IBM I: Viterbi Alignment

Jakub Waszczuk

December 2018

1 Preliminaries

As usual, we denote by:

- $\boldsymbol{f}$ – an input sentence of length n (actually $n+1$, if you take NULL into account)
- $\boldsymbol{e}$ – an output sentence of length m
- $A(m, n)$ – the set of all possible alignments between $\boldsymbol{e}$ and $\boldsymbol{f}$ (which only depends on the lengths n and m)

2 Viterbi Alignment

Definition 1. *Given $\boldsymbol{f}$ and $\boldsymbol{e}$, we define the **Viterbi alignment** as the alignment with the highest probability according to the underlying probabilistic model:*

$$\hat{a} = \arg\max_{a \in A(m,n)} P(a \mid \boldsymbol{e}, \boldsymbol{f}) \tag{1}$$

How to determine the Viterbi alignment for a given sentence pair? This depends on the underlying model. Here we only consider IBM-1, for which an exact solution can be easily determined. This is in contrast to, e.g., IBM-3 or IBM-4, for which computing Viterbi alignments is infeasible. As a result, heuristics (e.g., *hill climbing*) which only approximate Viterbi alignments have to be used.

Let's now get back to IBM-1. In this model, the probability of alignment a given $\boldsymbol{f}$ and $\boldsymbol{e}$ is represented by the following formula:

$$P(a \mid \mathbf{e}, \mathbf{f}) = \prod_{i=1}^{m} P(a(i) \mid \mathbf{e}, \mathbf{f}) \tag{2}$$

Where $P(a(i) \mid \mathbf{e}, \mathbf{f})$ is the probability of a particular alignment point $a(i)$:

$$P(a(i) \mid \mathbf{e}, \mathbf{f}) = \frac{P(e_i \mid f_{a(i)})}{\sum_{j=1}^{n} P(e_i \mid f_j)} \tag{3}$$

Both Eq. 2 and Eq .3 were introduced in the second lecture on IBM-1.

Proposition 1. *The Viterbi alignment for a given pair $(\boldsymbol{e}, \boldsymbol{f})$ is the one which maximizes $P(a(i) \mid \boldsymbol{e}, \boldsymbol{f})$ for each output position $i \in \{1, \ldots, m\}$ independently.*

Proof. Let's assume that this is not true. Then, for some sentence pair $(\boldsymbol{e}, \boldsymbol{f})$ and the corresponding Viterbi alignment $\hat{a}$, there would exist an output position i aligned with the input position $\hat{a}(i) = k$ such that:

$$P(a(i) = k \mid \boldsymbol{e}, \boldsymbol{f}) < P(a(i) = k' \mid \boldsymbol{e}, \boldsymbol{f})$$

for some other input position $k' \in \{0, \ldots, n\} \colon k' \neq k$. But then, the probability $P(\bar{a} \mid \boldsymbol{e}, \boldsymbol{f})$ of the alignment $\bar{a}$ defined as:

$$\bar{a}(j) = \begin{cases} \hat{a}(j) & \text{if } j \neq i \\ k & \text{if } j = i \end{cases}$$

would be even higher than the probability of $\hat{a}$. This, however, is contradictory with the assumption that $\hat{a}$ is the Viterbi alignment. □

There is also another, more generic way to understand why the choice of the alignments for the individual output positions can be performed independently. Namely, for any two output positions $i, j \colon i \neq j$, the corresponding two random variables $a(i)$ and $a(j)$ are conditionally independent[1] given $\boldsymbol{f}$ and $\boldsymbol{e}$.

This independence has an intuitive interpretation – whatever the value we choose for $a(i)$, it does not impact the probability of $a(j)$ in any way. Consequently, we can maximize their probabilities independently from each other.

Proposition 2. *Let X, Y be two random variables, conditionally independent given another variable Z. Then, the following holds for any value z of Z:*

$$\max_{x,y} P(x, y \mid z) = \max_{x} P(x \mid z) \times \max_{y} P(y \mid z) \tag{4}$$

Proof. Formally, the conditional independence of X, Y, given Z, means that for any values x, y, z of X, Y, Z:

$$P(x, y \mid z) = P(x \mid z) \times P(y \mid z)$$

Therefore:

$$\begin{aligned} \max_{x,y} P(x, y \mid z) &= \max_{x,y} P(x \mid z) \times P(y \mid z) \\ &= \max_{x} \left(\max_{y} P(x \mid z) \times P(y \mid z) \right) \\ &= \max_{x} P(x \mid z) \times \max_{y} P(y \mid z) \end{aligned}$$

The last transformation above is possible because $P(x \mid z)$ does not depend on y and, therefore, can be extracted out of the inner $\max_y$. □

[1] This conditional independence stems Eq. 2, but we are not going to prove it formally here.