

Extracting Data Elements from Punjabi Language query

Harjit Singh^{1*}, Ashish Oberoi²

¹ APS Neighbourhood Campus, Punjabi University Patiala, Punjab, India

² School of Engineering, RIMT University, Mandi Gobindgarh, Punjab, India

**Corresponding Author: h jit@live.com, Tel.: +91-98551-79078*

Available online at: www.ijcseonline.org

Accepted: 23/Dec/2018, Published: 31/Dec/2018

Abstract— Databases act as repository of data for various applications. In today's scenario every organization has the database to store data and database management systems to access that data. SQL (Structured Query Language) is the language of database management system which is a specific language used to write statements against Relational Database Management Systems to retrieve and manipulate data. But the common user asks query in his/her natural language such as Punjabi Language. That natural language query is not understandable to the computer and hence RDBMS cannot process that query. That natural language query can be processed through Natural Language Processing to understand what kind of data the user wants to retrieve. From that natural language query, we need to retrieve data from the database. To automate the process, various data elements need to be extracted from the Punjabi Language query. These data elements include Entity, Attributes, Condition etc. This paper explains the process of extracting Data Elements from the Punjabi Language query.

Keywords— Natural Language Processing, Punjabi Language Processing, Data Element Extraction, Structured Query Language (SQL)

I. INTRODUCTION

Internet is the leading source of data and is utilized by variety of users world wide. The access to internet by common users is increasing at a fast rate and more and more data is being transferred from one device to another and from clients to servers. Now a days, almost everyone has a smart phone which provided easy access to data and information directly or indirectly to common users. This vast collection of digital data is stored and organized in structured format to make it simple to explore and maintain. An RDBMS (Relational Database Management System) is a software package that is usually used to organize the storage of vast collection of data in a structured format. All RDBMS Systems use a special query language named SQL (Structured Query Language) to access and store data to and from databases. SQL is a fixed format query language and it is not easy to learn such an unnatural language by common users. It puts the limits on direct access of data stored in databases by common users. Accessing data indirectly through an application limits the way data can be accessed. It may not fulfill all requirements of end users. The user can easily specify his/her data requirements in natural language. The Natural Language sentence needs to be processed to extract Data Elements such as Entity, Attributes, and Conditions etc. to access the required data from the database.

In this paper, Section II is about NLP introduction, Section III provides details about adopted methodology, Section IV provides implementation and testing and Section V concludes the research work. It will motivate further research in this field.

II. NATURAL LANGUAGE PROCESSING (NLP)

Those languages which are spoken by common people to communicate with each other are called Natural Languages. But computers are unable to understand those languages. The languages that computers are able to understand are called computer languages. Artificial Intelligence (AI) is a vast field of research to make machines or computers artificially intelligent as the human beings are naturally. One branch of AI is NLP (Natural Language Processing), whose aim is to make the machines understand Natural Languages so that the machines can communicate with people using peoples' natural languages [1].

Under NLP, lot of research work has been done on English language, which is obvious because it is a global language. Also the computer languages developed so far are based on English, so it is convenient to process English language and make machines understand it. But in country like India, common people do not communicate in English, their native

languages like Punjabi, Hindi, Gujarati etc. are an easy way for their communication [2].

Translation is the process of converting a source language text into a target language text. The process seems easy for a person who has the knowledge of both source and target languages. But when computers are thought about doing the same job, it is very complex task to perform perfectly [3]. It is because translating just word by word substitution will not be perfect translation. Until the computers do not understand the meaning of source language text, perfect translation in target language is not possible. A single word of source language may have different meanings according to the situation in which that word is used [4]. It means the most challenging job is the first part of translation i.e. makes the computers understand the meaning of source language text. There are two paradigms to produce a target language text having the same meaning as the source text [5]:

A. Rule-Based Approach

This approach is also known as Knowledge Based Machine Translation. Bilingual dictionaries along with grammars are used to get linguistic information for translation process. The source to target language dictionaries and grammars should be available to adopt this approach. These resources provide syntactic, semantic and morphological information to the translation system. A set of linguistic rules are applied on source language text to produce target language text [6].

Some most commonly used modules in Rule based translation include Morphological Analyzer for morphological processing and POS (Part-of-Speech) tagger for assigning grammatical tags. Rule based approach for translation can be direct translation, interlingua translation or transfer based translation [7].

Rule based direct translation is a word by word substitution method of translation followed by grammatical adjustments wherever required. This approach is unidirectional translation approach where no intermediate representation is used in the process i.e. source words are converted directly to target words [8].

Rule based interlingua translation makes use of a language independent intermediate representation of the source language text to represent its meaning. It works in two phases. In first phase, source language text is converted to intermediate representation and in second phase intermediate representation is converted to target language text. Multilingual translation systems are generally based on this approach [9].

Rule based transfer translation also use intermediate representation like Interlingua translation approach but the intermediate representation used in this approach is not fully

independent of languages involved. The intermediate representation used there is partially dependent on source language and target language, and partially dependent on their structural differences [10].

B. Corpus-Based Approach

A Corpus is a huge collection of knowledge data containing parallel translations of both the languages involved in the process. The system obtains translation knowledge from this raw text data available in the bilingual corpus. There can be Statistical based or Example based Corpus approach [11].

The Statistical based approach analyzes bilingual corpora to retrieve some parameters which we use in statistical models for translation process. Initially, every sentence of the target language in corpora is assumed as a translated sentence of the source language sentence. Every assumed translated sentence is assigned with a probability and the highest probability sentence is picked up as the final target language sentence. Searching the huge corpora arises performance issues, so some approaches like heuristics are used to reduce search space [12].

Example based approach uses analogy as the basic idea. A set of source sentences along with the corresponding target translations using a point to point mapping are provided to the system as examples. The system understands those example translations by training, so these are called training sets from where the system gets the knowledge. The trained system becomes capable to translate similar sentences [13].

In this paper Rule-Based Approach is used and some set of rules are defined to extract data elements from the Punjabi Language query sentence.

III. METHODOLOGY

When the user enters his/her query in Punjabi Language, the Punjabi language sentence need to be processed to extract Entity information from the natural language sentence. A Punjabi language sentence entered to query some data from the relational database will contain all the required information to extract data from the database. Suppose a Punjabi language sentence:

ਉਹਨਾਂ ਕਰਮਚਾਰੀਆਂ ਦਾ ਨਾਮ ਅਤੇ ਪਤਾ ਦੱਸੋ ਜਿਹਨਾਂ ਦੀ ਤਨਖਾਚ 35000 ਤੋਂ ਵੱਧ ਹੈ।

The above sentence contains all the needed information that is sufficient to extract required data from the database. But this sentence need to be processed to extract that Entity information about which the data is demanded such as Entity, attributes and condition (if any). It involves following steps shown in Figure 1.

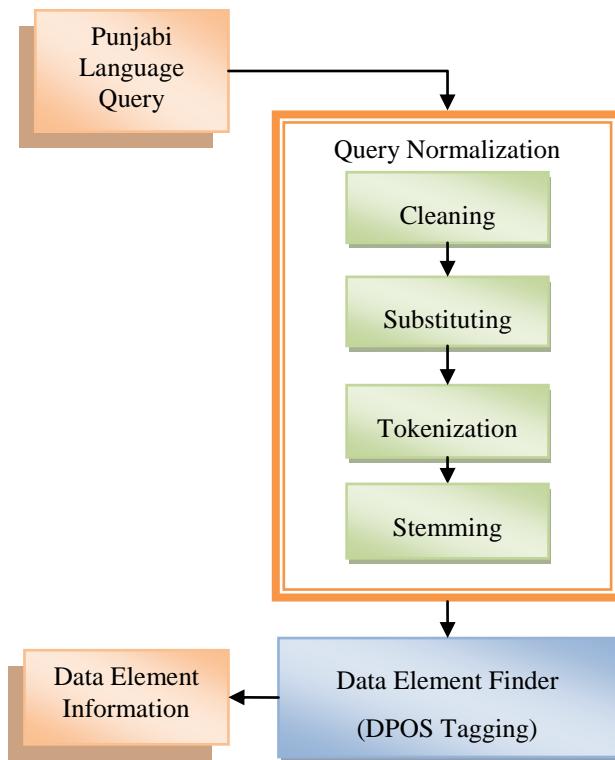


Figure 1. Data elements extraction

A. Query Normalization

It is a pre-processing phase and is presented in Harjit Singh et al. [1]. First of all the Punjabi language sentence need to be cleaned i.e. any special character and noise from the query sentence is removed. Special characters may be comma (,), semicolon (;), bar (|), question mark (?), exclamation (!), dot (.), colon (:) etc. The next step for normalization is substituting alternative simple words in place of more complex words or multiword expressions using substitution database tables. For example, the word ਸਿਰਨਾਵਾਂ is replaced with synonym and simple alternative word ਪਤਾ. More synonyms and multiword expressions are added to the substitution database tables presented in Harjit Singh et al. [1] to simplify the 'Data Element Finder' phase. For example, ਦੱਸੇ is replaced with ਦੱਸ and multiple expressions such as ਜਿਸ ਦਾ, ਜਿਸ ਦੀ, ਜਿਸ ਦੇ, ਜਿਸ ਦੀਆਂ, ਜਿਹਦਾ, ਜਿਹਦੇ, ਜਿਹਨਾਂ ਦਾ, ਜਿਨ੍ਹਾਂ ਦਾ, ਜਿਹਨਾਂ ਦੀਆਂ, ਜਿਹੜਿਆਂ ਦਾ etc. are replaced with single word ਜਿਸਦਾ. A total of 408 such non-noun substitutions are identified and substitution database tables are upgraded. Two separate substitution database tables are used i.e. Noun-Substitution Table and NonNoun-Substitution Table. Third step is to tokenize the query and divide the sentence to individual words for further processing. Forth

step of normalization is stemming i.e. removing any suffixes attached to any tokens. Restructuring and upgrading of substitution database tables resulted in more accurate results in normalization phase. The above sentence after normalization provides a set of 12 following tokens shown in Table 1.

Table 1: Query Normalization

ਉਹਨਾਂ
ਕਰਮਚਾਰੀ
ਦਾ
ਨਾਂ
ਤੇ
ਪਤਾ
ਦੱਸ
ਜਿਸਦਾ
ਤਨਖਾਚ
35000
ਵੱਧ
ਹੈ

B. Data Element Finder(DPOS Tagging)

Now the set of tokens are scanned in a number of ways to find Data Element tokens which are directly related to data retrieval. These tokens shall correspond to Entity, attributes and conditions (if any). DPOS (Data Part of Speech) tagging is the step that requires traversing the token set back and forth a number of times. As the tokens are stored in an array, it makes the traversal easier.

It is a rule based approach and a number of rules are generated to extract data elements from normalized Punjabi query sentence. For example, it is found that the word which appears before the word "ਦਾ" corresponds to the entity about which the information is sought in the query. For example in the above Punjabi language sentence "ਕਰਮਚਾਰੀ" appears before "ਦਾ", so "ਕਰਮਚਾਰੀ" is the entity about which information is demanded. So this token of query sentence is considered as <ENTITY>. Now what information is sought about this <ENTITY>?

The Normalized sentence structure of Punjabi language analysed and it is found that those attribute words appear after the word "ਦਾ". The demanded information may be a single attribute word such as "ਨਾਂ" or it may be multiple property words such as "ਨਾਂ ਤੇ ਪਤਾ". Multiple attribute words list ends with the attribute word specified after the word 'ਤੇ'. These tokens of query sentence are considered as

<ATTRIBUTE>. If no specific property word is found in the sentence, it means that whole available information is sought.

In Normalized Punjabi language query, a condition may appear in the sentence after the word "ਜਿਸਦਾ". ਜਿਸਦਾ is a condition token. The word after the condition token is the attribute word on which condition is applied followed by the attribute value and then the comparison word "ਵੱਧ", "ਘੱਟ" etc. If no comparison word is given in the sentence then "equal to" is assumed. These tokens of query sentence are considered as <CONDITION-ATTRIBUTE>, <CONDITION-VALUE> and <CONDITION-OPERATOR>. There may be multiple conditions in the same Punjabi language sentence. These are separated with words "ਤੇ" or "ਜਾਂ" and are considered sequential list like <CONDITION-ATTRIBUTE-1>, <CONDITION-ATTRIBUTE-2> and so on. Table 2 shows the extracted data elements.

Table 2: Considered Data Elements

ਕਰਮਚਾਰੀ	<ENTITY>
ਨਾ	<ATTRIBUTE-1>
ਪਤਾ	<ATTRIBUTE-2>
ਤਨਖਾਰ	<CONDITION-ATTRIBUTE-1>
35000	<CONDITION-VALUE-1>
ਵੱਧ	<CONDITION-OPERATOR-1>

IV. IMPLEMENTATION AND TESTING

The above methodology is implemented using Visual Studio.NET 2010 with C#.NET as programming language and SQL Server as backend database. The screen shot of interface of the implementation is shown in Figure 2.

In the implementation, when a data element is identified it is tagged with a small tag i.e. {EN} for ENTITY, {AT1} for ATTRIBUTE-1, {CA1} for CONDITION-ATTRIBUTE-1, {CO1} for CONDITION-OPERATOR-1, {CV1} for CONDITION-VALUE-1, {LO1} for LOGICAL-OPERATOR-1 and so on.



Figure 2: Implementation Interface

Also Punjabi comparison words are replaced with their correspond symbols such as > for ਵੱਧ, < for ਘੱਟ, >= for ਵੱਧ-ਜਾਂ-ਬਰਾਬਰ, <= for ਘੱਟ-ਜਾਂ-ਬਰਾਬਰ and logical connectors (ਜਾਂ, ਤੇ) are also replaced i.e. OR for 'ਜਾਂ', AND for 'ਤੇ'. The conditions are also realigned with the format <ATTRIBUTE><OPERATOR><VALUE>.

The implementation is tested using 585 Punjabi language query sentences with various sentence formats and from different domains. To automate and fasten the testing process, the Punjabi language query sentences are stored in a Unicode text file. These queries are read one by one on

pressing the button  shown in Figure 2. The current query number is shown as a label on the button. Some of the test queries and the result of extracting data elements by tagging them are shown in Table 3.

Table 3: Some Test queries and the tagged data elements

Punjabi Language Query	Tagged Data Elements
ਉਹਨਾਂ;,, ਵਿਦਿਆਰਥੀਆਂ/?.} ਦਾ	ਵਿਦਿਆਰਥੀ{EN} ਨਾ{AT1}
ਨਾਮ{ ਅਤੇ ਬਉ-ਪਤਾ ਦੱਸੋ ਜਿਹਨਾਂ ਦੇ ਅੰਕ 50 ਨਾਲੋਂ ਜਿਆਦਾ ਹਨ।	ਪਤਾ{AT2} ਅੰਕ{CA1} >{CO1} 50{CV1}
ਉਹਨਾਂ, ਮਰੀਜ਼? ਦਾ {ਨਾਮ} {ਬਿਮਾਰੀਆਂ} ਅਤੇ ਬਉ-ਪਤਾ ਦੱਸੋ ਜਿਹਨਾਂ ਦੀ ਉਮਰ 50 ਨਾਲੋਂ ਜਿਆਦਾ ਹੈ	ਮਰੀਜ਼{EN} ਨਾ{AT1} ਬਿਮਾਰੀ{AT2} ਪਤਾ{AT3} ਉਮਰ{CA1} >{CO1} 50{CV1}
ਉਹਨਾਂ ਕਰਮਚਾਰੀਆਂ ਦਾ ਨਾਮ ਅਤੇ ਪਤਾ ਦੱਸੋ ਜਿਹੜੇ 35000 ਤੋਂ ਵੱਧ ਜਾਂ ਬਰਾਬਰ ਤਨਖਾਰ ਲੈ ਰਹੇ ਹਨ	ਕਰਮਚਾਰੀ{EN} ਨਾ{AT1} ਪਤਾ{AT2} ਤਨਖਾਰ{CA1} >= {CO1} 35000{CV1}
ਉਹਨਾਂ ਕਰਮਚਾਰੀਆਂ ਦਾ ਨਾਮ ਅਤੇ ਪਤਾ ਦੱਸੋ ਜਿਹਨਾਂ ਦੀ ਤਨਖਾਰ 35000 ਤੋਂ ਵੱਧ ਜਾਂ ਬਰਾਬਰ 25 ਅਤੇ ਪਤਾ ਦੱਸੋ ਜਿਹੜੇ 25000 ਤੋਂ ਘੱਟ ਤੇ ਸ਼ਹਿਰ ਮੂਨਕ ਜਾਂ ਤਨਖਾਰ 25000 ਤੋਂ ਘੱਟ ਹੈ	ਕਰਮਚਾਰੀ{EN} ਨਾ{AT1} ਪਤਾ{AT2} ਤਨਖਾਰ{CA1} >{CO1} 35000{CV1} OR{LO1} ਉਮਰ{CA2} <{CO2} 25{CV2} AND{LO2} ਸ਼ਹਿਰ{CA3} = {CO3} ਮੂਨਕ{CV3} OR{LO3} ਤਨਖਾਰ{CA4} <{CO4} 25000{CV4}

Any implementation is efficient if it produces higher percentage of correct results for various inputs. The testing for accuracy percentage is done with 585 Punjabi queries with different sentence formats and from different domains. The results of testing are as follows:

Total number of input queries, $T_n = 585$

No. of outputs with correctly tagged data elements, $C_n = 549$

Accuracy percentage = $(C_n/T_n) \times 100 = 549/585 \times 100 = 93.8\%$

Erroneous outputs were due to incorrect tagging to words that were not identified by the system due to their absence in the database tables.

V. CONCLUSION

Almost all Relational Database Management Systems use Structured Query Language (SQL) to store and retrieve data from databases. It means that to work with these databases, a user must have the knowledge of SQL which limits the use of databases by normal users. The Natural Language sentence needs to be processed to extract Data Elements such as Entity, Attributes, Condition etc. to access the required data from the database. This paper presented the process of extracting Data Elements from the Punjabi Language query, so that this information can be used for further processing to retrieve data from the database. The methodology implemented in C#.NET and SQL Server database, provided 93.8% accuracy for Punjabi language queries related to different domains with different sentence formats.

REFERENCES

- [1] Joseph, Sethunya & Sedimo, Kutlwano & Kaniwa, Freeson & Hlomani, Hlomani & Letsholo, Keletso, "Natural Language Processing: A Review", Natural Language Processing: A Review, **6**, 207-210, 2016
- [2] Reshamwala, Alpa & Mishra, Dhirendra & Pawar, Prajaka, "Review on Natural Language Processing", IRACST – Engineering Science and Technology: An International Journal (ESTIJ), **3**, 113-116, 2013
- [3] Srivastava, Siddhant & Shukla, A & Tiwari, Ritu, "Machine Translation : From Statistical to modern Deep-learning practices", arXiv preprint, arXiv:1812.04238, 2018
- [4] Sundar Ram R, Vijay & Lalithadevi, Sobha, "Overview of Verb Phrase Translation in Machine Translation: English to Tamil and Hindi to Tamil", Conference: the 10th annual meeting of the Forum for Information Retrieval Evaluation, **6-10**. DOI: 10.1145/3293339.3293341, 2018
- [5] T.K, Bijimol & , Professor & Abraham, John, "A Study of Machine Translation Methods An Analysis of Malayalam Machine Translation Systems", Conference: NCILC-14, At CUSAT, Cochin, 2014
- [6] Sinhal, Ruchika & Gupta, Kapil, "Machine Translation Approaches and Design Aspects" IOSR Journal of Computer Engineering, **16**, 22-25, DOI:10.9790/0661-16122225, 2014
- [7] Fadiel Alawneh, Mouiad & Sembok, Tengku, "Rule-Based and Example-Based Machine Translation from English to Arabic", Conference: Bio-Inspired Computing: Theories and Applications (BIC-TA), 2011 Sixth International Conference, **343 – 347**, DOI: 10.1109/BIC-TA.2011.76, 2011
- [8] M. D. Okpor, "Machine Translation Approaches: Issues and Challenges", IJCSI International Journal of Computer Science Issues, Vol. **11**, Issue **5**, No **2**, September 2014
- [9] Shantanoo Dubey, "Survey of Machine Translation Techniques", International Journal of Advance Research in Computer Science and Management Studies, Special Issue, Volume **5**, Issue **2**, February 2017
- [10] Krishnamurthy, Parameswari, "Development of Telugu-Tamil Transfer-Based Machine Translation System: An Improvisation Using Divergence Index", Journal of Intelligent Systems, DOI: 10.1515/jisys-2018-0214, 2018
- [11] Dash, Niladri & Ramamoorthy, L., "Corpus and Machine Translation", In book: Utility and Application of Language Corpora, pp, **193-217**, DOI: 10.1007/978-981-13-1801-6_12, 2019
- [12] Mahata, Sainik Kumar & Mandal, Soumil & Das, Dipankar & Bandyopadhyay, Sivaji, "SMT vs NMT: A Comparison over Hindi & Bengali Simple Sentences", Proc. of ICON-2018, Patiala, India, pages **175–182**, December 2018
- [13] Carl, Michael & Way, Andy & Daelemans, Walter, "Recent Advances in Example-Based Machine Translation", Computational Linguistics, **30**, 516-520, DOI: 10.1162/0891201042544866, 2004

Authors Profile

Mr. Harjit Singh received the MCA (Master in Computer Applications) degree from IGNOU (Indira Gandhi National Open University), New Delhi, India and M.Phil.(CS) degree from Global Open University, Nagaland, India. He is working as Assistant Professor (Senior Scale) in Computer Science at Punjabi University Neighbourhood Campus Dehla Sehan, Sangrur, India. He is pursuing Ph.D. degree from RIMT University, Mandi Gobindgarh (Punjab). His current research interests include Natural Language Processing, Machine Translation, Artificial Intelligence.



Dr. Ashish Obroi is working as Professor in Computer Science and Engineering at School of Engineering, RIMT University, Mandi Gobindgarh, Punjab, India. He completed his Ph.D. in Computer Science and Engineering from Maharishi Markandeshwar University, Mullana, India. He is skilled and expertise in Image Processing, Medical Imaging, Image Segmentation, Diagnostic Imaging and Image Reconstruction.

