MT & S2S Models

Symbolic MT

By

Varun Prashant Gangal

Language Technologies Institute
CMU

Submitted On
March 23, 2017

Model

Alignment Model

We train IBM Model 1 with 10 iterations of EM on the training set. Note that we do not train alignments bidirectionally, but only unidirectionally. We keep a minimum frequency of 5 for English and 3 for German for UNKing. The language model used is always a bigram model, with $\alpha_1 = 0.15$.

Translation FST

Since IBM Model1 leaves aspects such as reordering and fertility unspecified (or rather, different alignments with different reordering and fertility are equivalent under the model as long as the word pairs matched are the same), we can capture some of our biases through the translation FST. We experiment with some word-word FST models (note that none of these approaches use phrases)

Non-Phrase Based Models

1. **OneToOne**: This approach needs to generate an English word for every German word it sees. The probabilities used are the Model1 probabilities. Note that this does not let us generate multiple German words for one English word. Surprisingly, though, this approach does not so badly on the test set, getting a BLEU score of 13.24. One reason for this is unlike the other approaches, it maintains a net hypothesis length close to the net reference length.

2. **GeneralModel1**: Here, we transition to a state corresponding to an English word (by generating an English word). In this state, any number of German words can be consumed along the self-directed arcs (incurring associated log probability costs). This model does worse than OneToOne, getting a BLEU score of 12.15. This is primarily due to the shorter length hypotheses this tends to produce

3. **FertilityConstraints**: Here, our state keeps track of how many German words we have let an English word generate. After generating two German words, one cannot generate further German words from the same English word. Note that we have not incorporated fertility into our alignment model estimation, and this is merely a heuristic we employ at testTime. This improves the BLEU to 12.44

4. **FertilityConstraints With Downweighing**: Here, we penalize (by an exponent on the probability), the generation of the second German word from an English word. This encourages the model to maintain a 1:1 ratio of lengths (roughly) as far as possible. This improves the BLEU score to 13.19 with a fertility constraint of 2

Results - Non Phrase Models

Model Type	BLEU
OneToOne	13.24
General	12.15
FertilityConstraints	12.44
FertilityConstraints+Downweighing (2)	13.08
FertilityConstraints+Downweighing (2.5)	13.19

Results - Phrase Models

Model Type	English Length	German Length	BLEU
Phrases+WordBackoff	2	2	15.20
Phrases+WordBackoff	2	3	15.23
Phrases+WordBackoff	3	3	15.44
Phrases+WordBackoff	3	6	15.58
Phrases+WordBackoff+UNKStrip	3	6	15.62
Phrases+WordBackoff+UNKStrip	4	6	15.64
Phrases+FertilityBackoff+UNKStrip	3	6	15.63

Phrase Based Models

We use the phrase-FST like in the lecture notes. However, we additionally chain a non-phrase word-level FST (with the same start and end state), to prevent cases where there is no valid path in the FST as a result of unseen phrases. Since the phrases usually have greater probabilities, the word-level FST only acts as a check for the unseen case (We included this after we noticed some of the inputs giving NULL English sentences as output). We refer to this model as Phrases+WordBackOff. We also observe that stripping away UNKs leads to a (very slight) improvement in BLEU score. Making the word-backoff part of the FST better does not improve the BLEU much.

Effect of Phrase Length

Increasing the phrase length ceases to improve BLEU after a point (greater than a length of 4)

Phrase Extraction

We extract phrases using the algorithm given in the readings. However, we notice that extracting phrases from the large training set leads to an explosion of phrase-pairs - 0.2 million with a length limit of 2 and 0.5 million with a length limit of 3, which make our FST's decoding inefficient at test time (due to a large number of states). Conceptually as well, it feeds in the training set's errors twice into our model (once while learning Model1, and the second through the alignments). Hence, we learn our alignments on the validation set.

Conclusion

The Phrases+WordBackOff+UNKStrip with English Phrase Length 4 and French Phrase Length 6 performs the best amongst all our approaches.

Code

The code for this assignment is accessible at https://github.com/vgtomahawk/MT_SYMBOLIC_CMU