Web Semantics: Science, Services and Agents on the World Wide Web 37–38 (2016) 132–151

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

Building event-centric knowledge graphs from news

Marco Rospocher [a], Marieke van Erp [b,*], Piek Vossen [b], Antske Fokkens [b], Itziar Aldabe [c], German Rigau [c], Aitor Soroa [c], Thomas Ploeger [d], Tessel Bogaard [d]

[a] Fondazione Bruno Kessler, Trento, Italy
[b] Vrije Universiteit Amsterdam, The Netherlands
[c] The University of the Basque Country, Donostia, Spain
[d] SynerScope B.V., Helvoirt, The Netherlands

ARTICLE INFO

Article history:
Received 7 April 2015
Received in revised form
21 October 2015
Accepted 22 December 2015
Available online 12 January 2016

Keywords:
Event-centric knowledge
Natural language processing
Event extraction
Information integration
Big data
Real world data

ABSTRACT

Knowledge graphs have gained increasing popularity in the past couple of years, thanks to their adoption in everyday search engines. Typically, they consist of fairly static and encyclopedic facts about persons and organizations – e.g. a celebrity's birth date, occupation and family members – obtained from large repositories such as Freebase or Wikipedia.

In this paper, we present a method and tools to automatically build knowledge graphs from news articles. As news articles describe changes in the world through the events they report, we present an approach to create Event-Centric Knowledge Graphs (ECKGs) using state-of-the-art natural language processing and semantic web techniques. Such ECKGs capture long-term developments and histories on hundreds of thousands of entities and are complementary to the static encyclopedic information in traditional knowledge graphs.

We describe our event-centric representation schema, the challenges in extracting event information from news, our open source pipeline, and the knowledge graphs we have extracted from four different news corpora: general news (Wikinews), the FIFA world cup, the Global Automotive Industry, and Airbus A380 airplanes. Furthermore, we present an assessment on the accuracy of the pipeline in extracting the triples of the knowledge graphs. Moreover, through an event-centered browser and visualization tool we show how approaching information from news in an event-centric manner can increase the user's understanding of the domain, facilitates the reconstruction of news story lines, and enable to perform exploratory investigation of news hidden facts.

1. Introduction

Knowledge graphs have gained increasing popularity in the last couple of years, thanks to their adoption in everyday search engines (e.g., Google, Bing). A knowledge graph is a knowledge-base of facts about entities (e.g., persons, organizations),[1] typically obtained from structured repositories such as Freebase and Wikidata, or extracted from encyclopedic knowledge such as Wikipedia. For instance, given a famous person, knowledge graphs typically cover information such as her birth date and birth place, her relatives and the major events and activities that made her famous. However, only a small part of what happens in the world actually makes it into these databases. There are many events that are not considered important enough to be included or may not directly involve famous people that have entries. Furthermore, current repositories tend to represent the actual state of the world and do not focus on the dynamics and the changes over time. More fluid information as reported in the growing stream of daily news tends to get lost in current knowledge graphs and our fading memories, but it can be of great importance to information professionals needing to reconstruct somebody's past or the massive history of complete industries, regions or organizations. There is thus a need for a different type of structured database constructed around events rather than entities and entity-focused actual facts. Capturing this dynamic knowledge requires to consider events as the unit for storing knowledge regardless of the fame of the people involved.

* Corresponding author.
E-mail address: marieke.van.erp@vu.nl (M. van Erp).

[1] The description of the latest release of DBpedia is an illustrative example as it states the following: "The English version of the DBpedia knowledge base currently describes 4.58 million things [...] including 1,445,000 persons, 735,000 places [...], 411,000 creative works [...], 241,000 organizations [...], 251,000 species and 6000 diseases". Events are not mentioned. http://blog.dbpedia.org/?p=77 *Last accessed: 7 April 2015.*

http://dx.doi.org/10.1016/j.websem.2015.12.004
1570-8268/

In this paper, we present a method and an open source toolkit to automatically build such *Event-Centric Knowledge Graphs (ECKGs)* from news articles in English and Spanish, Italian and Dutch. We define an Event-Centric Knowledge Graph as a Knowledge Graph in which all information is related to events through which the knowledge in the graph obtains a temporal dimension. In a traditional KG, information is often centered around entities. One can then find RDF triples (subject, predicate, object) where the subject and object are often entities, and any information about events is generally captured through the predicate. In ECKGs the subject of triples is typically the event related to entities and bound to time. This will allow specialists to reconstruct histories over time and networks across many different people and organizations through shared events. Dynamic trends and regional changes can be made visible abstracting from individuals and reasoning over the temporal aspects.

Consider the following example on the company Porsche. In DBpedia, the entry for the company Porsche provides triples that state what type of companies it is, what cars it makes, what management it has, etc. It does not list the history of deals, the market events, the changes in managements, nor the successes and failures over a longer period of time. On 15 October 2015, the Wikipedia entry of the same company does give a brief history in natural language, including how it was fully acquired by Volkswagen in 2009 but obtained 100% voting rights within the Volkswagen group in 2013 by buying back 10% stake from Qatar Holding. This history is not represented as structured data in DBpedia. If we next look at the Wikipedia page for Qatar Holding, we also find a brief history in natural language text that is not represented as structured data in the corresponding DBpedia entry. Interestingly, the history of Qatar Holding mentions that it currently still holds about 17% stake in the Volkswagen Group and Porsche. It does not mention that 10% of this stake was sold back to the Porsche family in 2013. Apparently, this event was important for the Porsche SE history but not for the Qatar Holding history. As events are first class citizens in our ECKGs (similar to entities in many other KGs), these selling and buying events are represented as a single event in which Porsche loses an asset and Qatar Holding acquires one, regardless of the perspective of the two companies and their relevance for either one. We leave it up to the user to order events in time, place and around participants to reconstruct storylines or histories from a complete representation of all events reported in the news.

From a representational point of view, in our ECKGs every event is a node of our knowledge graph and is uniquely identified by an URI, on which various properties can be asserted via triples. This provides a homogeneous representation of events, differently from what happen in other resources: e.g., in DBpedia, an analogous representation is applied only for *named* events[2] such as http://dbpedia.org/resource/2009_Japanese_Grand_Prix, while a minimal number of *smaller* events without established name are captured by properties such as http://dbpedia.org/property/acquired.

By exploiting state-of-the-art Natural Language Processing (NLP) techniques, we automatically extract information about the *events* mentioned in millions of news articles, together with the information on the event participants, time and location. All the extracted content is organized in an ECKG in a structured representation grounded in Semantic Web best practices. Moreover, these pieces of information are linked to available linked data resources (e.g. whenever possible, entities participating in events are linked to their DBpedia referent, otherwise an entity instance in our knowledge base is created) as well as to the actual textual occurrences from which they were extracted. Determining event identity and anchoring events to time eventually results in the representation of long-term developments and story-lines, where events are connected through bridging relation such as cause or co-participation. These "histories" reconstructed from news capture changes in the world instead of static properties and facts in traditional knowledge graphs.

To construct an ECKG, we have identified four main information extraction challenges: (i) proper modeling of the expression of information in text and the referential value of the expression in the formal semantic ECKG model; (ii) correctly extracting and interpreting the information contained in a news article, according to the ECKG data model; (iii) linking the extracted information to established linked data repositories (e.g., DBpedia); (iv) establishing referential identity for entities and events across different expressions and mentions within and across different sources (e.g., same entity or event mentioned in different news articles), potentially in different languages.

Our approach tackles all four challenges, as demonstrated in the four knowledge graphs that we built in several distinct domains. The text corpora from which we have constructed our ECKGs range from a few hundred to millions of news articles. The individual modules in our processing pipeline all perform at the level of or exceed the current state-of-the-art in natural language processing technology. Our ECKGs can then be used to answer queries that are difficult to answer using traditional KGs or the unprocessed source documents, as is the current de facto standard for information professionals. To the best of our knowledge, we are the first to automatically build ECKGs from large, unstructured news article text collections. Furthermore, our method also works cross-lingually, enabling integration of ECKGs extracted from different languages.

In this paper, we combine the contributions reported in several publications on the NewsReader project from the perspective of ECKGs. These contributions cover:

1. a definition of Event-Centric Knowledge Graphs (Section 1)
2. a formal semantic representation for ECKGs that includes reference to the original source (Section 3)
3. a method and open source tools for the extraction of Event-Centric Knowledge Graphs in four languages (Section 4)
4. four openly available ECKGs (Section 5)
5. a first assessment of the quality of automatically created ECKGs (Section 6).

The paper is organized as follows. In Section 2, we describe the background and related work. In Section 3, we describe how we modeled the information we extracted. In Section 4, we describe our processing pipeline. In Section 5, we describe our four use cases, namely general news, the FIFA world cup, and the global automotive industry, and news articles about the Airbus A380 in different languages. In Section 6, we report a first assessment of the accuracy of the ECKGs automatically created with our approach. In Section 7, we describe event-centric information access using the SynerScope tool, and report on additional applications and investigations enabled by our ECKGs. In Section 8, we discuss our approach and conclusions.

2. Background and related work

Knowledge Graphs (KGs) are used extensively to enhance the results provided by popular search engines (e.g. Google Knowledge Graph,[3] Microsoft's Satori[4]). These KGs are typically powered by

[2] Entities of type http://dbpedia.org/ontology/Event, in many cases corresponding to sports events or military conflicts.

[3] http://www.google.com/insidesearch/features/search/knowledge.html *Last accessed: 7 April 2015.*

[4] http://blogs.bing.com/search/2013/03/21/understand-your-world-with-bing/ *Last accessed: 7 April 2015.*

structured data repositories such a Freebase,[5] DBpedia [1], Yago [2], and Wikidata [3] (which are in itself KGs as well), that traditionally focus on encyclopedic facts and knowledge, i.e., they contain information such as the name and/or surname of some famous person, her date and place of birth and her professional activities. Dynamic information, such as the most recent events reported in the news involving that person, are generally not captured in these resources, and thus missing from most KGs.

This is partially due to the shortage of resources offering structured content about events. Indeed, only a few Linked Data resources describing events are available: Last.FM[6] and EventMedia.[7] Last.FM is an RDF version of the Last.FM website,[8] containing information about events, artists, and users. EventMedia is an aggregate of three public event directories (last.fm, eventful, and upcoming) and two media directories (flickr, YouTube). Events are represented using the LODE ontology,[9] while media are represented with the W3C Ontology for Media Resources.[10] It is interlinked with DBpedia, Freebase, Geonames[11] and it also contains links to numerous related web pages from MusicBrainz,[12] Last.fm, Eventful[13] Upcoming,[14] and Foursquare.[15] In the realm of biomedical research, knowledge bases such as Bio2RDF [4] and Open PHACTS [5] have sprung up. However, these resources are constructed from already structured data, whereas our ECKGs are constructed from plain text sources. Our approach differs in the sense that it can work with any news article and is thus not constrained to a specific web site or domain.

The past few years have seen an increasing research interest in supporting the automatic construction of knowledge graphs, although much of this effort was devoted toward the development of statistical models to infer new facts about entities in the graph (see [6]). Some prominent projects have been proposed to extract knowledge bases from semi-structured resources such as Wikipedia (cf. DBpedia [1], Freebase [7] or Google Knowledge Vault [8]), but the extracted information is centered on collecting facts around entities rather than events. In Ontos News Portal,[16] persons, organizations, locations, as well as some facts about these entities are automatically extracted from news articles. The Ontos News Portal differs from our approach, because event extractions are not explicitly addressed and only shallow natural language processing techniques are applied to extract content, resulting in a shallow grouping of news stories by topics and entities. [9] presents an approach to organize news articles around stories, which imply events, but the process relies on co-occurrence of words and phrases in a sequence of news articles rather than deep Natural Language Processing (NLP) techniques and does not yield a knowledge base as a resulting structure but a chain of news articles that form a story. EVIN [10] is an approach for automatically extracting named events from news article, while in our approach we perform a deeper analysis of text to extract any kind of event mentioned in the text, also those that have not received proper names.

The automatic extraction of facts and events from news articles has been addressed using more advanced NLP techniques commonly known as Open Information Extraction systems that are not tuned to a particular type of event or entity or domain. Examples of such systems are TextRunner [11] and NELL [12]. A correct interpretation of a text requires the detection of the event mentions and the participants that play a role in these events, including time and location expressions. [13] was the first to demonstrate the appropriateness of Semantic Role Labeling (SRL) for the identification of event frames in Information Extraction. [14] presented the use of semantic roles to extract events and their relations as defined by TimeML [15]. SRL has also been used to extract events from Wikipedia [16], to build an Open Information extractor [17] and for mining event-based commonsense knowledge from the Web [18]. The XLike project [19][17] is probably the closest related to the NewsReader project. In this project information from news articles in several languages is extracted, and converted to a common semantic representation. However, it differs from our approach in that similar news articles are clustered, thus distilling one representative macro event for each cluster obtained.[18] A similar clustering approach is applied also in [10]. In NewsReader, we perform a much more fine-grained extraction of events, *machine reading* each news article, and thus identifying (possibly multiple) events within it: this enables to capture events which, though not mentioned in some headlines, may be crucial to take well-informed actions in professional decision-making contexts.

There are already tools to convert the output from NLP processing to Semantic Web formats, with the most prominent tools being NLP2RDF[19] and Fred [20]. Our approach differs from these efforts, because after executing an advanced NLP processing pipeline, we perform an additional cross-document cross-lingual integration step to move from text mentions to instances. This additional step goes beyond what the aforementioned tools currently offer. The added benefit of our approach is that our instance representation allows us to aggregate information across many different sources, even in different languages showing complementarity and differences across these sources and the provenance of the information provided.

Our method will be presented in Section 4, after a description of our representation schema for event-centric knowledge in Section 3.

3. Event-centric knowledge representation

The event-centric data in our ECKG is meant to represent long-term developments and story lines by anchoring events in time and place and linking them to entities. Indeed, following [21] we define events as *things that happen*, consisting of four components:

1. an event action component describing what happens or holds true
2. an event time slot anchoring an action in time describing when something happens or holds true
3. an event location component specifying where something happens or holds true
4. a participant component that gives the answer to the question: who or what is involved with, undergoes change as a result of or facilitates an event or a state.

At the same time, the representation schema we employ needs to relate this data to detailed representations of events that are the output of deep linguistic analyses of texts, accumulating information about the same event from different sources and over time. We consider both the changes in the world and the news reporting on these changes as streams that are not fully aligned.

5 http://www.freebase.com/ *Last accessed: 7 April 2015.*
6 http://datahub.io/dataset/rdfize-lastfm *Last accessed: 7 April 2015.*
7 http://eventmedia.eurecom.fr/ *Last accessed: 7 April 2015.*
8 http://www.last.fm *Last accessed: 7 April 2015.*
9 http://linkedevents.org/ontology/ *Last accessed: 7 April 2015.*
10 http://www.w3.org/TR/mediaont-10/ *Last accessed: 7 April 2015.*
11 http://www.geonames.org *Last accessed: 7 April 2015.*
12 http://musicbrainz.org/ *Last accessed: 7 April 2015.*
13 http://www.eventful.com *Last accessed: 7 April 2015.*
14 http://www.upcoming.org *Last accessed: 7 April 2015.*
15 http://www.foursquare.com *Last accessed: 7 April 2015.*
16 http://news.ontos.com/ *Last accessed: 14 October 2015.*
17 http://www.xlike.org/ *Last accessed: 7 April 2015.*
18 See e.g., the XLike Event Registry (http://eventregistry.org/) *Last accessed: 7 October 2015.*
19 http://nlp2rdf.org/ *Last accessed: 7 April 2015.*

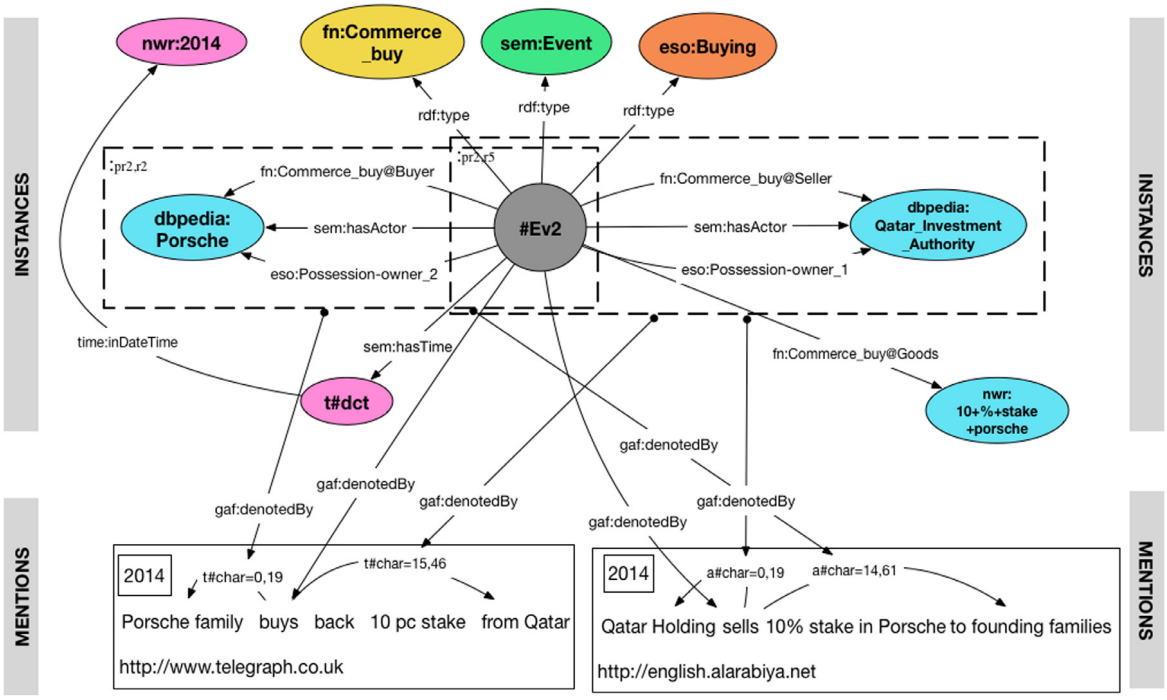

Fig. 1. Illustration of the representation of an event two main participants, links to ontologies and source text.

Event knowledge is usually spread over many different news articles. Over time, more information is provided or the perspective on the events in the world changes. For example, the first news article that reports on Qatar selling their 10% stake to the Porsche family does not mention the amount of money transferred. When this becomes known, it is published at a later point in time in other articles that update the information. In order to obtain a comprehensive representation of events, we thus need to be able to aggregate all information on the same event from many different sources that complement but possibly also contradict each other. In addition to establishing that these sources report on the same event (event identity), we thus also need to aggregate new information and represent conflicts.[20]

Given these objectives and the wide variation of information that may be found in text, we have established the following functional requirements for our representation schema:

1. It should define event identity across mentions and expressions in different sources.
2. It should define entity identity and time and place identity.
3. It should be able to handle complementing and conflicting information (when sources contradict each other).
4. It should provide the provenance of information (to compare between sources and allow users to assess the reliability of information).
5. It should be easy to relate the model to other structured repositories and ontologies for identifying background information as well as for facilitating reasoning.
6. It should be formally defined to allow for reasoning, e.g. to abstract from instances to generalize to categories or to derive implications of events.
7. The definitions that we use should be generic enough to capture a wide variation of events.

In addition, we aim for the following non-functional requirements which support the functional requirements outlined above:

8. The representations we used should build upon existing models as much as possible.
9. The data will be represented in RDF extended with Named Graphs.

In this section, we describe how we represent the information we extract from our texts in an Event-Centric Knowledge Graph with respect to the requirements listed above using the Grounded Annotation Framework (GAF, [23]) and the Simple Event Model (SEM, [24]). GAF and SEM complement each other: GAF provides links between events and the sources where they are mentioned and SEM models the events themselves, their participants, location and time. Section 3.1 explains how GAF provides a natural way to model coreference and provenance information fulfilling our first four functional requirements. In Section 3.2, we illustrate how SEM's simplicity and flexibility fulfill the last three requirements.

The explanations in Sections 3.1 and 3.2 are illustrated using the following example, consisting of two article titles published by different papers on the same date:

1. Porsche family buys back 10 pc stake from Qatar (source: http://www.telegraph.co.uk).
2. Qatar Holding sells 10% stake in Porsche to founding families (source http://www.english.alarabiya.net).

Fig. 1 provides a simplified illustration of how the interpretation of these titles is presented in our model. Fig. 2 provides part of the RDF triples that provide the same information. Figs. 1 and 2 indicate the outcome of a perfect analysis of the sentences and are intended to illustrate the structure of our model. The differences between this perfect interpretation and the actual output will be discussed in Section 4.

[20] The possibility of revealing what different sources state about the same event allows users to make a better informed judgment about the reliability of information. Users can furthermore examine differences between information coming from different sources (i.e. where do they contradict or complement each other?). The details of capturing different perspectives are beyond scope of this paper. It is mentioned here, because it forms one of the main motivations for the model. A description of this part of the model can be found in [22].

```
1  @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
2  @prefix time: <http://www.w3.org/TR/owl-time#> .
3  @prefix eso: <http://www.newsreader-project.eu/domain-ontology#> .
4  @prefix gaf: <http://groundedannotationframework.org/gaf#> .
5  @prefix nwrontology: <http://www.newsreader-project.eu/ontologies/> .
6  @prefix sem: <http://semanticweb.cs.vu.nl/2009/11/sem/> .
7  @prefix fn: <http://www.newsreader-project.eu/ontologies/framenet/> .
8
9  <http://www.newsreader-project.eu/instances> {
10     <http://www.telegraph.co.uk#ev2>
11             a sem:Event , fn:Commerce_buy , eso:Buying ;
12             rdfs:label "buy" , "sell";
13             gaf:denotedBy <http://www.telegraph.co.uk#char=15,19> , <http://
                   english.alarabiya.net#char=14,19> .
14
15     <http://dbpedia.org/resource/Porsche>
16             rdfs:label "Porsche" , "founding family" ;
17             gaf:denotedBy <http://www.telegraph.co.uk#char=0,7> , <http://
                   english.alarabiya.net#char=33,40> , <http://english.alarabiya.
                   net#char=44,61> .
18
19     <http://www.newsreader-project.eu/data/cars/non-entities/10pc+stake>
20             rdfs:label "10pc stake", "10 \% stake in Porsche" ;
21             gaf:denotedBy <http://www.telegraph.co.uk#char=25,35>, <http://
                   english.alarabiya.net#char=20,40> .
22
23     <http://dbpedia.org/resource/Qatar>
24             rdfs:label "Qatar", "Qatar Holdings" ;
25             gaf:denotedBy <http://www.telegraph.co.uk#char=41,46> , <http://
                   english.alarabiya.net#char=0,13> .
26
27     <http://www.telegraph.co.uk#dct>
28             a time:Interval ;
29             rdfs:label "2014" ;
30             gaf:denotedBy <http://www.telegraph.co.uk#dctm> , <http://english.
                   alarabiya.net#dctm> ;
31             time:inDateTime <http://www.newsreader-project.eu/time/2014> .
32 }
33
34 <http://www.telegraph.co.uk#dt2> {
35     <http://www.telegraph.co.uk#ev2>
36             sem:hasTime <http://www.telegraph.co.uk#dct> .
37 }
38
39 <http://www.telegraph.co.uk#pr2,rl2> {
40     <http://www.telegraph.co.uk#ev2>
41             sem:hasActor <http://dbpedia.org/resource/Porsche> ;
42             eso:possession-owner_2 <http://dbpedia.org/resource/Porsche> ;
43             fn:Commerce_buy@Buyer <http://dbpedia.org/resource/Porsche> .
44 }
45
46 <http://www.telegraph.co.uk#pr2,rl4> {
47     <http://www.telegraph.co.uk#ev2>
48             sem:hasActor <http://www.newsreader-project.eu/data/cars/non-
                   entities/10pc+stake> ;
49             fn:Commerce_buy@Goods <http://www.newsreader-project.eu/data/
                   cars/non-entities/10pc+stake> .
50 }
51
52 <http://www.telegraph.co.uk#pr2,rl5> {
53     <http://www.telegraph.co.uk#ev2>
54             sem:hasActor <http://dbpedia.org/resource/Qatar> ;
55             eso:possession-owner_1 <http://dbpedia.org/resource/Qatar> ;
56             fn:Commerce_buy@Seller <http://dbpedia.org/resource/Qatar> .
57 }
58
59 <http://www.newsreader-project.eu/provenance> {
60     <http://www.telegraph.co.uk#dt2>
61             gaf:denotedBy <http://www.telegraph.co.uk#char=30,35> , <http://
                   english.alarabiya.net#char=24,29> .
62
63     <http://www.telegraph.co.uk#pr2,rl2>
64             gaf:denotedBy <http://www.telegraph.co.uk#char=0,19> , <http://
                   english.alarabiya.net#char=14,61> .
65
66     <http://www.telegraph.co.uk#pr2,rl4>
67             gaf:denotedBy <http://www.telegraph.co.uk#char=15,35> , <http://
                   english.alarabiya.net#char=14,40> .
68
69     <http://www.telegraph.co.uk#pr2,rl5>
70             gaf:denotedBy <http://www.telegraph.co.uk#char=15,46> , <http://
                   english.alarabiya.net#char=0,19> .
71 }
```

Fig. 2. Example of instances, the relations between them and the link to their mentions in TriG.

3.1. The grounded annotation framework

One of the main properties of our ECKG model is that we make a clear distinction between *instances*, which represent events and other entities in the world (which may or not exist and may or may not have happened) and *mentions*, which represent, in our case textual, expressions that refer to these events and entities. The Grounded Annotation Framework (GAF, [23]) allows us to indicate which *mentions* refer to a specific *instance* through the `gaf:denotedBy` relation. We will explain the distinction using our example.

Our example sentences both express the same event: Porsche (or the Porsche family) buying Porsche stakes back from Qatar. This event is represented by the *instance* labeled *#Ev2* in Fig. 1. Both the expressions *buy* from Example 1 and *sell* from Example 2 refer to this event. This is indicated by the `gaf:denotedBy` arrows going from the instance *#Ev2* to the tokens *buy* and *sell* from the original text. The triples in line 13 of Fig. 2 show how this relation is expressed in RDF. Linking two mentions to the same instance through GAF directly reflects that they both refer to the same thing, i.e. GAF provides a natural way to model coreference [23]. The same principle is applied to the participants and time of the event. The `gaf:denotedBy` arrows for the participants and time have been omitted from the images for simplification, but the corresponding triples can be found in the RDF example in Fig. 2. Lines 16 and 17 link `dbpedia:Porsche` to the labels and tokens mentioning it in the text, lines 20 and 21 provide this information for the 10% stakes, lines 24 and 25 for Qatar Holdings and lines 29 and 30 indicate where the time of the event is mentioned. Querying for triples involving a URI, regardless of whether it represents an event or some other entity, thus provides an aggregated overview of all information about this entity identified in the corpus.

Because we also want to represent what each source says exactly about an event, we link the relations we identified between an event and its participants back to the source. This is slightly more complicated than linking entities to a mention, because (1) we must link a triple back to its source rather than a simple URI and (2) a relation between concepts is typically expressed by a relation between words rather than a simple expression. We therefore assign identifiers to the linguistic relations between words we identify in text. The `gaf:denotedBy` relation indicates which linguistic relation in the text expresses the semantic relation between two instances. There are several ways to make statements about triples in RDF. In our model, we use named graphs as introduced in RDF 1.1 [25].

Statements are placed in the same named graph based on shared provenance. A statement representing the relation between an event and participant will typically end up in its own named graph, because a specific linguistic relation generally expresses only one relation between an event and participant. This is also the case in our example: the relations between the buying event and its participants are each in separate named graphs that only contain information about these specific relations. To illustrate, the box around `dbpedia:Porsche` and *#Ev2* represents the named graph which is also described by the triples in lines 39–44 in Fig. 2. The relation between Porsche and the event is expressed by two linguistic relations between *buy* and *Porsche family* in sentence (1) and between *sell* and *founding family* in sentence 2, labeled *t#char=0,19* and *a#char=14,61*, respectively.[21] The `gaf:denotedBy` arrows connect the named graph to these mentions. The equivalent information in RDF can be found in lines 63 and 64 of Fig. 2.

[21] Note that the URI of a mention actually refers to an RDF resource, which can be described by assertions stating (the URI of) the news article containing the mention, its starting and ending character offset in the text, additional linguistic processing attributes, etc. These assertions are omitted in the example here described, but can be seen in the video referred in Section 5.1, where an ECKG instantiating the representation here discussed is accessed.

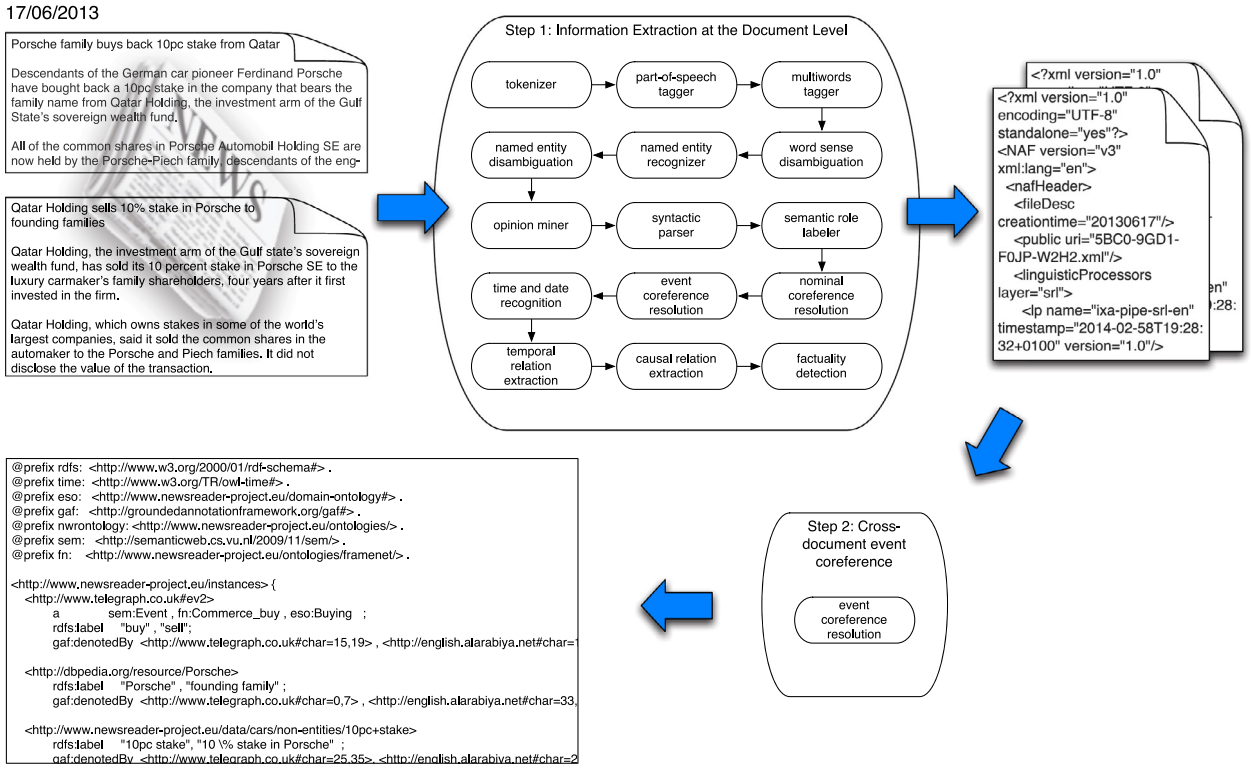

Fig. 3. Overview of the processing pipeline. On the left hand side, a set of documents comes in, which, in Step 1, is processed by the NLP pipeline. The NLP pipeline outputs a set of NAF documents containing the information extracted by the NLP pipeline per document. Then, in step 2, the cross-document event coreference aggregates the information from the individual NAF documents into an event-centric RDF representation.

GAF allows us to fulfill the first four requirements for our representation schema, i.e. it defines identity between events, entities, time and place (by referring alternative mentions to the same URI), it allows us to handle complementary and conflicting information (information is aggregated at instance level, yet appropriately separated at mentions level) and provides provenance information (it links information back to the news articles where they are mentioned).[22]

3.2. Event model

The previous subsection explained how we fulfill the first four functional requirements for our representation schema. The remaining requirements are concerned with how we represent events and their relations at the instance level of our model.

We use an extended version (SEM+, as described in [27]) of the Simple Event Model (SEM [24]) as a basis to model events. SEM is among the most flexible event models that is easy to adapt to different domains making it a suitable candidate for our purposes. SEM provides a generic framework to represent *who* did *what when* and *where* fulfilling the fifth requirement for our representation schema. These generic relations are compatible with more explicit relations. We can thus easily extend SEM representations with information from other ontologies fulfilling requirement six. In addition, we can make the relation between a complex event and its subevents explicit and indicate causal relations. Temporal relations between events are modeled through their associated times.

The information that our pipeline extracts from text is much more detailed than the generic model provided by SEM+. Our NLP pipeline, described in Section 4, also links events to FrameNet frames [28] and classes defined in the Event and Situation Ontology (ESO) [29]. FrameNet provides descriptions of events and their participants and is based on the theory of Frame Semantics. In our example, the FrameNet relations between the event and its participants indicate that Porsche is the *Buyer* and Qatar Holdings the *Seller*, where SEM does not distinguish what exact role each participant plays. The ESO relations model the difference in situation before and after the event. In our example, Qatar Holdings is owner_1, the owner before the sale takes place. Porsche is owner_2, which is defined as the owner after the event takes place. ESO provides this information for all events where ownership changes (e.g. stealing, buying, donating). In cases where a buying or selling event are not identified as the same or where one source claims something was bought and the other source that it was stolen, ESO allows us to infer that in both cases an object started in the possession of Owner A and ended with Owner B. Our formal representation in RDF allows us to link our information to ontologies such as ESO fulfilling the seventh requirement.

In the next section, we describe the process of identifying event and participant mentions in text and how we establish which mentions refer to the same instance.

4. Method and tools

Fig. 3 shows a schematic overview of our processing pipeline. There are two main steps to our data processing methodology: information extraction at the document level and cross-document event coreference. The document information extraction step is performed by a Natural Language Processing (NLP) pipeline that extracts mentions of events, actors and locations from the text and interprets time expressions. The document level processing

[22] Here 'provenance' refers to the basic capability of linking content to the news article where it was extracted. Standard provenance vocabularies such as [26] can be employed to complement this information (e.g. who authored the source, how the mention was extracted).

generates an interpretation for each mention and stored the results in the so-called Natural Language Processing Annotation Format (NAF,[23] [30]). Each news article is represented as a single NAF file. The cross-document event-coreference processing reads all the NAF files for a stream of news and crystallizes the mentions into instances, effectively forming a bridge, through GAF links, between the document (NLP) and instance (SW) levels. The output of the second step is in RDF-TRiG format according the GAF and SEM schemas as explained in Section 3.

In the remainder of this section, we explain these steps in our pipeline and conclude with details on our implementation.

4.1. Information extraction from documents

The information extraction pipeline extracts entities and events from raw text of newspaper articles [31]. The processing chain consists of several NLP modules that perform the required steps, as described below. In total, the system includes pipelines for English, Spanish, Dutch and Italian. Obviously, each pipeline relies on a very different set of linguistic modules adapted to perform a particular task in a language. Currently, the English pipeline is composed of 15 modules,[24] the Spanish pipeline integrates 11 modules,[25] the Dutch pipeline is composed of 14 modules,[26] and the Italian pipeline is composed of 11 modules from TextPro.[27]

The modules adopt a simple and well-known data-centric architecture, in which every module is interchangeable for another one as long as it reads and produces the required data format. This data-centric approach relies on NAF: an interchange format for representing linguistic annotations. NAF has evolved from the KYOTO Annotation Framework (KAF, [32]) and it is compliant with the Linguistic Annotation Format (LAF, [33]). It is a stand-off layered format for representing linguistic analysis at many levels. To facilitate easy processing and keeping track of the provenance of the system, we mark up the raw text with metadata such as title and publication date using the NAF format. Furthermore, every module in the pipeline adds an element to the header indicating the version of the module that was used and a timestamp as well as a layer encoding the information that was extracted from the text.

As English is the main focus of this paper, we here describe the English pipeline. More details about this pipeline, as well as about the Spanish, Dutch and Italian pipelines can be found in [34]. The information extraction processing starts with the **tokenizer** which splits the text into sentences and words. The **Part-of-Speech tagger** adds information on the type to each word, for example indicating whether it is a noun or a verb. The **Multiwords tagger** detects multiword expressions in WordNet, partly resolving ambiguity. These modules are all based on ixa-pipes [35].[28] The **Word Sense Disambiguation** module ranks the different meanings of words depending on its context. Then, the **Named Entity Recognizer** (NER) detects named entities and tries to categorize them as person, location, organization or miscellaneous names. The **Named Entity Disambiguation** (NED) module attempts to resolve the named entities against a knowledge base (in this case DBpedia), in order to link them to an entity instance. This module is followed by the **Opinion Miner**, which is aimed at detecting opinions (whether there is a positive or negative sentiment toward something), opinion holders (who has the opinion) and opinion targets (what is the opinion on). The **Syntactic Parser** aims to detect the syntactic structure of the sentence, e.g. the subject and object of the clause. The **Semantic Role Labeler** (SRL) detects the semantic arguments to predicates, e.g. who is the agent in *How far can you go with a Land Rover?*. The **Nominal Coreference Resolution** and **Event Coreference Resolution** modules compute which entities and events are the same within the document respectively. Then, the **Time and Date Recognizer** detects temporal expressions so that events can be organized on a timeline by the **Temporal Relation Detection** [36] module and causally linked by the **Causal Relation Detection** module. Finally, **Factuality Detection** is employed to determine which events have taken place or will probably take place, and which are denied or mentioned in a speculative manner.

These NLP modules were developed within the NewsReader project, except for the SRL and NED modules, which are based on third-party tools. Our modification to these tools was to make them work with the NAF format. The SRL module is based on the MATE tool [37], a state of the art system for dependency parsing and semantic role labeling. The NED module is based in DBpedia Spotlight [38], a general wikification system. We adapted DBpedia Spotlight to consider only named entity mentions and configured the tool to use the whole document as disambiguation context. Both modules will be described in more detail below.

For evaluation of the NLP modules we provide a package that helps to automate the evaluation process and ensures reproducibility of results.[29] The NLP modules have been evaluated in standard benchmark datasets and on a manually annotated gold standard on news articles (based on Wikinews) and perform comparable to or exceed the state-of-the art. A detailed analysis of the evaluation of all the modules in the four languages is out of the scope of this paper, and a complete description of the evaluation procedure and results can be found in [39]. In the remainder of this section, we highlight the performance of the NER and NED and SRL tasks as these are the key components for extracting events and participants from text. We will report their evaluation figures and detail their performance on the example sentences presented in Section 3, namely "Qatar Holding sells 10% stake in Porsche to founding families" and "Porsche family buys back 10pc stake from Qatar". These two sentences illustrate how the same event data can be packaged in different ways and represent a challenge to our software to detect the identity. When performing well, the processing should result in a single RDF representation for both.

Named entity recognition and disambiguation

In both examples, the **Named Entity Recognizer** correctly categories *Qatar Holding*, *Porsche* and *Qatar* as organization but the **Named Entity and Disambiguation** module fails to disambiguate correctly the entity mention *Qatar*. In the first example it correctly links the entity mentions to http://dbpedia.org/resource/Qatar_Investment_Authority (confidence 1.0) and http://dbpedia.org/resource/Porsche (confidence 0.99) dbpedia entries.

```
<entity id="e1" type="ORGANIZATION">
  <references>
    <!--Qatar Holding-->
      <span>
        <target id="t1" />
        <target id="t2" />
      </span>
  </references>
  <externalReferences>
```

23 http://wordpress.let.vupr.nl/naf/.

24 An online demo of the system is available at http://ixa2.si.ehu.es/nrdemo/demo.php.

25 http://ixa2.si.ehu.es/nrdemo_es/demo.php.

26 http://kyoto.let.vu.nl/nwrdemo_nl/demo.

27 http://hlt-services2.fbk.eu:8080/nwrDemo/nwr.

28 http://ixa2.si.ehu.es/ixa-pipes/.

29 The package is available at https://github.com/newsreader/evaluation.

```
10      <externalRef reference="http://dbpedia.org/resource/
              Qatar_Investment_Authority" confidence="1.0" />
11    </externalReferences>
12  </entity>
13  <entity id="e2" type="ORGANIZATION">
14    <references>
15      <!--Porsche-->
16      <span>
17        <target id="t8" />
18      </span>
19    </references>
20    <externalReferences>
21      <externalRef reference="http://dbpedia.org/resource/Porsche" confidence
              ="0.9931053" />
22      <externalRef reference="http://dbpedia.org/resource/Porsche_family"
              confidence="0.0068067894" />
23      <externalRef reference="http://dbpedia.org/resource/Porsche_911" confidence
              ="3.0875213E-5" />
24      ...
25    </externalReferences>
26  </entity>
```

In the second example it correctly disambiguates the *Porsche* entity with the http://dbpedia.org/resource/Porsche dbpedia entry, but, although understandable, it fails to link *Qatar* to the Qatar_Investment_Authority and links it to the http://dbpedia.org/resource/Qatar entry.

```
1   <entity id="e1" type="ORGANIZATION">
2     <references>
3       <!--Porsche-->
4       <span>
5         <target id="t1" />
6       </span>
7     </references>
8     <externalReferences>
9       <externalRef reference="http://dbpedia.org/resource/Porsche" confidence
              ="0.9998754" />
10      <externalRef reference="http://dbpedia.org/resource/Porsche_Design_Group
              " confidence="7.304302E-5" />
11      <externalRef reference="http://dbpedia.org/resource/Porsche_911"
              confidence="2.8215642E-5" />
12      ...
13    </externalReferences>
14  </entity>
15  <entity id="e2" type="LOCATION">
16    <references>
17      <!--Qatar-->
18      <span>
19        <target id="t8" />
20      </span>
21    </references>
22    <externalReferences>
23      <externalRef reference="http://dbpedia.org/resource/Qatar" confidence
              ="0.99899685" />
24      <externalRef reference="http://dbpedia.org/resource/Energy_in_Qatar"
              confidence="7.498067E-4" />
25      <externalRef reference="http://dbpedia.org/resource/
              Qatar_national_basketball_team" confidence="1.9039169E-4" />
26      ...
27    </externalReferences>
28  </entity>
```

When evaluating our Named Entity Recognizer on a standard benchmark dataset (CoNLL 2003 [40]) as well as on a domain specific corpus that was created within NewsReader, our system outperforms current state-of-the-art systems such as [41] and [42] with an F_1 score of 90.2 on the CoNLL 2003 dataset and an F_1 score of 68.67 on the NewsReader corpus.

For Named Entity Disambiguation, we evaluated our system against the CoNLL/AIDA [43] and TAC 2011[30] benchmarks, as well as our NewsReader corpus. On CoNLL/AIDA we achieve a precision of 79.67 and a recall of 75.94. On TAC 2010 we achieve a precision of 79.77 and recall of 60.68. On the NewsReader corpus, we achieve an F_1 score of 68.58.

Semantic role labeling

The **Semantic Role Labeler** annotates the events *sells* and *buys* with the PropBank concepts [44] predicates *sell.01* and *buy.01* respectively. In addition to these PropBank concepts, the module can add many more classes that are available in the *Predicate Matrix* (v1.1) [45].[31] The *Predicate Matrix* is a new lexical resource that integrates multiple sources of predicate information including FrameNet [28], VerbNet [46], PropBank [44], WordNet [47] and ESO [29]. Although this resource is still far from being complete, it contains many more alignments than SemLink [48].[32]

The enrichment with concepts from the *Predicate Matrix* provides semantic interoperability across different predicate models but also across different languages. For example in the representation of the first sentence in Fig. 4, the Predicate Matrix assigns *sell.01* to the mention of the predicate from PropBank, as well as external references to other sources such as the VerbNet class *give-13.1* and subclass *give-13.1-1*, the FrameNet frame *Commerce_sell*, the WordNet synset *ili-30-02244956-v* and *ili-30-02242464-v* and the ESO type *Selling*. According to the SEMANTIC ROLE LABELING system *Qatar Holding*is the *A0* of the *selling* event. According to the Predicate Matrix this argument corresponds to the VerbNet role *Agent*, FrameNet *Seller* and ESO *possession_owner_1*. Similarly, *10% stake in Porsche*is the *A1* and *to the founding families* the *A2* of the *selling* event. In the Predicate Matrix the first role corresponds to the VerbNet *Theme* or FrameNet *Goods* and the second role corresponds to the VerbNet *Recipient*, FrameNet *Buyer* or ESO *possession_owner_2*.

In Fig. 5, the **Semantic Role Labeler** assigns *buy.01* to the predicate 'buys' from PropBank, as well as external references to the VerbNet class *get-13.5.1*, the FrameNet frame *Commerce_buy*, the WordNet synset *ili-30-02207206-v* and *ili-30-02646757-v* and the ESO type *Buying*. It also annotates *Porsche family* as the *A0* of the *Buying* event. According to the Predicate Matrix this argument corresponds to the VerbNet role *Agent*, FrameNet *Buying* and ESO *possession_owner_2*. Similarly, *10pc stake* is the *A1* and *from Qatar* the *A2* of the *Buying* event. In the Predicate Matrix the first role corresponds to the VerbNet *Theme* or FrameNet *Goods* and the second role corresponds to the VerbNet *Source* and FrameNet *Means*.

In this case, the **Semantic Role Labeler** is able to extract similar semantic representations from two very different sentences. In both cases, the current English pipeline is quite close to fully recognizing the same event expressed from two different perspectives. That is, that the *Porsche family*is *Buying* *10% stake of Porsche*from *Qatar*. The elements from this representation are combined to form the semantic representation in RDF, which we will discuss in the next section. This still remains a challenge, since the labeling of the roles, the meaning of the predicates and the spans of the roles and entities still need to be matched somehow.

When evaluating the Semantic Role Labeler on the CoNLL 2009 [49] standard benchmark dataset we obtain an F_1 score of 84.74.

4.2. Cross-document event coreference

The NLP processing of documents results in the interpretation of a single textual source (i.e. document) expressed in NAF. The text is treated as a sequence of tokens that are interpreted by the various modules. The same events and the same entities can be mentioned several times within such a sequence. Information on each of these mentions can be incomplete: one sentence may make reference to the time and place of an event, while another sentence may specify the actors involved. If we consider a large set of textual sources,

30 http://www.nist.gov/tac/.

31 http://adimen.si.ehu.es/web/PredicateMatrix.

32 http://verbs.colorado.edu/semlink/ *Last accessed: 7 April 2015.*

```
<!--t3 sells : A0[t1 Qatar] A1[t4 10] A2[t9 to]-->
<predicate id="pr1">
  <!--sells-->
  <span>
    <target id="t3" />
  </span>
  <externalReferences>
    <externalRef resource="PropBank" reference="sell.01" />
    <externalRef resource="VerbNet" reference="give-13.1" />
    <externalRef resource="VerbNet" reference="give-13.1-1" />
    <externalRef resource="FrameNet" reference="Commerce_sell" />
    <externalRef resource="PropBank" reference="sell.01" />
    <externalRef resource="ESO" reference="Selling" />
    <externalRef resource="EventType" reference="contextual" />
    <externalRef resource="WordNet" reference="ili-30-02244956-v" />
    <externalRef resource="WordNet" reference="ili-30-02242464-v" />
  </externalReferences>
  <role id="rl1" semRole="A0">
    <!--Qatar Holding-->
    <span>
      <target id="t1" />
      <target id="t2" head="yes" />
    </span>
    <externalReferences>
      <externalRef resource="VerbNet" reference="give-13.1@Agent" />
      <externalRef resource="FrameNet" reference="Commerce_sell@Seller" />
      <externalRef resource="PropBank" reference="sell.01@0" />
      <externalRef resource="ESO" reference="Selling@possession-owner_1" />
    </externalReferences>
  </role>
  <role id="rl2" semRole="A1">
    <!--10 % stake in Porsche-->
    <span>
      <target id="t4" />
      <target id="t5" />
      <target id="t6" head="yes" />
      <target id="t7" />
      <target id="t8" />
    </span>
    <externalReferences>
      <externalRef resource="VerbNet" reference="give-13.1@Theme" />
      <externalRef resource="FrameNet" reference="Commerce_sell@Goods" />
      <externalRef resource="PropBank" reference="sell.01@1" />
    </externalReferences>
  </role>
  <role id="rl3" semRole="A2">
    <!--to founding families-->
    <span>
      <target id="t9" head="yes" />
      <target id="t10" />
      <target id="t11" />
    </span>
    <externalReferences>
      <externalRef resource="VerbNet" reference="give-13.1@Recipient" />
      <externalRef resource="FrameNet" reference="Commerce_sell@Buyer" />
      <externalRef resource="PropBank" reference="sell.01@2" />
      <externalRef resource="ESO" reference="Selling@possession-owner_2" />
    </externalReferences>
  </role>
</predicate>
```

Fig. 4. NAF example showing the enrichments made to the sentence *Qatar Holding sells 10% state in Porsche to founding families* through the PredicateMatrix links.

```
<!--t3 buys : A0[t1 Porsche] AM-ADV[t4 back] A1[t5 10pc] A2[t7 from ]-->
<predicate id="pr2">
  <!--buys-->
  <span>
    <target id="t3" />
  </span>
  <externalReferences>
    <externalRef resource="PropBank" reference="buy.01" />
    <externalRef resource="VerbNet" reference="get-13.5.1" />
    <externalRef resource="FrameNet" reference="Commerce_buy" />
    <externalRef resource="PropBank" reference="buy.01" />
    <externalRef resource="ESO" reference="Buying" />
    <externalRef resource="EventType" reference="contextual" />
    <externalRef resource="WordNet" reference="ili-30-02207206-v" />
    <externalRef resource="WordNet" reference="ili-30-02646757-v" />
  </externalReferences>
  <role id="rl2" semRole="A0">
    <!--Porsche family-->
    <span>
      <target id="t1" />
      <target id="t2" head="yes" />
    </span>
    <externalReferences>
      <externalRef resource="VerbNet" reference="get-13.5.1@Agent" />
      <externalRef resource="FrameNet" reference="Commerce_buy@Buyer" />
      <externalRef resource="PropBank" reference="buy.01@0" />
      <externalRef resource="ESO" reference="Buying@possession-owner_2" />
    </externalReferences>
  </role>
  <role id="rl3" semRole="AM-ADV">
    <!--back-->
    <span>
      <target id="t4" head="yes" />
    </span>
  </role>
  <role id="rl4" semRole="A1">
    <!--10pc stake-->
    <span>
      <target id="t5" />
      <target id="t6" head="yes" />
    </span>
    <externalReferences>
      <externalRef resource="VerbNet" reference="get-13.5.1@Theme" />
      <externalRef resource="FrameNet" reference="Commerce_buy@Goods" />
      <externalRef resource="PropBank" reference="buy.01@1" />
    </externalReferences>
  </role>
  <role id="rl5" semRole="A2">
    <!--from Qatar-->
    <span>
      <target id="t7" head="yes" />
      <target id="t8" />
    </span>
    <externalReferences>
      <externalRef resource="VerbNet" reference="get-13.5.1@Source" />
      <externalRef resource="FrameNet" reference="Commerce_buy@Means" />
      <externalRef resource="PropBank" reference="buy.01@2" />
    </externalReferences>
  </role>
</predicate>
```

Fig. 5. NAF example showing the enrichments made to the sentence *Porsche family buys back 10pc stake from Qatar* through the PredicateMatrix links.

we will also find many references across these sources that overlap and complement each other: today's news mentions the victims, tomorrow's news reveals who did it. To go from these mention-based representations in NAF to an instance representation in SEM, we go through a number of steps resolving co-reference across mentions (see [50] for a detailed description of our approach).

- Within-document co-reference
  - entity coreference
  - event coreference based on the same lemma or WordNet similarity score
- Cross-document co-reference
  - clustering events of the same global type within the same time constraints
  - event coreference of events within the same cluster based on overlap of participants and places

The NLP modules already identify entities in text and where possible assign a URI to each of them. The Entity Coreference Module uses the available information to decide which entities refer to the same instance but also resolves anaphoric expressions. Likewise, we can find participant relations for entities not only for cases where there is a direct reference to the entity's name, but also when he or she is mentioned differently. Each entity URI is used to represent a unique entity instance. If these entities overlap with coreference sets, all mentions within the coreference set are added to the entity instance as mentions. If we have a unique URI (e.g. `dbpedia.org/resource/Porsche`), it is used to

identify the entity, otherwise, we generate an URI from the words that refer to the entity (e.g. `data/cars/entities/Richard\_Aboulafia`). Phrases that are not detected as entities but still play an important role are represented as so-called *non-entities*. The URI is also based on the expression and distinguished from entities, e.g. `nwr:data/cars/non-entities/10+\%25+stake+in+Porsche`. Entity instances across documents can share the same URI, regardless of the fact that they are based on an external LOD resource or through a newly generated URI. They get a single representation in RDF-TRiG with GAF:DENOTEDBY links to all the places in the NAF files where they were mentioned. For each instance, we also provide the expressions detected by the NLP modules as labels.

In the remainder of this subsection, we show how the different instance representations for our Porsche–Qatar example are generated.

Entity linking

```
1  @prefix nwr: <http://www.newsreader-project.eu/> .
2  @prefix dbp: <http://dbpedia.org/resource/> .
3  @prefix gaf: <http://groundedannotationframework.org/gaf#> .
4  @prefix owl: <http://www.w3.org/2002/07/owl#> .
5  @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
6
7      dbp:Qatar
8              rdfs:label "Qatar" ;
9              gaf:denotedBy <http://www.telegraph.co.uk#char=41,46> .
10
11     dbp:Porsche
12             rdfs:label "Porsche" ;
13             gaf:denotedBy <http://www.telegraph.co.uk#char=0,7> , <http://
                   english.alarabiya.net\#char=33,40> .
14
15     dbp:Qatar_Investment_Authority
16             rdfs:label "Qatar Holding" ;
17             gaf:denotedBy <http://english.alarabiya.net#char=0,13> .
```

Lines 12–13 show that the different mentions of *Porsche* are merged into a single representation with gaf:denotedBy links to the character offsets of both sources, whereas *Qatar* but *Qatar Holding* (lines 8–9 and 16–17) are not merged due to the different DBpedia URIs that were recovered by the entity linking. Through the URI, we can access any background knowledge that is available for these entities in DBpedia. This information is not repeated here in the RDF-TRiG representation. The other concepts involved in the events of the examples can be represented as so-called non-entities, where the software was not able to map *founding family* to the *Porsche family* or *Porsche*:

```
1  @prefix nwr: <http://www.newsreader-project.eu/> .
2  @prefix gaf: <http://groundedannotationframework.org/gaf#> .
3      nwr:data/cars/non-entities/10+\%25+stake+in+Porsche
4              rdfs:label "10 \% stake in Porsche" ;
5              gaf:denotedBy <http://english.alarabiya.net#char=20,40> .
6
7      nwr:data/cars/non-entities/to+founding+families
8              rdfs:label "to founding family" ;
9              gaf:denotedBy <http://english.alarabiya.net#char=41,61> .
10
11     nwr:data/cars/non-entities/10pc+stake
12             rdfs:label "10pc stake" ;
13             gaf:denotedBy <http://www.telegraph.co.uk#char=25,35> .
```

Note that we create a domain (*nwr:data/cars/non-entities*) for each dataset processed. This means that similar phrases can become coreferential across resources but we cannot further interpret these concepts. We thus cannot tell one 10% *stake* from the other. We also see that small differences in descriptions (e.g. *10 % stake* versus *10pc*) already result in mismatches. For these concepts, we do not have any further knowledge except for the labels. It remains a challenge to further interpret these concepts that we aim to address in future work.

Dates and times

The document creation time and any normalized time-expression in NAF are represented as instances in RDF-TRiG using the owl-time vocabulary:

```
1  @prefix nwr: <http://www.newsreader-project.eu/> .
2  @prefix time: <http://www.w3.org/TR/owl-time#> .
3  @prefix gaf: <http://groundedannotationframework.org/gaf#> .
4  @prefix owl: <http://www.w3.org/2002/07/owl#> .
5  @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
6
7      <http://www.telegraph.co.uk#dct>
8              a time:Interval ;
9              rdfs:label "2014" ;
10             gaf:denotedBy <http://www.telegraph.co.uk#dctm> , <http://english.
                   alarabiya.net#dctm> ;
11             time:inDateTime nwr:time/2014> .
12
13     <http://www.telegraph.co.uk#>
14             a time:Interval ;
15             rdfs:label "2014" ;
16             gaf:denotedBy <http://www.telegraph.co.uk#dctm> , <http://english.
                   alarabiya.net#dctm> ;
17             time:inDateTime nwr:time/2014 .
18
19     <http://www.newsreader-project.eu/time/2014>
20             a time:DateTimeDescription ;
21             time:unitType time:unitDay ;
22             time:year "2014"^^<http://www.w3.org/2001/XMLSchema#gYear> .
23
24     wikinews:Airbus_parent_EADS_wins_£13_billion_UK_RAF_airtanker_contract.
           en#tmx5
25             a time:Interval ;
26             rdfs:label "month" ;
27             gaf:denotedBy
28             wikinews:
                   Airbus_parent_EADS_wins_£13_billion_UK_RAF_airtanker_contract
                   .en#char=1014,1018 ,
29             wikinews:
                   Airbus_parent_EADS_wins_£13_billion_UK_RAF_airtanker_contract
                   .en#char=1019,1024 ;
30             time:inDateTime nwr:time/200802 .
31
32     nwr:time/200802
33             a time:DateTimeDescription ;
34             time:month "--02"^^<http://www.w3.org/2001/XMLSchema#
                   gMonth> ;
35             time:unitType time:unitDay ;
36             time:year "2008"^^<http://www.w3.org/2001/XMLSchema#gYear> .
```

In these examples, *dct* on line 7 stands for **document-creation-time** and *dctm* on lines 10 and 16 for a mention of a **document-creation-time** in a source. We see here that the document-creation-time of both our examples gets a distinct URI but refer to the same time:inDateTime value. In addition, we show a representation for a time expression *month* that occurred in a Wikinews article and has been normalized to another time:inDateTime value: *200802*. The time:inDateTime values get separate representations according to owl-time to allow for reasoning over time.

Event linking

As for entities and time, we need to create instances for events. In the case of events however, we (usually) do not have an external URI. Events are less tangible and establishing identity across mentions is a difficult task. We follow an approach that takes the compositionally of events as a starting point [51]. The compositionality principle dictates that events are not just defined by the action (or the relation or property) but also by the time, place and participants. For that, we use an algorithm that compares events for all these properties [52].

We first establish coreference relations across events within the same document. As a starting point, we take the predicates of the Semantic Role Label layer and chain all predicates with the same lemma or that have a similarity score in WordNet above 2.0 [53] into a single coreference set. This represents an instance of an event within the same document. We assume that all the participant information and time anchors are spread over different mentions

of the same event within the document. We then create a so-called Composite Event Object (CEO) by aggregating the participants and time-expressions from all the coreferential mentions within the same source. The participants of the events are based on the RDF instance representation of the entities detected within the same source and therefore can have different mentions across the document as well. Their mentions within the document are matched with the span of the roles of the predicates to determine that the entity plays a role in the event. The final CEOs are SEM objects with RDF instance representations for events, participating entities (and non-entities) and their time anchoring. We store the CEOs for each NAF file in time-description folders based on the time-anchoring. A single NAF file thus can have multiple CEOs that are stored in different time-description folders. Note that events without an explicit normalized time-anchor are linked to the document-creation-time.

In a second step, we compare all the CEOs from the same time-description folder to establish cross-document coreference. We already know that these events have the same time-anchor. CEOs are matched according to the following criteria:

1. The action or process of two CEOs should have the same lemma as a label or the same WordNet reference as a subClassOf relation;
2. They should share at least one actor-participant, where we match participants though their URI;
3. If both CEOs have a place–participant, the URIs of at least one place–participant should match;

The matching of the CEOs can easily be adapted to get looser or stricter matches. For example, we require for speech-act type of events that the actor-participant with the *Speaker role* should match, whereas for cross-lingual comparison of events, we allow for looser WordNet matches of events instead of lemma matches.

If there is a match, we merge the information of one CEO into the information of another CEO, where we keep the unique URI of the first CEO as the identifier of the event instance. When we merge the information, participants and time relations are adapted to the shared URI as a subject. Since one CEO can partially differ from the other, we aggregate information across the CEOs. When they contain the same information, we just update the mentions of the relations. We iterate recursively over all the CEOs that need to be compared until no new matches arise. This results in chaining of CEOs for cases where a merge of CEOs creates the condition for another CEO to match.

For the *buy/sell* example discussed before, our system does not generate a match because both the labels and the WordNet synsets differ. Likewise, we generate the following instance representations of the events:

```
1  @prefix eso:<http://www.newsreader-project.eu/domain-ontology#> .
2  @prefix gaf: <http://groundedannotationframework.org/gaf#> .
3  @prefix nwr: <http://www.newsreader-project.eu/> .
4  @prefix owl: <http://www.w3.org/2002/07/owl#> .
5  @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
6  @prefix sem: <http://semanticweb.cs.vu.nl/2009/11/sem/> .
7  @prefix fn: <http://www.newsreader-project.eu/ontologies/framenet/> .
8
9  <http://www.telegraph.co.uk#ev2>
10   a sem:Event , fn:Commerce_buy , eso:Buying ,
11                                  nwr:ontologies/wordnet3.0/ili-30-02207206-v , nwr:
                                        ontologies/wordnet3.0/ili-30-02646757-v ;
12   rdfs:label "buy" ;
13   gaf:denotedBy <http://www.telegraph.co.uk#char=15,19> .
14
15 <http://english.alarabiya.net#ev1>
16   a sem:Event , fn:Commerce_sell , eso:Selling ,
17                                  nwr:ontologies/wordnet3.0/ili-30-02244956-v ,
                                        nwr:ontologies/wordnet3.0/ili
                                        -30-02242464-v;
18   rdfs:label "sell" ;
19   gaf:denotedBy <http://english.alarabiya.net#char=14,19> .
```

The representation is similar to the entity representation except that for each event instance we generate its URI and we type it according to classes of the ontologies we adopted, to allow for reasoning over the events. Note that a more abstract matching of *sell* and *buy*, e.g. through the FrameNet hierarchy, would result in a merge.

The next example is taken from the automotive corpus and shows that closer events as expressed by mentions of *get*, *purchase* and *buy* are merged by the system into a single event instance (line 9) with various mentions across two documents through the gaf:denotedBy predicate (line 12):

```
1  @prefix eso: <http://www.newsreader-project.eu/domain-ontology#> .
2  @prefix gaf: <http://groundedannotationframework.org/gaf#> .
3  @prefix nwr: <http://www.newsreader-project.eu/> .
4  @prefix owl: <http://www.w3.org/2002/07/owl#> .
5  @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
6  @prefix sem: <http://semanticweb.cs.vu.nl/2009/11/sem/> .
7  @prefix fn: <http://www.newsreader-project.eu/ontologies/framenet/> .
8
9  nwr:data/cars/2003/01/01/4HYF-CF10-TX4X-W2GG.xml#ev19
10           a sem:Event , fn:Getting , eso:Getting , eso:Buying , fn:Grasp , fn:
                   Commerce_buy, nwr:ontologies/wordnet3.0/ili
                   -30-01433294-v , nwr:ontologies/wordnet3.0/ili
                   -30-02210855-v , nwr:ontologies/wordnet3.0/ili
                   -30-02208265-v , nwr:ontologies/wordnet3.0/ili
                   -30-02359340-v , nwr:ontologies/wordnet3.0/ili
                   -30-02207206-v , nwr:ontologies/wordnet3.0/ili
                   -30-02646757-v ;
11           rdfs:label "get" , "buy" , "purchase" ;
12           gaf:denotedBy nwr:data/cars/2003/01/01/4HYF-CF10-TX4X-W2GG.
                   xml#char=508,511 , nwr:data/cars/2003/01/01/4HYF-CF10-
                   TX4X-W2GG.xml#char=602,608 , nwr:data/cars/2003/01/01/4
                   HYF-CF10-TX4X-W2GG.xml#char=714,722 , nwr:data/cars
                   /2003/01/01/4RTF-FY80-TX4X-W14X.xml#char=506,509 ,
                   nwr:data/cars/2003/01/01/4RTF-FY80-TX4X-W14X.xml#
                   char=600,606 , nwr:data/cars/2003/01/01/4RTF-FY80-TX4X-
                   W14X.xml#char=712,720 .
```

Once the identity of events is established, we output the relations between event instances and participants and event instances and their time anchors. The triples exploit a selection of the role relations from NAF as properties with the event instances as subjects and the entity instances as objects. The roles represent different levels of abstraction (SEM–PropBank–ESO–FrameNet) that can be exploited in the reasoning. Lines 11–42, for example, correspond to the example given in Fig. 4, where on line 18 Qatar Holding is identified as having the A0 role, which is here described on line 21 by the triple `<http://english.alarabiya.net#ev1> nwr:ontologies/propbank/A0 dbp:/Qatar_Investment_Authority`

```
1  @prefix dbp: <http://dbpedia.org/resource/> .
2  @prefix eso: <http://www.newsreader-project.eu/domain-ontology#> .
3  @prefix gaf: <http://groundedannotationframework.org/gaf#> .
4  @prefix nwr: <http://www.newsreader-project.eu/> .
5  @prefix owl: <http://www.w3.org/2002/07/owl#> .
6  @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
7  @prefix sem: <http://semanticweb.cs.vu.nl/2009/11/sem/> .
8  @prefix fn: <http://www.newsreader-project.eu/ontologies/framenet/> .
9
10
11 <http://www.telegraph.co.uk#pr1,rl1> {
12     <http://www.telegraph.co.uk#ev1>
13           sem:hasActor dbp:/Porsche ;
14           nwr:ontologies/propbank/A2 dbp:/Porsche .
15 }
16
17 <http://english.alarabiya.net#pr1,rl1> {
18     <http://english.alarabiya.net#ev1>
19           sem:hasActor dbp:/Qatar_Investment_Authority ;
20           eso:possession-owner_1 dbp:/Qatar_Investment_Authority ;
21           nwr:ontologies/framenet/Commerce_sell@Seller dbp:/
                   Qatar_Investment_Authority ;
22           nwr:ontologies/propbank/A0 dbp:/Qatar_Investment_Authority .
23 }
24
25 <http://english.alarabiya.net#pr1,rl2> {
26     <http://english.alarabiya.net#ev1>
```

```
27              sem:hasActor nwr:data/cars/non−entities/10+\%25+stake+in+Porsche
                    ;
28              nwr:ontologies/framenet/Commerce_sell@Goods
29                  nwr:data/cars/non−entities/10+\%25+stake+in+Porsche ;
30              nwr:ontologies/propbank/A1
31                  nwr:data/cars/non−entities/10+\%25+stake+in+Porsche .
32  }
33
34  <http://english.alarabiya.net#pr1,rl3> {
35      <http://english.alarabiya.net#ev1>
36              sem:hasActor nwr:data/cars/non−entities/to+founding+families ;
37              eso:possession−owner_2 nwr:data/cars/non−entities/to+founding+
                    families ;
38              nwr:ontologies/framenet/Commerce_sell@Buyer
39                  nwr:data/cars/non−entities/to+founding+families ;
40              nwr:ontologies/propbank/A2
41                  nwr:data/cars/non−entities/to+founding+families .
42  }
43
44  <http://www.telegraph.co.uk#dt1> {
45      <http://www.telegraph.co.uk#ev1>
46              sem:hasTime <http://www.telegraph.co.uk#dct> .
47  }
48
49  <http://www.telegraph.co.uk#pr2,rl2> {
50      <http://www.telegraph.co.uk#ev2>
51              sem:hasActor   dbp:Porsche ;
52              eso:possession−owner_2  dbp:Porsche ;
53              nwr:ontologies/framenet/Commerce_buy@Buyer dbp:Porsche ;
54              nwr:ontologies/propbank/A0>       dbp:Porsche .
55  }
56
57  <http://www.telegraph.co.uk#pr2,rl5> {
58      <http://www.telegraph.co.uk#ev2>
59              sem:hasActor dbp:/Qatar ;
60              nwr:ontologies/framenet/Commerce_buy@Means dbp:/Qatar ;
61              nwr:ontologies/propbank/A2 dbp:/Qatar .
62  }
63
64  <http://www.telegraph.co.uk#pr2,rl4> {
65      <http://www.telegraph.co.uk#ev2>
66              sem:hasActor nwr:data/cars/non−entities/10pc+stake ;
67              nwr:ontologies/framenet/Commerce_buy@Goods nwr:data/cars/non−
                    entities/10pc+stake ;
68              nwr:ontologies/propbank/A1 nwr:data/cars/non−entities/10pc+stake .
69  }
```

The representation shows that the identification of entities and events is crucial to the density of the representation that we can achieve. Following a stricter approach leads to distinct representations close to the individual mentions, whereas a looser approach will result in the merge and aggregation of instances and their relations. In our module, we can vary the degree and methods of similarity obtained for each event component. For the event mentions, we can choose overlap of words used to make reference, overlap of WordNet synsets, similarity of synsets within the WordNet graph, similarity according to other ontologies used (FrameNet, ESO), or similarity across to word-embeddings. We can use combinations of these measures and vary the thresholds. In the case of participants, we can use words overlap, URI identify but also meronymy relations between for example locations or temporal expressions. Furthermore, the API allows you to choose the number and types of event–participant relations that need to be matched. Requiring very rich and specific FrameNet roles to match across CEOs will generate high precision output but also few merges. We are investigating the optimal level of granularity through an empirical and application-driven evaluation. Optimal settings may vary across different genres of texts (having more or less variation in expression, less or more metaphoric meanings) or different topics (involving abstract or concrete events).

Another important aspect of the representation is that the relations are all embedded in named-graphs, see Section 3. By creating URIs for the stated relation, we can express various properties of the relations. This is for example used to express the provenance of the relation as shown below where the system generated the GAF links to the mentions of the relations identified by named-graphs:

```
1  <http://www.newsreader−project.eu/provenance> {
2      <http://www.telegraph.co.uk#pr2,rl5>
3              gaf:denotedBy <http://www.telegraph.co.uk#char=15,46> .
4      <http://www.telegraph.co.uk#pr3,rl6>
5              gaf:denotedBy <http://www.telegraph.co.uk#char=0,35> , <http://
                   english.alarabiya.net#char=24,40> .
6      <http://english.alarabiya.net#dt1>
7              gaf:denotedBy <http://english.alarabiya.net#char=14,19> .
8      <http://english.alarabiya.net#pr1,rl1>
9              gaf:denotedBy <http://english.alarabiya.net#char=0,19> .
10     <http://english.alarabiya.net#pr1,rl2>
11             gaf:denotedBy <http://english.alarabiya.net#char=14,40> .
12     <http://english.alarabiya.net#pr1,rl3>
13             gaf:denotedBy <http://english.alarabiya.net#char=14,61> .
14 }
```

In addition to these GAF links, the system can also generate provenance links to authors and owners of sources.

The RDF representations of the text interpretation are loaded into the KnowledgeStore (see Section 4.4), which allows for storing and querying of the ECKG.

4.3. Cross-lingual event extraction

As mentioned in the previous subsections, our processing results in language neutral representations for instances of entities and normalized time expressions. The entity linking for Dutch and Spanish, for example, generates URIs for both the language specific DBpedia resource and the English DBpedia resource. Also the roles assigned to predicates have been harmonized across the different languages. The only thing left is the representation of the event. For this, we can rely on the WordNet synset identifiers that are shared across English, Dutch and Spanish. Our software can thus take NAF files produced from English, Spanish and Dutch text and carry out the same analysis as explained in Section 4.2 for cross-lingual event extraction. In that case, events with the same time-anchors are matched with respect to their WordNet synsets and the participant URIs are merged if there is sufficient matching. Below, we show an example from the Wikinews dataset for which we translated the English articles and processed them by the respective language-processors. We then applied the cross-document interpretation to the NAF files for the original English texts and the Dutch and Spanish translations to generate a unified RDF output.

```
1  nwr:data/wikinews/1380_World_largest_passenger_airliner_makes_first_flight#
       evt1
2      a
3              sem:Event, fn:Bringing, fn:Motion, fn:Operate_vehicle,
4              fn:Ride_vehicle, fn:Self_motion;
5      rdfs:label
6              "volar", "fly", "verlopen", "vliegen" ;
7      gaf:denotedBy wikinews:english_mention#char=202,208>, wikinews:
               english_mention##char=577,580>, wikinews:dutch_mention##char
               =1034,1042>, wikinews:dutch_mention#char=643,650>, wikinews:
               dutch_mention#char=499,505>, wikinews:dutch_mention#char
               =224,230>, wikinews:spanish_mention#char=218,224>, wikinews:
               spanish_mention#char=577,583> ;
8      sem:hasTime nwrtime:20070391;
9      sem:hasPlace dbp:Frankfurt_Airport, dbp:Chicago , dbp:
               Los_Angeles_International_Airport, nwr:data/airbus/entities/
               Chicago_via_New_York;
10     sem:hasActor dbp:Airbus_A380, nwr:data/airbus/entities/Los_Angeles_LAX ,
               dbp:Frankfurt, nwr:data/airbus/entities/A380−machines.
```

In this example, line 6 shows the event mention labels from the different languages, and the gaf:denotedBy links on line 7 show where the different mentions originate from. Similarly, also the participants and locations were merged as is indicated by lines 9 and 10.

4.4. Implementation

Building ECKGs from large quantities of text, such as millions of news articles (see Section 5.3), requires designing solutions that are able to run distributed programs across a large cluster of machines.

Table 1
Statistics of the Event-Centric Knowledge Graphs built in the four scenarios.

	WikiNews	**FIFA WorldCup**	**Cars (Ver. 2)**	**Airbus Corpus**
Topic	General news	Sport, football	Automotive industry	Airbus A380
News providers	http://en.wikinews.org	LexisNexis, BBC The Guardian	LexisNexis	http://wikinews.org/
Language	English	English	English	English, Dutch, Spanish
Populated in	February 2015	May 2014	December 2014	February 2015
News articles	18,510	*212,258*	1,259,748	90 (30 EN, 30 NL, 30 ES)
Mentions	2,629,176	76,165,114	205,114,711	6,415
Events	624,439	9,387,356	25,156,574	2,574
Entities	45,592	858,982	1,967,150	934
Persons	*19,677*	*403,021*	*729,797*	*71*
in DBpedia	9,744	40,511	128,183	19
Organizations	*15,559*	*431,232*	*947,262*	*806*
in DBpedia	6,317	15,984	60,547	774
Locations	*10,356*	*24,729*	*290,091*	*57*
in DBpedia	7,773	16.372	88,695	53
Triples	105,675,519	240,731,408	535,035,576	95,994,233
from Mentions	*9,700,585*	*136,135,841*	*439,060,642*	*19,299*
from DBpedia	*95,974,934*	*104,595,567*	*95,974,934*	*95,974,934*
distilled from	DBpedia 2014	DBpedia 3.9	DBpedia 2014	DBpedia 2014

To process the required quantity of news articles in a timely manner, we designed and implemented a NLP pipeline that scales up with the number of documents through parallelization.

The processing chain is meant to be run and deployed into a cluster of machines. The current implementation relies on virtual machines (VM) that contain all the required modules to analyze the documents. Virtualization is a widespread practice that increases the server utilization and addresses the variety of dependencies and installation requirements. Besides, it is a 'de-facto' standard on cloud-computing solutions, which offer the possibility of installing many copies of the virtual machines on commodity servers. In our architecture, all NLP modules along with their dependencies are installed into a single VM, which is then copied and deployed into clusters of computers.

We use Apache Storm[33] to integrate and orchestrate the NLP modules of the processing chain.[34] Storm is an open source, general-purpose, distributed, scalable and partially fault-tolerant platform for developing and running distributed programs that process continuous streams of data. Storm allows setting scalable clusters with high availability using commodity hardware and minimizes latency by supporting local memory reads and avoiding disk I/O bottlenecks. The main abstraction structure of Storm is the *topology*, which describes the processing node that each message passes through. The topology is represented as a graph where nodes are processing components, while edges represent the messages sent between them.

Documents are sent to a single VM containing all the NLP processing modules, which are executed one module after another. Thus, the complete analysis for each document is performed inside a single VM. Each module receives a NAF document, creates annotations on top of it, and passes the enriched NAF to the next module. Partially analyzed NAF documents are stored and distributed among the cluster machines using a NoSQL database (mongoDB[35]). The current version of the VM containing our pipeline is available from http://bit.ly/1hHVwVc. All modules are freely available through https://github.com/newsreader. We have also developed a set of scripts with the aim of automatically create a fully working cluster for distributed NLP processing. We call these scripts "VM from scratch", as they create and configure the required virtual machines. The scripts are available from github repository https://github.com/ixa-ehu/vmc-from-scratch.

The source news, the NAF files, and the RDF content resulting from the conversion to SEM, are all uploaded into the KnowledgeStore[36] [54,55], a scalable, fault-tolerant, and Semantic Web grounded storage system that, thanks to a tight interlinking between the news, the structured content of the NAF documents, and the corresponding RDF entities and facts, enables to jointly store, manage, retrieve, and semantically query, both structured and unstructured content. We processed 1.26 million articles about the Global Automotive Industry (described in Section 5.3) in approximately 11 days. The cross-document coreference and the conversion to SEM 9 days.

The resulting ECKGs are presented in the next section.

5. Knowledge graphs

In this section, we present the four different Event-Centric Knowledge Graphs we have generated using a variety of text sources. Table 1 presents the overall statistics of each ECKG. In the remainder of this section, the motivations behind, and peculiarities and possible uses through a set of queries of each ECKG are described. The queries were chosen to illustrate the advantages of an event-centric approach, compared to entity- or document-centric approaches. Note that each ECKG also contains a subset of RDF triples obtained from DBpedia: this content complements the information automatically extracted from news articles with background knowledge facts (e.g., entity types, general facts about entities), to favor the exploitation of the ECKG in applications.

5.1. Wikinews

Wikinews[37] is a free multilingual open news source operated and supported by the Wikimedia foundation. We chose to use this source as it enables us to link entities and events across different language as well as its broad coverage. For English we cleaned the Wikinews dump from 16 January 2014. This resulted in 18,510 news articles which we then processed using the pipeline described in Section 4. A summary of the extracted content is reported in Table 1. The original news corpus, the intermediate results of the processing, as well as the resulting ECKG extracted can be downloaded[38] or directly access via a dedicated KnowledgeStore installation.[39]

33 https://storm.apache.org/ *Last accessed: 7 April 2015.*
34 http://storm.incubator.apache.org/ *Last accessed: 7 April 2015.*
35 https://www.mongodb.org *Last accessed: 7 April 2015.*
36 http://knowledgestore.fbk.eu.
37 http://www.wikinews.org *Last accessed: 7 April 2015.*
38 http://www.newsreader-project.eu/results/data/.
39 http://knowledgestore2.fbk.eu/nwr/wikinews/ui—a demonstration video explaining how to use the KnowledgeStore for accessing the ECKG can be accessed at https://youtu.be/YVOQaljLta4.

entity	frequency
dbpedia:John_McCain	59
dbpedia:Hillary_Rodham_Clinton	41
dbpedia:Democratic_Party_(United_States)	31
dbpedia:United_States	24
dbpedia:Republican_Party_(United_States)	21
dbpedia:George_W._Bush	16
dbpedia:Federal_government_of_the_United_States	16
dbpedia:United_States_Congress	15
dbpedia:Mitt_Romney	14
dbpedia:United_States_Armed_Forces	13

Fig. 6. Top 10 entities co-occurring with Barack Obama.

One thing that is difficult to query for in the raw text or in KGs such as DBpedia, but can be queried easily in our ECKG is: which entities are most often participating in events where President Barack Obama is also involved?[40] Here we find that, out of the mentions in total of Barack Obama in our corpus together with another entity, John McCain co-participates 59 times, Hillary Rodham Clinton 41 times, Democratic Party co-occurs 31 times, United States 24 times, and The Republican Party 21 (see Fig. 6). Such information can be useful to information specialists interested in the interactions between different players in a domain.

In total, Barack Obama was involved in 1292 events mentioned in the corpus: he was mostly involved in statement events e.g. giving speeches (235 times), text creation e.g., signing bills (127 times) and requesting or asking for something (66 times).

Another interesting example we found in the dataset is the mention of Apple Corps Ltd, the multimedia corporation founded by The Beatles, losing the court case against Apple Inc., the computer company.[41] The two mentions of Apple companies are correctly disambiguated and linked to the appropriate DBpedia resources in our dataset. An excerpt of the ECKG content about this particular event is shown in Fig. 7 (as accessed via the publicly available KnowledgeStore installation).

5.2. FIFA world cup

For a public Hackathon in June 2014,[42] we built a knowledge graph around a popular topic in that period, namely the FIFA World Cup. LexisNexis[43] provided us with 200,000 articles about football, dating from January 2004 to April 2014. This data was complemented by content scraped with permission from the BBC and Guardian websites.

The articles were processed using an initial version of the processing pipeline described in Section 4. The resulting ECKGs are split into three TRiG files, one containing the contextual events,[44] one containing the grammatical events,[45] and one containing the source events.[46] Source events are so-called speech-acts (*say*, *announce*) and cognitive events (*think*, *believe*) that mostly introduce sources of information in the news rather than the events that took place in the world. Grammatical events are events such *stop*, *start*, *take place* that express properties of other events but do not refer to distinct events themselves. All remainder events are contextual event and are assumed to refer to what happened in the world about which the news reports. Due to licensing issues, we are not authorized to make the original news article from which the ECKGs were extracted available.

For information professionals, it is often interesting to investigate the social network of an entity. A structured version of the text corpus allows for network visualizations such as the one presented in Fig. 8. In this visualization, Sepp Blatter and David Beckham are represented by the blue nodes at the center as these are the entities the visualization focus. The people that have interacted with both Blatter and Beckham are represented in green, with the size of the bubble indicating in how many events these persons co-occurred in the ECKG. The people that only interacted with one of our core entities are represented by pink nodes on the outside of the network. By clicking on a node, one can bring up some network statistics on the chosen entity at hand. The fact that this visualization could be created by a developer in one afternoon, shows the power of structured data over raw keyword search. The visualization can be explored interactively at: http://stevenmaude.github.io/newsreader-network-vis/#.

5.3. Global automotive industry

Building an ECKG could also help to investigate large and complex global developments over many years, e.g spanning the financial and economic crisis of the last decade. For this purpose, we constructed an ECKG from financial and economic news of the last ten years. In particular, we focused on the global automotive industry domain: this was one of the economic sectors heavily hit by the crisis, and it is also a complex global market involving not only industry but also governments and well-organized labor unions. This makes the ECKG an interesting playground for researchers from different fields, as well as information specialists within companies or other organizations. The documents were selected by first performing a query on the LexisNexis News database by selecting specific keywords related to car companies (“Alfa Romeo”, “Aston Martin”, “BMW”, etc.[47]) spanning the time period between 2001 and 2013. This initial query retrieved around 6 million news documents, which are further filtered by considering only documents that are between 1000 and 3000 characters. Through analysis of our corpus, we found that this is about half to two times the length of a standard news article. This will filter out really short and really long articles such as biographies, which also make up a large part of the original dataset. As a result, we got 126 million documents from 3111 different sources. All documents were converted from the original NITF format[48] that is used at LexisNexis to NAF and processed with the processing pipeline described in Section 4. The ECKG is split into three sets of TRiG files, one containing the source events,[49] the grammatical events,[50] and the contextual events.[51] Note that the ECKG does not include the content of the source news documents as we are not authorized to redistribute them due to licensing issues

40 Note that we are interested in entities co-participating in an event rather than in entities co-occurring in the same document, with the first set being a subset of the second.

41 http://en.wikinews.org/wiki/Beatles'_Apple_Corps_sues_Apple_Computer *Last accessed: 7 April 2015.*

42 http://www.eventbrite.com/e/kick-off-newsreader-and-hack-100000-world-cup-articles-tickets-2848605255 *Last accessed: 7 April 2015.*

43 http://www.lexisnexis.nl/.

44 Available for download from: http://kyoto.let.vu.nl/worldcup/othertrigs.tgz (2.6 GB).

45 Available for download from: http://kyoto.let.vu.nl/worldcup/grammaticaltrigs.tgz (498 MB).

46 Available for download from: http://kyoto.let.vu.nl/worldcup/speechtrigs.tgz (396 MB).

47 The full query spans almost one a4 and is described in [56].

48 http://www.iptc.org/site/News_Exchange_Formats/NITF/.

49 Available for download from: http://kyoto.let.vu.nl/cars/trigs/source_trigs.tgz.

50 Available for download from: http://kyoto.let.vu.nl/cars/grammatical_trigs.tgz (745 MB).

51 Available for download from: http://kyoto.let.vu.nl/cars/context_trigs.tgz (4.3 GB).

Resources mentioning the entity (1 out of 1) - 1 mentions total

resource ID	dct:created	dct:title	# mentions
<../Apple_Corps_loses_court_case_against_Apple_Computer>	2006-05-08T00:00:00Z	Apple Corps loses court case against Apple Computer	1

Triples describing the entity (11 out of 11)

subject	predicate	object
<../Apple_Corps_loses_court_case_against_Apple_Computer#ev11>	rdf:type	sem:Event
<../Apple_Corps_loses_court_case_against_Apple_Computer#ev11>	rdfs:label	lose
<../Apple_Corps_loses_court_case_against_Apple_Computer#ev11>	sem:hasTime	<../Apple_Corps_loses_court_case_against_Apple_Computer#tmx1>
<../Apple_Corps_loses_court_case_against_Apple_Computer#ev11>	sem:hasActor	dbpedia:Apple_Inc.
<../Apple_Corps_loses_court_case_against_Apple_Computer#ev11>	sem:hasActor	dbpedia:Apple_Corps
<../Apple_Corps_loses_court_case_against_Apple_Computer#ev11>	sem:hasPlace	dbpedia:High_Court_of_Justice
<../Apple_Corps_loses_court_case_against_Apple_Computer#ev11>	<../A0>	dbpedia:Apple_Corps
<../Apple_Corps_loses_court_case_against_Apple_Computer#ev11>	<../A1>	dbpedia:Apple_Inc.
<../Apple_Corps_loses_court_case_against_Apple_Computer#ev11>	<../AM-LOC>	dbpedia:High_Court_of_Justice
<../Apple_Corps_loses_court_case_against_Apple_Computer#ev11>	<../hasFactBankValue>	CT+
<../Apple_Corps_loses_court_case_against_Apple_Computer#ev11>	gaf:denotedBy	<../Apple_Corps_loses_court_case_against_Apple_Computer#char=31,35>

Fig. 7. Apple Corps loses court case against Apple Computer.

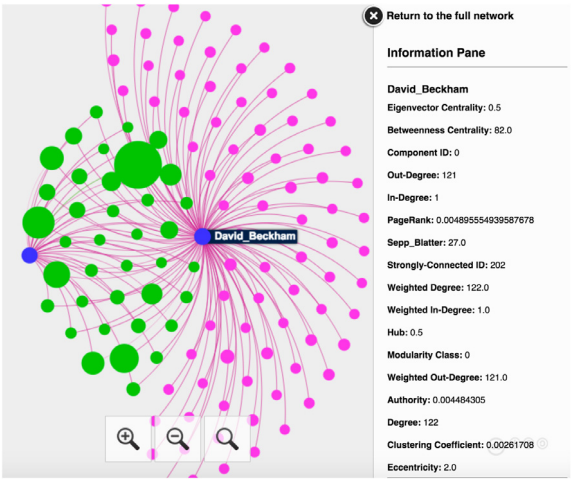

Fig. 8. Visualization of the network of interactions involving Sepp Blatter and David Beckham in the FIFA World Cup ECKG. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(although the articles can be looked up in the LexisNexis database by users with a subscription).

Despite the fact that this ECKG has not yet been derived from all relevant articles about the global automotive industry, it already enables us to answer some questions about the domain that cannot be answered through simple keyword search. We can for example retrieve the different events and the locations and times they took place at to obtain insights into the localities of the global automotive industry. If we visualize these results on a map ordered by year as shown in Fig. 9, we can see that there is little mention of India in our dataset in 2003, but in 2008 it is an important player in the Global Automotive Industry. When further looking into these results, we find that in 2008 Tata Motors, an Indian car manufacturer, launched the Nano, an affordable car and bought Jaguar Land Rover from Ford.

A more in-depth example of how the ECKG can be exploited to create a reconstruction around an entity is presented in Section 7.

5.4. Airbus corpus

The Airbus ECKG is the result of a preliminary experiment we performed to show how our tools and methodology for constructing ECKGs work with a cross-lingual corpus. We remark that, to the best of our knowledge, no state of the art tool is capable of producing such kind of event-centric structured representation of the content of news articles in different languages.

We created the corpus starting from English Wikinews articles. We selected 30 articles, about a certain topic (the Airbus A380), and spread over a period of five years. These articles were manually translated in Spanish and Dutch by professionals, obtaining a cross-lingual corpus of 90 news articles (30 for each language). In performing the translation, sentences between the translated and original news article were also aligned (to favor further comparisons). We applied the single-document processing pipelines (Section 4.1) for English, Spanish and Dutch to all 30 news articles for each language, and then we performed the cross-document cross-lingual processing (Sections 4.2 and 4.3) on the resulting NLP annotated files.

Since the documents of the corpus are manually translated, we expect the same content to be present across the three languages. Thus, if our cross-lingual processing methodology to build ECKGs is fully interoperable and generates the same quality across the languages, we expect to obtain exactly the same representation of events across the different languages (e.g., one single event, having mentions from English, Dutch, and Spanish articles). Indeed, our processing is able to produce such kind of output in several cases: an excerpt of event (flying) co-referred from news articles in different languages is shown in Fig. 10.

A closer look at the ECKG produced actually revealed that the extracted Spanish events cover approx. 26% of the events extracted from the English news articles, while for Dutch the coverage is approx. 11%. This is due to several reasons: e.g. different approaches for the language specific pipelines, different coverage of the resources (e.g. the Spanish wordnet is expanded from the English WordNet while the Dutch wordnet is built independently). Further note that we only measure to what extent we obtain the same information as from the English processing. If there is a difference this does not necessarily imply that it is wrong. Despite the fact that current processing, on this particular cross-lingual event extraction aspects, has a fair amount room for improvement, we remark that without cross-lingual capabilities, the resulting knowledge graph would consist of three disjoint sets of events, one for each language, even though the content of the news articles from which these events were extracted is exactly the same in the three languages.

6. ECKG quality evaluation

In this section, we present a first evaluation of the quality of the ECKG built with our approach. Due to the lack of a proper gold

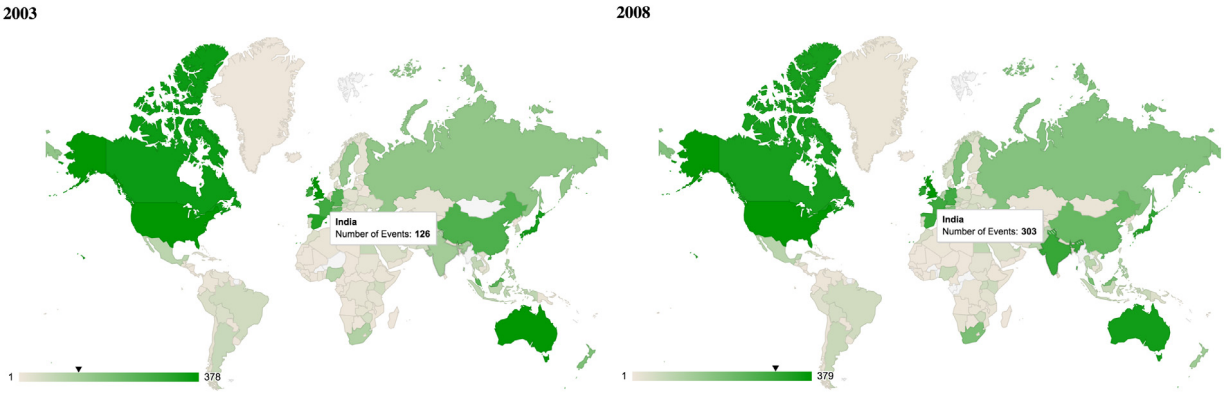

Fig. 9. Overview of number of events taking place in particular locations (from Global Automotive Industry ECKG).

predicate	object
rdf:type	sem:Event
rdf:type	eso:Motion
rdf:type	eso:Translocation
rdf:type	eso:Transportation
rdfs:label	flight
rdfs:label	fly
rdfs:label	vliegen
rdfs:label	volar
gaf:denotedBy	<../A380_makes_maiden_flight_to_US.en#char=174,180>
gaf:denotedBy	<../A380_makes_maiden_flight_to_US.en#char=19,25>
gaf:denotedBy	<../A380_makes_maiden_flight_to_US.en#char=202,208>
gaf:denotedBy	<../A380_makes_maiden_flight_to_US.en#char=566,572>
gaf:denotedBy	<../1816_Airbus_wins_Qatar_Airways_order_worth_15bn_dutch_utf8.nl#char=626,633>
gaf:denotedBy	<../6475_Singapore_Airlines_to_be_compensated_for_A380_delays_dutch_utf8.nl#char=1150,1157>
gaf:denotedBy	<../8658_Aer_Lingus_buys_twelve_new_long-haul_Airbus_jets_dutch_utf8.nl#char=1550,1558>
gaf:denotedBy	<../8658_Aer_Lingus_buys_twelve_new_long-haul_Airbus_jets_dutch_utf8.nl#char=1731,1737>
gaf:denotedBy	<../8658_Aer_Lingus_buys_twelve_new_long-haul_Airbus_jets_dutch_utf8.nl#char=2886,2893>
gaf:denotedBy	<../10021_First_Airbus.txt.es#char=254,258>
gaf:denotedBy	<../10566_Indonesia_transport.txt.es#char=1211,1216>
gaf:denotedBy	<../10566_Indonesia_transport.txt.es#char=1302,1307>
gaf:denotedBy	<../10566_Indonesia_transport.txt.es#char=1946,1951>

Fig. 10. Example of event co-referred from news articles in different languages.

standard to compare with, we relied on human judgment for the triples describing some randomly sampled events of the graph. A similar approach was applied to evaluate YAGO2 [57], a large knowledge graph automatically extracted from Wikipedia pages.

We conducted an evaluation of the ECKG extracted from a corpus consisting of 120 news document from Wikinews, performing an assessment of the event-related triples of the ECKG. We sampled 100 events from the resulting ECKG, splitting them in 4 groups of 25 events each. For each event, we retrieved all its triples, obtaining 4 subgraphs (labeled: S_1, S_2, S_3, S_4) of approx. 260 triples each. Each subgraph was submitted to a pair of human raters, which independently evaluated each triple of their subgraph. The triples of each subgraph were presented to the raters grouped by event, and for each event the link to its corresponding mentions in the text were provided, so that raters were able to inspect the original text to assess the correctness of the extracted triples. In total, 8 human raters evaluated a total of 1043 triples of the ECKG, with each triple independently evaluated by two raters.

Raters were given precise criteria to follow for evaluating their subgraph. For instance, in case of an event extracted from many event mentions, raters were instructed to first assess if all its mentions actually refer to the same event: if at least one of these mentions is referring to an event different than the other ones, all triples of the resulting instance have to be considered incorrect.[52] This is a quite strict and potentially highly-penalizing criterion, if considered in absolute terms from a triple perspective: one "wrong" mention out of many coreferring mentions, potentially contributing with few triples to the event, may hijack all the triples of the corresponding event. There were for example several instances in which 4 mentions were identified by the pipeline as referring to the same event instance, of which 3 were indeed referring to the same instance. Due to our strict evaluation method, all four mentions were considered incorrect. Performing a pairwise evaluation would have been less strict, but as our goal is to accurately extract event-centric knowledge graphs from text, and in particular to obtain correct structured description of events, we believe this criterion goes in this direction.

Table 2 presents the resulting triple accuracy on the whole evaluation dataset, as well as the accuracy on each subgraph

[52] A similar criterion was adopted for cases where something was wrongly identified as an event.

Fig. 11. SynerScope accessing the cars ECKG extracted from 1.26M news articles.

Table 2
Quality triple evaluation of ECKG extracted from Wikinews.

	S_1	S_2	S_3	S_4	**All**
Triples	267	256	261	259	**1043**
Accuracy	0.607	0.525	0.552	0.548	**0.551**
κ	*0.623*	*0.570*	*0.690*	*0.751*	

composing it, obtained as average of the assessment of the each rater pair. For each subgraph, the agreement between the rater pair is also reported, computed according to the Cohen's kappa coefficient (κ).

The results show an overall accuracy of 0.551, varying between 0.525 and 0.607 on each subgraph. The Cohen's kappa values, ranging from 0.570 and 0.751, show a substantial agreement between the raters of each pair. Drilling down these numbers on the type of triples considered—*typing* triples (rdf:type), *annotation* triples (rdfs:label), *participation* triples (properties modeling event roles according to PropBank, FrameNet, and ESO), the accuracy on annotation triples is higher (0.772 on a total of 101 triples), while is slightly lower for typing (0.522 on 496 triples) and participation triples (0.534 on 446 triples). Indeed, further drilling down on participation triples, the accuracy is higher for PropBank roles (0.559) while is lower on FrameNet (0.438) and ESO roles (0.407), which reflects the fact that the SRL tool used is trained on PropBank, while FrameNet and ESO triples are obtained via mapping.

Looking at the event candidates in the evaluation dataset, 69 of them (out of 100) were confirmed as proper events by both raters. Of the 17 candidate coreferring events (i.e., those having multiple mentions), only 4 of them were marked as correct by both raters (i.e., both raters stated that all mentions were actually referring to the same event) while in a couple of cases an event was marked as incorrect because of one wrong mention out of 4, thus causing all the triples of the event to be marked as incorrect. To remark the aforementioned strict evaluation criteria adopted, we note that ignoring all coreferring events (and their corresponding triples) in the evaluation dataset, the triple accuracy rises to 0.697 on a total of 782 triples.

7. Applications

In this section, we show how an ECKG can be exploited to reconstruct a story through visualization of events. Traditional KGs can be queried for co-occurrences or relationships between entities. We show that an event-centric approach allows the user to order events in which particular entities are involved on a timeline and reconstruct a story-line as reported in the sources.

7.1. Visual exploration of ECKGs

An example of an application that can exploit an ECKG is SynerScope,[53] an advanced visual analytics application. This application is designed to work with the basic elements of a graph: nodes and links. Graphs can be visualized as hierarchically bundled edges or in a temporally ordered view. More visualizations are available for the other attributes of nodes and links, such as maps, scatter plots, and word clouds.

All interactions (such as selecting or drilling down into parts of the graph) are coordinated across the visualizations. This allows users to interact with data in one dimension, and see the results in other dimensions. This principle enables the user to explore correlations between different facets of the data.

Fig. 11 shows a screenshot of SynerScope accessing the cars ECKG (Section 5.3). The top left visualization shows who interacted with who specifically, the top middle visualization shows the same interactions through time, while the bottom visualization shows the type of interactions in a word cloud. The user has selected a few events using the timeline, the other visualizations have automatically applied an equivalent selection showing who was involved and what the type of event was. The top right panel shows one of the original news article this ECKG was derived from.

In SynerScope, the ECKGs produced by our methodology are visualized as 'interaction pairs', meaning that actors and places are nodes which are linked by events. Two actors that participate in the

53 http://www.synerscope.com/.

same event are essentially connected to each other by that event. An actor can also be connected to a place where an event occurred. This allows users to explore who did what, with whom, where, and when. A user can for example explore the interactions the car manufacturer Peugeot had with other entities. The user can select Peugeot using the network visualization and ask SynerScope to show her who Peugeot interacted with by expanding our selection to the events Peugeot is connected to. These events in turn are connected to other actors, as well as places, both will become immediately visible in the network visualization. Here the user can see that Peugeot, as a French organization, interacted many times with the French government as well as that Peugeot interacted with Iran and Iranian companies. If the user wants to take a more detailed look at these specific interactions, she can select all interactions between Peugeot and the Iranians and drill down to show them in isolation.

Users can explore timelines of interactions between multiple actors, see the number of events that occurred in a certain place, find events of a specific type, and so forth. In the Peugeot example, a user can select events in this timeline from the earliest event to the latest, seeing the types of events change over time. Early events involving both Peugeot and Iran indicate an amicable relationship where Iranian companies are producing Peugeot vehicles under license. Later, the events indicate see Peugeot becoming hesitant to further invest in Iran. Ultimately, the timeline shows that the sustained production of Peugeot vehicles in Iran is becoming difficult as export and import restrictions effected for Iran by the international community. In addition, after Peugeot was taken over by General Motors, the sources for the selected events show pressure being applied to stop interactions with Iran because Peugeot is now part of an originally American organization.

Thanks to the event-centric nature of the data and the affordances of SynerScope, the kind of analytic workflow described above only takes a few minutes and can easily be applied to other cases as well. Performing a similar analysis in a standard document-based information system involves manually inspecting many documents to assess their relevance to the user's query, which takes considerably more time without the summarizing visualizations. End-user evaluations in which ECKG is pitted against a state-of-the-art document-based search system loaded with the sources from which the ECKGs is extracted are currently underway.

7.2. Community exploitation

We are also assessing the usefulness of the ECKGs produced by our tools and methodology, by inviting data journalists and web developers to play with them during Hackathons. Thus far, our ECKGs, together with the original documents from which they were extracted, were used in three Hackathons. In all three events, the ECKGs were made available through a dedicated KnowledgeStore installation.

In the first Hackathon,[54] 40 people, a mixture of Linked Data enthusiasts and data journalists, gathered for one day to collaboratively develop web-based applications exploiting the content of the ECKG extracted from the FIFA World Cup corpus (see Section 5.2). Ten web-based applications, implemented in different programming languages, were developed in roughly 6 working hours. For people not familiar with Semantic Web technologies such as RDF and SPARQL, the NewsReaderSimple API [58].[55] This is a HTTP ReST API developed in python that uses JSON and is easily accessible from JavaScript, and where each method implements a SPARQL query template instantiated at runtime with the actual parameters passed to the method, and fired against the KnowledgeStore that hosts the ECKG and original news article. Each application was developed with a focused purpose: among them, to determine which teams some football player had played during his career (by looking at transfer events); to discover which football teams were most commonly associated with violence; to determine people and companies related to gambling; and, to establish the popularity of people, companies, and football teams in different locations.

In the other two Hackathons,[56] a total of 50 people gathered to build enhanced applications and conduct exploratory investigation exploiting the Global Automotive Industry ECKG (see Section 5.3): among them, an in-depth analysis of the age of CEOs when they get hired or fired from companies, analysis of most dangerous cars around, car companies with high cars recall rate, sentiment analysis of car brands,[57] and so on.

8. Discussion and conclusions

In this paper, we defined the concept of Event-Centric Knowledge Graphs (ECKGs). We then presented a model to represent ECKGs, and a method and an open source toolkit to automatically build ECKGs after which we presented four different ECKGs extracted from different corpora consisting of news articles. ECKGs focus on capturing the dynamic information conveyed by events ("what, who, when, where") mentioned in the news, thus complementing the static encyclopedic content typically available in traditional knowledge graphs. As such, ECKGs are capable of covering long-term developments and histories on potentially hundreds and thousands of entities. To the best of our knowledge, we are the first to automatically build ECKGs from large, unstructured news article text collections.

Our approach stands on the shoulders of deep NLP techniques, such as Entity linking and Semantic Role Labeling, whose output at the text mention level is reinterpreted and abstracted at the level of event and entity instances in a Semantic Web grounded representation schema, independently of the news article language. Our processing pipelines, consisting of tools that we all released Open Source, are developed according to Big Data computation paradigms, enable us to efficiently process large text collections resulting in the construction of ECKGs in four use cases from corpora ranging from a few hundred to millions of sources. One of these ECKGs is constructed from text sources in different languages, showcasing the cross-lingual application of our approach. We described a concrete commercial application (SynerScope) that exploits the content contained in our ECKGs, and discussed several kinds of explorations of the information made possible by the output of our tools, concretely showing the usefulness of the ECKGs we are able to produce. Still, we reckon that there is room for improvement and there are several challenges to be addressed in our future work.

Each module of our processing pipelines delivers state-of-the-art or better performance on its task. Nevertheless, there are some key tasks for our approach, such as entity resolution (including both entity linking and nominal co-reference) and semantic role labeling, on which an improvement of the performance (e.g. coverage of more expressions and more complete predicate representations) would positively impact the ECKGs we produce. This improvement, especially in cross-lingual settings, could follow from the further enrichment and coverage extension of resources such as the Predicate Matrix, which align several predicate-based

[54] Held in London, 10th June 2014—http://www.newsreader-project.eu/newsreader-world-cup-hack-day/.

[55] Accessible at: https://newsreader.scraperwiki.com was made available. The code available at https://bitbucket.org/scraperwikids/newsreader_api_flask_app.

[56] Held in (i) Amsterdam, 21st January 2015—http://www.newsreader-project.eu/amsterdam-hackathon-recap/, and (ii) London, 30th January 2015—http://www.newsreader-project.eu/london-hackathon/.

[57] http://tinyurl.com/pdedwto.

resources such as PropBank, FrameNet, VerbNet, WordNet, and ESO. Furthermore, we are looking into utilizing additional metadata about the news sources such as the type of publication and the author, or even additional markup such as found in Wikinews, to enrich and improve the quality of the extracted information.

As mentioned in Section 4.2, the proper identification of co-referring entities and events highly impacts the density of the representation at the instance level that our approach produces: no or minimal coreference resolution of entities and events would lead to an instance representations close to mentions, whereas a highly co-referring approach will result in the merge and aggregation of several mentions in few instances and their relations. We plan to further investigate the optimal aggregation level of mentions into instances. We will also address the challenge posed by the proper representation and interpretation of non-entities, i.e. those concepts, other than proper named entities, typically involved in events (e.g. quantities or unspecified entities).

Due to the ambiguity of natural language, our ECKGs are inevitably affected by noise and redundancies introduced when automatically processing a text (cf. Section 4—e.g. redundant content, linking to the wrong DBpedia referent, mismatches between role and type of entities participating in an event). We are therefore investigating *knowledge crystallization* techniques, i.e. the use of reasoning techniques exploiting background knowledge to clean the ECKGs that we produce, as to remove redundant content and noise. More general, the information automatically extracted from our pipelines at the mention level could be considered as "evidence" of facts mentioned in text, and only when the amount of evidence exceeds an appropriate threshold or certain conditions are met, these facts should be considered as *first class citizens* to be included in the ECKGs.

So far, we applied our approach to build ECKGs all at once, i.e. processing all news articles of a text collection as part of a single batch of textual resources. However, with news articles being produced daily, a more realistic processing scenario would require to produce ECKGs in an incremental manner: starting from a (potentially empty) ECKG, events and entities extracted from daily news have to be interpreted also in light of *past* events and entities, already processed and contained in the ECKG. Despite the technical challenges, this could potentially require to revise (or even retract) the content already stored in the ECKG, for instance due to the need of revising the aggregation of mentions in instances following the new content extracted from today news. We plan to experiment with these options in streaming scenarios.

In the paper, we discussed the capabilities and usefulness of our approach, reporting on applications (e.g. SynerScope) and in-depth news investigations made possible by our ECKGs. We also presented the results of a first evaluation of the accuracy of the triples in our ECKGs. We continue to improve the output of the individual modules, as well as perform end-user evaluations to assess the quality of the ECKG and its usefulness in professional decision making.

Acknowledgment

This research was funded by European Union's 7th Framework Programme via the NewsReader project (ICT-316404).

References

[1] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, S. Hellmann, Dbpedia—a crystallization point for the web of data, J. Web Semant.: Sci. Serv. Agents World Wide Web 7 (3) (2009) 154–165.
[2] F.M. Suchanek, G. Kasneci, G. Weikum, Yago: a core of semantic knowledge, in: Proceedings of the 16th International Conference on World Wide Web, 2007, pp. 697–706.
[3] D. Vrandeččić, M. Krötch, Wikidata: a free collaborative knowledgebase, Commun. ACM 57 (10) (2014) 78–85.
[4] M.-A.N. Francois Belleau, N. Tourigny, P. Rigault, J. Morissette, Bio2rdf: Towards a mashup to build bioinformatics knowledge systems, J. Biomed. Inform. 41 (5) (2008) 706–716.
[5] A.J.G. Gray, P. Groth, A. Loizou, S. Askjaer, C. Brenninkmeijer, K. Burger, C. Chichester, C.T. Evelo, C. Goble, L. Harland, S. Pettifer, M. Thompson, A. Waagmeester, A.J. Williams, Applying linked data approaches to pharmacology: Architectural decisions and implementation, Semant. Web J. 5 (2) (2014) 101–113.
[6] A. Bordes, E. Gabrilovich, Constructing and mining web-scale knowledge graphs: Kdd 2014 tutorial, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'14, ACM, New York, NY, USA, 2014, pp. 1967–1967. http://dx.doi.org/10.1145/2623330.2630803. URL http://doi.acm.org/10.1145/2623330.2630803.
[7] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, ACM, 2008, pp. 1247–1250.
[8] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, W. Zang, Knowledge vault: a web-scale approach to probabilistic knowledge fusion, in: KDD'14 Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 601–610.
[9] D. Shahaf, C. Guestrin, Connecting the dots between news articles, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'10, Washington, DC, USA, 2010, pp. 623–632.
[10] E. Kuzey, J. Vreeken, G. Weikum, A fresh look on knowledge bases: Distilling named events from news, in: J. Li, X.S. Wang, M.N. Garofalakis, I. Soboroff, T. Suel, M. Wang (Eds.), Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, CIKM 2014, ACM, Shanghai, China, 2014, pp. 1689–1698. http://dx.doi.org/10.1145/2661829.2661984. URL http://doi.acm.org/10.1145/2661829.2661984.
[11] O. Etzioni, M. Banko, S. Soderland, D.S. Weld, Open information extraction from the web, Commun. ACM 51 (12) (2008) 68–74. http://dx.doi.org/10.1145/1409360.1409378. URL http://doi.acm.org/10.1145/1409360.1409378.
[12] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. H Jr., T. Mitchell, Toward an architecture for never-ending language learning, in: Proceedings of the Conference on Artificial Intelligence, AAAI, AAAI Press, 2010, pp. 1306–1313.
[13] M. Surdeanu, S. Harabagiu, J. Williams, P. Aarseth, Using predicate-argument structures for information extraction, in: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, vol. 1, ACL'03, Association for Computational Linguistics, Stroudsburg, PA, USA, 2003, pp. 8–15. http://dx.doi.org/10.3115/1075096.1075098.
[14] H. Llorens, E. Saquete, B. Navarro, Tipsem (english and spanish): Evaluating crfs and semantic roles in tempeval-2, in: Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval'10, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 284–291. URL http://dl.acm.org/citation.cfm?id=1859664.1859727.
[15] J. Pustejovsky, J. Castaño, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, G. Katz, Timeml: Robust specification of event and temporal expressions in text, in: Fifth International Workshop on Computational Semantics, IWCS-5, 2003.
[16] P. Exner, P. Nugues, Using semantic role labeling to extract events from wikipedia, in: Proceedings of the Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web, DeRiVE 2011. Workshop in conjunction with the 10th International Semantic Web Conference, 2011.
[17] J. Christensen, Mausam, S. Soderland, O. Etzioni, Semantic role labeling for open information extraction, in: Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading, FAM-LbR'10, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 52–60. URl http://dl.acm.org/citation.cfm?id=1866775.1866782.
[18] S.-H. Hung, C.-H. Lin, J.-S. Hong, Web mining for event-based commonsense knowledge using lexico-syntactic pattern matching and semantic role labeling, Expert Syst. Appl. 37 (1) (2010) 341–347. http://dx.doi.org/10.1016/j.eswa.2009.05.060.
[19] L. Padró, Željko Agić, X. Carreras, B. Fortuna, E. García-Cuesta, Z. Li, T. Štajner, M. Tadić, Language processing infrastructure in the xlike project, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC2014, 2014.
[20] V. Presutti, F. Draicchio, A. Gangemi, Knowledge extraction based on discourse representation theory and linguistic frames, in: A. ten Teije, J. Völker, S. Handschuh, H. Stuckenschmidt, M. d'Acquin, A. Nikolov, N. Aussenac-Gilles, N. Hernandez (Eds.), Knowledge Engineering and Knowledge Management, in: Lecture Notes in Computer Science, vol. 7603, Springer, Berlin, Heidelberg, 2012, pp. 114–129. http://dx.doi.org/10.1007/978-3-642-33876-2_12.
[21] A. Cybulska, P. Vossen, Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution, in: Proceedings of the 9th Language Resources and Evaluation Conference, LREC2014, Reykjavik, Iceland, 2014.
[22] Z. Beloki, G. Rigau, A. Soroa, A. Fokkens, K. Verstoep, P. Vossen, M. Rospocher, F. Corcoglioniti, R. Cattoni, S. Verhoeven, M. Kattenberg, System Design, Version 2, Deliverable 2.2, NewsReader Project, 2015.
[23] A. Fokkens, M. van Erp, P. Vossen, S. Tonelli, W.R. van Hage, L. Serafini, R. Sprugnoli, J. Hoeksema, Gaf: A grounded annotation framework for events, in: Proceedings of the 1st workshop on Events: Definition, Detection, Coreference, and Representation at the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL2013, Association for Computational Linguistics, Atlanta, GA, USA, no. ISBN: 978-1-937284-47-3, 2013.

[24] W.R. van Hage, V. Malaisé, R. Segers, L. Hollink, G. Schreiber, Design and use of the Simple Event Model (SEM), J. Web Sem. 9 (2) (2011) 128–136.
[25] R. Cyganiak, D. Wood, M. Lanthaler, Rdf 1.1 Concepts and Abstract Syntax, Tech. Rep., W3C, 2014. URL http://www.w3.org/TR/rdf11-concepts/.
[26] L. Moreau, P. Groth, Provenance: An Introduction to PROV, in: vol. 3 of Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan Claypool, 2013.
[27] Z. Beloki, G. Rigau, A. Soroa, A. Fokkens, P. Vossen, M. Rospocher, F. Corcoglioniti, R. Cattoni, T. Ploeger, W.R. van Hage, System Design, Version 1, Deliverable 2.1, NewsReader Project, 2014.
[28] C.F. Baker, C.J. Fillmore, J.B. Lowe, The berkeley framenet project, in: Proceedings of the 17th International Conference on Computational linguistics, vol. 1, Association for Computational Linguistics, 1998, pp. 86–90.
[29] R. Segers, P. Vossen, M. Rospocher, L. Serafini, E. Laparra, G. Rigau, Eso: A frame based ontology for events and implied situations, in: Proceedings of MAPLEX 2015, Yamagata, Japan, 2015. URL https://dkm-static.fbk.eu/people/rospocher/files/pubs/2015maplex.pdf.
[30] A. Fokkens, A. Soroa, Z. Beloki, N. Ockeloen, G. Rigau, W.R. van Hage, P. Vossen, NAF and GAF: Linking linguistic annotations, in: Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation, Reykjavik, Iceland, 2014, p. 9. URL http://sigsem.uvt.nl/isa10/ISA-10_proceedings.pdf.
[31] R. Agerri, X. Artola, Z. Beloki, G. Rigau, A. Soroa, Big data for natural language processing: A streaming approach, Knowl.-Based Syst. 79 (0) (2015) 36–42. http://dx.doi.org/10.1016/j.knosys.2014.11.007. URL http://www.sciencedirect.com/science/article/pii/S0950705114003992.
[32] W. Bosma, P. Vossen, A. Soroa, G. Rigau, M. Tesconi, A. Marchetti, M. Monachini, C.L. Aliprandi, Kaf: a generic semantic annotation format, in: Proceedings of the GL2009 Workshop on Semantic Annotation, 2009.
[33] N. Ide, L. Romary, É.V. de La Clergerie, International standard for a linguistic annotation framework, in: Proceedings of the HLT-NAACL 2003 Workshop on Software Engineering and Architecture of Language Technology Systems, SEALTS, Association for Computational Linguistics, 2003.
[34] R. Agerri, I. Aldabe, Z. Beloki, E. Laparra, M.L. de Lacalle, G. Rigau, A. Soroa, A. Fokkens, R. Izquierdo, M. van Erp, P. Vossen, C. Girardi, A.-L. Minard, Event Detection, Version 2, Tech. Rep., NewsReader Project, 2015.
[35] R. Agerri, J. Bermudez, G. Rigau, IXA pipeline: Efficient and ready to use multilingual NLP tools, in: Proceedings of the 9th Language Resources and Evaluation Conference, LREC2014, Reykjavik, Iceland, 2014.
[36] P. Mirza, A.-L. Minard, HLT-FBK: a complete temporal processing system for QA TempEval, in: Proceedings of the Ninth International Workshop on Semantic Evaluation, SemEval'15, 2015.
[37] A.B. Anders, B. Bohnet, L. Hafdell, P. Nugues, A high-performance syntactic and semantic dependency parser, in: Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations, COLING'10, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 33–36. URL http://dl.acm.org/citation.cfm?id=1944284.1944293.
[38] J. Daiber, M. Jakob, C. Hokamp, P.N. Mendes, Improving efficiency and accuracy in multilingual entity extraction, in: Proceedings of the 9th International Conference on Semantic Systems (I-Semantics), 2013.
[39] M. van Erp, P. Vossen, R. Agerri, A.-L. Minard, M. Speranza, R. Urizar, E. Laparra, I. Aldabe, G. Rigau, Deliverable d3.3.2: Annotated Data, Version 2, Tech. Rep., NewsReader Project, 2014.
[40] E.F. Tjong Kim Sang, F.D. Meulder, Introduction to the conll-2003 shared task: Language-independent named entity recognition, in: Proceedings of CoNLL-2003, Edmonton, Canada, 2003, pp. 142–147.
[41] J.R. Finkel, T. Grenager, C. Manning, Incorporating non-local information into information extraction systems by gibbs sampling, in: Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics, ACL 2005, 2005, pp. 363–370.
[42] L. Ratinov, D. Roth, Design challenges and misconceptions in named entity recognition, in: Proceedings of CoNLL'09, 2009.
[43] J. Hoffart, M.A. Yosef, I. Bordin, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, G. Weikum, Robust disambiguation of named entities, in: Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, 2011, pp. 782–792.
[44] M. Palmer, D. Gildea, P. Kingsbury, The proposition bank: An annotated corpus of semantic roles, Comput. Linguist. 31 (1) (2005) 71–106. http://dx.doi.org/10.1162/0891201053630264.
[45] M. López de Lacalle, E. Laparra, G. Rigau, Predicate matrix: extending semlink through wordnet mappings, in: The 9th edition of the Language Resources and Evaluation Conference. Reykjavik, Iceland, 2014.
[46] K. Kipper, Verbnet: A broad-coverage, comprehensive verb lexicon (Ph.D. thesis), University of Pennsylvania, 2005. URL http://repository.upenn.edu/dissertations/AAI3179808/.
[47] C. Fellbaum (Ed.), WordNet. An Electronic Lexical Database, The MIT Press, 1998.
[48] M. Palmer, Semlink: Linking propbank, verbnet and framenet, in: Proceedings of the Generative Lexicon Conference, 2009, pp. 9–15.
[49] A. Björkelund, L. Hafdell, P. Nugues, Multilingual semantic role labeling, in: Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task, CoNLL'09, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 43–48. URL http://dl.acm.org/citation.cfm?id=1596409.1596416.
[50] M. Rospocher, A.-L. Minard, P. Mirza, P. Vossen, T. Caselli, A. Cybulska, R. Morante, I. Aldabe, Deliverable d5.1.2: Event Narrative Module, Version 2, Tech. Rep., NewsReader Project, 2015.
[51] W.V. Quine, Events and reification, in: Actions and Events: Perspectives on the Philosophy of Davidson, Blackwell, 1985, pp. 162–71.
[52] A. Cybulska, P. Vossen, Bag of events approach to event coreference resolution. supervised classification of event templates, Int. J. Comput. Linguist. Appl. 6 (2) (2015) 9–24.
[53] C. Leacock, M. Chodorow, Combining local context with wordnet similarity for word sense identification, 1998.
[54] F. Corcoglioniti, M. Rospocher, R. Cattoni, B. Magnini, L. Serafini, Interlinking unstructured and structured knowledge in an integrated framework, in: 7th IEEE International Conference on Semantic Computing, ICSC, Irvine, CA, USA, 2013.
[55] F. Corcoglioniti, M. Rospocher, R. Cattoni, B. Magnini, L. Serafini, The knowledgestore: a storage framework for interlinking unstructured and structured knowledge, Int. J. Semant. Web Inf. Syst. 11 (2) (2015) 1–35. http://dx.doi.org/10.4018/IJSWIS.2015040101. URL http://www.igi-global.com/article/the-knowledgestore/136832.
[56] P. Stouten, R. Kortleven, I. Hopkinson, Deliverable d8.1: Test Data and Scenarios, Tech. Rep., NewsReader Project, 2013.
[57] J. Hoffart, F.M. Suchanek, K. Berberich, G. Weikum, YAGO2: A spatially and temporally enhanced knowledge base from wikipedia, Artificial Intelligence 194 (2013) 28–61.
[58] I. Hopkinson, S. Maude, M. Rospocher, A simple API to the KnowledgeStore, in: Proceedings of the ISWC Developers Workshop colocated with the 13th International Semantic Web Conference (ISWC'14), Riva del Garda, Italy, 2014.