically different from modern-day writing meaning that the model was never able to sufficiently learn some of the linguistic patterns required to correctly translate even sentences that do not include unfamiliar vocabulary. Hence, it is not surprising to find a difference of >15 BLEU, which is consistent with the evaluation disparities often observed in the literature, for instance in Cheng et al. (2017). This finding is especially significant, because it demonstrates the extent to which the lack of parallel data for under resourced languages is a major bottleneck in MT performance - not only are there less than 200,000 available parallel sentences to learn on as FLORES+ is an evaluation benchmark datasets, but these sentences are highly unrepresentative of modern-day text and substantially hinder MT models.

JW300	FLORES
(Muhammad 's) eyes did not deceive him , nor did they lead him to falsehood . Do you not see that I fill up the measure , and am the best of hosts ? But we will certainly benefit from now considering examples of Jehovah's unsearchable greatness .	Since Pakistani independence from British rule in 1947, the Pakistani President has appointed "Political Agents" to govern FATA, who exercise near-complete autonomous control over the areas. You can mark the passing of time yourself by observing the repetition of a cyclical event. A cyclical event is something that happens again and again regularly. Unless you are a diplomat, working overseas generally means that you will have to file income tax in the country you are based in.

Table 4: Random sample of sentences from JW300 and FLORES data sets in English. Parallel Hausa training data is not representative.

Secondly, we find that Hausa performs better with supervised alignment, whereas French performs better on unsupervised alignment. This is intuitive given that our unsupervised embedding alignment algorithm is known to assume linguistic similarity between languages. The Hausa language and English have significant linguistic differences due to their origins from different language families. Hausa belongs to the Afro-Asiatic language family, specifically its Chadic branch, while English is a Germanic language within the Indo-European family just like French.

Hence, the algorithm was likely much more successful in learning the best alignment. There are an additional two likely reasons the unsupervised alignment method outperformed the premade dictionary for French. Firstly, our unsupervised alignment algorithm rotated the top 300,000 n-grams for n<4, meaning mappings of sequences of words were captured in the unsupervised alignment method, whereas the French dictionary that the supervised alignment method had mostly unigrams. Secondly, rotating the entire French language in an unsupervised way preserves the semantic relationships between words within a language. A dictionary, on the other hand, enforces one (potentially dubious) translation on every source word, meaning the model becomes dependent on the quality of the dictionary and some semantic nuances get easily lost. Hence, when our PBSMT model is far less flexible in exploring alternative translations in the phrase table as similar words that are not in the dictionary end up underrated. Nonetheless, performance using supervised alignment is substantially better for Hausa and close for French simply because there is a guarantee that the word-by-word translations are correct without relying on a rotation in latent space. Our method of generating our own bilingual dictionary using Google Translate API calls on the vocabulary was thus successful in improving performance.

Thirdly, we find that Hausa simply performs substantially worse across all translation models than French, even when trained on similarly-sized monolingual data. The ineffectiveness of our unsupervised alignment algorithm in aligning Hausa and English embeddings explains this result for the unsupervised alignment approach, whereas the quality of our premade Hausa dictionary likely explains the disparity under the supervised alignment approach. While using Google Translate to manually generate our own dictionary improved performance relatively to relying on unsupervised alignment, it was substantially worse than the pre-established French dictionary that was used, which included multiple translations for the same French word and included some bigrams and trigrams due to the use of a French tokenizer in building the vocabulary. Crucially, we struggled to source a tokenizer for Hausa, meaning that our dictionary and vocabulary where constructed with a tokenizer trained on English and likely ended up essentially using space delimiters. Hence, word-for-word translations were often not meaningful in the context of the entire sentence to an extent that the language model was unable to correct for. Moreover, upon inspection by a native Hausa speaker, the Google Translate translations were very low quality, highlighting even further that the inability for established MT tools to apply accurately to Hausa is a major impediment in MT performance in addition to data availability.

5 Conclusion and Further Work

Overall, our research furthers the study of MT models for under-resourced languages. We firstly show that pre-trained word embeddings (FastText) for Hausa substantially limits the performance of MT models by developing our own embeddings that outperform it. Secondly, we show that the quantity of data is not the only or main limitation in fully supervised models for under-resourced languages - rather, the frequent need to resort to religious text translations rather than more representative and modern parallel corpora significantly worsens MT performance.

We also derive a number of novel insights about semi-supervised methods. Firstly, manually generating a bi-lingual dictionary using existing models like Google Translate to assist with alignment can be an effective strategy to overcome the lack of published dictionaries. Whether to align word embeddings through learning a rotation matrix in an unsupervised way and validating against a dictionary, or adhering more strictly to this dictionary depends primarily on (1) the quality of the dictionary and (2) the level of linguistic similarity to the target language. We conclude that among the models that was tried here, our semi-supervised method using a synthetically generated dictionary to create aligned embeddings and PBSMT with a language model is the best model for Hausa, notably outperforming our fully supervised approach when evaluated on a representative dataset.

There are a number of limitations to our results that future work can explore further. Most notably, in spite of our findings, we did not develop any Hausa MT model producing a strong BLEU score, and future work should aim to refine our process to improve performance. For instance, training a separate BPE tokenizer for Hausa to generate more semantically meaningful tokens, especially given that Hausa words are often constructed out of multiple subwords and punctuation.

Additionally, there are other techniques established in the literature that were not used in this study that future research may consider integrating. For example, iterative back translation is a key component of almost all unsupervised MT literature, but it was not used here due to computational constraints and because our project is mostly concerned with comparing high-level approaches rather than maximizing performance. Nonetheless, it is likely that including back translation may disproportionately benefit some approaches over others; for instance, back translation is known to be particularly useful in improving unsupervised models because their noisy translation benefit the most from iterative smoothing Marchisio et al. (2020). Our fully supervised model's strong performance on translating French suggests that its poor performance with Hausa is likely due to the non-representative nature of our parallel Hausa data, meaning that using back-translation to synthetically generate low-quality but abundant parallel data would be particularly useful in significantly improving the less supervised approaches.

In addition to integrating back-translation, future work should consider other ways to leverage modern transformer-based models to improve performance. Replacing out PBSMT model's KenLM with a transformer-based model such as Aya, for instance, would likely be effective. Other unsupervised methods such as using LLMs to generate zero-shot translations that are then similarly amplified through back translations have received state-of-the-art results can also be explored as a way to leverage modern models for better performance Han et al. (2021).

A final observation is that some of the conclusions made in this paper may be premature without without more rigorous experimentation. For example, the claim that our fully supervised model's poor performance is attributable solely to the training set could be made with more confidence if the same experiment was run using French JW300 data validated against French Wikipedia-derived data. Similarly, the low BLEU scores of our unsupervised alignment model cannot be attributed to the linguistic dissimilarity between Hausa and English without trying the same model on more languages of different levels of linguistic similarity and similar quality dictionaries for validation. Nonetheless, our work provides useful insights in better understanding the needs for MT models on under-resourced languages.

References

Idris Abdulmumin and Bashir Shehu Galadanci. 2019. hauwe: Hausa words embedding for natural language processing. In *2019 2nd International Conference of the IEEE Nigeria Computer Chapter (NigeriaComputConf)*. IEEE.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data.

Jesse Michael Han, Igor Babuschkin, Harrison Edwards, Arvind Neelakantan, Tao Xu, Stanislas Polu, Alex Ray, Pranav Shyam, Aditya Ramesh, Alec Radford, and Ilya Sutskever. 2021. Unsupervised neural machine translation with generative language models only.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based neural unsupervised machine translation.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work?

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data.

A Appendix (optional)

Firstly, it appears that our experimentation could have benefited from one more, fully unsupervised model that follows the method of unsupervised alignment, but without validation against a dictionary (thereby rendering a method that truly relies solely on monolingual corpora only). However, in experimentation, we found that this alone yielded extremely poor results for Hausa, yielding a BLEU score of . This was likely because English and Hausa are so linguistically different that the model found a rotation that appeared to align some words, but the alignment was incorrect, meaning the model was enforcing a dictionary feeding in nonsensical translations. This was not replicated for French because as the adversarial training is computationally expensive, and our poor results with Hausa was already intuitively consistent with our knowledge of the language and the well-established limitation of the method that it is ineffective for linguistically distinct languages.

Regarding our decision to use our own model as a baseline, Abdulmumin and Galadanci (2019) has a Hausa-English model, but it was not presented or published at a reputable conference and uses dubious evaluation metrics and results. Hence, we opted to omit it from the paper and instead only reference reputable models.

Additionally, it is wort acknowledging that much of the most recent literature on unsupervised MT most closely resembles Han et al. (2021) where the majority of the unsupervised MT pipeline is handled through prompting LLMs. We could have explored this as a route, but used more traditional architectures and methods as we felt it was more in the scope of the requirements of the course and the project.

Regarding partner contributions, Zouberou ran most of the experiments and handled most of the coding and dataset sourcing. Sajid conducted most of the literature review to understand the model and ran the fully supervised models.