Improving Sequence-to-Sequence with Adaptive Beam Search

Assignment 2

Yu-Hsiang Lin Shuxin Lin Hai Pham
Language Technologies Institute
Carnegie Mellon University
{*yuhsianl*, *shuxinl*, *htpham*}@andrew.cmu.edu

Abstract

We plan to apply new techniques such as combining dynamical beam size with trainable beam search to boost up the training and test-time decoding of the sequence-to-sequence model. We review the related approaches, and report our work in replicating state-of-the-art results.

1 Introduction

Since its invention, sequence-to-sequence (Seq2Seq) model (**?**) has been a go-to model for many translation-related tasks, especially since the advent of attention model (**??**). Despite its great successes in many domains, how to train and decode seq2seq model is still an open problem because of the drawback of traditional maximum likelihood training which is, most of the cases, unable to find the maximum-a-posteriori of a to-be-decoded single sentence over the whole corpus.

Amongst many heuristic approaches to remediate that problem, *greedy search* and *beam search* are probably the most popular. While greedy search is known for its lightweight, elegant characteristics, beam search is generally better in practice by considering not only the best-scored word at each time step but maintaining a window of best words. In this project, we will be addressing the disadvantages of previous approaches for seq2seq using beam search and proposing an improvement for it in training and decoding phases. We also present our results on the Name Entity Recognition task.

Table 1: Comparison of the F-score between our experiment and (**?**) at the NER task. Fixed attention is used, and beam size is 3.

	Greedy	Beam 3	Beam 3 Adaptive	Beam 6	Beam 6 Adaptive	Beam 9	Beam 9 Adaptive	Soft Beam
F-score	58.09	57.69	57.71	57.76	57.71	57.76	57.71	
Total beam #	48,571	145,713	92,727	291,426	126,759	437,139	182,785	
Avg. beam #	1	3	1.95	6	3.16	9	4.86	
Time (sec)	22	76	61	132	73	178	92	
Goyal F-score	54.92	51.34						56.38

Table 2: Comparison of the F-score between our experiment and (**?**) at the NER task. Fixed attention is used, and beam size is 3.

	RL Beam 3	RL Beam 6	RL Beam 9
F-score	57.66	57.61	57.52
Avg. beam #	1.17	2.00	3.06