Multimodal Document Alignment : Feature-based Validation to Strengthen Thematic Links

Dalila Mekhaldi
Computational Linguistics Research Group
University of Wolverhampton
Wolverhampton, UK
dalila.mekhaldi@gmail.com

Denis Lalanne
Departement d'Informatique
Université de Fribourg
Fribourg, Switzerland
denis.lalanne@unifr.ch

ABSTRACT: *In this paper, we present a validation approach of detected alignment links between dialog transcript and discussed documents, in the context of a multimodal document alignment framework of multimedia events (meetings and lectures). The validation approach consists in an entailment process of the detected alignment links. This entailment process exploits several features, from the structural level of aligned documents to the linguistic level of their tokens. The implemented entailment strategies were evaluated on several multimodal corpora. The obtained results prove that the choice of the relevant entailment strategy depends on the types of documents that are available in the corpus, on their content, and also on the nature of the corpus.*

Keywords: Thematic alignment, Validation of links, Structural / contextual / linguistic features of document

Received: 2 November 2009, Revised 29 December 2009, Accepted 2 Janurary 2010

1. Introduction

Multimedia applications have emerged in our life in several domains. This includes the academic domain by involving several multimedia technologies and electronic resources in the learning process in form of databases, eBooks, educational games, etc. In our daily meetings, several multimedia resources are being also exploited, which have been extended in space (i.e. teleconference meetings) thanks to video/audio streaming and to Internet technology. Another example is the social domain where billion of resources are available in form of text (blogs, etc.), videos, images, etc. The variousness in quantity and type of available multimodal information resources in all these applications (textual documents, images, video recording, etc.) requires specific processing in order to archive this information and make it available in an intelligent and useful way after the event. Therefore, many studies and projects are focusing nowadays on finding the best techniques and practices for intelligent archiving and browsing of this multimodal information.

In our previous studies, we have been interested in the development of techniques that facilitates the archiving and access to multimodal information of multimedia applications, especially meetings and lectures. The main technique we used consists in a mash-up of the various modalities, based on their textual content as being a relevant mean that is used by humans to express their queries. Therefore, we have implemented a multimodal document alignment framework that establishes relationships between the documents and resources of these events, more specifically between the dialog transcript and the printable documents being discussed (Mekhaldi 2006, Mekhaldi 2007).In addition to an efficient archiving of the material of multimedia events, our assumption behind the multimodal document alignment is that classical printable documents, called also static documents, play a central role in the thematic structure of discussions (Mekhaldi 2006). Further, we believe static documents could provide a natural and thematic mean for browsing and searching through multimedia meeting repository

(Ingold, von Rotz, Behera, Mekhaldi and Popescu 2005, Rigamonti, Lalanne and Ingold 2007). The current studies for building browsers for multimedia meeting corpora are often based on low-level visual features of image and video, which lack of semantic information.

Other research projects use language understanding techniques or text caption derived from OCR, in order to create more powerful indexes and search mechanisms. The development of our multimodal document alignment techniques (Mekhaldi 2006) follows the idea of linking documents with other multimedia meeting data in order to enhance browsing capabilities. Our multimodal alignment framework covers the detection of relationships between resources at several dimensions of their content, semantic (called thematic alignment), lexical (called quotation alignment) and structural (called reference alignment). This detection of alignment links between documents benefits from various methods and techniques from natural language processing, information retrieval and document image analysis. FaericWorld (Rigamonti, Lalanne and Ingold 2007) and FriDoc (Ingold, von Rotz, Behera, Mekhaldi and Popescu 2005) are multimodal browsers that allow navigating through a large multimedia corpus of meetings, and in a meeting respectively. For instance, FriDoc allows replaying a meeting using documents as structure vectors to multimedia content (see Fig. 1). Thanks to the alignment techniques (Mekhaldi 2006), all the representations are synchronized, meaning they all have the same time reference, and clicking on one of them causes all the components to visualize their content at the same time. Clicking on a document part positions audio/video clips at the time when it was discussed, positions the speech transcription at the same time, and displays the document that was projected. Several user evaluations performed on our document centric meeting browsers have shown the usefulness of document alignment as a way to improve question answering while replaying a meeting.

However, the achievement of the previously described goals, i.e. an efficient archiving, browsing and navigation in multimedia events, depends drastically on the quality of the alignment links being generated between the multimodal documents. Although the high performance of our multimodal document alignment framework, the automatic detection of alignment links between documents has generated ambiguous links which need to be identified and filtered out. From another side, many relevant links were missed out by the alignment process. In both cases, specific processing should be performed on the detected alignment links in order to prune them, which we call in this study *entailing alignment links*. Therefore, the aim of this study is to present several entailment methods that deal with ambiguous alignment links as well as those missing. The implemented entailment methods exploit several features of the aligned multimodal documents, specifically structural, contextual and linguistic features.

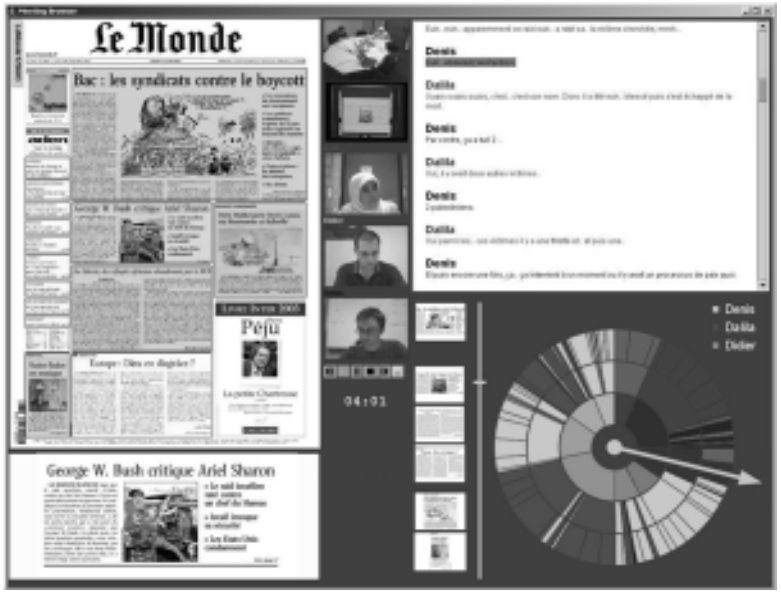


Figure 1. FriDoc multimodal browser, example of a press review meeting

This paper is organised as follow, in the next section we present a survey of some alignment studies in different domains, and how do they use additional features of aligned data in order to prune the bad alignment relationships. In section 3, the

related works multimodal corpora that were targeted in our study are described. A brief summary of our previous work on multimodal document alignment is presented in section 4. Finally in section 5, the entailment methods that were implemented to prune our multimodal alignment links are presented and evaluated.

2. Pruning alignment links

The study made by Snyder and Barzilay (Snyder and Barzilay 2007) addressed the task of aligning database with corresponding text, by linking individual database entries with sentences that verbalize the same information. In order to minimize the number of incorrect labels predicted, local ranking models, treated in this work as a pruning step, were used. Therefore, intelligent threshold selections were made based on the ranking scores as well as on relations between entries. In another study that aligns documents with similar content across two sets of bilingual comparable corpora from daily news texts (Vu, Aw and Zhang 2009), candidate links between both corpora are pruned based on two distinct filters, temporal-based and linguistic-based filters. In the temporal-based filter, a date-window was defined in order to limit the set of target candidate documents to those published in a close date to the source document. In the linguistic-based filter, only the target candidates containing the translation of at least one of the title words of the source document are considered. The experiments performed showed an improvement of the alignment scores when these two filters are considered. In a study related to optimal sequence alignment (Davidson 2001), heuristics have been used in order to find the optimal path in the dynamic programming matrix, by pruning unnecessary areas of the matrix. The used pruning heuristics are based on an algorithm that searches forward from a start state, always expanding the most promising node first. Each node on the frontier of the search is stored in an open list and sorted according to a computed cost. Once expanded, all its children are added to the frontier. To expand a given node n, its children are placed into the open list. If any children have a cost value higher than the lowest cost found so far, then it will be pruned by not placing it into the open list. Finally, the node n is placed onto the closed list. When there are no longer nodes left on the open list, the best path found to the goal will be the optimal path. Pruning of the dynamic programming matrix in RNA sequences alignment was also considered in the study of Havgaard et al. (Havgaard, Torarinsson and Gorodkin 2007), where it was used to discards any sub-alignment that does not have a score above a length-dependent threshold. This pruning method was useful mainly when there was not enough sequence similarity to make the necessary alignments. It was empirically proved that this dynamic pruning increased the processing speed whilst the algorithm retains its good performance.

3. Multimodal corpora

Two main multimedia applications have been chosen as basis for our multimodal document alignment framework, meetings and lectures, from four different domains and in two different languages. This includes three distinct meeting corpora, one French and two English (section 3.1), and an English scientific conference presentation corpus (section 3.2).

3.1. Meetings corpora

Three meeting corpora were chosen in our study, respectively press reviews, movie club and furniture proposal meetings. The press review meetings were considered in our evaluation as being a multimedia event, where static documents are present and used during almost the entire meeting (Fig. 2). In the Smart meeting room in Fribourg (SMR), twenty two press review meetings were recorded, where several participants discussed various newspaper articles. Documents that are either discussed or projected are captured, and then made available in the form of PDF files, linear text version and XML logical structure. The speech transcript is available in XML form. During the twenty two recorded meetings, several French newspapers were used, Le Monde (France), Le Devoir (Canada), Le Soir (Belgium) and La Presse (Tunisia), where the main articles are discussed. In eighteen meetings, only one document was used. In the four remaining meetings up to four documents were presented. In each discussed document in this corpus (and in all other corpora), the weight of important segments (titles, subtitles, etc.) was taken into account by multiplying their tokens according to their importance, e.g. tokens of the main title are multiplied three times. Since each document is composed of many thematically heterogeneous articles, two main scenarios were defined and followed by speakers, stereotyped (without any interruption from other participants) and non-stereotyped (with more interactivity between participants).

The second meeting corpus considered in our study is a movie club corpus (about decision-making on movies to display) containing one meeting, which was recorded at the Idiap Smart Meeting Room (Moore 2002). Eight documents are available in this corpus, the agenda of the meeting, two other documents presenting a selection of movies, three posters for a selected

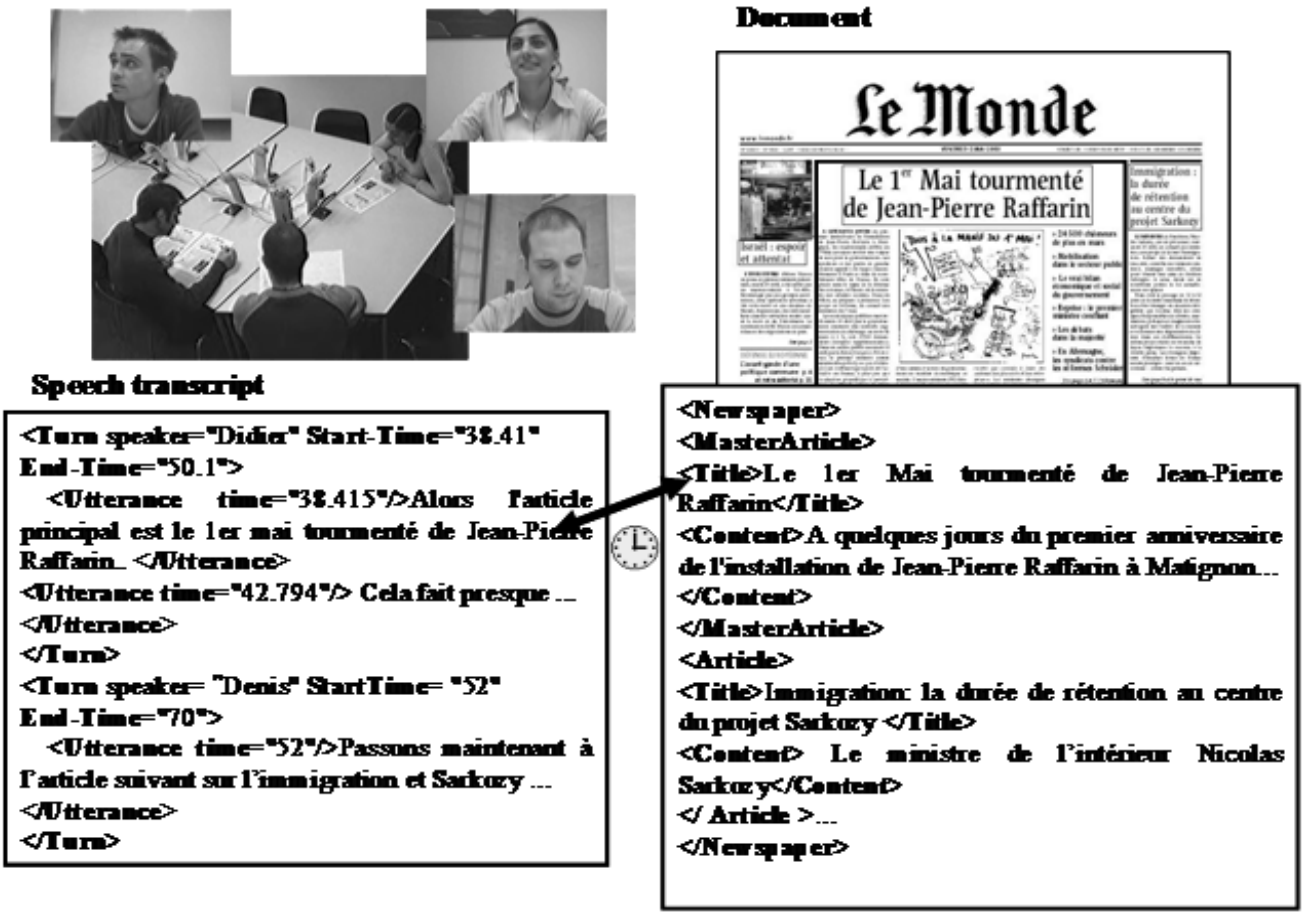


Figure 2. Press review meeting

movie, and two slideshows. Since the eight documents were presented and discussed sequentially by the different speakers, we decided to combine them within one file.

The third meeting corpus considered in our study consists in a furniture proposal corpus (decision-making about furniture to buy), which was registered by ISSCO Research group (Popescu, Georgescul, Clark and Amstrong 2004) at the Idiap Smart Meeting Room (Moore 2002). In this corpus, the furniture that should be chosen for a reading room was discussed. Two main meetings were considered in this corpus, where several resources were available. In the first meeting, a slideshow document was available, which presented the motivation for this meeting, as well as the constraints for choosing the furniture. In the second meeting, three slideshow documents were presented, the first one correponding to the agenda of the meeting, the other two slideshows presenting some ideas for choosing the furniture (furniture snapshots, prices, etc.). Moreover, three static documents were considered in this meeting, which represented more general ideas about how to choose the furniture for a meeting room. Similarly to the movie club corpus, and since the slideshows are discussed independently from the static documents, all the static documents and slideshows were combined within one file.

3.1 Scientific conference presentations corpus

Our multimodal alignment was also tested in the context of the SMAC project (Smart Multimedia Archive for Conferences, Fig. 3) (Abou Khlaed, Le Meur, Scheurer, Bourillot, Lalanne, von Rotz, Ingold and Baron 2006), where the material of CHEP'04, a scientific conference in physics of particles, was considered (Mekhaldi 2007). Within this corpus, eight scientific presentations were selected, in which three resources were available: scientific papers, slideshows and speech transcript. Due to the lack of clarity of attendees' speech (e.g. the audience when asking questions), only the presenters' speech was considered.

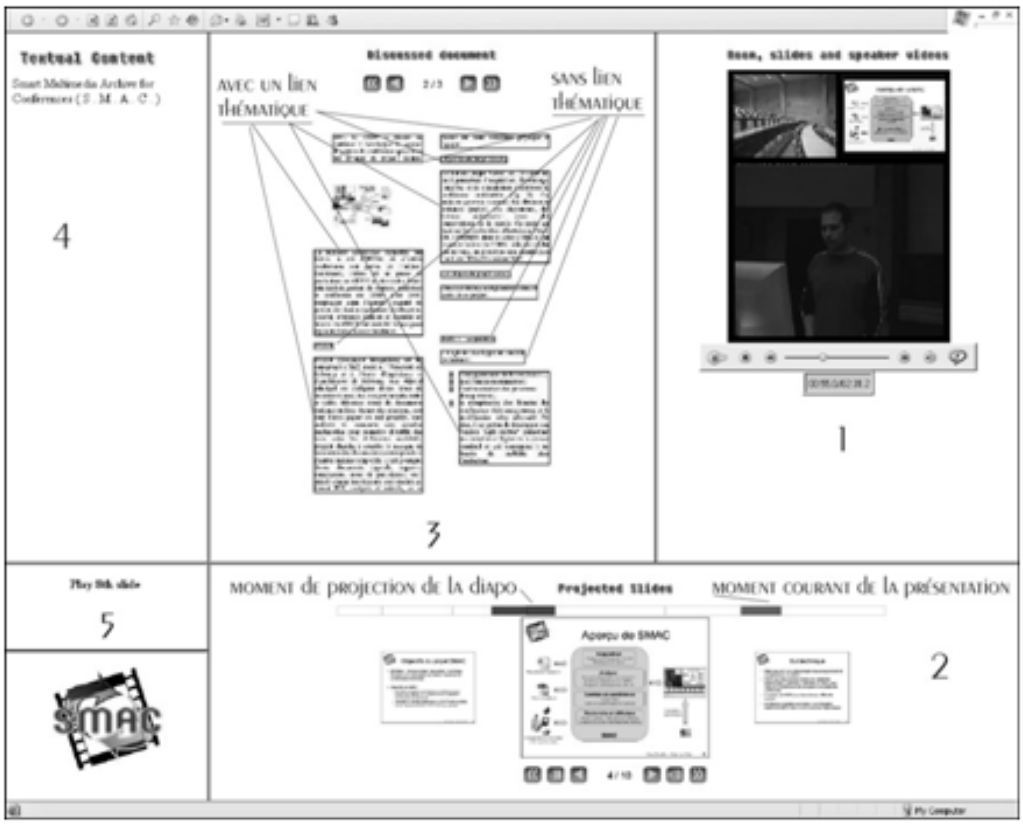


Figure 3. Browsing interface for scientific conferences (SMAC project)

In order to detect alignment relationships between the various documents, the latter should be first decomposed into segments according to several criteria. The static documents are segmented into logical blocks and into sentences. A logical segmentation of a static document corresponds to a labelling process of its components using layout rules, which indicates blocks function and meaning at the page level (Bloechle, Rigamonti, Hadjar, Lalanne and Ingold 2006). This labelling includes title, summary, sections, etc., and the respective slides in case of slideshows. This logical structure of documents was manually extracted in our work. From another side, the speech transcript is segmented into speaker turns and utterances. A speaker turn is defined as being a speech part that belongs to one speaker without interruption from other speakers. Speaker utterances are small units that compose a turn, which correspond to the smallest homogeneous parts within a turn. The transcriber tool (Barras, Geoffrois, Wu and Liberman 1998) was used in our work in order to manually transcribe dialogs and generate speaker turns and utterances.

An overview of the segmentation of the various corpora are respectively shown in Tables 1, 2, 3 and 4. For each of the evaluated corpora, appropriate segmentation methods are used. For press reviews, the speech transcript was segmented respectively into turns and utterances. The static documents were segmented into sentences and into logical blocks, where a logical block corresponds to a newspaper article. The segmentations considered for each of the furniture proposal, the movie club and the scientific presentation corpus, are the logical blocks of documents and slideshows which correspond to document sections and slides respectively. The speech transcript in these three corpora was segmented into speaker utterances.

4. Multimodal document alignment

Our multimodal document alignment aims to establish relationships between several documents that are recorded (speech), discussed (static documents) or displayed (slideshows) during a meeting or a lecture. Studying the relations that might exist between these documents covers many types of links based mainly on their thematic similarity, quotations from or references to static documents.

Static documents		Speech transcript			
Documents	Logical blocks	Sentences	Duration	Turns	Utterances
29	379	2772	240 min	1888	2936

Table 1. Statistics of the press review corpus

Static documents		Speech transcript	
Documents	Logical blocks	Duration	Utterances
8	48	48 min	1348

Table 2. Statistics of the movie club corpus

Static documents		Speech transcript	
Documents	Logical blocks	Duration	Utterances
7	51	64 min	1810

Table 3. Statistics of the furniture proposal corpus

Static documents		Slideshows	Speech	
Average length	Logical blocks	Slides	Duration	Utterances
6 pages	178	324	237 min	1952

Table 4. Statistics of the scientific conference corpus

4.1. Thematic alignment

The thematic alignment of static documents with speech transcript is defined as being a matching process between their respective segments that share the same themes. This alignment type is mainly based on the computation of the thematic similarity between compared segments, in order to elicit which segments share the same theme (Mekhaldi 2006). For instance, the following speaker utterance and document logical block are thematically similar, since both of them are about "the 1st of May and Raffarin":

- <utterance> Alors l'article principal est le 1er mai tourmenté de Jean-Pierre Raffarin.</utterance>
- <title> Le 1er Mai tourmenté de Jean-Pierre Raffarin. </title>

Following the above process, the segments of the documents being discussed will be linked to the speech transcript segments that are thematically similar and vice versa, generating two different directions of thematic alignment. In order to compute the thematic similarity between two segments, each segment should be first processed by removing its stop words and then reducing the remaining words into their stems. Later on, existing similarity metrics are applied on the vector representations of the respective segments. In our study, three similarity metrics which are based on term co-occurrences were used, Cosine, Jaccard and Dice Given two segments S1 and S2 from the document and the speech transcript respectively, and Wt,S1 the term frequency associated to a term t in the segment S1, the similarity between S1 and S2 is computed according to the following formulae that generates a similarity value between zero and one:

$$Co\sin e(S_1, S_2) = \sum_{t=1}^{N} W_{t,S_1} * W_{t,S_2} \Big/ \sqrt{\sum_{t=1}^{N} W_{t,S_1}^2 * \sum_{t=1}^{N} W_{t,S_2}^2}$$

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

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

In our experiments, the term weight $Wt,_{S1}$ was considered as being the number of occurrences of term t in segment S_1. Later on, the TF.IDF coefficient was chosen to associate weights to terms.

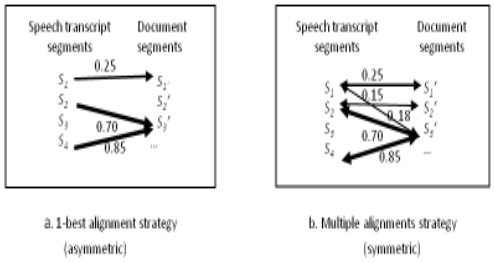
In addition to the three similarity measures used, two other functions have been defined in our work in order to measure the overlap between the compared segments, *membership* and *ownership* functions. The membership of a segment S_1 in a segment S_2 measures the percent of terms of S_1 being present in S_2, whilst the ownership measures the percent of terms of S_2 being present in S_1:

$$Membership\ (S_1, S_2) = |S_1 \cap S_2| \,/\, |S_1|$$

$$Ownership\ (S_1, S_2) = |S_1 \cap S_2| \,/\, |S2|$$

These two functions constitute complementary ways of measuring the thematic similarity between textual segments. A use case of these functions is presented in section 5.1.

After computing the similarity between the respective document segments, only links with significant similarity value are considered. In order to filter out the insignificant links, two strategies were defined based on the similarity value (Mekhaldi 2006), respectively the 1-best and the multiple alignments strategies. In the 1-best alignment strategy, only the link having the highest similarity value for each source segment is considered (e.g. in Fig. 4.a, $S_{3'}$ is the most similar segment for S_2, with a score of 0,70). In this strategy, the alignment relationship is asymmetric. In the multiple alignments strategy, all the relevant links between a source segment and the target segments are preserved (e.g. S_1 in Fig. 4.b is linked to $S_{1'}$ and $S_{3'}$). The selection of the relevant links in this strategy is based on filtering out the insignificant links which have a similarity value less than a determined similarity threshold (defined at 0.10 after observation of the scores). If the same similarity threshold is used in both directions within this strategy, then the alignment relationship will be symmetric, i.e. the same links are detected from speech transcript to the documents and vice versa.



a. 1-best alignment strategy (asymmetric)

b. Multiple alignments strategy (symmetric)

Figure 4. Selection strategies of thematic links

4.2. 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 thematic alignment is based on a thematic similarity of pairs of segments, quotation alignment takes into account lexical similarity and term order within compared segments (Mekhaldi, Lalanne, and Ingold 2005). Quotation alignment detection is deterministic and thus can be used to strengthen the thematic alignment links. In order to retrieve significant quotations, the minimal size of a quotation was fixed at three terms. After removing stop words from segments, and reducing the tokens into their stems, our quotation detection algorithm compares each speaker utterance with all document sentences. The matched pairs of sequences, having at least three terms in common in the same order, are considered as quotation/quoted pairs. An example of quotation alignment is shown in the following example:

• <utterance> donc le premier article, c'est un galion au large de Nieuport, un trésor archéologue et historique exceptionnel sous les eaux belges. </utterance>
• <sentence>Un galion au large de Nieuport, un trésor archéologue important</sentence>

4.3. Reference alignment

The third type of alignment, which links documents discussed during events to speech transcript, relates to the references given by speakers on the document logical blocks. An example of this alignment type is shown if Fig. 5. This alignment, studied in collaboration with the University of Geneva (Popescu and Lalanne 2006), is defined as the relationship that is

```
<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. Extract from a reference alignment file of a press review meeting

established between a speaker segment containing a referring expression, and the document segment being referred to. This kind of relationship is frequent in meetings in which static documents contain various articles (e.g. newspapers), in meetings dealing with various documents, or during presentations when the speaker refers to various slides.

Reference detection is required to reinforce the links between speech transcript and the documents discussed or projected, especially if their respective segments are not thematically similar. Reference alignment might also be useful in order to identify the meeting scenario, and how the various documents or document parts (e.g. newspaper articles) are chained. However, references still have a low frequency, compared to thematic links.

In order to detect reference alignment links, a list of all pairs (referring expressions, referred entities) should be extracted (Popescu and Lalanne 2006). 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 document logical blocks. The reference detection algorithm has two main steps: detecting referring expressions within the speech transcript, and then matching them with the corresponding logical blocks. At the first stage, a list of patterns is created. These patterns correspond to entities’ names (e.g. “article”, “table”, “figure”, etc.), or entities’ description, such as the position (e.g. “the first”, “the last”, etc.). After that, patterns rules are applied to speaker utterances. 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.), such as the expression “this article” detected in utterance 14 in Fig. 5, then it is matched with the current document element. If the referring expression is non-anaphoric, for instance the expression “les radios généralistes” detected in utterance 13 in Fig. 5, 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 and Lalanne 2006).

4.4 Annotation of Alignment Links

The evaluation of each of the detected alignment types consists of comparing automatically generated alignment links to the corresponding manual ground truth that contains all possible alignment links. The manual annotation was performed by an expert annotator at several levels of granularity according to the evaluated corpus, at utterances and/or turns for the speech transcript, and at sentences and/or logical structure for the document. The manual annotation of a segment from a source file involves detecting the id(s) of the best segment(s) from the target file that could be aligned with it (i.e. the best match according to an expert annotator). The comparison of the generated alignment links with the manual ground truth consists then of associating a binary score to each detected alignment link: “1” if it is correct, i.e. it matches with the manual ground truth alignment link, and “0” otherwise.

In order to measure and evaluate the performance and quality of the thematic and quotation alignment processes, recall, precision and the efficiency measure F were employed:

recall = |automatic $\cap$ manual|/|manual|
precision = |automatic manual|/|automatic|
F = 2 . (precision. recall) / (precision + recall)

where automatic is the set of detected links by our alignment algorithm, and manual is the set of links of the manual ground truth. From another side, the reference alignment links were evaluated using the accuracy measure.

After the detection and evaluation of the various alignment types between static documents and speech transcript, a validation step is necessary in order to disambiguate and filter out insignificant links, and to add missing links. In the current study, we focus on a validation or entailment approach that was defined based on various features of the corpora in order to validate the generated thematic links.

5. Entailing thematic links

When a source segment is thematically aligned with more than one target segment, the multiple alignment link strategy should be used, in which non-relevant alignment links are truncated based on a similarity threshold. In order to improve the alignment results obtained with this strategy, several validation or entailment methods were defined and used to prune the generated thematic links.The entailment of a target T with a hypothesis H (denoted T $\Rightarrow$ H) is defined as being the process of making a decision whether H could be deduced from T or not (Dagan and Glickman 2004).

Textual entailment is an example of this research approach which is exploited in natural language processing field (question answering, summarization, etc.), whereas T and H correspond to two textual segments.

The entailment approach that is used in this study is based on other features. Five entailment methods were defined in order to prune thematic alignment links based on several features of the aligned documents, mainly structural, contextual and linguistic. Structural features take into account the general structure of aligned documents and their superposition (section 5.1). Contextual features are related to the aligned segments and the relationships between them (section 5.2, 5.3 and 5.4). Finally, the linguistic features correspond to token characteristics (e.g. syntactic, lexical or semantic), or to higher level semantic features, especially shared named entities and terminology between segments (section 5.5) which is more significant for specific domain corpora.

5.1. Thematic alignment levels-based entailment

Our first entailment method is based on the structure of aligned documents. This entailment method was applied to the results of the thematic alignment between static document and speech transcript in the press review corpus, at two levels of granularity, logical block/ turn, and sentence/ utterance alignments. These two levels are illustrated in Fig. 6.a by a hierarchical representation. Furthermore, the generated alignment links at each of the two levels are symmetric when using the multiple alignment strategy to select relevant links. These two criteria, the hierarchy of levels and the symmetry of links at each level, were exploited in order to prune incoherent alignment links (Fig. 6.b, 6.c and 6.d). Supposing that we have an alignment link between an utterance utt and a sentence sent (respectively an alignment link between a turn T and a logical block L), 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 the following three entailment strategies at the logical block/turn level (Mekhaldi, Lalanne and Ingold 2005), where Sim(T, L), Memb(T, L) and Align(T, L) are respectively the similarity value, the membership value and the alignment decision between turn T and logical block L. θ_1 is a similarity threshold, and θ_2 and θ_3 are two thresholds for membership scores:

$$S_1 : \left[Sim(T,L) \geq \theta_1 \wedge \left(\exists (utt,sent) \subseteq (T,L), Sim(utt,sent) \geq \theta_1\right)\right] \Rightarrow Align(T,L) = 1$$

$$S_2 : \left[Sim(T,L) \geq \theta_1 \wedge \left(\exists (utt,sent) \subseteq (T,L), Sim(utt,sent) \geq \theta_1\right) \wedge Memb(T,L) \geq \theta_2\right] \Rightarrow Align(T,L) = 1$$

$$\left[Sim(T,L) \prec \theta_1 \wedge \left(\exists (utt,sent) \subseteq (T,L), Sim(utt,sent) \geq \theta_1 \wedge Memb(utt,sent) \geq \theta_3\right)\right] \Rightarrow Align(T,L) = 1$$

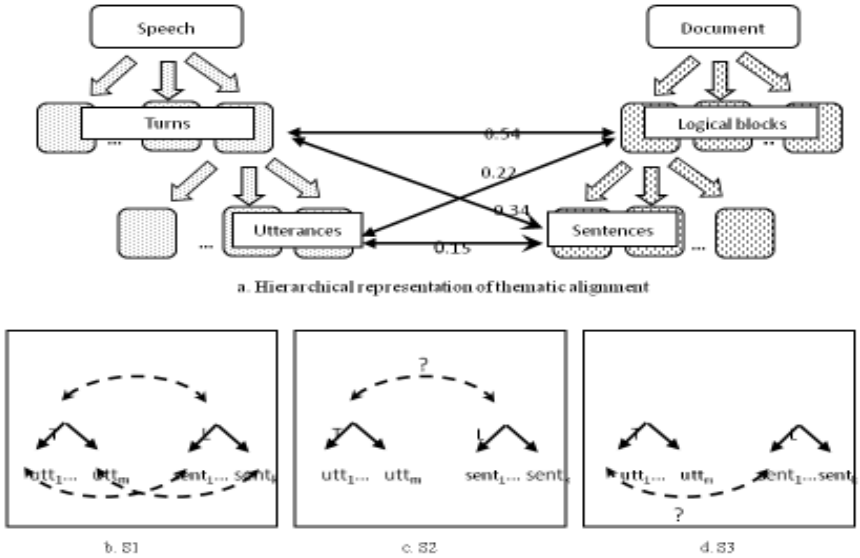


Figure. 6. Entailment based on thematic alignment levels grouping

Our assumption in the first entailment strategy S1 is that only links between turns and logical blocks whose descendants are thematically linked via at least one link, are considered as correct links and thus are preserved. For instance, the link between T and L in Fig. 6.b is preserved, since their descendants are linked.

In the second entailment strategy S2, the links between turns and logical blocks whose descendants are not linked, are suspected to be false links, and thus should be removed (Fig. 6.c). This kind of link incoherence is due to two main reasons. First, utterances of the turn T might have small sizes, therefore they share few terms with sentences of the logical block L, 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 therefore will be neglected. When the utterances and sentences are grouped in T and L respectively, the number of shared terms between T and L increases, and thus their similarity will be significant. The second reason for the incoherence of links between levels is that there is no shared theme between utterances of turn T and sentences of logical block L, even if there are few terms shared between them. Therefore, when utterances and sentences are grouped in T and L respectively, the number of shared terms between the latter increases which generates a false similarity link between them. According to strategy S2, and before removing the incoherent link between turn T and logical block L, other parameters are checked, such as the membership value between these segments in comparison to a threshold $\theta 2$.

In the last entailment strategy S3, if two descendant segments are linked but not their ascendants (Fig. 6.d), then their membership value should be checked in order to build a link between their ascendants' segments. Thus, in Fig. 6.d, a new link will be created between turn T and logical block L, if the membership value between utt1 and sent1 exceeds a threshold $\theta 3$.

The three entailment strategies S_1, S_2 and S_3 for levels grouping were applied successively on the thematic alignment results of the twenty two press review meetings, as being the corpus in which many alignment levels exist. The alignment pairs grouped in this entailment method were logical block/ turn using Cosine metric and sentence/ utterance using Jaccard metric (as being the metrics generating the best scores), both without then with the consideration of the TF.IDF. However, the evaluation of this entailment method focused on turns/logical blocks only, since it is less subjective from point of view of manual annotation. The results of this entailment method are presented in Table 5, and discussed in the following sections.

5.1.1 Using alignment results with Cosine

The initial average values for recall, precision and F, using the Cosine metric without TF.IDF, are 0.63, 0.55 and 0.56 respectively with a confidence interval of ±0.03 for the three measures. After applying the strategy S_1, those same values have significantly 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 one of these mono-utterance turns, T might be aligned with a similar logical block L, but its single utterance might not be aligned with any sentence from L. According to strategy S_1, the

Initial (Cosine)	Thematic alignment levels grouping			Alignment types merging	
		S1	S2	S3	
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. Entailment results based on thematic alignment levels grouping and alignment types merging respectively (logic/turn level for press review corpus)

link between T and L should be removed, ignoring its correctness.

In order to avoid the removal of correct links, strategy S_2 was used, where the membership value of the turn T in its similar logical block L is taken into account, before removing the link between them. If the value of the membership function of T in L exceeds the defined threshold θ_2, then the link between them is preserved, otherwise, it is removed. In order to observe the effect of the threshold on the alignment results, θ_2 was set at different values ranging from 0% to 100%. We realized that its ideal value is 20%, where the F metric reaches its maximal value, insignificantly increasing from 0.56 to 0.58±0.03, whereas the recall significantly decreased from 0.63 to 0.56±0.03, and the precision significantly increased from 0.55 to 0.65±0.03. The improvement of the precision indicates that many false links are being removed, whereas almost all the correct links are being preserved.Finally, we were interested in the insertion of new thematic links between turns and logical blocks, by applying the third strategy S_3 (Fig 6.d). The threshold θ_3 was set at different values ranging from 0% to 100%, whilst q_2 was fixed at 20% (the ideal value under S_2), in order to keep the correct links preserved by strategy S_2. When θ_3=75%, the F value insignificantly increased from 0.58 to its maximal value 0.59±0.03, the recall value significantly increased from 0.56 to 0.61±0.03 and the precision value significantly decreased from 0.65 to 0.61±0.03 (Table 5).

The final F value obtained for the logical block/turn alignment (0.59) after using the three entailment strategies S_1, S_2 and S_3, is the same as the F value obtained with Cosine+TF.IDF metric. These results prove that the grouping of thematic alignment levels has a similar effect as the consideration of TF.IDF coefficient on disambiguation and pruning of generated thematic links.

5.1.2 Using alignment results with Cosine+TF.IDF

When the three entailment strategies, S_1, S_2 and S_3 are applied on the alignment results obtained with Cosine+TF.IDF, no one of them succeeded in improving the F value, which remained stable at 0.59. This might be due to the performance and the efficiency of the TF.IDF coefficient in detecting and covering most significant thematic links.

5.2 Multiple document alignment-based entailment

In the scientific conference corpus, slideshows are introduced as a third data resource, in addition to static documents and speech transcript. Once the thematic alignment results are obtained for each document combination (documents/slideshow, slideshow/speech and documents/speech), and in order to exploit the rich information available when this third resource is considered (i.e. slideshow), the alignment pairs from the various document are grouped, according to the source segment for each alignment pair (Mekhaldi 2006, Mekhaldi 2007). In Fig. 7.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 thematic alignment pairs might be exploited in order to entail the detected links, as well as to add missing links within each document alignment pair.

Based on the cycle structure, two entailment strategies were defined in the alignment grouping. In the first entailment strategy, only links that construct a complete cycle with three arcs are preserved (e.g. (i, j), (j, k) and (i, k) in Fig. 7.a). The other links, such as (i+1, j+1), are removed. In the second entailment strategy (Fig. 7.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. Thus, in Fig. 7.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).

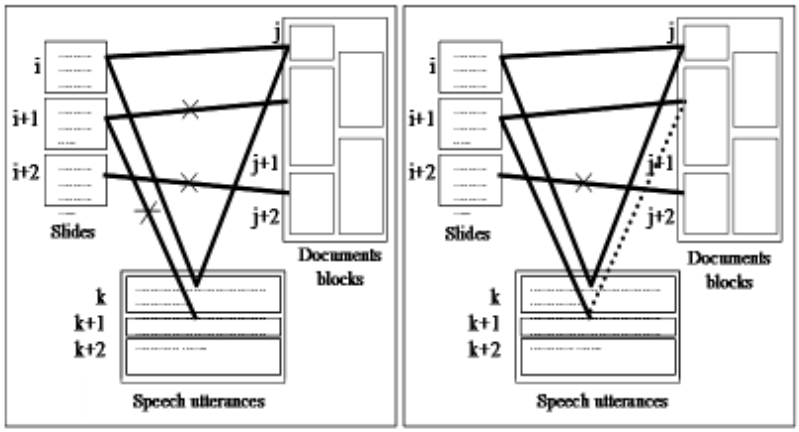


Figure 7. Multiple document alignment based entailment

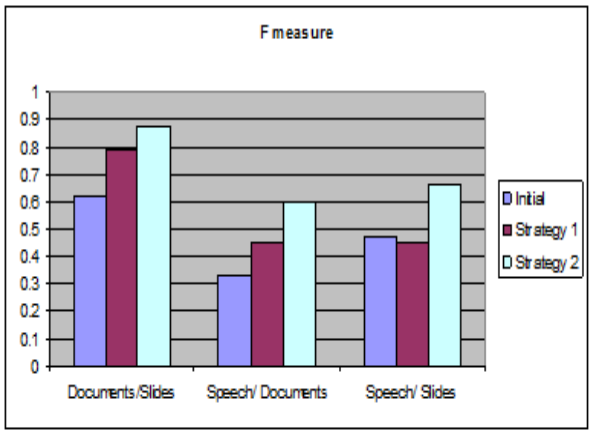


Figure 8. Improvement of F value after grouping alignments (conference corpus)

This entailment method was evaluated on the scientific conference corpus that contains 3 resources (static documents, speech transcript and slideshows), where the thematic links are generated by Cosine+TF.IDF. After having performed the first strategy of the entailment process, i.e. keep only the links that belong to a cycle, many false thematic links between the three resources were removed, which led to a significant increase in the F value from 0.62 to 0.79 ±0.04 for document/slideshow alignment, and from 0.33 to 0.45 ±0.02 for document/speech alignment. For the slideshow/speech alignment, only few correct links were removed, which insignificantly decreased the F value from 0.47 to 0.45 ±0.02.

When the second entailment strategy was used, which adds the missing arc to each path having two arcs (Fig. 7.b), several missing thematic links were added. Thus, the F values were significantly increased from 0.79 to 0.88 ±0.04 for document/slideshow, from 0.45 to 0.60 ±0.02 for document/speech, and from 0.45 to 0.66 ±0.02 for slideshow/speech alignments (Fig. 8).

These satisfactory results of the multiple document alignment-based entailment shows that each time a new resource is integrated into the thematic alignment process, better results are obtained. Furthermore, combining the thematic alignment results of the three modalities (documents/slideshow, slideshow/speech and documents/speech) drastically improved the obtained results for each individual pair, and thus the synchronization of the static documents with multimedia data.

5.3 Neighbouring-based entailment

Our third entailment method is based on the thematic similarity between neighbour segments of the source and target segments being aligned. Given a detected alignment link between a source segment S_i and a target segment S_j (Fig. 9), an entailment score between S_i and S_j is computed as being the average of the thematic similarity values obtained for the 8 potential links represented by dashed links in Fig. 9:

$$Entail(S_i, S_j) = \frac{\sum_{k=j-1}^{j+1} Sim(S_{i-1}, S_k) + Sim(S_i, S_{j-1}) + Sim(S_i, S_{j+1}) + \sum_{k=j-1}^{j+1} Sim(S_{i+1}, S_k)}{8}$$

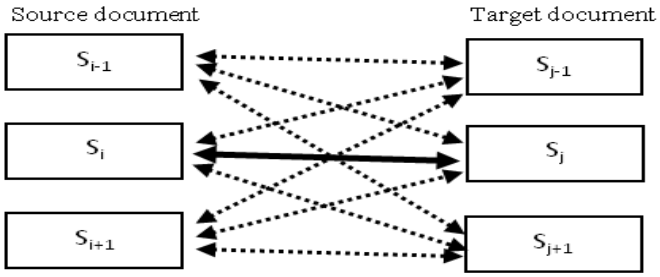


Figure 9. Thematic alignment links observed in the neighboring-based entailment

where $Sim(S_i, S_k)$ corresponds to the Cosine+TF.IDF similarity value between segments S_i and S_k.

The obtained entailment score between segments S_i and S_j is compared to a threshold θ that is varied dynamically, in order to decide whether to keep the alignment link between them or to consider it as non-relevant link and thus reject it.

The neighbouring-based entailment method was performed on the three corpora that contain slideshows, either combined with static documents (in case of movie club and furniture proposal corpora), or considered as a third data resource (scientific conference corpus). The results of the evaluation of this method, based on F measure, are presented in Table 6. In contrary to what we assumed, the consideration of contextual information for pruning thematic alignment links insignificantly improved the results for most of the document pairs. Therefore, the F value insignificantly increased from 0.62 to 0.64±0.02 for slideshow/speech alignment, and from 0.47 to 0.50±0.06 for document/slideshow alignment in the scientific conference corpus. Similarly, document/Speech alignment in the movie club corpus insignificantly increased from 0.37 to 0.40±0.06. The F values for the document/ speech alignment in the scientific conference and the furniture proposal corpora are still stable at 0.33 and 0.40 respectively. However, our assumption is that these preliminary results might be improved if the metric used to compute the neighbouring entailment score is enhanced by attributing various weights to the various neighbour links.

Corpus		**Scientific conference**			**Movie club**	**Furniture proposal**	**Press reviews**
Documents pair		Document/ Slideshow	Slideshow/ Speech	Document/ Speech	Speech/ Document	Speech/ Document	Speech/ Document
Alignment pair		Logic/ logic	Logic/ utterance	Logic/ utterance	Logic/ utterance	Logic/ utterance	Logic/ turn
Number of alignment links		307	1718	1630	293	187	1160
Cosine+TF.IDF (initial)		0.62	0.47	0.33	0.37	0.40	0.59
Entailment method	Multiple document alignment	0.79	0.45	0.45	-	-	-
	Neighbouring	0.64	0.50	0.33	0.40	0.40	-
	Entity Matching	0.46	0.35	0.29	0.40	0.37	0.43

Table 6. F measure results of three entailment methods (multiple document alignment-based, neighbouring-based, and entity matching-based). Empty cells correspond to non-appropriateness of the method to the corpus

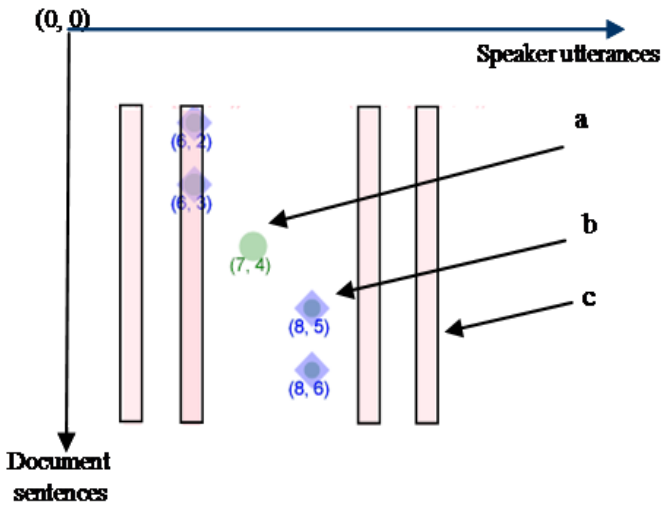


Figure 10. Merging various document/speech transcript alignment types: a. a thematic link; b. a quotation and a thematic link; c. a reference link

5.4. Alignment types merging- based entailment

The method defined in the current section, for entailing thematic alignment links, is based on merging the various alignment types detected between static documents and speech transcript (i.e. thematic, quotation and references). Therefore, the thematic links detected between respective segments are pruned according to the existence of other alignment typesbetween these segments. This means that if a speech segment is thematically linked to a document segment, then this thematic link is preserved only if there is a quotation link and a reference link between the two segments. The merging of the three alignment types might help also to measure their complementary.

This entailment method was evaluated only on the press review corpus, as being the corpus in which the three alignment types were identified and studied. Only the speech transcript/document alignment is concerned by this entailment method, since the speech transcript is the source modality for the quotation and reference alignments. Therefore, each of the three alignment processes was independently performed from speech transcript to documents. Later on, the results of the three alignment processes are merged according to their source unit, if it is a speaker turn or utterance (Table 5).

In order to visualize the results generated by the merging of the various alignment types, an SVG tool was implemented (Fig. 10). Within this tool, the speech transcript and documents are represented respectively by the X and Y axis. 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 number of sentences.In some cases, quotation alignment might be considered to be a special kind of thematic alignment (e.g. Fig. 10.b), which means that the two alignment types should be coherent. When such an assumption is considered, i.e. entailing thematic links according to the existence of quotation links, the F value (for turns/logical blocks multiple alignment strategy) remains stable at 0.59 for both alignment variants (Cosine, Cosine+TF.IDF), which proves that most of the detected quotations in our corpus correspond to already detected thematic links. From another side, we assumed that a reference link does not necessarily have to correspond to a thematic link. In Fig. 5, utterances 13 and 14 are linked thematically to sentences of the 6th logical block, even though there is a reference in utterance 13 to the 5th logical block. Therefore, merging reference alignment with the other alignment types does not change the scores. Nevertheless, reference alignments may add other information for users, such as how speakers chained up the various document articles, or various documents in case of multi-document meetings.

After merging the results of the three alignment types, thematic, quotations and references, the final recall, precision and F values are 0.69, 0.58 and 0.59 respectively for the alignment variant with Cosine (Table 5), and 0.55, 0.72 and 0.59 respectively for the alignment variant with Cosine+TF.IDF, with a confidence value of ±0.03 for the respective measures. Even though the merging of the three alignment types does not significantly improve the F value, it is still useful in order to highlight their complementarity. For instance in Fig. 10, the 6th speaker utterance is thematically similar to the 2nd and the 3rd document

sentences, and it contains two respective quotations from them. Moreover, it contains a reference link to the document logical block containing these two sentences.

5.1. Named entity- based entailment

Our last entailment method is based on the matching of named entities (NE) between aligned segments. Our assumption is that the existence of common NE between two compared segments means that they might be thematically similar, and therefore a thematic alignment link should be created between them (if it was not yet generated by the thematic alignment process).

The NE matching between two segments is defined on the basis of a NE similarity score that is computed according to the following formula, where NE_{Si}, NE_{Sj} are the named entities detected within segments S_i and S_j respectively.

For example, the following utterance and sentence share one movie title "The Big Lebowski", thus their NE similarity score is one:

· <utterance id="9" StartTime="28.6" EndTime="32.8"> Okay, I'll go with <NE>***The Big Lebowski***</NE>, since posters are already ready </utterance>

• <sentence id="15">Friday 29th April, 20h, <NE>The Big Lebowski</NE>, by Joel Coen </sentence>

It should be noted that other NE similarity formulae have been tried out, which consisted in considering the denominator either as |NESi| or |NESi?NESj|. However, all formulae have generated similar scores.

The detection of named entities in respective segments is based on their matching with predefined gazetteers. Four distinct gazetteers containing named entities for the four respective corpora were manually created. The press review meetings gazetteer contains names of famous personalities in politic, sport, etc., as well as names of countries, cities. The movie club gazetteer contains names of movies, actors, directors, etc. The furniture proposal gazetteer contains names of furniture items. Finally, the scientific corpus gazetteer was created based on the main scientific keywords that are used in each presentation, extracted respectively from the scientific paper and the slideshow for each presentation.

First, the thematic links in each corpus are detected, and then NE similarity between all speech transcript segments and documents segments is computed. Later on, the generated thematic alignment links are compared to the NE links, in order to add missing thematic links. Therefore a threshold θ that aims to filter candidate NE links according to their similarity values was considered. If the NE similarity score between two segments (Si, Sj) overcomes threshold θ that is dynamically varied from 10% to 100%, then a new thematic link is established between them.

The obtained results of this entailment method are presented in Table 6. As shown the F value insignificantly increased from 0.37 to 0.40±0.06 for the movie club corpus. For the furniture proposal corpus, the F value insignificantly decreased from 0.40 to 0.37±0.07, which might be due to the insertion of many noisy alignment links, since many furniture item names are shared between various document logical blocks. The non-effectiveness of the entailment process to add relevant thematic links is explained by the nature of this corpus, where 50% of documents were not discussed, and those discussed had a poor textual content (only names of furniture items). The F value for the press review corpus was also insignificantly decreased from 0.59 to 0.57±0.03. The non-efficiency of the entailment process to add pertinent thematic alignment link in this corpus might be due also to its nature where the documents are totally discussed and have heterogeneous content. Finally for the scientific conference corpus, the NE based entailment has also generated many noisy links, which significantly decrease the F value from 0.62 to 0.46±0.05 for document/ slideshow, from 0.47 to 0.35±0.02 for slideshow/ speech, and from 0.33 to 0.29±0.02 for document/speech alignment. The main reason for this negative effect of the entailment process is the regular distribution of the used named entities (i.e. keywords) over all the documents, which generates ambiguous links between the speech transcript and the document segments. This generation of noisy links might be resolved in the future if the ambiguity of named entities across documents is resolved by considering contextual information of the matched segments.

6. Conclusion

In this paper, a study about the validation of thematic alignment links, detected between speech transcript and discussed static documents in the context of meetings and lectures, is presented. The validation approach is based on five distinct entailment

methods that exploit several features of the multimodal corpora, both linguistic and non-linguistic, at various levels of granularity of the aligned documents.

The evaluation performed showed that the defined entailment methods have different effect on the different corpora, depending on their content. Therefore, some of the methods have significantly improved the thematic alignment scores, such as the method based on merging thematic alignment levels which emphasized the complementarity between structures of aligned documents. The entailment method based on multiple document alignment has also improved the scores which stressed the importance of considering additional information resources in our alignment process. However, the method based on the merging of various alignment types did not have any effect on the scores of the thematic alignment, which prove that the detected alignments are already complementary. The two methods that exploit neighbouring features and named entity matching also did not have significant effect on the scores, where the scores were highly influenced by the corpora nature. These two last methods need to be tuned in the future by considering new features, for instance by associating appropriate weights to the various neighbour links in the neighbouring-based entailment, and disambiguating the links in the entity-based method. Other tasks are planned in the future, mainly tuning filtering thresholds of the various entailment methods without knowledge of the ground truth and the nature of the corpora, and the validation of alignment links from other types (i.e. quotations and references). Furthermore, new methods will be tested for the thematic alignment, for instance those based on Latent Semantic Analysis and on dynamic programming. Finally, the detected alignment links between multimodal documents will be used in order to generate an automatic summary of the multimedia event.

References

[1] Abou Khaled, O., Le Meur, J-Y., Scheurer, R., Bourillot, D., Lalanne, D., von Rotz, D., Ingold, R., Baron, T (2006). SMAC Project: *SMAC - Smart Multimedia Archive for Conferences*. *In:* Flash Informatique (FI1/06), p. 3-10, Ecole Polytechnique Fédérale de Lausanne EPFL, Switzerland.

[2] Barras, C., Geoffrois, E., Wu, Zhibiao, Liberman, M (1998). Transcriber: a Free Tool for Segmenting, Labelling and Transcribing Speech. *In:* Proc. of International Conference on Language Resources and Evaluation (LREC'98), p. 1373-1376, Spain.

[3] Bloechle, J-L., Rigamonti, M., Hadjar, K., Lalanne, D., Ingold, R (2006). XCDF: A Canonical and Structured Document Format. *In*: Proc. of the 7th IAPR International Workshop on Document Analysis Systems (DAS'06), New Zealand.

[4] Dagan, I., Glickman, O (2004). Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability. *In*: *P*roc. of PASCAL Workshop on Learning Methods for Text Understanding and Mining, France.

[5] Davidson, A (2001). A Fast Pruning Algorithm for Optimal Sequence Alignment. *In:* Proc. of the 2nd IEEE International Symposium on Bioinformatics & Bioengineering (BIBE'2001) p. 49-56, Maryland, USA.

[6] Havgaard, J. H., Torarinsson, E., Gorodkin, J (2007). Fast Pairwise Structural RNA Alignments by Pruning of the Dynamical Programming Matrix. *PLoS Computational Biology*, 10(3) p. 1896-1908.

[7] Lalanne, D., Ingold, R., von Rotz, D., Behera, A., Mekhaldi, D., Popescu-Belis, A (2005). Using Static Documents as Structured and Thematic Interfaces to Multimedia Meeting Archives. *In:* Proc. of Machine Learning for Multimodal Interaction (MLMI'05) p. 87-100, LNCS 3361.

[8] Mekhaldi, D., Lalanne, D., Ingold, R (2005). From Searching to Browsing through Multimodal Documents Linking. *In:* Proc. of the 8th International Conference on Document Analysis and Recognition (ICDAR'05) p. 924-928, Seoul, Korea.

[9] Mekhaldi, D (2006). A Study on Multimodal Document Alignment: Bridging the Gap between Textual Documents and Spoken Language, *PhD Thesis*, N° 1521, Fribourg, Switzerland.

[10] Mekhaldi, D (2007). Multimodal Document Alignment: towards a Fully-indexed Multimedia Archive. *In:* Proc. of Multimedia Information Retrieval Workshop, SIGIR'07, the Netherlands.

[11] Moore, D (2002). The IDIAP Smart Meeting Room. Technical report, IDIAP-Com, Switzerland.

[12] Popescu-Belis, A., Georgescul, M., Clark, A., and Armstrong, S (2004). Building and Using a Corpus of Shallow Dialogue Annotated Meetings. *In: Proc. of LREC'04*, p. 1451-1454, Portugal.

[13] Popescu-Belis, A., Lalanne, D (2006). Detection and Resolution of References to Meeting Documents. *In: Proc. of Machine Learning for Multimodal Interaction (MLMI'06)*, p. 64-75, LNCS 3869.

[14] Rigamonti, M., Lalanne, D., Ingold, R (2007). FaericWorld: Browsing Multimedia Events through Static Documents and Links. *In: Proc. of International Conference on Human-Computer Interaction (Interact'07),* p. 102-115, Brazil.

[15] Snyder, B., Barzilay, R (2007). Database-Text Alignment via Structured Multilabel Classification. *In:* Proc. of International Joint Conference on Artificial Intelligence (IJCAI'07), p. 1713-1718 India.

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

[17] Vu, T., Aw, A. T., Zhang, M (2009). Feature-Based Method for Document Alignment in Comparable News Corpora. *In: Proc. of* European Chapter of the Association for Computational Linguistics (*EACL'09),* p. 843-851, Greece.