Glossary

Attention Mechanism

In Natural Language Processing (NLP), the Attention Mechanism is an algorithm used to calculate the relational weight between elements in a sequence of elements (most often words).

Close-ended

A closed-ended question is designed to allow a limited amount of responses.

Encoder-Decoder

In Machine Learning (ML), Encoder-Decoder is two Neural Networks (NNs) that work in pair. The Encoder generates a fixed-size output vector from any sized vector input. And the Decoder generates from the Encoder output a vector that could be any size.

Generative

In the context of the Thesis, we are using the generic word Generative as the ability concept of an algorithm able generating outputs in a meaningful but unpredictable manner from an input, which includes Language Model (LM)s and Generative Models.

Generative Model

In Machine Learning (ML), Generative Models are generating random outputs from a single input by using the probability of observing the output based on the input. In other words, it models the probability of observation for a given target.

Hop

In Question Answering (QA) Systems, a Hop is a quantitative measure of the number of combinations necessary between indirectly related pieces of information to provide an answer.

Knowledge Base

In Information Systems (IS), a Knowledge Base is a Knowledge Representation using a Linked Data database for storing and interlinking structured and unstructured data using a standard.

Glossary

Knowledge Graph

In Information Systems (IS), a Knowledge Graph is a Knowledge Base (KB) organized as a graph using semantics.

Language Model

In Natural Language Processing (NLP), a Language Model is a Model trained to provide likelihood probabilities of the following sequence of words in addition at providing the probability for each sequences of words.

Linked Data

In Information Systems (IS), Linked Data is a structured interlinked database mainly used for semantic queries.

Model

In Machine Learning (ML), a model is the representation of the assumptions made by the algorithm during the training phase. Models are used to output a result based on a provided input and the learned patterns.

Model Fine-Tuning

In Machine Learning (ML), Fine Tuning a Model is the technic of using a trained Neural Network (NN) model as a base and tune it for a specific task.

Multi-Hop

In Question Answering (QA) Systems, a Multi-Hop implies that the answer is within multiple Hop of the question. In other words, the answer requires a combination of different information to be answerable. Generally, extra qualifying Subject-Predicate-Object Tuple (SPO) are separating the question Subject and the answer Object.

Named-Entity Linking

In Natural Language Processing (NLP), Name-Entity Linking extends the Named-Entity Recognition by providing an unique identifier to each word allowing a mapping in various databases (useful in translations).

Named-Entity Recognition

In Information Extraction (IE), Named-Entity Recognition is a technic used to extract from unstructured text words predefined in a vocabulary.

Open Domain

In Information Retrieval (IR), the support of Open Domain questions provides a no restrictions for the theme of the question asked.

Part of Speech

In Natural Language Processing (NLP), Part of Speech is a technic used to categorize words that behave syntactically similar.

Part of Speech Tagging

In Natural Language Processing (NLP), The Part of Speech Tagging is extending the Part of Speech by addition a word tag depending on the context the word is in (the neighboring words).

Sequence-to-Sequence

In Machine Learning (ML), a Sequence-to-Sequence or Seq2Seq is an Encoder-Decoder Neural Network (NN) that for a given sequence of elements as input, outputs another sequence of elements.

Single-Hop

In Question Answering (QA) Systems, a Single-Hop implies that the answer is within a single Hop of the question. Generally, a unique Predicate separates the question Subject and the answer Object.

Transformer

In Natural Language Processing (NLP), a Transformer is a Sequence-to-Sequence (Seq2Seq) architecture using the Attention Mechanism..

Word Embedding

In Natural Language Processing (NLP), the Word Embedding is a technic for word representation as vectors in an embedding matrix. Additionally, it has often the particularity of preserving the semantical analogies of word-vectors.

Zero-Shot Learning

In Machine Learning (ML), Zero-Shot Learning is technic used to solve tasks without training on examples.

Acronyms

AGI
Artificial General Intelligence.

AI
Artificial Intelligence.

ANI
Artificial Narrow Intelligence.

BERT
Bidirectional Encoder Representations from Transformers.

CNN
Convolutional Neural Network.

DL
Deep Learning.

GPT-2
Generative Pre-Training 2.

GS
Generative System.

ICT
Information and Communications Technologies.

IE
Information Extraction.

IR
Information Retrieval.

IS
Information Systems.

KB
Knowledge Base.

Acronyms

KG
Knowledge Graph.

LM
Language Model.

ML
Machine Learning.

MN
Memory Network.

MRR
Mean Reciprocal Rank.

MRU
Master Research Units.

MT
Master's Thesis.

NL
Natural Language.

NLP
Natural Language Processing.

NLU
Natural Language Understanding.

NN
Neural Network.

POC
Proof of Concept.

QA
Question Answering.

RNN
Recurrent Neural Network.

Seq2Seq
Sequence-to-Sequence.

SOTA
State of the Art.

SPO
Subject-Predicate-Object Tuple.

Abstract

In the scope of this study, we realize the Proof of Concept (POC) of an Open Domain and Close-ended Question Answering (QA) chatbot able to output comprehensive Natural Language (NL) generated sentences using the Wikidata Knowledge Base (KB).

To achieve the concept, we explore the extraction, and the use of sub-knowledge graphs from the Wikidata Knowledge Base (KB) to answer Natural Language (NL) questions conversationally and using the sub-graphs as context holder. Additionally, we are extracting Subject-Predicate-Object Tuples (SPOs) paths from the graph and enricher them using a Language Model (LM) to provide NL answers.

The POC architecture uses a combination of State of the Art (SOTA) and industry-used Models with a fine-tuning strategy for the NLP. As a motivational target, we are using a Zero-Shot Learning approach, by using pure algorithmic for the graph manipulation and answer extraction.

Finally, we evaluate the answers and compare the results with State of the Art (SOTA) QA systems on Single-Hop and Multi-Hop QA datasets.

Keywords: Machine Learning (ML), Natural Language Processing (NLP), Single-Hop, Multi-Hop, Question Answering (QA), Wikidata, Wikipedia, Knowledge Graph (KG), Knowledge Base (KB), Word Embedding, Part of Speech Tagging, Named-Entity Recognition, Named-Entity Linking, Language Model (LM), Model Fine-Tuning, Graphs, Sub-Knowledge Graphs, Transformer, Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-Training 2 (GPT-2), Information Extraction (IE), Spacy, GloVe, DeepCorrect, PyTorch, Python, Chatbot, Conversational, Information Retrieval (IR), Queries