CHANGING THE TOPIC: HOW LONG DOES IT TAKE?

Mary O'Kane and P.E.Kenne

The University of Adelaide
South Australia 5005
Australia
{mok,pek}@dvcr.adelaide.edu.au

ABSTRACT

We examine the frequency of topic change in Australian court dialogue with a view to automatically changing language models.

1. INTRODUCTION

A formal knowledge of the dynamics of how the focus of attention (mircotopic) changes in spoken dialogue is an important feature of spoken language systems and is particularly useful in automatic processes for switching fine-grain language models. This issue has been studied in text-based systems for some time [1] but relatively little is known about it in verbal dialogue.

2. DATA AND THE PROBLEM

In this paper we investigate the frequency of microtopic change in the setting of the (somewhat stylised) dialogue used in the Australian Federal court system.

In order to study the phenomenon of rate of change of mircrotopic, we examined (in a semi-automated study) several days of transcripts from two Australian Federal Court cases. Table 1 gives details about the sizes of these cases. (In this study we only examined the training data from these two cases.)

Figure 3 gives an excerpt from case c1 and illustrates some of the difficulties likely to be encountered in studying topic change in court interactions: the assignment of topic is essentially arbitrary, depending on the level of detail desired, and the topic changes may occur very frequently, particularly in interactions between a lawyer and a witness. In this figure, questions and statements from the lawyer are preceeded by $>$ and responses from the witness are preceeded by $<$.

As a first step, we examined the distribution of utterance lengths within both cases. A general observation made from a (limited) subset of the transcripts indicated that lawyers often introduced a change of focus by a longer utterance. Figures 1 and 2 show the distribution of utterance lengths.

	c1	c2
Training size (words)	180K	250K
Test size (words)	20K	38K
% Coverage	85	95

Table 1: Training/test set details

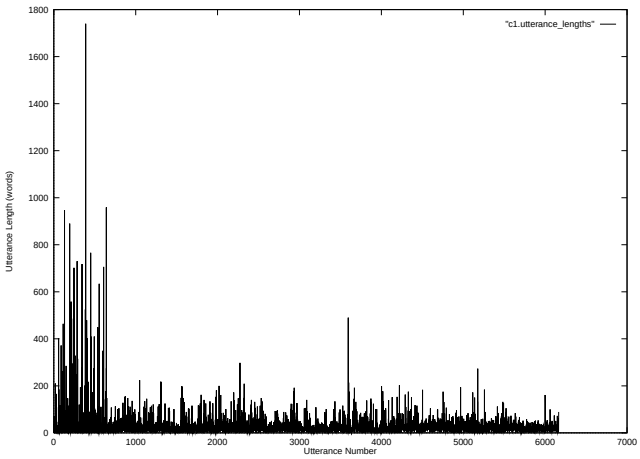

Figure 1: Utterance length versus utterance number, case c1

Unfortunately, utterance length does not provide a reliable indicator of topic change (see Kenne and O'Kane [2]). Inspection of the transcripts shows that utterance length is slightly more likely to signal topic change when a lawyer is questioning a witness than when interacting with the judge. The unreliability of utterance length as a predictor is illustrated by interactions such figure 3.

Another approach in discovering microtopics is by removing the common words from transcript, using previous findings [4]. The words that are left are examined automatically in a series of overlapping windows for high frequency of occurrence. These words are the potential indicators of the current microtopic and they may serve to localise the microtopic within a small number of windows. This approach

was tried by removing the 50, 200 and 1000 most frequent words from the transcripts. The remainder of the transcript was examined for high frequency words occuring in a series of overlapping windows.

Table 2 shows the results for case c2, using a window size of 600 words, with an overlap of 200 words. It is possible to follow microtopic change with time and also observer related topics appearing within a given window. Some of the results from this table may be used to provide (additional) semantic grouping. Experiments with systems such as wordnet (Miller *et al* [3]) applied to the transcripts to generate semantic groupings reveal the limitations of domain-specific training data.

We also found that the microtopic change occurs generally every few hundred words, and from table 2 the longest run is approximately 3200 words. Inspection of the transcripts shows that such long runs appear not to be common and that an upper bound appears to be approximately 4000 words.

This study indicates that spoken language systems need to be able to anticipate microtopic change every few hundred words and that microtopics recycle. After the microtopic has reappeared once or twice it is also possible to predict local ordering of microtopics.

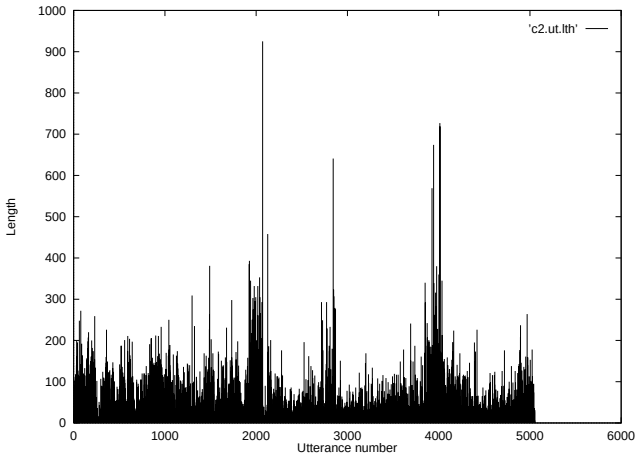

Figure 2: Utterance length versus utterance number, case c2

3. REFERENCES

1. G. Hirst. *Anaphona in Natural Language Understanding: a Survey*. Springer-Verlag, Berlin, New York, 1981.
2. P.E. Kenne and M. O'Kane. Topic change and local perplexity in spoken legal dialogue. In *Proceedings International Conference on Spoken Language Processing 96*, 1996.
3. G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K.J. Miller. Wordnet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–245, 1990.
4. M. O'Kane P.E. Kenne and H. Pearcy. Derivation of a large speech and natural language database through alignment of court recordings and their transcripts. In *Proceedings ICSLP94*, pages 1819–1822, 1994.

ROTHMAN > Has the union any involvement in the
superannuation area ?
< We have . The union , along with some of the other building
unions , have sponsored the Building Unions Superannuation
Scheme Queensland .
> Yes . Now , that is a separate scheme from the Buildings Unions
Superannuation Scheme that operates in other states ?
< It is , yes .
> Yes . And it's confined in its operation to Queensland .
Is that right ?
< It is .
> And does it come - does it have any relationship to the off-site joinery
shopfitting area ?
< It does , yes .
> And what is that ?
< Well , the - BUSS Queensland very early on adopted the 3%
occupational super so they - it modified its deed to accept
the 3% occupational super . So there's always been - for a long
time there's been those two standards for 3 per cent occupational and the
on-site component .
> So do I take it that when you say two standards , there is a standard
which is an over-award standard for the on-site construction ?
< Yes .
> And there is also what I will call the standard 3% superannuation ?
< Yes .
> Which operates by award provision in the off-site area . Is that right ?
< Correct .
> And do you know - if you don't say so - do you know if there are
employers that pay in in relation to off-site area , into that
superannuation scheme ?
< Well , I know that there is .
> Now , Mr Trohear , you are also a member of the divisional conference
and divisional executive at a national level of the BWIU division of the
CFMEU . Is that right ?
< I am , yes .
> And prior to its amalgamation , you were in fact a member of the
national conference and national executive of the BWIU . Is that
right ?
< I have been .
> And you are also a member of the national conference and national
executive of the CFMEU ?
< I am .
> Yes . And you are that by virtue of your - you hold those positions by
virtue of your position in Queensland , that is the Queensland divisional
branch secretary . Is that right ?
< Yes .
> You were in the courtroom this morning when I opened in relation to
the formation of the CFMEU . Do you recall that ?
< Yes .
> Now , can you just explain to the commission , briefly , the purpose of the
formation of the CFMEU , in terms of industry unionism ?

Figure 3: Excerpt of transcript of case c1

Postion	Top 50 removed	Top 200 removed	Top 1000 removed
216001	sea	grass	survey photographs halodule
216401	seagrass	suspended	solids reduction
216801	seagrass	dredging	reduction
217201	seagrass	turbidity	solids
217601	seagrass	beds seasonal	seasonal
218001	seagrass	sediment	seasonal
218401	sediment	sediment	values
218801	seagrass	detritivores	detritivores
219201	seagrass detritivores	detritivores	detritivores
219601	do dr	tests testing	florence's
220001	do dr tests bay	tests	replication
220401	those	magnetic island	cleveland quantity mainland cape
220801	currents magnetic island modelling	currents magnetic modelling	westerly
221201	modelling	modelling	patterson
221601	seagrass	dispersion turbidity beds	quantity
222001	seagrass	beds	quantity
222401	seagrass beds	beds	cyclones
222801	light	light	cyclones
223201	ward	ward	miller
223601	seagrass	aware conclusion	relied quantity florence's miller
224001	done	am	scenario
224401	seagrass	beds	expertise articles quantitative
224801	growth	growth	growth
225201	bay	growth	growth
225601	seagrass	didn't	computer journals
226001	seagrass	particles	leaves smaller
226401	particles	particles	membrane
226801	size	size	penetration
227201	light	light	penetration
227601	seagrass light	light	salinities intensity blowouts
228001	know	decline	decline
228401	know	grass	seeds
228801	grass sea	grass	blades
229201	grass sea	grass	blades
229601	know	grass organisms beds	previous
230001	know	light	dark
230401	know	light	dark
230801	seagrass cyclone know	cyclone	sewerage
231201	seagrass	discharge '90 june	'90
231601	case	discharge proceed	substances journey tyres
232001	doesn't know seagrasses	doesn't seagrasses	she
232401	seagrass	seagrasses	decline
232801	know	sediment traps	traps
233201	point know effects	effects	traps anchorage
233601	area	coming people government	root
234001	seagrass	aware	lympus
234801	prawn	cairns	cairns
235201	seagrass know	prawn	positive cairns involving types
235601	seagrass	beds trawlers	positive cod
236001	seagrass	lot	trawling energy paddock lumps

Table 2: Keywords for c2 high frequency words removed