DCU at the NTCIR-9 SpokenDoc Passage Retrieval Task

Maria Eskevich, Gareth J.F. Jones

Centre for Digital Video Processing
Centre for Next Generation Localisation
School of Computing
Dublin City University
Ireland

December, 8, 2011

Outline

Retrieval Methodology
- Transcript Preprocessing
- Text Segmentation
- Retrieval Setup

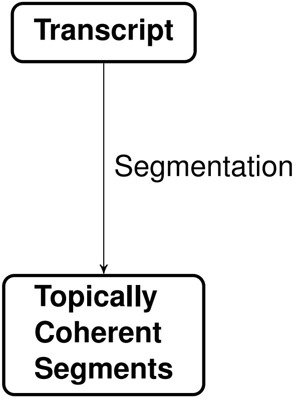
Results
- Official Metrics
- uMAP
- pwMAP
- fMAP

Conclusions

Retrieval Methodology

Transcript

Retrieval Methodology

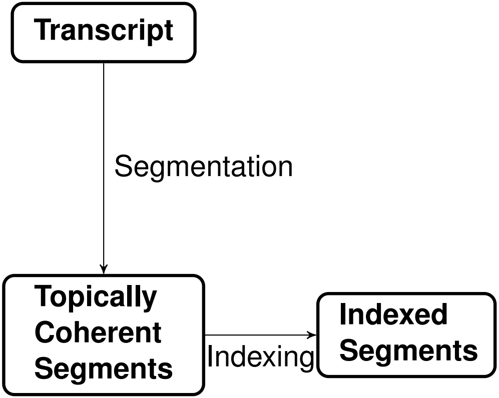


Transcript

Segmentation

Topically Coherent Segments

Retrieval Methodology



Transcript

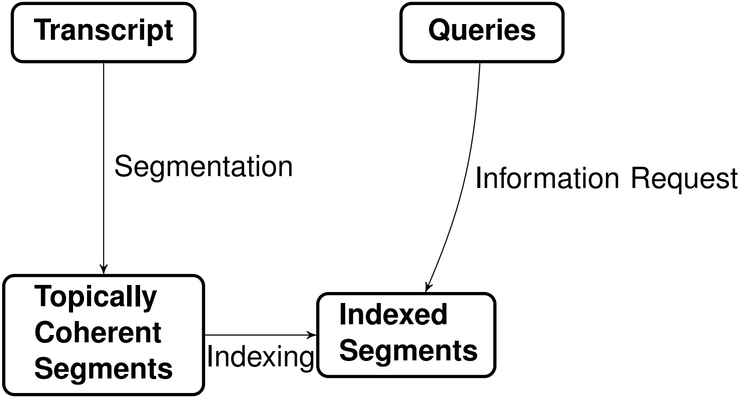
Segmentation

Topically Coherent Segments

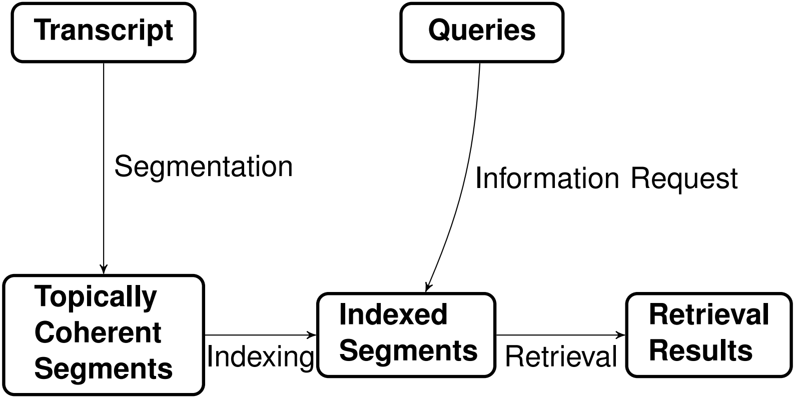
Indexing

Indexed Segments

Retrieval Methodology



Retrieval Methodology



Transcript

Queries

Segmentation

Information Request

Topically Coherent Segments

Indexing

Indexed Segments

Retrieval

Retrieval Results

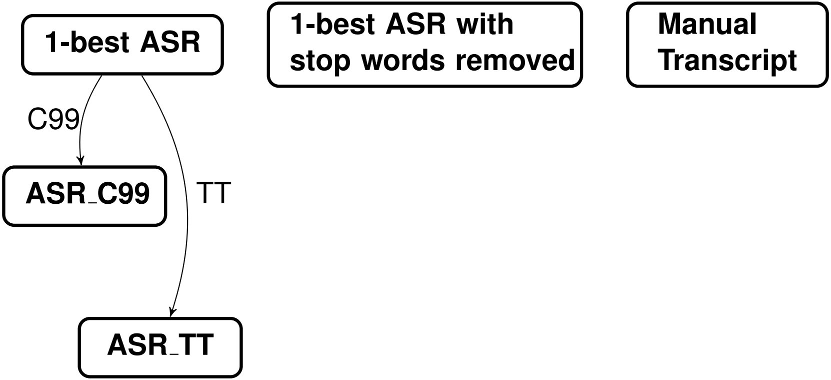
6 Retrieval Runs

1-best ASR

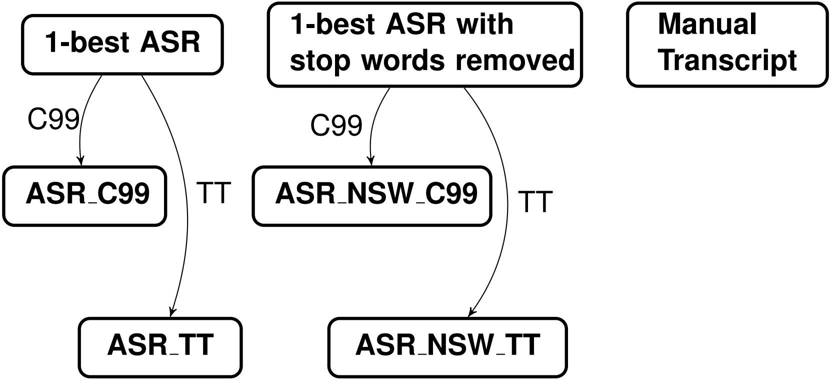
1-best ASR with stop words removed

Manual Transcript

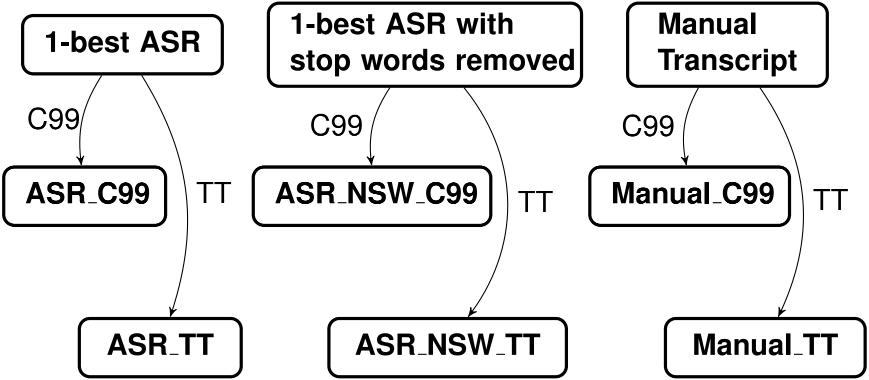
6 Retrieval Runs



6 Retrieval Runs



6 Retrieval Runs



6 Retrieval Runs

ASR_C99

ASR_NSW_C99

Manual_C99

ASR_TT

ASR_NSW_TT

Manual_TT

Transcript Preprocessing

- Recognize individual morphemes of the sentences: ChaSen 2.4.0, based on Japanese morphological analyzer JUMAN 2.0 with ipadic grammar 2.7.0

Transcript Preprocessing

- Recognize individual morphemes of the sentences: ChaSen 2.4.0, based on Japanese morphological analyzer JUMAN 2.0 with ipadic grammar 2.7.0
- Form the text out of the base forms of the words in order to avoid stemming

Transcript Preprocessing

- Recognize individual morphemes of the sentences: ChaSen 2.4.0, based on Japanese morphological analyzer JUMAN 2.0 with ipadic grammar 2.7.0
- Form the text out of the base forms of the words in order to avoid stemming
- Remove the stop words (SpeedBlog Japanese Stop-words) for one of the runs

Text Segmentation

Use of the algorithms originally developed for text:
Individual IPUs are treated as sentences

- **TextTiling:**
  - Cosine similarities between adjacent blocks of sentences
- **C99:**
  - Compute similarity between sentences using a cosine similarity measure to form a similarity matrix
  - Cosine scores are replaced by the rank of the score in the local region
  - Segmentation points are assigned using a clustering procedure

Retrieval Setup

SMART information retrieval system extended to use language modelling with a uniform document prior probability.

Retrieval Setup

SMART information retrieval system extended to use language modelling with a uniform document prior probability.

A query *q* is scored against a document *d* within the SMART framework in the following way:

$$P(q|d) = \prod_{i=1}^{n} (\lambda_i P(q_i|d) + (1 - \lambda_i) P(q_i))$$

Retrieval Setup

SMART information retrieval system extended to use language modelling with a uniform document prior probability.

A query q is scored against a document d within the SMART framework in the following way:

$$P(q|d) = \prod_{i=1}^{n}(\lambda_i P(q_i|d) + (1 - \lambda_i)P(q_i))$$

where

- $q = (q_1, \ldots q_n)$ is a query comprising of n query terms,
- $P(q_i|d)$ is the probability of generating the i^{th} query term from a given document d being estimated by the maximum likelihood,
- $P(q_i)$ is the probability of generating it from the collection and is estimated by document frequency

Retrieval Setup

SMART information retrieval system extended to use language modelling with a uniform document prior probability.

A query q is scored against a document d within the SMART framework in the following way:

$$P(q|d) = \prod_{i=1}^{n} (\lambda_i P(q_i|d) + (1 - \lambda_i) P(q_i))$$

where

- $q = (q_1, \ldots q_n)$ is a query comprising of n query terms,
- $P(q_i|d)$ is the probability of generating the i^{th} query term from a given document d being estimated by the maximum likelihood,
- $P(q_i)$ is the probability of generating it from the collection and is estimated by document frequency

The retrieval model used $\lambda_i = 0.3$ for all q_i, this value being optimized on the TREC-8 ad hoc dataset.

Outline

Retrieval Methodology
- Transcript Preprocessing
- Text Segmentation
- Retrieval Setup

Results
- Official Metrics
- uMAP
- pwMAP
- fMAP

Conclusions

Results: Official Metrics

Transcript type	Segmentation type	uMAP	pwMAP	fMAP
BASELINE		0.0670	0.0520	0.0536
manual	tt	**0.0859**	**0.0429**	**0.0500**
manual	C99	0.0713	0.0209	0.0168
ASR	tt	**0.0490**	**0.0329**	**0.0308**
ASR	C99	0.0469	0.0166	0.0123
ASR_nsw	tt	0.0312	0.0141	0.0174
ASR_nsw	C99	0.0316	0.0138	0.0120

Results: Official Metrics

Transcript type	Segmentation type	uMAP	pwMAP	fMAP
BASELINE		0.0670	0.0520	0.0536
manual	tt	**0.0859**	**0.0429**	**0.0500**
manual	C99	0.0713	0.0209	0.0168
ASR	tt	**0.0490**	**0.0329**	**0.0308**
ASR	C99	0.0469	0.0166	0.0123
ASR_nsw	tt	0.0312	0.0141	0.0174
ASR_nsw	C99	0.0316	0.0138	0.0120

- Only runs on the manual transcript had higher scores than the baseline (only uMAP metric)

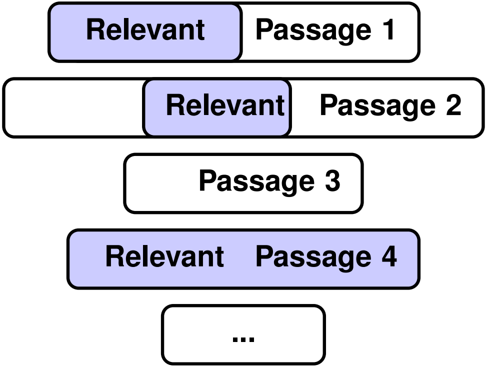
Results: Official Metrics

Transcript type	Segmentation type	uMAP	pwMAP	fMAP
BASELINE		0.0670	0.0520	0.0536
manual	tt	**0.0859**	**0.0429**	**0.0500**
manual	C99	0.0713	0.0209	0.0168
ASR	tt	**0.0490**	**0.0329**	**0.0308**
ASR	C99	0.0469	0.0166	0.0123
ASR_nsw	tt	0.0312	0.0141	0.0174
ASR_nsw	C99	0.0316	0.0138	0.0120

- Only runs on the manual transcript had higher scores than the baseline (only uMAP metric)
- TextTiling results are consistently higher than C99 for all the metrics for manual and ASR runs

Time-based Results Assessment Approach

For **each run** and **each query**:

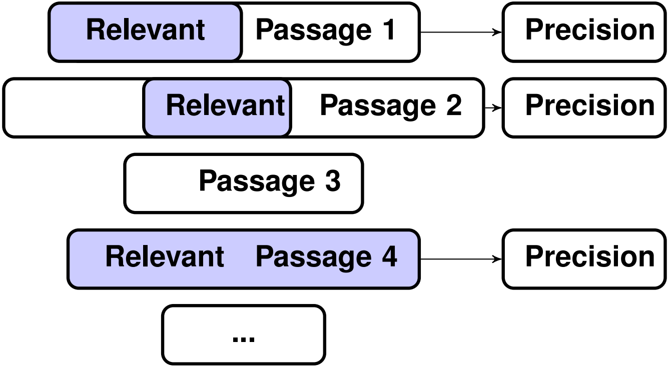


where:

$$\mathit{Precision} = \frac{\mathit{Length\ of\ the\ Relevant\ Part}}{\mathit{Length\ of\ the\ Whole\ Passage}}$$

Time-based Results Assessment Approach

For **each run** and **each query**:

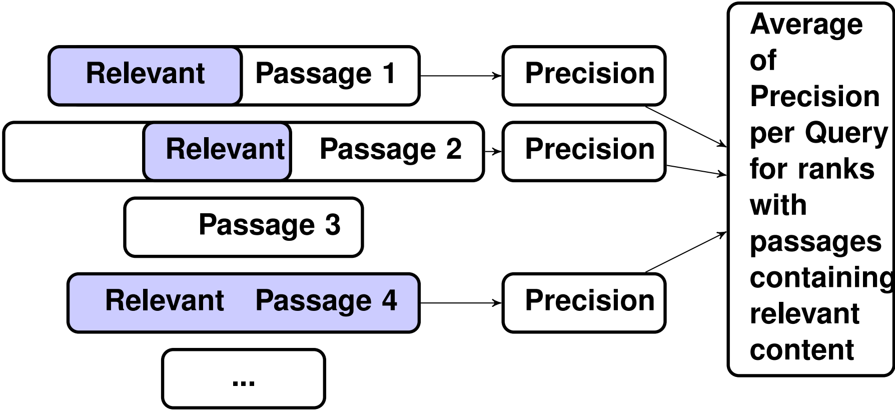


where:

$$Precision = \frac{Length\ of\ the\ Relevant\ Part}{Length\ of\ the\ Whole\ Passage}$$

Time-based Results Assessment Approach

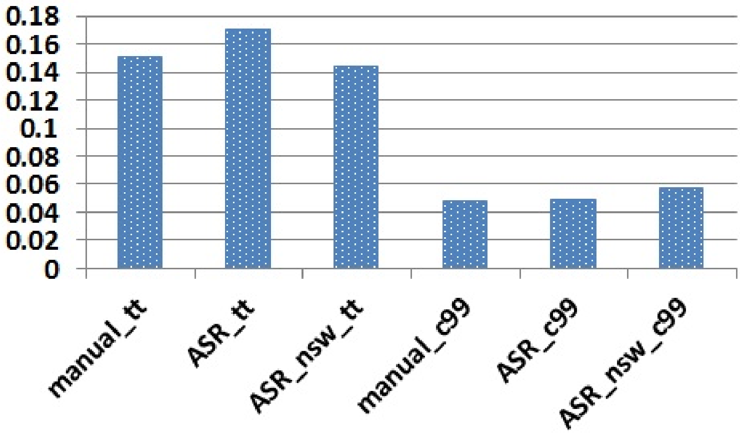
For **each run** and **each query**:



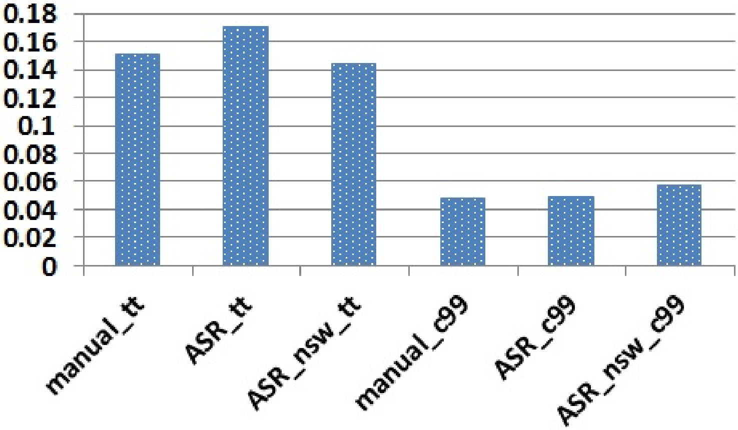
where:

$$Precision = \frac{Length\ of\ the\ Relevant\ Part}{Length\ of\ the\ Whole\ Passage}$$

Average of Precision for all passages with relevant content



Average of Precision for all passages with relevant content



- TextTiling algorithm has higher average of precision for all types of transcript, i.e.topically coherent segments are better located

Results: utterance-based MAP (uMAP)

Transcript type	Segmentation type	uMAP
BASELINE		0.0670
manual	tt	**0.0859**
manual	C99	0.0713
ASR	tt	**0.0490**
ASR	C99	0.0469
ASR_nsw	tt	**0.0312**
ASR_nsw	C99	0.0316

Results: utterance-based MAP (uMAP)

Transcript type	Segmentation type	uMAP
BASELINE		0.0670
manual	tt	**0.0859**
manual	C99	0.0713
ASR	tt	**0.0490**
ASR	C99	0.0469
ASR_nsw	tt	**0.0312**
ASR_nsw	C99	0.0316

- The trend ‘manual $>$ ASR $>$ ASR_nsw’ for both C99 and TextTiling is not proved by the averages of precision

Results: utterance-based MAP (uMAP)

Transcript type	Segmentation type	uMAP
BASELINE		0.0670
manual	tt	**0.0859**
manual	C99	0.0713
ASR	tt	**0.0490**
ASR	C99	0.0469
ASR_nsw	tt	**0.0312**
ASR_nsw	C99	0.0316

- The trend ‘manual > ASR > ASR_nsw’ for both C99 and TextTiling is not proved by the averages of precision
- Higher average values of the TextTiling segmentation over C99 are not reflected in the uMAP scores

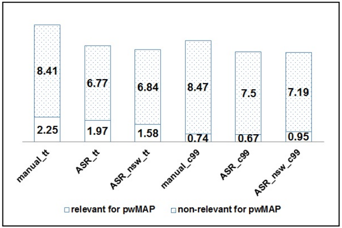
Results: utterance-based MAP (uMAP)

Transcript type	Segmentation type	uMAP
BASELINE		0.0670
manual	tt	**0.0859**
manual	C99	0.0713
ASR	tt	**0.0490**
ASR	C99	0.0469
ASR_nsw	tt	**0.0312**
ASR_nsw	C99	0.0316

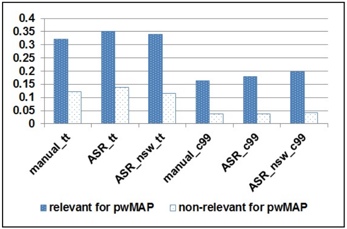
- The trend 'manual $>$ ASR $>$ ASR_nsw' for both C99 and TextTiling is not proved by the averages of precision
- Higher average values of the TextTiling segmentation over C99 are not reflected in the uMAP scores
- For some of the queries runs on C99 segmentation have better ranking of the segments with relevant content

Relevance of the Central IPU Assessment

Number of ranks
taken or not taken
into account by pwMAP

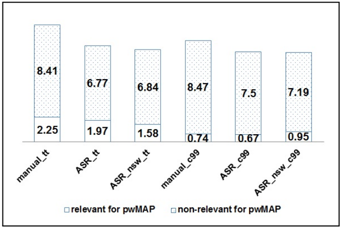


Average of Precision
for the passages at ranks
that are taken or not taken
into account by pwMAP

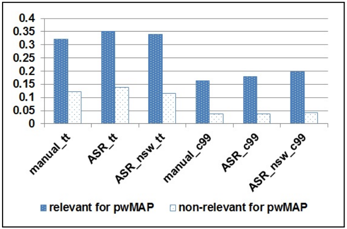


Relevance of the Central IPU Assessment

Number of ranks
taken or not taken
into account by pwMAP



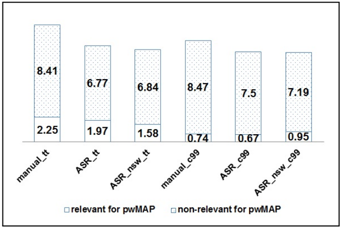
Average of Precision
for the passages at ranks
that are taken or not taken
into account by pwMAP



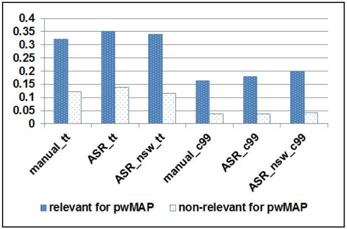
- TextTiling has higher numbers of segments that have central IPU relevant to the query

Relevance of the Central IPU Assessment

Number of ranks
taken or not taken
into account by pwMAP



Average of Precision
for the passages at ranks
that are taken or not taken
into account by pwMAP



- TextTiling has higher numbers of segments that have central IPU relevant to the query
- Overall the numbers of the ranks where the segment with relevant is retrieved is approximately the same for both segmentation techniques

Results: pointwise MAP (pwMAP)

Transcript type	Segmentation type	pwMAP
BASELINE		0.0520
manual	tt	**0.0429**
manual	C99	0.0209
ASR	tt	**0.0329**
ASR	C99	0.0166
ASR_nsw	tt	**0.0141**
ASR_nsw	C99	0.0138

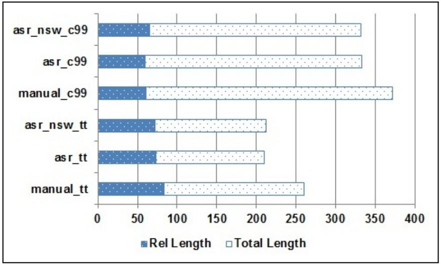
Results: pointwise MAP (pwMAP)

Transcript type	Segmentation type	pwMAP
BASELINE		0.0520
manual	tt	**0.0429**
manual	C99	0.0209
ASR	tt	**0.0329**
ASR	C99	0.0166
ASR_nsw	tt	**0.0141**
ASR_nsw	C99	0.0138

- TextTiling segmentation puts better topic boundaries for relevant content and have higher precision scores for the retrieved relevant passages

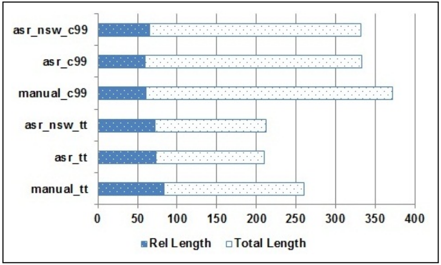
Average Length of Relevant Part and Segments (in seconds)

Center IPU is relevant



Average Length of Relevant Part and Segments (in seconds)

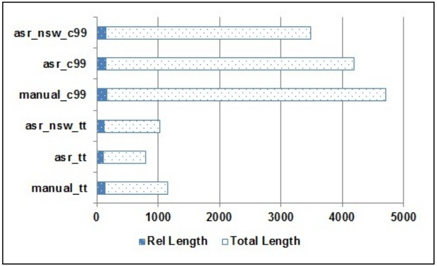
Center IPU is relevant



- **Center IPU is relevant:** Average length of the relevant content is of the same order for both segmentation schemes, slightly higher for TextTiling

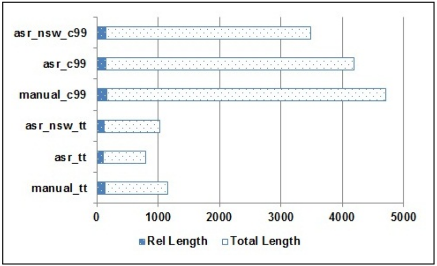
Average Length of Relevant Part and Segments (in seconds)

Center IPU is not relevant



Average Length of Relevant Part and Segments (in seconds)

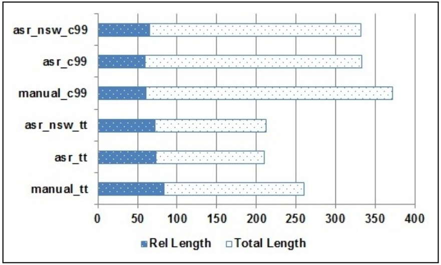
Center IPU is not relevant



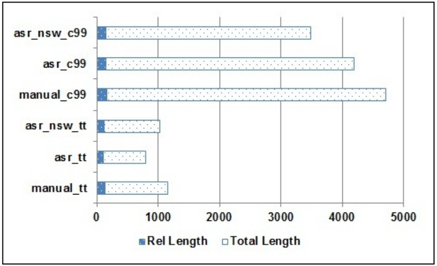
- **Center IPU is not relevant:** Average length of the relevant content is higher for C99 segmentation, due to the poor segmentation it correlates with much longer segments

Average Length of Relevant Part and Segments (in seconds)

Center IPU is relevant



Center IPU is not relevant



- **Center IPU is relevant:** Average length of the relevant content is of the same order for both segmentation schemes, slightly higher for TextTiling
- **Center IPU is not relevant:** Average length of the relevant content is higher for C99 segmentation, due to the poor segmentation it correlates with much longer segments

Results: fraction MAP (fMAP)

Transcript type	Segmentation type	fMAP
BASELINE		0.0536
manual	tt	**0.0500**
manual	C99	0.0168
ASR	tt	**0.0308**
ASR	C99	0.0123
ASR_nsw	tt	**0.0174**
ASR_nsw	C99	0.0120

Results: fraction MAP (fMAP)

Transcript type	Segmentation type	fMAP
BASELINE		0.0536
manual	tt	**0.0500**
manual	C99	0.0168
ASR	tt	**0.0308**
ASR	C99	0.0123
ASR_nsw	tt	**0.0174**
ASR_nsw	C99	0.0120

- Average number of ranks with segments having non-relevant center IPU is more than 5 times higher
- Segmentation technique with longer poor segmented passages (C99) has much lower precision-based scores

Outline

Retrieval Methodology
- Transcript Preprocessing
- Text Segmentation
- Retrieval Setup

Results
- Official Metrics
- uMAP
- pwMAP
- fMAP

Conclusions

Conclusions

- TextTiling segmentation shows better overall retrieval performance than C99:

Conclusions

- TextTiling segmentation shows better overall retrieval performance than C99:
  - Higher numbers of segments with higher precision

Conclusions

- TextTiling segmentation shows better overall retrieval performance than C99:
    - Higher numbers of segments with higher precision
    - Higher precision even for the segments with non-relevant center IPU

Conclusions

- TextTiling segmentation shows better overall retrieval performance than C99:
  - Higher numbers of segments with higher precision
  - Higher precision even for the segments with non-relevant center IPU
  - High level of poor segmentation makes it harder to retrieve relevant content for C99 runs

Conclusions

- TextTiling segmentation shows better overall retrieval performance than C99:
  - Higher numbers of segments with higher precision
  - Higher precision even for the segments with non-relevant center IPU
  - High level of poor segmentation makes it harder to retrieve relevant content for C99 runs
- Removal of stop words before segmentation did not have any positive effect on the results

Thank you for your attention!

Questions?