Timelines: Entity-centric Event Extraction from Online News

Jakub Piskorski
Polish Academy of Sciences
Warsaw, Poland
jpiskorski@gmail.com

Vanni Zavarella, Martin Atkinson, Marco Verile
Joint Research Centre of the European Commission
Ispra, Italy
{firstname.surname}@ec.europa.eu

Abstract

Automatically extracting structured information on events from online sources for the purpose of intelligence gathering has been acknowledged to be of paramount importance by various organisations worldwide .

This paper reports on an ongoing endeavour on developing a tool that for a given target entity of interest extracts from a stream of online news articles structured information on events this entity participated in or in whose context it was mentioned. Furthermore, other entity-related relations that hold are extracted and whenever applicable the events are anchored on a time-scale and classified, which all together constitutes a *target-entity event timeline*. The paper first briefly introduces the timeline extraction task and then gives an overview of the core extraction engine and event browsing functionalities. The results of a rudimentary evaluation of the quality of the extracted information is also provided. The paper is accompanied with a demo of the tool.

1 Introduction

Media monitoring, as a general service, remains a necessary task for most large organizations to keep track of any new developments in their domains as well as to assess impact and reputation. For that purpose, the European Commissions Joint Research Centre (JRC) developed the Europe Media Monitor[1] (EMM) [RAG+17], a fully automatic software system that gathers and analyses an average of 300,000 online news articles per day in up to 70 languages and is serving several EU institutions and agencies, EU member states authorities and international organisations. Sources identification is based on domain[2] experts advice and include international, national, regional, and local media outlets as well as specialized and institutional web sites. News are initially classified in more than 6000 categories according to complex keywords-based multilingual combinations (up to hundreds of keywords) and further filtered according to other metadata like source origin, language, ranking, geo-locations, entities, etc. The initial unmanageable dataflow is therefore reduced to very specialized feeds, fine-tuned for each team of analysts, and that can be screened in a reasonable amount of time in order to generate high quality media reports timely delivered to decision makers.

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia (eds.): Proceedings of the Text2Story'20 Workshop, Lisbon, Portugal, 14-April-2020, published at http://ceur-ws.org

[1] `https://emm.newsbrief.eu/`

[2] Domains range from socio-political media monitoring, to epidemic intelligence, security and conflict-related early warning, science and innovation.

However, media analysts need to move from a selection of relevant news to a more compact and abstract representation of the underlining extracted data including summaries, dashboards, and timelines that visually and quantitatively support the narrative of their reports. Moreover, ad-hoc briefings often require searching past data to gather background information about events and related entities. Media analysts are often shifting their scope into media intelligence looking for insights in massive multilingual textual collections like those produced by the on-line media and therefore difficult to be explored without automated instruments able to extract and index entity-related knowledge. The increasing efforts in the fight against disinformation also brought to the inception of new methodologies to support fact-checking activities like the identification and linking of reported past events related to public figures or organizations including the possibility to trace back to their sources.

In order to pursue these challenges, JRC started building a Media Analysis Capability (MAC) by developing new tools for both automated information extraction and visualization with the existing EMM engine. This paper reports on an ongoing endeavour in this context, namely, on the development of a tool that for a given entity of interest extracts from online news structured information on events this entity participated in and other entity-related relations that hold, and whenever applicable anchoring thereof on a time-scale. Similar-in-nature research work on the extraction of entity-centric information from online news has been reported in [DMBZ10, STSO16, ROR13, MBM19, DJT17].

Due to the application-oriented context of EMM and its highly multilingual nature we deploy methods which use as little linguistic sophistication as possible (i.e., we avoid using deep linguistic processing techniques and ones that require significant amount of time in order to create language-specific resources) and exploit Open Information Extraction (OIE) techniques [Mau16] to facilitate scalability and portability. Furthermore, the structure of the output is motivated by the practical utility of the results by the end-users rather than strictly sticking to the various formats and tasks established by the scientific community, e.g. instead of identifying precisely which semantic role the target entity has in an event, we only differentiate between the entity being participant or being mentioned in the context of the event, and leaving further interpretation to the end user. The main drive behind reformulating some extraction subtasks and introducing more lenient definition was also to find the best trade-off between accuracy/correctness of the returned results and fine-grained-ness of the system response. It is important to emphasize at this stage that we do not necessarily exploit (and it was not our intention) the latest state-of-the-art NLP techniques and toolkits due to the fact that the presented tool is to be used in an operational set-up, in whose context using, for instance, knowledge-based approaches for certain tasks has clear advantages over statistical approaches as discussed for instance in [Dah17], not to mention the multilinguality aspect which prohibits the use of a vast majority of available state-of-the-art NLP tools, which exhibit either black box character or would require a significant amount of work to cover high number of languages.

The work reported here is mostly related to research on OIE [BCS+07, EFC+11, Mau16, CSE11, Bor18, GWH+19, CSCXO09], Knowledge Harvesting [WHS16, RSH+16, GHMS14], temporal event reasoning [MSA+15, CV16, LAAR17, WSY17] and extraction of narratives from text [DBL19].

The rest of the paper is structured as follows. Section 2 briefly introduces the task of Timeline Extraction. Next, Section 3 provides the description of the core Timeline Extraction engine. Subsequently, the web-based tool for event browsing and visualization is presented in Section 4. The results of a rudimentary evaluation of the current version of the tool are provided in Section 5. Finally, Section 6 provides future outlook.

2 Timeline Extraction Task

Given a set of documents and a target entity of interest (e.g. a person reported in online media), our Timeline Extraction task consists of extracting a list of time-ordered tuples of the form: `<entity, event, entityRole, startTime,endTime,<relEntities>>`. The target entity is defined by the user in a form of so called entity profile, which consists of a canonical name of the entity (e.g., *Donald Trump*), related name variants (e.g., *President Trump*, *D. Trump*), and gender information[3]. An `event` is defined here as any kind of situation, involving the target entity `entity`, that happens or occurs, being punctual or lasting for a period of time, as well as those predicates describing states or circumstances in which something obtains or holds true, based on the TimeML standard specifications [SLK+06]. `entityRole` is an attribute that specifies whether the target entity is a: (a) participant of the event or, (b) non-participant of the event, but it is mentioned in the same sentence. Furthermore, `startTime` and `endTime` are optional slots for recording time-related information, whereas `<relEntities>` is a set of additional entities matched within the same sentence where the event was detected.

[3] which is exploited in the context of co-reference resolution

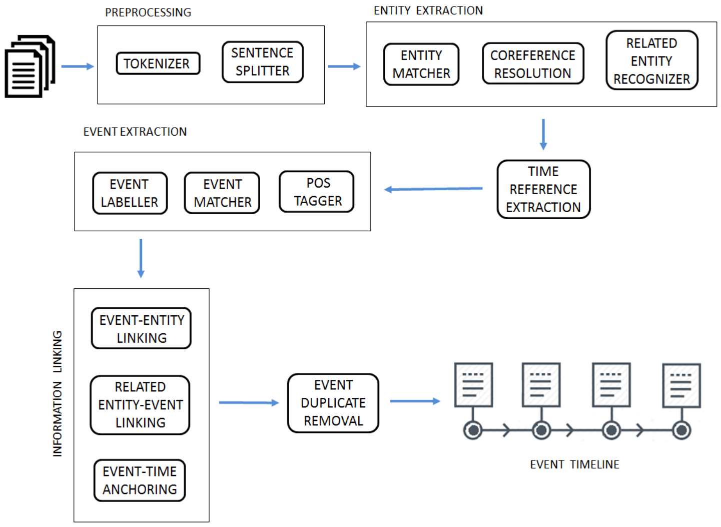


Figure 1: The architecture of the core Timeline Extraction engine.

The definition of Timeline Extraction sketched above is more lenient (mainly driven by the application scenario) compared to some standards adopted within the scientific community, e.g. [MSA$^+$15]. In particular, the target entity matching encompasses also the detection of additional entities, possibly of a different type than the target one, that contain the target entity in their name (e.g. *Elon Musk Foundation* wrt *Elon Musk*) based on the assumption that they are semantically related to the target entity and thus of interest to a media analyst. Furthermore, no full-fledged semantic role labelling is performed, i.e., the specific role the target entity fills with respect to the event (e.g. Agent, Patient) is not detected. Similarly, no specific role is assigned to the related entities associated with an event.

3 Core Timeline Extraction Engine

The process of extraction of a timeline (currently only for English) is broken down into several steps described below. The overall architecture of the core Timeline Extraction engine is depicted in Figure 3. While in general the Timeline Extraction engine can be applied on an arbitrary collection of documents, in the case of processing news articles a pre-selection thereof is made through retrieving from news repository the ones which include mentions of the target entity using the information stored in the entity profile mentioned above and some additional meta-data criteria (see further details in Section 4).

3.1 Pre-processing

Each of the input set of documents is first tokenized and split into sentences using the toolkit described in [Pis19]. All subsequent steps process each single sentence separately, unless specified differently.

3.2 Entity Extraction

The entity extraction process consists of three steps, namely, matching named mentions of the target entity first, followed by a lightweight co-reference resolution, and finally, recognition of other entity mentions in the relevant context.

Target Entity Matching is carried out by means of: (a) exact matching of the canonical form of the target entity or a variant thereof (either provided by the user or automatically generated), (b) expanding the exact matches into bigger units based on capitalization of adjacent tokens and use of a stop-word list, e.g., the exact match of *Melinda Gates* in the text *The Bill & **Melinda Gates** Foundation has plans to ...* is expanded to ***The Bill & Melinda Gates Foundation***, and (c) fuzzy matching through using a user-defined string distance metric, where by default, based on our empirical observations[4], a weighted version of the *Longest Common*

[4] Running tests on ca. 20 entities and news articles in English

Substrings metric introduced in [PWS09] is deployed. Although the expanded NE mention in the example above has a different NE type it is semantically related to the target entity, and thus potentially interesting for the analysis of target-entity related events.

Co-reference Resolution is deployed in order to boost the target entity detection recall. It simply resolves pronominal anaphora in the local context by searching the closest preceding target entity that matches Number, Gender, Person agreement constraints with the pronoun, unless another compatible entity is interposed in between. We limit the local context only to the sentence where the event mention was detected since going beyond the scope of the sentence has been reported by various research groups to deteriorate the co-reference resolution accuracy [LAAR17].

Related Entity Recognition is carried in three steps, the last one being optional. First, three multilingual lexico-semantic resources in the respective order are applied on the unconsumed part of text to recognize entities: (i) JRC Variant Names database (ca 4 mln entries, mostly person names) [EJS17], (ii) a collection of multi-word named entities from BabelNet [NP12] (ca. 6.8 mln entries) semi-automatically derived using the method described in [CJSP17], and (iii) toponyms (only populated places) from the GeoNames[5] gazetteer (ca. 1.4 mln entries). Next, simple patterns are deployed to combine previously recognized entities, e.g. based on the occurrence of conjunctions between entities. Finally, guessing heuristics are used on the yet unconsumed parts of the text, e.g., sequences of capitalized tokens filtered using a multilingual stop word list (ca 20K entries) are considered as related entities.

3.3 Time Reference Extraction

The identification of time references is performed separately on the full document text, by deploying a cascade of finite-state grammar rules for detecting temporal expressions according to the TIMEX3 tag specification of the TimeML framework [PCI+03], combined with a language-independent algorithm for resolving under-specified temporal expressions and assigning them a normalized time value according to the TimeML standard. The deployed module uses a shallow, rule-based approach we described in detail in [ZT13]. The normalization module performs some basic discourse analysis by keeping a reference time anchor throughout the text, falling back to document creation date in case no suitable anchor time is available. We process both punctual and durative time references, although we use only the left boundary date of a durative time reference for temporal ordering.

3.4 Event Extraction

The sentences in which mentions of the target entity have been found are subsequently morphologically analyzed using the full-form MULTEXT [Erj10] morphological lexicon for English (around 90k entries) and disambiguated using the Stanford part-of-speech tagger [TM00]. In particular, we exploit MULTEXT lexical resources since the same tagset is used across many languages.

Event Matcher applies then a small set of part-of-speech-based regular expression patterns (similar to the ones described in [FSE11]) to identify simple verb groups (VG), light verb constructions (LVC, i.e., multi-word expressions composed of a verb and a noun, with the noun carrying the semantic content of the predicate), as well as verb nominalizations (NVG, e.g., "*the election of*"). These VGs, LVCs and NVGs constitute event trigger (phrase) candidates and are subsequently checked against an event stop-phrase list to eliminate implausible event triggers. This latter resource has been constructed using frequency analysis of event phrase candidates in a 1.6 mln document English news corpus, while for matching VGs, LVCs and NVGs we deploy finite-state grammars that exploit a combination of surface forms and part-of-speech information, e.g., `V (Part)? Adv (Part | "to" | Adp) -> LVC` (a Verb form followed by optional particle and adverb and by a disjunction of possible PoS items or the "*to*" string). The recursive application of rule patterns in cascades allows to parse some form of complex verb phrases, including modal constructs (i.e. negation), and aspectual constructs (e.g."*started detaining*"). Verb nominalizations were integrated into the MULTEXT lexicon by filtering a subset of ca. 1220 entries from the NOMLEX-plus-clean.1.0 lexicon [MRM+04], namely, only the nominalizations that could be mapped to an existing verb entry in our original morphology. All candidate verb groups are finally validated against POS tag sequences from the Stanford tagger.

Event Labeller tags event descriptions with coarse-grained (person-centric) categories, including, i.a.: Birth, Death, Family (events involving family members), Education, Participation (reflecting attending events),

[5] http://www.geonames.org/

Statement (making statements, opinions, endorsements, etc.), Creation (creating artifacts), Ownership (obtaining, selling goods), Engagement (work relationships, involvement in initiatives), Interaction (meeting and interacting with other entities), Law-related (trials, accusations, violations), Location-related (events reflecting an entity visiting a location), Achievement (receiving prizes), Health-Condition, etc. For instance, event description *expressed the opinion in the meeting with* is tagged with Statement and Interaction labels. The definition of the aforementioned event categories emerged directly from application-oriented information needs.

Currently, the classification of event descriptions is done via matching key phrases (ca. 3300)[6] in event descriptions to the specific categories (e.g., mapping *met with* to Interaction category) and exploitation of a subset the lexico-semantic patterns for extracting binary relations presented in [NWS12].

3.5 Information Linking

In the Information Linking Phase, first, the participation relation of the target entity in the various events in the local context (sentence) is established. Next, related entities are linked to the events as well, and finally, events are anchored in time.

Target Entity-to-Event Linking exploits simple heuristics for guessing the target entity to be a participant of a detected event, using the techniques described in [FSE11]. It searches for the closest (not yet linked) event mention to the target entity, within the same sentence. It applies a rule cascade such that, if event mention is directly adjacent, then the target entity is assigned the Participant role. Otherwise, a weighted token distance measure is applied, penalizing events on the left over the ones on the right on the target entity, as well as the occurrence of intervening relative clause markers, propositional markers (such as commas) or related entities in between. The heuristic copes with coordinated verb phrase sequences as well, allowing the mapping of two or more verb groups to the same target entity.

Related Entity-Event Linking module tags all other entities detected within the same sentence as related to the event. In future, a more restrictive definition of entity-to-event relatedness will be introduced.

Event-Time Anchoring module maps each temporal expression in a sentence to an event mention (or a set of event mentions, in case of coordinate verb phrases) within the same sentence, and consequently assigns the corresponding time reference to the event. The heuristic applied here is similar to the Target Entity-to-Event Linking, it uses a weighted token distance for guessing the most likely event a time expressions is modifying, promoting events preceding a target time reference over the subsequent ones.

3.6 Event Duplicate Detection

Event Duplicate Removal module first removes obvious duplicates through identifying and ”merging” all event mentions into one provided that they have been: (a) triggered by the same event phrase in identical sentences (potentially included in different documents), (b) time-anchored to the same date, and (c) extracted from documents, whose creation date is identical. Optionally, more sophisticated techniques can be deployed on demand. For instance, all events with identical event mentions are grouped, and subsequently clusters are computed for each such group using the *Longest Common Substrings* string similarity metric, and then all events within each of the clusters are merged into one event template. For the linking of non-obvious event mentions together into one event template we exploit an SVM-based model that exploits a range of short text semantic similarity metrics as features (e.g., named-entity overlap, hypernym overlap, n-gram overlap) specifically designed for the event linking task, which we describe in more detail in [PŠZA18]. The rationale behind parametrisation of the event duplicate detection is mainly due to the different end-user needs, e.g., in certain scenarios, having access to different reporting on the same event might be the preferred option.

The various outputs produced by the different modules of the timeline extraction engine are exemplified in a simplified form in Table 1, whereas the final output for the extracted event in JSON format is shown in Figure 2.

4 Event browsing and visualization

In order to provide the core timeline extraction engine with a set of input documents a dedicated component with a complex interface and syntax is deployed in order to retrieve from EMM repository all potentially target

[6] They were created based on frequency analysis of event descriptions extracted from a corpus of 1.6 million news articles in English in 2017, i.e., mapping most frequent and ‘reliable’ word n-grams to the respective categories.

Table 1: The output (in a simplified form) produced by the respective modules for a given input sentence. The indices in square brackets illustrate both the co-reference links generated by the pronominal anaphora resolution module (row 2) and the linking from events to temporal expressions and target entities.

Preprocessing	*Donald Trump made no reference to signing a waiver that officially delays any move of the U.S. Embassy from Tel Aviv to Jerusalem, but the White House confirmed he signed the waiver Wednesday.*
Entity Extraction	***Donald Trump*** *[1] made no reference to signing a waiver that officially delays any move of the U.S. Embassy from Tel Aviv to Jerusalem, but the White House confirmed* ***he*** *[1] signed the waiver Wednesday.*
Time Reference Extraction	***Donald Trump*** *[1] made no reference to signing a waiver that officially delays any move of the U.S. Embassy from Tel Aviv to Jerusalem, but the White House confirmed* ***he*** *[1] signed the waiver* ***Wednesday (29-11-2017)****.*
Event Extraction	***Donald Trump*** *[1] made no reference to signing a waiver that officially delays any move of the U.S. Embassy from Tel Aviv to Jerusalem, but the White House confirmed* ***he*** *[1]* ***signed*** *the waiver* ***Wednesday (29-11-2017)****.*
Information Linking	***Donald Trump*** *[1] made no reference to signing a waiver that officially delays any move of the U.S. Embassy from Tel Aviv to Jerusalem, but the White House confirmed* ***he*** *[1]* ***signed*** *[1] the waiver* ***Wednesday(29-11-2017)*** *[1].*

entity-related news articles that were published within a specified time window, in a specified languages and meeting other criteria based on meta-data information associated with the articles. For document querying purposes so called entity profile is used (mentioned earlier), which primarily consists of a list of the canonical name of the entity and its potential variants). The user may decide whether and to what extent to include this information in the document retrieval process depending on specific application scenario, e.g., enforce that only documents containing the mention of the canonical name will be retrieved to increase precision, etc.

Once the input documents are selected the timeline engine is run and the returned output is then used to populate an event database for further querying and analysis.

A screenshot of the web-based interface of the application for searching and analysing the extracted events is presented in Figure 4. The interface is composed of 3 main panels: a search panel (top), a list panel (left), and, a detail panel (right). The search panel provides a number of different ways to filter out the very large number of events in order to focus on the information that is most relevant for the user. It provides a couple of free text search fields (the snippet – sentence from which the event was extracted, and description – phrase triggering the event) and other constrained search fields (e.g. event type, date, tense, etc.) that are all combined using `AND` logic. Once the Search Events button is pressed (even without putting any constraints into the search panel fields), the list panel is filled on a paged basis showing an overview of the available events that match the values from the search panel.

In the list panel there is the text snippet, event description and time reference related to the event. The view also shows highlights within the text snippet of the metadata extracted for each element: the event phrase in red, the other entities in green and the target entity in yellow. The detail panel shows more fine-grained information on the event: the text snippet, the full event description and other metadata that was extracted.

5 Evaluation

For the purpose of carrying out an evaluation of the quality of the system response we have randomly selected 100 events for the target entity **Donald Trump** from a set of ca. 60K of events automatically extracted from a corpus of 100K news articles in English language gathered by EMM in a one month period. 97% of all extractions were correctly identified as event mentions, out of which 81% were references to factual events. 70% of the event mentions were tagged with at least on generic event category by the Event Labeller, where in 91% of cases the category was correct. Next, we also evaluated the target-entity to event linking, i.e., assessing whether the target entity is correctly tagged as a participant of the event detected. We carried out strict and lenient evaluation, where in the latter case, the target entity is considered to be a participant of the event also when it appears in a

```
{ "eventMatch":
    { "docID": 16462,
      "creationDate": "7-12-2017",
      "startPosition": 6,
      "endPosition": 25,
      "startPositionInDoc": 4248,
      "endPositionInDoc": 4267,
      "confidence": 1.0,
      "eventStartDate": "29-11-2017"
      "eventEndDate": "29-11-2017"
      "textSnippet": "Trump made no reference to signing a waiver that officially delays any move of
                      the U.S. Embassy from Tel Aviv to Jerusalem, but the White House confirmed he
                      signed the waiver Wednesday."
      "eventDescription": "signed"
      "eventCategory": [ "CREATION" ]
      "targetEntityParticipation": true
      "nominalized": false
      "coordinated": false
      "tense": "PAST"
      "aspect": "UNSPECIFIED"
      "modality": "null"
      "relatedEntities": [ { "entity":"U.S. Embassy", "cat":"OTH", "matchType":"EXACT-NAMED-MATCH", },
                           { "entity":"Tel Aviv", "cat":"OTH", "matchType":"EXACT-NAMED-MATCH" },
                           { "entity":"Jerusalem", "cat":"OTH", "matchType":"EXACT-NAMED-MATCH", },
                           { "entity":"White House", "cat":"OTH", "matchType":"EXACT-NAMED-MATCH", } ]
    }
  "targetEntityMatches": [ { "form":"Trump", "matchType":"EXACT-NAMED-MATCH" },
                           { "form":"he(Donald Trump)", "matchType":"PRONOMINAL" } ]
}
```

Figure 2: The final output in JSON format (simplified) for the event extracted from the text shown in Figure 1.

noun phrase or clause that constitute the actual argument of the event, e.g., *Donald Trump* would be considered as a participant of the ***watch*** event in the sentence *35 million TV viewers* ***watch*** *Donald Trump's acceptance speech at GOP convention.* The figures for both strict and lenient evaluation were relatively close to each other, and whose values were 76% and 74% resp.

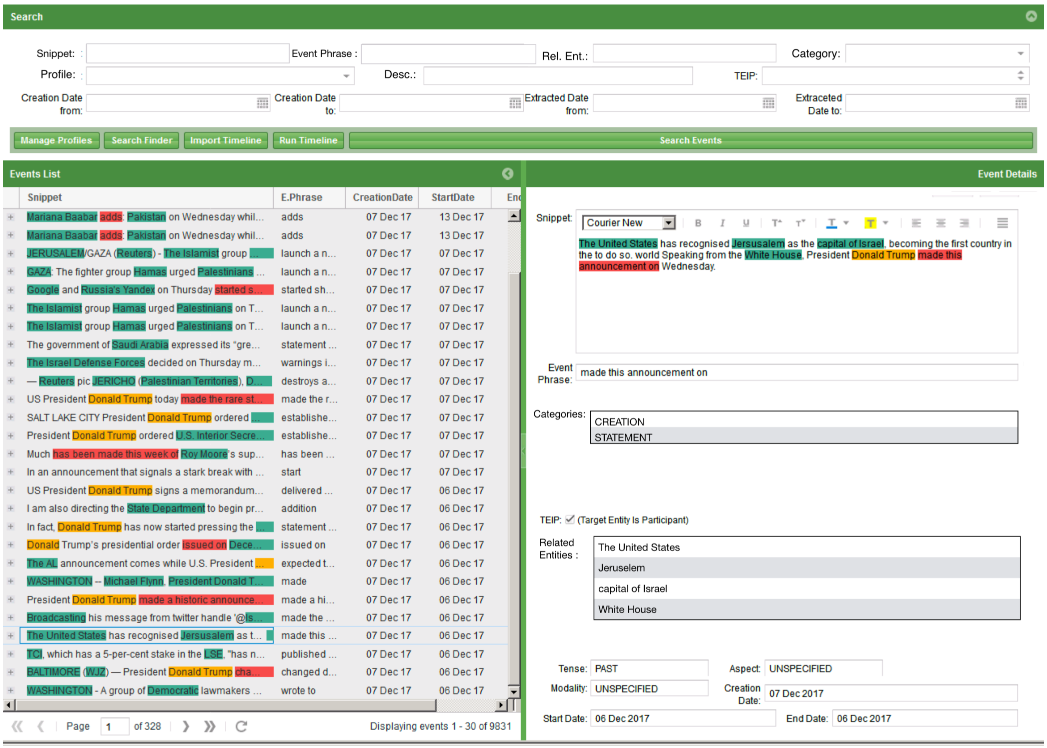
Most likely thanks to high coverage of the lexical resources exploited both precision (96%) and recall (91%) of Related-Entity extraction were high. Finally, 33% of the events were anchored to a time reference, whereas 63% of those references were correct. This is due to the fact that, while event-mapped only at a sentence level, time references are computed by some form of discourse analysis spanning the full document text. Some other aspects of the quality of the extraction needs yet to be evaluated, e.g., event duplicate detection, and recall of event detection. The results of the evaluation are summarized in Table 2.

Table 2: Evaluation results.

category	performance
Event detection accuracy (all)	97%
(only factual)	81%
Fraction of events assigned at least one category	70%
Event categorisation accuracy	91%
Target-Entity-Event-Linking accuracy (strict)	76%
(lenient)	74%
Related Entity Extraction precision	96%
recall	91%
Event-Time Anchoring coverage	33%
precision	63%

6 Future Outlook

The architecture and methods deployed in the presented tool lay the basis for future enhancements. The language specific resources deployed are generally shallow and some are suitable to be acquired semi-automatically through ML algorithms. Moreover the system integrated some modules (e.g. Entity Matching, Temporal Normalization) that are already multilingual. Therefore the extension of the tool across languages is envisaged.



Snippet	E.Phrase	CreationDate	StartDate
Mariana Baabar adds: Pakistan on Wednesday whil...	adds	07 Dec 17	13 Dec 17
Mariana Baabar adds: Pakistan on Wednesday whil...	adds	07 Dec 17	13 Dec 17
JERUSALEM/GAZA (Reuters) - The Islamist group ...	launch a n...	07 Dec 17	07 Dec 17
GAZA: The fighter group Hamas urged Palestinians ...	launch a n...	07 Dec 17	07 Dec 17
Google and Russia's Yandex on Thursday started s...	started sh...	07 Dec 17	07 Dec 17
The Islamist group Hamas urged Palestinians on T...	launch a n...	07 Dec 17	07 Dec 17
The Islamist group Hamas urged Palestinians on T...	launch a n...	07 Dec 17	07 Dec 17
The government of Saudi Arabia expressed its “gre...	statement ...	07 Dec 17	07 Dec 17
The Israel Defense Forces decided on Thursday m...	warnings i...	07 Dec 17	07 Dec 17
— Reuters pic JERICHO (Palestinian Territories), D...	destroys a...	07 Dec 17	07 Dec 17
US President Donald Trump today made the rare st...	made the r...	07 Dec 17	07 Dec 17
SALT LAKE CITY President Donald Trump ordered ...	establishe...	07 Dec 17	07 Dec 17
President Donald Trump ordered U.S. Interior Secre...	establishe...	07 Dec 17	07 Dec 17
Much has been made this week of Roy Moore's sup...	has been ...	07 Dec 17	07 Dec 17
In an announcement that signals a stark break with ...	start	07 Dec 17	07 Dec 17
US President Donald Trump signs a memorandum...	delivered ...	07 Dec 17	06 Dec 17
I am also directing the State Department to begin pr...	addition	07 Dec 17	06 Dec 17
In fact, Donald Trump has now started pressing the ...	statement ...	07 Dec 17	06 Dec 17
Donald Trump's presidential order issued on Dece...	issued on	07 Dec 17	06 Dec 17
The AL announcement comes while U.S. President ...	expected t...	07 Dec 17	06 Dec 17
WASHINGTON -- Michael Flynn, President Donald T...	made	07 Dec 17	06 Dec 17
President Donald Trump made a historic announce...	made a hi...	07 Dec 17	06 Dec 17
Broadcasting his message from twitter handle '@Is...	made the ...	07 Dec 17	06 Dec 17
The United States has recognised Jersusalem as t...	made this ...	07 Dec 17	06 Dec 17
TCI, which has a 5-per-cent stake in the LSE, "has n...	published ...	07 Dec 17	06 Dec 17
BALTIMORE (WJZ) — President Donald Trump cha...	changed d...	07 Dec 17	06 Dec 17
WASHINGTON - A group of Democratic lawmakers ...	wrote to	07 Dec 17	06 Dec 17

Page 1 of 328 Displaying events 1 - 30 of 9831

Snippet: Courier New
The United States has recognised Jersusalem as the capital of Israel, becoming the first country in the to do so. world Speaking from the White House, President Donald Trump made this announcement on Wednesday.

Event Phrase: made this announcement on

Categories: CREATION
STATEMENT

TEIP: ☑ (Target Entity Is Participant)

Related Entities : The United States, Jeruselem, capital of Israel, White House

Tense: PAST Aspect: UNSPECIFIED
Modality: UNSPECIFIED Creation Date: 07 Dec 2017
Start Date: 06 Dec 2017 End Date: 06 Dec 2017

Figure 3: Web-based interface for browsing and visualisation of event information.

Furthermore, we currently work on improving various modules, including, i.a., (a) exploring ML approaches to improve the quality of linking of target entities to events ('participation' relation extraction), (b) improving the model for computing event duplicates and linking related events, not necessarily being duplicates, and (c) exploiting event trigger contextual information in order to improve event categorisation. In the long term, i.e., once the above improvements are implemented, we aim at implementing a cross-lingual linking of timelines as well. Finally, elaboration of methods for semi-automated inferring of domain-specific event categories, using clustering and distributional similarity metrics is envisaged too.

References

[BCS+07] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. Open information extraction from the web. In *Proceedings of IJCAI 2007*, pages 2670–2676. Morgan Kaufmann Publishers Inc., 2007.

[Bor18] Emanuela Boros. *Neural Methods for Event Extraction*. PhD thesis, 2018.

[CJSP17] Sophie Chesney, Guillaume Jacquet, Ralf Steinberger, and Jakub Piskorski. Multi-word entity classification in a highly multilingual environment. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 11–20. ACL, 2017.

[CSCXO09] Daniela Barreiro Claro, Marlo Souza, Clarissa Castell Xavier, and Leandro Oliveira. Multilingual open information extraction: Challenges and opportunities. *Information-an International Interdisciplinary Journal*, 10(7), 2009.

[CSE11] Janara Christensen, Stephen Soderland, and Oren Etzioni. An analysis of open information extraction based on semantic role labeling. In *Proceedings of the Sixth International Conference on Knowledge Capture*, pages 113–120. ACM, 2011.

[CV16] Savelie Cornegruta and Andreas Vlachos. Timeline extraction using distant supervision and joint inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1936–1942, 2016.

[Dah17] Daniel Dahlmeier. On the challenges of translating NLP research into commercial products. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–96. Association for Computational Linguistics, 2017.

[DBL19] Proceedings of Text2Story - 2nd workshop on narrative extraction from texts, co-located with the 41st european conference on information retrieval, Text2Story@ECIR2019, 2019.

[DJT17] Yijun Duan, Adam Jatowt, and Katsumi Tanaka. Discovering typical histories of entities by multi-timeline summarization. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT 17, pages 105–114. Association for Computing Machinery, 2017.

[DMBZ10] Gianluca Demartini, Malik Muhammad Saad Missen, Roi Blanco, and Hugo Zaragoza. TAER: Time-aware entity retrieval-exploiting the past to find relevant entities in news articles. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1517–1520, 2010.

[EFC$^+$11] Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam Mausam. Open information extraction: The second generation. In *Proceedings of IJCAI 2011*, pages 3–10. AAAI Press, 2011.

[EJS17] Maud Ehrmann, Guillaume Jacquet, and Ralf Steinberger. JRC-Names: Multilingual entity name variants and titles as Linked Data. *Semantic Web*, 8(2):283–295, 2017.

[Erj10] Toma Erjavec. Multext-east version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proceedings of LREC 2010*. ELRA, 2010.

[FSE11] Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. ACL, 2011.

[GHMS14] Luis Galrraga, Geremy Heitz, Kevin Murphy, and Fabian M Suchanek. Canonicalizing open knowledge bases. In *Proceedings of the 23rd Acm International Conference on Conference on Information and Knowledge Management*, pages 1679–1688. ACM, 2014.

[GWH$^+$19] Kiril Gashteovski, Sebastian Wanner, Sven Hertling, Samuel Broscheit, and Rainer Gemulla. {OPIEC}: An open information extraction corpus. In *Automated Knowledge Base Construction (AKBC)*, 2019.

[LAAR17] Egoitz Laparra, Rodrigo Agerri, Itziar Aldabe, and German Rigau. Multi-lingual and Cross-lingual timeline extraction. *Knowledge-Based Systems*, 133:77–89, 2017.

[Mau16] Mausam. Open information extraction systems and downstream applications. In *Proceedings of IJCAI 2016*, pages 4074–4077. AAAI Press, 2016.

[MBM19] Daniele Metilli, Valentina Bartalesi, and Carlo Meghini. Steps towards a system to extract formal narratives from text. In *Proceedings of Text2Story - 2nd Workshop on Narrative Extraction from Texts, Co-Located with the 41st European Conference on Information Retrieval, Text2Story@ECIR 2019, Cologne, Germany, April 14th, 2019*, pages 53–61, 2019.

[MRM$^+$04] Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. Annotating noun argument structure for NomBank. In *LREC*, volume 4, pages 803–806, 2004.

[MSA$^+$15] Anne-Lyse Myriam Minard, Manuela Speranza, Eneko Agirre, Itziar Aldabe, Marieke van Erp, Bernardo Magnini, German Rigau, and Ruben Urizar. Semeval-2015 task 4: Timeline: Cross-document event ordering. In *9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 778–786, 2015.

[NP12] Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

[NWS12] Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. PATTY: A taxonomy of relational patterns with semantic types. In *Proceedings of EMNLP and CoNLL 2012*, pages 1135–1145. Association for Computational Linguistics, ACL, 2012.

[PCI$^+$03] James Pustejovsky, Jos M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. TimeML: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34, 2003.

[Pis19] Jakub Piskorski. Corleone: Core linguistic entity online extraction: Revised and enhanced version. Technical Report, Joint Research Centre of the European Commission, 2019.

[PŠZA18] Jakub Piskorski, Fredi Šarić, Vanni Zavarella, and Martin Atkinson. On training classifiers for linking event templates. In *Proceedings of the Workshop Events and Stories in the News 2018*, pages 68–78, Santa Fe, New Mexico, U.S.A, August 2018. Association for Computational Linguistics.

[PWS09] Jakub Piskorski, Karol Wieloch, and Marcin Sydow. On knowledge-poor methods for person name matching and lemmatization for highly inflectional languages. *Information retrieval*, 12(3):275–299, 2009.

[RAG$^+$17] Steinberger Ralf, Martin Atkinson, Teofilo Garcia, Erik van der Goot, Jens Linge, Charles Macmillan, Hristo Tanev, Marco Verile, and Gerhard Wagner. EMM: Supporting the analyst by turning multilingual text into structured data. In *Transparenz Aus Verantwortung: Neue Herausforderungen Fr Die Digitale Datenanalyse*. Erich Schmidt Verlag, 2017.

[ROR13] Ridho Reinanda, Daan Odijk, and de M Rijke. Exploring entity associations over time. In *Proceedings of TAIA 2013*. ACM, 2013.

[RSH$^+$16] Thomas Rebele, Fabian Suchanek, Johannes Hoffart, Joanna Biega, Erdal Kuzey, and Gerhard Weikum. YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames. In Paul Groth, Elena Simperl, Alasdair Gray, Marta Sabou, Markus Krtzsch, Freddy Lecue, Fabian Flck, and Yolanda Gil, editors, *The Semantic Web ISWC 2016: 15th International Semantic Web Conference*, pages 177–185. Springer International Publishing, 2016.

[SLK$^+$06] Roser Saur, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. TimeML annotation guidelines. 1(1):31, 2006.

[STSO16] Pedro Saleiro, Jorge Teixeira, Carlos Soares, and Eugnio C. Oliveira. TimeMachine: Entity-centric search and visualization of news archives. In *Proceedings of ECIR 2016*, volume 9626 of *LNCS*, pages 845–848. Springer, Springer, 2016.

[TM00] Kristina Toutanova and Christopher D Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 63–70. ACL, 2000.

[WHS16] Gerhard Weikum, Johannes Hoffart, and Fabian M Suchanek. Ten years of knowledge harvesting: Lessons and challenges. *IEEE Data Engineering Bulletin*, 39(3):41–50, 2016.

[WSY17] Yaguang Wu, Haichun Sun, and Chungang Yan. An event timeline extraction method based on news corpus. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, pages 697–702, 2017.

[ZT13] Vanni Zavarella and Hristo Tanev. FSS-TimEx for TempEval-3: Extracting temporal information from text. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 58–63, 2013.