Automated Assessment of Answer Relevance in Question Answering Systems: A Machine Learning Approach

EL ALAOUI Zakaria - zakaria.elalaoui@edu.uiz.ac.ma

Abstract

This paper presents a neural approach for scoring the relevance of answers to French language queries in question-answering systems. We propose a hierarchical architecture combining CamemBERT with a specialized neural network featuring dimensional reduction, dropout regularization, and layer normalization. The model is trained on balanced positive and negative pairs from the French Question Answering Dataset (FQuAD), using Binary Cross-Entropy with Logits Loss and mixed precision training. Our implementation produces continuous relevance scores between 0 and 1, effectively capturing semantic relationships between questions and potential answers while maintaining computational efficiency.

Keywords: Natural Language Processing, Relevance Scoring, CamemBERT, French Language Processing, Neural Networks, Question Answering Systems, Mixed Precision Training

1. Introduction

The rapid growth of question answering (QA) systems has created an increasing need for automated methods to evaluate answer quality and relevance. While significant progress has been made in generating answers to user queries, the automated assessment of answer relevance remains a crucial challenge in maintaining QA system quality. This paper presents our contribution to addressing this challenge through the development of a machine learning model for scoring answer relevance in the context of French language question answering systems. Our work focuses on developing a machine learning model that can automatically evaluate the relevance of potential answers to user queries, using the French Question Answering Dataset (FQuAD) as our primary data source. This task is particularly significant as it addresses the growing demand for French language natural language processing (NLP) tools and contributes to the broader field of multilingual QA systems.

2. Data Preparation and Preprocessing

Our approach to developing a relevance scoring system began with careful preparation of the training data. The initial dataset, derived from FQuAD (French Question Answering Dataset), consisted of question-article pairs in French. To create a robust training set, we implemented a balanced sampling strategy that generated both positive and negative examples:
This approach ensured a balanced dataset with an equal number of positive and negative examples, which is crucial for training an unbiased model.

1. **Positive Pairs:** We retained the original question-article pairs from the dataset, assigning them a relevance label of 1 to indicate high relevance.
2. **Negative Pairs:** For each question, we created corresponding negative examples by randomly sampling articles from the dataset that were not originally paired with that question, assigning them a relevance label of 0.

3. Model Architecture and Training Details

3.1 Neural Network Architecture

Our relevance scoring model employs a hierarchical architecture that combines pre-trained language representations with task-specific layers. The model architecture consists of three main components:

3.1.1 Base Language Model

We utilize CamemBERT as our foundation model, specifically the camembert-base variant, which provides contextual embeddings for French text. CamemBERT's architecture includes:

- Pre-trained transformer layers optimized for French language understanding
- Output dimension: 768 (hidden_size)
- Contextual representation of input tokens

3.1.2 Feature Processing Layers

The model incorporates several layers for processing the CamemBERT embeddings:
Key components include:

- **Dense Layer:** Reduces dimensionality from 768 to 256 units, enabling the model to learn task-specific representations
- **Dropout Layer:** Applies 20
- **Layer Normalization:** Stabilizes training by normalizing the activations
- **Output Layer:** Single-unit layer producing raw logits for relevance scoring

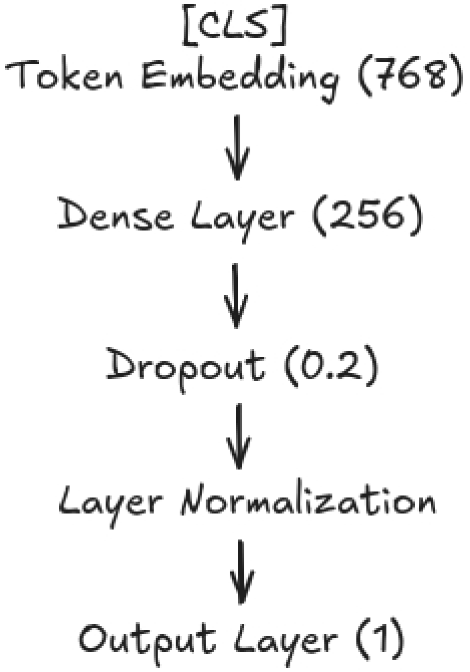

Figure 1: Feature Processing Pipeline.

3.1.3 Forward Pass Implementation

1. Extract token embeddings from CamemBERT.
2. Select the [CLS] token embedding for sentence-pair representation.
3. Apply dropout for regularization
4. Normalize the activations
5. Generate final relevance logits

3.2 Training Configuration

The model is trained with the following optimization setup:

3.2.1 Loss Function

Binary Cross-Entropy with Logits Loss (**BCEWithLogitsLoss**)

- Combines sigmoid activation and binary cross-entropy
- Provides numerical stability and appropriate for binary relevance scoring

3.2.2 Optimization

Optimizer: **AdamW**

- Learning rate: **2e-5**
- Handles weight decay correction

3.2.3 Training Enhancements

Mixed Precision Training

- Utilizes both FP16 and FP32 to reduce memory usage
- Accelerates training while maintaining stability

The model processes pairs of questions and potential answers, producing a raw logit score that is then transformed into a probability through the sigmoid function during inference, resulting in relevance scores between 0 and 1.

4. Inference Pipeline

For prediction, we implemented a pipeline that:

1. Tokenizes input question-article pairs.
2. Processes them through the model
3. Applies sigmoid activation to output logits
4. Returns a relevance score between 0 and 1