INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

R&D SHOWCASE 2020

Linguistically Informed Hindi-English Neural Machine Translation

ABSTRACT

Hindi-English Machine Translation is a challenging problem, owing to multiple factors including the morphological complexity and relatively free word order of Hindi, in addition to the lack of sufficient parallel training data. We propose a method to employ additional linguistic knowledge which is encoded by different phenomena depicted by Hindi to reduce data sparsity. We generalize the embedding layer of the state-of-the-art Transformer model to incorporate linguistic features like POS tag, lemma and morph features.

We compare the results obtained on incorporating this knowledge with the baseline systems and demonstrate significant performance improvement. We observe that although the Transformer NMT models have a strong efficacy to learn language constructs, the usage of specific features further help in improving the performance.

Adding Linguistic Input Features

Let $E \in \mathbb{R}^{m\times K}$ be the word embedding matrix for the standard Transformer encoder with no input features where m is the word embedding size and K is the vocabulary size of the source language. Therefore, the m-dimensional word embedding $e(x_i)$ of the token x_i (one-hot encoded representation i.e. 1-of-K vector) in the input sequence $x=(x_1,x_2,...,x_n)$ can be written as

$$e(x_i) = Ex_i$$

We generalize this embedding layer to some arbitrary number of features |F| as

$${e'}(x_i) = \text{concat}(E_{j}x_{ij}) \; \forall \; |F|$$

where $E_j \in {\mathbb{R}}^{m_j \times K_{j}}$ are the feature embedding matrices with m_j as the feature embedding size and K_j as the vocabulary size of the j^{th} feature.

Dataset

Dataset	Sentences	Tokens
IITB Train	1,528,631	21.5M / 20.3M
IITB Test	2,507	62.3k / 55.8k
IITB Dev	520	9.7k / 10.3k

The embedding layer size of the word or subword feature is set to bring the total size to 512.

Features	Emmbedding Sizes	
	all	single
Subword tags (IOB tagging)	6	5
Pos tags	10	10
Morph Features	20	20
Lemma	100	150
Word or subword	*	*

Results

System (Word Based)	BLEU
Word baseline	17.13
POS tags	17.51 (+0.38)
Lemma	**17.65 (+0.52)**
Morph features	17.44 (+0.31)
All features	17.32 (+0.19)

System (Subword based)	BLEU
Subword baseline	18.47
IOB tags	18.64 (+0.17)
POS tags	19.11 (+0.64)
Lemma	17.99 (-0.48)
Morph features	19.02 (+0.55)
IOB, POS tags and Morph features	**19.21 (+0.74)**
All features	18.34 (-0.13)

Authors: Vikrant Goyal, Pruthwik Mishra and Dipti Misra Sharma   *Research Center Name: LTRC*