

ABSTRACT

Previous work on topic changes in discourse analysis and automatic topic segmentation reported that so-called cue phrases, such as *okay, anyway, alright*, etc. provide valuable information on identifying topic boundaries. The present study investigates what kind of expressions are associated with topic changes, and how they are used in multi-party meeting conversations. This is done by obtaining word collocations from the utterances around topic boundaries and graphically representing the inter-relationship between them as a network.

Keywords: collocation, topic change, meeting, multi-party conversation

1 Introduction

Previous work on topic changes in discourse analysis and automatic topic segmentation has found that cue phrases provide valuable information on the structure of discourse [1, 2, 3]. Cue phrases are ‘words and phrases that directly signal the structure of a discourse [p501]’ [4]. In some topic segmentation algorithms, utterance-initial cue phrases (and sometimes other lexical cues) are used together with other topic-change indicators to detect topic boundaries [5].

However, to our knowledge there has not been any work investigating what sort of expressions (formulaic phrases or co-occurred words) are statistically-significantly associated with topic boundaries. The current paper addresses this gap by identifying word collocations from the utterances around topic boundaries and representing them in a network.

In this study, both grammatical and lexical single words that have statistical significance in relation to topic boundaries were extracted to investigate their co-occurrence patterns. These single words are referred to as ‘cue words’ in this study.

2 Database

The ICSI (International Computer Science Institute) Meeting Recorder Dialogue ACT (MRDA) Corpus [6] is used in this study.

- hand-annotated version of the ICSI Meeting Corpus [7]
- 75 naturally occurring multi-party meetings, each approximately one hour in length
- 53 different speakers appear in the corpus, with an average of approximately six speakers per meeting
- 679 topic change locations (*tc*) annotated in the MRDA Corpus
- a stream of dialogue is segmented in terms of *utterances*, each of which constitutes prosodically one unit

An example of the MRDA Corpus is given in Table 1.

Time	SP	DA	AP	Transcript
442.938-447.028	c3	s	25b.26a	it's ics- uh icsi has a format for frame-level representation of features.
447.808-448.338	cB	s'bk	26b	o.k.
448.22-448.67	c3	fh		um ==
448.388-452.688	cB	s'bu	26b+.27a	that you could call - that you would tie into this representation with like an i.d.
451.177-451.527	c3	s'aa	27b	right .
452.755-453.065	c3	s'aa'r	27b+	right .
453.255-457.595	c3	s	27b+..28a.29a	or - or there's a - there's a particular way in x.m.l to refer to external resources .
453.742-454.122	cB	fh		and ==
457.809-458.249	cB	s'bk	28b	o.k.
458.453-461.423	c3	s:s'co	27b+++29a+	so you would say refer to this external file .

Table 1: An example of the MRDA Corpus. SP = speakers; DA = dialogue acts; AP = adjacency pairs.

3 Word Collocation: Methodology

This study uses five steps to identify word collocations that reveal what expressions are associated with topic changes.

3.1 Step 1

The diacritic and punctuation marks, such as ‘==’, ‘.’ and so on (refer to Table 1) were removed from the transcribed texts, and then they were tokenised. No stemming algorithm was applied in Step 1.

3.2 Step 2

The frequencies of the tokenised words from Step 1 were calculated. After Step 1, the total number of words was 732918, and the number of different words was 12591. Low frequency words were removed for further analysis. Although 10, 50 and 100 were arbitrarily set as thresholds in the original study, the results based on the threshold of 50 are given in this paper. By setting 50 as the threshold, 1035 words were selected in Step 2.

3.3 Step 3

Steps 3 and 4 are for word collocations around topic boundaries. Step 3 involves the identification of those cue words that show a correlation with topic boundaries. To identify these cue words, *Yates' χ^2* test was employed in this study. *Yates' χ^2* test is defined in Formula 1.

$$Yates = \frac{n(ab - bc) - n/2)^2}{(a + b)(c + d)(a + c)(b + d)} \quad (1)$$

Yates' χ^2 test is based on a 2 x 2 contingency table which shows the frequencies of occurrence of all combinations of the levels of two dichotomous variables, in a sample of size N ($= 1035$ at Step 3). In Step 3, the two dichotomous variables are a given word w_i ($i = 0, 1, \dots, N - 1$) and *topic boundary* (TB). The combinations of the levels of those two variables are (w_i, TB) , $(\sim w_i, TB)$, $(w_i, \sim TB)$ and $(\sim w_i, \sim TB)$ as can be seen in Table 2. In the current study, the utterances around topic changes are those that occur within approximately 10 seconds of the utterances assigned a topic change DA.

	w_i	$\sim w_i$	total
TB	a	b	a + b
$\sim TB$	c	d	c + d
total	a + c	b + d	n

Table 2: 2 x 2 contingency table for *Yates' χ^2* test (Step 3).

In Table 2, a, b, c and d are the frequencies of the occurrence of (w_i, TB) , $(\sim w_i, TB)$, $(w_i, \sim TB)$ and $(\sim w_i, \sim TB)$, respectively. Using Table 2 and Formula 1, those words whose χ^2 value rejected the hypothesis under a 0.005-level confidence (the rejection criterion is $\chi^2 \geq 7.8794$) were selected as cue words ($W^{TB} = \{w_0^{TB}, w_1^{TB}, \dots, w_{N-1}^{TB}\}$, N = the total number of cue words in Step 3). The total number of cue words in Step 3 was 140. Table 5 contains the 15 cue words with the highest *Yates' χ^2* values.

3.4 Step 4

Step 4 investigates the dependency of any two cue words in W^{TB} in the utterances around the topic boundaries. That is, all possible combinations of two cue words (w_i^{TB}, w_j^{TB}) of W^{TB} (refer to the matrix given in Table 3) were tested in terms of their dependency ($\chi_{i,j}^2$) using the contingency table given in Table 4. The number of all possible combinations of two cue words is 9870.

	w_0^{TB}	w_1^{TB}	...	w_{N-1}^{TB}
w_0^{TB}	$\chi_{(0,0)}^2$			
w_1^{TB}	$\chi_{(0,1)}^2$	$\chi_{(1,1)}^2$		
...				
w_{N-1}^{TB}	$\chi_{(0,N-1)}^2$	$\chi_{(1,N-1)}^2$...	$\chi_{(N-1,N-1)}^2$

Table 3: A matrix showing all possible co-occurrence patterns of cue words. $N=140$

In Table 4, a, b, c and d are the frequencies of the occurrence of (w_i^{TB}, w_j^{TB}) , $(\sim w_i^{TB}, w_j^{TB})$, $(w_i^{TB}, \sim w_j^{TB})$ and $(\sim w_i^{TB}, \sim w_j^{TB})$, respectively ($i = j = \{0, 1, 2, \dots, 139\}$). Using Table 4 and Formula 1, those two cue words whose χ^2 value rejected the hypothesis under a 0.005-level confidence were selected as the collocations significantly associated with topic changes. 660 collocations were derived in Step 4.

	w_i^{TB}	$\sim w_i^{TB}$	total
w_j^{TB}	a	b	a + b
$\sim w_j^{TB}$	c	d	c + d
total	a + c	b + d	n

Table 4: 2 x 2 contingency table for *Yates' χ^2* test (Step 4).

3.5 Step 5

[8] define *collocation* as ‘an expression consisting of two or more words that correspond to some conventional ways of saying things.’ 660 collocations were identified in Step 4 as being significantly associated with topic boundaries. However, if one tries to graphically present the network of the 660 collocations derived from Step 4, the network becomes too difficult to comprehend. Furthermore, since we obtained these collocations on the basis of *utterances* of varying lengths, there is a possibility that the 660 collocations may include some pseudo-collocations which are not consistent semantically and/or morpho-syntactically in terms of the co-occurrence pattern of two cue words. Since the aim of this paper is to identify expressions that are significantly associated with topic boundaries, we need to eliminate these pseudo-collocations.

We eliminated pseudo-collocations by calculating the mean (μ) and the standard deviation (sd) of the distance between collocated cue words ($|w_i^{TB} - w_j^{TB}|$) in the utterances around topic changes. That is, if the μ of any given collocated cue words is large—which means that they do not appear closely to each other—those two cue words are less likely to be semantically and/or morpho-syntactically cohesive. Likewise, if the sd of any given collocated cue words is large, there is no consistent co-occurrence pattern, and they do not form a fixed expression.

By observing when the network starts making sense by changing the cutoff parameters, a μ and a sd of 4.5 were set as the cutoff parameters to eliminate pseudo-collocations. By setting the cutoff parameters accordingly, the 68 collocations given in Table 6 were selected as genuine collocations. A graph drawing package developed by AT&T called ‘NEATO’ [9] was used to graphically present the inter-relationship of these collocations in the form of a network.

4 Results

The results for the cue words are shown in §4.1, and the results for the collocations of the cue words in §4.2.

4.1 Cue Words

w^{TB}	χ^2
o.k	499.4
agenda	154.2
talk	134.2
about	132.7
alright	129.6
anyway	110.6
let's	105.7
so	103.5
uh	92.3
been	74.3
um	73.1
yeah	67.7
you	59.9
last	56.5
else	54.1

Table 5: Cue words.

4.2 Collocations and Network

Table 6 contains 68 collocations selected in Step 5.

w_i^{TB}	w_j^{TB}	μ	sd	χ^2	w_i^{TB}	w_j^{TB}	μ	sd	χ^2				
1	that'd	great	810	2.0	6	0.0	35	let's	o.k	13	1.9	12	1.1
2	anything	else	671	1.0	25	0.0	36	guess	i	13	1.6	82	2.1
3	go	ahead	582	1.0	17	0.0	37	yeah	um	13	2.6	21	2.2
4	i've	been	406	1.5	34	1.5	38	let	let	12	3.0	3	1.7
5	me	let	188	1.1	17	0.4	39	you	talk	12	4.1	15	2.9
6	oh	yeah	141	1.8	32	1.9	40	why	i	12	3.1	13	2.3
7	about	talk	131	2.6	72	3.7	41	had	i	12	3.4	48	3.6
8	thank	you	111	1.0	12	0.0	42	but	it's	12	3.1	30	3.3
9	playing	i've	67	2.0	5	0.0	43	working	been	12	1.5	7	1.1
10	next	week	52	1.0	10	0.0	44	yeah	so	12	2.3	36	2.7
11	oh	o.k	47	1.3	19	0.8	45	it	been	12	3.6	5	2.0
12	week	last	47	2.0	11	3.6	46	else	uh	11	2.5	4	1.7
13	playing	been	39	1.0	5	0.0	47	next	you	11	4.2	4	2.2
14	on	working	38	2.3	23	2.6	48	it's	so	11	3.3	28	3.9
15	i'll	ahead	34	3.3	3	2.3	49	that	been	11	3.9	18	2.6
16	that'd	good	33	2.3	3	0.5	50	is	meeting	11	4.0	10	3.2
17	mean	i	31	2.2	129	3.7	51	oh	so	10	3.1	7	2.3
18	should	we	26	2.6	72	4.2	52	um	o.k	10	2.7	19	1.9
19	better	it's	24	4.2	5	4.0	53	a	talk	10	3.5	16	2.5
20	mention	wanted	21	2.5	4	1.0	54	working	i've	10	2.8	5	1.3
21	done	we're	20	1.5	11	0.6	55	list	agenda	10	3.3	3	4.0
22	thanks	o.k	20	2.6	3	1.5	56	move	should	10	2.0	4	1.4
23	of	here	20	4.4	15	3.5	57	meeting	know	9	4.3	3	4.0
24	it's	about	19	3.7	4	3.2	58	thing	other	9	1.7	33	3.3
25	send	i'll	18	1.0	4	0.0	59	was	yeah	9	4.0	12	3.6
26	it's	great	18	1.1	6	0.4	60	ahead	o.k	9	4.0	4	4.0
27	is	mean	17	4.3	11	4.1	61	did	we	9	2.7	18	2.1
28	but	but	16	2.5	4	0.5	62	on	talk	9	4.4	5	3.2
29	talked	about	15	2.0	18	2.4	63	since	that	9	2.6	3	2.0
30	but	anyway	15	1.5	14	1.2	64	do	let's	9	2.6	20	3.4
31	eh	uh	14	1.2	4	0.5	65	do	should	9	3.9	34	3.8
32	a	couple	14	2.2	33	2.9	66	let's	alright	8	2.3	3	1.5
33	move	on	14	1.0	8	0.0	67	what's	the	8	3.2	13	2.3
34	let's	let's	14	3.0	6	1.2	68	wanted	you	8	3.3	9	3.0

Table 6: μ and sd = mean and standard deviation of the distance between w_i^{TB} and w_j^{TB} .

It is evident from Table 6 that many of the 68 collocations are semantically and/or syntactically self-explanatory (e.g. the highest 5 collocations: *that'd, great*), *(anything, else)*, *(go, ahead)*, *(i've, been)*, *(let, me)*.

The inter-relationships of the 68 collocations given in Table 6 are graphically presented in the form of a network in Figure 1. Judging from the readability of this network, the cutoff parameters used to remove pseudo-collocations appear to be empirically appropriate. Figure 1 shows one large cluster and eight small clusters comprised of 68 nodes (collocations) and 69 edges (arcs connecting two nodes) in total. There are some pivotal nodes (= cue words) which have more than one edge, such as *i, o.k, its, been, talk*, etc. The network given in Figure 1 allows us to identify various expressions that consist of multiple words.

5 Discussion

By analysing naturally-occurring conversations using the methodology of discourse analysis, [2] reports that a so-called *summary assessment* often appears at the end of a topic. The *summary assessment* is characterised as an utterance contributing little, if any, new information on the topic concerned. Our results support Howe's findings as several assessment expressions can be identified from the network, such as *(that'd, great/good)*, *(it's great/good)*, etc.

Some collocations, such as *(i've, been, working/playing, on)*, are used to report on progress made, or an action taken on a topic (e.g. *so i've been uh working still on the spectral subtraction; i've been exploring a parallel v.a.d without neural network*). The collocation *(we, did)* is also frequently used to report what was done on a topic (e.g. *this is the things [sic] that we did in the last three months*). The collocation *(last, week)* is often accompanied with the utterances of reporting (e.g. *i spent the last week understanding some of the data; so what happened since um last week is ...*). A similar collocation is *(next, week)*, which is used to talk about the action that will be taken on a topic (e.g. *and then continue with this next week*).

The verbs *talk* and *mention*, which are semantically similar, have complicated collocational relationships with other cue words, as can be seen in the collocations *(talk/talked, about, (a, couple))*, *(talk, on)*, *(you, wanted, mention/talk)*, etc. Depending on the subject of the utterance, the actual utterances in which these collocations occur have mainly two functions. If the subject of the sentence is the speaker (first person singular), then the speaker may use them to get the conversation ‘floor’ and ‘talk about’ something (e.g. *so other topics i wanted to talk about are ...*). If it is uttered by the chair of a meeting and the subject of the utterance is a second person, the chair may use them to induce a given participant to ‘talk about’ a given item (e.g. *you had, you wanted to talk about the ...*). One of the characteristics of multi-party meeting conversations is the involvement of a chair whose remarks can strongly influence the direction/progress of a meeting. The collocation *((o.k, go, ahead)*, which has a very high *Yates' χ^2* value (*(go, ahead)=582*), is mostly limited to the chair of a meeting in the current data, and unique to multi-party meeting conversations.

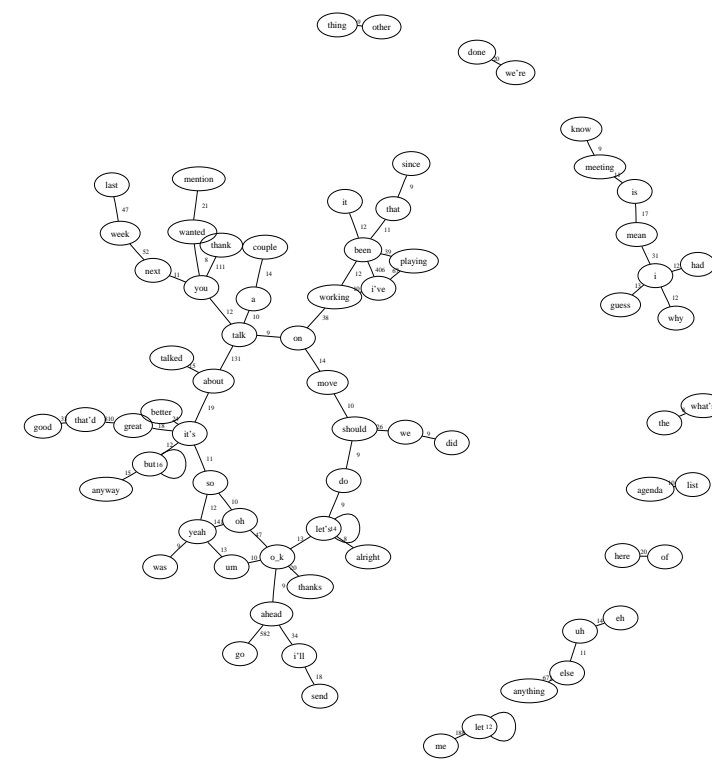


Figure 1: Network of identified 68 collocations.

In the current data, the collocations *(what's, the)* and *(anything, else)* are mostly used as questions. [2] notes that questions (both yes-no and wh-questions) often initiate new topics. However, our results show that only ‘what’ questions (not ‘how’, ‘where’, etc) are significantly associated with topic changes. The collocation *(anything, else)* is the second strongest topic boundary indicator ($\chi^2 = 671$), and is constantly used as a question in our data.

In our data, there are some expressions which directly manipulate the course of conversations, such as *(o.k/alright, let's, do)* and *(we, should, do/move)*. Our results show that the collocation *(let, me)* is also frequently used to obtain the floor of conversation. [2] mentions that changes in topic are a result of the collaborative and negotiative activity of participants, rather than imposition from one speaker. Perhaps due to this negotiative and collaborative nature, expressions like *(o.k/alright, let's, do)* and *(we, should, do/move)* were not reported in [2], in which only two-party conversations were analysed. However, in order to carry out meetings efficiently, these expressions often need to be used (but not always) by the chair. That is, the expressions which overtly attempt to change the direction of a conversation appear to be unique to multi-party meeting conversations.

6 Conclusion

In this paper, we have demonstrated by means of word collocations that there are a large number of expressions significantly associated with topic changes. These expressions have various functions. We have also argued that some expressions are unique to multi-party meetings, as they have not been reported in previous studies based on naturally occurring two-party conversations.

As future research, we plan to investigate if there are any expressions which are specifically correlated with the opening or ending of a topic. It will also be interesting to see how collocational information, as opposed to cue words, improve the performance of automatic topic segmentation systems.

7 Acknowledgements

This study was financially supported by the College of Asia and the Pacific, the Australian National University.

References

- [1] B. Grosz and C. Sidner. 1986. Attention, intentions and the structure of discourse. *Computational Linguistics*, 12(3):175-204.
- [2] M. Howe. 1991. *Topic changes in conversation*. Ph.D thesis, University of Kansas.
- [3] R. Passonneau and D. Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103-139.
- [4] J. Hirschberg and D. Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501-530.
- [5] M. Gallely, K. McKeown, E. Foster-Laussier and H. Jing. 2003. Discourse segmentation of multi-party conversation. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 562-569.
- [6] E. Shriberg, R. Dhillon, S. Bhatag, J. Ang and H. Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. *Proceedings of 5th SIGdial Workshop on Discourse and Dialogue*, 97-100.
- [7] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg and A. Stolcke. 2001. The meeting project at ICSI. *Proceedings of the 1st international conference on Human language technology research*, 1-7.