Programming Language Conversion using NLP

ABHIJIT BANERJEE[1], MADHUBAN MUKHERJEE[1], APARAJITA BANERJEE[1], MD. AALISHAN RAZA[2] , SUCHETA BHOWMICK[1],SAKSHI BHAGAT[1], SUDIPTA BASU PAL[3]

[1]Department of CST

University of Engineering and Management ,Kolkata

India

[2]Department of CSIT

University of Engineering and Management ,Kolkata

India

[3]Department of CST and CSIT

University of Engineering and Management,Kolkata

India

banerjeeabhijit111@gmail.com, madhubanmukherjee77@gmail.com,
aparajitab535@gmail.com,aalishan69raza@gmail.com,suchetabhowmick99@gmail.com,
sakshibhagat9873@gmail.com,sudipta_basu68@yahoo.com

Abstract- Natural language processing strives to build machines that understand and respond to text or voice data—and respond with text or speech of their own—in much the same way humans do. NLP models language computationally and deals with linguistic features of computation. Once a computer learns to do mathematical calculations it can perform many complex and big calculations much faster than humans. Similarly once computer starts to understand the human languages it can process all aspects of that language much faster than humans also opening a large number of possibility. So it cuts down on employment as one computer is capable of giving an output 10 times faster than a human can. So it benefits the employer not only financially but also by giving extremely accurate and faster outcome. Here we lay out an overall architecture to explain the overall processing. So now we take a look at the two general classes of systems they are special-purpose system and general-purpose system, explaining how they differ and their relative advantages and disadvantages. After that we point at the few remaining problems that require additional research. Finally, we conclude by discussing when natural language processing technology can be practically used at various levels .We also discuss about when it will become commercially practical, and what will be the cost to practically use this technology.

The techniques specifically developed for analysing and understanding the inner-workings and representations acquired by neural models of language is EMNLP 2018 BlackboxNLP. The approach includes: investigating the impact on the performance of neural network on systematic manipulation of input and also testing whether the interpretable knowledge can be decoded from intermediate representations to propose modifications to make the knowledge state or generated output more and also to examine the performance of networks on simplified or formal languages.

In the following report we aim to convert a program of a given language to an equivalent program of another language. For that we have taken help of NLP that is Natural Language Processing. By using Natural Language Tool Kit, we have successfully identified the variables, datatypes, operators, keywords, indentations. We have also discussed various aspects and domains of NLP and some real-world applications of it.

Keywords:Language Conversion, NLP, Keyword detection, valid Identifier checking, operators' identification, checking constants,Datatypes identification, NLTK, tokens

I. INTRODUCTION

NLP orNatural Language Processingis a branch of AI that gives the machine the ability to read, understand and derive meaning from human languages. Every day we exchange data via social media or other devices. These data is extremely useful and data experts implement this data to machine so that they can mimic human linguistic behaviour and it saves so much time and effort. It involves programming techniques to create a model that can understand the language just like normal human beings. It can even classify the contents and even generate and create new composition in human based language.

we don't even realize the wide use of NLP. we basically use it daily, like while using autocorrect method in mobile phones or checking if any document is going to be plagiarised or not. The Prolong language11 was originally invented for NLP applications. Its syntax is especially suited for writing grammars, although, in the easiest implementation mode rules must be phrased differently from those intended for a *yacc*-style parser. Top-down parsers are easier to implement than bottom-up parsers but are much slower.

A. Application of NLP

Talking about the use of Natural Language Processing involves how it has evolved in the era of technology. The basic aim of NLP is not only to understand the single word to word but also to have the capability to understand the context based on syntax, grammar, etc of those words, sentences, and paragraphs to give the desired result out of it.

In short, NLP gives the machines the ability to read, understand, drive meaning from the text and often generate the text from various languages.

NLP-enabled software assists us in our daily lives in various ways, for example:

- **Personal Assistants** (Speech recognition)**:** Siri, Cortana, and Google Assistant
- **Machine Translation**: Google Translator
- **Grammar check:** Grammarly app
- **Autosuggestion** (Sentiment analysis)**:** In search engines, Gmail, Developer's IDE
- Making **Chat bots**

B. Working Process of NLP

The working process of NLP not as easy as it may seem. Basically, it works in 3 steps,

Speech recognition → NLU → NLG

```
                     |      |

                           |  (Natural Language Generation)

(Natural Language Understanding)
```

Computers do not directly understand the words and sentences which belong to human languages. The computer only understands binary numbers as 0s and 1s. So, we had to initially develop a way for computers to understand the words. Word representation is a widely used implementation for this problem. Word representation is a technique to represent a word with a vector and each word has its unique vector representation. Text and word representation are essential for making computers understand words and thus we need to encode words into a format understandable by the computers. One-hot encoding is one such technique used to convert categorical data into numerical data. The numerical data is then used by the algorithms to learn and predict.

In short, we encode the input into numerical form and train our neural network model with that. And the we decode the output of the NN to get our desired result.

C. Encoding language into numbers

We can encode the language into numbers in many ways. The most common is to encode by letters. we generally use ASCII or Unicode value for that. But due to the presence of antigram the same number represents two words in a different order, which might make building a model to understand the text a little difficult. A better alternative might be to use numbers to encode entire word instead of the letters within them.

II. TASKS AND TECNIQUES OF NLP

There are different NLP techniques that helps us to convert the human language into machine understandable language such as

- Stemming
- Lemmatization
- Tokenization
- Stop words removal.
- Word Sense Disambiguation
- Part of Speech Tagging

A. Stemming

Stemming is a process of reducing similar words to their stem word.

E.g.: - History, Historical – Histori

B. Lemmatization

Lemmatization is the process of mapping words into their meaningful base structure.

E.g.: - Reach, Reaching, Reached, Reaches: - Reach

C. Tokenization

Tokenization breaks a sentence into words and turn them into tokens.

D. Stop words Removal.

Removal of words from sentences which do not contain any valuable information is known as Stop words removal. E.g.: - a, the, is, are etc.

Word Sense Disambiguation: It is used to determine if the Same words can have different meanings in different sentences.

Part of Speech Tagging: Part of Speech (POS) tagging is well-known in NLP, which is used to label each word in a sentence or it can be a paragraph with its appropriate part of speech. Part of speech includes verbs, adverbs, adjectives, pronouns, etc.

III. PROPOSED IDEA

The main goal of this project is *Conversion of Different programming languages*.

In recent days, learning multiple programming languages is really time consuming. If there were a system that could easily convert a general programming language into another one, then working in different fields would’ve been much easier. Suppose a person only knows python and he need to work in Java for a certain project. Now this kind of compiler would make it easier. The person will enter his code in python and the compiler will turn it into Java.

Now this concept was proposed earlier but making it work is not that easy. As we know that different programming languages have their own syntax and executing process, so converting them is not only time consuming but also it needs a lot of skills.

IV. APPLICATION OF NLP IN PROGRAMMING LANGUAGE CONVERSION of NLP

Here we will try to change a block of code from one language to another by the use of NLP. For simplicity we are choosing two Object Oriented Programming languages, otherwise it would be difficult to change a Object Oriented Programming language to a procedural language and vice versa.so for this report we are choosing python and Kotlin. As of now we are approaching the problem as mentioned below:

Steps to be followed: -

1. Just like part of speech tagging we must come up with a method that can tag key words, variables, constants, different types of operators (conditional, logical, mathematical).

2. If we can tag and tokenize the block of code like this it would be a lot easier for the computer to understand.

3. Then we will feed this tokenized encoded input to our trained neural network model.
4. After getting the output from the NN we have to decode it to get our desired result.

5. The NN will only change the words that has been tagged as keywords or operators.

6. The variable and constants will remain as it is. while the key words, operators and syntax will be changed as necessary.

7. We have to follow sequence to sequence conversion so that the main structure of the code does not get changed.

A. Experimental Setup and Result Analysis:

Theoretical approach for solving this kind of problem: -

Let us consider one simple python code of adding two integers-

Input program file:-

```
a=10

b=2

x=a+b

if(x>=25):

        print(x)

 else:

       print(x+10)
```

First, we will tag the variables and constants that is a, b, x ,10,2,25. these variables and constants value won’t change after conversion.

Then we will tag the key words i.e. print, if, else. we will tokenize it and feed it to the NN which will give us the output as the version of this keywords in the desired language.

After this by using neural machine translation we have to correctly change the syntax of the code if necessary.

And then after decoding we will get our desired output.

Expected Kotlin code: -

B. Keyword detection

Output:-

```
a is an identifier

= is an operator

10 is a constant

b is an identifier

2 is a constant

x is an identifier

+ is an operator

if is a keyword

bracket

>=  is an operator

25 is a constant

bracket

: is used for indentation

print is a keyword

else is a keyword
```

C. Challenges:

Implementing this Program is not that easy as we don't know if it'll actually work practically or not. The challenges that we have faced while implementing the process are quite a few:

1. Removing ambiguity- Ambiguity is an intrinsic characteristic of human conversations.Ambiguity is one of the biggest challenges in NLP. When trying to understand the meaning of a word we consider several different aspects, such as the context in which it is used.
2. Improvement of the performance of individual analysers, specially at the semantic/pragmatic level.
3. Definition of new tasks, such as the detection of exclusivity, parallelism/concurrency, decision points, or iteration of tasks.
4. Detecting comment lines and text lines separately.

V. PROPOSED SOLUTION

Solving these problems at once is not a piece of cake. Therefore, our goal is to find different solutions for the problems that are rising throughout the implementation process.

1. segmenting text into meaningful groups.
2. identifying individual tokens within a sentence.
3. Word/phrase order variation.
4. The identification of problem-specific information and its transformation into structured form.
5. Determining relationships between entities or events.

VI. Conclusion and Future scope

Natural Language processing is one of the fastest growing technologies in todays world. It is like a blessing to the mankind as it makes every single task which required human involvement solvable in a very minimum time. But again every good thing has a dark side too so this natural language processing has also cut down on employment as it benefits the employer by doing the task in an accurate and faster manner. As its processing speed is ten times faster than human so an employer always prefer this technology. Innumerous number of research and development in this technology makes it a tremendously strong upcoming field which has a dire need of skilled professionals. With the exponential growth of multi-channel data like social or mobile data, businesses need solid technologies to assess and evaluate customer sentiments. So far, businesses have been happy analyzing customer actions, but in the current competitive climate, that type of customer analytics is outdated.

Natural language processing is an AI-complete problem. It is same as solving central artificial intelligence problem whose main aim is to make computers as intelligent as people so that they can think and solve problems like humans .They can perform all the activities that humans can perform at a much accurate and faster level and makes it more efficient than humans. Also it can perform tasks that humans cannot. Growth of Artificial intelligence basically predicts the growth of natural language processing in future. Through natural language computers or machines or devices understanding of human language will increase and they will be able to collect understand and the information online and apply what they learned in the real world. Combined with natural language generation, computers will become more and more capable of receiving and giving useful and resourceful information or data.

NLP can be used in areas where technology is not customer or human-facing like voice assistants and chatbots. NLP is one of the largest growing technologies in the field of data science. It can be also used to decipher meaning from unstructured data. NLP will find more and more applications in everyday technology as and when Ai and need for applications and humans to inter-communicate increases .

Till now we only have been able to detect the keywords, identifiers, constants, datatypes, indentation, operators etc. In future we aim to convert these tokenized strings into its respective code form to get the equivalent code in target programming language.

REFERENCES

[1] AI and Machine Learning for coders By Laurence Moroney

[2] https://www.qblocks.cloud/blog/natural-language-processing-machine-translation#:~:text=Natural%20Language%20Processing%20(NLP)%20is,the%20same%20way%20humans%20do.

[3] https://academic.oup.com/jamia/search-results?cqb=[{%22terms%22:[{%22filter%22:%22Keywords%22,%22input%22:%22Natural%20language%20processing%22}]}]&qb={%22Keywords1%22:%22Natural%20language%20processing%22}&page=1&searchType=advanced&adv=true

[4] https://dl.acm.org/doi/abs/10.5555/3340

[5] https://apps.dtic.mil/sti/pdfs/ADA567972.pdf

[6] https://arxiv.org/pdf/1904.04063

[7]https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=eeace1d14e266a5cd44fe781a874c662928602fd

[8] https://www.ibm.com/in-en/topics/natural-language-processing#:~:text=IBM%20Watson%20Discovery-,What%20is%20natural%20language%20processing%3F,same%20way%20human%20beings%20can.

[9]https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1

[10]https://realpython.com/python-keywords/

[11]https://www.digitalocean.com/community/tutorials/python-keywords-identifiers

[12]https://www.w3schools.com/python/python_datatypes.asp

[13]https://www.geeksforgeeks.org/python-operators/

[14]https://www.nltk.org/