CIS 526 Homework 4: Reranking

James Yang

March 27, 2015

1 Introduction

The quality of a machine translation model can be evaluated using a variety of metrics, most popular being the BLEU metric. However, it is not uncommon for several translation models to produce overall mediocre results, but individually produce decent translations under certain circumstances. As a result, it is necessarily advantageous to train a discriminative model for evaluating the best of multiple translations of a given sentence using a variety of features that these models have in common. In this homework, we were given a document for which each sentence has 100 different machine translated sentences. We were tasked with choosing the best of these translations at the sentence level to produce a higher BLEU score of a given corpus. In this document, we discuss the gradient ascent method used to train the discriminative model and the traversed feature space.

2 Gradient Ascent

Our goal in gradient ascent is to find the parameters $\theta \in \mathbb{R}^n$ that produce the best translation $\hat{e}$ within a set of translations $E(f)$ for a particular sentence f with a set of features $h(e, f) \in \mathbb{R}^n$ such that

$$\hat{e} = \operatorname*{argmax}_{e \in E(f)} (\theta \cdot h(e, f))$$

For each f_i in the document, we evaluate $\hat{e}_i$ to obtain the best translation set $\hat{E} \mid \hat{e}_i \in \hat{E}$, where the quality is obtained by the BLEU score, $\text{BLEU}(\hat{E}, \theta)$ which restructures our problem to

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \left(\text{BLEU}(\hat{E}, \theta)\right)$$

Based on this problem, we iteratively search for the best set of parameters θ with a learning rate η

$$\theta_{k+1} = \theta_k + \eta \nabla \frac{\partial \text{BLEU}(\hat{E}, \theta_k)}{\partial \theta_k}$$

We are given that $\text{BLEU}(\hat{E}, \theta)$ is an unknown function that, while is known to have a global maximum, has many local maxima. As a result, a typical gradient ascent method is abandoned, and instead we take a carpet bomb approach where we evaluate the BLEU score at a given θ_k as well as around a hypercube of length r_k all θ_k where θ_k is the center of that hypercube. The parameters that give the best result is assigned as the next set of parameters θ_{k+1}. If no greater value is found, we shrink the neighborhood by evaluating a hypercube of length $r_{k+1} = 0.5r_k$ around that same $theta_k$. This is repeated until convergence or maximum number of resolution change.

3 Feature Space

The original feature space provided in the default code consists of the log probabilities of the language model $p(f)$, translation model $p(e|f)$, and lexical translation model $p_{lex}(e|f)$. To this feature space I added the length of the translated sentence $|e|$. Based on improved results, the BLEU score tended to favor longer translations which makes sense as the BLEU score evaluates the number of matched n-grams, and the probability of matching more n-grams increases with more words.

4 Results

Table 1 shows the results of the gradient ascent as well as the difference obtained by adding a meaningful feature.

Parameters $\{p(f), p(e\|f), p_{lex}(e\|f), \|e\|\}$	**BLEU Score**
{-1, -0.5, -0.5, 0}	0.2735
{-5.625, -3.125, -7.125, 0}	0.28325
{-7.7207, -7.2207, -5.3457, 7.748}	0.29246

Table 1: Results on sample data.

We can see that the gradient ascent method did indeed help provide a better score. Not only that, but adding another feature also helped the reranking. It is assumed that with more meaningful features, good translations can actually be linearly separated from the rest of the translation set and result in higher scores. It is also assumed that with a smoothed BLEU score, better results can be found locally as a smoothed BLEU score may remove many local maxima.