JRIRD at the NTCIR-15 QA Lab-PoliInfo-2 Task: An Abstractive Dialog Summarization System for Japanese Assembly Minutes

Kazuma Kadowaki The Japan Research Institute, Limited

Abstract

I developed a system for topic-aware summarization of assembly member speeches. It consists of:

(1) a pre-processor

(2) a BERT-based sentence extractor (that predicts a topic-aware importance of each sentence); and

(3) a UniLM-based summary generator (whose summary length is controllable).

My model achieved the best performance among all the participants in the Dialog Summarization subtask.

INTRODUCTION

Purpose

Generate a *topic-aware* short summary of Tokyo Metropolitan Assembly minutes, in order to fact-check and to understand speakers' policies

Task

Input: speaker's name, entire speech, topic, desired length

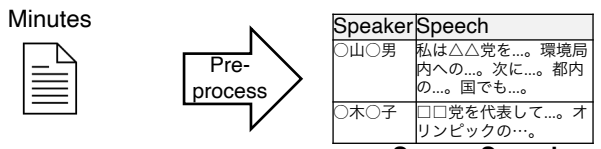
Output: a topic-aware summary of the speech

Challenges

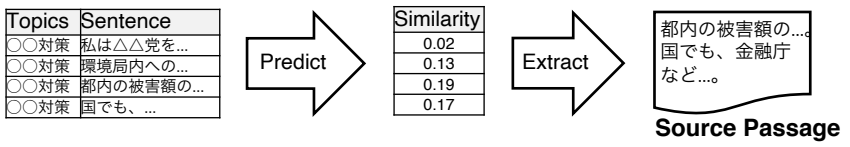
- Very long, multiple-topic speeches
- Minutes without annotation (not segmented, no importance scores)
- Maximum numbers of characters specified for each summary

MY APPROACH

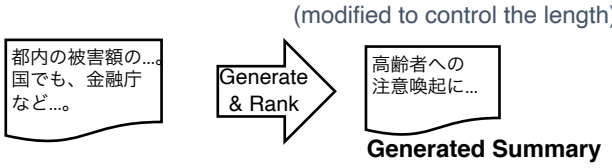
Pre-processor — Retrieves an entire "source speech"



Sentence Extractor — Extracts a "source passage" by predicting ROUGE-1 scores using BERT-based regression



Summary Generator — Generates an abstractive summary from the passage using UniLM (modified to control the length)



RESULTS

Models submitted

- ID 185: trained only using the datasets from the task organizers
- ID 189: trained also using my own dataset from different years

Results

- Achieved better performance in most of the metrics
- Adding my dataset further contributed to the performance

	Content		Well-formed	Non-twisted		Sentence goodness	Dialog goodness
	X = 2	X = 0		All	Evaluable		
ID 185	1.014	0.900	1.830	1.220	1.581	1.042	0.848
ID 189	**1.082**	**0.975**	**1.858**	**1.316**	**1.712**	**1.129**	**0.937**
Baseline	0.748	0.671	1.582	1.011	1.658	0.730	0.488

		Recall							F-measure						
		N1	N2	N3	N4	L	SU4	W1.2	N1	N2	N3	N4	L	SU4	W1.2
Surface Form	ID 185	0.503	0.221	0.134	0.087	0.415	0.252	0.199	0.373	0.158	0.096	0.061	0.303	0.174	0.193
	ID 189	**0.517**	**0.241**	**0.146**	**0.093**	**0.429**	**0.267**	**0.206**	**0.387**	**0.175**	**0.106**	**0.069**	**0.317**	**0.188**	**0.202**
	Baseline	0.405	0.130	0.076	0.046	0.338	0.169	0.160	0.308	0.099	0.058	0.036	0.253	0.123	0.159
Stem	ID 185	0.511	0.224	0.137	0.091	0.421	0.258	0.202	0.379	0.161	0.098	0.064	0.308	0.178	0.196
	ID 189	**0.526**	**0.247**	**0.152**	**0.098**	**0.437**	**0.277**	**0.210**	**0.394**	**0.180**	**0.110**	**0.073**	**0.323**	**0.194**	**0.206**
	Baseline	0.425	0.144	0.087	0.055	0.355	0.185	0.171	0.323	0.109	0.066	0.042	0.266	0.134	0.169
Content Word	ID 185	0.298	0.128	0.061	0.024	0.281	0.154	0.180	0.215	0.090	0.041	0.017	0.202	0.091	0.158
	ID 189	**0.321**	**0.149**	**0.077**	**0.034**	**0.302**	**0.171**	**0.192**	**0.237**	**0.109**	**0.056**	**0.027**	**0.222**	**0.106**	**0.172**
	Baseline	0.244	0.105	0.051	0.024	0.233	0.123	0.150	0.185	0.079	0.038	0.019	0.177	0.080	0.139

DISCUSSIONS

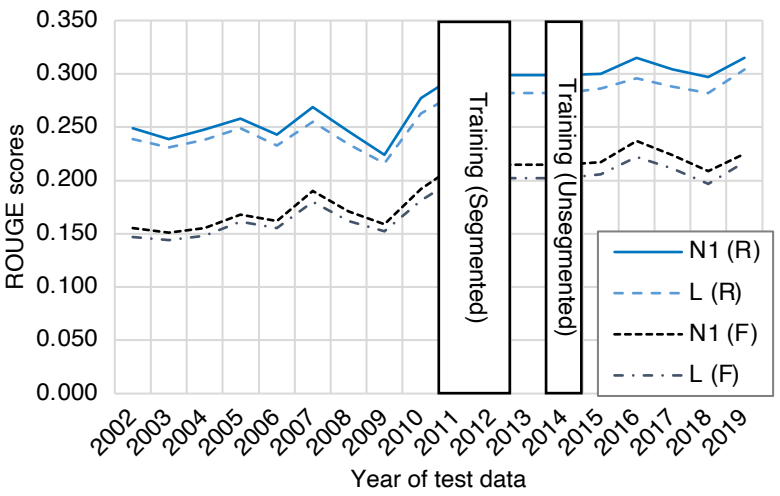
Performance of each module

- Each module extracted 51.7% and 75.7% of the available content words successfully

Text	Output by	Recall w.r.t. ref. summaries	Characters per summary
Source speech	Pre-processor	0.818	4,895.59
Source passage	Sentence extractor	0.423	117.65
Generated summary	Summary generator	0.320	57.76
Reference summary	-	-	38.69

Model generalization

- Robust enough for changes in topics discussed
- Future work: Mitigate/detect performance degradation



Human evaluation

- No system seems to be always helpful to fact-check
- Future work: Revise the task settings

	Content	Well-formed	Non-twisted	Sentence goodness	Dialog goodness
Grade A	29.7%	88.0%	60.5%	42.8%	29.3%
Grade B	38.0%	9.8%	10.6%	27.3%	35.0%
Grade C	26.9%	2.2%	28.9%	29.9%	35.6%
Grade X	5.3%				

CONCLUSIONS

Contributions

- I developed an assembly minutes summarizer, which consists of a BERT-based extractor and a UniLM-based generator
- My models achieved the best performance, and would generalize for future meetings
- The length of a generated summary can be controlled

Future work

- Add a mechanism to consider a context
- Apply my models to other real-world tasks (including business conversations)
- Revise the task settings for fact-checking
- Investigate summarization from noisy minutes generated by ASR systems

Copyright © 2020 The Japan Research Institute, Limited