A Theory of Unsupervised Translation for Understanding Animal Communication

Anonymous Author(s)
Affiliation
Address
email

Abstract

Unsupervised translation generally refers to the challenging task of translating between two languages without parallel translations, i.e., from two separate monolingual corpora. In this work, we propose an information-theoretic framework of unsupervised translation that can be well suited even for 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.

We identify two conditions that combined allow for unsupervised translation: (1) there is access to an *prior* distribution over the target language that estimates the likelihood that a sentence was translated from the source language; and (2) most alterations of translations are deemed implausible (i.e., unlikely) by the prior. We then give an (inefficient) algorithm which, given access to the prior and enough unlabeled source examples as input, outputs a provably accurate translation function. Surprisingly, our analysis suggests that the amount of source data required (information theoretically) for unsupervised translation is not significantly greater than that of supervised translation, i.e., the standard case where one has parallel translated data for training.

To support the viability of our theory, we propose a simplified probabilistic model of language: the *random sub-tree language* model, in which sentences correspond to paths in a randomly-labeled tree. We prove that random sub-tree languages satisfy conditions (1-2) with high probability, and are therefore translatable by our algorithm.

Our theory is motivated by a recent initiative to translate whale communication using modern machine translation techniques. The recordings of whale communications that are being collected have no parallel human-language data. We are further motivated by recent *empirical* work, reported in the machine learning literature, demonstrating that unsupervised translation is possible in certain settings.

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.

Submitted to Information-Theoretic Principles in Cognitive Systems Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

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 three 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. We believe that a language model such as GPT-3 can be adapted to obtain such a prior, as described in Appendix E.

- *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.

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: data requirements of unsupervised translation are not significantly larger than those of supervised translation, except for certain *plausible ambiguities* which unsupervised translation cannot resolve. This surprising conclusion is at odds with recent empirical studies Marchisio et al. [2020], Kim et al. [2020], that conclude that state-of-the-art unsupervised translators are outperformed by their supervised counterparts. To the best of our knowledge, our work is the first *theoretical* work which analyzes the conditions that allow for such a translation. We discuss this discrepancy in Appendix A.

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.

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]. A summary of related work is deferred to Appendix A.

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 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.

While x can contain raw audio and other metadata such as location (see Appendix F), for now consider 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.

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

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 $t \in \Theta$.[2] 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 $f_t \circ \mu \coloneqq \mu(f_t^{-1}(y))$.

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.

A formal description of the framework can be found in Appendix B.

3 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 3.1 (η-statistically accurate prior)**.** *For a given $\eta > 0$,*

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

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 3.2 (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 3.1 and 3.2 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, 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 3.3. *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. 3.1 holds for η, and Cond. 3.2 holds for ϵ, K, γ and Π_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*

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

A detailed discussion and proof of Thm. 3.3 are deferred to Appendix C.

[2] Any determinstic sentence-to-sentence is likely to be somewhat inaccurate. We leave it to future work to build on Appendix F and generalize our theory to the agnostic case, that is, when no such f_t exists.

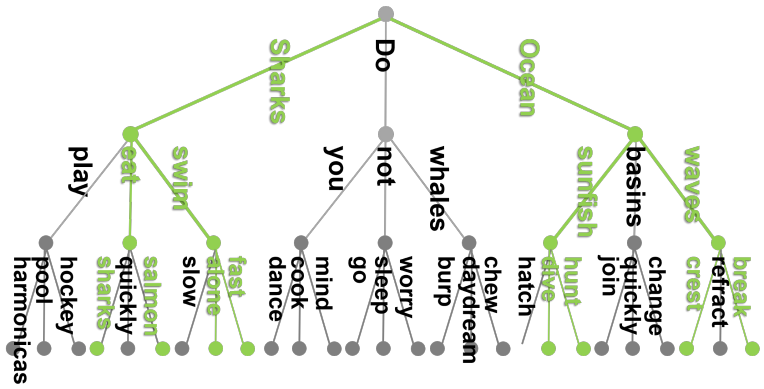

Figure 1: An example of a language tree G with $b = 3$, and the sub-tree H with $a = 2$ illustrated in green.

To compare with the supervised setting, the classic Occam's Razor bound shows that Cond. 3.1 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. Appendix D explains this, other generalization bounds, and their relationship to unsupervised translation. Of course, the challenge in our setting is that the learner is not given access to ground-truth labels y_i.

4 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 Appendix G we show that, with high probability over the generation of the source language and the prior, Conditions 3.1 and 3.2 are satisfied for an appropriate choice of parameters. This, we hope, supports the naturalness of our conditions. Finally, by applying Thm. 3.3, 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 $\mathrm{supp}(f_t \circ \mu) \not\subseteq \mathrm{supp}(\rho)$, and to the case of non-uniform μ.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. URL https://arxiv.org/abs/2204.02311.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Benjamin Doerr. Probabilistic tools for the analysis of randomized optimization heuristics. In *Natural Computing Series*, pages 1–87. Springer International Publishing, nov 2019. doi: 10.1007/978-3-030-29414-4_1. URL https://doi.org/10.1007%2F978-3-030-29414-4_1.

Shane Gero, Hal Whitehead, and Luke Rendell. Individual, unit and vocal clan level identity cues in sperm whale codas. *Royal Society Open Science*, 3(1):150372, 2016.

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.

William E Nagy and Richard C Anderson. How many words are there in printed school english? *Reading research quarterly*, pages 304–330, 1984.

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/.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.346. URL https://aclanthology.org/2020.emnlp-main.346.

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.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. Facebook ai's WMT21 news translation task submission. In Loïc Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, Tom Kocmi, André Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation, WMT@EMNLP 2021, Online Event, November 10-11, 2021*, pages 205–215. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.wmt-1.19.

A Related work

Project CETI. The sperm whale data collection effort began with a longitudinal dataset from a community of whales off the coast of Dominica that revealed interesting communication findings, such a dialects and vocal clans Gero et al. [2016]. The new CETI effort is to collect custom-build passive bioacoustic arrays (installed in Fall 2022) covering a 20×20 kilometer area where these whale families reside (collecting over 3TB/month) in tandem with on-whale robotic acoustic and video tags,

underwater (robotic swimming fish) and areal drones as well as other observation techniques in effort to augment rich contextual communication data.

Andreas et al. [2022] presents 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, like GPT-3, 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. How much data is required for translation and generation? Generative language models are notoriously data-hungry [Brown et al., 2020, Chowdhery et al., 2022] with PaLM being trained on 780 billion tokens.

Unlike with human languages, the project cannot rely on typical translation techniques that use defined inputs which motivated us to refine machine learning techniques that use unsupervised methods to translate human languages without relying on discrete inputs.

Sperm whales were chosen, in particular, due to their highly-developed neuro-anatomical features, cognitive abilities, and social structures. Furthermore, their discrete click-based communication make for an excellent starting point for machine learning tools that can be applied to other animals in the future.

In terms of the amount of data that can be collected (i.e., sample inputs in the source language using the terminology of the current paper), the current estimates are as follows. The data will be collected in a 20×20 km area off the shore of Dominica which sees the presence of up to 400 whales depending on the season. They are vocalizing continuously, with about 75–80% being echolocation clicks (searching for prey) and other ≈ 20–25% being the coda clicks (that are thought to constitute their main means of communication). Typical coda transmission rate is 0.25/sec and the typical coda size is 5 clicks. With these numbers in mind, an upper bound on the amount of data that can be collected is an order of 150M clicks/year.

Goal-oriented communication. It is interesting to contrast our work with the work on goal-oriented communication which was introduced in the work of Juba and Sudan Juba and Sudan [2008] and extended jointly with Goldreich 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 problem 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. Intuitively, we use assumed knowledge of a prior on the translation as a way to overcome incorrect translations and in this sense the two projects are similar: finding a way to eliminate candidates is a key stepping stone.

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 (see Section 1). This discrepancy points to an important future direction, namely, generalizing our framework to the computational setting (e.g., by studying an efficient analogue of Algorithm C.1).

Recent progress in language models. Language modeling refers to the task of predicting a missing word (or words) from a given. In recent years, large language models, such as GPT [Brown et al.,

2020], BERT [Devlin et al., 2018] and PaLM [Chowdhery et al., 2022] were shown to achieve state-of-the-art performance on many tasks in natural language processing (NLP) such as text generation, summariziation, or (supervised) MT. These models are indeed large, with hundreds of billions of parameters, and are pre-trained on hundreds of billions of tokens. We note that (supervised) neural machine translators use far less parameters; for example, Facebook AI's multilingual model achieved state-of-the-art performance on its benchmarks using only five billion parameters [Tran et al., 2021].

A pre-trained language model can be *fine-tuned* or *prompted* to solve a specialized task. Of particular interest to us is the latter method, as we discuss a prompting-based approach to getting a good prior (Appendix E).

B Framework and assumptions

B.1 Notation and basic definitions

We use $\log \coloneqq \log_2$ to denote the base-2 logarithm. For two distributions p and q, the *cross-entropy* between p and q is defined as

$$H(p, q) \coloneqq \mathop{\mathbb{E}}_{x \sim p} \left[-\log \left(q(x) \right) \right],$$

with $\log(0)$ is defined as ∞ for convenience. For any distribution p with finite support $\mathrm{supp}(p)$, we write

$$p(A) \coloneqq \Pr_{x \sim p}[x \in A].$$

For $f \colon \mathcal{X} \to \mathcal{Y}$ and $S \subseteq \mathcal{X}$, we write $f(S) \coloneqq \{f(x) \mid x \in S\}$. We use $f \colon \mathcal{X} \hookrightarrow \mathcal{Y}$ to denote a 1–1 function, in which $f(x) \neq f(x')$ for all $x \neq x'$. We use $A \sqcup B$ to denote the union between two disjoint sets A and B.

Throughout this paper we give several concrete examples of textual translations. To avoid confusion, we use sans-serif font to denote a word or sentence from the source or target language. So, for example, we can say that Pottwal is German for sperm whale.

B.2 A framework for unsupervised translation

Next, we describe a mathematical framework for analyzing unsupervised translation. In this framework, a *source language* is a distribution μ over a set of possible sentences $\mathcal{X}$. Similarly, a *target language* is a distribution ν over a set of possible sentences $\mathcal{Y}$. When clear from context, we may associate each language with its corresponding set of possible sentences.

A *translator* is a mapping $f \colon \mathcal{X} \hookrightarrow \mathcal{Y}$, where for simplicity we consider only 1–1 functions. Of particular interest is the *ground truth translator*, denoted f_t. Such a translator induced a *(ground truth) translated language*, which is the distribution τ over $\mathcal{Y}$ obtained by taking $f_t(x)$ for $x \sim \mu$. As you will soon see, most of our theorems and analysis will be stated in terms of τ, rather than μ.

Loss functions. For any set $\mathcal{Y}$, a (bounded) *loss function* is any $\ell \colon \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ such that $\ell(y, y) = 0$ for all $y \in \mathcal{Y}$. Intuitively, $\ell(y, y')$ is a measure of the difference in meaning between two sentences $y, y' \in \mathcal{Y}$. Of particular notoriety is the *0-1 (zero-one) loss*,

$$\ell_{01}(y, y') \coloneqq \begin{cases} 1 & y \neq y' \\ 0 & y = y'. \end{cases}$$

Parametrized functions (translators). We will commonly use a fixed and known set of 1–1 functions $\{f_\theta \colon \mathcal{X} \hookrightarrow \mathcal{Y}\}_{\theta \in \Theta}$ parameterized by a set of *parameters* Θ. We will call either $\theta \in \Theta$ or its corresponding function f_θ a *translator*. For simplicity, we assume that $\mathcal{X}, \mathcal{Y}$ (and hence $\mathcal{F}$) are all finite and give learning bounds in terms of $\log |\Theta|$, though Appendix D discusses infinite and over-parameterized generalizations.

C Conditions 3.1 and 3.2 imply translatability

We describe an unsupervised learning algorithm that, assuming Conditions 3.1 and 3.2 hold, learns an accurate translator with high probability. The inputs to our simple algorithm, called UNFISHY, are

Translator A	Translator B	Translator C
Have you seen mom? I just returned from the ocean basin. $p' = 4.0 \times 10^{-18}$	Have you seen any orcas today? I just got back from the reef. $p' = 8.9 \times 10^{-18}$	goose Translate good's 99. Hat out hat dsjghwetquetzvc!!! $p' = 2.0 \times 10^{-48}$
At the reef, there were a lot of sea turtles. $p = 7.3 \times 10^{-26}$	At the reef, there were a lot of sea turtles. $p = 1.5 \times 10^{-22}$	CETI bicycle Shufi underwater OMG and. $p = 3.2 \times 10^{-72}$

Figure 2: An example of three translations and their probabilities, denoted by p, computed using the GPT-3 API. The probabilities of just the first paragraphs are also shown as p'.

m iid samples $x_1, x_2, \ldots, x_m \sim \mu$ in the source language and a prior distribution ρ over the target language. In the context of our main motivation, these are m samples of Whalish communication and a language model over English; indeed, Whalish communication data is scarce while English language models of increasing size and accuracy are regularly being released.

Following a maximum-likelihood approach, the algorithm works by selecting the lossless translator parameters whose translations are the least "fishy" according to the prior distribtuion ρ, minimizing a variant of the 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.

C.1 Intuition via a concrete example

To understand the intuition behind UNFISHY, we use a hypothetical example illustrated in Fig. 2, where there are only three possible translators $|\Theta| = 3$. Let us suppose that B always produces accurate and natural translations, and further that they each work paragraph-by-paragraph, as modern supervised translation algorithms operate within some limited context window.

First consider only the first paragraphs output by the three translators and their corresponding probabilities p' as computed using the GPT-3 API, which we use as a simple approximation to ρ for this example. Translator C, whose output is complete gibberish, is eliminated due to its low probability assessment of $\approx 10^{-48}$ from GPT-3. However, the first paragraphs of A and B's output have similar probabilities $\approx 10^{-18}$ and are not enough to rule out the ambiguity ocean basin$\leftrightarrow$reef. Next, consider the full two-paragraph translations. The outputs of A and B are both grammatical, but B is more coherent (less fishy) because A abruptly switches from talking about the basin, the bottom of the ocean, to a reef. This is reflected in the fact that B's probability $\approx 10^{-22}$ is greater than that of A, $\approx 10^{-26}$, thus UNFISHY would output B. In fact, one can imagine A as a set of *isolated adversaries*[3] where each adversary is trying to mistranslate a paragraph, knowing the ground truth translation of their paragraph, while attempting to maintain the plausibility of the entire translation. If only the first-paragraph adversary mistranslates reef to ocean basin, then the translation lacks coherence and is unlikely. If the adversaries are in cahoots and coordinate to all translate reef to ocean basin, they would generate: Have you seen mom? I just returned from the ocean basin. At the basin, there were a lot of sea turtles. which has low probability $\approx 10^{-25}$, presumably because encoded in GPT-3's training data is the knowledge that there are no turtles deep in the ocean near the basin. While the adversary could also decide to change the word turtle to something else when it appears near basin, eventually it would get tangled up in its web of deceit. The results in this paper do not require paragraph-by-paragraph translators but instead require translators restricted in their effective number of parameters.

Since one can safely assume that whales are not talking about bicycles or Santa, we can refine the prior distribution ρ to one does not cover all of English but focuses on the plausible English translations $y \in Y$ of Whalish communications, so $\rho(y)$ is an estimate of the plausibility of y as a translation of $x \sim \mu$. For whales, this prior can be very broad as one can only speculate about

[3] It is not the case that we believe there is an actual adversary proposing bad translators. This standard adversarial perspective of theoretical computer science is based on the premise that an unsupervised translation algorithm will work in practice if it can identify a correct translator even in the presence of adversarially chosen alternatives. Since there are so many possible translators that we must avoid, it helps to think of an adversary as selecting the most likely alternative.

communication content. In Appendix E, we describe how such a prior ρ can be derived from existing language models such as GPT-3 [Brown et al., 2020] or PaLM Chowdhery et al. [2022].

C.2 Formal statement and analysis

Next, we define the UNFISHY algorithm[4] and analyse its performance when Conditions 3.1 and 3.2 are satisfied. UNFISHY takes three inputs. First is a prior probability distribution ρ over $\mathcal{Y}$, which is a prior distribution over plausible translations in the target language. One may simply take $\rho = \nu$ for an existing language model ν over $\mathcal{Y}$, but Appendix E discusses how to use prompts to improve a prior. The second input is m unlabeled examples $x_1, x_2, \ldots, x_m$ drawn iid from an (unknown) but fixed *source distribution* μ over $\mathcal{X}$. The third input is a plausibility threshold $K > 0$.

The UNFISHY algorithm itself is simple and is nearly the maximum-likelihood algorithm discussed in the introduction, except that we limit the penalty due to small $\rho(f_\theta(x))$.

Algorithm C.1 (UNFISHY)**.** *The* UNFISHY *algorithm takes as input black-box access to distribution ρ over $\mathcal{Y}$, and is explicitly given samples $x_1, \ldots, x_m \in \mathcal{X}$, a parameter family Θ, and a threshold $K > 0$. It runs as follows.*

- *Using exhaustive search, find $\hat{\theta} \in \Theta$ such that*

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{m} \sum_{i=1}^{m} \min\left(-\log \rho\left(f_\theta\left(x_i\right)\right), K\right)$$

- *Output $\hat{\theta}$.*

Theorem C.2 (Thm. 3.3, restated)**.** *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. 3.1 holds for η, and Cond. 3.2 holds for ϵ, K, γ and Π_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*

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

Proof. Fix $m = 8 \cdot \gamma^{-2} \cdot (\log(|\Theta|) + \log(1/\delta))$. Algorithm C.1 minimizes the following empirical objective over $\theta \in \Theta$:

$$\hat{\ell}(\theta) \coloneqq \frac{1}{m} \sum_{i \leq m} \hat{\ell}_i(\theta), \text{ where } \hat{\ell}_i(\theta) = \min\left\{-\log \rho(f(x_i)), K\right\}.$$

Define the "bad" θ's to be $D = \{\theta \in \Theta \mid \operatorname{err}(f_\theta) \geq \epsilon\}$, each of these corresponding to a $\pi \in \Pi_\Theta$. Since the min with K can only reduce the empirical objective, then by Cond. 3.1,

$$\mathbb{E}[\hat{\ell}(t)] \leq H(\tau, \rho) \leq \eta \leq \gamma K/2.$$

Further, for any $\theta \in D$, $\mathbb{E}[\hat{\ell}(\theta)] > \gamma K$ by Cond. 3.2. Now, since $|D| \leq |\Theta| - 1$, for Algorithm C.1 to select any $\theta \in D$, it must be the case that:

$$\hat{\ell}(t) \geq \frac{3\gamma K}{4} \text{ or } \hat{\ell}(f) \leq \frac{3\gamma K}{4} \text{ for some } f_\theta \in D.$$

Each of these $|D| + 1 \leq |\Theta|$ quantities is an average of m bounded random variables $\hat{\ell}_i(\theta) \in [0, K]$. Thus, by one-sided Chernoff bounds, the probability that any of them is off by $\geq \gamma K/4$ of their mean is at most $|\Theta| e^{-2m(\gamma/4)^2} < \delta$. □

[4] Note that whales are *not* fish.

Translation	Probability
I just ate a giant squid.	1.5×10^{-9}
I just ate a giant cheeseburger.	$\mathbf{1.9 \times 10^{-8}}$
A sperm whale said: I just ate a giant squid.	$\mathbf{6.8 \times 10^{-14}}$
A sperm whale said: I just ate a giant cheeseburger.	1.2×10^{-17}

Figure 3: Without using a prompt, the sentence I just ate a giant cheeseburger. is more likely, but using the prompt A sperm whale said:, the sentence I just ate a giant squid is much more likely.

D Bounds for supervised translation

In this work, for ease of presentation, we have interpreted $\log|\Theta|$ as an estimate of the number of examples needed to learn for supervised translation. This is based in the classic Occam bound for supervised learning:

Theorem D.1 (Realizable Occam bound)**.** *Let $\mathcal{X}$, $\mathcal{Y}$, be sets, $\mathcal{D}$ be a joint distribution over $\mathcal{X} \times \mathcal{Y}$, $\ell\colon \mathcal{Y} \times \mathcal{Y} \to [0,1]$ be a loss, and $f_\theta : \mathcal{X} \to \mathcal{Y}$ be a family of functions parameterized by $\theta \in \Theta$, and $\mathrm{err}(f_\theta) \coloneqq \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(y, f_\theta(x))]$. For any $\epsilon, \delta > 0$, $m \geq \ln(|\Theta|/\delta)/\epsilon$:*

$$\Pr_{(x_1,y_1),\ldots,(x_m,y_m)\sim\mathcal{D}^m}\left[\mathrm{err}(f_\theta) \leq \epsilon \textit{ for all } \theta \in \Theta \textit{ such that } \sum_i \ell(y_i, f_\theta(x_i)) = 0\right] \geq 1 - \delta.$$

This means that if you choose any classifier $\theta \in \Theta$ with zero training loss $\sum_i \ell(y_i, f_\theta(x_i)) = 0$ (which surely exists in the realizable case where there exists f_t with $\mathrm{err}(f_t) = 0$), then with high probability you will also have $\mathrm{err}(f_\theta) \leq \epsilon$. This is why $O(\log|\Theta|)$ can be thought of as the number of examples required for supervised translation when it is assumed that there is some perfectly accurate translator, since we view supervised translation as simply classification with many classes $\mathcal{Y}$.

Now, in some cases, one can learn with many fewer examples, which is important when parameters are real-valued and $|\Theta| = \infty$ or when the model is over-parameterized. However, one can analyze even these cases in our model using the following "trick." Suppose one has a supervised translation algorithm SUPER that takes m labeled examples (x_i, y_i) as input and outputs $\theta \in \Theta$. A simple observation in these cases is that one could use an initial set of m unlabeled examples, $x_1, \ldots, x_m$, to define a subset of translators:

$$\overline{\Theta} \coloneqq \left\{\textsc{Super}\big((x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\big) \mid y_1, y_2, \ldots, y_m \in \mathcal{Y}\right\}.$$

That is, we restrict attention to the set of possible translators that we could output for any given ground-truth translations, then it is not difficult to see $\log|\overline{\Theta}| \leq m \log|\mathcal{Y}|$. If we assume that one of these is accurate and natural, then restricting the attention of UNFISHY to this set will suffice, and Theorem 3.3 means that number of examples required is $O(\log|\overline{\Theta}|) = O(m \log|\mathcal{Y}|)$ which is a linear blowup in the number of examples m used for supervised translation. To make this formal, one would want to only assume that one of the translators had error close to zero and was nearly η natural–this is left to future work.

E Good priors through prompts

The most direct way to improve a prior is to train (or fine-tune) the language model on a dataset that includes a large number of articles relevant to the source language, e.g., volumes of oceanic research, for whales.

Another general strategy for creating a prior is to use prompting, a popular technique for dealing with large language models such as GPT-3. In particular, given a *prompt* string s, one can define $\rho(y) \propto \nu(s\,y)$, that is the prior is the distribution of text that that language model generates after (conditional on) the text beginning with s. Figure 3 illustrates some toy examples of how even a simple prompt like A sperm whale said: can help focus on translations that are more likely for a sperm whale to say, and eliminate irrelevant translations.

Background prompts. More interestingly, the problem of AI-based prompt generation has recently attracted attention [e.g., Shin et al., 2020]. There is natural and potentially powerful idea that an unsupervised translator, in addition to outputting translations, would automatically generating a *background prompt* that increases the intelligibility of many translations. Suppose, for example, across numerous communications, the unsupervised translator determines that sperm whales measure time in "number of naps" and that typical nap duration varies by age. Then, rather than having to repeatedly explain this in each translation, it can be explained once in a background prompt that is automatically inserted before each translation. For example, following the prompt

$s =$ Sperm whales measure time in numbers of naps, but each whale's typical nap duration depends on their age. So if a whale's nap duration is 9 minutes, then 4 naps is about 36 minutes (though whales tend to exaggerate times). A sperm whale said:

would make translations like Wow, I just dove for 6 naps or How many naps ago was that? both more likely and more understandable.

F Lossy translation

Many things can be included in textual transcriptions suitable for translation. For instance, one can distinguish the speakers, e.g., "Whale 1: ... Whale 2: ... Whale 1: ..." if the source data is annotated with speaker identifiers. Some aspects of the way one is speaking can be transcribed, e.g., "Whale 1 (speaking quickly): ... Whale 2 (voice disguised): ..." It may be possible to encode these textually in x.

However, if x is encoded in a more flexible format, such as arbitrary binary files, one can include as much raw information as possible, including the source audio recording, to provide the translator with as much context as possible. In that case, lossless translation will no longer be possible, because one cannot compute the raw x from a textual translation.

Given the possible benefits of such annotations, we propose an extension of our theory to the lossy setting. A natural generalization of the maximum-likelihood approach is as follows:

$$\min_{\theta \in \Theta} \frac{1}{m} \sum_{i=1}^{m} -\log \rho(f_\theta(x_i)) - \frac{1}{\lambda} \log \phi_\theta(x_i \mid y = f_\theta(x_i)).$$

Here $\phi_\theta \colon \mathcal{X} \times \mathcal{Y} \to [0,1]$ is a probabilistic inverse ("randomized backtranslation") of f whose parameters are also encoded in θ. Note that the family $\{(f_\theta, \phi_\theta) \mid \theta \in \Theta\}$ must satisfy that for all $y \in \mathcal{Y}$, $\sum_{x \in f^{-1}(y)} \phi_\theta(x \mid y) = 1$, though it is no longer required that f_θ be 1–1. Note that, for simplicity, we ignored the plausibility thresholding ($\min(\cdot, K)$) used by Algorithm C.1.

As λ decreases to 0, the optimal solution would assign infinite loss to any $f_\theta \colon \mathcal{X} \hookrightarrow \mathcal{Y}$ and ϕ_θ that are not perfect inverses, where there is some x (with positive probability under μ) such that $\phi_\theta(x \mid y = f_\theta(x)) < 1$. Thus, for sufficiently small λ, the algorithm is exactly the minimum cross-entropy (maximum likelihood) algorithm discussed in the introduction. For arbitrarily large λ, the algorithm will collapse to always outputting the most likely $y \in \mathcal{Y}$ under ρ. For example, everything could be translated to Hello regardless of its contents.

For intermediate values of λ, the chosen translator trades off naturalness in the form of $\rho(f_\theta(x))$ versus information loss which is inversely related to $\phi_\theta(x_i \mid y = f_\theta(x_i))$. This trade-off makes it challenging to define and analyze the success of a lossy unsupervised translation algorithm. Nonetheless, the algorithm is intuitive.

G RT: A tree-based probabilistic model of language

In this section we introduce a stylized model of natural language. In this model, a random tree G is generated by labeling the nodes of a tree with random words, with in the G corresponding to plausible sentences (according to the prior ρ). A translated language τ (or equivalently, a source language μ) is then derived from sentences corresponding to paths in a random sub-tree $H \subseteq G$. We then prove that, with high probability, a prior ρ and translated language τ sampled by this process satisfy Conditions 3.1 and 3.2, and therefore that they can allow for translation using Algorithm C.1. We conclude by discussing instantiations of this model with various parameters.

G.1 Random sub-tree languages

The *random sub-tree language* (RT) model is a randomized process for generating a tree-based source language μ, translated language τ, and prior ρ. It is parameterized by a finite vocabulary set $\mathcal{W}$, depth $n \in \mathbb{N}$, and arities $a, b \in \mathbb{N}$ such that $1 \leq a \leq b \leq |\mathcal{W}|/4$. For simplicity, the possible target sentences are taken to be the set of all possible n-grams over $\mathcal{W}$, namely, $\mathcal{Y} \coloneqq \mathcal{W}^n$. Again for simplicity, we also assume $|\mathcal{X}| = |\mathcal{Y}|$ so that each $f_\theta \colon \mathcal{X} \hookrightarrow \mathcal{Y}$ is a bijection. To generate an RT, first full b-ary, depth $n+1$ tree G is constructed. Labeled edges shall correspond to words, and paths shall correspond to sentences, as follows:

1. Starting from the root node and proceeding in level-order traversal, for each node v: sample b words from $\{w_1, \ldots, w_b\} \in \mathcal{W}$ uniformly at random and *without* replacement; label each of v's child edges with one of the sampled words.
2. The labels on each path from the root to a leaf $(y_1, \ldots, y_n)$ corresponds to a *plausible sentence*, giving a set of *plausible sentences*
$$P \coloneqq \{y \mid y = (y_1, \ldots, y_n) \text{ labels of a path in } G\}.$$
3. A sub-tree $H \subseteq G$ is obtained by sampling uniformly at random a out of b of the children at each level of the tree, in level-order traversal. The set of *translated sentences* is analogously defined as
$$T \coloneqq \{y \mid y = (y_1, \ldots, y_n) \text{ labels of a path in } H\} \subseteq P.$$

The prior ρ distributes uniformly at random over plausible sentences P. The translated language τ is uniform over T. We let $(\rho, \tau) \sim \mathrm{RT}(\mathcal{W}, n, a, b)$ denote the sampling of a prior and translated language obtained via this randomized process. When the parameters $\mathcal{W}, n, a, b$ are clear from context we simply write RT.

Finally, a source language distribution μ (over $\mathcal{X}$) is obtained by taking any $g \colon \mathcal{Y} \to \mathcal{X}$ and letting μ distribute uniformly over $g(T)$. This is the same as sampling $y \sim \mu$ by first sampling $y \sim \tau$, and then taking $x \coloneqq g(y)$. Crucially, g is not known to the unsupervised translation algorithm.

G.2 Translatability of a random sub-tree language

Next, we will argue that a random sub-tree language sampled in this process satisfies Conditions 3.1 and 3.2 with appropriate choice of parameters. While we make no assumptions on Π, it is important that the plausible sentences P are sampled *after* the possible permutations Π are chosen. Otherwise, if an adversary were able to choose a permutation $\pi \colon \mathcal{Y} \to \mathcal{Y}$ based on P, then they could arbitrarily permute P, resulting in a permutation π with high expected loss but no change to likelihood according to the prior ρ—thereby violating Cond. 3.2.

Theorem G.1 (Thm. G.3, simplified)**.** *Fix any $\epsilon, \delta \in (0,1)$, vocabulary set $\mathcal{W}$, and tree arities $a \leq b \leq |\mathcal{W}|/4$. Then for tree depth $n = O(\log_a(1/\epsilon), \log_a \log_2(1/\delta))$ and any parameter family Θ such that $\log_a \log_2 |\Theta| = O(n)$, any $(\rho, \mu) \sim \mathrm{RT}(\mathcal{W}, n, a, b)$ satisfy Cond. 3.1 for $\eta = n \log_2 b$ with certainty. Additionally, with probability at least $1 - \delta$, $(\rho, \tau) \sim \mathrm{RT}(\mathcal{W}, n, a, b)$ satisfy Cond. 3.2 for the 0-1 loss $\ell = \ell_{01}$, the given ϵ, $\gamma = O(\epsilon)$, and $K = O(n \log_2 b)$.*

You can think of Thm. G.1 as pointing to the required sentence length n and an upper bound on the number of parameters $|\Theta|$ so that Conditions 3.1 and 3.2 hold with high probability. As we we know from Thm. 3.3, and state clearly next, these in turn imply translatability of a random sub-tree language. Before we prove the theorem (in Appendix G.3), let us discuss its implications on random sub-tree languages.

Corollary G.2. *Fix any $\epsilon \in (0,1)$, vocabulary set $\mathcal{W}$, and tree arities $a \leq b \leq |\mathcal{W}|/4$. Then for tree depth $n = O(\log_a(1/\epsilon))$ and any parameter family Θ such that $\log_a \log_2 |\Theta| = O(n)$, with probability at least 99%, for $(\rho, \mu) \sim \mathrm{RT}(\mathcal{W}, n, a, b)$,* UNFISHY *run on prior ρ, threshold $K = O(n \log_2 b/\epsilon)$, and $m = O(\log_2 |\Theta|/\epsilon^2)$ iid samples from μ, outputs a translator $f_{\hat{\theta}}$ with $\mathrm{err}(f_{\hat{\theta}}) < \epsilon$.*

Note that the probability is over a prior ρ and source language μ sampled from RT*, as well as the randomness used by* UNFISHY.

Corollary G.2 is obtained by applying Thm. 3.3 to Thm. G.1, with $\delta = 1/200$ (for both theorems). To better understand Corollary G.2, let us consider two possible families of translators and suppose we are trying to apply Corollary G.2 to get bounds on the sample complexity m needed to translate a source language $\mu \sim \mathrm{RT}(\mathcal{W}, n, a, b)$ with constant error ϵ.

The first family is $\Theta_{\max}$, a huge family of translators; we take the maximal size for which Corollary G.2 holds, that is, we take $|\Theta_{\max}| = 2^{a^n}$. Then Corollary G.2 requires $m = O(a^n) = O(|T|)$ samples, which is proportional to the size of $\mathrm{supp}(\mu)$—that is, the number of possible sentences in the source language! So, translation requires an extremely large amount of samples for a minimally-restricted family of parameters $\Theta_{\max}$. In other words, *no free lunch*.

On the other hand, let us consider the family of *literal translators* Θ_{lit}. A literal translation, also known as a word-for-word translation, works by translating each word in a sentence separately, ignoring the surrounding context. The number of literal translations is the same as the number of word-to-word permutations, i.e., $|\Theta_{\text{lit}}| = |\mathcal{W}|!$. So Corollary G.2 gives a sample complexity bound of $m = O(\log \mathcal{W}!) = O(|\mathcal{W}| \log |\mathcal{W}|)$. The number of words in the English language is about $88,500$ [Nagy and Anderson, 1984], giving a sample complexity on the order of one million sentences. Each sentence should be of length on the order of $n \approx \log_a \log_2 |\Theta| \approx 10$.[5]

Literal translators are notoriously bad.[6] But recall our main motivation of translating whales to English, and note that even a literal Whalish–English translator would mark an astounding breakthrough. Indeed, Corollary G.2 describes the conditions for when unsupervised learning can find the optimal literal translator (barring computational constraints).

G.3 Proof of Thm. G.1

Thm. G.1 is an instantiation of the following theorem.

Theorem G.3. *Consider the 0-1 loss $\ell = \ell_{01}$ and any vocabulary set $\mathcal{W}$, tree arities $a \le b \le |\mathcal{W}|/4$, tree depth $n \in \mathbb{N}$. Then any ρ, τ sampled from $\mathrm{RT}(\mathcal{W}, n, a, b)$ satisfy Cond. 3.1 for $\eta = n \log b$, i.e.,*

$$\Pr_{\rho,\tau\sim\mathrm{RT}}\left[H(\tau,\rho) \le n \log b\right] = 1.$$

Additionally, fix $\epsilon, \delta \in (0,1)$, such that $n \ge \log_a \frac{\log(4|\Pi|/\delta)}{64\epsilon}$. Then ρ and τ satisfy Cond. 3.2 for ϵ, $\gamma = \epsilon/16$ and any $\mathrm{K} \ge n \log b$ with probability $\ge 1-\delta$, i.e.,

$$\Pr_{(\rho,\tau)\sim\mathrm{RT}}\left[\forall \pi \in \Pi \quad \mathop{\mathbb{E}}_{y\sim\tau}[\ell_{01}(y,\pi(y))] \ge \epsilon \implies \Pr_{y\sim\tau}[\rho(\pi(y)) < 2^{-K}] > \frac{\epsilon}{16}\right] \ge 1-\delta. \tag{1}$$

To prove Thm. G.3 we will two lemmas. The first is a known variant of the Chernoff bound for so-called 0-negatively correlated random variables.

Lemma G.4 (Chernoff bound for 0-negatively correlated random variables)**.** *Let $Z_1, \ldots, Z_n$ be* 0-negatively correlated *Boolean random variables, that is, they satisfy*

$$\forall I \subseteq [n] \quad \Pr\left[\forall i \in I \; Z_i = 0\right] \le \prod_{i\in I} \Pr\left[Z_i = 0\right].$$

Then, letting $Z \coloneqq \sum_{i=1}^n Z_i$ it holds that

$$\Pr\left[Z \le \frac{\mathbb{E}[Z]}{2}\right] \le e^{-\frac{\mathbb{E}[Z]}{8}}$$

Lemma G.4 follows from [Doerr, 2019, Theorem 1.10.24(a)] with $\delta \coloneqq 1/2$, $a_i \coloneqq 0$, $b_i \coloneqq 1$ and $Y_i \coloneqq X_i$. That theorem, in turn, is simply Theorem 1.10.10 for 0-negatively correlated random variables.

The second lemma is a combinatorial argument that we prove below.

[5] This is equivalent to about 20 copies of *Infinite Jest* by David Foster Wallace.

[6] Translating literally from Hebrew to English, potato is land apple, and delusional is living in a movie.

Lemma G.5. *Given any $\pi : \mathcal{Y} \hookrightarrow \mathcal{Y}$, it is possible to partition $A = \{y \in \mathcal{Y} \mid \pi(y) \neq y\}$ into four disjoint sets $A = A_1 \cup A_2 \cup A_3 \cup A_4$ such that, for each $i = 1, 2, 3, 4$, the following two conditions hold:*

$$\pi(A_i) \cap A_i = \emptyset$$

$$\forall z \in \mathcal{W}^{n-1} \; |\{y \in A_i \mid y \textit{ begins with } z\}| \leq \frac{|\mathcal{W}|}{2}$$

Proof of Lemma G.5. We will partition A greedily to achieve the two conditions. Begin with four empty sets $A_1 = A_2 = A_3 = A_4 = \emptyset$. For each $y \in A$ in turn, assign it to one of the 4 sets as follows.

1. Let i be the index of the set A_i such that $\pi(y) \in A_i$ has already been assigned to. If $\pi(y))$ has not yet been assigned, let $i = 1$.
2. Similarly, if $\pi(y) \in A_j$ has been assigned, let j be its index otherwise $j = 1$.
3. Let z be the first $n-1$ elements of x. Let k be the index of the set A_k such that $|\{y \in A_i \mid y \text{ begins with } z\}| \geq |\mathcal{W}|/2$ if there is such a set, otherwise $k = 1$. Note that there can be at most one such k because there are at most $|\mathcal{W}| - 1$ other elements beginning with z that have been assigned already.

Thus $S = \{1, 2, 3, 4\} \setminus \{i, j, k\} \neq \emptyset$ and we can assign x to any (say the minimum) element in that set. By induction, we preserve the two properties stated in the lemma.

□

Next, we prove Thm. G.3.

Proof. For Cond. 3.1, $H(\tau, \rho) = n \log b$ follows immediately from the fact that ρ is uniform over P of size $|P| = b^n$ and that τ is supported on $T \subseteq P$.

We now analyze the probability that Cond. 3.2 holds. Equivalently to Eq. (1), we let $\delta = 4|\Pi|e^{-\epsilon a^n/64}$ and show that

$$\Pr_{(\rho,\tau)\sim\mathrm{RT}}\left[\exists \pi \in \Pi \quad \mathop{\mathbb{E}}_{y\sim\tau}[\ell_{01}(y, \pi(y))] \geq \epsilon, \; \Pr_{y\sim\tau}[\rho(\pi(y)) < 2^{-K}] < \frac{\epsilon}{16}\right] \leq \delta.$$

Note that by definition of ℓ_{01}, $\mathbb{E}[\ell_{01}(y, \pi(y))] = \Pr[y \neq \pi(y)]$. Additionally, for any $y \in \mathcal{Y}$, $\rho(\pi(y)) < 2^{-\mathrm{K}}$ occurs if and only if $\pi(y) \notin P$, because $\rho(\pi(y)) = n^{-b} \geq 2^{-K}$ for any $\pi(y) \in \mathcal{P}$, and 0 elsewhere. Therefore, by the union bound over $\pi \in \Pi$, it suffices to show that, for each $\pi \in \Pi$,

$$\Pr_{(\rho,\tau)\sim\mathrm{RT}}\left[\Pr_{y\sim\tau}[y \neq \pi(y)] \geq \epsilon, \; \Pr_{y\sim\tau}[\pi(y) \notin P] < \frac{\epsilon}{16}\right] < \frac{\delta}{|\Pi|}. \tag{2}$$

Fix any $\pi : \mathcal{Y} \hookrightarrow \mathcal{Y}$. Eq. (2) can be restated in terms of *ambiguous sentences* $A \coloneqq \{y \in \mathcal{Y} \mid \pi(y) \neq y\}$ and *implausible ambiguities* $B \coloneqq \{y \in \mathcal{Y} \mid \pi(y) \notin P\}$:

$$\Pr_{(\rho,\tau)\sim\mathrm{RT}}\left[\tau(A) \geq \epsilon, \; \tau(B) < \frac{\epsilon}{16}\right] < \frac{\delta}{|\Pi|}.$$

Using Lemma G.5, partition A into four disjoint sets $A = A_1 \sqcup A_2 \sqcup A_3 \sqcup A_4$ such that for each $i \in [4]$, the following two conditions hold:

$$\pi(A_i) \cap A_i = \emptyset \tag{3}$$

$$\forall z \in \mathcal{W}^{n-1} \quad |\{y \in A_i \mid y \text{ begins with } z\}| \leq \frac{|\mathcal{W}|}{2}. \tag{4}$$

By the union bound, it suffices to show that,

$$\forall i \in [4] \quad \Pr_{(\rho,\tau)\sim\mathrm{RT}}\left[\tau(A_i) \geq \frac{\epsilon}{4}, \; \tau(B) < \frac{\epsilon}{16}\right] < \frac{\delta}{4|\Pi|},$$

because $\tau(A) = \sum_{i=1}^{4} \tau(A_i)$, so if $\tau(A) \geq \epsilon$ then it must follow that $\tau(A_i) \geq \epsilon/4$ for some i. It suffices to show that this holds conditioned on any value of $A_i \cap T$:

$$\forall V \subseteq A_i \quad \Pr_{(\rho,\tau)\sim \mathrm{RT}} \left[\tau(A_i) > \frac{\epsilon}{4}, \tau(B) < \frac{\epsilon}{16} \mid A_i \cap T = V\right] < \frac{\delta}{4|\Pi|}. \tag{5}$$

Thus, fix any $V \subseteq A_i$. We proceed by analyzing two cases, based on $|V|$ versus $|T| = a^n$.[7]

Case 1: $|V| < \frac{\epsilon}{4}|T|$. Then Eq. (5) holds with probability 0, because conditioning on $A_i \cap T = V$, we have

$$\tau(A_i) \coloneqq \frac{|A_i \cap T|}{|T|} = \frac{|V|}{|T|} < \frac{\epsilon}{4}.$$

Case 2: $|V| \geq \frac{\epsilon}{4}|T|$. For each $v \in V$ we define a Boolean random variable $Z_v = \mathbf{1}_{v \in B}$, i.e.,

$$Z_v = \begin{cases} 1 & v \in B \\ 0 & v \notin B. \end{cases}$$

We claim that the Z_v's satisfy the condition of Lemma G.4, which follows inductively from the fact any subset of Z_v's being 0 only makes it *less* likely for another Z_v to be 0, because the way that the Z_v's are chosen is such that there are exactly b 1's among the Z_v's (and other symmetric random variables which we have not named).

Therefore, if we define $Z \coloneqq \sum_{v \in V} Z_v$ and $\zeta \coloneqq \mathbb{E}[Z \mid A_i \cap T = V]$, Lemma G.4 gives,

$$\Pr\left[Z < \frac{\zeta}{2} \,\middle|\, A_i \cap T = V\right] \leq e^{-\zeta/8}. \tag{6}$$

We conclude by bounding both sides of Eq. (6) to obtain Eq. (5).

- For the right hand side of eq. (6), note $\Pr[Z_v = 1 \mid A_i \cap T = V] \geq 1/2$ for each $v \in V$, we will use our two conditions. First, fix v and let

  $$Q = \{y \in \mathcal{Y} \mid y \text{ and } \pi(v) \text{ agree on the first } n-1 \text{ words}\}.$$

  Eq. (3) implies that $\pi(v) \notin A_i$, and eq. (4) implies that $|Q \setminus A_i| \geq |\mathcal{W}|/2$. Thus there are $\geq |\mathcal{W}|/2$ other elements in $Q \setminus A_i$ that are as likely to be in P as $\pi(v)$, and P contains exactly $b \leq |\Sigma|/4$ elements, thus the probability that $\pi(v)$ is in P is at most:

  $$\frac{|\mathcal{W}|/4}{|\mathcal{W}|/2} \leq \frac{1}{2}.$$

  Since $\pi(v) \notin P$ is equivalent to $Z_v = 1$, this means that $\Pr[Z_v = 1 \mid A_i \cap T = V] \geq 1/2$, and hence:

  $$\zeta = |V| \cdot \mathbb{E}[Z_v \mid A_i \cap T = V] = |V| \cdot \Pr[Z_v = 1 \mid A_i \cap T = V] \geq \frac{|V|}{2}.$$

  Because we are in the case that $|V| \geq \frac{\epsilon}{4}|T|$, we have

  $$\zeta \geq \frac{|V|}{2} \geq \frac{1}{2} \cdot \frac{\epsilon}{4} \cdot |T| = \frac{\epsilon}{8} \cdot a^n. \tag{7}$$

  Therefore the right hand side of eq. (6) satisfies

  $$e^{-\zeta/8} \leq e^{-\epsilon a^n/64} = \frac{\delta}{4\,|\Pi|},$$

  which is indeed the right hand side of eq. (5). The equality above holds by our assumption that $n \geq \log_a \frac{\log(4|\Pi|/\delta)}{64\epsilon}$.

[7] Note that while T is randomly sampled, $|T| = a^n$ always, and therefore the cases analysis is valid.

- For the left hand side of eq. (6), note that

$$Z \coloneqq \sum_{v \in V} Z_v \coloneqq \sum_{v \in V} \mathbf{1}_{v \in B} = |V \cap B|.$$

and therefore, using eq. (7), the left hand side of eq. (6) satisfies

$$\Pr\left[Z < \frac{\zeta}{2} \;\middle|\; A_i \cap T = V\right] \geq \Pr\left[\frac{|V \cap B|}{|T|} \leq \frac{\epsilon}{16} \;\middle|\; A_i \cap T = V\right].$$

We claim that $V \cap B \subseteq B \cap T$: Due to the conditional event, we have $V \cap B = A_i \cap T \cap B \subseteq A \cap T \cap B$. However, we argue that $A \cap T \cap B \subseteq T \cap B$, i.e., that $T \cap B \subseteq A$. Indeed, by definition, if $y \in T \cap B$ then $y \in T$ but $\pi(y) \notin P$, and since $P \supseteq T$, this implies that $\pi(y) \neq y$; that is, that $y \in A$, as needed.

We have just shown that $V \cap B \subseteq B \cap T$, and therefore

$$\begin{aligned}\Pr\left[\frac{|V \cap B|}{|T|} \leq \frac{\epsilon}{16} \;\middle|\; A_i \cap T = V\right] &\geq \Pr\left[\frac{|B \cap T|}{|T|} \leq \frac{\epsilon}{16} \;\middle|\; A_i \cap T = V\right] \\ &= \Pr\left[\tau(B) \leq \frac{\epsilon}{16} \;\middle|\; A_i \cap T = V\right].\end{aligned}$$

This shows that the left hand side of eq. (6) upper bounds that of eq. (5), and concludes the proof.

□