# Video Summaries for Science Talks

**Paulo César Dias Figueiredo**

Thesis to obtain the Master of Science Degree in

## Information Systems and Computer Engineering

Supervisors: Doctor David Manuel Martins de Matos
Doctor Ricardo Daniel Santos Faro Marques Ribeiro

### Examination Committee

Chairperson: Doctor António Manuel Ferreira Rito da Silva
Supervisor: Doctor David Manuel Martins de Matos
Member of the Committee: Doctor Bruno Emanuel da Graça Martins

**October 2015**

# Acknowledgements

Foremost, I would like to thank my supervisor David Martins de Matos for the continuous support, motivation, and criticism. His guidance and knowledge allowed me to pursue this research area in an adventurous manner.

I would like to thank my co-supervisor Ricardo Daniel Ribeiro for the help and support throughout my research.

I would like to thank all the people that invested time in the process of evaluation, their meaningful input assisted me in the last steps of this work.

I would like to thank my colleagues at $L^2F$ for the stimulating discussions and ideas that enriched my work. I thank my colleague Ana Marta Aparício for helping me write those papers and being there in those late night works.

Last but not least, I would like to thank my family, and all my friends. A special thanks to Susana Ferreira for accompanying me throughout my academic years, for all those valuable discussions, and endless friendship.

Lisboa, November 2, 2015

Paulo César Dias Figueiredo

For my family

# Resumo

Apresentações sobre ciência variam desde conferências a aulas e documentários. No entanto, apesar de alguns destes formatos conterem conteúdo importante, como as aulas, a sua apresentação não é sempre a mais apelativa. Por outro lado, documentários e outros programas de entretenimento podem apresentar informação consolidada beneficiando esteticamente de efeitos visuais e sonoros. Motivados por estes aspetos, ilustramos os tópicos de uma aula de física através da utilização de documentários. Temos como preocupações principais o processo de seleção de conteúdo e aspetos de coerência.

O processo de criação utiliza uma transcrição sumarizada de uma aula de física e seleciona conteúdo topicamente relacionado a partir de uma coleção de documentários. O problema da coerência é tratado como consequência do processo de seleção de conteúdo, e por meio de técnicas de composição de vídeo através da sua organização. Neste sentido, exploramos dois métodos distintos para composição: um mecanismo de ranking para selecionar os segmentos mais importantes; e, incorporar aspetos de fluidez ao modelar estes segmentos como curvas suavizadas e selecionar aquelas cujas pontas são geometricamente mais próximas.

Num estudo preliminar nós avaliamos o comportamento de filmes e documentários em sumarização, usando como referência o comportamento conhecido de artigos de notícias. Mostramos que a sumarização de documentários tem melhor desempenho em comparação com filmes, comparando os sumários gerados contra sumários abstrativos usando as métricas ROUGE. Para os vídeos gerados sobre ciência, os nossos dados são o resultado de julgamento subjetivo humano. Composição com ranking atinge melhores resultados do que curvas simpliciais para manter a relação de tópicos da aula original e qualidade geral. Concluímos que nenhum dos métodos realmente gere coerência.

# Abstract

Scientific talks range from conference presentations to lectures and documentaries. Even though some formats have high-grade content, such as lectures, their presentation is not always the most appealing. On the other hand, documentaries and other entertainment-driven programs can present consolidated information benefiting from aesthetically assured visual and audible effects. Motivated by these aspects, we illustrate the topics of a physics lecture through the use of documentaries. Our main concerns are the process of content selection and coherence aspects.

The creation process uses a summarized transcript of a physics lecture to select topic-related content from a collection of documentaries. The issue of coherence is addressed as a consequence of the content selection process, as well as by means of video composition techniques via content organization. In this sense, we explore two distinctly-purposed methods for composition: a ranking mechanism to select the most important segments; and, incorporating aspects of fluidity by modeling these segments as smooth curves and choosing those with geometrically closest end to start points.

In a preliminary study we assess the behavior of films and documentaries in summarization, using the well–known behavior of news articles as reference. We show that summarization of documentaries achieves higher performance in comparison to films, comparing the generated summaries against abstracts using the ROUGE metrics. For science talks videos, our results concern data from human subjective judgments. Composition with ranking achieves higher ratings than simplicial curves for topic-relation to the lecture and overall quality. We conclude that none of the methods actually addressed coherence.

# Palavras Chave
# Keywords

## *Palavras Chave*

Sumarização Automática

Sumarização de Texto

Sumarização Multi-Documento

Sumarização de Vídeo

Seleção de Conteúdo

Coerência

## *Keywords*

Automatic Summarization

Text Summarization

Multi-Document Summarization

Video Summarization

Content Selection

Coherence

# Index

# List of Figures

# List of Tables

# 1 Introduction

*Science is more than a body of knowledge. It is a way of thinking; a way of skeptically interrogating the universe with a fine understanding of human fallibility.*

*– Carl Sagan*

Scientific talks have become an indispensable forum of communication to disseminate ideas and research to different groups of people. There are several means through which these talks can take place, varying from symposia, colloquia, or other types of conference presentations to lectures and audio-visual media, such as documentaries. However, even though some formats have high-grade content, such as lectures, their presentation is not always the most appealing. In particular, they are suited to transmit systematic knowledge instead of engaging the audience. On the other hand, documentaries and other entertainment-driven programs can provide an enhanced, or stylized, view of different aspects of science. As such, they can provide consolidated information benefiting from aesthetically assured visual and audible effects inherited from filmmaking.

Since the primordials of filmmaking, the entertainment industry has focused their efforts on the production of cinematic portrayals of reality. A practice that was later baptized as documentary (Grant and Sloniowski 1998). Today, they continue to portray these aspects and give cinematic embodiment to historical events, argumentation, and research. In particular in the context of science, documentary filmmaking became a vehicle through which not only basic principles were illustrated, but also the latest scientific discoveries. Still, documentaries are entertainment-driven programs, with an overall intent to educate the general public. For this reason, they are showcased in an organized and conceptually-understandable form. Many of the strategies and narrative structures employed in fictional films are commonly used for documentary films as a mechanism for addressing these real-life issues (Nichols 1991). As such, documentaries incorporate concepts of character development, continuity, as well as montage editing.

Lectures, on the other hand, are formal talks, and their primary concern is the transmission of information to an audience (Bligh 2000). They are strictly structured around a syllabus and present information concerning a specific issue with a set of topics. In recent years, there was an unprecedented rise in modern learning methods. Several academic institutions started recording lectures and making these videos available to the general public through OpenCourseWare (OCW) and other e-learning platforms. The proliferation of such content by institutions such as Massachusetts Institute of Technology (MIT) OCW, Open Yale Courses, Berkeley Webcasts, and other online repositories (VideoLectures.net), has stemmed numerous research to interpret, search, and retrieve information from these videos.

Given these resources, our focus is on the development of a computational method that illustrates the main subjects addressed in a physics lecture, through the use of a collection of scientific documentaries. These type of courses involve the study of various interdisciplinary areas. Broadly, they focus on the analysis of nature, on understanding the behavior of the universe, or real-world phenomena (Feynman, Leighton, and Sands 1963). Naturally, the fundamental principles taught can be exemplified in a variety of ways. Thereafter, scientific documentaries can provide the content to exemplify them. Our process is oriented for human consumption. For this reason, several elements regarding the understandability of the multimedia artifact produced must be taken into account, such as the process of selecting content and aspects of coherence. In that sense, the video produced should be composed by a sequence of segments that correspond to parts that make-up a narrative, following some intent (Branigan 1992). In this work, as usual in text-to-text generation, the creation process is guided by an original document, the input. In this manner, content selection is driven by the text stream, which corresponds to the subtitles of the input document. We focus on obtaining a concise, yet diverse, representation of the lecture comprising different topics. This task is accomplished via summarization. Coherence is addressed through video composition techniques, as well as by ensuring topic similarity between the lecture and the segments retrieved from documentaries, with the purpose of obtaining a thematically-coherent video.

In this thesis, we present work concerning the generation of video summaries. We review automatic summarization as an approach for content selection. As such, we present generic summarization algorithms, common terminology in the field, as well as evaluation methods broadly categorized into intrinsic and extrinsic approaches. The issue of coherence is

addressed by methods for predicting local measures of coherence through the relatedness of adjoining text units in a high-dimensional space, uncovering topical patterns in collections of data, and preserving sequential information in text data. We describe our approach for computing lecture-driven science talks videos, considering the above aspects. For evaluation, we present a preliminary study intended to assess the performance of generic summarization algorithms when applied to the domain of films and documentaries. Then, we progress towards the evaluation of lecture-driven science talks videos involving human subjective judgments.

## 1.1 Generation of Video Summaries

Multimedia artifacts intended for human consumption typically involve techniques for retrieving, or selecting, appropriate content from the input, as well as dealing with the organization of the segments that compose it. As such, aspects regarding the understandability, and coherence, should be taken into consideration in the generation process. There is a considerable body of work in this context, where techniques such as summarization are used only as part of the content selection process. In addition, depending on the purpose and domain, coherence can be achieved via several criteria. In automatic music video generation procedures, clips from a video can be assembled to match a piece of music, through synchronization (Xu, Shao, Maddage, and Kankanhalli 2005; Yoon, Lee, and Byun 2009). Other approaches, such as automatic generation of movie trailers, compose the artifact with a representation of the movie's action-related content to obtain a genre-specific, and therefore coherent, trailer (Chen, Rizvi, Özsu, and Tamer 2003).

### 1.1.1 Content Selection

Content can be extracted according to different purposes. Techniques for video summaries are typically explored in a genre-specific manner focusing on low-level features (Li and Merialdo 2014). In order to circumvent the complex task of full semantic understanding, there is work developed for generic video summarization based on modeling the viewer's attention (Ma, Lu, Zhang, and Li 2002; Ma, Hua, Lu, and Zhang 2005). The model captures how the user's attention is attracted to motion, objects, audio, and language, when watching a video program. These elements are integrated to generate a user attention curve for a given video

sequence. This curve is composed of a time series of attention values associated to each frame of the video. Through smoothing, as well as normalization, crests of the curve are identified, which represent the video segments that are presumed to captivate the user's attention. In this manner, content selection is performed based on the peaks of those crests.

Furthermore, a common challenge associated to summarizing general videos is the lack of prior knowledge concerning its main subject matter. Whereas most approaches follow a somewhat conventional scheme for determining important content, based on content frequency (Zhuang, Rui, Huang, and Mehrotra 1998), or non-redundancy (Cong, Yuan, and Luo 2012; Zhao and Xing 2014), more recent work has addressed this issue by using title-based summarization (Song, Vallmitjana, Stent, and Jaimes 2015). Building on the conjecture that the video title is carefully chosen to provide an accurate description regarding the video's main subject, title-based image search results can be collected and used to select representative shots from the video.

For entertainment-driven programs, such as movies, internal and external textual descriptions, can be exploited to aid the selection of important content. In work concerning semantic-based adaptive movie summaries, Wikipedia is used as an external knowledge base to lessen the gap between a user statement and the corresponding movie's subtitles (Ren, Misra, and Jose 2010). The system receives as input a user description, which is used to locate a movie item from Wikipedia. After the video's subtitles are collected, the generative probabilistic topic model LDA is used to estimate topic distributions in both subtitles and the Wikipedia item. As a result of matching both topic distributions, story boundaries in the video stream are identified. A saliency-based summarization system then selects perceptually attractive segments from each topic for summary composition, taking into account prior requirements on summary length. Other approaches for movie summarization also exploit subtitles (text), along with information from the audio and the visual streams, integrating cues from these sources in a multimodal saliency curve (Evangelopoulos, Zlatintsi, Potamianos, Maragos, Rapantzikos, Skoumas, and Avrithis 2013). Auditory saliency is determined by cues that compute multifrequency waveform modulations, visual saliency is calculated using intensity, color, and orientation values, and textual saliency is obtained through Part-Of-Speech (POS) tagging.

Informational presentations, such as lectures, are often accompanied with slides, which provide digested lecture material. The amount of available instructional videos in academic in-

stitutions motivated efforts for automatic creation of online-video presentation summaries (Ju, Black, Minneman, and Kimber 1998; Balagopalan, Balasubramanian, Balasubramanian, Chandrasekharan, and Damodar 2012). In spite of several proposed methods for summarizing video (Truong and Venkatesh 2007), the intrinsic nature of lecture recordings renders most techniques inappropriate. This is due to the amateur nature of these types of videos, on the account of their lack of variation in color, and absent, or very few, transitions in shots. In this sense, most of the useful information is comprised within the audio channel. Some techniques proposed, take this into consideration and exploit information available in the audio signal, such as pitch and pause information, knowledge concerning slide transitions in the presentation, as well as information regarding access patterns of previous users (He, Sanocki, Gupta, and Grudin 1999). More recent work, focused on the extraction of metadata information in this domain (Balasubramanian, Doraisamy, and Kanakarajan 2015). The process extracts topic-based segments, which represent portions of the lectures devoted to a particular topic and utilizes features from audio transcripts, as well as slide content from the video stream.

### 1.1.2 Coherence

Even though objective measures to assess coherence are lacking, this property is an essential part of consumer-oriented multimedia artifacts. In these terms, most approaches make use of subjective user studies to determine the quality of the artifact (Xu, Shao, Maddage, and Kankanhalli 2005). Moreover, existing methods assess aspects of coherence in different ways, including criteria such as clarity, conciseness, and an overall sense of quality. Especially in the context of extractive summarization techniques, the problem of coherence is widely known (Paice 1990).

In the domain of music video generation, the visual content plays a fundamental role in complementing the auditory melody of the song. In some methods, the video is composed through the use of web images, where the song's lyrics are used as a means to provide the keywords for searching, and retrieving, relevant image data (Wu, Xu, Qiao, and Tang 2012). Coherence can be addressed by aligning the images, which are related to the lyrics, with the audio of a given song. The relationship between image and music is accomplished using a ranking mechanism, where each music segment is compared, via a similarity function, to all the images, in order to obtain the most similar image candidates. In addition, to make the video

smoother, a similarity constraint is used to ensure that the comparison of successive images is in a reasonable range. Other approaches, bridge music and video by utilizing the affective content carried out in the multimedia signal, with the purpose of providing an emotionally-coherent music video (Wang, Yang, Wang, and Jeng 2012).

In audio-video presentations, such as lectures, the focus is on the context of the information presented, as well as the flow between the segments that compose the video (He, Sanocki, Gupta, and Grudin 1999). In such approaches, these properties are mostly ensured as a consequence of the content selection process. In informational talks, the audio carries much of the content in the presentation. Therefore, pitch activity can be explored to identify the speaker's emphasis in the presentation and pause information, to avoid including segments that start in the middle of a sentence. User studies typically involve studying the importance of the segments chosen for the summary, the amount of material covered, as well as other aspects regarding context and fluidity of the segments which compose the video.

From the perspective of entertainment-driven programs, efforts have been made to generate film trailers, which content is selected concerning the most emotional film segments (Irie, Satou, Kojima, Yamasaki, and Aizawa 2010). Impressive speech and video segments are extracted by using an affective content analysis technique. The trailer is composed by a set of shots, arranged to maximize the impact of the sequence, along with the film's theme music, logo, and impressive speech segments to complement the visual content of the video. For automatic generation of documentaries, some approaches encode constraints and rules, which follow a set of strategies based on the language of cinematography from annotated audio commentaries (Rocchi and Zancanaro 2003). Discourse structure is used to plan video segmentation, camera movements, and transition effects among shots, to reflect the rhythm of the audio commentary. Using rhetorical structure theory, coherence is addressed by the relations among portions of text in the nodes of a rhetorical tree (Mann and Thompson 1988). In order to produce a cohesive video sequence, where each segment is related to others and part of a whole, new information and previous discourse history are considered in the rhetorical strategy to build the video. Moreover, other methods proposed building a documentary concerning matter-of-opinion issues (Bocconi, Nack, and Hardman 2008). With most content being mainly expressed by people being interviewed, the material is organized around a particular point of view. From an annotated repository, a semantic graph is built to relate concepts, which are

annotations associated to media items, and edges between them, which represent argumentation relations of two possible polarities: positive or negative. Depending on the purpose, the documentary can be composed by choosing statements that support the argument's point of view (positive edges), or claims that negate, or contradict, the point of view (negative edges).

Furthermore, coherence can be established by means of composition techniques for video production, a common concern in consumer-oriented applications. To convey a story, the succession of video segments that make-up the artifact, should be comprised of a series of related events (Branigan 1992). As such, some techniques have been proposed to yield a coherent presentation, by composing a piece of video along a storyline, having as main concerns the flow of information (Ahanger and Little 1998). The video segments are orchestrated according to aspects such as: temporal continuity, which characterizes the sequencing of segments in time, to incorporate cause-effect relations; thematic continuity, which represents the flow of information conveyed between consecutive segments, ensured by a reasonable range of values in a similarity function (e.g., theme progression is considered static if values are above a threshold); and, content progression, defined as the sum of concepts encompassed in the composition and duration at which they are displayed.

## 1.2 Thesis Contributions

The outcome of this work is a contribution to the field of summarization and multimedia generation for human consumption. Two papers were produced and submitted. The first paper entitled "Summarization of Films and Documentaries Based on Subtitles and Scripts" was submitted to the Pattern Recognition Letters journal. The second paper entitled "Generation of Multimedia Artifacts: An Extractive Summarization-based Approach" was submitted to the European Workshop on Natural Language Generation (ENLG).

In the first paper (Aparício, Figueiredo, Raposo, de Matos, and Ribeiro 2015), we assess the performance of generic text summarization algorithms applied to films and documentaries, using the well–known behavior of summarization of news articles as reference. We use three different datasets, the first composed of news, the second of fictional films subtitles and scripts, and the last one of non-fictional documentaries' subtitles. The evaluation is carried out using the standard ROUGE metrics comparing system-generated summaries against news abstracts,

plot summaries, and synopses.

In the second paper (Figueiredo, Aparício, de Matos, and Ribeiro 2015), we explore methods for content selection and address the issue of coherence in the context of the generation of multimedia artifacts. We use audio and video to present two case studies: film tributes, and lecture-driven science talks. We perform content selection on the basis of centrality- and diversity-based summarization. Coherence is addressed in film tributes by the emotional content of a song, and for science talks by topical similarity between the lecture and documentaries. In addition, composition techniques for the production of multimedia artifacts are addressed as a means of organizing content.

## 1.3 Thesis Structure

This document is organized in the following manner:

- Chapter 2 reviews related work in automatic summarization as part of the content selection process. Section 2.1 introduces common terminology in the field, and categorizes summarization approaches. A special attention is devoted to generic summarization algorithms (Section 2.2), and methods for evaluation in this field (Section 2.3).
- Chapter 3 details methods for coherence in the context of multimedia artifacts.
- Chapter 4 details our procedure for generating lecture-driven science talks. Section 4.1 presents the two stages of our model for dealing with thematic diversity. Section 4.2 introduces video composition by means of two different techniques: a ranking mechanism; and simplicial curves.
- Chapter 5 presents a preliminary study intended to assess the performance of generic summarization algorithms when applied to films and documentaries, based on subtitles and scripts. Section 5.1 presents the datasets used, composed of news, films, and documentaries. Section 5.2 details our experimental setup, followed by the results and discussion (Section 5.3).
- Chapter 6 details the evaluation performed for our experiments. Section 6.1 presents the datasets used, composed of lectures, and documentaries. Section 6.2 details the experi-

mental setup. Section 6.3 presents the results obtained for science talks videos. Section 6.4 presents a general discussion of our procedure and results.

- Chapter 7 ends the document with the conclusion, and directions for future work.

# 2 Content Selection

In this section we review content selection methods. In such settings, summarization algorithms represent the best approaches for dealing with the task of identifying important content. Whether it be for selecting similar content from an information source, dealing with the issue of redundancy, or incorporating diversity.

Automatic summarization started motivated by the problem of condensing large amounts of scientific documents (Luhn 1958). Even though it began more than 50 years ago, several efforts have been made to solve the summarization problem (Paice 1990). Early systems were characterized by surface-level approaches, focusing primarily on statistical features in text data such as term frequency and distribution (Rath, Resnick, and Savage 1961); position of sentences in text (Baxendale 1958); and, cue-phrases (Edmundson 1969). Today, many of these techniques are still in use. Vast amounts of online data are encouraging the application of summarization to different domains, as well as new exploratory research.

News has been the focus of summarization for a long time. Numerous systems have emerged and still pervade, such as NewsInEssence (Radev, Blair-goldensohn, Zhang, and Raghavan 2001) and Newsblaster (McKeown, Barzilay, Evans, Hatzivassiloglou, Klavans, Nenkova, Sable, Schiffman, and Sigelman 2002). In particular for speech summarization, most work has concentrated on broadcast news. Text-based summarization techniques are often applied to spoken documents in order to identify important segments (McKeown, Hirschberg, Galley, and Maskey 2005). However, speech has particular properties that set this field apart (Furui 2007). Fundamental challenges associated to speech summarization include coping with speech recognition errors; disfluencies; and difficulties in identifying sentence boundaries. To deal with these problems, it is common to use speech-specific information, for example acoustic/prosodic features (Maskey and Hirschberg 2005), or speech confidence scores to minimize word error rates (Zechner and Waibel 2000). However, for planned speech, generic approaches to summarization seem to achieve performance results comparable to methods us-

ing speech-related features (Penn and Zhu 2008).

From the audio signal perspective, summarization has proven its extensive applicability. Specifically for music, most summarization methods are based on techniques for segmentation (e.g. tonality, and recurrent structure) and thumbnailing (Chai 2006; Jun and Hwang 2013). Identifying the representative parts of a musical piece is a key issue for browsing, searching, and recommending digital media content. Discovering the music's structure from the analysis of the signal is based on work developed for automatic speech recognition (Hunt, Lennig, and Mermelstein 1980; Peeters 2003). The techniques proposed typically search for repetitive parts in the melody, which is only appropriate for certain musical genres. In general, pop/rock songs are good representatives, and for that reason, are frequently the subject of this process (Logan and Chu 2000). Recurrent parts of a song can be discovered using a similarity matrix (Foote 1999). Since its introduction, it has been widely used to visually represent the structural information of a piece of music, due to the visual resemblance of repeating elements. Several features can be derived from the signal to compute similarity, for instance: Mel-frequency cepstral coefficients (MFCCs); chromagram; or scalar features (Rabiner and Juang 1993; Bartsch and Wakefield 2005; Glaczynski and Lukasik 2011). To compute the similarity between signal observations, distance metrics such as the Euclidean, or cosine distance, can be employed. By grouping similar segments into clusters, the song's structure is uncovered. A summary can be composed by selecting some of the groups identified, the choice relies on user preferences or time constraints (Peeters, La Burthe, and Rodet 2002). Whenever the goal of the summary is not oriented for human consumption, generic algorithms have shown to be useful in improving the performance of tasks such as music genre classification (Raposo, Ribeiro, and de Matos 2015).

Digital video is a rich multimodal information source which encompasses: speech; audio; text; and images. The tremendous rise of visual content has motivated a multitude of applications for image and video indexing and retrieval (Hu, Xie, Li, Zeng, and Maybank 2011), as well as summarization (Ajmal, Ashraf, Shakir, Abbas, and Shah 2012b). Several methods have been proposed to produce concise representations of this data. The multimodal nature of video makes summarization-related tasks more complex than summarizing text documents (Money and Agius 2008). Still, the developed techniques target video data from a variety of domains: news, sports, documentaries, movies, lectures, and home videos (Truong and Venkatesh 2007).

Most approaches condense the content of a video in a domain specific manner and for that reason, do not generalize adequately over different genres (Priyamvada R Sachan 2014). Existing methods are usually limited to exploring low-level features for video, for instance: color, shape, and motion; or acoustic characteristics for audio, such as loudness, and pitch. However, the content and structure of the video, do not correspond, directly, to strategies based solely on low-level features. This restricts their application to specific genres. Strategies proposed to work for general videos are often complex, and typically solve this problem by constructing hierarchical structures that describe the video content at various levels (Zhu, Fan, Elmagarmid, and Wu 2003). Nonetheless, promising results have been reported, especially for sport highlights (Anjum, Ali, Hassan, and Adnan 2013), movies (Tsai, Kang, Lin, and Lin 2013), audio-video presentations (He, Sanocki, Gupta, and Grudin 1999), and home videos (Lienhart 1999b). As it happens with music, the major task in video summarization is segmentation. A video can be segmented into different units (Koprinska and Carrato 2001), such as: (i) frames, which are the still images of the video stream; (ii) shots, defined as a series of consecutive frames recorded by a single camera, and (iii) scenes, characterized by sets of consecutive shots that, typically, describe a particular event. One type of video segmentation is achieved by key-frame extraction, which is the process of selecting characteristic frames from the video that contain meaningful information concerning its content. This approach is used to speed up the process of finding and retrieving relevant data, for example, searching for the right footage in the BBC's Natural History library (Gibson, Campbell, and Thomas 2002). In general, shot segmentation, usually denoted by shot-boundary detection, is the first step in key-frame extraction methods. Possible approaches consist of selecting the first frame of each shot (Tonomura, Akutsu, Otsuji, and Sadakata 1993), or selecting the first and last frames (Ueda, Miyatake, and Yoshizawa 1991). In order to incorporate the dynamics in the visual content of the shot, other approaches group key-frames into visually similar clusters (Arman, Hsu, and Chiu 1993). Most early shot boundary techniques focused primarily on cut detection, where abrupt changes occur in a single frame (Boreczky and Rowe 1996; Lienhart 1999a). More recent work, has focused on detecting gradual transitions, typical of cinematic movies, such as fades, where slow changes in brightness occur, or dissolves, where the final images of a shot are dimmed and overlapped with the next shot (Baber, Afzulpurkar, Dailey, and Bakhtyar 2011). Apart from the commonly used visual features, audio information can be embedded in this process, improving the performance of summarization systems. In sports, this has attained some degree of success, improving rel-

evance measures by means of whistle detection, goals, and faults (Raventós, Quijada, Torres, and Tarrés 2015). Also, text data, which is considered a high-level feature, can be used, if available, as a complement to low-level features, due to the rich content information it contains (Luo, Tang, Liu, and Zhang 2003). Furthermore, in the context of consumer-oriented video summarization, it is common to preserve the temporal nature of the summary. In the same manner as it happens for key-extraction summarization methods, sets of representative segments are chosen, usually, from the output of shot-boundary detection algorithms (Lu and Grauman 2013). Important segments that will compose the summary, can be selected on account of the motion activity present (Nam and Tewfik 2002), spatio-temporal features (Laganière, Bacco, Hocevar, Lambert, Païs, and Ionescu 2008), or even the attention of the user, which is modeled by incorporating what attracts viewer's attention in the images of the video program (Ma, Lu, Zhang, and Li 2002).

## 2.1 Taxonomy

There are several aspects that can be used to categorize automatic summarization approaches. Summaries can be discriminated based on factors such as: length; their focus user group; the number of information sources; or the relation between the summary's content and its information source. In this section, we present a brief overview concerning the nature of summaries. We follow terminology firmly established in the field (Nenkova, Maskey, and Liu 2011; Truong and Venkatesh 2007), especially focusing on text and video. The aforementioned categorization, results in several dualities which are presented, along with research applications, as well as current difficulties associated to them.

Foremost, the necessity of any kind of summarization arises with overabundance of content along with insufficient time to consume it. Although several definitions of what characterizes a summary exist, elaborate characterizations still lack consensus (Das and Martins 2007). Nonetheless, a summary can be broadly defined as: "a text that conveys important information and that is at least half of the original text(s), and usually, significantly less than that" (Radev, Hovy, and McKeown 2002), where text can represent any media content form. Moreover, the goal of automatic summarization systems is to explore an information source(s), then retrieve, and present the most important content from it, in a condensed form, that suits the user's needs (Mani and Maybury 2001). Next, we present the major dualities in this field:

- *Extractive* versus *abstractive*: Extractive summarization methods are characterized by re-using information as-is from the source. In text-based approaches it involves concatenating extracts taken from the information source to compose a summary, such as words, sentences, or paragraphs, as they appear. Most automatic systems use linguistic or statistical methods to identify important segments (Vanderwende, Suzuki, Brockett, and Nenkova 2007; Murray, Renals, and Carletta 2005; Nenkova, Vanderwende, and McKeown 2006), while others rely on machine learning techniques (Svore, Vanderwende, and Burges 2007). A common drawback associated with computer-produced extracts is the lack of coherence, which represents how the text is connected together, logically structured, and how it is semantically meaningful. Although coherence is not easily objectified, it can be understood as "the feeling that a text hangs together, that it makes sense, and is not just a jumble of sentences" (McCarthy 1991). Dangling anaphoric and cataphoric references, which refer back, or ahead, respectively, to some entity, are common issues of this practice. Due to the segments being extracted out of context, joining them together may result in a misleading interpretation of these dangling references. A possible solution that addresses these issues involves the inclusion of adjacent segments until all references are resolved, however including non-relevant material results in less focused and larger summaries (Paice 1990). Opposite to producing extracts, abstractive summarization methods create summaries by re-generating, or paraphrasing, important content from the information source. Typically, abstractive-based systems require advanced language generation and compression, or fusion, techniques in order to generate new sentences, and as a result improve the focus of the summary (Knight and Marcu 2000; Barzilay and McKeown 2005). The major challenge associated to abstractive systems is the representation problem. These systems' capabilities are constrained by the richness of their representations, as well as their ability to generate a fluent output from an abstract representation (Genest and Lapalme 2011).

- *Informative* versus *indicative*: In what concerns the style of the summary, indicative summaries provide just enough information to help the reader determine what the text is about, without conveying specific content. These summaries are commonly used as the end part of an information retrieval system to improve performance in browsing and search-related tasks (Kan and Klavans 2002). Informative summaries, on the other hand, are intended to replace the original text by retaining only essential content. Due to their

wide applicability, most work in summarization deals with producing these type of summaries (Kan, McKeown, and Klavans 2001; McKeown, Passonneau, Elson, Nenkova, and Hirschberg 2005; Murray, Renals, and Carletta 2005).

- *Generic* versus *query-oriented*: Generic summarization does not make assumptions about the audience for which it is intended. These summaries place an equal emphasis on distinctive information contained in the source, with the purpose of providing a balanced coverage (Gong and Liu 2001). Alternatively, in query-oriented summarization the goal is to summarize only the information that is relevant for a specific query. These approaches, which often are the extension of information retrieval techniques, are particularly useful for search engines (Turpin, Tsegay, Hawking, and Williams 2007; Varadarajan and Hristidis 2006), or question-answering tasks (Shi, Melli, Wang, Liu, Gu, Kashani, Sarkar, and Popowich 2007; Berger and Mittal 2000). Often, methods devised for query-oriented summarization can be adapted for generic summarization, usually, by taking the whole input as query (Carbonell and Goldstein 1998).

- *Single-* versus *multi-document*: Single-document summarization is the process of summarizing only one information source. Most early work dealt with producing summaries of a single document, in the context of news (Hovy and Lin 1998; McKeown, Hirschberg, Galley, and Maskey 2005; Svore, Vanderwende, and Burges 2007), meeting recordings (Murray, Renals, and Carletta 2005), or lectures (Fujii, Kitaoka, and Nakagawa 2007). As research progressed, more work started focusing on exploring the application of summarization techniques to sets of related documents. The first systems applied multi-document summarization to clusters of news articles covering the same event (McKeown, Barzilay, Evans, Hatzivassiloglou, Klavans, Nenkova, Sable, Schiffman, and Sigelman 2002; Radev, Otterbacher, Winkel, and Blair-Goldensohn 2005), or web pages (Radev, Fan, and Zhang 2001). In such applications, the existence of several documents with similar information, introduces redundant content, which must be handled. A common approach in text documents involves measuring similarity between all pairs of passages, and then cluster related passages to identify common themes of information (Radev, Allison, Blair-Goldensohn, Blitzer, Çelebi, Dimitrov, Drabek, Hakim, Lam, Liu, Otterbacher, Qi, Saggion, Teufel, Topper, Winkel, and Zhang 2004; Radev, Jing, and Budzikowska 2000; Ribeiro and de Matos 2011).

- *Static* versus *dynamic*: In what concerns video, the major duality lies on the presentation of stationary images (static summarization), or moving images (dynamic summarization). As it happens with indicative summaries, static summarization serves the purpose of facilitating search-related tasks, for example, browsing and navigating in image retrieval applications (Bailer, Mayer, Neuschmied, Haas, Lux, and Klieber 2003; Rautiainen, Ojala, and Seppanen 2004). Such approaches consist of extracting a collection of frames (story board) that represent the video source (Borth, Ulges, Schulze, and Breuel 2008). Apart from information retrieval tasks, other methods aim at creating story boards for entertainment-driven purposes, such as producing comic book style summaries (Uchihashi, Foote, Girgensohn, and Boreczky 1999). These static-type summaries are limited in terms of their descriptive ability, since they do not incorporate the dynamic audio-visual content of the original video. On the other hand, dynamic summarization approaches intend to provide a condensed version of the original video. Typically, by extracting segments that preserve the important content of the video. Popular practices include movie trailers (Sundaram and Chang 2001), sports (Coldefy, Bouthemy, Betser, and Gravier 2004), and news (Smith and Kanade 1997).

## 2.2 Generic Summarization Algorithms

In this section, we present unsupervised approaches to summarization. Several generic algorithms have been developed and successfully applied in different areas, such as text, music, video, and speech. We review generic, language-, and domain-independent methods based on: uncovering latent structure (LSA); centrality (LexRank and Support Sets); and diversity (Maximal Marginal Relevance (MMR) and Graph Random-walk with Absorbing StateS that HOPs among PEaks for Ranking (GRASSHOPPER)).

### 2.2.1 Latent Semantic Analysis (LSA)

LSA is a mathematical technique based on the Singular Value Decomposition (SVD). This approach aims at inferring contextual usage of text based on word co-occurrence. In generic summarization settings (Gong and Liu 2001), LSA is often used to uncover important topics in collections of documents without requiring external lexical resources, i.e., online thesauri

or dictionaries. This technique follows the notion that the occurrence context of a particular word provides information that can be used to assess meaning. LSA produces relations between words and sentences that correlate with the way humans make associations or discover semantic similarity (Landauer and Dutnais 1997; Landauer, Foltz, and Laham 1998).

In order to use LSA for text summarization, the input document is represented through a $t \times n$ term-by-sentences matrix A, where rows represent unique words ($t$) of the input document, and columns represent sentences ($n$). Then, the SVD is applied to A, resulting in its decomposition (Equation 2.1): $U$, a $t \times n$ matrix of left singular vectors (its columns); $\Sigma$, an $n \times n$ diagonal matrix of singular values; $V^T$, an $n \times n$ matrix of right singular vectors (its rows).

$$A = U\Sigma V^T \tag{2.1}$$

Some of the singular values in the diagonal matrix $\Sigma$ are too small and can be discarded, by setting them to zero. Therefore, by keeping the first, non-zero entries, of singular values of $\Sigma$, we reduce it to $\Sigma_k$, which represents the $k \times k$ submatrix of $\Sigma$. $U$ and $V^T$ are also reduced to $U_k$ and $V_k^T$ in order to have $k$ columns and $k$ rows, respectively. Therefore, $A$ is approximated to: $A_k = U_k \Sigma_k V_k^T$, where $A$ is the same as before, an $m \times n$ matrix, $U$ is $m \times k$, $\Sigma$ is $k \times k$ and $V$ is $k \times n$.

When the SVD is applied to matrix $A$, the transformation can be seen from two perspectives: as a mapping from one dimensional space to another, resulting in dimensionality reduction; and, as a semantic structure derived by aggregating words and sentences in similar contexts. If a pattern of words occurs very often in a document, this pattern will be captured and represented by one of the singular vectors, which represent topics or concepts, following the notion that similar words appear in similar contexts. From the summarization perspective, however, there are some limitations (Steinberger and Jezek 2004). The number of sentences to be included in the summary is equal to the number of dimensions used. Thus, as the number of dimensions increase, the higher the chances of including less significant sentences. Second, a sentence with high values, but not the highest, will not win in any dimension, and therefore, will not be chosen for the summary. As a result, an enhanced version of LSA was proposed (Ježek and Steinberger 2008), in which, the score of each sentence $s_i$ is computed, as

given by Equation 2.2, where $v(i,k)$ corresponds to the $k$-$th$ element of the $i$-$th$ sentence vector, and $\sigma(k)$ is the corresponding singular value of the matrix. For each sentence in the $V$ matrix, its components are multiplied by the corresponding singular values. This favors index values in the matrix $V$ that correspond to the highest singular values (the most important topics).

$$\text{Score}(s_i) = \sqrt{\sum_{k=1}^{n} v(i,k)^2 \times \sigma(k)^2} \tag{2.2}$$

### 2.2.2 LexRank

LexRank (Erkan and Radev 2004) is a centrality-based method based on Google's PageRank for ranking web pages (Brin and Page 1998). A graph is built using sentences, represented by TF-IDF vectors, as vertexes, and the cosine similarity determines how they connect. An edge is created if the similarity score exceeds some threshold. Then, the calculation described by Equation 2.3 ($d$ is a damping factor which ensures convergence, $N$ is the total number of vertexes, and $S(V_i)$ is the score of the $i$-$th$ vertex) is computed for each vertex until convergence, which happens when the error rate between two successive iterations is lower than a certain value for each vertex.

$$S(V_i) = \frac{(1-d)}{N} + d \times \sum_{V_j \in \text{adj}[V_i]} \frac{\text{Sim}(V_i, V_j)}{\sum_{V_k \in \text{adj}[V_j]} \text{Sim}(V_j, V_k)} S(V_j) \tag{2.3}$$

### 2.2.3 Support Sets

Documents are typically composed by a mixture of subjects, normally involving a main and various minor issues. Support sets are defined based on this observation (Ribeiro and de Matos 2011). Important content is determined by creating a support set for each passage, by comparing it with all others from the source. The most semantically-related passages, determined via geometric proximity, are included in the support set. In this manner, groups of related passages are uncovered, each one representing a subject. Summaries are composed by selecting the most relevant passages, i.e., the ones present in the largest number of support sets.

Given a segmented information source $I \triangleq p_1, p_2, \ldots, p_N$, support sets $S_i$ associated with each passage $p_i$ are defined as indicated in Equation 2.4, where $\text{Sim}$ represents a similarity

function, and $\epsilon_i$ is a threshold.

$$S_i \triangleq \{s \in I : \mathrm{Sim}(s, p_i) > \epsilon_i \wedge s \neq p_i\} \tag{2.4}$$

The most important passages are selected based on Equation 2.5.

$$\arg \max_{s \in U_{i=1}^{n} S_i} |\{S_i : s \in S_i\}| \tag{2.5}$$

### 2.2.4 Maximal Marginal Relevance (MMR)

MMR is a commonly adopted method for query-focused summarization (Carbonell and Goldstein 1998). A linear combination of relevance and novelty is established by configuring the model, which iteratively selects the documents that result from applying Equation 2.6.

$$\arg \max_{S_i} \left[ \lambda \left( \mathrm{Sim}_1 \left( S_i, Q \right) \right) - (1 - \lambda) \max_{S_j} \mathrm{Sim}_2 \left( S_i, S_j \right) \right] \tag{2.6}$$

Where $\mathrm{Sim}_1$ and $\mathrm{Sim}_2$ represent similarity metrics. $S_i$ represents the non-selected documents and $S_j$ the previously selected documents, $Q$ is the query and $\lambda$ is the parameter that configures between relevance, $\lambda(\mathrm{Sim}_1(S_i, Q))$, and novelty, $(1 - \lambda)(\max_{S_j} \mathrm{Sim}_2(S_i, S_j))$. Therefore, when $\lambda = 1$, the summary will be composed by the standard relevance list. On the other hand, for $\lambda = 0$, maximal diversity ranking is obtained. A good practice is to first observe the information space surrounding the query with $\lambda \simeq 0.3$, and then focus on the important parts applying MMR with $\lambda \simeq 0.7$ (Carbonell and Goldstein 1998). Furthermore, the MMR approach can generate generic summaries by considering the input sentences centroid as a query (Murray, Renals, and Carletta 2005; Xie and Liu 2008).

### 2.2.5 Graph Random-walk with Absorbing StateS that HOPs among PEaks for Ranking (GRASSHOPPER)

GRASSHOPPER (Zhu, Goldberg, Gael, and Andrzejewski ), is a re-ranking algorithm that maximizes diversity and minimizes redundancy. It is based on random walks in an absorbing Markov chain, and can be seen as an alternative to MMR because it addresses diversity as well.

The algorithm takes three parameters as input: a weighted graph $W$, a probability distribution $r$ (user-defined prior ranking) and $\lambda \in [0, 1]$ that balances the weighted graph and the prior ranking. $W$ is a $n \times n$ matrix, where $n$ is the number of sentences, and the weights are defined using a similarity measure (e.g., cosine distance). $r$ is a user-defined ranking (e.g., sentence position), defined by $r = (r_1, \ldots, r_n)$, where $\sum_{i=1}^{n} r_i = 1$, and $r_i$ is the probability of sentence $i$. When there is no prior ranking, a uniform distribution can be used, where each sentence has the same probability.

Sentences are ranked by applying the teleporting random walks method, which is based on the $n \times n$ transition matrix $\tilde{P}$ (calculated by normalizing the rows of $W$):

$$P = \lambda \tilde{P} + (1 - \lambda)\, \mathbf{1}\mathbf{r}^{\top} \tag{2.7}$$

The first sentence to be scored is the one with the highest stationary probability $\arg\max_{i=1}^{n} \pi_i$ according to the stationary distribution of $P$: $\pi = P^{\top}\pi$. Sentences already selected are not visited again, by defining $P_{gg} = 1$ and $P_{gi} = 0, \forall i \neq g$. The expected number of visits is given by an $N$ matrix, as defined by Equation 2.8, where $N_{ij}$ is the expected number of visits to the sentence $j$, if the random walker began at sentence $i$.

$$N = (I - Q)^{-1} \tag{2.8}$$

The average of all possible starting sentences is obtained to get the expected number of visits to the $j$-$th$ sentence, $v_j$. The sentence to be selected is the one given by the Equation 2.9.

$$\arg \max_{i=|G|+1}^{n} v_i \tag{2.9}$$

## 2.3 Evaluation Methods

Evaluation methods started to gain interest soon after the first automatic summarization systems were proposed (Edmundson 1969). Assessing the quality of a summary is an ambiguous, as well as complex, task, and for that reason, consensus has yet to be established for any particular method. There are several challenges associated to summary evaluation, which en-

compass: aspects of readability; informativeness; and coherence (Nenkova 2006; Spärck Jones 2007). Typically, for text-based approaches, summarization involves a machine producing an output for generic, or user-oriented, purposes. In most cases, even if the output is meant to answer a question, the system may generate a good summary that is different from what other human summarizers might produce, as a correct output. In addition, given that the summary is a condensed representation of the original source, the compression rate should be taken into account. Since human evaluation is expensive, most of the work has been focused on automatic procedures to determine summary quality (Mani 2001).

Methods for summarization can be broadly classified into two categories: in intrinsic evaluation procedures, the output of the system is assessed by itself, usually by comparison with gold standards, while in extrinsic evaluation, the produced summary is judged on the basis of its ability to help complete specific tasks. Several intrinsic models have been introduced, for instance, precision and recall (Jing, Barzilay, McKeown, and Elhadad 1998), relative utility (Radev and Tam 2003), the pyramid method (Nenkova, Passonneau, and McKeown 2007), and the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metrics (Lin 2004). Although widely employed, particularly, in the context of extractive summarization, they still require human summaries, or annotation. Nonetheless, these methods are, in general feasible in contrast with human evaluation. It was estimated that evaluating the Document Understanding Conference would require 3000 hours of human labor (Lin 2004). Even though the majority of approaches fall under the intrinsic category, ideally, summarization results should be assessed in a task-based setting (Nenkova 2006). However, their expensive, as well as time-consuming, setup, makes them not suitable to most cases.

While text-based summarization evaluation has attained some degree of success, there is a great deal of work in video summarization that does not include any evaluation procedures. And, most of the work that does, has been, mostly, of the intrinsic type (Taskiran and Delp 2004). For event-based content, such as sports programs, the summaries can be judged on the basis of their coverage, in juxtaposition to the source video. For instance, using precision and recall for goal, referee, and penalty box detection (Ekin, Tekalp, and Mehrotra 2003; Ferman and Tekalp 2003). Other approaches involve user studies to judge the quality of the generated summaries. For example, developing surveys based on questions concerning the original content of documentaries, educational videos, and presentations (Taskiran, Pizlo, Amir, Pon-

celeon, and Delp 2006). Furthermore, due to the difficulty in assessing the quality of moving-images, efforts have been made to transfer the problem of video summarization evaluation to the text domain. Since text is a natural domain for expressing semantic information, some approaches include annotating subshots, after video segmentation, with text descriptions. Given a video summary, a text representation of it is generated, and afterwards, text-based evaluation metrics can be used to compare this representation, against human-written ground truth summaries (Yeung, Fathi, and Fei-Fei 2014).

### 2.3.1 Intrinsic Methods

In this section, we present automatic methods to determine the quality of a summary. Precision and recall are standard measures, specially, for information retrieval, and are often combined with the F-Score. ROUGE is a recall-oriented adaptation of the BLEU approach (Papineni, Roukos, Ward, and Zhu 2002; Lin and Hovy 2003), for machine translation, and has been shown to correlate with human evaluation (Lin 2004).

#### 2.3.1.1 Precision and Recall

Most summarization systems extract the most representative sentences in the input and join them together to form a summary. In such settings, if a person is asked to choose the sentences that best convey the meaning of the text, the precision metric (Equation 2.10) can be used to determine the fraction of system generated sentences which are correct, while recall (Equation 2.11), would represent the fraction of sentences chosen by the person that were correctly identified by the system.

$$\text{Precision} = \frac{|S_A \cap S_H|}{|S_A|} \tag{2.10}$$

$$\text{Recall} = \frac{|S_A \cap S_H|}{|S_H|} \tag{2.11}$$

$S_A$ represents the sentences selected by the automatic method, and $S_H$ represents the sentences chosen by the human summarizer. These metrics are often complemented by the F-Score (Equation 2.12), which is a composite measure that combines both precision and recall:

$$\text{F-Score} = \frac{(\beta^2 + 1) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \tag{2.12}$$

where the weight $\beta$, allows configuring the F-Score to favor recall, $\beta > 1$, or precision, $\beta < 1$.

#### 2.3.1.2 ROUGE

ROUGE (Lin 2004) is a collection of recall-oriented metrics, in contrast to the precision-biased BLEU method. In ROUGE, the content quality of the automatic summaries is determined on the basis of word-overlap, between system generated and human-made, reference summaries. The most commonly used method inside this collection is the ROUGE-N. Apart from favoring recall, it is intended to have a good coverage of the reference summaries by counting the number of overlapping $n$-grams between the computer generated and the human-made summary. In addition, the ROUGE-SU metric measures the overlap of skip-bigrams, which is any pair of words in their sentence order, with the addition of unigrams as counting unit.

Formally, ROUGE-N is given by Equation 2.13, where $\text{gram}_n$ represents the length of the $n$-gram, and $\text{count}_{\text{match}}(\text{gram}_n)$ represents the maximum number of $n$-grams of a candidate summary that co-occur with a set of reference summaries. ROUGE-SU, instead of comparing sequences of $n$ consecutive words, takes into account bigrams and unigrams. It allows the insertion of words between the first and last elements of the bigram according to a pre-defined distance.

$$\text{ROUGE-N} = \frac{\sum_{S \in \text{RS}} \sum_{\text{gram}_n \in S} \text{count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \text{RS}} \sum_{\text{gram}_n \in S} \text{count}(\text{gram}_n)} \tag{2.13}$$

### 2.3.2 Extrinsic Methods

The purpose of extrinsic evaluation procedures is to measure the impact of summarization on some well-defined job. A variety of different tasks can be considered (Mani 2001), for instance: measuring the summary's efficiency on completing certain instructions; determining the suitability of the summary for finding information in large collections of data; or, measuring

the impact of the summary on another system. However, in addition to the mere difficulty of selecting the appropriate task, ensuring its unambiguity, so as to have some agreement among users performance, can be challenging (Nenkova 2006; Spärck Jones 2007).

In the paradigm of relevance assessment, several work has been done to determine the influence of a summary, tailored according to some query, to a given task. Usually, in the context of information retrieval, the subject is presented with a description and a document, and then, judges the relevance of the document to that description, based solely on its summary (Tombros and Sanderson 1998). The system's performance is usually measured on the basis of accuracy, as well as time (Jing, Barzilay, McKeown, and Elhadad 1998; Mani and Bloedorn 1999). Specifically for video summarization, other evaluation methods involve users matching video summaries to the key-frames of the source video (Christel, Smith, Taylor, and Winkler 1998).

In cases where input documents of summarization systems contain unstructured and disfluent text, the resulting summary is expected to contain the same noisy data. For spontaneous meeting speech (Murray, Kleinbauer, Poller, Renals, Kilgour, and Becker 2008), an evaluation was carried out where the task consisted on satisfying a complex information need: understanding why a given decision was made and assessing the impact of automatic speech recognition errors on users' performance. By means of questionnaires and human judgments, the findings concluded that although speech-related errors affected users' satisfaction, the summaries were regarded as useful tools for browsing meeting data.

Another approach to extrinsic evaluation is assessing the impact of automated summaries on other systems. For instance, improving accuracy scores in classification procedures. In the context of music, generic summarization algorithms were applied to obtain small important segments from different parts of a song (Raposo, Ribeiro, and de Matos 2015). Due to summarization being done without requirements of coherence, the music summaries produced were able to retain more relevant and less redundant information, making them more suitable for machine-oriented tasks. Thus, building on this idea, the automated summaries were able to improve the performance of a Portuguese music style Fado Genre classifier, on tasks that rely on processing just a portion of the whole audio signal.

## 2.4 Summary

In this chapter, we surveyed methods for content selection in the context of automatic summarization. We presented a brief overview concerning the nature of summaries, categorization approaches, and common terminology. Next, we reviewed generic, language, and domain independent summarization algorithms. Lastly, we presented evaluation methods broadly classified into intrinsic and extrinsic approaches.

# 3 Topical Coherence

In this section, we explore methods for coherence. We revisit LSA (introduced in Section 2.2.1) as a technique for characterizing coherence in the derived high-dimensional semantic space. We present LDA (Blei, Ng, and Jordan 2003), a generative probabilistic topic model, as a means of obtaining thematically coherent content. And, present the Locally Weighted Bag of Words (Lowbow) framework (Lebanon, Mao, and Dillon 2007) to preserve sequential information by modeling text as simplicial curves.

## 3.1 Using LSA for Local Measures of Coherence

As previously described, LSA is a statistical model originally developed for inferring contextual usage of text. Even though its popular use lies with text-based, generic summarization, approaches (Steinberger and Jezek 2004), it is often applied to different domains. For example, certain video applications employ LSA to reduce the semantic gap between low-level video features and semantically meaningful spatio-temporal entities (Hohl, Souvannavong, Merialdo, and Huet 2004), while other image retrieval applications focus on keyword extraction for personalized sports video (Zhang, Zhang, Xu, and Lu 2007). Furthermore, representing entities in a high-dimensional semantic space has proven useful in the context of different problems. As commonly associated to automatically-produced summaries, in particular with extracts, there is a lack of coherence-related characteristics in the output text, such as fluency and organization. To this end, LSA has been proposed for predicting local measures of coherence, through the relatedness of text units (e.g., sentences), in the derived semantic space (Foltz, Kintsch, and Landauer 1998). Although several aspects contribute to coherence, the underlying intuition of the model, is the observation that related sentences in a text, contain a high number of semantically related words. As such, semantic relatedness corresponds to the extent that adjacent text units are discussing semantically similar information.

In this framework, the input data is collected in a frequency matrix, where rows corre-

spond to text units (e.g., words) and columns represent the linguistic context (e.g., documents). After the application of SVD to the sparse frequency matrix (equation 2.1), the latent structure is uncovered. From this procedure, a sentence can be represented by a vector in the high-dimensional semantic space, composed by the mean of the vectors of its words (represented in the matrix of left singular vectors $U_k$). As a result, local measures of coherence can be determined, by assessing the similarity of adjacent sentences, through the cosine of their means as defined by Equation 3.1, where $\mu(\vec{S_i}) = \frac{1}{|S_i|}\sum_{\vec{u}\in S_i}\vec{u}$, and $\vec{u}$ is the row vector in matrix $U_k$ which corresponds to the word $u$. In this way, local coherence amounts to the degree of semantic similarity between sentences $S_1$ and $S_2$. Furthermore, an overall measure of text coherence can be determined by averaging the cosines of all pairs of adjacent sentences in the text.

$$\mathrm{Sim}(S_1, S_2) = \mathrm{Cos}(\mu(\vec{S_1}), \mu(\vec{S_2})) = \frac{\sum_{j=1}^{n}\mu_j(\vec{S_1})\mu_j(\vec{S_2})}{\sqrt{\sum_{j=1}^{n}(\mu_j(\vec{S_1}))^2}\sqrt{\sum_{j=1}^{n}(\mu_j(\vec{S_2}))^2}} \tag{3.1}$$

## 3.2 Latent Dirichlet Allocation (LDA)

As more multimedia content is available, it becomes more difficult to organize, search, and understand these vasts amounts of data. Topic models provide methods to automatically explore large electronic archives. The basic steps underlying these models encompass: (i) uncovering the hidden topical patterns that pervade a collection; (ii) annotating, with thematic information, the documents according to those topics; and, (iii) using the annotations to manipulate (e.g., organize, search) the collection. LDA is an unsupervised, statistical topic model, for collections of discrete data (Blei, Ng, and Jordan 2003). The generic applicability of the model extends its use for a variety of tools and domains (Hu 2009), for instance document modeling and classification for text-based applications; object categorization and segmentation for image-based tools; or, key-finding and modulation tracking (key shifts) for music.

In probabilistic modeling, the data is treated as observations that arise from a generative probabilistic process. Topics represent the hidden variables, i.e., the structure to be found in the data. For documents, the hidden variables reflect the thematic structure of the collection, where words (observed random variables) are generated by a stochastic procedure. LDA explains document collections based on mixtures of topic distributions. In this sense, LDA follows the intuitive notion that documents exhibit multiple topics. Each topic is characterized

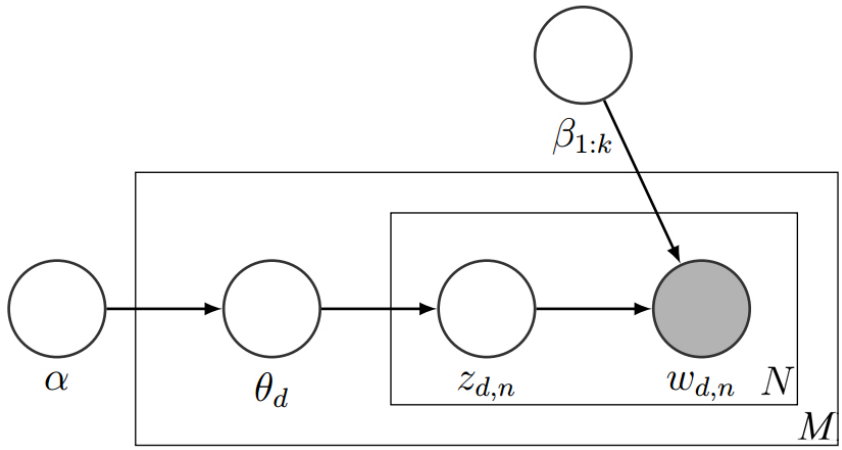



Figure 3.1: LDA represented using plate notation. Shaded nodes, and unshaded nodes, represent observed and unobserved (latent) variables, respectively.

as a distribution over words in a fixed vocabulary and each document is a random mixture of topics. Given the trained model, is it possible to find related documents by assessing the similarity of their topic mixtures.

Figure 3.1 illustrates the model using plate notation, where each plate represents a repetition of sampling steps, with the lower right corner variable referring to the number of samples: the inner $N$ plate denotes the collection of words in a document; while the outer plate $M$ represents the documents inside the collection. Arrows between plates represent conditional dependencies. Topics are represented by $\beta_{1:k}$, where $k$ is the number of topics; $\beta_k$ is a distribution over the vocabulary; $\theta_d$ is a $k$-dimensional variable representing the topic probabilities in document $d$; $z_{d,n}$ is the assigned topic for the word $n$ in a document $d$; lastly, the observed words in a document are given by $w_{d,n}$. The parameters $\alpha$ and $\beta$, are symmetric Dirichlet priors, sampled only once, and therefore constants in the model. The Dirichlet prior on the topic distribution $\theta$ can be interpreted as a smoothing topic distribution, where the amount of smoothing is determined by the parameter $\alpha$. In a probability distribution simplex, lower values (e.g. $\alpha < 1$) forces the topics to the corners of the simplex, resulting in sparsity, therefore it becomes more probable to draw topic distributions favoring few topics. On the other hand, higher $\alpha$ values, leads to more smoothing, moving the topics to the center of the simplex. In a similar manner, $\beta$ provides smoothing for the word distribution in every topic, greater values mean that each topic is more probable to contain a mixture of most words, and fewer words for lower values.

The iterative process for document generation works as follows, for each document $d$ in the collection:

1. A $k$-dimensional topic weight vector $\theta_d$ is drawn, modeled by the Dirichlet random variable $\alpha$, where $k$ corresponds to the number of topics, which is assumed to be known. The vector determines which topics are most likely to appear in the document.
2. For each word $n$ in document $d$:
   (a) A topic $z_n$ is drawn from the multinomial distribution $\theta_d$.
   (b) Given the chosen topic $z_n$, a word is drawn from the corresponding probability distribution over the vocabulary, $\beta$. From this procedure, each word in the document is generated by a different, randomly chosen, topic.

A new document can be situated in the trained model using variational inference. From this procedure, the variational Dirichlet parameters $\gamma$ and $\phi$ are learned. $\gamma$ represents the distribution of topics of the new document given the previous $k$-topic model. $\phi$ represents the distribution of topics for each word of the new document given the previous $k$-topic model. Variational inference can be seen as an optimization problem by simplifying the original model, where some dependencies are removed to decouple $\alpha$ and $\beta$, by losing the edges between $\theta$, $z$, and $w$.

## 3.3 Simplicial Curves

In this approach, we consider a new modeling process for representing documents as smooth curves in a multinomial simplex. The main idea behind the Lowbow framework (Lebanon, Mao, and Dillon 2007) is preserving sequential content information in text, as opposed to the bag-of-words assumption. In Lowbow, collections of word histograms, throughout the document, keep track of occurring changes, in contrast to $n$-gram models, that represent frequency, independently of the patterns position.

Fundamentally, this framework aims at temporally smoothing the original word sequence, using a local smoothing kernel. In geometrical terms, the collection of word histograms is equivalent to parametrized curves in the simplex. Following this model, differential geometry

provides the means to change properties of continuity, as well as smoothness in the generated curves. As such, they depend on the word sequence and a smoothing kernel ($\sigma$), which interpolates between a single-point bow representation ($\sigma \rightarrow \infty$) and the original word sequence representation ($\sigma \rightarrow 0$). By varying $\sigma$ between 0 and $\infty$, the Lowbow representation captures sequential detail at different resolutions.

## 3.4 Summary

In this chapter, we explored methods for coherence. We revisited LSA for predicting local measure of coherence, through the relatedness of text units in the derived semantic space. LDA is introduced for uncovering hidden topical patterns in collections of text data, as a means of obtaining thematically coherent content. Lastly, we overview the Lowbow framework for preserving sequential content information in text.

# 4 Lecture-driven Science Talks

In this chapter, we describe our approach for the automatic generation of lecture-driven science talks. We use scientific documentaries to illustrate the main subjects addressed in a physics lecture. As usual in informational talks, the audio stream carries most of the important information disclosed. For that reason, our creation process is governed by text transcriptions of lectures and subtitles of documentaries. We build on the conjecture that multiple topics are covered in the video lecture. Documentaries provide the means to showcase them. Given that our process is oriented for human consumption, several elements regarding the understandability of the multimedia artifact must be taken into account. In that sense, we structure an entertainment-driven artifact having as main concerns: content selection and aspects of coherence. Furthermore, in conjunction with the proposed architecture, we address the central issues associated to the identification of meaningful and diverse segments, representative of the topic spans within the lecture. Moreover, from the documentary collection perspective, uncovering the hidden topical patterns that enable the selection of topic-related content. We address coherence as a consequence of the content selection process and by means of video composition techniques, using two distinctly-purposed methods: centrality-based summarization as a mechanism for ranking and simplicial curves for incorporating aspects of fluidity.

Figure 4.1 depicts the architecture of the computational method developed for generating lecture-driven science talks. Our approach takes as input the lecture's subtitles, segmented at sentence-level. The primary concern at this phase, is to obtain a concise, yet diverse, representation of the lecture. Following this line of reasoning, we employ GRASSHOPPER (Section 2.2.5), a diversity-based summarization approach. The sentences returned by the algorithm are representative of particular groups (centrality), as well as how different they are from each other (diversity). We focus on maintaining the timeline order of the input lecture, with the intention of preserving the original, cause-effect, theme relations, in the final video.

Hereafter, we take an additional step to aid the subsequent process of extracting topic-

related content from documentaries. We explore the application of LSA (introduced in Section 2.2.1), as a measure of local coherence (see Section 3.1), which enables spanning the sentences returned from GRASSHOPPER, to include other, semantically-related, sentences, concerning the same subject. In this manner, we obtain locally-coherent groups, by comparing adjoining units of text (e.g., sentences) in the derived, high-dimensional, semantic space (see Equation 3.1). The degree of semantic relatedness is characterized by similarity metrics, such as the cosine distance. Highly-coherent discourse is represented by higher values, while lower values represent the lack of contextual and co-occurring information. In the subsequent phase, the identified locally-coherent groups of sentences, are used for querying the collection of documentaries in a 2-step process. First, at document-level, obtaining a subset of topic-related documentaries. Then, using this subset at sentence-level, obtaining topic-related sentences.

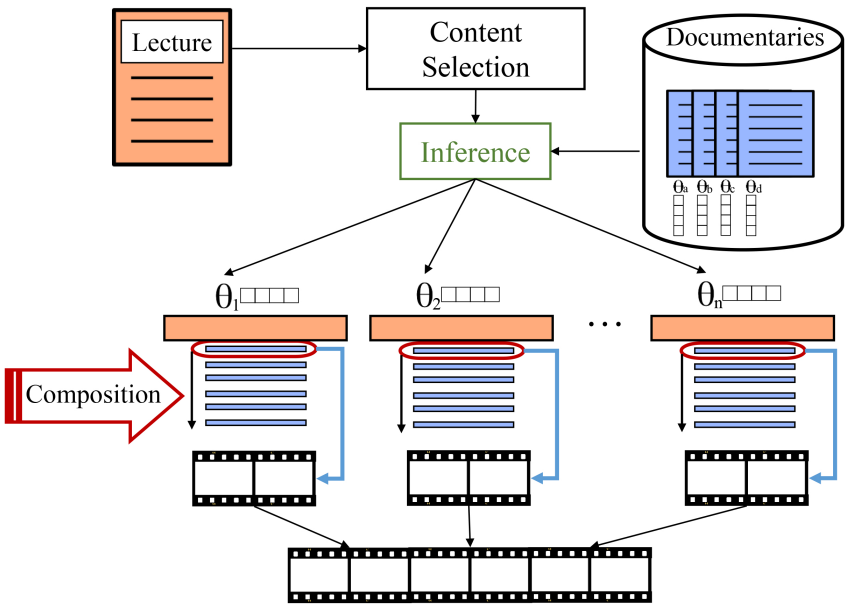



Figure 4.1: Proposed architecture for automatic generation of lecture-driven science talks. $\theta$ represents topic mixtures.

## 4.1 2-Stage Model

Here we introduce our approach for dealing with the thematic diversity within the collection of documentaries. We apply a generative probabilistic topic model, namely LDA (introduced in Section 3.2), to discover latent topics within the collection. In light of this, a given text

segment can be inferred in the estimated model to obtain its topic-segment relations.

We strategize a refinement process for content selection given the summarized lecture, which is twofold: we train a LDA model at document-level, to obtain a subset of topic-related documentaries; then, we train a new model using this subset, segmented at sentence-level, to allow for a more fine-grained selection. In the first stage of our approach, the topic information embedded in the documents can help the context understanding process by guiding the next stage. As such, we select the most topic-related documentaries to the lecture and then, proceed to select the most topic-related sentences.

In that sense, we explore the thematic structure of the collection of documentaries, by the use of the hidden variables uncovered by the model. In this process, the latent topics are discovered by identifying groups of words, in the collection, that are related by frequently occurring within documents. We characterize each documentary document, in the collection, by its topic weight vector $\theta_{d,k}$ (see Figure 3.1), which represents the topic proportions, or the amount of contribution for each of the $k$ topics. Variational inference is performed on the summarized lecture (unseen document), using the previously estimated model. From this procedure, the variational Dirichlet parameter $\gamma$ is learned, which represents the probability of the lecture being generated by the previous $k$-topic model. Therefore, topic relatedness can be estimated, by using similarity metrics (e.g., cosine distance), in the latent topic space. Accordingly, in the first stage of our process, similarity is performed based on topic mixtures at document-level. Thus, obtaining a subset of documentaries that are related, topic-wise, to the locally-coherent summarized lecture. In a similar manner, for the second stage, we compare each group of locally-coherent lecture sentences, with sentences from the documentary subset. As a result, we obtain the top-$n$ candidate sentences from documentaries, for each lecture group segment. This process forges the basis for the subsequent video composition techniques.

## 4.2 Video Composition

Composition is one of the main, and fundamental, aspects of video production. There are several elements involved in producing a fluid and well-composed video, which include: the sequencing of segments in time; the progression of content, or flow of information; as well as many other visual and audible factors, contingent on the overall purpose of the video. Our

dynamic approach for video composition is selectively focused on choosing the candidates that best present the original content. In light of this, given the top-$n$ candidate documentary sentences for each lecture segment, we explore two distinctly-purposed methods for video composition: a ranking mechanism and simplicial curves.

Another aspect to consider is the perceived impact of fragmentation in the final video owing to sentence-level composition. Notice that the text stream is directly mapped to the video using subtitles. As previously described for the lecture, we also apply LSA for measuring local coherence, by expanding the chosen candidate sentences to include other semantically-related sentences. This process constitutes the final step in the ranking mechanism. On the other hand, it is applied beforehand for simplicial curves, to improve the flow of information between the sequences of locally-coherent groups of sentences.

### 4.2.1 Ranking

Given the top-$n$ topic-related candidates for each lecture segment, we want to choose the sentence that best represents the group. For this reason, we use the Support Sets algorithm (introduced in Section 2.2.3), in order to let sets of candidate sentences compete among them and, in this way, obtain a sentence ranking. A support set is created for each candidate of the group, by computing the similarity between each candidate and the remaining ones. The geometrically closest candidates are selected for the support set. The most central candidates are the ones that occur in the largest number of support sets. In the current method, the final video is composed by the best ranked candidate sentences of documentaries. However, since the same sentence can be present in more than one lecture segment, which would result in redundant content in the final artifact, only unused candidates are chosen. Thus, if the best ranked candidate sentence already was selected, the next in rank is used.

### 4.2.2 Simplicial Curves

In the current work, we hypothesize that aspects of fluidity are incorporated in the video by minimizing the geometric distance between the end and initial segments of successive curves. Thus, given two sets of top-$n$ candidates corresponding to consecutive lecture segments, a curve is generated for each candidate, using a simplex composed from the collection of documentaries. We start by performing all comparison combinations between each of the final curve

segments of the first set and the initial curve segments of the second set. Hence, the best match of candidates is determined by the geometrically closest end to start points of these curves. After selecting the first two candidates, we proceed to compare the ending points of the second curve with the start points of the curves generated for the top-$n$ candidates corresponding to the next lecture segment. The process is repeated until a candidate is selected for each lecture segment. As in the ranking method, we circumvent the inclusion of repeated information. If the best match corresponds to a candidate that already was selected, the next geometrically closest is used.

## 4.3 Summary

In this chapter, we presented our approach for the automatic generation of lecture-driven science talks videos. We identify meaningful and diverse segments from the input lecture's subtitles, and obtain topic-related content from a collection of documentaries. Coherence is addressed as a consequence of the content selection process, as well as by means of video composition techniques via content organization, namely centrality-based summarization for ranking, and simplicial curves for fluidity.

# 5 Preliminary Study: Summarization of Films and Documentaries

Automatic summarization has been widely explored in different domains and media. While the input media has varied from the original work on text summarization (Luhn 1958; Edmundson 1969) to current work on speech (Maskey and Hirschberg 2005; Zhang, Chan, and Fung 2010; Ribeiro and de Matos 2012) and video summarization (Ajmal, Ashraf, Shakir, Abbas, and Shah 2012a), the application domain has not shown the same amount of diversity since most of the work concentrates on news (Barzilay, Elhadad, and McKeown 2002; Radev, Otterbacher, Winkel, and Blair-Goldensohn 2005; Ribeiro and de Matos 2007; Hong, Conroy, Favre, Kulesza, Lin, and Nenkova 2014), meetings (Murray, Renals, and Carletta 2005; Garg, Favre, Reidhammer, and Hakkani-Tür 2009), and lectures (Fujii, Kitaoka, and Nakagawa 2007), i.e., in general, informative sources. Nevertheless, different application areas within the entertainment industry are gaining significant attention. For instance, summarization of literary short stories in the interest of helping readers decide whether the story is worth reading (Kazantseva and Szpakowicz 2010), application of generic summarization algorithms to music (Raposo, Ribeiro, and de Matos 2015), summarization of books (Mihalcea and Ceylan 2007), or inclusion of character analyses into movie summaries (Sang and Xu 2010).

In this study, we progress in a similar direction, creating extractive, single-document, text summaries for films, and documentaries. In that sense, we assess the performance of these entertainment programs, and how they compare to the popular, and well-known, news domain, in summarization. We make a clear distinction between films and documentaries. Typically, in filmmaking, documentaries are related to some construction of reality (Grant and Sloniowski 1998), and are, therefore, seen as inherently more truthful than fictional films. Even though they do not fundamentally differ regarding narrative structure, the collection of collected documentaries address, almost entirely, scientific subjects. On the other hand, the collected films tell a story based on fictive events. Thus, we consider films as being fictional, and documentaries as non-fictional. As such, we study the parallelism between the informational content carried out by their subtitles and scripts. In that sense, we use five text-based summarization approaches

(described in detail in Section 2.2), namely: LSA (Gong and Liu 2001), LexRank (Erkan and Radev 2004), Support Sets (Ribeiro and de Matos 2011), MMR (Carbonell and Goldstein 1998) , and GRASSHOPPER (Zhu, Goldberg, Gael, and Andrzejewski ). Additionally, we follow an intrinsic approach for evaluation. We compare the generated automatic summaries against human-made abstracts, composed of plot summaries and synopses, for films and documentaries.

## 5.1 Datasets

We use three distinct datasets in our experiments: newspaper articles, films, and documentaries. They have different characteristics with regard to their extrinsic nature, such as the average document size, and their underlying semantic properties. The fictional nature of films manifests itself, to a great extent, as dialog between characters of the film. At the other end of the spectrum, non-fictional documentaries consist predominantly of monologue. The inclusion of the newspaper articles dataset serves as reference, to comparatively assess the performance of the generic text summarization algorithms in the different domains. As previously shown, most text summarization approaches have been widely tested on this type of data.

We present a brief review of the properties of the datasets used in the experiments, along with plot summaries and synopses for films and documentaries, and human-made reference summaries for newspaper articles. Usually, a plot summary consists of a description, concise and thorough enough for the reader to get a sense of what happens during a film. A plot synopsis is commonly much longer than a plot summary, more detailed, and it may contain important details concerning the turn of events in the story.

### 5.1.1 Newspaper Articles

TeMário (Pardo and Rino 2003) is composed by 100 newspaper articles written in Brazilian Portuguese, covering several subjects, for instance “world”, “politics”, and “foreign affairs”. For each newspaper article, there is a human-made abstract summary used as reference. Table 5.1 sums up the properties of this dataset.

Table 5.1: TeMário corpus characterization.

| | | AVG | MIN | MAX |
|---|---|---|---|---|
| #Sentences | News Story | 29 | 12 | 68 |
| | Summary | 9 | 5 | 18 |
| #Words | News Story | 608 | 421 | 1315 |
| | Summary | 192 | 120 | 345 |

### 5.1.2 Films

A total of 30 films were collected, with an average of 4 plot summaries (minimum of 1, maximum of 7), and 1 plot synopsis per film. Table 5.2 presents a brief overview of the properties of the film's subtitles, scripts, and the concatenation of both. Not all the information present in the scripts was used: all characters' dialogs were removed in order to make them more similar to a plot summary. Table 5.3 presents a brief analysis of the number of plot summaries per film.

Table 5.2: Properties of subtitles and scripts.

| | | AVG | MIN | MAX |
|---|---|---|---|---|
| #Sentences | Subtitles | 1421 | 773 | 3000 |
| | Script | 1870 | 738 | 3721 |
| | Script + Subtitles | 3291 | 1819 | 5389 |
| #Words | Subtitles | 9337 | 4663 | 20749 |
| | Script | 17165 | 6311 | 30980 |
| | Script + Subtitles | 27257 | 14309 | 41751 |

Table 5.3: Properties of plot summaries and synopses.

| | | AVG | MIN | MAX |
|---|---|---|---|---|
| #Sentences | Plot Summaries | 5 | 1 | 21 |
| | Plot Synopses | 120 | 6 | 399 |
| #Words | Plot Summaries | 102 | 17 | 438 |
| | Plot Synopses | 2261 | 236 | 7110 |

### 5.1.3 Documentaries

A total of 98 scientific documentaries were collected for this experiment. Table 5.4 presents a brief overview of the properties of their subtitles.

In addition, we collected 223 human-made plot summaries. Upon screening each one of them, four different classes of summaries were identified and labeled (Table 5.5). Namely, out of 223 human-made summaries, 143 were classified as "Informative", 63 as "Interrogative", 9

Table 5.4: Properties of documentaries subtitles.

| | AVG | MIN | MAX |
|---|---|---|---|
| #Sentences | 340 | 212 | 656 |
| #Words | 5864 | 3961 | 10490 |

as "Inviting", and 8 as "Challenge". Each class of summaries has a different nature: "Informative" summaries contain factual information about the program; "Interrogative" summaries typically contain questions that are meant to arouse the curiosity of potential viewers, such as "What is the meaning of life?"; the "Inviting" class represents any invitation to embark on a journey such as "Got time for a 24 year vacation? Then consider a journey to our most distant planets (...)" or "Pack a good spacesuit and plenty of oxygen and prepare to be amazed". Lastly, the "Challenge" class entices the viewer on a personal basis, such as "are you ready for (...)", or "a deadly asteroid is on a collision course with Earth (...), are we next?", or "Don't try this at home!". We chose "Informative" summaries due to their resemblance to the sentences extracted by the summarization algorithms. On average, there are 2 informative plot summaries per documentary (minimum of 1, maximum of 3).

Table 5.5: Properties of the documentary plot summaries.

| | | AVG | MIN | MAX |
|---|---|---|---|---|
| #Sentences | Informative | 4 | 1 | 18 |
| | Interrogative | 4 | 1 | 19 |
| | Inviting | 6 | 2 | 11 |
| | Challenge | 5 | 2 | 9 |
| #Words | Informative | 82 | 26 | 384 |
| | Interrogative | 103 | 40 | 377 |
| | Inviting | 146 | 63 | 234 |
| | Challenge | 104 | 59 | 192 |

## 5.2 Experimental Setup

In our experiments, the baseline is established by analysis of news summarization, using the TeMário corpus. For news articles, a summary was generated with the average size (in terms of sentences) of the corresponding human-made abstract ($\approx 31\%$ of their size). For each film, two summaries were generated with a size equal to (i) the average number of sentences of its human-made plot summaries ($\approx 1\%$ of the their size), and (ii) the same length of the corresponding synopsis. In addition, for this domain, three types of input were considered: script,

subtitles, script+subtitles. Each documentary was summarized with the same average number of sentences of its corresponding human-made plot summaries ($\approx 1\%$ of the documentary's size).

The evaluation was carried out using ROUGE (Lin 2004) (introduced in Section 2.3.1.2). As such, content quality of summaries is based on word overlap between generated summaries and their references. ROUGE-N favors recall and represents the fraction of words chosen by humans that are correctly identified by the automatic summarization algorithms. ROUGE-SU measures the overlap of skip-bigrams, which is any pair of words in their sentence order, with the addition of unigrams as counting unit. We limited the maximum gap length of skip-bigrams to 4.

Several parametrizations of the generic algorithms were used, we present only those that achieved the best results. In this manner, we evaluated subtitles and scripts against plot summaries and synopses to define an optimal performance reference. In the following section, we present the averaged ROUGE-1, ROUGE-2, and ROUGE-SU4 scores, and the performance of each summarization algorithm, as a ratio between the score of the generated summaries and this reference (relative performance). In addition, we assess the level of agreement among the manual summaries used for evaluation for each domain, similarly, using the averaged ROUGE scores.

## 5.3 Results and Discussion

News articles intend to answer basic questions about any particular event: who, what, when, where, why, and often how. This form of structure is sometimes addressed as the "inverted pyramid", where the most essential, and important, information comes first. Typically, the first sentences provide a good overview of the entire article and are more likely to be chosen for composing the final summary. Although documentaries follow a narrative structure similar to fictive films, they can be seen as more closely related to news than films, with regard to their intrinsic informative nature. In spite of the different nature of news articles, films, and documentaries, the summaries created by humans produce similar scores for all of them. As it can be observed from Figure 5.1, the abstractive, or "ideal", summaries used in our experiments, share comparable behavior. Note that documentaries achieve higher recall scores than news

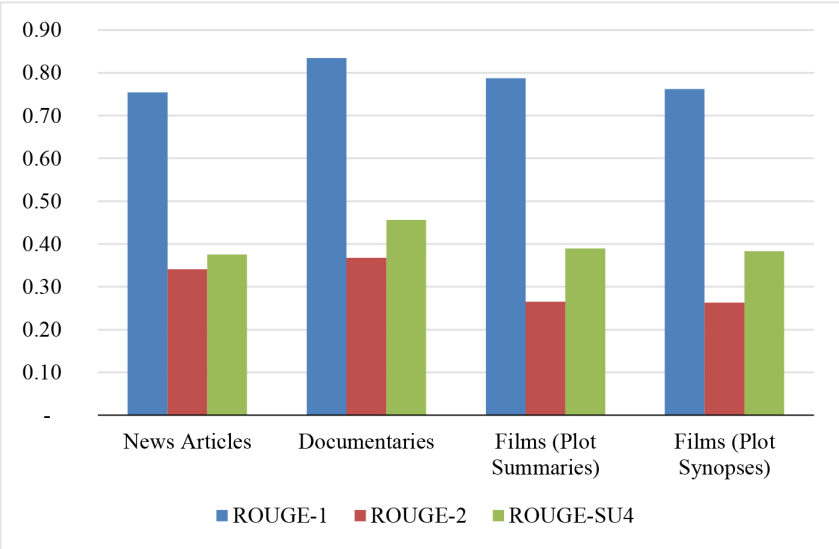



Figure 5.1: ROUGE recall scores using the original documents against manual references for news articles, plot summaries for documentaries, and plot summaries and synopses for films.

articles and films, using the original subtitles documents against the corresponding manual plot summaries.

Figure 5.2 presents an overview of the performance of each generic summarization algorithm across all domains. The results concerning news documents, using the TeMário (Pardo and Rino 2003) dataset, were the best out of all three datasets used in the experiments. There is one fundamental difference between TeMário and the films and documentaries datasets. The news summaries were created using a compression rate of, approximately, 31% of the size of the original articles (Table 5.1), in terms of sentences, which is significantly lower than films and documentaries. This leads to higher recall scores. Nonetheless, we can observe the differences in behavior between these domains (Figure 5.2). Notably, documentaries achieve the best results for plot summaries, with a summary compression rate of 99%, in comparison with films, using scripts, subtitles, or the combination of both. Additionally, as depicted in Figure 5.2, the experiments conducted combining scripts and subtitles for films, on a general note, do not improve results, with an exception for the Support Sets algorithm for ROUGE-1.

Overall, LSA performed consistently better for news articles and documentaries, which is in accordance with the work presented for meeting recordings, where the best summarizer was LSA (Murray, Renals, and Carletta 2005). One possible reason for these results is that LSA

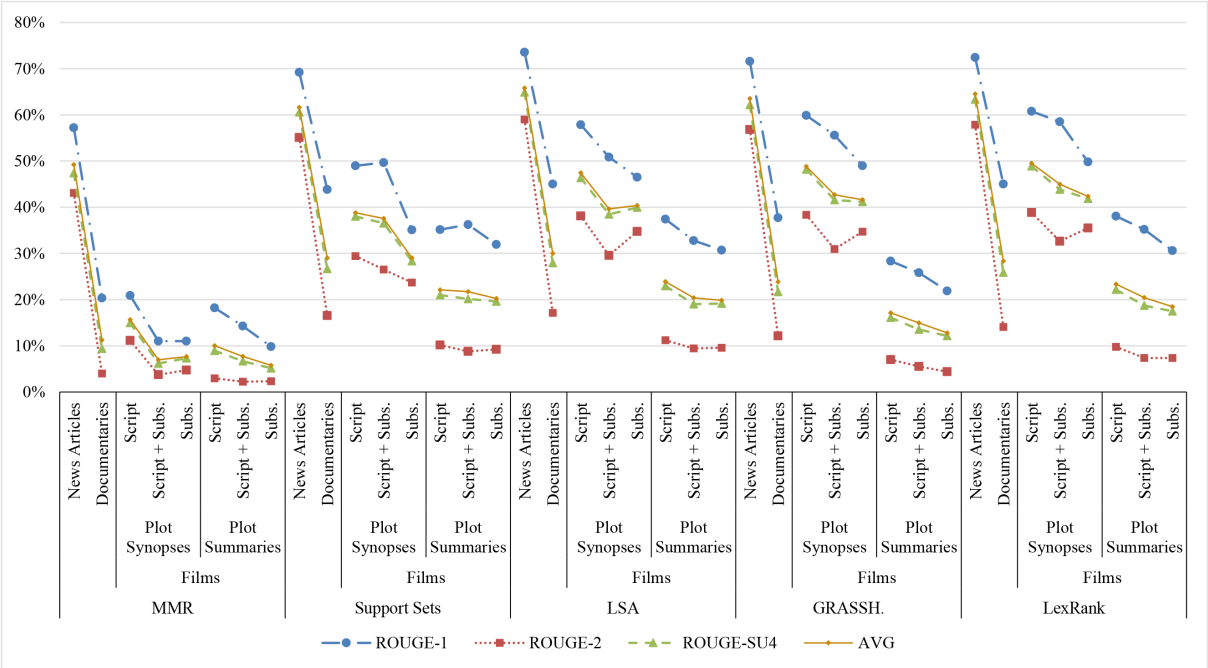



Figure 5.2: Relative performance for all datasets. For news articles, MMR was parametrized with $\lambda = 0.50$; Support Sets used Manhattan distance and Support Set Cardinality = 2. For documentaries the relative performance was measured against plot plot summaries: MMR used $\lambda = 0.75$; Support Sets used the cosine distance and threshold = $50\%$. For films the relative performance was measured against plot synopses and plot summaries: MMR used $\lambda = 0.50$; and Support Sets used the cosine distance and threshold = $50\%$.

tries to capture the relation between words in sentences. By inferring contextual usage of text based on these relations, high recall scores, apart from ROUGE-1, are produced for ROUGE-2 and ROUGE-SU4. For films, LexRank was the best performing algorithm for subtitles, scripts and the combination of both, using plot synopses, followed by LSA and Support Sets for plot summaries. Furthermore, MMR has the lowest scores for all metrics and all datasets. We observed that sentences closer to the centroid typically contain very few words, thus leading to shorter summaries and the corresponding low scores. Interestingly, by observing the average of ROUGE-1, ROUGE-2, and ROUGE-SU4, it is possible to notice that it follows very closely the values of ROUGE-SU4. The results from our experiments suggest that ROUGE-SU4 reflects the scores from both ROUGE-1 and ROUGE-2, being able to capture the concepts derived from either unigrams as well as bigrams.

In conclusion, considering plot summaries, non-fictional documentaries achieved higher results in comparison with fictional films. However, in general, the highest ROUGE score

for these two domains is achieved using films scripts against plot synopses. However, note that synopses have a significant difference in terms of sentences in comparison with plot summaries. As shown in Table 5.3, the average synopsis has 120 sentences, while plot summaries have, on average, 5 sentences for films, and 4 for documentaries (Table 5.5). This gives synopses a clear advantage in terms of recall, due to the high count of words. Additionally, it is possible to observe in Figure 5.2 that all algorithms behave similarly for both subtitles and scripts.

## 5.4 Summary

In this chapter, we assessed the performance of generic text summarization algorithms when applied to films and documentaries, using the well–known behavior of news articles summarization as reference. We presented the datasets used in our experiments which consist of (i) news articles, (ii) film scripts and subtitles, and (iii) documentary subtitles. The standard ROUGE metrics are used for evaluation. We show that the best performing algorithms are LSA, for news articles and documentaries, and LexRank and Support Sets, for films. Additionally, despite the different nature of films and documentaries, we show that their relative behavior is in accordance with the behavior obtained for news articles.

# 6 Multi-document Summarization of Science Talks

As previously described, summarization is a multi-faceted problem. The absence of ideal, or correct, summaries poses several difficulties in assessing their quality. Different human summarizers tend to create distinct summaries, when using the same source of information. Despite well-known limitations, automatic measures for summary evaluation are frequently employed. They provide a means for understanding several defining characteristics of automatic summaries (e.g., informativeness), typically based on how well they compare to a gold-standard. This enables us to determine the methods and systems that are best suited for specific scenarios.

Building upon the previous study, for lecture-driven science talks, we progress towards evaluation involving human subjective judgments. Our work involves many elements not easily objectified in current automated evaluation strategies. We focus on illustrating the main topics addressed in a physics lecture, using a collection of scientific documentaries. In this context, summarization represents only part of the content selection process. Even though several techniques have been proposed to summarize video (Truong and Venkatesh 2007), formal talks, such as lectures, contain the most useful information comprised within the audio channel (He, Sanocki, Gupta, and Grudin 1999). The amateur nature of such recordings renders most video summarization techniques seldom applicable, due to: very few, or non-existent, shot transitions; generally bad video quality, lacking color variation; as well as, absent video composition and production techniques. On the other hand, documentaries, while maintaining some portrayal of factual nature, follow narrative structures, production techniques, along with other cinematography practices (Grant and Sloniowski 1998). On account of their entertainment-driven nature, their content is typically presented in a style appropriate for a general audience. Since our approach is directed for human consumption, we assess several factors that require a deeper understanding of the output on the basis of user studies. As it happens with subjective measures, effective quality assessment is firmly grounded on the clear factor specifications of the task being performed. Broadly, we measure the adequacy of the selected content, by assess-

ing how much of the topic information in the lecture is portrayed in the final video, along with overall elements of coherence, as well as overall quality. In addition, we compare two different methods for video composition regarding their inclusion of hypothesized properties of fluidity, and composition, in the generated video.

In this chapter, we present the datasets collected and used in our experiments. Then, we describe the experimental setup, followed by the results and discussion of science talks videos concerning a user study.

## 6.1 Datasets

In this section, we present a brief overview of the datasets used for our experiments: lectures; and documentaries. The inclusion of the lectures dataset, serves to provide a general idea of the properties of the input document that is used to guide the subsequent phases of the content selection process. In addition, building upon the collection of documentaries employed in the previous study (Section 5.1.3), we present the properties of the complete collection.

### 6.1.1 Lectures

A total of 119 physics lectures were collected, we present a brief overview of their corresponding transcripts (Table 6.1). They represent part of the educational material archived at MIT OCW, and correspond to several one semester courses, namely: the early universe; classical mechanics; electricity and magnetism; and, vibration and waves.

Table 6.1: Properties of the lectures transcripts.

| | AVG | MIN | MAX |
|---|---|---|---|
| #Sentences | 497 | 32 | 915 |
| #Words | 8366 | 496 | 12900 |

### 6.1.2 Documentaries

A total of 277 documentaries were collected. Table 6.2 presents the properties of their subtitles. They constitute, almost entirely, scientific-related material. The predominant subject is correspondent to matters of the universe, for instance: its formation and development, the

chemistry and physics of interstellar objects, or examination of the cosmos, involving the study of the stars, planets, comets, and galaxies. As well as, the physics embodied in the fundamental forces of the universe. Other, more marginal subjects inside the collection, encompass the history of science, and the manifestations of life on Earth.

Table 6.2: Properties of documentaries subtitles.

| | AVG | MIN | MAX |
|---|---|---|---|
| #Sentences | 345 | 20 | 2080 |
| #Words | 6141 | 1719 | 39110 |

## 6.2 Experimental Setup

Following the proposed architecture (Section 4), we summarize the input lecture using GRASSHOPPER, with a summary compression rate of 99%, to obtain meaningful, and diverse sentences, representative of the topic spans within the lecture. In general terms, the average document (Table 6.2), is condensed to 5 sentences, in which the original sentence order is preserved.

Then, we explore the application of LSA for local measures of coherence (Section 3.1), and span the sentences of the summarized lecture to include other, semantically-related, sentences, concerning the same subject. We use Armadillo's (Sanderson 2010) implementation of the SVD operation, and apply it to a terms by documents matrix, built through TF-IDF weights, using the lecture dataset. As previously mentioned, the magnitude of the singular values indicates the importance degree of a particular combination pattern of words within the document. Thus, in the derived semantic space, we choose the number of dimensions whose values are greater than half of the highest singular value in the diagonal matrix $\Sigma$ (Steinberger and Jezek 2004). Additionally, the latent matrix of left singular vectors $U$ represents the importance degree of each word, for each salient topic/concept. In this sense, we represent each sentence of the original lecture, by the mean of the vectors of its words, represented in the $U$ matrix, using the previously determined dimensions. And, assign a coherence score for each lecture sentence, by assessing its similarity to adjoining sentences computing the cosine of their means, as given by Equation 3.1. Then, for each sentence of the summarized lecture, we iteratively include adjoining sentences if their assigned local coherence score is greater than a threshold (which we fixed to 0.8), in this way, forming locally-coherent groups.

In the context of our 2-stage process (Section 4.1), we train a LDA model for the documentary dataset, at document-level. Using its vocabulary, we choose training the model for 100 topics, with the hyperparameter $\alpha = 0.3$. These choices depend on a several number of factors, regarding the nature of the text collection, in general terms, some approaches suggest the following rule of thumb: $\alpha = 50/T$, where $T$ represents the number of topics (Steyvers and Griffiths 2007). Then, we perform variational inference to the locally-coherent summarized lecture, using the estimated model. We compare the topic mixture of the inferred lecture, using cosine distance, with those estimated for the documentary dataset. From this, we compose a subset of topic-related documentaries, whose values are greater than a third of the highest topic mixture similarity value. In the subsequent stage, we train a 10-topic model for this subset, segmented at sentence-level, with $\alpha = 0.5$. In a similar process, inferring the summarized lecture in the estimated model, we assess topic mixture similarity: for each of the lecture's locally-coherent group of sentences, we obtain the top-10 candidate, topic-related, documentary sentences.

For video composition, concerning the ranking mechanism, Support Sets used the cosine distance, and threshold $= 50\%$ (method described in Section 4.2.1). Moreover, simplicial curves were parametrized with $\sigma = 0.01$, and Euclidean distance determines the geometrically closest end to start segments of curves, which correspond to the first, or last, 10 words of the corresponding sentence. The procedure is described in detail in Section 4.2.2. The final step consists of spanning the chosen candidate sentences employing LSA, as previously described, applying SVD to a terms by documents matrix, now using the documentary dataset. Each group of locally-coherent sentences is included in the final video by considering the portion from the initial timestamp, corresponding to the first sentence in the group, to the last sentence's timestamp.

Additionally, joining audio/video segments from different sources results in the final audio stream lacking loudness consistency, as well as, abrupt video transitions amongst different documentaries. Thus, regarding the presentation of the video, we begin and end each video segment, corresponding to a group of locally-coherent sentences, with fade-in/fade-out effects, respectively. All video processing tasks (e.g., segmentation, re-enconding, and concatenation) were accomplished via FFMPEG (Bellard, Niedermayer, Sierralta Pulento, et al. 2003). All processing tasks for audio manipulation (e.g., dynamic range compression) were accomplished via Sound eXchange (SoX) (Bagwell 1991).

### 6.2.1 Pre-Processing

As it happens with large collections of text, the most common textual elements (stopwords) constitute noise for text-processing tasks, and for that reason are usually filtered out. In this context, text often follows an empirical law, in which the frequency of occurrence of any word in the dataset is inversely proportional to its rank in the frequency list: the second most frequent occurring word would correspond to half the frequency of the first in the list, and so forth, following the Zipfian distribution (Kurumada, Meylan, and Frank 2013). Figure 6.1, depicts the frequency of the top 100 words in the documentary dataset. The most frequent words constitute a small amount of the total number of words in the dataset, and greatly surpass all others. In this sense, we detected stopwords according to two different strategies: we choose the top $\approx 0.01\%$, according to frequency of occurrence (Figure 6.1); as well as, inspect the matrix $U$ using LSA, for the words that have the largest index values in the first dimensions.

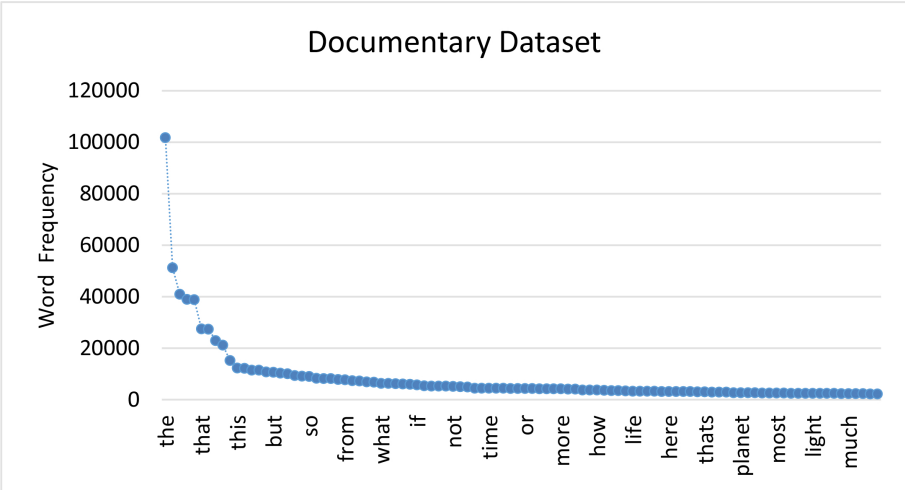



Figure 6.1: Word frequency of the documentary dataset: the words illustrated in the horizontal axis correspond to a subset of the total frequency displayed.

Furthermore, we apply a stemming procedure to the vocabulary used for LDA. This step, allows the topic model to consider similar meaning words that have the same root, as the same entity, by the process of removing their commoner morphological and inflectional, endings. Otherwise, these words would be considered as separate entities in the model with reduced importance. We apply the Porter's stemmer algorithm, a conservative approach for stemming that utilizes suffix stripping (Porter 1980). As an example, Figure 6.2 illustrates the top most

| "Asteriods" | "Black Holes" |
|---|---|
| asteroid | hole |
| asteroids | holes |
| impact | gravity |
| hit | galaxy |
| objects | centre |
| atmosphere | mass |
| object | black |
| scientists | physics |
| solar | horizon |
| meteorites | wormhole |

(a) Model trained without stemming.

| "Biology" | "Solar System" |
|---|---|
| cell | jupit |
| discoveri | system |
| dna | atmospher |
| gene | neptun |
| genet | pluto |
| caus | solar |
| diseas | surfac |
| bacteria | orbit |
| work | saturn |
| human | uranu |

(b) Model trained with stemmed text data.

Figure 6.2: Top 10 words for random topics in the documentary dataset (manually labeled).

probable words, for randomly chosen topics, in the 100-topic model trained for the documentary dataset at document-level. Before stemming (Figure 6.2a), the distribution over topics places similar probabilities for semantically similar words such as plurals: "asteroid" and "asteroids", for the topic "Asteroids"; or, "hole" and "holes", for the topic "Black Holes". On the other hand, by reducing topically similar words to their root form, the overall number of words reduces, which makes the text data less sparse. As such, the most probable words for each of the generated topics represent more accurate descriptions of the thematic structure of the collection (Figure 6.2b).

## 6.3 Results

For our study, we developed two questionnaires for collecting data from human subjective judgments (see Appendix A). We present videos regarding 4 physics lectures, and for each, we generated two videos on the basis of two different video composition techniques: a ranking mechanism; and simplicial curves. In this sense, the first two videos of the first questionnaire were composed using ranking, and the last two, simplicial curves. In the second questionnaire, guiding the video generation process using the same physics lectures, the first two videos were composed using simplicial curves, while the remaining two using ranking. The form's estimated time of completion is of 30 minutes. The average length of each video is of 5 minutes.

We collected a total of 49 answers, consisting of 2 groups of 26, and 23 anonymous answers, for the first and second questionnaires, respectively. In the following sections, we present an

analysis of the first part of the questionnaires regarding the users background, followed by, an analysis of the lecture-driven video results.

### 6.3.1 User Analysis

The gender population is predominantly male, consisting of 34 male, and 15 female answers. The demographic age consists mostly of users ranging [18-25] years of age – 29 answers; followed by the next age group of [26-49] – 12 answers; and the last group of 50, or more, years of age – 8 answers. In regard to the level of education, participants were asked their highest degree earned: the majority of answers constitutes a total of 41% for bachelor's degree; followed by 27% for master's degree; 20% for high school; and lastly, 12% of the respondents earned a level of education corresponding to doctorate or higher. In addition, we asked if the academic degree obtained is related to the subject of physics, of which 53% answered "yes", and 47% answered "no". Even though the question is ambiguous (e.g., some engineering courses might be considered tangent to the subject), it helps us understand the users point of view in this matter.

Furthermore, we assess the users' interest in the field of science, using a scale of 1 ("no interest") to 10 ("frequently follow scientific matters"). Prior to this question, we present a vague description of the scientific field as: "any matters concerned with the exploration of nature by construction of knowledge, and discovery of new facts". In Figure 6.3, the box plot illustrates that, in general, the users are highly interested in these matters, with median of 8. Figure 6.4 represents how frequently the users watch documentaries: almost half (49%) of the respondents reported watching documentaries on a monthly basis; 27% watch on a weekly basis; and 6% daily. Also, a small portion of users reported to occasionally (8%), or never (10%) watch documentaries.

### 6.3.2 Science Talks Videos

We precede each video presented in the questionnaire with a brief overview of the lecture that guided the process of content selection. After each video, we ask users to enumerate at least 3 topics that best characterize the content displayed and validate the results accordingly. In this sense, we assess the quality of the videos by asking the users to evaluate (i) if the topics of the video are related to the previously presented lecture summary; (ii) coherence, broadly defined as an overall sense of understandability, taking into account if the clips that compose

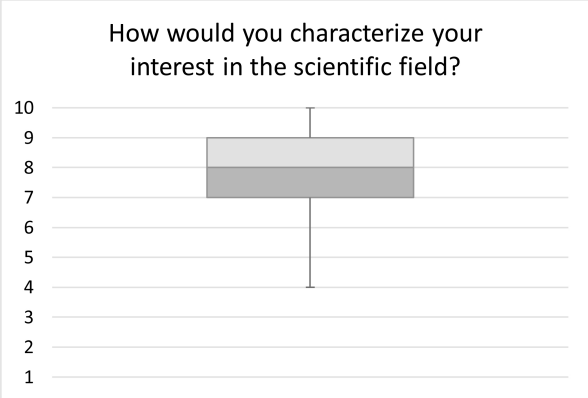



Figure 6.3: Users interest in the scientific field, characterized by a scale from 1 ("no interest") to 10 ("frequently follow scientific matters").

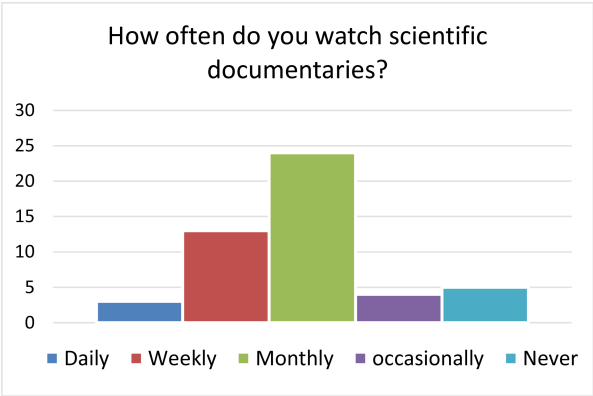
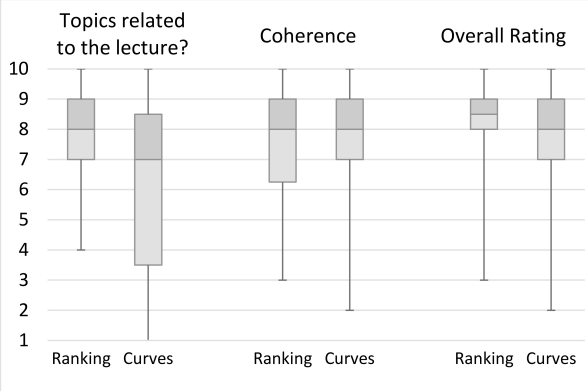


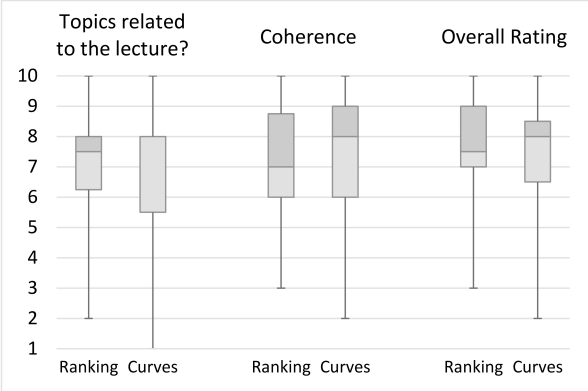

Figure 6.4: Results in terms of the number of users concerning the frequency of watching scientific documentaries.

the video are related, fit well together, and function as a whole; (iii) and, overall quality of the video. In light of this, we also compare the previously proposed methods for video composition, ranking and simplicial curves. Additionally, we perform Wilcoxon signed-rank tests to detect if there are differences between the scores attributed to both of these methods. Briefly, in this procedure, the absolute values of the differences between observations are ranked: the smallest gets rank of 1; then the next larger is ranked 2; and so forth, in which ties are attributed a rank equal to the average of the ranks they span. The ranks of all differences are added separately for positive and negative values. The smaller of these two sums is the test statistic W. We use the Z-value (standard score) to evaluate if the hypothesis that regards no difference between distributions falls in the region of rejection. Accordingly, we present the results for the Z-value, p-value, and the corresponding confidence interval.

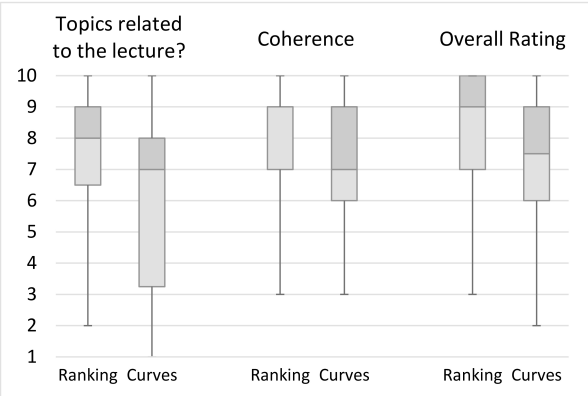
Figure 6.5 illustrates the results obtained regarding the videos produced for each lecture. The response options for these questions used a rating scale from 1 (lowest score) to 10 (highest score). For the first question, which concerns the topic relation between the video and lecture, it is possible to observe from the ranking results in Figure 6.5a that the distributional characteristic of the group of scores, is comprised mostly between 7 and 9 (median: 8, minimum: 4, maximum: 10). For curves, the results are lower with data being more skewed for the second quartile group (ranging from the first quartile to the median). Hence, 25% of the responses from users, which considered the video to not be significantly related to the lecture, tend to have dispersed opinion results that range from the score 3.5 to 7. For the second lecture (Figure 6.5b), concerning the same question, the results are slightly higher for curves than ranking, with the



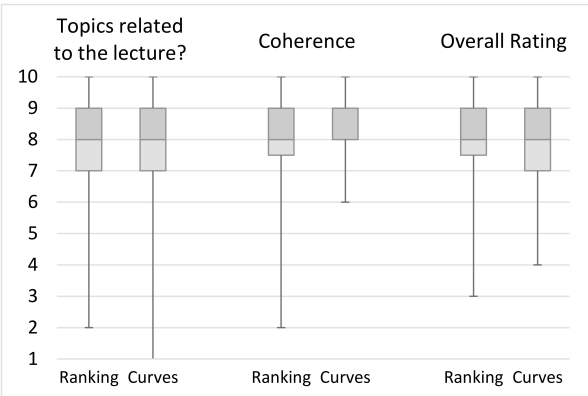

(a) Lecture 1: Basic Concepts of Newtonian Mechanics.




(b) Lecture 2: Black-Body Radiation and the Early History of the Universe.




(c) Lecture 3: The Cosmic Microwave Background Radiation.




(d) Lecture 4: Inflationary Cosmology Is Our Universe Part of a Multiverse.

Figure 6.5: Results for science talks videos generated for the four lectures in the questionnaires. For each question, we present the relative ratings of each composition technique, ranking and simplicial curves.

upper quartile equivalent to the median with score 8, but with higher dispersion of data. For the third lecture (Figure 6.5c), the difference between methods is more significant, with ranking achieving higher scores, similarly to the first lecture. For the forth lecture (Figure 6.5d), there is little statistical variability between the methods, which show symmetrical first, second, and third quartiles. In conclusion, we performed the Wilcoxon test using the scores for ranking and curves, which indicates that the results achieved with ranking are statistically significantly higher than those of curves (Z-value: -4.3101; p-value: 0; significant results at $p \leq 0.01$).

In what concerns results pertaining coherence, the overall patterns of response are closely related for both composition techniques. The first (Figure 6.5a), and forth lecture (Figure 6.5d), achieve the same median scores of 8. Regarding the second quartile group, response data is

less skewed for curves than ranking. Especially regarding the forth lecture, users tend to agree for curves mostly between 8 and 9 of rating (minimum: 6, maximum: 10). On the other hand, comparing the second (Figure 6.5b), and third lecture (Figure 6.5c), it is possible to observe mixed results concerning the median scores for both methods. While for the second lecture the median score is higher for curves (median: 8) in comparison to ranking (median: 7), for the third lecture, ranking achieves a median score of 9 in comparison to 7, for curves. We performed the Wilcoxon signed-rank test using the scores of both composition methods, to determine if there is a statistical difference regarding the distribution of scores. The results indicate that there is no difference between the methods (Z-value: -0.1042; p-value: 0.92034; the result is not significant at $p \leq 0.05$).

Concerning the overall quality rating, there are no clear requirements for the answer besides the users own personal opinion, which is biased by their background and interest in these matters. As shown for the first (Figure 6.5a) and forth lecture (Figure 6.5d), there is a high level of agreement amongst users for ranking, which is denoted by the small range between the first and third quartile. With the exception of the second lecture, composition by ranking seems to provide more favorable results according to the users overall rating. Notably, concerning the third lecture for ranking, 50% of the total respondents have a distributional score comprised within the rating of 9 and 10. Additionally, we perform the Wilcoxon test using the scores for both composition techniques, which indicates that the results achieved with ranking are statistically significantly higher than those of curves (Z-value: -2.4082; p-value: 0.00798; significant results at $p \leq 0.01$).

Furthermore, users were asked to identify the topics that best characterize each video presented. The most frequent topic is shared across both composition methods for the videos generated for each lecture. In accordance with the results concerning topic relation of the video and lecture, the generated videos for the first (Figure 6.5a) and third lecture (Figure 6.5c), present more variability regarding the most frequent topics. For the first lecture, the most frequently identified topics in the video generated using ranking encompass: gravity, Einstein's general relativity, and Newton's laws. Accordingly, for the video using curves, the most frequent topics are: gravity, string theory, mass, and wormhole. For the third lecture, the most frequent topics for ranking are: big bang, inflation, background radiation, and universe. Accordingly, for curves the most frequent topics are: big bang, fundamental particles, matter, and neutrino.

On the other hand, the videos for the second (Figure 6.5b), and forth lecture (Figure 6.5d) share most common topics. Namely, the topics black holes, and radiation, are shared for both composition methods for the second lecture, with the exception of speed of light for ranking, and ozone layer for curves. For the forth lecture, the most frequent topics are shared between methods: big bang; universe; and multiverse.

Overall, results suggest that video composition performed using ranking is better for retaining the topics of the original document. Also, the produced videos are more in accordance with users' satisfaction. In addition, regarding aspects of coherence, we were not able to detect statistical differences regarding the rating between the videos composed according to ranking and curves. In regard to the open questions, most critics from users were positive, remarking that the videos enabled good consolidation of information, regarding different sources, about the same subject. Other viewers, mentioned some abrupt transitions between parts of the video, while others suggested presenting a smooth transition to the end of the videos, instead of abruptly terminating. Specifically for the second lecture using ranking, viewers identified the segment of the video regarding ozone layer radiation as being less coherent with the rest of the video that refers to black-body radiation. Other viewers, without prior knowledge of the composition methods used, identified the videos composed with ranking better in terms of the knowledge being transmitted, than the videos composed using curves.

## 6.4 Discussion

The generation of multimedia artifacts oriented for human consumption involves several concerns, which encompass techniques for content selection and coherence. In the context of entertainment-driven programs, the viewers overall satisfaction is greatly influenced by the presentation and organization of content. In our work, we hypothesized that aspects of fluidity and flow of information are incorporated in the final video by minimizing the geometric distance between the end and initial segments of successive curves, generated for documentary candidate sentences. The results of our statistical analysis regarding the rating of users for both composition techniques, led us to conclude that none of the methods actually manages coherence, as it is perceived by viewers. Topical coherence is being dragged from the process of content selection. The documentary sentences that compose the video are always chosen from each top-$n$ candidate sentences that are topic-related to the lecture. In addition, LSA is used

to expand the chosen candidates, to include other semantically-related sentences. In this way, we minimize the perceived impact of fragmentation owing to sentence-level substitution, since the final video is composed by a set of locally-coherent groups of sentences. Still, modeling sequence using simplicial curves achieves very high results with considerable user agreement for the first and forth lectures (Figure 6.5). However, these results are not very consistent, considering the remaining lectures. On the whole, composition using Support Sets for establishing a ranking yields better results with less distributional variability, regarding the topic relation between the video and lecture, and overall rating. We parametrized the algorithm with threshold at 50%, which allows the inclusion of half of the input for each support set created. We determine a sentence ranking by letting each set of top-$n$ candidate sentences compete. As such, the top sentences are the most central (important) in the group, concerning information more similar to all other candidates. And, for this reason, the selected sentences will constitute a video more related to the topics conveyed by the lecture. On the contrary, for composition using simplicial curves, some of the sentences selected can be less topic related in comparison, but promote flow of information regarding the video as a whole. This explains the more dispersed ratings for some lectures regarding the results from Figure 6.5.

As previously mentioned, several viewers pointed out a particular video segment from the second lecture as being out of context, which concerns the ozone layer and UV radiation. This is a recurring problem associated to relying on the bag-of-words assumption for LDA. Word order is often critical to capture the meaning of text in various contexts. Nonetheless, it is possible provide, and introduce, more context using contiguous sequences of words, as opposed to representing text as unigrams, as it is the case for our topic model. In this work, words are considered as separate entities in the model with different degrees of importance. As such, the top most probable words of the topic that originated the video segment out of context are: radiation, black, cosmic, gamma, and burst. Through the process of inferring the lecture segment in the model, the candidate documentary sentence regarding ozone radiation and the UV rays in the atmosphere is given high probability for the previous topic. As a result, the sentence is included in the top-$n$ candidates for that lecture segment. In our experiments, we mitigated this problem by representing the documents as bigrams, which provide more context and resolve similar issues. This process was not included for the evaluation procedure. However, in this case, the bigrams ozone radiation, or UV radiation, will not be likely to occur with high probability in the previous topic. This ambiguity is resolved for other cases, for

instance consider the following examples from the model trained using bigrams: big bang; speed light; black hole; magnetic field; dark energy. In these examples, considering the words from the bigrams as separate concepts, disrupts the LDA model from obtaining more accurate, and less ambiguous, topic related content, resulting in the occasional inclusion of segments out of context in the video.

In previous experiments, one of the issues identified, and resolved, was the pace at which the video progressed. The final artifact was the result of concatenating video clips that corresponded to sentence-level segments. As such, the video included a large number of jump cuts, corresponding to fast pace changes of image and subject. Besides the introduction of fade-in/fade-out effects to perform gradual transitions between segments, we explored the application of LSA for local measures of coherence. The assigned cosine score by which adjoining sentences are included was fixed at 0.8. This threshold will ensure the inclusion of highly-coherent discourse, however, depending on the context of the documentary sentence, several sentences can be included spanning minutes in the final video. Without imposed limits, the final video can lack overall structure, by having a particular segment spanning over half of the video, while the remaining being fast paced. Nonetheless, by this process we can incorporate important visual concepts that would be discarded otherwise, due to the constriction of presenting only speech-related segments.

## 6.5 Summary

In this chapter, we presented the evaluation for lecture-driven science talks videos involving human subjective judgments. We presented the datasets used, which consist of lectures and documentaries, and the experimental setup carried out. The evaluation involved an analysis of the participants in the study, followed by the results of our videos comparing ranking and curves for composition. We show that video composition using ranking is better for retaining the topics of the original document. The videos produced were also rated higher than curves concerning overall quality. For coherence, there was not any statistical differences between the methods. Lastly, the chapter ends with a discussion of the results, including critics from the participants in the study.

# 7 Conclusion

In this thesis, we explored methods for the generation of video summaries in the context of multimedia artifacts for human consumption. We presented a computational method for illustrating the main topics addressed in a physics lecture using a collection of scientific documentaries. We structure an entertainment-driven artifact having as main concerns: content selection; and aspects of coherence. Lectures are informational talks, and therefore contain most of the important information comprised within the audio stream. For that reason, we use the audio transcript of a physics lecture to guide our creation process. As such, we obtain a concise, and diverse representation of the lecture, and select topic-related content from the subtitles of documentaries. We addressed the issue of coherence as a consequence of the content selection process, and by means of video composition techniques via content organization. In this sense, we compared two different methods for composition regarding their inclusion of hypothesized properties of fluidity, and overall well-composedness. Furthermore, in order to minimize the perceived impact of fragmentation in the final video, owing to sentence-level composition, we expand each of the topic-related sentences to include other adjoining semantically-related sentences, and thereby form locally-coherent sets.

We conducted a preliminary study intended to assess the behavior of films and documentaries in summarization. We analyzed the impact of the five summarization algorithms on three datasets. The newspaper articles dataset was used as a reference. The other two datasets, consisting of films and documentaries, were evaluated against plot summaries, for films and documentaries, and synopses, for films. Despite the different nature of these domains, the abstractive summaries created by humans, used for evaluation, share similar scores across ROUGE metrics. The best performing algorithms are LSA, for news and documentaries, and LexRank and Support Sets, for films. Overall, considering plot summaries, non-fictional documentaries achieved higher results in comparison with fictional films. For diversity-based algorithms, GRASSHOPPER achieves the best results in comparison with MMR for all domains and ROUGE metrics. Additionally, it is possible to observe that all algorithms behave simi-

larly for both subtitles and scripts. As previously mentioned, the average of the scores follows closely the values of ROUGE-SU4, suggesting that ROUGE-SU4 is able to capture concepts derived from both unigrams and bigrams.

For lecture-driven science talks videos, we progressed towards evaluation involving human subjective judgments. We collected data from two questionnaires regarding a total of 49 answers. The videos were generated using 4 physics lectures on the basis of composition using the Support Sets algorithm for ranking, and simplicial curves. We measured how much of the topic information in the lecture is portrayed in the final video; coherence broadly defined as an overall sense of understandability; and the overall quality of the video. The participants in the user study shown considerable interest in scientific matters with median score of 8 (scale of 1 ("no interest") to 10 ("frequently follow scientific matters"), and 82% of the total participants watch documentaries, at least, on a monthly basis. Overall, the ranking mechanism is better for retaining the topics of the original lecture, and producing videos more in accordance with users' judgments of overall quality. For coherence, we did not detect statistical differences between the composition techniques. However, while modeling sequence using the curves method is able to achieve higher results with more user agreement for two of the lectures, ranking achieves more consistent distributional scores in general. In light of these results, selecting the most central (important) segments from a group of topic-related content, is a better strategy to present the topics of the original lecture, and incorporate other properties that positively influenced the viewers overall measures of quality.

## 7.1 Main Contributions

This thesis presented a computational method for the automatic generation of lecture-driven science talks.

In a preliminary study we show that despite the different nature of films and documentaries, their relative behavior is in accordance with that obtained for news articles. LSA performed consistently better for news articles and documentaries. For films, LexRank was the best performing algorithm. MMR has the lowest scores for all metrics and all datasets. The results from our experiments suggest that ROUGE-SU4 is able to capture the concepts derived from either unigrams as well as bigrams.

Building upon this study, for lecture-driven science talks we assess the quality of the videos on the basis of ranking and simplicial curves for content organization. We show that video composition performed using ranking is consistently better for retaining the topics of the original lecture. Moreover, concerning the overall quality rating, the generated videos have higher ratings than simplicial curves with less distributional variability. We show that none of the composition techniques actually addressed coherence, as it is perceived by viewers.

## 7.2 Future Work

In consideration of the identified critics, a particular video segment was pointed out as being out of context. We show that building the vocabulary as unigrams for the LDA model may introduce ambiguity, by considering words that may be related as separate entities. In this case, UV radiation and ozone radiation were fitted by the model in a topic that mostly concerns cosmic background radiation. Contextual information can be incorporated by using contiguous sequences of words. Exploring the effects of different $n$-grams in the first stage of our process (topic model trained at document-level), might be beneficial for mitigating this ambiguity. As a result, documentaries such as the one concerning the effects of UV radiation, would not be included in the subset of documentaries in the second stage of the process (topic model trained at sentence-level).

Furthermore, the text stream is directly mapped to the video using subtitles. Occasionally, the time interval corresponding to the sentences of the subtitles does not encompass the speech that it is portraying, or starts and ends abruptly, without padding. This issue can be resolved by utilizing the underlying audio stream to provide continuity cues. For that, a data-drive voice-activity detector based on long short-term memory recurrent neural networks can be used (Eyben, Weninger, Squartini, and Schuller 2013).

# Bibliography


Ahanger, G. and T. D. C. Little (1998). Automatic composition techniques for video production. *IEEE Transactions on Knowl. and Data Eng. 10*(6), 967–987.

Ajmal, M., M. Ashraf, M. Shakir, Y. Abbas, and F. Shah (2012a). Video summarization: Techniques and classification. In *Computer Vision and Graphics*, Volume 7594, pp. 1–13. Springer Berlin Heidelberg.

Ajmal, M., M. H. Ashraf, M. Shakir, Y. Abbas, and F. A. Shah (2012b). Video summarization: Techniques and classification. In *Proc. of the 2012 Intl. Conf. on Computer Vision and Graphics*, ICCVG'12, Berlin, Heidelberg, pp. 1–13. Springer-Verlag.

Anjum, M., S. Ali, M. Hassan, and M. Adnan (2013). Video summarization: Sports highlights generation. In *Multi Topic Conf. (INMIC), 2013 16th Intl.*, pp. 142–147.

Aparício, M., P. Figueiredo, F. Raposo, D. M. de Matos, and R. Ribeiro (2015). Summarization of Films and Documentaries Based on Subtitles and Scripts. *CoRR abs/1506.01273*.

Arman, F., A. Hsu, and M.-Y. Chiu (1993). Image processing on compressed data for large video databases. In *Proc. of the First ACM Intl. Conf. on Multimedia*, MULTIMEDIA '93, New York, NY, USA, pp. 267–272. ACM.

Baber, J., N. Afzulpurkar, M. Dailey, and M. Bakhtyar (2011). Shot boundary detection from videos using entropy and local descriptor. In *Digital Signal Processing (DSP), 2011 17th Intl. Conf. on*, pp. 1–6.

Bagwell, C. (1991). Sox - sound exchange, the swiss army knife of audio manipulation. accessed 2015/09/21.

Bailer, W., H. Mayer, H. Neuschmied, W. Haas, M. Lux, and W. Klieber (2003). Content-based video retrieval and summarization using mpeg-7.

Balagopalan, A., L. Balasubramanian, V. Balasubramanian, N. Chandrasekharan, and A. Damodar (2012). Automatic keyphrase extraction and segmentation of video lectures. In *Technology Enhanced Education (ICTEE), 2012 IEEE Intl. Conf. on*, pp. 1–10.

Balasubramanian, V., S. Doraisamy, and N. Kanakarajan (2015). A multimodal approach for extracting content descriptive metadata from lecture videos. *Journal of Intelligent Information Systems*, 1–25.

Bartsch, M. A. and G. H. Wakefield (2005). Audio Thumbnailing of Popular Music using Chroma-based Representations. *IEEE Transactions on Multimedia 7*(1), 96–104.

Barzilay, R., N. Elhadad, and K. McKeown (2002). Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research 17*(1), 35–55.

Barzilay, R. and K. R. McKeown (2005). Sentence fusion for multidocument news summarization. *Computational Linguistics 31*(3), 297–328.

Baxendale, P. (1958). Machine-made index for technical literature—an experiment. *IBM Journal of Research Development 2*(4), 354–361.

Bellard, F., M. Niedermayer, J. J. Sierralta Pulento, et al. (2003). Ffmpeg multimedia system. accessed 2015/09/21.

Berger, A. and V. O. Mittal (2000). Query-relevant summarization using faqs. In *Proc. of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, Stroudsburg, PA, USA, pp. 294–301. Association for Computational Linguistics.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Bligh, D. A. (2000). *What's The Use of Lectures?* (1 ed.). Jossey-Bass.

Bocconi, S., F. Nack, and L. Hardman (2008). Automatic generation of matter-of-opinion video documentaries. *Web Semant. 6*(2), 139–150.

Boreczky, J. S. and L. A. Rowe (1996). Comparison of Video Shot Boundary Detection Techniques. In *Storage and Retrieval for Still Image and Video Databases IV*, Number SPIE 2664, Los Angeles, California.

Borth, D., A. Ulges, C. Schulze, and T. Breuel (2008). Keyframe extraction for video tagging & summarization. In *Informatiktage*, Volume S-6 of *LNI*, pp. 45–48. GI.

Branigan, E. (1992). *Narrative Comprehension and Film*. Routledge.

Brin, S. and L. Page (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proc. of the 7th Intl. Conf. on World Wide Web*, pp. 107–117.

Carbonell, J. and J. Goldstein (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proc. of the 21st Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR '98, New York, NY, USA, pp. 335–336. ACM.

Chai, W. (2006). Semantic Segmentation and Summarization of Music: Methods Based on Tonality and Recurrent Structure. *IEEE Signal Processing Magazine 23*(2), 124–132.

Chen, L., S. J. Rizvi, M. T. Özsu, and M. Tamer (2003). Incorporating audio cues into dialog and action scene extraction. In *In Proc. of SPIE Conf. on Storage and Retrieval for Media Databases*, pp. 252–264.

Christel, M. G., M. A. Smith, C. R. Taylor, and D. B. Winkler (1998). Evolving video skims into useful multimedia abstractions. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, CHI '98, New York, NY, USA, pp. 171–178. ACM Press/Addison-Wesley Publishing Co.

Coldefy, F., P. Bouthemy, M. Betser, and G. Gravier (2004). Tennis video abstraction from audio and visual cues. In *Multimedia Signal Processing, 2004 IEEE 6th Workshop on*, pp. 163–166.

Cong, Y., J. Yuan, and J. Luo (2012). Towards scalable summarization of consumer videos via sparse dictionary selection. *Multimedia, IEEE Transactions on 14*(1), 66–75.

Das, D. and A. F. T. Martins (2007). A survey on automatic text summarization. Technical report, Literature Survey for the Language and Statistics II course at Carnegie Mellon University.

Edmundson, H. P. (1969). New methods in automatic abstracting. *Journal of the Association for Computing Machinery 16*(2), 264–285.

Ekin, A., A. Tekalp, and R. Mehrotra (2003). Automatic soccer video analysis and summarization. *Img. Processing, IEEE Transactions on 12*(7), 796–807.

Erkan, G. and D. R. Radev (2004). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research 22*(1), 457–479.

Evangelopoulos, G., A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis (2013). Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia 15*(7), 1553–1568.

Eyben, F., F. Weninger, S. Squartini, and B. Schuller (2013). Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE Intl. Conf. on*, pp. 483–487.

Ferman, A. and A. Tekalp (2003). Two-stage hierarchical video summary extraction to match low-level user browsing preferences. *Multimedia, IEEE Transactions on 5*(2), 244–256.

Feynman, R., R. Leighton, and M. Sands (1963). *The Feynman Lectures on Physics* (Second ed.), Volume 1. Boston: Addison-Wesley.

Figueiredo, P., M. Aparício, D. M. de Matos, and R. Ribeiro (2015). Generation of multimedia artifacts: An extractive summarization-based approach. *CoRR abs/1508.03170*.

Foltz, P. W., W. Kintsch, and T. K. Landauer (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes 25*(2-3), 285–307.

Foote, J. (1999). Visualizing music and audio using self-similarity. In *Proc. of the Seventh ACM Intl. Conf. on Multimedia (Part 1)*, MULTIMEDIA '99, New York, NY, USA, pp. 77–80. ACM.

Fujii, Y., N. Kitaoka, and S. Nakagawa (2007). Automatic extraction of cue phrases for important sentences in lecture speech and automatic lecture speech summarization. In *Proc. of INTERSPEECH 2007*, pp. 2801–2804. ISCA.

Furui, S. (2007). Recent advances in automatic speech summarization. In *Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications) - RIAO 2007, 8th*

*Intl. Conf., Carnegie Mellon University, Pittsburgh, PA, USA, May 30 - June 1, 2007. Proc., CD-ROM.*

Garg, N., B. Favre, K. Reidhammer, and D. Hakkani-Tür (2009). ClusterRank: A Graph Based Method for Meeting Summarization. In *Proc. of INTERSPEECH 2009*, pp. 1499–1502. ISCA.

Genest, P.-E. and G. Lapalme (2011). Framework for abstractive summarization using text-to-text generation. In *Proc. of the Workshop on Monolingual Text-To-Text Generation*, MTTG '11, Stroudsburg, PA, USA, pp. 64–73. Association for Computational Linguistics.

Gibson, D. P., N. W. Campbell, and B. T. Thomas (2002). Visual abstraction of wildlife footage using gaussian mixture models and the minimum description length criterion. In *ICPR (2)*, pp. 814–817. IEEE Computer Society.

Glaczynski, J. and E. Lukasik (2011). Automatic Music Summarization: A "Thumbnail" Approach. *Archives of Acoustics 36*(2), 297–309.

Gong, Y. and X. Liu (2001). Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In *Proc. of the 24th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 19–25. ACM.

Grant, B. K. and J. Sloniowski (1998). *Documenting the Documentary: Close Readings of Documentary Film and Video*. Wayne State University Press.

He, L., E. Sanocki, A. Gupta, and J. Grudin (1999). Auto-summarization of audio-video presentations. In *Proc. of the Seventh ACM Intl. Conf. on Multimedia (Part 1)*, MULTIMEDIA '99, New York, NY, USA, pp. 489–498. ACM.

Hohl, L., F. Souvannavong, B. Merialdo, and B. Huet (2004). Using structure for video object retrieval. In P. Enser, Y. Kompatsiaris, N. O'Connor, A. Smeaton, and A. Smeulders (Eds.), *Img. and Video Retrieval*, Volume 3115 of *Lecture Notes in Computer Science*, pp. 564–572. Springer Berlin Heidelberg.

Hong, K., J. M. Conroy, B. Favre, A. Kulesza, H. Lin, and A. Nenkova (2014). A repository of state of the art and competitive baseline summaries for generic news summarization. In *Proc. of the Ninth Intl. Conf. on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pp. 1608–1616.

Hovy, E. and C.-Y. Lin (1998). Automated text summarization and the summarist system. In *Proc. of a Workshop on Held at Baltimore, Maryland: October 13-15, 1998*, TIPSTER '98, Stroudsburg, PA, USA, pp. 197–214. Association for Computational Linguistics.

Hu, D. J. (2009). Latent dirichlet allocation for text, images, and music.

Hu, W., N. Xie, L. Li, X. Zeng, and S. Maybank (2011). A Survey on Visual Content-Based Video Indexing and Retrieval. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 41*(6), 797–819.

Hunt, M., M. Lennig, and P. Mermelstein (1980). Experiments in syllable-based recognition of continuous speech. In *Acoustics, Speech, and Signal Processing, IEEE Intl. Conf. on ICASSP '80.*, Volume 5, pp. 880–883.

Irie, G., T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa (2010). Automatic trailer generation. In *Proc. of the Intl. Conf. on Multimedia*, pp. 839–842. ACM.

Ježek, K. and J. Steinberger (2008). Automatic text summarization (the state of the art 2007 and new challenges). *Proc. of Znalosti*, 1–12.

Jing, H., R. Barzilay, K. McKeown, and M. Elhadad (1998). Summarization evaluation methods: Experiments and analysis. In *AAAI Symposium on Intelligent Summarization*, pp. 60–68.

Ju, S., M. Black, S. Minneman, and D. Kimber (1998). Summarization of videotaped presentations: automatic analysis of motion and gesture. *Circuits and Systems for Video Technology, IEEE Transactions on 8*(5), 686–696.

Jun, S. and E. Hwang (2013). Music segmentation and summarization based on self-similarity matrix. In *Proc. of the 7th Intl. Conf. on Ubiquitous Information Management and Communication*, ICUIMC '13, New York, NY, USA, pp. 82:1–82:4. ACM.

Kan, M.-Y. and J. L. Klavans (2002). Using librarian techniques in automatic text summarization for information retrieval. In *Proc. of the 2Nd ACM/IEEE-CS Joint Conf. on Digital Libraries*, JCDL '02, New York, NY, USA, pp. 36–45. ACM.

Kan, M.-Y., K. R. McKeown, and J. L. Klavans (2001). Domain-specific informative and indicative summarization for information retrieval. In *Proceedings of the Document Understanding Conference*, New Orleans, U.S.A., pp. 19–26. ACM.

Kazantseva, A. and S. Szpakowicz (2010). Summarizing short stories. *Computational Linguistics 36*(1), 71–109.

Knight, K. and D. Marcu (2000). Statistics-based summarization - step one: Sentence compression. In *Proc. of the Seventeenth National Conf. on Artificial Intelligence and Twelfth Conf. on Innovative Applications of Artificial Intelligence*, pp. 703–710. AAAI Press.

Koprinska, I. and S. Carrato (2001). Temporal video segmentation: A survey. In *Signal Processing: Img. Communication*, pp. 477–500.

Kurumada, C., S. C. Meylan, and M. C. Frank (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition 127*(3), 439 – 453.

Laganière, R., R. Bacco, A. Hocevar, P. Lambert, G. Païs, and B. Ionescu (2008). Video summarization from spatio-temporal features. In P. Over and A. F. Smeaton (Eds.), *TVS*, pp. 144–148. ACM.

Landauer, T. K. and S. T. Dutnais (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review 104*(2), 211–240.

Landauer, T. K., P. W. Foltz, and D. Laham (1998). An introduction to latent semantic analysis. *Discourse processes 25*(2-3), 259–284.

Lebanon, G., Y. Mao, and J. Dillon (2007). The locally weighted bag of words framework for document representation. *Journal of Machine Learning Research 8*, 2405–2441.

Li, Y. and B. Merialdo (2014). Multimedia maximal marginal relevance for multi-video summarization. *Multimedia Tools and Applications*, 1–22.

Lienhart, R. (1999a). Comparison of automatic shot boundary detection algorithms. In *Storage and Retrieval for Image and Video Databases*, Number SPIE 3656, pp. 290–301.

Lienhart, R. (1999b). Dynamic video summarization of home video. In *Storage and Retrieval for Media Databases 2000*, Volume 3972 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pp. 378–389.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summ. Branches Out: Proc. of the ACL-04 Workshop*, pp. 74–81. Association for Computational Linguistics.

Lin, C.-Y. and E. Hovy (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proc. of the 2003 Conf. of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, Stroudsburg, PA, USA, pp. 71–78. Association for Computational Linguistics.

Logan, B. and S. Chu (2000). Music summarization using key phrases. In *In Proc. IEEE ICASSP*, pp. 749–752.

Lu, Z. and K. Grauman (2013). Story-driven summarization for egocentric video. In *Proc. of the 2013 IEEE Conf. on Computer Vision and Pattern Recognition*, CVPR '13, Washington, DC, USA, pp. 2714–2721. IEEE Computer Society.

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development 2*(2), 159–165.

Luo, B., X. Tang, J. Liu, and H. Zhang (2003). Video caption detection and extraction using temporal information. In *Img. Processing, 2003. ICIP 2003. Proc. 2003 Intl. Conf. on*, Volume 1, pp. I–297–300 vol.1.

Ma, Y.-F., X.-S. Hua, L. Lu, and H.-J. Zhang (2005). A generic framework of user attention model and its application in video summarization. *Multimedia, IEEE Transactions on 7*(5), 907–919.

Ma, Y.-F., L. Lu, H.-J. Zhang, and M. Li (2002). A user attention model for video summarization. In *Proc. of the Tenth ACM Intl. Conf. on Multimedia*, MULTIMEDIA '02, New York, NY, USA, pp. 533–542. ACM.

Mani, I. (2001). Summarization evaluation: An overview. In *Proc. of the Third Second Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization, NTCIR-2, Tokyo, Japan, March 7-9, 2001*.

Mani, I. and E. Bloedorn (1999). Summarizing similarities and differences among related documents. *Information Retrieval 1*(1-2), 35–67.

Mani, I. and M. T. Maybury (2001). Automatic summarization. In *Association for Computational Linguistic, 39th Annual Meeting and 10th Conf. of the European Chapter, Companion Volume to the Proc. of the Conf.: Proc. of the Student Research Workshop and Tutorial Abstracts, July 9-11, 2001, Toulouse, France.*, pp. 5.

Mann, W. C. and S. A. Thompson (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text 8*(3), 243–281.

Maskey, S. R. and J. Hirschberg (2005). Comparing Lexical, Acoustic/Prosodic, Structural and Discourse Features for Speech Summarization. In *Proc. of the 9$^{th}$ EUROSPEECH - INTERSPEECH 2005*, pp. 621–624.

McCarthy, M. (1991). *Discourse Analysis for Language Teachers*. Cambridge University Press.

McKeown, K., J. Hirschberg, M. Galley, and S. Maskey (2005). From text to speech summarization. In *2005 IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing, ICASSP '05, Philadelphia, Pennsylvania, USA, March 18-23,2005*, Volume 5, pp. 997–1000. IEEE.

McKeown, K., R. J. Passonneau, D. K. Elson, A. Nenkova, and J. Hirschberg (2005). Do summaries help? In *Proc. of the 28th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR '05, New York, NY, USA, pp. 210–217. ACM.

McKeown, K. R., R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman (2002). Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proc. of the Second Intl. Conf. on Human Language Technology Research*, HLT '02, San Francisco, CA, USA, pp. 280–285. Morgan Kaufmann Publishers Inc.

Mihalcea, R. and H. Ceylan (2007). Explorations in automatic book summarization. In *EMNLP-CoNLL'07*, pp. 380–389.

Money, A. G. and H. Agius (2008). Video summarisation: A conceptual framework and survey of the state of the art. *Journal Vis. Comun. Img. Represent. 19*(2), 121–143.

Murray, G., T. Kleinbauer, P. Poller, S. Renals, J. Kilgour, and T. Becker (2008). Extrinsic summarization evaluation: A decision audit task. In A. Popescu-Belis and R. Stiefel-

hagen (Eds.), *Machine Learning for Multimodal Interaction*, Volume 5237 of *Lecture Notes in Computer Science*, pp. 349–361. Springer Berlin Heidelberg.

Murray, G., S. Renals, and J. Carletta (2005). Extractive Summarization of Meeting Recordings. In *Proc. of the 9th European Conf. on Speech Communication and Technology*, pp. 593–596.

Nam, J. and A. H. Tewfik (2002). Event-driven video abstraction and visualization. *Multimedia Tools Appl. 16*(1), 55–77.

Nenkova, A. (2006). Summarization evaluation for text and speech: issues and approaches. In *INTERSPEECH 2006 - ICSLP, Ninth Intl. Conf. on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*.

Nenkova, A., S. Maskey, and Y. Liu (2011). Automatic summarization. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts of ACL 2011*, HLT '11, Stroudsburg, PA, USA, pp. 3:1–3:86. Association for Computational Linguistics.

Nenkova, A., R. Passonneau, and K. McKeown (2007). The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech Language Processing 4*(2).

Nenkova, A., L. Vanderwende, and K. McKeown (2006). A compositional context sensitive multi-document summarizer: Exploring the factors that influence summarization. In *Proc. of the 29th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR '06, New York, NY, USA, pp. 573–580. ACM.

Nichols, B. (1991). *Representing Reality: Issues and Concepts in Documentary*. Bloomington: Indiana University Press.

Paice, C. D. (1990). Constructing literature abstracts by computer: Techniques and prospects. *Inf. Processing Manage. 26*(1), 171–186.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu (2002). Bleu: A method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, Stroudsburg, PA, USA, pp. 311–318. Association for Computational Linguistics.

Pardo, T. A. S. and L. H. M. Rino (2003). Temario: a corpus for automatic text summarization. Technical Report NILC-TR-03-09, Núcleo Interinstitucional de Linguística Computacional (NILC).

Peeters, G. (2003). Deriving musical structures from signal analysis for music audio summary generation: "sequence" and "state" approach. In U. K. Wiil (Ed.), *CMMR*, Volume 2771 of *Lecture Notes in Computer Science*, pp. 143–166. Springer.

Peeters, G., A. La Burthe, and X. Rodet (2002). Toward Automatic Music Audio Summary Generation from Signal Analysis. In *Proc. of the 3rd Intl. Society for Music Information Retrieval Conf.*, pp. 94–100.

Penn, G. and X. Zhu (2008). A critical reassessment of evaluation baselines for speech summarization of spontaneous conversations. In *Proc. of the 46th Annual Meeting of the Association for Computational Linguistics*.

Porter, M. (1980). An algorithm for suffix stripping. *Program 14*(3), 130–137.

Priyamvada R Sachan, K. (2014). Enhanced generic video summarization using large scale categorization. *Intl. Journal of Innovative Research in Computer and Communication Eng. Signal Processing Magazine 2*(4).

Rabiner, L. and B.-H. Juang (1993). *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Radev, D., T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Çelebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, and Z. Zhang (2004). MEAD — A platform for multidocument multilingual text summarization. In *Conf. on Language Resources and Evaluation (LREC)*.

Radev, D. R., S. Blair-goldensohn, Z. Zhang, and R. S. Raghavan (2001). NewsInEssence: A System For Domain-Independent, Real-Time News Clustering and Multi-Document Summarization. In *Proc. of the First Intl. Conf. on Human Language Technology Research*, pp. 1–4.

Radev, D. R., W. Fan, and Z. Zhang (2001). Webinessence: a personalized web-based multi-document summarization and recommendation system. In *NAACL 2001 Workshop on Automatic Summarization*, pp. 79–88.

Radev, D. R., E. Hovy, and K. McKeown (2002). Introduction to the special issue on summarization. *Computational Linguistics 28*(4), 399–408.

Radev, D. R., H. Jing, and M. Budzikowska (2000). Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In *Proc. of the 2000 NAACL-ANLPWorkshop on Automatic Summarization - Volume 4*, NAACL-ANLP-AutoSum '00, Stroudsburg, PA, USA, pp. 21–30. Association for Computational Linguistics.

Radev, D. R., J. Otterbacher, A. Winkel, and S. Blair-Goldensohn (2005). NewsInEssence: Summarizing Online News Topics. *Communications of the ACM 48*(10), 95–98.

Radev, D. R. and D. Tam (2003). Summarization evaluation using relative utility. In *Proc. of the Twelfth Intl. Conf. on Information and Knowledge Management*, CIKM '03, New York, NY, USA, pp. 508–511. ACM.

Raposo, F., R. Ribeiro, and D. M. de Matos (2015). On the Application of Generic Summarization Algorithms to Music. *IEEE Signal Processing Letters 22*(1), 26–30.

Rath, G. J., A. Resnick, and T. R. Savage (1961). The formation of abstracts by the selection of sentences. Part I. Sentence selection by men and machines. *12*(2), 139–141.

Rautiainen, M., T. Ojala, and T. Seppanen (2004). Cluster-temporal browsing of large news video databases. In *Multimedia and Expo, 2004. ICME '04. 2004 IEEE Intl. Conf. on*, Volume 2, pp. 751–754 Vol.2.

Raventós, A., R. Quijada, L. Torres, and F. Tarrés (2015). Automatic summarization of soccer highlights using audio-visual descriptors. *SpringerPlus 4*(1).

Ren, R., H. Misra, and J. Jose (2010). Semantic based adaptive movie summarisation. In S. Boll, Q. Tian, L. Zhang, Z. Zhang, and Y.-P. Chen (Eds.), *Advances in Multimedia Modeling*, Volume 5916 of *Lecture Notes in Computer Science*, pp. 389–399. Springer Berlin Heidelberg.

Ribeiro, R. and D. de Matos (2007). Extractive summarization of broadcast news: Comparing strategies for european portuguese. In *TSD*, Volume 4629, pp. 115–122. Springer.

Ribeiro, R. and D. de Matos (2012). Summarizing speech by contextual reinforcement of important passages. In *Proc. of PROPOR 2012*, pp. 392–402.

Ribeiro, R. and D. M. de Matos (2011). Revisiting Centrality-as-Relevance: Support Sets and Similarity as Geometric Proximity. *Journal of Artificial Intelligence Research*, 275–308.

Rocchi, C. and M. Zancanaro (2003). Generation of video documentaries from discourse structures. In *In Proc. of 9th European Workshop on Natural Language Generation*, pp. 95–102.

Sanderson, C. (2010). Armadillo: An Open Source C++ Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments. Technical report, NICTA.

Sang, J. and C. Xu (2010). Character-based movie summarization. In *Proc. of the Intl. Conf. on Multimedia*, pp. 855–858. ACM.

Shi, Z., G. Melli, Y. Wang, Y. Liu, B. Gu, M. Kashani, A. Sarkar, and F. Popowich (2007). Question answering summarization of multiple biomedical documents. In Z. Kobti and D. Wu (Eds.), *Advances in Artificial Intelligence*, Volume 4509 of *Lecture Notes in Computer Science*, pp. 284–295. Springer Berlin Heidelberg.

Smith, M. and T. Kanade (1997). Video skimming and characterization through the combination of image and language understanding techniques. In *Computer Vision and Pattern Recognition, 1997. Proc., 1997 IEEE Computer Society Conf. on*, pp. 775–781.

Song, Y., J. Vallmitjana, A. Stent, and A. Jaimes (2015). Tvsum: Summarizing web videos using titles. In *The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Spärck Jones, K. (2007). Automatic summarising: The state of the art. *Inf. Processing Manage. 43*(6), 1449–1481.

Steinberger, J. and K. Jezek (2004). Using Latent Semantic Analysis in Text Summarization and Summary Evaluation. In *Proc. of ISIM*, pp. 93–100.

Steyvers, M. and T. Griffiths (2007). *Probabilistic Topic Models*. Lawrence Erlbaum Associates.

Sundaram, H. and S.-F. Chang (2001). Condensing computable scenes using visual complexity and film syntax analysis. In *Multimedia and Expo, 2001. ICME 2001. IEEE Intl. Conf. on*, pp. 273–276.

Svore, K. M., L. Vanderwende, and C. J. C. Burges (2007). Enhancing single-document summarization by combining ranknet and third-party sources. In J. Eisner (Ed.), *EMNLP-CoNLL*, pp. 448–457. ACL.

Taskiran, C. and E. Delp (2004). Video summarization. In *Digital Img. Sequence Processing, Compression, and Analysis*, Computer Eng. Series, Chapter 8. CRC Press.

Taskiran, C., Z. Pizlo, A. Amir, D. Ponceleon, and E. Delp (2006). Automated video program summarization using speech transcripts. *IEEE Transactions on Multimedia 8*(4), 775–791.

Tombros, A. and M. Sanderson (1998). Advantages of query biased summaries in information retrieval. In *Proc. of the 21st Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR '98, New York, NY, USA, pp. 2–10. ACM.

Tonomura, Y., A. Akutsu, K. Otsuji, and T. Sadakata (1993). Videomap and videospaceicon: Tools for anatomizing video content. In *Proc. of the INTERACT '93 and CHI '93 Conf. on Human Factors in Computing Systems*, CHI '93, New York, NY, USA, pp. 131–136. ACM.

Truong, B. T. and S. Venkatesh (2007). Video abstraction: A systematic review and classification. *ACM Transactions Multimedia Computational Commun. Appl. 3*(1).

Tsai, C.-M., L.-W. Kang, C.-W. Lin, and W. Lin (2013). Scene-based movie summarization via role-community networks. *Circuits and Systems for Video Technology, IEEE Transactions on 23*(11), 1927–1940.

Turpin, A., Y. Tsegay, D. Hawking, and H. E. Williams (2007). Fast generation of result snippets in web search. In *Proc. of the 30th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR '07, New York, NY, USA, pp. 127–134. ACM.

Uchihashi, S., J. Foote, A. Girgensohn, and J. Boreczky (1999). Video manga: Generating semantically meaningful video summaries. In *Proc. of the Seventh ACM Intl. Conf. on Multimedia (Part 1)*, MULTIMEDIA '99, New York, NY, USA, pp. 383–392. ACM.

Ueda, H., T. Miyatake, and S. Yoshizawa (1991). Impact: an interactive natural-motion-picture dedicated multimedia authoring system. In S. P. Robertson, G. M. Olson, and J. S. Olson (Eds.), *CHI*, pp. 343–350. ACM.

Vanderwende, L., H. Suzuki, C. Brockett, and A. Nenkova (2007). Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing and Management 43*(6), 1606–1618.

Varadarajan, R. and V. Hristidis (2006). A system for query-specific document summarization. In *Proc. of the 15th ACM Intl. Conf. on Information and Knowledge Management*, CIKM '06, New York, NY, USA, pp. 622–631. ACM.

Wang, J.-C., Y.-H. Yang, H.-M. Wang, and S.-K. Jeng (2012). The acoustic emotion gaussians model for emotion-based music annotation and retrieval. In *Proc. of the 20th ACM Intl. Conf. on Multimedia*, MM '12, New York, NY, USA, pp. 89–98. ACM.

Wu, X., B. Xu, Y. Qiao, and X. Tang (2012). Automatic music video generation: cross matching of music and image. In *Proc. of the 20th ACM Multimedia Conf., MM '12, Nara, Japan, October 29 - November 02, 2012*, pp. 1381–1382.

Xie, S. and Y. Liu (2008). Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization. In *Proc. - ICASSP, IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, pp. 4985–4988.

Xu, C., X. Shao, N. C. Maddage, and M. S. Kankanhalli (2005). Automatic music video summarization based on audio-visual-text analysis and alignment. In *Proc. of the 28th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR '05, New York, NY, USA, pp. 361–368. ACM.

Yeung, S., A. Fathi, and L. Fei-Fei (2014). Videoset: Video summary evaluation through text. *CoRR abs/1406.5824*.

Yoon, J.-C., I.-K. Lee, and S. Byun (2009). Automated music video generation using multi-level feature-based segmentation. *Multimedia Tools Appl. 41*(2), 197–214.

Zechner, K. and A. Waibel (2000). Minimizing Word Error Rate in Textual Summaries of Spoken Language. In *Proc. of the 1st North American Chapter of the Association for Computational Linguistics Conf.*, pp. 186–193. Association for Computational Linguistics.

Zhang, J. J., R. H. Y. Chan, and P. Fung (2010). Extractive Speech Summarization Using Shallow Rhetorical Structure Modeling. *IEEE Transactions on Audio Speech and Language Processing 18*(6), 1147–1157.

Zhang, Y., X. Zhang, C. Xu, and H. Lu (2007). Personalized retrieval of sports video. In *Proc. of the Intl. Workshop on Workshop on Multimedia Information Retrieval*, MIR '07, New York, NY, USA, pp. 313–322. ACM.

Zhao, B. and E. P. Xing (2014). Quasi real-time summarization for consumer videos. In *2014 IEEE Conf. on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 2513–2520. IEEE.

Zhu, X., J. Fan, A. K. Elmagarmid, and X. Wu (2003). Hierarchical video content description and summarization using unified semantic and visual similarity. *Multimedia Systems 9*(1), 31–53.

Zhu, X., A. B. Goldberg, J. V. Gael, and D. Andrzejewski. Improving Diversity in Ranking using Absorbing Random Walks. In *Proc. of the 5th NAACL - HLT*. ACL.

Zhuang, Y., Y. Rui, T. Huang, and S. Mehrotra (1998). Adaptive key frame extraction using unsupervised clustering. In *Img. Processing, 1998. ICIP 98. Proc. 1998 Intl. Conf. on*, Volume 1, pp. 866–870 vol.1.

# I Appendices

# A User test Questionnaire

# MSc - Video Summaries for Science Talks

This form was developed in the context of my master's thesis in Information Systems and Computer Engineering in Instituto Superior Técnico (IST), entitled "Video Summaries for Science Talks". We aim to develop a computational method that illustrates the main subjects addressed in a physics lecture using a collection of scientific documentaries. Your cooperation in completing this form is highly appreciated. The estimated time of completion is of 30 minutes. Your answers are anonymous, therefore do not put your name anywhere in the form. Any comments and suggestions are welcome!

* Required

**What is your age?** *

- 17 or less
- 18 - 25
- 26 - 49
- 50 or more

**What is your gender?** *

- Male
- Female

**What is your level of education?** *

Select your level of education (or similar) from the following list.

**Is your academic degree related, in some way, to physics (case study's theme)?**

If you answered with an academic degree superior to middle school in the previous question, anwser this question.

- Yes
- No

**How would you characterize your interest in the scientific field?** *

The scientific field is characterized as any matters concerned with the exploration of nature by construction of knowledge, and discovery of new facts. 1- I have no interest. 10 - Frequently follow scientific matters.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**How often do you watch scientific documentaries?** *

- Daily
- Weekly
- Monthly
- Never
- Other:

Continue »

16% completed

# MSc - Video Summaries for Science Talks

* Required

## Group I - Video 1/4

The following questions refer to an automatically produced video documentary that aims to address the main topics of a particular physics lecture. Next, we present sentences that intend to provide a brief representation of the lecture that was used to produce the video documentary:

---

- Today we will explore the three Newton's laws.

- Things that move don't move along a straight line and do not continue to move and the reason is there is gravity.

- Since all objects in this lecture hall and the Earth fall with the constant acceleration which is 'g' we can write down that the gravitational force would be 'm' times this acceleration 'g'.

- Once your speed approaches the speed of light then Newtonian mechanics no longer works. Then you have to use Einstein's theory of special relativity. So this is only valid as long as we have speeds that are substantially smaller say than the speed of light.

---

### 1. Click the following link to view the video: [https://drive.google.com/file/d/0Bxj8h_rDRTi4ZU5sbGctcnZueGc/view?usp=sharing](https://drive.google.com/file/d/0Bxj8h_rDRTi4ZU5sbGctcnZueGc/view?usp=sharing)

### List of topics

| | | | |
|---|---|---|---|
| Constellations | Gravity | Surface | Burst |
| Nasa | Orbit | Mass | Neutrino |
| Launch | Travel | Atmosphere | Dwarf |
| Shuttle | Light | Einstein | Higgs Boson |
| Spacecraft | Speed | Dark Matter | Particle |
| Satellite | Fall | Particle | Matter |
| Mission | Object | Light | Force |
| Water | Force | Black Hole | Theory |
| Rock | System | Massive | Field |
| Discovery | Magnet | Matter | Atmosphere |
| Asteroid | Field | Parallel | Bacteria |
| Extinct | Solar | Supernova | Magnetism |
| String Theory | Storm | Explosion | Electricity |
| Big Bang | Power | Massive | Molecule |

**2. What was the above video about?** *

Specify any words that represent the video's topics (at least 3). Use, but do not be limited to, the above list for help.

**3. In general, do you consider that the topics of the video are related to the summary of the lecture presented in the beginning of this section?** *

1- There is no relation between this video and the lecture. 10- This video accurately represents the lecture.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**4. How do you rate this video regarding coherence?** *

Consider coherence as an overall sense of understandability, take into account if the clips that compose the video are related, fit well together, and function as a whole. 1- Not at all coherent. 10- Very coherent.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**5. Overall, how do you rate this video?** *

1- Very bad. 10- Very good.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**6. Do you have any comments or suggestions about this video?**

« Back  Continue »

33% completed

# MSc - Video Summaries for Science Talks

* Required

## Video 2/4

The following questions refer to an automatically produced video documentary that aims to address the main topics of a particular physics lecture. Next, we present sentences that intend to provide a brief representation of the lecture that was used to produce the video documentary:

---

- One of the most famous equation in physics is Einstein's 'e' equals 'm' 'c' squared.

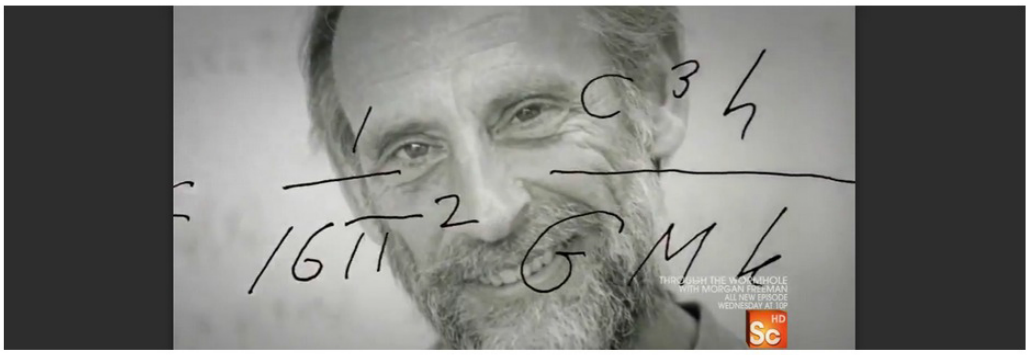
- Black-body radiation is a type of electromagnetic radiation within or surrounding a body in thermodynamic equilibrium with its environment, or emitted by a black-body (an opaque and non-reflective body) held at constant, uniform temperature. The radiation has a specific spectrum and intensity that depends only on the temperature of the body

- What is the literal sense of a black-body it's a body which is black in the sense that all radiation that hits it is absorbed. So it radiates black-body radiation.

- Black holes are near-perfect black-bodies, in the sense that they absorb all the radiation that falls on them.

---

### 1. Click the following link to view the video: [https://drive.google.com/file/d/0Bxj8h_rDRTi4VjQzMmJaQ0VVbFU/view?usp=sharing](https://drive.google.com/file/d/0Bxj8h_rDRTi4VjQzMmJaQ0VVbFU/view?usp=sharing)



### List of topics

| | | | |
|---|---|---|---|
| Constellations | Gravity | Surface | Burst |
| Nasa | Orbit | Mass | Neutrino |
| Launch | Travel | Atmosphere | Dwarf |
| Shuttle | Light | Einstein | Higgs Boson |
| Spacecraft | Speed | Dark Matter | Particle |
| Satellite | Fall | Particle | Matter |
| Mission | Object | Light | Force |
| Water | Force | Black Hole | Theory |
| Rock | System | Massive | Field |
| Discovery | Magnet | Matter | Atmosphere |
| Asteroid | Field | Parallel | Bacteria |
| Extinct | Solar | Supernova | Magnetism |
| String Theory | Storm | Explosion | Electricity |
| Big Bang | Power | Massive | Molecule |

**2. What was the above video about?** *

Specify any words that represent the video's topics (at least 3). Use, but do not be limited to, the above list for help.

**3. In general, do you consider that the topics of the video are related to the summary of the lecture presented in the beginning of this section?** *

1- There is no relation between this video and the lecture. 10- This video accurately represents the lecture.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**4. How do you rate this video regarding coherence?** *

Consider coherence as an overall sense of understandability, take into account if the clips that compose the video are related, fit well together, and function as a whole. 1- Not at all coherent. 10- Very coherent.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**5. Overall, how do you rate this video?** *

1- Very bad. 10- Very good.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**6. Do you have any comments or suggestions about this video?**

« Back  Continue »

50% completed

# MSc - Video Summaries for Science Talks

* Required

## Video 3/4

The following questions refer to an automatically produced video documentary that aims to address the main topics of a particular physics lecture. Next, we present sentences that intend to provide a brief representation of the lecture that was used to produce the video documentary:

---

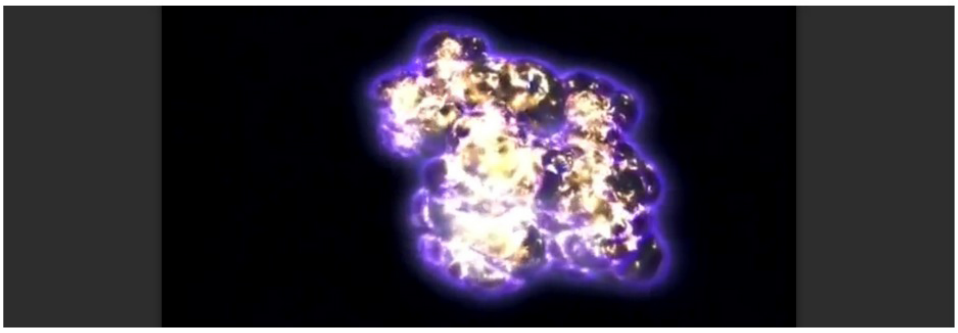
- Cosmic background radiation is electromagnetic radiation from the sky without discernible source. The origin of this radiation depends on the region of the spectrum that is observed. One component is the cosmic microwave background radiation. This component is redshifted photons that have freely streamed from an epoch when the Universe became transparent for the first time to radiation. Its discovery and detailed observations of its properties are considered one of the major confirmations of the Big Bang.

- I had suggested that we go on to talk about problems with the conventional Big Bang model. So far, we've been talking about the Big Bang theory without inflation.

- Assuming that physical influences are limited by the speed of light and according to everything that we know about the laws of physics it's true there is no way that the emission of that photon coming from that direction could have had any causal connection with the emission of the photon coming from the other direction.

---

### 1. Click the following link to view the video: https://drive.google.com/file/d/0Bxj8h_rDRTi4cWFweDFmbU1Pdnc/view?usp=sharing



### List of topics

| | | | |
|---|---|---|---|
| Constellations | Gravity | Surface | Burst |
| Nasa | Orbit | Mass | Neutrino |
| Launch | Travel | Atmosphere | Dwarf |
| Shuttle | Light | Einstein | Higgs Boson |
| Spacecraft | Speed | Dark Matter | Particle |
| Satellite | Fall | Particle | Matter |
| Mission | Object | Light | Force |
| Water | Force | Black Hole | Theory |
| Rock | System | Massive | Field |
| Discovery | Magnet | Matter | Atmosphere |
| Asteroid | Field | Parallel | Bacteria |
| Extinct | Solar | Supernova | Magnetism |
| String Theory | Storm | Explosion | Electricity |
| Big Bang | Power | Massive | Molecule |

**2. What was the above video about?** *

Specify any words that represent the video's topics (at least 3). Use, but do not be limited to, the above list for help.

**3. In general, do you consider that the topics of the video are related to the summary of the lecture presented in the beginning of this section?** *

1- There is no relation between this video and the lecture. 10- This video accurately represents the lecture.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**4. How do you rate this video regarding coherence?** *

Consider coherence as an overall sense of understandability, take into account if the clips that compose the video are related, fit well together, and function as a whole. 1- Not at all coherent. 10- Very coherent.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**5. Overall, how do you rate this video?** *

1- Very bad. 10- Very good.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**6. Do you have any comments or suggestions about this video?**

« Back  Continue »

66% completed

# MSc - Video Summaries for Science Talks

* Required

## Video 4/4

The following questions refer to an automatically produced video documentary that aims to address the main topics of a particular physics lecture. Next, we present sentences that intend to provide a brief representation of the lecture that was used to produce the video documentary:

---

- We discuss some of the implications of inflation and the multiverse we'll see that there are strong suggestions that the Big Bang was perhaps not really the beginning of existence but really just the beginning of our local universe often called a pocket universe.

- The Big Bang describes a number of important things and we'll be talking about this more as the course goes on.
It describes how the early universe expanded and cooled and we'll be spending a fair amount of time understanding the details behind those words.

- Later, we discuss more advanced topics like inflation.

---

## 1. Click the following link to view the video: https://drive.google.com/file/d/0Bxj8h_rDRTi4NTA5dFVpSml1VDg/view


## List of topics

| | | | |
|---|---|---|---|
| Constellations | Gravity | Surface | Burst |
| Nasa | Orbit | Mass | Neutrino |
| Launch | Travel | Atmosphere | Dwarf |
| Shuttle | Light | Einstein | Higgs Boson |
| Spacecraft | Speed | Dark Matter | Particle |
| Satellite | Fall | Particle | Matter |
| Mission | Object | Light | Force |
| Water | Force | Black Hole | Theory |
| Rock | System | Massive | Field |
| Discovery | Magnet | Matter | Atmosphere |
| Asteroid | Field | Parallel | Bacteria |
| Extinct | Solar | Supernova | Magnetism |
| String Theory | Storm | Explosion | Electricity |
| Big Bang | Power | Massive | Molecule |

**2. What was the above video about?** *

Specify any words that represent the video's topics (at least 3). Use, but do not be limited to, the above list for help.

**3. In general, do you consider that the topics of the video are related to the summary of the lecture presented in the beginning of this section?** *

1- There is no relation between this video and the lecture. 10- This video accurately represents the lecture.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**4. How do you rate this video regarding coherence?** *

Consider coherence as an overall sense of understandability, take into account if the clips that compose the video are related, fit well together, and function as a whole. 1- Not at all coherent. 10- Very coherent.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**5. Overall, how do you rate this video?** *

1- Very bad. 10- Very good.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**6. Do you have any comments or suggestions about this video?**

« Back  Continue »

83% completed

# MSc - Video Summaries for Science Talks

## Group - II

Consider all the videos presented and the context of this work.

**1. What is your opinion concerning the work being done?**

« Back  Submit

100%: You made it.

*Never submit passwords through Google Forms.*