Knowledge-Based Systems 000 (2016) 1–26

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

NewsReader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news

Piek Vossen[a,*], Rodrigo Agerri[b], Itziar Aldabe[b], Agata Cybulska[a], Marieke van Erp[a], Antske Fokkens[a], Egoitz Laparra[b], Anne-Lyse Minard[c], Alessio Palmero Aprosio[c], German Rigau[b], Marco Rospocher[c], Roxane Segers[a]

[a] Vrije Universiteit Amsterdam, the Netherlands
[b] IXA NLP Group, University of the Basque Country (UPV/EHU), Donostia-San Sebastián, Spain
[c] Fondazione Bruno Kessler, Trento, Italy

ARTICLE INFO

Article history:
Received 15 November 2015
Revised 20 June 2016
Accepted 7 July 2016
Available online xxx

Keywords:
Natural language processing
Semantic web
Knowledge resources
Event extraction
Cross-lingual interopearbility

ABSTRACT

In this article, we describe a system that *reads* news articles in four different languages and detects what happened, who is involved, where and when. This event-centric information is represented as episodic situational knowledge on individuals in an interoperable RDF format that allows for reasoning on the implications of the events. Our system covers the complete path from unstructured text to structured knowledge, for which we defined a formal model that links interpreted textual mentions of things to their representation as instances. The model forms the skeleton for interoperable interpretation across different sources and languages. The real content, however, is defined using multilingual and cross-lingual knowledge resources, both semantic and episodic. We explain how these knowledge resources are used for the processing of text and ultimately define the actual content of the episodic situational knowledge that is reported in the news. The knowledge and model in our system can be seen as an example how the Semantic Web helps NLP. However, our systems also generate massive episodic knowledge of the same type as the Semantic Web is built on. We thus envision a cycle of knowledge acquisition and NLP improvement on a massive scale. This article reports on the details of the system but also on the performance of various high-level components. We demonstrate that our system performs at state-of-the-art level for various subtasks in the four languages of the project, but that we also consider the full integration of these tasks in an overall system with the purpose of *reading* text. We applied our system to millions of news articles, generating billions of triples expressing formal semantic properties. This shows the capacity of the system to perform at an unprecedented scale.

1. Introduction

We massively communicate about the changes in the world through news and social media. This is mostly done in natural language. LexisNexis,[1] a multinational information broker, estimates that they archive over 1 million news articles from 30,000 different sources every working day and almost the same number of web pages. Their multilingual archive goes back several decades and contains billions of articles, biographies, and reports. Such an archive contains a wealth of knowledge and information on what happened in the world and what our perspective is on reported changes. Tapping into this knowledge would allow us to discover long-term developments at a global-scale. It would show us the global history and its impact as reported in all these media. However, this knowledge from text is currently only revealed through search and classification systems that give users a list of news articles in response to profiled queries. At most, such systems provide trending topics in time, typically exploiting the volume of news but they do not measure the volume of change in the world nor its impact. To find out what really happened, users are still forced to *read* news articles and social blogs from these clusters or result lists.

* Corresponding author.
E-mail address: piek.vossen@vu.nl (P. Vossen).
[1] http://www.lexisnexis.com.

http://dx.doi.org/10.1016/j.knosys.2016.07.013

In this article, we describe a system built in the NewsReader project[2] that does the reading for you in four different languages: English, Dutch, Spanish and Italian. It determines *what* happened, *who* was involved, and *where* and *when* it took place. The interpretation of natural language text is formally represented in RDF according to Semantic Web standards in the form of what we call Event-Centric-Knowledge-Graphs (ECKGs, [1]). Because we anchor events to time, we can extract longer sequences of events over time and discover the role of participants in history. It allows us to find networks of actors and implications of events on a large global scale and over longer periods of time. It provides the means to generalize from the level of individuals (people, companies, incidents) to classes and types (management, governors, industries and event types), discovering trends and patterns, or vice versa to specialize from general trends to personal stories.

By processing text to RDF, we make a fundamental distinction between the mentions of individuals and events in text and their more formal representation as instances in the reconstruction of a world. Many different sources will mention the same people, organizations and events many times but there is only a single world representation of instances to which these references should be linked. We defined the Grounded Annotation Framework (GAF, [2]) to bind mentions in text (and other modalities) to instance representations. GAF naturally resolves coreference of natural language expressions (mentions) within documents but also across documents and across languages. It also allows us to aggregate knowledge and information from these mentions, to deduplicate information, and to show the provenance and perspective on each piece of information from these sources. Whereas the Natural Language Processing (NLP) technology that we employ tries to interpret each mention semantically, a separate technology has been developed to aggregate these interpretations into a single RDF structure and represent this in a triple store for reasoning and exploitation.

The deep-reading technology developed by NewsReader is unique in its kind. It combines the most advanced NLP technology in four different languages to obtain interoperable semantic interpretations of text. Our four NLP pipelines perform named-entity-detection and linking, event and semantic role detection, temporal expression normalization and temporal relation detection. Interoperability across these modules is achieved through the Natural Language Annotation Format (NAF, [3]). Cross-source interoperability is achieved through the Simple Event Model (SEM [4]). The two formalisms are further integrated in the Grounded Representation and Source Perspective (GRaSP) framework. GRaSP allows us to formally model the sources, authors or owners of the text and their perspective, reflected by factuality and sentiment values, on what happened. GRaSP is compatible with the PROV-DM [5].

The NewsReader system currently exploits multiple knowledge bases, multilingual semantic resources and ontologies including episodic knowledge in the form of encyclopaedic facts about individuals. In addition, we also apply statistical language models for semantic classification tasks but also formal reasoners that take semantic representations of situations as input and derive implications from these situations through rules defined in our ontology. Finally, we generate new episodic knowledge on events and entities that is not represented in the background encyclopedia. This knowledge can be used to extend the same resources that are used for NLP, such as semantic resources on entities. Adding this knowledge results in better coverage and better interpretations when further (re-)processing the news.

The NewsReader system has been applied to millions of news articles and generated billions of RDF triples capturing event data for long periods of time (decades). We also processed news across four different languages, resulting in unified interoperable data across these languages. The knowledge resulting from the processed is stored in a dedicated KnowledgeStore that supports various APIs for semantic querying and exploitation of the data.

Our main contributions in this article are: 1) a formal model of semantic representations of mentions and instances in a uniform framework that is interoperable across languages and can handle interpretations across documents, 2) a way of modeling event data that allows for reasoning over episodic situations and deriving implications on individuals, 3) state-of-the-art performance for high-level semantic NLP modules in four different languages that exploit shared cross-lingual knowledge resources, 4) the capacity to process millions of news articles to generate episodic knowledge that extends existing resources in the Semantic Web and can be used to create new knowledge resources, and finally 5) a successful marriage between NLP and Semantic Web technology.

In this article, we describe the knowledge architecture and interaction with the text understanding process. We will first discuss the background and related work on semantic processing of text in Section 2 and explain the main differences of our approach. In Section 3, we give an overview of the system architecture, the representation formats and types of knowledge used. Next, we describe NLP pipelines for the four languages (Section 4) that use the knowledge and formats to generate interoperable semantic text representations. In Section 5, we explain the conversion of the mention-based NLP output to the instance representation in SEM, making reference to existing episodic and semantic knowledge. Section 6 explains how the semantic resources can be adapted to the domain to define these episodic situations more precisely. We explain how we deal with so-called *dark entities*, which are entities not (properly) represented in our knowledge resources, and how we built an event ontology to precisely model the implications of events for a domain. In Section 7, we explain the KnowledgeStore that stores the output of the reading process and we discuss the various data sets that we processed in the project. We also show evidence for the scalability of the system to deal with massive streams of news. The quality of the output is discussed in Section 8. We not only evaluated the most important NLP modules against standard data sets and our own data sets but also the RDF data that is derived from the output of these NLP modules. Finally, in Section 9 we discuss the status of the system and in Section 10 we provide some final conclusions.

2. Background and motivation

Knowledge bases are used extensively in NLP, from high-level tasks such as question answering [6] to low-level tasks such as spelling correction [7]. However, some NLP research aims at deep reading to understand the text with regard to the real world, as represented by the news [8]. Besides information extraction challenges that focus on textual news data, several projects have been devoted to summarize the news such as the European Media Monitor,[3] the Ontos News Portal,[4] and the XLike project [9].[5]

Within the field of information extraction, there are two main directions: closed information extraction, where the system is required to fill slots in a predefined template and open information extraction, where the concepts or types of relations that the system is required to extract are not predefined. Patwardhan and Riloff [10] is an example of closed information extraction. The authors use a machine learning approach to recognize events and

[2] http://www.newsreader-project.eu.

[3] http://emm.newsbrief.eu/NewsBrief/clusteredition/en/latest.html Last accessed: 12 November 2015.

[4] http://news.ontos.com/ Last accessed: 12 November 2015.

[5] http://www.xlike.org/ Last accessed: 12 November 2015.

associated roles which they evaluate on a dataset about terrorism attacks and disease outbreaks. Such systems often employ some form of background knowledge to limit the types of knowledge that are extracted. An example of an open IE system is NELL [11] which continuously crawls the web and tries to extract factoids about any possible topic. Because there is no gold standard dataset to evaluate such a system, they are often evaluated post-hoc. We do not know in advance what events and entities we may encounter in the news, the NewsReader system thus performs open information extraction. As Semantic Web technology also plays an important role in NewsReader, we can still employ the vast amount of open domain knowledge which is traditionally not the type of knowledge that is contained in the carefully curated knowledge bases that are employed in NLP.

The advent of easily accessible resources such as Wikipedia (and later DBpedia) made it possible to link any type of information. In contrast, previous efforts only linked specific databases, for example, databases with geospatial information (see Leidner et al. [12]). In Mihalcea and Csomai [13], Wikipedia pages were used as an index to identify interesting concepts in a text to ground. Milne and Witten [14] take the use of Wikipedia a step further and also take a part of Wikipedia that they use as training data for a machine learning approach for word-sense disambiguation. With linked open data (LOD), approaches that ground text mentions to LOD sources have taken flight [15] as sources becoming more and more available. One of these approaches, also used in NewsReader, is DBpedia Spotlight [16]. It establishes links between concepts in text and the DBpedia resource. In its slipstream, similar approaches either utilizing DBpedia or other generic databases such as Freebase have become available individually [17], allowing the user to choose which knowledge base to link to [15].

There are two main things that set the NewsReader project apart from prior work. First, there is a strong focus on episodic knowledge in the NewsReader project. While entities are important in this domain, events are the central focus of the knowledge base, and thus we do not only ground concepts and entities to external knowledge bases, but we also ground events and reason over them, i.e. we follow an event-centric approach. Second, NewsReader employs deep NLP as well as state-of-the-art Semantic Web technology, resulting in much more fine-grained analyses than projects that employ only shallow natural language processing or focus on a single NLP task. Our analysis targets all the content of the text and not a limited set of predefined properties. This allows us to make information explicit that is only implicitly present in the text sources and to store that as queryable information in a knowledge base. Likewise, we represent events externally from the text and aggregate the knowledge and information in the event from many different textual sources. In the next section, we outline the architecture of the system that makes this possible in more detail.

3. System and knowledge architecture

We divide the task of interpreting text and representing it in event-centric RDF in three main steps illustrated in Fig. 1. The first step applies various advanced linguistic analyses to single documents, the second translates the output of these analyses to RDF resolving mentions of information to an instance representation and the third and final step aggregates the RDF instance representations across different documents into a joined RDF representation. The steps are described in more detail below.

3.1. From text to RDF

Textual sources are processed one-by-one through a series of advanced NLP modules (Step 1). Each module applies a different interpretation task to the text and stores the interpretation in a separate layer in the Natural language processing Annotation Format (NAF) [3]. We built four pipelines to analyze text in four different languages. Although all pipelines contain language-specific modules, their output is interoperable through the uniform semantic representation in NAF. For each text, our system generates a separate NAF structure to represent the interpretation of the mentions in that text.

Modules apply the analyses to tokens or sentences in the order in which they appear in the raw text. In the first step of NLP processing, each token receives an identifier and its offsets in the original text are stored. Tokens are ultimately annotated with semantic interpretations such as references to concepts, events, entities, time expressions and relations. These can be mentioned several times throughout a text. If they have the same referent, they are grouped together in a coreference layer in NAF. These interpretation steps are described in Section 4.

In Step 2, NAF files are read as input and converted to RDF. The interpretations of events, entities, time-expressions and the relations between them are represented in RDF according to the Simple Event Model (SEM) [4]. SEM represents *what* is happening in the world, *who* is involved and *when* and *where* this happens. It thus represents instances, i.e. individual events, entities and time expressions in an assumed or real world. We link instances in SEM to their mentions in NAF where they originate from using *denotedBy* relations defined in the Grounded Annotation Framework (GAF). In other words, when an instance is *denotedBy* a specific mention, this means that this mention refers to the instance. GAF thus provides a natural way to capture coreference: mentions that corefer are all linked to the same instance in SEM. We will illustrate the relation between instances and mentions in Section 3.3.

Entities and time-expressions in SEM are related to events, which results in an event-centric representation. In Step 2, we incorporate coreference between mentions coming from the same document. In Step 3, we use our event-centric representation to establish which of the events and entities extracted from different documents are identical. If cross-document coreference is established, information from both documents is combined in a single representation. This leads to deduplication of shared information and aggregation of complementary information. In case of alternative views or conflicting information, we present the different perspectives of each mention. This procedure is further explained in Section 5. We illustrate the process and representation with an example below.

3.2. Two perspectives on a sale

Consider the following two examples of textual input. The examples are the titles of two news articles published on the same day: 17 June 2013:

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://.english.alarabiya.net)

Both example sentences express the same event information: Porsche (or the Porsche family) buying Porsche stakes back from Qatar, but they use different constructions and words. If our software interprets these sentences correctly, this should result in the same representation of content, which can then be merged across these sentences. To achieve this, our software first needs to find the participants in these sentences: *Qatar, Qatar Holding, Porsche family, founding families, 10% stake in Porsche* and *10pc stake*. Identity of the participants is established by assigning URIs to each mention. If these URIs are identical, the entities are also identical. Similarly, the mentions of actions *buy* and *sell* need to be de-

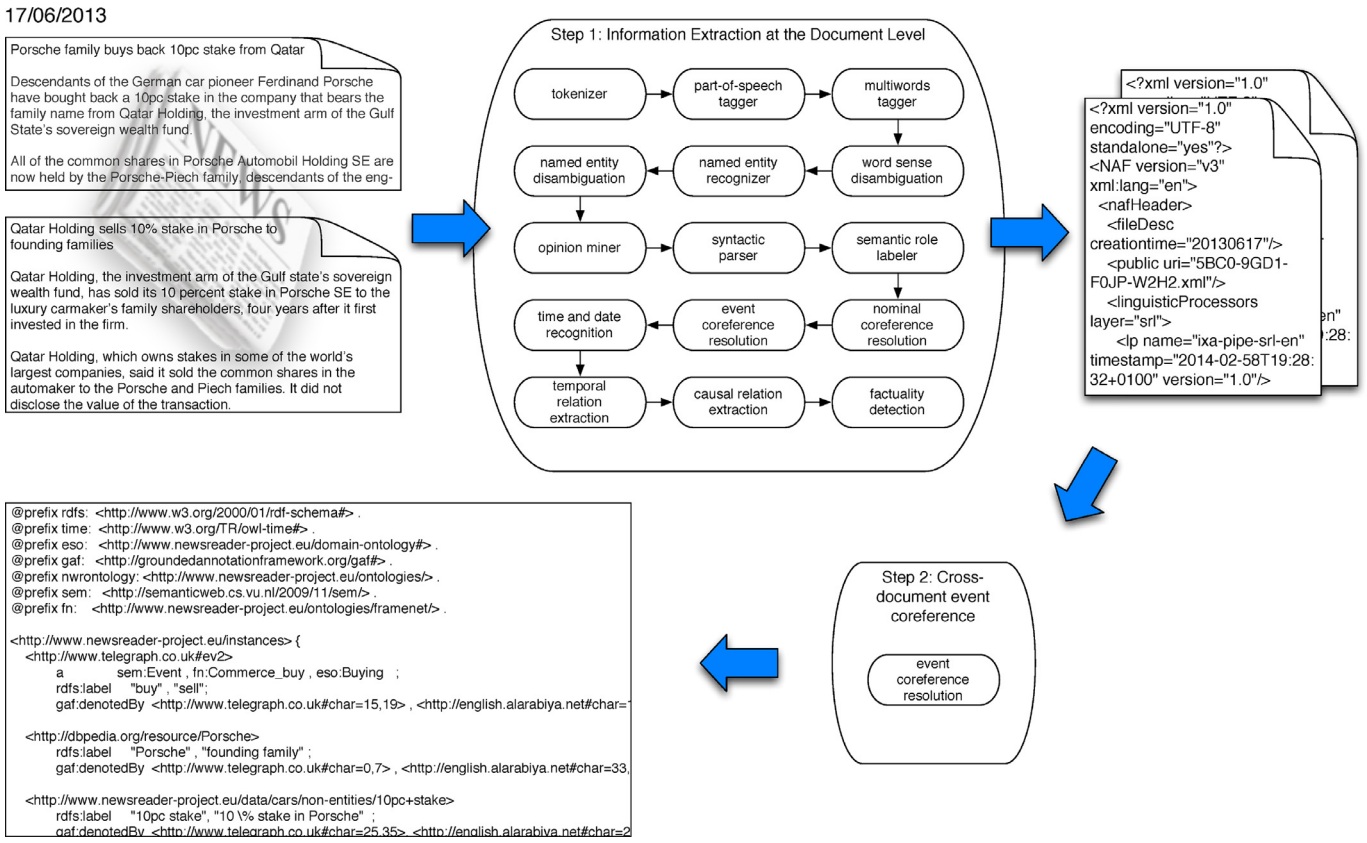

Fig. 1. Overview of the NewsReader system taking raw text as input, creating XML representations (NAF) for NLP output and creating RDF representations (SEM) from the XML representations.

tected, where identity can be established based on their relatedness in meaning and the fact that these events are reported to occur on the same day. Finally, the exact semantic relations between the participants and the actions are determined, since it makes a difference who is the *buyer* and who is the *seller*. Identity of the complete event is then based on the identity of the components of each mention, i.e. we only assume we are dealing with the same event if we are dealing with the same kind of event that takes place at the same time and in which the same participants play exactly the same role. The more overlap we find, the stronger the evidence that two texts discuss the same event.

3.3. Data representation

Fig. 2 is a simplified illustration of the final interpretation of the two examples in GRaSP.[6] The information from the two different sources is merged in a single RDF structure. The top level of the image shows the representation of the event in RDF according to the SEM model, it provides a representation of the *instances* involved. It shows the single representation of the event instance *#Ev2*. It is linked to the unique entities playing a role in the event (Porsche, Qatar and 10% stake), which are represented by their DBpedia [18] URIs and a generated URI for *10% stake*. Background information from DBpedia provides further information about the main participants (see Section 3.4 for more information about DBpedia). The event instance is further linked to other ontologies that indicate what kind of event we are dealing with (Commerce_buy/Buying). These types also define the possible roles of individual participants in such events, where some are filled in (e.g. Seller/owner_1, Buyer/owner_2) but others are not yet known, e.g. the amount of money paid for the 10% share.[7]

In our model, we require that all events are bound in time. In this example, the events are linked to the document creation time, which is 17 June 2013. *Qatar* selling *10%* to *the Porsche family* at another point in time is by definition another event and therefore involves another *10%*. Finally, the relation triples are presented within separate boxes that have their own identifiers. These boxes represent named graphs, which allow us to link each relation between a participant and an event separately to individual sources that mention them. If another source in the future states how much was paid for the 10%, the model allows us to fill in the missing information in the same picture at the instance level and we can still trace back which source provided what information when.

Representing the event as an instance in SEM implies that we express properties of the event rather than properties of entities, e.g.*#Ev2 sem:hasActor dbp:Porsche*. This effectively results in event centric knowledge graphs or ECKGs [1] as opposed to the entity-centric knowledge graphs that you find in resources such as DBpedia. The difference is illustrated in Figs. 3 and 4. In Fig. 3, the entity is the subject of the relation triple whereas *ownership* and *working as a key person* are properties and the objects are the values for the entity. In Fig. 4, we see that the same properties are represented as event instances in NewsReader in the subject position and the entities that participate are in the object position, whereas we use abstract roles for the predicates.

[6] Recall from the introduction, that GRaSP includes SEM, GAF relations to mentions and information on sources and perspectives.

[7] For reasons of space, we only represent all roles from FrameNet in the illustration. The Event and Situation Ontology (see Section 6.2) also provides information about all four elements and their roles.

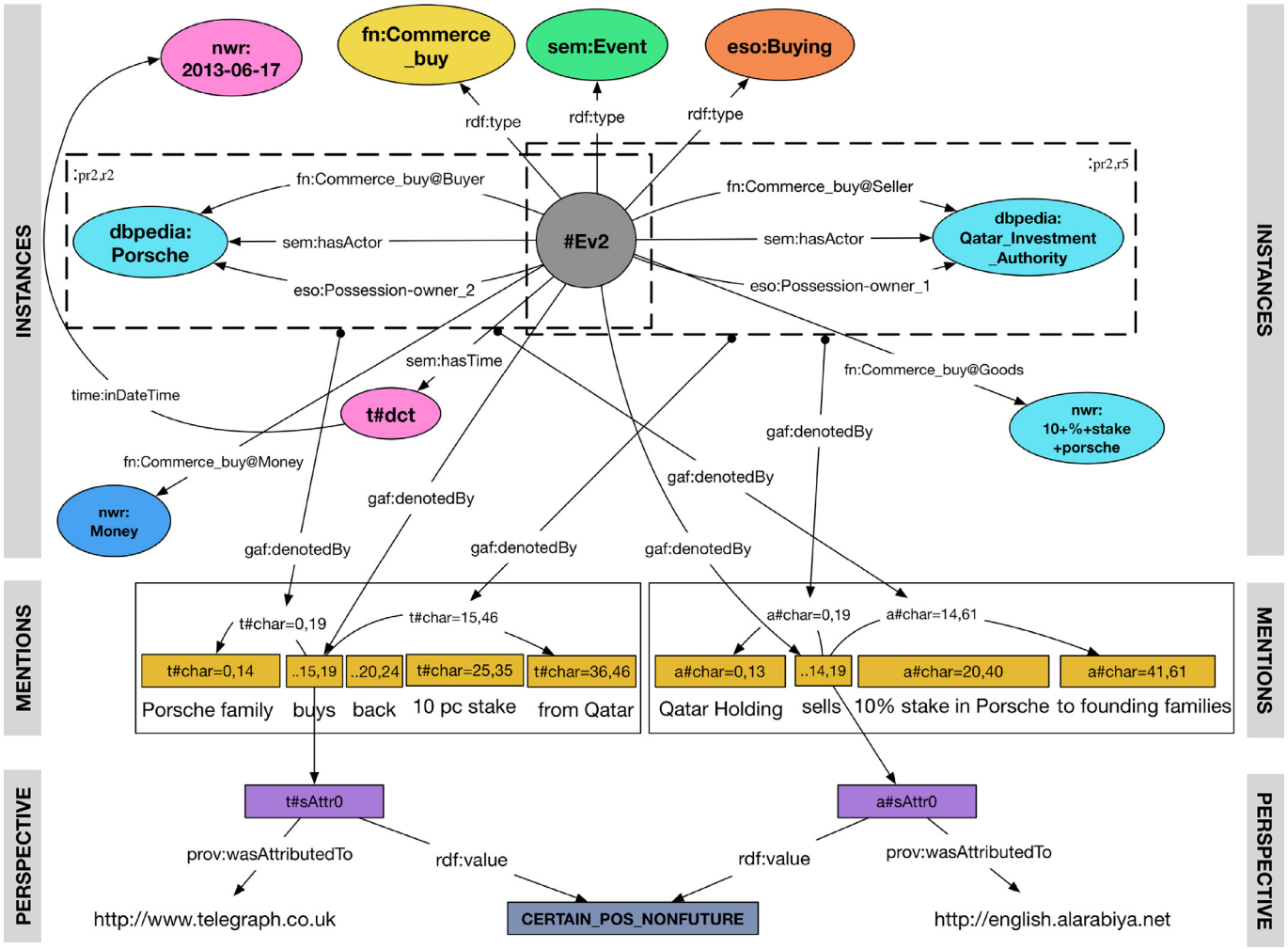

Fig. 2. Illustration of the event-centric representation of a single event involving two main participants, links to ontologies and links to their mentions in the source texts with the source perspective. It shows the modeling according to GRaSP.

```
dbp:Porsche
        dbp:keyPeople dbr:Martin_Winterkorn ,
        dbp:owner dbr:Volkswagen_Group .
```

Fig. 3. Entity centric RDF triples in DBpedia.

```
#owns23
        eso:owner dbr:Volkswagen_Group ,
        eso:property dbp:Porsche ,
        sem:hasBeginTime 2009.

#works25
        eso:employee dbr:Martin_Winterkorn ,
        eso:employer dbp:Porsche ,
        sem:hasBeginTime 2007 .
```

Fig. 4. Event centric RDF triples in NewReader.

The difference between event and entity centric representations is a small syntactic change but has major consequences. As you can see in Fig. 4, ECKGs can be bound to time and can involve multiple entities, whereas the entity centric information shows the latest published data only, i.e. the current CTO or management and not the past. ECKGs are also more expressive, because we can accommodate for an infinite number of event instances, each bound in time.

The boxes underneath the instance layer in Fig. 2 represent the mention layer with the original source texts. Elements in the upper instance boxes are linked to elements from the mention layer through *gaf:denotedBy* relations. These mention elements have unique identifiers that resolve to the specific character offsets in the source. We thus link instances of events or entities to the offset places in text where they are mentioned. Also the relations are linked to the places where they are mentioned, using the named graph identifiers of the relations, i.e. the two named graphs *:pr2,r2* and *:pr2,r5* are instances of relations.[8] The distinction between instances and mentions thus gives us immediate access to sources that talk about specific entities or events.

The formal representation of mentions in GRaSP is not only used to indicate where specific information comes from through GAF. We also use it to distinguish different perspectives on the same event. This is shown in the bottom box of Fig. 2. We consider choices about what information is included and left out as part of the perspective of a source. Therefore, the fact that these sources provide information about the sale of Porsche stakes is part of the source's perspective. Both sources state their belief that this took place and is true (NONFUTURE and POSITIVE) and both are also certain about it. This perspective is expressed by an attribution object that has the *rdf:value* CERTAIN_POSITIVE_NONFUTURE and is linked to the two sources that provide the information using the *prov:wasAttributedTo* relation.[9] The perspective layer allows us to model other opinions on events expressed in text as well, such as uncertainty, placing it in the future, negating it or expressing certain emotions. We can thus also represent the conflict of information if another source would deny the event took place. The

[8] The actual name is irrelevant as long as it is unique.

[9] *prov:* is shorthand for http://www.w3.org/ns/prov#.

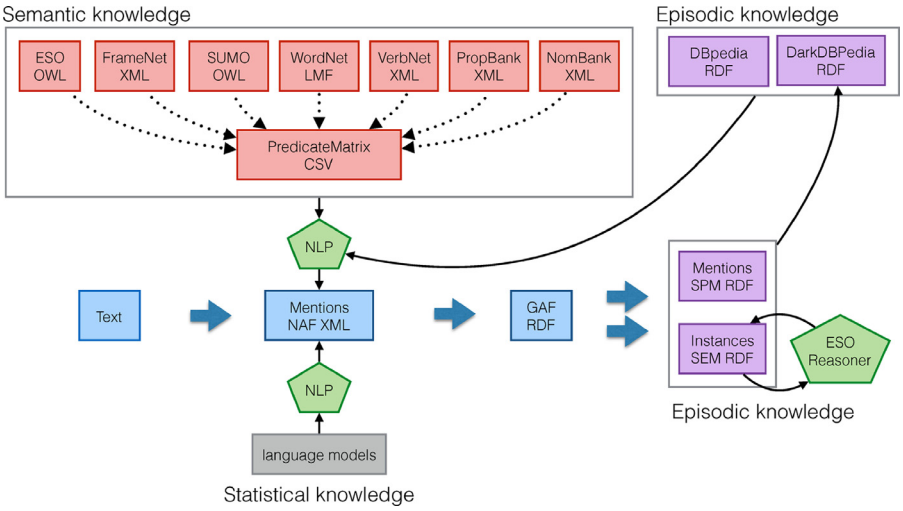

Fig. 5. Overview of different types of knowledge used and generated by the NewsReader processing pipeline.

model allows us to organize various perspectives from many different sources on the same event-centric representation of information. We will not discuss the GRaSP framework further in this article. More details can be found in Vossen et al. (2015) [19].

3.4. Resources and semantic frameworks

To achieve conceptual interoperability, the system relies heavily on shared semantic resources and shared formats. We shortly discussed the formats for representing interpretations of mentions and linguistic analyses (NAF) and instances (SEM) and the method of connecting the two (GAF) as well as the format for representing the perspective and provenance relations (GRaSP). Now we present the shared resources we use both to identify meaning of text and represent this meaning at the instance level.

Fig. 5 gives an overview of the different types of knowledge that play a role in the interpretation at different points in the process. The top of the image shows a range of lexical resources and ontologies that represent semantic knowledge. Semantic knowledge is represented in the form of concepts and relations. This knowledge is shared across languages through multilingual vocabularies. We use the following resources:

WordNet: WordNet [20] provides a lexical database where words are represented as groups of synonyms (synsets) that are organized in a hypernym hierarchy. Wordnets in other languages are used to establish cross-lingual connections to other semantic resources.

PropBank: PropBank [21] provides predicate-argument structures on top of the Penn TreeBank [22]. The annotations resulted in a lexicon of predicates and their argument structure.

NomBank: NomBank [23] is related to PropBank and provides argument structure of common nouns in the Penn TreeBank.

FrameNet: FrameNet [24] uses *semantic frames* to provide information about events (the frames) and the relations they invoke with participants (frame elements).

VerbNet: VerbNet [25] is a verb lexicon providing information on thematic roles and semantic restrictions on the verb's arguments. It is linked to WordNet and FrameNet.

AnCora: AnCora [26] includes a lexicon for Spanish and Catalan that includes argument structures for verbal and nominal predicates in those languages.

SUMO: The Suggested Upper Merged Ontology [27] is a formal ontology that defines concepts and relations between them. The full WordNet lexicon has been mapped to SUMO.

ESO: The Event and Situation Ontology [28] captures implications of events and maps these to FrameNet frames, SUMO types and roles. ESO is further described in Section 6.

PredicateMatrix: The PredicateMatrix[10] [29] is an automatic extension of SemLink[30]. It gathers all the resources listed previously plus some additional lexical knowledge coming from the **Multilingual Central Repository**[11] [31] such as **SUMO**[12] [27], **Top Ontology**[13] [32] or **WordNet domain**[14] [33]. These resources are connected automatically through a wide set of mappings. The current version of the PredicateMatrix contains 8495 predicates from PropBank and NomBank connected to 4704 synsets of WordNet, 554 frames of FrameNet and 55 different ESO classes. It also contains 23,386 roles of PropBank and NomBank mapped to 2343 frame-elements of FrameNet and 53 ESO roles. Thanks to the cross-lingual relations in wordnets, we have been able to map this PredicateMatrix data to Dutch and Spanish synsets and words as well.

These semantic resources are partially used to interpret tokens of text by the NLP modules. For example, the semantic role labeler uses PropBank as training data, whereas the Word-Sense-Disambiguation module assigns WordNet synsets to open class words in the text. Other semantic information is inserted into the representation of the tokens as a form of semantic typing through their association in the PredicateMatrix (see Section 4.2) which maps lemmas and WordNet synsets to FrameNet frames and classes in SUMO and ESO (see Section 4). The semantic typing of the interpreted textual elements is passed on to the instance representation in SEM providing the ESO and FrameNet interpretations of the *sale* event we saw in Fig. 2.

In addition to the above semantic knowledge, some NLP modules also use episodic knowledge:

DBpedia: DBpedia [34] is a database that provides structured information extracted from Wikipedia. The English variant currently provides information about more than 4 million *things*, including at least 1,445,000 persons, 735,000 places, 241,000 organizations classified in an ontology. According

[10] http://adimen.si.ehu.es/web/PredicateMatrix.
[11] http://adimen.si.ehu.es/web/MCR.
[12] http://www.ontologyportal.com/.
[13] http://adimen.si.ehu.es/web/WordNet2TO.
[14] http://wndomains.fbk.eu/.

to the DBpedia website, localized versions of DBpedia are available in 125 languages.[15] All these versions together describe 38.3 million *things*, out of which 23.8 million are localized descriptions of things that also exist in the English version of DBpedia. In addition, DBpedia is interlinked with many other data sources from various domains (life sciences, media, geographic government, publications, etc.), including Freebase[16] [35] and YAGO[17] [36], among many others.

Whereas we interpret an event as an instance of a *type* in the previous semantic ontologies (e.g. a FrameNet frame or ESO class), entities are interpreted directly as instances in DBpedia, which acts a register for instances. DBpedia then provides further background information about the entities it contains. However, not all identified entities mentioned in the news have a DBpedia entry. To be able to represent these instances, we need to create an artificial URI for each of them. Since there is no episodic knowledge about these entities and they make up a large proportion of all entities, we call them *dark entities* [37]. In Section 6, we explain how we can derive an extension to DBpedia from the news with these entities to improve further processing of the news. This is shows in Fig. 5 as an add-on DarkDBpedia. Fig. 5 also indicates that we apply a reasoner to the RDF output of our system to derive further episodic implications from the event-centric knowledge. Section 6 provides more details about this process as well.

The semantic knowledge and episodic knowledge[18] together provide the means to indicate what kind of events occur when and where in the world and who exactly is involved. As such, these knowledge resources are essential for semantic NLP and play a major role in our system. These knowledge resources are to some extent also available for other languages than English. Equivalence relations across different language wordnets allow us to share the knowledge framework across languages. This defines the conceptual interoperability of the interpretation of the text.

4. Event extraction

Event extraction demands both robust basic pre-processing – such as tokenization, Part of Speech tagging (POS), lemmatization – and advanced linguistic processing, such as – Named Entity Recognition and Classification (NERC), Syntactic Parsing, Coreference Resolution, Word Sense Disambiguation (WSD), Named Entity Disambiguation (NED), Semantic Role Labeling (SRL) and interpretation of temporal expressions. Although we already use NAF to harmonize the different results from the different linguistic modules, cross-lingual event detection additionally requires to perform all these tasks in a semantically compatible way. We have therefore developed cross-lingual pipelines for interpreting events and event components in text in a common language independent semantic representation. In order to achieve cross-lingual semantic interoperability, entities, event mentions and time expressions are projected to language independent knowledge representations. Thus, named entities are linked to English DBpedia entity identifiers thanks to DBpedia cross-lingual links while nominal and verbal event mentions are aligned to abstract event representations through the PredicateMatrix. Additionally, concepts for open class words are represented using the EuroWordNet Inter-lingual index [38]. Finally, time expressions are normalized following the ISO 24617-1 standard [39]. Several demonstrators exhibit the capability of the NewsReader cross-lingual event extraction pipeline,[19] which is, to the best of our knowledge, the first of its kind.

4.1. Newsreader NLP pipelines

The pre-processing required for event extraction in NewsReader consists of tokenization, lemmatization and POS tagging. For English and Spanish this is done by the IXA pipes [40][20] whereas for Dutch the morphological analysis is performed using Alpino [41].[21] For Italian the TextPro tool suite [42][22] is used. Other modules which perform additional processing (WSD, opinion mining, parsing, etc.) are described in Aggeri et al. [43]. Henceforth, we will focus on the modules used in the NewsReader pipeline related to entity and predicate processing for event extraction.

Entity processing for the four NewsReader languages is performed by a NED module that links the entities previously spotted by a NERC module to the corresponding DBpedia entries. Furthermore, general concepts that are considered to be relevant for NewsReader, although not strictly named entities, are detected by the Wikification module. For example, via Wikification the NewsReader pipeline would detect *dbp:Stock_market* as a relevant concept for the phrase *stock market*. The NewsReader NED and Wikification modules (for all four languages) are built on top of DBpedia Spotlight [44],[23] a Wikification tool for automatically annotating mentions of DBpedia resources in text, providing a solution for linking unstructured information sources to the Linked Open Data cloud through DBpedia. Spotlight also offers the option of performing only Named Entity Disambiguation given previously detected spots by another engine. In NewsReader, we use NERC modules for entity detection or *spotting* in each language. DBpedia Spotlight offers probabilistic models trained for the four languages covered by NewsReader.

In addition, the SRL module for English and Spanish detects PropBank predicates and roles of the sentences using the MATE tools [45]. It also provides the corresponding interpretations in FrameNet, VerbNet, WordNet and ESO using the PredicateMatrix. The Dutch SRL module is a Python reimplementation of SoNar SRL [46] for event predicates. As this SRL module does not handle nominalizations, a separate module detects the nominal predicates. Once the predicates and their roles have been detected a final process assigns FrameNet frames and elements to the predicates and roles. For Italian, due to the lack of annotated data, an SRL system has been developed based on dependency relations, events (output of the event recognition module) and PropBank-like frames (built automatically using the MultiSemCor English-Italian aligned corpus [47]). Once event extraction is performed by the SRL module, a separate module tries to establish which of the events corefer [48]. This module works the same for all the four languages.

Finally, some NLP modules perform time expression recognition and normalization, and temporal and causal relations extraction [49,50]. These tasks are based on the TimeML specification [39]. In addition to the extraction of temporal relation as defined in TimeML, the module identifies also temporal anchoring of events, i.e. the date (explicit in the text or not) when an event took place or will occur. The Spanish and Dutch module is based on HeidelTime [51], a multilingual temporal tagger. HeidelTime identifies temporal expressions based on language specific patterns. Identified temporal expressions are normalized and represented according to TIMEX3 annotations [52]. For English and Italian, a Time

[15] http://wiki.dbpedia.org/Datasets, accessed 30 November 2014.
[16] https://www.freebase.com/.
[17] http://bit.ly/1oEVkR7.
[18] There is also statistical knowledge used in the NLP process, e.g. for training language models. However, this is beyond the scope of this paper.
[19] http://www.newsreader-project.eu/results/demos/.
[20] http://ixa2.si.ehu.es/ixa-pipes/.
[21] http://kyoto.let.vu.nl/nwrdemo_nl/demo.
[22] http://hlt-services2.fbk.eu:8080/nwrDemo/nwr.
[23] https://github.com/dbpedia-spotlight.

Processing module has been developed in NewsReader (time expression extraction and normalization, event detection, temporal relation extraction and predicate time anchor) [49,53]. The time expression normalization is mainly carried out by the TimeNorm library implemented by [54] for English and adapted for Italian.

4.2. Cross-lingual interoperability

Cross-lingual semantic interoperability is achieved via the projection of entities, event mentions and time expressions to language independent knowledge representations.

4.2.0.1. Cross-lingual interoperability for entities. First, the NewsReader pipeline provides cross-lingual interoperability with respect to the named entities occurring in the text.

The NewsReader NED modules suggest a list of candidates for each entity. Based on the input language, the corresponding DBpedia is used to perform the semantic annotation. This means that the external references to DBpedia produced by each DBpedia Spotlight language module will be different. For instance, a mention to *New York* in an English document produces as external reference the identifier http://dbpedia.org/page/New_York. Similarly, a mention to *Nueva York* in a Spanish document produces as external reference the identifier http://es.dbpedia.org/page/Nueva_York. However, both identifiers are interoperable because there are cross-lingual links between both English and Spanish DBpedia entries. To make the interoperability explicit, we have modified our non English NED modules to also include the corresponding identifiers for English as external reference (if they exist). For example, for the mentions of *Nueva York* the English identifier http://dbpedia.org/page/New_York will also be added as external resource. This new feature allows us to work with cross-lingual links by linking the cross-lingual realizations of entities in different languages.

4.2.0.2. Cross-lingual Predicate Models. The event representation provided by a SRL system depends on the semantic resource used for training that system. Thus, each knowledge source of predicate information will contain different descriptions of the roles for each predicate. Our pipelines guarantee interoperability across language and predicate resources by integrating the PredicateMatrix within the SRL modules. The PredicateMatrix gathers multilingual knowledge bases that contain predicate and semantic role information (see Section 3.4).

Fig. 6 provides the output of our SRL module for the English sentence *Qatar Holding sells 10% stake in Porsche to founding families*. Our SRL module first processes the sentence providing predicates and role annotations from PropBank. Now, as PropBank is integrated into the PredicateMatrix, our SRL module can also obtain the corresponding predicate classes and roles for the rest of the predicate resources. Thus, *Qatar Holding* identified as *A0* role of the predicate *sell.01* by MATE tools corresponds to the *Buyer* role of a *Commerce_sell* frame according to FrameNet.

This also applies across languages. For example, the Spanish SRL module is trained using Ancora [26]. Thanks to the PredicateMatrix and the links included in Ancora to PropBank, we can also establish that the Spanish verb *vende* and its lemma *vender* is also aligned to the PropBank predicate *sell.01*, the *Commerce_sell* frame from FrameNet, and the rest of information included in the PredicateMatrix for this event description. Similarly, *El holding the Qatar* identified by the Spanish SRL module as *arg0* role of the Ancora dpredicate *vender.01* which also corresponds to the same *Buyer* role of the *Commerce_sell* frame. Thus, after our SRL module has processed the Spanish sentence, we obtain the same language-independent semantic representation as the one obtained from the English sentence. This same process was implemented for Dutch using a Dutch version of the PredicateMatrix and a SRL module for PropBank roles trained on SoNaR [46].

4.2.0.3. Normalization of time expressions. We normalize time expressions following the ISO 24617-1 standard [39]. For example, if temporal expressions such as *next Monday, tomorrow*, and *yesterday* in English or *ayer* and *el próximo lunes* in Spanish are referring to the same exact date (let's say *November 16th, 2015*), all these temporal expressions are normalized to the same TIMEX3 value corresponding to *2015-11-16*.

5. Event modeling

The pipelines described in Section 4 create rich NAF-XML interpretations for tokens in the text, which we call mentions. The interpretations are scattered over different layers in NAF, as the output of different modules, which partially express the same information and partially complement each other. To come to a formal representation of the reference of these interpretations, we derive a SEM representation at the instance level and relations between these instances. We define instances for the following data types by creating unique URIs: entities, non-entities, events and time expressions.

To create these SEM representations, we defined the IDAP procedure, which stands for Identification, Deduplication, Aggregation and Perspectivization:

1. **I**dentification: identity across mentions of entities and events is established on the basis of overlapping information;
2. **D**eduplication: overlapping information is only represented once in SEM RDF;
3. **A**ggregation: complementary information is combined in a single event-centric representation;
4. **P**erspectivation: differences and different viewpoints are modeled in GRaSP and are linked to the sources and mentions in the text;

This IDAP procedure is applied in two steps: 1) across different mentions in a single NAF file to create a SEM-RDF representation of instances for a source text, and, 2) across the SEM representations of different source texts to create a cross-document representation. Semantic identity of instances across mentions and across layers in NAF is crucial in this process. Semantic identity, among others, depends on the semantic knowledge resources that have been used in combination with the NLP processing. We explain this in more detail in the subsections below for the different types of instances.

5.1. Entities, dark entities and non-entities in SEM

Genuine entities are represented in the entity layer in NAF and often have a DBpedia URI that identifies them. Since this layer is mention-based, we can find the same URI across different entity phrases detected in the text. In SEM, we will represent this entity once through that URI and only extend the *gaf:denotedBy* links to these mentions. Consider the examples of the phrases *Didier Drogba* and *Didier Yves Drogba Tébily* in Fig. 7. They were detected as entities, but only the former is mapped to DBpedia. By taking the URI *dbp:Didier_Drogba* as an identifier for the instance, we thus automatically link the mentions of the tokens *t68, t69* and *t807* to the same instance but not the tokens *t2, t3, t4, t5*. However, NAF also provides a coreference layer established by a separate NLP module, which groups anaphora, noun phrases and entity phrases that are coreferential, as is shown in Fig. 8. We can see that both the token sets *t68, t69* and *t2. t3. t4. t5* are included in the same coreference set.

Likewise, the module can unify all the mentions from the entities with the same URI and the phrases from the coreference set

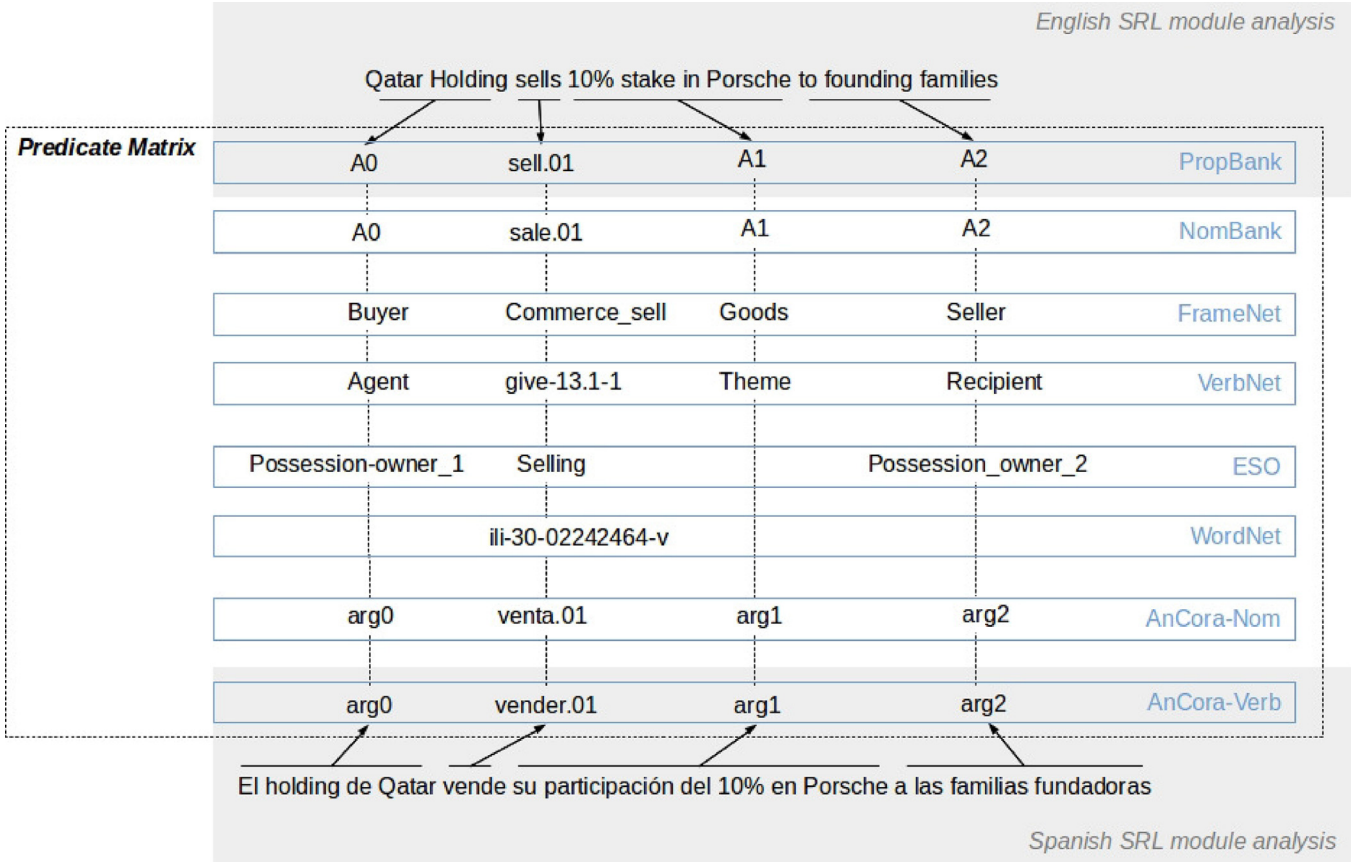

Fig. 6. Cross-lingual and semantic interoperability provided by the PredicateMatrix.

```
<entity id="e5" type="PERSON">
    <!--Didier Drogba-->
    <span><target id="t68"/><target id="t69"/></span>
  <externalReferences>
    <externalRef resource="spotlight_v1" reference="dbp:Didier_Drogba"
    confidence="1.0" reftype="en" source="en"/>
  </externalReferences>
</entity>

</entities>
  <entity id="e2" type="PERSON">
    <!--Didier Yves Drogba Tebily-->
    <span><target id="t2"/><target id="t3"/><target id="t4"/><target id="t5"/></span>
</entity>

<entity id="e64" type="PERSON">
    <!--Drogba-->
    <span><target id="t807"/></span>
  <externalReferences>
    <externalRef resource="spotlight_v1" reference="dbp:Didier_Drogba"
    confidence="1.0" reftype="en" source="en"/>
  </externalReferences>
</entity>
```

Fig. 7. Entities in NAF with and without a DBpedia URI.

into a single instance representation with *gaf:denotedBy* links to all these mentions. The result of this merge in RDF is shown in Fig. 9, where we show a subset of the *gaf:denotedBy* relations from the instance to all the mentions.

Not all entities without a URI can be resolved through coreference sets to other entities with an URI. For those cases, we create an artificial URI based on the phrase and represent it as an instance of the entity type that is assigned by the entity recognizer. The football players shown in Fig. 10, for example, were not resolved and are represented as instances through a URI based on their phrase and linked to the type *PERSON*. We discuss these so-called *dark entities* in more detail in Section 6.

```
<coref id="co2">
  <!--His-->
  <span> <target id="t129"/> </span>
  <!--Didier Drogba-->
  <span> <target id="t531"/> <target id="t532"/> </span>
  <!--Drogba-->
  <span> <target id="t429"/> </span>
  <!--he-->
  <span> <target id="t74"/> </span>
  <!--his-->
  <span> <target id="t9"/> </span>
  <!--Drogba-->
  <span> <target id="t333"/> </span>
  <!--Drogba-->
  <span> <target id="t353"/> </span>
  <!--Drogba-->
  <span> <target id="t202"/> </span>
  <!--Didier Drogba-->
  <span> <target id="t626"/> <target id="t627"/> </span>
  <!--Drogba-->
  <span> <target id="t766"/> </span>
  <!--Didier Drogba-->
  <span> <target id="t68"/> <target id="t69"/></span>
  <!--His-->
  <span><target id="t108"/></span>
  <!--his-->
  <span><target id="t84"/></span>
  <!--Didier Yves Drogba Tebily-->
  <span><target id="t2"/><target id="t3"/><target id="t4"/><target id="t5"/></span>
</coref>
```

Fig. 8. Nominal coreference set in NAF.

```
dbp:Didier_Drogba
        rdfs:label "Drogba" , "Didier Yves Drogba Tebily" , "his" ,
                "Didier Drogba" , "Didier Drogba 's" ;
        gaf:denotedBy
        <nwr:57R0-J5K1-JC86-C1N7.xml#char=11,36> ,
        <nwr:57R0-J5K1-JC86-C1N7.xml#char=694,697> ,
        <nwr:57R0-J5K1-JC86-C1N7.xml#char=2795,2808> ,
        <nwr:57R0-J5K1-JC86-C1N7.xml#char=2274,2280> , ...etc...
```

Fig. 9. Entity in SEM.

```
nwr:wikinews/entities/daHelton
        a nwrontology:PERSON ;
        rdfs:label "da Helton" , "de Helton" ;
        gaf:denotedBy :#char=1254,1263 , :#char=2323,2332 .
nwr:wikinews/entities/RaulMoreiles
        a nwrontology:PERSON ;
        rdfs:label "Raul moreiles" ;
        gaf:denotedBy :#char=1374,1387 .
nwr:data/cars/entities/Federal_Chamber_of_Automotive_Industry
        a nwrontology:ORGANISATION ;
        rdfs:label "Federal_Chamber_of_Automotive_Industries" ,
                "Federal_Chamber_of_Automotive_Industry" ;
        gaf:denotedBy :#char=3,43 , :#char=20,59 .
```

Fig. 10. Dark entities in SEM.

```
<nwr:non−entities/10+\%25+stake+in+porsche>
rdfs:label "10 \% stake in Porsche" ;
gaf:denotedBy <http://english.alarabiya.net#char=20,40> .

<nwr:non−entities/to+founding+families>
rdfs:label "to founding family" ;
af:denotedBy <http://english.alarabiya.net#char=41,61> .
```

Fig. 11. Non-entities in SEM.

In addition to these dark entities there are also other phrases that play an important role in the event but are not (and often should not be) detected as entities. For example, the sale between Qatar and the Porsche family involves *10% stake* which is detected by the semantic role labeler but it is not represented as an entity. Whenever important roles[24] cannot be assigned to any mention of an entity, we represent the concept as a so-called *non-entity*. The type NONENTITY is assigned to these instances. Fig. 11 shows a representation of two non-entities in SEM-RDF, which are derived from the Semantic Role Layer (SRL).

5.2. Temporal objects in SEM

Time objects and temporal relations play an important role in NewsReader. Without proper time anchoring, we cannot compare one event to another (see below). Time objects for delineating events are derived from the TIMEX3 layer in NAF, which is based on the TimeML formalism [39]. From these TIMEX3 elements, we derive two types of SEM instances: *time:Instant* and *time:Interval*. Instances of the type *time:Instant* have a *time:inDateTime* relation to a date object, whereas instances of *time:Interval* have a *time:hasBeginning* and/or *time:hasEnd* relation to a date. Dates are represented as separate instances of the type *time:DateTimeDescription* with values for the year, month and/or day according to owl-time.[25] In Fig. 12, we see two time expressions: *tmx0* and *tmx2* represented through URIs based on the documents in which they occur. The first is an *time:Instant* and has no mentions in the text because it represents the document creation time that is found in the meta data and not in the text. The second example *tmx2* is a *time:Interval* with mentions and also a label *week* that was normalized to a specific period of 7 days in July 1989.[26] We see that both time expressions have properties (*inDateTime, hasBeginning* and *hasEnd*) with dates as values. These dates are represented separately according to owl-time, allowing for temporal reasoning.

5.3. Events in SEM

Events usually do not end up as entries in DBpedia. Identity of events is also more complex than identity of entities. First of all,

```
# document creation time, which has no mentions
  <nwr:wsj_1013.xml#tmx0>
        a  time:Instant ;
        rdfs:label  "nwr:time/19890701" ;
        time:inDateTime  nwr:time/19890701 .

  # DURATION with begin and end point
  <nwr:4M1J−3MC0−TWKJ−V1W8.xml#tmx2>
        a  time:Interval ;
        rdfs:label  "week" ;
        gaf:denotedBy  nwr:4M1J−3MC0−TWKJ−V1W8.xml#char=822,825 ,
                nwr:4M1J−3MC0−TWKJ−V1W8.xml#char=834,839 ;
        time:hasBeginning nwr:time/19890701 ;
        time:hasEnd  nwr:19890707 .

  <nwr:time/19890701>
        a  time:DateTimeDescription ;
        time:day  "−−−01"^^<http://www.w3.org/2001/XMLSchema#gDay> ;
        time:month  "−−07"^^<http://www.w3.org/2001/XMLSchema#gMonth> ;
        time:unitType  time:unitDay ;
        time:year  "1989"^^<http://www.w3.org/2001/XMLSchema#gYear> .

  <nwr:time/19890707>
        a  time:DateTimeDescription ;
        time:day  "−−−07"^^<http://www.w3.org/2001/XMLSchema#gDay> ;
        time:month  "−−07"^^<http://www.w3.org/2001/XMLSchema#gMonth> ;
        time:unitType  time:unitDay ;
        time:year  "1989"^^<http://www.w3.org/2001/XMLSchema#gYear> .
```

Fig. 12. Temporal objects in SEM.

the identity of an event is the product of the identity of all its components [48]: the action, the participants, the place and the time. If *Qatar sells 10% of stake to Porsche* on another day, it is not the same event and probably not the same shares. All repetitive events on different dates are not identical and all events on the same date involving different participants are not identical. We can therefore not just use the words in the text that mention the event – as we did for dark entities and non-entities – to establish identity: it is too unlikely that one *sales* will be the same as another *sales* across different documents.

Another complicating factor is that all the information that uniquely defines an event in terms of its components is hardly ever mentioned in a single sentence [55]. News events are usually described using some narrative structure in which participants, time and place are given throughout the text. For example, the following article from Al Arabiya provides the following statements with respect to the *deal* in other sentences:

> "This transaction results as a logical step after the creation of the Integrated Automotive Group between Volkswagen and Porsche AG as finalized in 2012," Qatar Holding said in the statement. Neither party gave any details of the price paid for the stake. Porsche SE shares were trading at 60.76 euros per share on Monday, up 0.36 percent on the day. "This (sale by Qatar) is positive because the stake is returning to the hands of the Porsche/Piech families," Bamler added.

We therefore follow a two-step IDAP approach in which we first establish identity across mentions in a single document. Next, we aggregate the event properties across these mentions before we compare the events across different documents. Identity within a document is established through the mentions of the actions, whereas identity across documents is established by comparing all the event components. In both cases, identity results in deduplication and aggregation.

5.3.1. Event identity within a document

Identity within a single document is done by the NLP module for event-coreference. This module uses the predicates in the SRL layer as a starting point and applies the following heuristics:

1. It groups all predicates with the same lemma in an initial coreference set.

[24] To limit the amount of instances and triples, we only consider roles that have a FrameNet Element assigned. We consider these roles essential for understanding what the event is about.

[25] http://www.w3.org/TR/owl-time.

[26] In case of underspecification, time expressions can also point to months, quarters or years.

```
<coref id="coevent72" type="event">
  <span><target id="t371"/></span> <!--chased-->
  <externalReferences>
    <externalRef resource="WordNet-3.0" reference="eng-30-02001858-v"
    confidence="1.0" source="dominant_sense"/>
  </externalReferences>
</coref>

<coref id="coevent61" type="event">
  <span><target id="t345"/></span><!--shot-->
  <span><target id="t414"/></span> <!--shot-->
  <span> <target id="t684"/></span> <!--injured-->
  <externalReferences>
    <externalRef resource="Princeton WordNet 3.0" reference="eng-30-00069879-v"
    confidence="2.64" source="lowest_common_subsumer"/>
    <externalRef resource="WordNet-3.0" reference="eng-30-02055267-v"
    confidence="0.81" source="dominant_sense"/>
    <externalRef resource="WordNet-3.0" reference="eng-30-01134781-v"
    confidence="0.95" source="dominant_sense"/>
    <externalRef resource="WordNet-3.0" reference="eng-30-01137138-v"
    confidence="0.92" source="dominant_sense"/>
    <externalRef resource="WordNet-3.0" reference="eng-30-02484570-v"
    confidence="0.93" source="dominant_sense"/>
  </externalReferences>
</coref>
```

Fig. 13. Event coreference in NAF.

```
<nwr:55XK-XGX1-JBKJ-C3CF.xml#ev72>
        a sem:Event , nwrontology:contextualEvent , eso:Translocation ,
            fn:Cotheme , ili:i31747 ;
        rdfs:label "chase" ;
        gaf:denotedBy <nwr:55XK-XGX1-JBKJ-C3CF.xml#char=2074,2080> .

<nwr:59JB-GV01-JBSN-30SP.xml#ev84>
        a sem:Event , nwrontology:contextualEvent , fn:Hit_target ,
            ili:i27293 , ili:i27278 , ili:i34141 , fn:Shoot_projectiles ,
        ili:i22125 , fn:Use_firearm ;
        rdfs:label "shoot" , "injure" ;
        gaf:denotedBy <nwr:59JB-GV01-JBSN-30SP.xml#char=2215,2219> ,
        <nwr:59JB-GV01-JBSN-30SP.xml#char=2588,2595> .
```

Fig. 14. Event instances in SEM.

2. It collects the highest-scoring wordnet synsets of the mentions of these lemmas from the whole document; these form the dominant senses of the coreference set.
3. It merges all lemma-based coreference sets for which the WordNet Similarity [56] of their dominant senses scores above a threshold and it stores the lowest-common-subsumer that established the similarity.

In Fig. 13, we show a coreference set with a single predicate *chased* that was not merged with any other set and another set in which *shot* and *injured* were merged into a single set with the lowest common subsumer synset *eng-30-00069879-v, injure:1, wound:1* and a similarity score of 2.64.

Each coreference set becomes a potential event instance, where we use the mentions for the labels and the references to the WordNet synset as a subclass relation.[27] Furthermore, we collect a subset of the ontological labels assigned to each predicate in the SRL. The RDF result for the coreference sets in Fig. 13 then looks as shown in Fig. 14, where we created an artificial URI for the event instances: *ev72* and *ev84*, enriched with the semantic types obtained from the SRL, the labels from the mentions and the pointers to the mentions.

Once the instances for entities, non-entities and events have been established, we determine the relations between them. Since the events are derived from the SRL in NAF, we can go back to that layer to find the roles. Again, we match the span of any entity mention with the span of the roles to determine the entity for each role. If an event has more than one mention, as is the case for *shoot* and *injure* above we aggregate the role information across all

[27] We convert all reference to WordNet synset to InterLingualIndex concepts to allow for cross-lingual matching [57].

```
<http://english.alarabiya.net#ev1>
        a sem:Event , nwrontology:contextualEvent ,
            fn:Commerce_sell , eso:Selling , ili:i32963 , ili:i32953 ;
        rdfs:label "sell" ;
        gaf:denotedBy <http://english.alarabiya.net#char=14,19> ;
        sem:hasTime <http://english.alarabiya.net#tmx0> ;

        sem:hasActor dbp:Qatar_Investment_Authority ;
        eso:possession-owner_1 dbp:Qatar_Investment_Authority ;
        fn:Commerce_sell@Seller dbp:Qatar_Investment_Authority ;
        pb:A0 dbp:Qatar_Investment_Authority ;

        sem:hasActor <nwr:non-entities/10+\%25+stake+in+porsche> ;
        fn:Commerce_sell@Goods <nwr:non-entities/10+\%25+stake+in+porsche> ;
        pb:A1 <nwr:non-entities/10+\%25+stake+in+porsche> .

        sem:hasActor <nwr:non-entities/to+founding+families> ;
        eso:possession-owner_2 <nwr:non-entities/to+founding+families> ;
        fn:Commerce_sell@Buyer <nwr:non-entities/to+founding+families> ;
        pb:A2 <nwr:non-entities/to+founding+families> .

<http://www.telegraph.co.uk#ev2>
        a sem:Event , nwrontology:contextualEvent ,
            fn:Commerce_buy , eso:Buying , ili:i32788 , ili:i34901 ;
        rdfs:label "buy" ;
        gaf:denotedBy <http://www.telegraph.co.uk#char=15,19> ;
        sem:hasTime <http://www.telegraph.co.uk#tmx0> ;

        sem:hasActor dbp:Porsche ;
        eso:possession-owner_2 dbp:Porsche ;
        fn:Commerce_buy@Buyer dbp:Porsche ;
        pb:A0 dbp:Porsche ;

        sem:hasActor <nwr:non-entities/10pc+stake> ;
        fn:Commerce_buy@Goods <nwr:non-entities/10pc+stake> ;
        pb:A1 <nwr:non-entities/10pc+stake> ;

        sem:hasActor dbp:Qatar ;
        fn:Commerce_buy@Means dbp:Qatar ;
        pb:A2 dbp:Qatar .

<http://english.alarabiya.net#tmx0>
        a time:Instant ;
        time:inDateTime <http://www.newsreader-project.eu/time/20130617> .

<http://www.telegraph.co.uk#tmx0>
        a time:Instant ;
        time:inDateTime <http://www.newsreader-project.eu/time/20130617> .
```

Fig. 15. Sell and buy events in SEM without the entity representations.

these mentions. We do the same for anchoring the event to the timex expressions, which results in a *sem:hasTime* relation.

Fig. 15 then shows the complete representations of the event instances generated by our system from the two Qatar-Porsche titles, considering each document separately. We show the properties of each event and the SEM relations to entities, non-entities and time expressions. Since there are no time expressions in the titles, both events are anchored to the document creation times. Both events received different identifiers based on their source. This would be the result of Step 2 in the overall process before we match event descriptions across different sources. We see that we obtained triples by aggregating information, e.g. the SRL information is combined with the WSD output and the entity layer and the temporal information is obtained from the metadata. Processing more sentences and resolving event coreference across these sentences would result in further aggregation across mentions and richer descriptions. So far, we described the application of IDAP to single NAF files. In the next subsection, we explain how these aggregated event descriptions are used to establish event identity across NAF representations (Step 3 in the overall process).

5.3.2. Event identity across documents

For entities, non-entities and time objects, cross-document identity is established through the use of (partially) standardized URIs. As we explained above, event identity needs to be defined as a function of the identity of the components:

1. Identity of the action or process.
2. Identity of the temporal relation.
3. Identity of the participants and their roles.

We first check the action identity constraint. Action identity across documents is defined by the overlap of WordNet synsets or lemmas across the events, where WordNet overlap takes precedence over lemma overlap.[28] Secondly, we match the time relations of two event descriptions. In the same way as physical boundaries define shape and are most critical to keep entities apart, temporal boundaries delineate events. These boundaries can be defined at the granularity of a date, a month or a year. Finally, the matching of the event's participants is less strict and their rigidity varies per event type. For example, physical events such as motions are necessarily bound by location, while others, such as financial transactions, are not. Yet for speech-act events, such as *say* or *annouce*, it is crucial that the source of the event is identical. We therefore parameterized the matching of the participants so that we can apply different additional constraints for different types of events.

The Porsche-Qatar example is a real challenge in terms of cross-document event coreference. The results shown in Fig. 2 in Section 3 are therefore difficult to obtain. First of all, the predicates *buy* and *sell* do not share any WordNet synset nor any ontological class. Secondly, we can see that not all participants (*to+founding+families*) are resolved to entities or to the same entities (*dbp:Qatar* versus *dbp:Qatar_Investment_Authority*). To obtain a merge across these representations, we have to apply very loose constraints. The matching of the RDF event descriptions can therefore be tuned by setting constraints for each of the component matches or the combination of the component matches. If the parameters are very strict, hardly any cross-document event coreference is detected, if they are very loose a lot of event data is lumped together resulting in fewer event instances with strongly aggregated data.

6. Domain adaptation

We so far described the generic NewsReader system that can be applied to any text out-of-the-box. This system uses knowledge resources such as DBpedia for entities and lexical resources for events. In practice, any text data set contains specific data that is either not present in these resources or not properly modeled. In this section, we explain how we can overcome gaps in the coverage of DBpedia to link to unknown entities and how specific events can be modeled for a domain.

6.1. Dark entities

Dark entities are those entities for which no relevant information was identified in a given knowledge base / entity repository. First of all, there are cases where a resource is present, but it contains very little or no relevant information to further reason about the entity. For example at the time of writing this article, the DBpedia resource http://dbpedia.org/resource/Heinz_Branitzki detected in one of our datasets contains no information other than that it is an Entity of the type "Thing". In fact he has been an interim CEO of Porsche but this is not recorded in DBpedia. Querying the news for CEOs will thus never yield this person among the result. In addition there are many cases in which there is no resource at all present in the knowledge base. Finally, there may be entities linked to the wrong resource. All these cases, searching for types of entities will give partial and wrong results.

To get to grips with this problem, we have carried out an in depth analysis of a specific dataset in the technology domain. For this, we collected 43,000 articles from TechCrunch.com[29] and a dump of its accompanying structured CrunchBase database.[30] We analyzed the top 200 entity instances with DBpedia links and the top 149 instances without DBpedia links. We divided our analysis of the mentions with links into those that are out of domain, are the results of errors in the named entity recognition module, and others. For those cases in which the modules did not find a link, we classified the errors in the following categories: "entity not present in DBpedia", "spelling variation but present in DBpedia", "not an entity or an event", "conjunctions", and "others". In the remainder of this subsection, we describe our analyses as well as recommendations on how to overcome the domain specific errors.

Table 1
Types of mentions recognized in entity candidates occurring 50 or more times in the TechCrunch corpus.

Type	# Mentions	# Unique
Person	2156 (21.59%)	25
Organization	4222 (42.28%)	44
Location	84 (0.84%)	1
Product	2025 (20.28%)	21
Event	775 (7.76%)	8
Not an entity or event	724 (7.2%)	10

In the 43,000 TechCrunch articles, the NewsReader pipeline detected 807,088 entity mentions, which were aggregated into 212,611 unique entities. For 102,141 entities, links to DBpedia resources were identified, covering 608,801 (75.43%) of the entity mentions.

We inspected the links for the 200 most commonly occurring entity instances (representing 222,467 mentions) and found that for 185 of the entities (212,133 mentions) the correct link had been identified. We categorized the 15 cases (10,334 mentions) in which an incorrect link was provided as follows:

- **Out of domain link (10 instances)** If an entity outside the domain is prevalent in DBpedia, it may erroneously get chosen over the domain entity that is meant in the text but which is either not prevalent or not present in DBpedia. E.g. 'Box' is linked to http://dbpedia.org/resource/Box instead of http://dbpedia.org/page/Box_(company).
- **Error in Named Entity Recognition (5 instances)** Mostly incorrect entity boundaries, such as 'no.', which is linked to http://dbpedia.org/resource/Norwegian_language).

Many of the "out of domain" errors can be avoided by adapting the linking module to give preference over in-domain entities.

There were 198,287 mentions (110,470 unique) for which the named entity linking module could not find a link to DBpedia. We inspected the mentions that occurred more than 50 times in our dataset (covering 149 unique entities or 9986 mentions). The breakdown of these entities is presented in Table 1 and the statistics on the error analysis are presented in Table 2.

We included events labeled as entities in our analysis as they occurred in almost 8% of the cases. These are often nominal events such as "Startup Alley" and "Demo Day". They are very specific to the domain and behave very much like named entities. It is thus not surprising that they are labeled as such. Another important category in TechCrunch are products (20.28%), as the news articles in the technology domain is concerned with new product releases. Because the technology domain is made up of small startups that either make it or break, the companies and their products are less established and thus less likely to occur in DBpedia, which is the

[28] If the WSD makes a difference between firing guns and firing people, the different WordNet synsets will keep these two expressions separate despite these being the same words.

[29] http://techcrunch.com/.

[30] https://www.crunchbase.com/.

Table 2
Types of errors in entity mentions occurring 50 or more times in TechCrunch corpus.

Type	# Mentions	# Unique
Entity not present in DBpedia	7776 (77.86%)	85
Spelling variation/nickname	576 (5.76%)	4
Recall error	135 (1.35%)	2
Event	775 (7.76%)	8
Not an entity or event	724 (7.25%)	10

case in 77.86% or 85 unique entities of those analyzed. Alternatively, we found that a significant part of the entities could instead be linked to the CrunchBase resource (54 unique entities, 62.56% of the mentions that were not found in DBpedia). This resource also provides us with biographical information about persons and company histories, therefore linking entities to CrunchBase seems a viable option for this domain. The Spotlight tool allows for the use of such a customized database in addition to the generic resource DBpedia to improve recall and precision of linking for a domain.

Naming variations and nicknames are also found among the named entities, which often results in linking errors. We find for example “Mike Arrington” instead of “Michael Arrington” and “Zuck” as shorthand for “Mark Zuckerberg” as well as mentions such as “Ferriss” instead of “Tim Ferriss”. To investigate the in-document variation of the entity mentions, a rule-based consolidation script was devised that uses the string overlap between entity mentions to provide links to previously unlinked entity mentions. This results in 24,947 entity mentions that previously had no link assigned to become linked.

Overall, domain adaptation for entities is feasibly and can benefit greatly from the availability of additional domain resources. We summarize the main recommendations below:

Give preference to in-domain entities to overcome entities being linked to the most popular entities in the general domain;
Link to additional resources to overcome gaps in coverage of DBpedia;
In-domain coreference resolution to be able to link different variants of the same name.

Besides resolving acute issues within a domain, these steps can also result in the creation of a new resource with entities that were not registered before. For these entities surface forms (i.e. spelling variations), typing information and relationships to other entities can be modeled, effectively creating a Dark DBpedia.

6.2. Event implications

We explained how we derive RDF representations for the events from text in the previous Sections. However, events do not explain the implications of the changes they refer to. We can imagine that a report on a journey tells you about a person’s travel but that we need to understand the precise meaning of the event to infer where that person was at what point in time. We assume that for each domain and each application, there will be a specific set of implications that are important, e.g. the source, target and path of the journey, while other semantic information, the speed or manner of traveling, may be less relevant. Rather than trying to provide a full ontological definition of all the events in general, we tried to model the implications of events that matter for a specific domain or application. We developed the Event and Situation Ontology [58], that abstracts over the implications drawn by various fine-grained event expressions and the associated participants (through mapping to the PredicateMatrix). As such, it provides a typing system of events and a formal model to define the implications of these events and the entities affected by the event [58]. The model captures the implications that matter for the domain and data set and can be seen as a way of domain modeling of the events. Fig. 16 provides an example of the kind of knowledge that ESO captures. It shows that working relations can be derived from the implications of events, regardless of the precise way in which these relations started or ended. Rather than defining the full meaning of events such as *hire* or *fire*, ESO defines precisely the implication for the working relation.

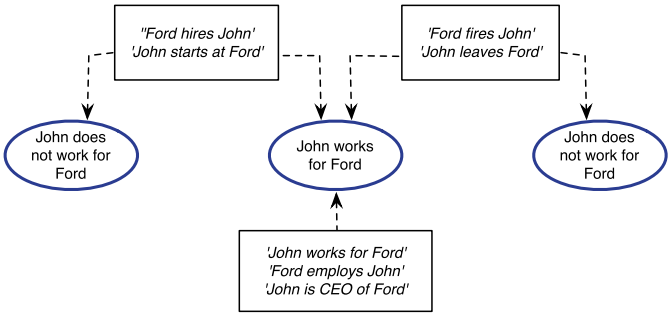

Fig. 16. Illustration of events and situations in ESO (events are in boxes, situations in circles).

Existing resources such as PropBank, VerbNet, FrameNet and NomBank provide the means to represent the role of individual participants of events. These resources mainly provide the information about the events expressed by lexical items and the participants they entail. FrameNet defines a limited set of sub-events and causal relations and VerbNet provides some fine-grained information about implications of certain events. Such resources define constraints at the lexical conceptual level, but this is not sufficient to reason about the implication that situations have on instances involved. NewsReader extends this conceptual-relational approach by capturing what specific events entail for situations that the text refers to. This implies that events and all the required entities need to be present in the representation of the text with their pertinent roles and that the temporal conditions are met before conclusions can be drawn on the implications.

Previous work has addressed applying deductive reasoning over frames [59] and the inference that can be deducted from events by defining pre- and post-situations [60]. However, to the best of our knowledge, no resource exists that provides the full picture of events, roles and implications for individual participants in such a way that it can be identified in text with semantic parsing technologies. Resources such as SUMO [61] and DOLCE [62] come close providing rich comprehensive specifications of the meaning of concepts. However, SUMO has a more generic focus and includes many classes and axioms that are not needed for our domain and it can not be coupled with a semantic parsing system. DOLCE, on the other hand, is too high level for our purposes. Because of these differences in focus, not all information needed to model changes in a specific domain, such as finance and commerce, is present. Moreoever, these resources do not consider the NLP platform and the lexical resources needed to derive the implications from textual expressions.

ESO is a hand-built OWL ontology[31] that represents events, their participants and relations between them on the instance level. It consists of 63 event classes, 65 ESO roles and 123 situation rule assertions.[32] ESO uses five basic components to capture this information:

[31] http://www.w3.org/2001/sw/wiki/OWL.
[32] The ESO ontology and documentation can be found here: https://github.com/newsreader/eso.

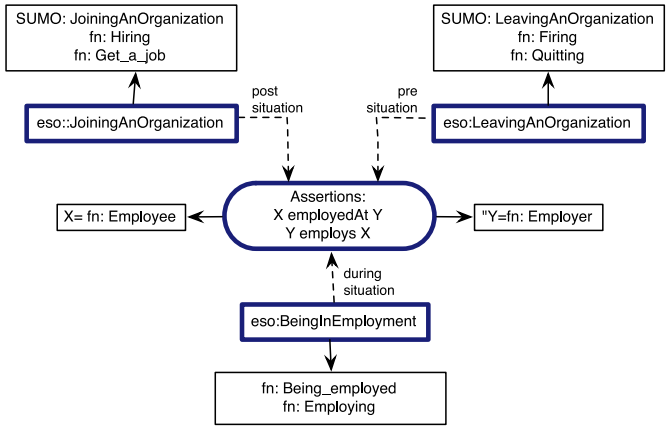

Fig. 17. Shared properties of related dynamic and static events [58].

1. **Event**: this class is the root of the taxonomy of event types. Any event detected in a text is an instance of some class of this taxonomy;
2. **DynamicEvent**: this is a subclass of Event for which dynamic changes are defined;
3. **StaticEvent**: this is another subclass of Event for "static" event types which capture more stable circumstances;
4. **Situation**: the individuals of this class are actual pre, post and during situations that are instantiated starting from the event instances detected in the text;
5. **SituationRule**: the individuals of this class encode the rules for instantiating pre/post/during situations when a certain type of event is detected.

Events are built upon those FrameNet frames that occur most frequently in a corpus consisting of 1.2 million news articles about the automotive industry and the financial domain processed by the NewsReader pipeline. Each ESO class is mapped to one or more FrameNet frames based on whether the pre and post situations defined in ESO hold for the frame. This means that more fine-grained distinctions or differences in perspectives that can be found in FrameNet are not maintained in ESO. For instance, frames representing *stealing, giving, supplying* are all mapped to *eso:Giving* regardless of how the change of ownership occurs. Likewise, ESO roles are mapped to (sets of) Frame Entities. In addition to FrameNet frames, ESO classes are linked to classes in SUMO whenever possible.

Fig. 17 illustrates the relations between dynamic events and static events. ESO captures the story of being employed somewhere using three classes: the dynamic events *eso:JoiningAnOrganization* and *eso:LeavingAnOrganization* and the static class *eso:BeingInEmployment*. ESO explicitly models that being in employment is a post situation of joining an organization. Furthermore, each pre-, post- and during-situation is defined by assertions that define the situation (here: *employed-At* and *employs*).

These explicit relations allow us to infer the reality of new events based on the events we identify in text. For instance, if we find that someone is employed somewhere, we know that this person joined the organization at some point in the past. If someone leaves an organization we know that this person was employed there, which in turn entails someone joined this organization earlier on. The next Section will explain how we draw such inferences.

6.2.1. Reasoning and inferencing

For all ESO classes, *eso:SituationRule* individuals are defined. These individuals trigger the pre-, during- or post-situation related to the class or set of classes it belongs to. Fig. 18 provides a (non-formal) overview of classes, mappings, assertions of the class *eso:JoiningAnOrganization*. In addition to the mapping to FrameNet and SUMO, assertions that distinguish the situation before the event and after the event are given. These assertions are linked to an event through the aforementioned *eso:SituationRules*.

JoiningAnOrganization subclassOf: IntentionalEvent
"The subclass of IntentionalEvent where someone starts working as an employee for some organization."

Class mappings:
closeMatch: fn:Hiring
closeMatch: fn:Get_a_job
broadMatch: sumo:JoiningAnOrganization

Role mappings:
employment-employee: fn:Employee
employment-employer: fn:Employer
employment-function: fn:Position
employment-value: fn:Compensation
employment-task: fn:Task
employment-attribute: -

Assertions:

pre situation	employment-employee	notEmployedAt	employment-employer
post situation	employment-employee	employedAt	employment-employer
	employment-employee	isEmployed	true
	employment-employee	hasFunction	employment-function
	employment-employee	hasTask	employment-task
	employment-employee	hasAttribute	employment-attribute
	employment-attribute	hasValue	employment-value

Example:

"Ford hired John as their new CEO for 100.000 euro a year."

pre situation	John	notEmployedAt	Ford
post situation	John	isEmployed	true
	John	employedAt	Ford
	John	hasFunction	new CEO
	John	hasAttribute	:abc124
	:abc124	hasValue	100.000 euro

Fig. 18. Non-formal transcription of the mappings, assertions and instantiation for the ESO class JoiningAnOrganization.

In our employment example, *eso:BeingInEmployment* has the rule *eso:during_BeingInEmployment*, and *eso:JoiningAn-Organization* has two specific individuals: *eso:pre_JoiningAn-Organization* and *eso:post_JoiningAnOrganization*. The class *eso:Situations* specifies how triples describing the situation must be defined. For each assertion, three annotation properties are provided, defining exactly what the role of the triple's subject, predicate and object are in the situation. For instance, the first two assertions of *eso:post_JoiningAnOrganization* are:

```
eso:post_JoiningAnOrganization_assertion1
    eso:hasSituationAssertionSubject    eso:employment
                                            -employee;
    eso:hasSituationAssertionProperty    eso:employedAt;
    eso:hasSituationAssertionObject    eso:employment
                                            -employer.

eso:post_JoiningAnOrganization_assertion2
    eso:hasSituationAssertionSubject  eso:employment
                                          -employee;
    eso:hasSituationAssertionProperty  eso:isEmployed
    eso:hasSituationAssertionObjectValue  true.
```

Using these assertions, it is possible to automatically infer that an event that belongs to the *eso:JoiningAnOrganization* involves an entity for which *eso:notEmployedAt* that organization holds true before the event occurred and *eso:EmployedAt* the same organization applies to the entity afterwards. Hence, we infer a post-

event situation that corresponds to the situation modeled for *eso:during_BeingInEmployment*.

In principle, only assertions involving participants that are identified by the semantic role labeling and PredicateMatrix interpretation are fired. We make an exception for ESO classes that express relative changes for one of the participants (e.g. *eso:Damaging, eso:Increasing*) where the changing attribute often remains implicit. For such values, an OWL existential restriction is placed on the roles in the assertion. This restriction will lead to the creation of a blank node when no explicit role for the participant is found. For a more elaborate explanation and extensive example, see Segers et al. (2016) [58].

We developed a reasoner tool (called *ESO reasoner*) for inferring situations from the detected event data.[33] The tool is structured as a dedicated processor of RDFpro[34] [63]. It combines OWL DL reasoning and a simple rule engine. For any event identified in the text, the module provides OWL reasoning to identify the ESO trigger rules (if any) to be applied on that event. Based on the roles attached to the event, it instantiates the corresponding implications according to the rules. As the ESO reasoner reads the rules it applies directly from the ESO Domain Ontology (i.e. rules are not hard-coded in the module), these rules can be revised or adapted without any adaptation of the module itself.[35]

7. KnowledgeStore and scalability

The NewsReader system consists of a series of software modules to process text and generate NAF and RDF output. The input text and the results of the processing are stored in a KnowledgeStore that supports reasoning and inferencing over the data and provides access to the data through various APIs. In this section, we explain the overall architecture and principles behind the KnowledgeStore and the capacity of the system to handle massive data. We describe the data sets processed in the project and the scalability issues of processing this data.

7.1. The KnowledgeStore

The KnowledgeStore[36] [64] is a framework that contributes to bridge the unstructured and structured worlds, enabling to jointly store, manage, retrieve, and semantically query both typologies of contents. First, the KnowledgeStore allows the user to store in its three interconnected layers all the typologies of content that have to be processed and produced when dealing with unstructured content and structured knowledge:

- the *resource layer* stores the unstructured news and their annotations;
- the *mention layer* identifies fragments of news denoting entities/events (e.g. a take-over event), relations between entity/event mentions (e.g., event participation), numerical quantities (e.g. a share price);
- the *instance layer* stores the structured descriptions of those instances extracted from resources and merged with available structured knowledge (e.g. Linked Data sources, corporate databases).

Second, as shown in Fig. 19, the KnowledgeStore acts as a shared data space supporting the interaction of the modules and tools (see Section 3), which can be roughly classified in:

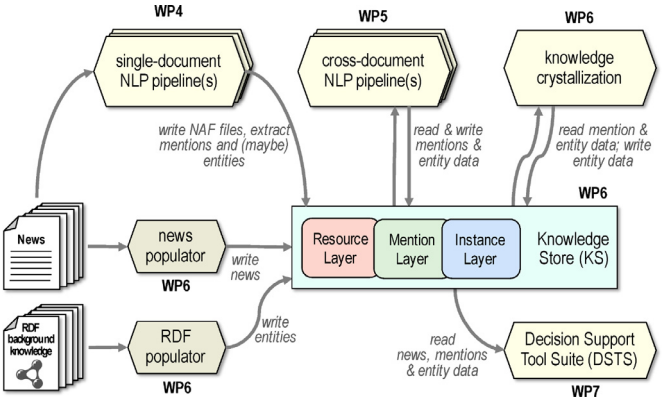

Fig. 19. The role of the KnowledgeStore in NewsReader.

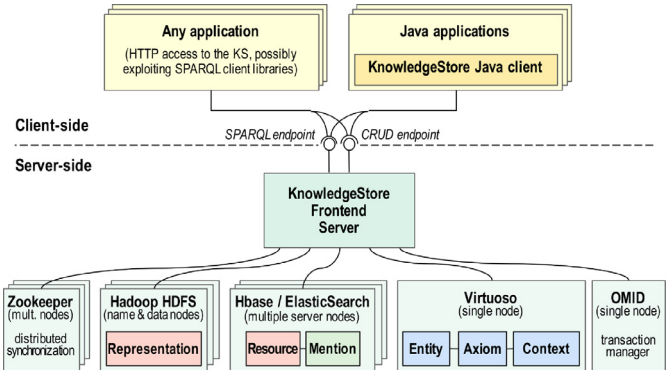

Fig. 20. KnowledgeStore architecture.

- *News* and *RDF populators*. These modules, developed as part of the NewsReader activities, enable the bulk loading of structured and unstructured content in the KnowledgeStore. The former processes a collection of linguistically annotated news documents injecting content in all three layers of the KnowledgeStore, while the latter augments the instance layer with Semantic Web compliant resources available in RDF repositories.
- *single-document NLP pipelines*. These pipelines (Section 3) work at the resource layer, and take care of processing a text document enriching it with linguistic annotations.
- *cross-document NLP pipelines*. These modules work at the mention and instance layers, exploiting the work of the NLP pipelines to instantiate, link, or enrich instances performing tasks such as cross-document coreference (see Sections 4 and 5).
- *Decision Support Tool Suite (DSTS)*. Finally, the decision support tool suite queries the KnowledgeStore – mainly the instance layer (although queries may also require to retrieve documents and mentions) – to obtain the information about events and narrative stories to be shown to users.

7.1.1. The KnowledgeStore architecture

As introduced in Fig. 19, the KnowledgeStore is a storage server: the other NewsReader modules are KnowledgeStore clients that utilize the services it exposes to store and retrieve all the shared content they need and produce. Fig. 20 shows the overall KnowledgeStore architecture, highlighting its client-server nature.[37]

[33] https://github.com/dkmfbk/eso-reasoner.
[34] http://rdfpro.fbk.eu/.
[35] An online demo for this tool is available on the project page: https://knowledgestore2.fbk.eu/reasoner/.
[36] http://knowledgestore.fbk.eu.
[37] Not shown in Fig. 20 are the additional tools and scripts for managing the complexity of software deployment in a cluster environment (potentially a cloud environment); they include, for example, the management scripts for infrastructure de-

7.1.1.1. Client side. The client side (upper part of Fig. 20) consists of a number of applications that access the KnowledgeStore through its two CRUD and SPARQL endpoints, either by direct HTTP interaction (for applications in any programming language), using the specifically developed Java client (for Java applications) or any of the available SPARQL client libraries for accessing the SPARQL endpoint, thanks to its standard-based nature.

7.1.1.2. Server side. The server side part of the architecture (lower part of Fig. 20) consists of a number of software components distributed on a cluster of machines that are accessed through a KnowledgeStore frontend server:

- the *Hadoop HDFS* filesystem provides a reliable and scalable storage for the physical files holding the representation of resources (e.g., texts and linguistic annotations of news articles);
- the *HBase* column-oriented store builds on the Hadoop filesystem to provide databases services for storing and querying semi-structured information about resources and mentions;
- the *Virtuoso* triple store stores and indexes crystallized axioms to provide services supporting reasoning and online SPARQL query answering, which cannot be easily and efficiently implemented in HBase or Hadoop;
- the *OMID* transaction manager[38] is used in combination with HBase to enforce the transactional guarantees of KnowledgeStore API operations;
- the *ZooKeeper* synchronization service is used to access and manage HBase nodes;
- the *KnowledgeStore frontend server* has been specifically developed to implement the operations of the two CRUD and SPARQL endpoints on top of the components listed above, handling global issues such as access control, data validation and operation transactionality;
- the *ElasticSearch* data store provides a database service for storing and querying semi-structured data; it can be used as an alternative of *HBase* and it does not need any additional software and/or server installed; in addition, it allows the user to install a stand-alone version of the KnowledgeStore , without the need of a multi-machine environment.

7.2. Data sets

The NewsReader system has been applied to various collections of news in different languages. In Table 3, we give some statistics for some of these data sets.

Wikinews[39] is a free multilingual open general 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. This generated over 2.6M mentions of events and entities, representing 624K event instances and 45K entities. We furthermore see how these entities are divided over persons, organizations and locations and how many of these have been mapped to DBpedia. We extracted 9.7M triples from the news and the KnowledgeStore contained over 95.9M triples with background knowledge in addition. The KnowledgeStore instance populated with Wikinews is publicly available.[40]

We also applied the system to 212K articles on the FIFA worldcup in 2014 in Brazil, provided by LexisNexis, the BBC and the Guardian. This larger set yielded over 76.1M mentions and over 9.3M instances. We extracted over 136M triples from the news, whereas 109M triples were extracted from DBpedia as background knowledge.[41]

The largest data set was provided by LexisNexis on the automotive industry, consisting of nearly 2.5M English articles spanning the period 2003 - 2015. This yielded over 842M mentions of events and entities, resolving to 42M event instances and 2.2M entities. In total, we extracted over 1.1 billion triples from the news in combination with 94 million triples from DBpedia.

The Airbus data set is the result of an experiment to show how our tools and methodology work with a cross-lingual corpus. We see here that we used 30 original English documents from Wikinews but also their translations to Spanish and Dutch as input to derive the instances representations and triples across all three sets of documents. The documents were processed by the pipelines in each language creating a set of interoperable NAF files. We then treated these NAF files in the same way as a set of NAF files from a single language trying to resolve cross-document event and entity coreference. In this data set we thus have mentions of the same events and entities across the three[42] languages that should ultimately resolve to the same entities and event instances in RDF. To the best of our knowledge, no state of the art tool is capable of producing such event-centric representations of the content of news articles in different languages. In Section 8, we provide more details on the data set and the experiment.

The final data set, Dutch House, was created for a pilot for the Dutch House of Parliament to process their archive created for a parliamentary enquiry on the bank-crisis in the Netherlands. We processed over 627K Dutch documents, yielding over 17,5M mentions of over 5.3M event instances and over 354K entities. In total over 122M triples were extracted from the data set and 188M triples were obtained from DBpedia as background information.

The data sets illustrate that the system can handle massive amounts of data on different domains and different languages without any further adaptation. The generic knowledge resources play an important role to obtain sufficient coverage for these data sets. In addition to processing the textual sources as such, we apply reasoning to the data to derive implications from the event data using ESO. For the cars data set for instance, the reasoner generated many additional triples representing 32M situations (15.5M pre-situations, 15.4M post-situations, 1M during-situations). In Section 8, we will discuss the quality of the data as produced by the generic system. Obviously, the generic system can be adapted to specific domains by extending DBpedia with lacking data, replacing DBpedia by other resources on entities or by adapting ESO to events that play an important role or implications that matter.

7.3. Scalability

In order to truly follow the news, we need to deal with massive amounts of data and elaborate linguistic analyses are computationally expensive. We collaborated with SURFsara to examine how much news our system can handle in a day with the current English pipeline. This research focused on processing a large batch of data as quickly as possible, i.e., rather than increasing the efficiency of components in the pipeline or reducing the processing

ployment (e.g. start-up & shut-down daemons, data backup & restoration and gathering of statistics).

[38] https://github.com/yahoo/omid.

[39] http://www.wikinews.org *Last accessed: 7 April 2015*.

[40] https://knowledgestore2.fbk.eu/nwr/wikinews/ui

[41] This data set was used for a public Hackathon in June 2014, http://www.eventbrite.com/e/kick-off-newsreader-and-hack-100000-world-cup-articles-tickets-2848605255

[42] The Italian pipeline was not yet completed at the time this corpus was created.

Table 3
Statistics of the Event-Centric Knowledge Graphs built during the project.

	WikiNews	FIFA WorldCup	Cars (Ver. 3)	Airbus Corpus	Dutch House
Topic	General News	Sport, Football	Automotive Industry	Airbus A380	Bank crisis
News Providers	Wikinews	LexisNexis, BBC The Guardian	LexisNexis	Wikinews	Dutch House of Parliament
Language	English	English	English	English, Dutch, Spanish	Dutch
Populated in	February 2015	May 2014	October 2015	February 2015	June 2015
News Articles	18,510	212,258	2,316,158	90 (30 EN. 30 NL. 30 ES)	627,341
Mentions	2,629,176	76,165,114	842,639,827	6415	17,583,997
Events	624,439	9,387,356	42,296,287	2574	5,383,498
Entities	45,592	858,982	2,263,156	934	354,857
Persons	19,677	403,021	895,541	71	28,799
in DBpedia	9744	40,511	126,140	19	949
Organizations	15,559	431,232	1,139,170	806	10,803
in DBpedia	6317	15,984	44,458	774	356
Locations	10,356	24,729	228,445	57	32,046
in DBpedia	7773	16,372	76,341	53	1056
Triples	105,675,519	240,731,408	1,240,774,944	95,994,233	310,961,410
from Mentions	9,700,585	136,135,841	1,146,601,954	19,299	122,665,094
from DBpedia	95,974,934	104,595,567	94,172,990	95,974,934	188,296,316
distilled from	DBpedia 2014	DBpedia 3.9	DBpedia 2015	DBpedia 2014	DBpedia 2015

time of the pipeline per document,[43] we optimized the overall process by analyzing large amounts of documents in parallel. This was done using Hadoop[44] [65]. We will provide a brief description of our approach and report on the processing setup and time for the largest dataset of nearly 2.5 million news articles. More details on these approaches can be found in Kattenberg et al. [66].

Hadoop is a framework that can distribute processing across clusters of machines. Hadoop Apache provides several libraries for developing parallel applications. However, in NewsReader we are dealing with a variety of existing applications that have different requirements. We therefore use the Cascading software library.[45] This library can handle complex workflows without reimplementing individual components. The Cascading architecture can combine any sequence of modules that take standard input and produce standard output,[46] as is the case for the NewsReader pipeline in which all modules in NewsReader read and produce NAF.

The largest dataset processed within NewsReader consisted of 2,498,633 documents. These documents were part of 11 years of news on the car industry provided by LexisNexis and were selected based on their length (1,000 - 4000 characters). The data was processed on SURFsara's Hadoop cluster consisting of 170 nodes, 1400 cores and 2 petabytes of data storage. Processing the entire corpus cost 198,134 CPU hours (the average time per document lying around 4.4 minutes). Hadoop is shared among users and we managed to process approximately 4000 documents per hour on average. In ideal circumstances, with the full Hadoop cluster being available, it would take 141.5 hours to process the full corpus, with an average of approximately 425,000 articles per day.

Regarding the KnowledgeStore , several tests assessing its scalability were performed. We assessed the performance both in data population and data retrieval. The complete analysis is described in Corcoglioniti et al. [64]. For data population, we analyzed both the impact of resource size (i.e., number of mentions per NAF file) and of the dataset size in the population of the resource and mention layers of the KnowledgeStore .[47] For the impact of resource size analysis, results show that the population rates inversely correlate with the average number of mentions per news article, while for the impact of dataset size, the population rate can be considered roughly constant during the whole population process, thus suggesting that consistent population performances can be achieved given the software infrastructure the KnowledgeStore builds on. For data retrieval, we tested the performances of the data retrieval operations offered by the KnowledgeStore (SPARQL queries and resource, mention and file retrieval) with different dataset sizes and numbers of concurrent clients. Adding new clients determines an increase of throughput with minor changes of the evaluation time up to a certain threshold, after which all the physical resources of the system (mainly CPU cores) are saturated, the throughput remains (almost) constant, and the evaluation time increases linearly as requests are queued for later evaluation. Concerning the effect of the dataset size on retrieval performances, a ~15 times increase in the number of news articles, from 81K news articles to 1.3M news articles, caused 'only' a ~2 times decrease in the throughput, from 21,126 to 10,212 requests/h for 64 clients. We believe all these findings are extremely significant for the practical adoption of the system, as all the evaluations were made on real-world data. Note that the tools for running the evaluation, in particular those for testing the data retrieval performances of a KnowledgeStore instance, are included in the KnowledgeStore source code, and documented on the KnowledgeStore web-site.[48]

The KnowledgeStore was also successfully exploited in two NewsReader Hackathons organized in Amsterdam and London in January 2015, as well as during two User Evaluations, held in Amsterdam in January 2015 and November 2015. During these events, the running KnowledgeStore instance was accessed through its API and the SPARQL endpoint, and it effectively handled a large amount of queries (117K), with peaks of 40 requests per second.

8. Evaluation

The NewsReader system consists of a cascade of NLP modules and a single high-level component that takes the output of the NLP

[43] Within NewsReader, we also worked on reducing processing time for a single document for a live stream setting. In this approach, we apply NLP modules in parallel where possible reducing the average processing time per document by half. The two approaches support different scenarios (dealing with a batch of document or an individual document as quickly as possible).

[44] https://hadoop.apache.org/.

[45] http://www.cascading.org

[46] The design and implementation of the Cascading system architecture were carried out by Mathijs Kattenberg from SURFsara.

[47] We ignored the population of the instance layer: the population of the resource and mention layers is around three order of magnitude slower than the population of the instance layer, and thus dominates and determines the overall population performances.

[48] https://knowledgestore.fbk.eu/test-tools.html

Table 4
Document level annotation in English (EN), Dutch (NL), Italian (IT), and Spanish (ES) in the MEANTIME corpus.

	EN	NL	IT	ES
# files	120	120	120	120
# sentences	597	597	597	597
# tokens	13,981	14,647	15,676	15,843
EVENT_MENTIONS	2096	1510	2208	2223
ENTITY_MENTIONS	2790	2729	2709	2704
TIMEX3	525	480	507	486
REFERS_TO	2983	2516	3054	3015
TLINK	1789	1516	1711	2186
CLINK	50	48	61	61
HAS_PART	1978	1930	1865	2152

modules as input to generate the ECKGs, which are RDF triples. Obviously, the quality of the different NLP modules determines to a large extent the quality of the final ECKGs in RDF. However to some extent, the semantic resources and models can also eliminate errors from the NLP modules that do not make any sense. In this section, we will first describe the evaluation results of the main NLP modules that produce semantic output: NERC, NED, SRL, TIMEX and next specific evaluations that focus on the ECKGs as the final output of the system which is built from the NLP output. First, we describe the specific evaluation data sets that were developed in the project for evaluation.

8.1. Evaluation data

We annotated two specific data sets for the evaluation of NewsReader. The MEANTIME corpus [67] was developed to test the high-level semantic NLP modules. The ECB+ corpus [68] on the other hand was developed for the evaluation of the cross-document event-coreference, which is the basis of the ECKGs.

8.1.1. MEANTIME

The NewsReader MEANTIME (Multilingual Event ANd TIME) corpus is a semantically annotated corpus of 480 English, Italian, Spanish, and Dutch news articles [67].[49] The English section of the corpus consists of articles from Wikinews (http://en.wikinews.org) about four topics: *Airbus and Boeing, Apple Inc., Stock market*, and *General Motors, Chrysler and Ford*. The Spanish, Italian and Dutch sections are translations of the English articles aligned at the sentence level. The texts have been manually annotated in each language at multiple levels, including entities, events, temporal information, semantic roles, and intra-document and cross-document event and entity coreference Table 4 presents statistics about the MEANTIME corpus.

8.1.2. ECB+

The Event Coreference Bank or ECB was developed by Bejan and Harabagiu (2010) [69] to test cross-document event coreference. It contains 43 different seminal events or so-called topics with about 10 to 20 different news articles reporting on this event. Across these articles, many mentions of events are coreferential. The ECB+ corpus [68] is an extended and re-annotated version of ECB, where we extended the ECB topics with texts about different event instances of the same event type. For example in addition to the topic of a specific celebrity checking into a rehab presented in ECB, we added descriptions of another event involving a different celebrity checking into another rehab facility. Likewise, we increased the referential ambiguity for the event mentions. Table 5 shows some examples of the seminal events represented in ECB+ with different event instances. Table 6 shows some statistics on the data, most notably 1983 coreference chains, corresponding to instance in the NewsReader terminology, group 6833 mentions of events. On average, 1.8 sentence per article was annotated.

8.2. NLP Modules

In Table 7, we show an overview of the results on standard data sets in the literature for the main NLP modules in the pipeline described in Section 4.1: NERC, NED, SRL and TIMEX detection and normalization. We provide the results for four languages (although for some languages, evaluation data is not available for every task). Every module is evaluated using the standard metrics and datasets for each task and compared with the state-of-the-art. All the NewsReader modules obtain state of the art performances for every task and language [70]. For NERC, we can see that NewsReader improves over the state-of-the-art for English, Spanish and Dutch and we used the state-of-the-art system from Evalita 2007 for Italian. For NED, the English and Spanish results perform lower than the state-of-the-art, but the use of DBpedia Spotlight in NewsReader was also motivated by its suitability for integration of a ready to use NED service for all four languages of the project and the option of easily building modules for disambiguation and wikification.

For NED, the evaluation metrics used in standard datasets and MEANTIME differ. In the case of TAC 2011 and TAC 2012, the measure used is the B-Cubed+ F_1[50] which measures the correctness of clusters. In contrast, we used a standard F_1 measure on the MEANTIME task, which makes the results difficult to compare. In addition, the MEANTIME NED results also depend on the performance of the NERC system since we applied NED to the automatically extracted entities. As mentioned in 4.1, we use the MATE tools [71] for English and Spanish SRL, which achieve the state-of-the-art on the CoNLL 2009 task for English. This module also obtains comparable results to the state of the art for Spanish in the same task. For the TIMEX detection and normalization there are only few results to compare with. NewsReader either sets the state-of-the-art or is very close to it.

The standard data sets and tasks mentioned above often provide both the training data and the test data. These results can be considered in-domain evaluations, where supervised machine learning is the most common approach. In Table 8 we therefore show the results of studying the performance of NewsReader NLP modules in out-of-domain data using the MEANTIME corpus [67].[51]

The text genre of MEANTIME is Wikinews, which is not that different from the standard datasets evaluated in Table 7. However, differences in the gold standard annotation of MEANTIME result in significant disagreements regarding the span of the annotations [67]. For example, named entity spans in MEANTIME differ from standard datasets such as CoNLL 2002 and 2003 as mentions include modifiers, for example articles: 'the United States' versus 'United States', or adjectives: 'new, faster iPhone' versus 'iPhone'. Regarding the SRL, the annotation in MEANTIME sets the relations between events as SLINKs. In other words, events are not annotated as roles of other events. However, these cases are taken into account by our SRL module because it has been trained with the CoNLL 2009 dataset. This is reflected in the performance of our SRL modules in MEANTIME.

The phrase based F_1 evaluation used in both in-domain and out-of-domain settings punishes any bracketing error as both false positive and negative. Thus, these span-related disagreements

[49] http://www.newsreader-project.eu/results/data/wikinews/

[50] The scorer is available at http://www.nist.gov/tac/2012/KBP/tools/.

[51] By out-of-domain, we mean that the models were trained on other data sets than the one tested.

Table 5
Overview of seminal events in ECB and ECB+, topics 1–10.

Topic	Seminal event type	Human participant		Time		Location		Number of documents	
		ECB	ECB+	ECB	ECB+	ECB	ECB+	ECB	ECB+
1	rehab check-in	T. Reid	L. Lohan	2008	2013	Malibu	Rancho Mirage	18	21
2	Oscars host announced	H. Jackman	E. Degeneres	2010	2014	–	–	10	11
3	inmate escape	Brian Nicols, 4 dead	A.J. Corneaux Jr.	2008	2009 Atlanta	court house Texas	prison	9	11
4	death	B. Page	E. Williams	2008	2013	LA	LA	14	10
5	head coach fired	Philadelphia 76ers M. Cheeks	Philadelphia 76ers J. O'Brien	2008	2005	–	–	13	10
6	"Hunger Games" sequel negotiations	C. Weitz	G. Ross	2008	2012	–	–	9	11
7	IBF, IBO, WBO titles defended	W. Klitchko H. Rahman	W. Klitchko T. Thompson	2008	2012	Germany	Switzerland	11	11
8	explosion at bank	–	–	2008	2012	Oregon	Athens	8	11
9	ESA changes	Bush	Obama	2008	2009	–	–	10	13
10	eigth-year offer	Angels M. Teixeira	Red Socks M. Teixeira	2008	2008	–	–	8	13

Table 6
ECB+ statistics.

ECB+	#
Topics	43
Texts	982
Action mentions	6833
Location mentions	1173
Time mentions	1093
Human participant mentions	4615
Non-human participant mentions	1408
Coreference chains	1958

make this setting extremely hard for models trained according to other annotation guidelines, as shown in Table 8.

For comparison, we also run the state-of-the-art systems on the MEANTIME data for NERC. The results show that also these systems suffer from the genre and annotation differences between the standard data sets and MEANTIME. However, the results clearly demonstrate that the Newsreader NERC module performs better in the MEANTIME out-of-domain evaluation settings than the state-of-the-art systems at this time.

8.3. Event-centric knowledge graphs

We performed four types of evaluations to test the quality of the ECKGs produced by NewsReader on top of the output of the NLP modules described in the previous subsection: 1) event coreference across different documents, 2) RDF triples extracted, 3) reasoning over event implications using ESO and 4) the cross-lingual interoperability of our reading technology. We will discuss these evaluations in the following subsections.

8.3.1. Cross-document event coreference evaluation

The MEANTIME data hardly contains any cross-document event coreference data since the news originates from a single source and is spread over time for specific entities. For the cross-document event-coreference, we therefore used the ECB+ data set that is specifically designed for this purpose. We compare the

Table 7
Benchmarking of NLP modules using standard metrics and datasets.

		English	Spanish	Dutch	Italian
NERC	Standard dataset	CoNLL 2003	CoNLL 2002	SoNaR	Evalita 2007
	SoA reference	Passos [72]	Carreras [73]	Desmet [74]	Zanoli [75]
	SoA F_1	90.90	81.39	84.91	82.10
	NewsReader F_1	91.36	84.16	87.72	82.10
NED	Standard dataset	TAC 2011	TAC 2012	N/A	N/A
	SoA reference	Barrena [76]	Monahan [77]		
	SoA F_1	81.55	62.22	N/A	N/A
	NewsReader F_1	62.90	50.00	N/A	N/A
SRL	Standard dataset	CoNLL 2009	CoNLL 2009	SoNaR	N/A
	SoA reference	Nugues [71]	Zhao [78]		
	SoA	85.63	80.46		N/A
	NewsReader F_1	85.63	79.91	74.02[a]	N/A
TIMEX detection	Standard dataset	TempEval3	TempEval3	N/A	Evalita 2014
	SoA reference	Lee [79]	Strötgen [51]		Mirza [53]
	SoA	83.10	85.33		82.70
	NewsReader F_1	84.71	85.33	N/A	82.70
TIMEX normalization	Standard dataset	TempEval3	TempEval3	N/A	Evalita 2014
	SoA reference	Lee [79]	Strötgen [51]		Manfred [80]
	SoA	82.40	85.33		70.90
	NewsReader F_1	72.16	85.33	N/A	68.40

[a] The Dutch SRL evaluation was carried out as follows. The e-magazines, magazines, press and newspaper portion of the SoNaR corpus were parsed by Alpino. Alpino's output was matched with the gold and for those instances that matched (leading to 47,889 instances in total), 10-fold cross-validation was carried out. The evaluation thus reflect semantic role classification and not the detection of predicates and roles.

Table 8
F_1 scores for out-of-domain benchmarking of NLP modules using MEANTIME.

	English	Spanish	Dutch	Italian
SoA reference	Stanford NER	Stanford NER	Desmet [74]	Zanoli [75]
SoA F_1	66.96	47.48	48.44	46.85
NERC	70.90	62.14	63.93	46.85
NED	64.22	65.87	51.44	60.37
SRL	34.78	29.68	26.76	31.62
TIME detection	80.50	78.30	50.20	85.70
TIME normalization	68.50	62.20	41.90	64.60

Table 9
Reference results macro averaged over ECB+ corpus as reported by Yang et al. [81] for state-of-the-art machine learning systems as compared to 3 NewsReader based systems: NWR−GOLD = the results of cross-document corefence using the gold-data for event detection, NWR−ARM = standard setting of NewsReader with at least one matching participant in any role (AR), time month match and action concept and phrase match with 30%, TEvalGOLD = cross-document results using event detection CRF module trained with TempEval 2013 gold data.

ECB+	MUC			BCUB			CEAFe			CoNLL
Topics 24–43	R	P	F_1	R	P	F_1	R	P	F_1	F_1
LEMMA	55.4%	75.10%	63.80%	39.60%	71.70%	51%	61.10%	36.20%	45.50%	53.40%
HDDCRP	**67.10%**	**80.30%**	**73.10%**	**40.60%**	**73.10%**	**53.50%**	**68.90%**	**38.60%**	**49.50%**	**58.70%**
NWR	42.58%	50.08%	46.03%	45.64%	44.99%	45.31%	46.94%	33.08%	38.81%	**43.38%**
TEvalGOLD	39.41%	59.89%	47.54%	40.32%	58.82%	47.85%	44.16%	38.02%	40.86%	**45.42%**
NWR-GOLD	44.97%	71.63%	55.25%	56.62%	80.47%	66.47%	75.62%	45.15%	56.54%	**59.42%**

NewsReader results with Yang et al. [81], who report the best results on ECB+ and compare their results to other systems that have so far only been tested on ECB and not on ECB+. Yang et al.use a distance-dependent Chinese Restaurant Process (DDCRP [82]), which is an infinite clustering model that can account for data dependencies. They define a hierarchical variant (HDDCRP) in which they first cluster event mentions and data within a document and next cluster the within document clusters across documents. Their hierarchical strategy is similar to our approach using event components, in the sense that event data can be scattered over multiple sentences in a document and needs to be gathered first. Our approach differs in that we use a semantic representation to capture all event properties and do a logical comparison, while Yang et al. as well as the other methods they report on are based on machine learning methods (both unsupervised clustering and supervised mention based comparison). Yang et al. also report on a lemma-baseline as proposed by Cybulska and Vossen [68], where all event mentions with the same lemma within and across documents are simply joined in a single coreference set.

Yang et al. test their system on topics 24–43 while they used topics 1–20 as training data and topics 21–23 as the development set. They do not report on topics 44 and 45. To compare our results with theirs, we also used topics 24–43 for testing. In Table 9, we give Yang's lemma baseline (LEMMA), Yang's best results (HDDCRP), and the out-of-the-box results for NewsReader results (NWR), in which at least one participant needs to match regardless of its role, the events need to have matching WordNet synsets for 30% and the time-anchors need to have the same month and year value. This NewsReader system is not trained on the ECB+ data set at all and just uses logical comparison of event data.

First of all, we can see that both Yang's *HDDCRP* and the lemma baseline outperform NewsReader system by 15 and 10 points in CoNLL F_1 score [83], which is the average of the F_1 scores for MUC [84], B3 [85], CEAF [86]. However, Yang et al. report that their system at first had an out-of-the-box accuracy for event detection of 56%. They therefore trained a separate Conditional Random Field (CRF) event detection system with event annotations of the first 20 topics (about half of the data set). This classifier has an accuracy of 95% on event detection and was used as the input for both the LEMMA baseline as HDDCRP. For comparison, the NewsReader system has an out-of-the-box accuracy of 67.1%, where events are detected by the MATE tool which is trained on PropBank data. Clearly, what events have been annotated and how they were annotated has a big impact on the results.

To see the impact of the event detection on the actual event-coreference results, we therefore add two other versions of the NewsReader system: 1) TEvalGOLD replaces the NewsReader event detection by a CRF classifier trained with SemEval 2013 - TempEval 3 gold data [87] and 2) NWR-GOLD used the gold-annotation of the event detection. The event detection accuracy of TEvalGOLD is 73.1% and the accuracy for NWR-GOLD is 97%.[52] We can see that TEvalGOLD performs 2 points higher on event-coreference and NWR-GOLD outperforms Yang et al.by almost 1 point even though the NWR-GOLD event detection accuracy is only a little higher than Yang's. Also note that the NewsReader event-coreference uses logical semantic comparison and is not trained on the ECB+ data set. We thus can expect its performance to be relatively stable across data sets, whereas Yang et al.s system is expected to perform significantly lower when applied to out-of-domain data.

8.3.2. Triple evaluation

Event coreference leads to RDF structures with triples for event relations with participants and time anchorings. It thus makes sense to evaluate the triples in addition to event coreference. The results reported here have been described in [1]. The evaluation was conducted on 100 randomly selected events extracted from the MEANTIME dataset. These events yielded 1043 triples of the RDF-SEM data, with each triple independently evaluated by two raters. Raters checked the triples against the original sources from which they have been extracted by resolving the *gaf:denotedBy* relations. A strict evaluation was applied: a mistake in any element of the triple qualifies the whole triple as wrong.

Table 10 presents the resulting triple accuracy on the whole evaluation dataset, as well as the accuracy on each subgraph composing it, obtained as average of the assessment of each rater pair.

[52] The reason that it is not 100% is because the NewsReader system could not process one of the evaluation files due to formatting problems.

Table 10
Quality triple evaluation of SEM-RDF extracted from MEANTIME.

	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*	

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 it is slightly lower for typing (0.522 on 496 triples) and participation triples (0.534 on 446 triples). Further drilling down on participation triples, the accuracy is higher for PropBank roles (0.559) while it 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 four 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 four, thus causing all the triples of the event to be marked as incorrect. To stress the aforementioned strict evaluation criteria, we note that, the triple accuracy rises to 0.697 on a total of 782 triples if we ignore all coreferring events (and their corresponding triples) in the evaluation dataset. Table 11 shows the details for both the full evaluation and the evaluation when the event coreference are ignored. Note that applying cross-document event coreference to MEANTIME does not make much sense since the news is spread over time and comes from a single source. It is therefore not surprising that cross-document coreference detection does more harm than good.

8.3.3. ESO

ESO and its implications are evaluated on MEANTIME. We performed a quality analysis on this corpus by passing it through the NewsReader pipeline, adding it to the KnowledgeStore and enriching the sets by applying the ESO reasoner. Table 12 provides a quantitative overview of the events and ESO classes that were found in the corpus. The pipeline identified 5443 distinct events, 2508 of which were linked to an ESO class. The "ESO events" included 444 events with inferred pre- and post-situations and 52 events that have a during-situation.[53]

We randomly selected one ESO event for low frequency classes against two ESO events for high frequency classes out of the events that inferred a situation. The total selection consisted of 43 events with pre- and post-situations and 9 with a during-situation leading to a total of 52 ESO events.[54] For these events, we checked the original sentence they were derived from and verified whether the ESO class and inferences made sense and the correct instances were identified. We found 37 events (71.1%) with a correct class label and 18 events (41.8%) with correct pre- and post-situations. The set of events with a during-situation was correct in 66.6% of the cases. Overall, 21 out of 52 inspected ESO events were found to be correct. Table 13 provides an overview of these results. The fact that the results for the pre- and post-situations (41.8%) are higher than the ESO roles assigned to the events (40% in the triple evalation) points to the phenomenon that NewsReader overgenerates events and roles. If these events and roles do not make sense as a combination according to ESO, they do not trigger a rule. Strict modeling of events through an ontology such as ESO thus results in crystalization of the extracted knowledge, i.e. only interpretations that lead to semantic closure within the model remain.

```
<http://DBpedia.org/resource/Airbus>
        rdfs:label "Airbus" ;
        gaf:denotedBy
        <nwr:1816_Airbus_wins_Qatar_Airways_order_worth_15bn.dutch#char=564,570> ,
        <nwr:1816_Airbus_wins_Qatar_Airways_order_worth_15bn.dutch#char=655,661> ,
        <nwr:1816_Airbus_wins_Qatar_Airways_order_worth_15bn.dutch#char=911,917> ,
        <nwr:1816_Airbus_wins_Qatar_Airways_order_worth_15bn.spanish#char=381,387> ,
        <nwr:1816_Airbus_wins_Qatar_Airways_order_worth_15bn.spanish#char=381,387> ,
        <nwr:1816_Airbus_wins_Qatar_Airways_order_worth_15bn.spanish#char=381,387> ,
        <nwr:1816_Airbus_wins_Qatar_Airways_order_worth_15bn.spanish#char=381,387> ,
        <nwr:1816_Airbus_wins_Qatar_Airways_order_worth_15bn.english#char=93,100> ,
        <nwr:1816_Airbus_wins_Qatar_Airways_order_worth_15bn.english#char=356,362> ,
        <nwr:1816_Airbus_wins_Qatar_Airways_order_worth_15bn.english#char=641,647> ,
        <nwr:1816_Airbus_wins_Qatar_Airways_order_worth_15bn.italian#char=641,647> ,
        <nwr:1816_Airbus_wins_Qatar_Airways_order_worth_15bn.italian#char=153,156> ,
        <nwr:1816_Airbus_wins_Qatar_Airways_order_worth_15bn.italian#char=872,878> .


nwr:ev#1
        a sem:Event , ili:i74228 , fn:Discussion , ili:i25503 , nwrontology:negociar.1.default , ili:i25518 ;
        rdfs:label "trattativa" , "negoziare" , "negotiation" , "negotiates" , "negotiate" , "negociar" , "
                onderhandelen" ;
        gaf:denotedBy
        nwr:8670_Aeroflot.english#char=9,19> ,
        nwr:8670_Aeroflot.italian#char=9,16> ,
        nwr:8670_Aeroflot.spanish#char=9,16> ,
        nwr:8670_Aeroflot.dutch#char=9,21> ,
        nwr:8670_Aeroflot.english#char=120,130> ,
        nwr:8670_Aeroflot.spanish#char=122,131> .
```

Fig. 21. RDF-TRiG representation of entities and events merged from English, Spanish and Dutch Wikinews.

8.3.4. Cross-lingual reading

The semantic interoperability of NewsReader makes it possible to test cross-lingual reading [88]. The MEANTIME data set contains translations of English articles to Dutch, Spanish and Italian by professional translators. In the ideal case that the pipelines for the four languages generate the same output, our conversion from NAF should generate the same RDF structure regardless of the input given, i.e. the RDF derived from the English NAF should be identical to the RDF derived from any of the translations or any combination. In this sense, cross-lingual processing is not different from merging interpretations across different NAF files in the same language, as we have discussed for the two titles of the English articles.

We therefore used the pipelines for English, Spanish, Italian and Dutch to process the MEANTIME data sets in the four languages. Next, we applied the IDAP procedure to each NAF file and across all the NAF files. In Fig. 21, we show a RDF-TRiG result for the entity *Airbus* and the event *negotiate* merged across English, Spanish, Italian and Dutch NAF files. Especially for the event, we see how different labels were merged and the ontological types are obtained across the different language NAF files. Both examples show mentions across the translations.

To carry out the cross-lingual evaluation, we compared the Spanish, Italian and Dutch data against the English data, calculating the coverage of the English mentions by the other languages for the above data. We cannot calculate true recall and precision since the English output cannot be seen as a gold standard. We

[53] Recall that situation rules are only triggered when the implied participants are present.

[54] The data and analysis can be found at https://github.com/newsreader/eso

Table 11
Detailed quality triple evaluation of SEM-RDF extracted from Wikinews with and without taking even-coreference into account.

	Group 1			Group 2			Group 3			Group 4			Overall	
	Trp	Avg	κ	Trp	Avg	κ	Trp	Avg	κ	Trp	Avg	κ	Trp	**Avg**
ALL	267	0.607	0.623	256	0.525	0.57	261	0.552	0.69	259	0.548	0.751	1043	**0.551**
TYPES	122	0.594	0.649	115	0.539	0.585	137	0.504	0.65	122	0.578	0.748	496	**0.522**
LABELS	28	0.768	0.7	26	0.788	0.661	24	0.729	0.684	23	0.804	0.862	101	**0.772**
ROLES	117	0.581	0.591	115	0.452	0.509	100	0.575	0.735	114	0.465	0.718	446	**0.534**
PROPBANK	39	0.628	0.629	39	0.423	0.633	30	0.583	0.796	28	0.554	0.928	136	**0.559**
FRAMENET	77	0.461	0.559	73	0.438	0.445	90	0.489	0.645	105	0.424	0.63	345	**0.438**
ESO	26	0.5	0.692	25	0.4	0.5	38	0.368	0.435	29	0.345	0.545	118	**0.407**
Without event coreference:														
ALL	207	0.732	0.447	188	0.601	0.491	184	0.731	0.683	203	0.7	0.625	782	**0.697**
TYPES	91	0.731	0.483	91	0.566	0.578	96	0.672	0.649	93	0.758	0.561	371	**0.663**
LABELS	23	0.87	0.617	20	0.925	0	17	0.912	0.638	19	0.974	0	79	**0.937**
ROLES	93	0.699	0.419	77	0.558	0.37	71	0.768	0.724	91	0.582	0.639	332	**0.678**
PROPBANK	32	0.734	0.472	25	0.56	0.516	20	0.825	0.828	23	0.674	0.901	100	**0.72**
FRAMENET	56	0.58	0.387	58	0.457	0.41	70	0.607	0.615	79	0.563	0.512	263	**0.548**
ESO	17	0.706	0.433	20	0.375	0.468	21	0.571	0.438	21	0.476	0.432	79	**0.557**

Table 12
ESO related statistics of the populated KnowledgeStore of the MEANTIME Corpus.

Component	Number
Events	5443
ESO events	2508
ESO events with ESO roles	736
ESO events with pre and post situations	444
ESO events with a during situation	52

Table 13
Results of the analysis of ESO events with during or pre/post situation assertions derived from the MEANTIME corpus.

ESO events with pre/post or during situation	495
Number of events inspected	52 (10.5%)
Number events insp. with a pre/post situation	43
Number events insp. with a during situation	9
Correct class label	37 (71.1%)
Correct pre and post situation(s)	18 (41.8%)
Correct during situation(s)	6 (66.6%)
Correct ESO events	21 (50%)

```
1  Triples in all 4 languages
2  ili-30-00975427-v;ili-30-00974367-v[announce]:hasActor:Boeing
3  ili-30-00975427-v;ili-30-00974367-v[announce]:hasActor:Airbus
4  ili-30-02646757-v;ili-30-02207206-v[buy]:hasActor:European_Union
5  ili-30-00761713-v[negociate]:hasActor:Aeroflot
6  ili-30-02244956-v;ili-30-02242464-v[deal;sell]:hasActor:Airbus
7  ili-30-00634472-v[conclude]:hasActor:Boeing
8  ili-30-00882948-v;ili-30-00875141-v[commend;advocate]:hasActor:Airbus
9  ili-30-00354845-v;ili-30-00358431-v[die;buy_the_farm]:hasActor:Steve_Jobs
10 ili-30-01734502-v;ili-30-00246217-v[duplicate;double]:hasActor:Apple_Inc.
11 ili-30-00975427-v;ili-30-00974367-v[announce]:hasActor:Starbucks
12 ili-30-00975427-v;ili-30-00974367-v;ili-30-00820801-v;ili-30-01010118-v[announce;declare]:
13   hasActor:United_States
14 ili-30-01182709-v;ili-30-02327200-v;ili-30-02479323-v[provide;furnish;issue]:
15   hasActor:General_Motors
16 ili-30-00975427-v;ili-30-00974367-v;ili-30-00820801-v;ili-30-01010118-v[announce;declare]:
17   hasActor:Ford_Motor_Company
18 ili-30-02244956-v;ili-30-02242464-v[deal;sell]:hasActor:Opel
19 ili-30-00975427-v;ili-30-00974367-v;ili-30-00820801-v;ili-30-01010118-v[announce;declare]:
20   hasActor:General_Motors
```

Fig. 22. Identical triples across different languages.

evaluated the results with respect to the entities, the events and the triples extracted. Table 14 gives the results.

For the entities, where we only consider entities matched to DBpedia, we see that all four languages generate more or less the same number of instances and mentions. The overlap of Spanish, Italian and Dutch with English is also very compatible, with macro and micro averaged coverage of 35.5 up to 40.3 and 37.1 up to 44.8 respectively. Obviously for events the coverage is lower and varies more. The Italian pipeline detected more events than English, whereas Spanish and Dutch detected about half of the events compared to English, both in terms of instances and mentions. Coverage results are nevertheless very similar across the languages, with Dutch performing a bit lower than the other languages. The Italian pipeline clearly over-generates events compared to the others. Overall the coverage of the events is lower than for entities. The latter applies even more for the overlap of triples. Although the amount of triples generated is just a bit lower than for English, all languages have very low coverage of the English triples. Obviously this is due to the constraint that the event, the entities and the roles need to match across the pipelines to have a positive coverage result.

Inspecting the results for the more frequent cases shows some interesting insights. First of all, the entity *United_States_dollar* with 146 mentions in English in the data set, turned out to be a systematic error in the English pipeline that is not mirrored by the other languages. The English pipeline erroneously linked mentions of the US to the dollar instead of the country. The second observation relates to the granularity of the mapping. For example in the case of the *airbus* data, *Boeing* is the most frequent entity in all four languages. The more specific entity *Boeing_Commercial_Airplanes* is however only detected in English and not in any of the other languages. This is due to the fact that the mappings across Wikipedia from the other language to English are at a more coarse-grained level.

For the events and triples there is no clear pattern emerging. The differences seem to relate to many different factors among which the difference in the resources used in the pipeline. Finally, Fig. 22 shows some examples of triples shared by all four languages. The above comparison is unique in its kind and provides an excellent basis for comparing semantic NLP pipelines across languages. In future research, we will put forward such data sets as tasks for cross-lingual semantic parsing to the community.

9. Discussion

The NewsReader system ultimately matches unstructured text with Semantic Web resources and standards. We rely heavily on knowledge resources in this process. Although our NLP systems perform at state-of-the-art level, the quality of the knowledge resources plays a major role, e.g. coverage of FrameNet and ESO determines the proportions of implications that we can derive. Knowledge resources are skewed in terms of the data given (pop-

Table 14
Entities, events and triples extracted for English, Spanish, Italian and Dutch Wikinews with proportion of coverage, measured as macro and micro coverage. I = instances, M = mentions, O = overlap, maC = macro-average over all document results, miC = microAverage over all mentions.

	English		Spanish					Italian					Dutch				
	I	M	I	M	O	maC	miC	I	M	O	maC	miC	I	M	O	maC	miC
DBPedia entities	376	2293	371	2069	1308	35.5	44.8	354	1507	963	37.1	41.2	394	1922	1041	40.3	43.7
Events	1309	3668	600	1672	879	26.4	24.1	1989	4024	822	32.6	23.0	686	1965	618	19.3	16.7
Triples	853	3410	355	1420	1423	3.1	3.1	572	2287	2301	0.6	0.6	420	1678	1706	0.7	0.7

ular entities have more data and are preferred due to overfitting) and still lack considerable amount of data (dark entities). Also the quality of these resources across languages varies considerably.

We applied deeper semantic evaluations of the end result of NewsReader both at the level of the triples extracted and the ESO types, roles and derived situations. Both evaluations are unique in its kind. We came to the interesting observation that semantic processing is on the one hand complex and error prone due to many dependencies across modules and resources but that there is also a crystalization effect. Crystalization means that errors that do not make sense or do not result in coherent pieces of information tend to get ignored in the final representation. As such we can see that F_1-measures of basic modules such as NERC that score higher than 80% on standard data sets (CoNLL) may drop to 70% when applied out of the training domain (MEANTIME) and drop further for semantic tasks that depend on this output to 43% (cross-document event coreference). However, our evaluation of the triples and the derived ESO situations still have accuracies around 55% and 50% respectively, even though they depend even more on the results of many submodules. This suggests that a strong conceptual model can filter out errors that do not make any sense in combination. Future systems thus should leave open alternative analyses, as is already done by many NewsReader modules, rather than selecting the highest scoring analysis. This leaves room for knowledge based approaches to select the most coherent interpretation based on a conceptual model, that can also be tuned towards a specific domain or user. Over-generating solutions may lead to low precision results for individual tasks, but it provides options for post-processing data in a knowledge-intense architecture.

To better understand the relations between the NLP modules and the quality of the final result, we carried out a specific study on the results for the SemEval 2015 Task 4: TimeLine: Cross-Document Event Ordering.[55] In this task, systems need to find all events in which an entity is involved and place them on a timeline given a set of documents. The MEANTIME data was specifically annotated for this task for 40 entities. The task requires almost all modules provide in NewsReader: detection of entities, events, roles, time expressions and temporal relations. It also requires cross-document identity.

It turned out that this task is extremely difficult, i.e. F_1 measures below 15%. We carried out an in-depth error analysis [89] in which we reversed the NewsReader pipeline to trace the modules responsible for the errors. This showed that most of the modules perform well although there is some piling up of errors from low-level modules to higher-level modules that depend on them. The main problem with respect to the quality is however that the high-level semantic modules (temporal relations, semantic-roles) rely too much on the sentence as a unit, while the relations and information is often not in a single sentence and in some cases not even in the document but based on world-knowledge. Recovering this information requires more intelligent reasoning over the information spread in the document. It also requires more intensive usage of knowledge in the processing than has been done so far.

[55] http://alt.qcri.org/semeval2015/task4/

Table 15
Mentions to instance reduction for 2.5M English news articles on the automotive industry from 2003 until 2015.

Type	Mentions	Instances	Ratio
Event	420,010,878	42,296,287	10.07%
Person	16,821,830	895,541	5.32%
Organization	23,841,719	1,139,170	4.78%
Location	11,839,365	228,445	1.93%

Cross-document event coreference turned out to be one of the major challenges for the future. Our evaluations show that cross-document event coreference performs below 50% but this is on an artificial task in which systematic two-fold ambiguity is created with about 10 different articles referring to each seminal event. It is difficult to estimate how this translates to realistic scenarios in which there can be thousands of news articles published on the same day that potentially refer to the same event. The possible impact of establishing event coreference on large data sets can be seen when we compare the ratio of mentions and instances. Table 15 shows these ratios for the processed 2.5M English news articles on the automotive industry. We can see that mentions of persons and organizations are reduced to 5% instances and locations to 2%, meaning that the former are mentioned 20 times on average in the news and locations 50 times.The difference between locations on the one hand and people and organizations on the other makes sense since news involves more different persons and organizations than different locations. If we look at the events, we see a lesser reduction to 10%, implying that events are mentioned 10 times on average in the news. We can consider the reduction of the persons, organizations and locations as the upper bound for coreference and the current event coreference reduction as the lower bound. It is difficult to judge of this reduction is right. In ECB+, most annotated events (about 95%) are not coreferential across documents, which means that there is only a single mention. The software therefore needs to be very conservative to establish coreference in order to perform. However, if we need to consider thousands of documents that report on the same event this may be very different. Furthermore, only 1.8 sentences per article have been annotated on average in ECB+. This may also reduce the degree of coreference, since articles may refer again to the same event in the remainder of the article.

It is interesting to realize that event coreference can be parameterized to a high degree depending on the type of news streams considered. By setting the constraints for establishing identity across events to a very strict level, e.g. precise date, same lemma for action, all participants matched with the same role, hardly any event mention will be coreferential. This will lead to many more events, little aggregation across sources and relatively sparse event-centric knowledge graphs. When we use very loose settings on the other hand, we will lump many events together, have less event-centric knowledge graphs with strong aggregation of relations and very rich graphs. The granularity of the event structures can also be seen as a user-driven parameter. For certain purposes, you may want to group all micro events in a single topical knowl-

edge graph with many subevents, whereas for others you may want to keep each of them separate.

So far, we mainly considered the ways in which knowledge resources play a role in NLP, with the overall goal of text understanding or deep reading. However, our system also generates massive data, especially episodic data on situations in which individuals are involved. The Semantic Web is mostly a collection of resources that express factual knowledge. Typically, semantic knowledge is more generic and less fluid. Whereas semantic knowledge defines what is possible according to our cognitive and cultural conceptualization, episodic knowledge defines what is actually the case. The NewsReader system thus uses semantic and episodic knowledge to learn from the news what is the case in the world. Specifically, our technology 'reads' about situations in which entities are involved that are included in DBpedia but in which these situations are not described. For example from the current data on the automotive industry, we extract 44,202 triples for the entity *dbp:Porsche* and 689 triples for the entity *dbp:Qatar_Investment_Authority* using over 10 years of English news. In DBpedia, we currently find 155 triples for *dbp:Porsche* and 70 triples for *dbp:Qatar_Investment_Authority*. Our technology can thus be applied to any textual source to generate new episodic knowledge that can be published to the Semantic Web. Cleaning, harvesting and crystalizing this knowledge is then a next step. We are however convinced that knowledge enhances knowledge and eventually suppresses noise. We demonstrated this already for the dark entities and for ESO. In both cases, we first applied the generic system to the data set to learn about the entities and the events that play a major role. By deriving the knowledge resources for these entities and events, the interpretation of the text in the domain can be improved efficiently without having to go through an expensive and painstaking annotation process.

10. Conclusions

In this article, we described the NewsReader system for deep reading of texts in four different languages. The system was designed to arrive at interoperable interpretations across different sources and across these languages. The high-level semantic processing relies heavily on multilingual and cross-lingual semantic knowledge resources. We also described our formal modeling of the interpretations of textual expressions. Our models (GAF, NAF and SEM) distinguish mentions from instances. We developed the IDAP procedure to derive situational representations for individuals from the interpretations of the textual mentions across various sources. We also demonstrated that we can derive the implications of these situations for individuals using a formal ontology (ESO) linked to the system, which operates within a KnowledgeStore environment that supports reasoning. All source code for the NLP pipelines, the cross-document RDF extraction, the ontologies and the KnowledgeStore are freely available through the Apache license on Github. Further instructions on downloading the source code and setting up the system are detailed on the NewsReader website.[56]

Overall, our NLP modules perform at the state-of-the-art level for the high-level tasks in all four languages. We have seen that the integrated results at the level that is required for an instance representation for situations according to Semantic Web standards still needs further improvements. Nevertheless, our system is a powerful platform for generating massive episodic knowledge that may eventually contribute to the Semantic Web. As such, we can deploy the system in a cyclic architecture in which textual resources are processed using the current semantic and episodic knowledge to produce more episodic knowledge. The newly learned episodic knowledge can be used to improve future processing of data, either directly or by deriving improved semantic knowledge from the massive data. Ultimately, we can thus create a machine that reads to learn and learns to read.

Finally, we have shown that we can create semantic data from textual sources across different languages. This demonstrates the capacity to build a platform for cross-regional and cross-cultural knowledge acquisition. This will also allow us to study different perspectives on the changes in the world, which opens up many new lines of research.

NewsReader made big progress on the integration of many high-level NLP tasks and Semantic Web technologies but also yielded new questions to be explored. Our data hides many stories and complex perspectives from different sources. Although, we developed a powerful system for creating huge knowledge graphs for events, we have only just started to explore how these events should be structured into storylines and larger structures. Nevertheless, such storylines are most natural to people to represent events and summarize the changes. In addition to generally improving the quality of our NLP systems, our future research thus also focuses on such larger and more complex structures.

[56] http://www.newsreader-project.eu/results/software/

Acknowledgement

We would like to thank the anonymous reviewers for their valuable feedback. The NewsReader project was co-funded by the European Union as project number: 316404, FP7 Work Programme Call FP7-ICT-2011-8 Objective Cooperation Research theme Information and Communication Technologies, challenge 4.4 - Area Intelligent Information Management. The creation of the larger datasets was carried out with the support of SURF Cooperative.

References

[1] M. Rospocher, M. van Erp, P. Vossen, A. Fokkens, I. Aldabe, G. Rigau, A. Soroa, T. Ploeger, T. Bogaard, Building event-centric knowledge graphs from news, J. Web Semant. (to appear). 10.1016/j.websem.2015.12.004

[2] 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, ISBN 978-1-937284-47-3. Atlanta, GA, USA, 2013.

[3] 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.

[4] 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.

[5] L. Moreau, P. Missier, K. Belhajjame, R. B'Far, J. Cheney, S. Coppens, S. Cresswell, Y. Gil, P. Groth, G. Klyne, T. Lebo, J. McCusker, S. Miles, J. Myers, S. Sahoo, C. Tilmes, PROV-DM: The PROV Data Model, Technical Report, 2012. URL http://www.w3.org/TR/prov-dm/.

[6] D. Shen, M. Lapata, Using semantic roles to improve question answering., in: EMNLP-CoNLL, 2007, pp. 12–21.

[7] G. Hirst, A. Budanitsky, Correcting real-word spelling errors by restoring lexical cohesion, Nat. Lang. Eng. 11 (01) (2005) 87–111.

[8] R. Grishman, B. Sundheim, Message understanding conference-6: a brief history., in: COLING, 96, 1996, pp. 466–471.

[9] 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.

[10] S. Patwardhan, E. Riloff, A unified model of phrasal and sentential evidence for information extraction, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, Association for Computational Linguistics, 2009, pp. 151–160.

[11] 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.

[12] J.L. Leidner, G. Sinclair, B. Webber, Grounding spatial named entities for information extraction and question answering, in: Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references-Volume 1, Association for Computational Linguistics, 2003, pp. 31–38.

[13] R. Mihalcea, A. Csomai, Wikify!: linking documents to encyclopedic knowledge, in: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, ACM, 2007, pp. 233–242.

[14] D. Milne, I.H. Witten, Learning to link with wikipedia, in: Proceedings of the 17th ACM conference on Information and knowledge management, ACM, 2008, pp. 509–518.

[15] G. Rizzo, R. Troncy, Nerd: a framework for unifying named entity recognition and disambiguation extraction tools, in: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2012, pp. 73–76.

[16] P.N. Mendes, M. Jakob, A. García-Silva, C. Bizer, Dbpedia spotlight: shedding light on the web of documents, in: Proceedings of the 7th International Conference on Semantic Systems, ACM, 2011, pp. 1–8.

[17] A. Sil, A. Yates, Re-ranking for joint named-entity recognition and linking, in: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, ACM, 2013, pp. 2369–2374.

[18] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, The Semantic Web (2007) 722–735.

[19] P. Vossen, F. Ilievski, A. Fokkens, T. Caselli, A. Cybulska, A.-L. Minard, P. Mirza, I. Aldabe, E. Laparra, G. Rigau, Deliverable D5.1.3: Event Narrative Module, version 3, Technical Report, NewsReader Project, 2015.

[20] C. Fellbaum (Ed.), WordNet. An Electronic Lexical Database, in: The MIT Press, 1998.

[21] P. Kingsbury, M. Palmer, From treebank to propbank., LREC, Citeseer, 2002.

[22] M.P. Marcus, M.A. Marcinkiewicz, B. Santorini, Building a large annotated corpus of english: the penn treebank, Comput. Ling. 19 (2) (1993) 313–330.

[23] A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, R. Grishman, The nombank project: An interim report, in: HLT-NAACL 2004 workshop: Frontiers in corpus annotation, 2004, pp. 24–31.

[24] C.F. Baker, C.J. Fillmore, J.B. Lowe, The berkeley framenet project, in: Proceedings of the 17th international conference on Computational linguistics-Volume 1, Association for Computational Linguistics, 1998, pp. 86–90.

[25] K.K. Schuler, VerbNet: A broad-coverage, comprehensive verb lexicon, Ph.D. thesis University of Pennsylvania, 2005.

[26] M. Taulé, M.A. Martí, M. Recasens, Ancora: multilevel annotated corpora for catalan and spanish., LREC, 2008.

[27] I. Niles, A. Pease, Towards a standard upper ontology, in: Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001, ACM, 2001, pp. 2–9.

[28] 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.

[29] 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.

[30] M. Palmer, Semlink: Linking propbank, verbnet and framenet, in: Proceedings of the Generative Lexicon Conference, 2009, pp. 9–15.

[31] A. Gonzalez-Agirre, E. Laparra, G. Rigau, Multilingual central repository version 3.0., in: LREC, 2012, pp. 2525–2529.

[32] J. Alvez, J. Atserias, J. Carrera, S. Climent, E. Laparra, A. Oliver, G. Rigau, Complete and consistent annotation of wordnet using the top concept ontology., LREC, 2008.

[33] L. Bentivogli, P. Forner, B. Magnini, E. Pianta, Revising wordnet domains hierarchy: Semantics, coverage, and balancing, in: Proceedings of COLING 2004 Workshop on Multilingual Linguistic Resources, 2004, pp. 101–108.

[34] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, S. Hellmann, Dbpedia-a crystallization point for the web of data, Web Semant. 7 (3) (2009) 154–165.

[35] 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.

[36] F.M. Suchanek, G. Kasneci, G. Weikum, Yago: a core of semantic knowledge, WWW conference, ACM Press, New York, NY, USA, 2007.

[37] M. van Erp, F. Ilievski, M. Rospocher, P. Vossen, Missing mr. brown and buying an abraham lincoln ? dark entities and dbpedia, in: Proceedings of The NLP & DBpedia Workshop, ISWC, Bethlehem, USA, 2015.

[38] P. Vossen (Ed.), EuroWordNet: A Multilingual Database with Lexical Semantic Networks, in: Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[39] J. Pustejovsky, K. Lee, H. Bunt, L. Romary, ISO-TimeML: an international standard for semantic annotation, in: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010.

[40] 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.

[41] G. van Noord, I. Schuurman, G. Bouma, Lassy Syntactische Annotatie, Technical Report, Technical Report 19455, University of Groningen, 2010.

[42] E. Pianta, C. Girardi, R. Zanoli, The textpro tool suite, in: Proceedings of LREC, 6th edition of the Language Resources and Evaluation Conference, in: LREC-08, Association for Computational Linguistics, Stroudsburg, PA, USA, 2008.

[43] 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.

[44] 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.

[45] A. Björkelund, 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, in: COLING '10, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 33–36. URL http://dl.acm.org/citation.cfm?id=1944284.1944293.

[46] O.D. Clercq, V. Hoste, P. Monachesi, Evaluating automatic cross-domain semantic role annotation, in: Proceedings of the 8th International Conference on Language Resources and Evaluation Conference (LREC-2012), Istanbul, Turkey, 2012, pp. 88–93.

[47] L. Bentivogli, E. Pianta, Exploiting parallel texts in the creation of multilingual semantically annotated resources: The multisemcor corpus, Nat. Lang. Eng. 11 (3) (2005) 247–261, doi:10.1017/S1351324905003839.

[48] A. Cybulska, P. Vossen, Semantic relations between events and their time, locations and participants for event coreference resolution, in: G. Angelova, K. Bontcheva, R. Mitkov (Eds.), Proceedings of Recent Advances in Natural Language Processing (RANLP-2013), INCOMA Ltd., Hissar, Bulgaria, 2013. ISSN 1313–8502

[49] P. Mirza, A.-L. Minard, Hlt-fbk: a complete temporal processing system for qa tempeval, in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado, 2015, pp. 801–805. URL http://www.aclweb.org/anthology/S15-2135

[50] P. Mirza, S. Tonelli, An analysis of causality between events and its relation to temporal information, in: Proceedings of the 25th International Conference on Computational Linguistics (COLING2014), Dublin, Ireland, 2014.

[51] J. Strötgen, J. Zell, M. Gertz, Heideltime: Tuning english and developing spanish resources for tempeval-3, in: Proceedings of the Seventh International Workshop on Semantic Evaluation, SemEval '13, Association for Computational Linguistics, Stroudsburg, PA, USA, 2013, p. 15—19.

[52] B.M. Sundheim, Overview of results of the muc-6 evaluation, in: Proceedings of a Workshop on Held at Vienna, Virginia: May 6–8, 1996, in: TIPSTER '96, Association for Computational Linguistics, Stroudsburg, PA, USA, 1996, pp. 423–442, doi:10.3115/1119018.1119073.

[53] P. Mirza, A.-L. Minard, FBK-HLT-time: a complete Italian Temporal Processing system for EVENTI-EVALITA 2014, in: Proceedings of the Fourth International Workshop EVALITA 2014, 2014.

[54] S. Bethard, A synchronous context free grammar for time normalization, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle, Washington, USA, 2013, pp. 821–826. URL http://www.aclweb.org/anthology/D13-1078.

[55] A. Cybulska, P. Vossen, "bag of events" approach to event coreference resolution. supervised classification of event templates, in: proceedings of the 16th Cicling 2015 (co-located: 1st International Arabic Computational Linguistics Conference), Cairo, Egypt, 2015.

[56] C. Leacock, M. Chodorow, Combining local context and WordNet similarity for word sense identification, WordNet: An Electronic Lexical Database 49 (2) (1998) 265–283.

[57] P. Vossen, F. Bond, J. McCrae, Toward a truly multilingual global wordnet grid, in: Proceedings of the 8th Global Wordnet Conference, 2016.

[58] R. Segers, E. Laparra, M. Rospocher, P. Vossen, G. Rigau, F. Ilievski, The predicate matrix and the event and implied situation ontology: Making more of events, in: Proceedings of GWC2016, 2016 (f.c.).

[59] J. Scheffczyk, A. Pease, M. Ellsworth, Linking framenet to the suggested upper merged ontology, in: Proceedings of FOIS 2006, 2006.

[60] S. Im, J. Pustejovsky, Annotating event implicatures for textual inference tasks, in: The 5th Conference on Generative Approaches to the Lexicon, 2009.

[61] I. Niles, A. Pease, Towards a standard upper ontology, in: Proceedings of FOIS-Volume 2001, ACM, 2001.

[62] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, L. Schneider, WonderWeb Deliverable D17, Technical Report, ISTC-CNR, 2002.

[63] F. Corcoglioniti, M. Rospocher, M. Mostarda, M. Amadori, Processing billions of RDF triples on a single machine using streaming and sorting, in: ACM SAC, 2015a, pp. 368–375.

[64] 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) (2015b) 1–35, doi:10.4018/IJSWIS.2015040101.

[65] T. White, Hadoop: the definitive guide: the definitive guide, "O'Reilly Media, Inc.", 2009.

[66] M. Kattenberg, Z. Beloki, A. Soroa, X. Artola, A. Fokkens, P. Huygen, K. Verstoep, Two architectures for parallel processing of huge amounts of text, in: Proceedings of LREC 2016, 2016.

[67] A.-L. Minard, M. Speranza, R. Urizar, B. Altuna, M. van Erp, A. Schoen, C. van Son, MEANTIME, the NewsReader Multilingual Event and Time Corpus, in: Proceedings of LREC 2016, 2016.

[68] 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.

[69] C.A. Bejan, S. Harabagiu, Unsupervised event coreference resolution with rich linguistic features, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 2010.

[70] M. van Erp, R. Segers, F. Illievsky, A. Fokkens, R. Agerri, M. Rospocher, I. Aldabe, E. Laparra, G. Rigau, D5.2.2 Domain model for financial and economic events, version 2, Technical Report, Newsreader Project, 2015.

[71] A. Björkelund, L. Hafdell, P. Nugues, Multilingual semantic role labeling, in: Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task, in: CoNLL '09, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 43–48. URL http://dl.acm.org/citation.cfm?id=1596409.1596416.

[72] A. Passos, V. Kumar, A. McCallum, Lexicon infused phrase embeddings for named entity resolution, in: Proceedings of the Eighteenth Conference on Computational Natural Language Learning, Association for Computational Linguistics, Ann Arbor, Michigan, 2014, pp. 78–86.

[73] X. Carreras, L. Marquez, L. Padro, Named entity extraction using AdaBoost, in: Proceedings of the 6th conference on Natural language learning-Volume 20, 2002, pp. 1–4.

[74] B. Desmet, V. Hoste, Fine-grained dutch named entity recognition, Language resources and evaluation 48 (2) (2014) 307–343.

[75] R. Zanoli, E. Pianta, EntityPro: exploiting SVM for Italian Named Entity Recognition, in: Intelligenza Artificiale - numero speciale su Strumenti per l'elaborazione del linguaggio naturale per l'italiano, 4, 2007, pp. 69–70.

[76] A. Barrena, A. Soroa, E. Agirre, Combining mention context and hyperlinks from wikipedia for named entity disambiguation, in: Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 101–105. URL http://www.aclweb.org/anthology/S15-1011.

[77] S. Monahan, D. Carpenter, Lorify: A knowledge base from scratch., TAC, NIST, 2012. URL http://dblp.uni-trier.de/db/conf/tac/tac2012.html#MonahanC12.

[78] H. Zhao, W. Chen, C. Kity, G. Zhou, Multilingual dependency learning: A huge feature engineering method to semantic dependency parsing, in: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task, Association for Computational Linguistics, Boulder, Colorado, 2009, pp. 55–60. URL http://www.aclweb.org/anthology/W09-1208.

[79] K. Lee, Y. Artzi, J. Dodge, L. Zettlemoyer, Context-dependent semantic parsing for time expressions, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 1437–1447.

[80] G. Manfredi, J. Strötgen, J. Zell, M. Gertz, HeidelTime at EVENTI: Tunning Italian Resources and Addressing TimeML's Empty Tags, in: Proceedings of the Fourth International Workshop EVALITA 2014, 2014.

[81] B. Yang, C. Cardie, P.I. Frazier, A hierarchical distance-dependent bayesian model for event coreference resolution, CoRR abs/1504.05929 (2015). URL http://arxiv.org/abs/1504.05929.

[82] D.M. Blei, P.I. Frazier, Distance dependent chinese restaurant processes, The Journal of Machine Learning Research 12 (2011) 2461–2488.

[83] S. Pradhan, L. Ramshaw, M. Marcus, M. Palmer, R. Weischedel, N. Xue, Conll-2011 shared task: modeling unrestricted coreference in ontonotes, in: Proceedings of CoNLL 2011: Shared Task, 2011.

[84] M. Vilain, J. Burger, J. Aberdeen, D. Connolly, L. Hirschman, A model theoretic coreference scoring scheme, in: Proceedings of MUC-6, 1995.

[85] A. Bagga, B. Baldwin, Algorithms for scoring coreference chains, in: Proceedings of the International Conference on Language Resources and Evaluation (LREC), 1998.

[86] X. Luo, On coreference resolution performance metrics, in: Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (EMNLP-2005), 2005.

[87] N. UzZaman, H. Llorens, L. Derczynski, M. Verhagen, J. Allen, J. Pustejovsky, Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations, 2013.

[88] P. Vossen, E. Laparra, I. Aldabe, G. Rigau, Interoperability for cross-lingual and cross-document event detection, in: Proceedings of the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation. EVENTS workshop at NAACL-HLT 2015, Denver, Colorado, 2015.

[89] T. Caselli, P. Vossen, M. van Erp, A. Fokkens, F. Ilievski, R.I. Bevia, M. Lê, R. Morante, M. Postma, When it's all piling up: investigating error propagation in an nlp pipeline, NLP Applications: completing the puzzle, Passau, Germany, 2015.