

# An analysis of the effectiveness of temporal mapping and speech recognition for content-based multimedia indexing

Matt Bouamrane and **Saturnino Luz**

Dept. of Computer Science, Trinity College Dublin, Ireland

**SMAP'06** ○ **Athens** ○ **Greece**

# Overview

- What has been done:
  - An evaluation of a strategy for browsing multimedia meeting records based on temporal mapping and automatic speech recognition
- Meeting records:
  - collaborative meetings through audio (speech) and text
  - metadata: time-stamped actions on the textual components speech segments
- An implementation
  - MeetingMiner

# Background

- Multimedia information retrieval: different challenges for different data sources.
- Compare retrieval from:
  - recorded data generated in media production environments (e.g. television news broadcast, films, etc) to
  - recorded unstructured data generated in spontaneous, highly interactive situations such as computer-supported on-line meetings
- Techniques used for the former cannot always be (reliably) applied to the latter...
- But we might be able to take advantage of the interactive nature of the latter.

# The focus of our work

- Retrieval of relevant information from recorded meetings.
- A pragmatic restriction: We focus on meeting recordings comprising two data streams:
  - Audio (recorded speech)
  - (collaboratively written) Text

# The focus of our work

- Retrieval of relevant information from recorded meetings.
- A pragmatic restriction: We focus on meeting recordings comprising two data streams:
  - Audio (recorded speech)  
continuous, time-based data
  - (collaboratively written) Text

# The focus of our work

- Retrieval of relevant information from recorded meetings.
- A pragmatic restriction: We focus on meeting recordings comprising two data streams:
  - Audio (recorded speech)  
continuous, time-based data
  - (collaboratively written) Text  
space-based data

# Existing approaches

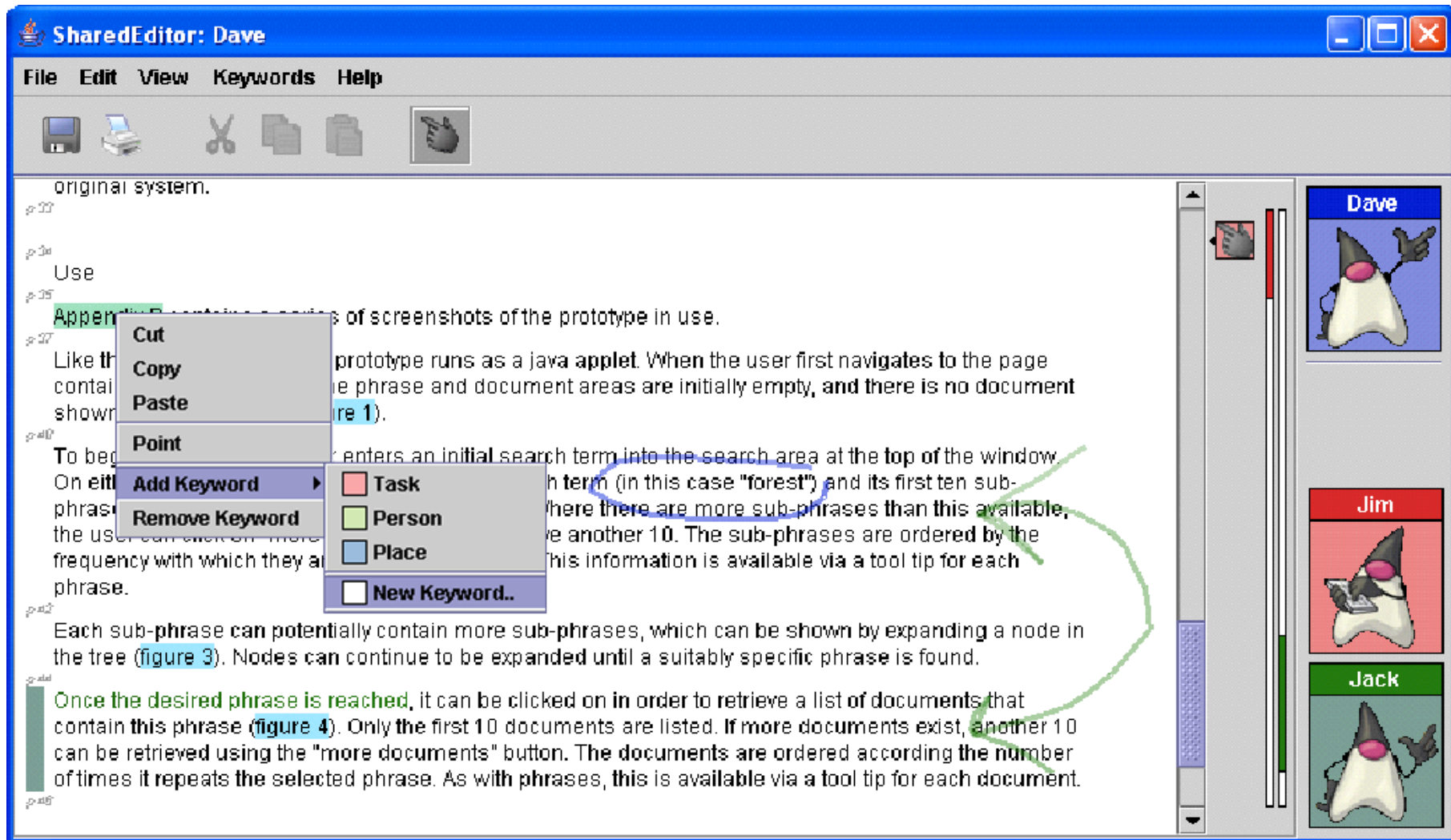
- “Meeting browsers” of various sorts, often combining:
  - meeting summarisation
  - topic segmentation
  - meeting visualisation
  - time-based (linear) access
- The prevailing paradigm: conversion of time-based data into a space-based form (e.g. via automatic speech recognition)
- Limitations:
  - focus on outcomes rather than processes
  - shortcomings of conversion technologies

# Our approach

- It consists of exploring interaction-generated metadata through
  - time-stamping of actions on space-based media objects (e.g. text paragraphs, drawings, tables etc)
  - synchronisation
- It is based on certain assumptions about user behaviour
- It could serve as a basis for uncovering semantic relationships between media streams

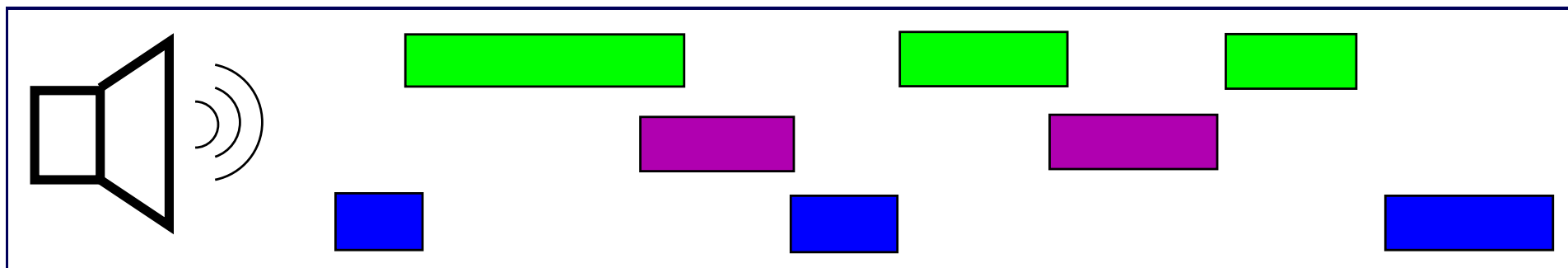
# A meeting capture environment

- REcording COLlaboraborative EDitor: RECOLED



# Annotation elements

- Segmentation:
  - Speech: “talk spurts” and silences
  - Text: paragraphs



```
<timestamp actionid='57' agent='2'  
  action='Insert' start='486' end='495' />  
  budget of 3000 from the  
<keyword type='entity'>student union</keyword>  
  ...
```

# Annotation elements: time stamps

```
<segment id='4.1'>
  <timestamp actionid='17' agent='2'
    action='ninsert' start='215' end='215' />
  <timestamp actionid='19' agent='2'
    action='Insert' start='215' end='217' />
  <timestamp actionid='20' agent='2'
    action='Delete' start='220' end='222' />
  <timestamp actionid='21' agent='2'
    action='Insert' start='221' end='221' />
  <timestamp actionid='22' agent='2'
    action='Insert' start='222' end='226' />
  <timestamp actionid='24' agent='1'
    action='Insert' start='231' end='231' />
  <timestamp actionid='57' agent='2'
    action='Insert' start='486' end='495' />
    budget of 3000 from the student union
</segment>
```

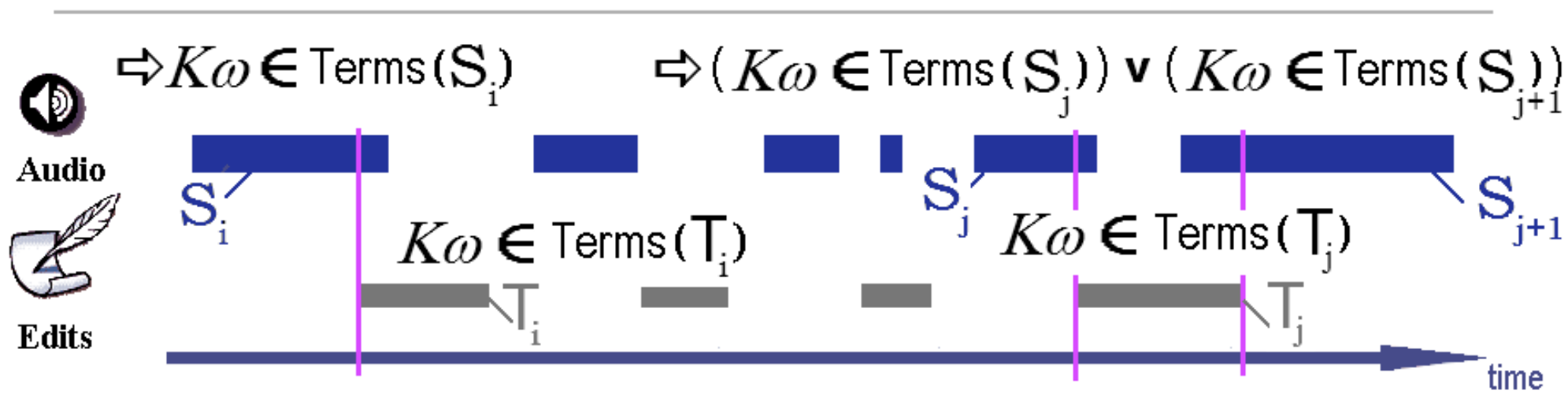
# Annotation elements: action descriptions

```
<action id="58" type="Insert" start="496" end="498"
  paragraphs="4.1.1" startOffset="0">
  maybe chga
</action>
<action id="59" type="Delete" start="498" end="499"
  paragraphs="4.1.1" startOffset="8">
  ga
</action>
<action id="60" type="Insert" start="499" end="502"
  paragraphs="4.1.1" startOffset="8">
  arge people
</action>
<action id="75" type="Gesture" start="1279"
  end="1283" paragraphs="16.4" startPar="16.4"
  points="(233,17),(222,14),(274,17),(233,17)">
  Booking
</action>
```

- Segments and timestamps are compiled into a “Multimodal Activity Matrix” (MAM)

# Assumptions

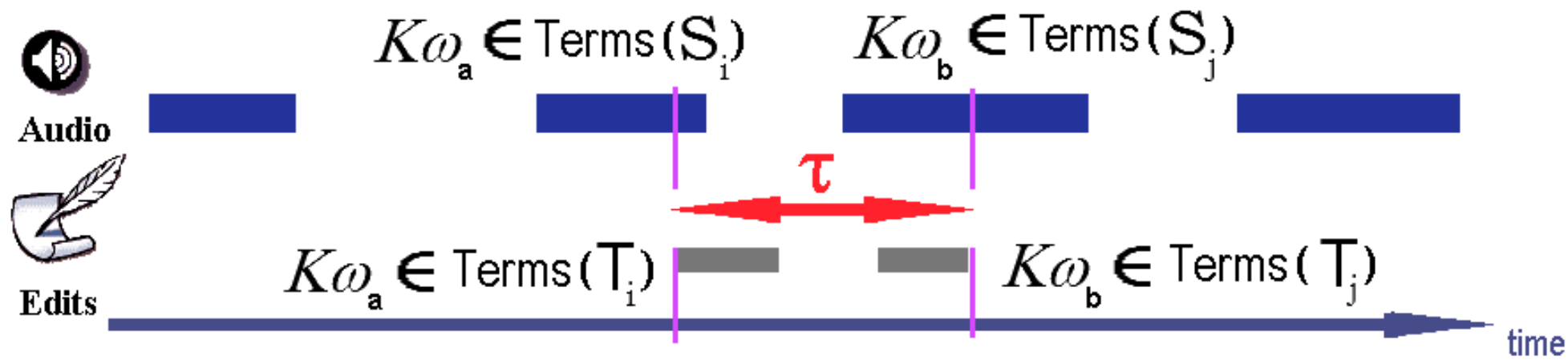
- Keyword occurrence: when participants manipulate (e.g: type, delete or point at) a specific word, they are likely to have uttered it shortly beforehand or alternatively, are about to utter it shortly afterwards



# Topic search assumption

- Semantic links: Two terms of a user query that occur in close temporal proximity are likely to be semantically related

if  $TS(K\omega_a, K\omega_b) \wedge (t(K\omega_b) - t(K\omega_a)) < \tau \Rightarrow \mathbf{sem}(K\omega_a, K\omega_b) = \mathbf{true}$



# The MeetingMiner

- Search and browsing based on the above assumptions.

The screenshot displays the MeetingMiner application window. The interface is divided into several sections:

- Top Bar:** Contains menu options (File, Edit, View, Keywords, Help) and playback controls (skip played, skip silence, cut, repeat, stop).
- View Modes:** Switches between Audio View and Paragraph View.
- Text View:** Shows a transcript of the meeting with highlighted keywords like "ticket" and "price".
- Paragraph List:** A list of paragraphs (p0 to p7.1) extracted from the audio, such as "Organising Music Gig in Dublin" and "The Band: Death Cab for Cutie".
- Keyword List:** A table showing the frequency of keywords found in the transcript.
- Participants Insertions:** A list of words inserted by participants, such as "March" and "18 max".
- Meeting Participants:** A list of participants with their names and avatars, including "Matt" and "Gerrard".
- Bottom Bar:** Contains playback controls and a timestamp "21.m - 30.s".

Remove	KEYwords	TF-ITF
Alphabet	Time	Freq.
	people	: 18
	piece	: 2
	play	: 5
	pm	: 3
	polar	: 2
	posters	: 1
	price	: 5
	probabky	: 2

Participants Insertions:

- March
- 18 max

Meeting Participants:

- Matt
- Gerrard

# Evaluation 1: keyword search

- Methodology: analytical evaluation
- Conditions: Temporal mapping, ASR and combined
- 20 randomly chosen keywords per meeting, from the following data:

Meeting ID	Number of Written Keywords in Meeting	Meeting Duration
A	116	55 min - 44 s.
B	140	1h - 20 min
C	48	37 min - 32 s.
D	61	47 min - 57 s.
E	85	44 min - 40 s.
<b>TOTAL</b>	450	4 h. - 25 min - 53 s.

# Measures

- Precision (with respect to speech segments retrieved):

$$\pi = \frac{|\{S_i | S_i \in KS(Kw) \wedge Kw \in Terms(S_i)\}|}{|KS(Kw)|}$$

- Recall (with respect to speech segments containing the targeted keywords):

$$\rho = \frac{|\{S_i | S_i \in KS(Kw) \wedge Kw \in Terms(S_i)\}|}{|\{S_j | Kw \in Terms(S_j)\}|}$$

# Results

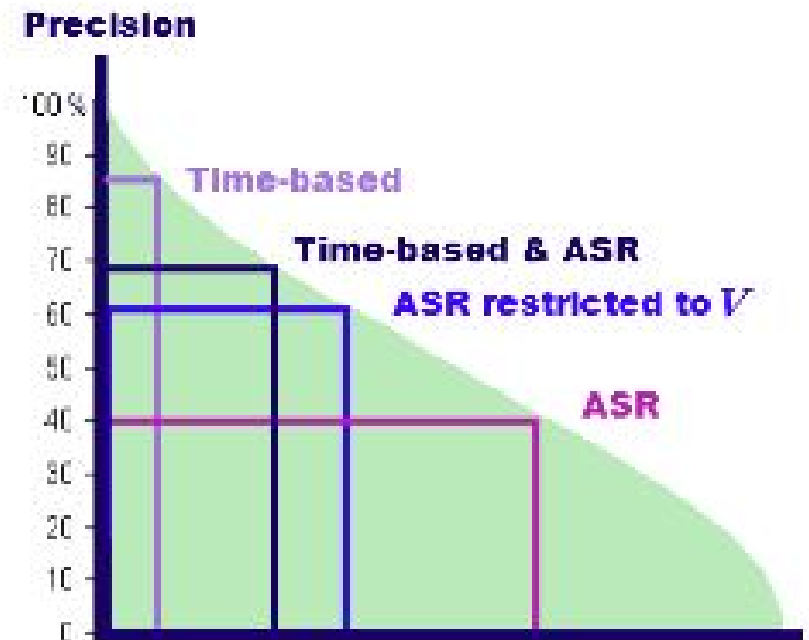
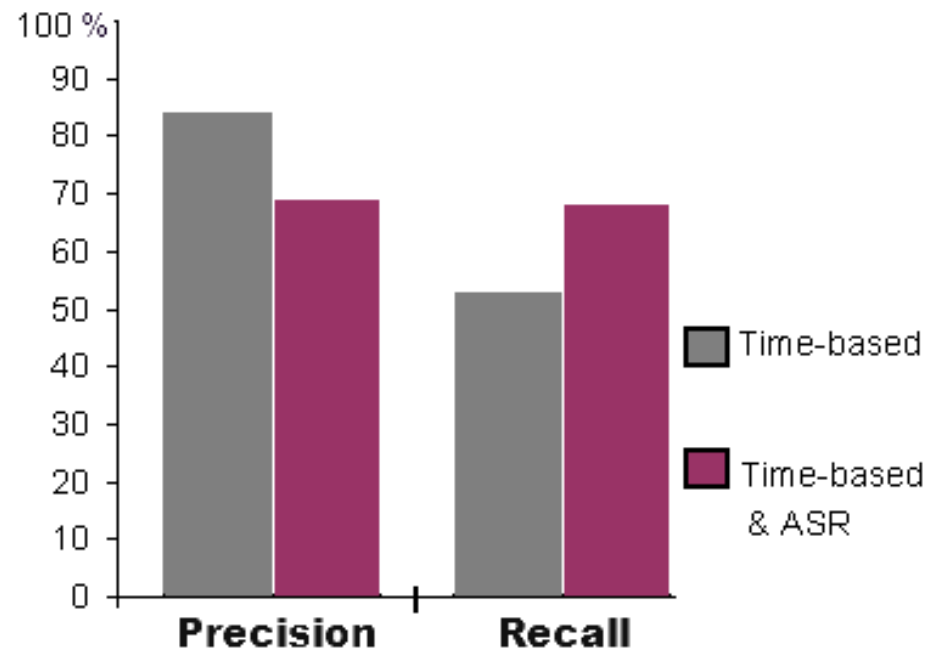
- Temporal mapping only:

Keywords	WTF	STF	Precision	Recall
Sum - $\Sigma$	338	1004		
Mean - $\mathcal{E}(X)$			0.844	0.531
Std. Error			3.2%	4.4%

- Segments found are of good quality but many relevant segments are missing.

# Combining temporal mapping and ASR

- With ASR (WER = 39.4%) we get better recall, at the expense of precision.



## Evaluation 2: topic search

- Method: search for “observations of interest”
- Similar to BET (Wellner et Al. 2005)

Speech	True hit	False alarm
the price of a ticket is 15 quid	X	
while he got the tickets, I went to check the price of ice-creams...		X

- Also consider misses and no match

# Results

- For a total of 48 topic searches we obtained:

Topic Search	TH	FA	Miss	No Match
48	61	7	45	5

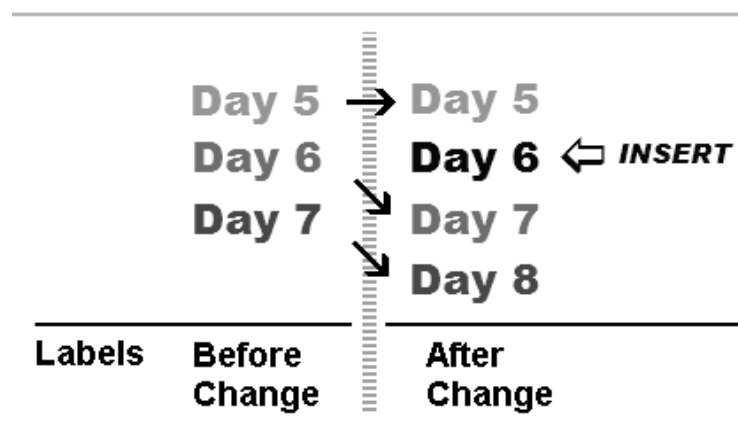
- But, for a given user query, if a match is returned it is relevant in 90.4% of cases.
- Further evaluation is needed to assess whether this improves browsing performance in real search tasks

# Some conclusions

- Retrieval through temporal mapping takes advantage of certain psycholinguistic characteristics of dialogue, particularly **priming** (repetitions, linguistic convergence).
- Tracking even a limited set of actions can suffice for identification of keyword clusters and aid ASR-assisted search
- Inverse time difference between query terms provide a **reliable** (if not **exhaustive**) indicator of semantic relatedness.

# Limitations of content-based indexing

- Consider the scenario below where participants are planning a trip and taking notes.
- Half-way through the meeting they decide to insert an extra day, so the activities that were planned for day 6 will now take place on day 7:



- The speech segments containing the discussion about day 7 now become hard to retrieve, even with perfect ASR!

# Work in progress

- Further exploration of “context” in the space-based medium:
  - E.g.: map from segments (rather than actions) and
  - explore the fact that, once created, text segments keep their identity throughout the meeting, regardless of content
- Comprehensive usability evaluation.