Assignment 02: Phrase based Machine Translation

Sai Krishna Rallabandi

March 26, 2017

Abstract

In this assignment, I present a full statistical machine translation pipeline, which involves building a language model as well as a phrase-based translation model.

1 Introduction

The task in this assignment was to build a phrase based machine translation system capable of generating English from German, which involved developing the following:

- An alignment module using unidirectional IBM Model 1 which aligns the German and English sentences.
- A phrase extraction algorithm to extract aligned phrases.
- Constructing a weighted finite state transducer.

2 System Overview

The idea is to follow a noisy channel approach with two components:

- A language model that assigns a probability p(e) for an English sentence e = e1...en where n is the length of the sentence. As a baseline, a bigram language model with interpolation was provided.
- An alignment/translation model that assigns a conditional probability based on the parallel corpus. The parameters of this model are estimated from the corpus using an Expectation-Maximation approach.

2.1 IBM Model 1

A unidirectional IBM model was implemented as baseline which didnot have null alignments. I then allowed null alignments in the model. I have extended this approach by using bidirectional alignments and then intersected the alignments from the models.

2.2 Phrase Extraction

I have implemented Algorithm 6 based on the pseudocode discussed to implement phrase extraction from the alignments produced by the IBM 1 model. I have tried experimenting with the length of the phrases extracted.

2.3 Decoding

After converting the phrases extracted from above to a WFST, the provided decoding module was used as is to calculate the BLEU scores.