An Introduction to the Summarization of Evolving Events: Linear and Non-linear Evolution

Stergos D. Afantenos[1], Konstantina Liontou[2], Maria Salapata[2], and Vangelis Karkaletsis[1]

[1] Software and Knowledge Engineering Laboratory
Institute of Informatics and Telecommunications,
National Center for Scientific Research (NCSR) "Demokritos"
{stergos,vangelis}@iit.demokritos.gr
[2] Institute of Language and Speech Processing

Abstract. This paper examines the summarization of events that evolve through time. It discusses different types of evolution taking into account the time in which the incidents of an event are happening and the different sources reporting on the specific event. It proposes an approach for multi-document summarization which employs "messages" for representing the incidents of an event and cross-document relations that hold between messages according to certain conditions. The paper also outlines the current version of the summarization system we are implementing to realize this approach.

1 Introduction

The exchange of information is of outmost importance for humans. Through the history of humankind it has taken many forms, from gossiping to the publication of news through dedicated media. More recently, the Internet has given a new perspective to this human faculty, making the exchange of information much more easy and virtually unrestricted.

Naturally this has caused some problems. Imagine, for example, that someone wants to keep track of an event that is being described on various news sources, over the Internet, as it evolves through time. The problem is that there exist a plethora of news sources making very difficult for someone to compare the different versions of the story in each source. Automatic text summarization is a solution to this information overflow problem. In this paper we propose a general framework for the automatic summarization of evolving events, *i.e.* the summarization of events that evolve through time.

A crucial question, that can possibly arise at this point, concerns the definition of the "event". In the Topic Detection and Tracking (TDT) research an event is described as "something that happens at some specific time and place" (Papka 1999, p 3; see also Allan et al. 1998). The inherent notion of time is what distinguishes the event from the more general term *topic*. For example, incidents which include hostages are regarded as topics, while a particular incident, such as the one concerning the two Italian women that were kept as hostages by an Iraqi group in 2004, is regarded as an event. In our discussion about "events" we will adopt this definition provided by the TDT research.

In the Multi-document Summarization community, a consensus that has emerged is that in order to summarize a set of related documents, one has to identify similarities and differences among the documents (Mani and Bloedorn 1999; Mani 2001; see also Endres-Niggemeyer 1998 and Afantenos, Karkaletsis, and Stamatopoulos 2005). Yet, no consensus has been reached concerning as to where those similarities and differences should be targeted. In our work we propose that the similarities and differences, at least for evolving events, should be viewed under two perspectives: *time* and *source*, through *cross-document relations*. We call *synchronic relations* those relations that are concerned with the similarities and differences, between the various sources, on the same temporal horizon and *diachronic relations* those relations that are more concerned with the evolution of an event as it is being described by one source.

Summarization of evolving events should not be confused with evolving summaries. Evolving summaries were originally proposed, but not implemented, by Radev (1999, p 149) as follows: "An evolving summary S_{k+1} is the summary of a story, numbered A_{k+1}, when the stories numbered A_1 to A_k have already been processed and presented in a summarized form to the user. Summary S_{k+1} differs from its predecessor, S_k, because it contains new information and omits information from S_k". What we propose, instead, is a framework which will enable the creation of summaries of evolving events.

Section 2 discusses the different kinds of evolution in terms of the time the incidents of an event are happening and in terms of the rate with which the various news sources are emitting their reports. Section 3 introduces the notion of messages which we use for representing the various incidents of an event. Section 4 discusses the two types of cross-document relations (synchronic and diachronic) which hold between messages. Section 5 outlines the system developed so far that realizes our approach, as well as other options we are currently investigating.

2 Kinds of Evolution

This work studies the summarization of events that evolve through time, as they are being described by various sources. In this study we came to the conclusion that we should distinguish between the evolution of an event in *time* and the *rate* of reporting about an evolving event from various sources.

Concerning the evolution of an event we distinguish between two types of evolution: *linear* and *non-linear* evolution. In linear evolution the major incidents of an event are happening in constant and possibly predictable quanta of time. This means that if the first incident q_0 happens at time t_0, then each subsequent incident q_n will come at time $t_n = t_0 + n * t$, where t is the constant amount of time with which the incidents are happening. In non-linear evolution, in contrast, we cannot distinguish any meaningful pattern in the order that the major incidents of an event are happening. This distinction is depicted in Figure 1 in which the evolution of two different events is depicted with the dark solid circles.

Linearly evolving events have a fair proportion in the world. They are related with human activities which occur at regular intervals. One such example can be the descriptions of various athletic events which occur regularly. In particular we have examined the descriptions of football matches (Afantenos et al. 2004). On the other hand, one can argue that most of the events that we find in the news stories are non-linearly evolving events. They can vary from political ones, such as elections or various international political issues, to airplane crashes or terroristic events. Currently we are investigating the domain of incidents which involve hostages.

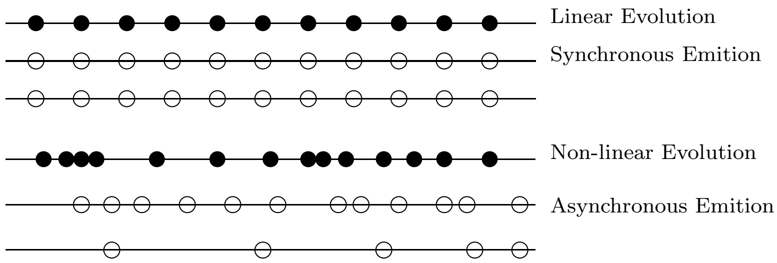

Fig. 1. Linear and Non-linear evolution

In terms of the reporting on an event from various sources we can distinguish between *synchronous* and *asynchronous* emission of reports. This distinction is depicted in Figure 1 with the white circles. In most of the cases, when we have an event that evolves linearly we will also have a synchronous emission of reports, since the various sources can easily adjust to the pattern of the evolution of an event. This cannot be said for the case of non-linear evolution, resulting thus in asynchronous emission of reports by the various sources.

In Figure 2 we represent two events which evolve linearly and non-linearly and for which the sources report synchronously and asynchronously respectively. The horizontal axis in this figure represents the number of reports per source on a particular event. The vertical axis represents the time, in minutes, that the documents are published. The first event concerns descriptions of football matches. In this particular event we have constant reports weekly, *i.e.* every 10800 minutes, from 3 different sources. The lines for each source fall on top of each other since they publish simultaneously. The second event concerns a terroristic group in Iraq which kept as hostages two Italian women threatening to kill them, unless their demands were fulfilled. In the figure we depict 5 sources. The number of reports that each source is making varies from five to twelve, in a period of time of about 23 days. As we can see from the figure, most of the sources begin reporting almost instantaneously, except one which delays its report for about twelve days. Another source, although it reports almost immediately, it delays considerably later reports.

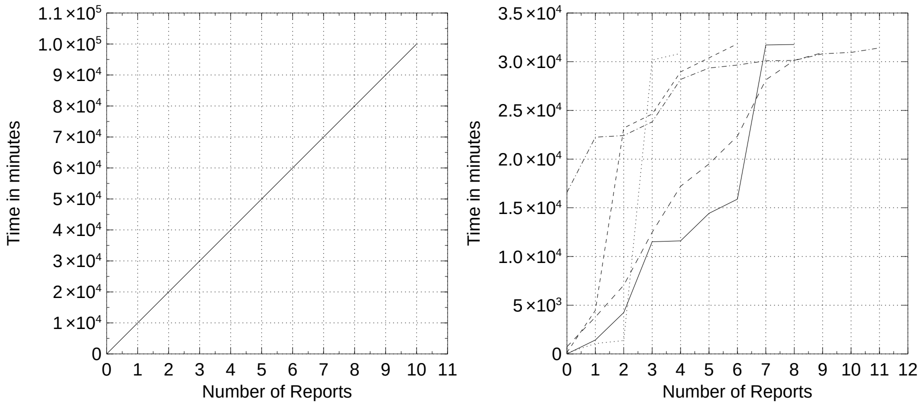

Fig. 2. Linear and Non-linear evolution

The linearity or non-linearity of an evolving event, as well as the rate of sources emission, affects our summarization approach which is based on the exploitation of the similarities and differences that exist synchronically and diachronically between the documents. The cross-document relations, and the way that they are affected by linearity, will be explained in more detail in section 4. In the following section we will concentrate on the notion of messages for representing the incidents of an event.

3 Messages

Each event is composed from various simpler incidents. For example, in the football domain, such incidents can be the performance of a player or a team, the goals that are achieved, the possible injuries of players, etc. In a domain with hostages such incidents can be the occupation of a building, the negotiations, the demands of the terrorists, the fact that they freed a hostage, etc.

We use messages to represent those incidents. Each message is composed of two parts: its *type* and a list of *arguments* which take their values from an *ontology* for the specific domain:[3]

$$\texttt{message\_type (arg}_1\texttt{, ... , arg}_n\texttt{)}$$
$$\text{where } \texttt{arg}_i \in \text{Domain Ontology}$$

[3]See Afantenos et al. (2004).

The message type represents the type of the incident, whilst the arguments represent the main entities that are involved in this incident. It is possible that some messages may be accompanied by some *constraints* on their arguments, which reflect various pragmatic constraints. These messages are similar structures (although simpler ones) with the templates used in the Message Understanding Conferences (MUC).[4]

Each message is also linked to a specific source and time. In other words, if we have a message m, then we have associated with it two extra pieces of information, m`.time` and m`.source`. Concerning the `source`, it is inherited by the document that contains the message. This cannot be said for the `time` as well, since the time of the incidents might be different from the emission time. This is expressed in the document by a temporal expression. Thus, in order to determine the `time` of a message we should interpret this expression in relation to the time of the publication of the document.

Linear		***Non-linear***	
performance (entity, in_what, time_span, value)		**negotiate** (entity$_1$, entity$_2$, about)	
entity	: Player or Team	entity$_1$	: Person
in_what	: Action Area	entity$_2$	: Person
time_span	: Minute or Duration	about	: Activity
value	: Degree		

Examples of messages' specifications, for a linear and a non-linear domain are shown in the above table. The arguments for each message come from the domain ontology. Thus, for example, the `Activity` argument in the second message corresponds to a set of activities which are defined in the ontology of the domain. The specifications for the first message come from the domain of football matches (Afantenos et al. 2004) and it represents the performance of a player or a team for a specific time-span and a specific action area (*e.g.* in the defense). The specifications of the second message come from the topic which is related with hostages, which we currently investigate. This message represents the fact that we have a negotiation between two entities concerning a specific activity (*e.g.* the release of some hostages).

4 Cross-document Relations

Cross-document relations hold between messages and are distinguished into *synchronic* and *diachronic*.

Synchronic relations try to identify the similarities and differences that two sources have, at about the same time. In the case of linear or synchronous evolution all the sources report in the same time. Thus in most of the cases the incidents described in each document refer to the time that the article was published. Yet, in some cases we might have temporal expressions in the text

[4] `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html`

that modify the time that a message might refer. In such cases, before establishing a synchronic relation, we should place this message in the appropriate time horizon. In the case of non-linear asynchronous evolution this phenomenon is predominant. Each source reports at irregular intervals, possibly mentioning incidents that happened long before the publication of the article, and which another source might have already mentioned in an article published earlier (see the second part of Figure 2). In this case we shouldn't rely any more to the publication of an article, but instead on the *time* tag that the messages have, which has been appropriately modified according to the temporal expressions found in the text. Once this has been performed, we should then establish a *time window* in which we should consider the messages, and thus the relations, as synchronic. This time window, depending on the domain, can vary from some hours to some days.

Diachronic relations, on the other hand, try to capture the similarities and differences, through time, that exist for an event as it is being described by the *same* source. In this sense, diachronic relations do not exhibit the problems of time that the synchronic relations do.

Cross-document relations, in our viewpoint, are domain dependent, since they represent pragmatic information which depends on the domain.[5] Examples of synchronic relations can be agreement, disagreement, elaboration, generalization, etc. Examples of diachronic relations can be positive or negative graduation, stability, continuation, repetition, etc.

In more formal terms, if we represent a relation r as a pair of messages $\langle m_1, m_2 \rangle$, where m_1 and m_2 are two messages, then a relation will be synchronic iff

$$m_1\texttt{.time} = m_2\texttt{.time} \text{ and } m_1\texttt{.source} \neq m_2\texttt{.source}$$

and diachronic iff

$$m_1\texttt{.time} > m_2\texttt{.time} \text{ and } m_1\texttt{.source} = m_2\texttt{.source}$$

We have to note that a relation has a directionality. As is evident, diachronically a relation can hold from a past time to a future time. In the case of a synchronic relation (*e.g.* agreement) a relation can have both directions, in which case we have in fact two relations.

In order to define a relation in a domain we have to provide a *name* for it, and describe the conditions under which it will hold. The name of the relation is in fact *pragmatic* information, which we will be able to exploit during the generation of the summary. The conditions under which a relation between two messages holds are represented in terms of values of their arguments, as well as their corresponding time and source.

Suppose, for example, that we have two identical messages. If they have the same temporal tag, but belong to different sources, then we have an *agreement*

[5]This does not mean that we do not believe that domain independent relations could not possibly exist. An example could be the relations agreement and disagreement, which can obviously be independent of domain.

relation. If, on the other hand, they have the same source but chronological distance one or higher, then we can speak, for example, of a *stability* relation. Thus we see that, apart from the characteristics that the arguments of a message pair $\langle m_1, m_2 \rangle$ should exhibit, the source and temporal distance also play a role for that pair to be characterized as a relation.

In Figure 3 we can see the difference, in terms of synchronic relations, between a domain which evolves linearly and has a synchronous emission of reports and a domain which evolves non-linearly and has asynchronous emission of reports. In the first case we have two identical `performance` messages (see the table of page 5), from two documents which have been published at the same time. Thus, and according to the specifications of the synchronic relations (Afantenos and Karkaletsis 2004), we have an *agreement* relation. In the second case we have two identical `negotiate` messages from documents that have different publication times. Yet, in the text that defines those messages, we have a temporal expression which modifies the `time` tag for one of the messages, making them refer on the same day. Thus, again we have an *agreement* relation, although the documents which contain the messages have not been published on the same day.

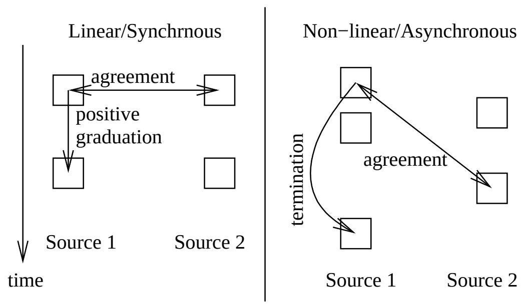

Fig. 3. Examples of synchronic and diachronic relations

In the same figure you can see two diachronic relations. In the linearly evolving case we have two `performance` messages

$$\texttt{performance (entity}_1\texttt{, in\_what}_1\texttt{, time\_span}_1\texttt{, value}_1\texttt{)}$$
$$\texttt{performance (entity}_2\texttt{, in\_what}_2\texttt{, time\_span}_2\texttt{, value}_2\texttt{)}$$

which have identical arguments, except that $\texttt{value}_1 < \texttt{value}_2$. In this case, and according to the specifications for the relations of the domain (Afantenos et al. 2004) we have a *positive graduation* diachronic relation. In the second case we have two different messages

$$\texttt{start (entity}_1\texttt{, activity}_1\texttt{)}$$
$$\texttt{end (entity}_2\texttt{, activity}_2\texttt{)}$$

where $\texttt{entity}_1 = \texttt{entity}_2$ and $\texttt{activity}_1 = \texttt{activity}_2$. In this case, according to the specifications, we have a *termination* diachronic relation. Note that in the first case we have a diachronic relation that holds between the same message types, while in the second case the diachronic relation holds between different message types. Also, in the first case the documents that contain the messages have distance one, *i.e.* the one follows immediately the other, while in the second case they have greater distance.

There may be also cases where an event is being described by one source but not from the others. Since we need at least two messages from different sources in order to have a synchronic relation, we will not connect that message with another one, thus possibly missing an important piece of information that a source is reporting. An *ellipsis* relation could be introduced to handle such cases.

5 Potential Computational Approaches

An initial study of a linearly evolving domain is presented in Afantenos et al. (2004). In Afantenos and Karkaletsis (2004) we present a system which automatically extracts the messages and the relations from the text. The messages extraction sub-system involves two processing stages, one for the identification of the messages' types and one for the filling in of its arguments. During the first stage a classifier is trained. The word lemmas and the Named Entities are used in the training vectors. The argument filling is performed using heuristics. The sub-system implementing the extraction of relations exploits the conditions under which a relation holds, as described in the specifications of each relation.

Currently we are investigating a topic which evolves non-linearly with asynchronous emission of reports, namely that of incidents involving hostages. For this topic, apart from performing the above experiments concerning the extraction of the messages and the relations, we are also implementing an algorithm which identifies the various temporal expressions in the text. This is essential since, as we have noted in sections 3 and 4 in order to identify the synchronic relation in a non-linearly evolving domain with asynchronous emission of reports, we should not rely anymore on the time an article was published. Instead we should recognize the time that a message is referring to, according to the temporal expressions which characterize this message.

Additionally, we plan to enhance our classification experiments, as well as the filling in of the messages' arguments, exploiting syntactic processing and incorporating WordNet.[6]

6 Concluding Remarks

This work has discussed the summarization of evolving events in terms of their evolution in time — linear, non-linear — and the source — synchronous, asynchronous. Of course, we are not the first to introduce the notion of time in

[6] `http://www.cogsci.princeton.edu/~wn/`

summarization. Allan, Gupta, and Khandelwal's (2001) work on temporal summarization is such a case. In their work they take the results from a TDT system for an event, and they put all the sentences one after the other in chronological order, regardless of the document that it belonged to, creating a stream of sentences. Then they apply two statistical measures, *usefulness* and *novelty*, to each ordered sentence. The aim is to extract those sentences which have a score over a certain threshold. This approach differs from ours in various ways. Firstly, they do not distinguish between the sources, while we try to incorporate in our system the different viewpoints that the various sources might have, and present them to the user. Also, they are not concerned with the evolution of the events; instead they try to detect novel information. Finally, we have an abstractive system, while they have an extractive one.

In terms of the source dimension, as far as we know, this has not been discussed elsewhere.

Another point that should be stressed concerns the use of the cross-document relations. In the past there have been several attempts to incorporate relations, in one form or another, for the creation of a summary. Radev (2000), for example, proposed the Cross-document Structure Theory (CST) which incorporated a set of 24 domain-independent relations that exist between various textual units across documents. In a later paper Zhang, Blair-Goldensohn, and Radev (2002) reduce that set to 17 relations and perform experiments with human judges. Those experiments revealed several interesting results. For example, human judges annotated only sentences, ignoring the other textual units (phrases, paragraphs, documents) that the theory suggests. Additionally, there was a rather small inter-judge agreement concerning the type of relation that connects two sentences. Nevertheless, Zhang, Otterbacher, and Radev (2003) and Zhang and Radev (2004) continue this work using Machine Learning algorithms to identify the cross-document relations. We have to note here that although some cross-document relations such as agreement and disagreement might be independent of the domain, we believe that in general cross-document relations do depend on the domain. Another difference with our work is that our relations concentrate on identifying the similarities and differences between the sources, in two different axes: *synchronically* and *diachronically*. In other words, we try to capture through those relations the points of difference between the sources, as well as the evolution of an event.

We are currently studying the summarization of non-linear events and extend our summarization system in order to improve the performance of the extraction sub-system.

References

Afantenos, Stergos D., Irene Doura, Eleni Kapellou, and Vangelis Karkaletsis. 2004, May. "Exploiting Cross-Document Relations for Multi-Document Evolving Summarization." Edited by G. A. Vouros and T. Panayiotopoulos, *Methods and Applications of Artificial Intelligence: Third Hellenic Confer-*

ence on AI, SETN 2004, Volume 3025 of *Lecture Notes in Computer Science*. Samos, Greece: Springer-Verlag Heidelberg, 410–419.

Afantenos, Stergos D., and Vangelis Karkaletsis. 2004, December. "Linear Evolving Summarization: The first Results." Technical Report 2004/6, I.I.T., N.C.S.R. "Demokritos", Athens, Greece.

Afantenos, Stergos D., Vangelis Karkaletsis, and Panagiotis Stamatopoulos. 2005. "Summarization from Medical Documents: A Survey." *Journal of Artificial Intelligence in Medicine*. In press.

Allan, James, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998, February. "Topic Detection and Tracking Pilot Study: Final Report." *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*. 194–218.

Allan, James, Rahuk Gupta, and Vikas Khandelwal. 2001. "Temporal Summaries of News Stories." *Proceedings of the ACM SIGIR 2001 Conference*. 10–18.

Endres-Niggemeyer, Brigitte. 1998. *Summarizing Information*. Berlin: Springer-Verlag.

Mani, Inderjeet. 2001. *Automatic Summarization*. Volume 3 of *Natural Language Processing*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Mani, Inderjeet, and Eric Bloedorn. 1999. "Summarizing Similarities and Differences Among Related Documents." *Information Retrieval* 1 (1): 1–23.

Papka, Ron. 1999. "On-line New Event Detection, Clustering and Tracking." Ph.D. diss., Department of Computer Science, University of Massachusetts.

Radev, Dragomir R. 1999. "Generating Natural Language Summaries from Multiple On-Line Sources: Language Reuse and Regeneration." Ph.D. diss., Columbia University.

———. 2000, October. "A Common Theory of Information Fusion from Multiple Text Sources, Step One: Cross-Document Structure." *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*. Hong Kong.

Zhang, Zhu, Sasha Blair-Goldensohn, and Dragomir Radev. 2002, August. "Towards CST-Enhanced Summarization." *Proceedings of AAAI-2002*.

Zhang, Zhu, Jahna Otterbacher, and Dragomir Radev. 2003, November. "Learning cross-document structural relationships using boosting." *Procceedings of the Twelfth International Conference on Information and Knowledge Management CIKM 2003*. New Orleans, Louisiana, USA, 124–130.

Zhang, Zhu, and Dragomir Radev. 2004, March. "Learning Cross-document Structural Relationships using Both Labeled and Unlabeled Data." *Proceedings of IJC-NLP 2004*. Hainan Island, China.