Machine Translation using Deep Learning: A Case Study for Indian languages Hindi and Odia

Shantipriya Parida
Postdoc@UFAL

Supervisor
Dr. Ondřej Bojar

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Contents

- **Machine Translation**
- Deep Learning
- Neural Machine Translation
- Case Study
  - Translating Short Segments with NMT: A Case Study in English-to-Hindi
  - CUNI NMT System for WAT 2018 Translation Task
  - OdiEnCorp: Odia-English and Odia Only Corpus for Machine Translation

Introduction

Automatic conversion of text/speech from one natural language to another.

Source Language →  → **Target Language**

Google Translate

SYSTRAN beyond language

History

- 7 January 1954 the first public demonstration of a Russian-English MT system held in New York at the head office of IBM (system having just 250 words and translating just 49 Russian sentences into English.
- The Cold War system producing rough translation of Russian scientific journals in order to intercept secret information.
- The early 70s the Russian-English project called SYSTRAN - an attempt to translate a vast body of terminology connected with the military.

Why Machine Translation?

Full Translation
- Domain specific, e.g., Weather reports

Machine-aided Translation
- Requires post-editing

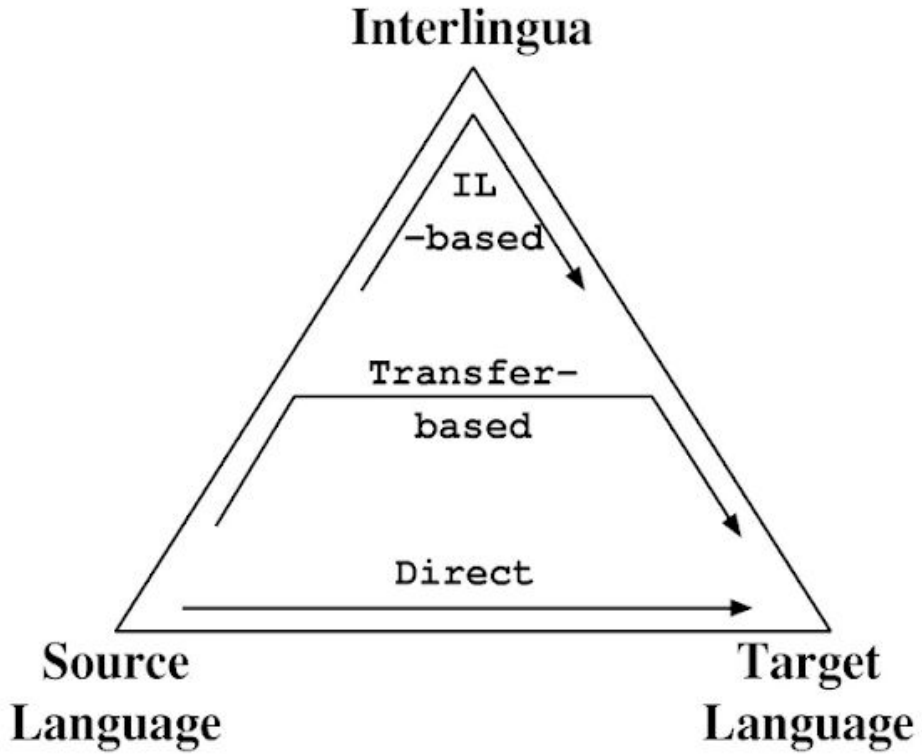
Cross-lingual NLP applications
- Cross-language information retrieval
- Cross-language Summarization

Testing grounds
- Extrinsic evaluation of NLP tools, e.g., parsers, pos taggers, tokenizers, etc.

Machine Translation Approaches

Grammar-based
- Interlingua-based
- Transfer-based

Direct
- Example-based
- Statistical
- Neural



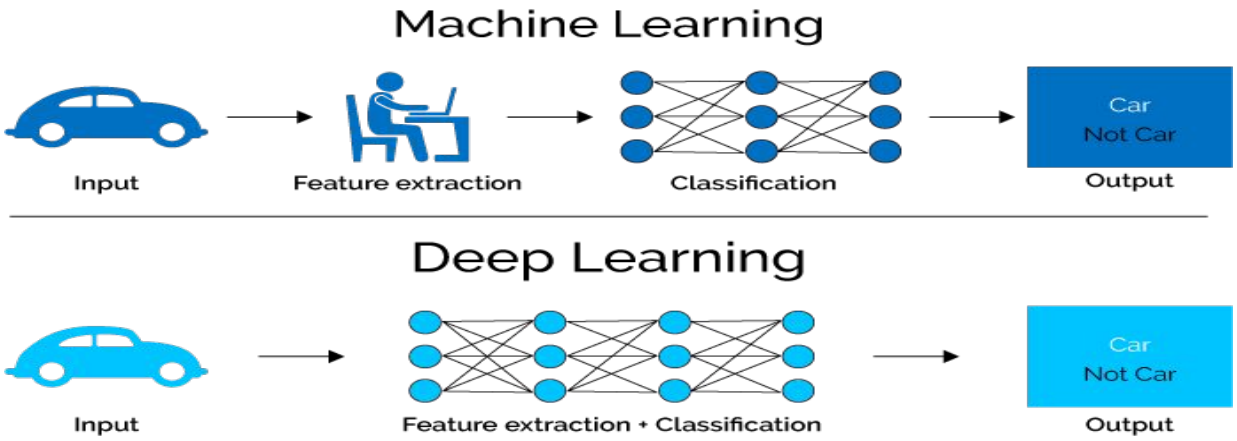
Modified from Vogel

Contents

- Machine Translation
- **Deep Learning**
- Neural Machine Translation
- Case Study
  - Translating Short Segments with NMT: A Case Study in English-to-Hindi
  - CUNI NMT System for WAT 2018 Translation Task
  - OdiEnCorp: Odia-English and Odia Only Corpus for Machine Translation

What is Deep Learning (DL) ?

- A machine learning subfield of learning **representations** of data.
- Exceptionally effective at **learning patterns**.
- Deep learning algorithms attempt to learn (multiple levels of) representation by using a **hierarchy of multiple layers**
- If you provide the system **tons of information**, it learns to respond in useful ways.



https://www.xenonstack.com/blog/static/public/uploads/media/machine-learning-vs-deep-learning.png

Contents

- Machine Translation
- Deep Learning
- **Neural Machine Translation**
- Case Study
  - Translating Short Segments with NMT: A Case Study in English-to-Hindi
  - CUNI NMT System for WAT 2018 Translation Task
  - OdiEnCorp: Odia-English and Odia Only Corpus for Machine Translation

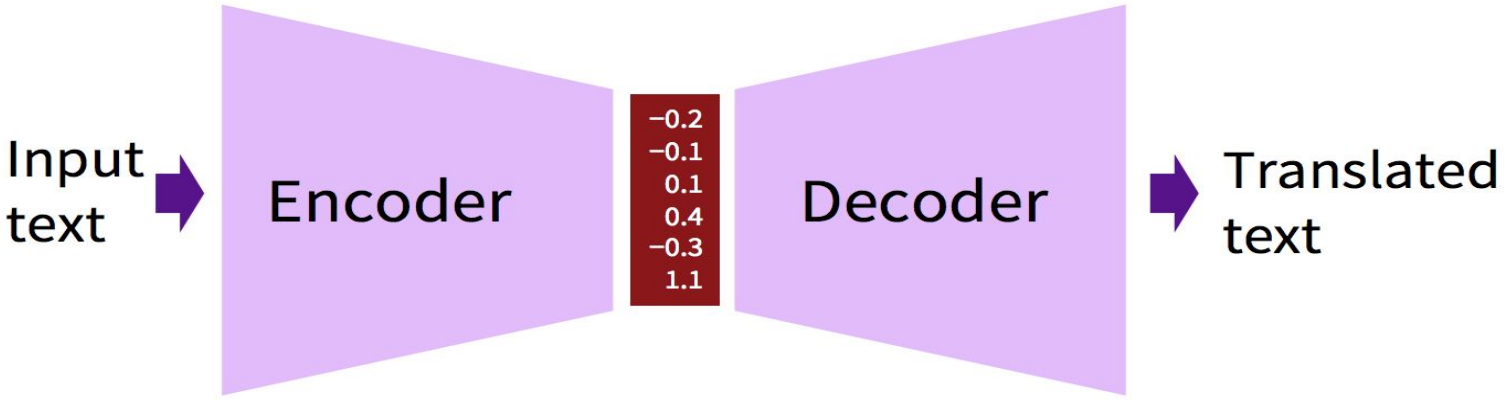
What is NMT ?

“Neural Machine Translation (NMT) is the approach of modeling the entire MT process via one big artificial neural network.”

(Manning, 2016)

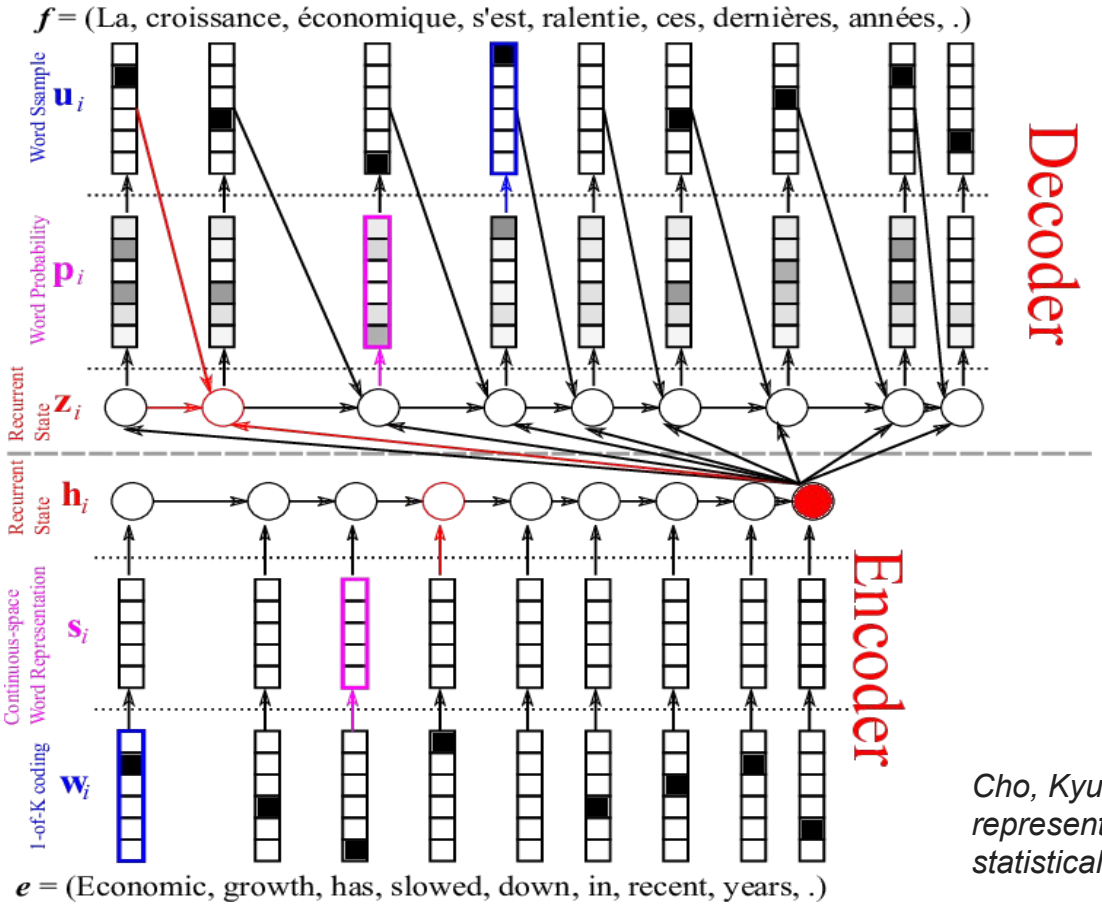
Neural Machine Translation

- Modeling the machine translation using neural networks
- Encoder for convert the input to a compact continuous representation.
- Decoder for language generation in target language



https://nlp.stanford.edu/projects/nmt/Luong-Cho-Manning-NMT-ACL2016-v4.pdf

Sequence 2 Sequence or Encoder-Decoder Model

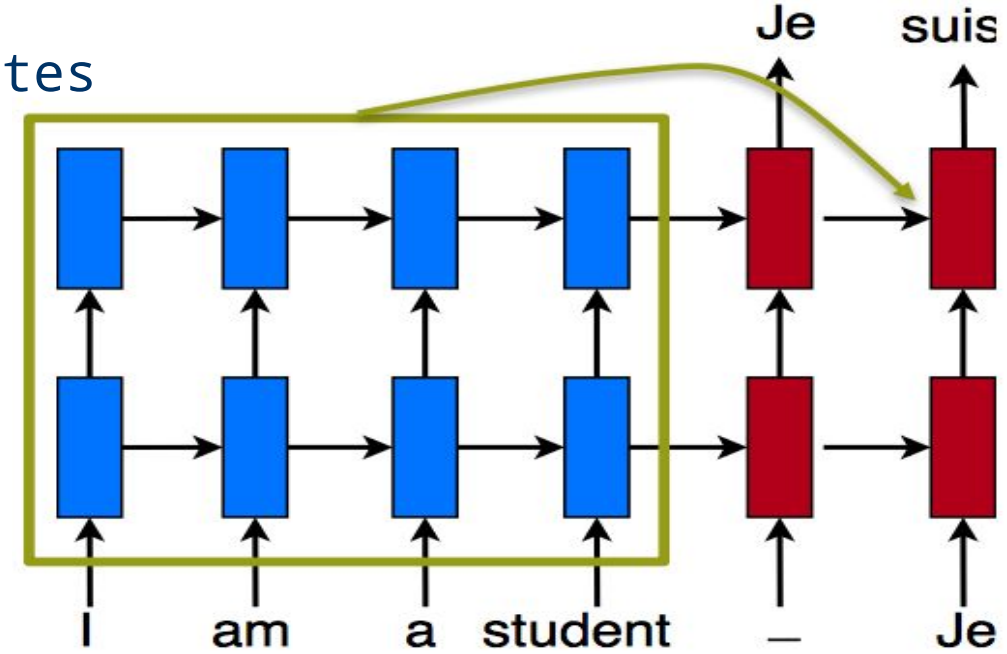


Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." EMNLP 2014

Attention Mechanism

Main idea: retrieve as needed

Pool of source states



Bahdanau D. et al. "Neural machine translation by jointly learning to align and translate." ICLR (2015)

Contents

- Machine Translation
- Deep Learning
- Neural Machine Translation
- **Case Study**
  - Translating Short Segments with NMT: A Case Study in English-to-Hindi
  - CUNI NMT System for WAT 2018 Translation Task
  - OdiEnCorp: Odia-English and Odia Only Corpus for Machine Translation

Translating Short Segments with NMT: A Case Study in English-to-Hindi

Visual Genome

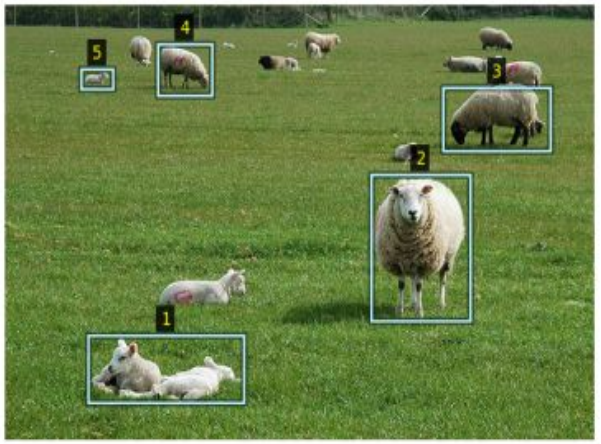
- Dataset of images, captions and relations potentially useful for many text and image processing applications.

- 108k images with 5.4M short captions in English

Motivation

- The Hindi version of Visual Genome would allow researchers to study multi-modal NLP for the world's fourth most spoken language.

Context Disambiguates



Caption 1: Two lambs lying in the sun.
Hindi MT: दो भेड़ के बच्चे सूरज में झूठ बोल रहे हैं
Gloss: Two baby sheep are **telling lies** in the sun.

Selected surrounding captions:
2. Sheep standing in the grass
3. Sheep with black face and legs
4. Sheep eating grass
5. Lamb sitting in grass.

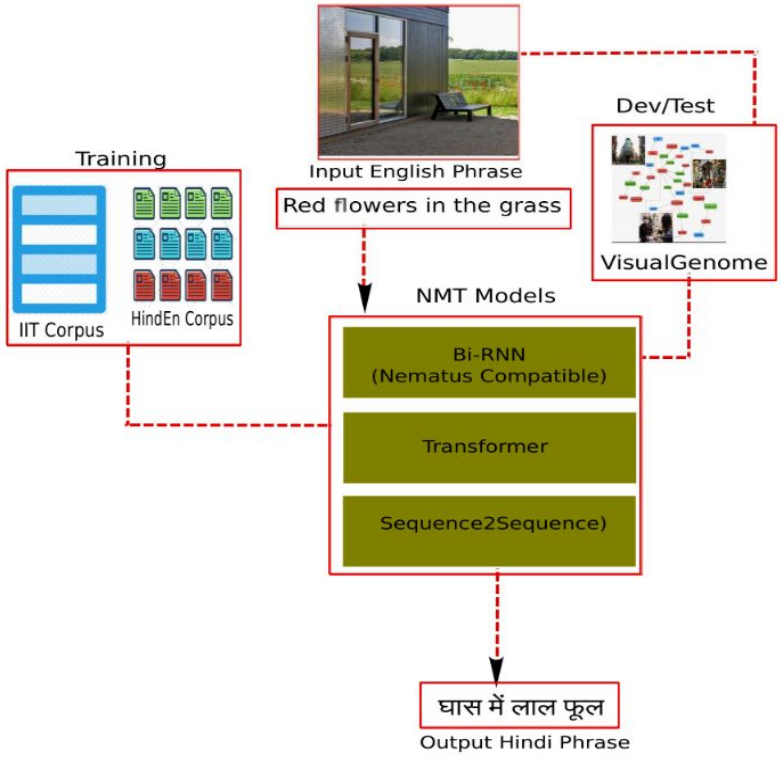
Translating Short Segments with NMT: A Case Study in English-to-Hindi

Experiments

- **Training and Evaluation Data:**

Dataset	#Sentences	#Tokens	
		En	Hi
Train (HindEnCorp)	273.9k	3.8M	5.6M
Train (IITB)	1492.8k	20.8M	31.4M
Dev (Visual Genome)	898	4519	6219
Test (Visual Genome)	1000	4909	6918

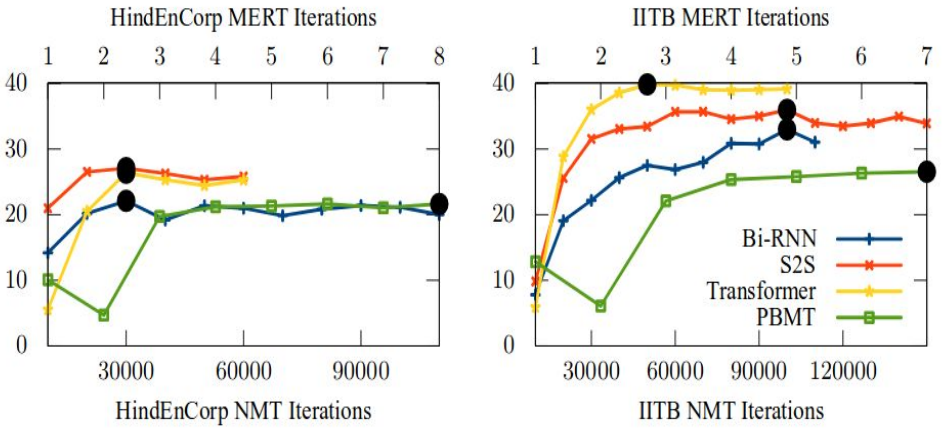
- **NMT Toolkit:** Marian (C++ implementation of several models).
- **MT Models** tested:
  - Marian's `nematus` Model (Bi-RNN), used shallow.
  - Marian's Sequence-to-Sequence (`s2s`) Model, used deep.
  - Marian's `transformer` Model.
- **Common Settings:** Tokenized with Moses tokenizer, joint BPE trained on HindEnCorp, 30k merge operations. Trained on four GeForce GTX 1080 Ti GPUs for 14 hours (best score).
- **Baseline:** Moses Phrase-Based MT with 5-gram language model.



Overall Experimental Setting

Translating Short Segments with NMT: A Case Study in English-to-Hindi

Results

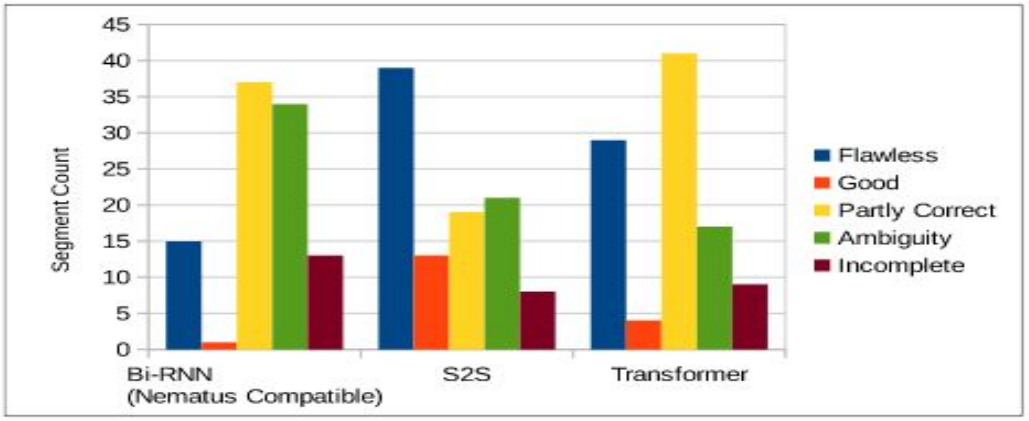


		Bi-RNN	S2S	Transf.	PBMT
HindEnCorp	BLEU	20.68	**26.45**	23.91	20.61
	chrF3	32.30	**39.52**	36.36	36.49
	nCDER	34.04	**40.91**	38.26	32.71
	nCharacTER	12.27	18.47	23.12	**29.05**
	nPER	41.76	49.05	47.01	**50.40**
	nTER	29.63	**35.70**	33.52	24.78
IITB Corpus	BLEU	31.78	32.81	**38.31**	25.06
	chrF3	42.63	44.50	**51.08**	43.09
	nCDER	44.49	44.91	**51.78**	37.54
	nCharacTER	-14.76	-47.00	25.07	**37.55**
	nPER	51.86	52.04	**59.60**	55.17
	nTER	40.62	41.44	**49.05**	32.76

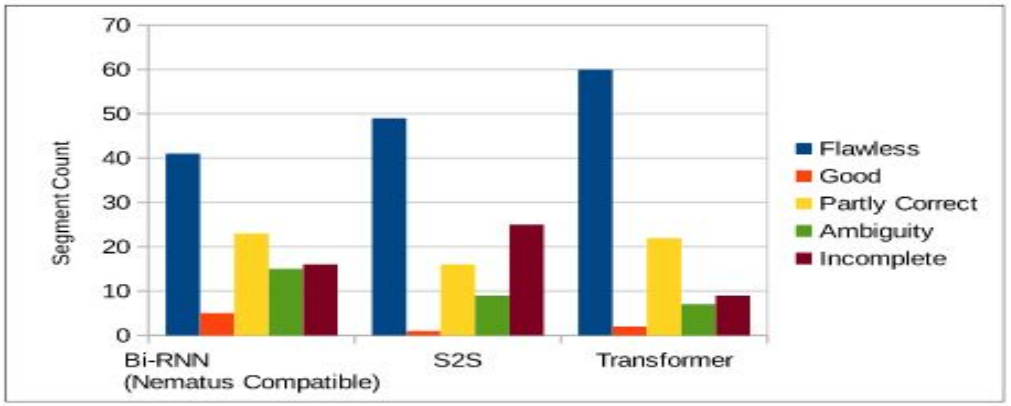
Learning curves in terms of BLEU of dev set.
Big black dots indicate which iteration was used for test set translation and evaluation

Results on the test set multiplied by 100.
Best model according to each automatic metric in bold.

Translating Short Segments with NMT: A Case Study in English-to-Hindi



(a) HindEnCorp-trained models



(b) IITB-trained models

Manual evaluation summary.

Flawless:
A car on a street
सडक पर एक कार
Gloss: A car on a street
A white and yellow passenger car
एक सफेद और पीला यात ◌ृ री कार
Gloss: A white and yellow passenger car
White part of the chair
कुर ◌ृ सी का सफेद भाग
Gloss: White part of the Chair
Partly Correct:
A man wearing white shorts
एक आदमी सफेद शॉर ◌ृ ट पहनना
Gloss: A man put on white short
(output does not convey the intended meaning in the target language)

Sample segment translations and their manual classification

Contents

- Machine Translation
- Deep Learning
- Neural Machine Translation
- **Case Study**
  - Translating Short Segments with NMT: A Case Study in English-to-Hindi
  - CUNI NMT System for WAT 2018 Translation Task
  - OdiEnCorp: Odia-English and Odia Only Corpus for Machine Translation

CUNI NMT System for WAT 2018 Translation Task

- Charles University NMT system submission to Workshop on Asian Translation (WAT) 2018, Hongkong, China.
- Participated in English-to-Hindi and Hindi-to-English translation task.
- Used transfer learning or domain adaptation techniques.
- Used English-Czech parallel corpus as additional corpus.
- Used Transformer model as implemented in Tensor2Tensor.

CUNI NMT System for WAT 2018 Translation Task

Method Description

- Train a general model followed by a domain specific model. In our case actual language pair (EN-HI).
- One constraint that is a shared vocabulary between language pairs of parent and child.
- Start with the parent model EN-CS translation as long as improves results on dev set then switch to child parallel corpus without any hyper parameter modification.
- For outputs not translated into the target language (identified by “langdetect”), use other output models.

CUNI NMT System for WAT 2018 Translation Task

Dataset

Set	#Sentences	#Tokens EN	CS	HI
Train (EN-CS)	40.1M	563.4M	490.5M	-
Train (EN-HI)	1.4M	20.6M	-	22.1M
TrainBack (EN-HI)	8.8M	161M	-	167M
Dev (EN-HI)	520	10656	-	10174
Test (EN-HI)	2507	49394	-	57037

Statistics of our Data

CUNI NMT System for WAT 2018 Translation Task

WAT 2018 Official Results

Team	Task	System	BLEU	Human	Note
CUNI	EN-HI	S3	17.63	**77.00**	
competitor	EN-HI	ConvS2S	19.69	69.50	Used external data
CUNI	EN-HI	S2	**20.07**	60.00	
competitor	EN-HI	ConvS2S	16.77	50.50	
competitor	HI-EN	CovS2S	**20.63**	**72.25**	Used external data
CUNI	HI-EN	S6	17.80	67.25	

WAT 2018 Official Automatic and Manual Evaluation Results for IITB corpora.

- **S2**: Transformer big, transfer learning from EN-CS 1M steps. We have not used any parallel data for EN-HI, we only used the back-translated EN-HI data, beam=8; alpha=0.8; averaging of last 8 models; stopped after 700k steps.
- **S3:** Transformer big, transfer learning from EN-CS 1M steps, followed by only back translation EN-HI for 300k steps, followed by genuine EN-HI for 500k steps, beam=8; alpha=0.8; averaging of last 8 models.
- **S6:** Transformer big, transfer learning from CS-EN 1M steps, only genuine HI-EN, beam=8; alpha=0.8; averaging of last 8 models; stopped after 230k steps. This model used primarily in back-translation but we also submitted it to the HI-EN task.

Contents

- Machine Translation
- Deep Learning
- Neural Machine Translation
- **Case Study**
  - Translating Short Segments with NMT: A Case Study in English-to-Hindi
  - CUNI NMT System for WAT 2018 Translation Task
  - OdiEnCorp: Odia-English and Odia Only Corpus for Machine Translation

OdiEnCorp: Odia-English and Odia Only Corpus for Machine Translation

- Language corpora provide citizen with NLP technologies readily available in other countries and support in education and cultural needs.
- Odia, classical Indian language spoken by 50 million speakers in five neighbouring states (AP, MP, WB, Jharkhand, Chhattisgarh) and one neighbouring country (Bangladesh).
- Odia language lacks sizable online contents.
- Big need and potential for Odia machine translation.

OdiEnCorp: Odia-English and Odia Only Corpus for Machine Translation

Data Sources

Parallel	Monolingual
English-Odia Parallel Bible	Odia Wikipedia
Government of Odisha Official Portal	Odia e-Magazines
Odia Digital Library	

OdiEnCorp: Odia-English and Odia Only Corpus for Machine Translation

Data Processing

- Extraction of Plain Text and De-Duplication
- Text Normalization
- Sentence Segmentation
- Sentence Alignment
- Manual Processing (correcting sentence alignment)

OdiEnCorp: Odia-English and Odia Only Corpus for Machine Translation

Corpus

Source	Sentences (Parallel)	Token English	Token Odia
Eng-Odia Parallel Bible	29069	756861	640157
Odisha Government Portal	122	1044	930
Odisha Govt Home Department Portal	82	367	327
Odia Digital Library (Odia Bibhaba)	393	7524	6233
Odia Digital Library (Odia Virtual Academy)	31	453	378
Total	29697	766249	648025

English-Odia Parallel Corpus Details

OdiEnCorp: Odia-English and Odia Only Corpus for Machine Translation

Corpus details

Source	Paragraphs	Sentences	Token
Odia wiki	30468	102085	1320367
AmeOdia	10605	27334	265681
Aahwan	10940	19582	248378
Odiasahitya	4390	4999	93922
Odiagapa	15295	67546	712960
Total	71698	221546	2641308

Odia Monolingual Corpus Details

OdiEnCorp: Odia-English and Odia Only Corpus for Machine Translation

Baseline Machine Translation

Dataset	#Sentences	#Tokens EN	#Tokens OD
Train	27136	706567	604147
Dev	948	21912	19513
Test	1262	28488	24365

Statistics of OdiEnCorp Parallel Data

Statistical MT Setup: Moses phrase-based MT system with n-gram language model trained with standard MERT optimization towards BLEU. Alignment based on lower case and stemmed to first 4 characters only.

NMT Setup: Transformer model as implemented in Tensor2Tensor version 1.4.2. "Big Single GPU" configuration. Batch Size 2300, sentence length 100 word pieces, learning rate 0.2, warm up steps 32000, beam size 8 and alpha (length penalty) 1.0 for decoding.

OdiEnCorp: Odia-English and Odia Only Corpus for Machine Translation

Results

Corpus	Task	System	BLEU Dev	BLEU Test
OdiEnCorp	EN-OD	SMT	-	6.49
OdiEnCorp	OD-EN	SMT	-	12.72
OdiEnCorp	EN-OD	NMT	4.29	4.10
OdiEnCorp	OD-EN	NMT	9.35	8.60

Results for Baseline System

Summary

- Based on NMT models we used, Transformer model is best and faster compared with shallow and deep S2S models.
- Transfer learning or domain adaptation, and usage of synthetic data found effective under low resource conditions.
- Based on the experiment using OdiEnCorp, we found phrase based machine translation (PBMT) performs considerably better in small data settings.


Thank You