Word Alignment Implementation

Wei Peng
University of Pittsburgh
wpeng1@andrew.cmu.edu

Abstract

We describe three models for word alignment in statistical translation and present experimental results. The best model achieves a AER of **0.19** and a BLEU score of 16.05 with a memory usage of 912M and a building time of 26m53s.

1 Models

In this section, we briefly introduce the mechanism of the aligners, particularly focusing on the model assumption and procedure of obtaining the best alignment.

1.1 Heuristic Aligner

Heuristic Aligner is not a probabilistic translation model, but rather matches up words based on association, using simple statistics calculated directly from the training corpora. Given a pair of sentences $\mathbf{e} = (e_1, e_2, \ldots, e_I)$ and $\mathbf{f} = (f_1 f_2, \ldots, f_J)$, the best alignment of f_j is given by

$$a_j = \underset{i=1,2,\ldots,n}{\arg\max} \left[c(e_i, f_j)/(c(e_i) \cdot c(f_j)) \right] \quad (1)$$

where the ratio in 1 denotes sentence-level co-occurence count divided by the product of unigram counts for a given English-French token pair.

1.2 Model1

IBM model1 (Och and Ney, 2003) is a generative model which breaks up translation process into smaller steps. It is the simplest possible lexical translation model due to the assumption that all alignments decisions are independent and the alignment distribution for each a_j is uniform over all source words and NULL. The model can be written as

$$p(\mathbf{a}, \mathbf{f}|\theta) = \prod_{j=1}^{J} p(a_j) p(f_j|e_{a_j}, \theta) \quad (2)$$

where $p(a_j) = \frac{1}{I+1}$, and $\theta = (\theta_{e,f})$ is the emission (translation) matrix. Then the best alignment of f_j is given by

$$a_j = \underset{i=1,2,\ldots,n}{\arg\max} \theta_{e_i, f_j}. \quad (3)$$

1.3 HMM

Typically, words are not distributed arbitrarily over the sentence positions, but tend to form clusters. The idea of HMM Aligner (Vogel et al., 1996) is to incorporate localization effect in matching up the words in parallel texts by making the alignment probabilities dependent on the differences in the alignment positions. More specifically, The HMM model can be written as

$$p(\mathbf{a}, \mathbf{f}|\mathbf{e}, \pi, \varphi, \theta) = p(a_1, \pi) \prod_{j=1}^{J} p(a_j|a_{j-1}, \varphi) p(f_j|e_{a_j}, \theta)$$

where $p(a_1 = l, \pi) = \pi_l / \sum_{l=1}^{|I|+1} \pi_i$ and

$$p(a_{j+1}|a_j, \varphi) = \begin{cases} (1-\epsilon)\varphi_{|i-i'|} / \sum_{k=0}^{|I|-1} \varphi_d, & a_j = i, a_{j+1} = i' \\ \epsilon, & a_j = i, a_{j+1} = \text{NULL} \\ (1-\epsilon)/|I|, & a_j = \text{NULL}, a_{j+1} = i \\ \epsilon, & a_j = \text{NULL}, a_{j+1} = \text{NULL} \end{cases}$$

The best alignment is given by

$$(a_1, \ldots, a_J) = \arg\max p(\mathbf{a}, \mathbf{f}|\mathbf{e}, \pi, \varphi, \theta). \quad (4)$$

2 Expectation Maximization (EM) Algorithm

Since $a_1, \ldots, a_J$ are unobserved, the parameters in Model1 and HMM are learned iteratively by EM algorithm. Details of EM can be found in (Borman, 2004). Generally, we have

$$\theta^{n+1} = \arg\max \mathbb{E}_{\mathbf{a}|\mathbf{f}, \theta^n} \left[\ln p(\mathbf{a}, \mathbf{f}|\theta) \right]. \quad (5)$$

The updating formulas for Model1 and HMM are provided in A.

3 Experiments

The data consists of sentence aligned French-English transcripts of the Canadian parliamentary proceedings. 10000 sentences are used for training. Alignment error rate (AER) is chosen to measure the aligning performance and is calculated as:

$$\left(1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}\right)\%. \qquad (6)$$

where A is a set of proposed edges, S is the sure gold edges, and P is the possible gold edges. All parameters are stored with array in log space.

3.1 Experiment Results

The intersected Model1 and the intersected HMM aligner are both trained with 10 epoches. We initialize π, φ and θ uniformly and set $\epsilon = 0.2$. The performance are shown below:

	Heuristic	Model1	HMM
Precision	0.37	0.87	0.92
Recall	0.41	0.63	0.71
ARE	**0.62**	**0.27**	**0.19**
BLUE	11.95	13.57	16.05
MemoryUsage	864M	891M	912M
TrainingTime	4s	57s	26m53s

Table 1: Model Performance

Our implementation of intersected1 HMM achieves a test set AER of 19% when trained with 10000 sentences, while our intersected Model1 implementation achieves 27%.

3.2 First Word Alignment

Since we learned $p(a_0|\mathbf{f})$ ($p(a_0|\mathrm{e})$) independently, instead of simply using φ, it would be interesting to check whether the result meets our guessing that $p(a_0 = 1|\mathbf{f})$ ($p(a_0 = 1|\mathrm{e})$) always dominates the others.

From 1 and 2, we find that the first word of a sentence of a language is always aligned to the first word of the other with high probability. For E2F, we have $p(a_0 = 1|\mathbf{f}) = 0.88$, and for for F2E, $p(a_0 = 1|\mathbf{e}) = 0.76$. All other alignments have very tiny probability. The distribution of $p(a_0|\mathbf{f})$ ($p(a_0|\mathrm{e})$) is closed to a zero-inflated Poisson distribution, which confirm to our intuition.

3.3 Transition

As we mentioned before, adjacent French words should be likely aligned to adjacent French words since they have similar syntactic structure. From 3 and 4, we find that transition with size one is most likely. Alignments do tend to clump together, with adjacent English words usually aligning to adjacent French words as expected. More remarkably, the probability of a transition is a monotone decreasing and roughly smooth function of the transition size. All of the phenomenon reflect that these two languages are closely related.

Additionally, since French sentences typically have longer length than the paired English sentences, the tail of the transition probability in 4 is slightly heavier than that in 3.

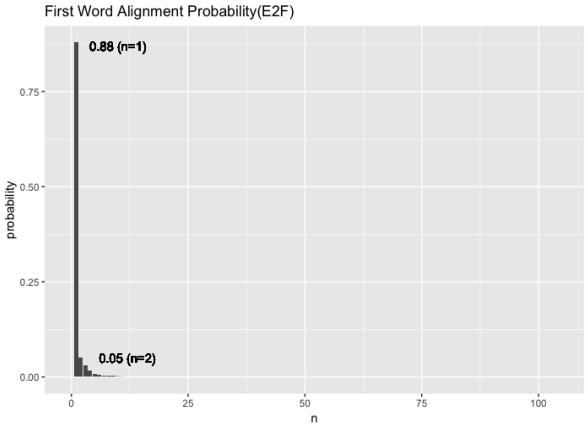

Figure 1: E2F First Word Alignment

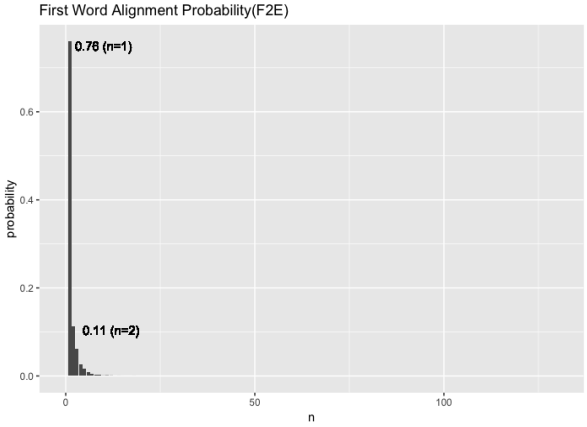

Figure 2: F2E First Word Alignment

3.4 Investigation & Error Analysis

For HMM aligner, to find the best alignment, we use Veterbi to select the one with the largest probability. Surprisingly, we find that E2F HMM Aligner can always aligned "." to "." and "?" to "?"" correctly, but F2E HMM Aligner can only successfully align "?" to "?" but fails to align "." to ".". Therefore, we tune our model slightly

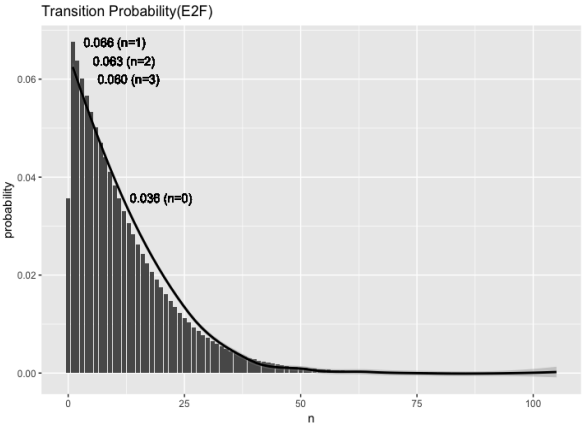

Figure 3: E2F Transition

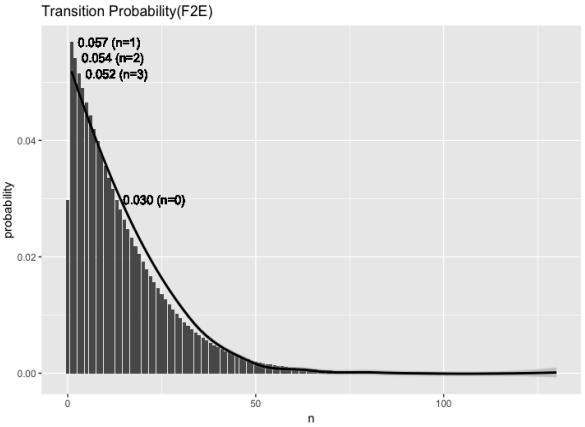

Figure 4: F2E Transition

by selecting the most likely alignment which successfully assigns "." to "." and "?" to "?". Consequently, the AER of the intersected HMM Aligner reduces from 0.23 to 0.19 by 0.04.

References

Sean Borman. 2004. The expectation maximization algorithm-a short tutorial. *Submitted for publication*, 41.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 836–841. Association for Computational Linguistics.

A Appendix: Updating formulas for Model1 and HMM

A.1 Model1

$$
\begin{aligned}
&\arg\max \mathbb{E}_{\mathbf{a}|\mathbf{f},\theta^t}\left[\ln\left(\prod_n\prod_{j=1}^{J_n} p(a_j^n)p(f_j^n|e_{a_j^n},\theta)\right)\right] \\
&= \arg\max \mathbb{E}_{\mathbf{a}|\mathbf{f},\theta^t}\left[\sum_n\sum_{j=1}^{J_n}\ln\left(p(f_j^n|e_{a_j^n},\theta)\right)\right] \\
&= \arg\max \sum_n\sum_{e,f}\sum_{i=1}^{I_n}\sum_{j=1}^{J_n}\ln(\theta_{e,f}) \\
&\qquad 1_{e_i^n=e,f_j^n=f}p(a_j^n=i|\mathbf{f},\theta^t) \\
&= \arg\max \sum_{e,f}\ln(\theta_{e,f})d_{e,f}^n(\theta) \\
&(\text{subject to } \sum_f \theta_{e,f}=1,\forall e)
\end{aligned}
\tag{7}
$$

where $d_{e,f}^n = \sum_n\sum_{i=1}^{I_n}\sum_{j=1}^{J_n} 1_{e_i^n=e,f_j^n=f}p(a_j^n = i|\mathbf{f}_n,\theta^t)$. Therefore, we have

$$\theta_{e,f}^{t+1} = \frac{d_{e,f}^t(\theta)}{\sum_{f'} d_{e,f'}^t(\theta)}$$

where

$$p(a_j^n = i|\mathbf{f}_n,\theta) = \frac{\theta_{e_i^n,f_j^n}1_{a_j^n=i}}{\sum_{i'=1}^{I}\theta_{e_{i'}^n,f_j^n}1_{a_j^n=i'}} \tag{8}$$

A.2 HMM

$$
\begin{aligned}
&\arg\max \mathbb{E}_{\mathbf{a}|\mathbf{f},\pi^t,\varphi^t,\theta^t}\left[\ln\left(\prod_n p(a_1^n,\pi)\right.\right. \\
&\left.\left. p(a_j^n|a_{j-1}^n\varphi)\prod_{j=2}^{J_n}\prod_{j=1}^{J_n}p(f_j^n|e_{a_j^n},\theta)\right)\right] \\
&= \arg\max \mathbb{E}_{\mathbf{a}|\mathbf{f},\pi^t,\varphi^t,\theta^t}\left[\sum_n[\ln p(a_1^n,\pi)+\right. \\
&\left.\sum_{j=2}^{J_n}\ln p(a_j^n|a_{j-1}^n,\varphi)+\sum_{j=1}^{J_n}\ln\left(p(f_j^n|e_{a_j^n},\theta)\right)\right] \\
&= \arg\max \mathbb{E}_{\mathbf{a}|\mathbf{f},\pi^t,\varphi^t,\theta^t}\left[\sum_n\sum_{j=1}^{J_n}\ln\left(p(f_j^n|e_{a_j^n},\theta)\right)\right. \\
&\left.+\sum_n\ln p(a_1^n,\pi)+\sum_n\sum_{j=2}^{J_n}\ln p(a_j^n|a_{j-1}^n,\varphi)\right] \\
&= \mathrm{I}+\mathrm{II}+\mathrm{III}
\end{aligned}
$$

Consider I, we have

$$\arg\max \mathbb{E}_{\mathbf{a}|\mathbf{f},\pi^t,\varphi^t,\theta^t}\left[\sum_n\sum_{j=1}^{J_n}\ln\left(p(f_j^n|e_{a_j^n},\theta)\right)\right]$$
$$=\arg\max\sum_n\sum_{e,f}\sum_{i=1}^{I_n}\sum_{j=1}^{J_n}\ln\left(\theta_{e,f}\right)$$
$$1_{e_i^n=e,f_j^n=f}p(a_j^n=i|\mathbf{f},\pi^t,\varphi^t,\theta^t)$$
$$=\arg\max\sum_{e,f}\ln(\theta_{e,f})d_{e,f}^n(\theta)$$
$$(\text{subject to }\sum_f\theta_{e,f}=1,\forall e)$$

thus we obtain a similar updating formula for θ as Model 1. Now consider II,

$$\arg\max \mathbb{E}_{\mathbf{a}|\mathbf{f},\pi^t,\varphi^t,\theta^t}\sum_n\sum_{j=2}^{J_n}\ln p(a_j^n|a_{j-1}^n,\varphi)$$
$$=\arg\max\sum_n\sum_{j=1}^{J_n}\sum_{i=1}^{I_n}\sum_{i=1}^{I_n}\sum_k\times 1_{|i-i'|=k}$$
$$\ln\left(\frac{\varphi_k}{\sum_{k=0}^{I-1}\varphi_k}\right)p(a_{j-1}^n=i,a_j^n=i'|\mathbf{f}_n,\pi^t,\varphi^t,\theta^t)$$
$$(\text{subject to}\sum_k\varphi_k=1)$$

Unfortunately, there is no close form for φ^{t+1}. Therefore, we calculate a heuristic solution by following:

1. Get

$$\varphi^{t+1,n}=\arg\max\sum_{j=1}^{J_n}\sum_{i=1}^{I_n}\sum_{i=1}^{I_n}\sum_k\times 1_{|i-i'|=k}$$
$$\ln\left(\frac{\varphi_k}{\sum_{k=0}^{I-1}\varphi_k}\right)p(a_{j-1}^n=i,a_j^n=i'|\mathbf{f}_n,\pi^t,\varphi^t,\theta^t)$$
$$=\arg\max\sum_{k=0}^{|I|-1}\left(\frac{\varphi_k}{\sum_{k=0}^{I-1}\varphi_k}\right)d_k^t(\varphi)$$

where $d_k^n(\varphi)=\sum_{j=1}^{J}\sum_{i=1}^{I}\sum_{i=1}^{I}\sum_k 1_{|i-i'|=k}p(a_{j-1}^n=i,a_j^n=i'|\mathbf{f},\varphi^t,\theta^t)$. Thus, we have

$$\varphi_k^{(t+1),n}=\frac{d_k^n(\varphi)}{\sum_{k'}d_{k'}^n(\varphi)}$$

2. Calculate

$$\varphi_k^{t+1}=\frac{\sum_n\varphi_k^{(t+1),n}}{\sum_k\sum_n\varphi_k^{(t+1),n}}$$

Now, consider III.

$$\arg\max\mathbb{E}_{\mathbf{a}|\mathbf{f}_n,\pi^t,\varphi^t,\theta^t}\sum_n\ln p(a_1^n,\pi)$$
$$=\arg\max\sum_n\ln\left(\frac{\pi_l}{\sum_{l=1}^{I+1}\pi_i}\right)p(a_1^n=l|\mathbf{f}_n,\pi^t,\varphi^t,\theta^t)\tag{9}$$

Unfortunately, again we can not get a close form of π^{t+1}. Therefore we get a heuristic solution by following:

1. Calculate $\pi_l^{(t+1),n}=p(a_1^n=l|\mathbf{f}_n,\pi^t,\varphi^t,\theta^t)$
2. Calculate

$$\pi_l^{(t+1)}=\frac{\sum_n\pi_l^{(t+1),n}}{\sum_l\sum_n\pi_l^{(t+1),n}}$$

What left is to find the formula for $p(a_{j-1}^n=i,a_j^n=i'|\mathbf{f},\pi^t,^t,\theta^t)$ and $p(a_j^n|\pi^t,\varphi^t,\theta^t)$. Unsurprisingly, we can get them from the standard Baum-Welch (Forward-Backward) method. Note that the p_i and φ_d are the global parameters of the HMM model. Instead π_l is indeed $\pi_l/\sum_{l=1}^{I+1}\pi_l$ and φ_d is indeed $\varphi_d/\sum_{d=0}^{I-1}\varphi_d$.

A.3 Forward-Backward

Forward

$$\alpha_i(1)=p(F_1=f_1,a_1=i|\varphi,\theta)=\pi_i\theta_{e_i,f_1}$$
$$\alpha_i(j)=p(F_1=f_1,\ldots,F_j=f_j,a_j=i|\pi,\varphi,\theta)$$
$$=\sum_{i'}\alpha_{i'}(j-1)\varphi_{|i-i'|}\theta_{e_i,f_j}\quad 1<j\leq n.\tag{10}$$

Backward

$$\beta_i(\eta)=1$$
$$\beta_i(j)=p(F_{j+1}=f_{j+1},\ldots,F_n=f_n|a_j=i,\pi,\varphi,\theta)$$
$$=\sum_{i'}\varphi_{|i-i'|}\theta_{e_{i'},f_{j+1}}\beta_{i'}(j+1)$$
$$\forall 1\leq j<n.\tag{11}$$

Update

$$\gamma_i(j)=p(a_j=i|\mathbf{f},\pi,\varphi,\theta)$$
$$=\frac{\alpha_i(j)\beta_i(j)}{\sum_i\alpha_i(j)\beta_i(j)}$$
$$\xi_{ii'}(j)=p(a_j=i,a_{j+1}=i'|\mathbf{f},\pi,\varphi,\theta)$$
$$=\frac{\alpha_i(j)\varphi_{|i-i'|}\theta_{e_{i'},f_{j+1}}\beta_{i'}(j+1)}{\sum_i\sum_{i'}\alpha_i(j)\varphi_{|i-i'|}\theta_{e_{i'},f_{j+1}}\beta_{i'}(j+1)}$$