English to French Translation (Using Deep Learning)

Submitted By:

Aashima
(021323001)

Semester Four

Session: 2023–27

Bachelor of Technology (B.Tech – AI ML)

Please feel free to reach out to me at the following

Email: aashimabhatia2005@gmail.com

LinkedIn: https://www.linkedin.com/in/aashima-bhatia-a30919300/

JAGANNATH UNIVERSITY NCR, HARYANA

College Dekho Assured (ImaginXp)
Jagannath Community College (Delhi)
Jagannath University, Bahadurgarh, Haryana

1 Introduction

This project develops an AI system for translating English sentences to French using deep learning. The system leverages a sequence-to-sequence model with LSTM networks to learn mappings between English and French sentence pairs. Trained on a dataset of 137,860 sentence pairs, it aims to provide accurate translations for simple sentences, supporting language learning and communication.

2 KeyLibraries

- **TensorFlow/Keras**: BuildsandtrainstheLSTM-basedsequence-to-sequence model.
- **Pandas**: Manages dataset loading and metadata handling (e.g., sentence pairs).
- **NumPy**: Supports numerical operations for data preprocessing.
- **NLTK**: Provides tokenization and stopwords for text processing.
- **Scikit-learn**: Splits data into training and test sets.
- **Matplotlib/Seaborn/Plotly**: Visualizes data and model performance.
- **WordCloud**: Generates word clouds for text analysis.

3 KeyFeaturesoftheProject

1. **Automated Translation**: Translates English sentences to French using a trained LSTM model.
2. **Sequence-to-Sequence Architecture**: Employs encoder-decoder LSTM for context-aware translations.
3. **Large Dataset**: Processes 137,860 sentence pairs for robust training.
4. **Text Preprocessing**: Includes tokenization, padding, and vocabulary creation.
5. **Model Evaluation**: Assesses translation accuracy on test data.
6. **Scalable Pipeline**: Handles large datasets with efficient batch processing.

4 Core Algorithm

4.1 DataPreparation

- LoadsEnglishandFrenchsentencesfrom `small_vocab_en.csv` and `small_vocab_fr.`
- Tokenizes sentences and creates vocabularies using `Tokenizer`.

- Pads sequences to fixed lengths (15 for English, 23 for French).
- Splits data into 90% training and 10% test sets.

4.2 ModelArchitecture

- **Encoder**: Embeddinglayer(256dimensions)followedbyLSTM(256units). •

Decoder: RepeatVectortoaligninput-outputlengths,LSTM(256units),and TimeDistributed Dense with softmax for word prediction.

- Compiles with Adam optimizer and sparse categorical cross-entropy loss.

4.3 Training

- Trains on batches of 1024 samples for 10 epochs.
- Uses validation split (10%) to monitor performance.

4.4 Prediction and Evaluation

- Predicts French translations by decoding model outputs.
- Evaluates translations qualitatively on test samples.

5 Conclusion

The project successfully implements an English-to-French translation system using LSTM-based deep learning, achieving 86.37% validation accuracy after 10 epochs. The model handles simple sentences well but struggles with complex structures and rare words. Future improvements could include attention mechanisms, larger datasets, and transformer models for enhanced accuracy and con- text understanding.

6 Links

- **Demo Link**: Not available.
- **LinkedInPost**:

https://www.linkedin.com/feed/update/urn:li:activity:7324430199721254