LANGUAGE DETECTION AND IDENTIFICATION USING NATURAL LANGUAGE PROCESSING TECHNIQUES

COMPILED BY: KULDEEP UPADHYAY

1. Introduction

In today's interconnected world, communication across different languages has become essential. People frequently interact, learn, and work globally; however, language differences often create communication barriers. Machine Translation is a field within Natural Language Processing (NLP) that focuses on automatically converting text from one language to another without changing the meaning or context.Traditional machine translation relied on dictionary lookup and grammar rules, which often produced incorrect or unnatural sentence structures. Modern systems, such as Google Translate, use Deep Learning to understand sentence context and generate fluent translations.In this project, we developed a Language Translation System using a Sequence-to-Sequence (Seq2Seq) Encoder-Decoder model with Attention Mechanism. This model reads the input sentence, understands the context, and produces meaningful translations in the target language.

2. Dataset Collection

For this study, we used a Parallel Corpus Dataset, which contains pairs of sentences where each sentence in the source language has its corresponding translation in the target

1. Dataset Type:

- Parallel Corpus Dataset containing pairs of sentences — each English sentence has a corresponding translation in the target language.

2. Languages Used:

- **Source Language:** English
- **Target Language:** (Select: Hindi / Kannada / Telugu / etc.)

3. Data Format:

- Two columns:
    - source_sentence — sentence in English
    - target_sentence — corresponding translated sentence in the target language

4. Preprocessing Steps:

- Converted all text to lowercase
- Removed special characters and extra spaces
- Performed tokenization (splitting sentences into individual words)
- Created vocabulary for both source and target languages
- Applied padding to ensure all sequences have equal length

5. Train-Test Split:

- 80% of the data used for training
- 20% of the data used for testing

3. Method Used

Model Type:

- Sequence-to-Sequence (Seq2Seq) Model with **Attention Mechanism**

The model consists of three major components:

1. Encoder

- Takes the input sentence in the **source language**
- Converts each word into **dense vector embeddings**
- Uses an **LSTM network** to capture the overall **meaning and context** of the input sequence

2. Attention Layer

- Enables the decoder to **focus on the most relevant words** from the input sentence during translation
- Enhances **translation accuracy and fluency**, especially for **long or complex sentences**

3. Decoder

- Generates the **translated sentence** in the target language **word by word**

- Utilizes previously generated words and the **context information** provided by the Attention mechanism to produce coherent and accurate translations

4. Comparative Analysis with Existing Methods

To understand the benefit of deep learning, we compared our LSTM-based approach with traditional classifiers:

```
K-Nearest Neighbors accuracy score :  0.524

Random Forest accuracy score :  0.927

MNB accuracy score :  0.981

CPU times: total: 2min 58s

Wall time: 2min 44s
```

Observations:

• **MNB achieved the highest accuracy (0.981),** outperforming Random Forest (0.927) and KNN (0.524).

• **Total computation time was around 3 minutes,** indicating efficient model execution.

5. Applications

Applications of this work include:

- **Education:** Language learning support
- **Travel Apps:** Communication in foreign countries
- **Customer Support:** Multilingual automated replies
- **Social Media:** Instant post and comment translation
- **Online Learning:** Multi-language content access

6. Conclusion

In this project, we successfully implemented a Neural Machine Translation (NMT) model using a Sequence-to-Sequence (Seq2Seq) Encoder-Decoder architecture with an Attention mechanism. The developed system effectively understands sentence structure, captures contextual meaning, and translates text accurately with smooth and natural phrasing. Compared to traditional translation approaches, this model demonstrates superior performance by offering better context understanding, higher translation accuracy, and improved fluency in the generated sentences.

References

1. IMDb Dataset: Language-Detection-using-NLP-and-ML
2. *Methods and Evaluation*. arXiv preprint. 2020. https://arxiv.org/abs/1912.08494
3. YouTube Project Reference: https://www.muratkarakaya.net/2022/11/seq2seq-learning-tutorial-series.html

Screenshots of Results

Dataset-(of 22000rows x 2columns)

	Text	language
0	klement gottwaldi surnukeha palsameeriti ning ...	Estonian
1	sebes joseph pereira thomas på eng the jesuit...	Swedish
2	ถนนเจริญกรุง อักษรโรมัน thanon charoen krung เ...	Thai
3	விசாகப்பட்டினம் தமிழ்ச்சங்கத்தை இந்துப் பத்திர...	Tamil
4	de spons behoort tot het geslacht haliclona en...	Dutch
...	...	...
21995	hors du terrain les années et sont des année...	French
21996	ใน พศ หลักจากที่เสด็จประพาสแหลมมลายู ชวา อินเ...	Thai
21997	con motivo de la celebración del septuagésimoq...	Spanish
21998	年月，當時還只有歲的她在美國出道，以mai-k名義推出首張英文《baby i like》，由...	Chinese
21999	aprilie sonda spațială messenger a nasa și-a ...	Romanian

22000 rows × 2 columns

OUTPUT:

[29]:
```
user = input("Enter a text")
data = cv.transform([user]).toarray()
output = model.predict(data)
print(output)
```

```
Enter a text    Text 23 la chirurgie comprenant principalement lablation de la tumeur la néphrectomie élargie comprenant le plus souvent la surrénale et les ganglions situés à proximité
['French']
```

[30]:
```
user = input("Enter a text")
data = cv.transform([user]).toarray()
output = model.predict(data)
print(output)
```

```
Enter a text    Text 30 فریک نامیده می‌شود ترکیبات زیادی در هر iii فروس نامیده می‌شوند و ترکیبات آهن ii آهن ترکیباتی را ایجاد می‌کند که عمدتاً در حالت‌های اکسیداسیون  و  هستند بطور سنتی ترکیبات آهن می‌باشد iii fecl و کلرید آهن ii feso یک از حالات اکسیداسیون وجود دارد که مثال‌هایی از آن شامل سولفات آهن
['Persian']
```