CHAPTER – 3

Study of related Tools and Technology

3.1 Parser

3.2 WordNet

3.3 Java

3.4 Database

3.5 Summary
3.1 PARSER

3.1.1 Introduction to Parser

The Syntactic level of Natural Language Processing has two purposes: (a) to check that a string of words (a sentence) is well formed and (b) to break a sentence into a structure that shows syntactic relationships between words [60]. A syntactic analyzer (known as a parser) does this using a lexicon processing and a set of syntax rules (the grammar). There are many publicly available Parsers such as Berkeley Parser, CMU Link Parser, Malt Parser, MST Parser, RASP Parser, and Stanford Parser [27] which can be used to parse the sentence.

We have used Stanford Parser since it contains a Java implementation of probabilistic Natural Language (English) Parsers and gives visualization of the parse tree using its GUI.

3.1.2 Stanford Parser

It is a statistical natural language parser developed by Stanford Natural Language Processing Group. It is used to parse input data written in either of languages such as English, German and Chinese. It is licensed under GNU GPL [47]. The parser runs under Windows/Linux/Unix/MacOS and requires Java Runtime Environment (JRE).

The Penn Treebank is widely used linguistic resource in Stanford Parser. It includes collection of text in four different formats; (a) Raw text (original) (b) Tagged with Part of Speech (POS) tag sets and its example is shown in figure 3.1 [91] (c) ‘Parsed’- that is marked up with constituent structure and (d) Combined – including both POS tags and constituent structure.
3.1.2.1 The Stanford Typed Dependency

The Stanford typed dependencies representation was designed to provide a simple description of the grammatical relationships in a sentence that can be easily understood and effectively used by people without linguistic expertise who wants to extract textual relations [60]. The dependency obtained from Stanford Parser can be mapped directly to graphical representation in which words in a sentence are nodes in a graph and grammatical relationships are labeled edges. This has been used to extract the relationship between subjects and objects when sentence is parsed. Stanford Dependency (SD) is triplets of: (a) name of the relation (b) governor and (c) dependent. The dependencies are binary relations. A grammatical relation holds

![Figure 3.1: Example of POS tag sets](image)

<table>
<thead>
<tr>
<th>Tag</th>
<th>Meaning</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJ</td>
<td>adjective</td>
<td>new, good, high, special, big, local</td>
</tr>
<tr>
<td>ADV</td>
<td>adverb</td>
<td>really, already, still, early, now</td>
</tr>
<tr>
<td>CNJ</td>
<td>conjunction</td>
<td>and, or, but, if, while, although</td>
</tr>
<tr>
<td>DET</td>
<td>determiner</td>
<td>the, a, some, most, every, no</td>
</tr>
<tr>
<td>EX</td>
<td>existential</td>
<td>there, there's</td>
</tr>
<tr>
<td>FW</td>
<td>foreign word</td>
<td>dolce, ersatz, esprit, quo, maître</td>
</tr>
<tr>
<td>MOD</td>
<td>modal verb</td>
<td>will, can, would, may, must, should</td>
</tr>
<tr>
<td>N</td>
<td>noun</td>
<td>year, home, costs, time, education</td>
</tr>
<tr>
<td>NP</td>
<td>proper noun</td>
<td>Alison, Africa, April, Washington</td>
</tr>
<tr>
<td>NUM</td>
<td>number</td>
<td>twenty-four, fourth, 1991, 14:24</td>
</tr>
<tr>
<td>PRO</td>
<td>pronoun</td>
<td>he, their, her, its, my, I, us</td>
</tr>
<tr>
<td>P</td>
<td>preposition</td>
<td>on, of, at, with, by, into, under</td>
</tr>
<tr>
<td>TO</td>
<td>the word to</td>
<td>to</td>
</tr>
<tr>
<td>UH</td>
<td>interjection</td>
<td>ah, bang, ha, whee, hmpf, oops</td>
</tr>
<tr>
<td>V</td>
<td>verb</td>
<td>is, has, get, do, make, see, run</td>
</tr>
<tr>
<td>VD</td>
<td>past tense</td>
<td>said, took, told, made, asked</td>
</tr>
<tr>
<td>VG</td>
<td>present participle</td>
<td>making, going, playing, working</td>
</tr>
<tr>
<td>VN</td>
<td>past participle</td>
<td>given, taken, begun, sung</td>
</tr>
<tr>
<td>WH</td>
<td>wh determiner</td>
<td>who, which, when, what, where, how</td>
</tr>
</tbody>
</table>
between a governor and a dependent. For example, the query inputted by the user can be “List of students who study in semester 3”. The POS and Stanford Dependencies representation will be in the form:

**POS Tagging:** list/VB stud/NN who/WP study/VBP in/IN semester/NN 03/CD

**Stanford Dependencies (SD):**
dobj(list, stud)
nsubj (study, stud)
rcmod (stud, study)
prep_in (study, semester)
um(semester, 03)

This SD can be mapped into a directed graph representation in which words of the sentence are nodes in the graph and grammatical relations are edge-labels. Figure 3.2 gives a graphical representation for the sentence “list of students who study in semester 3” [55].

*Figure 3.2: Graphical representation of ParseTree*
Chapter 3: Study of related Tools and Technology

There are different types of SD such as basic dependency, collapsed dependencies, collapsed dependencies with the propagation of conjunct dependencies and Collapsed dependencies preserving a tree structure. The figure 3.3 shows hierarchy of basic dependency [60].

We have utilized the Stanford Parser in two ways: (a) for developing a Multi-Liaison algorithm to find multiple subjects and multiple relationships from a sentence and (b) for developing a Natural – English Language Interface to Database (N-ELIDB) tool. These are discussed in chapter 4 and 5.

```
root - root
derp - dependent
aux - auxiliary
  auxpass - passive auxiliary
cop - copula
arg - argument
  agent - agent
comp - complement
  acomp - adjectival complement
  ccomp - clausal complement with internal subject
  xcomp - clausal complement with external subject
obj - object
  dobj - direct object
  iobj - indirect object
  pobj - object of preposition
subj - subject
  nsubj - nominal subject
    nsubjpass - passive nominal subject
  csubj - clausal subject
cc - coordination
conj - conjunct
expl - expletive
mod - modifier
  amod - adjectival modifier
  appos - appositional modifier
  advcl - adverbial clause modifier
  det - determiner
  predet - predeterminer
  preconj - preconjunct
  vmod - reduced, non-finite, verbal modifier
  mwc - multi-word expression modifier
```

Figure 3.3: Hierarchy of Typed Dependency
3.2 WordNet

WordNet is an on-line lexical database of English, developed by the Cognitive Science Laboratory at Princeton University under the guidance of George Miller. It consists of lexicographer files comprising nouns, verbs, adjectives and adverbs that are grouped into a set of synonyms (called synsets), each expressing distinct concepts [31]. Synsets are interrelated by means of conceptual (semantic) and lexical relations. Three kinds of relation are possible in WordNet: (a) noun relationship (b) verb relationship and (c) adjective and adverb relationships as shown in table 3.1, 3.2 and table 3.3 [21] respectively. WordNet is a useful tool for Computational Linguistic and NLP. Figure 3.4 represents an interface of WordNet. The N-ELIDB uses WordNet to match words semantically as discussed in chapter 4.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypernymy</td>
<td>From concepts to superordinates</td>
<td>Breakfast to meal</td>
</tr>
<tr>
<td>Hyponym</td>
<td>From concepts to subtypes</td>
<td>Meal to lunch</td>
</tr>
<tr>
<td>Has-Member</td>
<td>From groups to their Members</td>
<td>Faculty to professor</td>
</tr>
<tr>
<td>Member-Of</td>
<td>From members to their Groups</td>
<td>Copilot to crew</td>
</tr>
<tr>
<td>Has-Part</td>
<td>From wholes to part</td>
<td>Table to leg</td>
</tr>
<tr>
<td>Part-of</td>
<td>From parts to whole</td>
<td>Course to meal</td>
</tr>
<tr>
<td>Antonym</td>
<td>Opposites</td>
<td>Happy to sad</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relation</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypernym</td>
<td>From event to superordinate events</td>
<td>Fly to travel</td>
</tr>
<tr>
<td>Troponym</td>
<td>From events to their Subtypes</td>
<td>Walk to stroll</td>
</tr>
</tbody>
</table>
Table 3.3: Adjective and Adverb relations in WordNet

<table>
<thead>
<tr>
<th>Relation</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antonym</td>
<td>Opposite</td>
<td>Heavy to light</td>
</tr>
<tr>
<td>Adverb</td>
<td>Opposite</td>
<td>Quickly to slowly</td>
</tr>
</tbody>
</table>

Terminology used by WordNet [31]:
- Synsets: It is a set of synonym. It is a symmetric relationship between word forms.
- Antonymy (opposite form): It is also a symmetric semantic relationship between word forms.
- Hyponymy (sub-name) and its inverse Hypernymy (super-name): They are transitive relation between synsets.
- Meronymy (part-name) and its inverse Holonymy (whole-name): They are transitive semantic relations.
- Collation: It is a sequence of two or more words connected by spaces or hyphens, that go together to form a specific meaning such as “car-pool”, “blue-collar-job”.
- Baseform: It is formed by adding inflections.
- Lemma: It is a base form of a word or collocation. It is lowercase ASCII text of the word.

Usage of WordNet:
- To perform the tasks such as Word Sense Disambiguation, Machine Translation, Information Retrieval, Named Entity Recognition, etc.
- To derive the relationships between two terms t1 and t2, by exploring semantic information.
- To find the synonyms for a given term.
- To measure the distance between synsets in hypernymy structure.
• It is used in ontology mapping to calculate semantic similarity between two concepts-labels unlike Edit Distance that computes morphological similarity.

![WordNet 2.1 Browser](image)

*Figure 3.4: Screen shot of WordNet*

### 3.3 JAVA

#### 3.3.1 Java Platform

Java was developed originally by James Gosling at Sun Microsystems and released in 1995 as a core component of Sun Microsystems Java Platform [46]. Java as of 2014 is one of the most popular programming languages in use. Java is a set of several computer software products and specification from Sun Microsystems that together provides a system for developing application software and deploying it in a cross-platform computing environment. [46]. Java is used in a wide variety of computing platforms from embedded devices including mobile phones to enterprise servers and supercomputers on the high end. Java Applets are also used for improved and secured environment while browsing on the World Wide Web using desktop computers. The Java platform is not specific to any one operating
system but has an execution engine (called virtual machine) and a compiler with a set of libraries, which can be executed in various hardware and software systems. The library includes java card, java macro edition, java standard edition and java enterprise edition [46].

- **Java Card**: A technology that allows small Java-based applications (applets) to run securely on smart cards and similar small-memory devices.
- **Java Micro Edition (JME)**: It includes several library files which can be used for mobile devices, TV set-top boxes, printers, etc.
- **Java Standard Edition (JSE)**: It is used for general-purpose applications such as desktop PCs, servers and similar devices.
- **Java Enterprise Edition (JEE)**: It is used for multi-tier client-server enterprise applications. Figure 3.5 shows various components of editions of Java 2 platform [44].

![Figure 3.5: JSE platform (Source: www.netbeans.org/kb/docs)](image)

The Java platform consists of several programs, each of which provides a portion of its overall capabilities. For example, a Java compiler converts Java source code into Java bytecode (intermediate language) for Java Virtual Machine (JVM). The Java Runtime Environment (JRE), complementing the JVM, with Just-In-Time (JIT)
compiler converts intermediate bytecode into native code. The working strategy of JAVA is shown in the figure 3.6[44]. An executable set of libraries are also part of the Java Platform. The essential components in this platform are Java language compiler, the libraries and the runtime environment in which Java intermediate bytecode executes according to the rules laid out in the virtual machine specification [46].

3.3.2 Java Development Kit (JDK)

The JDK has its primary components as collection of programming tools which includes [46]:

- Appletviewer: This tool can be used to run and debug Java applets without a web browser.
- Apt: The annotation processing tool.
- Extcheck: A utility which can detect JAR- file conflicts.
- Idlj: The IDL to Java compiler.
- Jabswitch: The Java access bridge.
- Java: The loader for Java applications. This tool is an interpreter and can interpret the class files generated by the Java compiler.
• Javac: The Java compiler which converts source code into Java bytecode.
• Javadoc: The documentation generator which automatically generates documentation from source code comments.
• Jar: The achiever, which packages related class libraries into a single JAR file.
• JarSigner: The Jar file signing and verification tool.
• Jdb: The debugger.
• Jinfo: This utility gets configuration information from running java process or crash dump.
• JConsole: Java monitoring and management console, etc.

### 3.3.3 Java APIs

Java APIs is a set of classes included with Java Development Environment (JDE). These classes are written using Java language and run on Java Virtual Machine (JVM) [39]. For example, if we include the statement in the Java program “import javax.swing.*”, then the swing API provide access to graphical component used to build Graphical User Interface (GUI). The table 3.4 gives a partial list of Java APIs used in our present work.

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Package</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Java.awt</td>
<td>Contains all the classes for creating user interfaces and for painting graphics and images.</td>
</tr>
<tr>
<td>2</td>
<td>Java.beans</td>
<td>Contains classes related to developing beans</td>
</tr>
<tr>
<td>3</td>
<td>Java.io</td>
<td>Provides for system input and output through data stream, serialization and file systems</td>
</tr>
<tr>
<td>4</td>
<td>Java.sql</td>
<td>Provides the API for accessing and processing data stored in a data source (usually a relational database) using Java programming language.</td>
</tr>
<tr>
<td>5</td>
<td>Java.util</td>
<td>Contains the collection framework, legacy collection classes, event model, date and time facilities, internationalization and miscellaneous utility classes.</td>
</tr>
</tbody>
</table>
### 3.3.4 Java JAR Files

JAVA JAR (Java Archives) is a package format typically used to aggregate many Java class files and associated metadata and resources such as text, images, etc. JAR files are archive files built in ZIP file format and have jar file extension. A JAR file allows the Java runtime to efficiently deploy a set of classes and their associated resources. The elements in a JAR file can be compressed together with the ability to download an entire application. The package java.util.zip contains classes that read and write JAR files [46].

The present work uses following JAR files in the prototype implementation.

- Mysql-connector-java-5.1.18-bin.jar
- Stanford-parser-2010-02-26.jar
- Junit-4.1.jar
- Jwnl.jar
- Commons-logging.jar
- JDK1.6 (Default)
3.3.5 NetBeans

NetBeans Integrated Development Environment (IDE) is a free, open source, cross-platform IDE with built-in support for Java Programming Language and also for other languages like PHP, C++, HTML, etc. Figure 3.7 represents an interface of NetBeans IDE.

The NetBeans IDE is written in Java and can be run on windows, Linux, Solaris and other platform supporting compatible JVM [49]. NetBeans IDE 6.5 which was released in Nov 2008 and with extended Java features that supports Java Enterprise Edition, SOA, XML schema tools, etc. The Salient features of the NetBeans IDE platform are as follows:

- User Interface Management (e.g. menus and toolbars)
- User Setting Management
- Storage Management (saving and loading any kind of data)
- Windows Management
- Wizard Framework (supports step-by-step dialogues)
- NetBeans Visual Library
- Integrated development tools, etc.

![Figure 3.7: Screen shot of NetBeans IDE](image-url)
3.4 DATABASE

3.4.1 Database Management System (DBMS)

A database is a computer based record keeping system used to store, maintain and retrieve data. It is an organised collection of interrelated data. DBMS is a collection of interrelated files and set of programs which allows user to access and modify files. It provides convenient and efficient way to store and modify the information. DBMS deals with various types of data storage. It can be in the form of hierarchical model, network model, relational model or object-relational model.

Advantages of DBMS:

- It facilitates Data Management
- It reduces and controls Data Redundancy
- It avoids Data Isolation
- It provides Transaction support
- It provides Concurrency control
- It supports Backup and Recovery management
- It provides Security and Integrity
- It provides Performance Analysis
- It support data sharing across users and applications
- It supports Control of Physical storage, Audit etc.

3.4.2 Relational Database Management System (RDBMS)

Relational Database Management System (RDBMS) is a database management system that is based on relational model introduced by E.F.Codd. Relational database present data to the user as relations (collection of tables with each table consisting of a set of rows and columns) and also provide the relational operators to manipulate the data in tabular form. Two important characteristics of RDBMS are Entity Integrity and Referential Integrity.
3.4.3 MYSQL

MySQL is the world’s second most widely used open-source relational database management system (as of March 2014) [46]. MySQL supports SQL specification as shown in figure 3.8 [48]. The MySQL development project has made its source code available under the terms of the GNU- General Public License as well as under a variety of agreements.

MySQL is a popular choice for Linux, Apache, MySQL and Perl/PHP/Phython (LAMP). The official MySQL workbench which is a free, integrated environment developed by MySQL AB and it enables the user to graphically administer the MySQL database and visualize design database structures.

MySQL Workbench is available in two editions, the regular free and open source community edition, which can be downloaded from MySQL website and the proprietary Standard Edition, which extends and improves the feature set of Community Edition [48]. We have used SQL-YOG as GUI which is free and open source community edition.

![MySQL](image)

Figure 3.8: MySQL specification
Some of the features of MySQL Database are as follows [69]:

- It is written in C and C++.
- It works on different Platforms
- APIs are available for C, C++, Java, PHP, Perl, Phyton, etc.
- It is fully multi-threaded using kernel threads
- It has a fast thread based memory allocation systems
- It has highly optimized class library
- It supports fixed and variable length record
- It supports security mechanism based on privileges
- It handles large database
- The MySQL server supports connection based on TCP/IP sockets using named pipes, etc.
3.4.4 SQL-YOG

SQL-YOG is a Graphical User Interface (GUI) tool for MySQL relational database management system. It was developed by WebYog. It is available both as a freely available version and also as paid version. Figure 3.9 represents the user interface of SQL-YOG. Features of SQL-YOG are as follows [46]:

- It supports data manipulations (Insert, update and Delete) using spreadsheet like interface. It has an editor with syntax highlighting and various automatic formatting options.
- It has facilities like Visual Schema Designer, Visual Query Builder, and Query Formatter.
- It has connectivity options using MySQL API.
- It can import data from other databases.
- It supports SQL Scheduler and Reporting facility, etc.

![Figure 3.9: SQL-YOG User Interface](image)
3.5 SUMMARY

The chapter describes the concept of Parser, Stanford Dependencies and its hierarchy types. It also discusses the concept of WordNet, its terminologies and its usage in semantic matching of words. This chapter also discusses Java programming language, its features, Jar files, Java API, etc. Finally, it discusses the concept of DBMS, RDBMS, and MySQL database.