From Unstructured Data to Narrative Abstractive Summaries

Estela Saquete Boró
Department of Software and Computing Systems, University of Alicante,
Apdo. de Correos 99 E-03080, Alicante, Spain
https://www.dlsi.ua.es/eines/membre.cgi?id=cas&nom=stela
stela@dlsi.ua.es

Abstract

To provide with easy and optimal access to digital information, narrative summaries must have a coherent and natural structure. Depending on how a summary is produced, a distinction can be made between extractive and abstractive summaries. Using an abstractive summarization approach, the relevant information (e.g., who? what?, when?, where?,...) could be fused together, leading to the generation of one or more new sentences. However, in order to do this it is necessary to obtain and process the temporal information in a text. A very effective way is the generation of timelines starting from multiple documents so that the generation of summaries is supported by the generated timeline, without losing the relevant temporal information of the texts. In this proposal, a enriched timeline is generated automatically, and the process of generating abstractive summaries is presented using this timeline as a basis [1]. Finally, potential applications of the automatic timeline generation would be presented, as for example its application to Fake News detection.

2012 ACM Subject Classification Applied computing → Document management and text processing

Keywords and phrases Narrative summarization, Abstractive summarization, Timeline Generation, Temporal Information Processing, Natural Language Generation

Digital Object Identifier 10.4230/LIPIcs.TIME.2019.2

Category Invited Talk

Funding This research work has been partially funded by the projects PROMETEU/2018/089 and RTI2018-094653-B-C22.

1 Introduction

As human beings, we tend to organize the flux of happening in structured units known as events. Each event is a fact that occurs in the (real or imaginary) world with a specific structure (the event structure), and denotes processes, activities, states, achievements or accomplishments [7]. An event involves participants [3] and other components that complete the event such as time, place, instruments, patients, etc. In ISO TimeML Working Group[2], the event was defined as "something that can be said to obtain or hold true, to happen or to occur". Moreover, relating and ordering the information extracted from *different documents* is an essential task to obtain this knowledge. This cross-document processing improves the traditional single-document extraction and uses information redundancy to its advantage.

Hence, event ordering is a crucial task within Natural Language Processing. Cross-Document Event Ordering implies the accomplishment of three sub-tasks [10]. First of all, the *extraction of events* and related entities from texts, because it is necessary to know which events appear in each document, and which entities are related to each one of them. Then *temporal information processing* is required in order to extract the temporal expressions and the temporal relationships established between these events, determining thus which events happen at the same time. Finally, *cross-document event coreference* is needed in order to cluster all the mentions that refer to the same event, regardless of the words used to express

them. The final aim of combining event extraction and temporal information processing with cross-document event coreference enables us to automatically build ordered timelines of events from written texts.

To provide users with relevant information, summaries must provide a coherent and natural structure [5]. Text summarization allows to condense the relevant information of different documents (e.g. news) [6].

The main objective of this work is to demonstrate how the use of automatic timelines can benefit multiple NLP applications, including the generation of narrative abstractive summaries based on a natural time ordering of events from a set of documents (news in this case) that deal with the same real events, as it was described in depth in [1]. This approach has two main components: (i) a cross-document timeline generation module that extracts events related to the same entity from several texts (cross-document) and the time slot in which each event occurs, arranging them in a timeline; and (ii) an abstractive summarization module that transforms these time-ordered events into a single text with a time-based chronological narrative structure.

2 Cross-document Enriched TimeLine Extraction

As previously explained, given a set of documents and a set of target entities, the original task of Cross-Document Timeline Extraction consists of building an event timeline for a target entity from a set of documents [8].

As presented in [11], theoretically, the main idea of our approach is that two events *e1* and *e2* will be coreferent if they are not only temporal compatible ($e1_t = e2_t$) but also if they refer to the same facts (semantic compatibility: $e1_s \simeq e2_s$):

$$coref(e1, e2) \rightarrow (e1_t = e2_t) \wedge (e1_s \simeq e2_s)$$

Our proposal extends the approach by enriching the event clusters with all the arguments extracted from these events in the different documents where they are presented (see [11] for further details). The steps of this module are:

- **Temporal clustering.** The input is a set of plain texts, and, therefore, the events in those texts must be automatically extracted. Furthermore, considering that the final aim is building a timeline, temporal expressions and temporal links between events and times are required. For this reason, the first step is performing Temporal Information Extraction and Processing, and TIPSem system [4] is used for this purpose. Considering the premise that two events mentions referring to the same event happen at the same time, and using the temporal annotation of the input texts (TimeML), the temporal clustering algorithm performs two steps:

  - *Within-document temporal clustering*: For each document, the temporal information of each event is extracted. Each event is anchored to a time anchor[1] when a temporal SIMULTANEOUS/ BEGIN/ INCLUDES link exists between this event and a temporal expression. After this, two events will be considered part of the same cluster if they are temporally compatible. This means that: a) two events are anchored to the same time anchor, or b) two events have a temporal SIMULTANEOUS link between them.

[1] A time anchor is always a DATE (as defined in TimeML) and its format follows the ISO-8601 standard.

- *Cross-document temporal clustering*: Considering that in the previous step all the events of each document were assigned to a time anchor, in this step, this information is merged in a single timeline, in which all the events of the different documents are clustered together if they are happening at the same time.

Finally, the temporal clusters are chronologically ordered.

- **Semantic clustering.** The events are clustered using event type information and distributional semantic knowledge. Two or more event mentions in the same time slot could refer to the same real event. To detect these corefential events, we have applied a clustering process based on two kinds of semantic information: the first one is the event type and the second one is the distributional semantic similarity between event mentions. During the event extraction process, each event mention has been classified according to its type of event following TimeML standard: occurrence, perception, reporting, aspectual, state, intentional state and intentional action. All the event mentions with the same time slot have been regrouped after also considering the type of event assigned. Next, our approach clusters coreferential events (identifies all the events that share the same time slot and the same type of event) according to the compositional-distributional semantic similarity between them. The semantics of the event structure is represented as a compositional-distributional vector. These vectors are called contextual vectors.
  In our approach each event structure is formed, on the one hand by the event head and, on the other hand by the nouns, verbs and adjectives of the main arguments. All this information is extracted by applying Freeling as Part of Speech tagger and Semantic Role Labeling system. Following the additive model [9], these word vectors are added in a single compositional vector that represents the distributional meaning of the whole event structure.
- **Event cluster enrichment.** In the original concept of timeline, only events are grouped together in the clusters and not the arguments involved in those events that constitute a fundamental part of the information. Therefore, in this step, all the arguments (semantic roles extracted in the previous step with Freeling) of the events in each cluster are added to the timeline, enriching the information provided for each event.

3 Abstractive Summarization

The aim of this module is to produce a narrative abstractive summary with chronological information given an enriched timeline as input. As presented in depth in [1], this summary is generated employing NLG techniques. In particular, we employ a hybrid surface realization approach, based on over-generation and ranking techniques. For each of the enriched cluster of events from the enriched timeline, the next steps are applied:

- Argument selection: in case there is more than one argument for the same semantic role, a statistical selection of the arguments from the timeline is performed.
- Obtaining verb frames: information about the frames corresponding to the verbs of each event is obtained to generate a sentence without the need to resort to grammar specifications.
- Sentence Generation: for each of the frames obtained a sentence is generated, based on the frame structure.
- Sentence Ranking: a ranking is performed for selecting only one sentence representing a specific event (cluster of event mentions) in the timeline.

4 Conclusions and Further Work

In this paper an integrated approach is presented on two basic aspects. First, it is based on the fact that humans tend to apply chronological ordering of events in the summarizing process, which implies the need for timelines [11]. Second, when using an abstractive summarization approach, rather than an extractive one, the relevant information could be fused together, leading to the generation of more complete sentences, and thus, more comprehensible and effective summaries [1]. The proposal comprises two main modules: i) Enriched Timeline Extraction module, and ii) Abstractive Summarization module.

Enriched timeline generation has multiple applications in Natural Language Processing approach. Apart from summarization, the timeline with arguments can be used to detect contradictions in different media outlets, which is one the foundamentals when detecting mis- or disinformation. Furthermore, it can be used in fact-checking purposes or misleading headlines.

References

1 Cristina Barros, Elena Lloret, Estela Saquete, and Borja Navarro-Colorado. NATSUM: Narrative abstractive summarization through cross-document timeline generation. *Information Processing & Management*, February 2019. doi:10.1016/j.ipm.2019.02.010.

2 ISO TimeML Working Group. ISO TimeML TC37 draft international standard DIS 24617-1, 2008. URL: http://semantic-annotation.uvt.nl/ISO-TimeML-08-13-2008-vankiyong.pdf.

3 Heng Ji, Ralph Grishman, Zheng Chen, and Prashant Gupta. Cross-document Event Extraction and Tracking: Task, Evaluation, Techniques and Challenges. In *RANLP*, pages 166–172. RANLP 2009 Organising Committee / ACL, 2009.

4 Hector Llorens, Estela Saquete, and Borja Navarro-Colorado. Applying Semantic Knowledge to the Automatic Processing of Temporal Expressions and Events in Natural Language. *Information Processing & Management*, 49(1):179–197, 2013.

5 Elena Lloret and Manuel Palomar. Text Summarisation in Progress: A Literature Review. *Artif. Intell. Rev.*, 37(1):1–41, January 2012. doi:10.1007/s10462-011-9216-z.

6 Inderjeet Mani and Mark T. Maybury, editors. *Advances in Automatic Text Summarization*. MIT Press, Cambridge, MA, USA, 1999.

7 Inderjeet Mani, James Pustejovsky, and Robert Gaizauskas. *The Language of Time*. Oxford University Press, Oxford, 2005.

8 Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. MEANTIME, the NewsReader Multilingual Event and Time Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, May 2016.

9 Jeff Mitchell and Mirella Lapata. Composition in Distributional Models of Semantics. *Cognitive Science*, 34:1388–1429, 2010.

10 Borja Navarro and Estela Saquete. GPLSIUA: Combining Temporal Information and Topic Modeling for Cross-Document Event Ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 820–824, Denver, Colorado, June 2015. Association for Computational Linguistics.

11 Borja Navarro-Colorado and Estela Saquete. Cross-document Event Ordering Through Temporal, Lexical and Distributional Knowledge. *Know.-Based Syst.*, 110(C):244–254, October 2016. doi:10.1016/j.knosys.2016.07.032.