Machine Translation System for Low Resource Language Pairs

Introduction

Task of Machine Translation is to learn translation between languages. This work describe our efforts to develop machine translation systems for low resource language pairs (Bhojpuri, Magahi, Sindhi <=> English).

MT systems are usually build using rules, data driven or hybrid of both. In this work we look at data driven methods. Data driven MT system uses parallel sentences (i.e, x^{th} sentences in two languages show same meaning).

For the data driven system to learn translation, it requires sufficient amount of parallel text (bi-text), which is not always easy to get. Scarcity of parallel text can hinder data driven systems ability to give decent translations . This work outlines preprocessing , configuration of the two data driven methods and their result.

Data Set of Languages

Languages like Bhojpuri, Sindhi and Maghai are primarily spoken in northern India by around 50 million, 1.6 million, 12 million people respectively[2]. These languages are resources scarce, which obstructs us from building data driven models for language pairs involving these languages, Parallel and monolingual corpora for Bhojpuri, Magahi and Sindhi received from LoResMT shared task[1]. Monolingual data for English was taken from web[3]. We included training data to the monolingual corpus of each language for decent language model. Statistics of parallel and monolingual text used is given in table 1 and 2.

Language Pair	Train	Dev	Test
English - Bhojpuri	28999	500	250
English - Maghai	3710	500	250
English - Sindhi	29014	500	250

Table 1: Number of parallel sentences

Language	Number of Sentences
Bhojpuri	78999
Magahi	19027
Sindhi	102345
English	2410767

Table 2: We concatenate training data with monolingual data

Language Pair L1 - L2	Number of Unique Words			
	mc>=2		mc>=1	
	L1	L2	L1	L2
Eng - Bho	6710	8790	12684	19754
Eng - Mag	2946	3355	5650	6504
Eng - Sin	6726	7651	12127	15689

Table 3: Number of Unique words in training data for language pairs (Eng-English, Bho-Bhojpuri, Mag-Magahi, Sin-Sindhi), with minimum count (mc) >=2 and >=1 .

[1] The 2nd Workshop on Technologies for MT of Low Resource Languages (LoResMT 2019)
[2] Language - Census of India
[3] Leipzig Corpora Collection Download Page
[4] Indic NLP Library - Resources and tools for Indian language Natural Language Processing
[5] Moses, the Statistical machine translation system
[6] Nematus - Open-Source Neural Machine Translation in Tensorflow

System Description

Statistical and Neural Machine translation systems are created using Datasets. We use IndicNLP Toolkit4 to tokenize Bhojpuri, Maghai and Sindhi (train, dev, test and monolingual) as a first step. For English we utilize default Moses toolkit5 tokenizer to obtain clean tokenized text. We make use of Nematus toolkit6 to carry out our NN based experiments for all language pairs.

In Table 3, Columns show total number of unique words with minimum count (mc) 2 and 1 in training text for respective language pairs (L1-L2). One can observe that there is a significant increase in unique count between mc>=2 and mc>=1. Hence, vocabulary size increases significantly which affects learning due in low resource settings (because almost half of the vocab has frequency 1). Therefore, we explore Byte Pair Encoding (BPE) to handle rare words effectively.

Following are hyper-parameters we use in our NMT systems and rest were default as mentioned in Nematus, BPE Merge Operations: 5000, Hidden Layer Dimension of LSTM: 200, Loss: cross entropy, Optimizer: Adam, Beam Size (During Training): 4, Beam Size (During Testing): 10 , Size of Embedding Layer for NMT1: 50 , Size of Embedding Layer for NMT2: 200.

We make use of Moses toolkit5 for this paradigm. We also use GIZA++ to find alignments between parallel text and grow-diag-final-and method to extract aligned phrases. We utilize KenLM to train a trigram model with kneser ney smoothing on monolingual corpus of all languages and MERT is used for tuning the trained models (named as SMT in results).

From the experiments, We observe that SMT is consistently outperforming NMT in low resource settings.

Language (X to English)	Bhojpuri			Magahi			Sindhi		
Model	NMT1	NMT2	SMT	NMT1	NMT2	SMT	NMT1	NMT2	SMT
BLEU	10.12	12.09	17.03	1.86	3.03	9.71	19.11	26.68	31.32
Language (English to X)	**Bhojpuri**			**Magahi**			**Sindhi**		
Model	NMT1	NMT2	SMT	NMT1	NMT2	SMT	NMT1	NMT2	SMT
BLEU	6.19	10.5	10.69	1.63	1.83	9.37	17.43	25.17	37.58

Table 4: Performace of translation systems in terms of BLEU score

Publication:

Yadav, Saumitra, Vandan Mujadia, and Manish Shrivastava. "A3-108 Machine Translation System for LoResMT 2019." *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*. 2019.

Authors: Saumitra Yadav, Vandan Mujadia, Manish Shrivastava, Research Center Name: Machine Translation and NLP Lab, LTRC