Cross-Lingual Information Retrieval with Neural Network

—

- Shaun N., Sheroz S.

Table Of Contents

01

Introduction

What's our goal/idea?

Introduction

- What is cross-lingual information retrieval?
- **Objective**: Our goal is to facilitate cross-lingual information retrieval, enabling users to find relevant documents from a large corpus, regardless of their query language.
- **Idea**: A tool that aims to extract relevant documents from a big corpus based on a user-specified query where the query and relevant documents can be one of the 10 selected languages ('en': English, 'ar': Arabic, 'fi': Finnish, 'ja': Japanese, 'ko': Korean, 'ru': Russian', tr': Turkish, 'hu': Hungarian, 'nl': Dutch', de': German)

02

Necessity & Existing Tools

Why do we care (compared with the prior work or related applications)?

Language diversity is a challenge, hindering access to global knowledge. The need for such a tool is significant, and it can benefit a wide range of users.

Domain-Specific User

- Medical Researcher
- Linguist
- Legal Translator

Non-specialized User

- Travel Blogger
- Film Enthusiast
- Fitness Coach

02

Dataset

Description of the dataset
Columns in the dataset
Exploratory Data Analysis

- The dataset being used is Cross-lingual Open-Retrieval Question Answering Data

Query	Correct Answers / Document	InCorrect Answers / Document
wer hat these boots are made for walking gesungen	синатра, нэнси	hvem sang these boots are made for walking
wer hat these boots are made for walking gesungen	nancy sinatra	kto śpiewał "these boots are made for walking"
wer hat these boots are made for walking gesungen	nancy sinatra	quem cantava these boots are made for walking?
wer hat these boots are made for walking gesungen	nancy sinatra	vem sjöng these boots are made for walking
wer hat these boots are made for walking gesungen	nancy sinatra	ai đã hát bài hát these boots are made for wal...

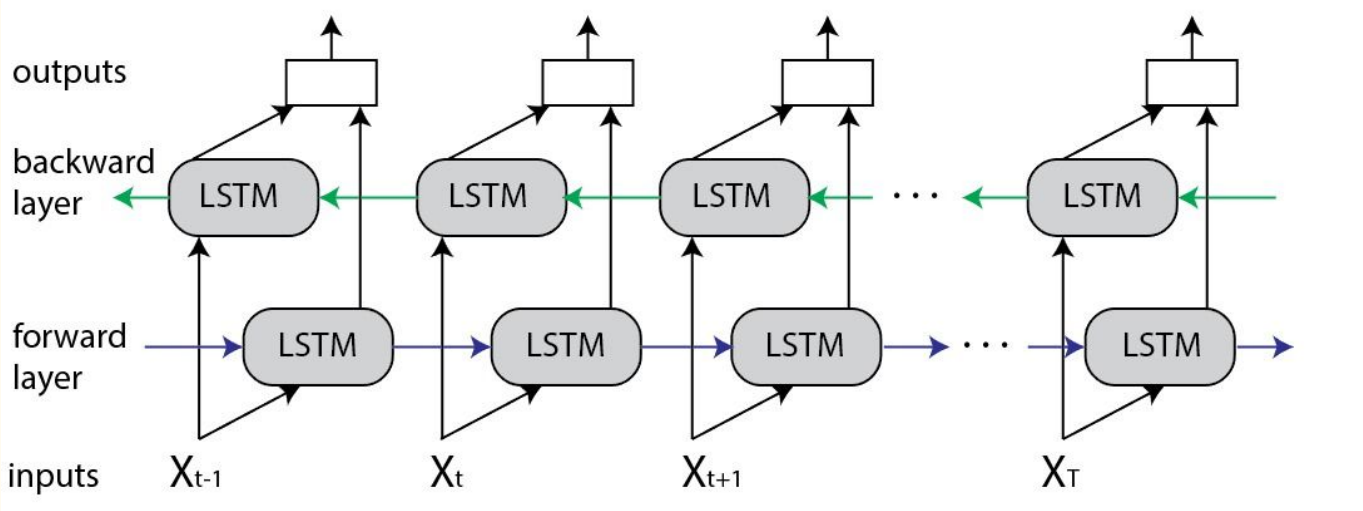
Column	Description
Query	Queries in various languages
Correct Answers /Document	Correct answers for the documents
Incorrect Answers /Document	Incorrect answers for the documents

04

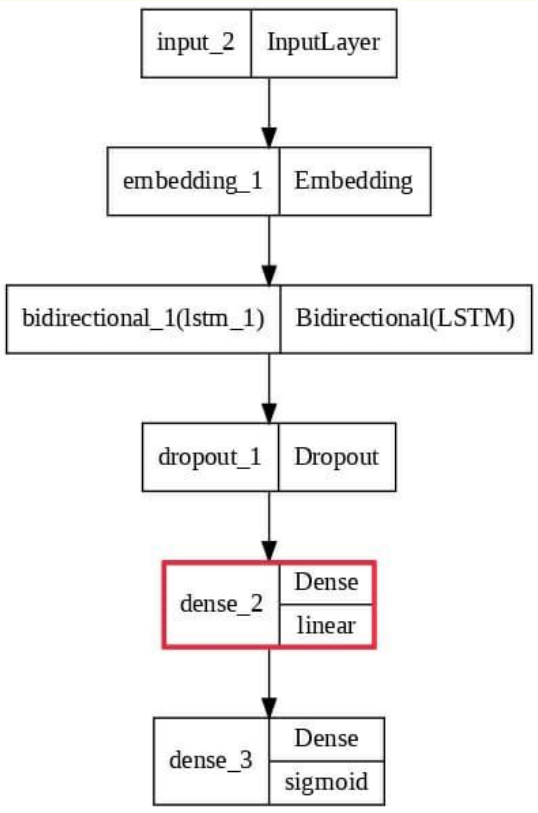
Implementation & Methodology

How your tool/application/algorithm would be different from the prior work?

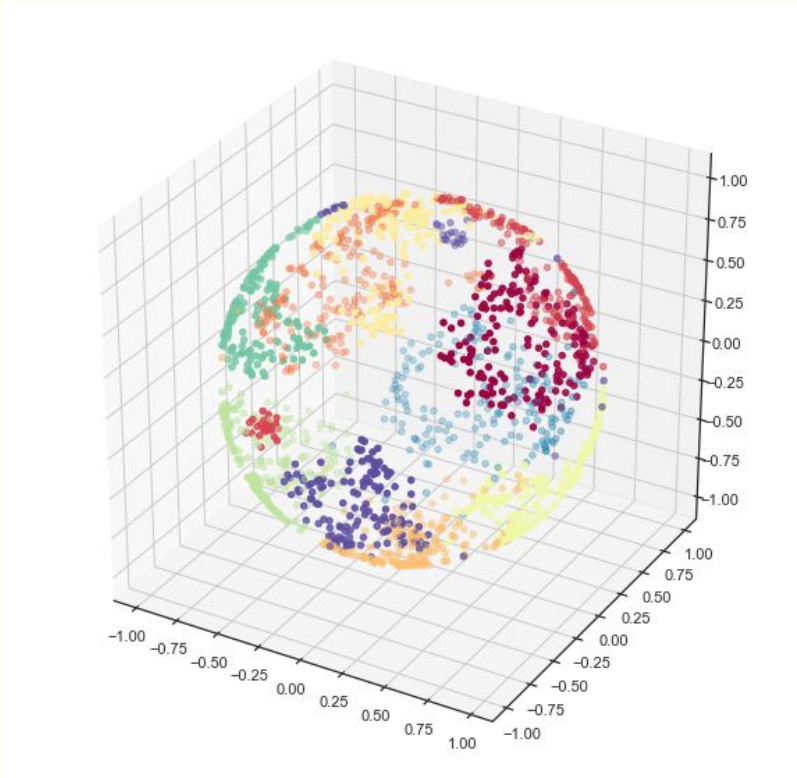
Bidirectional LSTM



Model Design



Embedding Space



05

Conclusion & Future Work

- The performance of the nearest neighbor search may vary across different methods (NearestNeighbors, AnnoyIndex, hnswlib). Comparing their recall rates can help identify the most suitable method.

	Index_IDS	NearestNeighbors	AnnoyIndex	HNSWLib	Best_Of_3	Neighbours
0	44	1	1	1	1	2
1	44	1	1	1	1	5
2	44	1	1	1	1	10
3	1156	1	1	1	1	2
4	1156	1	1	1	1	5
5	1156	1	1	1	1	10

- Sentence Tokenizer (BERT)
- Other search metrics such as Euclidean or Cosine Similarity

Questions?

Thank You!