SINDHI NAMED ENTITY RECOGNITION (SNER)

Dil Nawaz Hakro[1]

Maqsood Ahmed Hakro[2]

Intzar Ali Lashari[3]

Abstract:

Natural Language processing is one of the least advanced field of Artificial Intelligence (AI) due to the variety of languages and differences amongst the language rules. Named Entity Recognition (NER) is also a least advanced application of natural language processing and branch of information retrieval in which text is understood, classified, labeled and retrieved. A difference in language impose new challenges as some of the entity recognition systems for English perform at the level of human response. This paper introduces Sindhi Named Entity Recognition (SNER) as of its very first attempt for NER for this language. Sindhi is one of the seven languages of the history of the mankind. SNER is intelligently extract and classify entities from Sindhi text with pre-defined categories. Challenges and issues are analyzed and presented the solutions of these identified problems in Sindhi Named Recognition System. The system works on more than 200,000 words including Sindhi names, surnames, numbers, names of cities and other entities. The proposed system performs NER tasks successfully and presents 97% accuracy.

[1] Institute of Information and Communication Technology, University of Sindh, Jamshoro Pakistan

[2] Institute of Information and Communication Technology, University of Sindh, Jamshoro Pakistan

[3] Institute of Business Administration, University of Sindh, Jamshoro, Pakistan

Ambiguity handling is the next step along with the grammar understanding, which will lead to a Sindhi chat bot.

Keywords: *Entity Recognition, Sindhi, Human Language, Natural Language Processing*

Introduction

Artificial Intelligence is making machines more like human being and these machines learn and behave like human being. One of the branches of AI is natural language processing where a machines and computers are trained to understand and speak like human; called natural language understanding and natural language generation. Named Entity Recognition (NER) is one of the emerging branches of NLP having potential applications such as question answering, machine translation, information retrieval, clustering of text and others (Naji, 2011).

1.1 Introduction to Sindhi Language

Nearly 60 million speakers including 53 in Pakistan and 6 million in India depicts the importance of Sindhi language and the need of NLP applications. Sindhi is the member of Indo-Aryan languages and the Prakrit dialect is the base for the development of Sindhi. Sindhi is written in one of the popular three scripts namely Arabic, Devanagari and Roman whereas the Arabic script is widely employed by Sindh province in Pakistan and Devanagari script is used by Indian Sindhi dwellers. Sindhi script is considered one of the largest extension of the Arabic script as various languages such as Pashto, Urdu, Persian and Malay are the common examples of Arabic script extensions (Hakro,2015). Figure 1. depicts the alphabet of Sindhi script in various scripts.

1.2 Named Entity Recognition

As the name implies NER is searching meaningful entities such as names of the persons, names of the cities, various locations, date including months and years, weekdays, names of the organizations, cast names and other such type of information from the given text. The next job of NER is to classify these extracted entities from given text into various categories based on rule based approach or machine learning (Mansouri et al., 2008). The rule based approach heavily depends upon the grammar and linguistics knowledge which identifies the entities based on language rules. The other approach used for NER is based on machine learning approaches such as support vector machines, decision trees and hidden markov model for the extraction of entities from the sentences. Many of the south Asian languages (Goyal, 2008; Gali et al.,2008; Kumar et al.,2008), Indian languages (Sharma et al.,2011), such as Telugu (Srikanth and Murthy,2008), Bengali (Ekbal and Bandyopadhyay,2008) have their own NER systems. Some of the other languages like Urdu (Jahangir et al., 2012; Singh et al.,2012; Becker et al., 2002), Arabic (Benajiba et al.,2012; AbdelRahman et al., 2010; Abdul-Hamid and Darwish, 2010; Benajiba, 2009;), and other languages (Dey et al., 2014; Bandyopadhyay,2008) are enriched with their NER systems. There is no any work has been found regarding Sindhi named entity recognition, to be best of our knowledge.

Let us consider the example for Sindhi given in Table 1.

Table 1: An Example of a named entity in a sentence

Maqsood Ahmed and Sarang travelled to Ratodero to meet iqbal.
اقبال سان ملڻ لاءِ مقصود احمد ۽ سارنگ رتوديرو ويا

There are four named entities in the above Sindhi sentence, the named entity extractor would identify these entities and label them as shown in Table 2.

Table 2: A Sample of named entities classification

The Named Entity	**Tags**
(مقصود احمد, Maqsood Ahmed)	Person
(سارنگ, Sarang)	Person
(رتوديرو, Ratodero)	Location
(اقبال, Iqbal)	Person

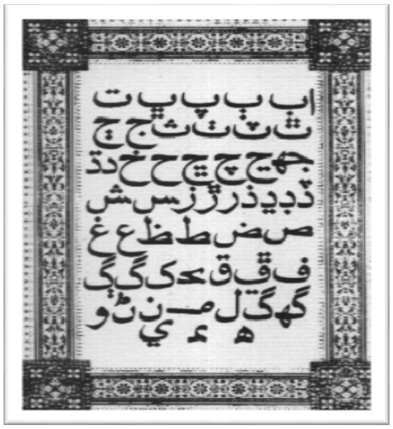


Figure 1: Alphabet of Sindhi Language (Arabic script)

2. Related Material

Shihadeh and Neumann (2012) presented a study on Arabic named entity recognition. A software has been developed which tokenizes, provides morphological analysis, tagging for the parts of the speech and other features. Gazetteer lookup method has been used as the first step for the improved performance for the classification of named entities. In Second step with the help of morphological analysis the affixes were removed for the sake of improved performance.

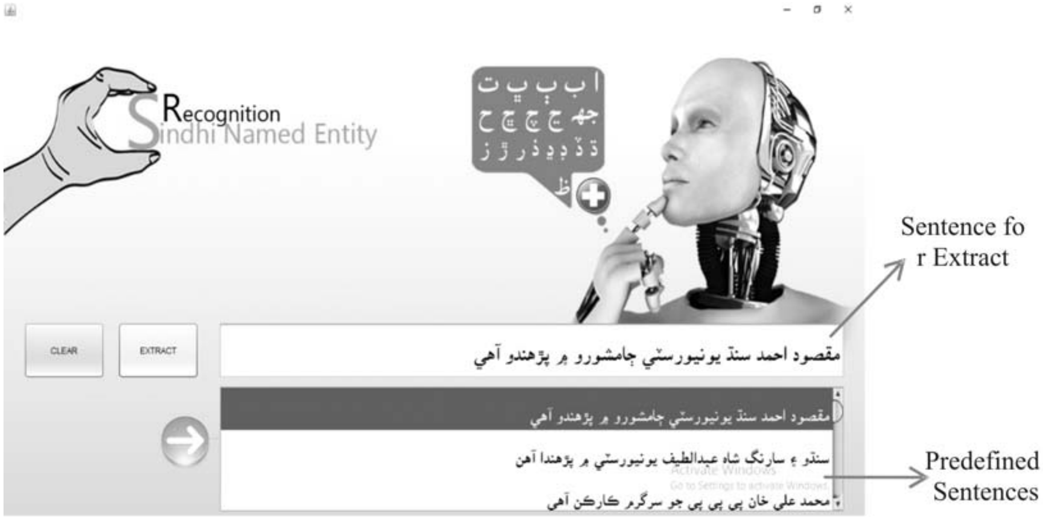
Asharef et al. (2012) presented an Arabic named entity recognition system in which experiments were performed on crime documents. A basic of crime analysis is a result of the entity recognition of such type of documents. The rule based approach has been used for their experiments on Arabic NER system. A formalized information and general crime indicator lists have been created. An Arabic named entities were annotated by matching corpus of crime domain. The number of Arabic NER entries were built and many of the syntactical rules were formalized, induced and applied to classify the Arabic entities of crime document based text. The system accuracy claimed is 90% and satisfactory performance along with effective approaches used.

Traboulsi (2009) presented a followed study of multilingual local generated grammar approach for the entity recognition of Arabic script. The approach has been applied on biochemistry literature in which commonly used and recursive phrases are checked. The approach has also been used to extract information like address expressions, date and time from the given letters. The methods have also been applied for the names extraction from various languages including, French, Turkish, Korean, Chinese, English and Portuguese. The local grammar approach has been used to extract Arabic names from the documents.

3. Materials and Methods

3.1 Sindhi Named Entity Recognition

Sindhi Named Entity Recognition (SNER) is an artificial intelligence based application which extracts the Sindhi Named entities from the text and classify into predefined classes or entities such as name of the person, numbers, years, months, place, organization, designation, brand, object, measurement abbreviation and time. The algorithm defined identifies the original entities with reference to the context of Sindhi language entities. Much of the work on various Sindhi application have been done such as Sindhi OCR (Hakro,2012; Hakro,2014,Hakro,2015), Sindhi text image database(Hakro,2015), Sindhi dictionary (Bhatti et al., 2014) and other applications (Chandio et al., 2016; Shah et al., 2004) but no any such work on Sindhi named entities has been found to the best of the authors knowledge. The ultimate goal of the research is to develop an entity recognizer in Sindhi and this study can be understood as the first step towards the Sindhi Chat bot. For this purpose, a Sindhi Entity recognition application has been created in java which identifies the Sindhi entities and can be executed on any machine. Successful identification will result in classification of the named entities and remaining are understood as non-entities. The interface of the custom built application is shown in Figure 1.



مقصود احمد سنڌ يونيورسٽي ڄامشورو ۾ پڙهندو آهي

مقصود احمد سنڌ يونيورسٽي ڄامشورو ۾ پڙهندو آهي

سنڌو ۽ سارنگ شاھ عبداللطيف يونيورسٽي ۾ پڙهندا آهن

محمد علي خان پي پي پي جو سرگرم ڪارڪن آهي

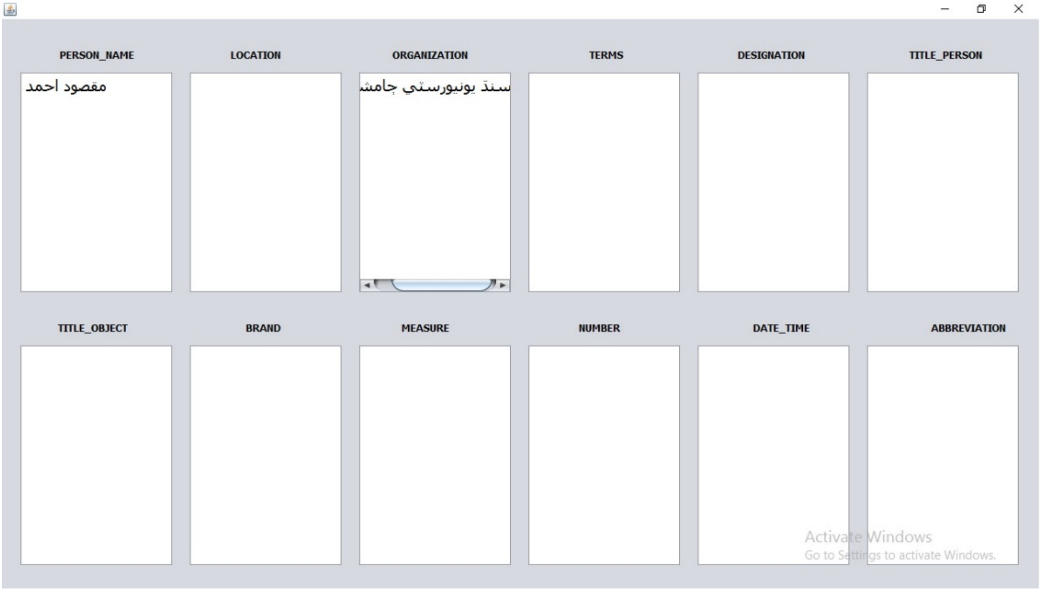
Sentence fo r Extract

Predefined Sentences

Figure 1: Custom built application for Sindhi Named Entity Recognition

The custom built application for Sindhi named entity recognition will take a single sentence as input as clearly shown in Figure 1. The application is capable of getting input of multiple sentences and identify the Sindhi entities and non-entities on a rule based approach based on repetitive tagging. For the sake of easiness and in situation where no resources are available, then many of the written examples have been given to user, so that the working mechanism of the SNER can be tested. Clear button clears the text area whereas the extract button triggers the automatic algorithm of SNER which extracts information and displays the result by identifying the entities and non-entities. Information extraction from the sentence along with results are shown in Figure 2 where a person name (مقصود) and the organization (سنڌ يونيورسٽي ڄامشورو) named entities have been extracted.

The SNER performs operation in two steps, namely NE detection (NED) where the text is identified forming the NE and the other is NE classification (NEC) which assigns the category of label to the found text span. The proposed system can recognize more than 222,000 entities including common names of the persons, popular brands (2000), countries along with their capital names, cities of Pakistan including Sindhi province and other entities.



PERSON_NAME	LOCATION	ORGANIZATION	TERMS	DESIGNATION	TITLE_PERSON
مقصود احمد		سنڌ يونيورسٽي ڄامش			
TITLE_OBJECT	BRAND	MEASURE	NUMBER	DATE_TIME	ABBREVIATION

The system is of an adoptive nature where the additional entities can be added in case when entity is not found then the system will ask the user if he/she wants to add the entity. On successful addition the entity is added and next time the entity will be the part of the system.

3.2 SNER Algorithm

The working mechanism of the SNER algorithm can be understood from the following example. The sentence " **سارنگ سائين دلنواز جو شاگرد آهي"** shown in Figure 3, which means Sarang is the student of sir Dil Nawaz**.** The process of creation of sets from the words is shown in Figure 4. The reason to create sets from the words is to handle the double word names. Such as two word names (مقصود احمد) and the other example is three words name (نصرت فتح علي) . After recognizing "آهي" as a non-Entity, the recognized non-entity "آهي" will be replaced with "- " sign and the remaining sentence will remain the same, without any change as shown in Figure 5.

آهي	شاگرد	جو	دلنواز	سائين	سارنگ

Figure 3: Sample Sentence given as input

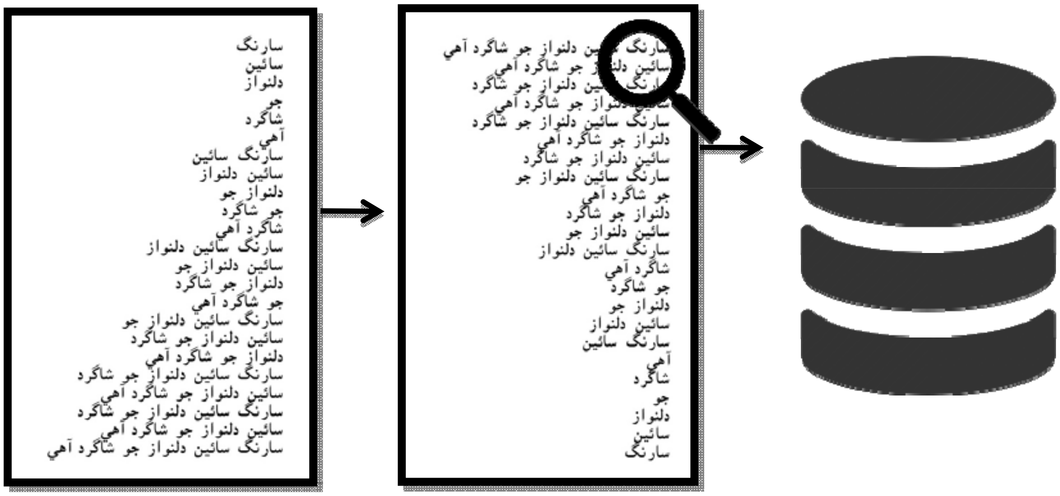


Figure 4: creation of sets from the given sentence (input)

-	شاگرد	جو	دلنواز	سائين	سارنگ

Figure 5: step wise creation of sets from the given sentence (input)

The iterative process of creating sets will continue until the entity is found and classified into predefined stored entities. The step wise process (remaining) is shown in Figure 6.

Figure 6: step wise creation of sets from the given sentence (input)

-	-	-	دلنواز	سائين	سارنگ

-	-	-	-	سائين	سارنگ

-	-	-	-	-	سارنگ

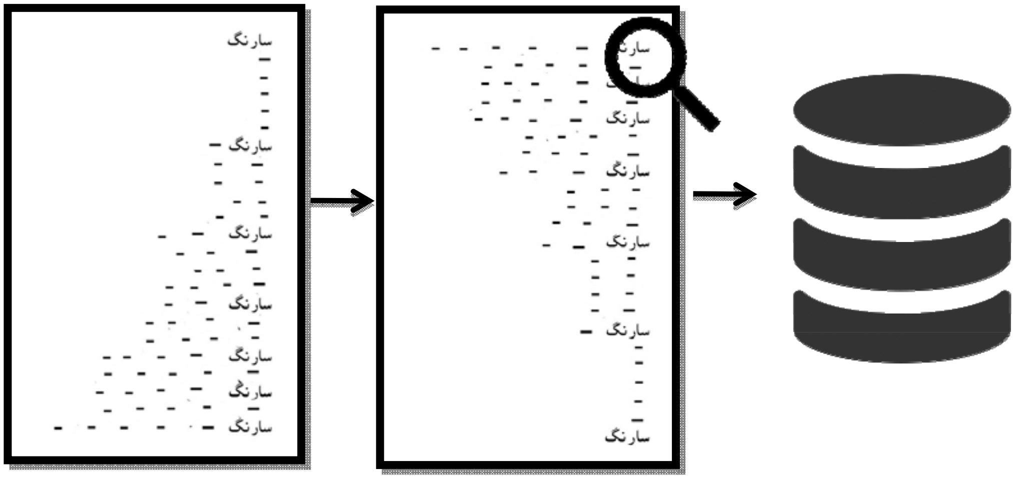
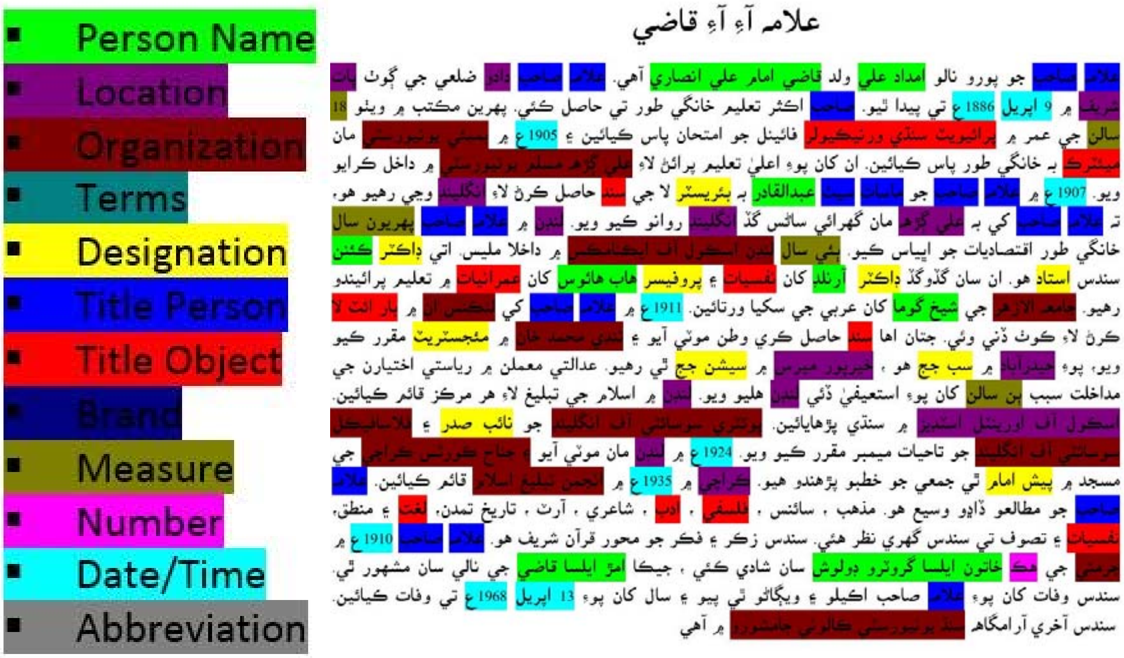


Figure 7: Graphical presentation of set creation from the given sentence (input)

Results and Discussion

To develop this NER system for Sindhi language, one must know Sindhi language so that problems and challenges can be studied properly, which may occur in the developing of NER system for Sindhi language. The proposed system successfully extracts entities and classify into categories with an accuracy of overall 97%. The remaining errors are due to the ambiguous entities which needs a separate algorithm to handle which is assumed as future work of this study. We took some problems/challenges which are mentioned in (Dev et al., 2014) and studied carefully to provide solutions of those problems. A comprehensive algorithm has been presented for named entity recognition of Sindhi Language which can easily handle the problems such as Free Word order, Nested Named Entities, and Compound Named Entities considered as one of the major problems. To enhance the accuracy of the system ambiguity problem will be handled in future as it needs a lot of efforts to make computer understand about words with more than two meanings. The system can understand, extract hundreds and thousands of the Sindhi named entities and capable of adding more entities in the case; when entities are not found. This adaptive nature of the SNER will pave the way for better accuracy and understanding more entities of Sindhi language. A sample of recognizing entities has been depicted in Figure 8.



علامہ آءِ آءِ قاضي

علامہ صاحب جو پورو نالو امداد علي ولد قاضي امام علي انصاري آهي. علامہ صاحب دادو ضلعي جي ڳوٺ پاٽ
شريف ۾ 9 اپريل 1886ع تي پيدا ٿيو. صاحب اڪثر تعليم خانگي طور تي حاصل ڪئي. پهرين مڪتب ۾ ويٺو 18
سالن جي عمر ۾ پرائيويٽ سنڌي ورنيڪيولر فائينل جو امتحان پاس ڪيائين ۽ 1905ع ۾ بمبئي يونيورسٽي مان
ميٽرڪ بہ خانگي طور پاس ڪيائين. ان کان پوءِ اعليٰ تعليم پرائڻ لاءِ علي ڳڙهہ مسلم يونيورسٽي ۾ داخل ڪرايو
ويو. 1907ع ۾ علامہ صاحب جو ساٿي سيٺ عبدالقادر بہ بئريسٽر لا جي سند حاصل ڪرڻ لاءِ انگلينڊ وڃي رهيو هو،
تہ علامہ صاحب کي بہ علي ڳڙهہ مان گهرائي ساڻس گڏ انگلينڊ روانو ڪيو ويو. لنڊن ۾ علامہ صاحب پهريون سال
خانگي طور اقتصاديات جو اڀياس ڪيو. ٻئي سال لنڊن اسڪول آف اڪنامڪس ۾ داخلا مليس. اتي ڊاڪٽر ڪئنن
سندس استاد هو. ان سان گڏوگڏ ڊاڪٽر آرنلڊ کان نفسيات ۽ پروفيسر هاب هائوس کان عمرانيات ۾ تعليم پرائيندو
رهيو. جامعہ الازهر جي شيخ ڳوما کان عربي جي سکيا ورتائين. 1911ع ۾ علامہ صاحب کي لنڪنس ان ۾ بار ائٽ لا
ڪرڻ لاءِ ڪوٽ ڏني وئي. جتان اها سند حاصل ڪري وطن موٽي آيو ۽ سنڌي محمد خان ۾ مئجسٽريٽ مقرر ڪيو
ويو، پوءِ حيدرآباد ۾ سب جج هو ، خيرپور ميرس ۾ سيشن جج ٿي رهيو. عدالتي معملن ۾ رياستي اختيارن جي
مداخلت سبب ٻن سالن کان پوءِ استعيفيٰ ڏئي لنڊن هليو ويو. لنڊن ۾ اسلام جي تبليغ لاءِ هر مرڪز قائم ڪيائين.
اسڪول آف اورينٽل اسٽڊيز ۾ سنڌي پڙهايائين. يونٽري سوسائٽي آف انگلينڊ جو نائب صدر ۽ فلاسافيڪل
سوسائٽي آف انگلينڊ جو تاحيات ميمبر مقرر ڪيو ويو. 1924ع ۾ لنڊن مان موٽي آيو ۽ جاح طورنس ڪراچي جي
مسجد ۾ پيش امام ٿي جمعي جو خطبو پڙهندو هيو. ڪراچي ۾ 1935ع ۾ انجمن تبليغ اسلام قائم ڪيائين. علامہ
صاحب جو مطالعو ڏاڍو وسيع هو. مذهب ، سائنس ، فلسفي ، ادب ، شاعري ، آرٽ ، تاريخ تمدن، لغت ۽ منطق،
نفسيات ۽ تصوف تي سندس گهري نظر هئي. سندس زڪر ۽ فڪر جو محور قرآن شريف هو. علامہ صاحب 1910ع ۾
جرمني جي هڪ خاتون ايلسا گروٽرو ڊولوش سان شادي ڪئي ، جيڪا امڙ ايلسا قاضي جي نالي سان مشهور ٿي.
سندس وفات کان پوءِ علامہ صاحب اڪيلو ۽ ويڳاڻو ٿي پيو ۽ سال کان پوءِ 13 اپريل 1968ع تي وفات ڪيائين.
سندس آخري آرامگاهہ سنڌ يونيورسٽي ڪالوني ڄامشورو ۾ آهي

Figure 8: SNER output for recognizing Sindhi Named Entities

Conclusion

The information extraction strategy known as entity extraction or named- entity recognition (NER) concerns itself with detecting and classifying certain elements in a string of text contained within in a natural language document in our case Sindhi language. An integrated SNER has been presented to identify and extract entities as places, quantities, time expressions, or names of persons or locations. The system successfully identifies and extracts Sindhi entities and presented up to 97% accuracy. The work described in paper concerns with Information Extraction (IE) and more specifically, named entity extraction in Sindhi language which leads to machine translation and other applications.

References

- Dey A., Abedin Md J. Purkayastha, B. S.,(2014), "A Comprehensive Study of Named Entity Recognition on Inflectional Languages", Int. J. of Advanced Research in Computer Sci. and Software Eng, Vol. 4(4) pp 95-104,.

- Goyal A., (2008), "Named Entity Recognition for South Asian Languages" in Proceedings of the IJCNLP-08 Workshop on NER for South and South-East Asian Languages, Hyderabad, India.

- Mansouri. A., Affendey, L.S, Mamat A. (2008), Named Entity Recognition Approaches, International J. of Computer Sci. and Network Security IJCSNS, vol. 8 (2), Pp 339-344.

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

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

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

- Asharef, M., Omar, N., Albared, M., Minhui, Z., Weiming, W., & Jingjing, Z. (2012). Arabic named entity recognition in crime documents. Journal of Theoretical and Applied Information Technology, 44(1), 1-6.

- Hakro, D. N., Memon, M. Awan S. A., Bhutto Z. A, Hameed, M. (2016), 'Isolated Optical Character Recognition', Sindh University Research Journal (Science Series) 48(4), 839-844.

- Ekbal, S. Bandyopadhyay, (2008), Bengali Named Entity Recognition Using Support Vector Machine, Workshop on NER for South and South East Asian Languages, IJCNLP 2008.

- Jahangir,F., Anwar, W., Bajwa, U. I. Wang, X. (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, pp 95–104, December 2012.

- Hakro, Dil Nawaz, (2015), "Enhanced Segmentation and Feature extraction approaches for Sindhi Optical Character Recognition", PhD thesis Dissertation submitted to Universiti Science Malaysia (USM), Malaysia.

- Gali K., Surana, H., Vaidya, A. Shishtla,P. Sharma, D., (2008), Aggregating Machine Learning and Rule Based Heuristics for Named Entity Recognition, Workshop on NER for South and South East Asian Languages, IJCNLP 2008.

- Naji, F. D. A. (2011). Towards Developing Automatic Name-Entity-Recognition System for Arabic Text (Doctoral dissertation, College of Computer and Information Sciences Department of Computer Sciences at the College of Computer and Information Sciences, King Saud University).

- Sharma, P. Sharma, U. Kalita, J., (2011), Named Entity Recognition: A survey of Indian Languages, special vol. Language in India, Problem of parsing in Indian Languages, pp 35-40, May 2011.

- Bandyopadhyay, S.(2008), Multilingual Named Entity Recognition" Proceedings of the IJCNLP-08 Workshop on NER for SSEAL, Asian Federation of NLP, Hyderabad, India, pp 3–4, January 2008.

- Kumar,S. Chatterji, S. Dandapat, S. Sarkar, S., (2008), A Hybrid Named Entity Recognition System for South and South East Asian Languages, Workshop on NER for South and South East Asian Languages, IJCNLP 2008.

- Shihadeh, C., & ünter Neumann, G. (2012). ARNE: A tool for named entity recognition from Arabic text. In Fourth Workshop on Computational Approaches to Arabic Script-based Languages (CAASL4), located at the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA) (pp. 24-31).

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

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

- Benajiba, Yassine. Diab, Mona., Rosso., 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.

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