CS224N Assignment 1: Statistical Machine Translation

Mihail Eric — meric@stanford.edu Victor Zhong — vzhong@stanford.edu

October 6, 2014

1 Introduction

In this report we show machine translation for French-English, Hindi-English,and Chinese-English using the models BaselineWordAligner, PMIModel, and IBMModels 1 and 2. We first describe alignment results and then describe features that improve translation quality.

2 Alignment

For every model we evaluate the alignment error rate for each translation pair. We train the model based on the training set (10K) and tune the model on the development set. We then evaluate the best hyperparameter setting on the test set.

2.1 Baseline

The baseline alignment model simply aligns every source word with the corresponding target word along the diagonal of the standard bilingual word alignment matrix. Here, we abbreviate Alignment Error Rate with AER.

	Fr-En	Hi-En	Ch-En
dev	0.7122	0.8713	0.9117
test	0.6865	0.8849	0.8816

Table 1: AER for BaselineWordAligner

2.2 Pointwise Mutual Information

Given a French-English sentence pair, denoted by $f = (f_1, f_2, ..., f_k)$ and $e = (e_1, e_2, ..., e_m)$, this model seeks to determine the optimal alignment for a pair of words $f_i \in f$, $e_j \in e$ satisfying:

$$a_i = \arg\max_i \frac{p(f_i, e_j)}{p(f_i)p(e_j)}.$$

	Fr-En	Hi-En	Ch-En
dev	0.7107	0.8451	0.8278
test	0.6917	0.8221	0.8093

Table 2: AER for PMI Model

A few notes regarding the performance and errors from this model. We use the following sentence to illustrate the alignment quatliy: "in recent years shanghai 's pudong has promulgated and implemented 71 regulatory documents relating to areas such as economics , trade , construction , planning , science and technology , culture and education , etc. , ensuring the orderly advancement of pudong 's development ."

The PMI Model tends to wrongly align functional words with content words. For example, for the sentence "the development of shanghai 's pudong is in step with the establishment of its legal system", "the" is aligned with the Chinese word for "development" and "of" is aligned with the Chinese word for "legal system". The wrongly aligns many words to one word, and has trouble dealing with phrases/words that are reordered across languages ("trade", "construction", "planning" are aligned to

"seventy-one" in Chinese). It has a tendency to align many content words to one word.

2.3 IBM Model 1

Given a French-English sentence pair, denoted by $f = (f_1, f_2, ..., f_m)$ and $e = (e_1, e_2, ..., e_n)$, and an alignment from English to French words denoted by $a_1, a_2, ..., a_n$, this model seeks to learn the following distribution: $p(e_1, ..., e_n, a_1, ..., a_n | f_1, ..., f_m, n) = \prod_{i=1}^{n} q(a_i|i, n, m)t(e_i|f_{a_i})$ where $t(e|f)$ denotes the probability of generating English word e from French word f and $q(j|i, n, m)$ is the probability of alignment variable a_i taking value j, conditioned on the lengths of the English and French sentences. IBM Model 1 assumes that the $q(\cdot|\cdot)$ parameters are a constant, so that all possible alignments for a given word have equal probability.

We train Model 1 using the EM algorithm to learn the parameters. Moreover, to ensure that the probability of null alignment did not vary with the length of an English sentence, we set the null alignment probability as a separate constant, denoted by p_{NULL}. The hyperparameters for this model are then the null probability and the number of EM iterations.

We train the model on $10,000$ training examples, and tune the hyperparameters on the validation set. We then evaluate the model with the optimal dev hyperparameters on the test set.The results for hyperparameter tuning across languages are provided below with the optimal dev parameters in bold. For space reasons, we place the French tuning results in the Appendix.

	Hindi				
	0.02	0.04	**0.06**	0.08	1
1	0.8451	0.8445	0.8486	0.852	0.8541
3	0.6392	0.6425	0.6429	0.6443	0.6557
10	0.5876	0.586	0.5834	0.5872	0.596
30	0.5876	0.5841	0.5796	0.5833	0.5842
100	0.5876	0.5841	0.582	0.5819	0.5837

Table 3: Model 1 Hindi Hyperparameter Tuning

	Chinese				
	0.02	0.04	**0.06**	0.08	1
1	0.8292	0.8274	0.8262	0.8267	0.8275
3	0.6216	0.6121	0.6148	0.6201	0.6246
10	0.5946	0.5828	0.5813	0.5848	0.5882
30	0.5898	0.5777	0.5739	0.577	0.5799
100	0.5878	0.576	0.5713	0.5733	0.5759
130			0.5784		

Table 4: Model 1 Chinese Hyperparameter Tuning

	Fr-En	Hi-En	Ch-En
dev	0.3441	0.5796	0.5713
test	0.3451	0.5837	0.5794

Table 5: AER for IBM Model 1

Model 1 also tends to wrongly align functional words with content words. For example, "of" is aligned with the Chinese word for "development". In the sentence , all commas are aligned to the Chinese word for "such as". The model wrongly aligns many words to one word; however, it is able to align content words that are reordered across languages ("trade", "construction" etc. are aligned correctly.).

2.4 IBM Model 2

Model 2 describes the same distribution as IBM Model 1 without assuming uniform distribution for alignment probabilities. That is, the $q(\cdot|\cdot)$ parameters are not constant and are treated as parameters to be estimated via EM. Again, we tune the null alignment probability and the EM iterations using the same dataset as Model 1 tuning. The results for hyperparameter tuning across languages are provided below with the optimal dev parameters in bold. Due to space constraints, we place the French tuning results in the Appendix.

	Hindi				
	0.02	0.04	0.06	0.08	**0.1**
1	0.652	0.652	0.652	0.661	0.673
3	0.587	0.584	0.582	0.584	0.587
10	0.587	0.584	0.582	0.578	0.580
30	0.585	0.579	0.575	0.573	0.572
100	0.582	0.575	0.573	0.573	0.572

Table 6: Model 2 Hindi Hyperparameter Tuning

	Chinese				
	0.02	0.04	**0.06**	0.08	1
1	0.632	0.628	0.635	0.640	0.643
3	0.589	0.578	0.580	0.585	0.588
10	0.583	0.571	0.571	0.573	0.576
30	0.582	0.570	0.568	0.569	0.571
100	0.582	0.570	0.567	0.567	0.568

Table 7: Model 2 Chinese Hyperparameter Tuning

	Fr-En	Hi-En	Ch-En
dev	0.3352	0.5721	0.5676
test	0.3407	0.5775	0.5742

Table 8: AER for IBM Model 1

Model2 performs similarly to Model1. However it does a better job of aligning functional words. In this case, it correctly aligns the first "in". Interestingly, Model2 also aligns many of the commas in this sentence to "NULL". All models struggle to align words such as "'s". Model2 does a surprisingly good job of recognizing functional words in English and aligning them to NULL. For example, many of the functional words in "the opening up to the outside of china 's construction industry began in the 1980 's ." are appropriately aligned to NULL in the Chinese sentence.

2.5 Machine Translation Feature Engineering

In the appendix we have included the BLEU score results of our Phrasal MT system for a number of feature sets, including combinations of these features. For each feature or set of features, we ran the system three times, due to the variability of the beam search decoding process, and computed an average BLEU score. All of the feature combinations that beat the baseline are highlighted in bold.

The TPS (TPS) feature, which was suggested in the assignment handout, captures the number of words in every rule in a given derivation, reflecting the intuition that small rules generally don't capture much bilingual information, while large rules are typically sparse.

The punctuation ratio (PR) feature, suggested by Green et. al. (2014) computes a ratio between the number of punctuation marks in the target phrase to the number of punctuation marks in the source phrase. This feature is meant to capture the intuition that use of punctuation marks varies greatly across languages and that false alignments typically align punctuation marks to words.

The phrase absolute diff (PAD) feature computes the absolute value of the difference between the number of words in the target phrase and the number in the source phrase. This feature is meant to reflect the notion that the quality of a translation rule is often inversely correlated to the difference in length from source to target phrases. The three largest BLEU score improvements came from including the PAD+PR features (+0.06), the TPS feature (+0.114), and the PAD+TPS features (+0.25).

When comparing the translation outputs to baseline, the TPS feature occasionally had more appropriate use of certain prepositions such as *of*. However, sometimes verbs in the translation of the system with this feature did not conjugate verbs appropriately (i.e. *had* vs. *having*).

The TPS+PR feature system had the advantage of occasionally including articles in sentences (e.g. *the* that the baseline system misses. Additionally, sometimes this system has more obfuscated word reorderings as compared to baseline. The PAD+TPS feature system also has somewhat better article inclusion as compared to baseline.

3 Appendix

	1	3	10	30	**100**	130
0.1	0.7423	0.4125	0.3598	0.3507	0.3542	
0.12	0.7433	0.4379	0.3684	0.3462	0.3446	
0.14	0.7413	0.4475	0.3816	0.3582	0.3441	0.3443
0.16	0.7475	0.4605	0.3995	0.3655	0.3555	
0.18	0.7484	0.4721	0.4055	0.3791	0.3643	
0.2	0.7462	0.4818	0.4111	0.3862	0.3746	
0.22	0.7484	0.4873	0.4223	0.3926	0.3869	

Table 9: Model 1 French Hyperparameter Tuning

	1	3	10	30	**100**	130
0.1	0.4261	0.3545	0.3487	0.3453	0.3491	
0.12	0.4509	0.3527	0.345	0.3384	0.3451	
0.14	0.4668	0.367	0.3462	0.3363	0.3359	
0.16	0.4821	0.377	0.3419	0.3392	0.3352	0.3403
0.18	0.4913	0.3914	0.3484	0.3359	0.3359	
0.2	0.4943	0.4014	0.3597	0.3435	0.3367	
0.22	0.4974	0.406	0.3661	0.3554	0.3461	

Table 10: Model 2 French Hyperparameter Tuning

Feature Name	Run 1	Run 2	Run 3	Average
Baseline	15.487	15.224	15.075	15.262
PR	15.385	15.244	15.148	15.259
TPS	15.493	15.282	15.354	15.376
TPS + PR	15.133	15.042	15.411	15.195
PAD	15.406	15.325	15.155	15.295
PAD^2	15.231	15.337	15.069	15.212
PAD + PR + TPS	15.263	15.446	15.014	15.241
PAD + TPS	15.654	15.547	15.335	15.512
PAD + PR	15.508	15.329	15.129	15.322

Table 11: Feature Sets with BLEU Scores