Transcript Segmentation Using Utterance Cosine Similarity Measure

Caroline Chibelushi, Bernadette Sharp[1] and Andy Salter[2]

[1] Staffordshire University, Beaconside, Stafford, ST18 0DG, UK
[2] University of Southampton, School of Civil Engineering & the Environment
Southampton SO17 1BJ, UK

Abstract. One of the problems addressed by the Tracker project is the extraction of the key issues discussed at meetings through the analysis of transcripts. Whilst the task of topic extraction is an easy task for humans it has proven difficult task to automate given the unstructured nature of our transcripts. This paper proposes a new approach to transcript segmentation based on the Utterance Cosine Similarity (UCS) method. Our segmentation approach is based on the notion of semantic similarity of utterances within the transcripts that measures the content similarity, semantic relationships, and use distance to differentiate same topics that appear in different context. The method is illustrated using one of the 17 transcripts in our study.

1 Introduction

The Tracker project is an attempt at dealing with the fundamental industrial problem of reducing rework in systems engineering projects. Though rework is often inevitable in large projects either because of changing requirements or changing priorities, we believe that a significant amount of rework arises as a result of communication failures between decision makers leading to inappropriate or incorrect decisions. By identifying the issues, and associated actions and decisions through the study of minutes of the meeting and a transcript record of these meetings we can begin to understand the causes of rework. If these minutes and transcript can provide some transparency in the decision making process we can minimise rework. This transparency is achieved by tracking the elements of decisions made by the participants during meetings. The elements can be identified through the linguistic analysis of the minutes and transcripts of meeting, and are expressed in terms of issues and sub-issues, associated actions and initiating or acting agents. By identifying the elements of the decision making process and discovering the semantic association between these elements we hope to identify causes of rework in these systems engineering projects.

This paper reports findings from a two-year study focusing on the transcripts of meetings, and describes our approach to extracting the component related to issues and sub-issues discussed at the meetings. An important stage in the extraction process is the automatic identification of issues and the boundaries between distinct

Chibelushi C., Sharp B. and Salter A. (2005).
Transcript Segmentation Using Utterance Cosine Similarity Measure.
In *Proceedings of the 2nd International Workshop on Natural Language Understanding and Cognitive Science*, pages 78-90
DOI: 10.5220/0002560900780090

issues in a given transcript. This process is known as text segmentation by the Topic Detection and Tracking (TDT) research community. This paper will describe our algorithm to transcript segmentation based on the *Utterance Cosine Similarity* measure (UCS) to identify related utterances within transcripts of meetings. This algorithm extends our previous work on the use of lexical chaining in tracking the major issues within a transcript as reported in [5].
The first part of this paper gives a brief description of the transcripts under study, and provides an overview of our approach in extracting the issues and sub-issues discussed in the meeting. The second part describes our algorithm in locating utterances that relate to the same issue discussed within the transcripts of meetings.

2 The Corpus

In organisation meetings, a transcript is usually a record of all uttered decisions, actions, discussions and arguments uttered by the participants. Our corpus consists of 17 different transcripts from three different meeting environments: industrial, organisational, and educational, involving a multi-party conversation and containing an exact and unedited record of the meetings and corresponding speakers. The meeting transcripts vary in size, ranging from 2,479 to 25,670 words. While previous research has focused on structured texts, broadcast news, and monologues which consist of cohesive stories, our transcripts, however, are multi-party conversation, and have no pre-set agendas. Consequently, the analysis of our transcripts poses an additional complexity due to their informal style, their lack of structure, their argumentative nature, and the usage of common colloquial words. These transcripts also contain incomplete sentences, sentences related to social chatting, interruptions, and references by participants made to visual context.

3 Detecting Issue Boundaries using the Utterance Cosine Similarity (UCS) Measure

Part of the difficulty in dealing with decisions, as reported by [11], is the problem of identification of the decisions in the first place. Many decisions are 'hidden' the only evidence of their existence appears in the form of actions. This research project seek to track issues, agents, needs, and actions as elements of the decision making process. Each of these elements can be part of one or more decisions. As we are concerned with extracting elements of decisions occurring within a meeting we therefore need to segment the transcripts to identify the issues, the needs, the actions, and the agents who initiated or undertook an action.
The automatic identification of boundaries where topics change in a given text is defined as text segmentation, and normally used in information retrieval and automatic summarisation tasks [8, 2]. Unlike the Topic and Detection Tracking initiative [1] which applies text segmentation to detect coarse-grained topic shifts in news stories we are interested in extracting issues and sub-issues discussed by the participants, and identifying their associated actions. Rather than taking the transcript

as the unit of analysis and then use queries to conduct analysis between transcripts, we use the utterance as the unit of analysis and conduct the analysis within transcripts, where an utterance is identified as a section of speech related to a participant and is continued until a new speaker is identified.

Text segmentation techniques tend to be either statistically or linguistically driven. Some statistical approaches are based on probability distribution [3], others use machine learning techniques, or treat text as an unlabeled sequence of topics using a hidden Markov model [17]. Other text segmentation approaches tend to use lexical chains to identify topic changes [16], or use clues which mark shifts to new topics [14]. A different approach is adopted by [12] who use decision trees to combine linguistic features extracted from spoken texts. In our project we have combined statistical and linguistic approaches to detect issues and associated sub-issues from our transcripts. We consider issues as equivalent to topics in the study of text segmentation. For the purpose of our project we define issue as a culminating point or matter leading to a decision. To detect these issues from our transcripts we extract nouns and compounds nouns and use lexical chaining to group them into semantically related clusters. While existing techniques rely on paragraph boundaries or used fixed length segments for text segmentation we base our initial segmentation on lexical chaining and utterance boundaries. As often there are more than one lexical chain identified within a segment so we use the Utterance Cosine Similarity (UCS) measure to identify the main lexical chain and refine the boundaries accordingly. Once the issues are detected then we track the verbs associated with these issues to capture their resolution. We also use linguistic patterns to determine who is to undertake what.

4 A framework for Transcript Segmentation

This section describes our framework for transcript segmentation using the Utterance Cosine Similarity (UCS), an adapted version of the Cosine Similarity; instead of measuring the similarity between a query and a document UCS measures the similarity between two utterances. As mentioned earlier due to the nature of our transcripts we cannot make use of paragraph boundaries or of a pre-set agenda to locate shift in topics being discussed, so our approach relies heavily on lexical cohesion based on word or phrase repetition. As [4] point out, cohesion can best be explained by focusing on how lexical repetition is manifested, in numerous ways, across pairs of sentences. The lexical cohesion is also used to measure similarity between segments. As the transcripts consist of a set of utterances the boundaries must include an entire utterance, and a segment can include one or more utterances.

Our framework for transcript segmentation can be visualised as consisting of three phases: pre-processing stage, vector representation stage, and segmentation stage using the Utterance Similarity Measure (UCS). These stages are described in figure 1.

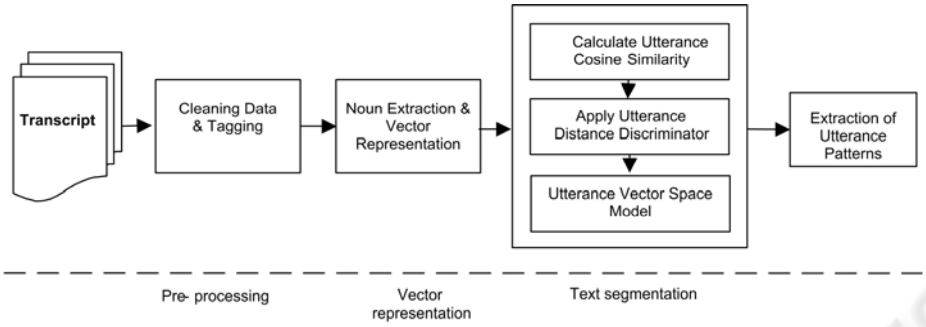

Fig. 1. Stages used in topic identification using UCS method

4.1 Pre-processing Stage

Given the nature of our transcripts the pre-processing stage consists of data cleaning and tagging tasks. Our data cleaning approach includes the removal of ambiguous, redundant and illegal characters, the removal of incorrect hyphenation, and the conversion of upper to lower case. In agreement with [9] and [6] we do not apply stemming as it has been shown to make no significant change to the processing of the transcript and has led to some syntactically motivated inflections being placed in an incorrect equivalent class. For example words like 'organ', 'organisation', and 'organism' will be placed in the same or equivalent class. This is because most stemming algorithms do not identify the morphemes, instead they simply remove common affixes from a word.
The second task involves the syntactic and semantic tagging of lexical items using WMATRIX which is a software tool developed by Lancaster University [13]. The outcome of the pre-processing stage is an XML tagged document which provides the basis for the utterance representation stage.

4.2 Vector Representation Stage

Content words, particularly nouns and proper nouns, introduce concepts, and are the means of expression for issues and sub-issues. Function words such as determiners, prepositions, conjunctions, relative pronouns, *etc*. support and coordinate the combination of content words into meaningful sentences. Though both word types are needed it is the content words which carry most weight in defining the actual topic of discourse [15], and capture the issues discussed in our transcripts. In agreement with [10] we argue that, when a concept named by a content word is topical for the document, the content word tends to be characterised by multiple occurrences within a segment of the document. Katz also states that, while a single occurrence of a topically used content word or phrase is possible, it is more likely that a newly topical entity will be repeated.

The tagging process used in the pre-processing stage splits compound words into their individual elements, for example the term 'user interface' will be regarded as 'user 'and 'interface'. The component words will not have the same semantic representation as the original compound word. To avoid this problem, our method recombines the separated component words and uses them as a compound noun (i.e. user_interface). The first stage produces an *Utterance Index Set* (UIS) which records the global frequency of nouns and compound nouns found in the transcripts. Only those nouns that score higher than a threshold value are selected, and the others are regarded as 'noisy' nouns and removed from the UIS. As our transcripts are verbal recording of meetings, we have found that transcripts contain high frequency of speech fillers (e.g. *thing, kind of*, and *sort of*). These noisy nouns are added to a stop list. It is interesting to note that the number of distinct content words is surprisingly small given the size of the corpus, as shown in table 1.

Table 1. The corpus

Transcript ID	Total no of words	participants	Total No. of content nouns	Transcript ID	Total no of words	participants	Total No. of content nouns
000403AL	2479	2	104	00ACT00 TR	12345	12	478
000BR00	20746	5	648	260702T R	11, 259	4	469
000GM0F	19977	9	828	230701T R	20, 567	6	459
120802TR	13962	9	365	00INT04	10,943	4	448
120901TR	12062	4	448	02INT04	9,384	2	376
290701TR	11471	10	499	03INT04	3261	2	106
200602TR	25670	7	734	000AU00	17753	10	796
070703TR	21003	12	461	Semlab02	8314	3	249
120902TR	22821	4	491				

The next step involves the construction of a term frequency vector for each utterance. An utterance consists of all the nouns and compound nouns and their position as spoken by one speaker in a turn. An utterance U_i is defined as $U_i = \{W_1 \ldots W_n\}$, whereby, W_i is a noun or compound noun as it appears in the utterance. A term frequency vector f_i is constructed for each utterance U_i by recording the frequency of occurrence of each of the members of the UIS in the utterance U_i. For example the utterance of U_{12} given below consists of 4 distinct nouns: size, board, laptop, and edge with their respective frequencies 1,3,1,1.

U_{12}: you can change the *size* of the *board* here in the *laptop*, just draw round the *edge* of the *board* and see where it appears on the *board*.

Thus for the utterance U_{12} the frequency vector is:

f_{12} = {1, 3, 1, 1}

As transcript 2 has 33 distinct nouns and compound nouns with frequency threshold >5, the final vector representation for the utterance U_{12} consists of 33 elements denoting the frequency of each distinct noun within that utterance as follows:

f_{12} = {1, 3, 1, 1, 0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0,0}

4.3 Segmentation using Utterance Similarity Measure (UCS)

We have adapted the cosine similarity measure which normally measures the similarity between various pairs of documents, to our transcripts analysis. We state that two utterances, U_i and U_j, are similar if the cosine of their frequency vectors is closer to 1. The UCS measure, denoted sim (U_i, U_j), is defined as

$$\text{sim}(U_i, U_j) = \cos(f_i, f_j) = \frac{\sum_k f_{ik} \times f_{jk}}{\sqrt{\sum_k f_{ik}^2 \times \sum_k f_{jk}^2}}$$

Where $0 \leq \cos(f_i, f_j) \leq 1$.

$\sum_k f_{ik} \times f_{jk}$ is the inner product of f_i and f_j, which measures how much the two vectors have in common. $\sqrt{\sum_k f_{ik}^2 \times \sum_k f_{jk}^2}$ is a vector length which is used to normalise the vectors.

Similar terms tend to occur in similar utterances, the angle between them will be small, and the cosine similarity measure will be close to 1. On the contrary, utterances with little in common will have dissimilar terms, the angle between them will be close to $\pi/2$ and the UCS measure will be close to zero. A UCS matrix is calculated by comparing every utterance with every other utterance in the transcript.

The UCS measure gives a value for the similarity between the content of two utterances, however it does not take context into consideration. It is therefore possible for two utterances to have similar content, indicated by a similarity value close to one, but be unrelated in context. To enable the method to differentiate utterances of different context, we apply an *Utterance Distance Discriminator* (UDD) which is defined as:

$$UDD_{ij} = \left| \frac{U_i - U_j}{U_{total}} \right| \quad \text{where } i = 1,\ldots n\text{-}1, \text{ and } j = 2, \ldots, n$$

U_i and U_j represent the occurrence of the utterance within the transcript and U_{total} is the total number of the utterances within the transcript. Through experimentation we have found that when the value of UDD_{ij} was less than 0.15 the utterances appeared to share similar context, and when it was greater than 0.15 it showed that though the utterances contained similar terms but they did not share the same context.

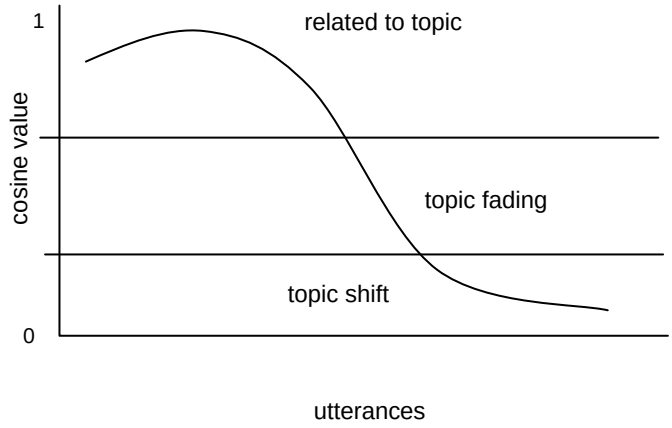

Fig. 2. The changes in cosine value indicating the progress of a topic

The boundaries between issues are identified by examining the patterns of the values recorded in the Utterance Cosine Similarity matrix. Our experiment showed that closely related utterances have a cosine value ranging from 0.6 to 1. Where the cosine value between adjacent utterances reduces to less than 0.6 this indicates a topic change. Cosine values between adjacent utterances which range from 0.1 to 0.2 indicates a topic shift. An example of the progression of a topic indicated by the changing cosine values is shown in figure 2.

5 Experimental Results

To date, we have experimented with 17 different transcripts from three different meeting environments: industrial, organisational, and educational. The results reported here are based on transcript 2 which contains a total of 11,471 words and 559 utterances, and involves a group of 10 participants. The topics discussed in the meeting included a demonstration of a tool designed to electronically capture text and diagrams from a whiteboard, a discussion related to employment of staff, the purchase of equipments, and project management issues.

In the first stage the transcript is pre-processed to check spelling, and to remove redundant and illegal characters (e.g. <, >), and redundant lexical items such as pause, er, um. The transcript is then syntactically and semantically tagged using WMATRIX.

USER$_y$ Um, various other things you can do, print it out, <unclear> your ***printer*** or whatever but, er and then re-size the ***board*** as well. And there's two other things here as well, which I haven't talked to you about <pause> but what you can do is, once you've got to a ***point*** in a ***meeting*** or you've recorded enough that you want on that ***screen*** then you either want to time stamp it for the ***record*** or you want to wipe the ***screen*** clear, you can do one of two things. Either make a new um, ***screen***, so if I clicked on that, what happens over here is, very difficult to see for you lot, but um, you get another <unclear> small representational <unclear> ***screen***, so you get another ***screen*** pop up which is blank and then this one copies what's on the white ***board*** but time stamps it.

Fig. 3. A typical utterance from transcript 2

The second stage extracts all nouns, recombines compound nouns appropriately, removes common words from the stop list, and generates the initial utterance index set (UIS). An output for this is the candidate nouns and compound nouns that are used for analysis. Transcript 2 has a total of 499 nouns and compound nouns. However, not all the nouns contribute to the identification of topics. Nouns and compound words were then filtered according to a threshold value of global occurrences in the transcript. From the results of analysing the 17 transcripts, a threshold value of 5 was identified as the optimum value. For transcript 2, 380 out of 499 nouns and compound words have a global occurrence of less than 5 and these were removed from the UIS. 77 out of 499 nouns were identified as high frequency nouns (nouns which have frequency higher than 5); 44 out of these 77 nouns (such as 'thing', 'stuff', and 'example') were removed and placed on the stop list. Finally, the UIS for transcript 2 consisted of 559×33 matrix, where 559 is the number of utterances and 33 represent the number of candidate nouns and compound nouns, listed below in the order of their first appearance in the transcript.

$UIS_{transcript2}$ = {*board, pen, screen, laptop, meeting, data, people, kit, project_management, bscw_server, project, staffing, capture, date, place, committee, list, system, information, site, problem, stream, decision, scanner, text, decision_capture, disc, audio, issue, theory, video, framework, interpretation*}.

The UIS is then used to construct the frequency vector for each utterance. The vector records the frequency of occurrence of each UIS member in the utterance. The results of the UCS are captured in Figure 4. Examination of the chart below indicates that there are a number of areas where the utterances contain similar content and context. These areas can be used to identify topic related utterances and topic boundaries and also for identifying topic related content words. The dark shade of the similarity values in area **A** of figure 4 indicates that the utterances are related to the same topic. In this case the topic is a demonstration of Mimio, a device for electronically capturing the information being recorded on a white board. As such the topic includes the UIS member *board* referring to the white board on which the writing occurs. The word *board* also appears in the utterances included in area **B** of figure 4 but in a different context. Here the utterances include "*since they have agreed to come on board*." referring to industrial collaborators agreeing to participate in the project. By

looking at the similarity values and UDD values in areas **A** and **B** our method can be used to separate the two words into different classifications and context.

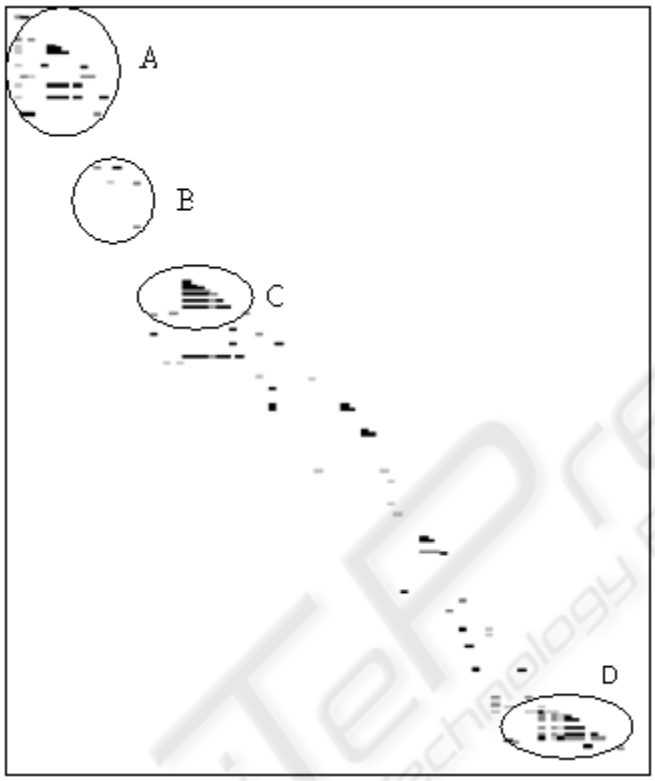

Fig. 4. Similarity matrix chart for transcript 2

Another example of the same word appearing in different topics is the word *meeting* which occurs in areas **C** and **D**. The topic of discussion at **C** includes the words *meeting, record,* and *information_capture*. These words are used in the context of recording the project meetings which could then be used in the development of a tool to assist in the capture of decisions. The word *meeting* is also used in the utterances appearing in area **D** but under a different context, that of finalizing the current meeting and reaching agreement on the date of the next meeting. Having produced the similarity matrix chart and located the areas of related utterances, it is possible to identify the topic boundaries.

5.1 Comparative analysis with C99

C99 is a text segmentation algorithm which has two key elements: a clustering strategy (divisive clustering) which
determines the location boundary and a similarity measure that use ranking scheme which linearises the cosine coefficient [7]. The application of the C99 algorithm to

transcript 2 produced 23 topic boundaries. However some of these boundaries were placed within sets of utterances that were related to the same topic. An example of a boundary being identified in the middle of a set of utterances is shown in figure 5, the boundary occurring between a question and an answer.

PR I don't know.
JC Would you see that as a question that's in your area?
PR Um.
JC Is that something that you would consider looking at?

==========

PR Yeah I think we should consider looking at it in the project. Haven't really thought much about that. The focus I was looking at the moment was……

Fig. 5. An example of incorrect segmentation using the C99 algorithm

Part of the reason for this incorrect identification of topic boundaries is the result of the C99 algorithm being intended for use on structured texts. Transcripts, by their nature, are more complex and unstructured. There can also be problems with the quality of the transcription itself, which is dependent on the skills of the transcribers. Using our UCS measure we are able to identify the topic boundaries by examining the patterns of the utterance similarities, placing a clear boundaries as in the case of the utterances U_{151} and U_{152} (figure 6).

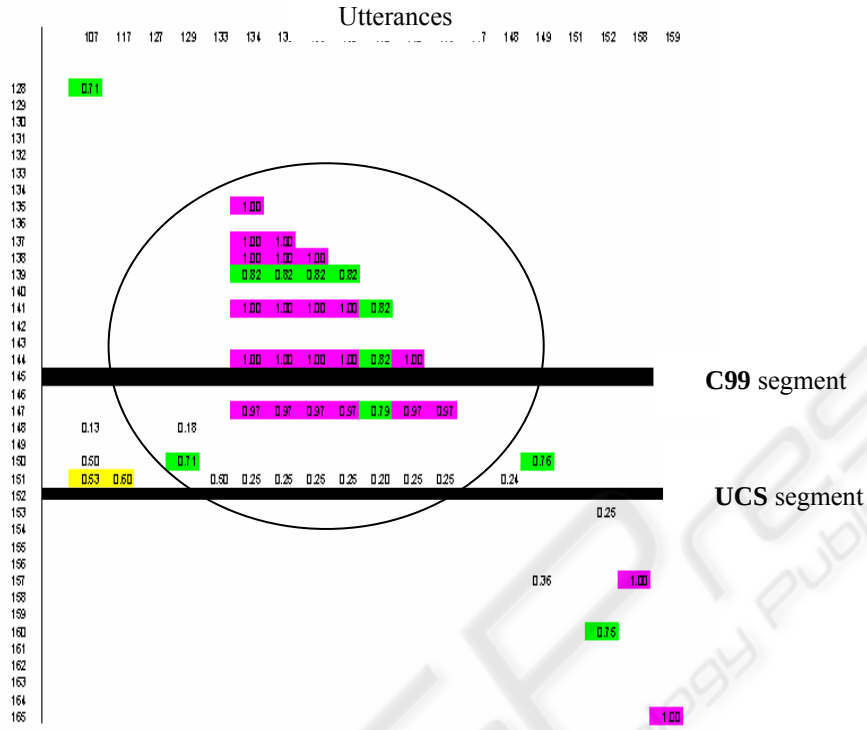

Fig. 6. C99 versus UCS segment

> U_{151} = {yeah, that's one of my actions too. To try and talk to those people or build a new advisory committee.}
>
> U_{152} = {ok , so do you want to move onto the next one ? status of the project, visibility to all members of the project. Does everybody know about the BCSW_server ? I think Fred, Joe and myself, Mary have all logged in. I'm not sure its a web based system at Lancaster where you can arrange directory and document, arrange meeting via this interface, so a lot of the shared information like the bid is on there as a document. So what happens is that anybody can, once you get a username and password, you can log into the system and invite someone else via e-mail. So I'll make sure that everybody gets invited, so the system e-mails you and then gives you some instruction as to how to log in, browse round the directory.}

Fig. 7. Utterances 151 and 152.

Figure 6 shows that the similarity values for utterances 134 to 150 vary between 0.70 to 1.00 indicating an area of topically related utterances. The similarity values between utterance 151 and other utterances vary from 0.20 to 0.60, indicating that a topic is fading away. Whereas the similarity values of utterance 152 and others are all zero indicating a topic shift. When applying the C99 algorithm, the topic change

occurred at utterance 144 which is in the middle of the topic. The topic being discussed in this part of the transcript included the content word 'meeting'. With the C99 boundary, 6 occurrences of the word *meeting*, with the same context, were located on the wrong side of the boundary. This problem is avoided by using the UCS method.

6 Conclusions and Future Development

This paper described a new approach to transcript segmentation combining techniques from both information retrieval and text mining fields. In our application transcript segmentation facilitates the tracking of issues in the transcript and allow us to extract the agents and the actions associated with a particular issue. This approach could be applied to analyse any unstructured text and extract salient information.

The Utterance Cosine Similarity method does not require the use of thesauri and corpora analysis and thereby it avoids the associated problems with domain specific terms. Our approach starts by building an Utterance Index Set (UIS) from the nouns used in the transcript. This UIS is then used to populate vectors for each transcript which can then be compared in order to identify similarities between utterances and therefore building a link between sets of utterances relating to a given topic. Problems in segmentation and the identification of topic boundaries are overcome by examining patterns in the utterance cosine similarity matrix.

In the process of developing the method we have discovered that (1) stemming does not always improve classification accuracy, (2) UCS can be used to obtain word/phrase classifications which adjust with the domain and contains more semantic relationships than those obtained from the thesaurus, (3) when people speak they use few synonyms or hypernyms related to the content words, instead there is evidence of word and phrase repetition, (4) it is possible to classify topics within transcripts without pre-defined classes. We have also identified possibilities for using the UCS method as a means for building ontologies which could be useful in areas like web engineering, question answering and text categorisation.

Further work is being conducted in order to develop an algorithm which can automatically determine the threshold value for the noun filter depending on the transcript content. There is also the need to refine the *Utterance Distance Discriminator* and investigate ways of improving the UCS method to improve its performance for transcripts comprising very long utterances as this may affect the threshold value which is used in the UDD.

Acknowledgements

The authors would like to thank Dr. F Choi for discussions about and the use of the C99 algorithm, Joel Di Manno for his work on the coding of the UCS and UDD algorithms used in the project, and Prof. H. Shah for her comments on the draft of this paper. This work was conducted under the auspices of the TRACKER project, UK EPSRC grant (GR/R12176/01).

References

1. Allan, J., Carbonell, J., Doddington, G., Yamron, J. & Yang, Y. 1998, 'Topic Detection and Tracking: Final Report', in *Proceedings of the DARPA Broadcast news Transcription and Understanding Workshop*.
2. Barzilay, R. & Elhadad, M. 1999, 'Using Lexical Chains for Text Summarization', in *Advances in Automatic Text Summarization*, eds. Mani, I.*, et al.*, MIT Press, Cambridge, MA, Madrid, Spain, pp. 111--121.
3. Beeferman, D., Berger, A. & Laffety, J. 1999, 'Statistical Models for Text Segmentation', *Machine Learning, Special Issue on Natural Language Processing,* vol. 34, no. 1-3, pp. 177-210.
4. Boguraev, B. K. & Neff, M. S. 2000, 'Discourse segmentation in aid of document summarization', in *Proceedings of 33rd Annual Hawaii International Conference on System Sciences, (HICSS)*, IEEE, Maui, Hawaii, pp. 778-787.
5. Chibelushi, C., Sharp, B. & Salter, A. 2004, 'A Text Mining Approach to Tracking Elements of Decision Making: a pilot study', in *Proceeding of the 1st International Workshop on Natural Language Understanding and Cognitive Science, NLUCS 2004*, ed. Sharp, B., INSTICC Press, Porto, Portugal, pp. 51-63.
6. Choi, F., Wiemer-Hastings, P. & Moore, J. 2001, 'Latent Semantic Analysis for Text Segmentation', in *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing.*, pp. 109 - 117.
7. Choi, F. Y. Y. 2000, 'Advances in domain independent linear text segmentation', in *Proceedings of NAACL00*, Seattle.
8. Green, S. 1997, *Automatically Generating Hypertext By Comparing Semantic Similarity*, University of Toronto, Technical Report number 366.
9. Hearst, M. 1994, 'Multi-paragraph Segmentation of Expository Text', in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico, pp. 9-16.
10. Katz, S. M. 1996, 'Distribution of Context Words and Phrases in Text and Language Modelling', *Natural language Engineering,* vol. 2, no. 1, pp. 15-59.
11. Mintzberg, H., Waters, J., Pettigrew, A. M. & Butler, R. 1990, 'Studying deciding: an exchange of views between Mintzberg and Waters, Pettgrew and Butler', *Organisation Studies,* vol. 11, no. 1, pp. 1-16.
12. Passoneau, R. & Litman, D. 1993, 'Intention-based Segmentation: Human Reliability and Correlation with Linguistic Cues.' in *Proceedings of the 31st Annual Meeting of the Association of Computational Linguistics (ACL-93)*, pp. 148-155.
13. Rayson, P. 2001, 'Wmatrix: a Web-based Corpus Processing Environment.' in *ICAME 2001 Conference*, Université Catholique de Louvain, Belgium.
14. Reynar, J. 1998, *Topic Segmentation: Algorithms and Applications*, University of Pennsylvania.
15. Richmond, K., Smith, A. & Amitay, E. 1997, 'Detecting Subject Boundaries within Text: A language independent statistical approach', in *Proceedings of the 2nd Conference on Emperical Methods in Natural Language Processing*, Rhode Island, USA.
16. Stokes, N. 2003, 'Spoken and Written News Story Segmentation using Lexical Chains', in *Proceedings of HLT-NAACL, Student Research Workshop*, Edmonton, pp. 49-54.
17. Yamron, J., Carp, I., Gillick, L., Lowe, S. & Mulbregt, P. V. 1998, 'A Hidden Markov Model Approach to Text Segmentation and Event Tracking', in *Proceedings of ICASSP'98, IEEE*, Seatle, WA, pp. 333-336.