Which Scores to Predict in Sentence Regression for Text Summarization?

TECHNISCHE UNIVERSITÄT DARMSTADT
AIPHES

Markus Zopf, Eneldo Loza Mencía, Johannes Fürnkranz | Research Training Group AIPHES

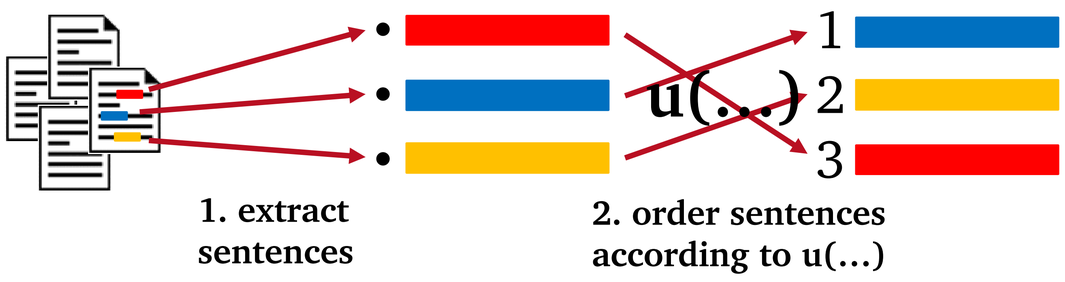
Summary

Sentence regression summarization system learn to predict ROUGE recall scores of individual sentences

We show that this choice leads to suboptimal performance

Learning to predict ROUGE precision scores of sentences performs consistently better in a wide range of experiments

Greedy Sentence Selection



3. select top n sentences (+ redundancy avoidance)

- function u(…) maps from sentences to utility scores
- sentence regression = learning function u(…)

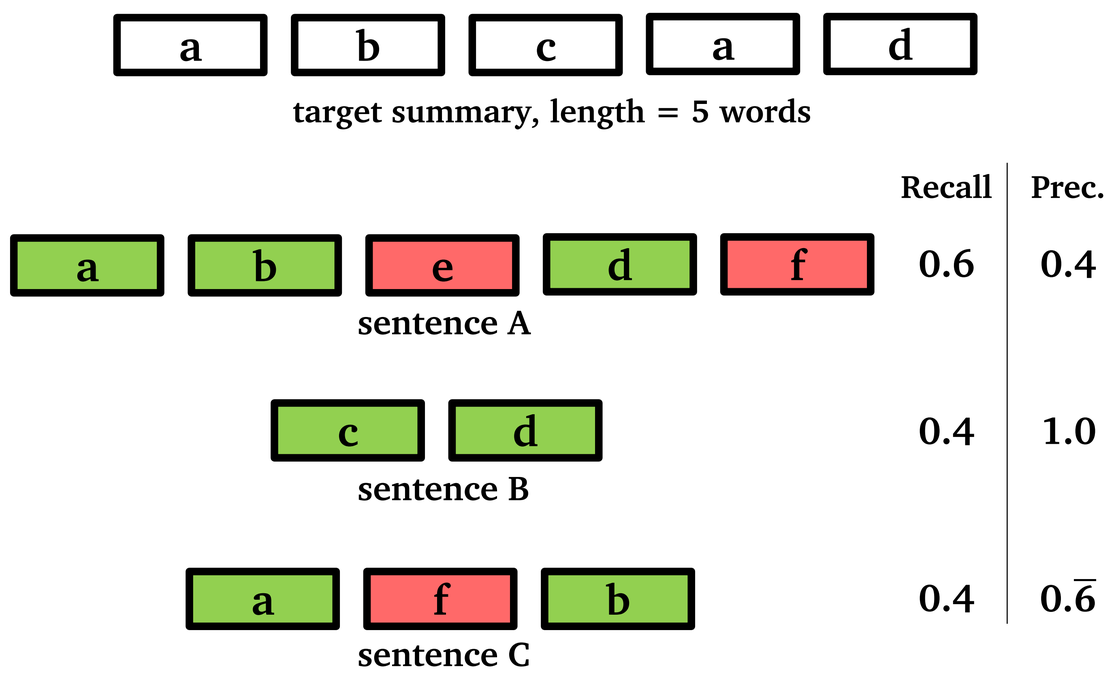
which u(…) should be learned?

- prior works learn to predict uni-/bigram recall of individual sentences since final summaries are evaluated with uni-/bigram recall (→ ROUGE-1/2)

Intuition: Precision vs. Recall

Hypotheses:

1. recall is biased towards long sentences
2. ordering according to precision leads to better summaries



in every step…

Recall:	Precision:
… put as much content as possible into summary	*… waste as little space as possible*
→ used by prior work	→ new paradigm

Experiments

Sentence Lengths

	tokens			sentences		
	D04	T08	T09	D04	T08	T09
R1 Rec	166	132	141	3.42	2.67	2.70
R2 Rec	160	129	132	4.26	3.46	3.55
R1 Prec	157	125	127	7.76	6.75	6.07
R2 Prec	157	129	126	7.10	6.13	6.09
max ADW	158	127	129	6.56	5.06	5.11
avg ADW	158	126	126	5.12	4.13	4.02
random	164	131	131	6.66	5.21	4.89

average number of tokens and sentences in the resulting summaries

Conclusion:

- ROUGE recall-based selection is biased towards long sentences

Optimal Prediction on German non-newswire data

	DBS		hMDS	
	R1	R2	R1	R2
R1 Rec	33.48	13.89	31.94	13.38
R2 Rec	38.67	21.77	40.67	24.39
R1 Prec	**42.20**	**25.55**	**43.25**	23.01
R2 Prec	37.01	23.12	41.65	**24.96**
random	23.27	04.23	20.63	02.36

ROUGE results for 2 German non-newswire corpora

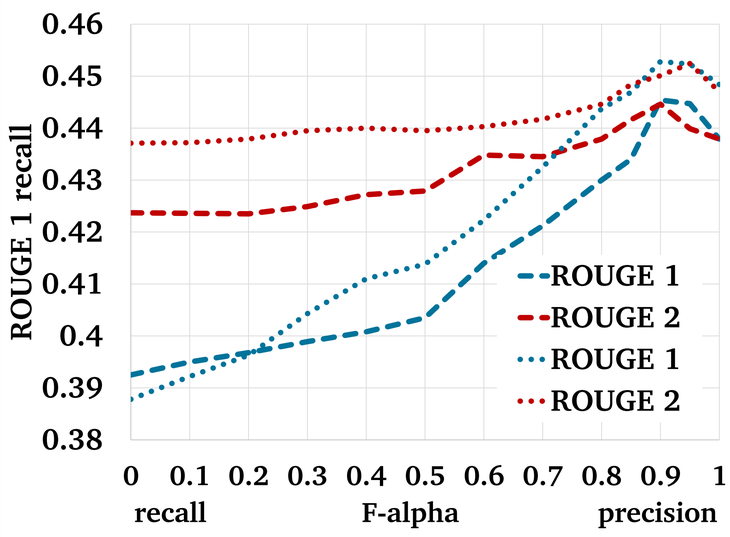
Conclusion:

- Selecting sentences according to ROUGE precision leads to better summaries

Optimal Prediction without Redundancy Avoidance

	DUC 2004		TAC 2008		TAC 2009	
	R1	R2	R1	R2	R1	R2
R1 rec	38.63	08.99	39.28	11.08	34.31	08.37
R2 rec	39.23	12.07	42.39	16.20	37.42	13.03
R1 prec	**41.29**	11.18	**43.56**	14.65	**39.45**	12.17
R2 prec	39.18	**12.73**	43.46	**18.19**	37.81	**13.64**
max ADW	37.60	10.13	42.55	15.46	34.56	11.05
avg ADW	38.50	09.62	40.97	12.43	35.48	09.34
random	31.76	04.66	29.58	04.60	29.88	04.63

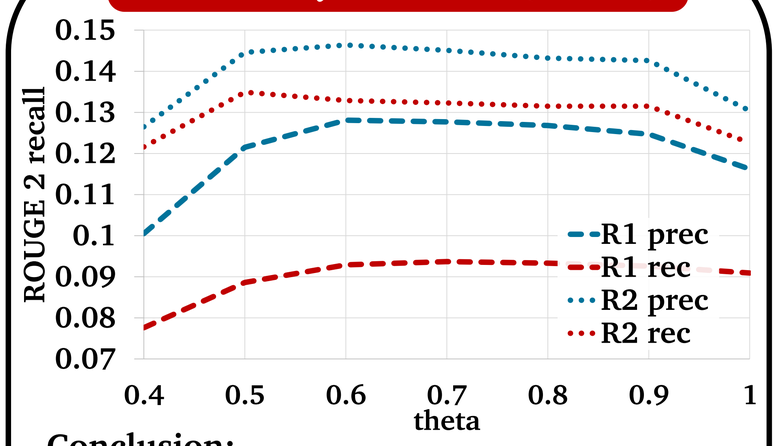
ROUGE recall scores of summaries in 3 English newswire corpora produced with different sentence selection scores



Conclusions:

- Selecting sentences according to ROUGE precision leads to better summaries
- Small fraction of recall improves results

Optimal Prediction With Redundancy Avoidance (DUC04)



Conclusion:

- Precision also leads to better results if redundancy avoidance is applied

Noisy Score Prediction

		DUC 2004		TAC 2008		TAC 2009	
		R1	R2	R1	R2	R1	R2
Uniform	R1 rec	36.78	07.43	35.70	08.00	36.04	08.27
	R2 rec	35.45	07.54	34.62	08.58	36.08	09.43
	R1 prec	**42.02**	10.45	**41.42**	11.75	**42.75**	12.83
	R2 prec	39.56	**11.16**	38.94	**12.64**	40.91	**14.29**
Gaussian	R1 rec	35.63	06.83	34.45	07.31	35.06	07.57
	R2 rec	33.39	06.04	32.76	06.93	32.88	07.98
	R1 prec	**41.70**	10.19	**41.41**	12.09	**43.06**	13.23
	R2 prec	38.41	**10.33**	38.27	**12.43**	40.15	**13.94**

ROUGE recall results for noisy score prediction

Conclusion:

- Precision also performs better, if scores can only be predicted approximately

This work has been supported by the German Research Foundation (DFG) as part of the Research Training Group Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) under grant No. GRK 1994/1

www.aiphes.tu-darmstadt.de