Sindh Univ. Res. Jour. (Sci. Ser.) Vol.49(3) 513-516 (2017)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Handling Ambiguities in Sindhi Named Entity Recognition (SNER)

D. NAWAZ[++], S. A. AWAN**, Z. A. BHUTTO*, M. MEMON*, M. HAMEED*

Institute of Information and Communication Technology, University of Sindh, Jamshoro, Pakistan

Received 3rd April 2016 and Revised 7th July 2017

Abstract: Natural language processing is considered the least advanced field of artificial intelligence and the main reason is the challenges in form of differences of accent, variations in script and other problems. A working Sindhi Named Entity recognition impose a challenging task, ambiguity; where the word can be a name of person and the day of the week. It is very hard to understand the difference between two meanings whereas the word is same in writing as well as in pronunciation. This research work presents the handling of ambiguities in Sindhi Entity Recognition (SNER). A rule based approach with the help of indexing has been used to handle ambiguities in SNER. The Sindhi Entity Recognition System recognizes ambiguous words of Sindhi language and produces corresponding meaning successfully. The SNER system clearly understands from the sentence and produces the accurate meaning corresponding to the ambiguous word and its available meaning.

Keywords: Entity Recognition, Information retrieval, Sindhi, Script, Ambiguity, Rule Base.

1. INTRODUCTION

Natural language processing is the understanding of the human languages by computers and generating human language structures by computers or machines. Named entity recognition is the branch of information retrieval where language texts are retrieved and named according to the predefined classes called entities. Named Entity Recognition (NER) is a automatic elements process of recognition the entities from given context and labeled with pre-defined tags like Person_ Name, Location, Measure and etc. I will not take string from sentence and match with database only but it will understand the complete sentence and takeout the actual usage meaning of that entity and extract it. Many of the related languages of Sindhi have their own named entity recognition systems such as Urdu has many of the works are available (Singh *et al.*, 2012; Jahangir *et al.*, 2012; Becker *et al.*, 2002). A lot of work is also available for Arabic entity recognition (Rahman *et al.*, 2010; Benajiba *et al.*, 2009; Abdul-Hamid and Darwish, 2010; Benajiba, 2009). Some other languages are also enriched with their respective entity recognition systems.

1.1 Introduction of Ambiguity:

Ambiguity is one of the toughest and difficult tasks in natural language processing field. Every Language has ambiguity problem and it is also difficult for person to find out the meaning of ambiguous words from sentences, but for computer program it is even ambiguous word but how computer recognizes the meaning of ambiguous words in sentence. Ambiguity: when ambiguous words occur in sentence or context is called ambiguity process. And ambiguous words are those words that have more than one meaning. Or word has multiple meanings is called ambiguous word. Some of the ambiguous words are shown in **(Table 1).**

Table 1: List of Ambiguous words in Sindhi Language

List of some ambiguous entities/words in Sindhi					
سومر	همت	سينگار	قيمت	امانت	شاهد
راحت	اقرار	نعمت	قائم	آزادي	انب
آچر	ضامن	عظيم	بهادر	پرهه	مرڪ
سنڌو	دعا	ضامن	ٿورو	سونهن	نواز
امير	برڪت	غلام	روشن	مراد	قدر
خوشبوءِ	احسان	پانهو	ضمير	شعبان	ڏنو
مدد	سَڃڻ	انعام	اخلاق	قربان	وسايو
روشن	اعتبار	اعتراض	جهڙڪ	احسان	رکيو

The words in table 1 are ambiguous in Sindhi language when occur in sentence will result a challenging job to recognize actual meaning of these words from sentence due to multiple meanings. Ambiguous words can be understood by following examples. The Sentence in Sindhi " مقصود کي راحت ملي"which may produce two meanings 1) Maqsood got relaxed and 2) Maqsood met with Rahat. Another

[++]Corresponding Author E-mail: dill.nawaz@gmail.com,
*Institute of Information and Communication Technology, University of Sindh, Jamshoro, Pakistan
**Benazir Bhutto Shaheed University, Cheel Chowk, Lyari Karachi, Sindh, Pakistan.
***Institute of Business Administration, University of Sindh, Jamshoro, Pakistan

ambiguity example is “چنيسر منهنجو شاهد آهي “ in which ”شاهد” which means witness as well as the name of person. In above sentences two ambiguous words “راحت” and ”شاهد” can be used with multiple meanings in sentence. “راحت” can be used as name as well as adjective and شاهد is name of a person and used as a noun meaning witness. So, it is a challenging job for a computer to understand that whether it is name of a person or it has been used in another sense, as a witness.

2. Related Material:

Traboulsi (2009) presented a local grammar based approach to extract Arabic entities from the biochemistry material. The other experiments were performed to extract address, time and other entities. The technique has been employed to extract English names and the approach has also been used in Korean, Turkish, French, Chinese and Portuguese news text material.

Alasiry (2015) proposed a query based entity recognition based on search. A novel method has been proposed for understanding candidate entities by segmenting query and grammar annotation. Another novel method based on seed expansion has also been proposed for classifying the entities. An additional characteristic of explanation analysis helps to refine the results for three main categories. These categories include verb, adjective and noun.

Alotaibi (2015) proposed a novel approach which introduced fifty classes instead of limited from three to ten. These classes are based on two levels. The fine grained 50 classes produce and advantage over the traditional low number of classes. Annotated or labelled data are required for training because the fine grained NER system is based on dual machine learning systems namely Conditional Random Fields and Maximum Entropy.

3. Sindhi Named Entity Recognition:

Sindhi Entity recognition system is an application which extracts information from the Sindhi sentences and classify entities into predefined classes such as the name, caste, place, date, place and others. Some other entities include numbers from zero to trillion, years, designation along with organizations, measurement, common local and international brands and others. Many of the words in Sindhi language have two or more meanings which make sentences even difficult to understand. These words are called ambiguous words.

4. Handling ambiguities in Sindhi Language

To understand two sentences are presented followed by the procedure that can handle the ambiguity problem in Sindhi Language. For example,” مقصود کي راحت ملي” is a Sindhi sentence in which an ambiguous word “راحت” is used. The word “راحت” can be used as name of a person as well as it can be used as adjective in sentence. But in above sentence راحت has used as name now let’s see how our program solve this ambiguity problem from this sentence.

To handle, this sentence is given as input and tokenization is performed followed by making sets from those tokens and store sets in one list. The list is then reversed and the matching process is started with the help of database. On the success of matching it is recognized entity and it is stored in another list called “Recognized List of Entities”. When stored successfully on the list, the word is automatically deleted from index or position and at the same time it is replaced with “-” sign. The process of tokenization, set preparation, storing in list, reversing of a list and other activities are repeated until all of the words have been replaced with “-” sign. Sindhi Named Entity Recognition (SNER) will recognize these four entities as shown in **(Fig. 1).**

Recognized Entities	Words in Sentence
Person_Name	مقصود
Non_Entitiy	کي
Person_Name	راحت
Non_Entity	ملي

Fig.1: Recognized entities in first example

The system found four entities as shown in **(Fig. 1).** Now these are recognized entities but still a dual meaning word can produce a wrong meaning ultimately the wrong sentence meaning. It is even harder to understand which word is the ambiguous one. There are 12 tables created in database, each table consists of thousand entities of each tag. These tags are created to understand the various forms of words so that ambiguity can be understood by the program. Each entity is dropped in these tables and if each entity found at more than 1 tables then the entity is considered ambiguous. In above example the system checked all entities in database and all entities found in just one table except the word “ راحت was found in two tables (Person_Name, Terms). The system has found only the ambiguous word but still unable to estimate the actual meanings of the ambiguous word yet. It knows only that this is an ambiguous word in sentence. The SNER system will apply rule based approach on this ambiguous word to find out the actual possibilities. The overall approach of handling ambiguities in SNER is shown in **(Fig. 2).**

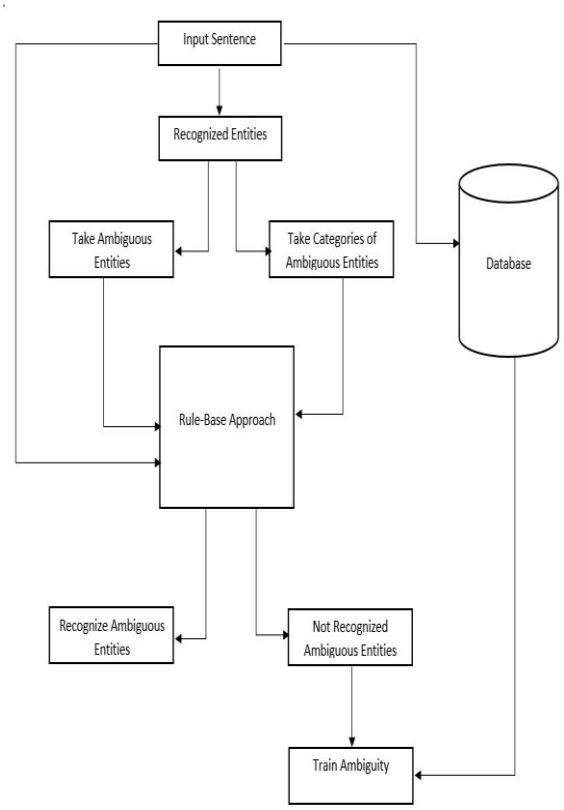

Fig.2: Handling ambiguities in SNER

4.1 Rule Based Approach:

In computer science rule based approach rules or principles are defined with specific form of data. According to those principles or rules, the system will act upon and predict results as shown in Figure 3. In our application , we have defined 13 tables for principles or rules. Resuming previous example "مقصود کي راحت ملي", system found "راحت" as ambiguous word and categorized as Person_Name and terms. The system will send this sentence "مقصود کي راحت ملي" , the index of ambiguous word and categories of that ambiguous word in a sentence to resolve the ambiguity problem as shown in **(Fig. 3).** A sample of recognizing entities has been depicted in **(Fig. 4).** From actual given sentence as in this case "مقصود کي راحت ملي" system will take five right words from the right position of "راحت" and five left words from the left position of "راحت". If either side (right/left), there are less than five words then system will tag "xyz" as entity to complete five words at right side and left side of ambiguous word. The counting and tagging process is shown in **(Table 2).**

Fig.3: Rule base approach table for Person_Name

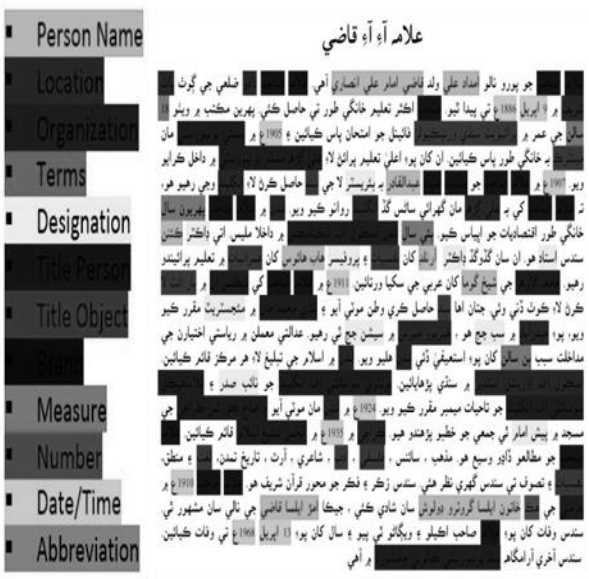

Fig. 4: SNER output for recognizing Sindhi Named Entities

The program will check in both Rule tables namely Person_Name , Terms and after careful calculation it will predict the value and the number is given as output as in our example the output is 17 for Person_Name and 12 for Terms.

Table 2: Left / Right tagging process

Actual Sentence: "مقصود کي راحت ملي"		
RULE BASE APPROACH	Left- Side Five Words	Right-Side Five Words
	L1 = ملي	R1 = کي
	L2 = XYZ	R2 =مقصود
	L3=XYZ	R3 = XYZ
	L4=XYZ	R4 = XYZ
	L5=XYZ	R5 = XYZ

The system will pick up the value as in our case 17 and the resolved entity has the greater chances to be recognized as Person_Name. So, word "راحت" in sentence "مقصود کي راحت ملي" is used as Person_Name.

5. RESULTS AND DISCUSSION

For the recognition of entities multiple classes along with supporting tables have been created so that the entities can easily classified. The additional job of handling ambiguities has been added to the already built application of Sindhi Named Entity Recognition (SNER). The ambiguities or ambiguous words create a challenging problem and handling of these ambiguous words have been tackled successfully. Our rule based approach solves the multiple number of ambiguous words in Sindhi and more than one hundred words (ambiguous) are successfully recognized by our approach. The system understands, processes and recognize successfully which was even a challenging job for a human being as the ambiguous words have remained a challenging problem for NLP researchers.

6. CONCLUSION

The research on Sindhi entity recognition is at infant level and lot of efforts are needed. Very little work is available in this area especially Sindhi Language. The SNER is a primer effort to mark our Sindhi language with other enriched languages of entity recognition. The system successfully recognizes the multiple ambiguous words in SNER using rule based approach. More ambiguous words will be tested and incorporated and the Machine learning approaches will be the additional feature of this system.

REFERENCES:

Abdel R., S. M. Elarnaoty, M. Magdy and A. Fahmy (2010). Integrated machine learning techniques for Arabic named entity recognition. IJCSI Int. J. Comput. Sci., 7: 27-36.

Abdul-Hamid, A. and K.Darwish, (2010). Simplified feature set for Arabic named entity recognition. Proceedings of the Named Entities Workshop. (NEWS' 10), ACM Press, USA., 110-115.

Alasiry, A. M. (2015). Named entity recognition and classification in search queries (Doctoral dissertation, Birkbeck, University of London).

Becker, D., B. Bennett, E. D. Davis, Panton, and K. Riaz. (2002), Named Entity Recognition in Urdu: A Progress Report. Proceedings of the International Conference on Internet Computing. IC 2002, Las Vegas, Nevada, USA, 757-761

Benajiba Y., (2009). Named entity recognition. Ph.D. Thesis dissertation, Universidad Polit´ecnica de Valencia, May, 44, 151-152.

Benajiba, Y., D. Mona. R. Paolo (2009), Arabic named entity recognition: A feature-driven study. The special issue on Processing Morphologically Rich Languages of the IEEE Transaction on Audio, Speech and Language Processing. 17(5), 926-934.

Chandio, A. A., M. Leghari, D. N. Hakro, S. Awan, A. H. Jalbani, (2016). A Novel Approach for online Sindhi Handwritten Word Recognition Using Neural Network, Sindh University Research Journal (Science Series) 48(1), 213-216.

Jahangir, F., W. Anwar, U. I. Bajwa, X. Wang, (2012), N-gram and Gazetteer List Based Named Entity Recognition for Urdu: A Scarce Resourced Language, Proceedings of the 10th Workshop on Asian Language Resources, Mumbai, India, 95–104,

Singh, U. P., V. Goyal, G. S. Lehal, (2012), "Named Recognition System for Urdu", COLING, Mumbai, India, 2507-2518.

Traboulsi, H. (2009). Arabic named entity extraction: A local grammar-based approach., IMCSIT'09. International Multiconference on Computer Science and Information Technology, Mragowo, Poland 139-143. IEEE.