iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45999

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

A Prototype of Android-Based Bi-Lingual Content Reference Tool

Soumya Majumdar
IIT Kharagpur

***Abstract:** Many dictionary applications (Online or offline) are available in the market but are not so helpful. They complicate the problem more than solve it in many cases; the dictionary refers to a complex solution, which leads to failure of referencing. Referencing tools are now not available in the market. A Bi-lingual tool can reference two contents of different languages in Paragraph, Line, and Word/Phrase levels.*

I. INTRODUCTION

India has 23 official languages, including 22 8th schedule languages and one additional official language, English. Those 22 vernacular languages are the native languages of India spoken by the Indian people. In the elementary level of Vernacular medium school, the medium of education is vernacular. That does not create many problems. But at a higher level, the medium of teaching is English. Unfortunately, we do not learn English well at the elementary level. So it creates problems at every level.

The problem usually arises because of having an impoverished vocabulary of the English language. We have two ways to overcome the problem; either learn the English language and enrich our vocabulary or use some tool which can reference the vernacular tongue with English vocabulary. If we have an English language content and the same content in vernacular language, then we can reference these two contents using some tool. In this way, we can understand essential topics written in English, and our fear of English will disappear. This tool is more helpful than a dictionary because we may not learn the proper use of English words. We need the knowledge of the practical use of words.

We built a prototype of an android based Bi-lingual (English-Hindi) content reference tool. In lower grade levels, the textbooks and medium of instruction delivery are in vernaculars. However, in upper-grade levels, English is the primary textbook language. As a result, many students find it challenging to cope with language changes. It happens due to the difference between these two languages.

This project aims at developing an android app for bi-lingual content reference to bridge the gap. Features of the app include-

(i) For a text segment selected by the student, the app will display similar text in another language.

(ii) The student will be able to select text in word phrase, sentence or paragraph level.

II. RELATED WORK

Bharati, A., V.Sriram, Krishna, A., Sangal, R., and S.M.Bendre proposed an algorithm for the alignment of sentences in bilingual corpora by lexical information[1]. The source and target language text will be divided into sub-parts called chunks. A chunk will be of two types- A noun chunk and a Verb chunk. A noun chunk is a non-recursive noun phrase. It consists of a determiner and an adjective followed by a noun. A verb chunk contains the main verb, supporting or auxiliary verbs and adverbs.

The objective of this algorithm is the identification of an appropriate translation for a specific sentence in the source language text among the sentences in the target language text, and it will be done by comparing of source sentence with the set of possible sentences that will be the translation of that source sentence.

A comparison score will be assigned for each such matching. This comparison score between the source and target sentences is determined by comparing the chunks of both sentences using an English-Hindi lexicon. Two chunks will be matched by first matching their headwords and then matching the support words of a chunk. Thus, the chunk matching will be done by looking inside a chunk, the words constituting the chunks. In a Noun chunk, all the words except the prepositions and postpositions are used to do chunk matching. In a Verb chunk, only the headword is used to do the chunk matching. Two words of a chunk can be matched by bi-lingual dictionary lookup, target-target dictionary lookup, numeric matching and phonetic matching. The alignment will be done after the comparison score is assigned, which can be obtained by matching the chunks. Different scoring functions can be used to calculate the score of the match.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

III. BACKGROUND

A. *Corpus*

A bi-lingual parallel corpus is needed for implementing our project. Because our project is totally dependable on parallel text. The parallel text consists of two texts in two different languages, where one is a translation of another. Examples of such parallel texts are- Loeb Classical Library and Clay Sanskrit Library. The multilingual Bible is also available, where one language is original, and one translation or several translations are for ease of study and comparison. The best example is Origen's Hexapla (Sixfold), where six versions of the Old Testament in different languages are side by side.

B. *Parallel Corpus*

A parallel corpus is a collection of texts. Each of which is translated into one or more other languages than the original[3]. Among these, the simplest case is where two languages only are involved- one of the corpora is an exact translation of the other. This type of corpus is called Bi-lingual Corpus. If more than two languages are involved, then this is called Multilingual Corpus.

The direction of translation is constant. Some texts in a parallel corpus can be translated from language A to language B. On the other hand; others can be translated in the reverse direction, i.e. from language B to language A. The direction of translation even may not be known.

At present, a parallel corpus is an object of interest because of the opportunity offered to align between original and translation and gain insights into the nature of translation.

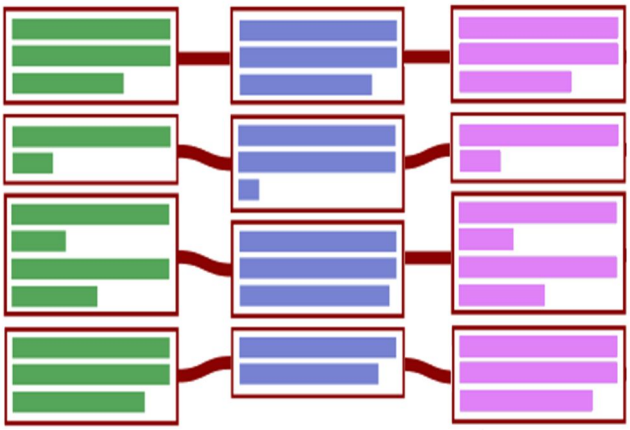


Figure 1: Multi-lingual corpus: here, one original language and two other languages exist

1) *Types of Parallel Corpora*

There are three types of parallel corpora-

a) *Noisy Parallel Corpus:* This type of corpus is not perfectly or poorly aligned.

b) *Comparable Corpus:* This type of corpus is built from non-sentence aligned and untranslated bilingual documents, but documents are topic-aligned.

c) *Quasi-comparable Corpus:* This type of corpus is built from very heterogeneous and non-parallel bilingual documents that may be or may not be topic-aligned.

Rarest Parallel Corpora are those corpora which contain translations of the same document in two or more languages. This corpus is aligned at least at the sentence level[3].

C. *Existing Bi-lingual Hindi-English Corpus*

1) *EMILLE/CIIL corpus[2]:* It is one of the first known corpora. It is created by a collaboration between Lancaster University and the Central Institute of Indian Languages, India, through the EMILLE project. It consists of texts in Hindi along with translations in Hindi, Bengali, Gujarati, Punjabi and Urdu. This corpus contains texts from different domains such as legal, education and health. This corpus consists of three components: monolingual, parallel and annotated corpora. It contains 92,799,000 words (including 2,627,000 words of arranged spoken data for Bengali, Gujarati, Hindi, Punjabi and Urdu). This parallel corpus consists of 200000 words in English and corresponding words in Hindi and other languages. A subset of this corpus was validated and released as part of the ACL (2005) shared task on word alignment.

2) *DARPA-TIDES Corpus:* It was released as a part of a language contest on SMT in 2002. After some manual refinement and cleaning, a subset of this corpus was released for the NLP Tools Contest Contest on SMT for English-Hindi by Venkatapathy in 2008. It is basically a dependency-based SMT system. The dependency-based framework in this corpus is best suited for translation between languages with free-word order, another characteristic of a few Indian languages like Hindi, Telugu and Marathi.
3) *EILMT and ILCI [2]:* Apart from the corpus mentioned above, the creation of large-scale multilingual parallel corpora for English, Hindi and other Indian languages has been part of two major projects-EILMT (English to Indian Languages MT) and ILCI (Indian Languages Corpora Initiative). Both projects basically focused on collecting two domains- health and travel. In the case of the EILMT, bilingual lexica have been created for both these domains and contain domain-specific term translations and multi-word expressions. On the other hand, ILCI provides parallel corpora with part-of-speech tags created by linguistic annotators. Both the projects are initiatives by the Department of Information and Technology (DIT) of India, handled by different participating institutions.

IV. CONTENT REFERENCES IN DIFFERENT LEVELS

A. Architecture of Our Model

Our project architecture will have a total of four types of modules. Paragraph Aligner, Sentence Aligner, Differentiators of Subject-Verb-Object, Word/Phrase Aligner. Paragraph Aligner takes some English and some Hindi paragraphs, and its output is aligned. Now for each paragraph, there are two sets of English and Hindi sets which are not aligned. So those two sets are the inputs of the Sentence aligner. The sentence aligner gives aligned sentences as output. Now each aligned sentence pair will go to the Subject-Object-verb differentiators. It differentiates the Subject, Object and Verb of English and Hindi sentences. It gives six outputs: Subject, Object and Verb of English sentence and Hindi sentence. Now there are three types of Word aligners. One word aligner will take the subject of a Hindi sentence and the subject of an English sentence and give aligned words or phrases contained in the subject as output. The other two Word aligners do the same job for Object and Verb. This project produces aligned paragraphs, sentences, and words or phrases.

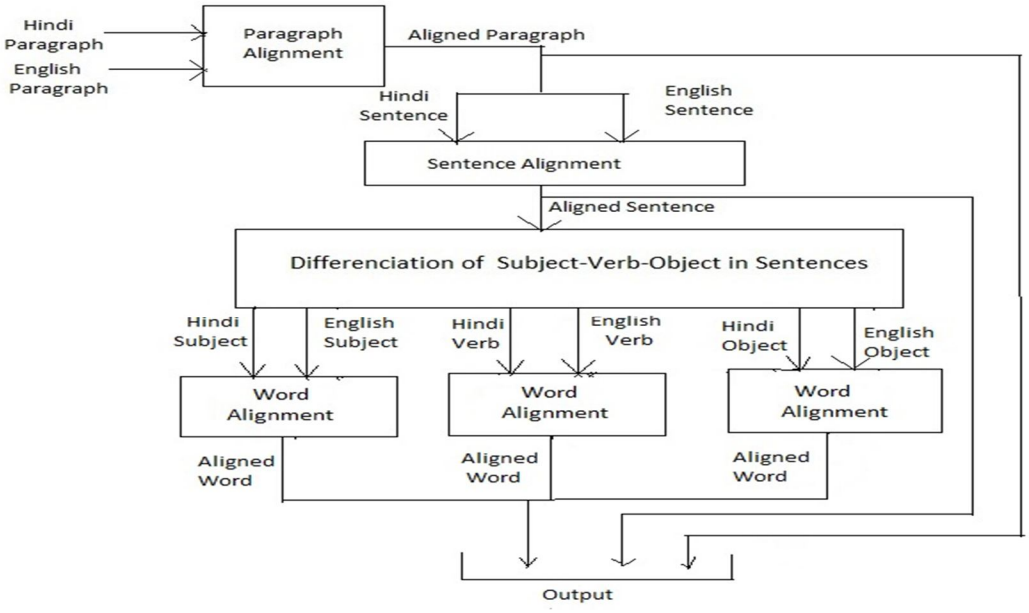


Figure 2: The architecture of Project

B. Content Reference at the Paragraph Level

We can reference two contents at the Paragraph level using the no of lines the paragraph has. Two same paragraphs on two different levels usually have the same no lines in them. But there may be some cases where two paragraphs are the same and other languages. In this case, ambiguity arises. So if this happens, we will check to stop words like numbers, punctuation, colon, semicolons etc. If matching is found between two paragraphs, then these two paragraphs will be aligned to each other.

C. Content Reference at Line Level

After aligning at the paragraph level, we need to align at the sentence level. It means sentences of alignment paragraphs will be aligned to each other. We proposed an algorithm for this job.

1) Proposed Algorithm for line alignment

We designed an algorithm which can align to map two sentences.

a) Input: Two paragraphs which have sentences in unaligned form.

b) Output: Two paragraphs which have sentences in aligned form.

Algorithm

- For a sentence in an English document-
  - Translate it to Hindi and tokenise it.
  - Now, tokenise each sentence in the Hindi document.
  - For each sentence in the Hindi document, match them. We will take a variable named matching score and initialise it by zero. If a common token is found, then matching is there. For each matching, the matching score will be incremented by 1. Repeat it for each sentence in the Hindi document.
  - The sentence with a maximum matching score will be aligned to the English sentence.
  - If more than one sentence exists with the highest matching score, then-
    - Take the length of each sentence with the highest matching score.
    - Take the length of Hindi translated English sentence.
    - Take the difference between the length of a Hindi sentence and Hindi translated English sentence. Repeat it for each sentence with the highest matching score.
    - This sentence will be aligned to an English sentence with less length difference than Hindi-translated English.
- Repeat step 1 for each sentence in the document.

2) Explanation and Result from Algorithm

This Algorithm aligns an English sentence to Hindi. Suppose there is an English sentence- "Children should be supported to carry out observation activities which require patience." And there are some Hindi sentences. Now the translation of this English sentence will be a Hindi sentence. This is a literal translation. So exact text should not be in Hindi documents. So we tokenise this Hindi translation and those Hindi sentences and do matching. For each Hindi text, there will be a matching score. The matching score will be initialised by 0, e.g.:- if there is no matching, the score will be zero. If any matching is there, the score will be implemented by 1. In this case, the matching score will be 0,1,3,0,3, respectively. The maximum matching score is 3 in this case, but we get it for two sentences, so we have to make the difference in their length and translated sentences. For the first sentence, it is 21, and for the second sentence, it is 1. Sentences with lengths with minimum length difference will be aligned. So 2nd sentence will be aligned with "Children should be supported to carry out observation activities which require patience.".

3) Time Complexity

If the number of sentences in each corpus is N, then the average time complexity of the proposed algorithm is- $O(N^2)$

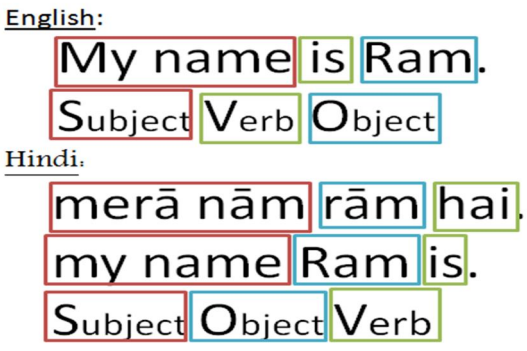


Figure 3: English and Hindi sentence structure-Example 1

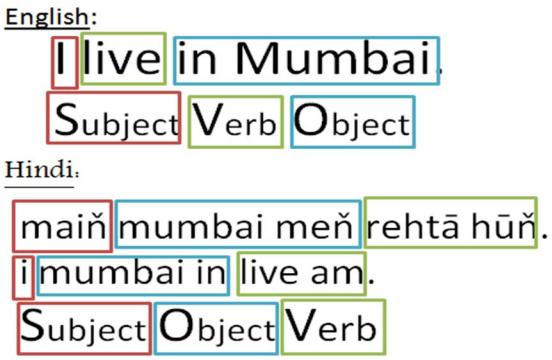


Figure 4: English and Hindi sentence structure-Example 2

D. Content Reference in Word level

In this level, we will align words in a sentence where sentences are already aligned. A sentence consists of words and phrases, which is an association of some words to convey specific meanings.

1) An Approach to Aligning Words

We know that a sentence consists of three things- Subject, Verb and Object. This Subject-Verb-Object triplet happens in a different order in different languages. In English, the order is Subject-Verb-Object; but in Hindi, the order is Subject-Object-Verb. In English, the verb exists in the middle of the sentence; but in Hindi, it exists at the end.

Now we will identify the subject, verb and object in English and Hindi sentences. Those two lines are the same but in different languages, so their subject, object and verb will be the same. So we can map them with each other.

Now in subject and object, there will be some parts of speeches like Noun, Adjective, and Adverb. We will tag those Parts of Speech and then mapped with each other. Because we know that if an English sentence has Part Of Speeches like Noun, Adjective, or Adverb, then the same sentence in Hindi has those Part of Speeches. So we can map them easily.

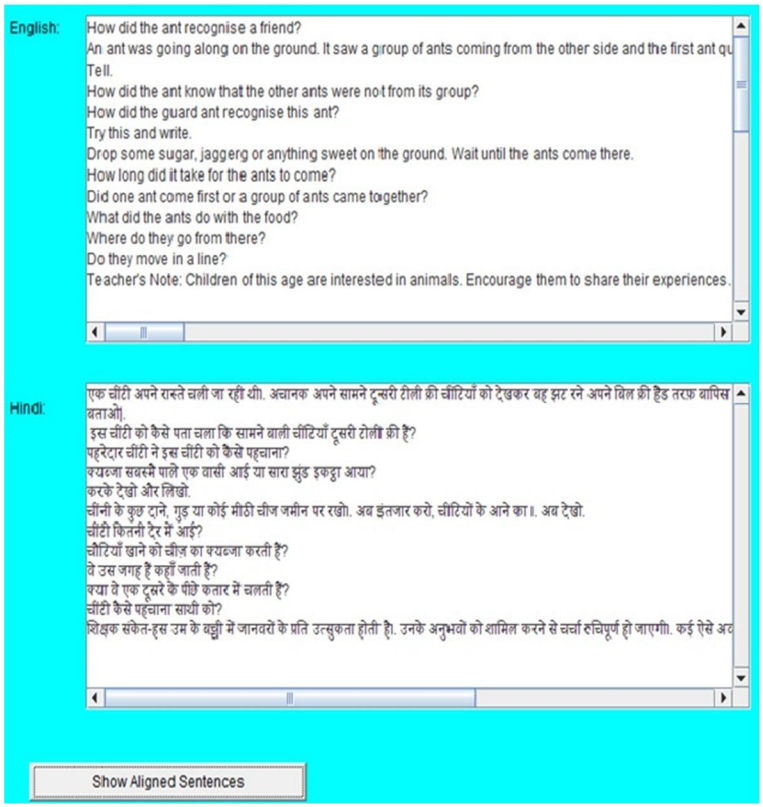


Figure 5: Two paragraphs before alignment

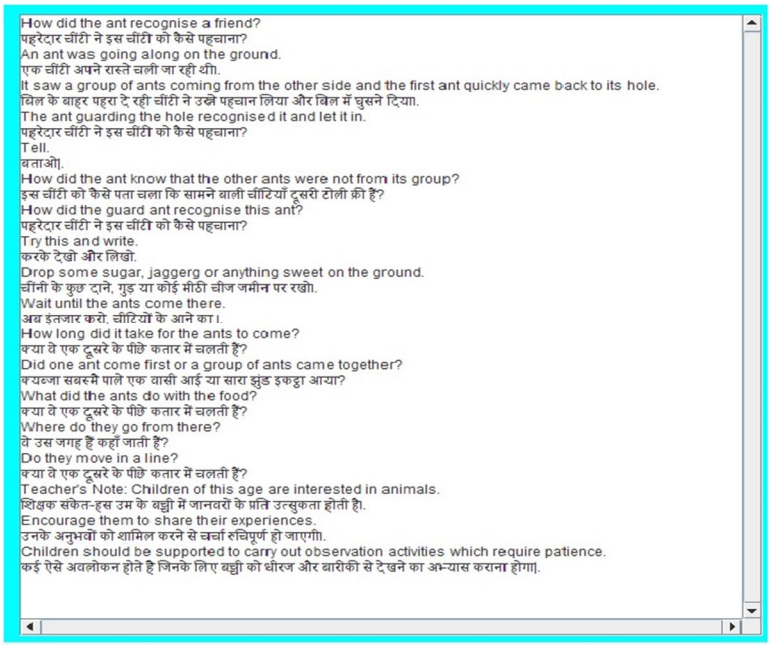


Figure 6: Two paragraphs after alignment

V. EXAMPLE AND RESULT

A. *Performance of Algorithm*

The algorithm's accuracy varies with different contents. For some content, accuracy is more. For some content, accuracy is less. Accuracy increases when one corpus is the proper translation of another corpus. The average accuracy of this algorithm is approximately 78%.

Sample	No of Sentences	Correctly Aligned	Accuracy
Sample 1	18	15	84%
Sample 2	17	12	71%
Sample 3	10	7	70%

Table 1: Performance algorithm on different sample

B. *DOM structure of the Aligned Contents*

<Chapter1>
<Paragraph1>
<Line1>
<WholeLine>
<English>...</English>
<Hindi>...</Hindi>
</WholeLine>
<Subject>
<English>...</English>
<Hindi>...</Hindi>
</Subject>
<Verb>
<English>...</English>
<Hindi>...</Hindi>

</Verb>
<Object>
<English>...</English>
<Hindi>...</Hindi>
</Object>
</Line1>

<Line2>
.....
.....
</Line2>
.....
.....
</Paragraph1>
.....
.....
</Chapter1>

```
<Line1>
<WholeLine>
<English>you did well.</English>
<Hindi>तुमने    अच्छा       किया।     </Hindi>
</WholeLine>
<Subject><English> you</English><Hindi> तुमने   </Hindi></Subject>
<Verb><English> did</English><Hindi> किया।     </Hindi></Werb>
<Object><English> well</English><Hindi> अच्छा     </Hindi></Object>
</Line1>
```

Figure 7: Aligned word in XML format

C. Factors Affecting Word Alignment

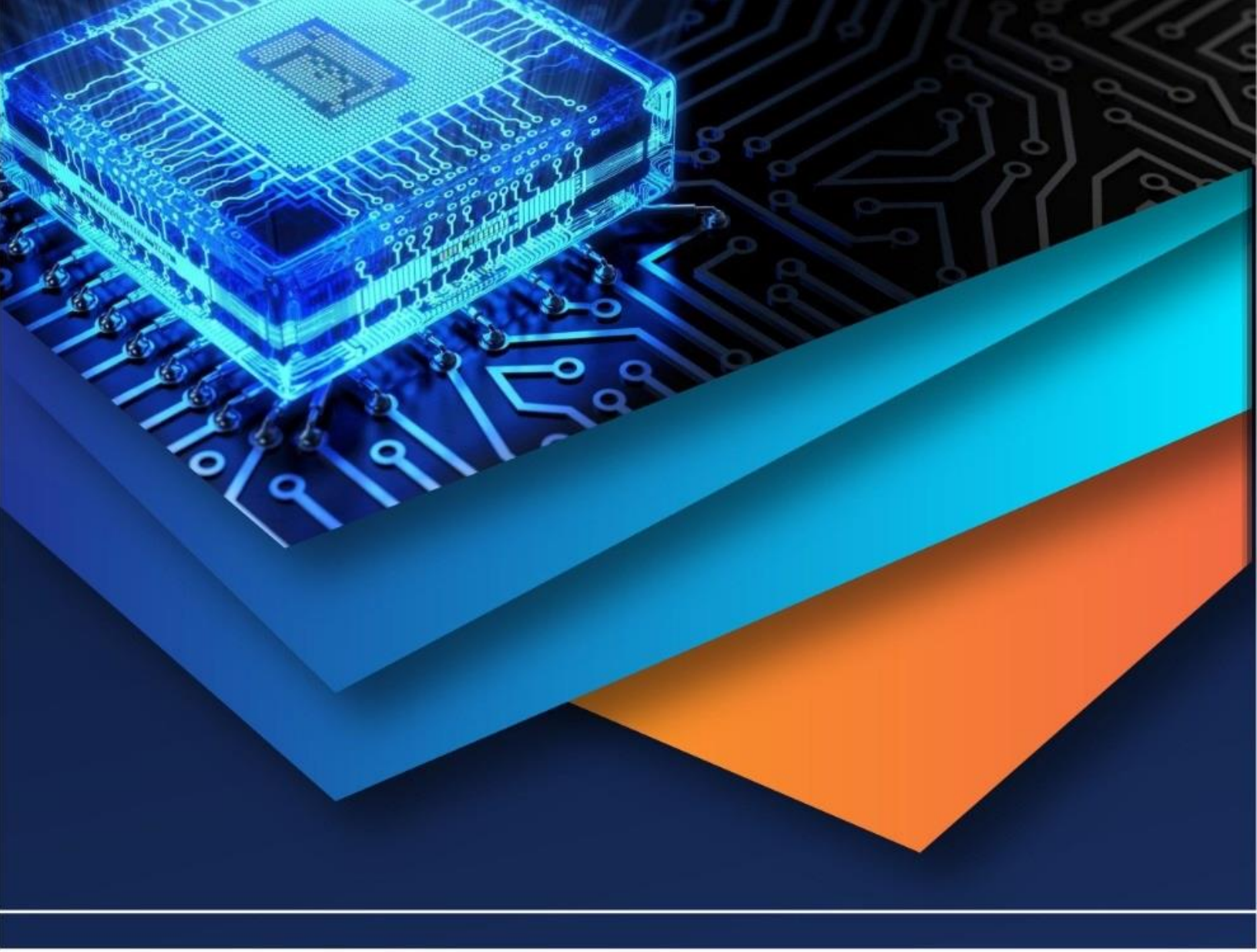
Proposed word aligning approach is based on Part of speech tagging of words in the sentences. So proper tagging of part of speech is important. There are good collections of POS taggers in the English language but not in the Hindi language. So POS tagging in Hindi doesn't give good results all time. It affects the word alignment job heavily because the success of this approach is based on appropriate POS tagging.

VI. CONCLUSION

Objective of this project is noble. After making this app, it will be helpful to students a lot. Where a dictionary app doesn't make a solution, our app can be a great solution for practical English learning because dictionaries may contain words that are no longer used in a practical sense. And on the other hand, practically used words may not get emphasis in the dictionary. Here our app will give students practical references. Students can learn new words and phrases and use this tool practically. They can also understand their study topic easily. This app can bridge the gap between English and other vernacular languages. Primarily this app is helpful for Hindi language people only. But later, it can be helpful for other vernacular language people also.

REFERENCES

[1] Bharati, A., Sriram, V., Krishna, A. V., Sangal, R., & Bendre, S. M. (2003). An algorithm for aligning sentences in bilingual corpora using lexical information. arXiv preprint cs/0302014.
[2] Yeka, J. R., Kolachina, P., & Sharma, D. M. (2014, May). Benchmarking of English-Hindi parallel corpora. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) (pp. 1812-1818).
[3] Jindal, K., & Goyal, V. (2010, July). Improved algorithm for automatic word alignment for hindi-punjabi parallel corpus. In International Conference on Data Engineering and Management (pp. 255-263). Springer, Berlin, Heidelberg.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY