$$A_{f\rightarrow e} - A_{e\rightarrow f} - A_{e\rightarrow f}(x^e) = x^f$$

$$E = \{x_1^e, .., x_i^e, .., x_{N_e}^e\}$$

$$F = \{x_1^f, .., x_i^f, .., x_{N_f}^f\}$$

We define Nbr(x, L, d) as the neighborhood in language L of size d (on either side) around word x in its sentence.

Each word x^f in the foreign vocabulary is associated with a dense vector x^f in R^ m, and each word x^e in English vocabulary admits at most T sense vector s, with the kth sense vector denoted as x_ k^e.

$$P(z_x = k \mid \beta_x) = \beta_{xk} \prod_{r=1}^{k-1} (1 - \beta_{xr})$$

$$\beta_{xk} \mid \alpha \overset{ind}{\sim} Beta(\beta_{xk} \mid 1, \alpha), \quad k = 1, \ldots .$$

β are the parameters determining the model probability on each sense for x^e (i.e., the weight on each possible value for z)

y^e — y^f

The English and foreign neighboring words are denoted by y^e and y^f

θ are model parameters (i.e. all embeddings) and α governs the hyper-prior on latent senses

Assume x^e has multiple se nses, which are indexed by the random variable z

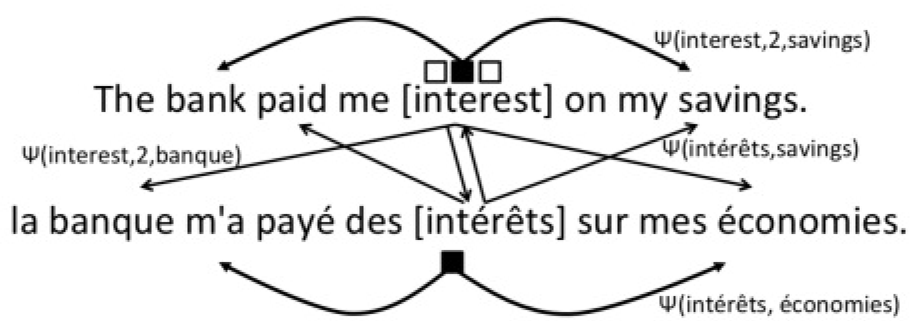
similar to the bilingual skip-gram

$$P(y^e, y^f \mid x^e, x^f; \alpha, \theta) \longrightarrow \int_\beta \sum\nolimits_z P(y^e, y^f z, \beta \mid x^e, x^f, \alpha; \theta) d\beta$$

$$P(y^e, y^f \mid z, x^e, x^f; \theta) =$$
$$P(y^e \mid x^e, x^f, z; \theta) P(y^f \mid x^e, x^f, z; \theta).$$

$$P(y \mid x^e, x^f, z = k; \theta) \propto \Psi(x^e, z = k, y) \Psi(x^f, y)$$
$$= \exp(\boldsymbol{y}^T \boldsymbol{x}_k^e) \exp(\boldsymbol{y}^T \boldsymbol{x}^f) = \exp(\boldsymbol{y}^T (\boldsymbol{x}_k^e + \boldsymbol{x}^f)),$$

$$P(y^e, y^f \mid z, x^e, x^f; \theta) \propto \Psi(x^e, z, y^e) \Psi(x^f, y^f)$$
$$\Psi(x^e, z, y^f) \Psi(x^f, y^e)$$



This modeling approach is reminiscent of (Luong et al., 201 5), who jointly learned embeddings for two languages l1 and l2 by optimizing a joint objective containing 4 skip-gr am terms using the aligned pair (x^e,x^f)– two predicting monolingual contexts l1 → l1, l2 → l2 , and two predicting crosslingual contexts l1 → l2, l2 → l1

learning
maximizing the log-likelihood

$$P(y^e, y^f \mid x^e, x^f; \alpha, \theta) =$$
$$\int_\beta \sum\nolimits_z P(y^e, y^f, z, \beta \mid x^e, x^f, \alpha; \theta) d\beta$$

for which we use variational approximation:
q(z, β) = q(z)q(β) = P(z,β | y^e,y^f,x^e,x^f,α)
q(z) = production of all q(z_i)
q(β) = production of all β_w^k

$$\theta \leftarrow \theta + \rho_t \nabla_\theta \sum_{k \mid z_{ik} > \epsilon} \sum_{y \in y_c} z_{ik} \log p(y \mid x_i, k, \theta) \tag{6}$$

Algorithm 1 Psuedocode of Learning Algorithm

Input: parallel corpus $E = \{x_1^e, .., x_i^e, .., x_{N_e}^e\}$ and $F = \{x_1^f, .., x_i^f, .., x_{N_f}^f\}$ and alignments $A_{e\rightarrow f}$ and $A_{f\rightarrow e}$, Hyper-parameters α and T, window sizes d, d' .
Output: θ, $q(\beta)$, $q(\mathbf{z})$

```
for i = 1 to N_e do    ▷ update english vectors
    w ← x_i^e
    for k = 1 to T do
        z_ik ← E_{q(β_w)}[log p(z_i = k|, x_i^e)]
    y_c ← Nbr(x_i^e,E,d) ∪ Nbr(x_i^f,F,d') ∪ {x_i^f}
    where x_i^f = A_{e→f}(x_i^e)
    for y in y_c do
        SENSE-UPDATE(x_i^e, y, z_i)
    Renormalize z_i using softmax
    Update suff. stats. for q(β) like (Bartunov et al., 2016)
    Update θ using eq. (6)
for i = 1 to N_f do    ▷ jointly update foreign vectors
    y_c ← Nbr(x_i^f,F,d) ∪ Nbr(x_i^e,E,d') ∪ {x_i^e}
    where x_i^e = A_{f→e}(x_i^f)
    for y in y_c do
        SKIP-GRAM-UPDATE(x_i^f, y)
procedure SENSE-UPDATE(x_i, y, z_i)
    z_ik ← z_ik + log p(y|x_i, k, θ)
```

Disambiguation:
Similar to (Bartunov et al., 2016), we can disambiguate the sense for the word x^e given a monolingual context y^e

$$P(z \mid x^e, y^e) \propto$$
$$P(y^e \mid x^e, z; \theta) \sum\nolimits_\beta P(z \mid x^e, \beta) q(\beta)$$