Augmented Multi-party Interaction
http://www.amiproject.org

Augmented Multi-party Interaction with Distance Access
http://www.amidaproject.org

State-of-the-art overview

Recognition of Discourse Segments in Meetings

Updated version 10.10.2007

Information Society Technologies

Sixth Framework Programme

1 Introduction

Discourse is the deliberation process of what can be said about a specific topic. In common terms, it has been described in wikipedia as follows:

> Discourse is communication that goes back and forth (from the Latin, discursus, "running to and fro"), such as debate or argument. (c.f., wikipedia, as of 2007 [1])

Discourse prevails in our daily communication across widely ranging mediums such as written and spoken language. Regardless of the communication medium in use, our cognitive system can perform discourse segmentation effectively. This involves grouping coherent sequences of successive units (e.g., sentences, speech acts, or speaker turns) into discourse segments, each encompassing meanings beyond what is literally expressed in the individual units. Examples of our capacity to discourse segmentation are that we are able to interpret referring expressions, such as definite descriptions and pronouns, and resolve ellipsis. Also, we are good at summarising the gist of a particular segment and referring back to the relevant segment later.

In previous work of conversational discourse research, discourse segments are determined either from its informational content or by its intentional coherence. The former is done by grouping successive conversational units that are similar in the semantic focus of their expressions (that is, “what have been said”) [33], or in the relations that link those units to each other, such as rhetorical predicates [27], coherence relations [37, 57], and conjunctive relations [33]. The latter is done by grouping units with similar expression styles (that is, “how the speaker expresses it”)[2], or more generally, the underlying meaning and implicature of these successive units as a whole (that is, “what does the speaker imply when saying it”) [25].

In fact, informationally and intentionally coherent segments are often posited as isomorphic, for example, in the hierarchical intention structure proposed in [30]. [53] have further provided empirical evidences on how most of the successive units that are intentionally similar are also informationally coherent.[3] Our cognitive system constantly monitors the phenomenon of informational coherence and that of intentional coherence in discourse – depending on the medium and the application, the system may choose to attend to one or both – to group successive units into discourse segments.

With respect to the different notions of discourse coherence, previous work has attempted different ways to determine discourse segments. On the one hand, informationally coherent conversational units have been captured by rendering inference-based formal methods (e.g. abduction) across propositional content in the discourse [55, 38]. These methods usually work as follows: Two adjacent units of discourse are considered at a time. If there exist a coherence relation (e.g., cause-effect, violated expectation, condition, similarity, contrast, elaboration, attribution, temporal sequence) [45, 71] between the situations described by the two units, then the two units can be concatenated as a coherent segment of discourse. Applying this algorithm successively to the whole discourse will result in a tree

[1] http://en.wikipedia.org/wiki/Discourse

[2] In [25]’s theory, we all have a “repertoire” we use to indicate different meanings. For example, I may always use the same gesture to let you know that I know what to do.

[3] [53] examined 183 sentences from general-interest magazines such as Reader’s Digest. Despite that being informationally coherent (in this case, semantically cohesive) does not necessary translate to being intentionally coherent and, in converse, being intentionally coherent does not necessarily translate to being informationally coherent, most of the informationally coherent segments and intentionally coherent ones do correspond to each other.

structure for this discourse. However, the inference-based methods often assume the existence of full-fledged knowledge bases and, as a result, have problems with scalability. Therefore, past research has also explored other measures of informational coherence, for instance, semantic cohesiveness (i.e., a device that carries unity over text-like string representations of conversation) [53].

On the other hand, intentionally coherent segments have been determined through deriving discourse structure from its pragmatic context. Various theoretical models of discourse structure, for example, those based on individual-based speech acts [62, 26, 30] and on collaborative plans [63, 29], have been proposed. To understand the pragmatic characteristics of intentionally coherent successive units, empirical studies have been also conducted to analyze dialogue context in terms of both verbal features, e.g., discourse connectives [49, 4, 43], and non-verbal features, e.g., turn-taking cue [61, 47], acoustics (pitch range, contour, timing, energy level) [28], intonation pattern and speech rate [64], hand gesture, eye gaze, and head nod [13].

However, the theoretical and empirical studies on discourse coherence have focused more on the coverage of linguistic phenomenon rather than the computability of the proposed models and features. More recently, researchers have attempted to develop an automatic machinery to find discourse segment boundaries. A majority of works draw on the lexically cohesive characteristics of the segments to find informationally coherent segments. One successful approach is to view discourse segmentation as a time series problem amenable to signal processing. For example, TextTiling, an unsupervised lexical approach proposed by [35], looks for significant patterns in a quasi-temporal representation of the successive text units, and finds significantly disruptive patterns (i.e., where lexical cohesion scores change noticeably) that indicate a topic shift. [65] have extended the TextTiling approach to hypothesize segments in broadcast news.

Many other approaches view discourse segmentation as a dimension deduction problem similar to multinomial principal component analysis (PCA). On this front, variants of clustering algorithms have been proposed to group lexically similar units. In particular, Latent Semantic Analysis (LSA) [18], probabilistic Latent Semantic Analysis (pLSA) [39], and, more recently, Latent Dirichlet Allocation (LDA) [6] have been proposed to map lexical units to their associated semantic groups (a.k.a. topics). The representation of a discourse is then divided into major segments with respect to the semantic group features of these successive lexical units.

In practice, segmentation optimization can be achieved by using graph-cutting techniques to find segmentation that minimises inter-partition similarity without compromising intra-partition similarity [60, 58, 15, 69]. The graph-based techniques have been further attempted on hierarchical topic detection (HTD). It aims at organizing an unstructured news collection in a directed acyclic graph (DAG) structure, reflecting the topics discussed. (For more details, please refer to the report of the hierarchical topic detection task of TDT 2004 and [67].)

Segmentation optimization can also be achieved by applying the Hidden Markov model (HMM) and its variants (e.g., aspect HMM (AHMM)). The HMM-based framework consists of two major steps. First, k topic models (i.e., semantic groups) are constructed from large corpus (such as Wall Street Journal articles and CNN transcribed broadcasts), each model $T^{(j)}$, $1 \leq k$, referring to a smoothed language model of one semantically similar group found by some automatic clustering technique (e.g., K-means). Then, with respect to each of the identified k topic models, the probability of a given discourse unit being generated by this topic model is calculated. The topic of the highest probability will then be selected as the topic label (i.e. semantic group feature) of the unit. The observation of a discourse unit is considered as a collection of L mutually independent words that are generated by a topic model z_t. More formally, $o_t = w_{t,1}, w_{t,2}, w_{t,3}, ...w_{t,L}$. The emission probability can be computed as follows:

$$P(o_t|z) = \prod_{n=1}^{L} P(w_i|z) \tag{1}$$

Transition probabilities among the topics and the self-loop probability are also calculated. Based on these probabilities, a search for the optimal segmentation will then be found by placing boundaries around where the associated topic of the current unit is different from that of the next unit [70, 5]. Various limitations of this supervised generative approach have been recognized. In particular, it requires sufficient labelled data for training representative topic models; the topics of a to-be-segmented discourse also have to fall within the range of the k topics which have associated models.

Finally, the machine learning approach has been further extended to combine cues that are central to the recognition of intentions and topical contents. Unlike previous works that use generative models, the intention-based segmentation works train discriminative models. Typically, in this framework the task is decomposed as a series of binary decisions: for each possible segment boundary site (i.e., the end of each discourse unit), the system extract the context of the site X. Given X, a pre-trained model $q(y|X)$ is then used to classify this site into a boundary class y, where $y \in YES, NO$. q can be learned from training data as a decision tree, i.e., a set of decision rules. For example, [28] and [49] have trained a decision tree to perform classification in spoken narratives, with respect to the acoustic contexts in discourse. q can also be a exponential model, i.e., a decision function which is parameterized by a set of weights for features in the context representation. [4] and [16] have achieved success on segmenting broadcast news by training exponential models with features that characterize both the information and the intentional coherence. The context X of each discourse unit is represented as a combination of these features, including the occurrence counts of topical words, that of discourse connectives in a neighbouring window, and the duration of pause.

2 Meeting Corpus

Spontaneous face-to-face dialogues in meetings violate many assumptions made by techniques previously developed for broadcast news (e.g., TDT and TRECVID), telephone conversations (e.g., Switchboard) [24] , and human-computer dialogues (e.g., DARPA Communicator) [20] . In order to develop techniques for understanding multiparty dialogues, smart meeting rooms have been built at several institutes to record large corpora of meetings in natural contexts, including ISL [8][4], CHIL ("Computers in the Human Interaction Loop"), LDC [17], NIST [22], ICSI [44], and in the context of the IM2/M4 project [51]. More recently, scenario-based meetings, in which participants are assigned to different roles and given specific tasks, have been recorded in the context of the CALO ("Cognitive Agent that Learns and Organizes") project (the Y2 Scenario Data) [9] and the AMI ("Augmented Multiparty Interaction") project [11].

the ICSI meeting corpus and the AMI meeting corpus, among the others, are the two corpora that contain discourse segmentation annotations. The ICSI meeting corpus (LDC2004S02) consists of the audio recording of seventy-five natural meetings in ICSI research groups. These meetings were recorded using close-talking far field head-mounted microphones and four desktop PZM microphones. The corpus includes manual orthographic transcriptions of all 75 meetings.

The AMI meeting corpus consists of the audio-video recordings of 173 meetings collected across

[4] The ISL Meeting Corpus contains 112 meetings collected at the Interactive Systems Laboratories at CMU during the years 2000-2001. The recorded meetings were either natural meetings, or artificial meetings, which were designed explicitly for the purposes of data collection but still had real topics and tasks. The duration of the meetings in this corpus ranges from eight to 64 minutes and averages at 34 minutes.

three sites, IDIAP, U of Edinburgh and TNO. This corpus also includes high quality, manually produced orthographic transcription for each individual speaker. It is different from the ICSI meeting corpus in several aspects. First, while all of the ICSI meetings are natural group meetings where participants needed to meet in real world, only 33 meetings of the AMI meetings are natural ones. Approximately two-thirds of AMI meetings (140 out of 173) are driven by a scenario, wherein four participants play the role of the project manager, marketing expert, industrial designer, and user interface designer in a design team, taking a design project from kick-off to completion. Second, in addition to audio recordings, the AMI meetings also come with video recordings recorded by individual and room-view video cameras, slides from a slide projector, the note-taking pen inputs, and input from an electronic whiteboard.

2.1 Structural Discourse Segmentation Annotation

One third of the ICSI meeting corpus (25 out of 75) comes with annotations of discourse segmentation.[5] The AMI project team have also produced discourse segmentation annotations for both the whole ICSI and AMI corpus. In these annotations, topic segmentation is used as a covering term of discourse segmentation, without differentiating information and intentional coherence. Annotators have the freedom to mark a topic as subordinated[6] wherever appropriate. Three human annotators used a tailored tool to perform topic segmentation in which they could choose to decompose a topic into subtopics, with at most three levels in the resulting hierarchy.

As it is expected that the preferred segmentation algorithm for predicting segment boundaries at different levels of granularity would be different, this research flattens the subtopic structure and consider only two levels of segmentation–top-level topics (TOP) and all subtopics (ALL). The top level of the structure signals either major topic shifts in discourse structure or serious abruption of the ongoing discussions. The second level of the structure signifies either a temporary digression or a discussion that is more focused on one aspect of the current major topic. Basic statistics of the topic segmentation annotations are reported in Table1. Compared to the ICSI corpus, the segmentation structure of the AMI corpus is much more shallower, with smaller difference between the number of TOP segments and that of ALL segments.

Take the topic segmentation annotation of a 60 minute meeting Bed003 in the ICSI corpus for example. In this meeting, the research team are discussing about the planning of an automatic speech recognition project. Four major topics, from "opening" to "general discourse features for higher layers" to "how to proceed" to "closing". Depending on the complexity, each topic can be further divided into a number of subtopics. For instance, "how to proceed" can be subdivided to 4 subtopic segments, "segmenting off regions of features", "ad-hoc probabilities", "data collection" and "experimental setup".

Average	TOP	ALL	Length
ICSI	6.96	17.2	40 mins
AMI	7.67	13.65	28 mins

Table 1: *Basic statistics of discourse segmentation annotations in the ICSI and the AMI corpus.*

Previous works have examined the reliability of human discourse segmentation annotations. [50] have reported that human annotators mostly agree with each other in the text segment boundaries

[5] In this annotation, Michel Galley et al.[21] have gathered together the majority codings from at least three coders per observation.

[6] In the AMI annotation, the subordinated topics can go down to two levels, while in the ICSI annotation, they can only go down to one.

they chose despite a margin of a few utterances. [54] have demonstrated the level of reliability of human segmentation annotations in spoken narratives is within a reasonable range.[7]

To establish reliability of the annotation procedures used for segmenting the meeting corpora, kappa statistics [10] have been calculated as a measurement of the agreement between the annotations of each pair of coders. We also reported on the overall segmentation error rate, Pk and WD. Pk [4] is the probability that two utterances drawn randomly from a document (in our case, a meeting transcript) are incorrectly identified as belonging to the same topic segment. WindowDiff (Wd) [56] calculates the error rate by moving a sliding window across the meeting transcript counting the number of times the hypothesized and reference segment boundaries are different.

Table 2 shows the average kappa statistics of the three pairs of coders on the top-level and sub-level segmentation respectively. [31] have reported kappa (pk/wd) of 0.41 (0.28/0.34) for determining the top-level and 0.45(0.27/0.35) for the sub-level segments in the ICSI meeting corpus. [42] have reported that the human annotators have achieved $\kappa = 0.79$ agreement on the TOP segment boundaries and $\kappa = 0.73$ agreement on the ALL segment boundaries. Do the kappa values shown here indicate reliable intercoder agreement? In computational linguistics, kappa values over 0.67 point to reliable intercoder agreement. But [19] have found that such interpretation does not hold true for all tasks. However, the low disagreement rate among codings in terms of the PK and WD scores can be used to argue for the reliability of the annotation procedure used in these studies.

Intercoder	Kappa	PK	WD
ICSI(TOP)	0.41	0.28	0.23
ICSI(SUB)	0.45	0.27	0.35
AMI (TOP)	0.66	0.11	0.17
AMI (SUB)	0.59	0.23	0.28

Table 2: *Intercoder agreement of annotations at the top-Level (TOP) and sub-Level (SUB) segments.*

A complete manual topic segmentation has been annotated for the ICSI meeting corpus and the AMI meeting corpus. In the ICSI corpus, topic labels were essentially free format. Annotators were asked to provide a free text label for each topic segment; they were encouraged to use keywords drawn from the transcription in these labels. However, to impose some level of consistency, some standard labels are also provided for annotating the off-topic discussions, such as “opening” and “chitchat”.

As for those AMI meetings that are scenario-driven, annotators are expected to find that most of the topics do recur. Therefore, they are given a standard set of topic descriptions that can be used as labels for each identified topic segment. Annotators will only add a new label if they cannot find a match in the standard set. The standard set of topic descriptions has been divided to three categories:

- Top segments refer to topics whose content largely reflects the meeting structure (e.g, presentation, discussion, evaluation, drawing exercise) and the key issues of the design task (e.g., project specs, user target group).
- ALL segments refer to parts of the top-level topics (e.g., project budget, look and usability, trend watching, components, materials and energy sources).
- Functional segments are those parts of the meeting that refer to either the varying process and flow of the meeting (e.g., opening, closing, agenda/equipment issues), or are simply irrelevant (e.g., chitchat).

[7] Seven annotators worked on segmenting the corpus, which consists of 20 narratives monologues about the same movie, taken from [14].

In addition to the manual transcriptions, these meeting corpora also come with ASR transcriptions. The ASR transcriptions were produced by [32], with an average WER of roughly 30%. The system used a vocabulary of 50,000 words, together with a trigram language model trained on a combination of in-domain meeting data, related texts found by web search, conversational telephone speech (CTS) transcripts and broadcast news transcripts (about 10^9 words in total), resulting in a test-set perplexity of about 80. The acoustic models comprised a set of context-dependent hidden Markov models, using gaussian mixture model output distributions. These were initially trained on CTS acoustic training data, and were adapted to the ICSI meetings domain using maximum a posteriori (MAP) adaptation. Further adaptation to individual speakers was achieved using vocal tract length normalisation and maximum likelihood linear regression. A four-fold cross-validation technique was employed: four recognizers were trained, with each employing 75% of the meetings as acoustic and language model training data, and then used to recognise the remaining 25% of the meetings.

3 Evaluation Metrics

3.1 Automatic Discourse Segmentation

To evaluate the performance of segmentation models, various metrics have been proposed in the field of text segmentation. The most typical example is accuracy. Previous work has shown that when class distributions display a high level of entropy, i.e. $P(c_i \mid T) \approx P(c_j \mid T), i \neq j$ for any two classes c and training data T, accuracy is an acceptable measure of quality for a classifier. But when class distributions are highly skewed, recall, precision and harmonic means of these like the F_β-score are better measures.

In fact, discourse segmentation is a typically class-imbalanced task. The number of linguistic units on which segmentation is based (like sentences) typically by far exceeds the number of actual topics. Consequently, optimizing a classifier for accuracy would automatically favor a majority classifier that labels all sentences as not initiating a new segment. Optimization for the classical notions of recall and precision would not work well here either: for instance, a discourse segmenter that always predicts a segment boundary close but not exactly corresponding to the ground truth prediction would produce zero recall and precision, while its performance can actually be quite good.

In respond to this problem, P_k and W_d were designed to overcome the limitations inherent in the use of precision and recall for discourse segmentation. [4] has defined the P_k measure as the probability that a randomly drawn pair of utterances are incorrectly predicted as coming from the same segment. Also, [56] have analyzed several weaknesses of the P_k measure and proposed an adapted metric WindowDiff (W_d). W_d is computed as the probability that the number of hypothesized and reference segment boundaries in a given window frame are different.

However, these specific measures like Pk and WindowDiff ([**?**]) compute recall and precision in a fixed-size window to alleviate this problem, but they do not penalize false negatives and false positives in the same way. For topic segmentation, false negatives probably should be treated on a par with false positives, to avoid undersegmentation. Recently, [23] proposed a new, cost-based metric called Pr_{error}:

$$Pr_{error} = C_{miss} \cdot Pr_{miss} + C_{fa} \cdot Pr_{fa} \tag{2}$$

Here, C_{miss} and C_{fa} are cost terms for false negatives and false alarms; Pr_{miss} is the probability that a predicted segmentation contains less boundaries than the ground truth segmentation in a certain interval of linguistic units (like words); Pr_{fa} denotes the probability that the predicted segmentation in a given interval contains less boundaries than the ground truth segmentation. We refer the reader to [?] for further details and the exact computation of these probabilities.

3.2 Automatic Discourse Labelling

To evaluate the automatically generated labels against reference labels in the meeting corpus, relevant candidate metrics can be found in the fields of “story boundary detection” studied in TDT [66], TRECVID [46], and summarization studied in DUC [34]. Since the discourse segmentation annotators of some of the meeting corpus (e.g., the ICSI corpus) are free in their choice of keywords for topic labels, automatic evaluation of topic label assignment is difficult nad has not been attempted. For those meeting corpus (e.g., the AMI meeting corpus) that have their discourse labels selected from a predetermind set, overall classification accuracy is calculated as f-score (F1) to evaluate the performance of discourse labelling components [41]. We loop over each discourse segment in the standardized set. For each label in a predertermined set, precision is then computed as the total number of the discourse segments that have been assigned correct labels divided by the total number of discourse segments in the groud truth data; Recall is computed as the total number of the discourse segments that have been assigned correct labels divided by the number of segments that have been hypothesized as this label. A psuedo algorithm is given as below.

1. Loop(1) over each topic in the predetermined set
2. recall= total number of segments that have been assigned correctly to this topic/ total number of reference segments of this topic
3. precision= total number of segments that have been assigned correctly to this topic/total number of segments hypothesized as this topic
4. End Loop(2)

4 Recognition of Discourse Segments in Meetings

The problem of how to divide unstructured meeting speech into a number of locally coherent segments is important for two reasons: First, empirical analysis has shown that annotating transcripts with semantic information (e.g., topics) enables users to browse and find information from multimedia archives more efficiently [2]. Second, because the automatically generated segments make up for the lack of explicit orthographic cues (e.g., story and paragraph breaks) in conversational speech, dialogue segmentation is useful in many spoken language understanding tasks, including anaphora resolution [30], information retrieval (e.g., as input for the TREC Spoken Document Retrieval (SDR) task), and summarization [72].

As mentioned in Section 1, previous works have adopted three major approaches to tackle the problem of discourse segmentation: lexical-cohesion based approaches, topic modelling approaches, and supervised learning approaches. The first two can be operated in an unsupervised fashion. In the field of meeting discourse segmentation, [21] have extended the lexical cohesion-based TextTiling approach (named as LCSeg), and [59] have adapted the topic modelling approach to combine different topics so as to make this approach generalize well to segment meetings.

[21] has also applied the supervised learning approach to combine the outputs from LCSeg, which indicate information coherence, and other conversational features, which indicate speaker intentions.

Results have shown that the latter approach which trains a segmentation model with features that are extracted from knowledge sources beyond words, such as speaker interaction (e.g., overlap rate, pause, and speaker change) can outperform LCSeg. In addition, [3] have also pointed out that, when participant behaviours, e.g., note taking cues, are aggregated into the segmentation model, the performance can be further improved.

[41] have extended the supervised learning work in two ways: First, to understand whether there exists a difference in the preferred approach for predicting topic segmentation at different levels of granularity, it applied approaches that have been proposed for predicting granular-level topic shifts to the problem of identifying segments at a finer level. Second, as perfect human transcripts are always available, it has explored the impact on performance of using ASR output as opposed to human transcription.

The examination of the effect of features on performance shows that predicting top-level and predicting subtopic boundaries are two distinct tasks: (1) the lexical cohesion-based approach alone can capture the finer-level topic shifts, (2) the supervised learning approach, which combines lexical cohesion and intention-indicative features, performs better on predicting granular-level segments than on finer-level ones, and (3) applying feature selection, such as filtering cue phrase features with statistical metrics, can improve the performance of (2) on predicting finer-level segments by 10.46%. The examination of the effect of ASR transcripts has shown that despite the inevitable errors in ASR transcriptions, the preferred approach for predicting granular-level and finer-level segments does not change.

The experiments of [21] and [41] are both run on the ICSI corpus (LDC2004S02) [44]. [40] have applied these approaches on the AMI corpus [12]. However, results have shown that LCSeg is less successful in identifying “agenda-based conversation segments” (e.g., presentation, group discussion) in the AMI meetings. This is not surprising since LCSeg considers only lexical cohesion, and agenda-based segments are typically signalled more by intentional coherence than by informational coherence.

In many other researches which consider segmentation, a variety of features have been identified as indicative of segment boundaries in different types of recorded speech. For example, [7] have shown that a discourse segment often starts with relatively high pitched sounds and ends with sounds of pitch within a more compressed range. [54] have identified that topic shifts often occur after a pause of relatively long duration. Other prosodic cues (e.g., pitch contour, energy) have been studied for their correlation with story segments in read speech [68, 48, 16] and with theory-based discourse segments in spontaneous speech (e.g., direction-given monologue) [36]. In addition head and hand/forearm movements are used to detect group-action based segments [52, 1].

Therefore, [40] have further extended previous work to combine more features that can be extracted from dialogue contexts and multimedia inputs. Results have improved on previous work by 8.8% for granular-level segmentation and 5.4% for finer-level segmentation. Analysis of the effectiveness of the various features shows that lexical features (i.e., cue words) are the most essential feature class to be combined into the segmentation model. However, lexical features must be combined with other features, in particular, conversational features (i.e., overlap, pause, speaker activity change), to train well performing models. Furthermore, the multimodal features are essential to achieve good performance of a combined model. This is mainly because (1) the presence of the non-verbal features in the model can balance the tendency of models trained with lexical cues alone to over-predict, and (2) there is an interaction effect between these non-verbal features.

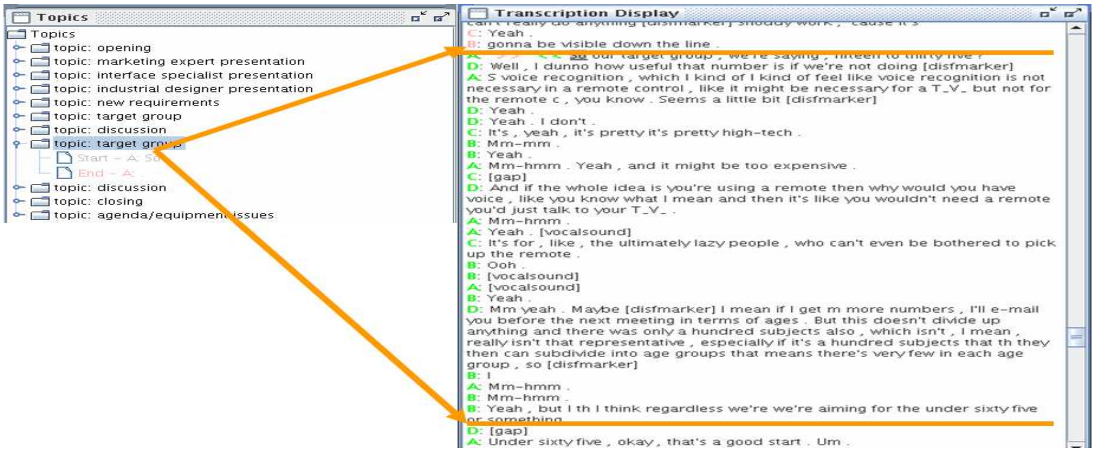
5 Application

The application needs of meeting speech segmentation is two-fold: On the one hand, the recognized discourse segments in meeting discourse form a quasi-summary for what have been transpired in a meeting and, in turn, provide the right level of details for users to interpret what the interlocutors are talking about in a meeting. Imagine the scenario that an industrial designer has missed a meeting and wanted to review the design team's discussion about the target user group. If the system can provide a discourse segment structure as shown in Figure 1, the users can then efficiently locate relevant information they are looking for (in this case, the segment about "target user group") from the list of segments. As evidenced in [2], discourse information does enable users to browse and find information from a meeting archive more efficiently. Moreover, when a recorded meeting has to be displayed on a mobile device, the recognised discourse segments can be used to construct an easy-to-grasp, thumb-nail view of the meeting. In short, discourse segmentation recognition has great potentials to enhance the current user interaction scheme of browsing and search.

On the other hand, discourse segment recognition benefits the development of other downstream meeting understanding applications. These applications include anaphora resolution [30], information retrieval (e.g., as input for the TREC Spoken Document Retrieval (SDR) task), summarization [72], and question answering. Take the application that needs to recover information for user queries for example, the recognized discourse segments can be used to guide the search of answer candidates toward those segments that are of topical relevance to the queries; the topical focuses of these segments can also serve as a means to rank the relevance of a list of answer candidates. The benefits of discourse segmentation on these applications would be even more evident when these applications have to be operated in an unfamiliar domain or in a foreign language environment.

6 Conclusion

We provided an overview of research in the area related to the recognition of discourse segmentation in meetings. The analyses concerns (1) the different notions of coherence central to discourse segmentation, (2) the features characteristics of coherence or the abruption of coherence, (3) the methods effective for finding discourse segments in text and spoken narratives, and (4) whether these methods and features can be effective for finding discourse segments in meetings.



Topics
- topic: opening
- topic: marketing expert presentation
- topic: interface specialist presentation
- topic: industrial designer presentation
- topic: new requirements
- topic: target group
- topic: discussion
- topic: target group
  - Start - A: So
  - End - A:
- topic: discussion
- topic: closing
- topic: agenda/equipment issues

can't really do anything [disfmarker] shoddy work , 'cause it's
C: Yeah .
B: gonna be visible down the line .
A: >> << So our target group , we're saying , fifteen to thirty five ?
D: Well , I dunno how useful that number is if we're not doing [disfmarker]
A: S voice recognition , which I kind of I kind of feel like voice recognition is not necessary in a remote control , like it might be necessary for a T_V_ but not for the remote c , you know . Seems a little bit [disfmarker]
D: Yeah .
D: Yeah . I don't .
C: It's , yeah , it's pretty it's pretty high-tech .
B: Mm-mm .
B: Yeah .
A: Mm-hmm . Yeah , and it might be too expensive .
C: [gap]
D: And if the whole idea is you're using a remote then why would you have voice , like you know what I mean and then it's like you wouldn't need a remote you'd just talk to your T_V_ .
A: Mm-hmm .
A: Yeah . [vocalsound]
C: It's for , like , the ultimately lazy people , who can't even be bothered to pick up the remote .
B: Ooh .
B: [vocalsound]
A: [vocalsound]
B: Yeah .
D: Mm yeah . Maybe [disfmarker] I mean if I get m more numbers , I'll e-mail you before the next meeting in terms of ages . But this doesn't divide up anything and there was only a hundred subjects also , which isn't , I mean , really isn't that representative , especially if it's a hundred subjects that th they then can subdivide into age groups that means there's very few in each age group , so [disfmarker]
B: I
A: Mm-hmm .
B: Mm-hmm .
B: Yeah , but I th I think regardless we're we're aiming for the under sixty five or something
D: [gap]
A: Under sixty five , okay , that's a good start . Um .

Figure 1: *Example of topic segmentation in a produce design meeting.*

The recognition of discourse segments requires recognition and tracking of lexical cohesion and other intention-indicative contexts, such as gesture/head movements, pitch, energy, rate of speech, and pause. There are many more works about inferring speaker intentions, for example, those in the line of human computer interaction research, we did not mention in this report. This is because our focus is on discourse segmentation. So we have only reviewed the discourse researches that are relevant to the recognition of segment boundaries.

Although recent research that used supervised learning approaches to combine various lexical cohesion and intention-indicative features have achieved success, it has at least two shortcomings: First, although these features are expected to be complementary to one another, few of them have studied how to systematically model the correlation among features in machine learned models. This has pointed out some possible improvements on applying some more sophisticated machine learning approaches, such as Conditional Random Fields, to overcome this shortcoming.

Second, training a well-performing discriminative model requires plentiful labelled data; yet, it is uncertain whether the trained model can be applied to segment meetings in a domain different from the labelled data. One solution is to apply unsupervised approaches. However, previous works in unsupervised meeting segmentation focus mainly on modelling word-related phenomenon, such as lexical cohesion and topical focus. Yet, we have seen in the supervised learning work that many other features beyond words, such as multimodal and dialogue contexts, are central to meeting segmentation. This is partially because meeting dialogues are spontaneous conversations in a multiparty environment, and naturally we have more communicative channels, such as body language, gaze engagement, gesture, and prosody, we can use to signal what we mean.

This has indicated the need of further investigation into how to combine multiple knowledge sources into the unsupervised approaches for meeting segmentation. To adaptively generalize the word-based approaches to combine multimodal features, two possible directions have thus arsed: (1) a more thorough empirical study about the synchronism mechanism between the intention-indicative features and the words, and (2) some novel ways to combine features in the current unsupervised segmentation approaches are also necessary.

References

[1] M. Al-Hames, A. Dielmann, D. GaticaPerez, S. Reiter, S. Renals, and D. Zhang. Multimodal integration for meeting group action segmentation and recognition. In *Proc. of MLMI 2005*, 2005.

[2] S. Banerjee, C. Rose, and A. I. Rudnicky. The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing. In *Proc. of the International Conference on Human-Computer Interaction*, 2005.

[3] S. Banerjee and A. Rudnicky. Segmenting meetings into agenda items by extracting implicit supervision from human note-taking. In *Proc. of IUI 2006*, 2006.

[4] D. Beeferman, A. Berger, and J. Lafferty. Statistical models for text segmentation. *Machine Learning*, 34:177–210, 1999.

[5] D. M. Blei and P. J. Moreno. Topic segmentation with an aspect hidden Markov model. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 2001.

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

[7] G. Brown, K. L. Currie, and J. Kenworthe. *Questions of Intonation.* University Park Press, 1980.

[8] S. Burger, V. MacLaren, and H. Yu. The isl meeting corpus: The impact of meeting type on speech style. In *Proceedings of the ICSLP 2002*, 2002.

[9] CALO. Cognitive agent that learns and organizes. $http://www.ai.sri.com/project/CALO$, 2006.

[10] J. Carletta. Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[11] J. Carletta et al. The AMI meeting corpus: A pre-announcement. In *Proceedings of 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, 2005.

[12] J. Carletta et al. The AMI meeting corpus: A pre-announcement. In S. Renals and S. Bengio, editors, *Springer-Verlag Lecture Notes in Computer Science*, volume 3869. Springer-Verlag, 2006.

[13] J. Cassell, Y. Nakano, T. Bickmore, C. Sidner, and C. Rich. Non-verbal cues for discourse structure. In *Association for Computational Linguistics Annual Conference*, 2001.

[14] W. L. Chafe. *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production.* Ablex Publishing Corporation, 1980.

[15] F. Choi, P. Wiemer-Hastings, and J. D. Moore. Latent semantic analysis for text segmentation. In L. Lee and D. Harman, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 109–117, 2001.

[16] H. Christensen, B. Kolluru, Y. Gotoh, and S. Renals. Maximum entropy segmentation of broadcast news. In *Proc. of ICASP*, Philadelphia USA, 2005.

[17] C. Cieri, D. Miller, and K. Walker. Research methodologies, observations and outcomes in conversational speech data collection. In *Proceedings of the Human Language Technologies Conference (HLT)*, 2002.

[18] S. Deerwester, S. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41 (6):391–40, 1990.

[19] B. Di Eugenio and M. G. Glass. The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101, 2004.

[20] M. Eskenazi, A. Rudnicky, K. Gregory, P. Constantinides, R. Brennan, C. Bennett, and J. Allen. Data collection and processing in the carnegie mellon communicator. In *Proceedings of Eurospeech*, 1999.

[21] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing. Discourse segmentation of multi-party conversation. In *Proc. of ACL 2003*, 2003.

[22] J. S. Garofolo, C. D. Laprun, M. Michel, V. Stanford, and E. Tabassi. The NIST meeting room pilot corpus. In *Proceedings of LRECâĂŹ04*, 2004.

[23] M. Georgescul, A. Clark, and S. Armstrong. Word distributions for thematic segmentation in a support vector machine approach. In *Proceedings of CoNLL*, pages 101–108, 2006.

[24] J. Godfrey, E. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research and development. In *Proceedings of ICASSP*, pages 517–520, 1992.

[25] H. P. Grice. Utterer's meaning and intentions. *Philosophical Review*, 1969.

[26] H. P. Grice. *Logic and conversation*, page 41âĂŞ58. New York: Academic Press, 1975.

[27] J. Grimes. *The thread of discourse*. The Hague, 1975.

[28] B. Grosz and J. Hirschberg. Some intonational characteristics of discourse structure. In *Proceedings of the International Conference on Spoken Language Processing*, 1992.

[29] B. Grosz and S. Kraus. Collaborative plans for group activities. In *Proceedings of IJCAI-93*, pages 367–373, Chambery, France, 1993.

[30] B. Grosz and C. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3), 1986.

[31] A. Gruenstein, J. Niekrasz, and M. Purver. Meeting structure annotation: Data and tools. In *Proc. of the SIGdial Workshop on Discourse and Dialogue*, 2005.

[32] T. Hain, J. Dines, G. Garau, M. Karafiat, D. Moore, V. Wan, R. Ordelman, and S. Renals. Transcription of conference room meetings: An investigation. In *Proc. of Interspeech 2005*, 2005.

[33] M. A. K. Halliday and R. Hasan. *Cohesion in English*. London: Longman, 1976.

[34] D. Harman and P. Over. The effects of human variation in duc summarization evaluation. In *Proceedings of the Workshop on Text Summarization Branches Out of ACL 2004*, 2004.

[35] M. Hearst. TextTiling: Segmenting text into multiparagraph subtopic passages. *Computational Linguistics*, 25(3):527–571, 1997.

[36] J. Hirschberg and C. H. Nakatani. A prosodic analysis of discourse segments in direction-giving monologues. In *Proc. of ACL 1996*, 1996.

[37] J. R. Hobbs. Coherence and coreference. *Cognitive Science*, 3:67ï£¡V90, 1979.

[38] J. R. Hobbs. Abduction in natural language understanding. In L. Horn and G. Ward, editors, *Handbook of Pragmatics*. Blackwell, 2004.

[39] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, 1999.

[40] P. Hsueh and J. Moore. Automatic topic segmentation and lablelling in multiparty dialogue. In *the first IEEE/ACM workshop on Spoken Language Technology (SLT) 2006*, 2006.

[41] P. Hsueh, J. Moore, and S. Renals. Automatic segmentation of multiparty dialogue. In *Proc. of EACL 2006*, 2006.

[42] P. Hsueh and J. D. Moore. Combining multiple knowledge sources for dialogue segmentation in multimedia archives. In *Proceedings of the 45th Annual Meeting of the ACL*, 2007.

[43] B. Hutchinson. Acquiring the meaning of discourse markers. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, 2004.

[44] A. Janin et al. The ICSI meeting corpus. In *Proc. of ICASSP 2003*, 2003.

[45] A. Kehler. *Coherence, Reference and the Theory of Gramma*, chapter A Theory of Discourse Coherence. CSLI Publications, Stanford, CA, 2002.

[46] W. Kraaij, A. Smeaton, P.Over, and J. Arlandis. Trecvid 2004 ï£¡ an introduction. In *TREC Video Retrieval Evaluation Online Proceedings*, 2004.

[47] S. C. Levinson. *Pragmatics*. Cambridge University Press, 1983.

[48] G. Levow. Prosody-based topic segmentation for mandarin broadcast news. In *Proc. of HLT 2004*, 2004.

[49] D. Litman and R. Passoneau. Combining multiple knowledge sources for discourse segmentation. In *Proc. of the ACL 1995*, 1995.

[50] W. Mann and S. Thompson. *Rhetorical structure theory: Toward a functional theory of text organization.* 1988.

[51] S. Marchand-Mailet. Meeting record modeling for enhanced browsing. Technical report, Computer Vision and Multimedia Lab, Computer Centre, University of Geneva, Switzerland, 2003.

[52] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(3):305–317, 2005.

[53] J. Morris and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1), 1991.

[54] R. Passonneau and D. Litman. Intention-based segmentation: Human reliability and correlation with linguistic cues. In *Proc. of ACL 1993*, 1993.

[55] F. C. N. Pereira and B. J. Grosz. *Natural Language Processing*. MIT Press, Cambridge, Massachusetts, 1994.

[56] L. Pevzner and M. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36, 2002.

[57] L. Polanyi. A formal model of discourse structure. *Journal of Pragmatics*, pages 601–638, 1988.

[58] J. Ponte and W. Croft. Text segmentation by topic. In *Proc. of the Conference on Research and Advanced Technology for Digital Libraries 1997*, 1997.

[59] M. Purver, P. Ehlen, and J. Niekrasz. Shallow discourse structure for action item detection. In *the Workshop of HLT-NAACL: Analyzing Conversations in Text and Speech*. ACM Press, 2006.

[60] J. Reynar. *Topic Segmentation: Algorithms and Applications*. PhD thesis, UPenn, PA USA, 1998.

[61] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696âĂŞ735, 1974.

[62] J. Searle. *Speech acts: An essay in the philosophy of language*. Cambridge University, Cambridge England, 1969.

[63] J. R. Searle. *Collective intentionality*. 1990.

[64] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communications*, 31(1-2):127–254, 2000.

[65] N. Stokes, J. Carthy, and A. Smeaton. Select: a lexical cohesion based news story segmentation system. *AI Communications*, 17(1):3–12, Jan. 2004.

[66] TDT-Evaluation. The 2002 topic detection and tracking (tdt2002) task definition and evaluation plan. Technical report, TOPIC DETECTION AND TRACKING (TDT2002), 2002.

[67] D. Trieschnigg and W. Kraaij. Hierarchical topic detection in large digital news archives: Exploring a sample based approach. *Journal of Digital Information Management*, 3(1), 2005.

[68] G. Tur, D. Hakkani-Tur, A. Stolcke, and E. Shriberg. Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics*, 27(1):31–57, 2001.

[69] M. Utiyama and H. Isahara. A statistical model for domain-independent text segmentation. In *Proceedings of the 28th Annual Meeting of the ACL*, 2001.

[70] P. van Mulbregt, J. Carp, L. Gillick, S. Lowe, and J. Yamron. Segmentation of automatically transcribed broadcast news text. In *Proceedings of the DARPA Broadcast News Workshop*, pages 77–80. Morgan Kaufman Publishers, 1999.

[71] F. Wolf and E. Gibson. Representing discourse coherence: a corpus-based analysis. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 134, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[72] K. Zechner and A. Waibel. DIASUMM: Flexible summarization of spontaneous dialogues in unrestricted domains. In *Proc. of COLING-2000*, 2000.