Notes: "A Multilingual View of Unsupervised Machine Translation"

Sun

March 30, 2020

1 Notes:

1.1 Summary:

1. By using a combination of black magic and clever optimization techniques, the authors proposed a system for unsupervised (no matching sentence pairs) translation between 2 or more languages.
2. Heavy bootstrapping combined with a clever pretraining gives the model the ability to learn entirely without labels
3. State of the art in certain datasets

1.2 High level:

One of the major issues with supervised machine translation is the abundance of data needed, all carefully labeled and matched. Most traditional methods seem to use matching sentence pairs: your data consists of sentences in different languages that have the same content. This is resource intensive, however, and even corpuses with that may have high overlap in terms of meaning (such as translated Wikipedia articles) do not fulfill this requirement.

The problem with unsupervised MT is that there is no clear way to do so. One could make arguments for pattern matching or similar, but the overall problem seems intractable from certain angles. Even beyond this, there is no "correct" objective. This paper addresses these issues largely by one clever argument: if a model translates from language $\mathcal{A}$ to language $\mathcal{B}$, then translates that same sentence from $\mathcal{B}$ to $\mathcal{A}$, no matter what, the sentence should not be fundamentally different. While this objective (back-translation) is commonly used, the derivation of this objective is novel and very morally correct. In addition, the paper presents a novel way to use data not directly associated with the end goal (such as optimizing English-French-Romanian translation to get better English-Romanian translations).

By applying this to the original problem, you can first pretrain the model on monolingual data (to get the model to be better than randomly initialized, but still bad), then take the models predictions as true (by translating from $\mathcal{A}$ to $\mathcal{B}$ and assuming that translation is accurate), the resulting model should then predict that the most likely translation back from $\mathcal{B}$ to $\mathcal{A}$ should be the original sentence.

By doing this "bootstrapped" training, it is possible for the model to produce legitimately good translations.

1.3 Low level:

1.3.1 EM Algorithm:

This paper heavily uses an algorithm from the late 70's called "Expectation-maximization" (EM). It is an iterative algorithm that has the ability to estimate unobserved (latent) values, $\mathbf{Z}$, while also optimizing a model.

Given a distribution, $\mathbf{X}$, with an underlying latent variable, $\mathbf{Z}$, and a vector of unknown model parameters, θ, with a likelihood function $L(\theta; \mathbf{X}, \mathbf{Z}) := p(\mathbf{X}, \mathbf{Z}|\theta)$, the maximum likelihood estimate (MLE) of the unknown parameters is achieved by maximizing the marginal likelihood of the observed data, e.g.:

$$L(\theta, \mathbf{X}) \equiv p(\mathbf{X}|\theta) \equiv \int p(\mathbf{X}, \mathbf{Z}|\theta) d\mathbf{Z} \tag{1}$$

This integral is often intractable, e.g.: due to computational constraints. The EM algorithm cleverly sidesteps this, breaking it into steps.

1. First, initialize θ randomly.
2. Estimate the conditional distribution of Z given X using θ, then compute the expectation with respect to $\mathbf{Z}$ given $\mathbf{X}, \theta$.
3. Replace the integral in (1) with the expectation given in step 2, then optimize θ using gradient ascent.
4. Repeat 2 and 3 until converged.

This algorithm has a guarantee to find a local maximum of L. Put more specifically, there are 2 main steps.

E step: Define $\mathcal{Q}(\theta|\theta_\mathbf{t})$ as the expected value of the log likelihood of θ w.r.t. the conditional distribution of $\mathbf{Z}$ given $\mathbf{X}$ and the current estimates of θ.

$$\mathcal{Q}(\theta|\theta_\mathbf{t}) := \mathbb{E}_{p(Z|X,\theta)}\big[\log L(\theta; \mathbf{X}, \mathbf{Z})\big] \tag{2}$$

(2) corresponds to estimating $\mathbf{Z}$.

M step: Optimize θ^t w.r.t. this new $\mathbf{Z}$.

$$\theta^{t+1} = \text{argmax}_\theta \mathcal{Q}(\theta|\theta^t) \tag{3}$$

Despite not directly increasing $\log p(\mathbf{X}|\theta)$, optimizing $\mathcal{Q}(\theta|\theta_\mathbf{t})$ does so indirectly.

1.3.2 Unsupervised machine translation using EM:

First, assume that we have 3 sets of monolingual data, $\mathcal{D}_x$, $\mathcal{D}_y$and $\mathcal{D}_z$, for languages X, Y, and Z respectively. The authors take the viewpoint that these datasets form a larger dataset, $\mathcal{D}_{x,y,z}$, of pairs (x, y, z) which are direct translations of each other. These pairs can be thought of as samples from random variables (X, Y, Z), and the resulting log-likelihood of the observed data is

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathcal{D}_x} + \mathcal{L}_{\mathcal{D}_y} + \mathcal{L}_{\mathcal{D}_z}$$

The goal is to learn a conditional translation model p_θ. The authors rewrite the log-likelihood as a marginalization over unobserved variables:

$$\begin{aligned}\mathcal{L}(\theta) = &\sum_{x \in \mathcal{D}x} \log \mathbb{E}_{(y,z)\sim(Y,Z)} p_\theta(x|y,z) \\ &+ \sum_{y \in \mathcal{D}y} \log \mathbb{E}_{(x,z)\sim(X,Z)} p_\theta(y|x,z) \\ &+ \sum_{z \in \mathcal{D}z} \log \mathbb{E}_{(x,y)\sim(X,Y)} p_\theta(z|x,y)\end{aligned} \tag{4}$$

However, this does not match our end translation goal. We cannot translate $z \to x$ without access to y, or $y \to x$ without z. The authors make the following assumption: given any variable in the (X, Y, Z) triplet, the remaining two are independent, e.g.: the conditioned variable details the content, while the given variables are direct translations in the respective language. As such, given that $p_\theta(x|y, z) \equiv p_\theta(x|y) \equiv p_\theta(x|z)$, we can rewrite (4).

$$\begin{aligned}\mathcal{L}(\theta) = &\sum_{x \in \mathcal{D}x} \log \mathbb{E}_{(y,z)\sim(Y,Z)} \sqrt{p_\theta(x|z) p_\theta(x|y)} \\ &+ \sum_{y \in \mathcal{D}y} \log \mathbb{E}_{(x,z)\sim(X,Z)} \sqrt{p_\theta(y|x) p_\theta(y|z)} \\ &+ \sum_{z \in \mathcal{D}z} \log \mathbb{E}_{(x,y)\sim(X,Y)} \sqrt{p_\theta(z|x) p_\theta(z|y)}\end{aligned} \tag{5}$$

However, note that these expectations are still intractable due to the possible number of sequences in a given language. The authors then prepare to use EM. They start by using Jensen's inequality:

$$\begin{aligned}\log \mathbb{E}_{(y,z)\sim(Y,Z)} p_\theta(x|y,z) &= \log \mathbb{E}_{(y,z)\sim(Y,Z)} \frac{p_\theta(x|y,z)}{p_\theta(y,z|x)} p_\theta(y,z|x) \\ &= \log \mathbb{E}_{(y,z)\sim p_\theta(y,z|x)} \frac{p_\theta(x|y,z)}{p(y,z|x)} p_\theta(y,z) \\ &= \mathbb{E}_{(y,z)\sim p_\theta(y,z|x)} \big[\log p_\theta(x|y,z) + \log p(y,z)\big] \\ &\quad + H(p_\theta(y,z|x)\end{aligned} \tag{6}$$

Where H is the entropy of a random variable. Since H is always non-negative, we can then bound this quantity from below.

$$\begin{aligned}\mathbb{E}_{(y,z)\sim(Y,Z)}p_\theta(x|y,z) &\geq \mathbb{E}_{(y,z)\sim p_\theta(y,z|x)}\big[\log p_\theta(x|y,z) + \log p(y,z)\big] \\ &= \frac{1}{2}\mathbb{E}_{y\sim p_\theta(y|x)}\log p_\theta(x|y) + \frac{1}{2}\mathbb{E}_{z\sim p_\theta(z|x)}\log p_\theta(x|z) \\ &\quad + \mathbb{E}_{(y,z)\sim p_\theta(y,z|x)}\log p(y,z)\end{aligned} \tag{7}$$

Which morally provides us with the back-translation terms. Applying this to (5) results in:

$$\begin{aligned}\mathcal{L}(\theta) \geq\ & \frac{1}{2}\mathbb{E}_{y\sim p_\theta(y|x)}\log p_\theta(x|y) + \frac{1}{2}\mathbb{E}_{z\sim p_\theta(z|x)}\log p_\theta(x|z) \\ &+\frac{1}{2}\mathbb{E}_{x\sim p_\theta(x|y)}\log p_\theta(y|x) + \frac{1}{2}\mathbb{E}_{z\sim p_\theta(z|y)}\log p_\theta(y|z) \\ &+\frac{1}{2}\mathbb{E}_{x\sim p_\theta(x|z)}\log p_\theta(z|x) + \frac{1}{2}\mathbb{E}_{y\sim p_\theta(y|z)}\log p_\theta(z|y) \\ &+\mathbb{E}_{(y,z)\sim p_\theta(y,z|x)}\log p(y,z) + \mathbb{E}_{(x,z)\sim p_\theta(x,z|y)}\log p(y,z) \\ &+\mathbb{E}_{(x,y)\sim p_\theta(x,y|z)}\log p(x,y)\end{aligned} \tag{8}$$

The back translation terms, e.g.: $\mathbb{E}_{y\sim p_\theta(y|x)}\log p_\theta(x|y)$ enforce accuracy of reciprocal translations. The joint terms, e.g.: $\mathbb{E}_{(y,z)\sim p_\theta(y,z|x)}\log p(y,z)$, vanish with the use of the EM algorithm.

To optimize (8) using the EM algorithm, there are two steps.

E step: at iteration t, we compute the expectations against the conditional distributions evaluated at the current set of parameters $\theta = \theta^t$. We approximate the expectation by using the mode instead, e.g.: $\mathbb{E}_{y\sim p_\theta(y|x)}\log p_\theta(x|y) \approx p_\theta(x|\hat{y})$, with $\hat{y} = \text{argmax}_y p_\theta(y|x)$. Put clearly, the expected value of x given y is approximately equal to x given the most likely y given x. In practice, the argmax is approximated using an iterative greedy decoding procedure (pick most likely first word, most likely second given first, most likely third given first and second, etc..).

M step: we choose the θ which maximizes the resulting terms after the **E** step. Considering the last three terms in (8) no longer depend on θ, they can be ignored. We optimize θ by preforming a single gradient update.

1.3.3 Auxiliary parallel data:

Another bit of novelty this paper presents is the ability to use parallel data. Assume you wish to translate from language X to Z, while also having a parallel corpus $\mathcal{D}_{x,y}$ that maps sentences from X to Y. We can then augment our log-likelihood as follows:

$$\mathcal{L}_{aug}(\theta) = \mathcal{L}(\theta) + \sum_{(x,y)\in\mathcal{D}x,y}\log\mathbb{E}_{z\sim Z}p_\theta(x,y|z) \tag{9}$$

Once again, by using the EM algorithm we can obtain and objective that can be optimized efficiently. The end result has a new kind of expression called $cross-translation$, e.g.: $\mathbb{E}_{z\sim p_\theta(z|y)}\log p_\theta(x|z)$. Intuitively, these terms ensure the model can translate from Y to Z, then Z to X.

1.3.4 Pretraining:

As mentioned previously, the model is pretrained as to make the starting embeddings useful. The MASS objective is used, which consists of masking randomly chosen segments of the text. The objective is to recreate the masked portion of the sentence.

1.4 Empirical results:

This model achieved state of the art in multiple tasks. The cross-translation terms improved performance in almost all cases where it was used. Pretraining alone gave competitive results in a couple sections.