Tracker Text Segmentation Approach: Integrating Complex Lexical and Conversation Cue Features

C. Chibelushi[1] and B. Sharp[2]

[1] School of Computing and IT, University of Wolverhampton
Technology Centre (MI Building), Room MI 317, City Campus - South
Wulfrana, Street, Wolverhampton. WV1 1SB, U.K.

[2] Faculty of Computing Engineering and Technology, Staffordshire University, U.K.

Abstract. While text segmentation is a topic which has received a great attention since 9/11, most of current research projects remain focused on expository texts, stories and broadcast news. Current segmentation methods are well suited for written and structured texts making use of their distinctive macro-level structures. Text segmentation of transcribed multi-party conversation presents a different challenge given the lack of linguistic features such as headings, paragraph, and well formed sentences. This paper describes an algorithm suited for transcribed meeting conversations combining semantically complex lexical relations with conversational cue phrases to build lexical chains in determining topic boundaries.

1 Introduction

The problem of text segmentation has been the recent focus of many researchers as more and more applications require the tracking of topics whether for summarization, classification and/or retrieval tasks of textual documents. Since 9/11 text segmentation became a common technique used to detect the threads contained in instant messaging and internet chat forums for various applications, including information retrieval, expert recognition and even crime prevention [3].Text segmentation can be carried out on audio, video, and textual data. The aim of segmentation is to partition a text into contiguous segments related to different topics. The increasing interest in segmenting conversations is mainly explained by the number of its applications as outlined in Table 1, whose granularity level of the segmentation depends on the size of the chosen units of analysis which varied from utterances to words to phrases.
In this paper we present an algorithm for text segmentation relevant to transcribed meetings involving a multi-party conversation. While previous research has focused mostly on structured texts, broadcast news, and monologues which consist of cohesive stories, our corpus consists of 17 manually transcribed meeting conversations. It includes incomplete sentences, sentences related to social chatting, interruptions, and references by participants made to visual context. Consequently, the analysis of our transcripts poses an additional complexity due to their informal style, the use of visual

Chibelushi C. and Sharp B. (2008).
Tracker Text Segmentation Approach: Integrating Complex Lexical and Conversation Cue Features.
In *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science*, pages 104-113
DOI: 10.5220/0001740501040113

Table 1. A review of the language processing applications to transcripts [5].

Application	Corpus used	Unit of analysis	Application	Corpus used	Unit of analysis
Automatic meeting understanding for a personal office assistance. Gruenstein *et al.* (2005)	Meeting	Utterance	Improving spoken language understanding to improve services in call centres. Begeja *et al.* (2004)	Telephone	Utterance
Information extraction from telephone dialogues for a decision support tool. Boufaden et al. (2001)	Telephone	Utterance	Automatic indexing of lecture speech. For browsing function for lectures and discussions. Kawahara *et al.* (2002)	Lecture speech	Phrases– discourse markers
Dialogue management. Strayer *et al.* (2003)	TRAINS	Utterance	Graphical dialogue annotation comparison tool for doing consensus annotation to reduce coding errors when formulating or verifying theories of dialogue. As training data for statistical models. Yang *et al.* (2002)	TRAINS	Phrase
Organizing and indexing meeting records for efficient access. Kazman *et al.* (1997b)	Meeting	Word	Segmentation of automatically transcribed broadcast news text for Information retrieval. Mulbregt *et al.* (1998)	News broadcast	Word
Segmenting news broadcasts for information retrieval and summarisation tasks. Stokes (2004)	News broadcast	Word	Discourse segmentation of spoken dialogue. To aid in the development of conversational systems. Flammia (1998)	Telephone	Word
Segmenting multi-party conversation to save as a pre-processing stage to areas like information retrieval, summarisation and dialogue understanding. Galley *et al.* (2003)	Meeting	Word	Automatic meeting record creation and access, a tool that can assure accuracy, originality and completeness of meeting records. Waibel *et al.* (2001)	Meeting	Words and phrases
Generation of concise summaries of spoken dialogues to aid in archiving, indexing, and retrieval of various records of oral communication. Zechner (2001)	Telephone	Word	Real-time Summarization of human-human spontaneous spoken language for Information extraction. Kameyama *et al.* (1996)	Human –human conversation	Words and phrases
Topic segmentation for Information retrieval tasks. Blei *et al.* (2001)	News broadcast	Word			
Speech-to-speech translation of spontaneous dialog to support verbal communication with foreign dialog partners in mobile situation. Wahlster (2001)	Human-human conversation	Words	Indexing and gisting meetings, a system that can automatically produce an outline of a meeting. Krisjansson *et al.* (1999)	Meeting	Word
Audio information Access from meeting rooms for archiving, indexing, retrieval and browsing spoken documents. Renals *et al.* (2003)	Meeting		Automatic decision detection. Hsueh *et al.* (2006)	Meetings	Word

cues, and the lack of macro-level text units such as headings, paragraphs as well as their spontaneous and often argumentative nature.

The motivation for our research project stems from the need to analyse a set of transcribed meetings with the view to track a set of decisions and their associated issues and actions discussed in the meetings in relation to software development. These elements are then fed into a database to provide a tracking system to support software development in identifying the decisions made at these meetings and gaining an understanding of the issues and decisions that may have led to any unnecessary rework.

In this paper we begin by reviewing the methodologies associated with text segmentation, and we describe our Tracker Text Segmentation (TTS) approach to segmenting transcribed meeting conversations. Finally we discuss the results and the limitations of our algorithm, and conclude our research outlining future research directions.

2 Previous Work

A review of the literature on text segmentation techniques reveals two distinct approaches: statistically based and linguistically driven methods [5, 14]. Some statistical approaches are based on probability distributions [2], machine learning techniques ranging from neural networks [4], to support vector machines [18] and Bayesian networks [21], while others treat text as an unlabelled sequence of topics using a hidden Markov model [24]. [8] developed a text segmentation tool called C99 which uses a divisive clustering algorithm developed by [20] to identify topic boundaries. The other text segmentation approach is derived from the lexical cohesion theory of [9] and uses terms repetition to detect topic changes [25, 19, 10] n-gram word or phrases [12], or word frequency [20, 1]. Some use lexical chains to identify topic changes [10, 22], or prosodic clues to mark shifts to new topics [13, 19]. However most lexical cohesion-based segmentation approaches use lexical repetition as a form of cohesion and ignore the other types of lexical cohesion such as synonym, hypernymy, hyponymy, meronymy [23]. A different approach is adopted by [16] who combine decision trees with linguistic features extracted from spoken texts.
The above segmentation methods are well suited for written and structured texts making use of their distinctive macro-level structures which are deficient in transcribed texts. In the study of our transcripts the topic boundaries are often fuzzy, some topics are re-visited at different stages of the meeting, and do not always follow the intended agenda, rendering the segmentation process a very challenging task. As a result we needed to develop a segmentation method which could handle the complexity and the lack of structure but building on the macro-level structures pertinent to transcribed texts such as the notion of utterance, the spontaneous speech cue phrases, and domain specific knowledge to build an effective semantic lexical chaining.

3 The Corpus

In our research project we used 17 transcripts recorded from three diverse meeting environments: industrial, organizational and educational, each involving a multi-party conversation and containing an accurate and unedited record of the meetings and corresponding speakers. The meeting transcripts which were varied in size, ranging from 2,479 to 25,670 words, were multi-party conversation, and some had no pre-set agendas. Consequently the analysis of these transcripts posed an additional complexity due to their informal style, their lack of structure, their argumentative nature, and the usage of common colloquial words. The transcripts also contain incomplete sentences, sentences related to social chatting, interruptions, and references by participants made to visual context. In this paper, a corpus with a total of 247238 words is used to illustrate our algorithm for confidentiality reasons.

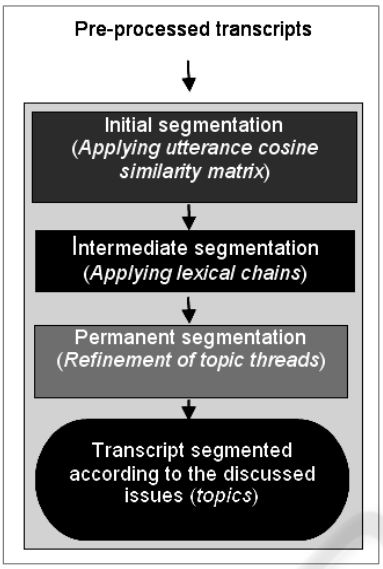

Fig. 1. A Three-Stage Segmentation process.

4 Tracker Text Segmentation Algorithm (TTS)

Our TTS algorithm, which builds on the concept of sliding window, uses the utterance as the base unit of analysis. The algorithm is context driven segmentation, combining lexical chaining method with more semantic complex types of lexical cohesion relationships between words in the transcripts in order to capture their sense relations, such as synonymy, hypernymy (ISA relation), hyponymy (kind-of relation), meronymy (part-of relation) and coordinate terms (e.g. computer and PC). Using WordNet and our extended version of WordNet these sense cases allow us to capture the hierarchical as well the transitivity relationships among the words in the transcripts and enhance the formation of lexical chains.
The study of these transcripts has led to the identification of major speech cue phrases used by the speakers to introduce new topics or highlighting new issues or a problems, examples of these are give in table 2. Prior to segmentation our transcripts have been subjected to pre-processing transcripts which involve tokenisation, POS tagging using WMATRIX, case folding, identification of compound concepts and removal of stop words.
There are three main stages performed by TTS: (i) initial segmentation, (ii) intermediate, and (iii) final segmentation (shown in Fig. 1).

4.1 Initial Segmentation

This stage involves the segmentation of the stream of transcribed meetings into topically cohesive items of discussion. It is based on the sliding window approach devel-

oped by [10] and later adopted by [19], which divides the text into multi-paragraph blocks and then using a vector space model it calculates the similarity of two consecutive blocks using the cosine value, a measure which has been widely used in Information Retrieval (IR) systems. Instead of paragraphs as the core base for segmentation our algorithm computes the similarity between utterances, referred to hereby as the Utterance Cosine Similarity (UCS). Thus instead of measuring the similarity between a query and a document as applied in IR systems, UCS measures the similarity between two utterances.

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 its frequency of occurrence within the transcript. Let us suppose a transcript consists of 33 distinct noun concepts, and one of its utterances is U_{12} which includes the four distinct concept nouns: *size*, *board*, *laptop*, and *edge*, its frequency vector representation will be denoted as follows.

> 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.

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}

In order to identify the similarity (sim) between two utterances, U_i and U_j, the cosine of their frequency vectors should be close or equal to 1. The UCS measure, denoted sim (U_i, U_j), is defined as follows:

$$\text{sim}(U_i, U_j) = \cos(f_i, f_j) = \frac{\sum_k f_{ik} \times f_{jk}}{\sqrt{(\sum_k f_{ik}^2) \times (\sum_k f_{jk}^2)}} \text{ where } 0 \le \cos(f_i, f_j) \le 1.$$

$\sum_k f_{ik} \times f_{jk}$ 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 product of the two vector lengths which is used to normalise the vectors.

The cosine similarity measure assumes that similar terms tend to occur in similar segments. In such instances, the angle between them will be small, and so the cosine similarity measure will be close to 1. Utterances with little in common will have dissimilar terms, the calculated angle between them will be close to π/2 and the UCS measure will be close to zero. A UCS matrix can then be prepared based on the comparison of each utterance with every other utterance in the transcript. An example of this matrix is shown in Fig. 2. The blank lines in Fig. 2 contain zero vectors; these zeros are removed for clarity. In our study after experimentation with our corpus the threshold value was set to five.

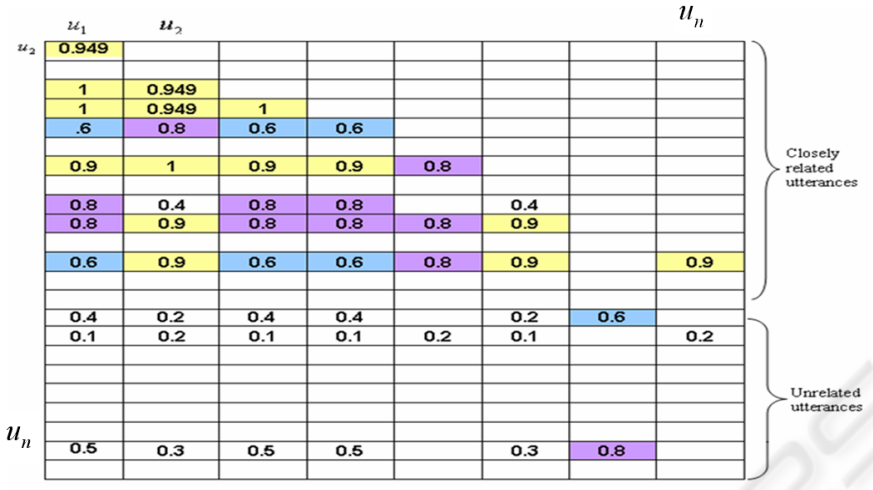

Fig. 2. A typical UCS Matrix.

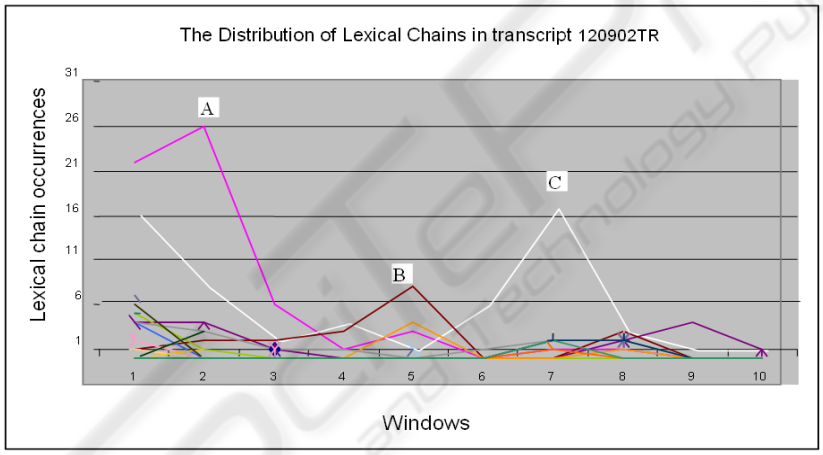

Fig. 3. Distribution of Lexical Chains within Transcript 120902TR.

4.2 Intermediate Segmentation

This stage builds the lexical chains which are generated through selected features and grouped based on their semantic senses relations as they appear in the transcript. Details of the algorithm to generate these chains are found in [5]. The frequency of each chain is examined based on the occurrences of each chain member in the window (Fig. 3). The highest frequency lexical chain is then identified and is used to extend the window or slide the window following the distribution of the topic chain members in the transcript (Fig. 4). As the window expands, it will reach a stage whereby the appearance of any of the members from that particular lexical chain fades away. This is the point where the *intermediate topic boundary* is identified. This step

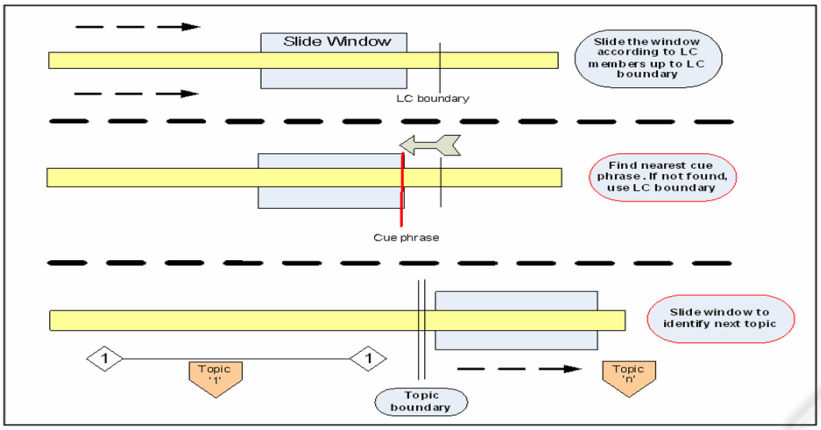

Fig. 4. Sliding Window Effect.

is based on the algorithm of [15] who states that 'a high concentration of chain-begin and end points between the two adjacent textual units is a good indication of a boundary point between two distinct news stories'.

4.3 Final Segmentation

The final segmentation refines further the new segments by searching for any speech cue phrases to confirm the topic boundary or re-examine the boundary of this segment. Unlike the domain independent cues used by [11] and the domain specific cues used by [19], our speech cue phrases were manually extracted from the corpus. An example of these cues is shown in Table 2.

Table 2. Speech Cue Phrases Extracted from our Corpus.

The reason we are having this meeting	So that's it, really for that mode
This meeting is about	can we get some business done then
The main issue is	Can we start with agenda items
The first problem is	We could jump over to
The first item	so do you want to move onto next one
The first agenda item	shall we whiz through onto
The first item on agenda	We seem to be down to
Tell you what before we finish	Any other Business
do you want to move on-to the next one	The other thing is

5 Evaluation and Results

The segmentation was evaluated by comparing TTS against the TextTiling and C99 methods. Three types of evaluation metrics were used, the P_k [1], P_k', and

WindowDiff [17]. The results were very encouraging and showed that TTS has outperformed both algorithms (Fig. 5).

TextTiling was the most underperforming algorithm for this corpus, possibly due to

1. its lexical cohesion-based algorithm which depends mainly on repetition. There are many cases in our transcripts where few consecutive utterances includes no word repetitions and consequently TextTiling identified them as four different topics;
2. its dependence on sentence-based structure and not utterance-based structure. The similarity measure used in TextTiling compares pair of sentences, and consequently relevant to structured and well punctuated texts but unsuitable for our ill-structured corpus;
3. the unsuitability of using a fixed window size.

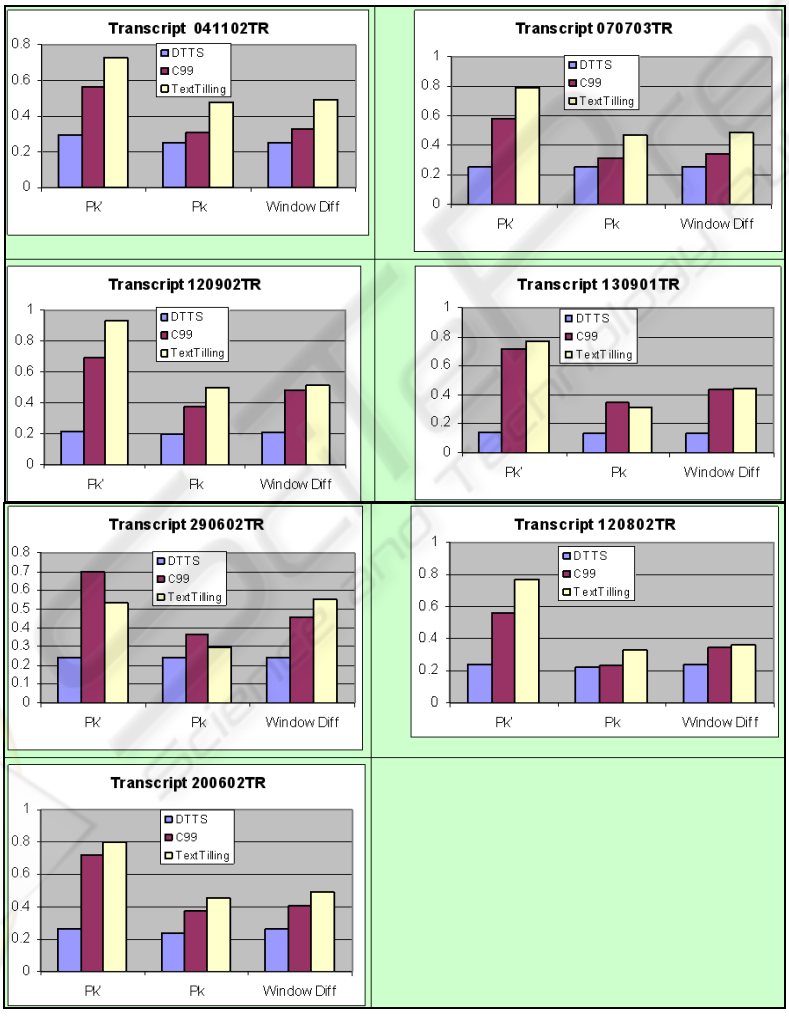

Fig. 5. Evaluation of TTS.

6 Conclusions

The TTS algorithm described in this paper is an iterative process that offers a great potential for analysing transcribed meetings involving a multi-party conversation. The study has extended the use of cosine similarity measure to transcribed texts and improved the performance of lexical chaining methods and text segmentation algorithms by including complex semantic relations and speech specific cue phrases.
Although the evaluation results highlighted the effectiveness of TTS compared to TextTiling and C99, there are few limitations related to the issue of compound words and the POS tagging system used. The identification algorithm of compound words developed in this study, has given, in some situations, unsatisfactory results, as not all the compound words were the result of combined nouns. Also some compound words in the corpus such as 'high voltage line' and 'natural language processing' were not automatically identified, partly due to the limitation of WMATRIX. Future work will attempt to resolve these problems.

References

1. Beeferman, D., Berger, A. and Laffety, J.: Text Segmentation Using Exponential Models, *Proceedings of the Proceedings of EMNLP-2* (1997).
2. Beeferman, D., Berger, A. and Laffety, J.: Statistical Models for Text Segmentation, *Machine Learning, Special Issue on Natural Language Processing*, Vol. 34, No. 1-3, (1999)177-210.
3. Bengel, J., Gauch, S., Mittur, E. and Vijayaraghavan, R.: Chattrack: Chat Room Topic Detection Using Classification, *Proceedings of the The 2nd Symposium on Intelligence and Security Informatics* (*ISI-2004)*, Tucson, Arizona, (2004) 266-277.
4. Bilan, Z. and Nakagawa, M.: Segmentation of On-line Handwritten Japanese Text of Arbitrary Line Direction by a Neural Network for Improving Text Recognition *Proceedings of the Proceedings of the Eighth International Conference on Document Analysis and Recognition*, (2005)157 - 161.
5. Chibelushi, C.: *Text Mining for Meeting Transcripts Analysis to Support Decision Management*, PhD thesis, Staffordshire University (2008).
6. Chibelushi, C., Sharp, B. and Salter, A.: Transcripts Segmentation Using Cosine Similarity Measure, In: B. Sharp (ed.), *Proceedings of the Proceedings of 2nd International Workshop on Natural Language Understanding and Cognitive Science* (*NLUCS2005*) *Collocated with ICEIS-2005*, Miami, USA (2005).
7. Choi, F., Wiemer-Hastings, P. and Moore, J.: Latent Semantic Analysis for Text Segmentation, *Proceedings of the Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing,* (2001)109 - 117.
8. Choi, F. Y. Y.: Advances in domain independent linear text segmentation, *Proceedings of the Proceedings of NAACL00*, Seattle (2000).
9. Halliday, M. and Hasan, R.: *Cohesion in English*, Longman, London (1976).
10. Hearst, M.: Multi-paragraph Segmentation of Expository Text, Proceedings of the Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, (1994)9-16.
11. Hirschberg, J. and Litman, D.: Empirical studies on the Disambiguation of Cue Phrases, *Computational Linguistics*, Vol. 19, No. 3, (1993) 501-530.

12. Kan, M., Klavans, J. L., and. McKeown, K. R.: Linear segmentation and segment relevance. In Proceedings of the Sixth Workshop on Very Large Corpora, (1998).
13. Levow, G.: Prosodic Cues to Discourse Segment Boundaries in Human-Computer Dia]logue, *Proceedings of the Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, M. Strube and C. Sidner, ACL Publisher, USA, (2004) 93-96.
14. Manning, C.: *Rethinking Text Segmentation Models: An Information Extraction Case Study*, University of Sydney (1998).
15. Okumura, M. and Honda, T.: Word Sense Disambiguation and Text Segmentation Based on Lexical Cohesion, *Proceedings of the Proceedings of the 15th International Conference on Computational Linguistics:(COLING-94)*, (1994) 775-761.
16. Passoneau, R. and Litman, D.: Discourse Segmentation by Human and Automated Means, *Computational Linguistics*, Vol. 23, No. 1, (1997)103-139.
17. Pevzner, L. and Hearst, M. evaluation Metric for Text Segmentation, *Computational Linguistics*, Vol. 28, No. 1, (2002)19-36.
18. Renjie, J., Feihu, Q., Xu, L. and Wu, G.: Detecting and Segmenting Text from Natural Scenes with 2-Stage Classification *Proceedings of the Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications:(ISDA'06)*, (2006). 819 – 824.
19. Reynar, J.: Statistical Models for Topic Segmentation, *Proceedings of the Proceedings of the Association for Computational Linguistics*, ACL, College Park, USA, (1999) 357-364.
20. Reynar, J.: *Topic Segmentation: Algorithms and Applications*, PhD Thesis thesis, University of Pennsylvania (1998).
21. Senda, S. and Yamada, K.: A Maximum-likelihood Approach to Segmentation-based Recognition of Unconstrained Handwriting Text, *Proceedings of the Proceedings of the Sixth International Conference on Document Analysis and Recognition*, (2001) 184 – 188.
22. Stokes, N.: Spoken and Written News Story Segmentation using Lexical Chains, *Proceedings of the Proceedings of HLT-NAACL, Student Research Workshop*, Edmonton, (2003) 49-54.
23. Stokes, N.: *Applications of Lexical Cohesion Analysis in the Topic Detection and Tracking Domain.*, PhD Thesis, University College Dublin (2004).
24. Yamron, J., Carp, I., Gillick, L., Lowe, S. and Mulbregt, P. V.: A Hidden Markov Model Approach to Text Segmentation and Event Tracking, *Proceedings of the Proceedings of ICASSP'98, IEEE*, Seatle, WA,:(1998) 333-336.
25. Youmans, G.: A New Tool for Discourse Analysis: The Vocabulary Management Profile, In: *Languages*, (1991)763-789.