Word Alignment using Bayesian Smoothing and Bidirectional Training

Nancy Wu, Matt Wang

1 Motivation

The motivation for choosing **IBM Model 1** and **EM algorithm**, enhanced by **Bayesian smoothing** and a **bidirectional approach**, was driven by its simplicity and effectiveness in handling word co-occurrences in bilingual texts. Bayesian smoothing was integrated to overcome the model's limitations with rare word pairs, ensuring robust and reliable translations across varying frequencies. The bidirectional technique was adopted to increase the symmetry and accuracy of alignments, considering translations from both language directions and their intersections, thus aligning with advanced methodologies and enhancing the practical applicability of the results.

2 Model & Algorithm

2.1 Initialization

2.1.1 Word count of occurrences:

We initialize three types of counts:

- **Count of occurrences of a French word f_i:**

$$\text{occur}(f_i) = \sum_{(f,e)} \mathbb{I}(f_i \in f)$$

- **Count of occurrences of an English word e_j:**

$$\text{occur}(e_j) = \sum_{(f,e)} \mathbb{I}(e_j \in e)$$

- **Co-occurrence count of a French word f_i and an English word e_j:**

$$\text{occur}(f_i, e_j) = \sum_{(f,e)} \mathbb{I}(f_i \in f \wedge e_j \in e)$$

Where $\mathbb{I}(\text{condition})$ is the indicator function defined as:

$$\mathbb{I}(\text{condition}) = \begin{cases} 1 & \text{if the condition is true} \\ 0 & \text{otherwise} \end{cases}$$

2.1.2 Word Probabilities:

The initial probability of a French word f_i being aligned to an English word e_j is initialized using frequency counts with **bayesian smoothing** parameter α to ensure non-zero probabilities even for rare word pairs:

$$p(f_i \mid e_j) = \frac{\text{occur}(f_i, e_j) + \alpha}{\text{occur}(f_i) + \alpha \cdot N_e}$$

Here, N_e is the total number of unique English words.

2.2 EM Algorithm

2.2.1 E-step: Expectation Calculation

In this step, we compute the expected count of the word pairs. For each pair of French and English sentences (f, e), the normalizing factor $Z(f_i)$ for a French word f_i is computed by summing over all alignments to English words e_j:

$$Z(f_i) = \alpha + \sum_{e_j \in E} p(f_i \mid e_j)$$

The expected count $c(f_i, e_j)$ of the alignment between f_i and e_j is calculated as:

$$c(f_i, e_j) = \frac{p(f_i \mid e_j)}{Z(f_i)}$$

For each word pair (f_i, e_j), we compute the accumulated expected contribution to the alignment:

$$\text{count}(f_i, e_j) \leftarrow \text{count}(f_i, e_j) + c(f_i, e_j)$$

The total expected count for each English word e_j is updated as:

$$\text{total\_count}(e_j) \leftarrow \text{total\_count}(e_j) + c(f_i, e_j)$$

2.2.2 M-step: Maximization of Probabilities

In the M-step, we update the conditional probability $p(f_i \mid e_j)$ for each word pair based on the expected counts calculated in the E-step. The updated probability is given by:

$$p(f_i \mid e_j) = \frac{\text{count}(f_i, e_j) + \alpha}{\text{total\_count}(e_j) + \alpha \cdot N_e}$$

2.3 Covergence Check & Termination

To check if the algorithm has converged, we calculate the change of the probability:

$$\Delta = \sum_{(f_i, e_j)} |p_{\text{new}}(f_i \mid e_j) - p_{\text{old}}(f_i \mid e_j)|$$

If $\Delta < \epsilon$, where ϵ is the convergence threshold, the algorithm stops.

If the result does not converge within the given maximum iteration, the algorithm stops as well.

2.4 Improve Alignment

After the final iteration, the algorithm outputs the alignment that has the highest probability for each French word:

$$\text{best\_alignment} = \arg\max_j p(f_i \mid e_j)$$

After computing the alignments in both directions (French-to-English and English-to-French), the alignments are symmetrized by taking the intersection of the two alignments:

$$\text{intersection} = \{(i,j) : (i,j) \in \text{alignment}_{f \to e} \cap \text{alignment}_{e \to f}\}$$

This ensures that the final alignment is consistent in both directions.

3 Results

In our exploration of machine translation models, we conducted extensive testing with various alignment models including IBM Model 1, IBM Model 2, a diagonal alignment model, a Hidden Markov Model (HMM), and a Bayesian model. Each model was assessed based on their alignment error rate (AER) and overall translation quality in a controlled experimental setting using a standardized bilingual corpus.

Despite the theoretical advantages of more complex models such as IBM Model 2, which incorporates local alignment context, and the HMM, which models alignment as a Markov process, our results indicated that these models did not outperform our baseline when implemented alone. The diagonal model, intended to enhance the probability of aligning words that appear in similar positions across texts, also did not show a significant improvement in performance compared to the baseline.

Surprisingly, the simplest approach, IBM Model 1, when augmented with Bayesian smoothing and a bidirectional alignment strategy, yielded the best results. The Bayesian smoothing effectively addressed the sparsity in the data by smoothing the probability distributions, thus preventing the model from assigning zero probabilities to rare alignments. This enhancement proved crucial in handling the diverse vocabulary present in our corpus. Moreover, the bidirectional method, which considers alignments from both source-to-target and target-to-source perspectives and then intersects them, significantly improved the symmetry and plausibility of the word alignments, which we ended up having the result of: AER rate: 0.322002, Precision rate: 0.62552, and Recall rate: 0.789941.