D5.1.1 Event Narrative Module, version 1

Deliverable D5.1.1

Version FINAL

Authors: Piek Vossen[1], Agata Cybulska[1], Egoitz Laparra[2], Oier Lopez de Lacalle[3], Eneko Agirre[2], German Rigau[2]

Affiliation: (1) VUA, (2) UPV/EHU, (3) IkerBasque

Building structured event indexes of large volumes of financial and economic data for decision making

ICT 316404

Grant Agreement No.	316404
Project Acronym	NEWSREADER
Project Full Title	Building structured event indexes of large volumes of financial and economic data for decision making.
Funding Scheme	FP7-ICT-2011-8
Project Website	http://www.newsreader-project.eu/
Project Coordinator	Prof. dr. Piek T.J.M. Vossen VU University Amsterdam Tel. + 31 (0) 20 5986466 Fax. + 31 (0) 20 5986500 Email: piek.vossen@vu.nl
Document Number	Deliverable D5.1.1
Status & Version	FINAL
Contractual Date of Delivery	December 2013
Actual Date of Delivery	January 30, 2014
Type	Report
Security (distribution level)	Public
Number of Pages	54
WP Contributing to the Deliverable	WP05
WP Responsible	VUA
EC Project Officer	Susan Fraser
Authors: Piek Vossen[1], Agata Cybulska[1], Egoitz Laparra[2], Oier Lopez de Lacalle[3], Eneko Agirre[2], German Rigau[2]	
Keywords: Event detection, event-coreference, reasoning, event components	
Abstract: This deliverable described the first results on modeling events. It extracts instance- of events and entities in a formal semantic representation from textual descriptions, according to the Grounded-Annotation-Framework developed in the project. Every instance of an event and entity and every relation receives a unique identifier and is linked to all the place in texts where they are mentioned. Coreference is the first important step to get from a presentation of mentions in text to a semantic representation of instances. The prototype clusters co-referencing event mentions, within and across documents, and outputs a unique list of event instances, merging information from different mentions. The system has been applied to two data sets: 63,811 English news articles provided by Lexis Nexis, on the car industry and published between 2003 and 2013 and 43,384 articles from the TechCrunch database with news about IT companies registered in Crunchbase. We also describe the preliminary ideas on deciding on the relevance and significance of the event data that is extracted.	

Table of Revisions

Version	Date	Description and reason	By	Affected sections
0.1	Nov 2013	Creation of document with structure		All
0.2	Dec 2013	Section on experiments on Bayesian model	Egoitz Laparra, German Rigau, Oier Lopez de Lacalle, Eneko Agirre	EHU
0.3	19 December 2013	Major revision and first draft version of the complete deliverable	Piek Vossen	VUA
0.4	26 December 2013	Overall editing and conclusion section	Eneko Agirre	EHU
0.5	6 January 2014	Major revision, included more statistics	Piek Vossen	VUA
0.5	8 January 2014	Internal review	Sara Tonelli	FBK
0.6	10 January 2014	Final editing and revision	Piek Vossen	VUA
0.6	29 January 2014	Comments and feedback	Agata Cybulska	VUA
2.0	30 January 2014	approval by project manager	Piek Vossen	VUA

Executive Summary

This deliverable describes the first cycle of T05.1 Event merging and chaining and T05.2 Event significance and relevance (21PM of effort, started on month 6 of the project). The prototype clusters co-referencing (identity) event mentions, within and across documents, and outputs a unique list of event instances, merging information from different mentions. We implemented different approaches: a baseline system using the lemmas or words only, a system using topic-clustering and machine learning from a large set of textual properties and a semantic approach that reasons over event components. The baseline system has been applied to two data sets and the result was imported into the Knowledge Store. The prototype also produces relevance ranking and selection of event instances.

Contents

List of Tables

1 Introduction

The goal of the NewsReader project[1] is to automatically process massive streams of daily news in 4 different languages to reconstruct longer term story lines of events. For this purpose, we extract events mentioned in news articles, the place and date of their occurrence and who is involved. At first, this processing is document based and the results are stored in the Natural Language Processing format (NAF, Beloki *et al.* (2014)) that was developed in the project. For each text file with news, we generate a corresponding NAF file that contains the events, the participants and the indications of the time and place. The software modules for this processing are described in the NewsReader deliverable D4.2.1 Event Detection, version 1 (Agerri *et al.* (2013)). The analysis of the news articles in Work Package 4 is done at the so-called mention level. This means that each description of an event in text is interpreted as a different event. No decision has been taken whether different events describe the same event. For example, the following fragments show 5 references to the same **decision** from two news articles in 2004:

- New Zealand Herald, Monday Apr 26, 2004:[2]
  - Schrempp may have suffered his own personal Waterloo on Friday when Daimler's board **voted** to pull the plug on troubled Japanese carmaker Mitsubishi Motors rather than pump in billions of euros to keep the company on financial life support.
  - The **decision** effectively kills Schrempp's dream of creating a global automotive giant by severing its Asian platform.
  - The Daimler CEO was conspicuously absent from a conference call to explain the **decision** to journalists.
- Automotive News, Monday Apr 26, 2004:[3]
  - The **decision** not to bail out Mitsubishi Motors Corp. raises fresh doubts about the future of DaimlerChrysler CEO Juergen Schrempp.
  - Warburton added: "It might have been easier to put further money into Mitsubishi, but yesterday's **decision** will strengthen Schrempp's position in the long run."

The first sentence introduces the **vote** event done by the Daimler's board and the next two sentences refer to this event as the **decision**, while providing more information on the implications. The fourth and fifth example come from another source referring

[1] FP7-ICT-316404 *Building structured event indexes of large volumes of financial and economic data for decision making*, www.newsreader-project.eu/

[2] http://www.nzherald.co.nz/business/news/article.cfm?c_id=3&objectid=3562563

[3] http://www.autonews.com/article/20040426/SUB/404260773/0/SEARCH#axzz2pKDsq4mB

to the same event also using the expression **decision**.[4] The event detection modules of Work Package 4 will extract each of these events separately and include participants and time/place references for each. It will create a semantic interpretation but not consider sameness.

Generalization over different **mentions** of the same event, and also from their participants, place and time results in a single representation of an **instance** with links to the mentions in the news. This is explained in the Grounded Annotation Framework (GAF, Fokkens *et al.* (2013)), which formally distinguishes between mentions of events and entities in NAF and instances of events and entities in the Simple Event Model (SEM, van Hage *et al.* (2011)) connected through **denotedBy** links to connect both representations. Work Package 5 of the NewsReader project deals with this next step in processing news by mapping mentions across NAF representations and representing them as instances in SEM. The main task for achieving this is called coreference. Coreference can be applied to entities and to events and it can involve mentions within the same document (intra-document coreference) and across documents (inter-document coreference). After determining coreference relations across mentions, we can **aggregate** the information from all the mentions and combine this at the instance level. These relations not only reflect participants, place and time relations between entities and events, e.g. the fact that the entity instance **Daimler's board** is a participant in the **decision** event, but also temporal and causal relations across different event instances, e.g. that **suffering by Schrempp** is the (possible) result of the **decision**. The modules developed in Work Package 5 take the output of Work Package 4 as the input. The final output (NAF+SEM) is stored in the Knowledge Store (Rospocher *et al.* (2013)) that is developed in Work Package 6 of NewsReader.

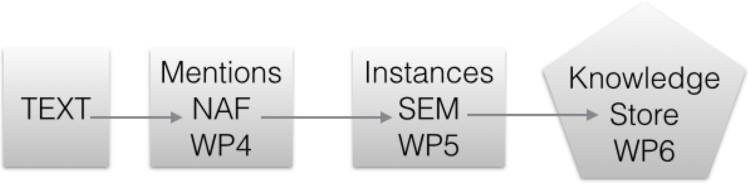


Figure 1: Input-output schema for Work Packages in NewsReader

This deliverable describes the first cycle of tasks T05.1 Event merging and chaining and T05.2 Event significance and relevance (21PM of effort, started on month 6 of the project). This first baseline prototype groups co-referencing event mentions, within and across documents, and outputs a unique list of event instances with URIs, merging information from different mentions. The prototype also produces a first relevance ranking and selection of event instances, aggregating the information produced in WP4 per mention.

[4] Note that the phrases **suffered his own personal Waterloo**, **raises fresh doubts about the future** and **Schrempp's position in the long run** refer to the same future but describe different implications of this decision and thus different versions of the future, which is far more difficult to determine

In the next section 2, we motivate our main strategy for the project based on state-of-the-art findings on event coreference. Our approach starts from the observation that variation and ambiguity of reference to events is highly constrained by the source, the place and time of publication. When considering event descriptions within the same source and/or with reference to the same place and date, a baseline that considers the lemma describing the event will already achieve high precision and reasonable recall. This lemma-based approach is described in more detail in section 3. In section 4, we describe two approaches to widen the recall of event-coreference considering other sources and wider scope of time. The first approach experiments with semantic similarity in combination with overlap of event components in the SEM representation of instances. The second experiment is a re-implementation of the Bejan and Harabagiu (2010) algorithm that can be applied to the NewsReader data. Section 5 describes the our first specifications for measuring relevance and significance of the event data. In section 6, we come to some conclusions and we look at the goals for the second year of the project.

2 Overall approach

Coreference is the first important step to get from a presentation of mentions in text to a semantic representation of instances.[5] Once coreference has been established, we can decide on the relations between events and the (re-)construction of longer story lines of events. Deciding on event relations and story lines is planned for the second year of the project. This deliverable reports on the work done for establishing co-reference relations.

The overall approach for creating an instance layer is based on a number of assumptions and findings. First of all, time and place are strict constraints for identifying events. Events can only exist within the same boundaries of time and place. The exact same action that repeats itself at the same place involving the same participants is still a different event instance if it takes place at different points of time, e.g. **John teaching mathematics at the University every Monday at 3:00pm** represents a series of different events although similar in the type of activity. This being said, events can stretch over a longer period of time and different events can (partially) overlap in time. Whether or not we are dealing with the same event or different events can therefore still be difficult to decide. Roughly, there are two approaches to event coreference:

1. description-based approaches that compare the wording and structure of each mention.

2. semantic-based approaches that compare the semantic components of the event instances.

Description-based approaches work very well for intra-document coreference. Throughout a single document, less variation is expected in the way the same event is mentioned and if so, variation is often linguistically marked as is the case of anaphoric references. In case of the inter-document coreference, especially when considering documents from a large variety of sources, events can be described in very different ways. A structural comparison is expected to be less successful since the styles and ways of describing are numerous and large volumes of training data are required to capture the variation. Another problem is that exactly the same or similar structural descriptions can still refer to very different events, e.g. **a car bombing in Madrid** and **a car bombing in Spain** have a similar structure but the first took place in 1995 and the second in 2009. Since place and time information is often not expressed in the same sentence or direct context of the event description, description-based approaches tend to assign a co-reference relation to such descriptions across document. Critical information, such as time and place, can often only be derived through semantic approaches that gather all the critical information at the

[5] Mentions are expressions in text that can refer to instances of events and entities. *Barack Obama* and *the president of the US* are two expressions that mention the same instance of an entity. *9/11* and *the attack on the World Trade center* are two expressions that mention the same instance of an event. In news articles, we typically find many references to the same instances of events and entities. These references are called mentions

instance level, possibly from many different mentions within the same document and use this to compare mentions across documents.

The state-of-the-art approach to cross-document event-coreference using descriptional properties is described by Bejan and Harabagiu (2010). They use topic clustering and machine-learning on a large variety of features and evaluate the results on the *EventCorefBank* (ECB)[6], which is a corpus with news articles annotated for events. The ECB contains 43 topics, 1744 event mentions, 1302 within-document events, and 339 cross-document events. A semantic-approach evaluated on the ECB corpus is described by Lee *et al.* (2012). The best performing system of Bejan and Harabagiu (2010) reports F-measures above 80% but Cybulska and Vossen (2013) show that a lemma-baseline (matching events within a topic solely on the basis of the same lemma) scores only 10% lower in F-measure and can easily be improved using simple heuristics for anaphora resolution and syntactic relations. Further studies on the ECB corpus from Cybulska and Vossen (Cybulska and Vossen (2014)) show that there is hardly any ambiguity across lemma mentions in the corpus as a a whole, let alone within a single topic, e.g. there is only one parliamentary election described in the whole corpus. Likewise, matching all occurrences of the lemma election to the same events gives extreme high precision and only a small effort is required to improve the recall. Within NewsReader, we expect that event-coreference is more complex when dealing with news over longer stretches of time and involving massive articles. We are therefore extending the ECB corpus with more events of the same type but referring to different instances to increase the ambiguity for lemma-based references. This type of complexity is more representative of the massive news streams that need to be analyzed in NewsReader (see Cybulska and Vossen (2014) for details).

Based on these findings, we defined a multi-stage approach for establishing event-coreference that is further described in this deliverable:

1. Stage 1: structural approach for intra-document mentions

2. Stage 2: structural approach for inter-document mentions within a tight temporal and topic cluster

3. Stage 3: A semantic approach for inter-document instances for more loose clusters of documents and across longer periods of time

The first and second stage start from the assumption that references within the same source and within tight temporal and topical clusters tend to use the same wordings to refer to the same event. Within these settings, we expect little ambiguity and little variation. The more we include sources over larger stretches of time and/or involve more places, the more powerful methods we need to establish valid coreference relations across event descriptions.

In Stage 1, we only create event co-reference representations within NAF for single news articles, i.e. across intra-document mentions. These results will have a relatively high

[6] http://www.hlt.utdallas.edu/∼ady

precision and recall. In Stage 2, we consider the NAF representations of sets of documents (inter-document mentions) that belong to the same time-span and topic-cluster. Currently, we used the publication date as the shared time-span but in the near future, we will use normalized timex expressions and topic-classification to define more fine-grained clusters. In this stage, we can combine data from the different mentions in each co-reference set to make a comparison. This results in larger coreference sets across documents that share the same time span, region and topic cluster. We expect that the ambiguity for similar event references remains limited within these tight clusters while the variation of mentions in the initial sets can be used to deal with variation across documents. At this level, we create a first representation of instances of events and participants in SEM with pointers to all sources with the mentions of these events. In Stage 3, which is planned for the second year of the project, we reason over these SEM representations to establish wider co-reference relations over longer time-spans. In this case, we either widen the matching of the participants within strict event matches or we widen the event references on the basis of strict participant matches. In any case, time and place information needs to be compatible as far as this information is available.

The complete approach is shown in figure 2. The news on a single day is first clustered for topic and within each topic for time and place, where the publication date is the ultimate fallback option to date events in case there is no other information on the time. Within a single source or news article, we can safely map events on the basis of the form of the mention in the majority of cases. Across sources but within the time, place an topic constraint, we should allow more loose mappings across events. The results of a single day form a graph of related event instances with pointers to various mentions. Eventually, we need to map these events graphs to the events stored in the KnowledgeStore that were processed in the past. These can be events that took place in the past or were speculated on for the future. This mapping is what we call **historical event-coreference**, since it is not just across sources but across temporal boundaries and historical (subjective) perspective.

This deliverable describes the first modules that have been developed for this approach: Stage 1 and 2. We developed a lemma-based intra-document approach followed by a cross-document coreference module that have been applied to two data sets:

- 63,811 English news articles provided by Lexis Nexis, on the car industry and published between 2003 and 2013
- 43,384 articles from the TechCrunch database with news about IT companies registered in Crunchbase

This processing resulted in a SEM representation for events, participants and their time points and place. The data structure has been imported in the Knowledge Store developed in Work Package 6. The lemma-based approach, described in 3, can be seen as a strong baseline system. In section 4, we describe two approaches to widen the recall of event-coreference. The first approach experiments with semantic similarity in combination with overlap of event components in the SEM representation of instances. The second

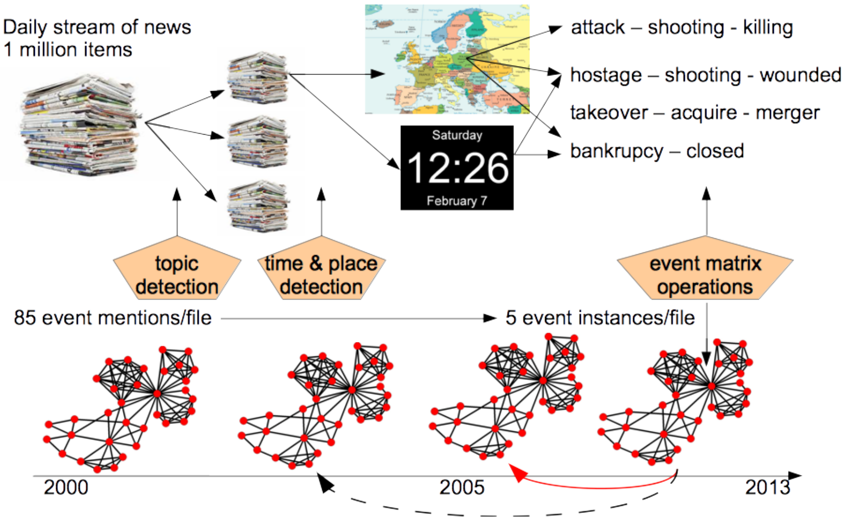


Figure 2: Historical event-coreference, relating topical event-instance of a single day to the past

experiment is a re-implementation of the Bejan and Harabagiu (2010) that can be applied to the NewsReader data.

3 Lemma-match baseline

3.1 Introduction

The lemma-baseline only considers lemma-matches for coreference relations between mentions of events. As explained in Cybulska and Vossen (2013), this gives very good results for intra-document-within -topic coreference in the ECB corpus: precision ranging from 83% till 91% and F-scores between 65% and 75%. In the next two sections we describe the first version of a baseline system that first creates event-coreference sets for each single NAF file of a news article and, secondly, takes a cluster of these NAF files to create inter-document coreference relations. The second step produces SEM as an output structure, which can directly be imported in the Knowledge Store.

3.2 Intra-document event coreference

The input for the intra-document event coreference module is the Semantic Role Layer (SRL) layer in NAF (see Deliverable 4.2.1 Agerri *et al.* (2013)), which specifies mentions of predicates (nominal, verbal and adjectival) in connection to arguments that have been detected within the same sentence. In the next (shortened) example, you see for 4 predicates involving the lemma "leave" that have been extracted with their roles according to a Propbank (Palmer *et al.* (2005)) classification from a single news article in the car industry data set (document id = 2004/4/26/4C7V-T4D0-0015-K19Y.xml):[7]

```
 <public publicId="4C7V-T4D0-0015-K19Y" uri="2004/4/26/4C7V-T4D0-0015-K19Y.xml"/>
<srl>
      <predicate id="pr4"> <!--left-->
      <externalReferences>
        <externalRef reference="leave.01" resource="PropBank"/>
        <externalRef reference="leave-51.2" resource="VerbNet"/>
        <externalRef reference="resign-10.11" resource="VerbNet"/>
        <externalRef reference="escape-51.1" resource="VerbNet"/>
        <externalRef reference="leave-51.2-1" resource="VerbNet"/>
        <externalRef reference="escape-51.1-1" resource="VerbNet"/>
        <externalRef reference="Departing" resource="FrameNet"/>
        <externalRef reference="Path_shape" resource="FrameNet"/>
        <externalRef reference="contextual" resource="EventType"/>
      </externalReferences>
      <span><target id="t23"/></span>
      <role id="rl11" semRole="R-AM-TMP"> <!--when-->
        <span><target head="yes" id="t21"/> </span>
      </role>
      <role id="rl12" semRole="A0"> <!--he-->
        <externalReferences>
          <externalRef reference="leave-51.2#Theme" resource="VerbNet"/>
          <externalRef reference="resign-10.11#Agent" resource="VerbNet"/>
          <externalRef reference="escape-51.1#Theme" resource="VerbNet"/>
          <externalRef reference="Departing#Theme" resource="FrameNet"/>
          <externalRef reference="Path_shape#Road" resource="FrameNet"/>
        </externalReferences>
```

[7] External reference links for predicates and their role elements provide first semantic typing of the elements. This typing is not used for the lemma-based approach but can be used in future extensions of the module to use semantic similarity.

```
      <span> <target head="yes" id="t22"/></span>
    </role>
    <role id="rl13" semRole="A1"><!--the top job in Hyundai 's Eastern sales region-->
      <externalReferences>
        <externalRef reference="leave-51.2#Initial_Location" resource="VerbNet"/>
        <externalRef reference="resign-10.11#Source" resource="VerbNet"/>
        <externalRef reference="escape-51.1#Initial_Location" resource="VerbNet"/>
        <externalRef reference="Departing#Source" resource="FrameNet"/>
        <externalRef reference="Path_shape#Source" resource="FrameNet"/>
      </externalReferences>
      <span>
        <target id="t24"/> <target id="t25"/> <target head="yes" id="t26"/><target id="t27"/>
        <target id="t28"/><target id="t29"/><target id="t30"/><target id="t31"/><target id="t32"/>
      </span>
    </role>
    <role id="rl14" semRole="AM-PNC"><!--to join Mitsubishi Motors North America-->
      <span>
        <target head="yes" id="t33"/><target id="t34"/><target id="t35"/>
        <target id="t36"/><target id="t37"/><target id="t38"/>
      </span>
    </role>
  </predicate>

  <predicate id="pr11"><!--left-->
    <externalReferences>.....</externalReferences>
    <span><target id="t72"/></span>
    <role id="rl29" semRole="A0"> <!--Tocci-->
      <externalReferences>
        <externalRef reference="future_having-13.3#Agent" resource="VerbNet"/>
        <externalRef reference="fulfilling-13.4.1#Agent" resource="VerbNet"/>
        <externalRef reference="keep-15.2#Agent" resource="VerbNet"/>
      </externalReferences>
      <span><target head="yes" id="t71"/></span>
    </role>
    <role id="rl30" semRole="A2"> <!--a company-->
      <externalReferences>
        <externalRef reference="future_having-13.3#Goal" resource="VerbNet"/>
        <externalRef reference="fulfilling-13.4.1#Recipient" resource="VerbNet"/>
        <externalRef reference="keep-15.2#Location" resource="VerbNet"/>
      </externalReferences>
      <span><target id="t73"/><target head="yes" id="t74"/></span>
    </role>
    <role id="rl31" semRole="A1"><!--with rising sales and a relatively happy dealer body-->
      <externalReferences>
        <externalRef reference="future_having-13.3#Theme" resource="VerbNet"/>
        <externalRef reference="fulfilling-13.4.1#Theme" resource="VerbNet"/>
        <externalRef reference="keep-15.2#Theme" resource="VerbNet"/>
      </externalReferences>
      <span>
        <target head="yes" id="t75"/><target id="t76"/><target id="t77"/><target id="t78"/>
        <target id="t79"/><target id="t80"/><target id="t81"/><target id="t82"/><target id="t83"/>
      </span>
    </role>
  </predicate>

 <predicate id="pr49"><!--leave-->
    <externalReferences>...</externalReferences>
    <span><target id="t341"/> </span>
    <role id="rl114" semRole="R-AM-CAU"><!--Why-->
      <span> <target head="yes" id="t338"/></span>
    </role>
    <role id="rl115" semRole="A0"> <!--you-->
      <externalReferences>... </externalReferences>
      <span><target head="yes" id="t340"/> </span>
    </role>
    <role id="rl116" semRole="A1"> <!--Hyundai-->
      <externalReferences>...</externalReferences>
      <span><target head="yes" id="t342"/> </span>
    </role>
  </predicate>

  <predicate id="pr50"> <!--leave-->
    <externalReferences>... </externalReferences>
    <span><target id="t356"/></span>
    <role id="rl117" semRole="A0"><!--the most difficult decision of my life to leave Hyundai-->
      <externalReferences>...</externalReferences>
      <span>
        <target id="t348"/><target id="t349"/><target id="t350"/><target head="yes" id="t351"/>
        <target id="t352"/><target id="t353"/><target id="t354"/>
        <target id="t355"/><target id="t356"/><target id="t357"/>
      </span>
    </role>
    <role id="rl118" semRole="A1"><!--Hyundai-->
      <externalReferences>... </externalReferences>
      <span><target head="yes" id="t357"/></span>
    </role>
  </predicate>
  ....
  </srl>
```

For such predicates with the same lemma within one and the same document, the module produces a single coreference set with the type ”event” and a unique identifier within the document, followed by span-element to point to the term identifiers in the text that represent the local mentions:

```
<public publicId="4C7V-T4D0-0015-K19Y" uri="2004/4/26/4C7V-T4D0-0015-K19Y.xml"/>
    <coref id="coe4" type="event">
      <span><target id="t23"/> </span>
      <span> <target id="t72"/></span>
      <span><target id="t341"/></span>
      <span><target id="t356"/> </span>
    </coref>
```

The function that creates these event-coreference sets is part of the Java library Event-Coreference.[8] It takes a NAF file with the semantic role layer (SRL) as input stream and adds the event-coreference sets to the coreference layer. The module has now been included into the Work Package 4 pipeline for producing NAF (see Deliverable Beloki *et al.* (2014)).

A similar baseline function was provided to create coreference structures for entities in NAF. Like the predicate in the SRL layer, the representation of entities is fully mention-based. In the next example taken from the same document, we see that 3 different entities are created for the same DBPedia URI, two of which have the same lemma:

```
<public publicId="4C7V-T4D0-0015-K19Y" uri="2004/4/26/4C7V-T4D0-0015-K19Y.xml"/>
    <entity id="e3" type="organization">
      <references>
        <span><!--Hyundai Motor America-->
          <target id="t15"/><target id="t16"/><target id="t17"/>
```

[8] It can be called through the function eu.newsreader.eventcoreference.naf.EventCorefLemmaBaseline

```
        </span>
      </references>
      <externalReferences>
        <externalRef reference="http://dbpedia.org/resource/Hyundai_Motor_Company" resource="spotlight_v1"/>
      </externalReferences>
    </entity>

    <entity id="e4" type="location">
      <references>
        <span><!--Hyundai-->
          <target id="t28"/>
        </span>
      </references>
      <externalReferences>
        <externalRef reference="http://dbpedia.org/resource/Hyundai_Motor_Company" resource="spotlight_v1"/>
      </externalReferences>
    </entity>

    <entity id="e18" type="person">
      <references>
        <span> <!--Hyundai-->
          <target id="t386"/>
        </span>
      </references>
      <externalReferences>
        <externalRef reference="http://dbpedia.org/resource/Hyundai_Motor_Company" resource="spotlight_v1"/>
      </externalReferences>
    </entity>
```

In the case of entities, we take any given URI as the basis for establishing coreference. If no URI is provided, we use the lemma as a key for identity. Matches result in a single coreference set, where the type of the first entity occurrence is taken as the type for the coreference set:

```
<public publicId="4C7V-T4D0-0015-K19Y" uri="2004/4/26/4C7V-T4D0-0015-K19Y.xml"/>
    <coref id="coentity3" type="organization">
      <span> <!--Hyundai Motor America-->
        <target id="t15"/><target id="t16"/><target id="t17"/>
      </span>
      <span> <target id="t28"/></span> <!--Hyundai-->
      <span><target id="t225"/></span> <!--Hyundai-->
      <span><target id="t265"/></span> <!--Hyundai-->
      <span> <target id="t284"/></span> <!--Hyundai-->
      <span> <target id="t342"/></span> <!--Hyundai-->
      <span> <target id="t357"/></span> <!--Hyundai-->
      <span>  <target id="t386"/></span> <!--Hyundai-->
      <span><target id="t440"/></span> <!--Hyundai-->
    </coref>
```

Future versions of the system will include other modules for entity coreference. Since these modules produce the same coreference layer in NAF, the current system does not need to be adapted to work with this output.

3.3 Cross-document event coreference

The second step in event-coreference produces an instance-based representation in SEM. For this purpose, it reads any collection of NAF files and extracts semantic instances from the coreference layers produced in the previous step. These coreference layers cover all the predicates and entities represented in NAF. We used the type attribute of the coreference element to create different semantic instances for events, actors and places. Furthermore, we add all semantic typing information expressed in the entity layer and the semantic role layer for these instances. Finally, we add all mentions of the instances through lemmas, where we quantify the use of a lemma to refer to the instance.

For the time elements, we took the superset of the publication date, all timex3 expressions, and all the roles in the semantic role layer with the role value ”AM-TMP”. Future versions of the system that produce normalized values for time expressions will result in more precise time indications grouped around these normalized values. For time objects, no typing is available and we only store the lemma references.[9]

Consider the following example. In the NAF representation of the following source file: 57DF-TK31-DXF1-N0P1.xml, we find a predicate structure in the SRL layer that refers to a purchase by the company *Ford*:

```
<predicate id="pr17">
  <!--purchased-->
  <externalReferences>
    <externalRef reference="purchase.01" resource="PropBank"/>
    <externalRef reference="obtain-13.5.2" resource="VerbNet"/>
    <externalRef reference="obtain-13.5.2-1" resource="VerbNet"/>
    <externalRef reference="Commerce_buy" resource="FrameNet"/>
    <externalRef reference="contextual" resource="EventType"/>
  </externalReferences>
  <span><target id="t111"/></span>
  <role id="rl36" semRole="AM-TMP">
    <!--In 2011-->
    <span><target head="yes" id="t107"/><target id="t108"/></span>
  </role>
  <role id="rl37" semRole="A0">
    <!--Ford-->
    <externalReferences>
      <externalRef reference="obtain-13.5.2#Agent" resource="VerbNet"/>
      <externalRef reference="Commerce_buy#Buyer" resource="FrameNet"/>
    </externalReferences>
    <span><target head="yes" id="t110"/></span>
  </role>
</predicate>
```

We also find a coreference set (type event) in which this predicate (t111) is a mention and another coreference set (type organization) in which Ford (t110) is a mention:

```
<coref id="coe16" type="event">
  <span><target id="t111"/> </span><span><target id="t140"/></span>
</coref>
<coref id="coentity3" type="organization">
  <span><!--Ford Motor Company--><target id="t29"/><target id="t30"/><target id="t31"/></span>
  <span><!--Ford--><target id="t80"/></span>
  <span><!--Ford--><target id="t97"/></span>
  <span><!--Ford--><target id="t110"/></span>
  <span><!--Ford--><target id="t186"/></span>
  <span><!--Ford--><target id="t351"/></span>
```

[9] In the current version of the system, the timex expressions have not been normalized. This means that expressions such as *last week* and *Monday, January 13th* are still not pointing to the same date.

```
    <span><!--Ford Motor Company--><target id="t372"/><target id="t373"/><target id="t374"/></span>
    <span><!--Ford--><target id="t400"/></span>
  </coref>
```

At the entity layer, we can find a URI to DBPedia that identifies the entity instance *Ford*. This URI can be used to represent the full coreference set of which this entity is a part, i.e. established through the overlapping span:

```
   <entity id="e8" type="organization">
    <references>
      <span><!--Ford--><target id="t110"/></span>
    </references>
    <externalReferences>
      <externalRef reference="http://dbpedia.org/resource/Ford_Motor_Company" resource="spotlight_v1"/>
    </externalReferences>
  </entity>
```

Based on all this information, we create an event instance for the mentions of *purchase* and an entity instance for the mentions of*Ford* and relations. The same is done not just for this file but also for instances extracted from other files. If these instances match, we merge the information. The resulting instance for the *purchase* event then looks as follows in the TRIG format:

```
<http://www.newsreader-project.eu/2013_1_1_57DG-05S1-DXF1-N197.xml#coe38>
        a               sem:Event , nwr:contextual , fn:Commerce_buy ;
        rdfs:label      "purchase:5" ;
        gaf:denotedBy  <http://www.newsreader-project.eu/2013_1_1_57DG-05S1-DXF1-N197.xml#char=1460,1468&word=w270&term=
```

The URI is based on the first mention in the first NAF file. The type relations are based on the types we find in the predicate elements in NAF in addition to the basic type sem:Event. We use all the predicate expressions that match with the mentions in the event coreference set. Here the types are restricted to FrameNet labels (Baker *et al.* (1998)) and the main NewsReader event types (grammatical, communication, cognition and contextual). This is done for all matching mentions across all the sources that are considered. The refs:label shows all the labels used in the mentions. Since the events are lemma-matched, there is only a single label in this example. The label is used 5 times, as indicated after the ":". The gaf:denotedBy holds the pointers to the mentions, in this case 5 mentions across 3 different sources.

In the case of *Ford*, we create an entity instance using the DBPedia URI:

```
dbp:Ford_Motor_Company
    a               sem:Actor , nwr:person , nwr:organization , <http://www.newsreader-project.eu/framenet/Statement#
    <http://www.newsreader-project.eu/framenet/Collaboration#Partners> ,
    <http://www.newsreader-project.eu/framenet/Collaboration#Partner_1> ,
    <http://www.newsreader-project.eu/framenet/Name_conferral#Entity> ,
    <http://www.newsreader-project.eu/framenet/Commerce_buy#Buyer> ;
    rdfs:label      "Ford Motor:1" , "Ford:5" , "Ford:8" , "Ford:20" , "Ford Motor Company Fund:2" ,
    "Ford Motor Company:8" , "Ford Motor Company:1" ;
    gaf:denotedBy
    <http://www.newsreader-project.eu/2013_1_1_57DG-05S1-DXF1-N197.xml#char=2681,2688&word=w491&term=t491> ,
    <http://www.newsreader-project.eu/2013_1_1_57DG-05S1-DXF1-N197.xml#char=2849,2853&word=w520&term=t520> ,
    <http://www.newsreader-project.eu/2013_1_1_57DF-TK31-DXF1-N0P1.xml#char=550,554&word=w97&term=t97> ,
    <http://www.newsreader-project.eu/2013_1_1_57DF-TK31-DXF1-N0P1.xml#char=629,633&word=w110&term=t110> ,
    <http://www.newsreader-project.eu/2013_1_1_57DF-TK31-DXF1-N0P1.xml#char=1051,1055&word=w186&term=t186> , etc...
```

In the same way as for the verbs, we collect the types from all mentions that intersect with roles that *Ford* takes in predicates in addition to the basic type sem:Actor or sen:Place. We see here that *Ford* takes the role of Speaker, Partner, Entity and Buyer in relation to various predicates. We now see a much larger variation of labels as compared to the event. This is because we use the DBPedia URI to establish coreference and not the lemma.

After creating a list of semantic objects (events, actors, places and times) for a single NAF file, we exploit the semantic role layer to establish relations between events and any of the other elements that have been accepted as event-components: participants, places and time expressions. In case there is no relation with a time expression, we create a relation between the event and the publication date. In this way, events are minimally anchored to the publication date as a default.

We create a unique URI for all instances (including the relations) based on the document URI and any available identifier. Once we extracted the object and relation instances of a single file, we compare these with the available instances in the cluster. If there is sufficient evidence that a new instance is the same as a stored instance in a cluster, then we merge the new instance with the given instance and copy all the new mentions to the stored instance. This is done for events, actors, places, dates and relations.

A first strict condition for merging is that the time of two event instances needs to be equal before they can be merged. If that condition is satisfied, events are compared in the same way as places and actors.[10] For all 3 types of objects, we have the option to match the lemmas of all the mentions and the semantic types of all the mentions. In the current baseline system, we first check if the overlap of the lemmas exceeds the threshold. If not and if a threshold is set for the semantic type match, we check if the overlap of the semantic types exceeds the threshold. The semantic matching depends very much on the granularity of the semantic classes that are associated with the mentions. We now use a range of types coming from the SemLink repository[11], which combines VerbNet (Kipper *et al.* (2006)), FrameNet (Baker *et al.* (1998)), WordNet (Fellbaum (1998)), NomBank (Meyers *et al.* (2004)) and PropBank (Palmer *et al.* (2005)). Future version of this function can also include other similarity measures (e.g. using wordnet) without a fundamental change in the architecture. If any of the thresholds is exceeded (or equal), we consider two instances to be equal, in which case the mentions of the candidate instance are merged. If below the threshold, we consider the new candidate as a new instance. The above examples for *purchase* and *Ford* are the result of merging such instances across the sources.

Relation instances are compared as well, where we compare the candidate relations with stored relations in terms of the involved objects and the type of relation. Note that the identifiers for objects for the candidate relations are already adapted given the previous process. In case of full equality, we merge the relation mentions with the given relation instance. If not, we create a new relation instance for the candidate within the cluster. For the example *purchase*, we thus get the following relations:

```
<http://www.newsreader-project.eu/2013_1_1_57DG-05S1-DXF1-N197.xml#pr86,rl174> {
```

[10] Places and individuals (not considering their role) are persistent over time whereas events are not.

[11] `http://verbs.colorado.edu/semlink/`

```
    <http://www.newsreader-project.eu/2013_1_1_57DG-05S1-DXF1-N197.xml#coe38>
            sem:hasActor  dbp:Ford_Motor_Company .
}

<http://www.newsreader-project.eu/2013_1_1_57DG-05S1-DXF1-N197.xml#docTime_28> {
    <http://www.newsreader-project.eu/2013_1_1_57DG-05S1-DXF1-N197.xml#coe38>
            sem:hasTime  tl:2013-01-01 .
}
```

The relations are represented as named graphs with a unique identifier that is based on the predicate-semantic role identifiers or the document time. These identifiers make it possible to state properties of the relations as is shown in the next example where we state the provenance[12] of the relation:

```
    <http://www.newsreader-project.eu/2013_1_1_57DG-05S1-DXF1-N197.xml#pr44,rl93>
            gaf:denotedBy  <http://www.newsreader-project.eu/2013_1_1_57DG-05S1-DXF1-N197.xml#/rl93> .
```

Below we show some more examples of instances that are stored in the resulting TRIG file inside a named graph. We create a separate named graph for each cluster.

```
nwr:instances {

    <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#coe3>
            a               sem:Event , nwr:cognition , nwr:contextual , fn:Departing , fn:Path_shape ;
            rdfs:label      "leave:7" ;
            gaf:denotedBy
            <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#char=21,27&word=w5&term=t5> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K19Y.xml#char=128,132&word=w23&term=t23> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K19Y.xml#char=385,389&word=w72&term=t72> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K19Y.xml#char=1688,1693&word=w341&term=t341> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K19Y.xml#char=1762,1767&word=w356&term=t356> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K1B6.xml#char=152,156&word=w25&term=t25> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4F0-0015-K1C1.xml#char=4134,4138&word=w821&term=t821> .


        dbp:Hyundai_Motor_Company
            a               sem:Actor , nwr:organization ,
             <http://www.newsreader-project.eu/framenet/Name_conferral#Entity> ,
             <http://www.newsreader-project.eu/framenet/Departing#Source> ,
             <http://www.newsreader-project.eu/framenet/Path_shape#Source> ;
            rdfs:label      "Hyundai Motor America:1" , "Hyundai:8" , "Hyundai Motor Co:1" , "Hyundai Motor Co.:1" ;
            gaf:denotedBy
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K19Y.xml#char=80,87&word=w15&term=t15> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K19Y.xml#char=88,93&word=w16&term=t16> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K19Y.xml#char=94,101&word=w17&term=t17> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K19Y.xml#char=148,155&word=w28&term=t28> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K19Y.xml#char=1133,1140&word=w225&term=t225> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K19Y.xml#char=1329,1336&word=w265&term=t265> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K19Y.xml#char=1427,1434&word=w284&term=t284> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K19Y.xml#char=1694,1701&word=w342&term=t342> ,  .


                <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K19Y.xml#coentity2>
            a               sem:Actor , nwr:person ;
            rdfs:label      "Michael Tocci:1" ;
            gaf:denotedBy
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K19Y.xml#char=38,45&word=w8&term=t8> ,
```

[12] We use provenance as defined by `http://www.w3.org/TR/prov-o/` to model properties of the sources of statements

```
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K19Y.xml#char=46,51&word=w9&term=t9> .


      dbp:Usnea  a             sem:Actor , nwr:person ,
            <http://www.newsreader-project.eu/framenet/Reporting#Authorities> ,
            <http://www.newsreader-project.eu/framenet/Request#Addressee> ,
            <http://www.newsreader-project.eu/framenet/Telling#Addressee> ;
            rdfs:label     "U.:6" ;
            gaf:denotedBy
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K1B1.xml#char=1653,1655&word=w293&term=t293> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K1B1.xml#char=4217,4219&word=w799&term=t799> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K19Y.xml#char=521,523&word=w97&term=t97> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K1B1.xml#char=4923,4925&word=w933&term=t933> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K1B7.xml#char=1546,1548&word=w306&term=t306> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4F0-0015-K1CN.xml#char=1279,1281&word=w240&term=t240> .


    <http://www.newsreader-project.eu/2004_4_26_4C7V-T4F0-0015-K1C1.xml#coentity38>
            a              sem:Actor , nwr:person ,
             <http://www.newsreader-project.eu/framenet/Departing#Theme> ,
<http://www.newsreader-project.eu/framenet/Path_shape#Road> ;
            rdfs:label     "Martin Leach:1" ;
            gaf:denotedBy
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4F0-0015-K1C1.xml#char=4121,4127&word=w819&term=t819> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4F0-0015-K1C1.xml#char=4128,4133&word=w820&term=t820> .

    dbp:Michigan  a        sem:Place , nwr:location ;
            rdfs:label     "Michigan:1" , "Mich:9" ;
            gaf:denotedBy
            <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#char=3749,3757&word=w680&term=t680> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K1B0.xml#char=4088,4092&word=w779&term=t779> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K1B1.xml#char=3572,3576&word=w677&term=t677> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K1B1.xml#char=4121,4125&word=w781&term=t781> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K1B1.xml#char=4793,4797&word=w906&term=t906> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4F0-0015-K1BK.xml#char=245,249&word=w49&term=t49> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4F0-0015-K1C1.xml#char=998,1002&word=w199&term=t199> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4F0-0015-K1C1.xml#char=3215,3219&word=w637&term=t637> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4F0-0015-K1CH.xml#char=914,918&word=w172&term=t172> .

    dbp:Stuttgart  a       sem:Place , nwr:location ;
            rdfs:label     "Stuttgart:1" , "stuttgart:1" ;
            gaf:denotedBy
            <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#char=3729,3738&word=w676&term=t676> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K1B1.xml#char=3794,3803&word=w723&term=t723> .

    tl:2004-04-26  a       sem:Time ;
            rdfs:label     "2004-04-26:16" ;
            gaf:denotedBy
            <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#nafHeader/fileDesc#creationtime> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7N-GTG0-0002-M1S5.xml#nafHeader/fileDesc#creationtime> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7S-VGW0-001P-V34C.xml#nafHeader/fileDesc#creationtime> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7T-HN30-01DF-W3YR.xml#nafHeader/fileDesc#creationtime> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K19P.xml#nafHeader/fileDesc#creationtime> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K19Y.xml#nafHeader/fileDesc#creationtime> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K1B0.xml#nafHeader/fileDesc#creationtime> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K1B1.xml#nafHeader/fileDesc#creationtime> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K1B6.xml#nafHeader/fileDesc#creationtime> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K1B7.xml#nafHeader/fileDesc#creationtime> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4F0-0015-K1BK.xml#nafHeader/fileDesc#creationtime> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4F0-0015-K1C1.xml#nafHeader/fileDesc#creationtime> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4F0-0015-K1CH.xml#nafHeader/fileDesc#creationtime> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4F0-0015-K1CN.xml#nafHeader/fileDesc#creationtime> .

   <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#rl139>
            a              sem:Time ;
            rdfs:label   "this month 's:1" ;
            gaf:denotedBy
            <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#char=1808,1812&word=w332&term=t332> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#char=1813,1818&word=w333&term=t333> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#char=1818,1820&word=w334&term=t334> .

    <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#rl176>
            a              sem:Time ;
            rdfs:label     "Friday 's:1" ;
            gaf:denotedBy
            <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#char=1025,1031&word=w174&term=t174> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#char=1031,1033&word=w175&term=t175> .

            }
```

Each instance has a unique URI (based on the first proposal in the cluster or on a DBPedia URI), one or more RDF.type relations, the set of labels based on the lemma mentions and a gaf:denotedBy relation to all the mentions in all the documents within the cluster. The RDF.type relations are based on the typing in the entity layers and the semantic role layers. The labels have been extended with a frequency number, e.g. ”leave:7” means that the lemma ”leave” was used 7 times. References to mentions are based on the URI for original news item followed by the offset position and length in the text, the word and term identifiers in the NAF representation of the text. For the time objects, we make a distinction between the document date, which has a reference to the meta data in the NAF header and time expressions found in the text itself, with references to text expressions. The latter are yet not normalized and thus have an artificial URI based on the first occurrence and the role identifier from which they were extracted.

The next examples illustrate the different types of SEM relations that we represent:

```
<http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#pr16,rl35> {
    <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#coe16>
            sem:hasActor  dbp:Mitsubishi_Motors .
}

<http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#pr30,rl60> {
    <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#coe29>
            sem:hasActor  dbp:Hyundai_Motor .
}

<http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#pr49,rl100> {
    <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#coe47>
            sem:hasActor  dbp:Michael_Schneider_conductor .
}

<http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#pr50,rl101> {
    <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#coe37>
            sem:hasActor  dbp:Jurgen_E_Schrempp .
}

<http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#pr75,rl154> {
    <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#coe8>
            sem:hasPlace  dbp:United_Kingdom .
}


<http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#pr79,rl162> {
    <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#coe66>
```

```
            sem:hasPlace  <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#coentity16> .
}

<http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#pr45,rl91> {
    <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#coe43>
            sem:hasPlace  <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#coentity6> .
}


<http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#docTime_2> {
    <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#coe91>
            sem:hasTime  tl:2004-04-26 .
}

<http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#docTime_3> {
    <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#coe53>
            sem:hasTime  tl:2004-04-26 .
}
```

As explained above, the provenance of these relations is expressed in a separate named graph through a gaf:denotedBy property:

```
nwr:provenance {

    <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#pr16,rl35>
            gaf:denotedBy
            <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#/rl35> .

    <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#pr16,rl36>
            gaf:denotedBy
            <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#/rl36>

    <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#pr34,rl68>
            gaf:denotedBy
            <http://www.newsreader-project.eu/2004_4_26_4C7M-RB90-01K9-42PW.xml#/rl68> ,
            <http://www.newsreader-project.eu/2004_4_26_4C7V-T4D0-0015-K19P.xml#/rl23> .
 }
```

The provenance layer can be extended through future modules to incorporate other properties such as factuality claims and opinions.

The inter-document coreference module has been applied to the set of 63,811 English documents from Lexis Nexis. These documents were first processed by the Natural Language Processing pipeline, creating a NAF file for each. We then divided the files in clusters on the basis of the publication date and processed each cluster. Table 1 shows the quantitative results of the processing collected per year. The rows give the NAF files for each year and the SEM files produced for specific days in those years. Furthermore, we provide the number of unique instances created per year, the number of mentions and the number of labels. We also provide the mentions per instance (M/I), sources per instance (S/I) and labels per instance (L/I) ratios. Sources are the different documents from which the instances are derived and the labels are the different words used to refer to them.

The total set thus contains over 1,7 million event URIs, over 445K actors, and 62K places. In this baseline result, we only used lemma-based matches. No threshold was set for concept-based matches. It took 2:54 hours to process all the files.

For events, we see that we have almost 3 text mentions per event on average, whereas we have 7 and 16 text mentions per instance for actors and places, respectively. We see a similar phenomenon for source mentions per instance, which indicates the average number of different sources making reference to the same instance within a single publication date.[13] This is due to the fact that in this first attempt, we did not relate the event instances across the different days. Obviously if we did this, it will result in a further reduction.[14]

To get an idea about the possible reduction we can get, we can consider those instances that have been mapped to DBPedia URIs that are stable across the current clusters. In table 7, we see the distribution of instances, mentions and labels for the DBPedia URIs. The unique number of instances is low and the ratio of text mentions and source mentions is higher than for the previous table 1: 21.43 text mentions on average per instance (compare 7 to 16 for actors and places), around 8.48 source mentions on average per instance (compare 2.36 for actors and 7.64 for places). The DBPedia results thus defines an upper bound for what could be achieved for those instances not mapped to DBPedia and for events. Nevertheless, we expect realistic figure to be lower than these.

The next tables (3, 4, 6 5) show the top-50 labels for events, actors, places and time references, spread over the different years. This clearly gives an idea about the content of the data set. These tables have not yet been differentiated for semantic subclasses, which is something we expect to do in the near future.

To get an idea about the real volume of entities involved, we collected all instances of actors and places with a DBPedia URI. In total, there are 41,089 unique DBPedia URIs, of which 36,051 actors (4% of the above total) and 11,249 places (6% of the above total). This is about the amount that we should expect if we further reduce the instances across the publication date. Table (7) gives the top-frequencies for the DBPedia URIs.

The top URIs are countries and car companies. The first persons occur lower on the list: dbp:Carlos_Ghosn (4,969 text mentions), and dbp:Alan_Mulally (4,026 text mentions). Since entities are more stable in time, the figures can be used as first estimates of the real volume of instances over the full period of 10 years.

Except for the quantitative overviews, we have no evaluation data yet for our approach. Evaluations will be carried out in the 2nd year of the project and will be described in the second version of this deliverable.

[13] This number does not indicate the unique number of sources in total but source mentions.

[14] We will start this in the second year of the project. In that case, lemma-based comparisons are no longer sufficient and more information is needed. For one thing, we need to normalize all time expressions and find a way to match these normalized time expressions across the clusters that are now based on the publication time. We will then also use other types of clustering, based on topics and the place information available for the individual events. The first prototypes for this type of processing are described in section 4.

	NAF files	SEM files	sem:Event					sem:Actor					sem:Place				
			Instances	Text Mentions	M/I	Source Mentions	S/I	Instances	Text Mentions	M/I	Source Mentions	S/I	Instances	Text Mentions	M/I	Source Mentions	S/I
2003	4,581	364	158,268	378,447	2.39	273,586	1.73	37,553	229,780	6.12	78,891	2.1	5,621	73,362	13.05	33,742	6
2004	4,754	364	158,514	393,448	2.48	284,180	1.79	37,738	238,481	6.32	82,940	2.2	5,610	74,238	13.23	35,890	6.4
2005	4,535	365	153,591	415,931	2.71	294,106	1.91	34,643	245,918	7.1	83,476	2.41	5,090	78,640	15.45	36,433	7.16
2006	6,532	365	179,341	593,505	3.31	411,705	2.3	48,339	354,122	7.33	118,691	2.46	6,868	121,783	17.73	53,649	7.81
2007	6,664	364	181,864	583,248	3.21	408,656	2.25	49,297	357,573	7.25	122,128	2.48	6,865	122,417	17.83	53,772	7.83
2008	6,138	366	165,183	499,298	3.02	353,739	2.14	42,997	303,979	7.07	103,207	2.4	5,746	107,502	18.71	47,958	8.35
2009	8,208	364	201,128	712,253	3.54	509,679	2.53	49,620	368,548	7.43	128,841	2.6	6,686	146,078	21.85	65,195	9.75
2010	5,402	364	160,829	460,179	2.86	324,266	2.02	35,925	255,568	7.11	86,001	2.39	4,825	86,112	17.85	39,596	8.21
2011	4,759	363	140,742	356,018	2.53	261,083	1.86	33,050	223,179	6.75	75,621	2.29	4,648	63,322	13.62	31,069	6.68
2012	9,457	366	216,299	651,715	3.01	480,877	2.22	57,015	419,557	7.36	131,986	2.31	7,290	133,822	18.36	59,879	8.21
2013	2,780	120	68,773	203,830	2.96	145,093	2.11	19,109	130,441	6.83	40,904	2.14	3,006	42,435	14.12	18,310	6.09
TOTAL	63,810		1,784,532	5,247,872	2.94	3,746,970	2.1	445,286	3,127,146	7.02	1,052,686	2.36	62,255	1,049,711	16.86	475,493	7.64

Table 1: Results for cross document coreference and aggregation to SEM for the car industry set

YEAR	Instances	Text Mentions	M/I	Source Mentions	S/I
2003	11,588	197,320	17.03	78,637	6.79
2004	11,385	202,543	17.79	83,550	7.34
2005	10,969	220,318	20.09	86,440	7.88
2006	13,499	318,364	23.58	121,661	9.01
2007	14,266	319,380	22.39	124,693	8.74
2008	12,614	274,439	21.76	108,229	8.58
2009	13,847	344,713	24.89	139,382	10.07
2010	10,292	224,288	21.79	89,302	8.68
2011	9,267	171,220	18.48	72,277	7.8
2012	12,594	328,068	26.05	127,171	10.1
2013	5,965	104,705	17.55	39,715	6.66
TOTAL	126,286	2,705,358	21.42	1,071,057	8.48

Table 2: Results for cross document coreference and aggregation to SEM for the DBPedia instances in the car industry set

Events	Total	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
have	119,575	9,633	9,470	10,543	13,570	13,351	11,522	15,550	9,863	8,259	13,542	4,268
sale	102,844	5,676	7,312	8,331	15,228	13,917	8,516	10,597	9,902	6,231	12,918	4,215
sell	60,100	4,021	4,268	4,671	7,287	7,289	5,820	8,445	5,822	3,758	6,682	2,037
include	49,746	3,045	3,619	3,672	5,786	5,895	4,550	6,505	4,992	3,234	6,491	1,957
make	48,114	3,555	3,722	4,060	5,367	5,877	4,998	6,514	3,984	3,163	5,148	1,726
plan	39,594	2,492	2,636	3,027	4,921	3,818	4,202	7,202	3,360	2,321	4,496	1,117
brand	36,005	1,739	2,124	2,431	3,867	4,475	3,251	4,998	4,189	2,714	4,710	1,507
model	34,422	2,648	2,844	3,362	3,702	3,587	3,176	3,593	3,374	2,417	4,279	1,440
share	31,294	1,852	2,175	2,873	5,176	4,242	2,816	3,811	2,799	1,739	2,923	887
see	27,378	2,120	2,198	2,230	2,916	3,178	2,593	3,432	2,435	1,709	3,364	1,203
expect	26,844	1,892	2,005	2,187	3,032	3,002	2,621	3,866	2,459	1,789	3,063	927
increase	26,782	1,464	1,644	1,964	3,085	3,004	2,370	2,715	2,793	1,947	4,492	1,304
offer	24,681	1,537	1,848	1,871	2,819	2,712	1,947	3,506	2,281	1,904	3,244	1,012
report	23,719	1,117	1,671	1,831	2,988	3,058	1,852	3,289	2,135	1,718	3,168	891
use	23,002	1,885	2,122	1,818	2,404	2,064	2,350	2,278	2,214	1,786	3,034	1,047
build	22,483	2,287	2,120	2,198	2,556	2,171	2,299	2,401	1,678	1,198	2,736	837
take	22,271	1,652	1,672	1,836	2,421	2,418	2,406	3,598	1,824	1,373	2,368	702
get	22,074	1,855	1,785	2,013	2,205	2,358	1,952	3,312	1,906	1,612	2,449	626
launch	21,488	1,515	1,465	1,895	2,037	2,369	2,085	2,462	1,896	1,673	3,115	976
product	20,985	1,618	1,587	1,615	2,175	2,247	1,825	2,631	2,084	1,452	2,895	856
percent	20,985	1,256	1,456	1,991	4,092	3,115	1,612	2,137	1,721	1,061	2,024	520
announce	20,927	1,051	1,139	1,604	2,457	2,220	2,229	3,100	1,935	1,329	2,954	907
buy	20,685	1,407	1,539	1,781	2,266	2,533	2,433	3,299	1,685	1,321	1,839	582
continue	20,510	1,132	1,370	1,478	2,218	2,051	2,042	2,865	2,195	1,409	2,863	887
show	19,849	1,664	1,802	1,956	1,967	2,105	2,241	1,778	1,646	1,482	2,363	845
start	19,416	1,579	1,382	1,555	2,123	2,020	1,984	2,205	1,856	1,232	2,677	803
produce	18,951	1,596	1,315	1,702	2,034	1,995	1,795	2,131	1,610	1,246	2,636	890
part	18,951	1,345	1,393	1,454	2,124	1,748	1,967	2,858	1,680	1,171	2,454	757
help	18,899	1,149	1,256	1,529	1,907	2,102	1,897	2,926	1,586	1,213	2,557	777
price	18,138	1,200	1,343	1,918	2,620	1,940	2,221	1,632	1,421	1,346	1,855	641
end	17,995	1,116	1,319	1,527	2,279	1,869	1,527	2,907	1,570	1,201	1,999	681
base	17,855	1,208	1,400	1,562	1,794	1,715	1,721	2,388	1,716	1,344	2,364	643
market	17,764	1,273	1,242	1,484	1,860	2,016	1,713	2,132	1,749	1,044	2,421	830
operation	17,719	1,296	1,264	1,264	2,056	2,045	1,759	2,817	1,543	1,000	1,874	800
stake	17,671	668	1,104	968	2,503	2,045	1,993	4,022	1,354	1,090	1,570	354
want	17,641	1,332	1,300	1,549	2,088	2,202	1,728	2,591	1,391	1,147	1,766	547
become	17,360	1,596	1,330	1,441	1,733	2,025	1,576	2,353	1,567	1,196	1,856	686
give	17,057	1,275	1,329	1,327	1,980	1,870	1,525	2,678	1,503	1,103	1,856	611
add	16,996	1,181	1,301	1,469	1,940	1,826	1,398	1,985	1,585	1,306	2,314	691
do	16,709	1,417	1,358	1,533	1,948	1,920	1,637	2,299	1,348	1,047	1,703	498
lead	16,403	1,085	1,184	1,135	1,841	2,146	1,473	2,116	1,495	1,166	2,087	675
rise	16,095	674	1,076	1,361	2,822	2,739	1,199	966	1,440	972	2,196	650
dealer	15,931	1,237	1,064	1,096	1,447	1,469	1,487	3,105	1,677	1,116	1,662	571
work	15,285	1,222	1,046	1,138	1,596	1,705	1,715	2,049	1,271	1,018	1,893	632
provide	15,128	899	1,064	890	1,131	1,347	1,514	2,238	1,538	1,266	2,461	780
production	14,936	956	1,047	1,348	1,587	1,374	1,427	1,751	1,417	1,144	2,225	660
design	14,870	1,287	1,338	1,320	1,599	1,351	1,142	1,600	1,338	995	2,273	627
drive	13,908	1,269	1,260	986	1,321	1,458	1,364	1,244	1,206	949	2,122	729
win	13,873	1,427	1,016	1,029	1,618	1,482	1,460	1,732	1,040	893	1,642	533
own	13,677	1,092	1,098	992	1,553	1,435	1,389	1,895	1,022	961	1,759	481

Table 3: The 50 most-frequent event labels across the years

Actors	Total	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Toyota	28,769	2,020	2,014	2,621	3,986	3,985	2,509	2,925	2,928	2,563	2,392	825
GM	26,985	1,453	1,232	2,258	5,228	2,676	2,324	6,707	2,112	1,073	1,540	382
Land Rover	18,236	1,270	1,808	1,137	2,002	3,399	3,401	1,496	819	579	1,707	618
BMW	16,126	1,339	1,675	1,383	1,546	1,650	1,265	1,573	1,590	1,266	2,234	605
Honda	10,913	1,019	978	1,041	1,246	1,406	1,070	919	994	829	1,185	225
Volkswagen	8,278	352	486	526	776	1,141	687	1,390	686	1,004	977	253
Aston Martin	7,652	466	520	415	1,471	1,043	504	454	511	420	1,354	494
VW	7,196	204	304	515	766	953	616	1,698	556	767	703	114
Audi	7,014	448	506	508	770	688	546	697	710	840	1,005	296
volkswagen	6,617	184	240	473	699	974	520	1,258	557	767	788	157
Hyundai	5,984	525	604	649	527	434	475	510	587	654	774	245
Mitsubishi	5,951	646	1,446	583	496	397	437	702	446	313	341	144
Range Rover	5,492	378	505	462	303	323	421	510	505	421	1,349	315
Chrysler	5,433	657	428	305	537	553	667	589	443	403	655	196
Jaguar Land Rover	5,217	3	5	3	2	98	873	1,104	193	264	2,032	640
Lincoln	5,190	398	520	247	703	782	345	585	545	156	689	220
Motors	5,094	393	392	440	501	405	451	1,064	547	382	417	102
Alfa Romeo	4,868	229	328	418	248	386	355	800	383	317	910	494
SUV	4,587	381	552	581	466	344	416	237	267	332	715	296
Ferrari	4,275	474	474	366	359	442	335	401	279	402	498	245
land Rover	3,952	255	580	227	435	648	521	274	121	208	509	174
Fiat	3,876	260	148	206	173	346	219	627	461	438	814	184
Porsche	3,775	212	215	256	229	448	463	695	424	455	294	84
Volvo	3,652	302	310	210	256	329	409	597	568	191	404	76
MG Rover	3,645	715	664	1,330	226	155	140	185	30	27	150	23
Nissan	3,559	315	324	322	518	306	300	183	283	512	356	140
Ford Motor Co.	3,203	291	321	326	621	482	300	316	256	99	172	19
PSA Peugeot Citroen	3,182	53	109	227	195	220	181	262	301	179	1,340	115
Tata Motors	3,178	5	3	31	20	259	1,034	617	223	145	714	127
Ford Motor Company	3,167	256	230	338	404	386	396	393	229	66	386	82
japanese	3,142	333	346	371	361	329	248	293	300	316	205	40
Motors Corp.	3,049	286	205	295	907	530	287	496	27	3	13	0
Toyota Motor Corp.	3,023	187	211	293	713	534	309	284	215	125	118	34
german	2,963	209	333	228	323	357	212	278	213	354	342	114
Volkswagen AG	2,900	71	94	201	243	408	293	544	321	350	304	71
Suzuki	2,856	339	232	155	369	244	246	219	272	504	210	66
Subaru	2,671	254	268	272	300	314	207	182	235	258	309	72
Mazda	2,636	296	226	206	182	224	184	212	241	192	399	274
Jaguar	2,615	209	185	242	206	358	317	313	185	133	317	150
Kia	2,544	226	171	335	208	193	116	266	295	259	385	90
McLaren	2,532	195	63	171	220	439	180	258	93	177	363	373
GMC	2,492	133	95	136	195	165	193	609	441	196	266	63
Peugeot	2,484	133	182	243	250	158	216	293	232	130	574	73
Sergio Marchionne	2,408	0	21	80	47	85	75	954	191	171	577	207
Ford Motor Co	2,393	182	191	274	441	447	227	228	195	69	100	39
italian	2,284	220	214	178	148	133	116	336	167	186	480	106
Renault	2,251	184	148	249	284	333	141	183	224	269	195	41
Mitsubishi Motors	2,243	192	1,048	182	164	70	92	177	157	29	55	77
Lexus	2,241	224	246	332	220	290	125	215	150	163	162	114
Power	2,228	173	289	260	289	223	187	221	185	138	214	49

Table 4: The 50 most-frequent actor labels across the years

Places	Total	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
China	23,432	1,237	1,418	1,716	2,130	1,782	1,842	2,260	2,700	1,666	5,214	1,467
Chrysler	21,854	556	662	579	777	3,510	1,979	7,371	1,388	1,004	2,897	1,131
Europe	17,219	1,170	1,377	1,643	1,894	1,674	1,655	2,125	1,419	924	2,719	619
Toyota	12,715	557	535	1,064	1,454	3,040	1,125	896	2,310	677	825	232
Japan	12,202	1,137	1,007	1,149	1,634	1,304	829	1,119	968	1,163	1,233	659
Fiat	11,545	356	326	432	376	475	512	4,560	695	621	2,361	831
UK	11,170	681	841	666	790	1,418	2,536	982	516	555	1,580	605
japanese	11,089	1,038	1,073	1,014	1,747	1,585	734	924	727	935	991	321
european	11,071	843	906	1,117	1,069	1,228	923	1,927	754	525	1,446	333
United States	11,040	932	983	1,105	1,426	1,126	1,005	1,241	1,008	727	1,145	341
Jaguar	10,759	723	1,041	787	1,748	2,013	2,205	537	347	343	820	195
India	10,380	103	117	325	648	1,052	1,711	1,368	1,101	1,167	2,172	616
US	10,373	430	711	846	1,091	1,317	1,133	2,199	797	480	959	410
Nissan	10,229	717	746	845	2,015	1,044	931	842	1,002	695	1,068	324
Germany	9,931	685	645	799	997	1,182	742	2,260	733	595	887	406
german	9,907	494	599	601	767	1,169	739	2,948	650	784	877	279
North America	8,753	582	657	806	1,206	987	881	947	726	427	982	552
Volvo	8,748	647	619	741	1,225	1,505	1,018	1,113	1,187	166	459	68
Detroit	8,711	790	618	601	1,027	1,144	929	1,186	659	542	877	338
Lincoln	8,045	795	922	330	1,406	1,256	604	786	710	225	794	217
chinese	7,698	319	433	687	591	514	526	1,157	1,075	522	1,424	450
Porsche	7,184	125	132	456	480	580	788	2,114	512	695	1,147	155
Chevrolet	7,027	416	220	511	749	1,223	801	1,085	843	499	474	206
Honda	6,781	490	357	567	697	775	662	1,035	566	535	793	304
Renault	6,492	301	409	594	1,808	674	588	480	752	323	474	89
Mazda	6,153	676	674	407	704	472	665	412	542	244	1,066	291
Audi	5,879	310	357	267	374	525	666	681	477	419	1,422	381
Opel	5,798	130	121	163	165	115	115	3,814	299	310	482	84
Saab	5,716	345	241	425	349	328	432	2,020	727	559	261	29
american	5,705	427	467	486	849	858	481	832	421	244	504	135
Russia	5,306	148	165	271	657	837	1,013	605	383	388	548	291
Canada	5,212	273	384	470	477	505	361	743	583	587	592	235
Dodge	5,172	402	201	352	537	836	561	942	355	184	649	153
british	5,162	360	399	527	474	711	1,009	384	178	223	692	205
north american	5,093	319	347	537	1,059	783	538	541	320	138	366	142
BMW	4,790	168	210	294	193	411	785	390	287	449	1,362	241
Italy	4,446	231	236	338	452	379	321	962	268	279	662	318
Australia	4,221	541	472	435	416	385	361	331	283	330	527	140
Lexus	4,080	357	389	378	383	590	340	366	405	256	432	184
Hyundai	3,977	275	345	260	414	290	254	461	583	368	586	141
Mich	3,903	307	248	305	567	386	390	534	362	206	445	153
italian	3,809	187	138	164	221	172	146	1,436	204	277	641	223
France	3,755	256	220	344	600	458	360	427	312	200	452	126
Chrysler Group	3,590	104	140	197	385	548	25	187	200	135	1,227	442
Mercedes	3,584	423	416	411	210	352	222	246	272	177	563	292
America	3,576	273	303	275	418	308	443	377	289	324	448	118
Sweden	3,121	159	166	220	257	279	410	662	556	174	192	46
Suzuki	3,015	179	172	212	300	154	160	603	278	666	190	101
Asia	3,008	196	212	209	273	525	290	313	350	159	402	79
swedish	2,990	72	51	122	221	253	442	1,155	476	119	66	13

Table 5: The 50 most-frequent place labels across the years

Time expressions	Total	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
when	19,828	1,704	1,638	1,656	2,170	2,318	1,926	2,468	1,790	1,243	2,333	581
now	17,510	1,342	1,412	1,419	1,674	1,809	1,839	2,297	1,559	1,215	2,235	709
last year	12,662	912	916	1,201	1,608	1,806	1,092	1,183	1,130	796	1,458	560
this year	10,685	948	886	1,007	1,331	1,180	899	1,028	864	812	1,294	436
already	8,970	655	616	747	964	954	1,033	1,365	743	583	994	316
still	8,900	672	652	703	891	1,030	982	1,338	885	589	894	264
annual	7,433	535	551	626	800	850	789	796	632	596	956	302
recently	7,230	459	454	560	836	727	887	976	581	449	1,032	269
today	7,148	469	518	437	651	559	627	991	556	586	1,330	424
current	5,972	402	415	477	552	584	623	959	635	381	726	218
currently	5,516	357	363	400	557	523	482	654	534	472	874	300
last month	5,321	269	338	347	994	955	377	673	485	254	483	146
former	5,247	388	430	448	713	785	486	658	475	297	438	129
future	4,624	334	402	370	414	564	461	664	413	297	538	167
last week	4,510	287	320	370	650	448	517	838	306	280	393	101
then	4,250	474	367	346	440	422	387	534	332	309	477	162
late	4,089	350	364	363	398	417	405	417	347	333	509	186
yesterday	4,066	324	295	446	510	531	399	687	301	173	312	88
recent	3,923	300	292	329	460	472	396	493	382	242	444	113
meanwhile	3,552	247	240	310	449	341	323	529	365	262	378	108
in 2006	3,327	46	190	280	365	1,175	530	256	191	78	175	41
in 2005	3,150	181	202	275	1,012	568	289	247	123	87	140	26
never	3,118	299	268	250	287	355	303	351	282	240	390	93
in 2007	2,995	32	37	184	272	472	947	469	211	131	183	57
again	2,946	224	187	242	334	406	284	340	302	220	304	103
in 2009	2,881	4	5	18	56	208	251	569	693	368	549	160
Wednesday	2,793	126	132	226	621	398	211	396	241	146	251	45
previously	2,706	141	158	169	328	371	272	385	258	199	305	120
Tuesday	2,652	130	151	211	418	350	215	458	279	122	258	60
in 2008	2,611	6	21	47	201	222	346	776	389	202	314	87
in 2004	2,605	209	327	915	375	214	144	208	76	51	73	13
Thursday	2,567	100	176	177	456	315	290	514	204	96	198	41
later	2,531	214	174	192	279	239	299	332	244	155	308	95
in 2010	2,476	5	8	15	38	93	292	344	336	780	478	87
Friday	2,472	116	143	188	632	271	176	398	169	141	186	52
in 2011	2,435	8	0	0	3	9	95	266	209	288	1,340	217
Monday	2,368	124	106	200	343	258	232	476	162	169	206	92
this week	2,358	162	152	184	321	252	200	420	197	163	251	56
next year	2,351	259	188	261	236	218	181	351	210	148	259	40
soon	2,336	207	185	203	211	177	317	298	209	168	285	76
earlier	2,263	123	139	164	331	274	173	355	244	152	260	48
first	2,256	166	172	162	275	269	190	302	212	138	271	99
yet	2,213	152	160	164	249	230	187	407	148	160	267	89
in 2003	2,109	256	632	306	230	134	145	85	91	66	128	36
always	2,038	150	213	158	213	250	153	212	182	152	271	84
ever	2,022	156	230	155	144	213	191	270	152	143	270	98
in 2000	1,933	205	245	236	207	276	272	193	87	60	122	30
previous	1,774	167	157	128	209	151	142	187	183	127	222	101
long	1,768	161	135	148	201	177	150	197	168	137	244	50
in 2002	1,730	579	238	187	148	141	110	83	68	60	92	24

Table 6: The 50 most-frequent time labels across the years

DBP	Text Mentions	Source Mentions	Labels	Types
dbp:Ford_Motor_Company	167,701	3,569	59	389
dbp:Toyota	59,975	3,378	56	285
dbp:Land_Rover	56,139	2,818	43	245
dbp:United_States	50,602	3,456	69	228
dbp:Chrysler	45,539	2,794	38	241
dbp:General_Motors	38,517	2,781	37	288
dbp:China	34,750	2,631	30	183
dbp:Europe	32,782	3,260	26	152
dbp:Japan	29,341	3,178	22	175
dbp:BMW	29,181	3,060	40	250
dbp:North_America	28,133	2,685	28	89
dbp:Fiat	27,785	2,144	48	210
dbp:Nissan_Motor_Company	27,158	2,682	36	202
dbp:Volkswagen	26,551	2,850	43	259
dbp:Honda	26,203	2,877	38	226
dbp:Germany	25,936	3,087	23	166
dbp:United_Kingdom	24,614	2,819	35	181
dbp:Jaguar_Land_Rover	19,547	853	17	173
dbp:Aston_Martin	18,601	1,771	23	154
dbp:Porsche	18,397	2,096	38	185
dbp:PSA_Peugeot_Citroen	17,273	959	21	129
dbp:India	15,856	1,847	21	116
dbp:Volvo	15,813	2,513	28	161
dbp:Volkswagen_Group	15,214	1,665	26	144
dbp:Audi	15,090	2,696	20	158
dbp:Jaguar_Cars	13,817	2,461	13	173
dbp:Range_Rover	13,673	1,209	27	150
dbp:Abraham_Lincoln	12,893	1,772	21	140
dbp:Daimler_AG	12,614	1,746	24	142
dbp:Alfa_Romeo	12,210	1,385	15	129
dbp:Italy	12,097	2,063	20	108
dbp:Mercedes_Benz	11,216	2,271	26	159
dbp:Renault	11,183	1,890	22	147
dbp:Saab	10,893	1,653	28	140
dbp:Mazda	10,368	2,221	26	129
dbp:Detroit	10,111	2,256	15	128
dbp:Chevrolet	9,939	1,694	21	94
dbp:MG_Rover_Group	9,573	726	11	136
dbp:Sport_utility_vehicle	9,498	2,258	18	141
dbp:Sweden	9,210	1,732	18	81
dbp:France	9,130	2,057	27	119
dbp:Suzuki	9,127	1,522	26	135
dbp:Canada	9,085	1,823	24	84
dbp:Russia	8,499	1,441	17	77
dbp:Tata_Motors	8,113	824	14	113
dbp:Ford_Motor_Credit_Company	8,075	675	11	71
dbp:Hyundai	7,993	2,130	8	161
dbp:Hyundai_Motor_Company	7,935	820	24	103
dbp:Opel	7,927	1,180	15	122
dbp:William_Clay_Ford_Jr	7,626	543	8	178

Table 7: The 50 most-frequent DBPedia URIs for the car industry set

4 Beyond lemma matching

4.1 Introduction

The lemma-based approach described in the previous section is limited to new items that are grouped in rather strict clusters. The reason for this is that lemmas become too ambiguous if the time and place constraints are lifted. On a single day, the number of **attacks** reported in the news are limited and thus mapping all mentions of the word **attack** has a high precision.

There are three major problems with the lemma-baseline:

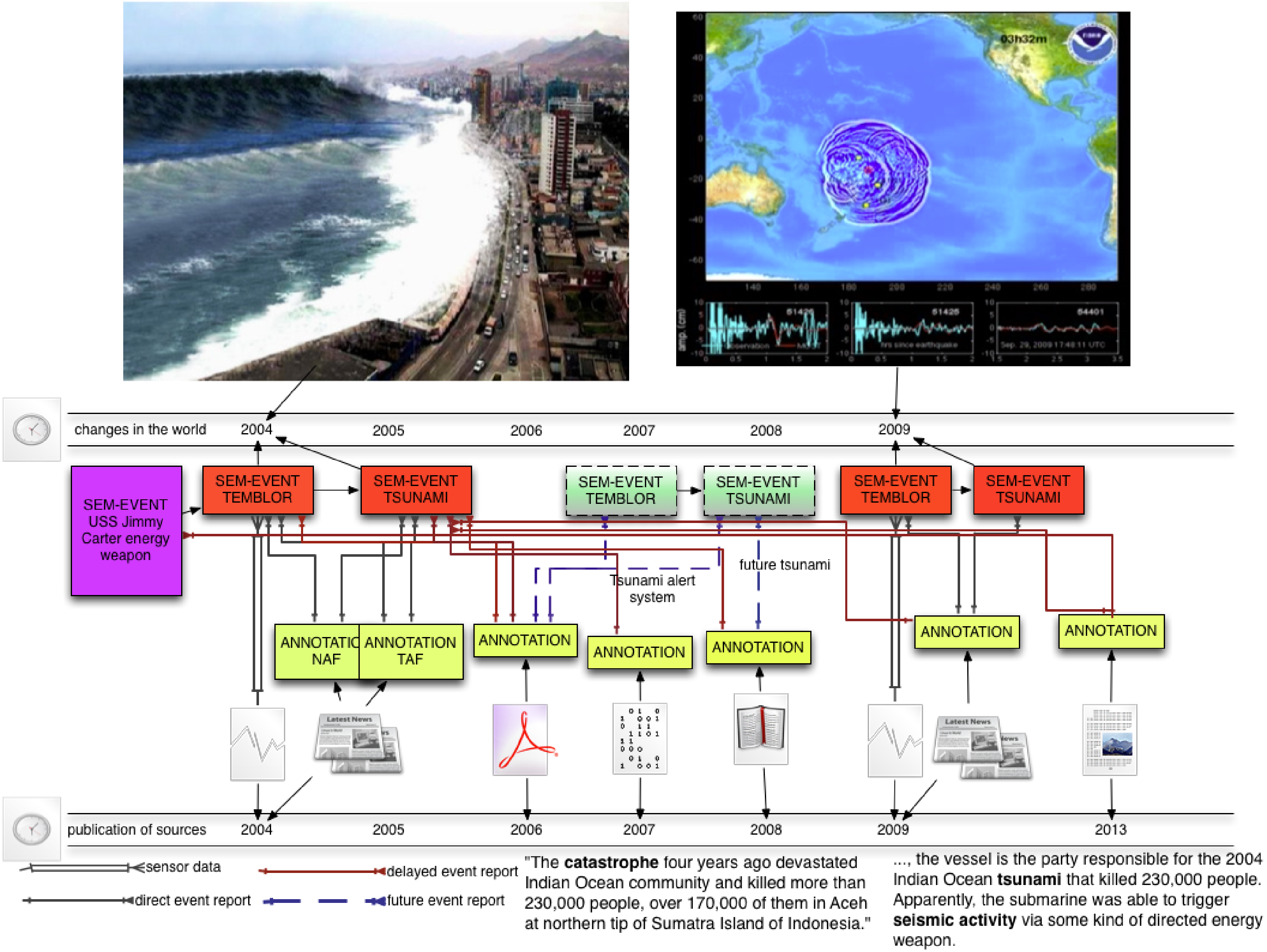
1. despite the strict time-based clusters, there may still be some ambiguity for lemmas across different events within the same time-frame, e.g. it is not unlikely that two different **attacks** happen on the same day.

2. it does not handle any variation in referring to events and their participants and therefore the recall remains low.

3. news articles do not only report on current events but also on past and future events.

The last point is crucial for interpreting news streams over longer periods of time. Very often, news articles give background information on past events or they give new information on events that took place earlier in time. In yet other cases, they talk about events in the future that did not happen yet but some day may happen. If the actual event reported matches a speculated event from older news, we need to match event descriptions across different publication dates. This situation is shown in image 3 that is taken from Fokkens *et al.* (2013). Here two earthquakes and tsunamis are shown on the upper time line that approximates the changes in the world. The lower time line represents sources of mentions of these events. Sensors can pick up an event exactly at the moment it happened, as was the case towards the end of 2004. News agencies report shortly after the event. Later in time, more publications are released with more details and knowledge about the event. In this actual case, some sources also start mentioning possible future events, in the context of a tsunami alert system. When a new earthquake and tsunami happens in 2009, picked up by a sensor, the news immediately refers back to the event in 2004 and the debate on the alert system. Finally, the picture shows a source in 2013 (a US veteran website) that introduces a new event before the 2004 disaster as the potential cause: the US marine vessel Jimmy Carter experimenting with a new energy weapon which causes the temblor instead of the tectonic plates.

Such mixtures of past, current and future events over longer periods of time are the rule rather than the exception in news. They also show a large variation in referring to events. In the next two sections, we therefore describe the work started in NewsReader to deal with these problems. This will continue in the second year of the project.



"The **catastrophe** four years ago devastated Indian Ocean community and killed more than 230,000 people, over 170,000 of them in Aceh at northern tip of Sumatra Island of Indonesia."

..., the vessel is the party responsible for the 2004 Indian Ocean **tsunami** that killed 230,000 people. Apparently, the submarine was able to trigger **seismic activity** via some kind of directed energy weapon.

Figure 3: Past and future event mentions in news streams

4.2 Experimenting with a Bayesian model

Bayesian models are another choice to face event coreference resolution. In particular, we are implementing the model presented by Bejan *et al.* (2009) and Bejan and Harabagiu (2010). This model follows the *Quinean theory* about event coreference (Quine, 1985), which states that two event mentions are coreferential if they share the same properties and participants. To characterize each mention of an event they proposed the following set of features:

- Lexical Features (LF)

  Head word, left and right surrounding words, left and right event mentions

- Class Features (CF)

  Part-of-Speech, event class, class of the head-word

- WordNet Features (WF)

  Synonymy relations, lexical-files

- Semantic Features (SF)

  Predicate argument structures (PropBank), semantic frames (FrameNet)

Bejan and Harabagiu (2010) included these features into an extension of the *hierarchical Dirichlet process* (HDP) model (Teh *et al.*, 2006) inspired from the proposal for entity coreference by Haghighi and Klein (2007). The application of the HDP to event coreference resolution allows to cluster the different mentions of events in a collection of documents. Each of the clusters obtained by the model represents an *instance of an event*, and all the mentions belonging to it would be correferent with each other. As HDP is an **unsupervised** and **non parametric** bayesian model the number of resulting clusters is potentially infinite, in other words, there is no need of estimating and setting manually the number of final event instances contained in the collection. In the extention proposed by Bejan and Harabagiu (2010), a *Dirichlet process* (Ferguson, 1973) is associated with each document, and each mixture component (i.e., event) is shared across documents. This means that the inferred distributions over the events describe clusters of coreferent mentions not only inside a single document but also **across** all the documents in the collection.

The performance of the model was firstly evaluated using the *ACE 2005* corpus (Walker *et al.*, 2006), but due to its lack of diversity of events, Bejan *et al.* (2009) developed a new corpus that also includes cross-document coreference: so-called *EventCorefBank* (ECB, see section 2). Table 8 shows the results of the model on the ECB with different settings of

	WD									CD								
Model configuration	B3			CEAF			PW			B3			CEAF			PW		
	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
HDP (LF)	81.4	98.2	89.0	92.7	77.2	84.2	24.7	82.8	37.7	63.8	97.3	77.0	84.9	54.3	66.1	27.2	88.5	41.5
HDP (LF+CF)	81.5	98.0	89.0	92.8	77.9	84.7	24.6	80.7	37.4	64.6	97.3	77.6	85.3	55.6	67.2	27.6	88.7	42.0
HDP (LF+CF+WF)	82.0	98.9	89.6	93.7	78.4	85.3	26.8	89.9	41.0	65.8	98.0	78.7	86.7	57.1	68.8	29.6	93.0	44.8
HDP (LF+CF+WF+SF)	82.1	99.2	89.8	93.9	78.2	85.3	27.0	92.4	41.3	65.0	98.7	78.3	86.9	56.0	68.0	29.2	95.1	44.4

Table 8: Results for within-document (WD) and cross-document (CD) coreference resolution on the ECB dataset.

features employing the coreference metrics: B3 (Bagga and Baldwin (1998)), CEAF (Luo (2005)) and the positive-link-identification, also known as *Pairwise (PW)*, a metric that computes P, R and F over all pairs of mentions in the same entity cluster.

Within the frame of NewsReader we plan to obtain an implementation of the HDP model using the output of the pipeline described on **WP4** to extract the set of features listed previously in order to replicate the results showed in (Bejan and Harabagiu, 2010). However, as the analysis performed by the tools of the pipeline provides a further and richer annotation of the documents, we also plan to use this analysis to include new features into the model.

4.3 Event coreference based on event components

In this section, we report on the work to deal with even larger variation in references to events and resolving ambiguity across a wider variety of events.

4.3.1 Starting points

Analysis of event mentions in textual data shows that descriptions of one and the same event can differ in specificity and granularity (compare: **two students taken hostage in Beslanian school** vs. **two people taken hostage in a classroom in Beslan Russia**). High level events, as war, are more general and abstract with longer time span and groups of participants; low level events, e.g. a shooting event, are rather specific with shorter duration, and individual participants (Cybulska and Vossen (2010)). In news texts, we frequently find both high and low level event descriptions. To still match these different descriptions, we applied an event model that consists of 4 components: a location, time, participant and an action slot (see van Hage *et al.* (2011) for the formal SEM model along the same lines).

In accordance with Quine (1985), we assume that coreference between elements of the contextual setting of events is crucial for solving event coreference itself. As explained before, time and place are the most important defining components. Coreference of events only makes sense for events within the same time and place. Furthermore, we claim that (linguistic) coreference is not an absolute notion. For example, **shooting** and **several shots** can refer to the same event and people may have different or vague intuitions about their identity (for a discussion of full and partial coreference see also Hovy *et al.*).

This approach employs a gradable notion of coreference with a continuum of non-disjoint events on which coreference of events (bombing vs. bombing attack) gradually transitions into other event relations such as scriptal (event vs. its subevent e.g. explosion as a step in the script of a bombing attack), is-a (bombing being a kind of attack) and membership relations (attack being a member of series of attacks). The gradual notion of confidence in coreference inversely correlates with semantic distance between two instances.

Semantic distance between instances of an event component can be determined, among others, by the kind of semantic relation between them. In text one comes across specific and general actions, participants, time expressions and locations; compare e.g. shooting, fighting, genocide and war, or participants: soldier vs. (multiple) soldiers vs. troops and multiple troops. The same holds for time markers as day, week and year and for locations: city vs. region vs. continent. Table 9 exemplifies instances of event components related through hyponymy and meronymy. Mentions of event components are either (partially) overlapping or disjoint.

Event Components	Is-a: from Class to Subclass	Inclusion: from Part-of to Member
Location	city to capital	Bosnia to Srebrenica
Participants	officer to colonel	army to soldier
Time	to Friday	week to Monday
Action	attack to bombing	series of attacks to attack

Table 9: Examples of event components related through hyponymy and meronymy, taken from Cybulska and Vossen (2013).

We developed a model for establishing gradual co-reference between event mentions based on the semantic similarity and granularity distance of the components that make up the event. Different components require different similarity metrics. Time and place have a different semantics than actions and participants. Since reasoning over time and place is more strict and can be done using the data in the Knowledge Store, we focussed on using loose similarity measures for actions and participants within a more strict time and place matching. Another reason for focussing on actions and participants is that specific time and place information is not always present in the sentence in which the event is mentioned.

Within this approach, we analyze semantic relations and semantic distance between two instances of each event component, to obtain a coreference score per component. We do not only take exact lemma-based matches of event mentions into account but we allow for soft matching based on shifts in levels of granularity and abstraction. Our intuition is that shifts vs. agreement in the level of granularity and in the level of abstraction play a crucial role in establishing coreference relations; obviously together with other coreference indicators such as lemma repetition, anaphora, synonymy and disjunction. Once semantic distance and granularity agreement is calculated for every component of an event pair, the separate scores are combined into a single score for an event pair indicating the likelihood of real world coreference as a whole. Through empirical testing, we can determine thresholds for establishing optimal coreference relations across events and their components.

The coreference module takes the NAF representation of text as input and uses the WordNet synsets assigned to the term layer to determine similarity matching between components (each component being represented by the head term of the phrase). Various similarity measures have been implemented in the Wordnet tools package. Wordnet tools is an opens-source package of functions that can be applied to any wordnet in WordNet-LMF format.[15] Instead of WordNet-based similarity, other measures can easily be integrated, such as distributional semantic vectors.

The module creates a separate matrix for each event component: action mentions, participants, places and time references. It first establishes a similarity score across all elements within the matrix. Potential co-reference sets are created for all mentions that exceed a preset threshold. This step is recall oriented and thus creates larger sets, while mentions can belong to more than one set. Next, we combine the components into a single event representation and check the overlap across the components of all the mentions in the same initial co-reference set. Within the module, we can set the weight for the overlap of each component.

In this way, we can fine-tune the system in various ways, through what we call event-equations. If two mentions of events have a greater semantic distance, e.g. **shooting** and **attack**, we can demand that the participants and/or the time and place should have a more strict matching, or the other way around, if participants are more distant, e.g. **British soldier** versus **Western alliance**, we can demand that the action, time and place need to be more strict.

In addition to the cumulative score of the similarity of the components, we can also measure the degree of component sharing. Event descriptions can vary in their richness. They can for example leave out the agent or the patient or do not specify the location or exact time. Within the candidate coreference sets, we can make further groupings for event mentions that share a high degree of components. We then boost the action coreference score for each shared participant, time and location. Since these participants, time and location mentions are also part of a coreference chain, we take the coreference score of each chain as a factor weight for sharing. For example, if two mentions of events each have a participant that is part of a participant coreference chain, we add the score of the participant coreference chain to the score of the event coreference relation between these two mentions. Likewise, overlap of participants with a high coreference score thus contribute more than overlap of participants with a low participant coreference score.

We used the following formula to model this factorization, in which membership to a coreference set of an event is initially based on the coreference score of the action mention but it is strengthened by the proportion that participants, time references or locations are shared with other mentions:

[15] Wordnet tools is freely available under a GPL license. It can be downloaded from: http://wordpress.let.vupr.nl/software/wordnettools/

$$Coref(m, E) = maxL\&C(m, E) + P(p) * P(t) * P(l) \tag{1}$$

In this formula, E is the set of mentions in the action coreference set, max LC is the highest similarity score for the mention m in the set E. The coreference score of action mention m equals the sum of the maximum coreference score max L&C, and proportion P of overlapping participants p (of m with the other members of the set) or times t or locations l, with other members of the set.

4.3.2 Experiments

We ran a number of experiments to see the effect of the above equation on the coreference relations in the stand-off annotation of events (Lee *et al.* (2012)) on top of the EventCorefBank (ECB) corpus, annotated with cross-document coreference between event mentions (see section 4.2 for more details on the ECB). The results described below were published in Cybulska and Vossen (2013).

To measure only the influence of time, location and participants on event coreference resolution, we used the set of event mentions from the evaluation data as a given set of events but without the coreference relations. The evaluation should not be skewed by the event extraction process itself. We thus measured the impact of the components on the ideal set of events. In addition to the given event mentions, we formulated patterns in the Kybot system[16] to find participants, places and time expressions.[17].

As the primary measure for matching of the action and participant component matrixes, we used the similarity method by Leacock and Chodorow (1998) as it has been implemented in Wordnet Tools. A second heuristic calculates distance in granularity. To determine granularity levels, we defined two semantic classes over synsets in WordNet: gran_person (e.g. *soldier*, *doctor*) denoting individual participants and gran_group referring to multiple participants (e.g. *army* or *hospital*). These two classes cover 36 WordNet hypernyms which map to 9,922 synsets. On top of agreement in granularity levels, we also account for lexical granularity clues within a level such as number and multiplications. At this point we make a rough distinction between one and multiple items within a concept type (e.g. gran_person). Difference in granularity level or number is treated as indication of a granularity shift and is turned into a distance measure. To better handle 43,415 participant mentions that were POS tagged as named entities, we decided to add an intermediate gran_instance class (for named entity participants that have no synsets such as person or organization names as John, or Doctors Without Borders) so that we can encourage number matching for our measurements of what granularity exclusively can contribute to event coreference. For agreement in semantic class level, two participant instances can maximally get 3 points. If there is 1 level difference between them (gran_person to gran_instance or gran_instance to gran_group) distance of 2 is determined. In case of participant pairs with gran_person and

[16] The Kybot system was developed in the FP7 KYOTO project but reimplemented for NewsReader. It can be downloaded from: git@github.com:cltl/KafKybot.git

[17] This work was done before the NewsReader pipeline was available. Now, the same process can be done directly on the NewsReader output described in Section 3

gran_group we have distance of 1. For number agreement we can maximally assign 2 points. If there is number disagreement, we assign 1 point. If there is both level type agreement as well as number agreement, a participant pair is given the maximum of 5 points. Since we aimed at measuring the influence of different event components on event coreference, we filter our action chains based on location and time compatibility. For locations and time expressions, very strict thresholds were used, to avoid matches as Monday and Tuesday, sharing a short path in the taxonomy and consequently a high L&C score. The same holds for the granularity and domain heuristics. This is why, for the time being, only lemma and synonym matches are used. In the future we will look into treating named entities differently, and apply similarity and granularity measurements to time expressions and locations that are not named entities. We will also consider employing geo and temporal ontologies containing named entities.

Heuristic	Event slot	MUC			B3			CEAF	BLANC			CoNLL
		R	P	F	R	P	F	F	R	P	F	F
LmB	All N&V	63.8	**82.8**	**71.2**	65.3	**90.6**	**75.0**	**65.9**	68.0	**84.1**	**71.1**	**70.7**
L&C	act.	**69.4**	72.4	69.5	**69.4**	73.3	68.9	58.7	**68.6**	71.8	67.5	65.2
act. L&C,time Lm	act. time	66.0	77.7	70.6	66.9	84.2	73.6	63.9	68.4	78.1	70.1	69.4
act. L&C, loc. Lm	act. loc.	66.3	77.4	70.6	67.4	83.0	73.4	64.1	**68.6**	77.3	70.0	69.3
act. L&C,part. Lm	act. part.	66.0	78.4	70.8	67.0	84.9	73.9	64.5	**68.6**	79.0	70.4	69.7
act. L&C, part. &C	act. part.	65.2	79.4	70.7	66.8	85.7	74.1	64.9	68.5	79.7	70.4	69.8
act. L&C, part.gran.	act. part.	66.5	77.8	70.4	67.6	81.7	72.2	62.5	68.3	77.9	69.4	68.2

Table 10: Coreference Evaluation on Cross-Document correference for the ECB data (all figures are macro averages). act.=action, part. = participant, loc. = location, gran. = granularity, LmB = lemma-baseline, L&C = Leacock & Chodorow

For the evaluation, the manual annotations of actions from the ECB corpus were used as key chains and were compared with the response chains generated for each topic by means of the above described heuristics. Since our goal was to evaluate the importance of coreference between other event components (than actions) for the task of event coreference resolution, we compare our evaluation results with system results based on action similarity only, i.e. when disregarding other event components. We also aimed at getting some insights into the contribution by shifts in hyponymy and granularity (soft matching). This is why we use a lemma baseline (LmB) that assigns coreference relation to all nouns and verbs that belong to the same lemma (strict matching). Table 10 presents coreference evaluation results achieved by means of the different heuristics: the L&C measure, granularity agreement as well as lemma match (Lm) in comparison to the baseline results (LmB) in terms of recall (R), precision (P) and F-score (F), employing the commonly used coreference evaluation metrics: MUC (Vilain *et al.* (1995)), B3 (Bagga and Baldwin (1998)), mention-based CEAF (Luo (2005)), BLANC (Recasens and Hovy (2011)), and CoNLL F1 (Pradhan *et al.* (2011)).

Compared to the lemma baseline, our approach using similarity of event actions only (second row in table 10), across majority of the evaluation metrics improves R with up to 6% while loses (2-17%) P, what is expected. As discussed in section 2, the baseline achieves remarkably good results. Within narrowly defined topics, such as news articles of the same day on a specific event, there is little variation and the same events are usually expressed by the same lemma (see above section 3). When comparing the contribution of participants, times and locations (all lemma matches for the sake of comparison) with the approach using exclusively action similarity, we see that the approach combining action and participant components achieved slightly better results (ca. 1% higher precision scores) than the two other approaches employing time and location slots. Altogether, the differences between the scores are in this case rather subtle. When analyzing these results one must keep in mind that these evaluation scores are conditioned by the fact that participant descriptions occur much more frequently in event descriptions than time and place markers. Out of the two different heuristics used in participant approaches; ca. 1% higher F-scores (a 2-4% improvement of precision) on most evaluation metrics were obtained with L&C similarity. Both participant approaches in most metrics improve the F-scores achieved by the action similarity heuristic; the granularity approach with ca. 1-4% and participant similarity with ca. 1-6%.

Compared to the lemma baseline (LmB), our best scoring approach of all (similarity with participant similarity) loses ca. 1% on F-score. It gains up to 2 points in recall, while generating output with ca. 4% lower precision. This small decline in F measure can be explained by the fact that we are dealing here with within topic coreference (although cross-document). Also, evaluation data seem to be biased towards coreference chains around smaller events. Evaluation corpora, including those annotated with cross-document coreference of events, (intentionally) tend to be composed around specific real world events, such as attacks or earthquakes, so that coreference chains are captured in a rather small time frame. The diversity of event instances from the same type of event class that happened in different time frames, places and with different participants is much lower in such a corpus than in realistic daily news streams. The relatively high scores achieved by the lemma baseline show the need for different event coreference datasets, where cross-document coreference is marked in text across different instances of particular event classes, e.g. describing two different wars that take place over longer stretches of time and include similar types of events. Only then the data will become more representative of the sampled population. We are currently extending the ECB corpus with more articles on events that belong to the same type, e.g. earthquakes and attacks, creating a more natural ambiguity for lemmas. For more details on the ECB+ corpus, see Cybulska and Vossen (2014).

For comparison, we give here to evaluation results achieved in related work as reported in the literature:

- Bejan and Harabagiu (2010): 83.8% B3 F, 76.7% CEAF F on the ACE (2005) data set and on the ECB corpus 90% B3 F, 86.5% CEAF F-score.

- Lee *et al.* (2012): 62.7% MUC, 67.7% B3 F, 33.9% (entity based) CEAF,71.7% BLANC F-score on the ECB corpus

- Che (2011): 46.91% B3 F on the OntoNotes 2.0 corpus by means of our best scoring approach, using action and participant similarity, coreference between actions was solved with an F-score of 70.7% MUC, 74.1% B3, 64.9% CEAFm, 70.4% BLANC F and 69.8 CoNLL F1.

Our lemma-baseline has F-measure between 65% and 75% (depending on the metrics), whereas the best results in the literature for the ECB by Bejan and Harabagiu (2010) are between 86% and 90%. It is not clear, if the high scores are due to the way singletons are treated, which can have a big impact on the scores. The cross-document results of Lee *et al.* (2012) and the results reported in section 4.2 for the baysian model are very similar to our baseline.

Considering that our approach neither considers anaphora resolution nor syntactic features, there is definitely room for improvement on event coreference resolution, including an approach that combines this problem with semantic matches of event components. For instance, the bayesian approach presented in the previous section performs better (cf. Table 8), and has the potential to incorporate different sources of knowledge which might be relevant to the task.

Conclusions: we have two different approaches that can be applied to obtain intra-document and inter-document coreference relations for events: one using a variety of structural and semantic features of mentions and one approach that reasons over event components. Both approaches can be combined by first creating coreferences on the basis of a baysian model using structural and semantic features of mentions and secondly reasoning over the components of these to refine or enlarge the initial sets. At any point, we can use the functions defined for the lemma-baseline to convert any set of coreferences to a SEM format that can be imported in to the KnowledgeStore.

Finally, it should be noted that we see coreference as a scalar notion. This means that we can tune thresholds to get coarse-grained or fine-grained coreference sets. This does not only result in lumping or splitting of data but we can also evaluate the effect in terms of semantic coherence and in terms of usability of the final user application.

5 Event Significance and Relevance

As explained in section 3, NewsReader generates massive amounts of events and relations, even at the instance level. Not all events are equally important and relevant within a news article but also from the perspective of the user to find a story. We define a story or narrative as a way of presenting events that are somehow connected through a plot. Some principles from plot theories in literature are very useful to model stories in news. The general view is that a plot structure always shows a development (rising action) towards some climax, after which there has to be a change or response (the falling action) and a final resolution. In news, we can see trendiness as the point of climax but there is also the explanation of how it came about (the rising action) and what the future perspective is (the falling action and resolution). According to Bremond (1966), Brooks (1992), Ryan (1991), plots can also be seen as schemas for human motivations and intentions of actions. These schemas further explain who was responsible for the climax event.

We are currently working out this model by translating properties of events and relations between them as features for the dramatic impact of an event. Dramatic impact can be defined by properties of the event itself or by any participant of the event. Events with participants that have impact, are automatically events with impact (e.g. anything Barack Obama does is important because Barack Obama is important), and the other way around, if the event has impact all participants will have impact from that moment on (e.g. an insignificant person involved in a dramatic disaster such as 9/11 inherits the impact from the event for the rest of his/her life). Measurable features for measuring the impact can be the following:

- trendiness: persons and events that frequently occur in the news, as reflected by the number of mentions and number of different news articles in which they are mentioned;
- strength of opinions on participants: sentiment analysis in social media on persons and events can be used to establish the arousal;
- role and function (events involving a decision maker with power, such as a CEO or President are important);
- past: participants with a 'backpack', involved in a previous event with impact, will carry this over to any new event;
- type of event with cultural impact status, such as wars, killings, disasters, scandals, fraud, corruption, bankruptcies;
- having impact on (many) socially weak and vulnerable people;
- states of important factors that develop towards critical values or show significant unexpected changes, e.g oil price, price of wheat, market shares, monopolies;

Scoring for these aspects can result in an overall relevance score for events and participants. The plot model can then be used to connect events that may be less relevant at first sight, to the events with impact because they fit some narrative plot or explanatory scheme. For example, a series of increases in the price of wheat in a period may seem insignificant at the time being but in the end result in a critical situation that forms a climax. The above model will be implemented an tested in the second year of the project.

Currently, our representation of text in NAF allows for a basic differentiation of events and participants in terms of the following aspects:

- the form of the mention of the event
- the type of event
- factuality of events
- provenance of the event

Mentions of events can have different structures or **forms**, as shown in the next examples:

- After a boom on the stock market that enticed many everyday people to invest their entire savings, the stock market crashed on October 29, 1929
- Sebi probing possible foul play in crashing of stock markets.
- Which was the reason for the crash of stock markets in India that year
- The Wall Street Crash of 1929, also known as Black Tuesday and the Stock Market Crash of 1929, began in late October 1929 and was the most devastating

Events can be expressed by the semantic main verb of a clause, a nomalization of a verb, a noun referring to an event and named events (Segers *et al.* (2011)). Reference by the main verb or clause is found in direct reporting styles, in which a lot of details are given on the participants through the syntactic arguments such as the subject and direct object. This by itself does not mark an event as important or relevant. If we nominalize an event or use a noun to refer to an event, this mean we start to talk about an event as a *thing*. Nominal reference is used to state something about an event, such as an opinion or some implication. This can be seen as a marking of importance of an event. Finally, the fact that names are given to events means that they had big impact. By giving an event a name, we give them a similar status as instances of people and objects in our world. By detecting the structure of the mentions and measuring the frequency of particular formal ways of mentions can thus indicate relevance of the event itself. The more an event is referred to with a name, the more important it is.

The second criterion, relates to the semantic type of the event. Currently, we distinguish 3 types of events in NewsReader:

1. grammatical events that do not represent instances of events directly but express properties of events or relations between events (e.g. aspectual, tense or causal relations).

2. speech acts or cognitive events that introduce sources that may be seen as provenance relations or as expressions of opinions.

3. contextual events that usually describe the actual changes in the world

To differentiate between these classes, we compiled a list of events that occur most frequently with a subject-verb or object-verb dependency in a domain set of 500 news articles. These articles were selected for their reference to a car company. The most frequent occurring verbs were manually checked as expressing a grammatical relation, a speech act or a cognitive event. The list of grammatical and speech act/cognitive expressions was used to type the mentions of events in the car data set. All event mentions outside this list are considered to be contextual. Table 11 shows the distribution of these types on the car industry data set.

YEAR	grammatical	speech-cognitive	contextual	other
2003	11,207	14,203	11785	43,328
2004	11,264	14,090	11,868	43,265
2005	11,101	14,197	11,511	41,964
2006	12,575	16,565	13,342	49,298
2007	12,632	17,324	13,367	49,713
2008	12,282	15,577	12,154	45,610
2009	14,094	19,728	14,595	56,010
2010	11,334	15,015	11,535	43,860
2011	10,094	13,037	10,243	37,699
2012	14,716	20,391	15,701	58,183
2013	4,679	6,565	4,828	18,422
TOTAL	125,982	166,699	130,935	487,367

Table 11: Differentiation of events in the car industry data sets for type of event

From the total set of events, the majority is contextual (53%). They represent the set of the most relevant events in the data. About 32% is a speech act or cognitive verb, whose subject can be seen as a source and the complement may contain a contextual event about some change in the world. They are mostly important as far as they can add to the provenance layer of the project. The grammatical verbs represent about 14% of the data and are most likely not relevant.

For all the contextual events, we also have a score for the factuality of the event per mention in NAF. A low factuality score is either based on the future tense of the main clause or it is the result of negation or uncertainty markers for events expressed in the present or past tense. By combining the event type with the factuality, we can differentiate the events in terms of factuality. Finally, the number of mentions of an event in the sources can be used as an indication or relevance but, more importantly, the number of sources confirming that instance of an event or a relation and possibly the type of sources as more precise provenance information on the relevance of the event. This information is now available in NAF and the SEM representation that we derive. In the second year of the project, we will translate this information to provenance and relevance values in the SEM representation so that they can be exploited more directly by the tools that access the data in the Knowledge Store.

6 Conclusions

This deliverable described the first results on modeling events. It extracts instance- of events and entities in a formal semantic representation from textual descriptions, according to the Grounded-Annotation-Framework developed in the project. Every instance of an event and entity and every relation gets a unique identifier and is linked to all the place in the texts where they are mentioned. Coreference is the first important step to get from a presentation of mentions in text to a semantic representation of instances. Once coreference has been established, we can decide on the relations between events and the (re-)construction of longer story lines of events. Deciding on event relations and story lines is planned for the second year of the project.

The prototype clusters co-referencing event mentions, within and across documents, and outputs a unique list of event instances, merging information from different mentions. The prototype also produces a relevance ranking and selection of event instances, aggregating the information produced in WP4 per mention. We defined a multi-stage approach for establishing event-coreference that is further described in this deliverable:

1. Structural approach for intra-document mentions

2. Structural approach for inter-document instances within a tight temporal and topic cluster

3. A semantic approach for inter-document instances for more loose clusters of documents and across longer periods of time

We reimplemented a state-of-the-art Bayesian approach to intra-document and cross-document event-coreference using descriptional properties Bejan and Harabagiu (2010), and a lemma-baseline (matching events within a topic solely on the basis of the same lemma) that scores only 10% lower in F-measure and can easily be improved using simple heuristics for anaphora resolution and syntactic relations Cybulska and Vossen (2013).

The lemma-based intra-document and cross-document coreference module that has been applied to two data sets:

- 63,811 English news articles provided by Lexis Nexis, on the car industry and published between 2003 and 2013

- 43,384 articles from the TechCrunch database with news about IT companies registered in Crunchbase

This processing resulted in a SEM representation for events, participants and their time points and place. The data were imported in the Knowledge Store developed in Work Package 6. We also describe the preliminary ideas on deciding on the relevance and significance of the event data that is extracted.

In the second year, the work on T05.1 Event Merging and Chaining will focus on improving the results on event co-reference for English, the extension to additional co-reference relations (subclass and meronymy), as well as to other languages and to cross-lingual co-reference relations. In addition, relations between event mentions will be derived. Especially, we will focus on historical event-coreference in which the news of a day is related to the news from the past as stored in the Knowledge Store.

The work on T05.2 Event Significance and Relevance will be completed, and information from narrative graphs and background models will be incorporated.

Tasks T05.3 and T05.4 will be initiated in the second year. Firstly, T05.4 Building Domain Model for Financial and Economic Events will produce a background model for the domain, based on the corpora gathered in the first year. Secondly, T05.3 Extraction of Narrative Graphs will induce the narrative stories (sequences of events) that are of relevance in the domain.

References

Rodrigo Agerri, Itziar Aldabe, Zuhaitz Beloki, Egoitz Laparra, Maddalen Lopez de Lacalle, German Rigau, Aitor Soroa, Marieke van Erp, Piek Vossen, Christian Girardi, and Sara Tonelli. Event detection, version 1. NewsReader Deliverable 4.2.1, 2013.

Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *Proceedings of LREC*, 1998.

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

Cosmin Adrian Bejan and Sanda M. Harabagiu. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010.

Cosmin Adrian Bejan, Matthew Titsworth, Andrew Hickl, and Sanda M. Harabagiu. Nonparametric bayesian models for unsupervised event coreference resolution. In *23rd Annual Conference on Neural Information Processing Systems (NIPS)*, 2009.

Zuhaitz Beloki, German Rigau, Aitor Soroa, Antske Fokkens, Piek Vossen, Marco Rospocher, Francesco Corcoglioniti, Roldano Cattoni, Thomas Ploeger, and Willem Robert van Hage. System design. NewsReader Deliverable 2.1, 2014.

Claude Bremond. The logic of narrative possibilities. *New Literary History*, 11:387–411, 1966.

Peter Brooks. *Reading for the Plot: Design and Intention in Narrative.* Harvard University Press, Cambridge, Mass, 1992.

A Unified Event Coreference Resolution by Integrating Multiple Resolvers, 2011.

Agata Cybulska and Piek Vossen. Event models for historical perspectives: Determining relations be-tween high and low level events in text, based on the classification of time, location and participants. In *Proceedings of LREC 2010, Valletta, Malta, May 17-23*, 2010.

Agata Cybulska and Piek Vossen. Semantic relations between events and their time, locations and participants for event coreference resolution. In *Proceedings of Recent Advances in Natural Language Processing (RANLP-2013)*, pages 156–163, 2013.

Agata Cybulska and Piek Vossen. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of LREC-2014*, 2014.

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

Thomas S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

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

Aria Haghighi and Dan Klein. Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007.

Eduard Hovy, Teruko Mitamura, Felisa Verdejo, and Andrew Philpot. Identity and quasi-identity re-lations for event coreference.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. Extending verbnet with novel verb classes. In *Fifth International Conference on Language Resources and Evaluation*, 2006.

Claudia Leacock and Martin Chodorow. Combining local context with wordnet similarity for word sense identification, 1998.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Sur-deanu, and Dan Jurafsky. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP-CoNLL)*, 2012.

Xiaoqiang Luo. On coreference resolution perfor-mance metrics. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. The nombank project: An interim report. In A. Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

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

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. Conll-2011 shared task: Modeling unre-stricted coreference in ontonotes. In *Proceedings of CoNLL 2011: Shared Task*, 2011.

Willard V. Quine. Events and reification. In *Actions and Events: Perspectives on the Philosophy of Davidson*, pages 162–71. Blackwell, 1985.

Marta Recasens and Eduard Hovy. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 4(17):485–510, 2011.

Marco Rospocher, Francesco Corcoglioniti, Roldano Cattoni, Bernardo Magnini, and Luciano Serafini. Interlinking unstructured and structured knowledge in an integrated framework. In *Proc. of 7th IEEE International Conference on Semantic Computing (ICSC), Irvine, CA, USA*, 2013. (to appear).

Marie-Laure Ryan. *Possible Worlds, Artificial Intelligence and Narrative Theory*. Bloomington: Indian University Press, 1991.

Roxane Segers, Marieke Van Erp, Lourens van der Meij, Lora Aroyo, Guus Schreiber, Bob Wielinga, J van Ossenbruggen, Johan Oomen, and Geertje Jacobs. Hacking history: Automatic historical event extraction for enriching cultural heritage multimedia collections. *Proceedings of the 6th International Conference on Knowledge Capture (K-CAP'11)*, 2011.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

Willem Robert van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. Design and use of the Simple Event Model (SEM). *J. Web Sem.*, 9(2):128–136, 2011. `http://dx.doi.org/10.1016/j.websem.2011.03.003`.

Marc Vilain, John Burger, John Aberdeen, Connolly Dennis, and Lynette Hirschman. A model theoretic coreference scoring scheme. In *Proceedings of MUC-6*, 1995.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. Ace 2005 multilingual training corpus, 2006.