MT HW2

Arijit Nukala, Claire Jung, and Bhavik Agarwal

September 21, 2023

The field of machine translation has consistently sought to improve the alignment of words between source and target languages. Proper word alignment not only aids in generating accurate translations but also enhances the understanding of linguistic nuances between languages. This quest for precision has led to the development and evolution of various models, which have evolved in complexity and accuracy iteratively. Among them, the IBM models represent the pinnacle of statistical machine translation and served as the gold standard of translation systems until the rise of neural machine translation.

1 IBM Model 1

1.a Motivation

The task of word alignment is a fundamental challenge in machine translation. Accurate word alignments are crucial for producing high-quality translations. IBM Model 1, a classic statistical model, has been widely used for word alignment tasks due to its simplicity and effectiveness and is one of the most prominent statistic translation methods.

1.b Technical Description of the Model

IBM Model 1, the first of the IBM translation models, establishes the foundation for more intricate alignment models. Its principle rests on a probabilistic generative procedure, with core assumptions to simplify alignment complexities.

1.b.1 Generative Process

The alignment procedure is mathematically represented by:

$$P(f, a|e) = \prod_i P(a_i = j||e|) \times P(f_i|e_j) \tag{1}$$

Where:

- f and e are the foreign and English sentences respectively.
- a represents the alignment of foreign words to English words.
- $P(f_i|e_j)$ denotes the translation probability of foreign word f_i given English word e_j.
- $P(a_i = j||e|)$ is the alignment probability for the i^{th} foreign word with the j^{th} English word.

A pivotal assumption in Model 1 is the uniform alignment probability. Each foreign word f_i aligns to precisely one English word e_j (or to a special NULL word). Therefore, $P(a_i = j||e|)$ is uniform, treating it effectively as a constant.

1.b.2 Expectation-Maximization (EM) in IBM Model 1

Given a parallel corpus, IBM Model 1 utilizes the Expectation-Maximization (EM) algorithm for parameter estimation. This iterative approach refines translation probabilities, alternating between:

- **E-step (Expectation):** With the current estimates of the translation probabilities, compute the expected count $c(e, f)$ of word pairs (e, f) over all sentence pairs in the corpus.

$$c(e, f) = \sum_{(e,f) \in \text{corpus}} \frac{t(f|e)}{\sum_{e'} t(f|e')} \tag{2}$$

- **M-step (Maximization):** Update the translation probabilities $t(f|e)$ by normalizing the expected counts.

$$t(f|e) = \frac{c(e, f)}{\sum_{f'} c(e, f')} \tag{3}$$

The EM algorithm guarantees non-decreasing likelihood of observing the data and often converges to a local maximum.

1.b.3 Limitations

The simplicity of IBM Model 1 is both its strength and weakness. The model assumes uniform alignment probability, which does not work well for long sentences and complex grammar. This motivated the exploration of more advanced models like IBM Model 2, which introduced alignment probabilities and offers improved alignment performance.

1.c Experimental Results

We conducted experiments on a bilingual dataset containing 1000 sentence pairs. Our implementation of IBM Model 1 achieved the following results:

- Precision: 0.608
- Recall: 0.799
- AER: 0.331

The Precision score measures the proportion of correctly aligned words out of all the words in the predicted alignments. Recall quantifies the proportion of correctly aligned words out of all the words that should have been aligned. AER is a composite metric that considers both Precision and Recall, providing an overall evaluation of alignment quality.

1.d Discussion and Conclusion

Our implementation of IBM Model 1 demonstrated promising results in word alignment. The Precision and Recall scores indicate that the model effectively captures word correspondences between languages. However, the AER suggests room for improvement in alignment accuracy. One limitation of IBM Model 1 is its assumption of one-to-many alignments for English words, which may not always hold in practice. While Model 1 is a valuable tool for word alignment tasks, there are improvements that can be made to increase performance.

2 IBM Model 2

2.a Motivation

Building upon the foundation laid by IBM Model 1, IBM Model 2 was developed to address the limitations faced by its predecessor, especially regarding the positional information of words in different languages.

2.b Technical Description of IBM Model 2

IBM Model 2 extends IBM Model 1 by relaxing the uniform alignment probability assumption, thus accounting for position-based alignment probabilities. This recognizes the fact that word order can significantly vary between languages and positions of words in one language can influence the alignment in another.

2.b.1 Generative Process

The alignment procedure for IBM Model 2 can be mathematically represented as:

$$P(f, a|e) = \prod_{i=1}^{l} \left(\frac{1}{(l+1)} \times P(f_i|e_{a_i}) \times a(a_i|i, l, m) \right) \tag{4}$$

Where:

- l and m are the lengths of the English and foreign sentences respectively.
- $a(a_i|i, l, m)$ denotes the alignment probability of the i^{th} foreign word to the a_i^{th} English word, given the sentence lengths.

The newly introduced term, $a(a_i|i, l, m)$, allows Model 2 to learn tendencies like the beginning words in one language aligning more often with the beginning words in another.

2.b.2 Parameter Estimation using Expectation-Maximization

For IBM Model 2, the EM algorithm has to estimate both the translation probabilities $t(f|e)$ and the alignment probabilities $a(j|i, l, m)$. The process alternates between:

- **E-step (Expectation):** Computes expected counts using current parameter estimates. Given the current probabilities, we determine the expected number of times a particular alignment is utilized.
- **M-step (Maximization):** Update the probabilities by maximizing the expected likelihood. The translation probabilities are updated in a similar manner to Model 1, but we also adjust the alignment probabilities based on observed alignments.

2.b.3 Limitations

While IBM Model 2 offers improvements over Model 1 by incorporating positional alignment probabilities, it still has limitations. The model assumes that alignment probabilities are only dependent on relative positions rather than the actual words themselves. This can be a drawback for languages with specific syntactical structures. Additionally, IBM Model 2 does not account for linguistic constructs like phrase-based alignments. Further advancements in the IBM Models series and other alignment models like the phrase-based models address these challenges.

2.c Experimental Results

Experiments were performed on a bilingual dataset containing 1000 sentence pairs. Our implementation of IBM Model 2 produced the subsequent outcomes:

- Precision: 0.632
- Recall: 0.825
- AER: 0.306

The precision, recall, and AER all show improvement over IBM Model 1.

2.d Discussion and Conclusion

A notable enhancement of IBM Model 2 over Model 1 is its capability to factor in position-based alignment probabilities. However, it's still constrained by its reliance on relative positions instead of context. While IBM Model 2 advances over its predecessor in capturing positional alignments, there remains potential for further enhancements.

3 Test Results

IBM Model 1 Statistics on English/French Text			
Sentences + Stemming	Precision	Recall	AER
100 No Stemming	0.330	0.391	0.651
100 Stemming	0.348	0.438	0.623
1000 No Stemming	0.434	0.586	0.517
1000 Stemming	0.465	0.654	0.475
10000 No Stemming	0.502	0.669	0.445
10000 Stemming	0.541	0.748	0.393
All No Stemming	0.581	0.737	0.369
All Stemming	0.608	0.799	0.331

IBM Model 2 Statistics on English/French Text			
Sentences + Stemming	Precision	Recall	AER
100 No Stemming	0.340	0.382	0.647
100 Stemming	0.363	0.429	0.616
1000 No Stemming	0.473	0.607	0.484
1000 Stemming	0.510	0.666	0.440
10000 No Stemming	0.581	0.757	0.363
10000 Stemming	0.627	0.796	0.319
All No Stemming*	0.598	0.812	0.316
All Stemming*	0.632	0.825	0.306

* Model 2 All results were run on 20,000 sentences due to memory errors when running on the dataset. Despite converting python dictionaries to numpy arrays to reduce memory usage, we were unable to run Model 2 across all data with hitting memory errors.

4 Discussion

4.a Effect of Dataset Size

From the results, it's evident that increasing the number of sentences in the training data significantly improves the alignment performance, especially in terms of Precision, Recall, and Alignment Error Rate (AER). This is expected since more data provides the models with a richer context, allowing for more accurate alignments.

4.b Impact of Stemming

Stemming, the process of reducing words to their base or root form, seems to consistently enhance the performance of both models across different dataset sizes. Particularly, when all sentences are considered, the application of stemming in IBM Model 2 resulted in an AER reduction from 0.316 (No Stemming) to 0.306. This is likely because reducing words down to their root form allows for more consistency over the corpus, especially when translating different tenses of words.