EXPERIMENT 7

Seq2Seq Model for English-to-Spanish Translation using LSTM

OBJECTIVE

To develop and evaluate a basic Sequence-to-Sequence (Seq2Seq) model based on LSTM networks to translate English sentences into Spanish.

DATA PREPROCESSING

Dataset:

- Source: `dataset.txt` (Tab-separated English-Spanish sentence pairs)

Steps:

- Convert text to lowercase
- Remove leading/trailing whitespace
- Filter out very short or overly long sentences

Tokenization:

- Tokenized on whitespace
- Added special tokens: `<sos>` (start of sentence), `<eos>` (end of sentence), `<pad>`, `<unk>`
- Constructed word-to-index and index-to-word dictionaries

Preparation:

- English (input) and Spanish (target) token sequences were converted to indices
- Target sequences wrapped with `<sos>` and `<eos>`
- Padding applied to ensure equal-length sequences

Split:

- 80% Training, 10% Validation, 10% Test

MODEL ARCHITECTURE (NO ATTENTION)

Encoder:

- Embedding Layer → LSTM Layer
- Returns final hidden and cell states

Decoder:

- Embedding Layer → LSTM initialized with encoder states → Dense → Softmax

Other Components:

- Weight Initialization: Xavier or uniform
- Activations: Tanh (LSTM), Softmax (output)
- No dropout or L2 regularization
- Loss: CrossEntropyLoss (ignore `<pad>`)

TRAINING CONFIGURATION

Parameter	Value
Epochs	200
Batch Size	64
Learning Rate	0.001
Optimizer	Adam
Teacher Forcing Ratio	0.5

RESULTS

Metric	Value
Final Loss	9.4202
BLEU Score	0.0878

SAMPLE TRANSLATION

Input	Predicted Output
I am happy	soy feliz.

INSIGHTS

- Slow convergence
- Struggles to generalize to longer or complex sequences
- Lacks context awareness, often outputs short or incomplete sentences