WikiTimes’s Knowledge Extraction and Enrichment Process
Technical Paper

Giang Tran, Mohammad Alrifai, Tu Ngoc Nguyen and Wolfgang Nejdl
L3S Research Center & Leibniz University Hannover
{gtran, alrifai, tunguyen, nejdl}@L3S.DE

ABSTRACT
The Wikipedia Current Events portal is a special part of Wikipedia that focuses on news events. Every day, a new page is added to this portal and news events of the day are collaboratively summarized and listed on the page. This is a very interesting part of Wikipedia that however has not been exploited so far by existing knowledge bases.

In this technical paper, we detail the automatic knowledge extraction and enrichment process on which our knowledge base WikiTimes is built. WikiTimes information derived through this process is very high quality, with an accuracy of about 97-100%. In addition, we describe heuristics for enriching extracted events with complementary information that is not directly provided in the portal such as location and categorization. Using these heuristics, we are able to provide enriched information with an accuracy of about 93-96% for location extraction, 91-94% for categorization and up to 98% for related story.

Keywords
Events, News, Wikipedia, Information Extraction, Temporal Index

1. INTRODUCTION
Temporal summarization is very beneficial for users seeking information on developing news stories, especially with the tremendous amount of news articles being published everyday.

There has been a rich body of research in multi-document summarization and event detection [10, 7, 6, 1, 2, 13, 4] to enable the automatic generation of news timelines. Although it has been demonstrated with quantitative evaluations that these automatic methods can achieve reasonable results, conducting qualitative evaluations and optimizations of the produced summaries remains a challenge.

Wikipedia’s Current Events portal (WCE) is a special part of Wikipedia that focuses on daily summaries of news events. Pages are created in the WCE portal for each day and short summaries on news events are inserted on a regular basis. Figure 1, for example, shows some events from the page of the 18th of September 2014 (http://en.wikipedia.org/wiki/Portal:Current_events/2014_September_18). In this example, updates on three developing stories (Islamic State of Iraq, Russian military intervention and Ebola virus epidemic in West Africa) are summarized and external links to news articles are provided.

On one hand, this linking between the daily updates and the news stories is very useful because it provides the context of the news update. On the other hand, it is currently not possible to retrieve the entire timeline of a news story due to the lack of an index of these daily updates. We found that some of the news stories of WCE exist already in DBpedia and YAGO, but the timeline summaries of those stories do not exist. Similarly, it is not possible to retrieve the entire timeline of a certain entity, although this information exists in the portal. The lack of a structured access to WCE, limits the extent to which its content is being exploited. To the best of our knowledge, the content of the WCE portal was not yet exploited by existing knowledge bases such as DBpedia and YAGO [9]. Our knowledge base WikiTimes bridges this gap by indexing the news events and linking them to their news stories and entities, thus opening up new opportunities for exploiting this data in many meaningful ways.

This technical paper presents our engineering efforts on extraction and enrichment process of WikiTimes system.

2. WIKITIMES SYSTEM
This section describes the key concepts and relations in the WikiTimes event representation model and gives a high overview of its extraction and storage system.
2.1 WikiTimes Concept

Our aim is to index all relevant information about news stories and their events. More specifically, we are interested in answering the **When**, **Where**, **Who** and **What** questions. While some of these questions are answered by existing knowledge bases, the **What happened** question remains the key open question that we are targeting in this work. WikiTimes answers this question by extracting the textual summary of news events from WCE.

Figure 2 gives an overview of the key concepts and their relations in the WikiTimes model, which we describe in more details in the rest of this section.

Following Linked Open Data practices, we use the base URL http://wikitimes.13s.de/rdf/resource/(wikitimes for short) for dereferencing all objects. Furthermore, we re-use open and widely used vocabularies from the semantic web such as Dublin Core, RDF, RDF Schema and OWL. We also map our model to the Event model of Schema.org and LODE as well as to the YAGO model, when applicable, for interoperability purposes.

2.1.1 Concepts

The WikiTimes knowledge base is centered around news events. We distinguish here between two concepts: events and stories.

**News Event**: A news event in WikiTimes is something that happened in the real world on a specific single day and was reported in the news. The restriction to single-day events is due to the fact that we extract daily events from WCE portal. Events can be part of a developing story (see the definition of news stories below). An event is characterized by a date and a short textual, manually created description, i.e. the summary. Events have also links to Wikipedia entities such as persons, locations or news stories. However, events themselves are not Wikipedia entities, and thus, cannot be found in existing knowledge bases such as DBpedia and YAGO. Events in WikiTimes are represented by the class `wikitimes:NewsEvent`, which is the same type as http://schema.org/Event.

Every `NewsEvent` object is identified by a unique URI in the form: `wikitimes:NewsEvent/ID`. We write:

```
s rdf:type wikitimes:NewsEvent
a rdf:type schema:Event
a dc:identifier wikitimes:NewsStory/ID
```

If a match is found, we link stories to DBpedia and YAGO. For example, to link story 123 to the entity Iraq_War we write:

```
wikitimes:NewsStory/123 owl:sameAs dbpedia:Iraq_War
wikitimes:NewsStory/123 owl:sameAs yago:Iraq_War
```

**Entity**: In addition to news stories and their events, another important concept in the WikiTimes model is the concept of a named entity. An entity could be a person, organization, location etc., simply any real-world entity that is described by a Wikipedia article. In WikiTimes we consider only entities that are linked to news events. Similar to news stories, many of WikiTimes entities exist in DBpedia and YAGO as well. The ordered list of all events that link to a specific entity forms the timeline of this entity. Entities are represented in WikiTimes by the class `wikitimes:Entity`, and each entity is identified by a URI in the form `wikitimes:Entity/ID`, where ID is an internal unique identifier. We write:

```
et rdf:type wikitimes:Entity
et owl:sameAs dbpedia:Berlin
et owl:sameAs yago:Berlin
```

2.1.2 Relations

The WikiTimes model captures several properties of the event, story and entity objects with respect to the **When**, **Where**, **Who** and **What** questions as illustrated in Figure 2. The values of the properties can be datatyped literals, URIs or other objects.

**Event relations**: Every `Event` object in WikiTimes has at least two relations: `happenedOnDate` and `summary`.

The `happenedOnDate` relation specifies the date of the event (when question) in the standard format YYYY-MM-DD. We map this relation to the relation `yago:happenedOnDate` in the yago model. It also corresponds to the relation `lode:atTime` in the LODE model. The Schema.org model has two different relations for dates: `startDate` and `endDate`. Since our events are associated to exactly one day, we map `happenedOnDate` relation to both `startDate` and `endDate`.

The `summary` relation of an event captures the textual summary that is extracted from WCE. This is the answer to the what question. We map this relation to the corresponding relation in Schema.org model, i.e. `sc:description`. However, there is no similar relation in the YAGO model.

The `involvesEntity` relation links an event to the entities that are mentioned in the textual summary (i.e. the who question). This corresponds to the relations `yago:linksTo`, `schema:attendee`, and `lode:involved` in the YAGO, Schema.org and LODE model, respectively.

The `happenedIn` relation specifies the location in which the event occurred (i.e. the where question). It corresponds to the relations `yago:location`, `schema:attendee`, and `lode:atPlace` in the YAGO, Schema.org and LODE model, respectively. To story, it indicates related location where the events occurred.

The `category` relation assigns an event into a news category (e.g. politics, armed conflicts, economics, sports etc.). Events in the

---

3 dc: http://dublincore.org/documents/2012/06/14/dcmi-terms/
4 rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
5 rdfs: http://www.w3.org/2000/01/rdf-schema#
6 owl: http://www.w3.org/2002/07/owl#
7 schema: http://schema.org/Event
8 lode: http://linkedevents.org/ontology/
9 http://dbpedia.org/resource/
10 http://yago-knowledge.org/resource/
WCE portal are typically tagged with some categorization information. This is, however, not based on a well-defined taxonomy. Tags are given in as a free text, which leads to a lot of duplicates and misspellings. In Section 3 we describe how we deal with this issue.

The mentionedIn relation links an event to one or more external web resource describing the event (e.g. news articles). It provides additional context and background to the event. The value of the relation is a wikitimes:Source object, which consists of a type (e.g. news article), a source name (e.g. BBC, or Reuters) and a URL to the online resource.

Furthermore, event objects can have a link to news stories. This is captured by the belongsToStory relation, which corresponds to the Schema.org relation schema:subEvent, which stands between two schema:Event objects. For example, to link the event Entity/51039 that is shown on Figure ?? to the story “2014 pro-Russian unrest in Ukraine”, which has the id NewsStory/1053, we write:

wikitimes:Event/51039 belongsToStory
wikitimes:NewsStory/1053

wikitimes:Event/51039 schema:superEvent
wikitimes:NewsStory/1053

Story relations

As mentioned earlier, news stories are a special type of entities. Therefore, each news story in WikiTimes has a Wikipedia page, which is specified by the relation wikitimes:Source, which consists of a type (e.g. news article), a source name (e.g. BBC, or Reuters) and a URL to the online resource.

Furthermore, event objects can have a link to news stories. This is captured by the belongsToStory relation, which corresponds to the Schema.org relation schema:subEvent, which stands between two schema:Event objects. For example, to link the event Entity/51039 that is shown on Figure ?? to the story “2014 pro-Russian unrest in Ukraine”, which has the id NewsStory/1053, we write:

wikitimes:Event/51039 belongsToStory
wikitimes:NewsStory/1053
wikitimes:Event/51039 schema:superEvent
wikitimes:NewsStory/1053

Each news story has a start and end date, which are the dates of the first and last events in its chronologically ordered timeline. We use schema:startDate and schema:endDate relations to capture these dates.

Similar to events, the news category of a story is specified by the relation category, and the story location by the relation happenedIn, which corresponds to schema:atPlace and yago:happenedIn relations.

Figure 3 gives a high level overview of the WikiTimes system and its components, which we briefly describe in this section.

2.2 Extraction, Indexing and Synchronization

We developed tools for fetching, parsing and extracting event information from the WCE pages of every month. Section 3 describes the extraction process in more details. The extracted information is then stored in a MySQL database. We created mappings of the database schema to an RDF schema and deployed D2R server on top of the MySQL database thus providing an RDF knowledge base with SPARQL endpoint.

Since news events summaries are added to WCE portal on a daily basis, WikiTimes has been setup to extract the information directly from the online Wikipedia on a regular (weekly) basis. A previous analysis of the revisions in the WCE pages revealed that most of the revisions to the daily pages are made within 5 days from creation time [11]. Based on this observation, we limit the backward synchronization of events to 1 week after the creation date.

2.3 Integration with external Knowledge bases

WikiTimes content is easily linked to existing Wikipedia-based knowledge bases such as DBpedia and YAGO. This is achieved by mapping all entities from WikiTimes (including news stories) to entities in these knowledge bases via the Wikipedia page URL. Moreover, when a match is found, we retrieve some of the metadata, which is not provided in WCE from YAGO into WikiTimes. For example, entity type information (e.g. person, location, organization etc) and story location information are not provided in WCE. We found in many cases that these information can be imported from DBpedia and/or YAGO.

2.4 Enrichment

We observed that for many events some information are either entirely (e.g. location) or partially (e.g. categories) missing in WCE. We developed some heuristic solutions for inferring this information based on the available information that has been extracted from WCE or imported from DBpedia and YAGO.

2.5 Interfaces to WikiTimes

The created knowledge base is publicly available at the URL http://wikitimes.l3s.de. We provide different ways for accessing the data. A SPARQL endpoint and RDF browsers can be found at http://wikitimes.l3s.de/rdf/. In addition, we also developed a RESTful endpoint for remotely querying and
downloading the WikiTimes data over the web in JSON or XML format.

Furthermore, we a graphical user interface with browsing and full-text searching functionalities is provided, including the ability to retrieve the entire timeline of a specific news story or a specific event.

3. KNOWLEDGE EXTRACTION

This section describes the extraction process of news events from the WCE portal. First, Section 3.1 briefly describes the structure of WCE pages and the challenges of parsing them. Next, Section 3.2 describes in more detail the extraction steps and the quality evaluation of their outcome.

3.1 The Source: WCE

WikiTimes’ main source of information is the Wikipedia Current Events portal (WCE), which is a special part of Wikipedia focusing on summaries of news events. It includes summaries of news events from the past 15 years.

3.1.1 WCE structure

The structure of the portal pages is rather simple. Users insert new events on a daily basis to the main page of the portal. Events of the day are inserted as a bullet list and described using a short text as shown in Figure 1 with links to external citations such as online news articles about the event. Entity mentions are linked to the respective Wikipedia article. Each event is typically assigned to a specific category (e.g. armed conflicts and attacks, Arts and culture, Business and economy etc.).

If a news event belongs to an ongoing story that has a Wikipedia article, a link is typically provided. In this case, a two level bullet list is used, where the first level is used for the link to the news story and the second level is used for the updates of the day (as shown in Figure 1).

Different than typical Wikipedia articles, this portal deploys an automatic mechanism for archiving the events. Starting from the 1st of January 2005, all events of the ending day are automatically archived into one Wikipedia page with a unique identifier of the form: Portal:Current_events/YYYY-month_dd (e.g. http://en.wikipedia.org/wiki/Portal:Current_events/2014_May_21). Before 2005, all events of the month were archived into one Wikipedia page with a unique identifier of the form: month/YYYY (e.g. http://en.wikipedia.org/wiki/May_2004).

3.1.2 Challenges

Our goal is to build a knowledge base with a high quality. Therefore, we had to deal with the lack of well-defined structure in the pages of the WCE portal and other editing errors introduced by the users. Some examples of the issues we had to deal with include the following:

- we observed some changes in the writing and formatting style and in the HTML structure of the monthly pages over the years (such as the date format or the introduction of categories)
- a lot of automatic cleaning was required due to the frequent occurrence of duplications and misspellings in the category and story labels rendering them useless for event indexing
- in many cases, editors use a link to a named entity (such as a country) in the place of a news story link
- over the years, links to the same entities have changed due to the merging or splitting of some Wikipedia pages. This is especially the case for Wikipedia pages of many news stories. It is quite common that several pages for the same stories are created in parallel, which are then merged at a later point of time. In other cases, when some pages get too long, Wikipedia editors tend to restructure them by splitting them into several pages.
- while some of the issues with the links to Wikipedia pages can be solved by following the redirection tags of the Wikipedia pages, we observe that in many cases, no such tags exist.

3.2 Extraction Process

We wanted to build a system that continuously and regularly collects events from WCE portal independent on the availability of Wikipedia dumps. Therefore, we extract the information from online WCE pages. As mentioned in Section 3.1, there has been two different modes (i.e. daily and monthly) of archiving the events in WCE over the past years. However, we observe that the content of the daily pages is exactly matching the content in the monthly pages. Therefore, we fetch the page of each month (starting from January 2000 and up to date) following the http://en.wikipedia.org/wiki/month_YYYY pattern (e.g. http://en.wikipedia.org/wiki/May_2004).

We developed tools for parsing the downloaded pages, extracting events information, and storing them in a MySQL database. As mentioned earlier, we observed that the format and HTML structure of the pages changed over time. Therefore, we developed specific parsers for specific periods accordingly. In fact we created 6 different parsers for the different time periods. However, it is worth mentioning that we observed a stability in the format and structure of WCE pages since May 2006.

Then, we analyzed the structure and format of these monthly pages to identify certain patterns. Based on the recognized patterns, we applied some rules for parsing the fetched pages in order to extract events and their metadata. Several scripts have been applied for different periods as the structure of the WCE changed several times over the past years. The core of the applied scripts remains the same: we firstly identify the beginning of the unordered bullet list of the events (<ul>) by parsing the body of the HTML code. Each list item (<li> ... </li>) is considered an event, unless it is an anchor text (between <a> and </a> elements) and has a sub-list; in which case it is considered a news story.

3.2.1 Extraction of news stories

As mentioned earlier, by news stories we refer to a special type of Wikipedia articles that are about news stories and are linked to one or more daily events in WCE. We identify news stories in WCE by following the following simple pattern: links to news stories are usually one level above the event description in the unordered item list of a day (see the examples in Figure 1). We extract the name of the story and the Wikipedia URL. Sometimes users provide different names for the same story and in some cases the names are misspelled or shortened. Therefore, we use the Wikipedia URL of the story to detect duplications. We also extract the name of the story from the URL instead of using the names provided by the users.

All events that are sub-listed underneath a story label are automatically linked to this story. This adds a fact of belongsToStory relation to each of those events and the corresponding subEvent relation to the news story object.

3.2.2 Extraction of entities

The links from events to Wikipedia entities are also extracted from the event description (before the removal of HTML tags).
Each entity is then linked to DBpedia and YAGO using the Wikipedia URL of the entity. Once a match is found, additional metadata about the entities such as the type of the entity (i.e., person, location, organization etc) is imported into Wikitimes.

An involvesEntity relation is then added from the event that mentions those entities to each of them.

3.2.3 Extraction of categories

Each list of events is typically preceded by the category name. We extract these categories when they exist. We then add a category relation from each of the events listed underneath the category label to this category. However, we observed some inconsistencies in the names of the categories. For instance, users sometimes use labels such as Armed conflict, Armed conflicts, Armed conflicts and attacks, which obviously refer to the same category. Sometimes, the names of the categories are also misspelled. We use some heuristics to detect such duplications and inconsistencies including stemming and tokenization techniques. Moreover, we also map the WCE categories to Open Calais topical categories\(^1\), which is recently widely used as a standard taxonomy (e.g.,[3]). The mapping process is done manually as the number of categories are not large and in particular we observed clear alignment between OpenCalais categories and what Wikipedia users described in WCE, for example, “(WCE) War Conflict and Attacks” \(\rightarrow\) “(OpenCalais) War_Conflict”, “(WCE) Politics” \(\rightarrow\) “(OpenCalais) Politics”, etc. The accuracy of this method is manually evaluated as described in Section sec:quality.

3.2.4 Resolving redirections

Redirections in WCE happen actually derived from the redirections occur in Wikipedia, when two story pages, after some time since their creation date, are considered to be merged into one. This problem is then very likely to be occurred when WCE users described an event and links it to a (will be a) redirected page. Later when the Wikipedia stories get more information and more events, they were merged into a single one (although in WCE, events are still labelled with different story links. For example, 2011-2012 Syrian uprising\(^12\), Syrian uprising (2011-present)\(^13\) and Syrian Civil War\(^14\). As redirections introduce noise and clearly affect the relation between events and their stories, we resolve them by parsing the Wikipedia articles of the stories and follows the redirection pattern there. We opt for this solution instead of extracting ones from existing knowledge bases because of the freshness quality (which is better) as the redirections are done in quite frequent manners.

3.3 Statistics

Table 1 shows the number of news events and stories and Table 2 shows their relations that we extracted from WCE for the period from the 1st of January 2000 to 15th of October 2014.

The numbers are growing day by day as the WikiTimes knowledge base is updated regularly once every week. There are in average about 20 to 40 new events added every day.

We observe that only about 16% of the events are linked to news stories and only about 47% of the events are labeled with a news category.

3.4 Quality Assessment

\(^1\)http://www.opencalais.com/documentation/calais-web-service-api/api-metadata/document-categorization
\(^12\)http://en.wikipedia.org/wiki/Syrian_uprising
\(^14\)http://en.wikipedia.org/wiki/Syrian_Civil_War

Quality of Wikitimes is our most concern as we are interested in (high) accuracy knowledge base. As there is no existing computer-processable ground truth that suits our purpose, we relied on the human assessment to measure the extraction quality of Wikitimes. The separated assessment for our enrichment quality will be further provided in Section 3.

Our assessment approach can be explained as follows: we randomly selected at least 1% of the instances of each relation type and asked a group of three experts to assess, independently, if the relations were extracted correctly from WCE. Each assessor was shown the textual description of an event and was provided with the URL to the corresponding WCE page to judge the correctness of the extraction process (to make it clear, we do not intend to evaluate whether information in Wikipedia is correct or not).

Note that we only evaluated relations which are the backbone of Wikitimes (that are event-centric) including happenedOnDate, category, involvesEntity and belongsToStory (link to Wikipedia story page). Other relations, such as name of the story or mentionedIn are more or less obvious. The quality assessment has revealed that the accuracy of the extraction process is very high, ranges between 97% and 100%.

This section describes the human quality assessment of the extracted information. Table 3 summarizes the statistical results of this assessment, with Wilson interval \(\alpha = 0.05\). Compared to belongsToStory and involvesEntity, our heuristic algorithms could not however achieve 100% accuracy for happenedOnDate and category. This mainly due to our heuristic extraction algorithms make mistakes in very rare cases where users used uncommon structures in HTML tag.

### Table 1: Size of Wikitimes (Objects) by extraction

<table>
<thead>
<tr>
<th>Object Name</th>
<th>Event</th>
<th>Story</th>
<th>Entity</th>
<th>Category</th>
<th>Source</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>53049</td>
<td>1202</td>
<td>30117</td>
<td>15</td>
<td>69882</td>
<td>4746</td>
</tr>
</tbody>
</table>

### Table 2: Size of Wikitimes (Facts) by extraction

<table>
<thead>
<tr>
<th>Relation</th>
<th>Domain</th>
<th>Range</th>
<th>Facts</th>
</tr>
</thead>
<tbody>
<tr>
<td>happensOnDate</td>
<td>Event</td>
<td>Story</td>
<td>10294</td>
</tr>
<tr>
<td>category</td>
<td>Event</td>
<td>Category</td>
<td>26282</td>
</tr>
<tr>
<td>mentionedIn</td>
<td>Event</td>
<td>Source</td>
<td>70930</td>
</tr>
<tr>
<td>involvesEntity</td>
<td>Event</td>
<td>Entity</td>
<td>186129</td>
</tr>
<tr>
<td>summary</td>
<td>Event</td>
<td>String</td>
<td>53049</td>
</tr>
<tr>
<td>happenedOnDate</td>
<td>Event</td>
<td>Date</td>
<td>53049</td>
</tr>
<tr>
<td>hasCategory</td>
<td>Story</td>
<td>Category</td>
<td>1202</td>
</tr>
<tr>
<td>name</td>
<td>Story</td>
<td>String</td>
<td>1202</td>
</tr>
<tr>
<td>sameAs</td>
<td>Story</td>
<td>Yago URI</td>
<td>302</td>
</tr>
<tr>
<td>sameAs</td>
<td>Entity</td>
<td>Yago URI</td>
<td>26299</td>
</tr>
</tbody>
</table>

### Table 3: Human quality assessment of event relations

<table>
<thead>
<tr>
<th>Fact</th>
<th>Evaluated</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>happenedOnDate</td>
<td>312</td>
<td>0.978 +/- 0.014</td>
</tr>
<tr>
<td>happensOnDate</td>
<td>100</td>
<td>1.000 +/- 0.000</td>
</tr>
<tr>
<td>belongsToStory</td>
<td>300</td>
<td>0.979 +/- 0.015</td>
</tr>
<tr>
<td>category</td>
<td>100</td>
<td>1.000 +/- 0.000</td>
</tr>
<tr>
<td>involvesEntity</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. KNOWLEDGE ENRICHMENT

This section contains three main subsections. Each section describes the problem, the proposed solution and the evaluation of the results.

4.1 Enrichment of happenedOnDate relation

Each event is associated with a location where it takes place in. Similarly, each story should be associated with a list of locations where its events occur. However, this location information is lacking in WCE. Our effort in Wikitimes aims at fulfilling these miss-
ing by inferring from the provided WCE data. We then structured events and stories with location information under the happenedIn relation. In detail, we are interested in entities of “location” type in the short description of an event. It is also necessary that each location is disambiguated and linked to one Wikipedia article because that provides additional information for any use. For example, one can link them to knowledge bases of entities such as Freebase or Yago, one can associate it with geo-spatial database Yahoo GeoPlanet, or one can summarize evolution of events in one country or city.

Approach.
We observed that entities are fortunately well annotated and disambiguated within event description by WCE users as showed in our previous analysis [11]. Based on top of that, we applied a heuristic algorithm (explained below) for determining location of events, after that the gained event-location relations are propagated to the story-location relation if that event belongs to the story. Our simple but effective heuristic algorithm for event location detection is based on an observation, that is, if in the description of an event, there appears one and only one entity which indicates a location/place, the event is very likely to occurred in that location. For examples:

A 4.5 magnitude aftershock strikes Haiti within days of the recent devastating earthquake.
An early morning bomb explodes before two courthouses in Athens; there are no injuries.

It is worth mentioning that to detect whether an entity is location, our process relies on the Yago entity types, which has been claimed to be highly precise in entity types and has a good cover of entities in Wikipedia.

Enrich happenedIn relation for Story.
We applied two-step process to enrich happenedIn for stories. First, propagation step: we add happenedIn relation for a story to any location of its events. Note that, we are interested in locations where its related events happened. There are obviously some major locations where most of events in the story happened and some are just related to its events. Secondly, extend from Yago knowledge step: we enriched the locations for a story by employing <Yago::happenedIn> property in Yago knowledge base. Basically, <Yago::happenedIn> introduces same concepts to the relation we use in Wikitimes.

Result and Discussion.
Up to the submission time, we obtained for happenedIn relation 20693 event-related facts, 2047 story-related facts from 729 stories. To evaluate how good the enrichment process is, we randomly sampled 300 facts and distributed to members in our lab for a manual evaluation. Overall, we obtained 0.93 +0.029 accuracy, with Wilson interval α = 0.05.

We have obtained a high precision on the detection of event happenedIn, while covering around 40% of events in Wikitimes. Note that this location information for atomic events is not available elsewhere in either WCE or Yago. By propagation of event’s location to their stories, we successfully cover around 72% of our stories. Regarding happenedIn for stories, we only successfully found a low coverage of 4% (40 stories) in Yago, among them 30 stories are previously found by the propagation step.

Comparing to Yago, Wikitimes provides richer happenedIn facts which come both from atomic events (not available in Yago) and stories (available in Yago). In fact, we discovered that Yago contained around 11K happenedIn facts, but many of them coming from stories (more accurately, story-like entities since they do not have story concept) that are not in Wikitimes. The reason could be that those stories are probably not important or influential enough to be reported in WCE, according to its user, for example “Konginkangas bus disaster” 15 or “Battle off the coast of Abkhazia” 16. We leave investigation into that issue for the future work.

4.2 Enrichment of category Relation
High quality event categorization is a valuable feature of our knowledge base. Recent events and stories have been manually annotated by users into general topics (i.e., sports, politics). The manual annotations are, however, available on WCE only for events that occurred after May, 2011 (26282 events 570 stories), leaving a haft of dataset (26967 events, 442 stories) being non-categorized. That motivates us to fill up category information for those non-categorized events and stories.

Approach.
Noticing that WCE users typically associate an event to one category which is likely to represent for the most relevant topic of event’s content. We thus designed Wikitimes category relation following this constraint, that is, only the most relevant category of an event is taken into account. We extend the multi-category relation in the future release.

Our approach is based on the main practice through our separate experiments that: (i) major number events of a story will belong to the same category; and (ii) if solely use event description, state of the art approaches (e.g., LibShortText 17 among other tools) for automatic classification do not yield high precision. While we opt for a knowledge base of the high quality, we follow two-step heuristic algorithm as follow.

Propagation process: we propagate the main category of each story that occupies more than 95% number of events in this story to all events that happened before May, 2011 and belongs to that story. We annotated categories for existing stories which ended before May, 2011 and, therefore, has no category in Wikitimes. We consider that practical task since the number of non-categorized stories is relatively small. The annotation process is done by 2 experts in our research lab, whose agreement κ = 0.95. For example, given a story name “Andijan massacre”, each judge is asked to visit its Wikipedia page to understand the story and then provide its category, which is clearly “War_Conflict” in this case.

Automatic process: to events which have no story, we simply applied automatic text categorization. That is a traditional short text classification problem, given a pre-defined categories in OpenCalais. To tackle this task, one can train the classification model (e.g, SVM) on Wikitimes data which consists of user annotated 26282 events. One could employ external resources, such as DMOZ 18 or similar text categorization articles. We tried various available algorithms for avoiding inventing the wheel. By empirical experiments, we found that using well-known OpenCalais classification system 19 can provide high precision especially when we applied heuristic rules on the top of its output. In particularly, we only took the category associated to a high backward confidence probability while limiting results to events that are classified to only one category. Conservatively, we selected only those having confidence probability equals to 1.0, the highest possible value returned by

\[\frac{\text{Precision}}{\text{Recall}}\]
OpenCalais classification system.

Result and Discussion.

Table 4 shows a summary of category facts we obtained from our empirical approach. By propagation of story category, we obtained 2571 events from non-categorized event pool. By applying heuristic rules with OpenCalais categorization on the rest of data, we obtained 8537 events more. Together with user-annotated events, Wikitimes provides 37453 events with category, hence, covers 70% of whole data.

| Method                           | #Gained Facts | #Evaluated | Precision  
|----------------------------------|---------------|------------|------------|
| Propagate by story category      | 2737          | 90         | 0.94±0.05  
| Heuristic rules + OpenCalais     | 8434          | 250        | 0.91±0.03  

Our user evaluation indicates high precision with more than 90% correct. The results on applying hybrid approach of heuristics rules with OpenCalais categorization resulted in slightly lower precision than that of propagation process, which possesses manual effort from WCE users.

4.3 Enrichment relatedStory Relation

We enrich relations between two stories to indicate whether they are connected to each other. For example, readers may find “Euromaidan” which is related to “2014 Crimean crisis” or “Automotive industry crisis of 2008–2009” is related to “Late 2000s recession”. Knowing related stories obviously enhances users experience while navigating through events and stories. We consider following relations among stories:

**relatedStory relation**: it is when two stories are related to each other such as one can be a part of the other or consequential part of the other or simply link to the another, for example: (part of ) “Inter-rebel conflict during the Syrian Civil War” and “Syrian Civil War”, ( cause ) “September 11 Attack” and “Iraq War”, among others. As it is a broad relation, we are also interested in its two sub-relations: (1) partOf relation: when a story happened to be a sub-story of longer (or broader ) stories; and (2) isRedirectedTo relation: when two stories, after some time, are considered to be merged into one as previously described in Section 3.2.4. For example, 2011–2012 Syrian uprising and Syrian uprising (2011–present) and Syrian Civil War. After redirection resolution, we keep the links among them for later use.

In the following section, we describe how we find related stories as well as our algorithm based on machine learning (ML) to estimate how strong the relations between two stories.

4.3.1 Finding Related Stories

We find related stories of the stories in WCE by parsing the Wikipedia article of the story, not excluding its Infobox. There are some heuristic triggers for determining which relations we will link to the discovered stories: (i) if a Wikipedia article redirects to other Wikipedia articles, we give them “isRedirectedTo” relation; (ii) if there appears pattern “[partOf [[story article]]]”, we link them with “partOf” relation; (iii) any stories found in the abstract of a Wikipedia article as well as the abstract of every paragraph, which often said where the source of main information of the para-

graph comes from by “mainstory syntax”, we made them of “relatedStory” relation; (iv) other stories found in body content of the Wikipedia article are associated with “relatedStory” relation to the given story.

4.3.2 Estimation of Confidence Score

Obviously, the aforementioned process for finding related stories always ensures us that the discovered stories have some sense of relations to the given story. However, one may question how strong the relation between 2 news stories is. Some others may be interested in strongly related stories to dig in depth while others may want to discover weak related stories to have broader insights. Let’s take an example of “Eruptions of Eyjafjallajökull” (2010) versus the very related story “Air travel disruption after 2010 Eyjafjallajökull eruption” focuses in air travelling related events, but another story, namely “Polish Air Force Tu-154 crash” is related to the eruption as the funeral of Polish President was affected. Similarly, “2010 Copiapó mining accident” is related but not strongly related to New York City Marathon (as one survival of the mining accident participated to the marathon).

Hence, having some degree of relation measurement between those stories are helpful for users. To estimate the story-story relation’s strength, we additionally measure a “confidence” score of the relation. The higher confidence score is likely to indicate stronger two related story pair. Our technique is based on machine learning (ML) approach: we extract relevant features to represent how events of two stories are related to each other, plus the similarity in content of their Wikipedia pages. After that we train a ML model on annotated data and then apply to un-scored relations. Due to space limitation, we briefly describe our features as the follow:

**Features.**

**Event-based features.** We extract a feature set from the event-based information that are categorized as follows. (1) Event-based word overlap, the story similarity is the ratio of word overlaps between two bags of words represent two set of events. (2) Event-based distributional difference which is the KL divergence between the two word distributions of the stories. (3) Event-based geometric distance, the Cosine distance of two vectors formed by the word distributions. And lastly, (4) Event to story similarity, in which we model each event (instead of story) as a bag of word and measure the average similarity between the word distribution of the story events to the other story.

**Entity-based features.** Similar to event-based features but here we represent each story as bag of entities instead.

**Wikipedia article-based features.** We fetch the Wikipedia pages of the stories and compute feature set similar to event and entities-based features (based on the description on Wikipedia pages and also the set of entities).

**Time-based features.** Typically, a story is a chain of events with different major phases. For example, in the beginning phase of the “Syrian Civil War” story (Jan-Apr.2011), main events were about people protests. Several months after that, for example, August 2011, as there were many clashes and riots between protesters and government, main events were about clashes and number of people killed. Later, there was an escalation phase when battles between anti-government groups like Free Syrian Army and the regime hap-
Gained Facts

100
83
1.00 +- 0.00
427
100
0.98 +- 0.03
418

1.00 +- 0.00

Evaluated

0.98 +- 0.03

42
87

tion. We distributed them among our users for manual evaluation.

The task is done by 2 authors, separately. With each pair, we are
given wither 0, 1, or 2 to measure strength of the relation between
them. It is 0 if two stories are very less related, like the afore-
mentioned example “2010 Copiapo mining accident” and “New
York marathon 2010”; 2 if two stories are very related, like “Arab
Spring” and “Syrian Civil war” or “Israel -Palestine conflict” and
“Gaza war”. Those 2 cases are easy to annotated. We given 1 to
the relation when they are not fallen into those two cases. \( \kappa = 0.92 \)
is observed between them.

Given the condition that the number of stories in our dataset cur-
tently is not big so it is hard for single models for learning and
later predicting on very much unseen data (stories that have not
happened yet!). Therefore, we choose ensemble model\(^3\) because
it first constructs meta estimator that fits a number of randomized
decision trees on various sub-samples of the dataset, then uses aver-
aging to improve performance and controls over-fitting. Our exper-
iment results in 0.82 by Spearman and 0.84 by Pearson correlation,
which are pretty good for a regression task. At the end, we rescale
the predicted score to [0, 1] for better intuition given any new story
pairs.

Results and Discussion.

Evaluation of relatedStory facts. In total, we obtained 2211
facts of relatedStory relation in which 427 are isRedirectedTo and
100 are partOf. Table 5 shows results of story relation enrichment
and human evaluation on the relations among stories. For evalua-
tion, users (we distributed the facts to colleagues in our labs) are
asked to read Wikipedia articles of given pairs of stories for back-
ground and answered whether two stories are related to each other
as if one refers to the other in Wikipedia articles in the defined
relations. Our results suggest very high precision, even though it is
not so surprised, and reliable answers for related stories question.

Table 5: Enrichment result of relatedStory relation

<table>
<thead>
<tr>
<th>Relation</th>
<th>#Gained Facts</th>
<th>#Evaluated</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>isRedirectedTo</td>
<td>427</td>
<td>100</td>
<td>1.00 +- 0.00</td>
</tr>
<tr>
<td>partOf</td>
<td>100</td>
<td>100</td>
<td>0.98+ -0.03</td>
</tr>
<tr>
<td>*relatedStory</td>
<td>1683*</td>
<td>100</td>
<td>1.00+ -0.00</td>
</tr>
<tr>
<td><strong>highly related story</strong></td>
<td>418**</td>
<td>83</td>
<td>0.98+-0.03</td>
</tr>
</tbody>
</table>

\*: count excluding isRedirectedTo and partOf; **: retrieval of highly related stories by
\( \geq 0.9 \) confidence score

Retrieval of highly related stories. In addition to the facts we
enriched, we would like to see how helpful our confidence score
is in scenario that users may look for highly related stories. We
retrieved the sample of 83 facts from 418 facts (20%) of which
each has confidence score from 0.9, excluding isRedirectedTo re-
lation. We distributed them among our users for manual evaluation.
Again, we also obtained very high answers to retrieval of highly
related stories on the sample, thus, claims the usefulness of confi-
dence score in this scenario. Few good examples include “Extraju-
dicial killings and forced disappearances in the Philippines”\(^31\) and
“Mauguidanau massacre”, or “Territorial disputes in the South
China Sea”\(^32\) and “Spratly Islands dispute”, whose strong con-
nection is not available in Inforbox or any similar structured infor-
mation as is in Wikipedia.

4.4 Summary

We have presented our efforts in enriching information of Wik-
times to extend what is extracted from WCE in term of location,
category and related stories aspect. For location, we obtained 20693
facts for events with around 93% accuracy overall. In addition, by
propagation to story level, we obtained 2047 facts for 729 stories
for indicating locations that are related to those stories. For cate-
gory, our heuristic based approach employs indirected information
in WCE and OpenCalais categorization system to assign category
for 11171 events which have no category yet. The accuracy of this
approach is more than 91%. In total, Wikitimes have round 37K
events with category information. We presented an approach to
identify related stories of a story which gained 2211 facts in to-
tal. We also presented a machine learning approach to score that
relatedness, which has demonstrated an excellent performance in
identify highly related stories.

5. CONCLUSION

We introduced WikiTimes, a summarization-centric knowledge
base of news events by the crowd. This paper presents our knowl-
edge extraction process of WikiTimes system. The event informa-
tion is extracted from Wikipedia’s Current Events portal, which is
a special part of Wikipedia that is valuable however not yet exploited
by existing knowledge bases so far. WikiTimes extends existing
knowledge bases such as DBpedia and YAGO by timeline summaries
with human readable texts wrt. a story or an entity. The content of WikiTimes is publicly accessible in different ways including
SPARQL endpoints, RESTful endpoints as well as graphical
user interfaces.

We showed that the quality of the extracted information is around
98%, which is confirmed with human assessments. That suggests
Wikitimes is a good source for many applications also as golden
standard datasets for improving and evaluating important event se-
lection and summarization.

6. REFERENCES

of new topics. In Proceedings of SIGIR’01, pages 10–18,
along a timeline. In Proceedings of SIGIR’04, pages
S. Das, S. Subramaniam, A. Rajaraman, V. Harinarayan,
and A. Doan. Entity extraction, linking, classification, and
VLDB Endow.
S. Siersdorfer, and W. Nejdl. Temporal summarization of
event-related updates in wikipedia. In Proceedings of the
\(^31\)http://en.wikipedia.org/wiki/Extrajudicial_killings
_and_forced_disappearances_in_the_Philippines
\(^32\)http://en.wikipedia.org/wiki/Territorial_disputes_in_the_South_China_Sea


