Decoding Whale Songs Using Backtranslation and Transformer-Based Models

Morgan Rivers
Department of Physics
Freie Universität Berlin
danielmorganrivers@gmail.com

August 4, 2024

Abstract

This paper proposes a novel method for decoding whale songs using backtranslation techniques combined with transformer-based language models. The framework involves training models to translate whale vocalizations into English and back, ensuring consistency and accuracy through reinforcement learning. By maximizing the likelihood of joint probability distributions, this approach aims to create a general model for cross-species language translation.

1 Introduction

Decoding whale songs presents a unique challenge in the field of animal communication. I believe communication with whales will both teach us for the first time a complete language of another species, as well as enable communication with many more animals. This will be critical to both understand the desires and steward the environment for these animals, as well as raise popular awareness of how humanity is affecting wild animals on this planet.

In this paper, I propose an unsupervised translation approach motivated by understanding animal communication. This involves using backtranslation techniques to create a model that can translate whale vocalizations into English and back, ensuring that the translations are contextually and semantically accurate.

I will use the term "Whale" to refer to the language Whales are speaking. For example, "she understands both Whale and English."

2 Related Work

Studies on neural networks, such as AlphaFold for molecular interactions, have shown the potential for applying similar techniques to decode whale songs.

One paper specifically about whales is interesting, and partially motivated this work: [6].

However, it lacks specificity and clarity for their favored approach, and tries to cover a lot of ground which is not entirely needed for simply trying to translate Whale.

Several studies have explored backtranslation techniques to improve neural machine translation. Notably, the work by Lambert et al. provides a method for improving unsupervised neural machine translation with semantically weighted backtranslation, which is particularly relevant for morphologically rich and low-resource languages [4]. Another significant contribution is by the Language Resource Association, which presents an approach for training data self-correction to enhance unsupervised neural machine translation [5].

3 Data Preparation

The initial data consists of a large corpus of whale vocalizations, encoded as text with accompanying environmental context. Each vocalization is attributed to a specific whale, with information about the time and surrounding whales. This data is formatted similarly to a theatrical script, with each entry requiring less than 2048 tokens.

4 Model Architecture

Three decoder-only transformer language models are utilized:

- M_w: Trained solely on whale vocalizations
- M_e: Trained solely on English text
- M_{we}: Trained on both whale vocalizations and English text

5 Training Process

5.1 English Context Fine-tuning

M_e is fine-tuned on relevant English context, including:

- English plays
- Scientific literature about whales
- Biology and science textbooks focusing on whales
- Specific context about the region and family of whales being studied

5.2 Whale Language Model Training

M_w is trained to predict the next tokens in the whale text, excluding environmental context unless dependent on whale actions.

5.3 Bilingual Model Training

M_{we} is trained on both whale vocalizations and English text, with a focus on maintaining proficiency in both languages.

6 Translation Methodology

6.1 Whale to English Translation

M_{we} is prompted with:

```
[Environmental context]
[Paragraph of Whale]
Please translate the above paragraph to English:
```

The resulting translation is denoted as *whale_to_english*.

6.2 English to Whale Back-translation

The *whale_to_english* translation is then used to prompt M_{we} again:

```
[Environmental context]
[Paragraph of Whale translated to English (whale_to_english)]
Please translate the above paragraph to Whale:
```

The result is denoted as *whale_to_english_back_to_whale*.

7 Penalty Functions

Three penalty functions are employed to refine the translation process:

- P_{we}: English likelihood penalty
- P_{we}: Whale context probability penalty
- P_{sd}: Semantic distance penalty

7.1 English Likelihood Penalty (P_{we})

P_{we} is calculated using M_e to evaluate the average probability per token of the *whale_to_english* paragraph. A high P_{we} indicates an unlikely English paragraph.

7.2 Whale Context Probability Penalty (P_{we})

P_{we} is determined by inserting *whale_to_english_back_to_whale* into the overall whale text and calculating the probability using M_w. A high P_{we} indicates low probability of the translated whale text appearing in that context.

7.3 Semantic Distance Penalty (P_{sd})

P_{sd} measures the semantic distance between the original whale paragraph and *whale_to_english_back_to_whale*. A high P_{sd} indicates significant divergence in meaning. This penalty is weighted higher than P_{we} to prioritize meaning preservation.

8 Putting it all together

The penalties are used for reinforcement learning applied to M_{we}. This process is repeated for all available paragraphs, potentially tens of thousands of times, effectively fine-tuning the model on a paragraph-by-paragraph basis. Here is a summary of the training algorithm above:

Algorithm 1 Backtranslation Algorithm

1: Initialize parameters θ
2: **for** each iteration **do**
3: Translate whale vocalizations to English
4: Calculate probability penalty Pwe
5: Translate English back to whale
6: Calculate contextual and semantic penalties $Pwew$ and Psd
7: Update parameters θ based on combined penalties
8: **end for**

8.1 Datasets

The datasets provided a comprehensive set of whale vocalizations and environmental contexts, essential for training and evaluating our models.

I would primarily be interested in using two whale datasets for this work: 1)

```
Bermant, P.C., Bronstein, M.M., Wood, R.J. et al. Deep Machine Learning
    Techniques for the Detection and Classification of Sperm Whale
    Bioacoustics. Sci Rep 9, 12588 (2019).
    https://doi.org/10.1038/s41598-019-48909-4

For more detailed whale datasets, we will use recordings from socially
    segregated, sympatric sperm whale clans in the
```

Atlantic Ocean \cite{rsos}. These recordings were conducted over a decade, from 2005 through 2015, covering
approximately 2000 km^{2} along the western coast of Dominica (15.30 degree N, 61.40degree W). Membership in social
units was designated based on photo-identification analysis and long-term spatio-temporal coordination among unit
members. Acoustic recordings were made during deep foraging dives and socializing at the surface, using various systems
with flat frequency responses across ranges of at least 2-20 kHz and sampling rates of 44.1 kHz or higher. For this
study, the focus will be on the EC1 clan, which predominates in the study area and neighboring waters \cite{rsos}.

Using LStM and GRU Rnns to classify codas, determine vocal clan, and recognize individual
whales. Using LSTM RNNs, we construct deep artificial neural networks, which we train to perform a num-
ber of classification tasks based on high-quality manually annotated datasets. Data used in this study comes
from two long-term field studies on sperm whales: (1) off the island of Dominica in the eastern Caribbean 31
; and
(2) across the ETP but with a focus on the Galapagos Islands 58 . The two study sites employed similar methods
common in the study of this species to detect, follow, and record sperm whales as well as similar analytical tech-
niques to define coda types, delineate vocal clans, and identify individual whales (details in 29,44,59
). The Dominica dataset contains ~9,000 annotated codas that were collected across 3,834 hours with whales on 406 days
over 530
days of effort from 2005-2016. This represents recordings from over 10 different social units who are members of
two vocal clans referred to as "EC1" and "EC2" 31,44
. The ETP dataset contains ~17,000 annotated codas collected
between 1985-2014 based on recordings of over 89 groups who are members of six vocal clans29,59
. The annotated
datasets contain inter-click interval (the absolute time between clicks in a coda, ICI) vectors classified according
to categorical coda type, vocal clan membership, and (in the case of the Dominica dataset) individual whale
identity using the methods outlined in the publications listed above. While codas tend to be comprised of 3-40
clicks 27 , we restrict the analysis to codas with at most 10 clicks using the Dominica dataset or at most 12 clicks
using the ETP dataset.

Abstract

The Dominica database is used to evaluate and compare the performance of sperm whale click detectors. The database consists of 3 hours of recordings of sperm whale echolocation clicks and 4 hours of sound recordings containing delphinid clicks and transients from different sound sources, but no sperm whale clicks. Data were collected in September 2023 approximately 2 km west off the coast of Dominica Island in the eastern Caribbean using a home-built recorder equipped with an M36 Geospectrum hydrophone with near-constant sensitivity in the frequency range of 2 kHz to 20 kHz and a sampling rate of 192 kHz. The recordings lasted 5 days, with the recorder deployed at a depth of 20 m and anchored to a free-drifting buoy on the surface. The water depth was about 1,600 m.
Instructions:

This database is divided into two folders: Signal parts: consist of recordings containing sperm whales' clicks. Noise parts: consist of continuous recordings containing delphinid clicks and transients from various noise sources, but no sperm whale clicks. Additionally, a mat file named "Annotations_Dominica" is included, which contains a cell array with 28,300 annotations of sperm whale echolocation clicks. The array is structured in two columns: one describes the names of the recordings and the other contains a vector of the arrival time of the annotated clicks (in seconds).

https://ieee-dataport.org/documents/dominica-dataset
(note: it is ~18Gb download, but it seems this is the only data source with consecutive clicks).
"We demonstrate that LSTM and GRU, RNN architectures used for speech recognition and text translation, are able to classify codas into recognizable types and to accurately predict the clan membership and individual identity of the signaler."

"In addition, our results show the feasibility of using "self-supervised" learning on proxy tasks and applying trained neural networks to label unseen coda data according to type, clan, and individual whale, which could drastically expedite the analysis of recorded sperm whale signals.
The datasets in this study collectively contain ~26,000 annotated sperm whale codas and represent the largest available labeled data of sperm whale coda signals. Using these datasets, this study exhibits the potential applications of ML to sperm whale vocalizations by demonstrating the ability of NNs to perform detection and classification tasks with high degrees of accuracy"
This dataset is what is run to classify, although:
https://github.com/dgruber212/Sperm_Whale_Machine_Learning/tree/master

2) Recordings of 1203 tagged clicks of a single whale from the Bahamas with 120 sec of noise-only recording and recordings of over 15,000 manually tagged clicks from multiple whales recorded on the Dominica Island together with 3.6 hours of noise-only data [2].

3)

```
Evidence of social learning across symbolic cultural barriers
in sperm whales
@misc{leitao2024evidencesociallearningsymbolic,
      title={Evidence of social learning across symbolic cultural
          barriers in sperm whales},
      author={Antonio Leitao and Maxime Lucas and Simone Poetto and
          Taylor A. Hersh and Shane Gero and David Gruber and Michael
          Bronstein and Giovanni Petri},
      year={2024},
      eprint={2307.05304},
      archivePrefix={arXiv},
      primaryClass={cs.SI},
      url={https://arxiv.org/abs/2307.05304},
}

"We model the internal structure of codas, in terms of rhythmic
    variations at the level of clicks,
by using variable length Markov chains (VLMCs). Our analytical pipeline
    is illustrated in Fig. 1C.
We build each VLMC in two main steps. We first convert codas, naturally
    represented as sequences
of continuous, absolute, inter-click intervals (ICIs), to sequences of
    discrete ICIs (dICIs), by dis-
cretizing time into bins. In this way, each dICI represents a narrow
    range of possible ICI values.
The bins have a fixed width (or resolution) deltat and thus implicitly
    correspond to the temporal
resolution of our representation (see Methods for details on the optimal
    choice of deltat). Note that
although ICIs have units of time (seconds), dICIs are (unit-less)
    symbols (e.g. A, B, C, etc.), rep-
resenting multiples of deltat (and so the smaller deltat, the more the
    symbols). For example, the shortest
ICIs will be mapped to the symbol A whereas longer ones will be mapped
    to symbols further down
the alphabet. Hence, each coda (a sequence of ICIs) is mapped to a
    sequence of discrete symbols
(a sequence of dICIs). The second part of the pipeline focuses on
    modeling the internal structure
of codas in terms of dICI sequences. Essentially, we want to estimate
    transition probabilities from
a dICI sub-sequence to the next dICI (Fig. 1B)."
```

So that means they have datasets which are 1. labelled by coda 2. sequential temporally. But I couldn't find anything that seemed like it matched that.

"The information about vocal style contained in subcoda trees is sufficient to recover the social
structure of sperm whales (social units and clans). We show this in two ways. First, we analyze a
dataset from sperm whales in Dominica (Dominica dataset) [21]. This dataset has rich annotations
(coda type annotations, identity of recorded whales, social relations of recorded whales) which
makes it particularly useful for validation"

BUT
following the link, this is just for \bibitem{rsos} (Socially segregated, sympatric sperm whale clans in the Atlantic Ocean).

"Socially segregated, sympatric sperm whale clans in the Atlantic Ocean" only seems to have one dataset:

https://datadryad.org/stash/dataset/doi:10.5061/dryad.53g73

that contains, e.g.:

codaNUM ICI1 ICI2 ICI3 ICI4 ICI5 ICI6 ICI7 ICI8 ICI9 typeName Unit
1 0.293 0.282 0.298 0.315 0 0 0 0 0 5 5R3 1

4931 rows of that.

There is no identity, social relationship!
So I don't see where that data comes from.

Further on, they say:

"Motivated by these results, we extend our analysis to a much larger dataset from the Pacific
Ocean (Pacific dataset) [11]. This dataset is more sparsely annotated because of the breadth of its
spatial coverage. We restricted our analyses to a well-sampled subset (n = 57 coda samples) of the
full Pacific dataset (see Methods for details). Coda samples are only labeled by the spatial position
at which they were recorded, but no information is available about the identity of the vocalizing
sperm whales (see Methods for details). In fact, each repertoire likely contains codas from multiple

```
individuals of a single clan. It has recently been shown that these coda
    samples can be divided
into seven vocal clans based on their coda usage [11]. We use those
    clans as a benchmark for the
following analysis.
Since there is no social unit-level information for this dataset, we fit
    a subcoda tree for each
repertoire (i.e., all of the codas recorded on a single day in a single
    region)."

Okay that's great, looks like there should be lots of good sequentially
    utterred codas for us.
[11] Hersh, T. A. et al. Evidence from sperm whale clans of symbolic
    marking in non-human
cultures. Proc. Natl. Acad. Sci. 119, e2201692119 (2022).

Indeed the data are here:

https://osf.io/2jhd4

This dataset looks good, although of course it's annoying we don't know
    which whale made each coda. On a day with a lot of data there might
    be 200 codas. That seems a bit odd. For the machine learning paper
    ((2) above), there were 28,300 annotations of sperm whale
    echolocation clicks over what appears to be 5 days. I guess this is
    because researchers were manually annotating! So most likely, the
    machine learning dataset above is what we start with.
```

Also see:

- Project CETI
- Project CETI Publications
- Project CETI Workshop - Decoding Communication in Nonhuman June 12-13, 2023 @ Simons Institute for the Theory of Computing

9 Validation

Validation is performed using unseen whale vocalizations, comparing the accuracy of *whale_to_english_back_to_whale* to that of the training data.

10 Test Case Refinement - Low resource human languages

There is a need to refine the methodology with a test case with a known solution. This will:

1. Identify the appropriate size of the language model needed.
2. Help to experiment and improve methodology for the test case, which should also improve the whale case.

It is important to use a language that has very little leakage into the training set for an English language model, making it reasonably analogous to the translation from whale.

The methodology to test and refine the method outlined above with low resource human languages is as follows.

Using an isolate language ensures there is no leakage of that language's grammar or words into English, preventing an easier translation due to prior knowledge.

The context of the recordings should be significantly different from the context in which English is spoken. This similarity to the different context of whale vocalizations makes the task more analogous.

Starting with a spoken language and using a similar approach as the SETI project is also beneficial. This includes grouping into phonemes and converting those phonemes into letters of the English alphabet.

The use of a language that is particularly complex and foreign to English speakers is more appropriate. The more complex and alien the language, the more analogous it is expected to be to whale vocalizations.

Additionally, certain features of whale communication can be incorporated into test languages. For example, human languages with clicks and tones may be useful due to whales' use of clicks and tone as markers of meaning.

10.1 Selection of Test Language

In selecting a test language, several criteria were considered to ensure the chosen language is appropriately analogous to whale vocalizations:

1. Linguistic Isolation: The language should be a linguistic isolate or significantly different from English in terms of structure and origins.
2. Extensive Documentation: There should be a substantial corpus of digitized texts or recordings available for study.
3. Cultural Distinction: The culture associated with the language should be distinct and relatively unrelated to Western cultures to avoid cultural leakage.
4. Availability of Audio Recordings: Ideally, the language should have historical audio recordings that capture it in a native, traditional context.

Several languages were considered based on these criteria:

Old Javanese (Kawi) An ancient language used in Java from the 8th to 16th centuries, with a rich literary tradition and a distinct Hindu-Buddhist culture. Significant corpus of digitized texts is available, including epic poems and religious texts.

Classical Nahuatl The language of the Aztec Empire, part of the Uto-Aztecan family, with extensive digitized texts including the Florentine Codex and Cantares Mexicanos.

Navajo An Athabaskan language with a complex structure and extensive historical recordings, including those made during World War II by Navajo Code Talkers. It has minimal leakage into mainstream English culture.

Ket A Yeniseian language spoken in Siberia, with a limited number of speakers and some recorded materials. It is an isolate with unique linguistic features and minimal exposure to Western culture.

Xhosa and Zulu Bantu languages with click consonants, spoken in South Africa. They have significant recordings and a rich cultural heritage, though there has been some cultural leakage into Western consciousness.

After careful consideration, **Navajo** was selected as the most suitable test language. Its polysynthetic structure, making it likely the most complex language to learn from a model that only knows english, extensive historical recordings, and minimal cultural leakage make it an ideal candidate for refining the translation methodology. Additionally, the availability of recordings in traditional contexts provides valuable data for testing the translation framework, and the ability of modern translation methods allows for easy validation of the final quality of the unsupervised translation framework.

10.2 Potential Datasets for Test Task

- The Endangered Languages Archive (ELAR)
- The World Oral Literature Project, which has collected recordings of endangered languages worldwide

These datasets allow testing when converting from a given spoken language to a series of letters corresponding to vocals, similar to Navajo. It helps establish a baseline for the ability of an unsupervised translation algorithm to effectively translate an unknown language into a known language.

11 Potential Challenges

1. Poor performance of M_w.
2. Insufficient context for accurate translation.
3. Tendency towards common whale utterances in back-translation.
4. Development of a system to "cheat" the translation process.

5. Unclear direction of penalties for efficient learning.
6. Overfitting to specific training data.
7. Risk of translating only sounds and grammar without capturing meaning.

12 Miscellaneous end notes

It might be better to have an ensemble of answers for a given translation whale_to_english_back_to_whale, and then boost the model with the best of these answers. That addresses 5, as it gives the model a direction to move the prediction.

Note: this ends up looking like an auto encoder! We are expanding into English language token space, and then going back to the original space.

Another way to help it start in the right direction is to use statistical matching: most common concepts are assumed to be all the whale words in English. So then you can give it context: given the above best guess translation, improve it to a better one. And then you pick the best answer closest to the input.

13 Conclusion

This approach presents a novel method for translating whale vocalizations to human language using transformer-based models and reinforcement learning. While challenges exist, the use of multiple models and carefully designed penalty functions offers a promising direction for future research in animal communication translation.

References

[1] Roee Diamant, Anthony Knapr, Shlomo Dahan, Ilan Mardix, John Walpert, and Steve DiMarco. *Themo: The texas a&m-university of haifa-eastern mediterranean observatory*. In 2018 OCEANS-MTS/IEEE Kobe Techno-Oceans (OTO), pages 1–5. IEEE, 2018.

[2] Link to implementation code and database [online]. available at: https://drive.google.com/drive/folders/1HkGZcpYbrft3pGtqdt45bUZXVcrJVrdm?usp=sharing.

[3] Luca M. Lambert, Hal Whitehead, Shane Gero, "Socially segregated, sympatric sperm whale clans in the Atlantic Ocean," *Royal Society Open Science*, 2016, available at http://dx.doi.org/10.1098/rsos.160061.

[4] Lambert et al., "Improved Unsupervised Neural Machine Translation with Semantically Weighted Back Translation for Morphologically Rich and Low Resource Languages," *ACL Anthology*, 2022, available at https://aclanthology.org/2022.acl-long.2.

[5] Language Resource Association, "Improving Unsupervised Neural Machine Translation via Training Data Self-Correction," *LREC-COLING 2024*, pages 8942–8954, May 2024, available at https://aclanthology.org/2024.lrec-coling.8942.

[6] "A Theory of Unsupervised Translation Motivated by Understanding Animal Communication," available at: https://arxiv.org/abs/2211.11081.