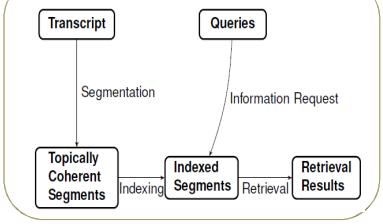
DCU at the NTCIR-9 SpokenDoc Passage Retrieval Task

Maria Eskevich, Gareth J. F. Jones
Centre for Digital Video Processing,
Centre for Next Generation Localisation
School of Computing, Dublin City University, Ireland

Retrieval Methodology

Overview

- Investigate application of content-based segmentation for spoken passage retrieval
- Segmentation using standard TextTiling and C99 algorithms from text
- Standard Japanese text processing applied with language modelling information



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
- Remove the stop words (SpeedBlog Japanese Stop-words) for one of the runs (NSW)

Segmentation

Transcripts are segmented using either:

TextTiling (TT):

- Cosine similarities between adjacent blocks of sentences

C99:

- Similarity between sentences cashed using a cosine similarity measure to form a similarity matrix
- Cosine scores replaced by the rank of the score in the local region
- Segmentation points assigned using a clustering procedure

Retrieval System

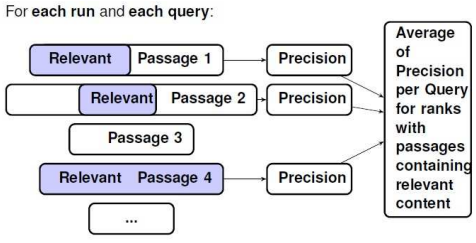
SMART information retrieval system extended to use language modelling with a uniform document prior probability

Retrieval Results

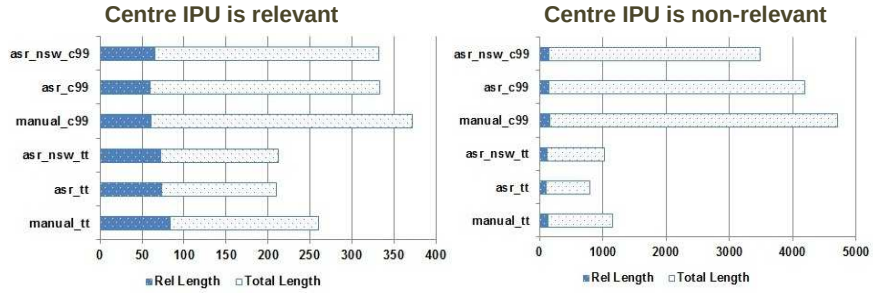
Transcript Type	Segmentation Type	uMAP	pwMAP	fMAP
BASELINE		0.0670	0.0520	0.0536
Manual	TT	0.0859	0.0429	0.0500
Manual	C99	0.0713	0.0209	0.0168
ASR	TT	0.0490	0.0329	0.0308
ASR	C99	0.0469	0.0166	0.0123
ASR_NSW	TT	0.0312	0.0141	0.0174
ASR_NSW	C99	0.0316	0.0138	0.0120

Results Analysis

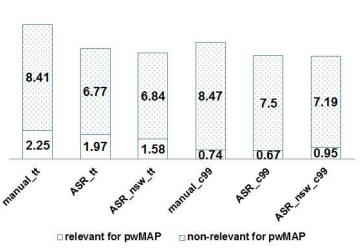
Calculation of Average of Precision (sec)



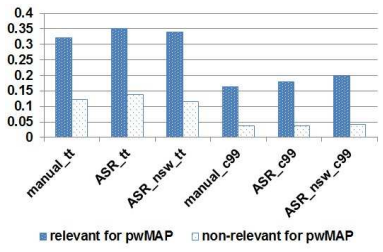
Average Length of Relevant Part and Segments



Number of ranks with centre IPU being relevant or not



Average of Precision for passages with centre IPU being relevant or not



Conclusions

- Only runs on the manual transcript had higher scores than the baseline (uMAP metric only)
- TextTiling results are consistently higher than C99 for all the metrics for manual and ASR runs
- TextTiling has higher average of precision (in seconds) for all types of transcript, i.e. it locates topically coherent segments better
- 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

This work is funded by a grant under the Science Foundation Ireland Research Frontiers Programme 2008 Grant No: 08/RFP/CMS1677
Further information: http://www.cdvp.dcu.ie/IISSCoS/