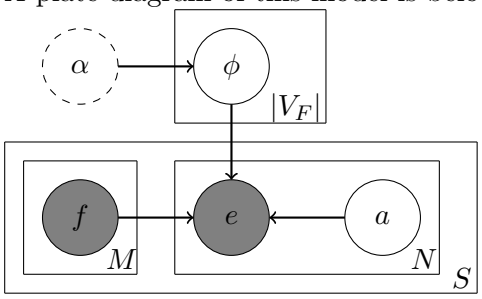
Variational IBM Model 1

August 31, 2016

We want to derive a variational approximation to IBM Model 1 given some bitext (F, E) consisting of S sentence pairs of French and English text. We will notate that $|F_s| = M_s$ and $|E_s| = N_s$. (Note: the fast_align paper uses the opposite defitions of M and N.) The model selects for each target word $e_{s,n}$ an alignment link $a_{s,n}$ in $[1, M_s]$ which indicates that the target word e is generated as a translation of $F_{s,a}$.

This model has two latent variables. The first is the set of alignments A, which we will assume have uniform prior over the valid range for each sentence. The second is Φ, the *translation table*, which gives the translation probability $p(e|f)$ that a source word f generates a target word e. We will assume that each ϕ_f is a $|V_E|$-dimensional Dirichlet with parameters α_{f_e}, all of which are initially set to α_0, a hyperparameter.

A plate diagram of this model is below.



Mathematically the model is defined as follows. We assume that F and E (and thus their lengths M and N) are observed, and that α is given as a hyperparameter.

$$\begin{aligned}
\phi_f &\sim \text{Dir}(|V_E|, \alpha_f) \text{ for } f \in V_F \\
a_{s,n} &\sim \text{Uniform}(1, M_s) \\
e_{s,n} &\sim \text{Categorical}(\phi_{F_{s,a_{s,n}}})
\end{aligned}$$

Recall that $\text{Dir}(\phi|K, \alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha)} \prod_i \phi_i^{\alpha_i - 1}$.

Thus we can write down the joint probability of the model as $p(F, E, \Phi, A) = p(\Phi|\alpha) \cdot p(A|F) \cdot p(E|F, A, \Phi)$. We seek the variational approximation $q(\Phi, A) \approx p(F, E, \Phi, A)$ which we will assume factors as $q(\Phi, A) = q_\Phi(\Phi) \cdot q_A(A)$.

Let's examine the log joint probablity of our model:

$$
\begin{aligned}
\ln p(F, E, \Phi, A) &= \ln p(\Phi|\alpha) + \ln p(A|F) + \ln p(E|F, A, \Phi) \\
&= \sum_f \big[\ln \Gamma(\sum_e \alpha_{f_e}) - \sum_e \ln \Gamma(\alpha_{f_e}) + \sum_e (\alpha_{f_e} - 1) \ln \phi_{f_e} \big] \\
&+ \big[\sum_s \sum_{n=1}^{N_s} - \ln M_s \big] \\
&+ \big[\sum_s \sum_{n=1}^{N_s} \ln \phi_{F_{s,a_{s,n}} e_{s,n}} \big]
\end{aligned}
$$

So if we want $\ln q_A(A) = \mathbb{E}_\Phi \big[\ln p(F, E, \Phi, A) \big]$ we have:

$$
\begin{aligned}
\ln q_A(A) &= \mathbb{E}_\Phi \Bigg[\sum_f \big[\ln \Gamma(\sum_e \alpha_{f_e}) - \sum_e \ln \Gamma(\alpha_{f_e}) + \sum_e (\alpha_{f_e} - 1) \ln \phi_{f_e} \big] \\
&+ \big[\sum_s \sum_{n=1}^{N_s} - \ln M_s \big] \\
&+ \big[\sum_s \sum_{n=1}^{N_s} \ln \phi_{F_{s,a_{s,n}} e_{s,n}} \big] \Bigg] \\
&= \mathbb{E}_\Phi \Bigg[\sum_s \sum_{n=1}^{N_s} \ln \phi_{F_{s,a_{s,n}} e_{s,n}} \Bigg] + C \\
&= \mathbb{E}_\Phi \Bigg[\sum_e \sum_f \text{Count}(e, f) \ln \phi_{f_e} \Bigg] + C \\
&= \sum_e \sum_f \text{Count}(e, f) \mathbb{E}_\Phi [\ln \phi_{f_e}] + C
\end{aligned}
$$

where $\text{Count}(e, f)$ is the number of times the word e is aligned to the word f in the corpus, according to A. According to the wiki page on the Dirichlet Distribution $\mathbb{E}_\Phi [\ln \phi_{f_e}] = \psi(\alpha_{f_e}) - \psi(\sum_e \alpha_{f_e})$ where ψ represents the digamma function.

$$
\begin{aligned}
\ln q_A(A) &= \sum_e \sum_f \text{Count}(e, f) \mathbb{E}_\Phi [\ln \phi_{f_e}] + C \\
&= \sum_e \sum_f \text{Count}(e, f) \big(\psi(\alpha_{f_e}) - \psi(\sum_e \alpha_{f_e}) \big) + C
\end{aligned}
$$

Now for $\ln q_\Phi(\Phi) = \mathbb{E}_\Phi\big[\ln p(F, E, \Phi, A)\big]$ we get:

$$\begin{aligned}
\ln q_\Phi(\Phi) &= \mathbb{E}_A\Bigg[\sum_f \big[\ln\Gamma(\sum_e \alpha_{f_e}) - \sum_e \ln\Gamma(\alpha_{f_e}) + \sum_e (\alpha_{f_e} - 1)\ln\phi_{f_e}\big] \\
&\quad + \big[\sum_s \sum_{n=1}^{N_s} - \ln M_s\big] \\
&\quad + \big[\sum_s \sum_{n=1}^{N_s} \ln\phi_{F_{s,a_{s,n}} e_{s,n}}\big]\Bigg] \\
&= \mathbb{E}_A\Bigg[\big[\sum_f \sum_e (\alpha_{f_e} - 1)\ln\phi_{f_e}\big] + \big[\sum_s \sum_{n=1}^{N_s} \ln\phi_{F_{s,a_{s,n}} e_{s,n}}\big]\Bigg] + C \\
&= \mathbb{E}_A\Bigg[\big[\sum_f \sum_e (\alpha_{f_e} - 1)\ln\phi_{f_e}\big] + \big[\sum_e \sum_f \mathrm{Count}(e, f)\ln\phi_{f_e}\big]\Bigg] + C \\
&= \sum_f \sum_e (\alpha_{f_e} - 1)\ln\phi_{f_e} + \mathbb{E}_A\Bigg[\sum_e \sum_f \mathrm{Count}(e, f)\ln\phi_{f_e}\Bigg] + C \\
&= \sum_f \sum_e (\alpha_{f_e} - 1)\ln\phi_{f_e} + \sum_e \sum_f \mathbb{E}_A\big[\mathrm{Count}(e, f)\big]\ln\phi_{f_e} + C \\
&= \sum_f \sum_e \Big(\alpha_{f_e} + \mathbb{E}_A\big[\mathrm{Count}(e, f)\big] - 1\Big)\ln\phi_{f_e} + C
\end{aligned}$$

From this it's clear that each ϕ_f is now a non-symmetric Dirichlet with parameters $\alpha_0 + \mathbb{E}_A[\mathrm{Count}(e, f)]$ for each $e \in V_e$.

This suggests the following simple algorithm for model 1:

1. Initialize $\alpha_{f,e}$ to α_0 $\forall f, e$
2. For each iteration:
3. Initialize $\alpha' = \alpha$
4. Loop over s and n. Let $e = E_{s,n}$. For each source word $f \in F_s$ add $\psi(\alpha_{f_e}) - \psi(\sum_{e' \in E_s} \alpha'_{f_e})$ to $\alpha'_{f,e}$.
5. Update $\alpha = \alpha'$
6. Repeat until convergence

See variational_ibm1.py for an implementation of this algorithm.