International Journal of scientific research and management
(IJSRM) ||Volume||4||Issue||11||Pages||4800-4802 ||2016||
Website: www.ijsrm.in ISSN (e): 2321-3418

Role of Java in Natural Language Processing for Indian Regional Languages

Harjit Singh
Punjabi University Neighbourhood Campus, Dehla Seehan (Sangrur)

ABSTRACT

Java is an object oriented paradigm based programming language designed for originality for embedded software but later extended for Internet application implementation. Java supports a powerful and robust feature set such as hardware/software neutral, distributed and multithreaded programming which makes this language a relevant tool for modern software development. To deal with diversity of natural languages, it supports the Unicode characters sets. It has finely developed library that provides text processing packages for natural language processing research. We may find numerous toolkits for English text processing but may fail to find a suitable Indian languages text processing toolkit. In the present situation, the efforts to process these languages are being done from scratch by the natural language processing researchers. In this paper, Java's appropriateness for processing Indian languages is discussed in detail from various aspects.

Keywords: Natural Language Processing, NLP for Indian Regional Languages. Java for Natural Language Processing.

I. INTRODUCTION

The emergence of new technology offers more opportunities to researchers to do something new in their area of research and development. In the beginning state, web was just a collection of static pages of information i.e. Hyper Text Markup Language (HTML). In early days of World Wide Web, some dynamic behavior is added to web pages by Richard Denny with the invention of CGI (Common Gateway Interface). Now the web pages were able to call programs residing on server. These programs were able to access data stored in databases from the server and the result was displayed on client computers. The technology was adopted by computational linguistics and used to provide online linguistic tools including mono and bilingual dictionaries. The purpose of this paper is not to promote java or compare it with other programming languages being used for Natural Language Processing. It is just an analysis of java in view of Natural Language Processing.

NLP REQUIREMENTS AND JAVA

Some requirements that need to be fulfilled and are expected by Natural Language Processing researchers are given below. Java is analyzed to determine how much it fulfills these requirements.

A. Platform Independence

Natural languages should be processed with keeping in mind the hardware and software neutrality. It will allow the code to perform its functions on any choice of operating system or chipset. Java's robust invention named Java Virtual Machine (JVM) perfectly suits this requirement (Figure 1). JVM is sometimes called as Java Runtime Environment (JRE). With JVM functionality, two phase compile procedure is followed by java to obtain native machine code. The compiler of java produce a middle level code called byte code which is the output code of compiler but not a complete machine code. The remaining task to complete it is done by java interpreter. It can be done by running the interpreter or by the JVM. Therefore program compilation and execution requires JRE installation which provides the Suitable Environment.

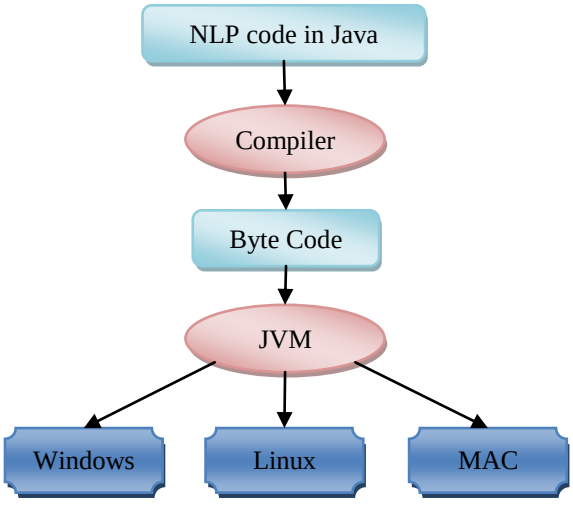

Figure 1

A new feature is Just-In-Time (JIT) compilation. This feature compiles the code when it is needed and saves it for reuse if need again, it overcomes the requirement to be compiled again.

B. Fast-Track Processing

Natural languages need to be processed by making use of large dictionaries and other datasets. These large datasets affect the speed of processing. So the processing model should be fast enough to quickly do its job. The facility of JIT compilation fulfills this requirement. The code once compiled need not be recompiled. Another feature is multithreading providing a way for simultaneous execution.

An application can be modeled for multithreaded execution. The total number of tasks can be divided and submitted to separate the threads for execution to cut off the total time of processing. For example, one thread can

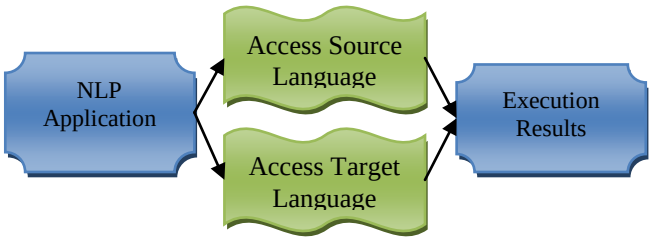

Figure 2

be used to access source language and another thread can be used to access target language in case of translation and transliteration (Figure 2). One thread can be given the task of accessing dictionary and another thread can be left free to interact with user responses.

C. Up to Date Linguistic Knowledge

Natural languages' linguistic knowledge is collected to store in large data files. Changing a database e.g. SQL Server to Oracle does not need any modifications in the Java's connection coding. The feature of establishing a bridged connection works at the top of API related to Java Database Connectivity (JDBC). The java program is totally detached from the database containing the linguistic knowledge (Figure 3). In this sense, we can make the knowledge set up to date any times without affecting the java applications using it. Therefore any number of applications accessing the shared database of knowledge can continue to work in parallel to the updating underlying database. One application can update the knowledge base while others are consuming it to perform NLP tasks. In real time those applications are accessing the data but it is abstracted to each of them.

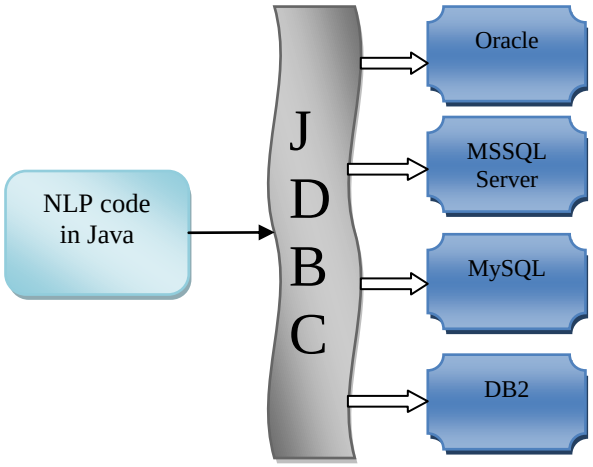

Figure 3

The users can change the connected database that may be of same format such as both databases belong to Oracle or it can be of different such as the new belongs to SQL Server. Java can handle each of them.

D. Knowledge Reusability

Linguistic knowledge set is a precious knowledge base that is very important resource for NLP Research. This type of dataset cannot be reserved for only one application. Although it is collected to be used for some specific application but it should not be bound to that one. It should be a separate entity of information shareable to any number of applications that need it.

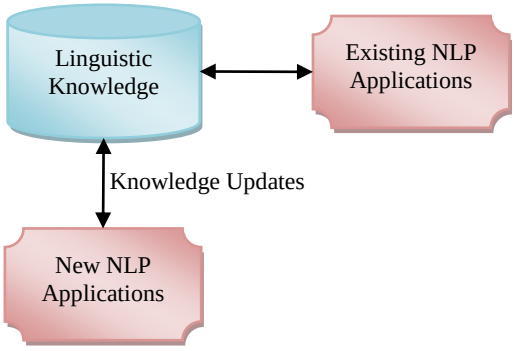

Figure 4

Since we already discussed previously that the linguistic knowledge is a sharable resource and is not bound to the java application, it can be used multiple times. Any freshly developed software can consume the available knowledge without the need to collect it from beginning stage. Practically this knowledge is preserved in tables of database. As per the requirement of freshly developed software, we can add more tables to the database without

affecting the performance of old applications. The functionality is not disturbed (Figure 4).

E. Provision of Online Availability

The solutions developed for NLP functionality can be more use worthy if these can be accessed over the Internet without installing offline. Actually due to the use of huge datasets, the NLP software will be so bulky that every machine may not incorporate it due to the limited configuration of resources. In the modern scenario, the cloud computing can provide such a robust platform to these bulky applications that these can be used on the go on any machine.

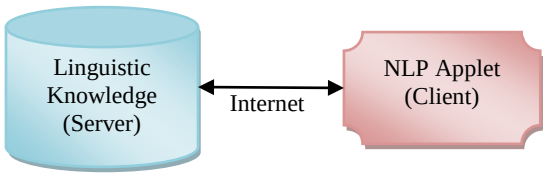

Figure 5

Java supports distributed programming and fits well for this situation. Java distributed components include a unique function component called applet. These are designed to be a part of web page and executed by java supported web browsers. These applets can provide the full functionality of an installed application by communicating with the server. The user uses the interface of the applet to communicate with server and server performs the heavy load processing as per the demand (Figure 5). The applets provide secure communication over the public network and are perfectly used in banking section. Before the applet invention, there were other techniques used in web pages i.e. CGI or JavaScript, but were not capable as applets.

II. CONCLUSION

In this paper the basic requirements of Natural Language Processing in a programming language have been discussed and it is analyzed that how java may be useful to fulfill those requirements. Although many tools are available to work with English text, but no such tool is available to work with Indian regional language processing. Java may provide platform independence to these applications with speedup processing capability. It may provide an architecture that allows easy updation of linguistic knowledge and that knowledge is reusable in further development of new Natural Language Processing applications. Java allows the availability of research solutions provided online without the need to install applications on individual computer systems.

REFERENCES

[1] Abhimanyu Chopra, Abhinav Prashar, Chandresh Sain, Natural Language Processing, International Journal Of Technology Enhancements And Emerging Engineering Research, Vol 1, Issue 4

[2] Prof. Langote Manojkumar S, Miss Kulkarni Sweta, Miss Mansuri Shabnam, Miss Pawar Ankita and Miss Bhoknal Kishor, Role of NLP in Indian Regional Languages, IBMRD's Journal of Management and Research Volume-3, Issue-2, September 2014

[3] Bharati,Akshar,Chaityanya Vineet and Sangal Rajeev,(1995),Natural Language Processing: A Paninian Perspective,Prentice-Hall of India.

[4] Gore Lata and Patil Nishigandha, English to Hindi-Translation System,Proceedings of Symposium on translation systems strans (2002).

[5] Cini Kurian, A Review of the Progress of Natural Language Processing in India, International Journal of Advances in Engineering & Technology, Volume 7, Issue 5 (Nov. 2014).

[6] Padariya Nilesh,Chinnakotla Manoj,Nagesh Ajay and Dawant Om P.,(2008),Evaluation of Hindi to English, Marathi to English and English to Hindi.

[7] http://www.slideshare.net/jhonrehmat/natural language processing.

[8] Natural Language Processing,www.myreaders.info /html/artificial_intelligence.html.

[9] Natural Language Processing-Computer science and engineering, www.cse.unt.edu/~rada/CSCE5290/Lectures/Intro.ppt

[10] NLP, https://www.coursera.org/course/nlp

[11] NLP, research.microsoft.com/en-us/groups/nlp/

[12] Dash, N S and B B Chaudhuri. "Why do we need to develop corpora in Indian languages", International Conference on SCALLA, Banglore, 2001

[13] Murthy, B K and W R. Despande. Language technology in India: past, present, and the future. In the Proceedings of the SAARC Conference on extending the use of Multilingual and Multimedia Information Technology (EMMIT'98). Pune, India

[14] Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah and Shata-Anuvadak: Tackling Multiway Translation of Indian Languages, LREC 2014, Rekjyavik, Iceland, 26-31 May, 2014

[15] R M K Sinha. " Machine Translation : An Indian Perspective " , Proceedings of the Language Engineering Conference (LEC'02)

[16] Vishal Goyal and Gurpreet Singh Lehal. "Web Based Hindi to Punjabi Machine Translation System", Journal of Emerging Technologies in Web Intelligence, Vol. 2, No. 2, May 2010, pg(s):148-151.

[17] Pushpak Bhattacharyya, Natural Language Processing: A Perspective from Computation in Presence of Ambiguity, Resource Constraint and Multilinguality, CSI Journal of Computing, Vol. 1, No. 2, 2012

[18] https://en.wikipedia.org/wiki/History_of_natural_language_processing