iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VI **Month of publication:** June 2017

DOI:

www.ijraset.com

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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Information Extraction Using Enhanced N-Ary Relation Based IR Algorithm

S. R. Tandan[1], Priyanka Tripathi[2], Rohit MIRi[3]
[1,2] *Dept of Computer Science and Engineering, Dr. C.V. Raman University, Kargi Road Kota, Bilaspur, CG, INDIA*
[3] *National Institute of Technical Teacher Training, Bhopal, MP, INDIA*

Abstract: In this paper, N-ary relation based open domain question answering system for Extraction Information from an oversized assortment of document against arbitrary questions has been presented. We proposed two algorithms to extract entity and relationship from string and to extract answer for queried question. Our proposed algorithm works on both online and offline mode with greater accuracy.

Keywords-Ary Relation, Open Domain, Question Answering System, Knowledge Base, Information Extraction, NLP

I. INTRODUCTION

Question answering system came into news when [1] IBM's question answering system, Watson, defeated the two greatest Jeopardy champions, Brad Rutter and Ken Jennings, by a significant margin. It is typically a computer program that can answer in natural language, of a natural language question.

B. Question Answering System

Question Answering (QA) is a computer science discipline of information retrieval and natural language processing (NLP). The discipline deals with building a system which can analyze a query in human's natural language and can answer automatically in the form of natural language. Question Answering System (QAS) deals with a variety of question types, such as what, how, when, where, what, hypothetical, cross-lingual, semantically constrained etc.

C. Classification of Question Answering System

Question answering system can be classified into various grounds. It can be classified as:

1) Based on Domain of Questions[2, 3]: Closed Domain Question Answering System: This kind of QAS deals with specific domain of topic of interest. e.g. question based on medical queries. These kinds of QAS are easy to implement [4]. Natural Language processing (NLP) system uses domain specific knowledge for extraction of answers. Sometimes closed domains refer to only specific type of question [5]. It may be descriptive rather than procedural.

Open Domain Question Answering System: These kinds of QAS are made to deal with almost any kind of questions [4, 6]. The system relies on the knowledge base which may provide desIRed information such as local text, web pages, other databases etc. It requIRes bigger knowledge base.

2) Based on Response: Question Answering System: eg. yahoo answer, forum, wiki answer et Automatic Question Answering System: eg. IBM Watson

3) Based on Interaction: Interactive Question Answering System Non-Interactive Question Answering System

4) Problem Decription and Solution strategy: We are making effort to overcome the following disadvantages of available solutions:

1) Relying on the offline data so that updated answer can be retrieve
2) Storing data locally is overhead
3) Speed is l
4) To overcome above Problem we propose following
5) Do not rely on offline data instead utilize World Wide Web.
6) To increase speed retrieve filtered page from a search engine it will save searching time complexity
7) Perform scoring locally and make simple. No need to depend on the heavy algorithm.
8) We have modeled N-ary query and answer model.

II. RELATED WORK

Question Answering System is studied by many researchers. Since it is not much older field, we find relatively small literature on it. Some important studies are as follows:

(Afader et al., 2013) [7] Studied question answering as a machine learning problem, and induce a function that maps open-domain questions to queries over a database of web extractions. The trained the function that will map a natural language question to a query over a database D. The precision was ~77%. (NIRanjan et al., 2012) [8] Introduced the Rel-grams language model, which is analogous to an n-grams model, but is computed over relations rather than over words.

(Alan Ritter et al., 2012) [9] Presented a scalable and open-domain approach to extracting and categorizing events from status messages of Twitter users. They used basian technique to extract event based information from twitter. The extraction is a 4-tuple representation of events which includes a named entity, event phrase, calendar date, and event type.(Etzioni et al., 2011) [10] Described the extraction of unstructured data based on structure of English grammar. They introduced Open Information Extraction paradigm which is basis for question-answering system.

(Day C. Wimalasuriya et al., 2010) [11] Introduced Ontology Based Information Extraction using different classification techniques support vector machines (SVM), maximum entropy models and decision trees have been used in IE.

(Thomas Lin et al., 2010) described information extraction as common sense and are denoted by f (a, b) where f is relation between attribute a and b. They have employed Open Information Extraction (Open IE) for relation extraction.

(Michele Banko et al., 2007) [34] performed experiments over a 9,000,000 Web page corpus that compare TextRunner with KnowItAll, a state-of-the-art Web IE system. Text Runner achieved an error reduction of 33% on a comparable set of extractions.

(Bill Dolanet al., 2004) [13] Described unsupervised techniques for acquIRing monolingual sentence-level paraphrases from a corpus of temporally and topically clustered news articles collected from thousands of web-based news sources. They employed two techniques: (1) simple string edit distance, and (2) a heuristic strategy that paIRs initial (presumably summary) sentences.

A. ARCHITECURE OF QUESTION ANSWERING SYSTEM

1) Component of a Question Answering System: Typically a QAS has following component

a) Question Classifier
b) Document retrieval Component
c) Filter
d) Answer extraction Component
e) Knowledge Base

B. Question Classifier

Question classifier module take input a question in natural language. It determines class and types of question and class and type of answer. After the analysis of question different NLP techniques are applied over the input.

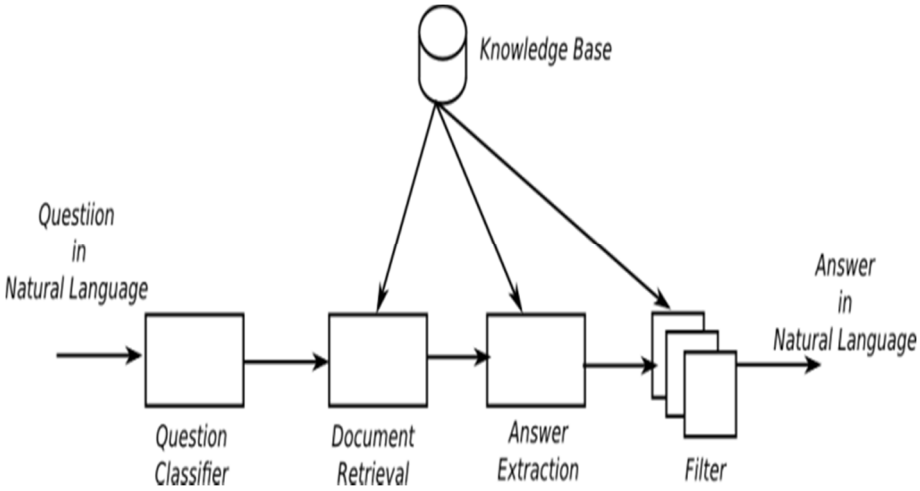


FIGURE: 1.1 ARCHITECTURE OF A TYPICAL QAS.

1) *Document retrieval Component:* Document retrieval module or component exploits search engines to find out relevant documents that may contain possible answer.
2) *Filter:* Filter module select highly relevant document and leave trivial and non relevant documents. Then it passes output to the Answer extraction component of the question answering system
3) *Answer extraction Component:* Answer extraction component looks for answer into text depending on different context, ambiguity removing techniques and scheme of the answer. Thus the system is able to answer most of the question
4) *Knowledge Base:* Knowledge Base is the database of the fact, document etc. It may be web document, text document or text inside database. The accuracy of the system heavily depends on the correctness and completeness of the knowledge base.

III. METHODOLOGY

A. Stemming

Stemming is that the term utilized in linguistic morphology and knowledge retrieval to explain the method for reducing inflected (or typically derived) words to theIR word stem, base or root type typically a word type. several search engines treat words with constant stem as synonyms as a sort of question enlargement, a method referred to as conflation. For stemming purpose we have used tools these are Pling stemmer and snowball tools. Pling Stemmer stems associate degree English noun (plural or singular) to its singular. [15]

B. POS tagging

In corpus etymology, grammatical kind labeling (POS labeling or POST), likewise known as grammar labeling or word-classification elucidation, is that the procedure of checking up a word in an exceedingly content (corpus) as with reference to a selected grammatical feature, taking under consideration each its definition, and conjointly its setting i.e. association with near And connected words in an expression, sentence, or section.[16]

C. NER

Named-element recognition (NER) (called substance ID, element piecing and element extraction) is a subtask of data extraction that looks to find and arrange components in content into predefined classifications, for example, the names of persons, associations, areas, articulations of times, amounts, fiscal qualities, rates, and so forth we have utilized Standford NER(Named Entity Recognizer) tools for this Purpose.[14]

D. WordNet

"WordNet is a semantic vocabulary for the English dialect. It aggregates English words into sets of equivalent words called synsets, gives short, general definitions, and records the different semantic relations between these equivalent word sets. The reason for existing is twofold: to deliver a mix of word reference and thesaurus that is all the more naturally usable, and to bolster programmed content investigation and computerized reasoning applications. WordNet recognizes things, verbs, modifiers and qualifiers in light of the fact that they take after diverse syntactic standards. [18]

E. NLP

Natural language processing (NLP) is a field of software engineering, man made brainpower(Artificial Intelligence), and computational etymology concerned with the collaborations in the middle of PCs and human (common) languages. All things considered, NLP is identified with the territory of human PC association. Numerous difficulties in NLP include common language understanding, that is, empowering PCs to get importance from human or characteristic language data, and others include regular language era.[17]

IV. PROPOSED WORK

(Afader et al., 2013) [7] The researcher has given all focuses on binary relation only. The complex questions have generally N-ary relation with more than two entity participating. If we consider more entities in the question the scoring becomes more precise. Our goal is to give a model for N-ary relation and faster algorithm to find better answer.

A. Models

The component of the question and answer are defined as follows:
Entity Set E = { e_1, e_2, e_3, . . . e_n}

Relation Set R = { $r_1, r_2, r_3, \ldots r_m$ }
Knowledge Base K = E × R ×E
The query model:
Query Q = ($e_{q1}, e_{q2}, e_{q3} \ldots e_{ql}$, r) where $e_{qi} \in E$ for i = 1, 2, 3, . . . l and r ∈ R
The answer model Answer A = {$e_{a1}, e_{a2}, e_{a3}, \ldots e_{aj}$}
Where $e_{qi} \in E$ and $e_{ai} \in A \Lambda \forall e_{qk} \in Q \Lambda r \in Q \mid (e_{ai}, e_{qk}, r) \in K$

B. Proposed Algorithms
We proposing two algorithms
extract entities and relation from string
To extract answer for queried question

1) Entity relationship extractor ERE(S)

a) Input: A string S = { $w_1, w_2, w_3 \ldots w_n$} Where w_i is word in the string and i = 1, 2, 3. . . . n

b) Output:
Knowledge Base Tuple ε = ($e_{s1}, e_{s2}, e_{s3} \ldots e_{n-1}$, r) Where e_{si} is either entity corresponding extracted from string S and r is Relation among entities extracted from string S

c) Algorithm

i) n ← number of words in input string
ii) St[n] is set of stemmer
iii) **for** i ← 1 to n
iv) St[i] ← StemOf(S[i])
v) **end for**
vi) Pt[n] is set of POS(Parts of Speech) Tag
vii) **for** i ← 1 to n
viii) St[i] POSTOf(St[i])
ix) end for
x) ε [n] is Knowledge base tuple
xi) j ← 1
xii) k ← n
xiii) for i ← 1 to n
xiv) if St[i] ← is noun then
xv) ε[j] = S[i]
xvi) j ← j + 1
xvii) else
xviii) ε[n] = S[i]
xix) k ← k -1
xx) end if
xxi) end for
xxii) return ε

C. AnswerExtraction (AE)

1) Input: A question String Qs = { $w_1, w_2, w_3 \ldots w_n$} Where wi is word in the string and i= 1, 2, 3. . . n
2) Output: Answer Entity Ae
3) Algorithm:

a) Query tuple ε_Q ← ERE(Qs)
b) P[ ] ← is set of text paragraph extracted by search engine
c) Plength ← length of P
d) HighestScore ← 0
e) **for** page in P

f) **for** String in page
g) Score ← 0
h) Entity Relation tuple ε_s ← ERE(String)
i) Score ← number of matched entity in ε_Q and ε_s
j) if HighestScore < Score then
k) HighestScore ← Score
l) Ae ← ε_s - ε_Q
m) end if
n) end for
o) end for
p) return A_e

D. Experiment

We have simulated experiment using Bing API free subscription of 5000 by default data set for online mode and for offline mode we are utilizing freely available database of English Wikipedia [19]. We processed it so that search engine can index it. We are locally installing open source freeware search engine INDRI to index the Wikipedia data.

For evaluation and testing we have used TREC data for both online and offline mode. TREC is good source of question and answer list.

TABLE I
THE CONFIGURATION OF THE SYSTEM

Configuration	Offline Mode	Online Mode
Data source	English Wikipedia dump	Bing API
Test data	TREC list of question and answer	TREC list of question and answer
Search Engine	INDRI	Bing Search engine
Stemmer	Pling Stemmer	Pling Stemmer
POSTagger	Standford POST	Standford POST
NLP Toolkit	Standford NLP	Standford NLP
Name entity recognizer	Standford NER	Standford NER
Wordnet Library	JWNL Wordnet Library	JWNL Wordnet Library
Processor	Intel Core i3	Intel Core i3
Clock rate	2.40GHz	2.40GHz
RAM	4GB	4GB

V. RESULT AND DISCUSSION

Based on the experimental work carried out for proposed N-ary relation based system in previous chapter. The work has been compared with existing question answering system for correctly classify the outcomes of developed approach.

A. Existing system handled Boolean query using REVERB database which is outdated and IRrelevant for complex query.
B. In existing system total 698 questions had been tested by creating total 37 question clusters.
C. Further, they did not mention that how many question correctly answered/not answered/ invalid questions.

D. In the existing system, proposed work observed that Wiki Answer are either too complex to be mapped to a simple relational query, or are not covered by the REVERB database. Further, approximately one thIRd of the missing answer is due to entity and relation recognition error.

Error Distribution Of Paralex On A Restricted Sample Of Question From The Wiki Answer Dataset

Incorrectly answered/ unanswered questions: (Existing System)

36%	Complex question
	Need N-ary or higher order relation (14%)
	Answer is a set instruction (13%)
	Need database operators e.g. Joins (9%)
32 %	Entity or relation recognition errors
	Entity recognition errors (13%)
	Relation recognition errors (12%)
	Entity and relation recognition errors (7%)
28 %	Incomplete Database
	Derived a correct query, but no answers.
4 %	Typos/ Inscrutable Questions

Developed system outperform for different group of query term (question). N-ary based approach easily deals with query which is written in complex form such “How long does it take to drive from Bilaspur to Raipur”

E. Error rate analysis

S.NO.	Category of Question	Existing System	Developed System
1	Complex Question	36 %	27 %
2	Entity –Relation Based Question	32 %	20 %
3	Incomplete Database	28 %	18 %
4	Inscrutable Question	4 %	17 %

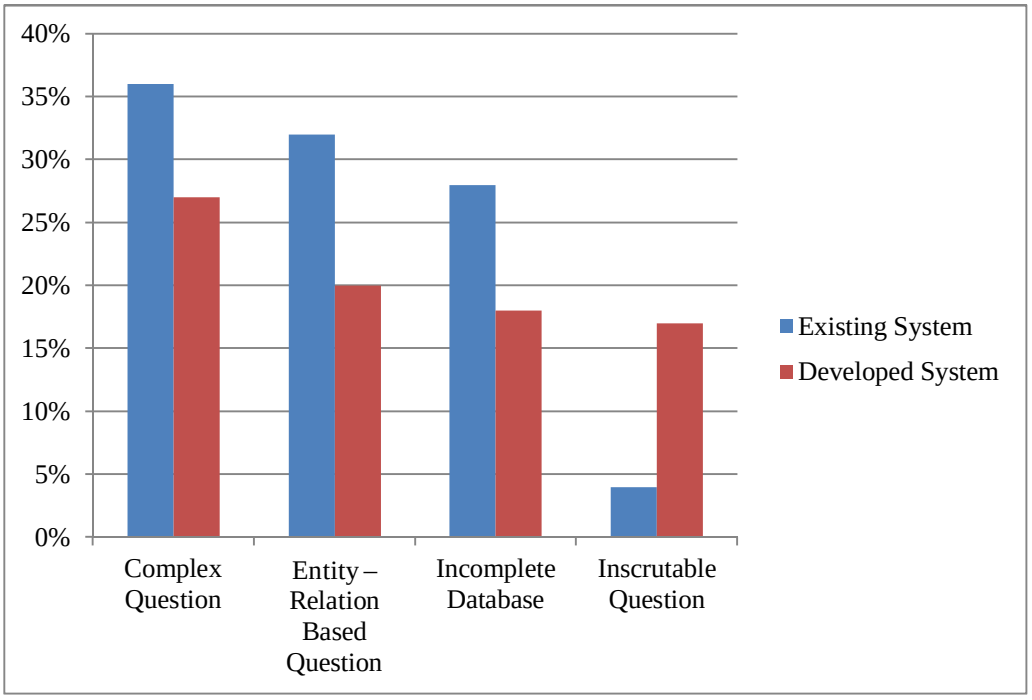


Figure 1.2 Performance comparison existing Vs Developed system

F. Developed System Performances Over Number Of Question Tested

Total number of question tested = 987

S.NO.	Category of Question	Correctly Answered	Not Correctly Answered	Total
1	Complex Question	201	75	276
2	Entity –Relation Based Question	203	51	254
3	Incomplete Database	213	47	260
4	Inscrutable Question	164	33	197

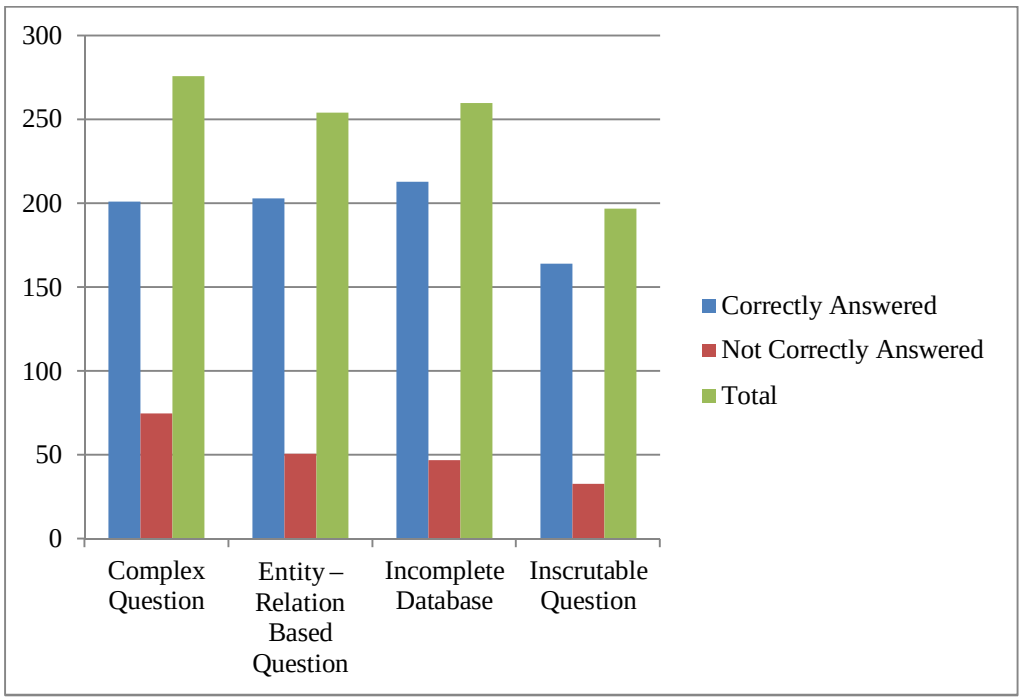


Figure 1.3 Developed system query correctly answered / unanswered

The performance of our system is satisfactory as is higher than performance of base paper the bench mark performance of binary relation based on paraphrased. In this thesis achieve higher performance by increasing size of Knowledge base.

VI. CONCLUSION

The proposed N-ary model for open domain Question answering system is novel approach for information extraction. Our proposed algorithm produced more accurate result for fIRed query term. We replaced answer searching and matching algorithm of Ephyra by our proposed algorithm. Implementation of algorithm is easy and reducing the time complexity.

REFERENCES

[1] Jeopardy watson, ibm. 2011. URL http://www.nytimes.com/2011/02/17/ science/17jeopardy-watson.html?_r=0.

[2] Alfonso Valencia Roser Morante, Martin Krallinger andWalter Daelemans. Machine reading of biomedical texts about alzheimer's disease. CLEF 2012 Evaluation Labs and Workshop, September 2012.

[3] L. HIRschman and R Gaizauskas. Natural language question answering. the view from here. Natural Language Engineering (2001), pages 275{300, 2001.

[4] Pampapathi R Galitsky B. Can many agents answer questions better than one.FIRst Monday, 2005.

[5] Boris Galitsky. Natural language question answering system: Technique of semantic headers. International Series on Advanced Intelligence, 2, 2003.

[6] J Lin. The web as a resource for question answering: Perspectives and challenges.Proceedings of the ThIRd International Conference on Language Resources and Evaluation (LREC 2002), 2002.

[7] AFader, L. Zettlemoyer and O. Etzioni, "Paraphrase-driven learning for open question answering." Sofia, Bulgaria, pp. 16081618, 2013

[8] Alan Ritter, Mausam, Oren Etzioni and Sam Clark, "Open Domain Event Extraction from Twitter", Knowledge Discovery and Data Mining, 2012.

[9] NIRanjan Balasubramanian, Stephen Soderland and Mausam , "Rel-grams: A Probabilistic Model of Relations in Text", Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction , 2012.

[10] Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland and Mausam, "Open Information Extraction: the Second Generation", International Joint Conference on Artificial Intelligence, 2011

[11] Daya C. Wimalasuriya and Dejing Dou., "Ontology-based information extraction: An introduction and a survey of current approaches.", J. Inf. Sci. 36, 3 (June 2010), 306-323.

[12] Michele Banko, Oren Etzioni, Stephen Soderland, and Daniel S. Weld. "Open information extraction from the web. Commun.", ACM 51, 12 (December 2008), 68-74.

[13] Bill Dolan, Chris QuIRk, and Chris Brockett. "Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources." In Proceedings of the 20th international conference on Computational Linguistics (COLING '04). Association for Computational Linguistics, Stroudsburg, PA, USA, , Article 350, 2004

[14] Named Entity Recognizer NER.http://en.wikipedia.org/wiki/Named-entity_recognition

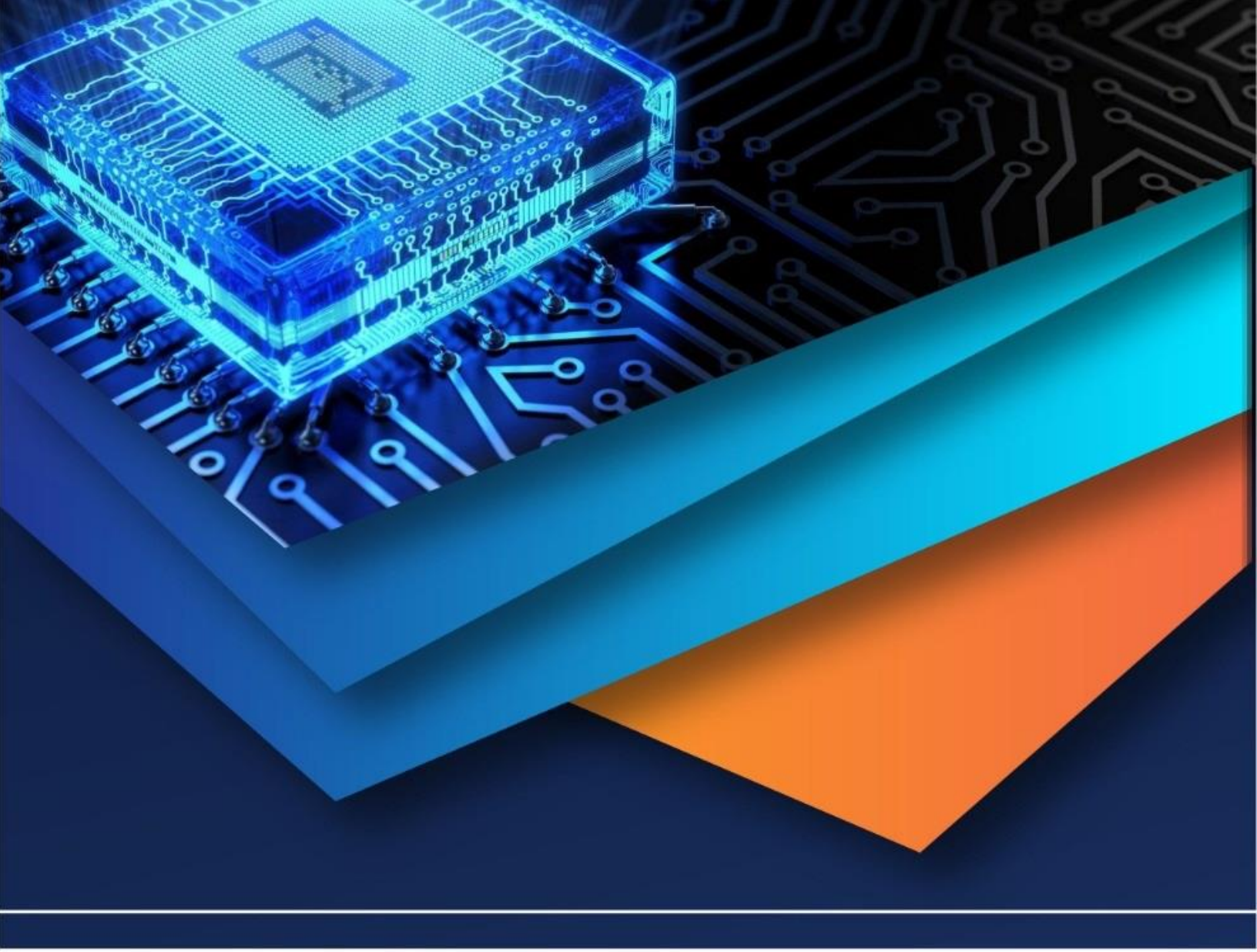
[15] Stemmer.http://en.wikipedia.org/wiki/Stemming

[16] POS Tagger. http://en.wikipedia.org/wiki/Part-of-speech_tagging

[17] Natural language processing. https://en.wikipedia.org/ wiki/Natural_language_processing

[18] WordNet. http://en.wikipedia.org/wiki/WordNet

[19] Wikipedia:Database Download https://en.wikipedia.org/wiki/ Wikipedia:Database_download



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY