KG+LLM: A Happy Marriage

Integrating LLMs with Knowledge Graphs

Emil Soltanov

Università degli Studi di Milano

January 2025

Outline

Background

- Knowledge Graphs (KGs): Structured ways of representing entities and relationships.
- Challenge: Traditional KG construction is labor-intensive.
- Solution: Automate KG construction using NLP models like LSTM, BERT, and RoBERTa.

Problem Statement

- **Current Challenge**:
  - Manual KG construction is time-consuming
  - Hard to maintain consistency
  - Difficult to scale
- **Research Question**: How can we effectively automate KG construction using language models?
- **Impact**: More efficient and scalable knowledge representation

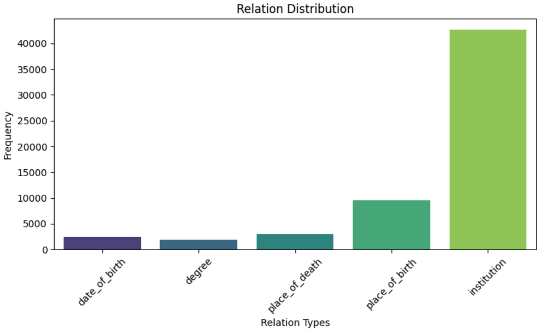
Objectives

1. Evaluate machine learning models for relation classification.
2. Construct and visualize Knowledge Graphs.
3. Identify strengths and limitations of each model.

Dataset Overview

- Relation types:
  - `date_of_birth`, `place_of_birth`, `place_of_death`, `institution`, `degree`.
- Preprocessing:
  - Normalization, Balancing, Tokenization.
- Dataset Link: Relation Extraction Corpus.

Dataset Distribution



- Total samples: 8,745
- Balanced across 5 relations
- Evidence snippets included

Models Used

- **LSTM**: Captures long-term dependencies.
- **BiLSTM**: Adds bidirectional context.
- **BERT**: Transformer-based, pretrained embeddings.
- **RoBERTa**: Robustly optimized BERT.

Model Configurations

- **LSTM/BiLSTM**:
  - Embedding dim: 256
  - Hidden dim: 256/512
  - Layers: 2
- **BERT/RoBERTa**:
  - Base models
  - Dropout: 0.4/0.5
  - Learning rates: 1e-5/3e-6

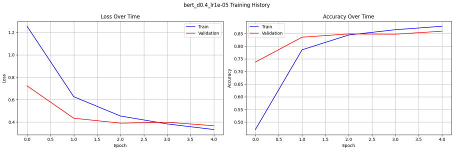
Training and Evaluation

- Metrics: Accuracy, Precision, Recall, F1-Score.
- Tools: PyTorch, Hugging Face, Matplotlib.
- Workflow:
  1. Data Splitting.
  2. Model Training and Tuning.
  3. Evaluation using classification reports and confusion matrices.

Model Performance

- BERT: Highest accuracy (86%).
- RoBERTa: Strong performance (85%).
- BiLSTM: Balanced performance (81%).
- LSTM: Efficient but less accurate (74%).

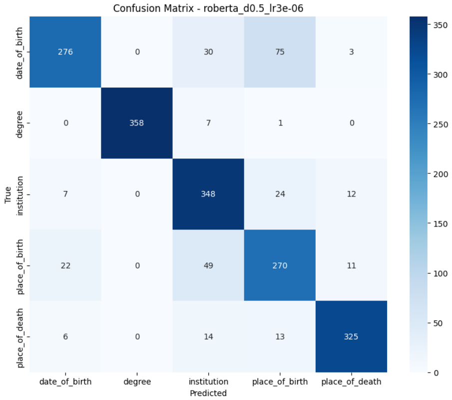
Training Progress



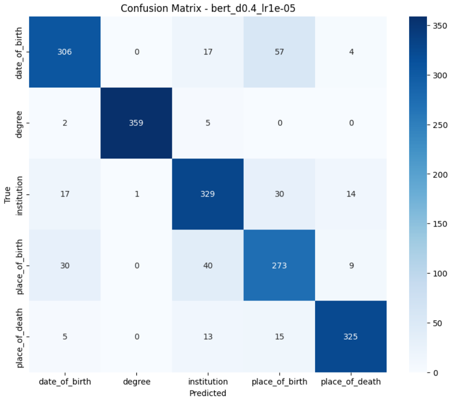
BERT Training

RoBERTa Training

Confusion Matrices

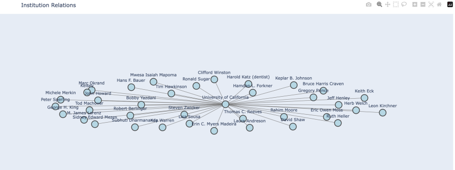


RoBERTa Confusion Matrix



BERT Confusion Matrix

Knowledge Graph Analysis



- Institution relations form key clusters
- Clear temporal patterns visible
- Geographic distribution evident
- Connected educational networks

Key Findings

- **Model Performance**:
  - Transformers excel at complex relations
  - LSTM/BiLSTM efficient for simple relations
- **Trade-offs**:
  - Accuracy vs Computational Cost
  - Model Size vs Training Speed
- **KG Insights**:
  - Rich institutional networks
  - Clear temporal patterns

Conclusion

- BERT and RoBERTa excel in accuracy and generalization.
- BiLSTM is a balanced option for resource-constrained setups.
- Knowledge Graphs can be effectively constructed using automated relation extraction.

Future Work

- Explore lightweight models like DistilBERT.
- Incorporate external data sources for enrichment.
- Optimize for scalability using distributed training.

Thank You!

Questions?