IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 3, March 2024

DOI: 10.17148/IARJSET.2024.11333

# Deep Cross Lingual Semantic Search For CLIR System

**G. Venkateswari[1], P. Chandrika[2], SK. Nausheen[3], P. Manasa Veena[4]**

MTech(P.hd), Computer science & Engineering, Bapatla women's Engineering College, Bapatla, INDIA[1]

BTech, Computer science & Engineering, Bapatla Women's Engineering College, Bapatla, INDIA[2-4]

**Abstract**: The amount of digital content available on the Internet has grown exponentially in recent years, and this rise has coincided with an increase in the number of non-English Internet users as a result of the Internet's globalization. This emphasizes how crucial it is to make materials available to people who wish to research things rather than restricted to the languages they are able to speak. For instance, those who wish to utilize the Internet to research medical information about their ailments (self-diagnosis) but are unable to access resources in their native tongue. Language barriers are overcome by Cross Lingual Information Retrieval (CLIR), which enables document searches in languages other than the query language.

**Keywords:** Cross-lingual Information Retrieval, Machine Translation, Consumer Health Search, NLP

## I. INTRODUCTION

The digital medical content available online has snowballed in recent years. This growth has the potential to improve experience with web medical Information Retrieval (IR) systems, which are more and more used for health consultations. Fox [2011] reported that about 80% of Internet searchers in the U.S. looked for health information online, and this number was expected to grow. The significant increase of non-English digital content on the Internet had been followed by an increase in looking for this information by internet searchers. Grefenstette and Nioche [2000] presented an estimation of language size in 1996, late 1999 and early 2000 for documents captured from the Internet. Their study showed that the English content had grown by 800%, German by 1500%, and Spanish by 1800% in the same period. Naturally, some information that a searcher is looking for might be available only in a language that they do not understand, which makes such information not accessible to those searchers. Cross-Lingual Information Retrieval (CLIR) comes to tackle this issue and to help internet searchers break language barriers and access valuable information that is not available in their language.

## II. BACK GROUND AND RELATED WORK

By enabling queries in a language other than the collection language, CLIR facilitates information searches for users. This makes it easier for searchers to access a wealth of material that is expressed in multiple languages by bridging the language gap.

The since the late 1990s, this challenge has drawn attention from the IR research community. The rise of the Internet was a strong indicator of the necessity for CLIR systems, as the amount of digital content available worldwide started to rise dramatically.

A CLIR system typically consists of two steps: the translation phase, which translates the document collection into the query language or the queries into the language of the document collection.

### A. Dictionary-based Query Translation

CLIR helps users find information by allowing queries in a language other than the collection language. By bridging the language divide, this facilitates searchers' access to a multitude of material represented in many languages. The IR research community has been paying attention to this topic since the late 1990s. The volume of digital content that became available globally began to expand quickly with the emergence of the Internet, which was a strong indicator of the need for CLIR systems. The translation phase of a CLIR system usually consists of two steps: either the queries are translated into the language of the document collection, or the document collection is translated into the query language.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 3, March 2024

DOI: 10.17148/IARJSET.2024.11333

## III. METHODOLOGIES

All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

### A. Query Translation Using Word Embeddings

Finding syntactic and semantic similarities in text and capturing word context in documents are the two main objectives of word embedding in natural language processing (NLP). To achieve this, a distributed representation of words as dense vectors is introduced.

Still, the word-embedding technique is not an original attempt at NLP.

The concept originated with Latent Semantic Analysis (LSA) [Deerwester et al., 1990]. LSA is regarded as the first method to represent words as vectors in a semantic space. The primary premise upon which LSA is predicated is that related words occur in the same textual segments (paragraphs).

### B. Section Headings

One type of machine learning model that produces sequential data as output and accepts sequential data as input is called the seq2Seq model. Machine translation systems relied on statistical techniques and phrase-based methodologies prior to the introduction of Seq2Seq models. Using phrase-based statistical machine translation (SMT) systems was the most widely used method. That failed to take into account global context and manage long-distance dependencies. Seq2Seq models leveraged neural network power, particularly recurrent neural networks (RNN), to overcome the problems. Google's publication "Sequence to Sequence Learning with Neural Networks" presented the idea of the seq2seq model. This research paper discusses the architecture that serves as the foundation for tasks related to natural language processing. Encoder-decoder models are what the seq2seq models are.

## IV. PROPOSED SYSTEM

XLM-RoBERTa is an XLM extension that uses the Transformer model's RoBERTa architecture for pre-training. The RoBERTa architecture, an enhanced version of BERT (Bidirectional Encoder Representations from Transformers), serves as its foundation. XLM-RoBERTa enhances the capabilities of XLM and integrates the progress from RoBERTa to attain superior outcomes in downstream NLP tasks and multilingual language comprehension. The main aim of XLM-RoBERTa is to build a more potent and successful cross-lingual language model by utilizing the large-scale pre-training and intensive hyperparameter tweaking of RoBERTa's pre-training technique.

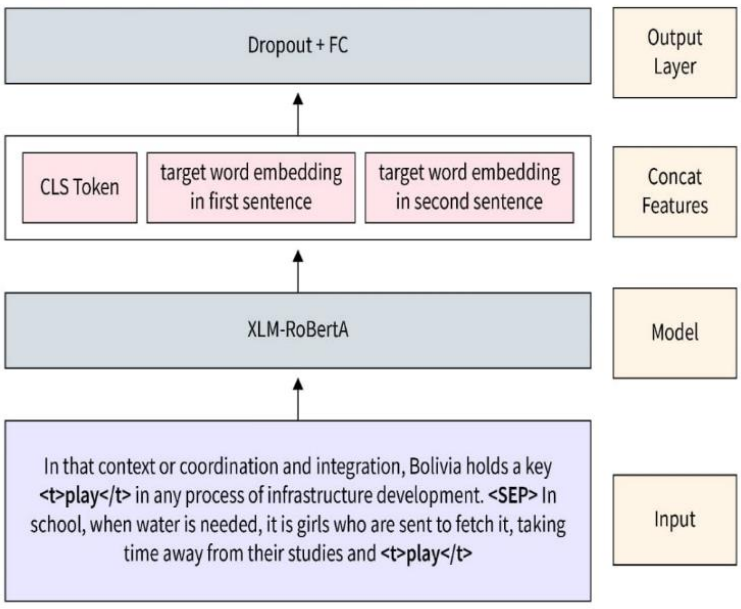



Fig: simple XLM-RoBERTa Structure

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 3, March 2024

DOI: 10.17148/IARJSET.2024.11333

## V. RESULTS

A. Training the model with Dataset

| | Tweet | Definitely English | Ambiguous | Definitely Not English | Code-Switched |
|---|---|---|---|---|---|
| 0 | Bugün bulusmami lazimdiii | 0 | 0 | 1 | 0 |
| 1 | Volkan konak adami tribe sokar yemin ederim :D | 0 | | | |
| 2 | Bed | 1 | | | |
| 3 | I felt my first flash of violence at some fool... | 1 | | | |
| 4 | Ladies drink and get in free till 10:30 | 1 | | | |
| 5 | @Melanynijholtxo ahhahahahah dm! | 0 | | | |
| 6 | Fuck | 0 | | | |
| 7 | Watching #Miranda On bbc1!!! @mermhart u r HIL... | 1 | | | |
| 8 | @StizZsti fino | 0 | | | |
| 9 | Shopping! (@ Kohl's) http://t.co/I8ZkQHT9 | 1 | | | |
| 10 | @Mizzh_ celos es tu segundo nombre | 0 | | | |
| 11 | @emilyroxxxx @melanienasson hahahahahah awww | 0 | | | |
| 12 | @BeGinkiTh otak ang semangat nak blajaq ! | 0 | | | |
| 13 | Quiero jugar tenis ♥ | 0 | | | |
| 14 | Yessss ^_^ | 1 | | | |

B. Training Data with and without using NLP

```
----------------------------------------------------
                Before Applying NLP
----------------------------------------------------

0                            Bugün bulusmami lazimdiii
1       Volkan konak adami tribe sokar yemin ederim :D
2                                                  Bed
3    I felt my first flash of violence at some fool...
4               Ladies drink and get in free till 10:30
5                     @Melanynijholtxo ahhahahahah dm!
6                                                 Fuck
7    Watching #Miranda On bbc1!!! @mermhart u r HIL...
8                                          @StizZsti fino
9             Shopping! (@ Kohl's) http://t.co/I8ZkQHT9
Name: Tweet, dtype: object
----------------------------------------------------
             After Applying NLP
----------------------------------------------------

0                         bugün bulusmami lazimdiii
1      volkan konak adami tribe sokar yemin ederim
2                                              bed
3    felt first flash violenc fool bump piti fool
4                           ladi drink get free till
5                   melanynijholtxo ahhahahahah dm
6                                             fuck
7            watch miranda mermhart u r hilari ❤💋
8                                    stizzsti fino
9                                        shop kohl
Name: Clean, dtype: object
```

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 3, March 2024

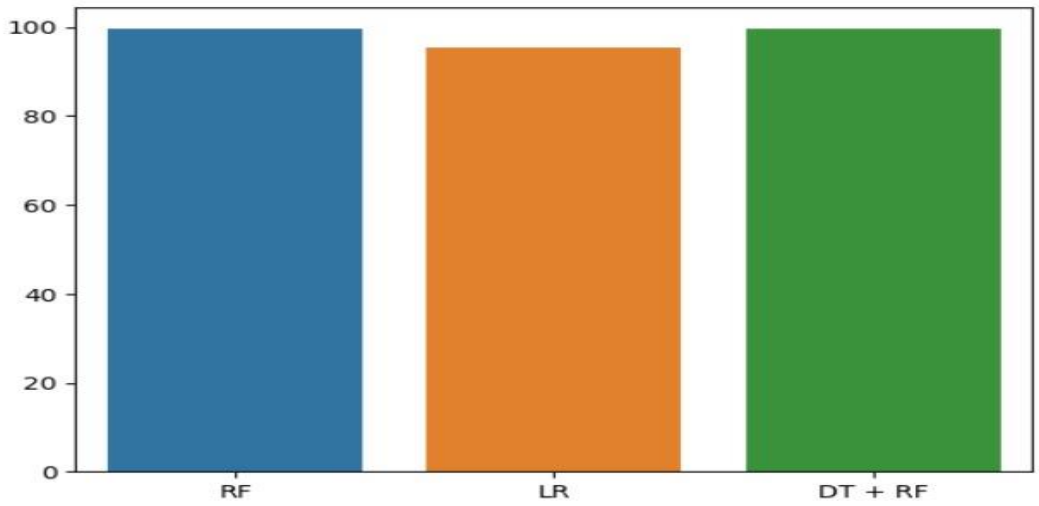
DOI: 10.17148/IARJSET.2024.11333

C. Encoding the Language Codes

```
Before Label Encoding
---------------------------------
0      TR
1      TR
2      NL
3      US
4      US
5      NL
6      US
7      GB
8      RS
9      US
10     MX
11     CA
12     MY
13     BR
14     US
Name: Country, dtype: object
---------------------------------
After Label Encoding
---------------------------------
0     114
1     114
2      82
3     119
4     119
5      82
6     119
7      41
8      99
9     119
10     77
11     20
12     78
13     16
14    119
Name: Country, dtype: int32
-----------------------------------------------------
                Before Applying NLP
```

D. Accuracy of Trained Models

```
1) Accuracy =  99.523866206404 %

2) Classification Report

              precision    recall  f1-score   support

           0       1.00      1.00      1.00      7810
           1       0.94      0.99      0.97       591

    accuracy                           1.00      8401
   macro avg       0.97      1.00      0.98      8401
weighted avg       1.00      1.00      1.00      8401


-----------------------------------------------------------
```



IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 3, March 2024

DOI: 10.17148/IARJSET.2024.11333

E. Vectorization

COUNT VECTORIZATION

```
(0, 3354)      1
(0, 3385)      1
(0, 13114)     1
(1, 25088)     1
(1, 12574)     1
(1, 223)       1
(1, 23992)     1
(1, 21990)     1
(1, 25943)     1
(1, 6810)      1
(2, 2284)      1
(3, 7921)      1
(3, 8103)      1
(3, 24998)     1
(3, 8213)      2
(3, 3393)      1
(3, 18551)     1
(4, 12868)     1
(4, 6537)      1
(4, 8336)      1
(4, 23586)     1
(5, 14650)     1
(5, 415)       1
(5, 6329)      1
(6, 8411)      1
:      :
(10498, 24187)         1
(10498, 8010)  1
(10499, 651)   1
(10499, 12620)         1
(10499, 13963)         1
(10499, 14022)         1
(10499, 23306)         1
(10499, 4658)  1
(10499, 17645)         1
(10499, 13710)         1
```

F. Language Translation

```
Enter the text to translate: iam student
Here are some examples of language codes:
1. English: `en`
2. French: `fr`
3. Spanish: `es`
4. German: `de`
5. Japanese: `ja`
6. Chinese (Simplified): `zh-CN`
7. Chinese (Traditional): `zh-TW`
8. Russian: `ru`
9. Arabic: `ar`
10. Italian: `it`
11. Portuguese: `pt`
12. Korean: `ko`
13. Hindi: `hi`
14. Turkish: `tr`
15. Dutch: `nl`
Enter the language code to translate to (e.g., 'fr' for French): de

Original text: iam student
Translated text: bereits Student
```

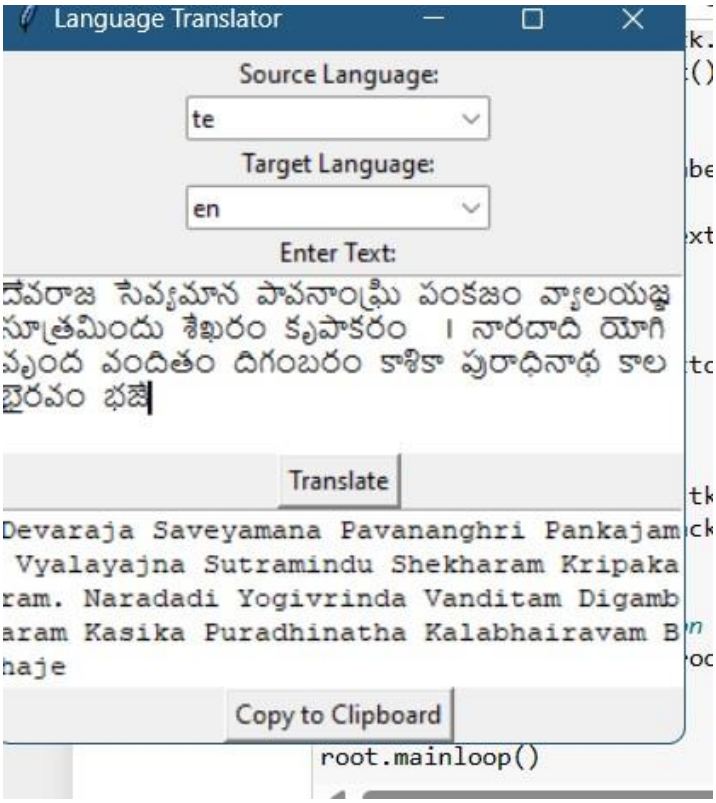
IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

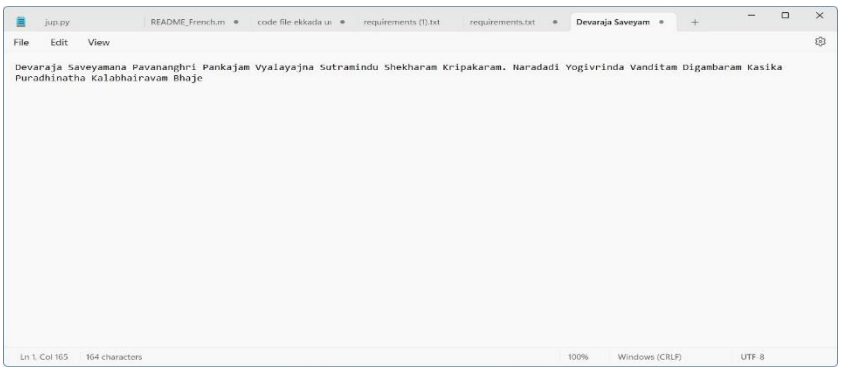
Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 3, March 2024

DOI: 10.17148/IARJSET.2024.11333

G. Translation to Different Language




H. Clip the Text




## VI. CONCLUSION

The CLIR eHealth IR tasks from 2013 to 2015 provided test collections that we could use for CLIR experiments in seven different languages: Czech, German, French, Spanish, Hungarian, Polish, and Swedish. On the other hand, our system's judgment rate was low and the human translations of the queries were lacking; therefore, in order to thoroughly assess our built systems, we carried out an extensive relevance assessment and manually translated the English inquiries to cover all the investigated languages. To facilitate more study into the problem, the extended test collection is accessible through the LINDAT repository and is licensed under the Creative Commons - Attribution-NonCommercial 4.0 license.

## REFERENCES

[1] Eneko Agirre, Llu'is Marquez, and Richard Wicentowski, editors. 2007. Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007).

[2] Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. Cloze driven pretraining of self-attention networks. arXiv preprint arXiv:1903.07785.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 3, March 2024

DOI: 10.17148/IARJSET.2024.11333

[3] Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second PASCAL recognising textual entailment challenge. In Proceedings of the second PASCAL challenges workshop on recognising textual entailment.

[4] Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth PASCAL recognizing textual entailment challenge.

[5] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In Empirical Methods in Natural Language Processing (EMNLP).

[6]William Chan, Nikita Kitaev, Kelvin Guu, Mitchell Stern, and Jakob Uszkoreit. 2019. KERMIT: Generative insertion-based modeling for sequences. arXiv preprint arXiv:1906.01604.

[7] Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment.

[8] Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In Advances in Neural Information Processing Systems (NIPS).

[9]Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In North American Association for Computational Linguistics (NAACL).

[10]William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In Proceedings of the International Work shop on Paraphrasing.

[11] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. arXiv preprint arXiv:1905.03197.

[12] Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing. https://data.quora.com/First Quora-Dataset-Release-Question Pairs

## BIOGRAPHY

**Mrs. G. Venkateswari**
M.Tech(Ph.D.), Asst.Professor, Head of Department
Department of Computer Science Engineering
Bapatla Women's Engineering College, Andhra Pradesh,
India

**P. Chandrika** B. Tech with Specialisation of Computer Science and Engineering in Bapatla Women's Engineering College, Bpatala

**SK. Nausheen** B. Tech with Specialisation of Computer Science and Engineering in Bapatla Women's Engineering College, Bpatala

**P. Manasa Veena** B. Tech with Specialisation of Computer Science and Engineering in Bapatla Women's Engineering College, Bpatala