Chapter 5
Multi-modal Summarization

Tsuneaki Kato

Abstract Multi-modal summarization is a technology that provides users with abridgments of topics of interest. Such abridgments consist of organized text and informative graphics. These summarizations have two roles. One is to assist the users to review and understand their topics of interest. The other is to guide users both visually and verbally in their exploratory search. To establish this technology, it was necessary to integrate several research streams. These included information access, information extraction, and information visualization; all of these technologies had been developing rapidly since the beginning of the twenty-first century. MuST was a workshop, the main theme of which was research on multi-modal summarization of trend information. It was not an evaluation workshop and did not present the participants with a specific task, because at the time when the workshop was conducted, multi-modal summarization was merely an agglomeration of yet-to-be-developed technologies that had not yet been fully synthesized. Rather than sharing a task, the MuST workshop shared a data set. Making an annotated corpus shared as its unifying force, the workshop encouraged cooperative and competitive researches on trend information. Several innovations emerged from the workshop. These covered trend information extraction, visualization as information access interface and as data analysis method, linguistic summary generation from charts, and trend mining.

5.1 Background

By the beginning of the twenty-first century, information access technologies had changed and diversified. What was being accessed had changed from entire documents to passages within documents, and thence to the information itself. Question answering, the motto of which was to return information itself rather than pages or documents, had already progressed to managing simple factoid questions, and was
expected to reply to increasingly complicated queries such as those that included causes and definitions.

Access methods had also changed. Exploratory and interactive search was being emphasized. Information gathering was no longer a one-shot interaction through which users described their interest precisely and in return obtained adequate relevant feedback; instead, the process had become continuous, wherein users browsed information that was gathered according to general descriptions and then identified aspects regarding which they need more detailed information. Through this process, users interactively accumulated information while simultaneously expanding their area of exploration.

Methods for displaying the information so obtained had also advanced from simple ranked lists to information visualization. Some visualization techniques helped users to represent their information requests visually, others helped them to interactively analyze and interpret the results. Such information visualization techniques for information access were new and had different characteristics from those for scientific visualization.

Information was no longer simply collected or retrieved. Advances now allowed it to be compiled and synthesized using information extraction and multi-document summarization, which were techniques that had matured during that period.

Some of the research fields, such as exploratory search and information visualization, that adopted such changes in that era closely interacted with each other. This was, however, not the case for many other fields. Although one could find a limited implementation of some aspects (Ahmad et al. 2004), at that time, it was not envisioned that anything similar to the recent disaster informatics system would arise; this system synthetically processes both numeric data and linguistic data, such as documents, and summarizes and visualizes that data according to the users’ requirements. There was, however, an expectation that interactions among, and fusions of, those research fields would bring about a number of fundamental innovations.

5.2 Applications Envisioned

These anticipated fusions could take many forms. One form could lead to a sophisticated question-answering system for responding to queries such as “How have oil and gasoline prices changed this year?” or “How bad were the typhoons last year?” The system would achieve this by compiling text and statistical data and then generating combinations of succinct text and information graphics. More advanced applications of such systems may include patent or research-map generation, which would show and explain the trends of patent applications or the publication of scholarly papers. These potential developments were subsequently pursued in another NTCIR workshop, which is briefly mentioned in Sect. 5.4.

This mechanism, which we termed multi-modal summarization, can be regarded as an effort to expand text summarization. While text summarization extracts important content from a body of real-world text and presents it in a condensed form, multi-
modal summarization also processes non-linguistic information such as numerical data and information graphics. Whereas multimedia presentation generation (Fasciano and Lapalme 1996; Roth and Mattis 1990 for example), which had been actively studied at the end of the last century, aimed to generate multimedia presentations from media-independent semantic representations; multi-modal summarization does not presume the existence of such well-formed semantic representations and grapples with the enormous amount of unstructured and uncoordinated information available in the real world.

Another form of fusion supports interactive and exploratory search. It interprets and guides users’ queries linguistically and visually, progressing from the abstract to the concrete and thence to the specific. For example, initially, one may be interested in the annual movement of the oil price but later become interested in the change at a specific point in time, and finally, decide to investigate the cause and effect of that change. It also supports users’ analysis of a series of events by showing various data from several viewpoints. The occurrence of typhoons is plotted on a geographic space and time scale and then linked to data on resultant damage and its associated verbal descriptions. At least two characteristics are required for such systems to be effective. Firstly, a framework is needed that seamlessly supports users throughout the information access process, from browsing an outline or summary to subsequent elaboration or specificity and to acquiring accurate information. Secondly, linguistic and non-linguistic information could be cooperatively employed in this process. Information need not be limited to text but may include non-linguistic information such as a series of numerical values. Non-linguistic modes could be utilized even during presentation, which would then lead onward to multi-modal presentation and information visualization.

The term, multi-modal summarization, is also used for the second technology, though the name does not adequately emphasize the significance of interactivity and relationship to exploratory search. These technologies share the name because these techniques have a common core that compiles useful and relevant information and presents it to users utilizing multiple modes, including text and visuals.

5.3 Multi-modal Summarization on Trend Information

The MuST was a workshop on multi-modal summarization focused on trend information (Kato et al. 2005, 2007a, b, 2008). Why did we focus on trend information? It was because a trend, which is a general tendency in the way a situation is changing or developing, is based on temporal statistical data and can be obtained by synthetically summarizing it, but not by simple enumeration. Trends are the first answers to users’ questions such as “How has the game machine industry performed since 2006?”, “How have oil and gasoline prices changed this year?”, and “How bad were the typhoons last year?” Each answer to those questions can be considered a summary...

---

1From Longman Advanced American Dictionary.
of all the information that users are interested in and a starting point for interactive and explorative information access.

The information from which trends are composed and the process of identifying trends have several interesting features. First, to obtain trends, it is necessary to compile information spanning a specific and extensive period. As they include significant redundancies, such compilations must be synthetic and well organized. Secondly, trends usually contain summaries of non-linguistic information, for example, statistical information such as time-series data and geometric data. Some statistics such as political party approval ratings and companies’ market share of a given product type are more complicated and have other dimensions. Each dimension could be an axis representing those statistics and bring different summarization methods. Thirdly, not only information such as reports on changes in statistical data, but also their interpretation, analysis of causes, and forecasts of impacts are important and should be included when defining trends.

As trend compilation requires sophisticated processes for handling complex and diverse information, it is an important research subject for multi-modal summarization aimed at supporting interactive and explorative information access.

5.3.1 Objective

The objective of the MuST workshop was to create an agora or arena where researchers from the several fields mentioned above could interact. The workshop prioritized trend analysis as its common theme because trends have interesting characteristics that are suitable as the starting point for exploratory search and as a subject for analysis. The MuST workshop promoted both cooperative and competitive research on trend information. It was not an evaluation workshop and thus identified neither a specific task nor evaluation measures.² For many, the workshop was motivated by a common evaluation. Sometimes the objective of the workshop was to enable large-scale evaluation, which required to employ the pooling method. It is beneficial to evaluate technologies on the common ground using standard measures. That, however, is only possible when technologies have matured or when they are focused on common objectives. Research on multi-modal summarization consists of many kernels of technologies still in development and not synthesized yet. Accordingly, each research group had its specific focus. In that situation, neither a common evaluation nor shared tasks were possible or stimulating. That is why we did not conduct an evaluation-oriented workshop. We needed another motivation to make the workshop cooperative and competitive, yet still, allow the participants to focus on their interests.

The MuST workshop was conducted a bit earlier than the IEEE VAST shared-task evaluation (IEEE symposium on VAST 2006). Although both were concerned

²In its third cycle, however, some evaluation tasks were set. Those tasks were considered as shared building blocks common to trend information summarization.
with visualization technology, they were different in nature. MuST addressed various problems, rather than a substantial single problem such as the one that IEEE VAST undertook. Rather, the policy of MuST was similar to that of the interactive track held in TREC 6 (Dumais and Belkin 2005), in which, through a common experiment, the participants conducted their own studies; such individual studies are more productive than a joint evaluation. During the MuST workshop, many technologies reflecting each participant’s interests were examined. Although they would be associated with each other later in the process, initially, they did not have the same goal.

5.3.2 Data Set as a Unifying Force

Instead of a common topic for evaluation, a data set provided a unifying force for the MuST workshop. The use of a shared resource, which motivated researchers to participate and to conduct several research missions, was the major characteristic of the workshop. The resources that were shared, the MuST data set, included the materials to be processed, the intermediate results acting as the organizational hub, and the eventual output design.

The core of the data set is annotated newspaper articles concerning statistics and a wide variety of topics. The topics were drawn from disparate social and economic domains, such as the oil industry, the personal-computer market, and car production; groups of events such as earthquakes and typhoons; and organizations such as Sony Corp. Linguistic descriptions of statistics and reports on events in articles were identified and annotated, as trends would be extracted from them. For example, trends in the personal-computer industry included statistics on shipment volume, shipment value, and market share of major manufacturers. Typhoon trends consisted of a review of typhoon-related events, such as their formation, landfalls, and related damage statistics.

Examples of English texts to which the annotation schema was applied are shown in Fig. 5.1, instead of the real data, which is in Japanese. Sentences mentioning selected statistics or events are annotated as unit elements. From the text of an unit element, phrases mentioning the name of the statistic (name element), the value of the statistic (val element), the relative values, which are associated with the statistic but are not the value itself (rel element), dates (date element), and other parameters (par element) are identified and annotated.

The annotation of the MuST data set represents the intermediate result of semantic and pragmatic analysis tuned to statistical and/or event information. In the summarization, extraction and analysis of important sentences are followed by rephrasing and sentence construction to eliminate redundancy and maintain consistency. Annotation corresponds to the output of extraction and analysis and the input to rephrasing and sentence generation. Using the terminology of the information extraction field,

---

3 Articles of the Japanese Mainichi newspapers from 1998 and 1999 were used.
Based on the July 6th report announced by the Oil Information Center, the price of gasoline (one liter, regular) reached a national average of ¥92 per liter, higher than last week's average price.

The oil price (Dubai Oil) has kept dropping since its peak last October, of about $20 per barrel, and fell to $12.50 in late January. After Dubai oil price rise temporarily, Dubai oil price because of tension due to the Iraq situation, but has been struggling recently at around $10, due to oversupply.

Medium-strength typhoon No. 10 struck Makurasaki-shi, Kagoshima, about 4:30 pm on the 17th, and will strike the vicinity of Shukumo-shi, Kochi the same night.

The domestic shipment volume was $4,391,000 for the first half of the year, which is a 34% higher than for the same period last year and marked the highest level for the half-year range.

Fig. 5.1 Examples of MuST data annotations on English text

This annotation completes named-entity recognition and temporal-expression analysis. For researchers who are interested in sentence extraction or text processing on named-entity recognition and temporal-expression analysis, annotation can be referred to as the gold standard of their process. It can also be used as training data if they take a machine learning approach. For researchers interested in rephrasing, sentence generation, and information visualization, annotation can be used as input data in which several fundamental analyses are already completed. In extreme cases, studies on information visualization from the text could be conducted without text processing. In this sense, the annotated articles behave as a hub for multi-modal summarization.

Multi-modal summarization requires several component technologies that are dispersed across many research fields. This makes it difficult to construct an integrated system. By using this data set, nevertheless, the participants can address their own subjects of interest. This is especially important for those studying elemental technologies. Moreover, participants from different communities can discuss their interests with each other using the data set as common ground and can contemplate how their studies or their modules fit into the framework. Of course, researchers...
having the same interest can use the data set as material for objective evaluation. To encourage and foster research through such interchanges was the objective of sharing this research resource and of the MuST workshop.

5.3.3 Outcome

Many research themes were pursued in the MuST workshop and several technologies emerged from it. These include extraction of statistics from texts as materials for trend summarization; visualization of statistical information extracted and/or collected; generation of text that explains statistical information; and trend mining that is a version of text mining, and attempt to find and visualize trends from huge document sets.

5.3.3.1 Trend Information Extraction

Information extraction on statistics from the text was a major sub-problem of trend summarization. Many participants had addressed this problem, which is the reason that this theme was pursued in the evaluation-workshop style at the final cycle of the MuST workshop.

The simplest form of information extraction is to obtain as many tuples as possible of three elements; the name of a statistic, the date, and a value for the statistic on that date, an example of which looks like this: (Dubai oil price, 1998/12/21, $12.50). That triplet constitutes points plotted on the chart depicting the changes or trends of a given statistical category. Many complicated problems would remain even if the date and numeral expressions could be extracted using techniques of named-entity recognition. Those difficulties are epitomized in the first passage shown in Fig. 5.1, “the price of gasoline (one liter, regular), . . . , reached a national average of 92 yen, 1 yen higher than last week’s average price.”

First, the names of statistics are long and complex; they are frequently abbreviated and may be expressed in more than one way. These are usually expressed as a noun phrase, but sometimes split into many phrases. That is the case in this example in which the name of the statistic discussed is a national average of pump price of gasoline (one liter, regular). A method to handle such complex names of statistics was proposed. It deconstructs statistic names into their components and categorizes those characteristics and functions. To identify the name in its entirety, the method first identifies each component by text-chunking and then assembles those components into one name (Mori et al. 2008).

Second, not all numerical expressions directly describe the statistical values. Some of them are comparative or relative expressions. In this example, “1 yen” is not a gasoline price itself but the differential of two prices. Such relative expressions must be distinguished from direct expressions of statistical values. On the other hand, using such comparisons, an additional triplet instance of the gasoline price, “last
week,” and “91 yen” could be obtained. Methods were proposed for distinguishing those expressions and using them to obtain additional triplet instances.

Besides, relative or context-dependent time expressions such as “last week” and cases where more than one statistic is mentioned in a sentence raised problems that are still to be solved.

Other research paid attention to extraction of information beyond simple triplets. Qualitative expressions, such as “peak” and “keep dropping” in the second passage in Fig. 5.1, were used cooperatively with numerical data representations for trend summarization. Descriptions of causes of events described such as “because of the tension of the Iraq situation” are useful for understanding context. Techniques were proposed for extracting and using such descriptions for summarization and visualization purposes.

5.3.3.2 Visualization

The interactivity of visualization was a major feature identified as an objective in the MuST workshop. Interactivity allows for interactive and exploratory search. Techniques were proposed that would assist users to analyze trends from various viewpoints and provide response mechanisms for new requests that emerged from such analysis.

Figure 5.2a shows an example of visualization as information access interface (Matsushita et al. 2004). A line chart was used as an information access interface. The chart as a whole represents the changes of a statistic of interest. The data points and segments are connected to the article that describes those statistics. Users can easily go back and forth between the chart and the articles as they are interconnected. This is a technique known as brushing (Scherr 2008). In another visualization shown in Fig. 5.2b, the line chart is augmented by schematic shapes that represent qualitative changes extracted from articles, such as “rebounding” and “continuing to increase” (Matsushita and Kato 2006). This chart can also be interconnected with textual materials. This is a typical example of multi-modal summarization.

For data analysis using visualization, a framework named a visualization cube was proposed (Takama and Yamada 2009, 2010). Events such as earthquakes which are characterized by time and geographical locations have their features represented as a cube, which allows the systematic manipulation of visual representation according to changes in the user’s viewpoint. That is, a user can, through intuitive operations, freely place earthquakes of interest on a topographical map or on a timeline. Figure 5.3 schematically shows this operation. Statistics can be handled similarly. Each statistic corresponds to one cube and the cubes can be stacked upon each other. This operation corresponds to drawing a stacked bar chart. Changing the granularity of the chart or focusing on a specific data range are also defined as operations of particular cubes. Thus, it is a visualized version of an OLAP cube (Codd et al. 1993) used in online data analysis.
5.3.3.3 Linguistic Summary Generation from Charts

Summarization can be done using linguistic expressions. A typical approach is to redact long documents into succinct phrases. In multi-modal summarization, series of numbers, tables, and charts can be verbalized. This makes it possible for complex numerical dynamics to be expressed in a short descriptive phrase such as “wild gyration.”

This method was proposed for generating paragraph-length documents to explain a line chart of a given set of statistics. (Kobayashi et al. 2007; Kobayashi and Okumura 2008). The method for determining such content is critical. The chart is segmented, and a description of the relevant values and a description of the shape of the segments...
are decided and then appropriately linked to the content. The sets of two types of texts, those for describing values and those for shapes, are stored and used in the system as linguistic knowledge that is drawn from the corpus of real-life human explanations.

### 5.3.3.4 Trend Mining

Some trend summarizations can be conducted with a broader perspective via a version of text mining, which we termed as trend mining, that reveals current trends. Keywords, such as names of statistics, are linked to relevant topics. The observation that certain keywords appear frequently in documents reveals a trend that specific subjects are topical. Moreover, the co-occurrence pattern of those keywords suggests their relationship. One proposal visualized the relationship of statistical terms by calculating their co-occurrence frequencies. Such patterns are characteristic of events and phenomena in the real world (Kawai et al. 2008). The dynamic network established in this way allows users to review the structure of complex and global problems. Reviewing this, the user can discover the structure of a given problem and other useful related factors, thus facilitating access to accurate information about it.
5.4 Implication

The MuST workshop was conducted from 2005 to 2008 at the NTCIR-5, 6, and 7 workshops. It was a pilot task at first, and then became a core task with an evaluation subtask. Research activities on multi-modal summarization and trends went beyond these workshops. For five years, since 2006, special theme sessions were held at annual conferences of the Japan Society for Artificial Intelligence (JSAI). These focused on information compilation (Kato and Matsushita 2006), which aimed at using multi-modal summarization as an interface for interactive information access. It was emphasized that linguistic and non-linguistic information should be managed and utilized seamlessly. In 2009, a special interest group of the same name was launched by the JSAI. In 2012, it was renamed to Interactive Information Access and Visual Mining, and its activities have continued to the present (SIG-AM 2020).

In the NTCIR workshops, at NTCIR-8, an evaluation task was conducted on interactive information access using visual information (Kato et al. 2011). The patent information mining task in NTCIR-8 also handled text data and numerical data and extracted some trends observed in patent information (Nanba et al. 2010).

It is doubtful whether the MuST workshop itself had any direct influence on subsequent research trends. The workshop, however, contributed to advancing research on information access. Explanatory search has since become a key research area. Visual interfaces are an important component of such research. The MuST workshop was a significant catalyst in these developments.

Acknowledgements The author thanks Mitsunori Matsushita and Noriko Kando, co-organizers of the MuST workshop, for their contribution to organizing the workshop. The author also thanks all the participants of the workshop for their valuable research efforts on multi-modal summarization.

References

Ahmad S, de Oliveria PCF, Ahmad K (2004) Summarization of multimodal information. In: Proceedings of LREC-2004, pp 1049–1052
Codd EF, Codd SB, Salley CT (1993) Providing OLAP to user-analysts: an IT mandate. Technical report, E. F. Codd and Associates
Dumais ST, Belkin NJ (2005) The TREC interactive tracks: putting the user into search. In: Voorhees EM, Harman DK (eds) TREC experiment and evaluation in information retrieval. The MIT Press, Cambridge, pp 123–152
Fasciano M, Lapalme G (1996) Postgraphe: a system for the generation of statistical graphics and text. In: Proceedings of 8th international workshop on natural language generation, pp 51–60
IEEE symposium on VAST (visual analytics science and technology) 2006 (2006). http://www.cs.umd.edu/hcil/VASTcontest06/
Kato T, Matsushita M, Kando N (2005) MuST: a workshop on multimodal summarization for trend information. In: Proceedings of the 5th NTCIR workshop meeting on evaluation of information access technologies, pp 556–563
Kato T, Matsushita M, Kando N (2007a) Expansion of multimodal summarization for trend information –report on the first and second cycles of the MuST workshop–. In: Proceedings of the 6th NTCIR workshop meeting, pp 235–242
Kato T, Matsushita M, Kando N (2007b) Fostering multi-modal summarization for trend information. In: Proceedings of KES2007, pp 377–386
Kato T, Matsushita M, Kando N (2008) Overview of MusT at the NTCIR-7 workshop –challenges to multi-modal summarization for trend information–. In: Proceedings of the 7th NTCIR workshop meeting, pp 475–488
Kato T, Matsushita M (2006) Toward information compilation (in Japanese). In: Proceedings of the 20th annual conference of the Japan society for artificial intelligence, ID3-2
Kato T, Matsushita M, Joho H (2011) Overview of the VisEx task at NTCIR-9. In: Proceedings of the 9th NTCIR workshop meeting, pp 526–532
Kawai H, Kunieda K, Yamada K, Saito H, Tsuchida M, Mizuguchi H (2008) Visualization for statistical term network in newspaper. In: Proceedings of the 7th NTCIR workshop meeting, pp 549–554
Kobayashi I, Watanabe C, Okumura N (2007) Intelligent information presentation based on collaboration between 2D chart and text-with an example of Nikkei stock average text and its 2D charts presentation (in Japanese). Trans Inform Proc Soc Jpn 48(3):1058–1070
Kobayashi I, Okumura N (2008) Text generation for explaining the behavior of 2D charts: with an example of stock price trends. In: Proceedings of the 7th NTCIR workshop meeting, pp 515–519
Matsushita M, Nakakoji K, Yamamoto Y, Kato T (2004) InTREND: an interactive tool for reflective data exploration through natural discourse. In: Proceedings of KES2004, vol 2, pp 148–155
Matsushita M, Kato T (2006) Statistical chart generation from multiple documents based on numerical data supplement and chart shape suggestion (in Japanese). Trans Jpn Soc Fuzzy Theory Intell Inform 18(5):721–734
Mori T, Fujioka A, Murata I (2008) Automated extraction of statistical expressions from text for information compilation (in Japanese). Trans Jpn Soc Artif Intell 23(5):310–318
Nanba H, Fujii A, Iwayama M, Hashimoto T (2010) Overview of the patent mining task at the NTCIR-8 workshop. In: Proceedings of the 8th NTCIR workshop meeting, pp 293–302
Roth SF, Mattis J (1990) Data characterization for intelligent graphic presentation. In: Proceedings of conference on human factors in computing systems, pp 193–200
Scherr M (2008) Multiple and coordinated views in information visualization. In: Trends in information visualization, vol. 38, pp 1–8
SIG-AM (special interest group on interactive information access and visual mining) (in Japanese) (2020). https://must.c.u-tokyo.ac.jp/sigam/
Takama Y, Yamada T (2009) Visualization cube: modeling interaction for exploratory data analysis of spatiotemporal trend information. In: Proceedings of IW2009 (WI-IAT2009), pp 1–4
Takama Y, Yamada T (2010) Application of visualization cube to analysis of interaction pattern in exploratory data analysis of spatiotemporal trend information. In: Proceedings of ISCI2010, pp 85–93

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.