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

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

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 their 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 $f_\theta \colon \mathcal{X} \to \mathcal{Y}$ 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 μ. 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 $f_\theta \colon \mathcal{X} \hookrightarrow \mathcal{Y}$ that are 1–1, meaning that it is possible to invert f_θ 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 $f_t \colon \mathcal{X} \to \mathcal{Y}$. Moreover, we assume that f_t is expressible in the class of possible translators, that is, that $t \in \Theta$. We can then define the *translated language distribution* $f_t \circ \mu$ over the target language $\mathcal{Y}$, which is obtained by sampling a sentence x from the source language μ and translating it with f_t; formally, we define $f_t \circ \mu := \mu(f_t^{-1}(y))$.

The final component of our framework is a *prior* distribution ρ over $\mathcal{Y}$. Ideally, the prior ρ should approximate the translated language distribution $f_t \circ \mu$. Looking ahead, access to a good prior will play a key role in our theorem, up next.

4 Sufficient conditions for unsupervised translatability

The first condition asserts that the prior ρ over the target language $\mathcal{Y}$ 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 (η-statistically accurate prior)**.** *For a given* $\eta > 0$*,*

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

The second condition concerns the *ambiguities* of a translator f_θ. Informally, these are sentences that are confounded by the translator $f_\theta(x) \neq f_t(x)$. Naturally, the impact an ambiguity has on the efficacy of a translator depends on the difference between the translations $\ell(f_t(x), f_\theta(x))$, 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_θ as $\mathbb{E}_{x \sim \mu}[\ell(f_t(x), f_\theta(x))]$.

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

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

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

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 ρ over the target language, and outputs a translator $f_{\hat{\theta}}$.

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

Theorem 4.3. *Let* μ, ρ *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* η*, and Cond. 4.2 holds for* ϵ, K, γ *and*

[1] This is just a terminological choice; our theory could be applied to x's that are paragraphs, documents, etc.

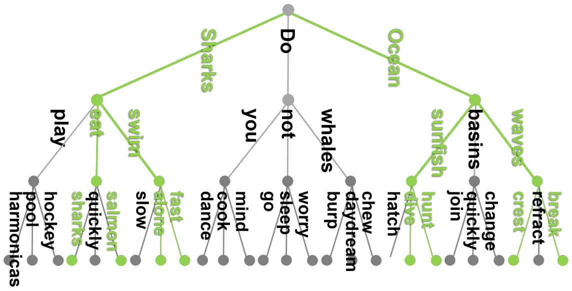

Figure 1: An example of a language tree G with $b = 3$, and the sub-tree H with $a = 2$ illustrated in green.

Θ. *Then, with probability at least* $1 - \delta$ *over*

$$m = O\left(\frac{\log|\Theta| + \log(1/\delta)}{\gamma^2}\right)$$

iid samples from μ*, the translator* $f_{\hat{\theta}}$ *output by* UNFISHY *satisfies*

$$\mathop{\mathbb{E}}_{x\sim\mu}\left[\ell(f_t(x), f_{\hat{\theta}}(x)\right] < \epsilon.$$

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 $m = O(\frac{\log|\Theta| + \log(1/\delta)}{\epsilon})$ iid *labeled* training samples $(x_i, f_t(x_i))$, one can find a classifier f_θ with error at most ϵ. 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 μ and priors ρ 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 ρ 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 μ 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 ρ, that is, when $\text{supp}(f_t \circ \mu) \not\subseteq \text{supp}(\rho)$, and to the case of non-uniform μ.

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/Hansjorg Wyss, Chris Anderson/Jacqueline Novogratz through The Audacious Project: a collaborative funding initiative housed at TED.

References

Jacob Andreas, Gašper Beguš, Michael M. Bronstein, Roee Diamant, Denley Delaney, Shane Gero, Shafi Goldwasser, David F. Gruber, Sarah de Haas, Peter Malkin, Nikolay Pavlov, Roger Payne, Giovanni Petri, Daniela Rus, Pratyusha Sharma, Dan Tchernov, Pernille Tønnesen, Antonio Torralba, Daniel Vogt, and Robert J. Wood. Toward understanding the communication in sperm whales. *iScience*, 25(6):104393, 2022. ISSN 2589-0042. doi: https://doi.org/10.1016/j.isci.2022.104393. URL `https://www.sciencedirect.com/science/article/pii/S2589004222006642`.

Emily Anthes. The animal translators. *The New York Times*, Aug 2022. URL `https://www.nytimes.com/2022/08/30/science/translators-animals-naked-mole-rats.html`.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Unsupervised neural machine translation, a new paradigm solely based on monolingual text. *Proces. del Leng. Natural*, 63:151–154, 2019. URL `http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6107`.

Oded Goldreich, Brendan Juba, and Madhu Sudan. A theory of goal-oriented communication. *J. ACM*, 59(2):8:1–8:65, 2012. doi: 10.1145/2160158.2160161. URL `https://doi.org/10.1145/2160158.2160161`.

Brendan Juba and Madhu Sudan. Universal semantic communication I. In Cynthia Dwork, editor, *Proceedings of the 40th Annual ACM Symposium on Theory of Computing, Victoria, British Columbia, Canada, May 17-20, 2008*, pages 123–132. ACM, 2008. doi: 10.1145/1374376.1374397. URL `https://doi.org/10.1145/1374376.1374397`.

Yunsu Kim, Miguel Graça, and Hermann Ney. When and why is unsupervised neural machine translation useless? In Mikel L. Forcada, André Martins, Helena Moniz, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof Arenas, Mary Nurminen, Lena Marg, Sara Fumega, Bruno Martins, Fernando Batista, Luísa Coheur, Carla Parra Escartín, and Isabel Trancoso, editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020, Lisboa, Portugal, November 3-5, 2020*, pages 35–44. European Association for Machine Translation, 2020. URL `https://aclanthology.org/2020.eamt-1.5/`.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018a. URL `https://openreview.net/forum?id=rkYTTf-AZ`.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 5039–5049. Association for Computational Linguistics, 2018b. URL `https://aclanthology.org/D18-1549/`.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. When does unsupervised machine translation work? In Loïc Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors, *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 571–583. Association for Computational Linguistics, 2020. URL `https://aclanthology.org/2020.wmt-1.68/`.

Antonio Valerio Miceli Barone. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 121–126, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-1614. URL `https://aclanthology.org/W16-1614`.

Sujith Ravi and Kevin Knight. Deciphering foreign language. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 12–21. The Association for Computer Linguistics, 2011. URL `https://aclanthology.org/P11-1002/`.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: masked sequence to sequence pre-training for language generation. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR, 2019. URL `http://proceedings.mlr.press/v97/song19d.html`.