Assignment 1:
Word-alignment IBM Model1 using Gibbs sampling

1 Task definition

Download the English-Czech sentence-aligned corpus `english-czech.tsv` from the course web pages. Your task is to infer a word-alignment, where each English word is aligned with just one Czech word. A Czech word can be aligned with zero, one, or more English words.

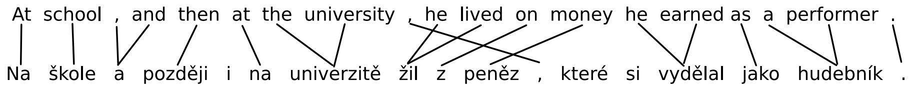
At school , and then at the university , he lived on money he earned as a performer .
Na škole a později i na univerzitě žil z peněz , které si vydělal jako hudebník .

Figure 1: English-to-Czech asymetric word-alignment

Implement the IBM Model1, which models a probability of an aligned English word e_i conditioned by a Czech word c_j. Assume a categorical distribution

$$p(e_i|c_j) \sim \text{Categorical}(\theta^{(c)})$$

For a model with $|C|$ possible Czech words, each of the translation distributions $\theta^{(c)}$ has $|E|$ components. Assume a symmetric Dirichlet prior for the distributions $\theta^{(c)}$.

$$\theta^{(c)} \sim \text{Dirichlet}(\boldsymbol{\alpha}), \qquad \boldsymbol{\alpha} = (\alpha, \dots, \alpha)$$

2 Gibbs sampling

2.1 Initialization

At the beginning, initialize the word alignment randomly. Align each English word in the corpus with a randomly selected word from the respective Czech sentence.

2.2 Sampling

Go through all the English words in a random order. For each such word e_i:

1. Compute the alignment probabilities for all possible Czech counterparts $c_j \in \{c_1 \ldots c_n\}$, based on all other alignment links that are currently in the corpus. Let's denote them as D_{-i}. The predictive probability for a new alignment link $[e_i, c_j]$ is computed as follows:

$$p([e_i, c_j]|D_{-i}) = \int p(e_i|c_j, \theta)p(\theta|D_{-i})d\theta = \frac{count([e_i, c_j]) + \alpha}{count([*, c_j]) + \alpha|E|},$$

   where $count([e_i, c_j])$ is the number of alignment links between the words e_i and c_j in the data D_{-i}, $count([*, c_j])$ is number of alignment links going from the words c_j in D_{-i}, and $|E|$ is a number of distinct English words.

2. Choose one Czech word c_j randomly according to the probability distribution $p([e_i, c_j]|D_{-i})$ and change the alignment link of e_i to c_j. Note that the newly chosen word can be the same as before.

Repeat the process in 20 iterations (20 passes through the data).

2.3 Results

- Display the current word-alignment after the 20th iteration in a suitable format, for example

  ```
  At{Na} school{škole} ,{a} and{a} then{později} at{na}...
  ```

- Based on the counts collected on the last 10 iterations, generate the English-Czech dictionary with the word pairs sorted according to $p(e_i|c_j)$. Do not include Czech words that occure less than five times in the data.

- Try different values of α. How does it affect the inference? What happens if $\alpha = 0$?

- Suggest a better prior distribution than symmetric. For example, boost the probability of alignment links between the equal words (e.g. proper names or numbers).