11-731 Assignment 2
Phrase Based Machine Translation

Varsha Embar (vembar)

Introduction

Statistical machine translation generate translations using statistical models. The parameters are estimated using bilingual corpora. The basic idea comes from information theory where a document is tranlated according to the probability distribution $P(e|f)$ where e is the target language and f is the source language. To model this distribution, Bayes Theorem is used as follows $P(e|f) \sim P(f|e)P(e)$ where the translation model $P(f|e)$ is the probability that the source string is the translation of the target string, and the language model $P(e)$ is the probability of seeing that target language string. Finding the best translation is done by picking up the one that gives the highest probability.

Phrase based machine translation aims to coeve the short coming of word based machine translation by translating whole sequences of words, where the lengths may differ. The IBM models were an early approach to SMT. The models are still very useful and the approach taken by there models are as follows:

- The models make direct use of the idea of alignments, and as a consequence allow us to recover alignments between French and English words in the training data.
- The parameters of the IBM models will be estimated using the expectation maximization (EM) algorithm.

In this assignment, we use IBM Model 1 to get our word alignments, perform phrase extraction and create a weighted finite state transducer to get our translation model. The details of each model is explained in the coming sections.

IBM Models

IBM Model 1 was used to get alignments and an attempt was made to also build IBM Model 2.

IBM Model 1 assumes that given a finite set of source words F and target words E, and maximum length of both sentences, it calculates the parameter $t(f|e)$ which is the conditional probability of generating French word f from English word e. It computes the probability of translation of the source sentence, it's alignment with target sentence given the target sentence and it's respective lengths.

$$P(f_1 \ldots f_n,\ a_1 \ldots a_n \mid e_1 \ldots e_m,\ n) = \prod \frac{1}{l+1}\ x\ t(f \mid e_a) = \frac{1}{(l+1)^n} \prod t(f \mid e_a)$$

We estimate these parameters using Expectation Maximization algorithm which is a two step process. We calculate the expectations of counts of latent variables using the current model parameters in the E step and update the model parameters using the counts in the M step.

IBM Model 2 on the other hand is based on reordering between sentences F and E essentially has a canonical word order..The simple alignment of Model 1 is replaced by a more sophisticated one which is learnt from the data. A slightly more complicated version of the same exists which uses HMM alignment considering the word order.

Phrase Extraction

The Phrase Extraction model memorizes multi-symbol strings, and translates this string as a single segment. This provides us the advantage that translations are done uniformly, in that articles like “a” and “an” can be correctly translated when done with the qualified noun as well. Also, it is very useful when multiple words are aligned to the same word between the source and target language.

Both IBM Model 1 and the Phrase Extraction modules were build by referring to the class notes which can be found in the reference section.

Weighted Finite State Automaton

In order to generate the actual translation, we have a large search space. One way of handling this is to use WFSTs. Each phrase extracted is represented as a path in a WFST which acts as a search graph. Each node represents a coverage vector, and each edge between nodes represents the translation of a particular phrase. The score of these edges would be the sum of the negative log phrase translation and reordering probabilities calculated during the phrase extraction phase.

Experiments

We perform experiments using IBM Model 1, followed by phrase extraction and WFST model to translate from german to english using the IWSLT data. We set a limit on the maximum length of phrases to 4. The model takes 4-5 hours to train on the train set which has 99412 samples and to test on a validation set of 887 samples and test set of 1565 samples. The BLEU scores of the experiment is shown in the table below.

Model	Validation	Test
Using IBM Model 1	17.83	17.23

Table 1. Results of experiment with IBM Model 1

The code for both IBM Model 1 can be found at https://github.com/vembar/11-731_Phrase_MT . Experiments on IBM Model 2 were not complete during the time of submission.

Conclusion

SMT, although a classical approach to MAchine Translation, is still of importance to us. The BLEU scores achieved on the dataset are comparable to those of the basic Encoder-Decoder model used in the first assignment. But in recent years, SMT has resurfaced where models that use the best of both worlds - that of deep MT models and statical MT models - seem to be the theme of the day.

References

- http://www.phontron.com/class/mtandseq2seq2017/mt-spring2017.chapter11.pdf
- http://www.phontron.com/class/mtandseq2seq2017/mt-spring2017.chapter13.pdf
- http://www.cs.columbia.edu/~mcollins/courses/nlp2011/notes/ibm12.pdf
- https://en.wikipedia.org/wiki/Statistical_machine_translation