Odisha Corporate Foundation

Odias in Machine Learning

Natural Language Processing and Application

Dr. Shantipriya Parida

AI Machine Learning Summer School
June 18 - July 2

Agenda

- Overview
  - Language and Resource
- NLP Applications
  - Text Classification
  - Named Entity Recognition
  - Topic Modelling
  - Language Model
  - Text Summarization
  - Machine Translation
- Case Studies
  - Case Study 1 -Corpus Development
  - Case Study 2 - Machine Translation
  - Case Study 3 - Treebank
  - Case Study 4 - OdiaBert
  -
- Conclusion

Some Facts

BY THE NUMBERS

There are over **7,000** languages worldwide.

Only 23 languages account for more than half of the world's population.

At least half of the world's population is bilingual.

2,400 of the world's languages are currently in danger of becoming extinct.

Papua New Guinea has the most languages, at **840**

Many linguists believe that language originated around 100,000 BC.

Spanish is the 2nd most spoken language in the world.

The English language contains the most words, with over **250,000**

MORE FUN FACTS

The first language spoken in outer space was Russian.

Other than English, French is the only language that is taught in every country.

Learning a second language can improve the memory and slow the process of aging.

About one language becomes extinct every two weeks.

ABOUT THE ALPHABET

74 Cambodian has the longest alphabet with 74 characters.

The word "alphabet" is formed from the first two letters of the Greek alphabet - alpha and beta.

11 The Papuan language of Rotokas only has 11 letters in its alphabet.

CULTURAL FACTS

The Bible is the most translated book, followed by Pinocchio.

There are over 200 artificial languages created for books, movies, and TV shows.

The Pope tweets in nine languages, but his Spanish account has the most followers.

The culinary and ballet worlds use mostly French words and terms.

The first printed book was written in German.

The average person only uses a few hundred words in daily conversation.

Physical contact during a conversation is completely normal when speaking Spanish.

Cryptophasia is a language phenomenon that only twins can understand.

21 Twenty-one countries have Spanish as their official language.

300 Over 300 languages are spoken in London alone.

4000 Spanish contains about 4,000 Arabic words.

Source: https://takelessons.com/blog/language-facts-z14

Some Facts

LANGUAGE IN EUROPE

The language of "La Gomera" spoken off the coast of Spain consists entirely of whistles.

24 — *There are about 24 official languages spoken throughout the European Union.*

French is the main foreign language taught in the UK.

Italy has many regional dialects, but the Florentine dialect was chosen as the national language.

Basque, a language spoken in the Pyrenees mountains, has no relation to any other known language.

German is the most spoken language in Europe.

20,000

Over 20,000 new French words are created each year.

German words can have three genders: masculine, feminine, and neuter.

LANGUAGE IN THE AMERICAS

Argentina has a lot of Welsh speakers, due to settlers inhabiting the Patagonia mountains.

The United States has no "official language." Most people just assume it's English.

30% — *About 30% of English words come from French.*

Italian is a minority language in Brazil.

More than 1.5 million Americans are native French speakers.

Hawaiians have over 200 different words for "rain."

The U.S. has the second highest number of Spanish speakers, after Mexico.

LANGUAGE IN AFRICA

Botswana has a language that is made up of five primary "click" sounds.

South Africa has the most official languages with 11.

About ⅔ of all languages are from Africa and Asia combined.

Kinshasa, the capital of the Congo, is the world's second largest French speaking city.

LANGUAGE IN ASIA

People who speak Chinese use both sides of the brain; English only uses the left side.

Hindi didn't become the official language of India until 1965.

Japanese uses three different writing systems: Kanji, Katakana, and Hiragana.

In Indonesian, "air" means "water."

Mandarin Chinese is the most spoken language in the world.

你好

Source: https://takelessons.com/blog/language-facts-z14

Some Facts

- How many facts (from above slides) already you knew ?.
- Do you have any interesting facts about languages (e.g. Indian languages) to share ?.

Off the Coast of India, Another Language Dies

By Ishaan Tharoor | Wednesday, Feb. 17, 2010

Нравится 243 Tweet | Share | Read Later

On some days, Boa Sr would sit silently in the jungle surrounding her home on one of India's Andaman Islands and gaze up at the sky. According to researchers who looked on, birds perched above would descend to the ground and inspect her; in turn Boa Sr spoke to them in her native tongue, calling them her ancestors and her friends. Her speech was rich with words of the natural world, words of the forest and the sea that some linguists suspect date back tens of thousands of years to the first migrations of man. Boa Sr was the last person alive to know them. In early February, she passed away, leaving behind no surviving siblings or children. As she died, so too did the language of her people.

Alok Das/ SURVIVAL INTERNATIONAL / AFP

Boa Sr, the last speaker of Bo, one of the 10 Great Andamanese languages, on the Andaman and Nicobar Islands

source: http://content.time.com/time/world/article/0,8599,1964610,00.html

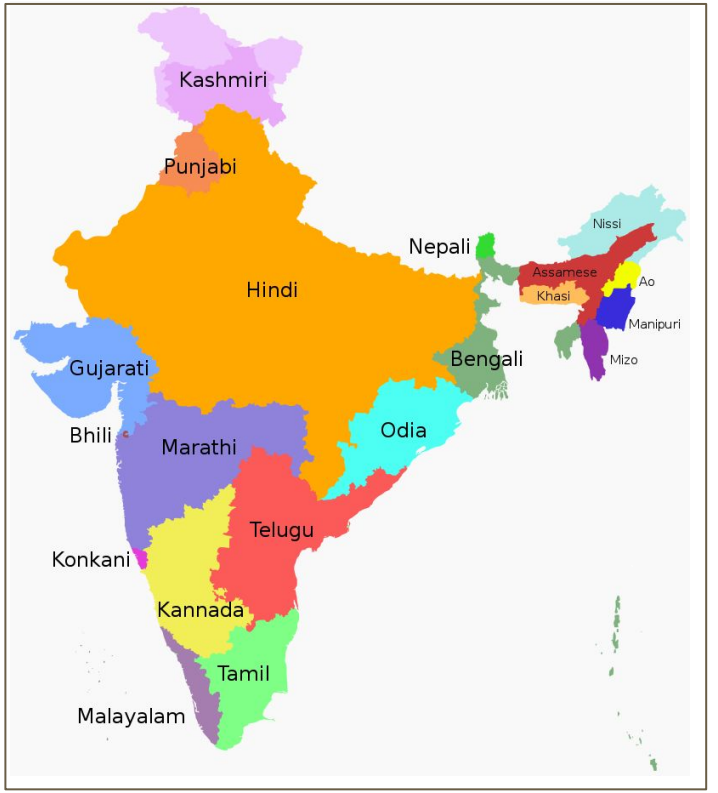
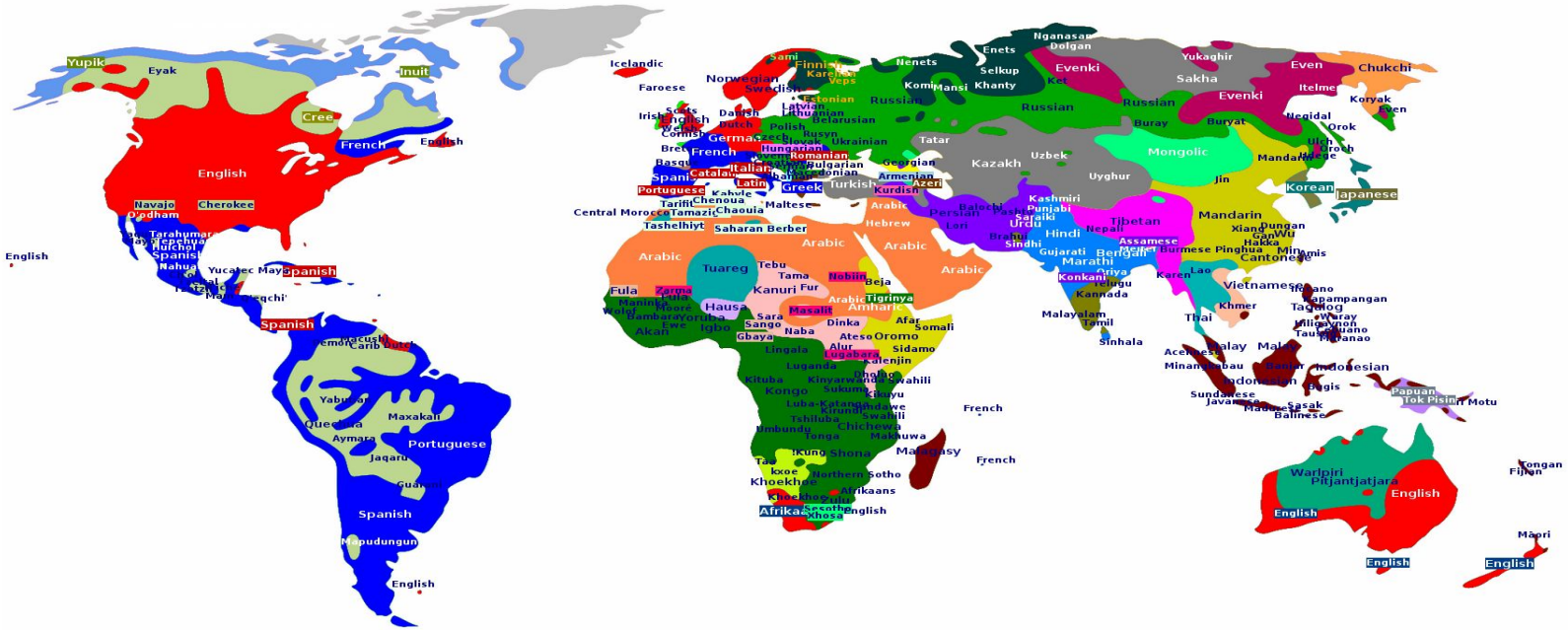


Image source: https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers_in_India#/media/File:Language_region_maps_of_India.svg

Multilinguality

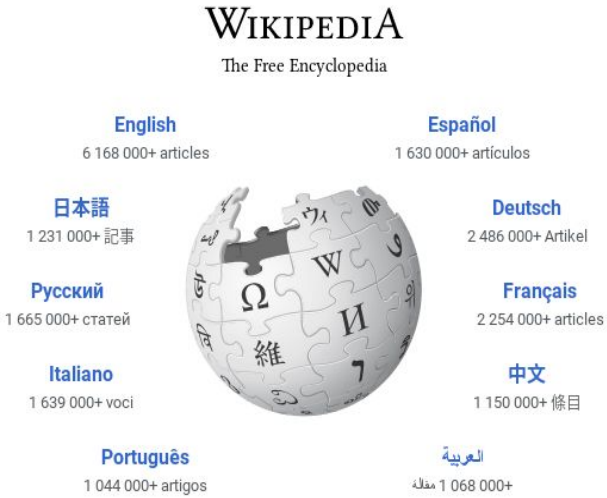
- The ethnologue.com website lists over 7000 languages in the world.



The geographical pattern of the major languages of the world.
Source: https://en.wikipedia.org/wiki/Template:Distribution_of_languages_in_the_world

Need for Language Resource

- Wikipedia has texts in 313 languages.
- Natural language technology development depends on large numbers of language resources (text / speech).
- Lack of language resources affects the development of natural language technologies.



Resource

- OdiEnCorp2.0 (Odia-English parallel corpus)
- OdiEnCorp 1.0 (Odia-English parallel and Odia monolingual corpus)
- Odia Treebank
- Hindi Visual Genome 1.1 (English to Hindi Multimodal dataset)
- Malayalam Visual Genome 1.1 (English to Malayalam Multimodal dataset)
- Bengali Visual Genome 1.0
- Hausa Visual Genome 1.0
- English->Hindi Machine Translation System
- Odia-NLP-Resource-Catalog

Supervised Autoencoder
https://github.com/idiap/sae_lang_detect

A Catalog for Odia Language NLP Resources

The purpose of this catalog is to provide a one-stop solution for the researchers looking for Odia NLP resources. This is a collective effort and any contribution to enriching Odia NLP resource are welcome. All contributors are listed on the CONTRIBUTOR list.

Table of Contents

- NLP Repositories
- Text Corpora
  - Parallel Translation Corpus
  - Monolingual Corpus
  - Lexical Resources
  - POS Tagged Corpus
  - Dialect Detection Corpus
- Models
  - Language Model
  - Word Embedding
  - Morphanalyzers
- Text Classification
- Libraries / Tools
- Speech Corpora
- Other Indian language NLP Resources

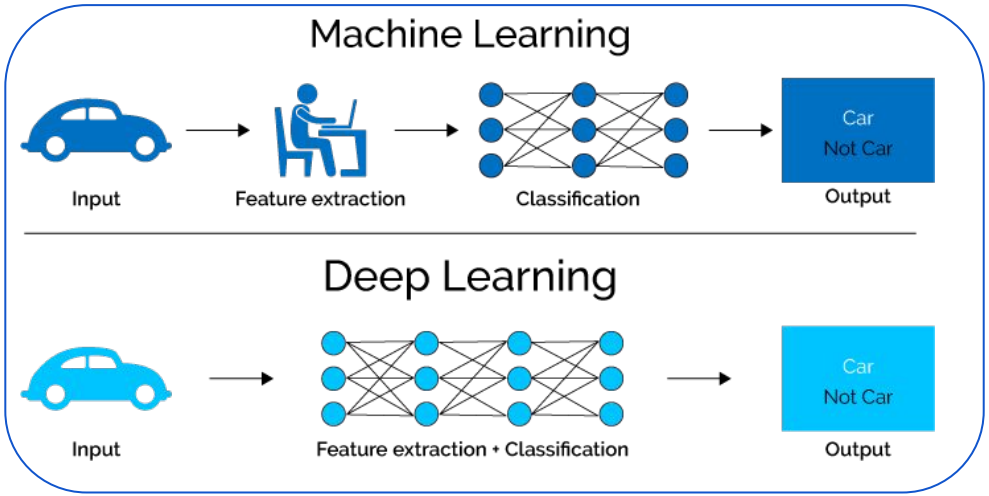
https://github.com/shantipriyap/Odia-NLP-Resource-Catalog

Overview

- Natural language processing (NLP) helps computers communicate with humans in their own language and scales other language-related tasks.
- NLP makes it possible for computers to read text, hear speech, interpret it, measure sentiment and determine which parts are important.

Deep Learning in NLP

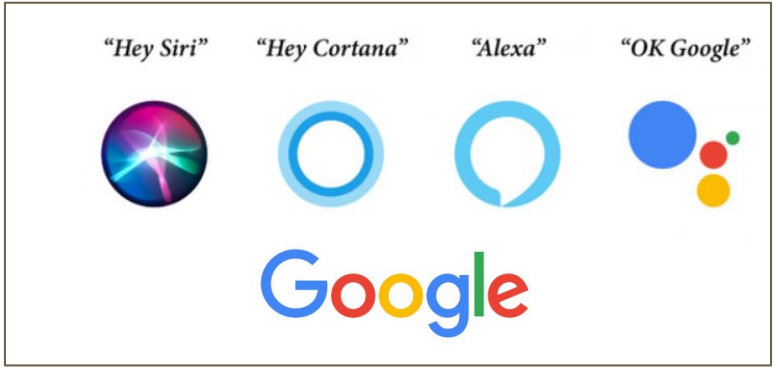
- A machine learning subfield of learning representations of data.
- Exceptionally effective at learning patterns.
- Deep learning algorithms attempt to learn (multiple levels of) representation by using a hierarchy of multiple layers.
- If you provide the system tons of information, it learns to respond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

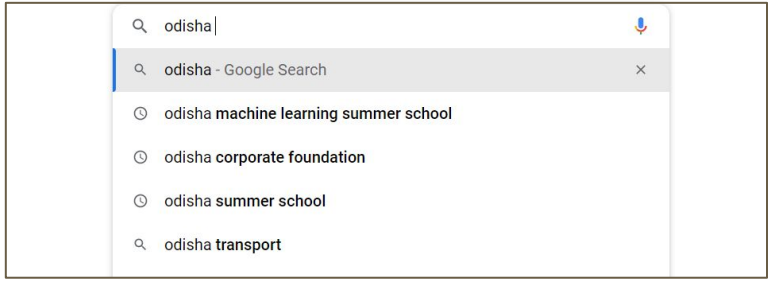


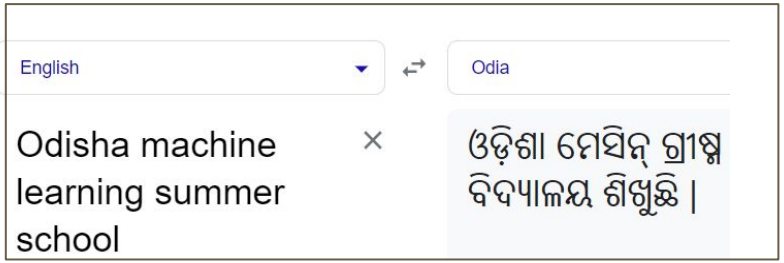
https://www.xenonstack.com/blog/static/public/uploads/media/machine-learning-vs-deep-learning.png

NLP in Daily Life

- Personal assistants: Siri, Cortana, and Google Assistant.
- Auto-complete: In search engines (*e.g.* Google).
- Spell checking: Almost everywhere, in your browser, your IDE (*e.g.* Visual Studio), desktop apps (*e.g.* Microsoft Word).
- Machine Translation: Google Translate.
- Chat bots.







Why NLP is difficult ?

- Natural language is highly ambiguous.
- Words can have several meanings and contextual information is necessary to correctly interpret sentences.
- Syntactic analysis (syntax) and semantic analysis (semantic) are the two primary techniques that lead to the understanding of natural language.

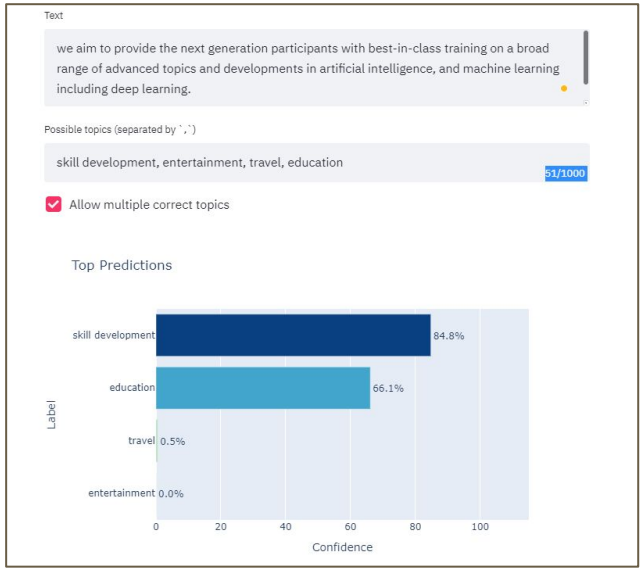
(a) Street sign advising of penalty.

(b) The penalty box is white lined.

Figure: An illustration of two meanings of the word “penalty” exemplified with two images

Text Classification

- Text classification is a task of NLP where the model needs to predict the classes of the text documents.
- In the traditional process, we are required to use a huge amount of labelled data to train the model, and also they can't predict using the unseen data.
- Adding zero-shot learning with text classification has taken NLP to the extreme. https://huggingface.co/zero-shot/
- Zero-shot text classification technique classify the text documents without using any single labelled data or without having seen any labelled text.



Zero-shot Text Classification (EN)

Zero-shot Text Classification (OD)

Named Entity Recognition

- Named entity recognition is a natural language processing technique that can automatically scan entire articles and pull out some fundamental entities in a text and classify them into predefined categories.
- Entities can be: Organizations, Quantities, Monetary values, Percentages, People's names, Company names, Geographic locations (Both physical and political), Product names, Dates and times, Amounts of money,Names of events.

Original Text

The program is planned for the summer holiday in Odisha (June-July 2022) and will be in virtual mode considering teaching by many international experts. The participants will be based on registration considering the eligibility and background of the participants. The maximum number of participants will be 100. The AI ML Summer School is a five days program and will be conducted on weekends (Saturday and Sunday) only with 3 lecturer sessions per day and each session will be 1.5 hours (1-hour Theory, 30 minutes Demo) except first and last sessions. However, students can execute the assigned mini-project on other days at their convenient time in groups.

Analysis Result

the summer holiday/DATE

Odisha/GPE

June-July 2022/DATE

100/CARDINAL

The AI ML Summer School/ORG

five days/DATE

Saturday/DATE

Sunday/DATE

3/CARDINAL

1.5 hours/TIME

1-hour/TIME

30 minutes/TIME

GPE: Geopolitical Entity (countries, cities, states), CARDINAL: number

Topic Modelling

- What is Topic Modelling ?.
  - Topic modeling is a statistical modeling approach to discover the abstract “topics” occurs in a collection of documents.
- Types of Topic Modeling
  - Unsupervised, and Semi-supervised
- . Application of Topic Modeling
  - text mining, text classification, machine learning, information retrieval, and recommendation engines.

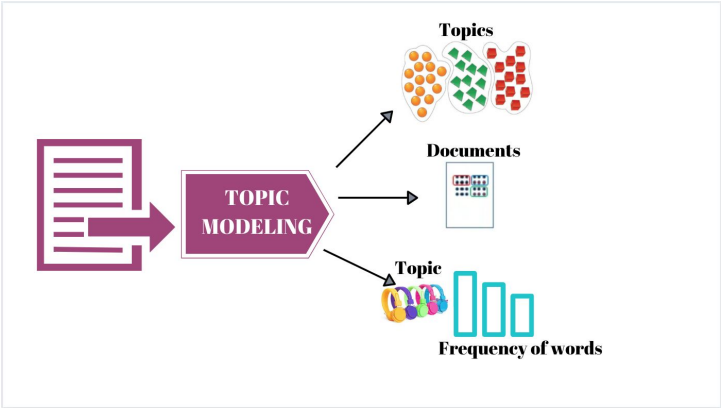
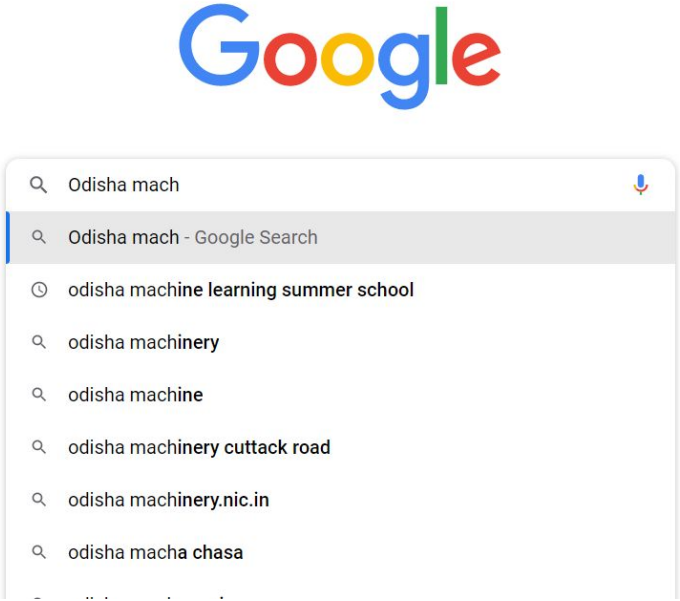


Fig: Topic Modeling

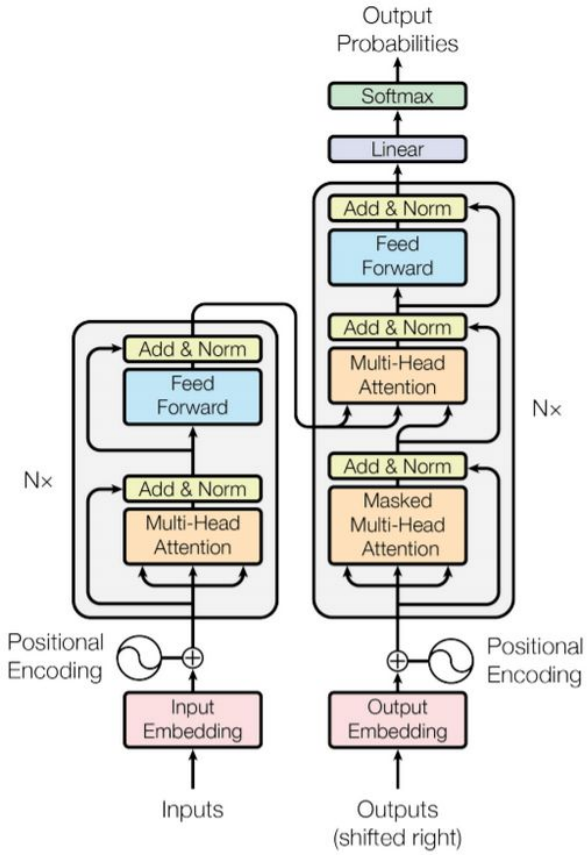
Language Model

- Language modeling is the task of predicting the next word in a sentence, given all previous words.
- It is an effective task for using unlabeled data to pretrain neural networks in NLP.
- Language models capture general aspects of the input text that is almost universally useful.



BERT

- BERT is an open-sourced NLP pre-training model developed by researchers at Google in 2018.
- BERT trains a language model that takes both the previous and next tokens into account when predicting.
- BERT uses the Transformer architecture for encoding sentences.
- BERT uses wordpieces (e.g. playing -> play + ##ing) instead of words.
- BERT is pre-trained on a large corpus of unlabelled text which includes the entire Wikipedia (that's about 2,500 million words) and a book corpus (800 million words).

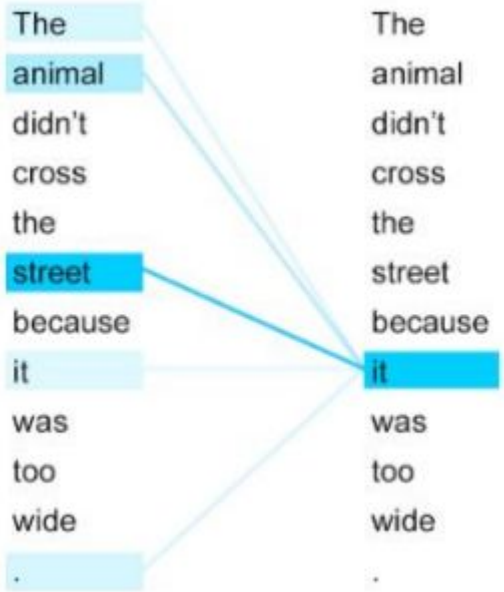


Transformer Model

Transformer Model

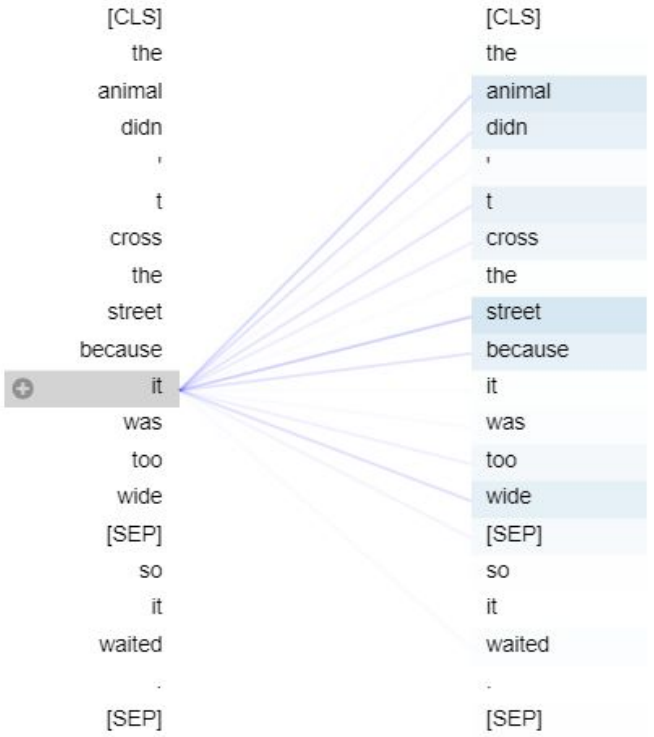
Self-Attention

- Can you figure out what the term “it” in this sentence refers to?
- Is it referring to the street or to the animal? It’s a simple question for us but not for an algorithm.
- Self-attention allows the model to look at the other words in the input sequence to get a better understanding of a certain word in the sequence.



BERT (Attention)

- In BERT, an attention mechanism lets each token from the input sequence (e.g. sentences made of word or subwords tokens) focus on any other token.
- The word “it” attends to every other token and seems to focus on “street” and “animal”.
- BERT uses 12 separate attention mechanism for each layer.



Visualization of attention values on layer 0 head #1, for the token “it”.

How BERT Works ?

Two strategies:

- Mask Language Model (MLM)
- Next Sentence Prediction (NSP)

Mask Language Model (MLM)

- BERT randomly masks words in the sentence and predicts them.
- Before feeding word sequences into BERT, 15% of the words in each sequence are replaced with a [MASK] token.
- The model then attempts to predict the original value of the masked words, based on the context provided by the other, non-masked, words in the sequence.

Input: The $[MASK]_1$ is not working. It's unable to $[MASK]_2$

Labels: $[MASK]_1$ = computer; $[MASK]_2$ = start.

How BERT Works ?

Next Sentence Prediction (NSP)

- In the training process, the model receives pairs of sentences as input and learns to predict if the second sentence in the pair is the subsequent sentence in the original document.
- Let's consider two sentences A and B, is B the actual next sentence that comes after A in the corpus, or just a random sentence? For example:

Sentence A = The computer is not working.

Sentence B = It's unable to start.

Label = IsNextSentence

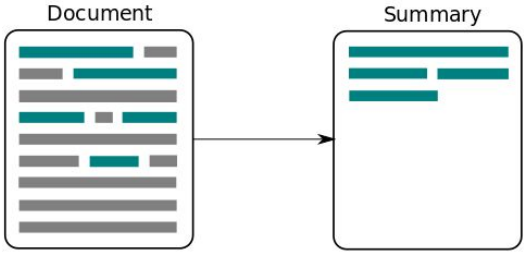
Sentence A = The computer is not working.

Sentence B = Coffee is very tasty.

Label = NotNextSentence

Text Summarization

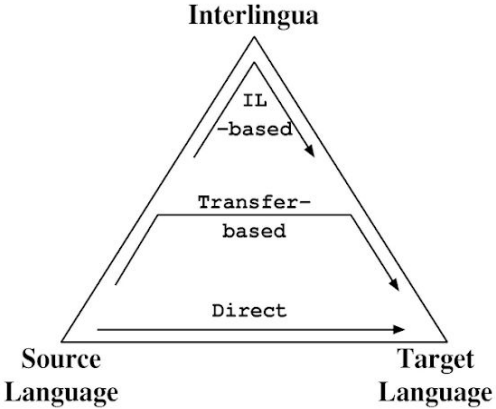
- Automatic text summarization aims to transform lengthy documents into shortened versions, something which could be difficult and costly to undertake if done manually.
- Two major approaches for automatic summarization are: extractive and abstractive.
- The extractive summarization approach produces summaries by choosing a subset of sentences in the original text.
- The abstract text summarization approach aims to shorten the long text into a human readable form that contains the most important fact from the original text



Machine Translation

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

- Machine translation approaches:
  - Grammar-based
    - Interlingua-based
    - Transfer-based
- Direct
  - Example-based
  - Statistical
  - Neural



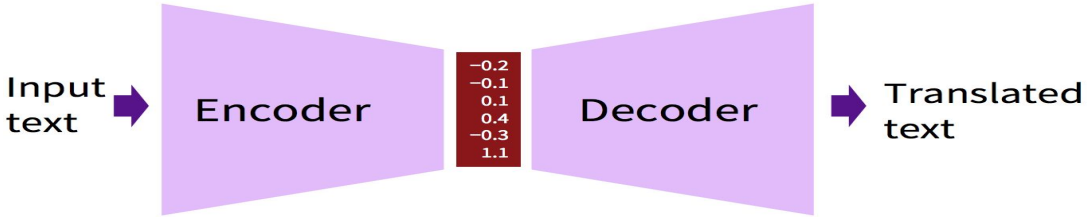
Google Translate

SYSTRAN
beyond language

Machine Translation

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.

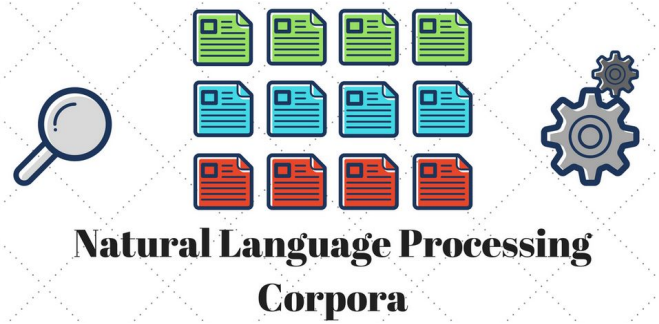


English to Odia

English (Source)	Odia (Translation)	Gloss/Remark
It is located on the bank of the River Sone which merges with River Ganges at Digha a few kilometers from Danapur.	ଏହା ଦାନପୁର ଠାରୁ କିଛି କିଲୋମିଟର ଦୂରରେ ଦିଘାଠାରେ ଗଙ୍ଗା ନଦୀ ସହ ମିଳିତ ହେଉଥିବା ସୋନ ନଦୀର କୂଳରେ ଅବସ୍ଥିତ ।	It is located on the bank of the river Sone which merges with river Ganges a few kilometer away from Danapur

Case 1 - Corpus Development

- Corpus (plural corpora) : A collection of linguistic data, either compiled as written texts or as a transcription of recorded speech.
- A corpus can be made up of everything from newspapers, novels, recipes and radio broadcasts to television shows, movies and tweets.
- In NLP, a corpus contains text and speech data that can be used to train AI and machine learning systems.
- Generally, the larger the size of a corpus, the better (prioritize quantity over quality).



Corpus - How to Build ?

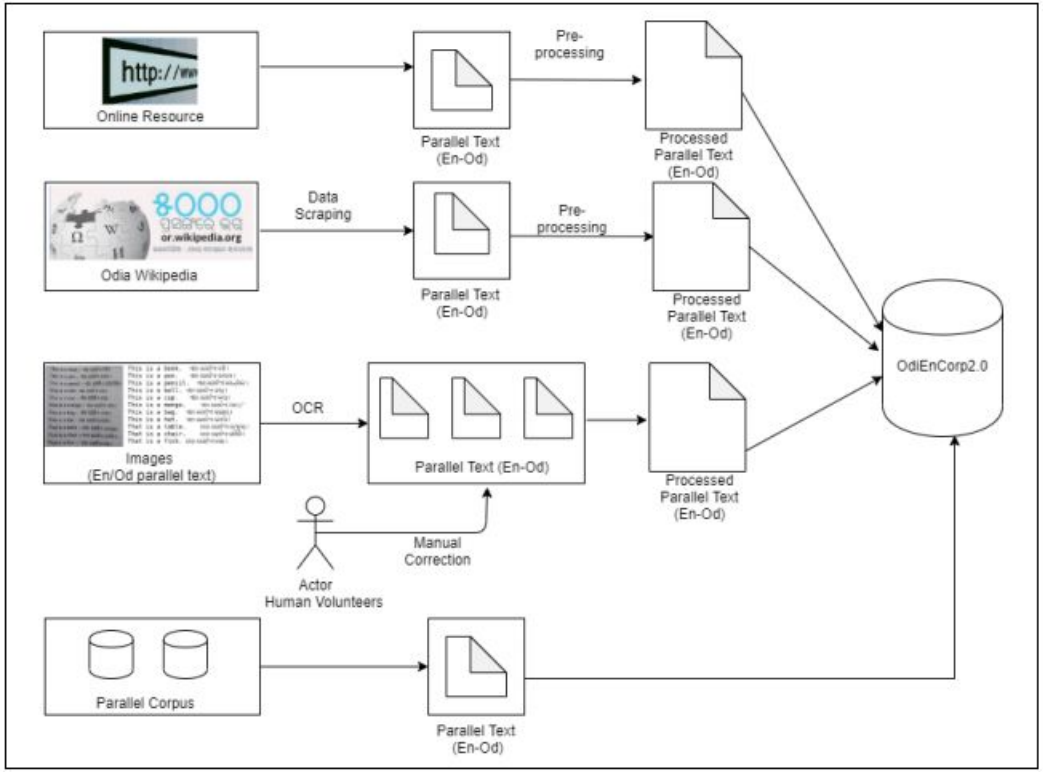
- Data Collection
  - Data type
    - Text/Image/Speech/Video
  - Identify source
    - Web, Social Media, Books, Recordings
  - Web scraping
    - Identify URLs (e.g. language, text, tags)
  - Bots
  - Optical Character Recognition (OCR)
  - Extract data
    - tools: Python, BeautifulSoup
- Data Processing
  - segmentation, alignment
    - Purnaviram, Hunalign
- Finalization and Release
  - Split train/dev/test set
  - Baseline
  - License
  - Release platform
  - Share/organize shared task
    - WMT, WAT, ICON, etc...

Image source:
https://medium.com/analytics-vidhya/web-scraping-and-coursera-8db6af45d83f

Image source: Image source:
https://medium.com/states-title/using-nlp-bert-to-improve-ocr-accuracy-385c98ae174c

Sample (OdiEnCorp)

- Data extracted from other online resources.
- Data extracted from Odia Wikipedia.
- Data extracted using Optical Character. Recognition (OCR).
- Data reused from existing corpora.



Sample (OdiEnCorp)

- Data Processing

- Extraction of plain text.
  - Python script to scrape plain text from HTML page.
- Manual processing.
  - Correction of noisy text extracted using OCR-based approach.
- Sentence segmentation.
  - Paragraph segmented into sentences based on English full stop (.) and Odia Danda (।) or Purnaviram.
- Sentence alignment.
  - Manual sentence alignment for Odia Wikipedia articles where text in two language are independent of each other.

Dataset	#Sentences	#Tokens EN	#Tokens OD
Train 2.0	69260	1340371	1164636
Dev 2.0	13429	157951	140384
Test 2.0	14163	185957	164532

Dataset Statistics.

Availability

OdiEnCorp 2.0 is available for research and non-commercial use under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, CC-BY-NC-SA at :

http://hdl.handle.net/11234/1-3211

Case 2 - OdiaTreebank

Universal Dependency Treebank for Odia Language

Shantipriya Parida[1], Kalyanamalini Sahoo[2], Atul Kr. Ojha[3], Saraswati Sahoo[4], Satya Ranjan Dash[5] and Bijayalaxmi Dash[6]

[1]Silo AI, Helsinki, Finland
[2]University of Lille, France
[3]Insight Centre for Data Analytics, DSI, NUI, Galway, Ireland
[4]Institute of Mathematics and Applications, India
[5]KIIT University, Bhubaneswar, India
[6]Ravenshaw University, Cuttack, India
shantipriya.parida@silo.ai, kalyanamalini.shabadi@univ-lille.fr, atulkumar.ojha@insight-centre.org,
sahoosaraswati455@gmail.com, sdashfca@kiit.ac.in, rudrabijayalaxmi@gmail.com

Abstract

This paper presents the first publicly available treebank of Odia, a morphologically rich low resource Indian language. The treebank contains approx. 1082 tokens (100 sentences) in Odia selected from "Samantar", the largest available parallel corpora collection for Indic languages. All the selected sentences are manually annotated following the "Universal Dependency (UD)" guidelines. The morphological analysis of the Odia treebank was performed using machine learning techniques. The Odia annotated treebank will enrich the Odia language resource and will help in building language technology tools for cross-lingual learning and typological research. We also build a preliminary Odia parser using a machine learning approach. The accuracy of the parser is 86.6% Tokenization, 64.1% UPOS, 63.78% XPOS, 42.04% UAS and 21.34% LAS. Finally, the paper briefly discusses the linguistic analysis of the Odia UD treebank.

The Odia Treebank available for research at:https://github.com/UniversalDependencies/UD_Odia-ODTB/tree/dev

Case 3 - BertOdia

Overview

- Building a language model is a challenging task in the case of low resource languages where the availability of contents is limited.
- We focus on building a general language model using the limited resources available in the low resource language which can be useful for many language and speech processing tasks.

- Our key contribution includes building a language-specific BERT model for this low resource Odia language and as per our best knowledge, this is the first work in this direction.

Case 3 - BertOdia

Data and Model

- Building a language model is a challenging task in the case of low resource languages where the availability of contents is limited.
- We focus on building a general language model using the limited resources available in the low resource language which can be useful for many language and speech processing tasks.
- Our key contribution includes building a language-specific BERT model for this low resource Odia language and as per our best knowledge, this is the first work in this direction.

Source	Sentences	Unique Odia tokens
OdiEnCorp2.0	97,233	1,74,045
CVIT PIB	58,461	66,844
CVIT MKB	769	3,944
OSCAR	1,92,014	6,42,446
Wikipedia	82,255	2,36,377
Total Deduped	430,732	11,23,656

Table . Dataset statistics.

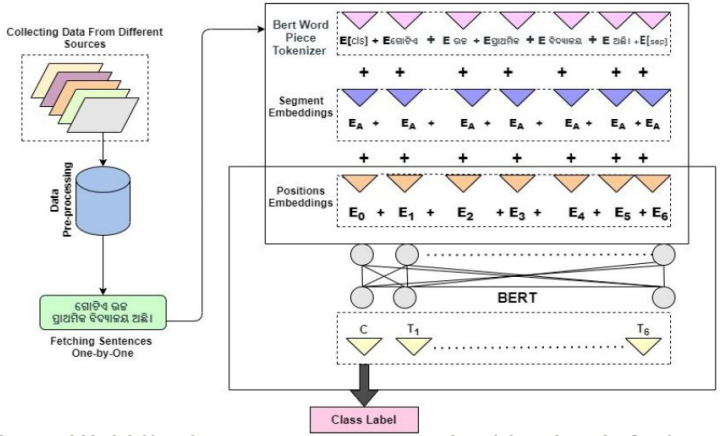


Fig: The Proposed Model: Visualisation of an our experimental model used for the Single Sentence Classification Task with Bert Embedding Layers.

Case 3 - BertOdia

BERT/RoBERTa Model Training

Parameter	BERT	RoBERTa
Learning Rate	5e-5	5e-5
Training Epochs	5	10
Dropuout Prob	0.1	0.1
MLM Prob	0.1	0.2
Self attention layer	6	6
Attention head	12	12
Hidden layer size	768	768
Hidden layer Activation	gelu	gelu
Total parameters	84M	84M

Table 2. Training Configurations

Model	Text Classification Accuracy
BertOdia	96.0
RoBERTaOdia	92.0
ULMFit	91.9

Table 3. BertOdia Performance

Case 3 - BertOdia

IndicGlue Task

- For the Cloze-style Multiple-choice QA task, we feed the masked text segment as input to the model and we fine-tune the model using cross-entropy loss.
- For the Article Genre Classification task we used the IndicGLUE dataset for news classification.

Model	Article Genre Classification	Cloze-Style multiple-choice QA
XLM-R	97.07	35.98
mBERT	69.33	26.37
IndicBERT base	97.33	39.32
IndicBERT large	97.60	33.81
BertOdia	96.90	23.00

Table: Comparison of BertOdia with IndicBERT. BertOdia was trained on 6% of the data of IndicBERT.

- The code and dataset are available at:

https://colab.research.google.com/gist/satyapb2002/aeb7bf9a686a9c7294ec5725ff53fa49/odiabert_languagemodel.ipynb

Mini-project

Odia Dialect Corpus

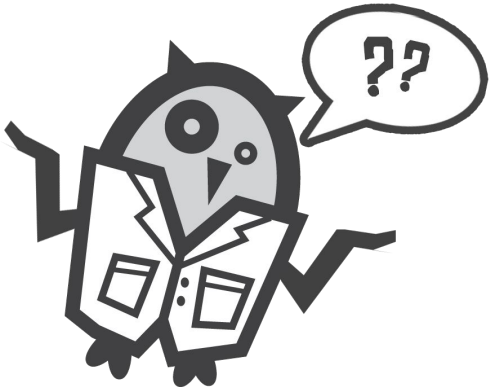
Link: https://odisha-ml.github.io/mini-projects/mp10/readme/

Problem Statement

As the written text follows the same Odia script, how to distinguish different dialects for many NLP applications (e.g. automatic identification of language for machine translation, text categorization) for different services. Can machine learning techniques help to solve this ?.

ଲିଖିତ ଓଡ଼ିଆ : ମୋର ଆଜି ବହୁତ କାମ ଅଛି
ଦେଶିଆ: ମର କେତେ କାମ୍
ଗଞ୍ଜାମୀ: ଆଜି କାମ ଟିକେ ଅଧିକା କା
ସମ୍ବଲପୁରୀ : ମୋର ଆଜି ଟିକେ ବେଶି କାମ ଅଛେ
ବାଲେଶ୍ୱରୀ: ମର ଆଜି ଢେର କାମ ଅଛି

Q&A



Contact information:

- Twitter: @Shantipriyapar3
- Web : shantipriya.me

Thank you