Multi-lingual and Cross-lingual TimeLine Extraction

Egoitz Laparra*, Rodrigo Agerri*, Itziar Aldabe*, German Rigau*

IXA NLP group, Computer Science Faculty UPV/EHU, Manuel Lardizabal 1, 20008 - Donostia, Basque Country

Abstract

In this paper we present an approach to extract ordered timelines of events, their participants, locations and times from a set of multilingual and cross-lingual data sources. Based on the assumption that event-related information can be recovered from different documents written in different languages, we extend the Cross-document Event Ordering task presented at SemEval 2015 by specifying two new tasks for, respectively, Multilingual and Cross-lingual Timeline Extraction. We then develop three deterministic algorithms for timeline extraction based on two main ideas. First, we address implicit temporal relations at document level since explicit time-anchors are too scarce to build a wide coverage timeline extraction system. Second, we leverage several multilingual resources to obtain a single, interoperable, semantic representation of events across documents and across languages. The result is a highly competitive system that strongly outperforms the current state-of-the-art. Nonetheless, further analysis of the results reveals that linking the event mentions with their target entities and time-anchors remains a difficult challenge. The systems, resources and scorers are freely available to facilitate its use and guarantee the reproducibility of results.

Keywords: Timeline extraction, Event ordering, Temporal processing, Cross-document event coreference, Predicate Matrix, Natural Language Processing

1. Introduction

Nowadays, Natural Language Processing (NLP) may help professionals to access high quality, structured knowledge extracted from large amounts of unstructured, noisy, and multilingual textual sources (Vossen et al., 2016). As the knowledge required is usually equivalent to reconstructing a chain of previous events, building timelines constitutes an efficient and convenient manner of structuring the extracted knowledge. However, yielding timelines is a high level

*Corresponding author
Email address: egoitz.laparra@ehu.eus (Egoitz Laparra)

task that involves information extraction at multiple tiers, including named entities, events or time expressions. Furthermore, it should also be considered that the information required to construct the timeline must be gathered from different parts of a document, or even from different documents. Thus, coreferential mentions of entities and events must be properly identified.

For example, a named entity can be mentioned using a great variety of surface forms (Barack Obama, President Obama, Mr. Obama, Obama, etc.) and the same surface form can refer to a variety of named entities [1]. Furthermore, it is possible to refer to a named entity by means of anaphoric pronouns and co-referent nominal expressions such as 'he', 'her', 'their', 'I', 'the 35 year old', etc. The same applies to event mentions, which can be verbal predicates or verbal nominalizations. Thus, the following two sentences contain different mentions of the same event, namely, that a gas pipe exploded, via the two different predicates 'exploded' and 'blast'. Furthermore, while Example (1) allows us to explicitly time-anchor the event via the temporal expression 'yesterday', that does not occur in the second example. In this context, building a timeline amounts to detecting and temporal ordering and anchoring the events in which a target named entity participates.

(1) A leak was the apparent cause of yesterday's gas *blast* in central London.

(2) A gas pipe accidentally *exploded* in central London. Only material damage was reported.

Several tasks from SemEval (Verhagen et al., 2007, 2010; UzZaman et al., 2013; Llorens et al., 2015) and other recent challenges as the 6th i2b2 NLP Challenge (Sun et al., 2013) have focused on temporal relation extraction. In these tasks, systems should detect events and time-expression as well as the temporal relations between them, discovering what events occur before, after or simultaneously with respect to others.

More recently, the task 4 of SemEval 2015 (Minard et al., 2015) proposed some novel differences regarding temporal information extraction. The goal of this task is to build timelines of event involving a target entity. The events belonging to a timeline must be recovered across documents and sort according to their time anchors. Thus, Semeval 2015 task 4 requires a more complete time anchoring than previous challenges.

We base this work on the SemEval 2015 Timeline extraction task to present a system and framework to perform Multilingual and Cross-lingual Timeline Extraction. This is based on the assumption that timelines and events can be recovered from a variety of data sources across documents and across languages. In doing so, this paper presents a number of novel contributions.

Contributions. The original Cross-document event ordering task defined for SemEval 2015 (main Track A) is extended to present two novel tasks for two

[1] For example, see the Wikipedia disambiguation page for 'Europe': `http://en.wikipedia.org/wiki/Europe_(disambiguation)`

languages (English and Spanish) on Multilingual and Cross-lingual timeline extraction, respectively. The tasks also generated publicly available annotated datasets for trial and evaluation. Additionally, two new evaluation metrics improve the evaluation methodology of the SemEval 2015 task to address both the multilingual and cross-lingual settings.

Interestingly, we also show that extracting just the temporal relations that explicitly connect events and time expressions produces incomplete timelines. We propose a method to discover implicit temporal relations that works at a document level and proves to obtain a more complete time-anchoring annotation. Furthermore, we show how to effectively leverage multilingual resources such as the PredicateMatrix (López de Lacalle et al., 2014) and DBpedia[2] to improve the performance in a more realistic setting of building cross-lingual timelines when no parallel data as input is available. We present a deterministic approach that obtains, by far, the best results on the main Track A of SemEval 2015 task 4. Our deterministic approach is fledged out via three different timeline extraction systems which extend an initial version presented in Laparra et al. (2015), including an adaptation of this system for Spanish and to allow the cross-lingual timeline extraction. To guarantee reproducibility of results we also make publicly available the systems, datasets and scripts used to perform the evaluations [3].

Next section reviews related work, focusing on the SemEval 2015 Timeline extraction task. Next, Section 3 describes the two new Cross-lingual and Multilingual Timeline extraction tasks. The construction of the datasets for the new tasks occupies Section 4 and Section 5 formulates the evaluation methodology employed in this work. In section 7 we report the evaluation results obtained by the systems previously presented in Section 6. Finally, Section 8 provides an error analysis to discuss the results and contributions of our approach while Section 9 highlights the main aspects of our work and future directions.

2. Related work

The present work is directly related to the SemEval 2015 task 4, Timeline: Cross-document event ordering (Minard et al., 2015). Its aim is to combine temporal processing and event coreference resolution to extract from a collection of documents a set of timelines of events pertaining to a specific target entity. The notion of event is based on the TimeML definition, namely, an event is considered to be a term that describes a situation or a state or circumstance that can be held as true (Pustejovsky et al., 2003b).

In fact, the Timeline extraction task is in turn quite close to the TempEval campaigns (Verhagen et al., 2007, 2010; UzZaman et al., 2013; Llorens et al., 2015). Briefly, the problem is formulated as a classification task to decide the type of temporal link that connects two different events or an event and a

[2] http://wiki.dbpedia.org/.

[3] http://adimen.si.ehu.es/web/CrossTimeLines

temporal expression. For that reason, supervised techniques have been the main approaches to solve the task. For example, Mani et al. (2006, 2007) trained a MaxEnt classifier using bootstrapped training data that was obtained applying temporal closure. Chambers et al. (2007) focused on event-event relations using the attributes learned from previous events. More recently, DŚouza and Ng (2013) proposed a combination of hand-crafted rules and semantic and discourse features. Laokulrat et al. (2013) obtained the best results in TempEval 2013 using predicate-role annotations, while Mirza and Tonelli (2014) described a set of simple features and proved to obtain better performances. Other recent works such as Chambers et al. (2014) have pointed out that these tasks cover just a part of all the temporal relations that can be inferred from the documents.

The SemEval 2015 timeline extraction task proposed two tracks, depending on the type of data used as input. The main track A for which only raw text sources were provided, and Track B, where gold event mentions were also annotated. For each of the two tracks a sub-track was also proposed in which the assignment of time anchoring was not taken into account for the evaluation. No training data was provided for any of the tracks.

Track A received three runs from two participants: the WHUNLP and SPINOZAVU teams. Both approaches were based on applying a pipeline of linguistic processors including Named Entity Recognition, Event and Nominal Coreference Resolution, Named Entity Disambiguation, and temporal processing (Minard et al., 2015). The SPINOZAVU system was further developed in Caselli et al. (2015). The results in this track proved the difficulty of the task. Besides that the chain of errors produced by the individual modules affected their final performance, these systems failed specially to anchor every event to a time expression.

The Track B approaches, represented by the two participants HEIDELTOUL and GPLSIUA, substantially differ from those of Track A because the event mentions pertaining to the target entity are already provided as gold annotations. Therefore, those systems focused on event coreference resolution and temporal processing (Minard et al., 2015). Two recent works have been recently published on Track B: an extension of the GPLSIUA system (Navarro-Colorado and Saquete, 2016), and a distant supervision approach using joint inference (Cornegruta and Vlachos, 2016). As all these systems depend on the event gold annotations, they cannot be directly applied in the Track A.

Track A is, in our opinion, the most realistic scenario as systems are provided a collection of raw text documents and their task is to extract the timeline of events for each of the target entities. More specifically, the input provided is a set of documents and a set of target entities (organization, people, product or financial entity) while the output should consist of one timeline (events, time anchors and event order) for each target entity.

Compared to previous works on Track A of the SemEval 2015 Timeline extraction task, our approach differs in several important ways. Firstly, it addresses the extraction of implicit information to provide a better time-anchoring (Palmer et al., 1986; Whittemore et al., 1991; Tetreault, 2002). More specifically, we are inspired by recent works on Implicit Semantic Role Labelling

(ISRL) (Gerber and Chai, 2012) and, specially, on Blanco and Moldovan (2014) who adapted ISRL to focus on modifiers, including temporal arguments, instead of core arguments or roles. Given that no training data is provided, we developed a deterministic algorithm for timeline extraction loosely inspired by Laparra and Rigau (2013). Secondly, we extend the monolingual approach to make it multi- and cross-lingual, which constitutes a novel system on its own. Finally, our approach outperforms every other previous approach on the task, almost doubling the score of the next best system.

3. Multilingual and Cross-lingual Timeline Extraction

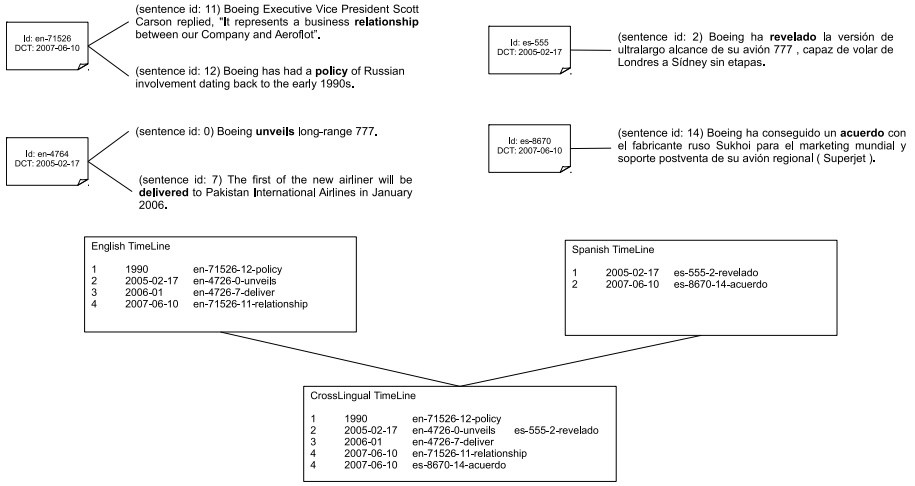

Figure 1: Example of multilingual and cross-lingual timelines for the target entity *Boeing*.

The Timeline Extraction definition was formulated as follows: "Given a set of documents and a target entity, the task is to build an event timeline related to that entity, i.e. to detect, anchor in time and order the events involving the target entity" (Minard et al., 2015). As we have already mentioned in the previous section, in this work we will focus on Track A (main track), which is the most demanding and realistic setting of the two: systems are given a set of raw text documents and the task is to extract the timelines. Furthermore, we provide two novel extensions to the original task:

- **Multilingual Timeline Extraction**: This task straightforwardly extends the SemEval 2015 task to cover new languages. Thus, a parallel set of documents and a set of target entities, common to all languages, are provided. The goal is to obtain a timeline for each target entity in each language independently.

- **Cross-lingual Timeline Extraction**: For this task, the timelines are built from source data in different languages identifying those event mentions that are coreferent across languages. However, unlike in the multilingual setting, every document in every language is considered together so that *a single cross-lingual timeline* is expected for each of the target entities.

These two new tasks are presented here for two languages, namely, English and Spanish. Figure 1 shows an example of both multilingual and cross-lingual timelines for the target entity *Boeing*. The left-hand side column corresponds to an English timeline extracted from four sentences in two different English documents. On the right-hand side is shown an Spanish timeline obtained from two sentences contained in two different documents. Words in bold refer to the event mentions that compose the timeline. Finally, the box in the bottom depicts a cross-lingual timeline built from sources in both English and Spanish. Coreferent events across languages, such as *unveils* and *revelado*, are annotated in the same row, while events that are simultaneous but are no coreferent appear in different rows. The events *relationship* and *acuerdo* (in the last two rows) provide such an example. The following section describes in more detail the procedure used to build the datasets for both the Multilingual and Cross-lingual Timeline Extraction tasks.

4. Data Annotation

In the original Timeline Extraction task at SemEval 2015 (Minard et al., 2015), the dataset was extracted from the raw text of the English side of the MeanTime corpus (Minard et al., 2016). Given that MeanTime is a parallel corpus that includes manual translations from English to Spanish, Italian and Dutch, it is straightforward to use its Spanish part for the Multilingual and Cross-lingual Timeline Extraction tasks.

4.1. Creation of multilingual and cross-lingual timelines.

In order to better understand the procedure to create the datasets for the multilingual and cross-lingual settings, a brief overview of the original annotation to create the gold standard timelines for English is provided. For full details of the original annotation, please check the SemEval 2015 task description (Minard et al., 2015). As already mentioned, the input to the task consisted of the target entities, the event mentions and the time anchors. In the following, each of these three aspects are described.

Target Entities. A set of target entities were selected that belong to type PERSON (*e.g. Steve Jobs*), ORGANISATION (*e.g. Apple Inc.*), PRODUCT (*e.g. Airbus A380*), and FINANCIAL (*e.g. Nasdaq*). The target entities must appear in at least two different documents and be involved in more than two events.

Events. The annotation of events was restricted by limiting the annotation to events that could be placed on a timeline. Adjectival events, cognitive events, counter-factual events, uncertain events and grammatical events were not annotated. Furthermore, timelines only contain events in which target entities explicitly participate as *Agent* or *Patient*.

Time anchors. A time anchor corresponds to a TIMEX3 of type DATE (Pustejovsky et al., 2003a). Its format follows the ISO-8601 standard: YYYY-MM-DD (i.e. Year, Month, and Day). The finest granularity for time anchor values is DAY; other granularities admitted are MONTH and YEAR (references to months are specified as YYYY-MM and references to years are expressed as YYYY).

As Minard et al. (2015) explain, once the corpus has been annotated with the required linguistic layers (entities, events and time anchors), the gold-standard timelines are automatically created by ordering the events according to their time anchors. A manual revision is performed afterwards so that events with the same time anchor are ordered based on textual information. Minard et al. (2015) computed inter annotator agreement using the Dice's coefficient (Dice, 1945) and data from three annotators. For entity and event mentions, they obtained 0.81 and 0.66 respectively and for entity coreferences 0.84.

Creation of multilingual timelines. The process described above was followed to create timelines in Spanish. In both cases, English and Spanish, timelines are represented in tabulated format. Each row contains one event representing an instance of an event occurring at a specific time. The first column of each row indicates the position of the event in the timeline. The second column specifies the time-anchor of the event. Additional columns in the row, if any, refer to the different mentions of that event in the dataset. Each event mention is identified with the document identifier, the sentence number and the textual extent of the mention. The document identifier is in turn composed of a prefix specifying the language in which the document is written and its numerical identifier. If two events have the same time-anchor but they are not coreferent, they are placed on different rows. An example of multilingual annotations for English and Spanish is provided by Figure 1.

Creation of cross-lingual timelines. To create the gold-standard timelines for the cross-lingual task, we automatically cross the manual annotations from the English and Spanish parallel corpora. The resulting timelines have the same format as the original ones. More specifically, when two mentions of the same event in two different languages refer to the same event then they are included in the same row. The automatic mapping of annotations to construct the cross-lingual timelines was manually revised. A brief example of a cross-lingual dataset is illustrated by the box at the bottom of Figure 1.

		Trial	Test			
		Apple Inc.	Airbus	GM	Stock	Total
English (SemEval-2015)	# documents	30	30	30	30	90
	# sentences	463	446	430	459	1,335
	# tokens	10,343	9,909	10,058	9,916	29,893
	# event mentions	178	268	213	276	757
	# event instances	165	181	173	231	585
	# target entities	6	13	12	13	38
	# timelines	6	13	11	13	37
	# event mentions / timeline	29.7	20.6	19.4	21.2	20.5
	# event instances / timeline	27.5	13.9	15.7	17.8	15.8
	# docs / timeline	5.7	5.2	4.1	9.1	6.2
Spanish	# documents	30	30	30	30	90
	# sentences	454	445	431	467	1,343
	# tokens	10,865	10,989	11,058	11,341	33,388
	# event mentions	187	222	195	244	661
	# event instances	149	163	147	212	522
	# target entities	6	13	12	13	38
	# timelines	6	13	11	13	37
	# event mentions / timeline	31.2	17.1	17.7	18.8	17.9
	# event instances / timeline	24.8	12.5	13.4	16.3	14.0
	# docs / timeline	5.5	4.8	3.7	8.5	5.8
Cross-lingual	# documents	60	60	60	60	180
	# sentences	917	891	861	926	2,678
	# tokens	21,208	20,898	21,116	21,257	63,271
	# events mentions	364	490	408	520	1,418
	# event instance	165	181	174	231	586
	# target entities	6	13	12	13	38
	# timelines	6	13	11	13	37
	# events / timeline	60.7	37.7	37.1	40.0	38.3
	# event chains / timeline	27.5	13.9	15.8	16.2	15.8
	# docs / timeline	11.5	10.0	8.2	17.6	12.1

Table 1: Counts extracted from the Multilingual and Cross-lingual gold datasets.

4.2. Task dataset

The English dataset released for the SemEval 2015 Timeline extraction task consists of 120 Wikinews[4] articles containing 44 target entities. The Wikinews articles are focused mostly on four main topics, 30 documents per topic. A split of 30 documents and 6 target entities (each associated to a timeline) are provided as trial data, while the rest is left as evaluation set: 90 documents and 38 target entities (each associated to a timeline). Similarly, the Spanish

[4] `http://en.wikinews.org`

dataset also contains 120 articles with 44 entities. The trial and test splits for this language are the same as in the English dataset. On the other hand, as the cross-lingual dataset arises from joining the English and Spanish datasets, it contains 240 articles containing same 44 target entities as in the English and Spanish datasets. In this case, the trial split includes 60 documents and 6 target entities while the test set contains the remaining 180 documents and 38 target entities. For all the cases, the trial data contains one ORGANISATION (*Beatles Apple corps.*) target entity, one PERSON (*Steve Jobs*), and 4 PRODUCT entities (*iPhone 3g, iPhone 4, iPod, iTunes*). With respect to the evaluation set, 18 entities are ORGANISATION (*Boeing, General Motors, Bank of America, ...*), 10 FINANCIAL (*CAC 40, Nasdaq Composite, Nikkei 225 ...*), 7 PERSON (*Louis Gallois, Barack Obama, Jim Press...*), and 3 of the PRODUCT class (*Airbus a380, Boeing 777, Boeing 787 dreamliner*). The four topics are the following: (i) Apple Inc. for the trial corpus; (ii) Airbus and Boeing; (iii) General Motors, Chrysler and Ford; and (iv) Stock Market.

Table 1 provides some more details about the datasets. It should be noted that although there are 38 target entities, 37 were used for the evaluation because one timeline contained no events. Furthermore, although the three evaluation corpora are quite similar, the timelines created from the Stock Market corpus contain a higher average number of events with respect to those created from the other corpora. Additionally, it can also be seen that the Stock Market timelines contain events from a higher number of different documents. It should also be noticed that although the English and Spanish corpora are parallel translations, the number of event instances and mentions in both cases are not exactly the same. This is due to the fact that some of the events from the English corpus cannot be expressed in Spanish with events that comply with the restrictions explained in Section 4.1. For example, in the sentence *"The iPhone 4 is slated for a U.S. release on June 24."*, *slated* can be included as an event mention in the *iPhone 4* timeline because *iPhone 4* is the *Object* of *slated*. However, in the corresponding translated sentence *"El lanzamiento en Estados Unidos del iPhone 4 está previsto para el 24 de junio."*, *iPhone 4* is not a participant role of *previsto*, i.e. the corresponding translation of *slated*.

5. Evaluation Methodology

The evaluation methodology proposed in SemEval 2015 was based on the evaluation metric used for TempEval3 (UzZaman et al., 2013). The metric aims at capturing the temporal awareness of an annotation by checking the identification and categorization of temporal relations. In order to do this, UzZaman et al. (2013) compare the graph formed by the relations given by a system ($Sys_{relation}$) and the graph of the reference (gold standard) annotations ($Ref_{relation}$). From these graphs, their closures ($Sys^{+}_{relation}$, $Ref^{+}_{relation}$) and reduced forms ($Sys^{-}_{relation}$, $Ref^{-}_{relation}$) are obtained. The reduced form is created by removing redundant relations (those that can be inferred from other relations) from the original graph.

At the original SemEval 2015 task, the following steps were proposed to transform the timelines into graphs of temporal relations:

1. Every time anchor is represented as a TIMEX3.
2. Each event is related to one TIMEX3 by means of the SIMULTANEOUS relation type.
3. If one event occurs before another one, a BEFORE relation type is created between both events.
4. If one event occurs at the same time as other event, a SIMULTANEOUS relation type links both events.

These steps are followed to obtain both $Sys_{relation}$ and $Ref_{relation}$. Figure 2 shows the resulting graph after applying these four steps to the cross-lingual timeline in Figure 1. The doted lines represent the implicit relations that will be part of the closure ($Sys^{+}_{relation}$, $Ref^{+}_{relation}$), while the grey lines represent the redundant relations absent in the reduced graph ($Sys^{-}_{relation}$, $Ref^{-}_{relation}$). For example, the SIMULTANEOUS relation between *en-unveils* and *es-revelado* can be inferred from the fact that both events are linked to the same TIMEX3 anchor via a SIMULTANEOUS relation.

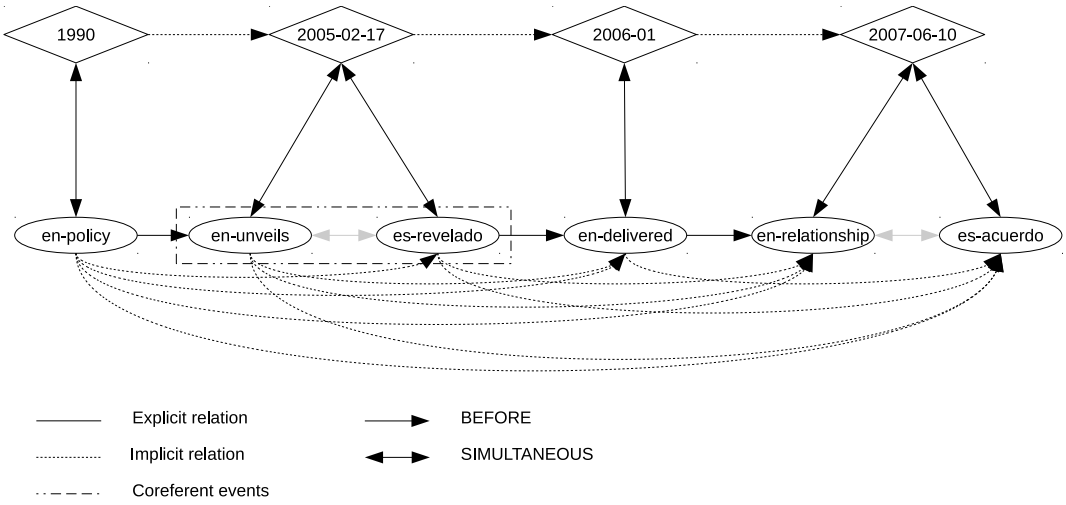

Figure 2: Time graph produced by original SemEval 2015 evaluation. Grey lines represent redundant relations.

In this setting, once the graphs representing the timelines are obtained and their closures and reduced forms derived, Precision and Recall metrics are calculated as follows:

$$Precision = \frac{|Sys^{-}_{relation} \cap Ref^{+}_{relation}|}{|Sys^{-}_{relation}|}$$

$$Recall = \frac{|Ref^{-}_{relation} \cap Sys^{+}_{relation}|}{|Ref^{-}_{relation}|}$$

Precision is calculated by counting the number of relations in the reduced system graph ($Sys^{-}_{relation}$) that can be found in the closure reference graph ($Ref^{+}_{relation}$) out of total number of relations in the reduced system graph ($Sys^{-}_{relation}$). Recall corresponds to the number of relations in the reduced reference graph ($Ref^{-}_{relation}$) that can be verified from the closure system graph ($Sys^{+}_{relation}$) out of the total number of relations in the reduced reference graph ($Ref^{-}_{relation}$). Final scores are based on the micro-average of the individual F_1 scores for each timeline, namely, the scores are averaged over the events of the timelines of each corpus. The micro-averaged precision and recall values are also provided.

However, it is important to note that this evaluation method does not distinguish coreferent events, namely, mentions of the same event, from those that simply occur at the same time (simultaneous). In this sense, in Figure 2, the same SIMULTANEOUS relation is used to connect two *coreferent events* such as *en-unveils* and *es-revelado*, and two events *en-relationship* and *es-acuerdo*, that simply occur at the same time (e.g., they are not coreferent). Hence, while this methodology is sufficient to check the temporal ordering of events, it is not adequate for cross-lingual timeline extraction, because it is crucial to identify that two event mentions refer to the same event across languages. In order to address this issue, this paper extends the original evaluation method from the Timeline Extraction SemEval 2015 task and proposes two alternative scoring methods:

- A **strict** evaluation where every single mention of every event is expected to be recovered, grouping adequately coreferent events. Thus, this evaluation demands good performance in both crosslingual event coreference and language-specific timeline extraction.

- A **relaxed** evaluation that assumes that a timeline can be complete if all the event instances are extracted even when not all the mentions of those instances are recovered. This way, we can test if a system is able to obtain more instances combining different languages than working just with a single one. In any case, the evaluation penalizes coreferent events that are identified as different instances.

In the following, we explain in detail how the graphs of temporal relations are built for each of these two evaluation methods.

5.1. Strict evaluation

In the strict evaluation method a timeline must contain every mention of the events that can be found in the document set. Moreover, event mentions referring to the same event should be identified and distinguished from those that simply occur at the same time. With this aim in mind, the following changes are proposed:

- Coreferent events are not linked via the SIMULTANEOUS relation but by means of a new IDENTITY relation.

- The IDENTITY relations are never removed from the reduced graphs. They are not redundant.

The strict temporal graph depicted in Figure 3 shows the graph obtained applying our new methodology. Whereas in the original graph in Figure 2 the coreferent events *en-unveils* and *es-revelado* are linked by a redundant SIMULTANEOUS relation, in Figure 3 a non-redundant IDENTITY relation links those two events.

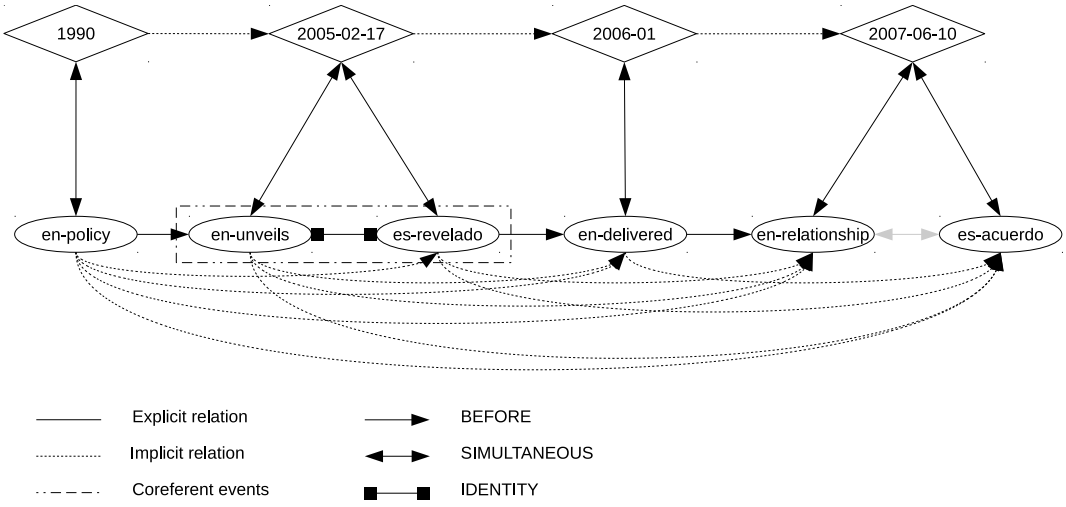

Figure 3: Time graph produced by *Strict evaluation*. Grey lines represent redundant relations.

Note that this method is more demanding in terms of precision because it adds the extra difficulty of distinguishing between IDENTITY and SIMULTANEOUS relations. Moreover, the set of temporal relations that must be captured is larger because the IDENTITY relations will not be removed when producing the reduced graphs. Thus, this also makes the task more demanding in terms of recall. That is why this evaluation method is named **strict evaluation**.

5.2. Relaxed evaluation

A second alternative stems from considering that, instead of using every event mention, a timeline could be composed of event types. Thus, coreferent events would be grouped as a single event by removing their temporal relations. The following changes are then performed with respect to the original SemEval 2015 evaluation:

- Every relation between coreferent events is removed.
- All the SIMULTANEOUS relations between coreferent events and a TIMEX3 anchor are reduced to a single relation.

These changes are explicitly shown by Figure 4. It can be seen that there is no relation linking the *en-unveils* and *es-revelado* coreferent events. Furthermore, the SIMULTANEOUS relations that connected those event with their TIMEX3 have been reduced to one, namely, they are now linked to the event type (or to every mention of one specific event).

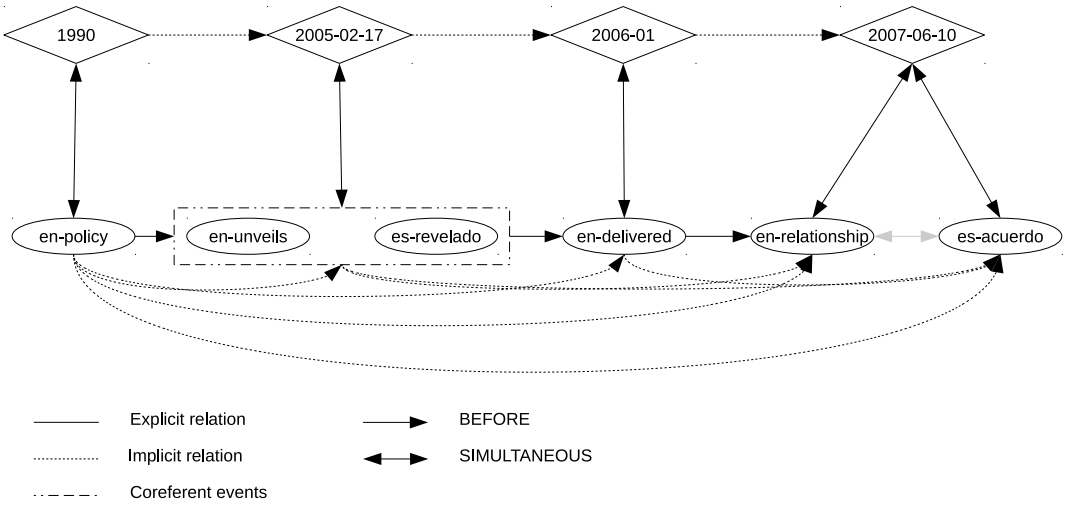

Figure 4: Time graph produced by *Relaxed evaluation*. Grey lines represent redundant relations.

In this method the number of relations that must be captured is smaller because detecting just one of the coreferent event mentions shall be enough. Thus, this evaluation is more **relaxed** in terms of recall. However, it is still required to properly detect coreferent events, otherwise they will be evaluated as different instances, consequently harming the precision.

6. Automatic Cross-lingual TimeLine extraction

This section presents our approach for timeline extraction, including both multilingual and cross-lingual systems. Given a set of documents and a target entity, a three step process is applied. First, the mentions of the target entity are identified. Second, the events in which the target entity is involved are selected. Finally, those events are anchored to their respective normalized time expressions. Once this process is completed, the events are sorted and the timeline built.

In the following we describe the three different systems for Timeline extraction applied to the tasks previously described. Section 6.1 introduces the baseline (BTE) system. BTE performs timeline extraction by combining the output of a NLP pipeline for both English and Spanish. The baseline system is then improved in section 6.2 by applying the algorithm presented in Laparra

et al. (2015) to perform document level time-anchoring (DLT). While both BTE and DLT can be used for multilingual timeline extraction, their performance in the cross-lingual setting is not as good as in the English and Multilingual tasks. Thus, in section 6.3 we propose a new approach to obtain interoperable annotations across languages from the same NLP pipelines used for BTE in section 6.1. We can then use this approach to identify coreferent event mentions across languages which is crucial to build cross-lingual timelines.

6.1. BTE: Baseline TimeLine Extraction

Detecting mentions of events, entities and time expressions in text requires the combination of various NLP tools. We apply the NewsReader NLP pipelines (Vossen et al., 2016) that includes, both for English and Spanish, Named Entity Recognition (NER) and Disambiguation (NED), Coreference Resolution (CR), Semantic Role Labelling (SRL), Time Expressions Identification (TEI) and Normalization (TEN), and Temporal Relation Extraction (TRE). Table 2 lists the specific tools used for English and Spanish.

	English	Spanish
NER	Agerri and Rigau (2016)	
NED	Daiber et al. (2013)	
CR	Agerri et al. (2014)	
SRL	Björkelund et al. (2009)	
TEI	Mirza and Minard (2014)	Strötgen et al. (2013)
TEN	Mirza and Minard (2014)	Strötgen et al. (2013)
TRE	Mirza and Tonelli (2014)	Llorens et al. (2010)

Table 2: English and Spanish NLP tools.

The extraction of target entities, events and time anchors is performed as follows:

(1) Target entity identification: The target entities are identified by the NER and NED modules. As the surface form of the candidate entities can vary greatly, we use the redirect links contained in DBpedia to extend the search of the events involving those target entities. For example, if the target entity is *Toyota* the system would also include events involving the entities *Toyota Motor Company* or *Toyota Motor Corp.* In addition, as the NED module is not alway able to provide a link to DBpedia, we also check if the wordform of the head of the event argument matches with the head of the target entity.

(2) Event selection: We take the output of the SRL module in order to extract the events occurring in a document. Given a target entity, we combine the output of the NER, NED, CR and SRL to obtain the events that have the target entity as filler of their ARG0 or ARG1. We also follow the specifications of the SemEval task and set some constraints to select certain events. Specifically, we avoid events that are within the scope of a negation or are related to modal verbs (except *will*).

(3) Time-anchoring: The time-anchors are identified using the TRE and SRL output. From the TRE, we extract as time-anchors the SIMULTANEOUS

relations between events and time expressions. On the other hand, we get from the SRL those ARG-TMP related to time expressions. In both cases we use the time expression produced by the TEI module. According to the tests performed on the trial data, the best choice for time-anchoring results from the combination of both options. For each time-anchor we use the TEN module to normalize the time expression.

6.2. DLT: Document Level Time-anchoring

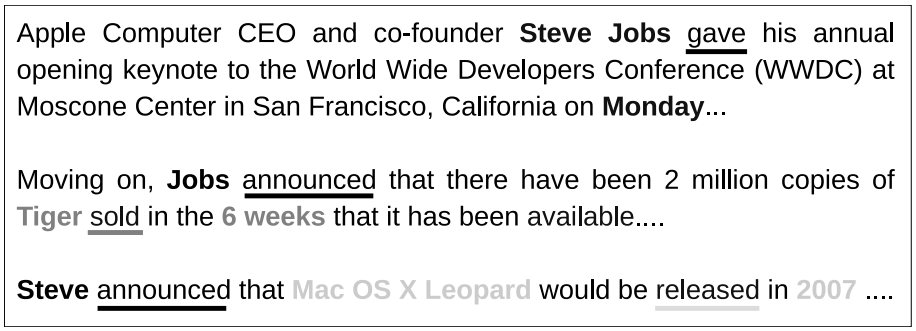

Figure 5: Example of time-anchoring at document level.

The NLP tools of the system presented in previous section 6.1 are not able to provide a time-anchor for every event involving a particular entity cause many of them are anchored to time implictly in the text. Therefore, events withou an explicit time-anchor are not captured as part of the timeline. This means that we have to be able to also recover those time-anchors that are implicitly conveyed in the text.

In Laparra et al. (2015) we devised a simple strategy to capture implicit time-anchors while maintaining the coherence of the temporal information in the document. The rationale behind the algorithm shown in Algorithm 1 is that the events involving a specific entity that appear in a document tend to take place at the same time as previous events involving the same entity (unless explicitly stated). For example, in Figure 5 every event related to *Steve Jobs*, such as *gave* and *announced*, are anchored to the same time expression (*Monday*) even though it is only explicitly conveyed for the first event *gave*. This example also illustrates the fact that for those other events that occur at different times, their time-anchor is also explicit, as it can be seen for the *Tiger* and *Mac OS X Leopard* entities.

The application of Algorithm 1 starts taking as input the annotation obtained by the NLP described in Section 6.1. For each entity a list of events ($eventList$) is created sorted by appearing order. Next, for each event in the list the algorithm checks whether that event is already anchored to a time expression ($eAnchor$). In that case, that time-anchor is included in the list of default time-anchors ($defaultAnchor$) for any subsequent events of the entity in the same verb tense ($eTense$). If the event does not yet have an explicit

time-anchor assigned, but the system has found a time-anchor for a previous event in the same tense ($defaultAnchor[eTense]$), the algorithm assigns that time-anchor to the current event ($eAnchor$). If none of the previous conditions hold, then the event is anchored to the **Document Creation Time** (DCT) attribute and sets this time-expression as the default time-anchor for any subsequent events in the same verbal tense.

Algorithm 1 Implicit Time-anchoring

```
1: eventList = sorted list of events of an entity
2: for event in eventList do
3:    eAnchor = time anchor of event
4:    eTense = verb tense of event
5:    if eAnchor not NULL then
6:       defaultAnchor[eTense] = eAnchor
7:    else if defaultAnchor[eTense] not NULL then
8:       eAnchor = defaultAnchor[eTense]
9:    else
10:      eAnchor = DCT
11:      defaultAnchor[eTense] = DCT
12:   end if
13: end for
```

The **DLT** system build the timeline by ordering the events according to the explicit and implicit time-anchors. Note that Algorithm 1 strongly depends on the tense of the mentions of events appearing in the document. As this information can be only recovered from verbal predicates, this strategy cannot be applied to events conveyed by nominal predicates. Consequently, for these cases just explicit time-anchors are taken into account.

6.3. CLE: Cross-Lingual Event coreference

As it has been already mentioned, cross-lingual timeline extraction crucially depends on being able to identify those events that are coreferent across languages (not only across documents). In order to address this issue, we propose a language independent knowledge representation for cross-lingual semantic interoperability at three different annotation levels.

First, we used interconnected links in the DBpedia entries to perform cross-lingual Named Entity Disambiguation (NED). The NED module used in the NLP pipeline for BTE provides links to the English and Spanish versions of the DBpedia. Thus, a mention of *U.S. Air Force* in English should link as external reference to the the identifier `http://dbpedia.org/page/United_States_Air_Force`. Similarly, a mention of *Fuerzas Areas americanas* in Spanish should produce as external reference the identifier `http://es.dbpedia.org/page/Fuerza_Area_de_los_Estados_Unidos`. As both identifiers are connected within the DBpedia, we can just infer that those two pointers refer to the same target entity regardless of the language in which the mentions of that entity are expressed.

Second, we obtain inter-operability across languages and Semantic Role Labeling annotations by means of the PredicateMatrix (López de Lacalle et al., 2016a,b). The event representation provided by our SRL systems are based on PropBank, for English, and AnCora (Taulé et al., 2008), for Spanish. The PredicateMatrix gathers knowledge bases that contain predicate and semantic role information in different languages, including links between PropBank and AnCora. Using these mappings, we can establish, for example, that the role *arg0* of the Spanish predicate *elegir.1* is aligned to the role *A0* of the PropBank predicate *select.01*.

Finally, the TEN modules normalize time expressions following the ISO 24617-1 standard (Pustejovsky et al., 2010). For example, if temporal expressions such as *last Sunday*, *February 29*, and *yesterday* in English or *ayer* and *el 29 de febrero* in Spanish are referring to the same exact date (let's say *February 29th, 2009*), then they will be normalized to the same TIMEX3 value corresponding to *2009-02-29*.

We can include these three levels of cross-lingual information to extend the multilingual system DLT presented in the previous section. When extracting the cross-lingual timeline for a given target entity, expressed as e_E and e_S in English and Spanish respectively, the system establishes that the English event p_E and the Spanish event p_S are coreferent if the following conditions are satisfied:

1. e_E and e_S are connected by DBpedia links to the same entity.
2. e_E plays the role r_E of p_E, e_S plays the role r_S of p_S, and r_E and r_S are linked by a mapping in the PredicateMatrix.
3. p_E is anchored to a TIMEX3 t_E, p_S is anchored to a TIMEX3 t_S and t_E and t_S are normalized to the same ISO 24617-1.

Figure 6 contains an example of two events that satisfy the previous conditions and are consequently identified as cross-lingually coreferent. The **CLE** system uses the same strategy as DLT to build timelines with the difference that cross-lingual coreferent events are identified.

7. Experimental Results

In this section we present a set of experiments in order to evaluate the three timeline extraction systems presented in the previous section: (i) the **BTE** baseline system based on the analysis given by a pipeline of NLP tools; (ii) the **DLT** algorithm that aims at capturing implicit time-anchoring at document level; and (iii) the **CLE** system to address cross-lingual event co-reference. The evaluations are undertaken for the original English SemEval 2015 task as well as for the Multilingual and Cross-Lingual Timeline Extraction tasks proposed in section 3. Every result is evaluated using the original SemEval 2015 metric as well as the *strict* and *relaxed* metrics introduced in section 5.

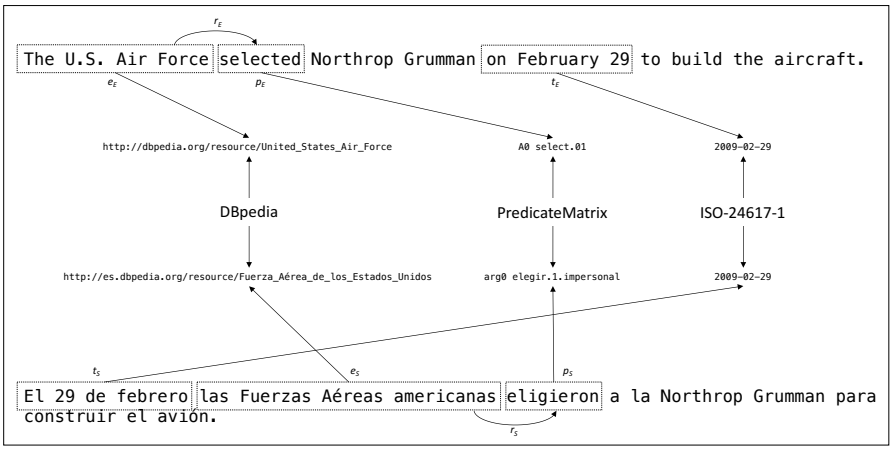

Figure 6: Example of event coreference through cross-lingual semantic resources.

7.1. Multilingual evaluation

In this setting we evaluate both BTE and DLT systems on the Track A (main track) of the TimeLine Extraction task at SemEval 2015 and on the Multilingual task described in section 3. Track A at SemEval 2015 had just two participant teams, namely, **WHUNLP** and **SPINOZAVU**, which submitted three runs in total. Their scores in terms of Precision (P), Recall (R) and F1 are presented in Table 3. We also present in italics additional results of both systems obtained after the official evaluation task (Caselli et al., 2015). The best run was obtained by the corrected version of **WHUNLP_1** with an F1 of 7.85%. The low figures obtained show the difficulty of the task.

System	P	R	F1
SPINOZAVU-RUN-1	7.95	1.96	3.15
SPINOZAVU-RUN-2	8.16	0.56	1.05
WHUNLP_1	14.10	4.90	7.28
OC_SPINOZA_VU	-	-	7.12
WHUNLP_1	14.59	5.37	7.85
BTE	**24.56**	4.35	7.39
DLT	21.00	**11.01**	**14.45**

Table 3: Results on the SemEval-2015 task

Table 3 also contains the results obtained by our systems. The results obtained by our baseline system, **BTE**, are similar to those obtained by **WHUNLP_1**. However, the results of the implicit time-anchoring approach (**DLT**) clearly outperforms our baseline and every other previous result in this task. This result

would imply that a full time-anchoring annotation requires that a temporal analysis be carried out at document level. As expected, Table 3 also shows that the improvement of DLT over BTE is much more significant in terms of Recall.

		English			Spanish		
Scorer	System	P	R	F1	P	R	F1
	BTE	24.56	4.35	7.39	12.07	4.25	6.29
SemEval-2015	**DLT**	21.00	11.01	14.45	12.77	8.60	10.28
	BTE	24.56	3.62	6.32	12.07	3.60	5.55
strict-evaluation	**DLT**	21.00	9.18	12.77	12.77	7.29	9.28
	BTE	24.12	5.32	8.71	11.55	5.18	7.15
relaxed-evaluation	**DLT**	19.39	12.95	15.53	11.47	9.72	10.52

Table 4: Results on the multilingual task.

Table 4 provides the results obtained by **BTE** and **DLT** in the Multilingual Timeline extraction setting using also the *strict* and *relaxed* evaluation metrics described in Section 5. Predictably, the strict evaluation is the most demanding, specially in terms of Recall. With respect to the results obtained using the relaxed scorer, precision is lower whereas recall is higher with respect to the other two metrics. Furthermore, **DLT** outperforms **BTE** whatever the language and the evaluation methodology. It is also remarkable that the results obtained for English are always better than the results for Spanish. This can be explained by the differences in the performances of the English and Spanish NLP modules.

7.2. Cross-lingual evaluation

The dataset for cross-lingual timelines contains 180 documents (see Section 4), of which half are Spanish translations of the other half written in English. This fact allows us to set different experiments by varying the percentage of documents written in each language that are provided as input. Three different experiments were performed in order to evaluate our systems on the Cross-lingual Timeline extraction task:

1. An experiment using the full set of documents in both languages available in the dataset (*Full data*).
2. A more realistic scenario where we get half of the documents in each language avoiding to include parallel translations (*50-50 split*).
3. An evaluation of the number of event instances recovered by our system depending on the number of documents in each language used as input (*Varying input per language*).

Full data. For the first experiment, we use as input the full collection (180 documents) independently of the language. As shown by Table 5, the results using the SemEval 2015 scoring method, as it was the case in the multilingual setting, the **DLT** system almost doubles the score of the baseline system **BTE**. Furthermore, **DLT** and **CLE** obtain exactly the same results because co-referent events

are not taken into account. However, the *strict* and *relaxed* scoring methods proposed in this work make it possible to distinguish between the performances of the two systems. Not surprisingly, the scoring by strict evaluation continues to be the lowest. Overall, **CLE** outperforms **DLT** being only in terms of precision (relaxed evaluation) or in both precision and recall (strict evaluation).

Scorer	System	P	R	F1
SemEval-2015	**BTE**	13.98	4.68	7.02
	DLT	14.96	10.74	12.50
	CLE	14.96	10.74	12.50
strict-evaluation	**BTE**	13.98	3.12	5.10
	DLT	14.96	7.14	9.67
	CLE	16.59	8.47	11.22
relaxed-evaluation	**BTE**	10.13	8.16	9.04
	DLT	9.75	17.70	12.57
	CLE	10.97	17.70	13.55

Table 5: Results on the cross-lingual task

50-50 split. As we believe that the availability of a set of parallel documents as input is not the most realistic scenario, we design another setting by choosing at random 50% of the documents in each language, namely, 45 documents for English and 45 for Spanish respectively. The resulting input set would contain 90 non-parallel documents in two languages without the mentions of the events that belong to documents not included in the final collection of 90 documents. Furthermore, we automatically generate not just one but $1,000$ different 50-50 input sets of 90 documents at random, namely, each of the thousand sets contain 45 documents in each language. The box-plots in Figure 7 show the results obtained by our systems in this experiment applying the strict and relaxed evaluation methodologies to the one thousand evaluation sets.

Following the trend of previous results, both **DLT** and **CLE** outperform the baseline system with **CLE** obtaining the best overall performance. The F1 score differences between **DLT** and **CLE** using both evaluation methods are significant with $p < 0.001$.[5] In any case, the results show that performance between **DLT** and **CLE** has reduced with regard to the results obtained in the previous experiment reported by Table 5. Our hypothesis is that as the set of input documents in this experiment has been halved, the number of coreferent mentions in the gold-standard is much lower, which means that the advantage of CLE over DLT is not that meaningful. The most remarkable variation can be observed in the Recall values obtained using the relaxed evaluation. This is not that strange if we consider that in the relaxed evaluation detecting only one mention of an event is enough.

[5]We have used the paired t-test to compare the $F1$ obtained by the systems.

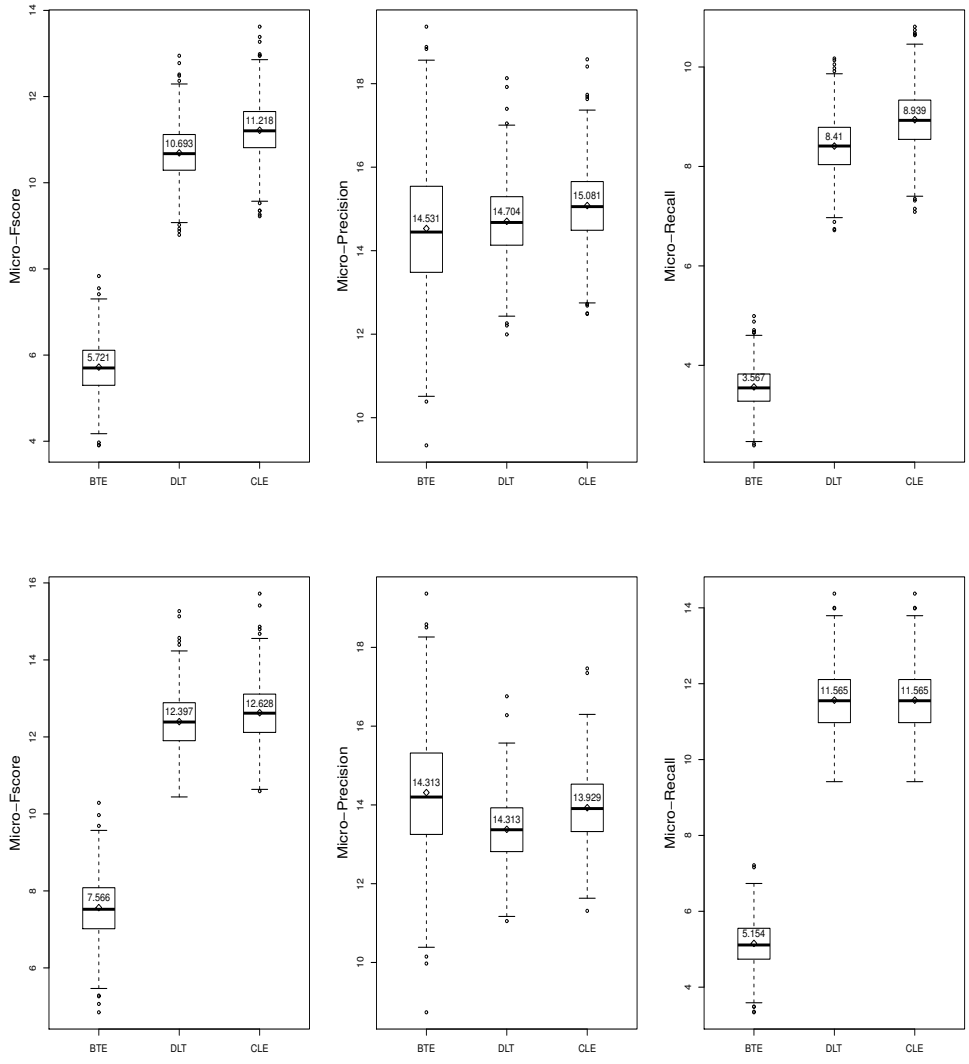

Figure 7: Evaluation 50-50. The top row results are calculated using the *strict* metric whereas the results at the bottom row refer to the *relaxed* evaluation method.

Varying input per language. This last experiment was designed to study how varying the number of documents per language affects the performance in the cross lingual setting. The line charts in Figure 8 show the results obtained varying the percentage of the documents being used.

On the left-hand side plot we show the results of experiments using a range of 5% to 95% documents for both languages (Spanish on top, English at the

bottom). Now, for each point in the range we randomly generate 30 input sets. For example, at the 10% Spanish and 90% English 30 different configurations are randomly generated each of which would contain 81 English documents and 9 Spanish documents (90 documents in total).

In the experiments reported by the central and right-hand size plots, we use the above method to generate 30 input sets for each point in the range, but with two important differences. Firstly, every document in one language is alternatively used (English in the central plot and Spanish on the right-hand side) and we increase the number of documents in the other language from 5% up to 95% (when the 180 documents are used). Secondly, in these two cases parallel documents are allowed.

For all three cases each point represents the *arithmetic mean* of the output given for the 30 different input document sets generated without replacement. The evaluation method used is *relaxed* due to the fact that we start with the full set of possible events. Thus, varying or increasing the number of documents in the other language does not in fact increase the number of events, just (possibly) the number of event mentions. Therefore, the *relaxed* method allows us to focus on studying whether adding parallel documents in other language improves the overall F1 score, paying particular attention to the Recall.

The results illustrate that the CLE F1 score keeps degrading as we include Spanish documents into the fold. This is somewhat explained by CLE results obtained in Table 4 where the performance of the Spanish system is much worse that its English counterpart. Particularly, according to the Recall in the left plot the Spanish pipeline extract fewer event instances that the English pipeline. However, the Recall in the middle and right plots shows that the system obtains more event instances when it combines multilingual parallel sources that with an isolated language. In other words, both English and Spanish pipelines recover event instances that are missed by the other pipeline.

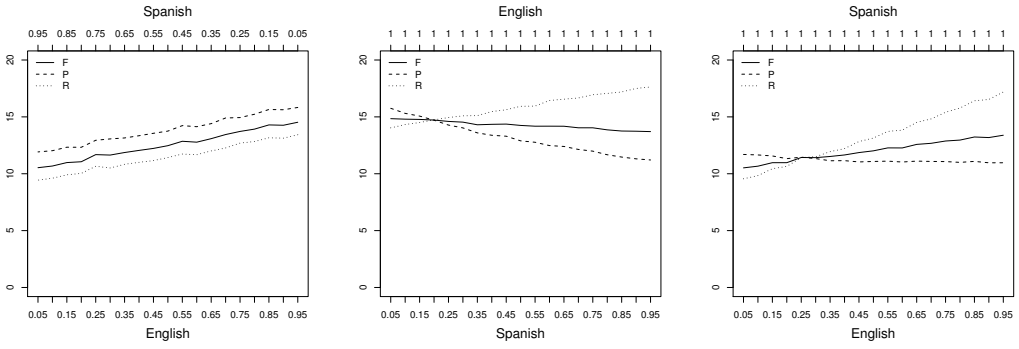

Figure 8: Varying the number of input documents per language. The y axis in each box represents the percantage of documents used for each language.

8. Error Analysis

As shown in Section 6.1, our baseline approach for timeline extraction (on which **DLT** and **CLE** build) is based on the output of a set of NLP modules. Now, although they are state-of-the-art tools on standard evaluation data, they still produce cascading errors, most notably when applied to out-of-domain data (see Table 8 at Vossen et al. (2016)). The aim of this section is to identify the main source of errors.

	English	Spanish	Cross-lingual
full (BTE)	6.04	8.43	8.20
full (DLT)	19.55	17.70	19.83
SRL	72.06	56.29	63.04
SRL+NER+NED	22.01	17.67	20.97
SRL+NER+NED+CR	23.95	17.67	22.25
SRL+TEI+TEN+TRE (BTE)	13.72	21.50	19.47
SRL+TEI+TEN+TRE (DLT)	46.16	53.21	50.43

Table 6: Percentage of events captured by the pipelines. The *full* rows correspond to SRL+NER+NED+CR+TEI+TEN+TRE.

In a first experiment we study the capability of our system for extracting those events that participate in the timelines, regardless of time ordering. The first two rows in Table 6 show that the **DLT** system is able to extract way more events than the **BTE** baseline system, however in both cases the percentage of events captured is still low. To study the causes of these figures we have repeated the same experiment with partial combinations of the NLP modules. As explained in Section 6.1, we use a SRL system to detect event mentions. Table 6 shows that for English the SRL module detects more events than for Spanish (72.06% vs 56.29%). This is largely due to the Spanish SRL not dealing correctly with verbal nominalizations.

In order to extract only those events that are linked to the target entity, we use the combined output of the SRL, NER, NED and CR tools (see Section 6.1). Table 6 shows that this is a very difficult step and that the percentage of events identified is rather low. Detecting and linking every mention of an entity is a very difficult task, specially in the case of pronouns. As it can be seen, the coreference module helps although not as much as it would have been expected.

The final two rows of Table 6 report on the results obtained when only events with a time anchor are included in a timeline. The number of events linked to a explicit time-anchor by our BTE baseline system is very low whereas looking at the implicit anchors in the DLT system helps to substantially improve the results. Notice that in this case the figures are higher for Spanish (21.50% and 53.21%) than for English (13.72% and 46.16%). This means that time modules for Spanish try to anchor more events that the English modules.

In a second experiment we study the quality of the time anchoring. Table 7 shows the accuracy of the time-anchors for the events that we know have been

correctly identified. It makes sense that the accuracy of **DLT** be much lower than just taking into account explicit time-anchors as **BTE** does. However, it should be noted that number of events extracted by the DLT system is much higher than BTE (as per Table 6), which means that accuracy for DLT is calculated over a much larger number of correctly identified event mentions. As can be seen, the English systems perform better than the Spanish systems. As explained above, the Spanish modules try to time-anchor more events and this fact can explain that they obtain a lower accuracy.

	English	Spanish	Cross-lingual
BTE	69.49	50.70	68.70
DLT	51.31	46.98	62.59

Table 7: Accuracy of the time-anchoring for extracted events.

9. Concluding Remarks

In this work we present a system to perform Multilingual and Cross-lingual Timeline Extraction (or Cross-document event ordering). In doing so, this paper presents a number of novel contributions.

Firstly, the original Cross-document event ordering task defined for SemEval 2015 (main Track A) has been extended to present two novel tasks for two languages (English and Spanish) on Multilingual and Cross-lingual timeline extraction respectively. The annotated datasets for trial and evaluation are publicly available.

Secondly, two new evaluation metrics improve the evaluation methodology of the SemEval 2015 task in two ways: (i) A new *strict* metric allows to evaluate timelines containing coreferent event mentions across both documents and languages; and (ii) a *relaxed* evaluation metric where event types (instead of mentions) can be considered, somewhat diminishing the importance of recall when evaluating the timelines.

Thirdly, three deterministic Timeline extraction systems have been developed to address the three tasks. In fact, we have empirically demonstrated that addressing implicit time-anchors at document level (DLT system) crucially improves the performance in the three tasks, clearly outperforming previously presented systems in the (main) Track A of the original Timeline Extraction task at SemEval 2015. Furthermore, we have shown how to effectively use cross-lingual resources such as the PredicateMatrix and DBpedia along with time normalization to improve the performance of the DLT system in the most realistic setting of building cross-lingual timelines without parallel data as input (see Figure 7).

Finally, we have analyzed the cascading errors produced by the NLP pipeline used to identify the entities, events and time-anchors. The results allow to conclude that the most difficult obstacles reside in detecting and resolving every

mention of entities related to the relevant mention events and the identification of time-anchors when they are not explicitly conveyed. These two aspects shall point out future work towards improving timeline extraction.

Acknowledgments

This work has been supported by the European projects QTLeap (EC-FP7-610516) and NewsReader (EC-FP7-316404) and by the Spanish Ministry for Science and Innovation (MICINN), SKATER (TIN2012-38584-C06-01) and TUNER (TIN2015-65308-C5-1-R).

References

Agerri, R., Bermudez, J., Rigau, G., 2014. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014).

Agerri, R., Rigau, G., 2016. Robust multilingual named entity recognition with shallow semi-supervised features. Artificial Intelligence 238, 63–82.

Björkelund, A., Hafdell, L., Nugues, P., 2009. Multilingual semantic role labeling, in: Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task, pp. 43–48.

Blanco, E., Moldovan, D., 2014. Leveraging verb-argument structures to infer semantic relations, in: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden. pp. 145–154.

Caselli, T., Fokkens, A., Morante, R., Vossen, P., 2015. SPINOZA_VU: An nlp pipeline for cross document timelines, in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado. pp. 786–790.

Chambers, N., Cassidy, T., McDowell, B., Bethard, S., 2014. Dense event ordering with a multi-pass architecture. Transactions of the Association for Computational Linguistics 2, 273–284.

Chambers, N., Wang, S., Jurafsky, D., 2007. Classifying temporal relations between events, in: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Prague, Czech Republic. pp. 173–176.

Cornegruta, S., Vlachos, A., 2016. Timeline extraction using distant supervision and joint inference, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016), pp. 1936–1942.

Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N., 2013. Improving efficiency and accuracy in multilingual entity extraction, in: Proceedings of the 9th International Conference on Semantic Systems (I-Semantics).

DŚouza, J., Ng, V., 2013. Classifying temporal relations with rich linguistic knowledge, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia. pp. 918–927.

Gerber, M., Chai, J., 2012. Semantic role labeling of implicit arguments for nominal predicates. Computational Linguistics 38, 755–798.

López de Lacalle, M., Laparra, E., Rigau, G., 2014. Predicate matrix: extending semlink through wordnet mappings, in: The 9th edition of the Language Resources and Evaluation Conference (LREC 2014). Reykjavik, Iceland.

Laokulrat, N., Miwa, M., Tsuruoka, Y., Chikayama, T., 2013. Uttime: Temporal relation classification using deep syntactic features, in: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, Georgia, USA. pp. 88–92.

Laparra, E., Aldabe, I., Rigau, G., 2015. Document level time-anchoring for timeline extraction, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015), Beijing, China.

Laparra, E., Rigau, G., 2013. Impar: A deterministic algorithm for implicit semantic role labelling, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), pp. 33–41.

Llorens, H., Chambers, N., UzZaman, N., Mostafazadeh, N., Allen, J., Pustejovsky, J., 2015. Semeval-2015 task 5: Qa tempeval - evaluating temporal information understanding with question answering, in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado. pp. 792–800.

Llorens, H., Saquete, E., Navarro, B., 2010. Tipsem (english and spanish): Evaluating crfs and semantic roles in tempeval-2, in: Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics. pp. 284–291.

López de Lacalle, M., Laparra, E., Aldabe, I., Rigau, G., 2016a. A multilingual predicate matrix, in: Proceedings of the 10th international conference on Language Resources and Evaluation (LREC 2016), Tartu, Estonia.

López de Lacalle, M., Laparra, E., Aldabe, I., Rigau, G., 2016b. Predicate matrix. automatically extending the semantic interoperability between predicate resources. Language Resources and Evaluation 50.

Mani, I., Verhagen, M., Wellner, B., Lee, C.M., Pustejovsky, J., 2006. Machine learning of temporal relations, in: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia. pp. 753–760.

Mani, I., Wellner, B., Verhagen, M., Pustejovsky, J., 2007. Three Approaches to Learning TLINKs in TimeML. Technical Report.

Minard, A.L., Speranza, M., Agirre, E., Aldabe, I., van Erp, M., Magnini, B., Rigau, G., Urizar, R., 2015. Semeval-2015 task 4: Timeline: Cross-document event ordering, in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado. pp. 778–786.

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

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

Mirza, P., Tonelli, S., 2014. Classifying temporal relations with simple features, in: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Gothenburg, Sweden. pp. 308–317.

Navarro-Colorado, B., Saquete, E., 2016. Cross-document event ordering through temporal, lexical and distributional knowledge. Knowledge-Based Systems 110, 244 – 254.

Palmer, M.S., Dahl, D.A., Schiffman, R.J., Hirschman, L., Linebarger, M., Dowding, J., 1986. Recovering implicit information, in: Proceedings of the 24th annual meeting on Association for Computational Linguistics, New York, New York, USA. pp. 10–19.

Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., Lazo, M., 2003a. The TIMEBANK corpus, in: Proceedings of Corpus Linguistics 2003, Lancaster.

Pustejovsky, J., Lee, K., Bunt, H., Romary, L., 2010. ISO-TimeML: An International Standard for Semantic Annotation, in: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta.

Pustejovsky, J., no, J.C., Ingria, R., Saur, R., Gaizauskas, R., Setzer, A., Katz, G., 2003b. Timeml: Robust specification of event and temporal expressions in text, in: in Fifth International Workshop on Computational Semantics (IWCS-5).

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

Sun, W., Rumshisky, A., Uzuner, O., 2013. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. Journal of the American Medical Informatics Association 20, 806–813.

Taulé, M., Martí, M.A., Recasens, M., 2008. Ancora: Multilevel annotated corpora for catalan and spanish., in: LREC 2008.

Tetreault, J.R., 2002. Implicit role reference, in: International Symposium on Reference Resolution for Natural Language Processing, Alicante, Spain. pp. 109–115.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., Pustejovsky, J., 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations, in: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, Georgia, USA. pp. 1–9.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., Pustejovsky, J., 2007. Semeval-2007 task 15: Tempeval temporal relation identification, in: Proceedings of the 4th International Workshop on Semantic Evaluations, Prague, Czech Republic. pp. 75–80.

Verhagen, M., Saurí, R., Caselli, T., Pustejovsky, J., 2010. Semeval-2010 task 13: Tempeval-2, in: Proceedings of the 5th International Workshop on Semantic Evaluation, Los Angeles, California. pp. 57–62.

Vossen, P., Agerri, R., Aldabe, I., Cybulska, A., van Erp, M., Fokkens, A., Laparra, E., Minard, A.L., Aprosio, A.P., Rigau, G., Rospocher, M., Segers, R., 2016. Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. Knowledge Based Systems 110, 60–85.

Whittemore, G., Macpherson, M., Carlson, G., 1991. Event-building through role-filling and anaphora resolution, in: Proceedings of the 29th annual meeting on Association for Computational Linguistics, Berkeley, California, USA. pp. 17–24.