Interoperability of cross-lingual and cross-document event detection

Piek Vossen
VU University Amsterdam
De Boelelaan 1105
Amsterdam, 1081HV, Netherlands
piek.vossen@vu.nl

Egoitz Laparra, Itziar Aldabe and German Rigau
The University of the Basque Country
Donostia-San Sebastián, Spain
egoitz.laparra@ehu.eus
itziar.aldabe@ehu.eus
german.rigau@ehu.eus

Abstract

We describe a system for event extraction across documents and languages. We developed a framework for the interoperable semantic interpretation of mentions of events, participants, locations and time, as well as the relations between them. Furthermore, we use a common RDF model to represent instances of events and normalised entities and dates. We convert multiple mentions of the same event in English, Spanish and Dutch to a single representation. We thus resolve cross-document event and entity coreference within a language but also across languages. We tested our system on a Wikinews corpus of 120 English articles that have been manually translated to Spanish and Dutch. We report on the cross-lingual cross-document event and entity extraction comparing the Spanish and Dutch output with respect to English.

1 Introduction

News reports on events in the world. Applying event extraction to many different news articles provides an interesting perspective on event-coreference, assuming that different sources in different languages report on the same events. These texts may partially provide the same and partly different information on these events. To deal with cross-document event coreference, it is necessary to make a formal difference between the mentions of an event in text and its representation as single event instance. Ideally, we want to be able to match event descriptions within a text, across texts and across languages into a single representation. The fact that different sources provide different information opens new perspectives to study the role of these sources in reporting on what happened in the world. When we consider news written in different languages this perspective becomes more complex but also more interesting.

For such a cross-document and cross-lingual perspective it is essential to define a semantically interoperable approach that can handle the large variation of event expressions within and across languages. In this paper, we report on a system to derive interoperable event representations across documents and across languages. In particular, we focus on English, Spanish and Dutch. Firstly, we developed Natural Language Processing (NLP) pipelines for interpreting mentions of events and event components in text in a uniform way and, secondly, we developed a method to derive instance representations for these interpretations in RDF that is agnostic for the linguistic forms of expression. We report on the evaluation of the systems on a publicly available corpus of English Wikinews articles that has been translated to Spanish and Dutch. We show the capability of our framework and system to perform cross-lingual event extraction from multiple documents, which is, to our knowledge, the first in its kind.

This paper is further structured as follows. In section 2, we describe relevant related work and in section 3, we describe our approach to aggregate event information across different mentions in RDF. In section 4, we explain the interoperability of the NLP pipelines in the three languages. The conversion of the NLP output to RDF is then explained in section 5. Finally, we present the evaluation results in section 6 and we conclude in section 7.

Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT 2015, pages 108–116,
Denver, Colorado, June 4, 2015.

2 Related work

In Unstructured Information Management Architecture (UIMA) (Ferrucci and Lally, 2004) interoperability is provided by platform independent data representations and interfaces. Information is represented in the Common Analysis Structure (CAS). In CAS, annotations are defined as typed objects. For each type, a set of features is defined and an is-a relation with its supertype from which features are inherited. The Subject of Analysis (Sofa) method is used to allow for multiple annotations of the same object. UIMA uses a layered stand-off representation for the annotations of text. A similar approach is followed in the OntoNotes project (Pradhan et al., 2007). In OntoNotes multiple layers of annotation defined in a relational database are combined to arrive at semantic interpretations. Both approaches focus on the generic annotation of text. Howewer, they do not specifically focus on the representation of events and they do not present events according to an RDF model independently of the text as a natural way of cross-document event representation.

The idea of using Linked Data and RDF to represent linguistic annotations for achieving interoperability among linguistic resources has been discussed previously (Chiarcos et al., 2012). Following Linked Data and RDF principles provides a way to address conceptual interoperability among resources, i.e. the ability of heterogeneous NLP resources and tools to talk and understand each other. (Ide et al., 2003) explicitly mention RDF as a possible format to provide semantic coherence in representations. The NLP2RDF initiative collects a number of efforts for representing NLP related information in RDF, including notable efforts such as Ontologies for Linguistic Annotation (OLiA) (Chiarcos, 2008). FRED (Presutti et al., 2012) also produces automatically RDF/OWL and linked data from natural language sentences, but its output is currently limited to English. Still, to our knowledge, there are relatively few implementations of RDF-compatible annotation formats that are actively used or produced by NLP modules. Notable exceptions are the NLP Interchange Format (NIF) (Hellmann et al., 2013), which is tightly linked to OLiA, UIMA Clerezza, and the conversion of GrAF to RDF by (Cassidy, 2010). NIF has the disadvantage that it is not easy to integrate its representations in NLP tools, as shown by user evaluations (Hellmann et al., 2013). Because linguistic annotations are linked to strings it is furthermore not practical for representing hierarchical structures. (Fokkens et al., 2014) presents a more detailed discussion of the formal representations of linguistic annotations.

Besides the formal representation of NLP output, our work relates to the representation of events and cross-document and cross-lingual event coreference. Cross-document event coreference so far has been addressed as a task, in which event markables are related to each other as coreference sets (Bejan and Harabagiu, 2010; Lee et al., 2012). For instance, the ECB corpus represents events and coreference relations using inline annotations in text and cross-document identifiers with offset references. Representation and evaluation of cross-document event-coreference is often done using scorers that use the CONLL-2011 format for expressing coreference (Pradhan et al., 2011). This format also exploits a simple token representation and identifiers. To the best of our knowledge, nobody really addressed the semantic representation of events as instances, exploiting interoperable semantic representations of event instances and entity instances according to Semantic Web practices.

3 The representation of event mentions and instances

Events can be defined as situations in the world in which certain entities participate, where this participation relation is bound in time and place. In text, we make reference to these events in many different ways. Each time we refer to an event in text, the expression through which we make reference can be seen as a mention of the event that we consider to be an instance of a mental representation or real-world event. Typically, there is a coreference relation between different mentions of the same event instance.

In many cases, mentions that refer to events are partial, i.e. not all the details about an event are given within a single sentence. For example, the next sentences from a Wikinews[1] article make reference to a single *flight* but the details are given in

[1] http://en.wikinews.org/wiki/A380_makes_maiden_flight_to_US

different expressions:

> A380 makes maiden flight to US. March 19, 2007. The Airbus A380, the world's largest passenger plane, was set to land in the United States of America on Monday after a test flight. One of the A380s is flying from Frankfurt to Chicago via New York; the airplane will be carrying about 500 people.

The main event is the *test flight* which is mentioned in the title, at the end of the first sentence and referred to again in the second sentence as *flying*. The *carrying* is a subevent of the main event, whereas one could argue whether *landing* is a subevent or a following event. The date of the event is given in the first sentence (*Monday*), which refers to *March 19, 2007*, the flight route is given in the second sentence and passengers are mentioned in the last clause: *500 people* through the implicit relation between *carrying* and *flying*.

Depending if *carrying* and *landing* are different events, we have in this example 5 mentions of 3 unique events. To aggregate the information for the *flight* event, we need to resolve coreference and combine the information from each coreferential mention into a single representation for the instance. To connect the mentions to the instance representation, we use the Grounded Annotation Framework (GAF) (Fokkens et al., 2013). Within GAF, instances are represented according to the Simple Event Model (SEM) (van Hage et al., 2011) using a unique URI and relations to actors, places and time. Furthermore, we use the *gaf:denotedBy* relation to point to the offset mentions of the event in the text. When applied to the above example, the event instance for the *flight* would be represented as follows, where we abstract from the specific roles of the actors and places:

```
:ev17Flight
rdfs:label "maiden flight", "test flight", "flying" ;
gaf:denotedBy
  wikinews:A380_makes_maiden_flight_to_US#char=19,25,
  wikinews:A380_makes_maiden_flight_to_US#char=174,180,
  wikinews:A380_makes_maiden_flight_to_US#char=202,208;
sem:hasTime wikinews:20070319;
sem:hasActor dbp:Airbus_A380, wikinews:500_people;
sem:hasPlace dbp:United_States, dbp:Frankfurt, dbp:Chicago,
 dbp:New_York.
```

Each of the actors, places and points in time is represented as an entity instance as well, with pointers to the mentions in the text. Below, we show the representation of *Airbus* as an example with 2 mentions in the same document:

```
dbp:Airbus
rdfs:label     "Airbus A380", "A380" ;
gaf:denotedBy
wikinews:A380_makes_maiden_flight_to_US#char=415,421,
wikinews:A380_makes_maiden_flight_to_US#char=1132,1138.
```

In this example, the actors, places and time points are aggregated from different mentions in a single representation. If another text is processed, we may detect more mentions of the same event and the same entities. In principle, this will lead to the same instance representation for the event where we only need to extend the *gaf:denotedBy* relations to the new mentions and if it happens aggregate more relations to other entities.

> A380 commercial route proving. 19-28 March 2007. Watch the A380 as it makes its first landings in the United States as part of a 12-day commercial route proving mission in 2007, performed in conjunction with Lufthansa. Follow the aircraft as it flies to New York, Chicago and Washington, D.C., as well as Hong Kong, Frankfurt and Munich.

This message partially overlaps with the previous one but also describes more stops on the route of the airplane. Establishing coreference across the two flights and the A380 results in a single event instance combining the data and pointing to different mentions across the two articles:

```
:ev17Flight
rdfs:label "maiden flight", "test flight", "flying", "flies" ;
gaf:denotedBy
  wikinews:A380_makes_maiden_flight_to_US#char=19,25,
  wikinews:A380_makes_maiden_flight_to_US#char=174,180,
  wikinews:A380_makes_maiden_flight_to_US#char=202,208,
  http://www.airbus.com/newsevents/events/mention#char=242,247;
sem:hasTime wikinews:20070319;
sem:hasActor dbp:Airbus_A380, wikinews:500_people;
sem:hasPlace dbp:United_States, dbp:Frankfurt,
   dbp:Chicago, dbp:New_York, dbp:Washington_D.C.,
   dbp:Hong_Kong; dbp:Munich.
```

Obviously, many similar events are reported which, however, do not refer to the same event instance. Consider the next news item[2] that reports on another *maiden flight* of the Airbus A380 in 2008:

> Qantas A380 arrives in LA after maiden flight. October 21, 2008. The first flight of an Airbus A380 by Qantas touched down in Los Angeles today, inaugurating the Australian carrier's service using the world's biggest commercial jet.

This flight involves a similar participant, *Airbus A380* and a different location: *Los Angeles*. The

[2] http://www.theage.com.au/articles/2008/10/21/1224351190665.html

main and only distinguishing feature is the date that makes it a different event from the previous example. Hence, it will get a different instance representation:

```
:ev18Flight
rdfs:label "maiden flight", "flight";
gaf:denotedBy
  http://www.theage.com.au/articles/2008/10/21/
 1224351190665.html#char=33,46,
  http://www.theage.com.au/articles/2008/10/21/
 1224351190665.html#char=72,78;
sem:hasTime wikinews:20081021;
sem:hasActor dbp:Airbus_A380;
sem:hasPlace dbp:United_States, dbp:Los_Angeles.
```

The above model allows us to combine the information across different mentions within and across documents of the same language. However, the model is also agnostic of the language in which the information is expressed. Likewise, we can use the same model to represent the information from texts in different languages. In order to achieve that, the processing of text across these languages needs to be semantically interoperable. Since we defined events as combinations of actions (or relations and properties), actors, places and time, we also need to achieve an interoperable interpretation of these elements across languages. This will be discussed in the next section.

4 The interoperable interpretation of event and entity mentions across languages

Detecting mentions of events, entities and time expressions in text in several languages requires the combination of various Natural Language Processing modules. Our framework and system obtains interoperable representations of the interpretation of events, the entities that play a role within these events as well as the time expressions associated to the events. The output of the language specific pipelines is represented in the Natural Language Processing Format (NAF) (Fokkens et al., 2014). NAF is a standoff layered format for many different annotations, such as tokens, entities, semantic role (SR) structures and time expressions, where the elements in the layers point to spans of terms. In the next examples, we show in NAF entities, a SR structure with a predicate and several of its roles, and a time expression for an English text. Each of the elements has a span element pointing to term identifiers that mark words and phrases in the text. We see in the first structure that the expression *United States* is detected as a named entity of the type LOCATION and is disambiguated to a DBpedia entry.[3] The SR element consists of a predicate and roles, where the predicate has references to various FrameNet frames (Baker et al., 1998) and WordNet synsets (Fellbaum, 1998) along with the predicate information included in the Predicate Matrix (Lacalle et al., 2014). The roles have a PropBank role (Palmer et al., 2005) and possibly one or more FrameNet elements.[4] Finally, the time expression *Monday* has been normalised by reference to a particular date.

```
<entity id="e3" type="LOCATION">
  <!--United States-->
   <span><target id="t28"/><target id="t29"/></span>
  <externalReferences>
     <externalRef confidence="0.94"
        reference="http://dbpedia.org/resource/United_States
        reftype="en" resource="spotlight_v1"/>
  </externalReferences>
</entity>

<predicate id="pr5"> <!--flying-->
  <externalReferences>
    <externalRef reference="fn:Bringing", "fn:Motion",
    "fn:Operate_vehicle", "fn:Ride_vehicle",
    "fn:Self_motion", "wn:ili-30-01451842-v",
    <externalRef reference="wn:ili-30-01847845-v",
    "wn:ili-30-01840238-v", "wn:ili-30-02140965-v"/>
  </externalReferences>
  <span><target id="t44"/></span>
  <role id="rl14" semRole="A1">
    <!--One of the A380s-->
    <externalReferences>
      <externalRef reference="fn:Bringing@Theme",
      "fn:Motion@Theme", "fn:Operate_vehicle@Vehicle",
      "fn:Ride_vehicle@Theme", "fn:Self_motion@Self_mover"/>
    </externalReferences>
    <span><target head="yes" id="t39"/><target id="t40"/>
              <target id="t41"/><target id="t42"/></span>
   </role>
   <role id="rl15" semRole="AM-DIR"> <!--from Frankfurt-->
    <span><target head="yes" id="t45"/><target id="t46"/></span>
   </role>
   <role id="rl16" semRole="AM-DIR"> <!--to Chicago-->
    <span><target head="yes" id="t47"/><target id="t48"/></span>
    </role>
    <role id="rl17" semRole="AM-MNR">  <!--via New York-->
     <span><target head="yes" id="t49"/><target id="t50"/>
     <target id="t51"/></span>
    </role>
</predicate>

<timex3 id="tmx2" type="DATE" value="2007-03-19">
  <!--Monday-->
  <span><target id="w33"/></span>
</timex3>
```

The English text from the first example above has been translated to Spanish and Dutch. The translations are shown in the next examples:

> El A380 hace su vuelo inaugural a los EEUU. 19 de marzo del 2007. El Airbus A380, el mayor avión de pasajeros del mundo, aterrizó el lunes en los Estados Unidos de América, tras un vuelo de prueba. Uno de los A380s volará de Francfort a Chicago pasando por Nueva York; el avión llevará unas 500 personas.

[3] We show here only the top-ranked DBpedia URI. The software also adds links to alternative DBpedia URIs

[4] We abbreviated the externelRef representation here and in the following examples by combining attribute values and separate them by commas for reasons of space

> Eerste vlucht van A380 naar V.S. 19-Mar-07. De Airbus A380, het grootste passagiersvliegtuig ter wereld, maakte zich maandag op om na een testvlucht te landen in de Verenigde Staten van Amerika . Een van de A380-machines vliegt van Frankfurt naar Chicago via New York en vervoert ongeveer 500 mensen.

Processing the translations through the Spanish and Dutch pipelines results in the following NAF elements, which are interoperable with the English output:

```
<entity id="e2" type="ORGANIZATION"> <!--EEUU-->
<span><target id="t9"/> </span>
<externalReferences>
  <externalRef confidence="0.99"
     reference="http://es.dbpedia.org/resource/Estados_Unidos"
     reftype="es" resource="spotlight_v1">
  <externalRef confidence="0.99"
     reference="http://dbpedia.org/resource/United_States"
     reftype="en" resource="wikipedia-db-esEn"/>
  </externalRef>
</externalReferences>
</entity>


<predicate id="pr3"><!--volará-->
<externalReferences>
 <externalRef reference="fn:Bringing,"fn:Motion",
   "fn:Operate_vehicle", "fn:Ride_vehicle", "fn:Self_motion"/>
 <externalRef reference="wn:ili-30-01451842-v",
   "wn:ili-30-01847845-v", "wn:ili-30-01840238-v",
   "wn:ili-30-02140965-v"/>
</externalReferences>
<span> <target id="t49"/> </span>
<role id="rl8" semRole="arg0"> <!--Uno de los A380s-->
<externalReferences>
 <externalRef reference="fn:Bringing@Agent", "fn:Motion@Theme",
   "fn:Operate_vehicle@Driver", "fn:Ride_vehicle@Theme",
   "fn:Self_motion@Source", "fn:Bringing@Theme",
   "fn:Operate_vehicle@Vehicle", "fn:Self_motion@Self_mover",
   "fn:Ride_vehicle@Vehicle", "fn:Operate_vehicle@Source"/>
</externalReferences>
<span> <target head="yes" id="t45"/> <target id="t46"/>
          <target id="t47"/> <target id="t48"/> </span>
</role>
<role id="rl9" semRole="arg3"> <!--de Francfort-->
<span><target head="yes" id="t50"/><target id="t51"/></span>
</role>
<role id="rl10" semRole="arg4"> <!--a Chicago-->
<span> <target head="yes" id="t52"/><target id="t53"/></span>
</role>
<role id="rl11" semRole="argM"> <!--pasando por Nueva York-->
<span> <target head="yes" id="t54"/>
<target id="t55"/><target id="t56"/><target id="t57"/></span>
</role>
</predicate>


<timex3 id="tx3" type="DATE" value="2007-03-19"> <!--el lunes-->
<span><target id="w30"/><target id="w31"/></span>
</timex3>


<entity id="e2" type="LOCATION">
<!--Verenigde Staten van Amerika-->
<span> <target id="t_29"/><target id="t_30"/>
   <target id="t_31"/><target id="t_32"/></span>
<externalReferences>
  <externalRef confidence="1.0"
    reference="http://nl.dbpedia.org/resource/Verenigde_Staten"
    reftype="nl" resource="spotlight_v1">
 <externalRef confidence="1.0"
    reference="http://dbpedia.org/resource/United_States"
    reftype="en" resource="wikipedia-db-nlEn"/>
 </externalRef>
</externalReferences>
</entity>


<predicate id="pr17"> <!--vliegt-->
<externalReferences>
       <externalRef confidence="0.95" reference="fn:Motion",
        "fn:Ride_vehicle", "fn:Self_motion",
        "fn:Operate_vehicle", "fn:Bringing"/>
        <externalRef reference="wn:ili-30-01451842-v"/>
</externalReferences>
<span><target id="t_38"/></span>
<role id="r26" semRole="AM-DIR"> <!--naar Chicago-->
 <span><target head="yes" id="t_41"/><target id="t_42"/></span>
</role>
<role id="r28" semRole="AM-DIR"> <!--via New York-->
 <span><target head="yes" id="t_43"/> <target id="t_44"/>
  <target id="t_45"/></span>
</role>
<role id="r54" semRole="A3"> <!--van Frankfurt-->
 <span><target head="yes" id="t_39"/><target id="t_40"/> </span>
</role>
<role id="r77" semRole="A0"> <!--Een van de A380-machines-->
<externalReferences>
 <externalRef reference="fn:Motion@Theme",
    "fn:Ride_vehicle@Theme", "fn:Ride_vehicle@Vehicle",
    "fn:Self_motion@Source", "fn:Operate_vehicle@Driver",
    "fn:Bringing@Agent"/>
</externalReferences>
 <span><target head="yes" id="t_34"/><target id="t_35"/>
  <target id="t_36"/><target id="t_37"/></span>
</role>
</predicate>

    <timex3 id="tmx5" type="DATE" value="2007-03-19">
      <!--maandag-->
      <span> <target id="w20"/> </span>
    </timex3>
```

First of all, note that the entity in Spanish and Dutch has been linked to the language-specific DBpedia URI but also to the cross-lingual and equivalent URI in English. In the SR layer, we see that predicates in Spanish and Dutch are matched with FrameNet frames and Wordnet synsets just as for the English SR structure. We can thus derive a similar SR structure across the three languages in the same way as we can map the DBpedia entity referenced to by the named entity expressions.[5] Finally, we can see that the time expressions detected have been normalised in the same way.

A similar output is generated for the all 120 news articles in the Wikinews corpus across different documents and languages.[6] In the next section, we explain how this output is converted into a unified RDF-SEM structure.

5 Event coreference across mentions

The NAF representations explained in the previous sections represent the cross-lingual and cross-document interoperable interpretation of entity, predicate and time mentions in text. In this section, we explain how we convert them to an RDF format using the SEM/GAF model. As explained in section

[5]Note that the English and Spanish predicate is aligned to multiple synsets, whereas the Dutch predicate only got a single synset assigned. This difference is the result of the different ways in which the SR modules have been implemented.

[6]On-line demos of the pipelines are available at `http://www.newsreader-project.eu/results/demos/`

3, different mentions of the same instance are represented only once. For entities and time expressions this is automatically achieved by the normalisation to DBpedia URIs and dates. When converting each mention of an entity or time expression to RDF, we create an URI on the basis of its normalised value. Within the RDF model, these data structures are automatically merged and the references to the mentions are combined, both for cross-document references and the cross-language references.

Obviously, for events this is more difficult. We follow an approach that takes the compositionally of events as a starting point (Quine, 1985). The compositionality principle dictates that events are not just defined by the action but also by the time, place and participants. For that, we use an algorithm that compares events for all these properties (Cybulska and Vossen, 2015). Currently, we compare first the events on the basis of the lemma of the predicates, the FrameNet frames and the Wordnet synsets.[7] From a cross-lingual perspective, it only makes sense to compare events according to language-neutral classes in FrameNet and WordNet.

The second important element is the time-reference. We relate all event mentions to time-expressions in the text, where we first consider the references in the same sentence, next the surrounding sentences (2 before, 1 after the current one) and finally the publication date of the news article. We then only compare events anchored to the same temporal reference.

Finally, note that the entity layer and the role layer are only indirectly aligned through their span references. Since the layers are generated by different software modules, we need to determine the expression in a role that is attributed to an entity in the entity layer. We match the output of the layers by intersecting the spans by calculating the Dice coefficient of the content words in each entity mention with the role mention. If the overlap is more than 75%, we assign the role to the entity. To be able to represent matching events through a shared URI across languages, we create an artificial URI from the set of WordNet synsets that were associated with the predicates in the SR from which they are derived. If predicates from different languages have been matched with intersecting synsets, we consider the actions to be similar. Note that this can be loosened to other similarity measures. In addition to event similarity, time and participants need to match in the same way as described for the cross-document case described before.

Below, we show the result of applying our cross-lingual and cross-document event extraction module to the Airbus A380 article in the three languages. Our current program creates two *flying* events from the first sentences. The first event is represented by a series of five WordNet synsets all related to *flying*. We see a series of RDF subclass relations for this event to various FrameNet frames. We also see labels in Spanish *volar*, English *fly* and Dutch: *verlopen* and *vliegen*.[8] Next, we see mentions from all three language texts and finally the aggregated relations. Some of these are detected as places and some as actors. Furthermore, we see some entities not matched to DBpedia for various reasons, such as *Chicago_via_New_York* and *Los_Angeles_LAX* coming from the Dutch processing. We can also observe that some places are detected as actors in the SR, with roles such as A3 or A4 instead of AM-DIR or AM-LOC. The same event was detected across the three languages and the relations have been merged in a single representation. The basis for the final merging is the fact that the events share WordNet references, all of them bound to the same point in time and also share at least one actor and place.

```
wn:ili-30-01451842-v;ili-30-01847845-v;ili-30-01840238-v;
ili-30-02140965-v;ili-30-01941093-v
   a  sem:Event, fn:Bringing, fn:Motion, fn:Operate_vehicle,
       fn:Ride_vehicle, fn:Self_motion;
   rdfs:label "volar", "fly", "verlopen", "vliegen" ;
   gaf:denotedBy
          wikinews:english_mention#char=202,208>,
          wikinews:english_mention##char=577,580>,
          wikinews:dutch_mention##char=1034,1042>,
          wikinews:dutch_mention#char=643,650>,
          wikinews:dutch_mention#char=499,505>,
          wikinews:dutch_mention#char=224,230>,
          wikinews:spanish_mention#char=218,224>,
          wikinews:spanish_mention#char=577,583> ;
   sem:hasTime nwrtime:20070391;
   sem:hasPlace
          dbp:Frankfurt_Airport, dbp:Chicago ,
          dbp:Los_Angeles_International_Airport,
          nwr:airbus/entities/Chicago_via_New_York;
   sem:hasActor
          dbp:Airbus_A380, nwr:airbus/entities/Los_Angeles_LAX ,
          dbp:Frankfurt,  nwr:airbus/entities/A380-machines.
```

The English pipeline generated an additional *flying* event that was not matched. Although there is a

[7] In fact, the Predicate Matrix provides many other mappings that could be used, such as PropBank, NomBank, VerbNet or SUMO

[8] *verlopen* is the result of an error by the word-sense disambiguation

match for the WordNet references and the time anchoring is the same, none of the actors and places match with the previous event.

```
wn:ili-30-01451842-v;ili-30-01847845-v;ili-30-01840238-v;
ili-30-02140965-v
 a sem:Event, fn:Bringing, fn:Motion, fn:Operate_vehicle,
    fn:Ride_vehicle, fn:Self_motion;
 rdfs:label "flight" ;
 gaf:denotedBy
          wikinews:english_mention##char=19,25,
           wikinews:english_mention##char=174,180,
           wikinews:english_mention##char=566,572;
 sem:hasTime nwrtime:20070391;
 sem:hasActor dbp:United_States_dollar, dbp:Qantas .
```

The complete system for processing text in English, Spanish and Dutch, as well the cross-document and cross-language coreference are available under an open source license and accessible through GitHub.[9] In the next section, we provide an initial evaluation of the cross-lingual processing.

6 Evaluation on the cross-lingual Wiki news corpus

We created an evaluation corpus from English Wikinews articles. We selected four different topics *Airbus*, *Apple*, *GM-Chrysler-Ford* and the *stock market*. For each topic, we selected 30 articles spread over a period of five years. The English corpus was manually annotated for various layers, including entities, events, time-expressions, event relations, and coreference relations. We translated the corpora also to Spanish and Dutch, where the sentences have been aligned.[10] The cross-lingual corpora allow for two types of evaluation: 1) we can evaluate the quality of the NLP modules in each language on each corpus, 2) we can apply the RDF-SEM extraction to the NAF output of each corpus independently and compare these structures. Currently, we report on the second evaluation. In the near future, we also plan to evaluate against the annotations in each language and across languages.

Since the corpora are manually translated, we expect that the same content is expressed in the three languages. Thus, if our cross-lingual NLP processing is fully interoperable and generates the same quality across the languages, we expect to obtain exactly the same events across the different languages. As such the translated corpus provides an excellent benchmark dataset for evaluating event extraction across languages. For the evaluation, we applied the pipelines for English, Spanish and Dutch to all 120 articles in each language. Next, we extracted the RDF representations from the NAF files in each topic. Since the final RDF representation is agnostic with respect to its textual realisation in the different languages, we can directly compare the extracted representations. In Table 1, we show the results averaged over the four different topics, where we compare the output from the Spanish and Dutch systems to the English output as a reference.[11]

	English		Spanish				Dutch			
	I	M	I	M	O	C	I	M	O	C
entities	318	4101	204	2209	1469	34.88	187	1313	1030	24.46
events	590	2402	323	1036	610	26.04	651	1545	281	11.88
triples	665	866	220	276	60	7.07	619	689	25	2.93

Table 1: Cross-lingual coverage of Spanish and Dutch RDF data compared to English.

Table 1 provides figures for the DBpedia entities, the events represented as WordNet synsets and triples where entities are related to the events either as actors or as places. For each language, we present the number of instances (*I*, unique URIs in the data) and the number of mentions (*M*) in triples. For Spanish and Dutch, we provide the overlap (*O*) and the micro-averaged coverage (*C*) of the English mentions. For entities, we can see that the mentions detected for Spanish is 34.88% of the English ones, while for Dutch this is 24.46%. We also see that detecting events and triples (which are combinations of events, entities and a SEM relation) is more difficult. Spanish coverage of the English events is 26.04% for events and only 7.07% coverage for full triples. In general, the Dutch system is performing less compared to Spanish. Obvious explanations for this behaviour are the different performance of the Spanish and Dutch pipelines, and the different coverage of the resources (both DBpedias and wordnets). As expected, the drop for the events and triples is bigger compared to entities. Detecting events correctly is more complex and challenging than disambiguating DBpedia references. Also recall that the comparison of events and triples is based on WordNet equivalences.

[9] http://github.com/newsreader

[10] Currently, the translated corpora is being annotated according to similar guidelines and cross-document coreference relations are added.

[11] Note that output that does not match with English is not necessarily incorrect. We are only measuring the coverage of one language with respect to the English data.

	English	Spanish	Dutch
Boeing	156	183	98
Airbus	107	81	37
European_Union	83	17	29
Indonesia	57	13	0
France	56	1	1
Boeing_Commercial_Airplanes	50	0	3
United_States_dollar	39	0	2
Government_Accountability_Office	36	0	0
Aer_Lingus	33	16	9
United_States_Air_Force	32	21	7
Boeing_747	30	0	0
Singapore	25	0	7
Airbus_A320_family	22	3	3
Toulouse	17	0	1
Northrop_Grumman	15	1	0
United_States_Armed_Forces	15	0	0
United_Kingdom	14	2	0
EADS	13	3	0
Sydney_Airport	12	0	0
United_States	11	7	23

Table 2: Entities most frequent in English data

We also inspected the results for the Airbus corpus by looking at the entities and events that are most frequently mentioned in the English output. Table 2 shows the top-frequent entities with the corresponding counts for Spanish and Dutch. We inspected obvious entities such as *Indonesia*, *France* and *Toulouse*. It turns out that the NLP modules did detect these entities: *Indonesia* (8 Spanish and 9 Dutch mentions), *France* (13 Spanish and 11 Dutch mentions), *Toulouse* (6 Spanish and 4 Dutch mentions) but that they were not linked to events and therefore not represented. This points to a difference and probably lower coverage of the semantic role module to connect entities to events in a uniform way. Another case is represented by *United_States*, *United_States_Air_Force(s)*, and *United_States_dollar*. The latter have high frequencies in English but none in Spanish and Dutch, while the former has even higher frequencies in Dutch. In this case, the English system makes a systematic mistake by not always resolving expressions such as *US* to the right URI but to the dollar, while the other systems do not make this mistake because their expressions are very distinct: *Estados Unidos de América* and *de Verenigde Staten van Amerika*. A final type of difference is illustrated by *Boeing* versus *Boeing_747*. Where the English module tends to prefer more specific entities, the other fall back to the more generic ones. Such metonymic mismatches are less of a problem.

Regarding events, the Dutch events are often linked to other meanings in WordNet that may also apply (e.g. *fly* 72 English, 34 Spanish and 8 Dutch but also *buy* 10 English, 17 Spanish and 16 Dutch). Furthermore whereas the English and Spanish module often provide more than one synset, the Dutch system only gives one, lowering the chances to intersect. In a future version, sets of closely related synsets will be generated for Dutch as well to solve the fine-grained sense matching problem. Another option is to fall back on more general event classes in the PredicateMatrix (e.g. VerbNet, FrameNet, etc.). Most of the other differences relate to small differences acorss systems and poor coverage of semantic resources in Spanish and Dutch.

7 Conclusions

We described a system for the cross-document and cross-lingual event and entity extraction that is unique in its kind. We use GAF to make a clear distinction between mentions and instances, where mentions of events and entities are interpreted according to an interoperable RDF framework that uses URIs, WordNet and FrameNet concepts, normalised time expressions and normalised relations between entities and events. We developed NLP pipelines in English, Spanish and Dutch that process text according to the shared framework. In addition, we developed software to convert the output of the NLP modules to the RDF representation of instances. We showed that we can represent the accumulated information from different articles and even across languages. We described the first evaluation results for our system.

The current system leaves room for improvement. The matching of entities across mentions and languages can be harmonised and the matching of events through WordNet concepts is not precise enough. In many cases, the background resources (DBpedia in different languages and wordnets in different languages) lack the proper mapping. Finally, the quality of the SR module needs to be improved to capture more expressions and harmonise the interpretations of these expressions. Nevertheless, the current work forms an excellent basis to flesh out these problems without the need to change the fundamental cross-lingual architecture. When the translated corpora are fully annotated, we will be able to further benchmark the NLP processing in the different languages and compare the results in terms of precision and recall independently of English.

Acknowledgements

We thank the anonymous reviewers for the feedback. This work was supported by the European Union's 7th Framework Programme via the NewsReader (ICT-316404) project and by the SKATER Spanish project (TIN2012-38584-C06-02) with European Regional Development Fund support.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING-ACL'98*, pages 86–90, Montreal, Canada.

Cosmin Adrian Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the ACL'2010*, Uppsala, Sweden.

Steve Cassidy. 2010. An RDF realisation of LAF in the DADA annotation server. In *Proceedings of ISA-5*, Hong Kong.

Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann. 2012. *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*. Springer, Heidelberg.

Christian Chiarcos. 2008. An ontology of linguistic annotations. *LDV Forum*, 23(1):1–16.

Agata Cybulska and Piek Vossen. 2015. "Bag of Events" Approach to Event Coreference Resolution. Supervised Classification of Event Templates. In *International Journal of Computational Linguistics and Applications (IJCLA)*. (to appear).

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

David Ferrucci and Adam Lally. 2004. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

Antske Fokkens, Marieke van Erp, Piek Vossen, Sara Tonelli, Willem Robert van Hage, Luciano Serafini, Rachele Sprugnoli, and Jesper Hoeksema. 2013. GAF: A grounded annotation framework for events. In *Proceedings of the first Workshop on Events: Definition, Dectection, Coreference and Representation*, Atlanta, USA.

Antske Fokkens, Aitor Soroa, Zuhaitz Beloki, Niels Ockeloen, German Rigau, Willem Robert van Hage, and Piek Vossen. 2014. NAF and GAF: Linking linguistic annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 9–16.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating NLP using Linked Data. In *Proceedings of the ISWC'2013*.

Nancy Ide, Laurent Romary, and Eric Villemonte de La Clergerie. 2003. International standard for a linguistic annotation framework. In *Proceedings of the HLT-NAACL 2003 Workshop on Software Engineering and Architecture of Language Technology Systems (SEALTS)*. Association for Computational Linguistics.

Maddalen Lopez De Lacalle, Egoitz Laparra, and German Rigau. 2014. Predicate matrix: extending semlink through wordnet mappings. In *Proceedings of LREC'2014*, pages 26–31, Reykjavik, Iceland.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *Proceedings of EMNLP-CoNLL'2012*.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31(1).

Sameer Pradhan, Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. Ontonotes: A unified relational semantic representation. *International Journal of Semantic Computing*, 1(4):405–419.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of CoNLL'2011*, pages 1–27.

Valentina Presutti, Francesco Draicchio, and Aldo Gangemi. 2012. Knowledge extraction based on discourse representation theory and linguistic frames. In *Knowledge Engineering and Knowledge Management*, pages 114–129. Springer.

Willard V. Quine. 1985. Events and reification. In *Actions and Events: Perspectives on the Philosophy of Davidson*, pages 162–171. Blackwell.

Willem Robert van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. 2011. Design and use of the Simple Event Model (SEM). *Journal of Web Semantics*, 9(2):128–136.