Department of Computer Science
University of Fribourg, Switzerland

A STUDY ON MULTIMODAL DOCUMENT ALIGNMENT: BRIDGING THE GAP BETWEEN TEXTUAL DOCUMENTS AND SPOKEN LANGUAGE

Thesis

Submitted to the Faculty of Science, University of Fribourg (Switzerland)
to obtain the degree of Doctor Scientiarum Informaticarum

Dalila MEKHALDI

from

Oran, Algeria

Thesis N° 1521
Imprimerie St Paul, Fribourg
2006

Accepted by the Faculty of Science of the University of Fribourg (Switzerland), on the recommendation of:

- Prof. Heinz Gröflin, University of Fribourg, Switzerland (Jury president)
- Prof. Rolf Ingold, University of Fribourg, Switzerland (Ph.D supervisor)
- Prof. Peter King, University of Manitoba, Canada (Reporter)
- Prof. Jacques Savoy, University of Neuchâtel, Switzerland (Reporter)
- Dr. Denis Lalanne, University of Fribourg, Switzerland (Reporter)

Fribourg, June 8^{th}, 2006

Ph.D Supervisor:

Prof. Rolf Ingold

Faculty Dean:

Prof. Marco Celio

Dedication

I would like to dedicate this thesis to my parents, as a small symbol of my gratitude for their invaluable trust in me, without which, this work would not be carried out.

J'aimerais dédier cette thèse à mes parents, afin d'exprimer ma gratitude pour la confiance qu'ils ont placé en moi, et sans laquelle ce travail n'aurait jamais vu le jour.

Acknowledgements

This thesis is the result of three and half years of work, whereby I have been accompanied and supported by many people. It is a pleasant aspect that I have now the opportunity to express my gratitude for all of them.

- First and foremost I would like to thank my thesis supervisors, Prof. Rolf Ingold and Dr. Denis Lalanne at the Department of Computer Science in Fribourg, for their never ended support that made this thesis possible. In addition to making technical research, I ought to mention that I learned a host of other things such as technical writing and presentation, through their scrupulous but veracious guidance of this thesis.

- I would like to thank both Department of Computer Science in Fribourg and the FNS for providing resources and funding for my research assistantship throughout this PhD.

- Thanks also to the members of my PhD committee who monitored my work and took effort in reading and providing me with valuable comments on this thesis: Prof. Heinz Gröflin, Prof. Peter King and Prof. Jacques Savoy.

- Special thanks to my parents, my brothers and sisters, who have been extremely understanding and supportive of my studies, and for their patience and encouragement while have been separated for many years. I feel very lucky to have a family that shares my enthusiasm for academic pursuits.

- I am also grateful to fellow colleagues in the Department of Computer Science for their assistance and companionships. Special acknowledgment is also given to Asmaa El-Hannani, Soraya and Ghita Kouadri Mostéfaoui, for the invaluable days and moments that we spent together in Fribourg.

- Last and but not least, I am greatly indebted to all my friends, especially to Karima Dilmi, Olfa Ben-Setira and Danielle Jmaa, who have been very supportive during the PhD period.

Abstract

This thesis proposes a multimodal alignment framework that bridges the gap between static documents and spoken language. This alignment aims mainly at linking static documents with temporal data, in order to exploit the multi-level structure of documents for indexing multimedia recordings of events. This novel multimodal alignment method, largely described in this thesis, is applied on two particular case studies, meetings and lectures.

Aligning static documents with the speech transcript of meeting dialogs consists in establishing relationships between them, according to their textual content, at various levels of granularity. The main relationships studied in this thesis are based on shared thematic content, quotations and references made by speakers to the static documents used during the meeting.

The thesis first introduces the novel concept of multimodal document alignment by putting it in perspective with related research works. Then, state-of-the-art methods for mining information from textual documents are presented and compared. Further, in addition to the detailed presentation of the alignment methods to detect thematic relationships, quotations and references between static documents and spoken language, several strategies for combining the three alignment types are presented and evaluated. These strategies aim at re-enforcing the cooperation of these distinct methods, and at resolving the inconsistency between the various levels of granularity of the documents being aligned.

Our multimodal alignment framework has been applied and evaluated on two corpuses, with very satisfactory results. The first corpus consists in a French press review meetings, recorded in our smart meeting room in Fribourg. The second one contains English scientific presentations, recorded at CERN. In this latter use case, the speech transcript of the speaker, slideshows and the scientific article presented are all aligned together.

In addition to this novel multimodal document alignment method, a complementary research axis has been investigated in this thesis: the bimodal thematic structuring of meetings. Based on a spatial and temporal clustering of the thematic alignment results, the proposed bimodal segmentation method generates simultaneously the thematic segmentation of the discussed static documents and the meeting dialogs. An evaluation has shown that the new bimodal method outperforms classical monomodal approaches such as TextTiling.

The satisfactory results obtained within this thesis prove the performance of our proposed multimodal document alignment solution, and that it supports meetings structuring, searching and browsing. These results highlight also the document usability for accessing multimedia data and its role in multimodal applications.

Résumé

Cette thèse propose une méthode d'alignement multimodal, qui permet d'établir le lien entre les documents statiques et le langage parlé, conformément à leur contenu textuel. Le but étant d'exploiter la structure multi-niveaux des documents statiques pour l'indexation des enregistrements multimédia. Cette nouvelle méthode d'alignement multimodal, largement détaillée dans cette thèse, est appliquée sur deux cas d'études particuliers, les réunions et les conférences.

La thèse présente d'abord le nouveau concept d'alignement multimodal de document, en le situant par rapport à des travaux de recherches dans le domaine. Elle présente par la suite l'état de l'art des méthodes d'extraction d'information des documents textuels, et leurs comparaisons. En plus de la présentation détaillée des méthodes d'alignement pour détecter les liens thématiques, les citations et les références entre les documents statiques et la transcription de la parole, plusieurs stratégies permettant de les combiner sont présentées et évaluées. Ces stratégies mettent en évidence la complémentarité de ces différentes méthodes d'alignement. Elles traitent l'incohérence qui peut surgir de par la structure multi-niveaux des documents alignés.

Notre méthode d'alignement multimodal a été évaluée sur deux corpus, avec des résultats très satisfaisants. Le premier corpus correspond à des réunions de revue de presse en français, enregistrées dans notre salle de réunion à Fribourg. Le deuxième contient des présentations scientifiques en anglais, enregistrées au CERN. Dans ce dernier cas d'études, la transcription de la parole, les transparents de la présentation et l'article scientifique présenté sont alignés les uns par rapport aux autres.

En plus de cette nouvelle méthode d'alignement multimodal de document, un axe de recherches complémentaire a été étudié dans cette thèse : la structuration thématique bimodale des réunions. Basée sur un groupement spatial et temporel des résultats d'alignement thématique, la méthode de segmentation bimodale proposée produit simultanément la segmentation thématique des documents statiques discutés et des dialogues de réunion. Une évaluation de cette nouvelle méthode a montré qu'elle surpasse des approches classiques monomodales, telles que la méthode "TextTiling".

Les résultats satisfaisants obtenus dans cette thèse prouvent que notre approche d'alignement multimodal de document est performante. Elle permet de structurer des réunions, et facilite la recherche et la navigation. Ces résultats mettent également en évidence la pertinence des documents pour l'accès à des données multimédia, ainsi que leur rôle dans des applications multimodales.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Documents are crucial vectors for communicating information in our daily life, such as in meetings and lectures, where they are present in various forms, for instance, agendas, budgets, meetings minutes, bloc notes and course supports. Moreover, documents constitute a good artefact for storing information.

In this thesis, a **document** is defined as being a set of information that is designed and presented as an individual entity, independently from the media through which it is being generated. This information corresponds to any written, printed, recorded, magnetic or graphic matter. Two types of documents have been defined: static documents and multimedia documents. A **static** document is a physical or electronic entity that contains any printed work, such as book or manuscript. It contains text, graphics, images or a mixture of them. Newspapers, scientific documents and magazines are examples of documents that contain mainly text, and have specific layout structures and complex logical structures. On the other hand, a **multimedia** document combines different types of multimedia data in time and space, like animations, audio and video streams. The multimedia data could be temporally organized by the user, which adds to the document a temporal structure.

Multimodal applications, such as conferences and meetings recording and analysis, constitute a new domain of interest for research in computer science (e.g. the IM2 Project [IM2]). Multimodal applications are particularly interesting because they deal with various modalities. A **modality** is defined as being a channel of communication through which the information is transported, such as vision, voice, gestures, facial expressions and body movement.

The major challenge of multimodal applications is the multimodal browsing and indexing. However, it is difficult to extract high level semantic indexes from multimedia documents, i.e. the semantic meaning within events, such as entities, places and concepts. From another side, static documents have been widely exploited for information retrieval tasks, such as searching and retrieval, thanks to their highly thematic and logical structures, as well as to their content that is easy to index.

Our hypothesis is that static documents can provide natural and thematic means to index, access and browse multimedia documents. For these reasons, links between static documents and multimedia documents should be built, in order to bridge the gap between them. The linking process could be made by enriching the static documents with temporal indexes. A manner to get the latter consists in retrieving them from the multimedia documents, for instance from the audio recording, slideshows or video recording. Nevertheless, it is crucial to define the criteria and basis of this retrieval of temporal indexes.

1.2 Goal

During a multimodal event (e.g. a lecture), static documents are either being discussed, quoted, referenced or enriched with notes. At the same time, the multimedia documents are being recorded. The temporal parallelism between static document utilization and multimedia document generation can help deducing the required temporal indexes for static documents.

The main goal of this thesis is to bridge the gap between static documents and multimedia documents within events (meetings and conferences), through a multimodal alignment between them [Mekhaldi et al. 2003] [Lalanne et al. 2004b], which fully benefits from the temporal parallelism between the two documents types. An **alignment** is a process that establishes relationships between two documents. An alignment is considered **multimodal** if the documents to align are acquired from different modalities. In our thesis, the speech transcript, as being a reliable trace of multimedia events, and the textual content of slideshows, which are projected and referenced by speakers during meetings and conferences, have been chosen for the alignment with the static documents that are discussed or presented.

Depending on the availability and reliability of the multimedia data within the multimodal event from one side, and on the type of content of static documents from another side, other alignment types are possible in order to reach our goal.

Slideshows and captured videos during conferences are useful if the presented static documents are rich with graphics and images. In such case, the matching will be based on the image similarity. In some multimodal events such as lectures, other multimedia data are available, for instance handwritten annotations on electronic whiteboard or on paper, which provides another way of alignment based on the matching of textual content.

1.3 Methods

Aligning two data resources requires the consideration of specific aspects that might be shared between them. One of the aspects is the shared themes between the speech transcript and the static documents discussed during the event. Therefore, according to the shared themes, a thematic relationship might be created between the two data resources. Detecting the thematic relationship might be achieved by a process called **thematic alignment** between the static documents and the speech transcript. The same process might be performed between static documents and slideshows. Other useful aspects might appear, when a particular speaker makes a quotation of a given part from the static documents, or makes a reference to a given document or document element (e.g. a title or a figure) [Popescu-Belis 2004b]. These two additional aspects, **quotations** and **references**, are also used to establish other particular relationships between the static documents and speech transcript. Based on the three relationships, thematic, quotations and references, the respective discussed parts of the static documents will be enhanced with the temporal indexes of the corresponding speech transcript parts.

1.4 Contribution of this Thesis

This thesis makes several contributions. It shows the collaboration as well as the interaction that might be made between the various modalities of a multimodal event, containing static data (static documents), and multimedia (speech recording, slideshows, video, etc.), which helps to build a multimedia archive. Another contribution consists in highlighting the role of static documents for structuring meetings, and for accessing and indexing multimedia archives.

1.5 Structure of this Thesis

In addition to the current chapter, this thesis is composed of seven other chapters:

- Chapter 2 presents an overview of the meeting/classroom projects that exist nowadays. In addition to the presentation of two classifications made for these projects, a novel classification is described, in which the concept of multimodal alignment is introduced.

- In chapter 3, many methods that deal with data mining are described, such as static document and speech segmentation methods, and information retrieval techniques (stemming, similarity metrics, TF.IDF metric, etc.).

- Our alignment framework is presented in chapter 4, where the pre-processing of data, and the three alignment types between static documents and the speech transcript (thematic, quotations and references) are presented.

- In chapter 5, the application of our alignment framework to press review meetings is presented and evaluated, with very satisfactory results. The thematic alignment has been tested in respect to two strategies, 1-best alignment and multiple alignments, at various levels of granularity (turns/logical blocks, utterances/ logical blocks, sentences/ turns, etc.). The quotation alignment links have been detected using a lexicographic based matching algorithm. Finally, the results of the two alignment types (thematic and quotations) have been merged with the reference alignment links. An evaluation of the role of the static documents within multimedia archives has been described in section 5.4 of chapter 5, through a user evaluation [Lalanne et al. 2004a].

- Later in chapter 6, another corpus, containing scientific conference presentations, has been used for experimenting with our alignment framework on multiple modalities, where slideshows, considered in this thesis as a second kind of multimedia documents, have been integrated into the thematic alignment process. In addition, the three multimodal alignment pairs (documents/speech, documents/slideshows and speech/slideshows) have been grouped, in order to validate the alignment links. The grouping process has drastically increased the thematic alignment scores, which proves that integrating new data resources within the thematic alignment process should facilitate the static documents synchronization.

- In chapter 7, the structure of static documents has been exploited in order to structure multimedia events such as meetings. Based on the thematic alignment with the speech transcript, a bimodal thematic segmentation of meetings has been defined. The new segmentation method consists in a clustering process of the thematic alignment results, which have been represented by a 2D graph. Afterward, the thematic structure of each modality, static documents and speech transcript respectively, is extracted from the generated clusters. The new bimodal segmentation method has shown good performance, when compared to other classic monomodal segmentation methods (TextTiling and baseline methods).

- Finally, chapter 8 concludes this thesis, presents the limitations of our solution and opens new perspectives.

Chapter 2

Related Research in Multimodal Document Alignment

2.1 Introduction

With the constant progression of technologies involved in multimodal applications, such as lectures, conferences and meetings, there have been various studies that deal with the data of these multimodal applications. The tasks most often considered are data recording, analyzing, indexing, summarizing and browsing.

This chapter targets the main research projects that deal with multimodal applications. It highlights the utilization of the alignment aspect within these projects, and shows the lack in the alignment studies that we want to accomplish in this thesis. The next section presents an overview of the main projects in this field, as well as some classifications of them according to different criteria. Section 2.3 focuses on the document alignment aspect within these projects. Section 2.4 presents our new definition of the alignment, the multimodal document alignment, which includes our proper classification of the previous projects.

2.2 Meeting/Classroom Projects: State of-the-Art

A number of research projects have been described, which study matters related to classroom lectures. These studies include the eClass project [Brotherton et al. 1998], iClass project [Macedo et al. 2004], DSTC project [Little et al. 2002] and Cornell project [Mukhopadhyay and Smith 1999]. Other research projects are interested

in meeting rooms like the FXPal project [Chiu et al. 2000a], Microsoft Research project [Cutler et al. 2002], ISL project [Schultz et al. 2002], etc. Conferences have been also the focus of some projects, like Indico project [Le Meur and Bourillot 2005]. Among all these studies, various modalities have been considered, for instance video and audio recordings, gestures, slides, pen stroke annotations on whiteboard, digital ink annotations on slides, etc. In the next paragraphs, some of these projects and their related tasks are described.

The distributed Meetings project (DM) at Microsoft Research [Cutler et al. 2002] has focused on three main tasks, meeting broadcasting, recording and browsing. In addition to the overview and whiteboard cameras, a RingCam camera is used in order to capture all participants simultaneously, in a panoramic way. A microphone array for beamforming and sound source localization are integrated in this RingCam. Using the data generated by the various cameras and the microphone array, several processing are performed by the meeting room server, such as sound source localization, speaker segmentation and clustering for archived meetings, person detection and tracking, and whiteboard processing for both live broadcasting and archived meetings. Several functionalities are presented by the DM remote client to the user, in parallel to a timeline, for both live and asynchronous viewing of meetings. The most important functionalities of this project are the panoramic view by the RingCam, the whiteboard image augmented by pen stroke time stamps, the speaker segmentation and filtering allowing an accurate access to a given speaker, and time compression to remove pauses and increase playback speed.

Many aspects related to meetings have been studied in the Interactive Systems Lab (ISL) at Carnegie Mellon University [Schultz et al. 2002] [Bett et al. 2000]. In this project, the dialog has been analyzed using Model Combined Based Acoustic Mapping (MAM), in order to use non-key words based features for accessing meetings. The main considered aspect in this task has been speakers' identities and dominance identification. Participants are identified using audio and video based recognizers. This is made by sound source localization, people segmentation, color appearance and face identification. Moreover, based on a ranked list of sentences, a meeting summary is generated using the Maximal Marginal Relevance algorithm (MMR). Finally, meeting capturing and other functionalities, required for manipulating, accessing and reviewing meeting data, are integrated in the meeting browser.

In the Jabber-2 system [Kazman and Kominek 1997] developed at the University of Waterloo, many problems related to multimedia meetings have addressed, such as

topics detection, temporal idioms identification, meeting monitoring, etc. Moreover, a user interface for meeting information retrieval is provided. The detection of meeting topics requires first the recognition of the recorded meeting dialogs, then the use of the Lexical Chain technique to recover semantic clusters of words. Furthermore, in order to detect semantic relationships between words, the WordNet thesaurus has been used. The identification of temporal idioms is based on the identification of participants and the analysis of the interaction between them, regarding periods of speech and pauses in an audio sample. Using identified topics and temporal idioms, and through a meeting monitoring, the user might know who was involved in what kind of idioms, as well as people who were not involved, according to a timeline. Meeting information retrieval tasks such as searching, browsing, gazing, etc. are also available in the user interface.

The eClass project [Brotherton et al. 1998], developed at Georgia Tech, targets the preservation of the record of the entire class activity. This project tends also to reduce the content generation effort. In order to generate a lecture record enriched by audio and video, an electronic whiteboard is employed. In this project, the various streams have been classified into three main sets according to their use, respectively simple, derived and control streams. Simple streams represent any stream that is played as it is after having been captured, such as audio and video. Derived streams are streams that are generated from other streams, such as phonemes that are derived from audio, or gestures that are derived from the video. Control streams are meta-streams that contain information about simple streams, for instance annotations related to audio or video recordings. Once the capturing process is accomplished, the generated streams (audio and video) are matched. The matching process is made at three levels, slide, pen stroke and word levels. At the slide matching level, the audio associated to a given slide might be accessed. At the pen stroke level, the audio corresponding to a given pen stroke might be accessed. Finally, at the word level, the audio corresponding to a particular written word might be retrieved.

The iClass project [Macedo et al. 2004] is a recording system for data generated during live presentations, as well as a system for linking related presentation sessions. The linking process consists first in composing a query, using the current session data such as slides content, text of visited Web pages, pen strokes and handwritten annotations, etc. Afterwards, the query is processed in order to identify the related presentations from previous sessions, using the Latent Semantic Analysis algorithm (LSA). Finally, the live session information is presented to the user as an XML

document including SMIL (Synchronized Multimedia Integration Language).

The Cornell project [Mukhopadhyay and Smith 1999] aims at automatically constructing structured multimedia presentations from live lectures and seminars. Thus, the capture made by an overview and a tracking camera has been considered as a passive capturing, i.e. without speaker assistance to aid the capture process. When synchronizing the various resources on a timeline, three synchronization modes have been distinguished, according to the nature of resources. The first mode, called timed-timed synchronization, concerns the matching of temporal media, such as combining the video streams captured by independent capture devices. This matching has been made by detecting one or more synchronization points within the streams, where the correspondence can be computed. The second mode, the timed-untimed synchronization, corresponds to the matching of slides with the video streams. This synchronization is based on the detection of slide changes within the video stream, then a matching process between the extracted slides and the corresponding video frames. Finally, the untimed-untimed synchronization is used for non-temporal data, for instance matching slides titles on the browser with the corresponding slides. The generation of structured multimedia presentations, where pairwise of data are synchronized, is finally made on three steps. First, the two captured video streams are synchronized, and then slides are synchronized with one of the video stream. Finally, the two video streams are combined, and a structured presentation integrating all resources is generated in the form of an HTML file.

At the FX Palo Alto Laboratory [Chiu et al. 2000a], a conference room has been built for multimedia capturing of various presentation styles, augmented by a note taking application called NoteLook. This conference room is equipped by a flush-mounted large screen, three computer controllable cameras and a video conference camera. The NoteLook application is running on a wireless pen based notebook computer in the room. In this application many facilities are presented to the user, such as annotating video images captured by the various cameras, incorporating these images into note pages, automatic note taking with digital ink, dragging of slides images, notes printing, etc. From another side, specific methods have been developed for identifying pertinent video segments that summarize the entire video [Girgensohn et al. 2001]. This is made by computing the importance score of video segments that is based on their duration and rarity. Furthermore, the relevant and useful parts of the video might be accessed through the manga interface, using video key frames, i.e. the images that are automatically extracted from video footage.

Finally, the generated images, either from video recordings or projected slides, are matched with the scanned paper handouts of slides, in order to generate hypermedia meeting records.

Anderson et al. have studied the relationship between digital ink on slides and speakers' audio recording [Anderson et al. 2004]. According to their method, the handwriting on slides is matched with the speech transcript of the speaker presentation, by combining the speech with the written phrases. Moreover, terms are recognized and matched on one modality, ink or audio, and the recognized words are disambiguated by triangulating across both modalities.

Tang and Kender [Tang and Kender 2005] have proposed a prototype for matching handwritten text in educational video recording with the corresponding textbook chapters. First, text frames in the video are extracted, and then the corresponding words are recognized. From another side, a small vocabulary of the table of content of course material is constructed. Finally the recognized handwritten words are used in order to query the table of content, in the form of a latent semantic analysis matrix.

Finally, Little et al. [Little et al. 2002] have been interested in exploiting meta-data, related to atomic components in a presentation recording and their spatio-temporal relationships, in order to generate meta-data for composite mixed-media digital objects. Lectures video footage and PowerPoint presentations have been used, in order to generate SMIL presentations. The digitized video footage and logging data have been analyzed in order to identify the slide changes. These temporal information have been then exploited in order to synchronize slides and digital video on a timeline, offering the possibility to users for browse and access accurately to a given slide and the related audio/video streams of the speaker.

After this overview on the existing multimodal projects, their classification is presented in the next section.

2.2.1 Meeting/Classroom Projects Classification

Meeting and Classroom projects have been classified in [Lalanne et al. 2003] according to the data and annotations used for browsing and retrieval tasks, in their corresponding user interface. Two main categories have been distinguished in this classification: document related annotations and speech related annotations. The eClass [Brotherton et al. 1998] and the Cornell [Mukhopadhyay and Smith 1999] projects

presented above are two examples of the first category. In the eClass project, captured slides have been considered as indexes for the meeting browser, in which the user can select a particular slide to access to the corresponding audio. In the Cornell meeting browser, the slides are employed in order to generate structured multimedia presentations. Thus, the user can jump to a particular slide, which gives an access to the corresponding video footage. On the other hand, the ISL project [Cutler et al. 2002] and Jabber-2 system [Kazman and Kominek 1997] browsers, presented above, belong to the speech related annotations category. In the ISL project, the dialog has been analyzed in order to create indexes to access to the spoken communication. Speaker identities and dominance, which are based on sound source localization and other audio and video features, have been used. The Jabber-2 system offers many facilities to the user, such as meeting monitoring, based on topic detection and temporal idioms identification. A speech recognition system and a thematic segmentation method have been used in order to detect various topics. The interaction between participants has been exploited in order to identify temporal idioms.

In the study made by Tucker and Whittaker [Tucker and Whittaker 2004], the meeting room and classroom projects have been classified according to the focus of the browser, i.e. the basis of navigation and presentation in the user interface. Thus, audio, video, artefact and discourse based interfaces have been distinguished. Kimber's meeting browser [Kimber et al. 1995] is an audio based interface that uses a visual index for the navigation, obtained by speaker segmentation. This way, the involvement of speakers in the meeting can be viewed. In the video based interface category, Foote's meeting browser [Foote et al. 1998] has used index points, computed from video and audio properties, to jump to meeting sections. The artefact based interface, which corresponds to other meeting data rather than audio and video recordings, such as slides and handwritten text, includes many projects. The most important one is the eClass project [Brotherton et al. 1998] that offers an interface of navigation within lectures recordings, allowing the user to play back the audio corresponding to a particular slide, pen stroke or written word on the whiteboard. FriDoc [Lalanne et al. 2004a], Fribourg Document centric meeting navigator, is also cited in their classification, as being an artefact based meeting browser, thanks to the document alignment techniques, presented in this thesis. Finally, the discourse based interface concerns derived elements and annotations of meetings. As an example in this category, the ISL project [Schultz et al. 2002] [Bett et al. 2000] where the

speech features, such as speaker identities and dominance, are exploited in order to enrich the browser with new functionalities. Likewise in Jabber-2 system, the detected meeting topics and temporal idioms, which are based on dialogs analysis, are used to monitor meetings [Kazman and Kominek 1997].

2.2.2 Place of Static Documents in these Projects

Many projects have been described and analyzed in the previous sections, as well as their classification according to their user interface, or to the data and the annotations used. Most of them tried to cover user needs and requirements, by exploiting the various data generated either during the meeting (audio, video, etc.) or after the meeting (handwritten annotations, speech transcript, etc.). Unfortunately, static documents, that might be present and discussed during meetings, have been rarely used, despite the richness of their information. This phenomenon might be explained by the fact that the static aspect of these documents gives the feeling that they could not serve for tasks, such as browsing, reviewing, etc., that have been historically related to multimedia data.

However in our view, static documents are required and useful for many tasks. Static documents that are discussed during meetings represent a set of vocabulary that might help for recognizing the speech. Static documents are structured, and thus they might help for structuring meetings, and also facilitate meeting topics detection. Static documents also constitute a solid basis for meeting browsers, where the user can access to a particular meeting part, video or audio that correspond to a particular part of the discussed document. Finally, the textual content of static documents can be useful for information retrieval, searching, etc.

Reaching the goals described previously requires the integration of static documents with the other temporal meeting data, since the time is the core of meetings. This integration process should synchronize multimedia documents and static documents, even though the latter are not time stamped. In this context, the current thesis work proposes a solution that consists in matching the textual content of static documents with dialogs transcript, since the latter is the textual trace of the multimodal applications, and the most related resource to discussed static documents. Thus, their respective related parts are detected and matched. This process is called document alignment with speech transcript.

Historically, aligning documents or data has been considered in many other research fields. Some of these fields are described in the next section.

2.3 Document Alignment and Related Researches

Alignment is a process that takes two data resources as input, and establishes relationships between them, at many levels. Even though it has been given various names, the alignment has been the subject in many research works, such as multilingual studies [Yu 2004] [Kay and Roscheisen 1993] [Church 1993] [Ghorbel et al. 2002], information retrieval [Chang and Lui 2001] [Petinot et al. 2004] [Jianying et al. 1999] [Denoyer et al. 2003] [Bagga and Biermann 2000] [Mihalcea and Hassan 2005], etc., and in bio-genetic studies [Higgins and Taylor 2000] [Smith and Waterman 1981] [Needleman and Wunsch 1970].

2.3.1 Alignment in Multilingual Studies

Yu [Yu 2004] has proposed a method for aligning texts and their counterparts in bilingual web sites, for stories written in English and Chinese. This method is based first on automatic translation of stories in Chinese (respectively in English). Then, using Blue's method, each English story (respectively Chinese) is compared with all the stories in the other language. Pairs with the highest scores are considered as final results of the alignment.

Kay and Roscheisen [Kay and Roscheisen 1993] have proposed a method for aligning a text with its translation at the word level, based on the similarity of words distribution. The assumption in this method is that aligning words could induce to aligning sentences.

In order to avoid noise in the detection of paragraphs boundaries, Church has been interested in aligning texts at the character level [Church 1993]. His method is based on the Cognate approach, i.e. the distribution of shared characters in the words.

Simard [Simard 1999] has extended a method for bilingual alignment to align three languages. Thus, a classic alignment method, based on dynamic programming, has been improved by finding the optimal path in an N-dimensional matrix.

Finally, Ghorbel et al. [Ghorbel et al. 2002] have studied the alignment of medieval manuscripts. The approach consists in enhancing classic methods used in multilingual alignment with linguistic properties such as similarity at the lexical, syntactic, semantic and morphological levels, as well as structural properties of texts, such as rhetorical structure.

The multilingual alignment is performed between parallel texts, similarly to our

document/speech transcript alignment. However, the advantage here is that one of the two texts is a translation of the other, which generally generates a list of parallel links.

2.3.2 Alignment in Information Retrieval

Document matching has been also the subject of many works in information retrieval field, including cross documents co-reference field [Bagga and Biermann 2000] [Chang and Lui 2001] [Kuper et al. 2003] [Petinot et al. 2004], documents classification field [Kloptchenko et al. 2003] [Jianying et al. 1999] [Le and Thoma 2000] and [Mihalcea and Hassan 2005], citations field [Zhang et al. 2004], etc. These works are detailed in the following sections.

2.3.2.1 Cross Documents Co-Reference

In the information retrieval area, the detection of relationships between documents segments that refer to the same person, event, place, etc., is called cross documents co-reference process.

Bagga and Biermann have studied the resolution of cross documents co-reference for entities and events [Bagga and Biermann 2000]. Thus, the document is parsed and all the sentences related to a given entity are extracted. Then for each pair of documents, a Vector Space Model based matching is used to compute similarities between extracted sentences.

Kuper et al. [Kuper et al. 2003] have been interested in indexing sport video recordings, by enriching them with information extracted from other resources and modalities (video, text, speech, etc). The extracted information is matched in order to complete the missing aspects of events among the various resources. Therefore, all pairs of resources are compared, by considering various aspects such as shared events and players involved in the scenes. Then the best matching for each scene is chosen, while ignoring the cross links. In order to connect various documents, all pairs of scenes are organized in sets, where each set groups all inter-connected scenes.

The extraction of information from similar web pages has been the focus of Chang and Lui study [Chang and Lui 2001]. The proposed system aims to automatically identifying record boundaries of web pages. A sequence matching algorithm has been used for detecting similar patterns among various web pages.

Informedia digital video library is a project that aims to create video digital libraries [Hauptmann 2005][Olligschlaeger 1999]. One of the main focuses of Informedia project was information extraction from broadcast TV news and documentary content. Several modalities have been exploited (speech, image and natural language) in order to provide several functionalities of video understanding (searching, retrieval, visualization, and summarization), either from one or several video recordings. One of the useful aspects for video understanding is entity identification (e.g. places, persons and organizations). Entity identification is performed in many steps. First, each word from the speech transcript or from the text extracted from the video is synchronized with the video, according to the time when it was mentioned. Then using the Viterbi algorithm, sentences are parsed in order to match each word with one of the entities. Additional tools have been used for the matching process between text and entities, such as the address coverage table used for the matching of places with their addresses.

WIP system (Knowledge-based presentation on Information) is an intelligent system for the production of multimedia presentations containing text, graphics, dialogue and annotation sequences [Wahlster 1993]. WIP has been mainly used for generating instruction, for the maintenance and repair of technical devices. WIP is considered as a goal-driven system, in which two main categories of users are distinguished, novice and expert users. Based on a common representation, a coordinated multimedia presentation is generated, according to the user parameters (language, document type, target group, etc.). The coordinated presentation is the result of the alignment between the various media, according to the referential relationship between them (e.g. the expression "the upper left part of the picture", in the case of a text/picture alignment).

2.3.2.2 Document Classification

The classification and archiving of documents constitutes another approach of documents alignment, in which documents are put under corresponding classes, according to defined criteria related to their content, structures, etc.

Routing submitted scientific papers to reviewers and area committees, where documents are matched using text clustering methods [Kloptchenko et al. 2003] [Mihalcea and Hassan 2005], is an example of documents linking according to their content. Afterwards, documents are put into the pertinent categories and conference tracks.

Document structures are also used in order to link documents, by grouping them according to their types (letters, magazines, scientific documents, etc.). Since each document type has its appropriate layout, the documents are compared at the layout level, and then put into distinct categories [Jianying et al. 1999] [Le and Thoma 2000]. Furthermore, both content and structures might be used to classify documents [Denoyer et al. 2003].

2.3.2.3 Citations

Documents might be also linked according to their fields, such as their titles, author's names, institutions and citations, especially in case of scientific documents.

Citeseer [Petinot et al. 2004] is a system in which the citations field is used in order to match scientific documents, by retrieving all documents available in the database that cites a query document.

In other methods, documents that are cited by the same documents (co-citation method), or those that share at least one citation (bibliographic method), are matched [Zhang et al. 2004]. All these methods are based on the comparison of documents at the titles, author's names, bibliography, etc. levels, in order to get the best correspondence.

2.3.2.4 Document Lexicographic Matching

Matching many texts, which might be obtained by various devices, has been studied in many projects. Kornfield et al. [Kornfield et al. 2004] were interested in matching handwritten text with corresponding ASCII transcript. First, the handwritten pages were segmented, then features were extracted from both handwritten and ASCII texts. An algorithm based on Dynamic Word Warping has been used to match handwritten page images with transcript words. The ASCII words and handwritten segmentation boxes are employed as samples for the time series. Later, this method has been applied at the line level, then at the whole page level.

The alignment in information retrieval field is focusing on specific entities, layout elements or lexicographic matching of texts within the documents. However, in our document/speech alignment, the overall content of both resources is exploited.

2.3.3 Bio-genetic

Matching specific kinds of data has been also performed in the bio-genetic field, where many researches are focusing on aligning similar genome sequences, in order to find biologically meaning relationships between organisms.

The method of Smith and Waterman [Smith and Waterman 1981] aims to find the best alignment between sequences. In this method, columns that are composed between respective positions of both sequences are weighted. Furthermore, columns with the same nucleotides are rewarded, otherwise they are penalized.

Needleman and Wunsch's algorithm [Needleman and Wunsch 1970] is based on the transformed matrix, representing the comparison of the two protein sequences. The algorithm aims to find the best alignment between sequences and the best score for the alignment. This is made by identifying the pathway having the highest score within the comparison matrix.

Finally, Higgins and Taylor [Higgins and Taylor 2000] have focused on the multiple alignments, i.e. aligning more than two protein sequences. In this method, the alignment distances between all pairs of sequences is used in order to build a guide tree. Afterward, the sequences are aligned, according to the guide tree branches.

The bio-genetic alignment method is based on the similarity of sequences. Motivated by this technique, our document/speech transcript thematic alignment process exploits the thematic similarity between respective parts of the static documents and the speech transcript, in order to align them.

After this overview on works focusing on documents and data alignment approach in various research areas, our multimodal document alignment is presented in the next section, where the alignment definition is extended to cover other data modalities.

2.4 Document Alignment vs. Multimodal Document Alignment

Multimodal meeting room and classroom applications are rich with various data types, such as video and audio recordings, slides, handwriting annotations, visited web pages, etc. These data might be exploited and represented to the user in the form of a user interface. In order to be able to jump and to play back a section of the

application, pairs of data should be synchronized or aligned. Thus, the traditional definition of documents alignment should be extended to cover other modalities, besides static documents.

2.4.1 Definition of Multimodal Document Alignment

In the context of multimodal events (meetings, scientific conferences, etc.), the multimodal alignment is a matching process between documents resulting from various modalities, static (discussed documents, handwritten documents, etc.) and multimedia (audio, video, slideshows, etc.). This matching consists in detecting and then establishing relationships between the respective parts of the documents being aligned. Thanks to this multimodal alignment, the temporal indexes of multimedia documents might be associated to the static documents.

2.4.2 Multimodal Alignment Types

According to the types of data involved in the alignment process (multimedia data, text, images, etc.), various types of alignment can be defined. The most important types of alignment are *temporal based*, *image based* and *textual content based alignment*. In the *temporal based alignment*, the shared temporal aspect of multimedia documents is exploited in order to synchronize them, such as synchronizing the video and audio recordings. In the *image based alignment*, methods that exploit similar images are used in order to match documents, such as the slide changes detection and slide identification methods, used for aligning slides with video recording. In the *textual content based alignment*, documents are compared according to their content. This latter type has four main varieties, *lexicographic*, *thematic*, *reference* and *quotation alignments*. Documents are matched in the *lexicographic alignment* according to a pure lexical similarity of their texts. The *thematic alignment* is based on detecting documents parts that share the same theme. In the *references alignment*, documents parts that refer to particular parts of other documents are highlighted. Finally, the *quotations alignment* is a special kind of *lexicographic alignment*, where the terms order in the sequences aligned is taken into account. According to these types of alignments, a new classification of the projects, described in section 2.2, is presented in the following section.

2.4.3 A New Classification of Meeting/Classroom Projects

The various alignment tasks, performed in the different projects seen in section 2.2, might be classified according to the types of data involved in these tasks. The eClass project [Brotherton et al. 1998] has exploited the temporal relationships between captured streams, in order to integrate them and facilitate the access of the user. Thus, the audio has been synchronized with slides, pen stroke and words. In another study, the eClass-Coweb linking project [Macedo et al. 2001], the eClass lectures material (slides content, whiteboard annotations, titles of web pages visited during lectures, etc.) has been matched with the material of the Coweb repository that contains web pages, created outside the classroom and related to the course (timetable, etc.). This alignment is based on a lexicographic matching between the textual content of eClass lectures and the titles of Coweb pages. In the iClass project [Macedo et al. 2004], live presentations have been linked with previous session presentations. For this reason, a query is composed of the material of the live presentation (slides text, handwriting, visited web pages text, etc.). Then, using a lexical matching process, the query is processed in order to detect correspondence between archived and live presentations. In the Cornell project [Mukhopadhyay and Smith 1999], two different alignment processes have been performed. On the other hand, the two captured video recordings are aligned according to their temporal indexes. However, the slides and one of the videos are aligned according to similar images, by detecting the slide changes in the video. One of the aims of the FXPAL project [Chiu et al. 2000b] is the generation of hypermedia meetings record. Thus, there has been a focus on linking meeting documents, especially the matching of similar paper handouts of presentation slides with video recording of the meeting or video projector. This matching process has used the image similarity in order to detect links. Moreover, the developed application for taking notes, NoteLook [Chiu et al. 2000a], generates web pages that link video recording with images and ink strokes, using temporal indexes. In the multimedia presentation archiving project in DSTC [Little et al. 2002], the temporal relationships between lectures video recordings and presentation slides have been exploited, in order to generate a SMIL browsing interface. In the Washington project [Anderson et al. 2004], digital ink on lecture slides has been linked with speech recording, based on a lexical alignment between recognized handwritten terms on slides and the speech transcript. Finally, Tang and Kender method, that aims to match handwritten text frames with textbook chapters [Tang and Kender 2005], has used a thematic matching process,

	Video	Audio	Slides	Temporal Annotation	Static Documents
Video	Cornell/ temporal	Informedia/ temporal	Cornell/image, DSTC/temporal, Tang/thematic, FXPAL/image, NoteLook/temporal	NoteLook/temporal	Tang/thematic Informedia/temporal
Audio			eClass/temporal	eClass/temporal, Washington/ lexicographic	
Slides			iClass/lexicographic		eClass-Coweb/ lexicographic, iClass/lexicographic
Temporal Annotation				iClass/temporal	eClass-Coweb/ lexicographic
Static Documents					eClass-Coweb/ lexicographic, iClass/lexicographic, WIP/references

Table 2.1: Multimodal document alignment classification

by extracting topic terms of the course documents (table of content, online syllabus, electronic slides), and then matching them with the recognized handwritten text in video frames. The temporal relationship between the video recording and words either from the speech transcript or extracted from the video, has been exploited in the Informedia project [Olligschlaeger 1999]. This temporal alignment between text and video has been exploited in order to extract event entities, useful for video understanding tasks such as events summarizing. The generation of multimedia presentations within the WIP system [Wahlster 1993] is based on the reference alignment between text and graphics.

Table 2.1 summarizes these various alignment types. In this table, the video data includes video recording, participant segmentation, etc. The audio data includes the audio recording and speech annotations, such as speech transcript, speaker segmentation, etc. Temporal annotations correspond to digital ink on slides, pen stroke annotation, etc. Finally, static documents correspond to printable documents, handwriting documents or web pages.

From a quick overview, it is clear that the majority of the existing projects

have used the temporal, image or lexicographic based alignment. However, with the exception of Tang and Kender work [Tang and Kender 2005], in which static documents are thematically aligned with video frames, and the reference alignment between text and graphics in the WIP system [Wahlster 1993], the thematic alignment, references and quotations have never been considered in multimodal applications. More specifically, the alignment of static documents with the speech transcript of the meeting dialogs, being the adequate pair for these three alignment varieties thematic alignment, references and quotations), has never been studied. Thus, the current thesis aims mainly to accomplish these remaining tasks, by studying all possible relationships that might exist between static documents, discussed and used during multimodal events, and the corresponding speech transcript.

2.4.4 Conclusion

This chapter has presented the state-of-the-art of the existing meeting/classroom projects. Moreover, two classifications of these projects have been described, based respectively on the data and annotations used for browsing/retrieval tasks, and the focus of the browsers. In addition, our classification method, which is based on multimodal document alignment aspect, has been presented. In the next chapter, the techniques and methods, that have been exploited in our document/speech transcript alignment framework, and their state-of-the-art are described.

Chapter 3

A Review on Multimodal Processing Techniques

3.1 Introduction

In the previous chapter, a literature survey has been conducted on the existing multimodal projects dealing with the multimodal document alignment issue. The current chapter describes the main methods that might be required for the processing of the data in our multimodal alignment, as well as their state-of-the art. Various fields are involved, document segmentation, speech segmentation, and information retrieval techniques. In the next section, state-of-the art methods for segmenting static documents are presented.

3.2 Static Document Segmentation

Many information retrieval tasks, such as searching, classification, etc. deal with collections of documents. In such cases, individual documents are manipulated as units for searching and retrieving information, within the entire collection. However, in some cases the same document is considered as a source or a target for multiple queries, such as in document alignment . Aligning two documents is based on the detection of the various links existing between their respective parts, which means that each of the two documents is manipulated as a collection of units. For this reason, both documents should be decomposed into significant segments, in order to facilitate the alignment process.

Document segmentation or decomposition is one of the most important research

axes, in the field of document image analysis. Documents segmentation has been utilized for many applications and systems, such as archiving documents collection, letter sorting, commercial document automation, document hyper-textualization, etc. Depending on the purpose and the objective of the application, the base of the segmentation might be different. For browsing tasks, the focus should be made on the geometrical segmentation of documents. For a content-based alignment, for instance, the functional segmentation, such as the logical segmentation, is more appropriate.

According to the relationships between the generated components after the segmentation, two main categories of segmentation might be defined, hierarchical components segmentation (e.g. physical and logical segmentations), and linear components segmentation (e.g. the syntactic and thematic segmentations). The main segmentation methods are studied in the following paragraphs.

3.2.1 Document Physical Segmentation

The physical structure of a document, called also structural or geometrical layout analysis [Matrakas and Bortolozzi 2000], consists in the description of the document as a set of physical components (e.g. pages, column, text blocks, paragraphs, text lines, words, etc.), as shown in Figure 3.1. Thus, the physical structure should contain at least one element from these components. Graphic objects, such as tables and figures, might also be considered as physical components. The consideration of all these types, or some of them, in the segmentation process depends highly on the objectives of the application. For instance, the focus might be made on column segmentation [Kopec and Chou 1994], as well as on page segmentation [Tokuyasu and Chou 2001] [Krishnamoorthy et al. 1993].

Discovering the physical structure of a document might be made by one of the three following approaches [Mao et al. 2003], top-down, bottom-up or hybrid approaches. In the top-down approach, which is an iterative approach, the entire document is analyzed and divided successively into smaller components, until the needed elements are obtained [Nagy et al. 1992] [Baird et al. 1990]. The bottom-up approach clusters pixels into connected components (e.g. characters), which are clustered into bigger components, e.g. words then lines then blocks [O'Gorman 1993] [Kise et al. 1998]. Finally, the hybrid approach combines the two approaches, bottom-up and top-down, such as the method of Pavlidis and Zhou

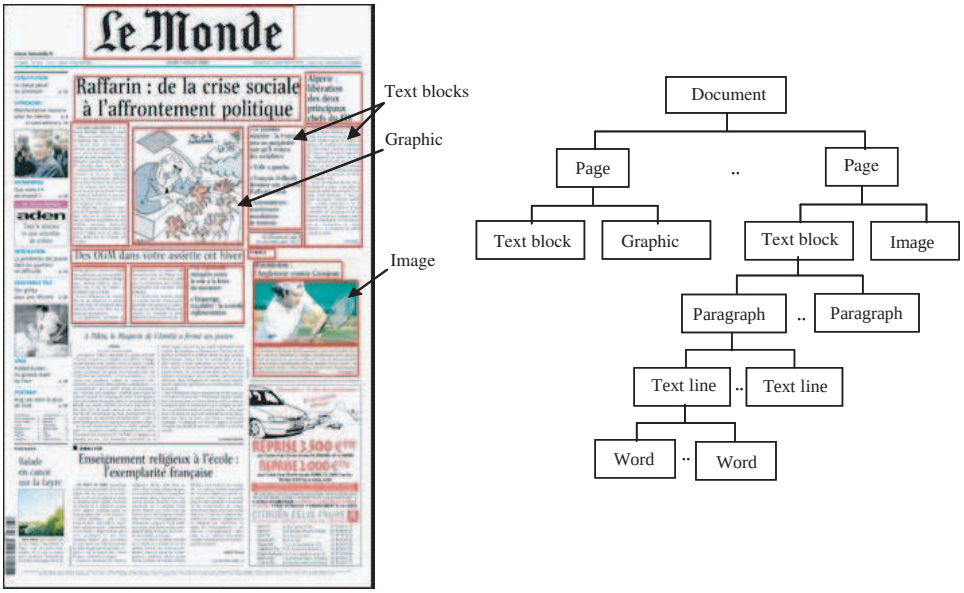


Figure 3.1: Physical structure of a document

[Pavlidis and Zhou 1992] that uses a split-and-merge strategy.

The generated physical layout might be represented by different ways. The main representations are: style parameter based representation, rule-based representation and formal grammar representation [Mao et al. 2003]. In the style-based representation, the objects size and gaps between them are used [Ishitani 1999] [Lin et al. 1997]. In the rule-based representation, the spatial relationships between components are described in the form of a tree [Tokuyasu and Chou 2001] [Kopec and Chou 1994]. Finally, the formal grammar based representation is a kind of rule-based representation, where the general rules satisfy a grammar, in contrast to the rule-based representation that might be arbitrary.

3.2.2 Document Logical Segmentation

The logical segmentation of a document is a labeling process of document components, using layout rules, which indicates blocks function and meaning at the page level (e.g. title, summary, content, etc., in Figure 3.2). Furthermore, in the logical segmentation the document is hierarchically organized, according to a model. For instance, a newspaper contains articles, and each article contains a title, author name,

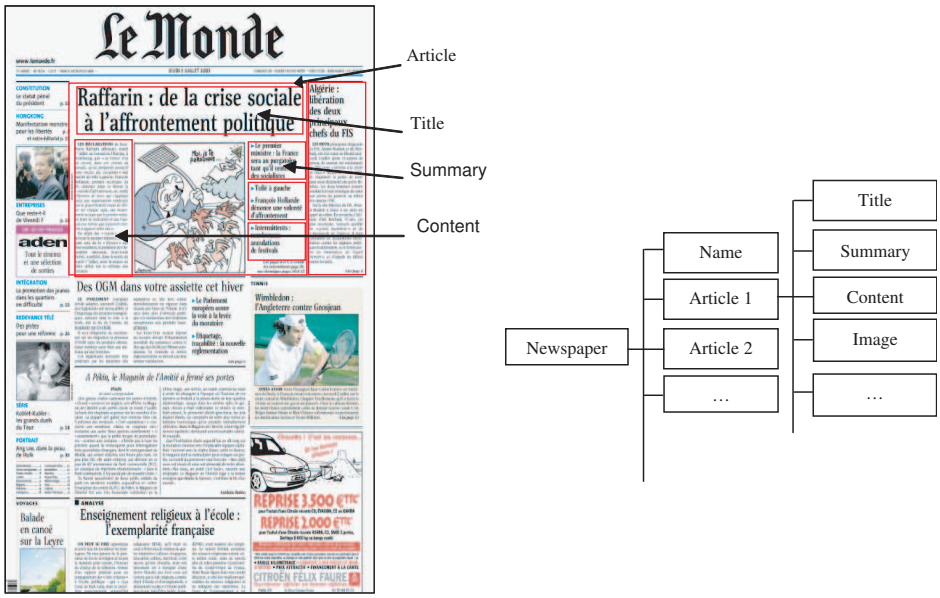


Figure 3.2: Logical structure of a document

pictures, content, etc. The logical segmentation might be derived from the physical segmentation [Fisher 1991] [Bloechle et al. 2006], as it might not [Lin et al. 1997] [Ishitani 1999]. The tree representation of the physical structure has been transformed into a logical tree in [Tsujimoto and Asada 1990], where logical labels are assigned to the various physical blocks, using a set of rules. The separations and frames within the document, considered as virtual physical blocks, are exploited with other rules to ensure the transformation process. In Fisher's method [Fisher 1991], the physical structure of documents is exploited in order to generate the logical structure. Using the three following rules, location cues, format cues and textual cues, the words are grouped into paragraphs and columns, and the paragraphs that are broken during formatting are reconstructed. Column boundaries and locations are identified according to the reader order of text blocks that is defined by the system. In the Dolores tool (Document Logical Restructuring) [Bloechle et al. 2006], the physical structure of documents is used in order to generate the logical structure, by labeling the various textual entities. Then these labels are projected in a document grammar based model, defined according to document types (e.g. newspaper). The logical structure might be generated using other techniques, independently from

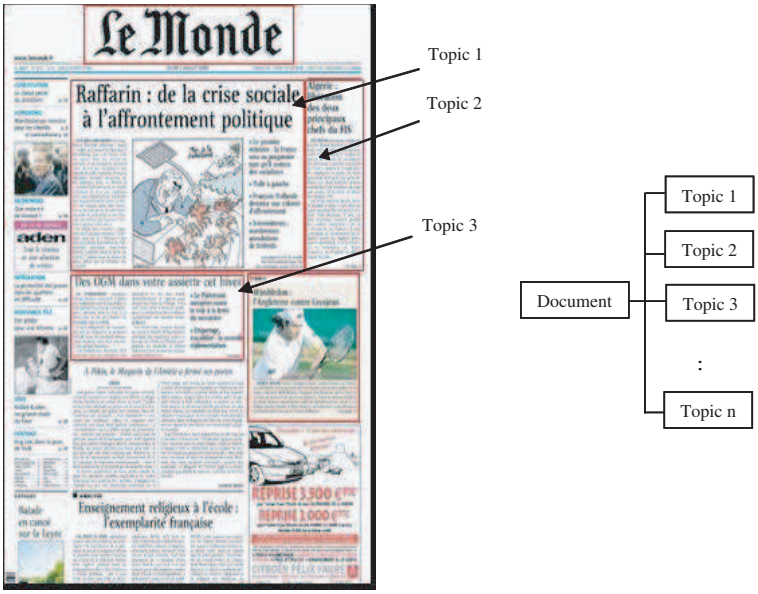


Figure 3.3: Thematic structure of a document

the physical structure. For instance, in the work of Lin et al. [Lin et al. 1997], the logical structure of book pages is extracted using content page information. The structure of various elements in a page (e.g. page number, header, footer, main text, etc.) are analyzed, and then matched with text lines, which are extracted from the page and performed by an Optical character Recognition system (OCR). Using interacting modules in [Ishitani 1999], such as typography analysis, objects recognition, objects grouping, etc., the document image is segmented, and text lines extracted are grouped and classified into logical components, based on heuristic rules. The objects that are incorrectly segmented might be corrected, according to the logical consistency among objects.

3.2.3 Document Thematic Segmentation

There is a strong relationship between the thematic segmentation and the thematic alignment studied in this thesis. The thematic or topic segmentation is a process that decomposes a text, according to the theme changes in case of multi-topic documents, like encyclopedias and newspapers (Figure 3.3). In case of mono-topic documents, like scientific articles, the document is segmented into sub-topic segments. Many

thematic segmentation methods exist, based on many strategies. The most known methods are the statistical word frequency method [Hearst 1994], the map representation method [Salton et al. 1996], the Local Context Analysis (LCA) based method [Ponte and Croft 1997], and the Lexical Cohesion Profile (LCP) [Kozima 1993].

The statistical word frequency method, called TextTiling [Hearst 1994], has been applied on magazine articles. TextTiling is based on the assumption that each topic area within a document is characterized by the existence of a specific vocabulary (except stop words), with a high frequency. Thus, any change of the frequency of a given word or a group of words might correspond to a change of theme. In order to detect the theme change, a high variation of group of words frequencies, among successive text segments, should be detected. For this purpose, the method decomposes the document into equal blocks, delimited by a window of a defined size. Then a similarity method (Cosine metric) is used to compute the thematic similarity between each two adjacent blocks. The similarity value between two blocks, which lies between 0 and 1, shows if they belong to the same theme or they do not. In Salton's method [Salton et al. 1996], the document (an encyclopedia) has been transformed into a map, where the various text segments (e.g. sentences) are represented by nodes. Any thematic similarity between two segments (superior to 0.2) is represented by an arc between the corresponding nodes. The final thematic segments are defined by merging all triangles having similar centroïds within the map. The map complexity might be used to deduce the complexity of the document structure. Maps representing documents with thematically homogeneous text are highly convex, but it is not the case for documents with various heterogeneous texts. The LCP method (Lexical Cohesion Profile) [Kozima 1993] has been applied on narrative text. The basic idea is that the words contained in a sentence might have semantic relationship between them. The segmentation process consists in fixing a window around a particular word, and computing the similarity of this word with all the other words within the window, using a semantic network. Then the mean of all words similarities in this window is computed and represented in a line-graph. This way, a high mean similarity of a segment reflects the strong relationship between their words. The lowest valleys of the generated line-graph, which correspond to the weak mean similarities of segments, are considered as possible locations of topic change. In Ponte and Croft's method [Ponte and Croft 1997], the LCA method (Local Context Analysis) is used to get semantic relationships between words of sentences, in order to segment the text by topic. The assumption is that the statistical word frequency

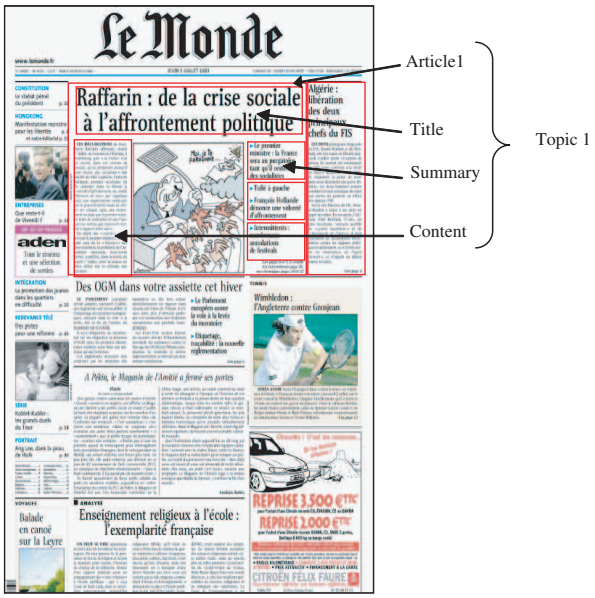


Figure 3.4: Relationship between various document structures

method itself might not be enough to detect relatively small topic segments, which might not share many common words. For this reason, segments are ranked according to the difference between their internal similarity and external similarity. This way, a segment with a high internal similarity and a low neighbor similarity might be considered as a thematic segment.

3.2.4 Document Syntactic Segmentation

The syntactic segmentation of a document consists in its decomposition into textual units, such as paragraphs and sentences, without the consideration of the geometric, logical or thematic characteristics. Sentence boundary detection is based on pure syntactic aspects, such as the capitalized word first letter, punctuation marks, etc. Paragraph segmentation might be obtained for instance by grouping sentences that are spatially adjacent.

3.2.5 Relationship between various Document Structures

Each document can be represented by various structures as described in the previous paragraphs. Therefore, a document might be seen as a multi-layered structure,

when all these structures are superposed (see Figure 3.4). We have noticed that two segments from two different structures might be overlapped, or one might be included in the other. For example, a logical block might be composed of many physical blocks. Similarly, a physical block (e.g. a page) might be composed of many thematic segments. From another point of view, a thematic segment might be shared by two physical blocks (e.g. the end of the first column and the beginning of the second column), or by many logical blocks (see Figure 3.4).

Aligning static documents with the speech requires also a segmentation of the latter. Some of the state-of-the-art methods for segmenting the speech are described in the next section.

3.3 Speech Segmentation

When compared to static documents, the speech is characterized by dialog related features, such as speaker and dialog acts. These features might be used to segment the speech, either alone, or combined with classic linguistic information, such as word frequency.

3.3.1 Speaker Turn/Utterance Segmentation

In multi-party meetings, the discussion is turned between various speakers. Thus, speaker change might be considered as a criterion for the segmentation. Depending on the scenario of the meeting, a speaker turn lies from a part of a sentence to a set of sentences. For this reason, turns are also segmented into smaller units that are called utterances, which correspond to the smallest homogeneous parts of a turn.

3.3.2 Speech Thematic Episodes Segmentation

Discourse features might be an alternative or a complement for classic linguistic features (e.g. statistic word frequency), which should reinforce and increase the performance of topic segmentation methods. Cue phrases that might indicate the flow and structure of a discourse (e.g. "ok", "now", etc.) have been considered in many works related to dialog topic segmentation [Litman and Passonneau 1995] [Hirschberg and Litman 1993].

The study achieved in [Hirschberg and Litman 1993] aims to disambiguate these cue phrases, in order to ensure that they indicate correctly topic change. The

prosodic information, such as pause duration and pitch curve, has been considered in this disambiguation.

In the work made by Litman and Passonneau [Litman and Passonneau 1995], a machine learning approach of discourse segmentation boundary with two classes (boundary and non-boundary), is proposed. Both prosodic and textual features of dialog have been exploited in this work. Thus, in a new utterance of a turn, if a particular sentence starts with a cue phrase word (e.g. "first", "also", "now", etc.), followed by another secondary cue phrase word, then this should be considered as a potential boundary. However, if there is a co-reference across this potential boundary, where the first noun phrase after the boundary is co-referent with a noun phrase located before the boundary, then this decreases its likelihood to be a boundary. Other prosodic information might also help, such as pause duration that might be a crossing to a new topic.

3.3.3 Other Speech Segmentations

In case of multi-party meetings, the various speakers' utterances might be analyzed in order to determine their role in the meeting. Thus, different kinds of dialog acts (DA) have been defined (question, answer, agreement, apology, etc.). DA segmentation and classification have been studied in many works [Zimmermann et al. 2005] [Popescu-Belis 2004a] [Ang et al. 2005].

Two different approaches for DA segmentation have been used in the study of Ang et. al [Ang et al. 2005]. First, a decision tree, with pause duration as a feature, has been used in order to estimate the conditional probability of boundary class, i.e. if it is a DA or not. The second approach is based on the hidden event language model (HE-LM), that predicts DA boundary or non-boundary event. DA classification is performed using Maxent, a Maximum Entropy classifier, which exploits textual features such as units' length, the first two words, the final two words, the initial words of the following DA, etc. Other prosody features have been also used, such as DA duration, average pitch, etc.

Our multimodal alignment framework is based on the comparison of the segments of the documents to align, according to various processes (thematic, quotation and reference alignment). Some techniques of text processing, that are involved in text comparison and that might be useful in our alignment methods, are described in the following sections.

3.4 Insignificant Elements Removal from Segments

The main parameter in the relationship between compared textual segments is the number of common key concepts, as well as their frequency. For this reason, a cleaning and reorganization of the segments content are required. These are performed by using stop words removal technique, highlighting words that are morphologically similar using stemming technique, and linking words that are semantically similar by exploiting a semantic network (thesaurus).

3.4.1 Stop Words Removal

Considered as poor descriptors of documents, and meaningless in terms of information retrieval, stop words (articles, prepositions and conjunctions, pronouns, auxiliary verbs, etc.) are actually neglected in many information retrieval tasks. The main aim is to reduce the number of common words between documents descriptors, and to save space and time. Stop words are called also noise words, or negative dictionary. However, in some cases, ignoring stop words could lead to conflicts and decrease the sense of the phrase. For instance, stop words that have more than one meaning (e.g. "a" in "a vitamin" and in "vitamin a" in English, or the French word "été" that means "been" and "summer"). There are also some stop words that are part of proper names (e.g. the preposition "on" in "Stoke-on-Trent" town). Moreover, grouping some individual stop words might produce meaningful expressions (e.g. "to be or not to be"). Stop word lists might be created in advance, such as the classic Brown list, or derived automatically from the document collection, according to words frequencies. Stop word lists depend highly on the collection of the documents to process. Therefore, a stop word list for a scientific document is not the same as a stop word list for a speech transcript. The latter category might contain specific speaker vocabulary (e.g. "umm", "okay", etc.). In order to maximize the performance of an information retrieval system, it is crucial to adapt the existing classic stop word lists to the content of the collection, or to merge them with other lists generated by other approaches, such as the Zipf law [Tsz-wai et al. 2005].

After removing all stop words from segments, there are still only significant terms that should reflect the semantics of the segments. For example, the segment "the alignment requires a processing of data" is reduced to "alignment requires processing data", where the articles "the" and "a" and the preposition "of" have been removed, and only the significant words remain.

Two compared segments might contain different words having the same morphology. Exploiting this morphological similarity between words should consolidate and reinforce the relationship between segments. For this reason, there exists a technique, called word stemming, in which, any suffixed or prefixed word is reduced to its stem, according to its morphological and inflectional rules.

3.4.2 Word Stemming

A stemmer algorithm can be defined as a conflation method that reduces any word to its stem root form, by removing its suffix and prefix (e.g. "retrieval", "retrieving" are stemmed into "retriev"). Thus, the key concepts in a retrieved document segment will be grouped by stem, then the stems frequencies will be considered when measuring the relationship between the two compared segments. Moreover, a segment will be indexed with all the linguistic roots of its terms, instead of being indexed by all the existing words, which reduces drastically the index size. The involved rules within a stemming algorithm vary from language to language.

The stemming methods have been used widely in various information retrieval fields, such as search engines, computational linguistic applications and other natural language processing tools, such as the Coveo Enterprise Search tool [Coveo]. Rather than regular stemmer algorithms, there exist multilingual stemmers that consider the morphological rules of several languages at the same time, which is very useful for documents with many languages. These multilingual stemmers are enriched by procedures that determine which rules of which language to apply, for a particular word. However, these procedures are still insufficient to avoid the ambiguity between languages, which decreases the precision of these multilingual stemmers, compared to monolingual stemmers.

Many stemming algorithms exist. However, selecting the most suitable one should take into account its performance and strength. The famous stemmer algorithms are Lovins (1968), Paice/Husk (1990), Krovetz stemmer (1993) and Porter (1980) [O'Neill and Paice 2001]. Lovins stemmer has been developed by Julie Beth Lovins at Massachusetts Institute of Technology in 1968. Even if this algorithm considers both information retrieval and linguistic areas of stemming, it does excel neither of them. Lovins rules are still insufficient, since they have been derived by processing and studying a word example. Moreover, the large rules set affects the stemmer performance and attempts to recode words. The Paice/Husk stemmer

is a conflation based iterative stemmer, that has been developed at the University of Lancaster in 1990 by Chris Paice. Krovetz stemmer has been developed at the University of Massachusetts in 1993. This is a linguistic morphological inflection stemmer, which makes it of a low level strength. For this reason, Krovetz stemmer is frequently used in conjunction with the other stemmers, in order to exploit its accuracy of removal of suffixes. Porter stemmer is a conflation stemmer, that has been developed by Martin Porter at the University of Cambridge in 1980. The Porter algorithm is linear step stemmer, having five steps. Within each step, the corresponding rule is applied, once its prefix/suffix matches a word and if its corresponding conditions are checked (i.e. number of characters). Finally, the detected prefix/suffix is removed, and the next step is performed.

Besides the stemming technique that reduces words to their stems, words might be also reduced to their normalized forms, thanks to the lemmatization technique. The latter is presented in the next section.

3.4.3 Word Lemmatization

Word lemmatization is a linguistic process that reduces any word to its lemma, by substituting its grammatical ending by the normalized ending. Thus, lemmas for nouns group singular and plural forms. Lemmas for adjectives group comparative and superlative forms. Finally, lemmas for verbs group the various verb tenses. The lemmatization process might be similar to the stemming process, especially for words that have a stem similar to their lemma (e.g. "reading" that is stemmed and lemmatized to "read"). However, for other words, the lemma might be different from the stem (e.g. "taking" that is stemmed to "tak", but lemmatized to "take").

The lemmatization process has been defined in order to cover all forms of a given word in an information retrieval system, such as search engines. The lemmatization process is involved in many natural language processing tasks, such as syntax parsing, machine translation, automatic indexing [Strömbäck 2005]. Thus, many lemmatization systems have been developed for many languages, such as the Morph system developed for English language [Strömbäck 2005] [Minnen et al. 2001].

The usefulness of the lemmatization in an information retrieval task depends on other factors, such as the implication of a thesaurus or not, since this latter needs to link words according to their normalized forms.

Considering the semantic relationships between words in a corpus (synonymous, hyponymy, etc.) requires the integration of a semantic network or a thesaurus. In the next paragraph, an overview on the main thesauruses is presented.

3.4.4 Thesaurus Integration

A thesaurus is defined as a set of different language resources, useful for a range of different language engineering purposes [Kilgarriff and Yallop 2000]. Thesauruses are mainly used for word sense disambiguation [Chakravarthy 1995] [Kwong 1998], as well as applications dealing with natural language, such as search engines and text retrieval systems. A thesaurus might be seen as a network connecting many words, according to various criteria (semantic, lexical, syntactic, etc.). Many thesauruses exist nowadays, some of them are for general use such as WordNet [WordNet] and Roget [Jarmasz and Szpakowicz 2001], and others are domain-specific thesauruses, such as UMLS, a medical thesaurus [UMLS], WebLaw, a law thesaurus [WebLaw], Getty, a thesaurus for geographic names [Getty], etc.

WordNet is a famous English lexical semantic thesaurus [WordNet], which has been developed at the University of Princeton. WordNet has been manually constructed, according to psycholinguistic principles. WordNet thesaurus is composed of many synsets, i.e. sets of synonyms, where each set corresponds to a concept, and groups all the words sharing the same sense. Thus, a word that has several senses occurs in many synsets. The synsets are hierarchically organized, reflecting hierarchical relationships (e.g. synonyms, hyponyms, antonyms, holonyms, etc.). WordNet thesaurus groups about 200.000 words, classified into four main categories, nouns (67%), verbs (12%), adjectives (18%) and adverbs (3%) [Jarmasz and Szpakowicz 2001].

The next paragraph describes another technique that should highlight key terms within documents, and thus maximizes the information retrieval system performance, which consists in associating weights to terms.

3.4.5 Term Weighting using TF.IDF Metric

Terms weighting should reflect the importance of terms within documents or document segments. Rather than the binary weighting (i.e. one for an existing terms in a document and 0 for a non-existing one), two basic functions exist in order to measure terms weight, the Normalized Term Frequency (TF) and the Inversed Document Frequency (IDF). The TF function measures the importance of a term

within a document. Therefore, a term with a high frequency becomes a descriptor or key term of this document.

The IDF function, initially called term specify [Robertson and Sparck 1976], measures the term importance in the context of the entire documents collection. Thus, a term having a high frequency in the entire collection should be insignificant and less important, since it can not be used to distinguish documents. Unfortunately, a term that appears in few documents is more significant, and thus it can be used as a descriptor. In other words, when the number of documents sharing a particular term is high, then the importance of this term is low, and vice versa.

In order to make a balance between the two functions, the weight of a term t in a document D is measured by the ratio of TF and IDF functions as follow:

$$W_{t,D} = TF.IDF_{t,D}$$

Measuring the relationship between two segments might be made by comparing them, using various techniques. The best known techniques are dynamic programming, latent semantic analysis (LSA), terms co-occurrences based similarity metrics, etc. The last technique is known to be one of the stronger and most precise methods. In the next section, state-of-the-art of various works related to these similarity metrics is presented.

3.4.6 Similarity Metrics

Determining how two documents, or documents segments, are thematically related requires a mechanism that translates this thematic relationship in terms of numeric scores, where a high score corresponds to a high closeness of the compared segments. This mechanism is known as the thematic similarity. In our work, we consider that two segments of a document are thematically similar, if their thematic similarity value is important. In the information retrieval field, finding similar text segments has been the base for many research works, such as the thematic segmentation of documents [Hearst 1994], the story link detection [Chen et al. 2004], the publication likelihood [Bani-Ahmad at al. 2005], etc. In Hearst's thematic segmentation method [Hearst 1994], the candidate thematic segments of a document are compared by the Cosine metric, in order to detect the change of topic between each two successive segments. To detect links between stories in the work conducted by Chen et al. [Chen et al. 2004], various similarity measures have been used, among time-

ordered sets of stories. In the study of Bani-Ahmed et al. [Bani-Ahmad at al. 2005], similarity metrics have been applied to compute similarity between documents content (title, abstract, body, etc.), in order to detect similar publications. Jeaon et al. [Jeon et al. 2005] have focused on detecting similar questions among FAQ (Frequently Asked Questions) pages, by computing similarity between the answers of questions.

Computing the similarity of two segments is made by defined measures between them, involving their common terms. Given two segments S_1 and S_2, and $W_{t,Si}$ the weight associated to a term t in the segment S_i, the known measures are computed according to the following formulas, where the generated similarity value is varying between 0 and 1:

$$Cosine(S_1, S_2) = \sum_{t=1}^{N} W_{t,S_1}.W_{t,S_2} / \sqrt{\sum_{t=1}^{N} W_{t,S_1}^2 . \sum_{t=1}^{N} W_{t,S_2}^2}$$

$$Jaccard(S_1, S_2) = \sum_{t=1}^{N} W_{t,S_1}.W_{t,S_2} / (\sum_{t=1}^{N} W_{t,S_1}^2 + \sum_{t=1}^{N} W_{t,S_2}^2 - \sum_{t=1}^{N} W_{t,S_1}.W_{t,S_2})$$

$$Dice(S_1, S_2) = 2. \sum_{t=1}^{N} W_{t,S_1}.W_{t,S_2} / (\sum_{t=1}^{N} W_{t,S_1}^2 + \sum_{t=1}^{N} W_{t,S_2}^2)$$

The existing similarity measures have been classified in many works. In the study of Michel [Michel 2000], three categories have been defined, cardinal, nominal and ordinal measures, where document collections are manipulated as sets. The cardinal measures are useful when there is no information available about the various documents of both collections. In such a case, collection cardinality is exploited (e.g. $|C_1|/|C_2|, |C_1|/(|C_1|+|C_2|)$, etc.). When documents of both collections are identified individually, so that common and non-common documents are known, then nominal measures might be used. These nominal measures make use of common documents between collections, computed by $|C_1 \cap C_2|$, or all documents, computed by $|C_1 \cup C_2|$. Using this second category, the previous similarity metrics are represented as follows:

$$Cosine(C_1, C_2) = |C_1 \cap C_2| / \sqrt{|C_1|.|C_2|}$$

$$Jaccard(C_1, C_2) = |C_1 \cap C_2| / |C_1 \cup C_2|$$

$$Dice(C_1, C_2) = 2.|C_1 \cap C_2| / (|C_1| + |C_2|)$$

In the last category, when the documents of both collections are personalized and totally ranked, the ordinal measures are used.

3.5 Conclusion

In this chapter, we have presented the main methods that are used for static documents segmentation, speech segmentation, and text processing. In the next chapter, our multimodal alignment framework is described. Therefore, the pre-processing of the data, that exploits some of the methods studied in this chapter, is described. Then, our three alignment types are studied in details: thematic, quotations and references.

Chapter 4

Multimodal Alignment of Document with Speech Transcript

4.1 Introduction

In the previous chapter, various methods related to document and speech analysis, as well as information retrieval techniques, required for text segments comparison, have been presented. The current chapter describes our alignment framework. In fact, the focus is made on aligning static documents with dialog recordings and with slideshows within multimodal events, such as press review meetings (Figure 4.1) and scientific conferences. Our alignment framework is composed of three main steps:

- The first step consists in a pre-processing of the multimodal data taken as input, printed documents, speech and slideshows (Figure 4.2).
- The second step corresponds to the alignment method, composed of three processes applied independently on the data, thematic, quotation and reference alignments (Figure 4.3).
- In the last step, the results of the three alignment processes are merged within a same structure (Figure 4.3).

Before aligning the static documents with speech or with the slideshows, the three documents should be prepared for the alignment framework. The next section

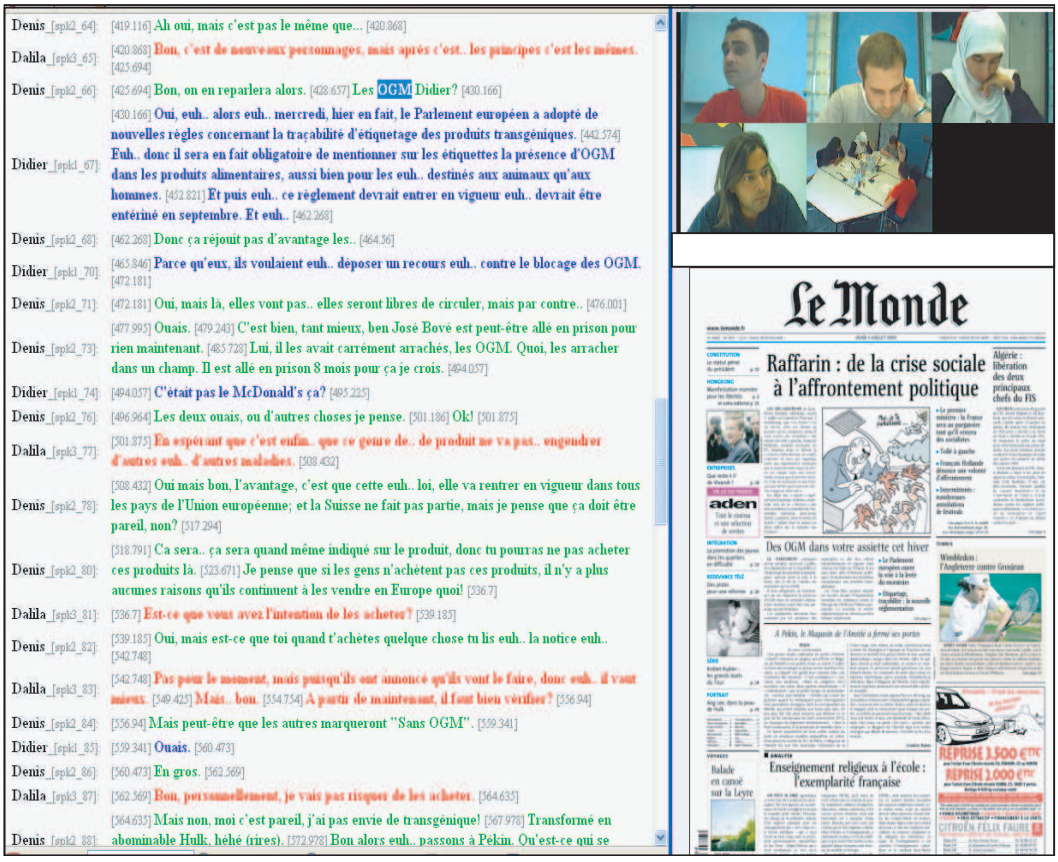


Figure 4.1: Press review meeting data: speech transcript (left hand), meeting participants (top right hand), and discussed document (bottom right hand)

presents briefly the methods, among those studied in chapter 3, that are required for our multimodal alignment framework.

4.2 Preparing Data for Alignment

Aligning static documents with speech transcript or with slideshows requires first a preparation of the textual data of these resources (Figure 4.2). The textual content of static documents and slideshows is extracted from the PDF files (Portable Document Format). Then the three resources are segmented, so that their respective segments could be manipulated individually, in order to detect the relationships between them. Finally, the text of the respective segments is processed. Segments

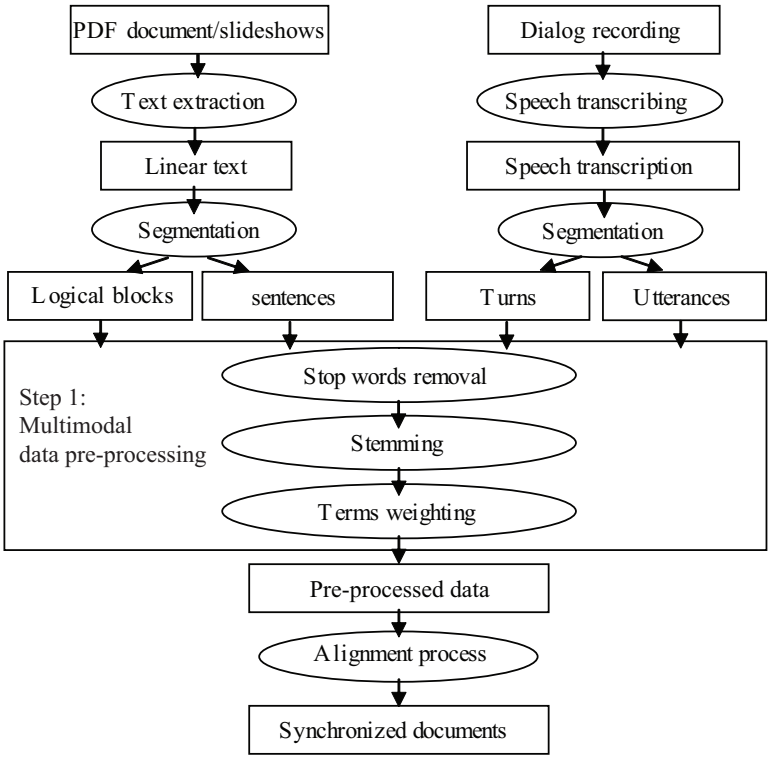


Figure 4.2: Multimodal alignment framework. Focus made on the pre-processing of data

processing consists in cleaning them by removing insignificant words (stop words) and part of words (prefixes and suffixes), and associating weight to the remaining terms, in order to highlight the important terms within each segment. The segmentation methods of static documents, speech transcript and slideshows, that have been employed in our work, are presented in the next paragraphs.

4.2.1 Static Document Segmentation

Various document structures have been studied in section 3.2 of chapter 3, physical, logical, thematic and syntactic structures.

In our thesis, some document physical components are not suitable for our alignment relationships, such as pages or columns. For this reason, the physical segmentation has not been considered in our alignment process.

The logical segmentation of documents highlights their functional segments (e.g. title, article, etc.), which is very significant for our alignment process. Thus, it has

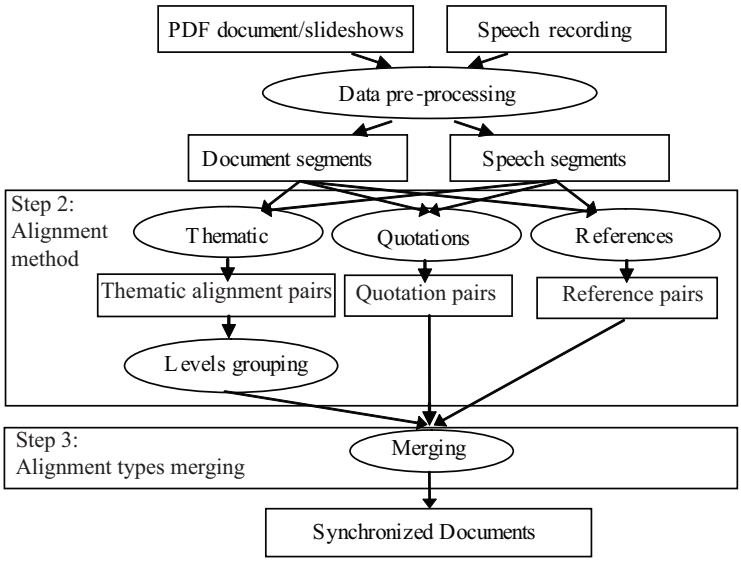


Figure 4.3: Multimodal alignment framework. Focus made on the alignment processing

been considered as one of the crucial structures to perform this process. Actually, the logical structure of documents that are used in our work, as it is described in chapters 5 and 6, has been manually extracted.

Even though there are a variety of thematic segmentation methods for documents (section 3.2.3 of chapter 3), these methods are not well adapted for our document types, for example newspaper pages, where the various articles have different sizes and different topics. For instance, TextTiling method [Hearst 1994] has been applied on magazine articles that focus on the same topic. The LCA approach [Ponte and Croft 1997] has been adopted for small thematic segments, such as newsbite feeds. Due to this document type dependency, the thematic segmentation has not been considered in our alignment process. However, the strong relationship that seems to exist between the thematic segmentation and the thematic alignment, that is studied in this thesis, leads to think that this alignment type might help in discovering the thematic segmentation of events, and thus the thematic segmentation of documents, as it is described in chapter 7.

Our document syntactic segmentation might generate either paragraphs or sentences. Document sentences are the smallest significant units in a document, which are equivalent to speech transcript utterances seen in section 3.3.1 of chapter 3. For

this reason, document segmentation into sentences has been chosen, combined with the logical segmentation, for our alignment process.

4.2.2 Speech Segmentation

In section 3.3 of chapter 3, various segmentation methods of the speech have been presented, turn/utterance, thematic episodes and dialog acts segmentations respectively. However, only the turn/utterance segmentation has been considered in our alignment framework.

Our alignment method of the static documents with speech recording requires the transcription of the latter. Many researches are made in order to build a powerful automatic speech recognizer system (ASR), such as CMU Sphinx [Sphinx], Scansoft [Scansoft] and HTK tools [HTK]. However, in order to avoid the effect of the noise on our alignment framework, we have been limited in this study to the manual transcription of the speech. Later in chapter 6, the robustness of our framework to noise is measured by simulating ASR systems. This simulation consists in adding noise to the manual speech transcript, according to a fixed Word Error Rate (WER) threshold.

The transcriber tool [Barras et al. 1998] has been used in our work in order to generate the turn/utterance structure, which combines speakers' turns and utterances. The turn/utterance structure has been decomposed into two different structures, respectively turn and utterance structures, in order to be used in our multimodal alignment framework.

4.2.3 Slideshows Segmentation

The segmentation methods of static documents can be applied on slideshows, in order to generate a set of segments that can be manipulated individually. As seen in section 4.2.1, the physical structure of a document is not really significant for our thematic alignment, especially for slideshows, since their geometric description is very variable. The thematic structure of slideshows is difficult to extract, due to the nature of their content that consists usually in a list of paraphrases. For all these reasons, we have fixed the logical structure as the ideal slideshows structure for our alignment, where each slide is considered as a logical block.

Once the static documents, speech and slideshows are segmented, their respective segments should be compared in a pairwise manner, in order to get relevant pairs of

alignment. However, the relevance of alignment results depends drastically on the strength of the relationships between compared segments. The information retrieval techniques, described in section 3.4 in chapter 3, aim to reinforce the links between compared segments. Among these techniques, those that have been chosen for our alignment framework (Figure 4.2) are briefly presented in the next section.

4.2.4 Data Representation

The various data resources, speech transcript, static documents and slideshows, are represented in the XML format, according to specific DTD definitions. For instance in Figure 4.4, two extracts from the speech transcript and the newspaper respectively, in the context of press review meetings, are presented. In the speech transcript extract, the $< Turn >$ element, that corresponds to a speaker turn, contains some attributes, such as the speaker name, the start time and the end time temporal indexes. Inside the $< Turn >$ element, there are $< Utterance >$ elements, where each utterance contains its own start time, as well as its textual content.

From another hand, in the extract of the document logical structure, the logical elements are presented in a hierarchical form, such as the $< MasterArticle >$ element that contains the elements $< Title >$, $< Content >$, etc. These various elements correspond to the different logical labels within the newspaper document.

4.2.5 Segments Processing

Before aligning document segments, these segments should be first cleaned from stop words (section 3.4.1 of chapter 3). In our case, for English and French data respectively, existing stop word lists have been enriched in order to construct a final list containing 322 words for the English corpus and 540 words for the French corpus. Since our corpus includes also speech transcript, the lists have been adapted with specific speech stop words (e.g. "umm", "okay" for English, and "ben", "emmh" for French). In Figure 4.4, after stop words removal, the utterance:

"Alors l'article principal est le 1er mai tourmenté de Jean-Pierre Raffarin"
is reduced to: "article principal 1er mai tourmenté Jean-Pierre Raffarin".

After stop words removal, the remaining words in each segment should be stemmed by removing the prefix/suffix parts (section 3.4.2 of chapter 3), so that the morphological similarity of words could be exploited when comparing segments. Porter is one of the most stronger and complete stemmers, which has been widely used for

word stemming [Saetre et al. 2005] [Abberley et al. 1998]. Therefore, it has been selected in our work for the English corpus. In order to process the French corpus, an update has been made on Porter rules by covering all prefixes and suffixes of the French language. Thus, the words "principal" and "article", in the first utterance in Figure 4.4, are stemmed to "princip" and "articl", respectively.

The detection of relationships between segments should be enhanced when the important terms within each segment are highlighted. Highlighting the important terms within each segments might be performed by the consideration of terms weight. The $TF.IDF$ metric, seen in section 3.4.5 of chapter 3, has been initially used for terms weighting in document collections. In our case, documents are segmented, and the generated segments are manipulated individually. For this reason, the $TF.IDF$ metric is computed for terms within segments in the context of the overall document, rather than the documents collection. Two main formulas have been used to compute the weight of a term t in a segment S_i:

$$TF.IDF_{t,S_i} = occ_{t,S_i}.\sum_{j=0}^{N} occ_{t,S_j}$$

$$TF.IDF_{t,S_i} = occ_{t,S_i}.log_2(N/n_t)$$

where N is the number of document segments, n_t is the number of segments containing the term t, and occ_{t,S_j} is the number of occurrences of a term t in a segment S_i.

After pre-processing the data of static documents, speech and slideshows, these resources should be pairwise aligned. In the following section, the alignment of the static document with speech transcript is described. However, the multiple documents alignment, involving the three resources, static documents, speech and slideshows, is presented in section 4.4.

4.3 Document/Speech Alignment Methodology

Aligning the printed documents discussed during an event (meeting, conference, etc.) with the speech transcript of this event consists in extracting the existing relationship between them, at variable granularity levels and meeting time. Since the two resources, documents and speech transcript, are being segmented, as seen

Speech transcript

```
<Trans>
  <Turn speaker="Didier" Start-Time="38.41" End-Time="50.1">
    <Utterance time="38.415">Alors l'article principal est le 1er mai tourmenté de Jean-Pierre  Raffarin..
    </Utterance>
    <Utterance time="42.794"> Cela fait presque ...
    </Utterance>
  </Turn>
  <Turn speaker= "Denis" Start-Time= "52" End-Time="70">
    <Utterance time="52">Passons maintenant à l'article suivant sur l'immigration et Sarkozy …
    </Utterance>
  </Turn>
   ...
</Trans>
```

Document

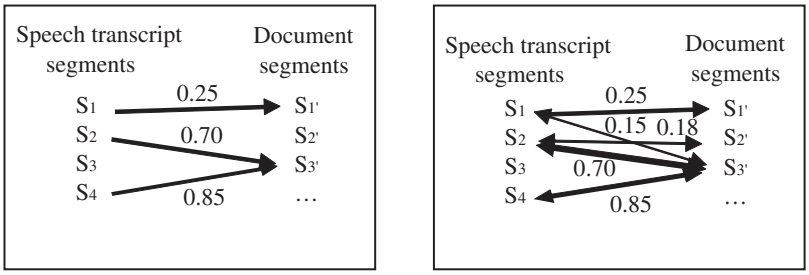
```
<Newspaper>
 <MasterArticle>
  <Title>Le 1er Mai tourmenté de Jean-Pierre Raffarin</Title>
  <Content>A quelques jours du premier anniversaire de l'installation de Jean-Pierre Raffarin à Matignon
  </Content>
 </MasterArticle>
  …
</Newspaper>
```

Figure 4.4: Extracts from the speech transcript of a press review meeting, and the logical structure of the discussed newspaper, respectively.

in section 4.2, their respective segments are compared by pairs. In other words, each segment from the source file (document or speech transcript) is compared with all the segments from the target file, in order to get the pertinent couples forming alignment pairs.

4.3.1 Various Document/Speech Transcript Alignment Types

The alignment of respective segments of the static document and speech transcript covers many types of alignment, which are extracted following several processes. Three main types of alignment between static documents and speech transcript have been defined and studied in this thesis: thematic, quotation and reference alignments [Mekhaldi et al. 2003] [Lalanne et al. 2004b]. In the next section, the thematic alignment is presented.



a. 1-best alignment strategy (asymmetric); b. Multiple alignments strategy (symmetric)

Figure 4.5: Thematic Alignment Strategies

4.3.1.1 Thematic Alignment of Document with Speech Transcript

A thematic alignment is defined as being a matching process between segments that share the same themes. Therefore, thematic links are built between the segments of the documents and the segments of the speech transcript. The thematic alignment process is mainly based on the computation of the thematic similarity between the compared segments, in order to elicit which document and speech transcript segments share the same theme. For instance in Figure 4.4, the two following speaker utterance, and the *Title* logical block in the document, are thematically similar:

- "Alors l'article principal est le 1er mai tourmenté de Jean-Pierre Raffarin".
- "Le 1er Mai tourmenté de Jean-Pierre Raffarin".

Following the above process, some segments from the documents being discussed will be linked to the thematically similar speech transcript segments, and some speech transcript segments will be linked to the similar documents segments. Therefore, two directions are distinguished in our thematic alignment process, from documents to speech, and vice versa.

In addition to the thematic linking with the speech transcript, document segments will be enriched by temporal indexes, expressing this way when these segments have been discussed. For instance, the document logical block *Title* in Figure 4.4 will be enriched by the temporal index of the corresponding speaker utterance.

Detecting the thematic links between segments of the respective resources is based on the computation of the thematic similarity between them. Therefore, the similarity metrics, described in section 3.4.6 of chapter 3, have been used.

4.3.1.1.1 Similarity Metrics: in order to compute the similarity between vectors representing the respective segments of documents and speech transcript in our thematic alignment process, three existing similarity metrics have been chosen in our work, respectively *Cosine*, *Jaccard* and *Dice*. All these similarity metrics are based on term co-occurrences within compared segments.

Given two segments S_1 and $S_{1'}$ from the document and the speech transcript respectively, and W_{t,S_i} the term frequency (TF) associated to a term t in the segment S_i, the similarity between S_1 and $S_{1'}$ is computed according to the following formulas, where the generated value varies between 0 and 1:

$$Cosine(S_1, S_{1'}) = \sum_{t=1}^{N} W_{t,S_1}.W_{t,S_{1'}} / \sqrt{\sum_{t=1}^{N} W_{t,S_1}^2 . \sum_{t=1}^{N} W_{t,S_{1'}}^2}$$

$$Jaccard(S_1, S_{1'}) = \sum_{t=1}^{N} W_{t,S_1}.W_{t,S_{1'}} / (\sum_{t=1}^{N} W_{t,S_1}^2 + \sum_{t=1}^{N} W_{t,S_{1'}}^2 - \sum_{t=1}^{N} W_{t,S_1}.W_{t,S_{1'}})$$

$$Dice(S_1, S_{1'}) = 2.\sum_{t=1}^{N} W_{t,S_1}.W_{t,S_{1'}} / (\sum_{t=1}^{N} W_{t,S_1}^2 + \sum_{t=1}^{N} W_{t,S_{1'}}^2)$$

The thematic similarity between two segments might be enhanced if the semantic relationship between words is considered. For this aim, the WordNet semantic thesaurus [WordNet] has been integrated into the similarity metrics.

4.3.1.1.2 WordNet Integration: in order to link words in our corpus according to their semantic relationships, WordNet [WordNet] has been chosen as being a general and large thesaurus for the English language. However, we did not get any pertinent thesaurus for the French corpus, except EuroWordNet [EuroWordNet], a multilingual thesaurus that includes several European languages, which is not free of charge for the French language. WordNet thesaurus corresponds to a set of files that describe relationships between words. These files might be accessed and loaded using specific libraries, such as the JWNL API (Java WordNet Library) [Corral 2005]. Within WordNet thesaurus, words are organized in four categories, nouns, verbs, adjectives and adverbs, representing four separated semantic networks. Each category is composed of many synsets (i.e. sets of synonyms) that group words having the same semantic meaning. Each word is characterized by one or several POS values (word Part Of Speech), depending on the number of the word meanings, either in one or several categories [Corral 2005].

Integrating WordNet thesaurus in similarity metrics means that the comparison of words should take into account their semantic relationship inside WordNet. Therefore, when two words are compared, they are searched respectively in the WordNet nouns category, verbs category, adjectives category then in adverbs category [Corral 2005]. After that, the relationship between the two words is determined. Since words are organized inside WordNet according to their meanings (i.e. in the form of synsets), one of the strategies that have been adopted for determining the relationship between two compared words consists in searching their respective first meanings [Corral 2005]. Then, a weight is associated to each relationship (synonyms, hyponyms, etc.), according to its type and to the depth of the path between the two words. Thus, if the two words are equals, synonyms, hypernyms or hyponyms, then the weight of their relationship corresponds to 1, 1, 0.8 and 0.8 respectively. Finally, the scores obtained for the respective pairs of compared words are taken into account in the similarity metrics.

After computing the similarity between all source segments (from the documents or the speech transcript) and target segments, only the pertinent alignment pairs are preserved. In the next section, the criteria for selecting these pertinent alignment pairs are described.

4.3.1.1.3 Selection of Pertinent Thematic Alignment Pairs: given that N and M are the respective segment numbers in the speech transcript and the documents, the number of the generated pairs, after computing the similarity, might reach NxM thematic alignment pairs. The similarity value of these pairs lies between 0 and 1, thereby possibly including many insignificant weak thematic links. In order to filter these insignificant links, we have fixed two strategies, taking into account the number of links that should be preserved for each source segment, the 1-best alignment and the multiple alignments strategies.

The 1-best alignment strategy consists in selecting the link with the highest similarity value for each source segment (e.g. in Figure 4.5.a, $S_{3'}$ is the most similar segment for S_2, with a score of 0,70). In this strategy, the alignment links are asymmetric. If a segment S_1 from the speech transcript is the most similar segment of a particular segment $S_{1'}$ from the document (i.e. S_1 should be aligned with $S_{1'}$), then the opposite is not necessarily true. In Figure 4.5.a, the segment $S_{3'}$ is the most similar for S_2 (0.70), even if S_4 is the most similar for $S_{3'}$ (0.85).

In the multiple alignments strategy, all the relevant links between a source seg-

ment and all the target segments are kept (e.g. S_1 in Figure 4.5.b is linked with $S_{1'}$ and $S_{3'}$). The selection of the relevant links in this strategy is based on filtering the insignificant links, which have a similarity value inferior to a determined threshold (fixed to 0.10 in our work). Within this alignment strategy, the generated alignment links are symmetric, i.e. the same links are detected in both directions, from the speech transcript to the documents and vice versa, if the threshold value is the same in both directions.

Besides the similarity measures, two functions have been defined in our work in order to measure the overlap between the compared segments, the *membership* and the *ownership* functions. The *membership* of a segment S_1 in a segment $S_{1'}$ measures the percent of terms of S_1 being present in $S_{1'}$. Whereas, the *ownership* measures the percent of terms of $S_{1'}$ being present in S_1:

$$Membership(S_1, S_{1'}) = |S_1 \cap S_{1'}|/|S_1|$$

$$Ownership(S_1, S_{1'}) = |S_1 \cap S_{1'}|/|S_{1'}|$$

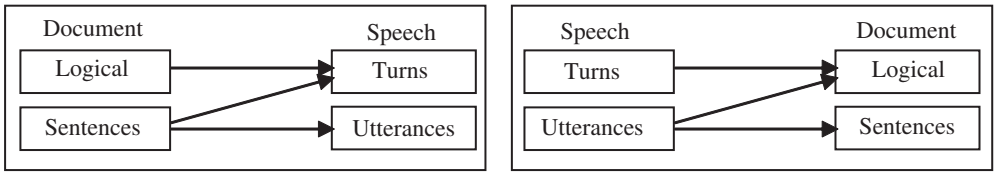
These two functions constitute another way for measuring the thematic similarity between textual segments, which might be complementary for the information given by the similarity metrics. An example of use of these two functions is shown in section 4.3.1.1.5.

4.3.1.1.4 Thematic Alignment and Multiplicity of Structures: the thematic alignment is applied between the segments of the documents and the speech transcript. The multiplicity of structures for both documents and speech transcript leads to a multiplicity of alignment levels. Thus, the following pairs can be constructed between the documents and the speech transcript:

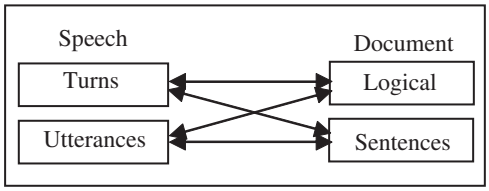
- Document physical/speech thematic, document physical/speaker turns, document physical/speaker utterances.

- Document logical/speech thematic, document logical/speaker turns, document logical /speaker utterances.

- Document thematic/speech thematic, document thematic/speaker turns, document thematic/speaker utterances.



a. 1-best alignment strategy (oriented)



b. Multiple alignments strategy (non-oriented)

Figure 4.6: Selected structure combinations, for both thematic alignment strategies

- Document Sentences/speech thematic, document sentences/speaker turns, document sentences/speaker utterances.

However, some of these combinations are not relevant and should be neglected, due to the non-significance of some structures in our alignment, such as the document physical structure, or to the non-availability of other structures, such as the document and speech transcript thematic structures. Moreover, the alignment strategy to be applied has an influence on the utilized structures. In some of the previous structure combinations, the size of the source segment might outweigh the size of the target segment, for instance when aligning documents logical blocks with speaker utterances. In that case, considering the 1-best alignment strategy, i.e. only one thematic link for each source unit, becomes unfair, mainly because the source segment is larger than the target segment, and thus it should be aligned with many target segments rather than only one. However, if the multiple alignments strategy is employed, then the significant links are more likely to appear and to be preserved, independently from the difference in size of compared segments. In Figure 4.6, the selected combinations of structures for our thematic alignment process, for both the 1-best and multiple alignments strategies are illustrated.

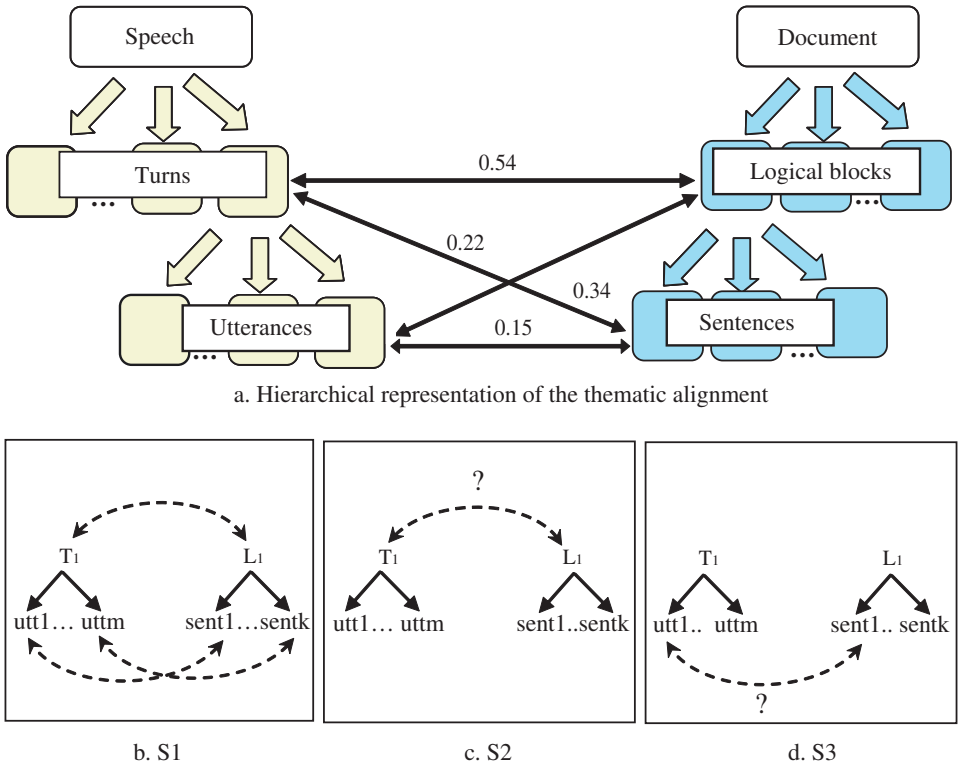


Figure 4.7: Thematic alignment levels and strategies for links validation

4.3.1.1.5 Thematic Alignment Levels: in Figure 4.7.a, the thematic alignment of documents with speech transcript is illustrated by a hierarchical representation, composed of two main levels, logical blocks/turns, and sentences/utterances levels, with the second level being a descendant of the first one. Furthermore, by using the multiple alignments strategy (Figure 4.5.b), the alignment links generated in each level becomes symmetric. These two criteria, the hierarchy of levels and the symmetry of links at each level, might be exploited in order to treat incoherent links (Figure 4.7). Supposing that we have an alignment link between a speaker utterance utt_i and a document sentence $sent_j$ (or an alignment link between a speaker turn T_i and a document logical block L_j, respectively). Then it is possible to check if their ascendant segments (respectively their descendant segments) are correctly aligned or not, i.e. if there is coherence between both levels or not.

In order to deal with the incoherence of alignment levels, we have defined three main validation strategies at the turns/logical blocks level [Mekhaldi et al. 2005]. In the first strategy S_1 (Figure 4.7.b), only the links between segments, where at least one link exists between their descendant segments, are preserved. For instance, the link between T_1 and L_1 in Figure 4.7.b is preserved, since their descendants are linked. Likewise, the link between turn 4 and logical block 19 in Figure 4.8 should be preserved, since there exist thematic links between utterance 11 and sentence 44, and between utterance 12 and sentence 45.

In the second strategy S_2, the links between segments, whose descendants are not linked, are considered as suspects to be wrong links (Figure 4.7.c). This kind of link incoherence is due to two main reasons. First, utterances of the turn T_1 might have small sizes, and thus they share few terms with sentences of the logical block L_1, even if the same theme is shared between these utterances and sentences. Thus, the detected similarity values between these utterances and sentences are weak and thus neglected. When the utterances are grouped in the form of a turn T_1, the number of shared terms between the latter and the logical block L_1 increases, and thus the similarity between the two segments, T_1 and L_1, will be significant. This case of link incoherence is illustrated in Figure 4.9. Therefore, turn 34 is thematically linked to logical block 9, however, there is no link between their descendant segments. The second reason of the incoherence of links between levels is explained by the fact that there is no shared theme between utterances of turn T_1 and sentences of logical block L_1, even if there are few shared terms between them. However, when the utterances and sentences are grouped respectively in the turn T_1 and logical block

Speech transcript

```
<Turn id="4">
  <Thematic With="logic">
    <logical_block id="19" similarity="0.59" membership="0.78" ownership="0.44"/>
  </Thematic>
  <utterances>
   <utter id="11"> Voilà. Euh.. la troisième semaine du procès Elf vient de commencer à
     Paris.
      <Thematic With="sentences">
        <sentence id="44"  similarity="0.16" membership="0.67" ownership="0.17"/>
      </Thematic >
    </utter >
    <utter id="12"> Euh.. durant les deux premières semaines le système de défense des
      principaux  prévenus ont été.. ont été profondément  déstabilisé.
       <Thematic With="sentences">
         <sentence id="45" similarity="0.30" membership="0.88" ownership="0.32"/>
       </Thematic >
    </utter>
      ...
  </utterances>
</Turn>
```

Document

```
<Logic id="19">
  <Thematic With="Turn">
    <turn id="4" similarity="0.59" membership="0.44" ownership="0.78"/>
  </Thematic>
  <sentences>
   <sent  id="44"> Justice : les surprises du procès Elf  La justice  marque des points au
     procès Elf, dont la troisième semaine devait s'ouvrir, lundi 31 mars, devant le tribunal
     correctionnel de Paris, avec le retour très attendu de Loik Le Floch-Prigent, l'ancien
     PDG du groupe pétrolier.
      <Thematic With="utterances">
        <utterance id="11" similarity="0.16" membership="0.17" ownership="0.67"/>
      </Thematic>
   </sent>
   <sent id="45">Les deux premières semaines d'audience ont profondément déstabilisé
     les systèmes de défense des principaux prévenus : M  Le Floch-Prigent, mais aussi
     Alfred Sirven, l'ancien directeur des affaires générales, et André Tarallo, l'ex-"M
     Afrique" d'Elf.
      <Thematic With="utterances">
        <utterance id="12" similarity="0.30" membership="0.32" ownership="0.88"/>
      </Thematic>
   </sent>
     ...
  </sentences>
</Logic>
```

Figure 4.8: Illustrative example for the validation strategy S_1

Speech transcript

```
<Turn id="34">
  <Thematic With="logic">
    <logical_block id="9" similarity="0.15" membership="0.4" ownership="0.01"/>
  </Thematic>
  <utterances>
    <utter id="50">Alstom...
      <Thematic/>
    </utter>
      <Thematic/>
    </utter>
    <utter id="51">C'est.. c'est une grosse société apparemment..
      <Thematic/>
    </utter>
    <utter id="52"> Euh.. elle contient euh.. 118 000 employés.
      <Thematic/>
    </utter>
  </utterances>
</Turn>
```

Document

```
<Logic id="9">
  <Thematic With="Turn">
    <turn id="34" similarity="0.15 membership="0.01" ownership="0.4"/>
  </Thematic>
  <sentences>
   <sent id="20">Alstom : l'ultimatum fait obligation à Paris de renoncer à toutes les
   mesures envisagées - entrée de l'Etat dans le capital  comme prêts subordonnés - si
   aucun accord avec Bruxelles n'est trouvé d'ici au  22 septembre
     <Thematic/>
   </sent>
   <sent id="21">Cette décision, qui a donné lieu à un vif débat au sein du collège des
     commissaires, place le ministère français des finances dans une situation  difficile
     <Thematic/>
   </sent>
   <sent id="22">Après la controverse sur les déficits publics, il ne veut pas se dérober à
cette demande
     <Thematic/>
   </sent>
   <sent id="23">Il craint donc  que certaines banques, notamment étrangères, finissent
par se retirer du montage, ce  qui menacerait la survie du groupe d'énergie et ferroviaire,
qui emploie 118 000 salariés
    <Thematic/>
   </sent>
      ...
 </sentences>
</Logic>
```

Figure 4.9: Illustrative example for the validation strategy S_2

Speech transcript

```
<Turn id="1">
 <Thematic With="logic">
   <logical_block id="16" similarity="0.11" membership="0.17" ownership="0.07"/>
   <logical_block id="17" class="Added after levels Grouping" />
 </Thematic>
 <utterances>
   <utter id="1">Bonjour à tous les deux. Encore une fois nous
   allons voir la une du Monde de ce mercredi 2 avril 2003.
     <Thematic/>
   </utter>
   <utter id="2">Pour commencer, Dalila nous présentera les grands points de l'actualité,
  ensuite je vous parlerai des mystères du syndrome respiratoire aigu sévère.
     <Thematic>
       <sentence id="41" similarity="0.19" membership="0.31" ownership="0.33"/>
     </Thematic>
   </utter >
   <utter="3">Euh.. ensuite euh.. Didier présentera en Macédoine la coexistence entre
  soldats américains, britanniques et français.
     <Thematic>
       <sentence id="51" similarity="0.3" membership="0.67" ownership="0.35"/>
     </Thematic>
   </utter>
     ...
 </utterances>
</Turn>
```

Document

```
<Logic id="17">
 <Thematic With="Turn"/>
   <turn id="1" class="Added after levels Grouping" />
   <turn id="4" similarity="0.59" membership="0.44" ownership="0.78"/>
 </Thematic>
 <sentences>
     <sent id="51">En Macédoine, la coexistence entre soldats américains, britanniques et
     français  L'opération s'appelle " Concordia ", et la ministre de la défense (UMP),
     Michèle Alliot- Marie, n'a pas voulu manquer ça
       <Thematic>
         <utterance  id="3" similarity="0.3" membership="0.35" ownership="0.67"/>
       </Thematic>
     </sent>
     <sent id="52">En ces temps de guerre irakienne et de discorde européenne autour de
     l'engagement américain contre Bagdad, aucun signe d'entente entre alliés n'est néglige
       <Thematic/>
     </sent>
      ...
  </sentences>
</Logic>
```

Figure 4.10: Illustrative example for the validation strategy S_3

L_1, the number of shared terms between T_1 and L_1 increases, which generate a wrong similarity link. According to strategy S_2, and before removing the incoherent link between turn T_1 and logical block L_1, other parameters are checked, such as the *membership/ownership* between these segments.

In the last strategy S_3, if two descendants segments are linked, but not their ascendants (Figure 4.7.d), then their *membership/ownership* values should be checked, in order to build a link between their ascendants segments. Thus, in Figure 4.10, a new link is created between turn 1 and logical block 17, since a thematic link has been detected between utterance 3 and sentence 51.

The three validation strategies defined in this section should improve, or at least validate, the discovered thematic links. In chapter 5, all these strategies are evaluated.

After the description of the thematic alignment process, the second type of alignment between documents and speech transcript, which is the quotation alignment, is described in the following section.

4.3.1.2 Quotation Alignment

Quotation alignment can be defined as a lexical matching of term sequences, between the speech transcript and the corresponding documents discussed during an event. Whereas the thematic alignment is based on a thematic similarity of pairs of segments, the quotation alignment takes into account the lexical similarity and the order of terms, within compared segments [Mekhaldi et al. 2005]. Quotation alignment detection is deterministic and thus can be used to strengthen the thematic alignment links. However, in some cases, the lexical matching of two sequences does not mean that they are thematically similar. For instance, the following speaker utterance and document sentence are not thematically similar, even though the first one contains the quotation "lundi 31 mars" from then second one:

- "En Europe, tout d'abord, l'indice européen a été publié euh.. **lundi 31 mars**, et il a baissé de 0.6 point"
- "L'Irak a vécu, **lundi 31 mars**, sa journée de bombardements la plus intense depuis le début de la guerre"

In order to retrieve significant quotations, the minimal size of a quotation has been fixed at three terms. After removing stop words from segments, and applying

the stemming technique on the remaining words (section 3.4.2 of chapter 3), our quotations detection algorithm compares each speaker utterance with all documents sentences. The matched pairs of sequences, having at least three terms in common, are considered as quotation/quoted pairs.

However, our quotations detection algorithm is serial, i.e. it treats speech transcript segments one after the other, which has some drawbacks, such as the overlap of quotations. For example, in the following speaker utterance:

- "donc le premier article, c'est **un galion au large de Nieuport, un trésor archéologue** et historique exceptionnel sous les eaux belges"

the sequence "trésor archéologue" is shared between the two quotations that are made from the two respective document sentences:

- "**Un galion au large de Nieuport, un trésor archéologue** important"
- "Ici, le trésor archéologique et historique exceptionnel est sous les eaux belges"

which leads to an overlap between the two quotations. When two quotations q_1 and q_2 overlap (Figure 4.11), the process does not detect the second quotation q_2, as long as q_1 is not totally detected. Thus, q_2 will be reduced to q_3, i.e. to the remaining part after detecting q_1. Furthermore, q_3 might be totally neglected, if its size is inferior to the defined threshold for a significant quotation, i.e. inferior to three terms. This drawback might be resolved if the predecessor terms of the 2^{nd} detected quotation within the same utterance (e.g. the predecessor terms of q_2 in Figure 4.11) are checked, whether they constitute an extension of this quotation or not.

There are some other limitations that lead to break quotations. For example, a speaker might not pronounce a given word correctly, or he might pronounce just part of a word, before repeating it (e.g. "the same infra.. infrastructure.."). Another limitation is caused by ambiguity between some stop words and non-stop words, such as the French word "son" that means "sound", and at the same time represents the pronoun (i.e. stop word) "its". This ambiguity causes the deletion of such words from the compared sequences and thus breaks the quotations. An evaluation of this quotation alignment is presented in the next chapter.

The third type of alignment, that links documents discussed during events with speech transcript, relates to the references given by speakers on the document logical

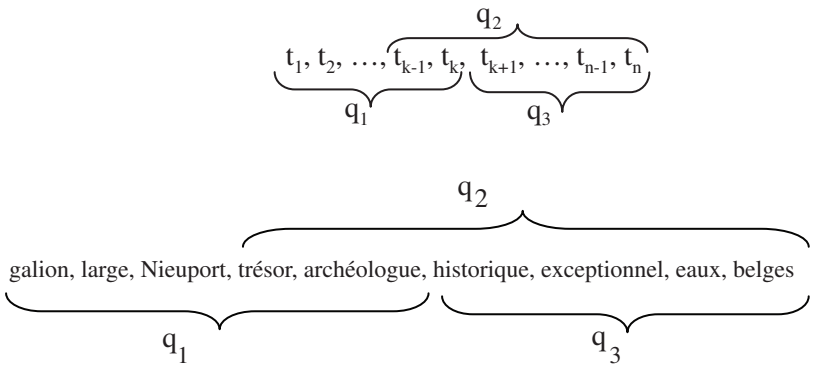


Figure 4.11: Quotations overlap

blocks. This alignment, which has been studied in collaboration with the University of Geneva [Popescu-Belis 2004b], is described in the next section.

4.3.1.3 Reference Alignment

What happens if a speaker makes a reference to a given part from the documents, using its logical label, such as "the title", "the author", etc.? In this type of relationship, the thematic alignment might not be useful. However, a specific processing is required [Popescu-Belis 2004b], in order to detect these references made on documents. A reference alignment between a speech transcript segment and a document segment can be defined as the relationship that is established between the speaker segment containing a referring expression, and the document segment being referred to. This kind of relationship is frequent in meetings, where the documents contain various articles (e.g. newspapers), in meetings dealing with various documents, or during presentations when the speaker refers to various slides, etc.

Detecting references is required in order to reinforce the links between the speech transcript and the documents discussed or projected, especially if there is no thematic similarity between their respective segments. Reference alignment might also be useful in order to know the meeting scenario, and how the various documents or

document parts (e.g. newspaper articles) are chained. However, references are still having a low frequency, comparing to the thematic links.

In order to detect the reference alignment links, the list of all pairs (referring expressions, referred entities) should be extracted [Popescu-Belis 2004b]. A referring expression is a sequence of terms in the speech transcript that refers to a document entity (e.g. "the main article", "in the first table", etc.). The referred entity might correspond to one or more documents logical blocks. The references detection algorithm has two main steps [Popescu-Belis 2004b]: detecting referring expressions within the speech transcript, and then matching them with the corresponding documents logical blocks. At the first stage, a list of patterns is created. These patterns correspond to entities names (e.g. "table", "figure", etc.), or entities description, such as the position (e.g. "the first", "the last", etc.). After that, patterns rules are applied on speakers' utterances [Popescu-Belis 2004b]. At the second stage, two categories of referring expressions are distinguished, anaphoric and non-anaphoric. If the referring expression is anaphoric (e.g. "the author", "the content", "it", etc.), then it is matched with the current document element. If the referring expression is non-anaphoric, then the *Cosine* similarity metric is used to get the most similar document logical blocks, by considering the referring expression terms, as well as the right/left context [Popescu-Belis 2004b].

After this investigation of the main alignment types that exist between the static documents and speech transcript, thematic alignment, quotations and references, another data resource is introduced into our thematic alignment process in the next section. This new resource consists in slideshows that might be presented during events, such as in scientific conferences.

4.4 Multiple Documents Alignment

The aim of the thematic alignment process is to establish thematic links between speech transcript and static documents, as well as temporal indexing of these latter. However, in some cases the speech transcript might not be available, or the Word Error Rate (WER) of its automatic transcription might be too high to be usable. In such cases, the availability of another multimedia data resource, such as slideshows that are presented by the speakers, is crucial and relevant.

Slideshows have become an omnipresent part during multimodal events, espe-

cially in scientific conferences. The consideration of slideshows in the multimodal thematic alignment process has many advantages. Slideshows might be used for pruning static documents/speech transcript thematic alignment links, by aligning them with the documents, as well as with the speech transcript. From another side, the temporal indexes of slideshows might be exploited in order to index static documents, if the speech transcript is not available. Slideshows are also useful for other tasks related to the automatic recognition of the speech, since the vocabulary of slides is frequently used by the speaker during its speech. The following paragraphs describes this variant of the thematic alignment, in which the slideshows are integrated and considered as a kind of static document, as well as a kind of multimedia document.

Before aligning the content of slideshows with the content of the other modalities, the pre-processing of the data is still required, similarly as seen in section 4.2. After the segmentation of the three resources, static documents, slideshows and speech transcript, the obtained segments are processed as usual, by removing stop words, reducing words to their stems, and then associating weights to these stems using the $TF.IDF$ metric. In the next paragraph, the thematic alignment of the three resources, static documents, slideshows and speech transcript is explained.

4.4.1 Aligning Thematically the three Resources

The thematic alignment process is applied in a pairwise manner to the different modalities, i.e. aligning static documents with slideshows, slideshows with speech transcript and then speech transcript with static documents (Figure 4.12). Document logical blocks, slides and speaker utterances have been selected respectively for this multiple documents alignment. The multiple alignments strategy, seen in section 4.3.1.1.3, has been employed, since several links might exist between each respective modality segments. For instance, a particular document logical block might be presented by two slides. Similarly, one slide might correspond to many speaker utterances, etc.

The results of the thematic alignment of the three modalities (documents/slideshows, slideshows/speech and documents/speech) might be combined. The strategies used for this aim are presented in the next section.

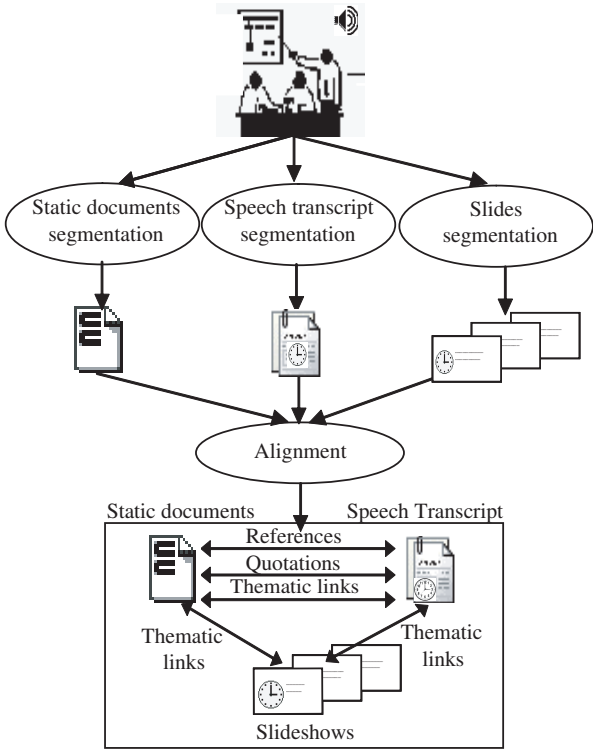
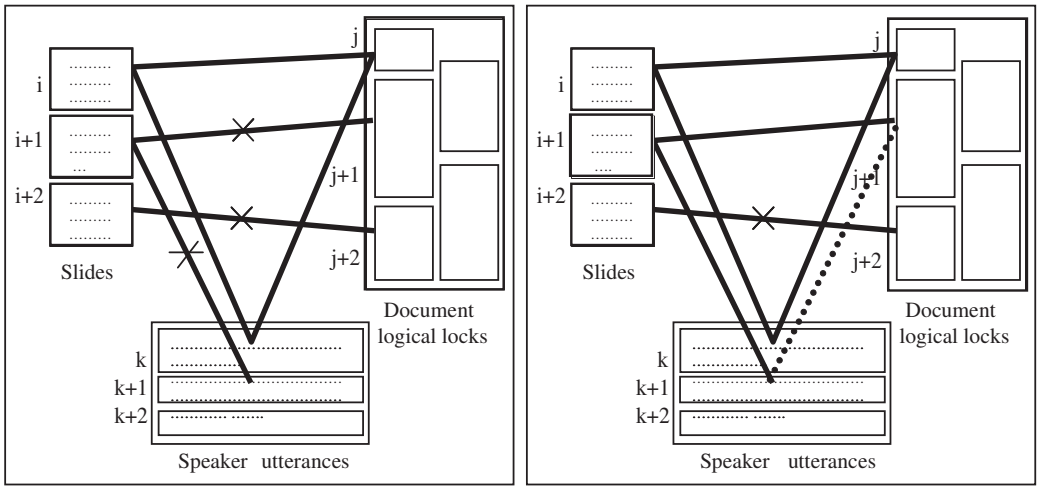


Figure 4.12: Multiple documents thematic alignment

4.4.2 Multiple Documents Alignment Grouping/Validation

Once the thematic alignment results are obtained for each combination pair (documents/slideshows, slideshows/speech and documents/speech), the various pairs can be grouped, according to the source segment for each pair. In Figure 4.13.a, the three pairs of alignment (i, j), (j, k) and (i, k) are grouped, which generates a cycle composed of three arcs (i, j, k). This grouping of the thematic alignment pairs might be exploited in order to validate the detected links, as well as to add the missing links, within each alignment pair. Based on the cycle structure, we have defined two validation strategies. In the first strategy, only the links that construct a complete cycle with three arcs are preserved (e.g. (i, j), (j, k) and (i, k) Figure 4.13.a). The other links, such as $(i + 1, j + 1)$, are removed. In the second strategy (Figure 4.13.b), and in addition to the detected cycles, all paths with two arcs are preserved and then completed with the missing arc, in order to accomplish a cycle.



a. Strategy 1: keep only cycles

b. Strategy 2: keep cycles and complete paths having two arcs

Figure 4.13: Strategies for grouping multiple documents alignment

Thus, in Figure 4.13.b a link is added between the document logical block $j+1$ and the speaker utterance $k+1$, in order to generate a cycle $(i+1, j+1, k+1)$.

The two grouping strategies defined have been used in the multiple documents thematic alignment, as being a way for exploiting the rich information available, when a third data resource is considered. An evaluation of these strategies is presented in chapter 6.

4.5 Conclusion

I this chapter, the methodology of our alignment framework has been introduced, including the three alignment types, thematic, quotations and references, as well as the various techniques and strategies involved. In the two next chapters, two case studies are presented, respectively press review meetings and scientific conference presentations. In these two case studies, the various alignment techniques and methods are applied and evaluated.

Chapter 5

Case Study 1: Press Review Meetings

5.1 Introduction

The various techniques, methods and strategies that are involved in the multimodal alignment framework of static documents with the speech transcript, as well as with slideshows, have been presented and described in the previous chapters. In order to measure the performance of our framework, two main multimodal applications have been chosen, press review meetings to perform static documents/speech alignment, and scientific conference presentations to perform the multiple documents/speech thematic alignment. The first corpus is studied in the current chapter, whereas the second one is evaluated in chapter 6.

5.2 Press Review Meeting Corpus

Among various kinds of meetings (administrative meetings, interviews, etc.), the press review meetings have been chosen for our evaluation, as being a multimodal event, where static documents are present and used during almost the entire meeting (Figure 5.1).

In our Smart meeting room in Fribourg [SMR], 22 press review meetings have been recorded. This document-centric meeting room is equipped by a video projector, microphones and cameras. Documents that are either discussed or projected are captured. The recorded meetings are archived in a media file server, in the form of directories. In each meeting directory, and besides the audio and video record-

Figure 5.1: Press review meeting

ings, captured documents are available in the form of PDF files, linear text version and XML logical structure. The speech is available in the form of an XML manual transcription, in addition to other speech annotations.

During the 22 recorded meetings, of a total duration of 4 hours, several French newspapers have been used, Le Monde (France), Le Devoir (Canada), Le Soir (Belgium) and La Presse (Tunisia), where the main articles are discussed. In 18 meetings, only one document is presented. In the four remaining meetings, 2, 3, 4 and 4 documents are presented respectively. During the 22 meetings, several participants discuss the various newspaper articles. The average number of participants is 4 per meeting.

Since each document is composed of many thematically heterogeneous articles, two main scenarios have been defined, a stereotyped and a non-stereotyped scenario.

Static documents			Speech transcript		
Number	Logical blocks	Sentences	Duration	Turns	Utterances
29	379	2772	4h	1888	2936

Table 5.1: Statistic of the 22 meetings

In the stereotyped meetings, each participant presents an entire article, without any interruption by the other participants. In the non-stereotyped meetings, there is more interactivity between participants, which means that several speakers are involved in presenting the same article. The interactivity between speakers consists in asking questions, commenting the content, etc. In order to compare the performance of stereotyped and non-stereotyped meetings, as well as their effect on the thematic alignment, the two scenarios, stereotyped and non-stereotyped, have been followed by the participants.

Before performing the alignment process on the data, it is crucial to segment the static documents and speech transcript, according to the methods described in section 4.2 of chapter 4. The results of the segmentation of the 22 meetings, as well as the ratio of segments according to various criteria (the type of scenario and the number of documents per meeting), are presented in Table 5.1, 5.2 and 5.3 . As seen in section 4.2.1 of chapter 4, the documents are first manually segmented into logical blocks that correspond to the different articles of the newspapers, and then automatically segmented into sentences. Using the transcriber tool [Barras et al. 1998], the speech is segmented into speaker turns that contain utterances, then this structure is decomposed into two structures, speaker turns and speaker utterances, respectively.

After the segmentation step, our alignment processes, thematic and quotations, are independently applied on the data. For both alignment processes, the source file and the target file (source segment and target segment, respectively) should be distinguished. For each process (thematic and quotations), the document segments and the speech transcript segments are taken as input, then two alignment files are generated, documents/speech transcript alignment and vice versa. The two alignment files contain the details of the detected links, such as the identifiers of similar segments and the similarity values in the case of thematic alignment (Figure 5.2), or the identifiers of the quoted document sentences in the case of quotation alignment (Figure 5.6). In the next section, our two alignment processes are respectively evaluated.

Stereotyped (9)				Non-stereotyped (13)			
Static documents		Speech transcript		Static documents		Speech transcript	
Logical blocks	Sentences	Turns	Utterances	Logical blocks	Sentences	Turns	Utterances
198	1375	234	617	181	1397	1954	2319

Table 5.2: Statistic according to the meetings scenario

Mono-document (18)				Multi-documents (4)			
Static documents		Speech transcript		Static documents		Speech transcript	
Logical blocks	Sentences	Turns	Utterances	Logical blocks	Sentences	Turns	Utterances
265	1639	1539	2354	114	1133	349	582

Table 5.3: Statistic according to the number of static documents per meeting

5.3 Evaluation

The Evaluation of our thematic and quotation alignment processes consists in comparing the generated links of each alignment type, with a corresponding manual ground truth that contains all possible alignment links. Within manual ground truth, the null alignments, as well as segments containing only stop words, are neglected. A binary score is attributed to each detected alignment link, "1"' if it is correct, i.e. it exists in the manual ground truth, and "0" otherwise.

In order to measure and evaluate the performance and quality of our thematic and quotation alignments, *recall*, *precision* and the efficiency measure F are employed. If we consider *auto* as the set of detected links by our algorithm, and *manual* as the set of links of the manual ground truth, then the evaluation measures are computed as follow:
recall = number of correct links detected / number of links that should be detected
$recall = |auto \cap manual|/|manual|$
precision = number of correct links detected / number of links detected
$precision = |auto \cap manual|/|auto|$
$F = 2\ .\ (precision\ .\ recall)\ /\ (precision + recall)$

In the next section, the experiments that have been carried out, starting by the thematic alignment, are described.

5.3.1 Thematic Alignment Results and Analysis

After segmenting documents and speech transcript, the various segments are processed, in order to remove stop words, reduce words to their stems, and attribute to them importance weights using the $TF.IDF$ metric. Among the two $TF.IDF$ formulas, seen in section 4.2.5 of chapter 4, the first formula has generated best scores ($TF.IDF_{t,S_i} = occ_{t,S_i} \cdot \sum_{j=0}^{N} occ_{t,S_j}$). Thus, it has been adopted for all our experiments. Using the various similarity metrics seen in section 4.3.1.1.1 in chapter 4 (*Cosine*, *Dice* and *Jaccard*), the similarity is computed between document segments and speech transcript segments, and vice versa. WordNet thesaurus [WordNet] is not considered, since the current corpus is French, and WordNet is only available in English.

Within the thematic alignment process, document segments and speech transcript segments are taken in the form of 2 XML files, and comparison is performed in both directions, then the input files are enriched by thematic alignment links. Thus, each XML element representing a segment within the document (speech transcript respectively), is enriched by new information indicating the identifiers of the similar speech segments (document segments respectively), and the value of this similarity. An example of the thematic alignment between the speech transcript and the static documents, for both directions, is shown in Figure 5.2. The number of target similar segments that are selected for each source segment depends on the chosen strategy (section 4.3.1.1.3 of chapter 4), if 1-best or multiple alignment strategy. In the next sections, the two strategies of thematic alignment are evaluated, using the various similarity metrics.

5.3.1.1 1-best Thematic Alignment

In this thematic alignment strategy (Figure 4.5.a), for each source segment, only the link with the highest similarity value is considered. The generated results in both directions, from documents to speech transcript and vice versa, are not symmetric. For this reason, the two directions should be evaluated independently.

It might happen frequently during a meeting that some terms appear in various documents articles, even though the topic is not the same. In such cases, the alignment process, which is based on terms co-occurrence, might be influenced by the false alignment links that are generated. Therefore, the $TF.IDF$ metric has been integrated into the similarity metrics, in order to associate weights to the different terms representing their discriminancy, according to terms frequency inside and

Speech transcript

<Utterance id="11"> Voilà. Euh.. la troisième semaine du procès Elf vient de commencer à Paris.
<Thematic>
<sentence id="44" similarity="0.16" membership="0.67" ownership="0.17"/>
</Thematic >
</Utterance>
<Utterance id="12"> Euh.. durant les deux premières semaines **le système de défense des principaux prévenus** ont été.. ont été profondément déstabilisé.
<Thematic>
<sentence id="45" similarity="0.30" membership="0.88" ownership="0.32"/>
</Thematic>
</Utterance>

Document

<Sentence id="44">
Justice : les surprises du procès Elf La justice marque des points au procès Elf, dont la troisième semaine devait s'ouvrir, lundi 31 mars, devant le tribunal correctionnel de Paris.
<Thematic>
<utterance id="11" similarity="0.16" membership="0.17" ownership="0.67"/>
</Thematic>
</Sentence>
<Sentence id="45">Les deux premières semaines d'audience ont profondément déstabilisé les **systèmes de défense des principaux prévenus** : M Le Floch-Prigent, mais aussi Alfred Sirven, l'ancien directeur des affaires générales, et André Tarallo, l'ex-"M Afrique" d'Elf.
<Thematic>
<utterance id="12" similarity="0.30" membership="0.32" ownership="0.88"/>
</Thematic>
</Sentence>

Figure 5.2: Example of thematic alignment between static documents and speech transcript

outside the segments. The structure combinations chosen for the 1-best alignment strategy are:

- Document sentences/speaker utterances, document sentences/speaker turns, document logical blocks/speaker turns.

- Speaker utterances/document sentences, speaker utterances/document logical blocks, speaker turns/document logical blocks.

The three similarity metrics (*Cosine*, *Jaccard*, *Dice*) have been used, without and then with the consideration of the $TF.IDF$ metric. Then, for each alignment pair, the confidence interval for the metric generating the best F value has been defined, in order to measure the statistical significance of the results. The confidence

interval is computed according to the following formula:

$$CI_{(1-\alpha)} = F \pm Z_{\alpha} \cdot \sqrt{F(1-F)/n}$$

Where F is the retrieved F value for a given metric, $(1-\alpha)$ is the confidence level fixed to 95%, Z_{α} is a constant fixed to 1.96, and n is the number of alignment links of the manual ground truth, used for the evaluation.

The results obtained, using the various structures of documents and speech transcript, are illustrated in Figure 5.3. As shown, the 1-best alignment of speaker utterances with document sentences (Figure 5.3.a) has provided the best score with the $Jaccard$ metric, where the F value is 0.68, with a confidence interval of 0.68 ± 0.03. The score obtained using $Cosine$ and $Cosine + TF.IDF$ metrics are included in this confidence interval, where the respective F values are 0.67 and 0.65. For the sentences/utterances alignment (Figure 5.3.b), the F value obtained is the same, i.e. 0.62, whether $Jaccard$ or $Dice$ metrics are being used, with a confidence interval of 0.62 ± 0.03. Nevertheless, the $Jaccard$ metric is still generally the most appropriate metric for aligning utterances with sentences, and vice versa.

The 1-best alignment of speaker turns with document logical blocks, using $Cosine$ metric (Figure 5.3.c), has generated the higher score with F value equals 0.62, where the confidence interval is 0.62 ± 0.03 that includes the $Cosine + TF.IDF$ score (0.60). These two scores (0.62 and 0.60) are drastically better than those obtained with other metrics ($Dice = 0.47$, $Dice + TF.IDF = 0.42$, $Jaccard = 0.47$, $Jaccard + TF.IDF = 0.37$). In the opposite direction, i.e. the logical blocks/turns alignment (Figure 5.3.d), $Jaccard$ metric is the most appropriate, generating F value of 0.60 with a confidence interval of 0.60 ± 0.07, including $Jaccard + TF.IDF$, $Dice$ and $Dice + TF.IDF$ values.

When aligning speaker utterances with document logical blocks (Figure 5.3.e), the best F value is obtained by using the $Cosine + TF.IDF$ metric (0.65), where the confidence interval is 0.65 ± 0.03, with a *precision* of 0.70. However, when aligning document sentences with speaker turns, as shown in Figure 5.3.f, $Dice + TF.IDF$ metric has generated the best score, where $F = 0.57$, and a confidence interval of 0.57 ± 0.05 that includes the $Jaccard + TF.IDF$ value (0.53).

Through these results, it is clear that the 1-best alignment strategy is not symmetric, since the scores obtained from the documents to the speech transcript are not the same as those obtained in the opposite direction. Moreover, the scores gen-

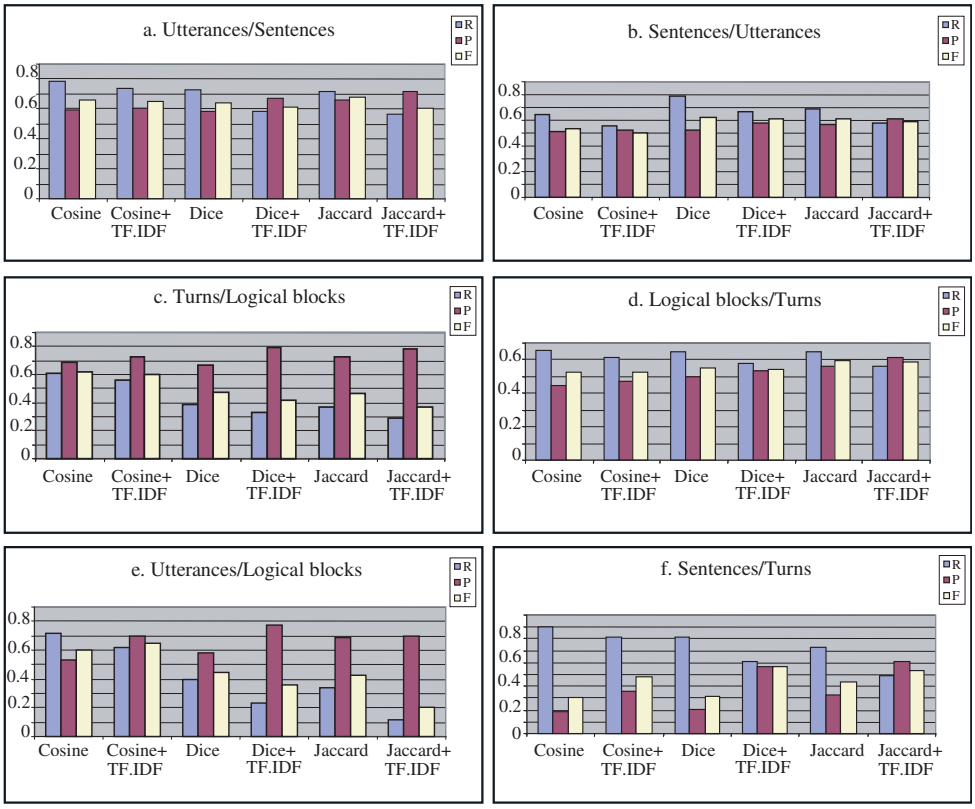


Figure 5.3: 1-best thematic alignment strategy results

erally increased when the $TF.IDF$ metric is integrated into the similarity metrics in general increases the scores.

When speaker segments (turns and utterances) are aligned with document segments (logical blocks and sentences), the scores are more or less better when the $Cosine$ or $Cosine{+}TF.IDF$ metrics are used. However, aligning document segments with speaker segments generates best scores either with $Jaccard$, $Jaccard{+}TF.IDF$, $Dice$ or $Dice + TF.IDF$ metrics.

There is a constraint that has affected the 1-best thematic alignment strategy results, which is the likeness of some articles content within a document, or within various documents in case of multi-document meetings. In that case, the alignment of one source segment, e.g. a speaker utterance, with a particular document sentence or logical block might be imperfect, which constitutes one of the limitations of the

1-best alignment strategy. This limitation is partially solved by using the multiple alignments strategy (section 5.3.1.2).

When aligning large segments with smaller ones, e.g. speaker turns with document sentences or logical blocks with utterances, the alignments scores are being drastically affected. The decrease of scores is mainly due to the important difference in the size of the compared segments, since a speaker turn corresponds normally not only to one but to several document sentences, especially in stereotyped meetings, where the same speaker presents an entire article. Similarly, a logical block, which corresponds to a newspaper article in this case study, should be aligned with many speaker utterances. Once again, we conclude that the 1-best alignment strategy is not the most appropriate, but a good application to validate our method. The multiple thematic alignments strategy, that is evaluated in the next section, should resolve the problem, where each source segment is matched with all the relevant target segments.

5.3.1.2 Multiple Thematic Alignments

In this thematic alignment strategy, many alignment links might be considered for each source segment, provided that they exceed a defined similarity threshold. If the similarity of a detected thematic link overcomes this threshold (fixed to 0.10), this link is considered as significant, otherwise it is removed. This way, the same thematic links are obtained in both directions, from documents to speech transcript and vice versa, which means that the multiple thematic alignments strategy is symmetric.

Due to the multiplicity of thematic links and the subjectivity of the manual ground truth alignments, only some combinations of structures have been evaluated within this strategy. Thus, turns/logical blocks, turns/sentences and logical blocks/utterances pairs have been considered, since their manual ground truth alignments are similar, from one human annotator to another.

The results of the evaluation of this thematic alignment strategy are shown in Figure 5.4. Comparing to other metrics, the $Cosine{+}TF.IDF$ metric has generated the best score for the turns/logical blocks alignment, where $F = 0.59$ (Figure 5.4.a), and a confidence interval of 0.59 ± 0.03 that contains the $Cosine$ value (0.56). The score obtained with $Cosine + TF.IDF$ metric equals the one of the logical blocks/turns alignment, using the 1-best alignment strategy. Using the other metrics ($Dice$, $Dice{+}TF.IDF$, $Jaccard$ and $Jaccard{+}TF.IDF$, respectively), the retrieved

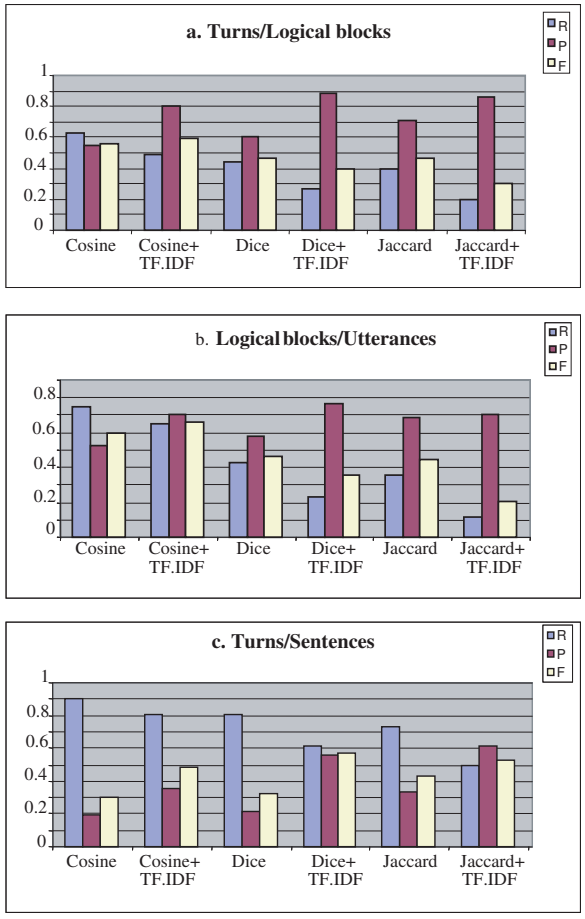


Figure 5.4: Multiple thematic alignments strategy results

alignment links are not interesting, which explains the low values of *recall* (0.44, 0.27, 0.40 and 0.20, respectively), and the low values of F (0.46, 0.39, 0.46 and 0.30 respectively), even though the *precision* values are high (0.61, 0.88, 0.71 and 0.86, respectively), which means that very few alignment links have been retrieved.

The multiple alignments of logical blocks with speaker utterances (Figure 5.4.c) is the same as in the opposite direction, i.e. the multiple alignments of utterances with logical blocks, due to the symmetry of this alignment strategy. From another hand, the generated scores are more or less the same as the ones of the utterances/logical blocks using 1-best alignment strategy (Figure 5.3.e). In this case, the best scores are obtained using $Cosine+TF.IDF$ metric, where $F = 0.66$, with a confidence interval

of 0.66 ± 0.03. The similarity of scores between logical blocks/utterances in the multiple alignments strategy and utterances/logical blocks in the 1-best alignment strategy makes sense, since one logical block might be aligned with many utterances (multiple), whereas each one of these utterances is likely to be aligned exclusively with this logical block (1-best). The small difference between the *recall* values (0.65 in Figure 5.4.c and 0.62 in Figure 5.3.e) is explained by the fact that there are some utterances that are linked with more than one logical block in the multiple alignments strategy (Figure 5.4.c). However, these links might be secondary links, and therefore might be neglected in the 1-best alignment (Figure 5.3.e), resulting in a lower *recall* value, compared to the multiple alignments strategy.

On the other hand, the turns/sentences multiple alignments (Figure 5.4.c), which is the same as sentences/turns multiple alignments (symmetry of links), has generated the best scores using $Dice+TF.IDF$ metric, where $F = 0.57$, and a confidence interval of 0.57 ± 0.05 including the $Jaccard+TF.IDF$ value (0.53). Similar to the previous case, the scores of the turns/sentences multiple alignments are more or less the same as the ones of the 1-best alignment, obtained in the other direction, i.e. sentences/turns (Figure 5.3.f), with a *recall* and *precision* sharing the same values (0.61 and 0.56 respectively).

Except the turns/sentences alignment, the $Cosine + TF.IDF$ metric has generated the best scores for the alignment pairs in this multiple alignments strategy. The consideration of the $TF.IDF$ metric is still improving the scores, similarly as seen in the 1-best alignment strategy.

Besides covering all the significant thematic links for a given source segment, the multiple alignments strategy has other advantages. This strategy might help in discovering the thematic structure of meetings, as it will be presented in chapter 7, and might help in pruning and correcting the generated alignment links. The latter aspect is detailed in the next section.

5.3.1.3 Multiple Alignments Grouping/Validation

As described in section 4.3.1.1.5 of chapter 4, two hierarchical alignment levels (turns/logical blocks and utterances/sentences) can be superposed, and their coherence can be checked, in order to validate, correct and prune the generated alignment links, especially in the ascendant level, i.e. turns/logical blocks level. Therefore, three validation strategies, S_1, S_2 and S_3, have been defined in section 4.3.1.1.5

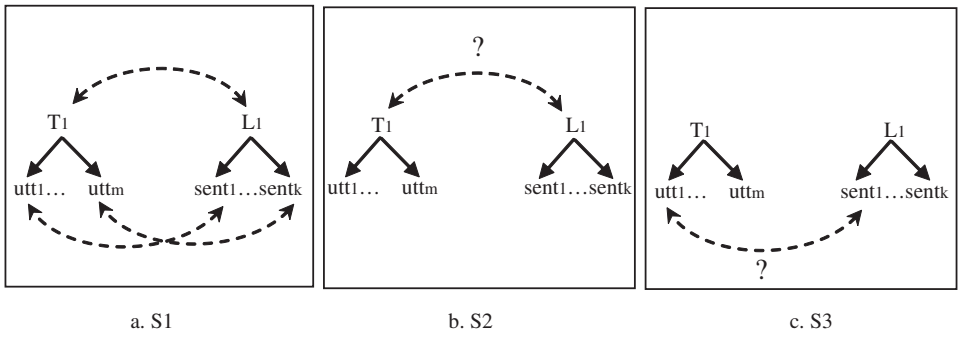


Figure 5.5: Thematic alignment levels and strategies for links validation

in chapter 4 (Figure 5.5), and have been applied successively on the alignment results of the 22 meetings. The selected alignment pairs that have been used for this validation process are the turns/logical blocks with the *Cosine* metric, and the utterances/sentences with the *Jaccard* metric, without then with the consideration of the $TF.IDF$ metric. The scores obtained for this grouping/validation process are presented in the following sections.

5.3.1.3.1 Using Alignment Results without TF.IDF: in this section, the thematic alignment results that are used have been obtained without the integration of the $TF.IDF$ metric. The first strategy S_1 for the correction of incoherent thematic links (Figure 5.5.a) consists in removing all the thematic links that are established between turns and logical blocks, which are not validated by their descendants, meaning that there is no thematic link between their descendant segments .

The initial average values for *recall*, *precision* and F, with the $TF.IDF$ not being integrated into the similarity metrics, are 0.63, 0.55 and 0.55 respectively. After applying the strategy S_1, those same values have decreased to 0.2, 0.22 and 0.2 respectively, which can be explained by the deletion of many correct links from turns, especially those composed of only one utterance. Given a particular turn T_i one of these mono-utterance turns, T_i might be aligned with a similar logical block L_j, but its single utterance might not be aligned with any sentence from L_j. According to the strategy S_1, the link between T_i and L_j should be removed, ignoring its correctness. A use case of this strategy is shown in Figure 4.8 in chapter 4.

In order to avoid the removal of correct links, another strategy S_2 has been

defined (Figure 5.5.b), where the *membership* value of the turn T_i in its similar logical block L_j is being taken into account, before removing the link between them (i.e. the ratio of terms of T_i that are present in L_j). A use case of this strategy is shown in Figure 4.9 in chapter 4. If the value of the *membership* function of T_i in L_j overcomes a defined threshold Th_1, then the link between T_i and L_j is preserved, otherwise, it is removed. In order to observe the effect of the threshold on the alignment results, Th_1 was set at different values ranging from 0% to 100%. We have realized that its ideal value is 20%, where the F metric reaches its maximal value, increasing from 0.56 to 0.58, whereas the *recall* decreases from 0.63 to 0.56, and the *precision* increases from 0.55 to 0.65. The improvement of the *precision* indicates that many false links are being removed, whereas almost all the correct links are being preserved.

Finally, our interest was the insertion of new thematic links between turns and logical blocks, by applying another strategy S_3 (Figure 5.5.c). In Figure 4.10 in chapter 4, an example of use of this strategy is illustrated. According to this strategy, new links can be added between turns and logical blocks that are not linked, by taking into account the weight of the links between their descendant segments. For this reason, the *membership* function in the utterances/sentences level has been considered. If the *membership* value of an utterance utt_i in a similar sentence $sent_j$ overcomes a threshold Th_2, then a new link is established between their parents. The second threshold Th_2 was set at different values ranging from 0% to 100%, while Th_1 was fixed at 20% (the ideal value), in order to keep the correct links preserved by the strategy S_2. When $Th_2 = 0.75$, the F value increases from 0.58 to its maximal value 0.59, the *recall* value increases from 0.56 to 0.61 and the *precision* value decreased from 0.65 to 0.61 (Table 5.4).

In addition to the improvement of the scores after the grouping/validation process with its three strategies, the final F value obtained (0.59) is the same as the F value obtained for the turns/logical blocks alignment, with the integration of the $TF.IDF$ within the similarity metrics.

5.3.1.3.2 Using Alignment Results with TF.IDF: after applying the three grouping/validation strategies on the alignment results obtained by integrating the $TF.IDF$ into the similarity metrics, no strategy succeeded in increasing the F value that remained stable at 0.59. This can be explained by the performance and the usefulness of the $TF.IDF$ metric to cover most significant thematic links.

	Initial	S_1	S_2	S_3	Alignment types merging
Recall	0.63	0.20	0.56	0.61	0.69
Precision	0.55	0.22	0.65	0.61	0.58
F	0.56	0.20	0.58	0.59	0.59

Table 5.4: Effect of alignment levels grouping strategies, and alignment types merging on the thematic links (alignment results without TF.IDF)

After the evaluation of the thematic alignment process, with its various strategies, the quotation alignment and the reference alignment are evaluated in the next sections.

5.3.2 Quotations Alignment: Results and Analysis

Our algorithm that detects the quotations between speech transcript and documents is based on a lexicographic matching of terms of their respective sequences (section 4.3.1.2 of chapter 4). The size of a significant quotation should be at least three terms.

This oriented alignment has been applied between speaker utterances and document sentences of the 22 meetings. The quotation alignment process compares each speaker utterance with the overall documents sentences. When a quotation is detected between a speaker utterance and a document sentence, the quotation sequence is highlighted inside the utterance, in the form of an XML element, and the identifier of the quoted document sentence, as well as the name of the document, are stored. An example of a quotation output file and the corresponding document file are shown in Figure 5.6.

Once the quotation detection algorithm is applied, the obtained values for *recall*, *precision* and F were 0.95 (Figure 5.7). All the quotations that have been missed out by our quotations algorithm are due either to an ambiguity between some terms and stop words within speaker utterances, or to the noisy speaker terms (e.g. repeating terms, imperfect pronunciation, etc.), or to the overlap of quotations, as seen in section 4.3.1.2 of chapter 4. The scores obtained varied from one meeting to another. For instance, for the meetings 4, 7, 17 and 20, the F values obtained were low, comparing to the other meetings (0.88, 0.91, 0.86 and 0.91, respectively). However the average scores are still very satisfactory.

Speech transcript

<Utterance id="1">Bonjour à tous les deux. Encore une fois nous allons voir la une du Monde de ce mercredi 2 avril 2003.
</Utterance>
<Utterance id="2">Pour commencer, Dalila nous présentera les grands points de l'actualité, ensuite je vous parlerai des mystères <quotation id="1">**syndrome respiratoire aigu sévère**</quotation>.
</Utterance >
<Utterance id="3">Euh.. ensuite euh.. Didier présentera <quotation id="2"> **en Macédoine la coexistence entre soldats américains, britanniques et français**.</quotation>
</Utterance>
…

<Quotations>
<quotation id="1" utter-id="2" sent-id="41" doc="file1.xml"/>
<quotation id="2" utter-id="3" sent-id="51" doc="file1.xml"/>
…
</Quotations>

Document

…
<Sentence id="41" >La propagation du SRAS**, le syndrome respiratoire aigu sévère** ou pneumonie atypique, a pris, en quelques semaines, une dimension planétaire
</Sentence>
<Sentence id="51">**En Macédoine, la coexistence entre soldats américains, britanniques et français.** L'opération s'appelle " Concordia ", et la ministre de la défense (UMP), Michèle Alliot- Marie, n'a pas voulu manquer ça
</Sentence>
…

Figure 5.6: Example of quotation alignment between static documents and speech transcript

5.3.3 Reference Alignment Results

The references detection process has been made in collaboration with the University of Geneva [Popescu-Belis 2004b]. This oriented alignment has been tested on the 22 meetings, more precisely between speaker utterances and document logical blocks, where 437 referring expressions have been detected. The evaluation of the algorithm detecting the references has focused on two aspects, first the detection of referred documents, and second the detection of the referred document elements. The accuracy of retrieved documents was 0.60, whereas the accuracy of the retrieved document elements was 0.32. An extract from a reference alignment file is presented in Figure 5.8, where the referring expressions, represented by $< er >$ XML element, are included inside the corresponding speaker utterances. The reference links are

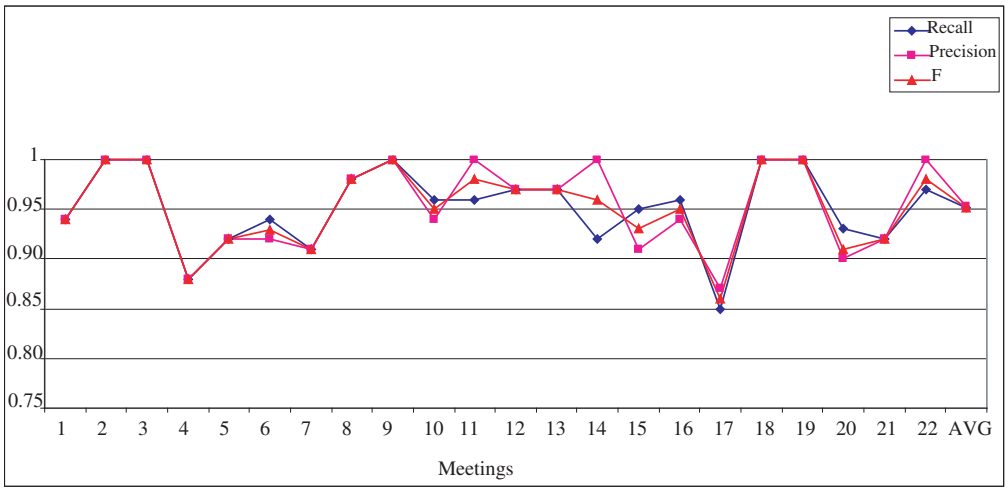


Figure 5.7: Quotation alignment evaluation

Speech transcript

```
…
<Utterance id="13">
  Here is the article about <er id=”10”> “les radios généralistes” </er>, but there is
  nothing important to say.. Let’s go to the <er id=”11”> last article, “une apocalypse
  aveugle” </er>
<Utterance>
<Utterance id="14">
  so the content of <er id=”12”> this article </er> is about …
</Utterance>
…
<References> …
 <ref id="10" utter-id="13" logicalBlock-id="5" doc="file.xml"/>
 <ref id="11" utter-id="13" logicalBlock-id="6" doc="file.xml"/>
 <ref id="12" utter-id="14" logicalBlock-id="6" doc="file.xml"/>..
</ References >
```

Figure 5.8: Extract from a reference alignment file

represented by $< ref >$ elements, containing the identifiers of the referring utterance $utter - id$, the identifiers of the referred document element $logicalBlock - id$ and the referred document name *doc*.

After the evaluation of each of the three processes, thematic, quotation and reference alignments, the merging of the results of these alignment types is presented in the following section.

5.3.4 Merging all Alignment Types

When studying the relationship between static documents and the speech transcript, all the alignment types between the two resources, i.e. thematic, quotations and references, have to be covered (see Figure 5.9). In this section, our interest is focusing on the grouping and merging of all the identified alignment types [Mekhaldi et al. 2005]. The merging of alignments types is crucial, in order to measure the complementary of links of various types. The merging might also help in the validation and correction of each type of links, taking into account the other types. Only the speech transcript is concerned by this merging process, since it is the source modality for the quotation and reference alignments.

After running the three alignment processes independently (thematic, quotation and reference alignments), our merging process combines all these links according to their source unit, if it is a speaker turn or an utterance. An extract from the final structure obtained after the merging process is shown in Figure 5.9. In some cases, the quotation alignment might be considered to be a special kind of thematic alignment, thus the two alignment types should be coherent. When such constraint is considered, i.e. correcting the thematic links by taking into account the quotation links, the F value (for turns/logical blocks multiple alignments) remains at 0.59, as shown in Table 5.4, for both alignment variants, without and with the $TF.IDF$ integration. This result proves that in our corpus, all the detected quotations correspond more or less to the detected thematic links.

On the other hand, the reference links do not have to correspond necessarily to the thematic links. In Figure 5.8, utterances 13 and 14 are linked thematically with sentences of the 6^{th} logical block, even though there is a reference in utterance 13 to the 5^{th} logical block. This is why no constraint is made between the thematic and the reference alignments. Thus, merging the reference alignment with the other alignment types does not change the scores. Nevertheless, the references may add other

information to the user, such as, how the speakers chained the various document articles, or the various documents in the case of multi-documents meetings.

After merging the results of the three alignment types, thematic, quotation and reference alignments, the final *recall*, *precision* and F values are 0.69, 0.58 and 0.59 respectively, for the alignment variant without $TF.IDF$ (Tab54), and 0.55, 0.72 and 0.59 respectively, for the alignment variant with $TF.IDF$.

In order to visualize the framework generated by the merging process of the various alignment types, an SVG tool has been implemented (Figure 5.10). Within this tool, the speech transcript and the documents are represented by the X and Y axis, respectively. The thematic links are represented by circles at the intersection of the corresponding utterances and sentences, and the quotations by diamonds. The references are represented by rectangles, where the height depends on the size of the referred logical block, in terms of the number of sentences.

Even though the merging of the three alignment types does not improve the scores of the thematic alignment, it is still useful in order to highlight the complementarity between the three alignment types. For instance in Figure 5.10, the 6^{th} speaker utterance is thematically similar to the 2^{nd} and the 3^{rd} document sentences, and it contains two respective quotations from them. Moreover, it contains a reference link to the document logical block that contains these two sentences.

In order to evaluate the usefulness of our multimodal alignment, as well as to measure the efficiency of static documents role in the browsing and navigation interfaces within multimedia archives, a user evaluation [Lalanne et al. 2004a] has been performed. This evaluation is described in the next section.

5.4 Assessment: User Evaluation of the Multimodal Alignment

The work performed in this task consists in a user evaluation of the thematic alignment of static documents with the speech transcript [Lalanne et al. 2004a]. Therefore, 8 users have been involved in this evaluation, for a period of about 15 to 20 minutes. The task consists in answering 8 questions, using the meeting browser FriDoc (Figure 5.11) [Lalanne et al. 2004a], where 4 questions are monomodal, using either the document or the speech, and 4 are multimodal using both resources. Depending

Speech transcript

```
<Turn id="1">
  <Thematic with="logic">
<logical_block id="22" similarity="0.15"/>
    …
  </Thematic>
  <utterances>
       <utter id="1">
 Didier will discuss the first article which is about “surprises du procès Elf”, then….
          <Thematic with=”sentences”>
               <sent-id="85" similarity="0.16"/>
                    …
          </Thematic>
          <Quotations>
               <quotation id="1" utter-id="2" sent-id="67" doc=" file.xml"/>     …
          </Quotations>
          <References>
              <er id="2" logicalBlock-id="1" doc="file.xml"> first article </er>
                   …
          </References>
       </utter >
                   …
   </utterances>
    …
</Turn >..
```

Figure 5.9: Final structure after alignment types merging

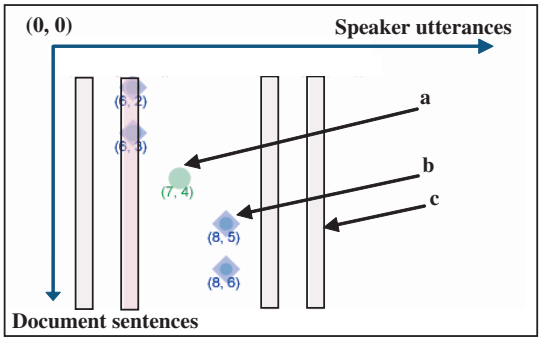


Figure 5.10: Merging various types of document/speech transcript alignment: a) a thematic link; b) a quotation and a thematic link; c) a reference link

Figure 5.11: User evaluation interface

on the exploitation of the documents alignment or not, two meeting browsers B_1 and B_2 have been implemented. B_1 benefits from document/speech alignments and B_2 is exactly the same meeting browser without the document/speech alignments enabled. Two main protocols were then defined according to the type of browsers (B_1/B_2, B_2/B_1), and balanced with two meetings (M_1/M_2).

When the documents alignment was utilized, the accuracy of the answers was 76%, where 70% of the multimodal questions were answered. When the documents alignment is not utilized, the accuracy of the answers was 66%, where 50% of the multimodal questions were solved. These results stress the importance of the static documents alignment role within multimedia archive interfaces, as well as the importance of considering all modalities that are available in such browsing tasks, instead of only one modality [Lalanne et al. 2004a].

5.5 Conclusion

The main objective of this chapter was to present the evaluation of the various alignment processes between static documents and the speech transcript of meeting dialogs. The satisfactory results that have been obtained reflect the performance of our alignment framework. The superposition of the thematic alignment levels has generated two new hierarchical structures, representing the complete information about the thematic alignments. Furthermore, the grouping of levels of the thematic alignment, turns/logical blocks and utterances/sentences, has reinforced the obtained results. The merging of the various alignment types (thematic, quotations and references) has shown that these types are complementary, and their consideration has covered all the links that might exist between static documents and speech transcript, even though the merging process does not increased the results of the thematic alignment.

In order to test our alignment framework on other corpuses, where several documents from different types are available, another case study is presented in the next chapter. This corpus consists in scientific conference presentations that provide various data, static documents, slideshows and speech recording.

Chapter 6

Case Study 2: Scientific Conference Presentations

6.1 Introduction

After presenting the press review meetings evaluation in the previous chapter, our thematic alignment process is tested in the context of another type of multimodal event, in the current chapter. An English scientific conference has been chosen as application, in order to assess the performance of the synchronization of slideshows with static documents (Figure 6.1). In this application, both slideshows and static documents (scientific articles in this application) are available. Moreover, the availability of the speech recording of the various presentations might be exploited in order to compare the performance of the slideshows and the speech as a mean for synchronizing static documents. Furthermore, we will see that the slideshows and the speech transcript might be also exploited together in order to improve this synchronization.

6.2 Conference Presentations Corpus

Our thematic alignment has been tested in the context of SMAC project (Smart Multimedia Archive for Conferences), where the material of CHEP'04, a scientific conference in the field of physic of particles, has been considered [Lalanne et al. 2005] [Abou Khaled et al. 2006]. Within this conference, 8 scientific presentations have been selected, in which the three resources are available (static documents, slideshows and speech recording).

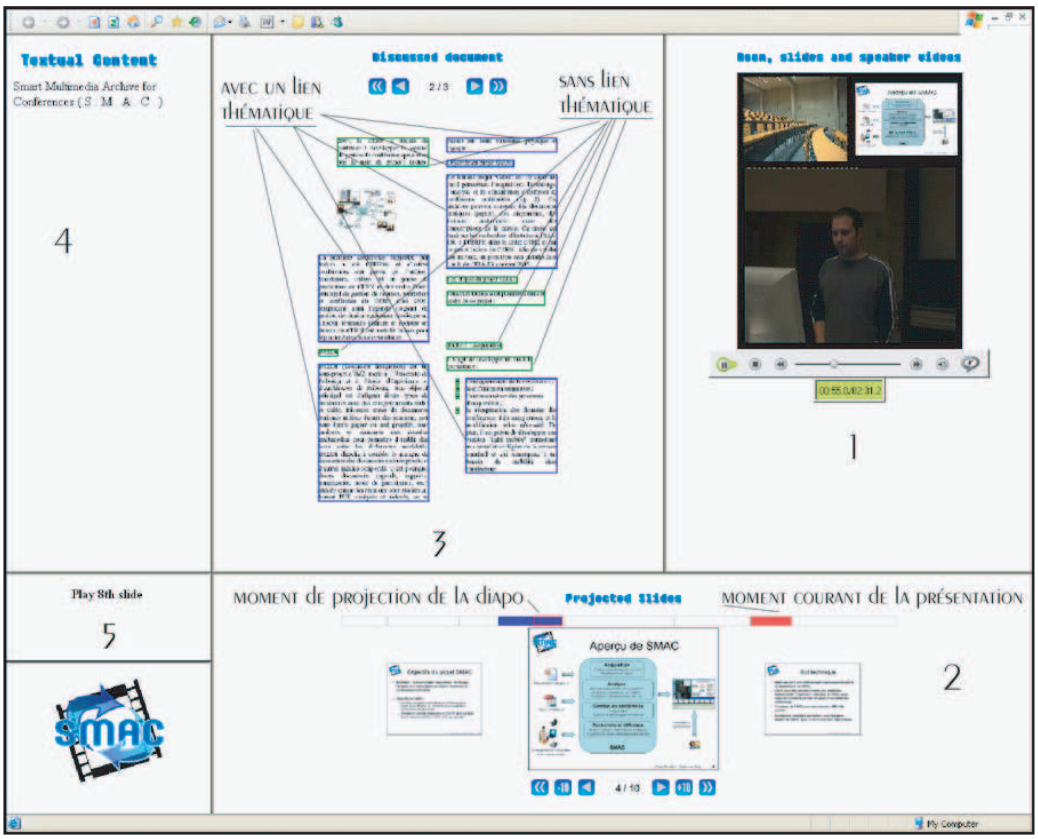


Figure 6.1: Archiving interface for conferences (SMAC project)

Documents		Slideshows	Speech	
Average length	Logical blocks	Slides	Duration	Utterances
6 pages	178	324	237 minutes	1952

Table 6.1: Statistic of the 8 presentations

6.2.1 Static Documents and Slideshows

The average length of static documents within this corpus is roughly of 6 pages. They might contain text as well as graphics (images, charts, diagrams, tables, etc.). The slideshows contain also text and graphics. Slides size and content change from one slide to another and from one slideshow to another. The slide size lies from one word per slide (e.g. a title), to many lines of text that might include mathematical formulas, etc. Usually, the outline of a slideshow keeps up the outline of the corresponding document. Sometimes, it might happen that the speaker skipped some slides. Therefore, these slides are not aligned with any speaker segment.

Both static documents and slideshows might be segmented into syntactic (sentences), logical, thematic, or physical structures. The logical structure, which corresponds to the various slides within slideshows, and the various sections within documents, is the most significant and suitable in this context. This structure has been extracted manually in this work, for both slideshows and static documents. Aligning slides sentences with documents sentences is not really interesting, since a slide sentence corresponds more or less to one document paragraph. The thematic segmentation is not interesting any more, due to the unicity of topic between the documents and the slideshows. Finally, the physical structure, which reflects the geometric characteristics, is more required for browsing tasks, rather than thematically aligning the content.

6.2.2 Speech Transcript

The total duration of the 8 speech recordings is 237 minutes. Since the main speaker in the presentation is generally the presenter, the consideration of speaker turns in the alignment process is not significant. Therefore, speaker utterances have been chosen to be aligned with slides and with document logical blocks. Due to the non-clearness of the speech of the attendees (e.g. the audience when asking questions), only the presenter utterances have been considered. The statistic of the generated data after the segmentation of the various resources is presented in Table 6.1.

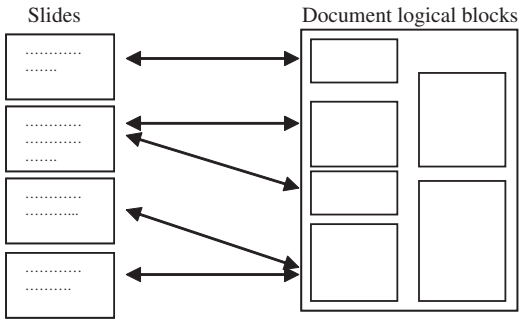


Figure 6.2: Multiplicity of links between slides and document logical blocks

6.2.3 Adding Noise to the Speech Transcript

In order to observe the effect of the Word Error Rate (WER) on our alignment process, and due to the non-availability of a robust Automatic Speech Recognition system (ASR), a method adding noise to the manual speech transcript has been implemented. Adding noise to a particular speaker utterance consists generally in removing terms (deletion), inserting terms (insertion) or substituting terms (modification). However, only the deletion has been considered in our work, so that we can control the WER with precision. Since the stop words are not considered in the detection of thematic alignment links, they have been excluded from the set of words to remove. The WER has been changed from 10% to 40% (Figure 6.4), which corresponds to the proportion of words to remove.

After segmenting the textual content of slideshows, static documents and speech transcript, the next step consists in retrieving the thematic alignment pairs, among the three resources. Before performing the similarity calculation, the various segments should be first cleaned from stop-words, and their words should be stemmed. Then, each term is weighted by the $TF.IDF$ metric. Finally, similarity metrics are applied on pairs of segments (documents/slideshows, documents/speech transcript and slideshows/speech transcript).

6.3 Evaluation

Similarly as seen in section 5.3 of chapter 5, the *recall*, *precision* and F measures have been used to evaluate the generated thematic alignment links of the three modality pairs, according to the prepared manual ground truths.

6.3.1 Thematic Alignment Results and Analysis

During presentations, a document section might be presented by many successive slides, and also by many speaker utterances. From another side, many successive document sections might be presented by only one slide (Figure 6.2). This means that several thematic links should be considered for each source segment, from the documents, the slideshows and the speech transcript. For this reason, the multiple thematic alignment strategy, seen in section 4.3.1.1.3 in chapter 4, has been selected. In this section, the evaluation is based on the speech transcript before addition of noise. The effect of the noise on the performance of our thematic alignment process is studied in the next section.

The thematic links that are considered as non-pertinent, and thus are removed, have been filtered using a similarity threshold value equal to 0.10. The generated pairs of alignment have been compared to the manual ground. The evaluation of the various thematic alignment pairs, using the three metrics, *Cosine*, *Dice* and *Jaccard*, without and then with the consideration of the $TF.IDF$, is presented in Figure 6.3, where $TF.IDF_{t,S_i} = occ_{t,S_i}.\sum_{j=0}^{N} occ_{t,S_j}$. As we can see in this figure, the best F value for each alignment pair is obtained using the $Cosine + TF.IDF$ metric. When the $TF.IDF$ is integrated in the *Cosine* metric, the F value has increased from 0.19 to 0.62 for the documents/slideshows alignment, from 0.22 to 0.33 for the documents/speech transcript alignment, and from 0.19 to 0.47 for the slideshows/speech transcript alignment. This is mainly due to the nature of the content of the static documents, slideshows and speech transcript. Since the three resources generally share the same topic and thus the same vocabulary, some terms appear regularly in their overall segments, such as the terms of the main title. Therefore, weighting the terms by integrating the $TF.IDF$ into the similarity metrics, highlights the important terms in each segment, and devalues those who have regular occurrences within all the segments.

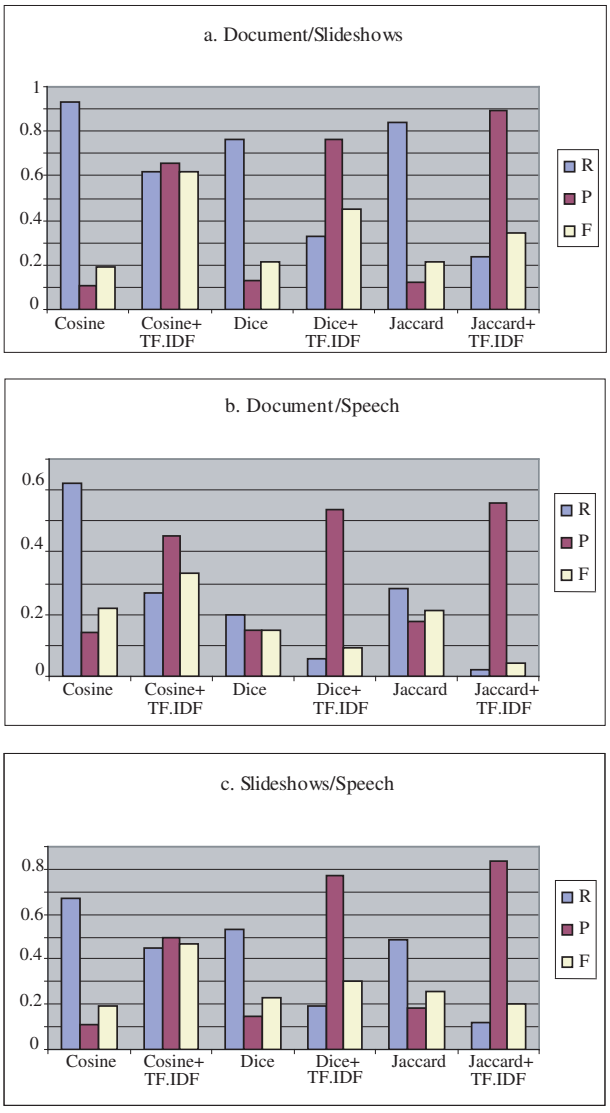


Figure 6.3: Thematic alignment of documents, slideshows and speech transcript

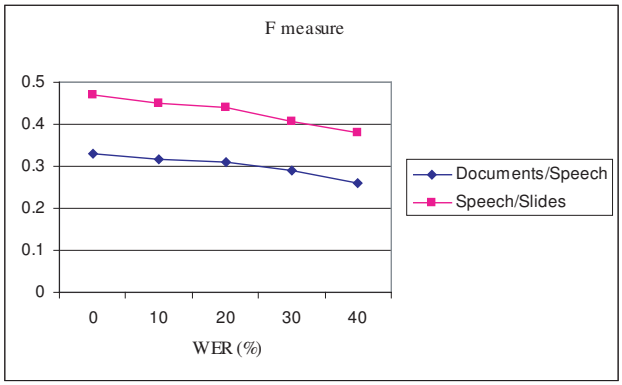


Figure 6.4: Effect of the noise on the thematic alignment

6.3.2 Effect of the Speech Noise on Alignment

In this evaluation, the alignment results using the $Cosine + TF.IDF$ metric have been used. When noise has been added to the speech transcript, the F value has been decreased from 0.47 to 0.38 for the document/speech alignment, and from 0.33 to 0.26 for the slideshows/speech alignment ($WER = 40\%$ in Figure 6.4). However, the effect of noise on our thematic alignment process is not too significant.

6.3.3 Validation of Thematic Alignment

The results of grouping/validation process of the multiple documents alignment, defined in section 4.4.2 of chapter 4, are shown in Figure 6.5. This grouping/validation process, that is composed of two validation strategies, has reinforced and improved the existing alignments for each modality pair. After having performed the first strategy of the validation process, i.e. keep only the links that belong to a cycle (Figure 4.13.a in chapter 4), many false thematic links have been removed, which has increased the F value from 0.62 to 0.79 for documents/slideshows alignment, and from 0.33 to 0.45 for documents/speech alignment. However, some correct links have been removed in the slideshows/speech alignment, which has decreased the F value from 0.47 to 0.45. Using the second strategy of the validation process, which adds the missing arc to each path having two arcs in order to construct a cycle (Figure 4.13.b in chapter 4), many missing thematic links have been added. Thus, the F values have been increased from 0.79, 0.45 and 0.45 to 0.88, 0.60

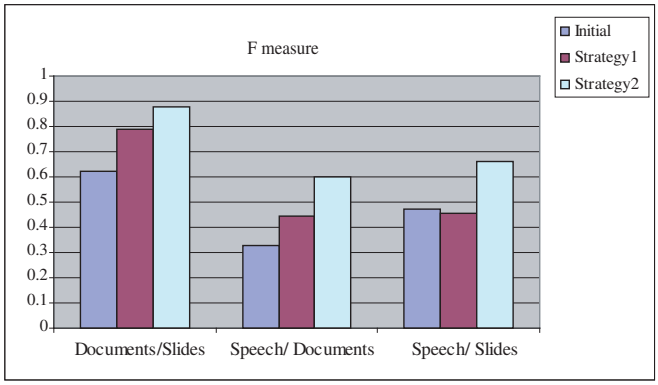


Figure 6.5: Improvement of F value after grouping alignments

and 0.66, for the documents/slideshows, documents/speech and slideshows/speech alignments, respectively. These satisfactory results obtained by the grouping of the various modality alignment pairs shows that each time a new resource is integrated into the thematic alignment process, better results are obtained.

6.3.4 WordNet Integration

In a recent study [Corral 2005], and in order to consider the semantic relationships between words of segments being compared with the similarity metrics (e.g. in Figure 6.6), the WordNet thesaurus [WordNet], studied in section 4.3.1.1.2 of chapter 4, has been integrated in the *Cosine* similarity metric. The semantic relationships between words that have been considered in this work are synonymous (having the same meaning), hypernymy (belonging to a higher class) and hyponymy (belonging to a lower class), with the respective weight values 1, 0.8 and 0.8. The enhanced *Cosine* similarity metric has been evaluated in the context of a categorization of news articles. The articles have been collected from various Web sites (Reuters, ABC and New-York Times). Therefore, 20 articles from 5 distinct categories (4 articles per category), have been chosen in order to be classified in a database composed of 5 categories, where each category contains 12 articles. In this evaluation, the categorization rate has been increased from 65% using the classic *Cosine* metric,

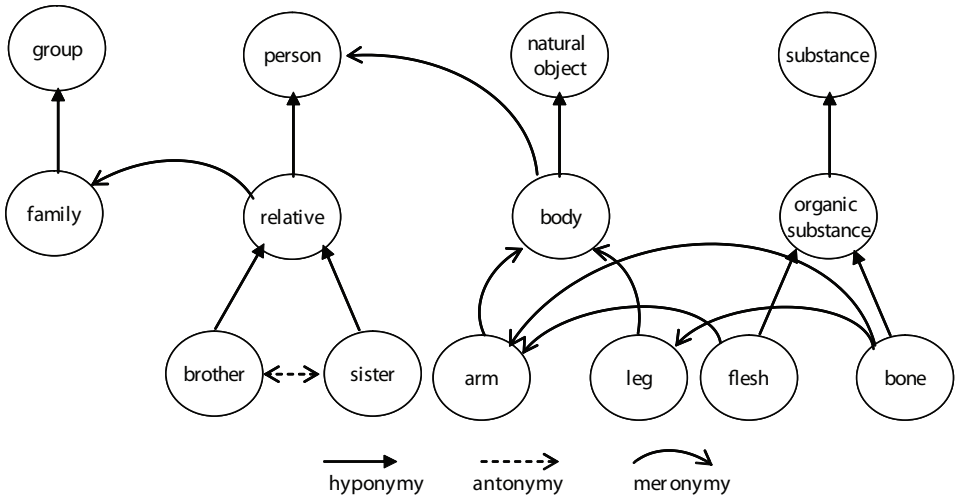


Figure 6.6: Some relationships in WordNet thesaurus

to 75% using *Cosine* metric integrating WordNet. These encouraging results have stimulated us to integrate WordNet thesaurus in our thematic alignment process. Unfortunately, the thematic alignment process of the 8 presentations has produced unconvincing results, especially for the documents/slideshows alignment, where the F value has been decreased from 0.62 to 0.43. These disappointing results can be explained by the fact that usually the slides vocabulary is taken from the presented documents, and thus, the consideration of other semantic relationships between words adds noise to the alignment (the *precision* has decreased from 0.65 to 0.39). Moreover, due to the nature of the application, i.e. the scientific field, terms are quite precise and share limited relationships, which is not the case in other fields, such as in the news classification application, tested before in [Corral 2005].

6.3.5 Annotator Evaluation

The same person has prepared all the thematic alignments of the manual ground truth, used for evaluating of the scientific conference corpus. Due to the relative subjectivity of these alignment annotations, an inter-annotator agreement has been measured through two persons' annotations, especially for static documents/slideshows thematic alignment. The obtained comparison result tends to prove that the task is straightforward, since the annotators' agreement rate is about 97%.

6.4 Conclusion

In the current chapter, an enhanced alignment framework is described. In this framework, slideshows captured during events are exploited, besides the speech transcript, in order to synchronize the corresponding static documents. The satisfactory results of our methods prove that slideshows constitute a considerable alternative to speech recordings, and a complementary resource for information. Slideshows help in associating temporal indexes to static documents, besides speech transcript that might not be robust enough and expensive, when the WER is considerable. Moreover, combining the thematic alignment results of the three modalities (documents/slideshows, slideshows/speech and documents/speech) improves drastically the obtained results for each pair independently, and thus the synchronization of the static documents with multimedia data. Furthermore, our multiple documents alignment method has been successfully integrated in the framework of SMAC project [Lalanne et al. 2005] [Abou Khaled et al. 2006].

After this evaluation of the multimodal alignment framework, another study, conducted in this thesis, is presented in the next chapter, which consists in the thematic segmentation of meetings. This new method is a bimodal segmentation method, in which the results of documents/speech transcript thematic alignment are exploited, in order to generate the thematic structure of meetings.

Chapter 7

Thematic Alignment vs. Thematic Segmentation of Meetings

7.1 Introduction

In the previous chapters, the alignment framework with its three types (thematic, quotations and references) has been described and experimented upon two sets of data, press review meetings and scientific presentations. In the current chapter, another aspect studied in our thesis is described, which consists in a bimodal thematic segmentation method of meetings [Mekhaldi et al. 2004a] [Mekhaldi et al. 2004b]. This research area has been considered in this thesis, since we have noticed that there is a strong relationship between the thematic alignment of static documents with meeting dialogs and the thematic segmentation of these meetings.

Our bimodal segmentation method exploits the results of the thematic alignment, in order to generate simultaneously the documents and the speech transcript thematic segmentations (Figure 7.1). The following sections describe this new segmentation method in details.

7.2 Bimodal Thematic Segmentation Method

In section 3.2.3 of chapter 3, various thematic segmentation methods have been described, especially TextTiling [Hearst 1994], Salton's method [Salton et al. 1996],

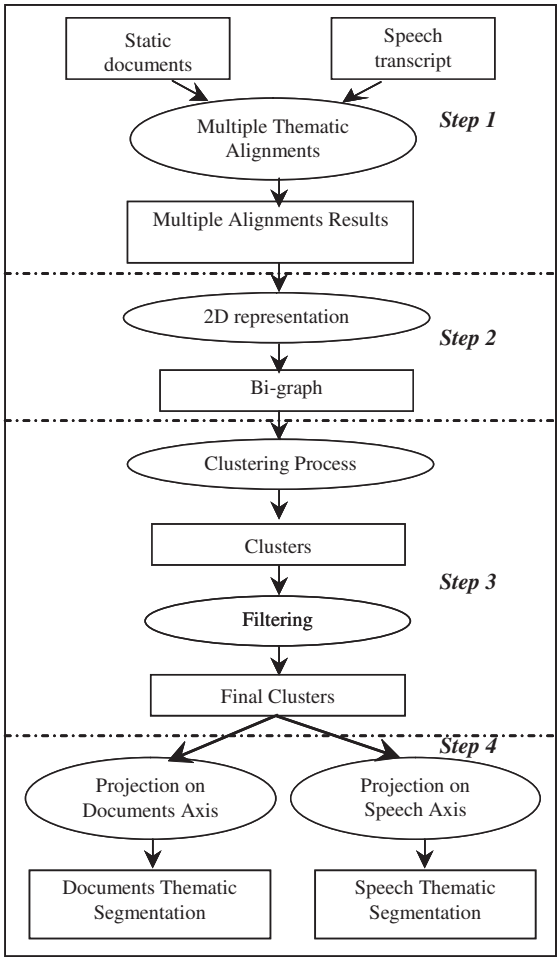


Figure 7.1: Illustrative diagram of our bimodal segmentation method

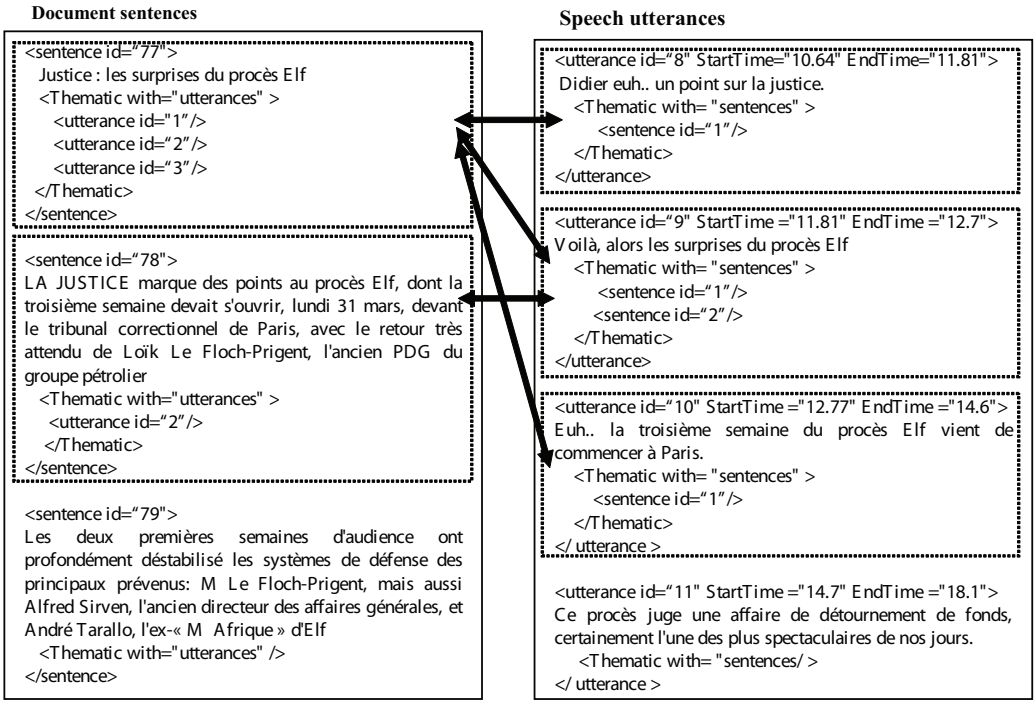


Figure 7.2: Multiple alignments strategy of document sentences with speakers' utterances

the LCP method [Kozima 1993] and the LCA method [Ponte and Croft 1997]. All these methods are monomodal, i.e. they have as input and output only one modalitiy.

Contrary to these monomodal methods, two modalities are used in our bimodal segmentation method, documents and speech transcript, in order to generate the respective thematic segmentations of both resources. The second characteristic of our segmentation method is that it is a simultaneous method, i.e. the segmentations of the two resources are generated in parallel. The third characteristic of our bimodal method is that it uses the results of another framework, i.e. the thematic alignment results, and thus, it might be seen as a validation step of these results.

Our bimodal segmentation method is carried out in four main steps (Figure 7.1). First, the results of the multiple thematic alignments strategy are organized in the form of a bi-graph between two sets, representing document sentences and speakers' utterances respectively. Second, this bi-graph is transformed into a 2D representation, where the alignment links are disposed within a 2D space (Figure

7.5), made by the two axes representing the documents and the speech transcript respectively. In the third step, a clustering process is applied on the nodes within the 2D representation, in order to extract the various clusters representing the meeting topics. The extracted clusters are filtered in order to keep only the significant ones. Finally, the final clusters are projected on both axes, in order to retrieve the corresponding thematic segments of both documents and the speech transcript. The next sections describe all these steps respectively, with more details.

7.2.1 Graphic Representation of Multiple Thematic Alignments

The multiple thematic alignments strategy generates all the significant thematic links, for each source segment, from the documents to the speech transcript and vice versa (e.g. in Figure 7.2). The list of the generated thematic links might be represented in the form of a bi-graph (Figure 7.3), where the segments in respective resources are represented by nodes organized in two sets, and each thematic link between two segments is illustrated by an edge between the corresponding nodes.

In order to keep all the generated information by the thematic alignment process, the various edges (i.e. thematic links) are weighted by their corresponding similarity values. In this context, we have chosen the alignment results of document sentences with speakers' utterances. The selected granularities, i.e. sentences and utterances, have been chosen as being the smallest and the most precise segments, when compared to other segments, such as document logical blocks and speakers' turns.

Within the bi-graph representation of Figure 7.3, there might be many denser regions of edges, with few incoming and outcoming links from and to the other regions, such as in the region composed of group A and group B. Therefore, we can expect that such regions correspond to the meeting topics that should be extracted.

7.2.1.1 Intersection Graphs and Denser Regions Extraction

Our first attempt to solve this thematic regions extraction problem was based on the intersection graphs [Golumbic 1997], which are generated by the projection of the bi-graph on each side, i.e. on documents then on speech transcript side. An intersection graph for the documents and another for the speech transcript are thus generated (Figure 7.4.b and 7.4.c). These two intersection graphs group each two units from one resource, being related to the same intermediate unit from the second resource,

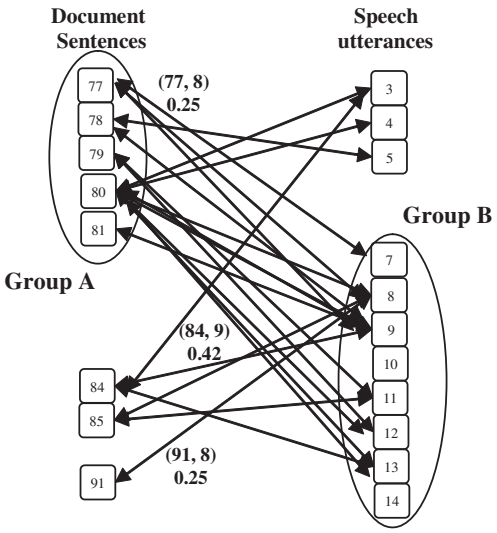


Figure 7.3: Bi-graph representing the results of the multiple alignments strategy

e.g. the document sentences 77 and 78 that are related to the speaker utterance 9 in Figure 7.4.a. According to the number of intermediate nodes to traverse, many intersection graphs of various levels might be obtained. For instance, the vertex (77, 78) in Figure 7.4.a belongs to an intersection graph of level 1, whereas the vertex (77, 81) belongs to an intersection graph of level 2, since the two nodes are connected via two intermediate nodes from the other resource (9 then 12).

Our assumption was that the denser regions, within the bigraph of Figure 7.3, could be isolated and then extracted, by fortifying the edges weight within the intersection graphs. In the beginning, we thought that the edges weight could be fortified by detecting the intersection graphs until a fixed n^{th} level. Unfortunately, we noticed that this method was not efficient. The main reason is that in each intersection graph, more or less all the nodes are related. Furthermore, the intersection graph representation does not preserve all the information offered by the multiple alignment results, such as details about the other resource of the thematic links (i.e. the second intersection graph), which might be crucial for the segmentation process.

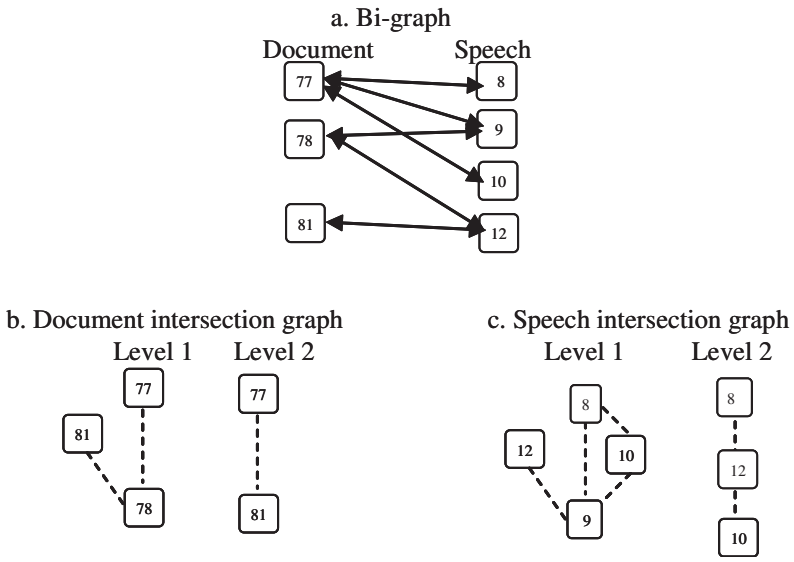


Figure 7.4: Exemple of intersection graphs

7.2.1.2 2D Representation and Denser Regions Extraction

Another way of illustrating the multiple alignments results might be made in the form of a 2D representation, where the document and the speech transcript files are represented by two axes, X and Y respectively (Figure 7.5). Therefore, the various edges of Figure 7.3, i.e. the thematic alignment links, are represented by nodes within the 2D representation. In this space, each node represents the intersection of the two corresponding segments from the documents and the speech transcript, respectively. The similarity value is illustrated in this representation by the node size. Thus, a big node means that the corresponding thematic link has an important similarity value, compared to other small nodes [Mekhaldi et al. 2004b].

This second way of representing thematic alignment results is more efficient and has many advantages, especially by preserving all alignment characteristics. Therefore, the spatial adjacency of documents or speech transcript segments, as well as the temporal adjacency of speech transcript segments i.e. their temporal chronology, are still presented. The denser regions that might represent the meeting topics, e.g. the region AxB in Figure 7.5, are well represented by a set of close nodes, which corresponds to a strong concentration of the thematic links between groups A and B within the bi-graph in Figure 7.3.

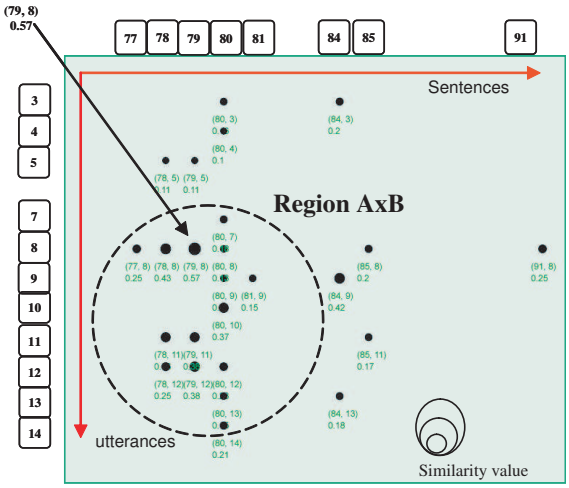


Figure 7.5: 2D representation of the results of the multiple alignment strategy

Much information can be derived from this 2D representation, such as:

- Are there many themes in the meeting?

- What are the main themes?

- When was a given document part discussed?

- In which temporal order was the meeting played? i.e. in which order were the documents parts discussed?

- Are there any overlapped thematic segments? i.e. are there any similar documents parts or speech transcript segments? etc.

We will see later that these information help us not only to detect the thematic segments of both documents and speech transcript, but also to detect the temporal links between the segments of the document. Therefore, the 2D representation of the thematic alignment might be exploited as an interface for navigation within meeting recordings space (Figure 7.11).

Once the various thematic links are laid out within this 2D space, the next step in our bimodal segmentation method consists in extracting the various denser regions. This step is detailed in the following section.

7.2.2 How to Extract the Meeting Topics?

From a quick overview on the 2D representation in Figure 7.5, it is clear that a clustering algorithm should be efficient for extracting the denser regions within this the 2D representation. In the next section, the clustering process that has been used in our work is described.

7.2.2.1 Clustering Methods: State-of-the-Art

Data clustering is a considerable research area, which is exploited in many fields. Starting with a large amount of data, this process consists in the creation, in an unsupervised way, of many data sets that are called clusters [Zhao and Karypis 2001] [Perner 2002] [Jain and Dubes 1988]. This categorization is based on the similitude between the various data. This similarity is computed via a distance metric (Euclidean, Manhattan, etc.). The clustering methods can be categorized as follows [Fasulto 1999]:

- Hierarchical methods, where the clusters are organized as a tree. We distinguish two kinds of hierarchical methods, agglomerative methods and divisive methods. In the agglomerative methods (or bottom-up), each object is initially assigned to one cluster, and then successively, the closest clusters are merged. In the divisive methods (top-down), all the data are initially in the same cluster, and then successively the resulting clusters are split, until the desired number of clusters is reached.

- Partitioning methods, where the data set is decomposed directly into a set of clusters, so that each datum belongs to only one cluster, e.g. the K-Means method [McQueen 1967]. The methods of this category are based on some criteria, such as the maximization of the similarity between the data within each cluster, as well as the maximization of the dissimilarity between the various clusters.

Even though many clustering methods are available, we have chosen the most common one in order to bootstrap our method, the K-Means method.

7.2.2.2 Graph Clustering with K-means

The objective of this clustering process is to define the denser clusters, and the most separated ones from the others. The existing standard K-Means procedure is

described as follows:

1. First, the K value, the number of clusters that should be obtained, must be defined.

2. K points are then generated randomly within the 2D graph, which are considered as the preliminary clusters centroïds.

3. Via a distance formula, e.g. the Euclidean distance, each point from the data set is assigned to the nearest centroïd.

4. After the clusters construction, the new centroïds coordinates are computed using an average formula, to find the mean of the X and Y values of the clusters' nodes.

Steps 3 and 4 are repeated until a stable state is reached, i.e. there is no important change in the positions of clusters centroïds between two successive iterations, according to an Error formula and a prefixed error threshold.

Even though the standard K-means algorithm is widely used [McQueen 1967], we have noticed that it has many drawbacks, which we have classified into two categories, inter-clusters and intra-clusters drawbacks. The inter-clusters drawbacks of K-means are:

- The number of final clusters that should be known in advance, which could not be fixed in our case, since the number of meetings topics changes from a meeting to another.

- The distance between the clusters is not significant, which means that two adjacent clusters are considered different, even though they belong to the same meeting topic.

- There is no criterion to filter insignificant clusters. More precisely, no threshold is considered for the density of significant clusters.

In the intra-clusters drawbacks, there is no consideration of:

- The clusters compactness, i.e. the distance between each cluster centroïd and its nodes.

- The clusters densities, which should include the nodes size, nodes number as well as their respective distance from the centroïd.

For all the previous reasons, the application of the standard version of the K-Means algorithm is not sufficient in our work. However, an improved version of K-Means is presented in [Looney 2002], where the mentioned intra-clusters and inter-clusters criteria are considered, but it is still considering neither the node weights nor the clusters density.

7.2.2.3 Graph Clustering with Extended K-means

In the extended K-Means version [Looney 2002], three main thresholds, which should be initialized by the user, are considered:

1. Defining the K centroïds of the clusters randomly is always followed by a merging method (linking process), which checks the non-closeness of the generated centroïds; otherwise, it merges the clusters with the closest centroïds. This merging method depends on the distance between the centroïds, using a defined threshold. Therefore, the first required threshold, in the extended K-Means algorithm, corresponds to the maximum distance that should not be exceeded, before merging the clusters.

2. The second threshold corresponds to the minimum number of nodes within a significant cluster. This value has been defined by observing a sample of clustering. In our work, we have fixed this threshold to two nodes per cluster, i.e. a meeting topic should contain at least two thematic links between documents and speech transcript segments. With these two thresholds, it is more practical to define a large K value, since it will be decreased if there are close or non-significant clusters.

3. The third threshold is used by the clustering validity measure that is required for stopping the clustering process. The convergence of the clustering algorithm to the best result is measured by the variance formula XB [Xie and Beni 1991], which checks the change of the centroïds positions.

With these three additional thresholds in the extended K-means version, the inter-clusters and intra-clusters limitations, seen in section 7.2.2.2, seem to be covered. Thus, this clustering algorithm has been integrated into our bimodal segmentation method

Before running the clustering process, our alignment results are organized in the form of vectors. Each alignment pair is represented by a vector containing

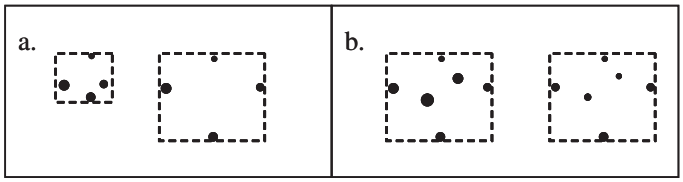


Figure 7.6: Effect of a. the distance between nodes, b. nodes size on the density of clusters. In both cases, the left hand cluster is more significant than the second one

three attributes, the two respective identifiers of the aligned segments from the documents and the speech transcript, as well as their corresponding similarity value. For instance, the vector (77, 8, 0.25) corresponds to the thematic link established between the document sentence 77 and the speaker utterance 8 in Figure 7.3.

The clustering algorithm takes in the input the list of vectors representing all the nodes, and generates a list of clusters. The distance between nodes, as well as clusters, has been computed with three various metrics, the *Euclidean*, *Manhattan* and *Chebychev* distance. However, the *Euclidean* distance has generated the best scores:

$$Euclidean(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2)}$$

$$Manhattan(x, y) = |x_1 - y_1| + |x_2 - y_2|$$

$$Chebychev(x, y) = Max_i|x_i - y_i|$$

Once the clustering process reaches a stable state, another step should be performed, which consists in removing weak clusters that are considered as non-significant.

7.2.2.4 Weak Clusters Filtering

Since the extended K-Means algorithm does not consider the nodes weights when constructing clusters, as seen in section 7.2.2.3, we have enriched our bimodal segmentation method by two filtering steps, in order to remove the weak clusters. The weak clusters correspond to clusters with small densities.

The first filtering step filters clusters according to their densities. The density of a cluster is computed as follow, using the nodes weights, nodes number, and nodes distances from the cluster centroïd:

$$Density = \sum_{i=1}^{Size} (W_i / Distance_i(C)).(Distance_{max}(C)).Size$$

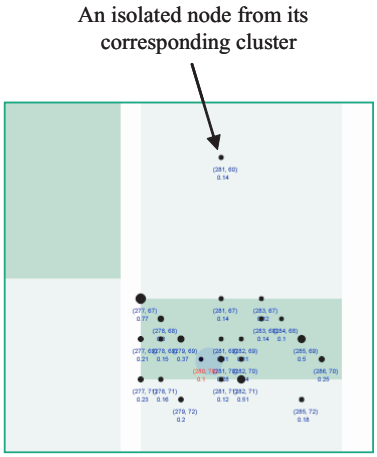


Figure 7.7: Transitional utterances effect on the clustering

$$Density_{relative} = (Density - Density_{min})/(Density_{max} - Density_{min})$$

Where $Density$ and $Density_{relative}$ are respectively the cluster density and relative density according to the overall clusters. $Size$ is the cluster size in terms of nodes, and C is its centroïd. $Distance_i(C)$ is the distance between a cluster centroïd C and a node i from this cluster. W_i is the weight of a node i. $Density_{min}$ and $Density_{max}$ are the minimum and maximum clusters densities, respectively.

According to this density formula, a cluster with a given number of nodes laid in a small surface, is more significant than a cluster with the same nodes but in a larger surface (Figure 7.6.a). In the same way, a cluster with heavy nodes is more significant than a cluster having the same surface and the same number of nodes but with lighter weights (Figure 7.6.b). Filtering the clusters having weak densities is based on a threshold, dynamically defined according to the average clusters density. The second filtering step, that we have defined, is more focusing on the inside each cluster, independently from the other clusters. It consists in removing the isolated nodes inside each cluster. This additional step is applied on the light nodes that are very far from the other nodes within the same cluster, taking into account their distance from the document and the speech axes (Figure 7.7). The gap of these isolated nodes is due in general to the transitional speakers' utterances between the speech thematic segments. Therefore, they cause an overlap of the speech thematic segments, by making them longer than what they should be (Figure 7.7).

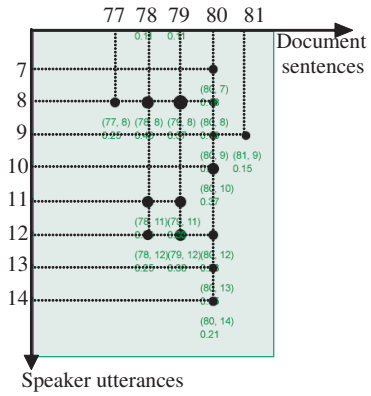


Figure 7.8: Cluster projection

After performing the two filtering steps on the generated clusters, the remaining clusters are considered as the various meeting themes, where each cluster links a speech thematic segment with a similar document thematic segment.

Once the data are clustered and the weak clusters are filtered, the next step of our bimodal segmentation method consists in extracting the thematic segments from the respective clusters. The following section describes this step in more details.

7.2.3 Segments Extraction via Clusters Projection

Each detected cluster, within the 2D representation, corresponds to the intersection of a series of document sentences with a series of speakers' utterances. These respective series might be seen as thematic segments for respective resources. In order to extract the various thematic segments, our proposed solution consists in the projection of the final clusters on both axes X and Y, representing the documents and the speech transcript (Figure 7.8). Therefore, the projection of the cluster AxB in Figure 7.5 generates two segments A = (77, 78, 79, .., 81) and B = (7, 8, .., 14), as shown Figure 7.8, where the coordinates correspond to the document sentences and speaker utterances identifiers respectively. This way, the projection of all final clusters generates the two thematic segmentations for the documents and the speech transcript, respectively. The evaluation of the obtained thematic segmentations, using our bimodal method, is presented in the next section.

7.3 Bimodal Segmentation Method Evaluation

The 22 meetings of the press review corpus (section 5.2 in chapter 5) have been tested in this experiment, with a total of 2936 speaker utterances and 2772 document sentences.

7.3.1 Comparison with other Methods

In order to bootstrap our bimodal segmentation method, it has been compared to two monomodal methods: the TextTiling method [Hearst 1994] seen in section 7.2, and a specific baseline segmentation method, appropriate for each one of the documents and the speech transcript. The documents baseline segmentation method consists in a reflexive segmentation method, which is based on a reflexive document alignment and clustering, i.e. aligning the document with itself, then utilize these alignment results for the clustering process. This baseline method should generate ideally a set of clusters that are laid on the diagonal of the 2D representation (Figure 7.9). The baseline segmentation method that has been considered for the speech transcript corresponds to the speakers' turns, i.e. each turn whose size overcomes a defined threshold (10 words) is considered as a thematic segment of the speech transcript.

7.3.2 Evaluation measure

In order to evaluate our bimodal segmentation method, the Beeferman P_k measure is used [Beeferman et al. 1999], in respect to a manual ground truth thematic segmentation for both resources, documents and speech transcript. The P_k measure computes the probability that a randomly chosen pair of segments, at a distance of k segments apart, is inconsistently classified, in respect to the ground truth. Thus, for a perfect segmentation, the P_k value should be equal to 0. For this experiment, the k parameter has been fixed to four segments (document sentences or speakers' utterances), which corresponds to the minimum size of a relevant thematic segment. This threshold is different from the second threshold seen in section 7.2.2.3 that corresponds to the number of thematic links per cluster. P_k measure has been chosen for our evaluation rather than the *recall/precision* measures, since these latter have many disadvantages [Pevzner and Hearst 2002] [Kehagias et al. 2003], especially, because they measure just the correctness of boundaries detection, without the consideration of the distribution within the content of the generated segments.

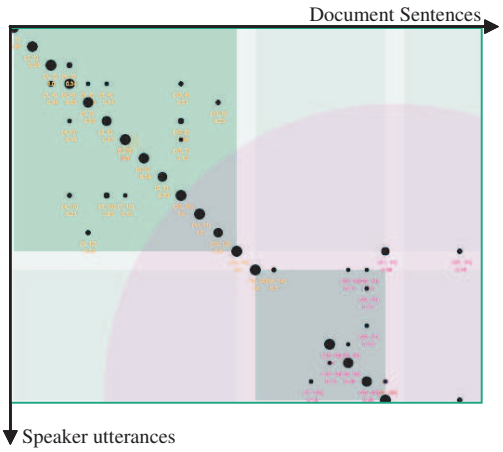


Figure 7.9: Reflexive clustering of documents

In the next paragraph, the thematic segmentation of the speech transcript is evaluated.

7.3.3 Speech Transcript Thematic Segmentation Evaluation

The two scenarios of meetings (stereotyped and non-stereotyped), seen in section 4.3 of chapter 4, are distinguished by the ration of speakers' utterances/turn. Therefore, if this ratio is superior to 2 then the meeting is considered as stereotyped, otherwise, it is considered as non-stereotyped. Actually, in the 9 stereotyped meetings, the speakers present the various articles rather than debate or discuss them, with few interactions or comments. In this category of meetings, there are 1375 document sentences, 617 speaker utterances and 234 turns, with an average ratio of 2.6 utterances per turn, and an average duration of 8 minutes per meeting. In the 13 non-stereotyped meetings, the participants have a high interactivity between them, and debate the different articles. This interactivity increases the number of turns, compared to the stereotyped meetings. Within this category, there are 1397 document sentences, 2319 speaker utterances and 1954 turns, with an average ratio of 1.4 utterances per turn. The average duration in this category is 12 minutes per meeting, which is quite high, when compared to stereotyped meetings.

As shown in Table 7.1, the results of the thematic segmentation, using the three methods (bimodal, TextTiling and baseline), varies according to the scenario of the

	Meetings	
	Stereotyped (9)	Non- stereotyped (13)
Bimodal	0.36	0.44
Texttiling	0.53	0.69
Speaker-Turns	0.44	0.61

Table 7.1: The average P_k value of the thematic segmentation of the speech transcript. P_k=0 for a perfect segmentation

tested meetings. With our bimodal method, the P_k value is in general satisfactory for the two meetings categories (stereotyped and non-stereotyped), in comparison to the TextTiling and the Speaker-Turns methods. However, our method is relatively more efficient for non-stereotyped meetings, which are closer to realistic meetings. Thus, it should be mentioned that in this category, the second filtering step for the isolated nodes, seen in section 7.2.2.4, have not been applied, because it does not improve the results. Indeed, in non-stereotyped meetings, the number of transitional utterances or comments rises, even in the middle of a theme, and thus a topic can be composed of many small turns (the average utterances/turn ratio is 1.4).

In general, our bimodal segmentation method is more accurate in detecting the exact number of thematic segments. This is the case neither for the TextTiling method nor for the Speaker-Turns method, which generate many extra segments. The accuracy of our method can be explained by the fact that using the document modality limits the number of possible themes, which constrains the segmentation, and thus helps in computing the exact number of the speech thematic segments. Another benefit of our bimodal method is the detection of the similar thematic segments within the speech transcript. The similarity of speech thematic segments is illustrated by the vertical alignment of some clusters, which is explained by the fact that many speech thematic segments are aligned with the same document thematic segment (e.g. segments S_A and S_B in Figure 7.10). In the next paragraph, the documents thematic segmentation is evaluated and discussed.

7.3.4 Documents Thematic Segmentation Evaluation

In this case, the 22 tested meetings have been classified into mono-document and multi-documents meetings, according to the number of documents that are discussed in each the meeting. In the 18 mono-document meetings, there are 1639 document

	Meetings			
	Mono-document (18)		Multi-documents (4)	
	Stereotyped (7)	Non-stereotyped (11)	Stereotyped (2)	Non-stereotyped (2)
Bimodal	0.32	0.28	0.24	0.30
Texttiling	0.64	0.60	0.49	0.55
Reflexive	0.45	0.48	0.51	0.63

Table 7.2: The average P_k value of the thematic segmentation of the documents. P_k=0 for a perfect segmentation

sentences, 2354 speaker utterances and 1539 turns. In the four multi-documents meetings, there are 1133 sentences, 582 speaker utterances and 349 turns. In this latter category, the various documents of a meeting have been merged in one file.

As shown in Table 7.2, the P_k values using our bimodal method are better, compared to the TextTiling method and the baseline segmentation method (reflexive alignment/clustering of the documents). The main reason is that our method is more accurate in detecting the thematic segments of the documents. Indeed, the TextTiling and the reflexive methods generate many extra segments. Moreover, our method detects similar articles within a document, or within various documents in the case of the multi-documents meetings. The similarity of articles is illustrated by a horizontal alignment of some clusters in our 2D representation, which means that many documents articles are aligned with the same speech thematic segment (e.g. segments D_A and D_C in Figure 7.10). Nevertheless, it should be mentioned that our bimodal method is only appropriate for the thematic segmentation of the documents that are almost fully discussed during the meetings, otherwise the documents are partially segmented, i.e. only the discussed thematic segments are detected.

7.3.5 Visualization of Meetings Themes

The results of the clustering process for a given meeting, presented above, is presented in Figure 7.11, using an SVG browser. The circle, around each cluster centroïd, represents the cluster density, where the radius increases according to the density value. The vertical bars (A_1, A_2, etc.) correspond to the manual thematic segmentation of the document, which corresponds to the various newspaper articles. The horizontal bars (S_1, S_2, etc.) correspond to the manual thematic segmentation

of the speech transcript. Furthermore, the rectangle obtained by the intersection of the bars A_i with S_j (the highlighted rectangles), represent the ground-truth meeting themes, which means that the document segment A_i is thematically similar to the speech segment S_j. This visualization technique is an efficient tool to check the validity of the multiple thematic alignments, and to check whether or not the generated thematic segments are lined up with the thematic segments of the ground truth. Moreover, the visualization gives an idea about the reading order of the various documents articles in the meeting, i.e. which article has been discussed first, etc.

7.3.6 Analysis

Despite the fact that the alignment similarity values have not been taken into account when assigning the nodes to clusters, the comparison of our bimodal method with the two monomodal methods, tends to prove that using various modalities improves considerably both the documents and speech transcript thematic segmentation. Moreover, our bimodal segmentation method represents an important advantage, which is the detection of all the potential thematic links between non-adjacent segments within a document (speech transcript respectively). The thematic similarity between non-adjacent segments appears when the segments of the document (speech transcript respectively) are being aligned with the same speech thematic segment (documents thematic segment respectively), especially when discussing several documents that contain similar articles (Figure 7.10).

However, the horizontal or the vertical alignment of the clusters may generate overlapped segments, when the corresponding clusters are projected. This aspect, which can have a negative effect on the results of the segmentation, is discussed in details in the next section.

7.3.7 Overlap of Generated Thematic Segments

During the segments extraction step, more specifically when projecting the denser clusters, we have faced a problem with the clusters that are aligned horizontally or vertically, i.e. clusters whose projection on the document axis or on the speech axis generates overlapped segments, as shown in Figure 7.10. Given two overlapped segments, two kinds of overlap have been distinguished:

- The two segments are totally overlapped (S_C contains S_A).

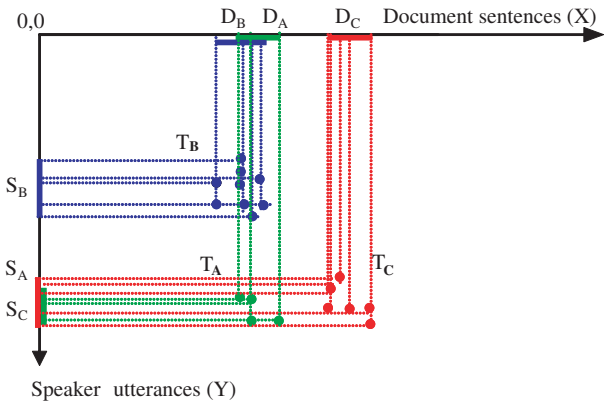


Figure 7.10: Example of overlap of thematic segments

- The two segments are partially overlapped (D_B with D_A).

In the first kind of the overlap (total overlap), only the largest segment is considered, because it is extremely probable that they have the same theme. In the second kind of the overlap (partial overlap), depending on the size of the common part, the overlapped segments are considered as two distinctive segments, taking into account a pre-defined tolerance margin, in number of segments. Depending on which axis the overlap is taking place, a different explanation can be made. The two following sections describe the overlap on the document axis and then on the speech axis, respectively.

7.3.7.1 Overlap on the Document Axis

This first type of overlap (e.g. D_B with D_A) can be explained by the fact that a specific topic T from the document is discussed twice during the meeting, in two non-adjacent moments, for instance, at the time corresponding to the segment S_B, then at the time corresponding to the segment S_A. Another explanation is that two distinct topics from the document (e.g. T_B and T_C corresponding respectively to segments D_B and D_C) are thematically similar, so the corresponding speech thematic segment (e.g. S_C) has been aligned with the two topics, which has generated two clusters (A and C). For example, the two following sentences come from two distinct articles, and have a distinct topic. One deals with the war in Iraq, and the second tackles the effect of Chirac's political position on his popularity in France. However, they share

some words (guerre, Iraq) and thus risk to be aligned with similar speech segments:

a- "La prolongation de la guerre en Irak affecte gravement la conjoncture économique, l'indice européen ..."

b- "Avec sa position forte sur la guerre en Irak, il s'est produit quelque chose entre Jacques Chirac et les Français"

7.3.7.2 Overlap on the Speech Axis

In this second type, the overlap of the speech segments (e.g. S_A with S_C) has two explanations. The first one is that there is a thematic similarity between two topics of the documents (e.g. T_A with T_C), such as in the multi-documents meetings, where the documents might share similar articles. The second explanation is that there is a topic that has been referenced at the same time when the another was discussed, as illustrated in the following speakers' utterances. In the first utterance, the journalist talks about Iraq after-Saddam, and whether it will be liberated or occupied. In the second utterance, another speaker refers to another article describing the point of view of Blair about this issue:

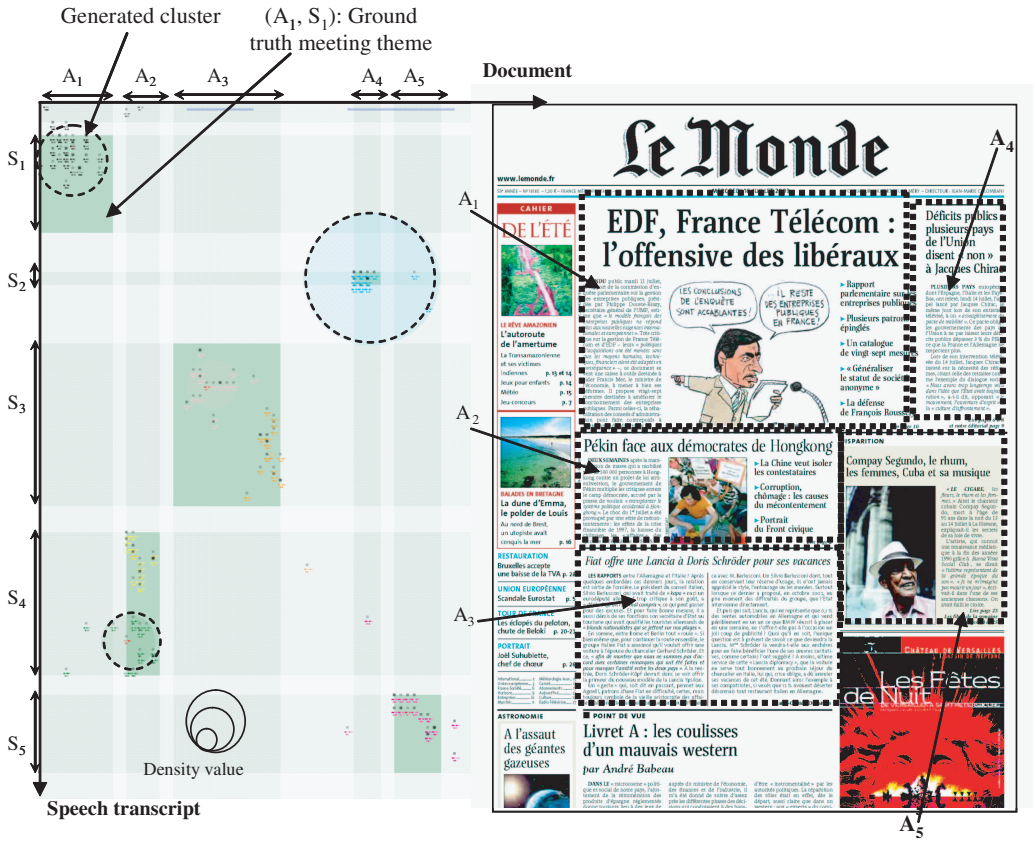
a- "Et.. il y a un article aussi dans lequel le journaliste réfléchit à l'après-Saddam. Euh.. voir qu'est-ce qui va se passer, si l'Irak sera occupé ou libéré"

b- "Justement j'ai un petit article sur ce point-là, donc selon Tony Blair, euh.. l'après-Saddam, c'est-à-dire l'Irak de l'après-Saddam va être géré par des irakiens"

Improving the segmentation scores of our bimodal method is conditioned by solving the overlap of segments. Several methods have been tested (e.g. the Gaussian) in order to solve jointly documents and speech segments overlap. However, with the second filtering step seen in section 7.2.2.4, which consists in the elimination of isolated nodes within clusters, the ratio of the overlapped segments has decreased, especially for stereotyped meetings. Hence, our bimodal segmentation performances have been increased.

7.4 Conclusion

In the current chapter, a bimodal thematic segmentation method of meeting dialogs has been described. This new method is based on the multiple thematic alignment



Manual Speech Transcript

<Thematic_Segment id= "S_1"> …
<block id= "6" StartTime="47.429" EndTime="61.062" speaker="spk2">
Alors euh.. mardi 15 juillet, hier, euh.. la commission d'enquête parlementaire en France a rendu un rapport euh.. sur la gestion des entreprises.. entreprises publiques. </block>
…
<block id= "8" StartTime="71.806" EndTime="88.664" speaker="spk1">
Euh.. ils critiquent très fortement euh.. France Télécom et EDF et leur gestion, et euh.. il est dit aussi que leur politique d'acquisition ont été menées sans que les moyens humains, techniques et financiers aient été adaptés en conséquence. </block> …
</ Thematic_Segment >
<block id= "26" StartTime="148.898" EndTime="151.674" speaker="spk3">
On passe à l'article suivant? Celui-là sera pas très long non plus.</block>
< Thematic_Segment id= "S_2">
<block id= "27" StartTime="151.674" EndTime="170.194" speaker="spk3">
….</block>
</ Thematic_Segment >
…

Figure 7.11: A preliminary interface for meetings navigation

of static documents with speech transcript. The satisfactory scores obtained with this bimodal method, when compared to other monomodal and baseline methods, prove that there is a strong relationship between the thematic alignment process and the thematic segmentation of meetings. Moreover, these results prove that the structured static documents can serve as a mean to structure meetings, rather than using the traditional audio and video features.

After presenting the work that has been conducted in this thesis, especially in chapter 4, 5, 6 and 7, the next chapter concludes our thesis.

Chapter 8

Conclusion

The major goal of this thesis was to bridge the gap between static documents and multimedia documents. This goal aims to exploit static documents for accessing, indexing and browsing multimedia documents, and from another side, exploit the multimedia documents for the temporal indexing of static documents. In order to reach this goal, our proposed solution consists in an automatic generation of links between the static documents and the other multimedia documents, a process that is called multimodal document alignment. The speech transcript has been chosen as the main multimedia document for this multimodal alignment. Later, slideshows have been integrated into our alignment framework. These two documents, speech transcript and slideshows, have been selected, as being both textual and multimedia documents, and thus alignable with static documents.

Our multimodal alignment of static documents with the speech transcript has covered three main alignment types, namely thematic, quotations and references. However, only the thematic alignment has been performed between static documents and slideshows. Each of the three types of alignment, thematic, quotations and references, is extracted following a different process, but all of them insure a temporal indexing of static documents. Furthermore, the three alignment types are useful for many information retrieval tasks, such as searching, browsing, indexing, etc. In addition, the thematic alignment has been exploited in this thesis in order to generate the thematic structure of meetings.

The multimodal document/speech transcript alignment has been adopted as a solution for the problem of the temporal indexing of static documents, and the index-

ing of multimedia documents using the static documents. The satisfactory results obtained on two use cases, respectively meetings and conferences, have proved the performance of the proposed solution for this problem, as well as its usefulness for event structuring and browsing. Moreover, and according to this multimodal alignment, a new classification has been proposed for the existing meeting/classroom projects, which is based on the types of data used for the synchronization.

Nevertheless, there have been some limitations, at several stages in this work. The alignment framework has been performed on speech transcript that is manually transcribed (chapter 5). In chapter 6, noise has been added to the manual speech transcript. Even if the effect of this noise on the thematic alignment was not significant, it is still required to experiment our framework on automatic speech transcript.

The logical structure of documents, that has been used in the alignment framework, has been also manually extracted. However, an automatic logical structure should be employed in the future, for both static documents and slideshows.

The algorithm detecting quotations between speech transcript and static documents should be improved, in order to deal with the problem of overlapped quotations, as seen in section 4.3.1.2 of chapter 4.

Furthermore, the integration of WordNet, a general semantic thesaurus, in the thematic alignment process has not increased the alignment results (chapter 6), which might not be the case if a domain-specific thesaurus is used in the future. This drawback might be explained by the fact that in the application used, conference lectures, exactly similar terms are used in all the modalities (scientific articles, slideshows, speech), and thus the use of synonyms, hypernyms and hyponyms, only introduces noise. However, when the WordNet thesaurus has been considered in the context of a classification task of news articles [Corral 2005], very satisfactory results have been obtained, which is encouraging for the future.

Even though the thematic alignment between static documents and slideshows has been studied in this thesis (chapter 6), the consideration of other alignment types should reinforce the links between them, such as the image based alignment that has been already studied in another thesis [Behera et al. 2005]. In this image based alignment, the content based analysis and the layout features of document images captured during event are exploited, in order to align images with original static documents.

Finally, our bimodal thematic segmentation method, described in chapter 7, has faced an overlap problem between the generated thematic segments, which has affected the results. For this reason, a solution should be considered in order to decrease the ratio of overlapped clusters. Moreover, the thematic structures obtained for both static documents and speech transcript should be exploited in our multimodal alignment framework, in addition to the other combination pairs of structures, as seen in section 4.3.1.1.4 of chapter 4.

Despite these limitations, our multimodal documents/speech alignment has opened the way to a new research field. This alignment has bridged the gap between three distinct research fields: information retrieval field [Mekhaldi et al. 2004a], document analysis field [Mekhaldi et al. 2003] [Mekhaldi et al. 2004c] [Mekhaldi et al. 2005], and multimedia data analysis field [Mekhaldi et al. 2004b] [Lalanne et al. 2004a].

Furthermore, it has highlighted new aspects such as the multi-layered document structures, the interactivity and cooperation between different modalities, meetings structure, bimodal segmentation, etc. For instance, the multi-layered document structure unfolds the multiple document structures (physical, logical, thematic, etc.), that can be used for improving document recognition and indexing.

In the future, our multimodal alignment framework should be extended to other multimodal events (e.g. lectures, administrative meetings, political debates and forums, legal debates, psychological meetings, weather news, etc.), which use additional static and multimedia documents (handwriting documents, visited web pages, handwriting on electronic whiteboard, budgets, tables, reports, law texts, etc.). The consideration of other modalities might be also interesting and enriching, as for example gestures and body movements, that might be aligned with documents, in order to index the video stream.

For long term perspectives, the multimodal document alignment might be used for other fields, such as multimedia search engines, which are specialized in searching and retrieving audio and video files. In this case, our multimodal alignment techniques might be exploited in order to retrieve the multimedia data linked to the retrieved documents. The idea is to fully benefit from cross-media indexing methods to enrich search and browsing strategies.

The digital interactive TV is another emerging field that could benefit from the multimodal document alignment. This new technology, that combines elements from

analogic TV and from Internet, offers to the user a digital broadcasting, enhanced by interactive services. With these services, the user might directly interact with the content by changing the narrative flow, searching and retrieval of media, etc. Such interactive services might benefit from the multimodal document alignment, especially when the program guide and the content annotations are exploited.

In our society nowadays, biometric systems are become challenging tools that are required for identifying and individualizing people. These systems exploit either one or several person characteristics (speech, signature, etc.), and from another side, identity documents (passports, identity cards, etc.) that contain information designed to distinguish people. Based on the data contained within these documents, the multimodal document alignment represents a pertinent solution for the identification and verification of persons within these biometric systems.

The multimodal alignment of various modalities (e.g. in a soccer game) might enhance the respective annotations and descriptors by various information (temporal, conceptual, etc.). This could be useful for many tasks within the multimodal systems, such as content planning, design of the system presentation, resolving conflict and inconsistency between various modalities, etc.

Bibliography

[Abberley et al. 1998] Abberley, D., Renals S., Cook, G.. Retrieval of Broadcast News Documents with the THISL System, in Proceeding of ICASSP, the International Conference of Acoustics, Speech and Signal Processing, Seattle, USA, 1998, pp. 3781-3784.

[Abou Khaled et al. 2006] Abou Khaled, O., Le Meur, J.Y., Scheurer, R., Bourillot, D., Lalanne, D., Von Rotz, D., Ingold, R., Baron, T.. SMAC Project: SMAC - Smart Multimedia Archive for Conferences, in Flash Informatique, Ecole Polytechnique Fédérale de Lausanne EPFL, Switzerland, FI1/06, January 2006, pp. 3-10.

[Anderson et al. 2004] Anderson, R., Hoyer, C., Prince, C., Su, J., Videon, F., Wolfman, S.. Speech, Ink, and Slides: The Interaction of Content Channels, in Proceedings of the ACM Multimedia, New York, USA, 2004, pp. 796-803.

[Ang et al. 2005] Ang, J., Liu, Y., Shriberg, E.. Automatic Dialog Act Segmentation and Classification in Multiparty Meetings, in Proceedings of ICASSP, the International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, Philadelphia, USA, 2005, pp. 1061-1064.

[Bagga and Biermann 2000] Bagga, A., Biermann, A. W.. A Methodology for Cross-Document Coreference, in Proceedings of JCIS, the 5^{th} Joint Conference on Information Sciences, North Carolina, USA, 2000, pp. 207-210.

[Baird et al. 1990] Baird, H. S., Jones, S. E., Fortune, S. J.. Image Segmentation by Shape-directed Covers, in Proceedings of ICDAR, the International Conference on Pattern Recognition, Atlantic City, USA, 1990, pp. 820-825.

[Bani-Ahmad at al. 2005] Bani-Ahmad, S., Cakmak, A., Ozsoyoglu, G, Al-Hamdani, A.. Evaluating Publication Similarity Measures, Bulletin of the Tech-

nical Committee on Data Engineering, IEEE Computer Society, Vol. 28, No 4, 2005, pp. 23-30.

[Barras et al. 1998] Barras, C., Geoffrois, E., Wu, Z., Liberman, M.. Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech, in Proceedings of LREC, the 1^{st} International Conference on Language Resources and Evaluation, Granada, Spain, May 1998, pp. 1373-1376.

[Beeferman et al. 1999] Beeferman, D., Berger, A., Lafferty, J.. Statistical Models for Text Segmentation, Machine Learning, Vol. 34, No 1/3, 1999, pp. 177-210.

[Behera et al. 2005] Behera, A., Lalanne, D., Ingold, R.. Influence of Fusion Strategies on Feature-based Identification of Low-resolution Documents, in Proceedings of DocEng, the ACM Symposium on Document Engineering, Bristol, UK, 2005, pp. 20-22.

[Bett et al. 2000] Bett, M., Gross, R., Yu, H., Zhu, X., Pan, Y., Yang, J., Waibel, A.. Multimodal Meeting Tracker, in Proceedings of RIAO, Recherche d'Informations Assistée par Ordinateur, France, 2000.

[Bloechle et al. 2006] Bloechle, J. L. , Rigamonti, M., Hadjar, K., Lalanne, D., Ingold, R.. XCDF: A Canonical and Structured Document Format, in Proceedings of DAS, the 7^{th} IAPR International Workshop on Document Analysis Systems, Nelson, New Zealand, 2006.

[Brotherton et al. 1998] Brotherton, J. A., Bhalodia, J. R., Abowd, G. D.. Automated Capture, Integration, and Visualization of Multiple Media Streams, in Proceedings of IEEE Multimedia, 1998, pp. 54-63.

[Chakravarthy 1995] Chakravarthy, A.. Sense Disambiguation Using Semantic Relations and Adjacency Information, in Proceedings of ACL, the 33^{rd} Annual Meeting of the Association for Computational Linguistics, Cambridge, Massachusetts, USA, 1995, pp. 284-286.

[Chang and Lui 2001] Chang, C. H., Lui, S. C.. IEPAD: Information Extraction Based on Pattern Discovery, in Proceedings of WWW, the 10^{th} World Wide Web Conference, Hong Kong, 2001, pp.681-688.

[Chen et al. 2004] Chen, F., Farhat, A., Brants, T.. Multiple Similarity Measures and Source-pair Information in Story Link Detection, in Proceedings of HLT-

NAACL, Human Language Technology Conference of the North American, Chapter of the Association for Computational Linguistics, Boston, USA, 2004, pp. 313-320.

[Chiu et al. 2000a] Chiu, P., Kapuskar, A., Reitmeier, A., Wilcox, L.. Room with a Rear View: Meeting Capture in a Multimedia Conference Room, IEEE Multimedia, Vol. 7, No 4, 2000, pp. 48-54.

[Chiu et al. 2000b] Chiu, P., Foote, J., Girgensohn, A., Boreczky, J.. Automatically Linking Multimedia Meeting Documents by Image Matching, in Proceedings of Hypertext'00, ACM Press, Texas, USA, 2000, pp. 244-245.

[Church 1993] Church, K. W.. Char_align: a Program for Aligning Parallel Texts at the Character Level, in Proceedings of ACL, the Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, 1993, pp. 1-8.

[Corral 2005] Corral, D.. Including a Thesaurus in Similarity Calculation, a Bachelor Thesis in Computer Science, University of Fribourg, Switzerland, November 2005.

[Cutler et al. 2002] Cutler, R., Rui, Y., Gupta, A., Cadiz, J., Tashev, I., He, L., Colburn, A., Zhang, Z., Liu, Z., Silverberg, S.. Distributed Meetings: a Meeting Capture and Broadcasting System, in Proceedings of ACM Multimedia, France, 2002, pp. 503-512.

[Denoyer et al. 2003] Denoyer, L., Vittaut, J.N., Gallinari, P., Brunessaux, S.. Structured Multimedia Document Classification, in Proceedings of ACM Symposium on Document Engineering, France, 2003, pp. 153-160.

[Fasulto 1999] Fasulto, D.. An Analysis of Recent Work on Clustering Algorithms, Technical Report, University of Washington, April 1999.

[Fisher 1991] Fisher, J. L.. Logical Structure Descriptions of Segmented Document Images, in Proceedings of ICDAR, the International Conference on Document Analysis and Recognition, France, 1991, pp. 302-310.

[Foote et al. 1998] Foote, J., Boreczky, G., Wilcox, L.. An Intelligent Media Browser Using Automatic Multimodal Analysis, in Proceedings of ACM Multimedia, Bristol, UK, 1998, pp 375-380.

[Ghorbel et al. 2002] Ghorbel, H., Coray, G., Linden, A.. SAM: System for Multi-criteria Text Alignment, in Proceedings of LREC'02, the International Conference on Language Resources and Evaluation, Las Palmas, 2002, pp. 404-410.

[Girgensohn et al. 2001] Girgensohn, A., Borczky, J., Wilcox, L.. Keyframe-based User Interfaces for Digital Video, IEEE Computer, 2001, pp. 61-67.

[Golumbic 1997] Golumbic, M.C.. Algorithmic Graph Theory and Perfect Graphs, 2^{nd} Edition, Hardcover, Publisher Academic Press, 1997, ISBN 0-444-51530-5.

[Hauptmann 2005] Hauptmann, A., Lessons for the Future from a Decade of Informedia Video Analysis Research, in Proceedings of CIVR, the International Conference on Image and Video Retrieval, National University of Singapore, Singapore, 2005.

[Hearst 1994] Hearst, M.. Multi-Paragraph Segmentation of Expository Text, in Proceedings of ACL, the 32^{nd} Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, USA, 1994, pp.9-16.

[Higgins and Taylor 2000] Higgins, D., Taylor, W.. Bioinformatics - Sequence, Structure and Databanks, Oxford University Press, 2000, ISBN 0199637903.

[Hirschberg and Litman 1993] Hirschberg, J., Litman, D.. Empirical Studies on the Disambiguation of Cue Phrases, Computational Linguistics, Vol. 19, No 3, 1993, pp. 501-530.

[Ishitani 1999] Ishitani, Y.. Logical Structure Analysis of Document Images Based on Emergent Computation, in Proceedings of ICDAR, the International Conference on Document Analysis and Recognition, India, 1999, pp. 189-192.

[Jain and Dubes 1988] Jain, A., Dubes, R.. Algorithms for Clustering Data, Hardcover, Publisher Prentice Hall College Div, 1988, ISBN 013022278X.

[Jarmasz and Szpakowicz 2001] Jarmasz, M., Szpakowicz, S.. Roget's Thesaurus: a Lexical Resource to Treasure, in Proceedings of NAACL, the Workshop on WordNet and Other Lexical Resources, Pittsburgh, 2001, pp. 186-188.

[Jeon et al. 2005] Jeon, J., Croft, B., W., Ho, L. J.. Finding Semantically Similar Questions Based on their Answers, in Proceedings of SIGIR, the International Conference on Research and Development in Information Retrieval, Salvador, Brazil, 2005, pp. 617-618.

[Jianying et al. 1999] Jianying, H., Ramanujan, K., Gordon W.. Document Classification using Layout Analysis, in Proceedings of the Workshop on Document Analysis and Understanding for Document Databases, Florence, Italy, 1999, pp. 556-560.

[Kay and Roscheisen 1993] Kay, M., Roscheisen, M.. Text-Translation Alignment, in Proceedings of ACL, Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, 1993, pp. 121-142.

[Kazman and Kominek 1997] Kazman, R., Kominek, J.. Supporting the Retrieval Process in Multimedia Information Systems, in Proceedings of the 30^{th} Annual Hawaii International Conference on System Sciences, Hawaii, 1997, pp 229-238.

[Kehagias et al. 2003] Kehagias, A., Pavlina, F., Petridis, V.. Linear Text Segmentation Using a Dynamic Programming Algorithm, in Proceedings of 10^{th} Conference of the European Chapter of the Association for Computational Linguistics, Budapest, Hungary, 2003, pp. 171-178.

[Kilgarriff and Yallop 2000] Kilgarriff, A., Yallop, C.. What's in a Thesaurus?, in Proceedings of LREC, the 2^{nd} Conference on Language Resources and Evaluation LREC, Athens, 2000, pp. 1371-1379.

[Kimber et al. 1995] Kimber, D.G., Wilcox, L.D., Chen, F.R., Moran, T.. Speaker Segmentation for Browsing Recorded Audio, in Proceedings of CHI, Conference on Human Factors in Computing Systems, Denver, Colorado, USA, 1995, pp. 212-213.

[Kise et al. 1998] Kise, K., Sato, A., Iwata, M.. Segmentation of Page Images using the Area Voronoi Diagram, Computer Vision and Image Understanding, Vol. 70, No 3, 1998, pp. 370-382.

[Kloptchenko et al. 2003] Kloptchenko, A., Back, B., Vanharanta, H., Toivonen, J., Visa, A.. Prototype-Matching System for Allocating Conference Papers, in Proceedings of 36^{th} Annual Hawaii International Conference on System Sciences, Big Island, Hawaii, 2003, 79-88.

[Kopec and Chou 1994] Kopec, G. E., Chou, P. A.. Document Image Decoding using Markov Source Models, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 16, No 6, 1994, pp. 602-617.

[Krishnamoorthy et al. 1993] Krishnamoorthy, M., Nagy, G., Seth, S., Viswanathan, M.. Syntactic Segmentation and Labeling of Digitized Pages from Technical Journals, IEEE Transactions in Pattern Analysis and Machine Intelligence, Vol. 15, No 7, 1993, pp. 737-747.

[Kornfield et al. 2004] Kornfield, E. M., Manmatha, R., Allan J.. Text Alignment with Handwritten Documents, in Proceedings of DIAL, Document Image Analysis for Libraries, San Jose, California, USA, 2004, pp. 195-211.

[Kozima 1993] Kozima, H.. Text segmentation Based on Similarity between Words, in Meeting of ACL, the Association for Computational Linguistics, Colombus, Ohio, USA, 1993, pp. 286-288.

[Kuper et al. 2003] Kuper, J., Saggion, H., Cunningham, H., Declerck, T., de Jong, F., Reidsma, D., Wilks, Y., Wittenburg, P.. Intelligent Multimedia Indexing and Retrieval through Multi-source Information Extraction and Merging, in Proceedings of IJCAI, the 18^{th} International Joint Conference on Artificial Intelligence, Acapulco, Mexico, 2003, pp. 409-414.

[Kwong 1998] Kwong, O. Y.. Bridging the Gap between Dictionary and Thesaurus, in Proceedings of COLING-ACL, the 36^{th} Conference on Association for Computational Linguistics, Montreal, Canada, 1998, pp. 1487-1489.

[Lalanne et al. 2003] Lalanne, D., Sire, S., Behera, A., von Rotz, D., Mekhaldi, D., Ingold, R.. A Research Agenda for Assessing the Utility of Document Annotations in Multimedia Databases of Meeting Recordings, in Proceedings of the 3^{rd} International Workshop on Multimedia Data and Document Engineering, in conjunction with VLDB, Germany, 2003, pp. 47-55.

[Lalanne et al. 2004a] Lalanne D., Ingold, R., von Rotz, D., Behera, A., Mekhaldi, D., Popescu-Belis, A.. Using Static Documents as Structured and Thematic Interfaces to Multimedia Meeting Archives, in Proceeding of MLMI, the International Workshop on Machine Learning for Multimodal Interaction, Martigny, Switzerland, 2004, pp.87-100.

[Lalanne et al. 2004b] Lalanne, D., Mekhaldi, D., Ingold, R.. Talking about Documents: Revealing a Missing Link to Multimedia Meeting Archives, in Proceedings of Document Recognition and Retrieval XI, IS&T/SPIE's International Symposium on Electronic Imaging 2004, San Jose California; 2004; Published in Doc-

ument Recognition and Retrieval XI SPIE Vol. 5296, ISBN/ISSN 0-8194-5199-1, pp. 82-91.

[Lalanne et al. 2005] Lalanne, D., von Rotz, D., Ingold, R.. IM2.DI, Intégration de Documents dans des Archives Multimédias de Réunions, in Flash Informatique, Ecole Polytechnique Fédérale de Lausanne, FI2/05, February 2005, pp. 15-18.

[Le and Thoma 2000] Le, D. X., Thoma, G. R.. Page Layout Classification Technique for Biomedical Documents, in Proceedings of SCI, World Multiconference on Systems, Cybernetics and Informatics, Orlando, FL, USA, Vol. 10, July 2000, pp. 348-52.

[Le Meur and Bourillot 2005] Le Meur, J-Y, Bourillot, D.. INDICO, un Logiciel de Pointe pour la Gestion de Conference, in Flash Informatique, Ecole Polytechnique Fédérale de Lausanne, FI2/05, February 2005, pp. 12-14.

[Lin et al. 1997] Lin, C. C., Niwa, Y., Narita, S.. Logical Structure Analysis of Book Document Images using Contents Information, in Proceedings of ICDAR, the International Conference on Document Analysis and Recognition, Ulm, Germany, 1997, pp. 1048-1054.

[Litman and Passonneau 1995] Litman, D. J., Passonneau R. J.. Combining Multiple Knowledge Sources for Discourse Segmentation, in Proceedings of ACL, the 33^{rd} Annual Meeting of the Association for Computational Linguistics, Montreal, Quebec, Canada, 1995, pp. 108-115.

[Little et al. 2002] Little, S., Geurts, J., Hunter, J.. Dynamic Generation of Intelligent Multimedia Presentations through Semantic Inferencing, in Proceedings of ECDL, the 6^{th} European Conference on Research and Advanced Technology for Digital Libraries, Rome, Italy, 2002, pp. 158-175.

[Looney 2002] Looney, C.. Interactive Clustering and Merging with a New Fuzzy Expected Value, Pattern Recognition, Vol. 35, No 11, August 2002, pp. 2413-2423.

[Macedo et al. 2004] Macedo, A. A., Camacho-Guerrero, J. A., Cattelan, R. G., Inacio, V. R., da Graça, C. P. M.. Interaction Alternatives for Linking Everyday Presentations, in Proceedings of ACM Hypertext, USA, 2004, pp. 112-113.

[Macedo et al. 2001] Macedo, A. A., da Graça, C. P. M., Camacho-Guerrero, J. A.. Latent Semantic Linking over Homogeneous Repositories, in Proceedings of DocEng, the ACM Symposium on Document Engenieer, Georgia, USA, 2001, pp. 144-151.

[Mao et al. 2003] Mao, S., Rosenfeld, A., Kanungo, T.. Document Structure Analysis Algorithms: a Literature Survey, in Proceedings of SPIE Electronic Imaging, Maryland, USA, 2003, pp. 197-207.

[Matrakas and Bortolozzi 2000] Matrakas, M. D., Bortolozzi, F.. Segmentation and Validation of Commercial Documents Logical Structure, in Proceedings of ITCC, International Conference on Information Technology: Coding and Computing, Las Vegas, Nevada, USA, 2000, pp. 242-246.

[McQueen 1967] McQueen, J.. Some Methods for Classification and Analysis of Multivariate Observations, in Proceedings of the 5^{th} Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, USA, 1967, pp. 281-297.

[Mekhaldi et al. 2005] Mekhaldi D., Lalanne D., Ingold R.. From Searching to Browsing through Multimodal Documents Linking, in Proceedings of ICDAR, the 8^{th} International Conference on Document Analysis and Recognition, Seoul, Korea, 2005, pp. 924-928.

[Mekhaldi et al. 2004a] Mekhaldi, D., Lalanne, D., Ingold, R.. Using Bimodal Alignment and Clustering Techniques for Documents and Speech Thematic Segmentation, in Proceedings of CIKM, the 13^{th} Conference on Information and Knowledge Management, Washington D.C., USA, 2004, pp. 69-77.

[Mekhaldi et al. 2004b] Mekhaldi, D., Lalanne, D., Ingold, R.. Thematic Segmentation of Meetings through Document/Speech Alignment, in Proceedings of 12^{th} Annual Conference ACM Multimedia 2004, Columbia University, New York, USA, 2004, pp. 804-811.

[Mekhaldi et al. 2004c] Mekhaldi, D., Lalanne, D., Ingold, R.. Unity is Strength: Coupling Media for Thematic Segmentation, in Proceedings of DAS, the 6^{th} IAPR International Workshop on Document Analysis Systems, Florence, Italy, 2004, pp. 559-562.

[Mekhaldi et al. 2003] Mekhaldi, D., Lalanne, D., Ingold, R.. Thematic Alignment of Recorded Speech with Documents, in Proceedings of DocEng, the ACM Symposium on Document Engineering, Grenoble, France, 2003, pp. 52-54.

[Michel 2000] Michel, C.. Cardinal, Nominal and Ordinal Similarity Measures in Information Retrieval Evaluation, in Proceedings of LREC, the 2^{nd} International Conference on Language Resource and Evaluation, Athens, 2000, pp. 1509-1513.

[Mihalcea and Hassan 2005] Mihalcea, R., Hassan, S.. Using the Essence of Texts to Improve Document Classification, in Proceedings of RANLP, the Conference on Recent Advances in Natural Language Processing, Borovetz, Bulgaria, 2005.

[Minnen et al. 2001] Minnen, G., Carroll, J., Pearce, D.. Applied Morphological Processing of English, Natural Language Engineering, Vol. 7, No 3, 2001, pp. 207-223.

[Mukhopadhyay and Smith 1999] Mukhopadhyay, S., Smith, B.. Passive Capture and Structuring of Lectures, in Proceedings of the 17^{th} ACM International Conference on Multimedia, Florida, USA, 1999, pp. 477-487.

[Nagy et al. 1992] Nagy, G., Seth, S., Viswanathan, M.. A prototype Document Image Analysis System for Technical Journals, IEEE Computer, Vol. 25, No 7, 1992, pp. 10-22.

[Needleman and Wunsch 1970] Needleman, S. B., Wunsch, C. D.. An Efficient Method Applicable to the Search for Similarities in the Amino Acid Sequences of Two Proteins, Journal of Molecular Biology, Vol. 48, 1970, pp. 444-453.

[O'Gorman 1993] O'Gorman, L.. The Document Spectrum for Page Layout Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, No 11, 1993, pp. 1162-1173.

[Olligschlaeger 1999] Olligschlaeger, A. M., HauptmannA., G. Multimodal Information Systems and GIS: The Informedia Digital Video Library, in Proceedings of ESRI User Conference,California, USA, 1999.

[Pavlidis and Zhou 1992] Pavlidis, T. Zhou, J.. Page Segmentation and Classification, Graphical Models and Image Processing Vol. 54, No 6, 1992, pp. 484-496.

[Perner 2002] Perner, P.. Data Mining on Multimedia Data, Edition and publisher Springer Verlag 2002, ISBN 3-540-00317-7.

[Petinot et al. 2004] Petinot, Y., Lee Giles, C., Bhatnagar, V., Teregowda, P. B., Han, H., Councill, I.. Citeseer-API: towards Seamless Resource Location and Interlinking for Digital Libraries, in Proceedings of CIKM, 13^{th} Conference on Information and Knowledge Management, Washington D.C., USA, 2004, pp. 553-561.

[Pevzner and Hearst 2002] Pevzner, L. Hearst M.. A Critique and Improvement of an Evaluation Metric for Text Segmentation, Computational Linguistics, Vol. 28, No 1, 2002, pp.19-36.

[Ponte and Croft 1997] Ponte, J. M., Croft, W. B.. Text Segmentation by Topic, in Proceedings of ECDL, the European Conference on Digital Libraries, Pisa, Italy, 1997, pp. 113-125.

[Popescu-Belis 2004a] Popescu-Belis, A.. Abstracting a Dialog Act Tagset for Meeting Processing, in Proceedings of LREC, the 4^{th} International Conference on Language Resources and Evaluation, Lisbon, Portugal, Vol. 4, 2004, pp. 1415-1418.

[Popescu-Belis 2004b] Popescu-Belis, A., Lalanne, D.. Reference Resolution over a Restricted Domain: References to Documents, in Proceedings of ACL Workshop on Reference Resolution and its Applications, Spain, 2004, pp. 71-78.

[Robertson and Sparck 1976] Robertson, S. E., Sparck-Jones, K.. Relevance Weighting of Search Terms, Journal of the American Society for Information Science, Vol. 27, No 3, 1976, pp.129-146.

[Saetre et al. 2005] Saetre, R., Tveit, A., Steigedal, T. S., Laegreid, A.. Semantic Annotation of Biomedical Literature using Google, in Proceedings of DMBIO, Data Mining and Bioinformatics, Singapore, 2005, pp. 327-337.

[Salton et al. 1996] Salton, G., Singhal, A., Buckley, C., Mitra, M.. Automatic Text Decomposition Using Text Segments and Text Themes, in Proceedings of Hypertext '96, the 7^{th} ACM Conference on Hypertext, Washington D.C., USA, 1996, pp. 53-65.

[Schultz et al. 2002] Schultz, T., Waibel, A., Bett, M., Metze, F., Pan, Y., Ries, K., Schaaf, T., Soltau, H., Westphal, M., Yu, H., Zechner, K.. The ISL Meeting Room System, in Proceedings of HSC, the Workshop on Hands-Free Speech Communication, Kyoto, Japan, 2001.

[Shriberg et al. 2000] Shriberg, E., Stolcke, A., Hakkani-Tur D., Tur, G.. Prosody-based Automatic Segmentation of Speech into Sentences and Topics, Speech Communication, Vol. 32, No 1/2, 2000, pp. 127-154.

[Simard 1999] Simard, M.. Text-Translation Alignment, Three Languages are better than Two, in Proceedings of SIGDAT, Conference on Empirical Methods in NLP and Very Large Corpora, University of Maryland, USA, 1999, pp. 2-11.

[Smith and Waterman 1981] Smith, T. F., Waterman, M. S.. The Identification of Common Molecular Subsequences, Journal of Molecular. Biology, Vol. 147, No 2, 1981, pp. 195-197.

[Strömbäck 2005] Strömbäck, P.. The Impact of Lemmatization in Word Alignment, a Master thesis in Computational Linguistics, Uppsala University, Sweden, 2005.

[Tang and Kender 2005] Tang, L., Kender, J., Educational video Understanding: Mapping Handwritten Text to Textbook Chapters, in Proceedings of ICDAR, the 8^{th} International Conference on Document Analysis and Recognition, Seoul, Korea, 2005, pp. 919-923.

[Tokuyasu and Chou 2001] Tokuyasu, T. A., Chou, P. A.. Turbo Recognition: a Statistical Approach to Layout Analysis, in Proceedings of SPIE Conference on Document Recognition and Retrieval, San Jose, California, USA, 2001, pp. 123-129.

[Tsujimoto and Asada 1990] Tsujimoto, S., Asada, H.. Understanding Multi-articled Documents, in Proceedings of ICDAR, the International Conference on Pattern Recognition, Atlantic City, New Jersey, USA, 1990, pp. 551-556.

[Tsz-wai et al. 2005] Tsz-wai, R. L., Ben, H. Iadh, O.. Automatically Building a Stopword List for an Information Retrieval System, Journal on Digital Information Management DIR, Special Issue on the 5^{th} Dutch-Belgian Information Retrieval Workshop, 2005.

[Tucker and Whittaker 2004] Tucker, S., Whittaker, S.. Accessing Multimodal Meeting Data: Systems, Problems and Possibilities, in Proceedings of MLMI, International Workshop on Machine Learning for Multimodal Interaction, Martigny, Switzerland, 2004, pp. 1-11.

[Xie and Beni 1991] Xie, X. L., Beni, G.. A Validity Measure for Fuzzy Clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 13, No 4, August 1991, pp. 841-847.

[Yu 2004] Yu, J. H.. Alignment of Bilingual Web Pages Based on the MT Evaluation Method of BLEU, in Student Workshop of ROCLING 14, Conference on Computational Linguistics and Speech Processing, Taipei, Taiwan, 2004.

[Wahlster 1993] Wahlster, W., André, E., Finkler, W., Profitlich, H. J., Rist, T., Plan-based Integration of Natural Language and Graphics Generation, Artificial Intelligence Vol. 63, 1993, pp. 387-427.

[Zhang et al. 2004] Zhang, B., André, M., Calado, P., Cristo, M.. Combining Structural and Citation-based Evidence for Text Classification, in Proceedings of CIKM, the 13^{th} Conference on Information and Knowledge Management, Washington D.C., USA 2004, pp. 162-163.

[Zhao and Karypis 2001] Zhao, Y., Karypis, G.. Criterion Functions for Document Clustering: Experiments and Analysis, Technical Report, University of Minnesota, USA, 2001.

[Zimmermann et al. 2005] Zimmermann, M., Liu, Y., Shriberg, E., Stolcke, A.. Toward Joint Segmentation and Classification of Dialog Acts in Multiparty Meetings, in Proceedings of MLMI, International Workshop Machine Learning for Multimodal Interaction, Edinburgh, UK, 2005, pp. 187-193.

[Coveo] Coveo tool. Available from http://www.coveo.com/

[EuroWordNet] EuroWordNet thesaurus. Available from http://www.illc.uva.nl/EuroWordNet

[Getty] Getty thesaurus. Available from http://www.getty.edu/research/conducting_research/vocabularies/tgn/

[HTK] HTK tool. Available from http://htk.eng.cam.ac.uk/links/asr_tool.shtml

[IM2] The IM2 project. Available from http://www.im2.ch/

[Fast 2005] Mine Text - Discover Gems, Distributed by FAST Search Best Practices (FAST SBP), October 2005.

[O'Neill and Paice 2001] O'Neill, C., Paice, C.. The Lancaster Paice/Husk Stemming Algorithm, 2001. Available from http://www.lancs.ac.uk/ug/oneillc1/stemmer/general/.

[Scansoft] Scansoft system. Available from http://www.scansoft.fr/

[SMR] The Smart meeting room recorded data. Available from http://diuf.unifr.ch/im2/

[Sphinx] CMU Sphinx system. Available from http://cmusphinx.sourceforge.net/html/cmusphinx.php

[UMLS] UMLS thesaurus. Available from http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html

[WordNet] WordNet thesaurus. Available from http://WordNet.princeton.edu/

[WebLaw] WebLaw thesaurus. Available from http://www.weblaw.edu.au/weblaw/