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

Indexed Segments

Queries

Retrieval

Information Request

Retrieval Results
Retrieval Methodology

Transcript → Segmentation → Topically Coherent Segments → Indexing → Indexed Segments → Queries → Information Request → Retrieval Results
Retrieval Methodology

Transcript

Segmentation

Indexed Segments

Queries

Information Request

Topically Coherent Segments

Indexing

Retrieval

Retrieval Results
6 Retrieval Runs

- 1-best ASR
- 1-best ASR with stop words removed
- Manual Transcript
6 Retrieval Runs

- 1-best ASR
- 1-best ASR with stop words removed
- Manual Transcript

- C99
- ASR_C99
- TT
- ASR_TT
6 Retrieval Runs

- 1-best ASR
- 1-best ASR with stop words removed
- Manual Transcript

1. **1-best ASR**
   - C99
   - ASR_C99
   - TT
   - ASR_TT

2. **1-best ASR with stop words removed**
   - C99
   - ASR_NS_C99
   - TT
   - ASR_NS_C99_TT

3. **Manual Transcript**
6 Retrieval Runs

1-best ASR

C99

ASR_C99

TT

ASR_TT

1-best ASR with stop words removed

C99

ASR_NSW_C99

TT

ASR_NSW_TT

Manual Transcript

C99

Manual_C99

TT

Manual_TT
6 Retrieval Runs

ASR_C99  ASR_NS_C99  Manual_C99

ASR_TT  ASR_NS_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

<table>
<thead>
<tr>
<th>Transcript type</th>
<th>Segmentation type</th>
<th>uMAP</th>
<th>pwMAP</th>
<th>fMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td></td>
<td>0.0670</td>
<td>0.0520</td>
<td>0.0536</td>
</tr>
<tr>
<td>manual</td>
<td>tt</td>
<td>0.0859</td>
<td>0.0429</td>
<td>0.0500</td>
</tr>
<tr>
<td>manual</td>
<td>C99</td>
<td>0.0713</td>
<td>0.0209</td>
<td>0.0168</td>
</tr>
<tr>
<td>ASR</td>
<td>tt</td>
<td>0.0490</td>
<td>0.0329</td>
<td>0.0308</td>
</tr>
<tr>
<td>ASR</td>
<td>C99</td>
<td>0.0469</td>
<td>0.0166</td>
<td>0.0123</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>tt</td>
<td>0.0312</td>
<td>0.0141</td>
<td>0.0174</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>C99</td>
<td>0.0316</td>
<td>0.0138</td>
<td>0.0120</td>
</tr>
</tbody>
</table>

- 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.
## Results: Official Metrics

<table>
<thead>
<tr>
<th>Transcript type</th>
<th>Segmentation type</th>
<th>uMAP</th>
<th>pwMAP</th>
<th>fMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td></td>
<td>0.0670</td>
<td>0.0520</td>
<td>0.0536</td>
</tr>
<tr>
<td>manual</td>
<td>tt</td>
<td>0.0859</td>
<td>0.0429</td>
<td>0.0500</td>
</tr>
<tr>
<td>manual</td>
<td>C99</td>
<td>0.0713</td>
<td>0.0209</td>
<td>0.0168</td>
</tr>
<tr>
<td>ASR</td>
<td>tt</td>
<td>0.0490</td>
<td>0.0329</td>
<td>0.0308</td>
</tr>
<tr>
<td>ASR</td>
<td>C99</td>
<td>0.0469</td>
<td>0.0166</td>
<td>0.0123</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>tt</td>
<td>0.0312</td>
<td>0.0141</td>
<td>0.0174</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>C99</td>
<td>0.0316</td>
<td>0.0138</td>
<td>0.0120</td>
</tr>
</tbody>
</table>

- Only runs on the manual transcript had higher scores than the baseline (only uMAP metric)
## Results: Official Metrics

<table>
<thead>
<tr>
<th>Transcript type</th>
<th>Segmentation type</th>
<th>uMAP</th>
<th>pwMAP</th>
<th>fMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td></td>
<td>0.0670</td>
<td>0.0520</td>
<td>0.0536</td>
</tr>
<tr>
<td>manual</td>
<td>tt</td>
<td><strong>0.0859</strong></td>
<td><strong>0.0429</strong></td>
<td><strong>0.0500</strong></td>
</tr>
<tr>
<td>manual</td>
<td>C99</td>
<td>0.0713</td>
<td>0.0209</td>
<td>0.0168</td>
</tr>
<tr>
<td>ASR</td>
<td>tt</td>
<td><strong>0.0490</strong></td>
<td><strong>0.0329</strong></td>
<td><strong>0.0308</strong></td>
</tr>
<tr>
<td>ASR</td>
<td>C99</td>
<td>0.0469</td>
<td>0.0166</td>
<td>0.0123</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>tt</td>
<td>0.0312</td>
<td>0.0141</td>
<td>0.0174</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>C99</td>
<td>0.0316</td>
<td>0.0138</td>
<td>0.0120</td>
</tr>
</tbody>
</table>

- 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:

- **Relevant** Passage 1
- **Relevant** Passage 2
- Passage 3
- **Relevant** Passage 4
- ...

where:

\[
Precision = \frac{\text{Length of the Relevant Part}}{\text{Length of the Whole Passage}}
\]
Time-based Results Assessment Approach

For each run and each query:

- Relevant Passage
- Precision

where:

$$\text{Precision} = \frac{\text{Length of the Relevant Part}}{\text{Length of the Whole Passage}}$$
Time-based Results Assessment Approach

For each run and each query:

For each run and each query:

- Passage 1
  - Relevant
  - Precision

- Passage 2
  - Relevant
  - Precision

- Passage 3
  - Relevant

- Passage 4
  - Relevant
  - Precision

... (more passages)

Where:

\[
\text{Precision} = \frac{\text{Length of the Relevant Part}}{\text{Length of the Whole Passage}}
\]
Average of Precision for all passages with relevant content

![Bar Chart]

- **manual_tt**
- **ASR_tt**
- **ASR_nsw_tt**
- **manual_c99**
- **ASR_c99**
- **ASR_nsw_c99**

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

<table>
<thead>
<tr>
<th>Transcript type</th>
<th>Segmentation type</th>
<th>uMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td></td>
<td>0.0670</td>
</tr>
<tr>
<td>manual</td>
<td>tt</td>
<td>0.0859</td>
</tr>
<tr>
<td>manual</td>
<td>C99</td>
<td>0.0713</td>
</tr>
<tr>
<td>ASR</td>
<td>tt</td>
<td>0.0490</td>
</tr>
<tr>
<td>ASR</td>
<td>C99</td>
<td>0.0469</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>tt</td>
<td>0.0312</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>C99</td>
<td>0.0316</td>
</tr>
</tbody>
</table>

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.
### Results: utterance-based MAP (uMAP)

<table>
<thead>
<tr>
<th>Transcript type</th>
<th>Segmentation type</th>
<th>uMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td></td>
<td>0.0670</td>
</tr>
<tr>
<td>manual</td>
<td>tt</td>
<td>0.0859</td>
</tr>
<tr>
<td>manual</td>
<td>C99</td>
<td>0.0713</td>
</tr>
<tr>
<td>ASR</td>
<td>tt</td>
<td>0.0490</td>
</tr>
<tr>
<td>ASR</td>
<td>C99</td>
<td>0.0469</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>tt</td>
<td>0.0312</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>C99</td>
<td>0.0316</td>
</tr>
</tbody>
</table>

- The trend ‘manual > ASR > ASR_nsw’ for both C99 and TextTiling is not proved by the averages of precision.
Results: utterance-based MAP (uMAP)

<table>
<thead>
<tr>
<th>Transcript type</th>
<th>Segmentation type</th>
<th>uMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td></td>
<td>0.0670</td>
</tr>
<tr>
<td>manual</td>
<td>tt</td>
<td>0.0859</td>
</tr>
<tr>
<td>manual</td>
<td>C99</td>
<td>0.0713</td>
</tr>
<tr>
<td>ASR</td>
<td>tt</td>
<td>0.0490</td>
</tr>
<tr>
<td>ASR</td>
<td>C99</td>
<td>0.0469</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>tt</td>
<td>0.0312</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>C99</td>
<td>0.0316</td>
</tr>
</tbody>
</table>

- 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)

<table>
<thead>
<tr>
<th>Transcript type</th>
<th>Segmentation type</th>
<th>uMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td></td>
<td>0.0670</td>
</tr>
<tr>
<td>manual</td>
<td>tt</td>
<td>0.0859</td>
</tr>
<tr>
<td>manual</td>
<td>C99</td>
<td>0.0713</td>
</tr>
<tr>
<td>ASR</td>
<td>tt</td>
<td>0.0490</td>
</tr>
<tr>
<td>ASR</td>
<td>C99</td>
<td>0.0469</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>tt</td>
<td>0.0312</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>C99</td>
<td>0.0316</td>
</tr>
</tbody>
</table>

- 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
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)

<table>
<thead>
<tr>
<th>Transcript type</th>
<th>Segmentation type</th>
<th>pwMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td></td>
<td>0.0520</td>
</tr>
<tr>
<td>manual</td>
<td>tt</td>
<td>0.0429</td>
</tr>
<tr>
<td>manual</td>
<td>C99</td>
<td>0.0209</td>
</tr>
<tr>
<td>ASR</td>
<td>tt</td>
<td>0.0329</td>
</tr>
<tr>
<td>ASR</td>
<td>C99</td>
<td>0.0166</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>tt</td>
<td>0.0141</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>C99</td>
<td>0.0138</td>
</tr>
</tbody>
</table>

TextTiling segmentation puts better topic boundaries for relevant content and have higher precision scores for the retrieved relevant passages.
Results: pointwise MAP (pwMAP)

<table>
<thead>
<tr>
<th>Transcript type</th>
<th>Segmentation type</th>
<th>pwMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td></td>
<td>0.0520</td>
</tr>
<tr>
<td>manual</td>
<td>tt</td>
<td>0.0429</td>
</tr>
<tr>
<td>manual</td>
<td>C99</td>
<td>0.0209</td>
</tr>
<tr>
<td>ASR</td>
<td>tt</td>
<td>0.0329</td>
</tr>
<tr>
<td>ASR</td>
<td>C99</td>
<td>0.0166</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>tt</td>
<td>0.0141</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>C99</td>
<td>0.0138</td>
</tr>
</tbody>
</table>

- 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

- `asr_nsw_c99`
- `asr_c99`
- `manual_c99`
- `asr_nsw_tt`
- `asr_tt`
- `manual_tt`

![Diagram showing average length of relevant content](image)
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

![Bar chart showing average length of relevant content compared to total length for different segmentation schemes.](chart.png)
Average Length of Relevant Part and Segments (in seconds)

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:

- 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)

<table>
<thead>
<tr>
<th>Transcript type</th>
<th>Segmentation type</th>
<th>fMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td></td>
<td>0.0536</td>
</tr>
<tr>
<td>manual</td>
<td>tt</td>
<td>0.0500</td>
</tr>
<tr>
<td>manual</td>
<td>C99</td>
<td>0.0168</td>
</tr>
<tr>
<td>ASR</td>
<td>tt</td>
<td>0.0308</td>
</tr>
<tr>
<td>ASR</td>
<td>C99</td>
<td>0.0123</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>tt</td>
<td>0.0174</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>C99</td>
<td>0.0120</td>
</tr>
</tbody>
</table>

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.
Results: fraction MAP (fMAP)

<table>
<thead>
<tr>
<th>Transcript type</th>
<th>Segmentation type</th>
<th>fMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td></td>
<td>0.0536</td>
</tr>
<tr>
<td>manual</td>
<td>tt</td>
<td>0.0500</td>
</tr>
<tr>
<td>manual</td>
<td>C99</td>
<td>0.0168</td>
</tr>
<tr>
<td>ASR</td>
<td>tt</td>
<td>0.0308</td>
</tr>
<tr>
<td>ASR</td>
<td>C99</td>
<td>0.0123</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>tt</td>
<td>0.0174</td>
</tr>
<tr>
<td>ASR_nsw</td>
<td>C99</td>
<td>0.0120</td>
</tr>
</tbody>
</table>

- 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?