A Theory of Unsupervised Translation for Understanding Animal Communication

Shafi Goldwasser
UC Berkeley & Project CETI
shafi.goldwasser@berkeley.edu

David F. Gruber
Project CETI
david@projectceti.org

Adam Tauman Kalai
Microsoft Research & Project CETI
adam@kal.ai

Orr Paradise
UC Berkeley & Project CETI
orrp@eecs.berkeley.edu

Abstract

Unsupervised translation refers to the challenging task of translating between two languages without parallel translations, i.e., from two separate monolingual corpora. We propose an information-theoretic framework of unsupervised translation that models the case where the source language is that of highly intelligent animals, such as whales, and the target language is a human language, such as English. In particular, there may be limited quantities of source data, the source and target languages may be quite different in nature, and few assumptions are made on the source language syntax.

We apply our theory to a stylized setting of tree-based languages. Our analysis suggests that the amount of source data required for unsupervised translation is not significantly more than that of supervised translation. Our analysis is purely information-theoretic; issues of algorithmic efficiency are left for future work.

We are motivated by an ambitious initiative to translate whale communication using modern machine translation techniques. The recordings of whale communication that are being collected have no parallel human-language data.

1 Introduction

The current amazing success of natural language translation by neural networks ultimately depends on the availability of parallel translated data from the source and target language, to be utilized as inputs to a translation system trained in a supervised manner.

While in the case of human languages some parallel data generally exist, in the case of translating animal communication, no parallel data exists. This makes existing methods of supervised language translation essentially irrelevant for understanding animal communication.

In light of this, we propose a theory of unsupervised translation. Our theory shows that, in lieu of parallel data, access to a good prior distribution over translations in the target language suffices for successful (unsupervised) translation, barring certain plausible ambiguities.

More concretely, we develop an information-theoretic framework for unsupervised translation. We establish two conditions under which unsupervised translation is (information-theoretically) possible.

- Access to good prior: There is access to a prior distribution over translations (in the target language) that assigns non-trivial probability to the outputs of a hypothetical ground-truth translator function. Crucially, no access to the ground-truth translator function itself is
needed. Intuitively, such a prior distribution allows to distinguish between plausible and implausible translations which in turn enables to invalidate incorrect candidate translators.

- **No implausible ambiguities:** any alteration (ambiguation) of translations will be deemed implausible (i.e., occur with low probability) by the prior.

We then show an (inefficient) algorithm that provably finds an accurate translator function (i.e., close to the ground truth translator), as long as the above conditions hold and a sufficient number of samples from the source language are available. To illuminate our conditions and argue for their “naturalness”, we also propose a randomized model for generating simple tree-based languages, and prove that these languages satisfy our conditions, and are therefore translatable with high probability.

The upper bounds we derive on the number of samples needed for unsupervised translation (when all conditions are met), lead us to the main message of this paper:

**Main message.** The data requirements of unsupervised translation may not be significantly larger than those of supervised translation, except for plausible ambiguities which unsupervised translation cannot resolve.

**On computational tractability.** Unsupervised translation involves solving a massive puzzle that is more computationally demanding than the supervised translation analog. This paper focuses purely on sample complexity bounds, leaving the algorithmic challenge (of efficiently learning from samples) for future work.

### 2 Related work

**Project CETI.** Andreas et al. [2022] present CETI’s initial scientific roadmap for understanding sperm whale communication, identifying the potential for unsupervised translation to be applied to whale communication. That roadmap, however, involves training a generative language model for whale communication. This is in contrast to our analysis suggesting that the number of Whalish samples necessary for translation is on the order of what is required for ordinary (supervised) MT. Our work may inform the amount and type of animal data collected with the the recent interest in translating the communication of various different animals [Andreas et al., 2022, Anthes, 2022].

**Goal-oriented communication.** It is interesting to contrast our work with the work on goal-oriented communication which was introduced by Juba and Sudan [2008] and extended in [Goldreich et al., 2012]. They study the setting of two communicating parties (one of which is trying to achieve a verifiable goal) using each a language completely unknown to the other. They put forward a theory of goal-oriented communication, where communication is not an end in itself, but rather a means to achieving some goals of the communicating parties. Focusing on goals provides a way to address “misunderstanding” during communication, as in when one can verify whether the goal is (or is not) achieved. Their theory shows how to overcome any initial misunderstanding between parties towards achieving a given goal. Our setting is different: Informally, rather than be a participant in a communication with someone speaking a different language, we wish to translate communications between two external parties speaking in a language unknown to us and there is no verifiable goal to aid us in this process.

**Unsupervised translation.** In unsupervised machine translation [Ravi and Knight, 2011], a translator between two languages is learned based only on monolingual corpora from each language. Unsupervised neural machine translation (UNMT) [Miceli Barone, 2016, Lample et al., 2018a, Artetxe et al., 2019, Lample et al., 2018b, Song et al., 2019] refers to neural network-based approaches to this task, for example those that employ a neural language model pre-trained on monolingual corpora.

Empirical evaluation of UNMT found that it is outperformed by its supervised setting, even when UNMT is trained on several orders of magnitude more data [Marchisio et al., 2020, Kim et al., 2020]. However, our theory shows that when our conditions are met, sample complexity should remain roughly the same between the supervised and unsupervised settings, barring computational constraints. This discrepancy points to an important future direction, namely, generalizing our framework to the computational setting.
3 A framework for unsupervised translation

We model a translator as a function \( f_\theta : \mathcal{X} \rightarrow \mathcal{Y} \) from source language \( \mathcal{X} \) to target language \( \mathcal{Y} \). Our goal is to learn \( \theta \in \Theta \), say, a realization of parameters in a neural network. Each language is thought of as a distribution over sentences. \(^1\) More specifically, the source language distribution over \( x \in \mathcal{X} \) by \( \mu \). For a sentence \( x \in \mathcal{X} \), you can think of \( \mu(x) \in [0, 1] \) as corresponding to how natural (or likely to occur) the sentence \( x \) is in the source language.

We focus on the case in which \( x \) is in a textual format. For further simplicity, we consider lossless translators \( f_\theta : \mathcal{X} \rightarrow \mathcal{Y} \) that are 1–1, meaning that it is possible to invert \( f_\theta \) and recover any \( x \in \mathcal{X} \) given \( y = f_\theta(x) \). This excludes a translator that always outputs, say, “A whale communication”, which is correct but imprecise.

We assume the existence of a ground-truth translator \( f_1 : \mathcal{X} \rightarrow \mathcal{Y} \). Moreover, we assume that \( f_1 \) is expressible in the class of possible translators, that is, \( t \in \Theta \). We can then define the translated language distribution \( f_1 \circ \mu \) over the target language \( \mathcal{Y} \), which is obtained by sampling a sentence \( x \) from the source language \( \mu \) and translating it with \( f_1 \); formally, we define \( f_1 \circ \mu := \mu(f_1^{-1}(y)) \).

The first condition asserts that the prior \( \rho \) is statistically accurate with respect to ground-truth translations of the source language (i.e., the translated language distribution). Statistical accuracy is measured with respect to the cross-entropy between distributions, denoted by \( H \). Stated formally,

**Condition 4.1 (\( \eta \)-statistically accurate prior).** For a given \( \eta > 0 \),

\[
H(f_1 \circ \mu, \rho) := \mathbb{E}_{x \sim \mu} \left[ -\log \rho(f_1(x)) \right] \leq \eta.
\]

The second condition concerns the ambiguities of a translator \( f_\theta \). Informally, these are sentences that are confounded by the translator \( f_\theta(x) \neq f_1(x) \). Naturally, the impact an ambiguity has on the efficacy of a translator depends on the difference between the translations \( \ell(f_1(x), f_\theta(x)) \), as well as the likelihood of the translation itself—which is simply the likelihood of encountering the source sentence \( \mu(x) \). With these in mind, we define the ambiguity of \( f_\theta \) as \( \mathbb{E}_{x \sim \mu} [\ell(f_1(x), f_\theta(x))] \).

The second condition asserts that any sufficiently-ambiguous translator should be deemed implausible (unlikely) by the prior \( \rho \):

**Condition 4.2 (No plausible ambiguities).** For \( \epsilon, \gamma, K \), with respect to \( \mu, \rho \), parameter family \( \Theta \) and translator family \( \{ f_\theta \}_{\theta \in \Theta} \):

\[
\forall \theta \in \Theta \quad \mathbb{E}_{x \sim \mu} [\ell(f_1(x), f_\theta(x))] \geq \epsilon \implies \Pr_{x \sim \mu} [\rho(f_\theta(x)) < 2^{-K}] > \gamma.
\]

We show that Conditions 4.1 and 4.2 imply unsupervised translatability in the information-theoretic (as opposed to the computational) sense. We do this by devising an algorithm, called UNFISHY, that takes as inputs \( m \) iid samples \( x_1, \ldots, x_m \sim \mu \) in the source language, and access to a prior \( \rho \) over the target language, and outputs a translator \( f_\theta \).

The algorithm works by selecting the lossless parameters whose translations are the least “fishy” according to the prior \( \rho \), and minimizing a variant of the maximum-likelihood objective \( \sum_i -\log \rho(f_\theta(x_i)) \) over \( \theta \in \Theta \). The algorithm is computationally inefficient. One may think of it as roughly a process of elimination, where for each source sample \( x_i \), we eliminate all translators \( \theta \) where \( \rho(f_\theta(x_i)) \) is implausibly small. We prove sample complexity upper bounds as follows:

**Theorem 4.3.** Let \( \mu, \rho \) be probability distributions over \( \mathcal{X} \) and \( \mathcal{Y} \), and let \( \epsilon, \gamma, \delta \in (0, 1) \) and \( \eta, K > 0 \) with \( \eta < \gamma K/2 \). Suppose that Cond. 4.1 holds for \( \eta \), and Cond. 4.2 holds for \( \epsilon, K, \gamma \) and

\(^1\)This is just a terminological choice; our theory could be applied to \( x \)’s that are paragraphs, documents, etc.
\[ \Theta. \] Then, with probability at least \( 1 - \delta \) over
\[ m = O \left( \frac{\log |\Theta| + \log(1/\delta)}{\gamma^2} \right) \]
iid samples from \( \mu \), the translator \( f_\theta \) output by UNFISHY satisfies
\[ \mathbb{E}_{x \sim \mu} \left[ \ell(f_t(x), f_\theta(x)) \right] < \epsilon. \]

To compare with the supervised setting, the classic Occam’s Razor bound shows that Cond. 4.1 alone implies that with probability at least \( 1 - \delta \), given \( m = O \left( \frac{\log |\Theta| + \log(1/\delta)}{\epsilon^2} \right) \) iid labeled training samples \((x_i, f_t(x_i))\), one can find a classifier \( f_\theta \) with error at most \( \epsilon \). Of course, the challenge in our setting is that the learner is not given access to ground-truth labels \( y_i \).

5 Random sub-tree languages

We instantiate our conditions with a simplified, tree-based probabilistic model of language. At a high level, a random tree with nodes labeled by words will give plausible sentences by tracing root-to-leaf path, and the source language is derived from paths in a random sub-tree. See Fig. 1 and the details that follow.

The random sub-tree language (RT) is a distribution over source languages \( \mu \) and priors \( \rho \) parameterized by a vocabulary \( \mathcal{W} \), tree arities \( a < b \) and tree depth \( n \). First, a \( b \)-ary tree of depth \( n \) is constructed, with nodes randomly labeled by words from \( \mathcal{W} \). The set of plausible sentences \( P \) is obtained by tracing root-to-leaf paths in this tree, and the prior \( \rho \) distributes uniformly over \( P \). Next, an \( a \)-ary sub-tree is derived by choosing \( a \) random children of each node, in level-order traversal. The source language \( \mu \) is uniformly distributed over paths in the sub-tree (back-translated from the target language to the source language).

In the full version of this paper, we show that, with high probability over the generation of the source language and the prior, Conditions 4.1 and 4.2 are satisfied for an appropriate choice of parameters. This, we hope, supports the naturalness of our conditions. Finally, by applying Thm. 4.3 we conclude that random sub-tree languages are translatable with high probability.

We believe that the random sub-tree model and its analysis can be generalized to settings when there is not a perfect prior \( \rho \), that is, when \( \text{supp}(f_t \circ \mu) \nsubseteq \text{supp}(\rho) \), and to the case of non-uniform \( \mu \).

Acknowledgements

We thank Madhu Sudan, Yonatan Belinkov and the entire Project CETI team, especially Pratyusha Sharma, Jacob Andreas, Gašper Beguš, Dan Tchernov and Michael Bronstein for illuminating discussions. This study was funded by grants from Dalio Philanthropies and Ocean X; Sea Grape Foundation; Rosamund Zander/Hansjörg Wyss, Chris Anderson/Jacqueline Novogratz through The Audacious Project: a collaborative funding initiative housed at TED.
References


