# Semantic Annotation of Bangla News Stream to Record History

**Conference Paper** · December 2015

DOI: 10.1109/ICCITechn.2015.7488135

CITATIONS
2

READS
200

**3 authors**, including:

Hanif Seddiqui
University of Chittagong
**38** PUBLICATIONS **340** CITATIONS
SEE PROFILE

Md. Hasan Hafizur Rahman
Comilla University
**11** PUBLICATIONS **26** CITATIONS
SEE PROFILE

# Semantic Annotation of Bangla News Stream to Record History

Md. Hanif Seddiqui
Dept. of Computer Science & Engineering
University of Chittagong
Chittagong - 4331, Bangladesh
Email: hanif@cu.ac.bd

Md. Nesarul Hoque
Dept. of Computer Science & Engineering
Port City International University
Chittagong, Bangladesh
Email: nesarul@portcity.edu.bd

Md. Hasan Hafizur Rahman
Dept. of Computer Science & Engineering
Comilla University
Comilla, Bangladesh
Email: hhr@cou.ac.bd

***Abstract*—Every day thousands of news articles are published in Bangla from several different sources on the web and this number is even increasing rapidly. On the contrary, the readers are often selective to read their desired news only. In this connection, classical Information Extraction $(IE)$ techniques are used to query with keywords from unstructured or semi-structured news contents to fulfill partial requirements. However, they cannot interpret sequences of events, relation among entities, inference some unveiled facts to facilitate further human analysis. To achieve this goal, semantic technology adds formal structure and semantics to the news stream. In this paper, we propose a system to analyze Bangla news content to annotate especially things, people and places with semantic technology automatically by extracting $what$ happened, $when$, $where$ and $who$ being involved in the news with the help of classical Natural Language Processing $(NLP)$ techniques. Furthermore, we relate news of today with the previous news to accumulate information over time. We present our proposed system of annotating Bangla news semantically and experiment with $SPARQL$ to inference integrated news from different sources over time and shows its effectiveness in querying specific information.**

## I. Introduction

Due to the proliferation of news articles all over the world, many readers like to view news from various news sources on a particular event to analyze their credibility. Credibility often depends on $where$, $when$, and $who$ being involved in a particular event. However, this is a formidable task in news processing due to a number of reasons including complexity of Natural Language Processing (NLP). Moreover, the exponential growth of the World Wide Web (WWW) with unstructured and semi-structured contents generally proliferates information jargon with noisy HTML tags, which have only effect on the style, not on the semantics. Therefore, semantic web has been coined by Sir Tim Berners Lee [1], the inventor of the WWW, to add formal structure and semantics (metadata and knowledge) to the web content for the purpose of more efficient management and access by means of ontology.

Ontology is defined as *"an explicit, formal specification of a shared conceptualization of a domain of interest"* [2]. The basic building block of semantic technology is called a *triple*: $\langle subject \rangle$ $\langle predicate \rangle$ $\langle object \rangle$. Although a triple is always simple with a little semantics, however, "A Little Semantics Goes a Long Way" [3]. Therefore, adding simple semantics in existing web pages by means of intrinsic or extrinsic data-store is still getting researchers attention. This process of adding semantics as metadata with the existing traditional WWW pages is often called as *semantic annotation*, a formal definition of the entities available in the text by means of external or background knowledge, which was proposed in [1]. The goal is to create annotations with well-defined semantics for the sake of interoperability and of converting web pages into "intelligent" documents. The benefit of semantic annotation is of two folds [4]: enhanced information retrieval and improved interoperability.

The semantic annotation often enables many new types of applications, such as highlighting, indexing and retrieval, categorization, smooth traversal between unstructured text and available relevant knowledge. Furthermore, information extraction and knowledge acquisition can be performed on the basis of extracting and analyzing of relationships among events, entities like $when$, $where$ and $who$ being involved at the events, situation descriptions, and so on.

Now-a-days there are a large number of Bangla news portals with millions of news already published. Extracting information regarding the basic questions of $what$ happened, $when$, $where$ and $who$ being involved is not a trivial task due to the unstructured HTML based content of the published news. To meet the challenge of extracting specific information from the Bangla news content, we need proper annotation to these pages. Therefore, our vision is to annotate the existing news pages automatically with basic building blocks of semantic technology, i.e. triples.

To envision semantic annotation (also called "semantic tagging" [5]) over large document collection of Bangla news contents, we use NLP modules ranging from cleaning, tokenization, lemmatization, part-of-speech tagging, named-entity recognition (NER), semantic role labeling, event and entity co-reference to factuality and opinions. We represent relations among events and persons involved, place and time as named graphs. These relations indicate *who* made what statement in *what* event, *when*, *where*, and *whom* about. We keep this background knowledge, which is external to the original sources of news content. This background knowledge is accumulated over time to record history. Furthermore, we extract and inference knowledge with a well-known SPARQL[1] [6], [7].

In this paper, we describe the first implementation of the

[1] http://www.w3.org/TR/rdf-sparql-query

978-1-4673-9930-2/15/$31.00 ©2015 IEEE

system and the primary results of annotating news stream along with basic SPARQL query. The rest of the paper is organized as follows. **Section II** focuses on the other existing related work of semantic annotation, while **Section III** articulates ontologies to be used in semantic annotation of news stream. **Section IV-A** describes different approaches of extracting news articles and cleaning the contents effectively for further processing by our proposed system of semantic annotation, while accumulation of background knowledge is articulated in **Section IV-B** along with some comprehensive examples. **Section V** includes experiments and evaluation to show the effectiveness of our proposed semantic annotation to record history of Bangla news stream. Concluding remark and some future directions of our work are described in **Section VI**.

## II. Related Work

The semantic annotation identifies people, organizations, and projects which are mentioned in a web news story, as well as including traditional metadata, such as the author's name and date of publication [8].The related work focuses on two relevant areas: semantic annotation techniques and news annotation.

### A. Semantic Annotation Techniques

In MIAKT (Medical Imaging and Advanced Knowledge Technologies) the annotations make the knowledge contained in unstructured sources (medical images such as x-rays) available in a structured form, allowing both accurate and focused retrieval and knowledge sharing for a given patient's case. Moreover, the annotation can be used to provide automated services [9].

The use of knowledge embodied in annotation is being investigated in domains as diverse as scientific knowledge [10], radio and television news [11], genomics [12], making web pages accessible to visually impaired people [13] and the description of cultural artifacts in museums [14].

There are two frameworks: Annotea and CREAM. The main format for Annotea is RDF and the kinds of documents that can be annotated are limited to HTML or XML-based documents[15], [16].

The CREAM framework [17] looks at the context in which annotation could be made and used as well as the format of the annotation themselves. CREAM have considered the possibility of annotating the deep web.

There are a number of automatic semantic annotation tools. Lixto is a web information extraction system which allows wrappers to be defined for converting unstructured resources into structured ones [18].

Armadillo is a system for unsupervised creation of knowledge bases from large repositories (e.g. the Web) as well as document annotation [19].

KnowItAll [20] automates extraction of large knowledge base of facts from the Web in a similar fashion to Armadillo. The most notable difference is the way the system assesses the plausibility of candidate extractions. This is done using the PMI (point-wise mutual information) measure rather than weighing multiple evidence from domain-specific oracles.

The SmartWeb project is also investigating unsupervised approaches for RDF knowledge base population [21]. Their approach resolves the issue of not having per-existing mark-up to learn from by using class and subclass names from the ontology to construct examples. The context of these examples is then learnt. In this way, instances can be identified which have similar contexts, but which may use different terminology to the ontology.

SemTag is another example of a tool which focuses only on automatic mark-up [22]. It is based on IBM's text analysis platform Seeker and uses similarity functions to recognize entities which occur in contexts similar to marked up examples.

KIM [23], [24], uses information extraction techniques to build a large knowledge base of annotations. The annotations in KIM are metadata in the form of named entities (people, places and so on) which are defined in the KIMO ontology and identified mainly from reference to extremely large gazetteers.

### B. News Annotation

Troncy et al. [25] focuses on the event linking of media directories like flickr, youtube and so on represented with the Media Ontology.

NewsReader [26], an European funded project, is closely relevant to our proposed system. It merges news of today with the previous news, creating a long-term history by means of a large knowledge graph. It is a standoff layered representation for the results of Natural Language Processing (NLP) modules ranging from tokenization, part-of-speech tagging, lemmatization, dependency parsing, named-entity recognition, semantic role labeling, event and entity co-reference to factuality and opinions.

Social media journalism [27] coined a framework for social media journalism of breaking news, a wealth of crowd-sourced data, in the form of text, video and image.

## III. News Ontology

For the efficient exchange of news, the International Press Telecommunication Council (IPTC) has developed the NewsML Architecture (NAR)[2], which provides the framework for the second generation of IPTC G2 standards. NAR is a generic model that defines four main objects such as $newsItem$, $packageItem$, $conceptItem$ and $knowledgeItem$ and the processing model associated with these structures in detail.

Although NAR architecture defines the basic concepts for representing the various news contents such as text, photo, audio, video, graphics and so on, a number of other standards are used in the media industry [28]. Photography capptured by journalist come with EXchangeable Image Format (EXIF)[3]. DIG35[4] is a specification of the International Imaging Association (I3A). It defines, within an XML Schema, metadata related to image parameters, information creation, content description (who, what, when and where), history and intellectual property

[2]http://www.iptc.org/NAR/

[3]http://www.digicamsoft.com/exif22/exif22/html/exif22_1.htm

[4]http://www.i3a.org/resources/dig35/

rights. Adobe's Extensible Metadata Platform (XMP)[5] describes a native RDF data model with the help of Dublin Core, basic rights and media management schemas for describing still images. PhotoRDF[6] focuses on the standardization of a set of categories for personal photo management using Dublin Core and a minimal RDF schema defining 10 terms for the dc:subject property. Moreover, Video can be decomposed and described using MPEG-7, the Multimedia Content Description ISO Standard [29], [30].

On the other hand, *rNews* is a data model for embedding machine readable publishing metadata in web documents. The IPTC recently adopted rNews 1.0, based somewhat on the NewsML-G2, *News Industry Text Format (NITF)* and hNews models but extending beyond those standards in certain ways. The rNews is designed in such a way that it aligns with the IPTC Photo Metadata standards. However, it has a very limited set of metadata for managing content unlike NITF. NITF uses the eXtensible Markup Language (XML) to define the content and structure of news articles.

Considering all standards for describing news contents or associated media, We have selected rNews[7] standard model of IPTC for their small, however, sufficient number of concepts and properties to fulfill our requirement for semantic annotation of Bangla news stream. Moreover, rNews is a semantically described ontology, which has a number of extra benefits over XML formated other standards. It defines 12 classes along with 72 properties in an RDF format. Among the 72 properties 57 are $DatatypeProperty$ (the range of the property is a data, not an object) and 15 are $ObjectProperty$ (the range of the property is another object). The model of rNews is articulated in the Fig. 1 with 12 distinct classes and a few properties.

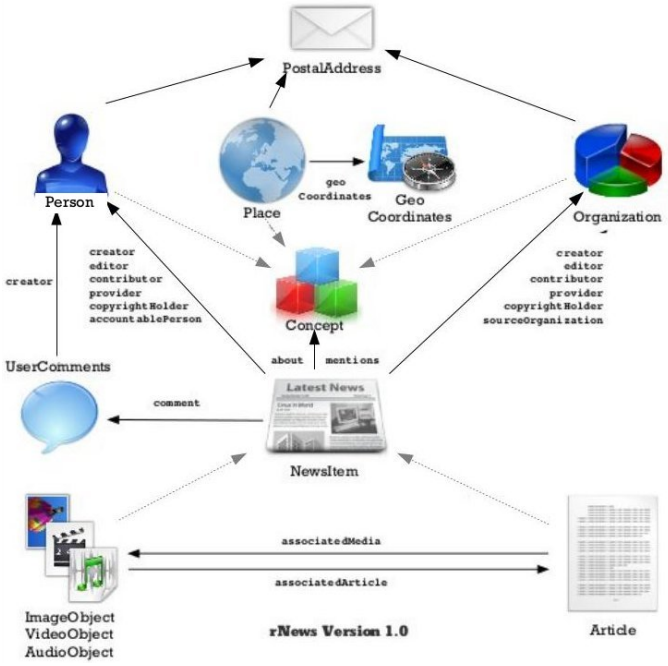


Fig. 1. The model of rNews

[5]http://www.adobe.com/products/xmp/
[6]http://www.w3.org/TR/photo-rdf/
[7]http://dev.iptc.org/rNews

The concept $NewsItem$ is defined with a number of properties, such as 'headline', 'description', 'datePublished', 'genre', 'alternativeHeadline', 'copyrightYear' and so on. $Article$ is a subclass of $NewsItem$. $Article$ is further defined with properties 'articleBody', 'wordCount', 'printPage', 'printColumn', 'printEdition', 'printSection' and so on. Moreover, the concept $Article$ has property relations with either $ImageObject$, $AudioObject$, $VideoObject$ or vice-versa. The property 'mentions'/'about' of $newsItem$ describes a $Person$, $Place$, or $Organization$. A $Place$ may be further associated with $GeoCoordinates$ concept. Either $Person$, $Place$, or $Organization$ may have 'address' as $PostalAddress$. Any $NewsItem$ may be 'comment'ed with $UserComment$ described by 'commentText', 'commentTime' and so on.

## IV. Overall System

The architecture of our proposed system, articulated in the Fig. 2, includes news content accumulation from accessible news sites on the web by introducing a crawler. It converts these crawled information into machine understandable representation by adding semantic structure. In our research we demonstrate a novel system that decomposes the whole process into two steps: pre-processing and main processing approach.

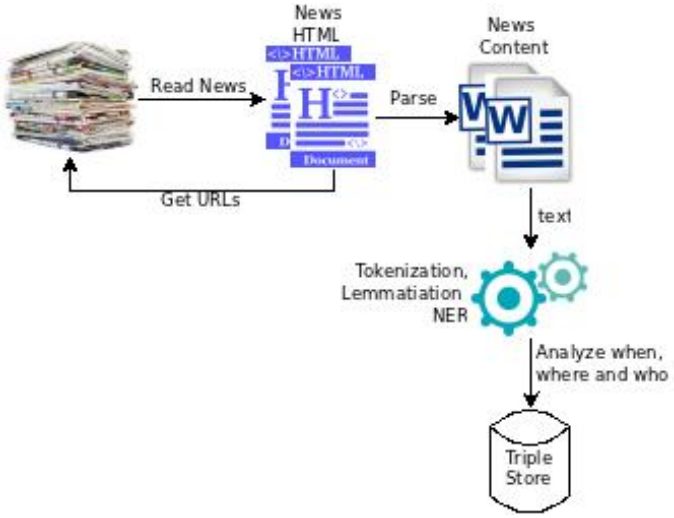


Fig. 2. Overall architecture of our system

### A. Preprocessing Approaches

Our preprocessing module comprises of three individual steps: crawling URLs, cleaning HTML text and parsing news content. Individual steps are described in the following subsections.

*1) Crawling URLs:* Our algorithm of finding news links uses Breadth-First-Search ($BFS$) to find necessary news articles efficiently in a news media as Depth-First-Search may sink into deeper HTML jargon. The algorithm is demonstrated in Fig. 3.

In the algorithm $crawler$ of Fig. 3, the queue, $q$ and a list, $visited$ are initialized with base url are demonstrated in line 1 and 2. Removing each url available from the front of a queue (as line 4), the text inside the url is processed by JSoup [31],

```
Algo. crawler(base_url)
     q:Queue
     visited:List

1.  enqueue base_url into q
2.  insert base_url into visited
3.  while(q is NOT empty)
4.    delete front_url from q
5.    process front_url to get html_text
6.    for each <a href='new_url'> ∈ html_text
7.      if(new_url not ∈ visited
           and value(new_url) ∋ bangla_text)
8.        enqueue new_url into q
9.        insert new_url into visited
10.       map new_url, bangla_text
11.     end if
12.   end for
13. end while
```

Fig. 3. Pseudo code of how our crawler works.

demonstrated at line 5, to retrieve further new urls. For each (as line 6) retrieved new url (as line 7), insert the url at the rare of the queue, $q$ and insert it into the list, $visited$ which is demonstrated at line 8 and 9 respectively. The process is continued until the queue, $q$ is empty (as line 3).

*2) Cleaning HTML:* An HTML based Web page typically contains many non-informative blocks called as noisy blocks such as navigation panels, copyright and privacy notices, advertisements and so on. The information contained in these noisy blocks can deteriorate the efficiency of information extraction. Therefore, we remove these noisy block to retrieve a clean Bangla news article.

*3) Parse News Contents:* After crawling and cleaning of HTML text to find Bangla blocks, we parse the blocks to identify news title, news publishing date, news reporter and the body of the news article using JSoup.

### B. Main Processing Approach

After preprocessing, our main process focuses on the analysis of the clean news content by a number of our own developed tools for Bangla processing. It includes tokenization, lemmatization and named entity recognition systems.

*1) Tokenization:* Tokenization is the process that aims to fracture the stream of characters into tokens delimited by white space, tab, new line and so on. As we are working on Bangla news document, there are lots of Bangla punctuations in the document. Moreover news document may contain Bangla as well as English digits. As meaningful Bangla words do not contain these characters we remove these. We also eliminate the single letter word from the document in this stage. Then the document becomes nothing but a bag-of-words. Bangla is a highly inflected language with relatively free or pragmatically free word i.e., Bangla (verb, noun, adjective) words are inflected from head words. We perform stemming on the tokenized words.

*2) Lemmatization:* As a part of extracting head word from a word variant, we use stemming. Stemming is an operation that splits a word into the constituent root part and affix without doing complete morphological analysis. Terms with common stems tend to have similar meaning, which makes stemming an attractive option in information retrieval applications. Another advantage of stemming is that it can drastically reduce the dictionary size used in various NLP applications, especially for highly inflected languages [32]. The last two decades has witnessed an immense escalation of Bangla web and digital text contents and is having an exponential growth rate. This has enhanced the need for the development of highly efficient IR systems and consequently good stemmers [33].

Moreover we create a light Bangla dictionary that help us stemming a word. This dictionary contains about 45000 words that frequently used in online Bangla news. It plays an important role while stemming a word.

*3) Named Entity Recognition (NER):* Named Entity Recognition for Bangla [34], [35], [36], [37] plays an important role in identifying answer of the questions *What*, *When*, *Where* and *Who* that mainly helps annotating news data for further information retrieval.

*What:* Our view is media centered data that includes images along with text data. Our system focuses on the title of news and caption of images to retrieve the answer to the question.

*When:* To annotate news content with our specific goal, we retrieve all available date/time data from the content. It helps identifying chronological event occurrence.

*Where:* Identifying the location centric view to answer this question is a tedious job. However, our geo-Bangladesh [38] dataset helps identifying most of the locations out of the news content.

*Who:* Bangla Named Entity Recognition (NER) system helps identifying people and organization names from Bangla news content.

## V. Experiment and Evaluation

Our system converts crawled information into machine understandable representation by adding semantic structure in $N-TRIPLE$ format. The fragment of semantic view of news information is portrayed in the Fig. 4 to comprehend the essence of semantic annotation in news stream.

To evaluate the effectiveness of our system, we performed a large number of experiments using semantic search engine, $SPARQL$ to retrieve specific information from the news stream. In this regard, the semantic query is given in the Fig. 5 that extracts the Uniform Resource Identifiers *(URI)* of news items along with related $URIs$ of primary entities of these news.

The result of this query $(Fig.\ 5)$ is portrayed in the Fig. 6.

In this stage, we can retrieve the mentioned people, places, organizations and so on of a news item by implementing these extracted $URIs$ from the query in Fig. 5. We demonstrated a semantic query in the Fig. 7 to extract URIs of people, places and organizations available in a specific news item.

<http://www.skeim.org/Person#100001> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://iptc.org/std/rNews/2011-10-07#Person>.
<http://www.skeim.org/Person#100001> <http://iptc.org/std/rNews/2011-10-07#name> "সাকমুদুল আলম@bn".
<http://www.prothom-alo.com/bangladesh/article/404230> <http://iptc.org/std/rNews/2011-10-07#about> <http://www.skeim.org/Person#100001>.
<http://www.skeim.org/Person#100002> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://iptc.org/std/rNews/2011-10-07#Person>.
<http://www.skeim.org/Person#100002> <http://iptc.org/std/rNews/2011-10-07#name> "মঞ্জুরুল আলম@bn".
<http://www.prothom-alo.com/bangladesh/article/404230> <http://iptc.org/std/rNews/2011-10-07#mentions> <http://www.skeim.org/Person#100002>.
<http://www.skeim.org/Person#101270> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://iptc.org/std/rNews/2011-10-07#Person>.
<http://www.skeim.org/Person#101270> <http://iptc.org/std/rNews/2011-10-07#name> "রাফিয়া হাসিনা@bn".
<http://www.prothom-alo.com/bangladesh/article/404230> <http://iptc.org/std/rNews/2011-10-07#mentions> <http://www.skeim.org/Person#101270>.
<http://www.skeim.org/Person#101271> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://iptc.org/std/rNews/2011-10-07#Person>.
<http://www.skeim.org/Person#101271> <http://iptc.org/std/rNews/2011-10-07#name> "জিসান@bn".
<http://www.prothom-alo.com/bangladesh/article/404230> <http://iptc.org/std/rNews/2011-10-07#mentions> <http://www.skeim.org/Person#101271>.
<http://www.skeim.org/Person#100032> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://iptc.org/std/rNews/2011-10-07#Person>.
<http://www.skeim.org/Person#100032> <http://iptc.org/std/rNews/2011-10-07#name> "শেখ হাসিনা@bn".

Fig. 4. Semantic view of bangla news stream

| ?newsURI | ?aboutURI |
|---|---|
| <http://www.prothom-alo.com/bangladesh/article/404230> | <http://www.skeim.org/Person#100001> |
| <http://www.bd-pratidin.com/entertainment/2014/05/06/4086> | <http://www.skeim.org/Person#101027> |
| <http://www.bd-pratidin.com/entertainment/2014/05/06/3984> | <http://www.skeim.org/Person#101032> |
| <http://dainikamadershomoy.com/2014/08/12/161415.html> | <http://www.skeim.org/Organization#10197> |
| <http://dainikamadershomoy.com/2014/08/13/161649.html> | <http://www.skeim.org/Organization#10135> |
| <http://www.banglanews24.com/beta/fullnews/bn/352237.html> | <http://www.skeim.org/Person#101297> |

Fig. 6. The result of the query that is given in Fig. 5

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX  iptc: <http://iptc.org/std/rNews/2011-10-07#>
PREFIX skeimPerson: <http://www.skeim.org/Person#>
PREFIX skeimPlace: <http://www.skeim.org/Place#>
PREFIX skeimOrganization: <http://www.skeim.org/Organization#>

SELECT  ?newsURI   ?aboutURI
 WHERE
      {
         ?newsURI     iptc:about    ?aboutURI.
         }
```

Fig. 5. SPARQL query to retrieve the $URIs$ of all news along with their related $URI$ of primary entities

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX  iptc: <http://iptc.org/std/rNews/2011-10-07#>
PREFIX skeimPerson: <http://www.skeim.org/Person#>
PREFIX skeimPlace: <http://www.skeim.org/Place#>
PREFIX skeimOrganization: <http://www.skeim.org/Organization#>

SELECT  ?entityURI
 WHERE
      {
          <http://www.prothom-alo.com/bangladesh/article/404230>
                          iptc:mentions    ?entityURI.
         }
```

Fig. 7. Query to retrieve URIs of people, places and organizations of a news item

The snippet of the feedback result from the search engine for the query $(Fig.\ 7)$ is portrayed in the Fig. 8.

These $URIs$ $(Fig.\ 8)$ are used to retrieve corresponding

| ?entityURI |
|---|
| <http://www.skeim.org/Person#100001> |
| <http://www.skeim.org/Person#100002> |
| <http://www.skeim.org/Person#101271> |
| <http://www.skeim.org/Place#10086> |
| <http://www.skeim.org/Place#10001> |
| <http://www.skeim.org/Place#10016> |
| <http://www.skeim.org/Organization#10001> |
| <http://www.skeim.org/Organization#10235> |
| <http://www.skeim.org/Organization#10236> |
| <http://www.skeim.org/Person#100032> |

Fig. 8. Query to retrieve URIs of people, places and organizations of a news item

name and type from news stream and demonstrated in the Fig. 9. The search result of this query $(Fig.\ 9)$ is articulated in the Fig. 10. Moreover, we implement a large number of experiments to extract important knowledge pieces from news stream along with our generic knowledge repository, $Geo-Bangladesh$ on the web to fetch geo-coordinates of a location of Bangladesh. In this connection , we depict the fragment of data structure of $Geo-Bangladesh$ in the Fig. 11 that serves spatial data of Bangladesh when initiate demands from our news items.

## VI. Conclusion

Our system extracts information from available news sites on the web and add semantic annotation after natural language processing. Then we performed a large number of experiments using semantic search engine to retrieve specific information of a news items along with the inferencing on related news follow-up from accessible news portals to analyze news credibility issues effectively. Moreover, we integrate compatible

| Subject | Predicate | Object |
|---|---|---|
| <http://www.skeim.org#20> | <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> | <http://www.skeim.org#division>. |
| <http://www.skeim.org#20> | < http://www.w3.org/2000/01/rdf-schema#label> | "Chittagong@en". |
| <http://www.skeim.org#20> | <http://www.w3.org/2000/01/rdf-schema#comment> | "Information about Chittagong division@en". |
| <http://www.skeim.org#20> | < http://www.w3.org/2003/01/geo/wgs84_pos#lat> | "22.330391 @en". |
| <http://www.skeim.org#20> | <http://www.w3.org/2003/01/geo/wgs84_pos#long> | "91.82518000000004@en". |
| <http://www.skeim.org#20> | < http://www.w3.org/2000/01/rdf-schema#partOf> | "http://www.skeim.org#Bangladesh". |
| <http://www.skeim.org#2019> | <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> | <http://www.skeim.org#district>. |
| <http://www.skeim.org#2019> | < http://www.w3.org/2000/01/rdf-schema#label> | "Comilla@en". |
| <http://www.skeim.org#2019> | <http://www.w3.org/2000/01/rdf-schema#comment> | "Information about Comilla district@en". |
| <http://www.skeim.org#2019> | < http://www.w3.org/2003/01/geo/wgs84_pos#lat> | "23.455959 @en". |
| <http://www.skeim.org#2019> | <http://www.w3.org/2003/01/geo/wgs84_pos#long> | "91.18203689999996@en". |
| <http://www.skeim.org#2019> | < http://www.w3.org/2000/01/rdf-schema#partOf> | "http://www.skeim.org#20". |

Fig. 11. Data Structure of our Knowledge-base, $Geo - Bangladesh$

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX  iptc: <http://iptc.org/std/rNews/2011-10-07#>
PREFIX skeimPerson: <http://www.skeim.org/Person#>
PREFIX skeimPlace: <http://www.skeim.org/Place#>
PREFIX skeimOrganization: <http://www.skeim.org/Organization#>

SELECT  ?name   ?type
 WHERE
      {
        <http://www.skeim.org/Person#100001>        iptc:name    ?name.
        <http://www.skeim.org/Person#100001>        rdf:type     ?type.
        <http://www.skeim.org/Place#10016>          iptc:name    ?name.
        < http://www.skeim.org/Place#10016>         rdf:type     ?type.
       <http://www.skeim.org/Organization#10236>    iptc:name    ?name.
       <http://www.skeim.org/Organization#10236>    rdf:type     ?type.
         }
```

Fig. 9. Semantic Query to retrieve name and type of each URIs

| ?entityURI | ?name | ?type |
|---|---|---|
| <http://www.skeim.org /Person#100001> | মাকসুদুল আলম@bn | <http://iptc.org/std/rNews/ 2011-10-07#Person> |
| <http://www.skeim.org /Place#10016> | বাংলাদেশ@bn | <http://iptc.org/std/rNews/ 2011-10-07#Place> |
| <http://www.skeim.org /Organization#10236> | বাংলাদেশ পাট গবেষণা ইনস্টিটিউট@bn | <http://iptc.org/std/rNews/ 2011-10-07#Organization> |

Fig. 10. Query Result that is given in Fig. 9

information from diverse knowledge repositories to address the semantic interoperability issues in a faster and efficient way. The semantic information integration reflects its usability and effectiveness through the adequate qualitative results in terms of concepts and news entities. In future, we map our contents based on available places in a news item to increase search ability and evaluate our system based on quantitative analysis.

## REFERENCES


[1] T. Berners-Lee, J. Hendler, O. Lassila *et al.*, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 28–37, 2001.

[2] R. Studer, V. Benjamins, and D. Fensel, "Knowledge Engineering: Principles and Methods," *Journal of Data & Knowledge Engineering*, vol. 25, no. 1-2, pp. 161–197, 1998.

[3] J. Hendler, "The dark side of the semantic web," *IEEE Intelligent Systems*, vol. 22, no. 1, pp. 2–4, 2007.

[4] V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, and F. Ciravegna, "Semantic annotation for knowledge management: Requirements and a survey of the state of the art," *Web Semantics: science, services and agents on the World Wide Web*, vol. 4, no. 1, pp. 14–28, 2006.

[5] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. A. Tomlin, and J. Y. Zien, "Semtag and seeker: Bootstrapping the semantic web via automated semantic annotation," in *Proceedings of the 12th International Conference on World Wide Web (WWW03). Budapest, Hungary.*, 2003.

[6] L. Dodds, "Introducing sparql: Querying the semantic web," *Retrieved on April*, vol. 20, p. 2006, 2005.

[7] P. McCarthy, "Search rdf data with sparql: Sparql and the jena toolkit open up the semantic web, in developerworks," 2005.

[8] C. Welty and N. Ide, "Using the right tools: enhancing retrieval from marked-up documents," *Computers and the Humanities*, vol. 33, no. 1-2, pp. 59–84, 1999.

[9] K. Bontcheva and Y. Wilks, "Automatic report generation from ontologies: the miakt approach," in *Natural Language Processing and Information Systems*. Springer, 2004, pp. 324–335.

[10] N. S. Friedland, P. G. Allen, G. Matthews, M. Witbrock, D. Baxter, J. Curtis, B. Shepard, P. Miraglia, J. Angele, S. Staab *et al.*, "Project halo: Towards a digital aristotle," *AI magazine*, vol. 25, no. 4, p. 29, 2004.

[11] M. Dowman, V. Tablan, H. Cunningham, and B. Popov, "Web-assisted annotation, semantic indexing and search of television and radio news," in *Proceedings of the 14th international conference on World Wide Web*. ACM, 2005, pp. 225–234.

[12] F. Rinaldi, G. Schneider, K. Kaljurand, J. Dowdall, C. Andronis, A. Persidis, and O. Konstanti, "Mining relations in the genia corpus," in *Proceedings of the Second European Workshop on Data Mining and Text Mining for Bioinformatics*, 2004, pp. 61–68.

[13] P. Plessers, S. Casteleyn, Y. Yesilada, O. De Troyer, R. Stevens, S. Harper, and C. Goble, "Accessibility: a web engineering approach," in *Proceedings of the 14th international conference on World Wide Web*. ACM, 2005, pp. 353–362.

[14] R. Schroeter, J. Hunter, and D. Kosovic, "Vannotea: A collaborative video indexing, annotation and discussion system for broadband networks," in *Knowledge capture*. ACM Press (Association for Computing Machinery), 2003, pp. 1–8.

[15] J. Kahan, M.-R. Koivunen, E. Prud'Hommeaux, and R. R. Swick, "Annotea: an open rdf infrastructure for shared web annotations," *Computer Networks*, vol. 39, no. 5, pp. 589–608, 2002.

[16] M.-R. Koivunen, "Annotea and semantic web supported collaboration," in *Invited talk at Workshop on User Aspects of the Semantic Web (UserSWeb) at European Semantic Web Conference*, 2005, pp. 5–16.

[17] S. Handschuh and S. Staab, "Cream: Creating metadata for the semantic web," *Computer Networks*, vol. 42, no. 5, pp. 579–598, 2003.

[18] R. Baumgartner, S. Flesca, and G. Gottlob, "Visual web information extraction with lixto," in *VLDB*, vol. 1, 2001, pp. 119–128.

[19] F. Ciravegna, S. Chapman, A. Dingli, and Y. Wilks, "Learning to harvest information for the semantic web," in *The Semantic Web: Research and Applications*. Springer, 2004, pp. 312–326.

[20] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, "Unsupervised named-entity extraction from the web: An experimental study," *Artificial Intelligence*, vol. 165, no. 1, pp. 91–134, 2005.

[21] P. Buitelaar and S. Ramaka, "Unsupervised ontology-based semantic tagging for knowledge markup," in *Workshop on learning in web search at 22nd international conference on machine learning, ICML*, vol. 5, 2005, pp. 26–32.

[22] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, K. S. McCurley, S. Rajagopalan, A. Tomkins *et al.*, "A case for automated large-scale semantic annotation," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 1, no. 1, pp. 115–132, 2003.

[23] B. Popov, A. Kiryakov, D. Ognyanoff, D. Manov, A. Kirilov, and M. Goranov, "Towards semantic web information extraction," in *Human Language Technologies Workshop at the 2nd International Semantic Web Conference (ISWC2003)*, vol. 20, 2003.

[24] B. Popov, A. Kiryakov, D. Ognyanoff, D. Manov, and A. Kirilov, "Kim-a semantic platform for information extraction and retrieval," *Natural language engineering*, vol. 10, no. 3-4, pp. 375–392, 2004.

[25] R. Troncy, B. Malocha, and A. T. Fialho, "Linking events with media," in *Proceedings of the 6th International Conference on Semantic Systems*. ACM, 2010, p. 42.

[26] P. Vossen, G. Rigau, L. Serafini, P. Stouten, F. Irving, and W. Van Hage, "Newsreader: recording history from daily news streams," in *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014), Reykjavik, Iceland, May*, 2014, pp. 26–31.

[27] B. R. Heravi and J. McGinnis, "A framework for social semantic journalism," in *First International IFIP Working Conference on Value-Driven Social & Semantic Collective Intelligence (VaSCo), at ACM Web Science*, 2013.

[28] M. Hausenblas, S. Boll, T. Bürger, O. Celma, C. Halaschek-Wiener, E. Mannens, and R. Troncy, "Multimedia vocabularies on the semantic web," *W3C Multimedia Semantics Incubator Group Report*, 2007.

[29] P. Salembier and O. Avaro, "Mpeg-7: Multimedia content description interface," in *Workshop on MPEG*, vol. 21, 2000, pp. 20–21.

[30] H. Agius, "Mpeg-7: Multimedia content description interface," *Encyclopedia of Multimedia*, pp. 475–483, 2008.

[31] J. Hedley, "jsoup: Java html parser," 2010. [Online]. Available: http://jsoup.org/

[32] M. Z. Islam, M. N. Uddin, M. Khan *et al.*, "A light weight stemmer for bengali and its use in spelling checker," 2007.

[33] S.-B. Cho and J.-H. Lee, "Learning neural network ensemble for practical text classification," in *Intelligent Data Engineering and Automated Learning*. Springer, 2003, pp. 1032–1036.

[34] S. Banerjee, S. K. Naskar, and S. Bandyopadhyay, "Bengali named entity recognition using margin infused relaxed algorithm," in *Text, Speech and Dialogue*. Springer, 2014, pp. 125–132.

[35] S. Morwal, N. Jahan, and D. Chopra, "Named entity recognition using hidden markov model (hmm)," *Int. J. Nat. Lang. Comput.(IJNLC)*, vol. 1, no. 4, pp. 15–23, 2012.

[36] P. Sharma, U. Sharma, and J. Kalita, "Named entity recognition: A survey for the indian languages," *Parsing in Indian Languages*, p. 35, 2011.

[37] A. Ekbal and S. Bandyopadhyay, "Named entity recognition using support vector machine: A language independent approach," *International Journal of Electrical and Electronics Engineering*, vol. 4, no. 2, pp. 155–170, 2010.

[38] M. H. H. Rahman, S. Chakraborty, and M. H. Seddiqui, "Machine understandable information representation of geographic related data to the administrative structure of bangladesh," in *16th International Conference on Computer and Information Technology (ICCIT), 8-10 March 2014, Khulna, Bangladesh*, 2014, pp. 236–241.