Example Tasks for Discrete Bayesian Inference

David Mareček

November 29th, 2017
• Word-alignment - acquire a dictionary based on parallel texts
• Word-clustering - group words into clusters
Word alignment

You are given a set of text documents in language A and a set of their translations into a language B.

- Assume you can easily split the documents into sentences.
- Assume that there is 1-to-1 mapping between sentences in documents of A and B.
- Assume you do not have any external dictionary mapping individual words.

The task is to align (connect) the corresponding words between A and B and build a probabilistic dictionary of translations $p(w_a|w_b)$. 
Alignment connections - problems

- There may be different number of words in the corresponding sentences.
- The alignment between words do not need to be 1-to-1.

We want to model $p(w_a|w_b)$. 
More possibilities:
- For each word in $A$, find a counterpart word in $B$.
- For each word in $B$, find a counterpart word in $A$.
- Find all links between $A$ and $B$. 
Predictive distribution

\[ p(w_a|w_b) = \frac{\alpha + \text{count}(w_a, w_b)}{\alpha \cdot |W_a| + \text{count}(w_b)} \]