DBpedia based Ontological Concepts Driven Information Extraction from Unstructured Text

Adeel Ahmed
Department of Computer Science
Shaheed Zulfikar Ali Bhutto Institute of Science and Technology
Karachi, Pakistan

Dr. Syed Saif ur Rahman
Business Intelligence and Reporting Team
WiseTech Global
Sydney, Australia

***Abstract*—In this paper a knowledge base concept driven named entity recognition (NER) approach is presented. The technique is used for information extraction from news articles and linking it with background concepts in knowledge base. The work specifically focuses on extracting entity mentions from unstructured articles. The extraction of entity mentions from articles is based on the existing concepts from DBPedia ontology, representing the knowledge associated with the concepts present in Wikipedia knowledge base. A collection of the Wikipedia concepts through structured DBpedia ontology has been extracted and developed. For processing of unstructured text, Dawn news articles have been scrapped, preprocessed and thereby a corpus has been built. The proposed knowledge base driven system shows that given an article, the system identifies the entity mentions in the text article and how they can automatically be linked with the concepts to the corresponding entity mentions representing their respective pages on Wikipedia. The system is evaluated on three test collections of news articles on politics, sports and entertainment domains. The experimental results in respect of entity mentions are reported. The results are presented as precision, recall and f-measure, where the precision of extraction of relevant entity mentions identified yields the best results with a little variation in percent recall and f-measures. Additionally, facts associated with the extracted entity mentions both in form of sentences and Resource Description Framework (RDF) triples are presented so as to enhance the user's understanding of the related facts presented in the article.**

Keywords*—Ontology-based information extraction; semantic web; named entity recognition; entity linking

I. INTRODUCTION

The text contained in unstructured documents, such as news articles or scientific literature, is often replete with many different persons, organizations, places, time, spatial information, etc. These relevant subjects, generally referred to as entity mentions in unstructured text are cited in form of words or phrases. The information provided about all such entity mentions within the article may vary depending upon the context of the article. For example, an article discussing about a ministerial meeting may not elaborate on the profile or background information about each person attending the meeting. Similarly the article may cite a number of entities as organizations and places without necessarily explicating their background information. The lesser the information or facts mentioned about some entity mentions, the greater the chances that user or more specifically a reader may end up searching for background information on some of the mentions over web. Knowledge base, such as Wikipedia serves as the guide to background information to a large collection of concepts to which users could potentially relate their looked up entity mentions on internet. These concepts can also be associated with an equivalent unique hyperlink in Wikipedia. This leads to the problem of extracting entity mentions from unstructured text and linking the same to background information in Wikipedia. This is addressed as a knowledge base concept driven named entity recognition (NER) — information extraction technique, addressing both entity extraction or entity identification or entity chunking and entity linking. Subsequent to this, additionally relevant information from within news article in form of sentences and associated RDF triples is identified and presented.

The Information Extraction (IE) is defined as the task of extracting raw text from natural language based document [1]. The IE systems are responsible for processing of text from input document(s) to separate useful raw text from noisy and irrelevant text by eliminating irrelevant pieces of words or phrases in an attempt to establish further meaning of extracted terms as entities and associate relevant relationships amongst them [2]-[6]. The output as in form of textual data can either be used directly for the purposes of presenting it to the user, stored for further database oriented tasks, used for natural language processing or information retrieval tasks and applications.

NER is defined as the task associated with identification of specific terms or phrases referred to as entity mentions. The entity mentions are representative of names such as persons, organizations, places, date, time, locations, etc. It is one of the subtasks associated with information extraction which helps identify mentions to its one of known categories or classes as mentioned previously. The said task helps address natural language processing and associated information retrieval tasks as well.

Wikipedia serves as the most popular free encyclopedia on internet. It is a voluminous information resource providing users with background information on various different topics across a wide variety of disciplines. However, for the purpose of referring to concepts in Wikipedia, an open community DBpedia knowledge base representative of the Wikipedia resources to the extent of 4.58 million things is used. DBpedia provides with an ontology of classes representing available knowledge about vast number of concepts across Wikipedia pages. These concepts about different resources over

Wikipedia are categorized under classes such as thing, agent, food, place, etc. However, extraction of concepts classified as persons, a sub-class of agents associated with Pakistan is set as the focus here. The knowledge within unstructured wikipedia articles is stored in form of over 1.8 billion RDF triples, classified under different ontology classes. In this paper, the Wikipedia concepts are collected using the DBpedia ontology for further extracting the entity mentions from unstructured text.

The daily Dawn, the most popular and leading newspaper in Pakistan is used as unstructured news article text collection. As this research work focuses on domain-specific extraction of entity mentions from news articles, therefore it was aimed to develop news article corpus from Dawn newspaper website by web scrapping the news archive. This provides a wide variety of news article categories published over several years. However, for this research, articles published over 15 months in year 2015 and 2016 have been collected and preprocessed.

Having extracted and linked the entity mentions to concepts in the knowledge bases and extracted the associated facts, other type of entity mentions associated with the persons in the news articles, i.e. places, organizations, time, etc. can also be extracted. Moreover, this much-needed information could help extract relevant cross-document information, perform cross-lingual information extraction, identify a series of spatio-temporal events and generate summaries.

The subsequent part of the paper is organized as follows: Section II provides the details on related work associated with use of ontology in terms extraction, named entity recognition approaches, entity linking and facts extraction. Section III discusses the over details about the system, including problem definition, Wikipedia concepts collection, news archive corpus collection, news articles preprocessing, knowledge base concept driven named entity recognition (NER) — information extraction technique for extraction and linking entity mentions to concepts and facts extraction in form of sentences and RDF triples from news articles. Section IV discusses the results presenting extracted entity mentions from news articles along with the mapped Wikipedia URLs or DBpedia URIs and the related metrics, measuring the relevance and accuracy of retrieved results in terms of precision, recall and f-measure. Finally, in Section V the work is concluded and the case for future work is presented.

II. Related Work

The ontology is defined as conceptualization of a particular domain [7]. The ever evolving size of unstructured text and the information present in text could help identify both new facts and thereby any shortcomings in existing ontologies. The information present in text helps identify a relevant ontology and later using information extracting methods identify new instance information which could populate the existing ontology. Raghu A., Srinivasan R. and Rajagopalan [8] in their research have identified as to how ontology concepts can guide extracting relevant information from general text. However, the selected approach uses a variety of ontologies created by humans followed by which it identifies the appropriate ontology and thereby enabling to extract the information in form of triples from unstructured text. Another system called KnowRex [9] uses the ontology-based approach to extract common properties as in form of semantic information from unstructured text documents. This emphasizes the use of concepts as a guide to extract relevant information in form of triples. However, the said approach focusing on an ontology defining the encyclopedia is used so as the extracted keywords can be linked with the concepts with the background knowledge available in encyclopedia.

In text processing, NER is referred to as task for designating specific keywords or tokens, phrases from within text to their identifying entity classes such as persons, locations and organizations. Many NER systems would use either entropy [10] based supervised learning techniques, user driven rules or random fields [11]. However such systems because of their heavy dependence on voluminous corpora and tagged or labeled data lead to divergence from addressing specific domain [12].

In this regard, to identify the entity mentions, other NER systems emphasize on using NER systems based on syntactic features, knowledge base and structured repositories for specific domains such as academic literature to effectively increase the precision and recall measures of NER [13]. Roman P., Gianluca D. and Philippe Cudre-M. have proposed an n-gram inspection and classification based NER approaches and evaluating the same based on part-of-speech tags, decision trees and co-location statistics. Their NER approach evaluates to 85% accuracy in respect of scientific collections and easily surpasses the efficiency of other NERs based on maximum entropy. However, their NER proved to perform better when the use of external knowledge base such as DBLP was taken into account. The NER in general has mostly been applied on news article text in respect of identification of names of persons, company or locations. However there are a few exceptions where NER has been used for collections which are more domain-specific and these include extraction of genes, drugs and protein entities [14], [15]. Therefore, the motivation is to use an existing ontology as a guide to make use of concepts associated with knowledge base along with an implementation of NER over domain-specific collections for extraction of entity mentions from within unstructured text. But for this, domain specific unstructured news corpus specific to Pakistan was built.

In regards to entity linking systems ZenCrowd [16], learning to link with Wikipedia [17] and Wikify [18] by Mihalcea and Csomai have been proposed for a variety of entity linking problems. Furthermore, the research into extraction of facts in general and temporal in particular from semi-structured and structured Wikipedia articles would be relevant to identify facts from within unstructured text [19].

III. System Overview

A detailed overview of the problem, internal working of the proposed system, the related technologies and tools in carrying out the said research work, collecting processing of concepts from Wikipedia, building a news article corpus and extraction of facts from unstructured news articles are presented.

A. Problem Definition

In this paper, the task of identification of entity mentions in a specific domain of unstructured text and the association, commonly known as entity linking of domain specific concepts present in DBPedia ontology are addressed. To understand this further, it is aimed to identify and relate some important keywords known as entity mentions with the existing background knowledge present in Wikipedia knowledge base. All such background information about a specific concept mainly appears in form of a wiki page on Wikipedia. The same concept, representative of wiki page is a unique resource and is assigned a unique resource identifier (URI). So all such Wikipedia concepts are organized in a structured form under a variety of classes under DBPedia ontology. To undertake this task, concepts extraction from DBpedia followed by the identification of entity mentions from text articles is performed.

B. Framework

The framework is built on underlying three modules, namely Wikipedia concepts collection, corpus collection, concepts driven name entity recognizer and facts extractor. The overall architecture of the DBpedia concepts driven information extraction workflow is shown below in Fig. 1.

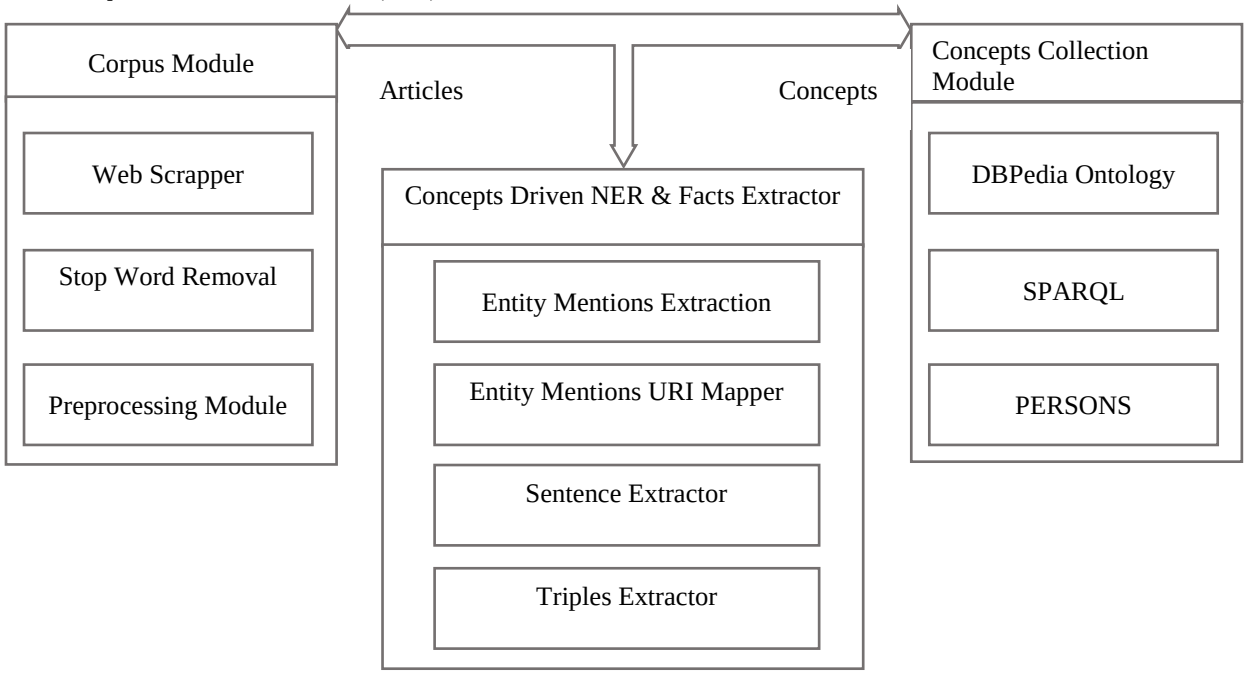

Fig. 1. Architecture of concept driven information extraction and entity linking.

C. Wikipedia Concepts Collection

A simple protocol and RDF Query Language (SPARQL) over DBpedia person class to extract concepts have been used. For the purpose of this research, three types of persons, i.e. politicians, singers and cricketers associated with Pakistan have been collected. Each person is described by a unique Uniform Resource Identifier (URI) in DBpedia, representing an equivalent person concept in Wikipedia URL.

The attributes collected in respect of person are categorized into required and complementary classes. As part of the underlying requisite research, the required class of data which include name of a person and person representing URI is taken into consideration. The complementary class of attributes includes birth name, birth date, death date, occupation, nationality and citizenship.

The extraction of persons associated with Pakistan is collected on nationality and citizenship. This is done because sometimes concepts are not all defined by the same attribute rather users chose to refer to one or more parameters to associate a person with a country. This in turn led to duplication of some persons where the persons based on their unique URI have been filtered.

The concept collection is based on persons defined in English language. The concept extraction is performed through OpenLink Virtuoso SPARQL endpoint. For example, a sample SPARQL query for extracting politicians is shown in Fig. 2 below:

```
" prefix category: <http://dbpedia.org/resource/Category:>
prefix dcterms: <http://purl.org/dc/terms/>

select distinct ?name ?person ?birthName ?birthDate ?deathDate
?occupation ?nationality  ?citizenship
where {
 ?person foaf:name ?name .
 optional { ?person a dbo:Politician . }

 ?person dcterms:subject category:Pakistani_politicians .

 optional { ?person dbo:nationality ?nationality .
          filter regex(?nationality, "Pakistani") .   }

 optional { ?person dbo:birthName ?birthName . }

 optional { ?person dbo:nationality ?nationality . }

 optional { ?person dbo:citizenship ?citizenship .
          filter regex(?citizenship, "Pakistan") .    }

 optional { ?person dbo:birthDate ?birthDate   .    }

 optional { ?person dbo:deathDate ?deathDate   . }

 optional { ?person dbp:occupation ?occupation . }

 filter (langMatches(lang(?name),"en"))
}
order by ?person  "
```

Fig. 2. SPARQL query.

TABLE I. PERSON CONCEPTS

Type of Person	Total
Politicians	933
Crickters	279
Singers	72

As a result, the total number of persons of each type of persons extracted is shown in Table 1 above.

D. Articles Corpus Collection

In this paper, a news corpus for the extraction of entity mentions from unstructured text has been built. The news articles are collected from the daily Dawn newspaper archive. The corpus collected is comprised of 11 categories, namely Pakistan, Sport, Entertainment, Blogs, Business, Magazine, Multimedia, Newspaper, World, Home & Others. A total of approximately 17030 articles, published over a period of 15 months between January 2015 and March 2016 have been collected. The number of articles collected in respect of each category is shown in Table 2. A web scrapper in Java for building news archive corpus is built. In this paper, for the extraction of entity mentions from unstructured news articles, three categories of news articles, i.e. Pakistan, Sport and Entertainment are processed.

TABLE II. ARTICLES CORPUS

	Category	No. of Articles
1	Pakistan	7993
2	Sport	1094
2	Entertainment	85
4	Blogs	422
5	Business	23
6	Magazine	435
7	Multimedia	34
8	Newspaper	4569
9	World	1794
10	Home	31
11	Others	550

E. News Articles Preprocessing

The articles are preprocessed for the entity mention extraction phase. An excerpt from a preprocessed article is shown in Fig. 3. Stop words are removed to decrease the noise of the common words appearing in the unstructured text. Moreover, any punctuation marks including apostrophe ('s), commas have been removed to facilitate the precise extraction of entity mentions from the articles. The output of preprocessed documents is temporarily stored before it is made ready for the named entity recognition in the subsequent phase. The preprocessing decreases the size of the text to the considerable limit and making the entity recognition phase considerably faster. The preprocessing is performed in KNIME.

```
"
2014 highs lows Pakistan hockey Sport 2014 highs
lows Pakistan hockey Umer Bin Ajmal easy blame
financial restraints poor performances shown hockey
field — factors legit — skill ability account.
"
```

Fig. 3. Preprocessed article excerpt.

F. Knowledge Base Concept Driven Name Entity Recognizer

NER is the task associated with identifying terms or phrases in the text that precisely represents names of entities such as persons, locations, organizations, etc. These terms or phrases are referred to as concepts. In this paper, a knowledge base centric DBpedia based ontological concepts driven named entity recognition approach specific to persons for identification of entity mentions in the news articles is used. A concept in wiki pages is referred to as resource and is accordingly classified as an ontological class or sub-class. The approach uses a concept representing class of primary attributes concept name and the associated resource URI in DBpedia i.e. <concept, DBpediaURI>. For example, a concept of a person with a concept name "Shaikh Rasheed Ahmad", classified as a class of agent, person, politician and Pakistani has "http://dbpedia.org/resource/Shaikh_Rasheed_Ahmad" DBped-ia resource URI. The underlying system uses a non-exact matching dictionary driven tagger and the text article as an input to associate the concepts with the term and phrases in the articles. Subsequently a bag of words list is created with named entities recognized as persons and others, followed by the filtering of entities named as person. The output generated is represented as entity mention, count of each entity mention and the DBpedia URI transformed into an equivalent wiki page URL by the mapper i.e. <person entity mention, count of person, Wiki URI>. The underlying system is implemented in KNIME.

G. Facts Extractor

To enhance a user's overall reading experience, not only the persons in the article have been identified and thereby linked with the relevant background wiki knowledge concepts but also the relevant facts of the identified concepts have been

presented from within the article in form of sentences and RDF triples. A sentence represents a collection of words or phrases, an unstructured in its form is easily comprehensible by a human. However, an equivalent representation of the same sentence in form of a structured representation commonly known as triple, consisting of three constituents i.e. subject, predicate and object is what a machine can comprehend for processing and querying over unstructured text. An example of a triple is shown in Fig. 4 below.

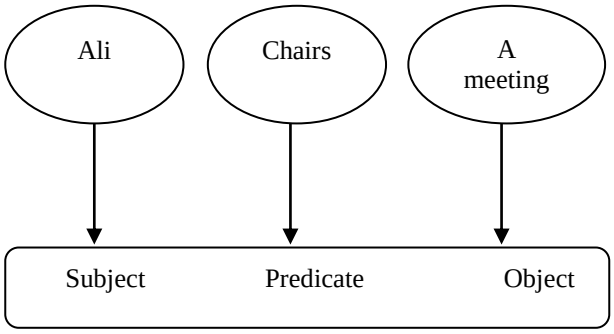

Fig. 4. An exmaple of a triple.

Given a set of extracted entity mentions of an article extracted in the previous step and associated article as input, the relevant sentences and their associated triples are extracted.

IV. Results

In the following section, the performance of experimental setup over the news article data set is presented and thereby the findings and the related measures are elaborated.

A. Experimental Setting

Based on the proposed NER technique above, the empirical evaluation and the relevant findings are presented in the following sections. The person concepts collected across three types from DBpedia is used to test how they map on to the terms or key phrases over three news article categories including Pakistan, Sport and Entertainment. The total instances of three person types include 933 politicians, 279 cricketers and 72 singers. In this paper, the primary setup and findings are based on extracting entity mentions from single articles, where in the detailed findings in terms of precision, recall and f-measure are presented. However, additionally the system is tested on multiple articles as a whole to find the resultant entity mentions in general. The system was built and tested in KNIME.

1) Dataset Description: The system is tested on three different set of news articles from Pakistan, Sport and Entertainment categories as shown in Fig. 5, published in daily dawn newspaper between January 2015 and March 2016.

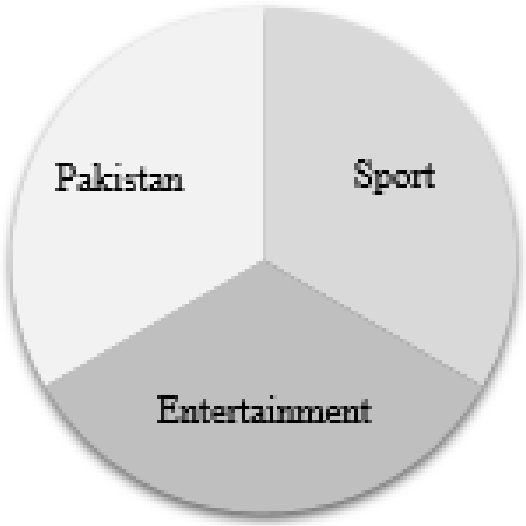

Fig. 5. News article categories.

B. Experimental Results

1) First Set of Articles: First experimental evaluation was based on extracting entity mentions from within Pakistan news articles.

TABLE III. Extraction of Politcian Mentions

ArticleID	Person	N	WikiURI
2015-01-01-6	"Abdul Rashid Godil"	1	"/Abdul_Rashid_Godil"
2015-01-01-15	"Benazir Bhutto"	1	"/Benazir_Bhutto"
	"Nawaz Sharif"	1	"/Nawaz_Sharif"
2015-01-02-17	"Asif Ali Zardari"	1	"/Asif_Ali_Zardari"
	"Chaudhry Aitzaz Ahsan"	1	"/Chaudhry_Aitzaz_Ahsan"
	"Mian Raza Rabbani"	1	"/Mian_Raza_Rabbani"
	"Nawaz Sharif"	1	"/Nawaz_Sharif"
	"Nisar Ali Khan"	1	"/Nisar_Ali_Khan"
	"Pervaiz Rashid"	1	"/Pervaiz_Rashid"
2015-01-04-8	"Faisal Raza Abidi"	1	"/Faisal_Raza_Abidi"
	"Shaikh Rasheed Ahmad"	1	"/Shaikh_Rasheed_Ahmad"
2015-01-07-50	N/A	-	-

For this purpose, system was run over 5 articles separately and extracted a total of 11 entity mentions. A total of 4 out of 5 articles resulted in extraction of entity mentions. The maximum number of entity mentions identified was 6 and the minimum number of entity mentions resulted was zero. The only duplicate entity mention across 5 different articles identified was "Nawaz Sharif". The resultant entity mentions are detailed in Table 3, where the ArticleID represents the date and the article number, person represents the entity mentions, N represents the number of entity mentions identified and WikiURL represents only the truncated part of complete wikiURL representing concept equivalent of entity mentions in text. For example, a complete wikiURL generated for an entity mentions "Abdul Rashid Godil" appeared in the actual output as "https://en.wikipedia.org/wiki/Abdul_Rashid_Godil".

2) On manual inspection of two such articles, performance measures precision, recall and f-measure were computed indicating, the faction of retrieved entity mentions relevant to concepts in wikipedia, the fraction of entity mentions successfully retrieved and the harmonic mean of precision and recall values respectively, as shown in Table 4 below.

The said approach does not result in precision less than 100%, reflecting that no irrelevant entity mentions are generated which are beyond the concepts predefined in wiki pages. However, the recall varies from 28.5% to 33%. This reflects that there are certain person entity mentions in the articles which are not extracted correctly.

On further manual inspection, it was identified from the contents of article "2015-01-01-6" that two false negatives were "Syed Khursheed Shah" and "Pervez Khattak". This is precisely because their names on wiki pages appeared with different spellings, i.e. "parvez khattak" and "Syed Khurshid Ahmed Shah" and moreover, the later name was not classified under Pakistani nationality or citizenship. The resultant average values all three measures, namely precision, recall and f-measure is plotted in graph shown in Fig. 6.

TABLE IV. EVALUATION RESULTS FOR POLITICIANS MENTIONS

ArticleID	Precision	Recall	F-measure
2015-01-01-6	100%	33%	50.00
2015-01-02-17	100%	28.5%	44.44

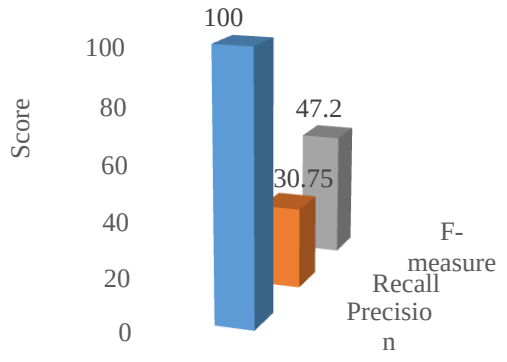

Fig. 6. Average politician mentions scores.

TABLE V. EXTRACTIONS OF CRICKETER MENTIONS

ArticleID	*Person*	*N*	*WikiURI*
2016-03-03-6	"Misbah-ul-Haq"	1	"/Misbah-ul-Haq"
	"Shahid Afridi"	1	"/Shahid_Afridi"
	"Younis Khan"	1	"/Younis_Khan"
2016-02-24-96	"Anwar Ali"	1	"/Anwar_Ali_(cricketer,_born_1987)"
	"Misbah-ul-Haq"	1	"/Misbah-ul-Haq"
	"Umar Akmal"	1	"/Umar_Akmal"
2015-01-02-17	"Abdul Qadir"	1	"/Abdul_Qadir_(cricketer)"
	"Haroon Rasheed"	1	"/Haroon_Rasheed"
	"Javed Miandad"	1	"/Javed_Miandad"
	"Sarfraz Nawaz"	1	"/Sarfraz_Nawaz"
	"Younis Khan"	1	"/Younis_Khan"

TABLE VI. EVALUATION RESULTS FOR CRICKETER MENTIONS

ArticleID	Precision	Recall	F-measure
2016-02-24-96	100%	60%	75.00

1) Second *Set of Articles*: The second set of results was experimentally evaluated over entity mentions representing Cricketers appearing in Sport news articles. This was tested over 3 articles, each run separately and thereby extracted 11 person entity mentions in total. All three articles resulted in extraction of entity mentions.

A maximum of 4 and a minimum of 3 persons were identified as Cricketers whose background Wikipedia concepts existed as in the form of structured DBpedia ontology. At least two persons appeared to be extracted twice from two different articles, namely, "Misbah-ul-Haq" and "Younis Khan". The outcome of the extraction of entity mentions from these articles in respect of 3 articles is presented in Table 5 above.

The precision and recall measures for Cricketers appearing in one of the news article undertaken on manual inspection are shown in Table 6 above.

2) Third Set of Articles: A third type of articles associated with entertainment category was processed and evaluated over 6 articles for entity mentions representing Singers in DBpedia. This resulted in extraction of a total of 6 person entity mentions. None of the articles was found to have returned zero results. One of the artists "Ali Zafar" appeared twice in results across two different articles. The results modeled after Table 3 (see Section IV(B-a)) are shown in Table 7 below.

Similarly, precision, recall and f-measure of one article 2015-01-28-29 from Table 7 was measured as 100%, 20% and 33.33, respectively. The measures computed in respect of all three categories news articles for Politicians, Cricketers and Singers is shown in Fig. 7 below.

TABLE VII. EXTRACTIONS OF SINGER MENTIONS

ArticleID	Person	N	WikiURI
2015-01-28-29	"Abida Parveen"	1	"/Abida_Parveen"
2015-02-23-14	"Ali Zafar"	1	"/Ali_Zafar"
2015-03-21-34	"Nusrat Fateh Ali Khan"	1	"/Nusrat_Fateh_Ali_Khan"
2015-01-26-17	"Waqar Ali"	1	"/Waqar_Ali"
015-01-31-18	"Sajjad Ali's"	1	"/Sajjad_Ali"
2015-08-22-25	"Ali Zafar"	1	"/Ali_Zafar"

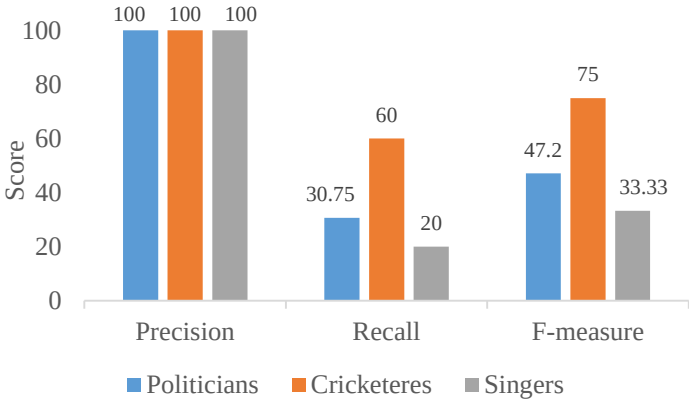

Fig. 7. Comparative Evaluation Results for All Three Entity Mentions.

3) Persons Extracted from Three Categories: The overall number of persons extracted over the entire test collection is measured for all three set of articles. A total of 4130 politicians were recognized from within Pakistan categorized news articles, of which 295 persons were unique. These extracted entity mentions represent 31.61% of 933 concepts collected from Wikipedia, which stands at approximately one third of the total number of politicians from Pakistan. However, this does not necessarily mean that the precision of the system is low, rather it just highlights that some of the politicians appearing in Wikipedia are not much referred or discussed in news articles. Similarly, 1790 cricketers, of which 114 unique were extracted from Sports articles, representing 40.8% of 279 concepts mapped onto entity mentions within news articles. For the third news article category entertainment, only 6 entity mentions mapped on to 72 concepts from Wikipedia.

4) Sentence Extraction: To make sense of the existing article in respect of the entity mentions extracted as persons, the relevant facts are extracted in form of sentences. This task is performed using Stanford NLP. A sample article along with input entity mention resulted in extraction of sentences, shown in Fig. 8.

5) Triples Extraction: Facts so extracted in form of sentences represents the knowledge about the entity in the article. Therefore, it is pertinent to keep track of the existing facts about such entities and convert them from unstructured sentence based representation to a more structured form as in RDF form. The sentences are converted in form of triples so as this may facilitate querying over the news articles for person entity mentions which are linked with concepts in wiki pages. This would help extract facts representing knowledge from within news articles which can be potentially used and compared with the knowledge extracted from linked wiki pages for different practical purposes. Fig. 9 lists the relevant triples generated in respect of an entity mention "Nawaz Sharif" from article "2015-01-02-17":

"

Prime Minister Nawaz Sharif on Friday is chairing a Multi-Party Conference (MPC) at the Prime Minister's House in Islamabad.

"

Fig. 8. Sentences extracted w.r.t a person entity mention.

"

<Prime_Minister_Nawaz_Sharif> <is_chairing/2015-01-01> <conference_at_Prime_Minister_'s_House>.

<Prime_Minister_Nawaz_Sharif> <is_chairing/2015-01-01> <multi-party_conference_at_Prime_Minister_'s_House_in_Islamabad.>.

"

Fig. 9. TriplesS extracted w.r.t a person entity mention.

V. CONCLUSIONS

A knowledge base concept driven named entity recognition information extraction technique for extraction and linking entity mentions to concepts was presented. The said technique was implemented in KNIME. The Wikipedia concepts representing three different set of persons from Pakistan was collected using existing DBpedia ontology classes through OpenLink Virtuoso SPARQL endpoint and tested the same over the Dawn news article corpus across three domain-specific news articles Pakistan, Sports and Entertainment. All in all the proposed technique resulted in 100% precision, that is, all entity mentions were correctly identified as persons however the recall varied from 20% to 60%, suggesting that some of the entity mentions were present in the articles however they could not be identified. Finally, information relevant to entity mentions was extracted and StanfordNLP was used for identifying sentences and their associated triples from unstructured news articles.

As part of future work, this work can potentially be improved to improve recall measure. Although StanfordNER was used for named entity recognition over 3, 4 and 7 class models, post-tagging, co-referencing and produced intermediary results which could be compared with the technique presented in paper. Therefore, the exhaustive comparison of all such results with the other techniques formulates the basis of a separate study wherein additional features can be taken into account to reach at conclusive comparison and establish advantages of the technique discussed in paper. Moreover, the persons identified with exact similar names belonging to two different or same disciplines must be disambiguated by taking into account the current classification of article and the associated facts cited in unstructured text. The n-gram based technique could be implemented to identify the entity mentions appearing with only first and last names in the article. The work is planned to

be extended to take into account the supervised techniques such as HMM, maximum entropy models, and training CRF based StanfordNER with concepts from knowledge bases for identification of persons and other type of entities such as people, places and organizations. For the purposes of ranking of the entity mentions, tf-idf could be used to identify the relevant candidate entities for linking with background information otherwise too many hyperlinks within text could potentially affect the overall reading experience.

REFERENCES

[1] L. Xiao, D. Wissmann, M. Brown, S. Jablonski. Information extraction from the Web: System and Techniques. Applied Intelligence, vol. 21, pages 195-224, 2004.

[2] O. Etzioni, M. Banko. Open information extraction from the web. In Communications of the ACM 51(12): 68-74, 2008.

[3] R. Feldman, Y. Aumann, , M. Finkelstein-Landau, E. Hurvitz, Y. Regev, A. Yaroshevich. A Comparative Study of Information Extraction Strategies . In Proceedings of CICLing, pages. 21–34, 2002, Mexico City, Mexico.

[4] C.H. Chang, M. Kayed, M.R. Girgis, K.F. Shaalan. A Survey of Web Information Extraction Systems. IEEE Transactions on Knowledge and Data Engineering 18(10), pages 1411–1428, 2006.

[5] J. Jiang. Information extraction from text. In Mining Text Data, pages 11–41, 2012.

[6] W.T. Balke. Introduction to information extraction: Basic notions and current trends. In Datenbank-Spektrum 12(2), 81–88 , 2012.

[7] T.R.Gruber.,A Translation Approach to Portable Ontologies, Knowledge Acquisition, 5(2):199-220, 1993.

[8] R. Anantharangachar, S. Ramani, S. Rajagopalan. Ontology Guided Information Extraction from Unstructured Text. Int. J. of Web & Sem. Tech. 4(1), 19–36, 2013.

[9] W.T. Adrian, N. Leone, M. Manna. Ontology-driven information extraction. arXiv preprint arXiv:1512.06034, 2015.

[10] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05, pages 363{370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[11] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In Proceedings of the 6th Workshop on Very Large Corpora, pages 152-160, 1998.

[12] T. Poibeau and L. Kosseim. Proper name extraction from non-journalistic texts. In Computational Linguistics in the Netherlands, pages 144–157, 2001.

[13] R. Prokofyev, G. Demartini, and P. Cudré-Mauroux. Effective named entity recognition for idiosyncratic web collections, Proceedings of the 23rd international conference on World wide web, 2014, Seoul, Korea.

[14] B. Settles. ABNER: An open source tool for automatically tagging genes, proteins, and other entity names in text. Bioinformatics, 21(14):3191-3192, 2005.

[15] Y. feng Lin, T. han Tsai, W. chi Chou, K. pin Wu, T. yi Sung, and W. lian Hsu. A maximum entropy approach to biomedical named entity recognition. In Proceedings of the 4th ACM SIGKDD Workshop on Data Mining in Bioinformatics, pages 56-61, 2004.

[16] G. Demartini , D. E. Difallah, P. Cudré-Mauroux. ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking, Proceedings of the 21st international conference on World Wide Web, April 16-20, 2012, Lyon, France.

[17] D. Milne , I. H. Witten. Learning to link with wikipedia, Proceedings of the 17th ACM conference on Information and knowledge management, October 26-30, 2008, Napa Valley, California, USA.

[18] R. Mihalcea, A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In Proceedings of the 16th ACM Conference on Information and Knowledge management (CIKM'07), pages 233-242, 2007, Lisbon, Portugal.

[19] E. Kuzey , G. Weikum. Extraction of temporal facts and events from Wikipedia, Proceedings of the 2nd Temporal Web Analytics Workshop, April 17-17, 2012, Lyon, France.