From TimeLines to StoryLines: A preliminary proposal for evaluating narratives

Egoitz Laparra, Itziar Aldabe, German Rigau
IXA NLP group, University of the Basque Country (UPV/EHU)
{egoitz.laparra,itziar.aldabe,german.rigau}@ehu.eus

Abstract

We formulate a proposal that covers a new definition of StoryLines based on the shared data provided by the *NewsStory workshop*. We re-use the SemEval 2015 Task 4: Timelines dataset to provide a gold-standard dataset and an evaluation measure for evaluating StoryLines extraction systems. We also present a system to explore the feasibility of capturing StoryLines automatically. Finally, based on our initial findings, we also discuss some simple changes that will improve the existing annotations to complete our initial StoryLine task proposal.

1 Introduction

The process of extracting useful information from large textual collections has become one of the most pressing problems in our current society. The problem spans all sectors, from scientists to intelligence analysts and web users. All of them are constantly struggling for synthesizing the relevant information from a particular topic. For instance, behind this overwhelmingly large collection of documents, it is often easy to miss the important details when trying to make sense of complex stories. To solve this problem various types of document processing systems have been recently proposed. For example, generic and query-focused multi-document summarization systems aim to choose from the documents a subset of sentences that collectively conveys a query-related idea (Barzilay et al., 1999). News topic detection and tracking systems usually aim at grouping news articles into a cluster to present the events related to a certain topic (Allan, 2002). Timelines generation systems create summaries of relevant events in a topic by leveraging temporal information attached or appearing in the documents (Swan and Allan, 2000; Shahaf and Guestrin, 2010; Matthews et al., 2010; Mazeika et al., 2011; Do et al., 2012). TimeLines differ from other narrative structures like (Chambers and Jurafsky, 2008; Chambers and Jurafsky, 2009) in that the time-anchors of the events are required for Timelines construction. Although TimeLine systems present the sequence of events chronologically, linear-structured TimeLines usually focus on a single entity losing comprehensive information of relevant interactions with other participants. Thus, some other systems try to construct maps of connections that explicitly captures story development (Shahaf et al., 2013) or complex storylines (Hu et al., 2014).

Following this research line, we propose a cross-document StoryLine task based on the shared data provided by the workshop organizers. The approach extends the TimeLines evaluation task carried out in SemEval 2015[1] (Minard et al., 2015). The aim of the TimeLine task is to order on a TimeLine the events in which a target entity is involved (cf. Section 2). In contrast, our approach explores the inner interactions of these TimeLines. As a result, we define a StoryLine as a group of interacting TimeLines. For instance, given *Apple Inc.* as the news topic, Figure 2 presents a StoryLine built from *Steve Jobs* and *iPhone 4* TimeLines. It shows how an interaction of two TimeLines is highlighted when events are relevant to both TimeLines. In this way, a StoryLine groups together the events corresponding to multiple but interacting TimeLines. In the same way, if two additional entities interact with each other and they do not interact with *Steve Jobs* and *iPhone 4* TimeLines, two separate StoryLines would be derived from the *Apple Inc.* topic, each one corresponding to the set of interacting entity TimeLines.

The contributions of this research are manifold.

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

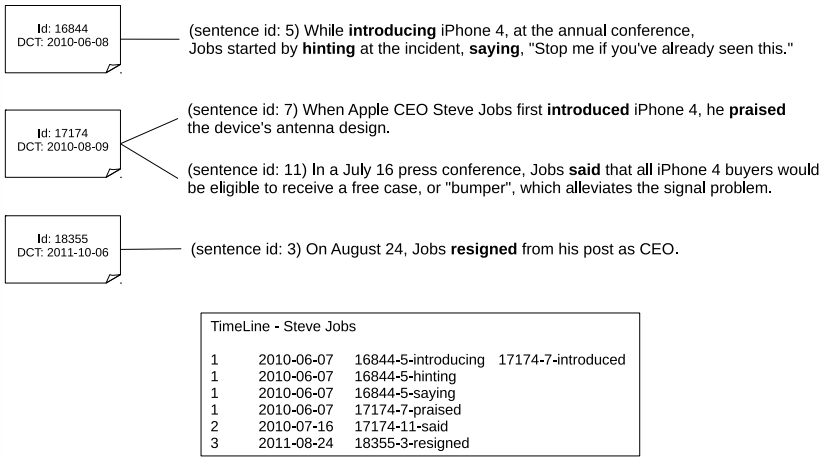

Figure 1: Example of the *Steve Jobs* TimeLine.

First, we devise a proposal that covers a new definition of StoryLines based on the existing proposal for TimeLines. We provide gold-standard StoryLines and we re-use the evaluation metric proposed in SemEval-2015 to evaluate StoryLines. We also present a very basic system that tries to capture the StoryLines that appear in the original documents of the TimeLines task. Finally, based on our initial findings, we discuss some initial improvements that can be addressed in the existing annotations and evaluation system to complete our initial StoryLine task proposal.

2 TimeLines

The aim of the Cross-Document Event Ordering task is to build TimeLines from English news articles (Minard et al., 2015). Given a set of documents and a set of target entities, the TimeLines task consisted of building a TimeLine for each entity, by detecting the events in which the entity is involved and anchoring these events to normalized times. Thus, a TimeLine is a collection of ordered events in time relevant for a particular entity.

Figure 1 shows the TimeLine extracted for the target entity *Steve Jobs* using information from 3 different documents. The events in bold form the TimeLine that can be placed on a TimeLine according to the task annotation guidelines (Minard et al., 2014). TimeLines contain relevant events in which the target entity participates as ARG0 (i.e agent) or ARG1 (i.e. patient) as defined in Prop-Bank (Palmer et al., 2005). Events such as adjectival events, cognitive events, counter-factual events, uncertain events and grammatical events are excluded from the TimeLine.[2] For example, the events *introducing*, *hinting* and *saying* from sentence 5 in document 16844 are part of the TimeLine for the entity *Steve Jobs* but the events *started* and *Stop* are not. *Steve Jobs* participates as ARG0 or ARG1 in all the events, but *started* is a grammatical event and *Stop* is an uncertain event. Thus, according to the SemEval annotation guidelines, they are excluded from the TimeLine. In addition, each event is placed on a position according to the time-anchor and the coreferring events are placed in the same line (see *introducing* and *introduced* events in documents 16844 and 17174 respectively).

The main track of the task (Track A) consists of building TimeLines providing only the raw text sources. The organisers also defined Track B where gold event mentions were given. For both tracks, a sub-track in which the events are not associated to a time anchor was also presented. The StoryLines proposal here presented follows the main track approach.

3 A Proposal for StoryLines

In this section we present a first proposal for a novel evaluation task for StoryLines. We propose that a StoryLine can be built by merging the individual TimeLines of two or more different entities, provided that they are co-participants of at least one relevant event.

In general, given a set of related documents, any entity appearing in the corpus is a candidate to take

[2] A complete description of the annotation guidelines can be found at `http://www.newsreader-project.eu/files/2014/12/NWR-2014-111.pdf`

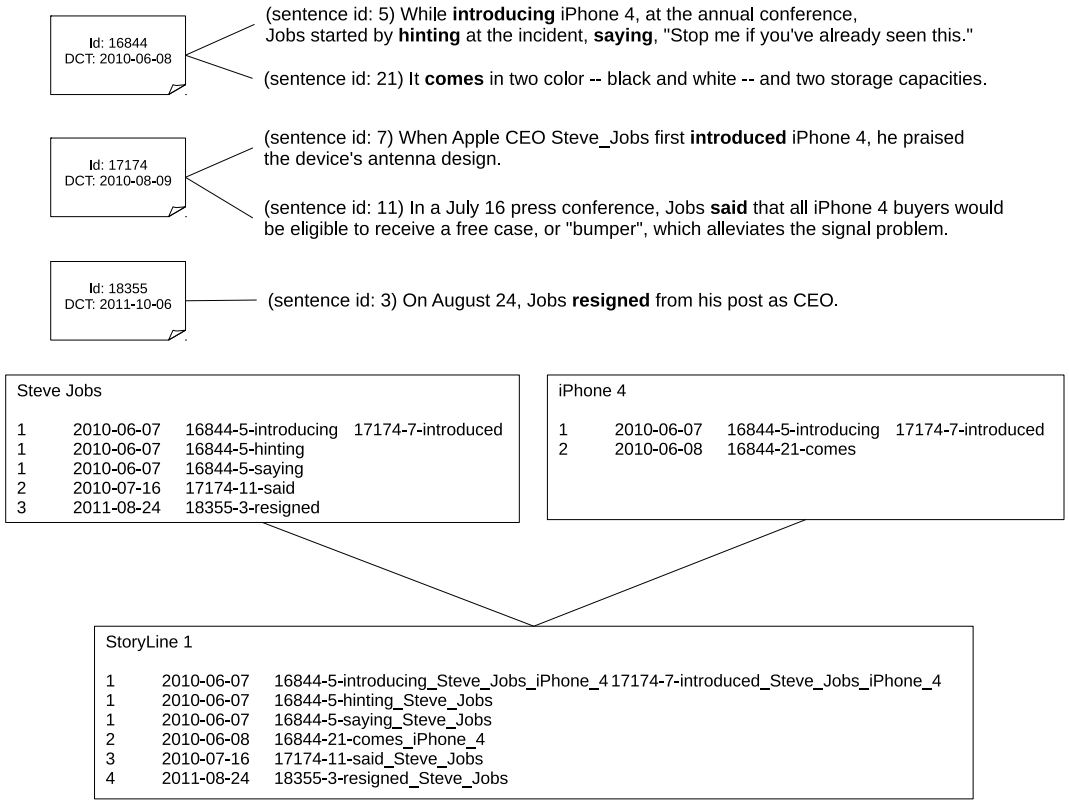

Figure 2: Example of a StoryLine merging the TimeLines of the entities *Steve Jobs* and *Iphone 4*.

part in a StoryLine. Thus, a TimeLine for every entity should be extracted following the requirements described by SemEval-2015. Then, those TimeLines that share at least one relevant event must be merged. Those entities that do not co-participate in any event with other entities are not considered participants of any StoryLine.

The expected StoryLines should include both the events where the entities interact and the events where the entities selected for the StoryLines participate individually. The events must be ordered and anchored in time in the same way as individual TimeLines, but it is also mandatory to include the entities that take part in each event.

Figure 2 presents graphically the task idea. In the example, two TimeLines are extracted using 5 sentences from 3 different documents, one for the entity *Steve Jobs* and another one for the entity *Iphone 4*. As these two entities are co-participants of the events *introducing* and *introduced*, the TimeLines are merged in a single StoryLine. As a result, the StoryLine contains the events of both entities. The events are represented by the ID of the file, the ID of the sentence, the extent of the event mention and the participants (i.e. entities) of the event.

3.1 Dataset

As a proof-of-concept, we start from the dataset provided in SemEval-2015. It is composed of 120 Wikinews articles grouped in four different corpora about Apple Inc.; Airbus and Boeing; General Motors, Chrysler and Ford; and Stock Market. The Apple Inc. set of 30 documents serve as trial data and the remaining 90 documents as the test set.

We have considered each corpus a topic to extract StoryLines. Thus, for each corpus, we have merged the interacting individual TimeLines to create a gold standard for StoryLines. As a result of this process, from a total of 43 TimeLines we have obtained 7 gold-standard StoryLines. Table 1 shows the distribution of the StoryLines and some additional figures about them. *Airbus*, *GM* and *Stock* corpora are similar in terms of size but the number of gold StoryLines go from 1 to 3. We also obtain 1 StoryLine in the *Apple Inc.* corpus, but in this case the number of TimeLines is lower. The number of events per StoryLine is quite high in every corpus, but the number of interacting events is very low. Finally, 26 out of 43 target entities in SemEval-2015 belong to a gold StoryLine. Note that in real StoryLines all interacting

	Apple Inc.	Airbus	GM	Stock	Total
timelines from SemEval	6	13	11	13	43
storylines	1	2	1	3	7
events	129	135	97	188	549
events / storyline	129	67.5	97	62.7	78.4
interacting-events	5	12	2	11	30
interacting-events / storyline	5	6	2	3.7	4.3
entities	4	9	4	9	26
entities / storyline	4	4.5	4	3	3.7

Table 1: Figures of the StoryLine gold dataset.

entities should be annotated whereas now we only use those already selected by the TimeLines task.

3.2 Evaluation

The evaluation methodology proposed in SemEval-2015 is based on the evaluation metric used for TempEval-3 (UzZaman et al., 2013) which captures the temporal awareness of an annotation (UzZaman and Allen, 2011). For that, they first transform the TimeLines into a set of temporal relations. More specifically, each time anchor is represented as a TIMEX3 so that each event is related to the corresponding TIMEX3 by means of the SIMULTANEOUS relation. In addition, SIMULTANEOUS and BEFORE relation types are used to connect the events. As a result, the TimeLine is represented as a graph and evaluated in terms of recall, precision and F1-score.

As a first approach, the same graph representation can be used to characterize the StoryLines. Thus, for this trial we reuse the same evaluation metric as the one proposed in SemEval-2015. However, we already foresee some issues that need to be addressed for a proper StoryLines evaluation. For example, when evaluating TimeLines, given a set of target entities, the gold standard and the output of the systems are compared based on the F1 micro average scores. In contrast, when evaluating StoryLines, any entity appearing in the corpus is a candidate to take part in a StoryLine, and several StoryLines can be built given a set of related documents. Thus, we cannot compute the micro-average of the individual F1-scores of each StoryLine because the number of StoryLines is not set in advance. In addition, we also consider necessary to capture the cases in which having one gold standard StoryLine a system obtains more than one StoryLine. This could happen when a system is not able to detect all the entities interacting in events but only some of them. We consider necessary to offer a metric which takes into account this type of outputs and also scores partial StoryLines. Obviously, a deeper study of the StoryLines casuistry will lead to a more complete and detailed evaluation metric.

3.3 Example of a system-run

In order to show that the dataset and evaluation strategy proposed are ready to be used on StoryLines, we follow the strategy described to build the gold annotations to implement an automatic system. This way, we create a simple system which merges automatically extracted TimeLines. To build the TimeLines, we use the system which currently obtains the best results in Track A (Laparra et al., 2015). The system follows a three step process to detect events, time-anchors and to sort the events according to their time-anchors. It captures explicit and implicit time-anchors and as a result, it obtains 14.31 F1-score.

Thus, for each target entity, we first obtain the corresponding Timeline. Then, we check which TimeLines share the same events. In other words, which entities are co-participants of the same event and we build StoryLines from the TimeLines sharing events. This implies that more than two TimeLines can be merged into one single StoryLine.

The system builds 2 StoryLines in the *Airbus* corpus. One StoryLine is derived from the merging of the TimeLines of 2 target entities and the other one from the merging of 4 TimeLines. In the case of the *GM* corpus, the system extracts 1 StoryLine where 2 target entities participate. For the *Stock* corpus, one StoryLine is built merging 3 TimeLines. In contrast, in the *Apple* corpus, the system does not obtain any StoryLine. We eval-

uated our StoryLine extractor system in the cases where it builts StoryLines. The evaluation results are presented in Table 2.

Corpus	Precision	Recall	Micro-F
Airbus	6.92	14.29	4.56
GM	0.00	0.00	0.00
Stock	0.00	0.00	0.00

Table 2: Results of the StoryLine extraction process.

Based on the corpus, the results of our strategy vary. The system is able to create StoryLines which share data with the gold-standard in the *Airbus* corpus, but it fails to create comparable StoryLines in the *GM* and *Stock* corpora. Finding the interacting events is crucial for the extraction of the StoryLines. If these events are not detected for all their participant entities, their corresponding TimeLines cannot be merged. For that reason, our dummy system obtains null results for the *GM* and *Stock* corpus.

However, this is an example of a system capable of creating StoryLines. Of course, more sophisticated approaches or approaches that do not follow the TimeLine extraction approach could obtain better results.

4 Conclusions and future work

We have proposed a novel approach to define StoryLines based on the shared data provided by the NewsStory workshop. Basically, our initial approach extends the pilot TimeLines evaluation task carried out recently in SemEval 2015. Our proposal defines a StoryLine as a group of interacting entity TimeLines. In particular, a StoryLine grups together the events corresponding to multiple but interacting TimeLines. Thus, several separate StoryLines can be derived from a news topic, each one corresponding to a set of interacting entity TimeLines.

As a proof-of-concept, we derive a gold-standard StoryLine dataset from the gold standard TimeLines provided by the pilot SemEval-2015 task. We also present a very basic system that tries to capture the StoryLines that appear in the original documents of the TimeLines task. As the same graph representation is valid for both TimeLines and StoryLines, we directly apply to our StoryLines the evaluation measure and system provided by the TimeLine pilot SemEval-2015 task. The gold StoryLines datasets are publicly available.[3]

Based on our initial findings, we foresee two major issues that need to be addressed. First, given a set of documents, the gold standard StoryLines require to annotate all the named entities participating in the StoryLine. That is, annotating the relevant events and entities interacting in the documents. Second, our proposal still needs to devise a more complete evaluation metric for properly evaluating StoryLines.

Acknowledgments

We are grateful to the anonymous reviewers for their insightful comments. This work has been partially funded by SKaTer (TIN2012-38584-C06-02) and NewsReader (FP7-ICT-2011-8-316404), as well as the READERS project with the financial support of MINECO, ANR (convention ANR-12-CHRI-0004-03) and EPSRC (EP/K017845/1) in the framework of ERA-NET CHIST-ERA (UE FP7/2007-2013).

References

James Allan. 2002. Introduction to topic detection and tracking. In *Topic detection and tracking*, pages 1–16. Springer.

Regina Barzilay, Kathleen R McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 550–557. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio, June.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 602–610, Suntec, Singapore.

Quang Xuan Do, Wei Lu, and Dan Roth. 2012. Joint inference for event timeline construction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 677–687. Association for Computational Linguistics.

[3] `http://adimen.si.ehu.es/~laparra/storylines.tar.gz`

Po Hu, Min-Lie Huang, and Xiao-Yan Zhu. 2014. Exploring the interactions of storylines from informative news events. *Journal of Computer Science and Technology*, 29(3):502–518.

Egoitz Laparra, Itziar Aldabe, and German Rigau. 2015. Document level time-anchoring for timeline extraction. In *Proceedings of ACL-IJCNLP 2015*, page to appear.

Michael Matthews, Pancho Tolchinsky, Roi Blanco, Jordi Atserias, Peter Mika, and Hugo Zaragoza. 2010. Searching through time in the new york times. In *Proc. of the 4th Workshop on Human-Computer Interaction and Information Retrieval*, pages 41–44. Citeseer.

Arturas Mazeika, Tomasz Tylenda, and Gerhard Weikum. 2011. Entity timelines: visual analytics and named entity evolution. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2585–2588. ACM.

Anne-Lyse Minard, Alessandro Marchetti, Manuela Speranza, Bernardo Magnini, Marieke van Erp, Itziar Aldabe, Rubén Urizar, Eneko Agirre, and German Rigau. 2014. TimeLine: Cross-Document Event Ordering. SemEval 2015 - Task 4. Annotation Guidelines. Technical Report NWR2014-11, Fondazione Bruno Kessler. `http://www.newsreader-project.eu/files/2014/12/NWR-2014-111.pdf`.

Anne-Lyse Minard, Manuela Speranza, Eneko Agirre, Itziar Aldabe, Marieke van Erp, Bernardo Magnini, German Rigau, and Ruben Urizar. 2015. Semeval-2015 task 4: Timeline: Cross-document event ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 778–786, Denver, Colorado, June 4–5.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, March.

Dafna Shahaf and Carlos Guestrin. 2010. Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–632. ACM.

Dafna Shahaf, Jaewon Yang, Caroline Suen, Jeff Jacobs, Heidi Wang, and Jure Leskovec. 2013. Information cartography: creating zoomable, large-scale maps of information. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1097–1105. ACM.

Russell Swan and James Allan. 2000. Automatic generation of overview timelines. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–56. ACM.

Naushad UzZaman and James Allen. 2011. Temporal evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 351–356, Portland, Oregon, USA.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 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)*, SemEval '13, pages 1–9, Atlanta, Georgia, USA.