NLA Project Scope - Hybrid MT

Souvik Banerjee (20171094)
Zubair Abid (20171076)

14th March 2020

1 Project Overview

In this project, we will build a Hybrid Machine Translation system using a Syntax-based SMT system using Joshua, and an NMT system: Neural Machine Translation with Universal Visual Representation, ICLR 2020. We compare the results against a baseline SMT system and a baseline NMT system.

The language pair used is English-French, with English-Bengali to be explored if we get the time. We chose to not continue with an English-[Indian Language] pair to begin with due to insufficient data.

2 Broad Project Plan

2.1 Tasks

The project comprises two baseline systems (one SMT, one NMT), two independent models, and a hybrid system of the two. Comparison will be done across all, with BLEU scores as the primary metric.

2.2 Dataset

We will use the same dataset used in the NMT with Universal Visual Representation paper: the WMT'14 EN-FR translation task.

3 Implementation

3.1 Baseline Models

- **Statistical Machine Translation:** We will use the Moses Statistical MT System for the SMT Baseline.
- **Neural Machine Translation:** We will use an implementation of Neural Machine Translation by Jointly Learning to Align and Translate, ICLR 2015.

3.2 Syntax-based Statistical Machine Translation

We will use the Joshua pre-trained language models to estimate performance, and make further improvements if found necessary.

3.3 Neural Machine Translation with Universal Visual Representation

We will implement Neural Machine Translation with Universal Visual Representation. The paper presents a universal visual representation learned over a monolingual corpora with image annotations, thus theoretically overcoming the lack of large-scale bilingual sentence-image pairs.

3.4 Hybrid System

The basic idea for the hybrid system is to use the best of both worlds. We plan on incorporating both SMT and NMT in two ways. First method involves going through one epoch of NMT and then use the probability distribution from this to eliminate least probable cases in SMT, this results in the decoder working much faster. Second method involves using the probability translation table of SMT and keeping them as base probability for NMT thus cutting down on epochs.

3.5 Further Development

If we have time, we will try EN-BN translation.

4 Timeline

- **First Evaluation:** Implementation of the baseline model
- **Second Evaluation:** Implementation of the two independent models
- **Final Evaluation:** Implementation of the hybrid model, performance comparison