Gold Summary: Redpath has ended his eight-year association with Sale Sharks. Redpath spent five years as a player and three as a coach at sale. He has thanked the owners, coaches and players for their support.	Salience	Content	Novelty	Position	Prob.
Bryan Redpath has left his coaching role at Sale Sharks with immediate effect.	0.1	0.1	0.9	0.1	0.3
The 43 - year - old Scot ends an eight-year association with the Aviva Premiership side, having spent five years with them as a player and three as a coach.	0.9	0.6	0.9	0.9	0.7
Redpath returned to Sale in June 2012 as director of rugby after starting a coaching career at Gloucester and progressing to the top job at Kingsholm .	0.8	0.5	0.5	0.9	0.6
Redpath spent five years with Sale Sharks as a player and a further three as a coach but with Sale Sharks struggling four months into Redpath's tenure, he was removed from the director of rugby role at the Salford-based side and has since been operating as head coach .	0.8	0.9	0.7	0.8	**0.9**
'I would like to thank the owners, coaches, players and staff for all their help and support since I returned to the club in 2012.	0.4	0.1	0.1	0.7	0.2
Also to the supporters who have been great with me both as a player and as a coach,' Redpath said.	0.6	0.0	0.2	0.3	0.2

Figure 2: Visualization of SummaRuNNer output on a representative document. Each row is a sentence in the document, while the shading-color intensity is proportional to its probability of being in the summary, as estimated by the RNN-based sequence classifier. In the columns are the normalized scores from each of the abstract features in Eqn. (6) as well as the final prediction probability (last column). Sentence 2 is estimated to be the most salient, while the longest one, sentence 4, is considered the most content-rich, and not surprisingly, the first sentence the most novel. The third sentence gets the best position based score.

	Rouge-1	Rouge-2	Rouge-L
Lead-3	21.9	7.2	11.6
LReg(500)	18.5	6.9	10.2
Cheng *et al* '16	22.7	8.5	12.5
SummaRuNNer-abs	23.8	9.6	13.3
SummaRuNNer	**26.2**±0.4*	**10.8**±0.3*	**14.4**±0.3*

Table 1: Performance of various models on the **entire Daily Mail test set** using the **limited length recall** variants of Rouge with respect to the abstractive ground truth at **75 bytes**. Entries with asterisk are statistically significant using 95% confidence interval with respect to the nearest model, as estimated by the Rouge script.

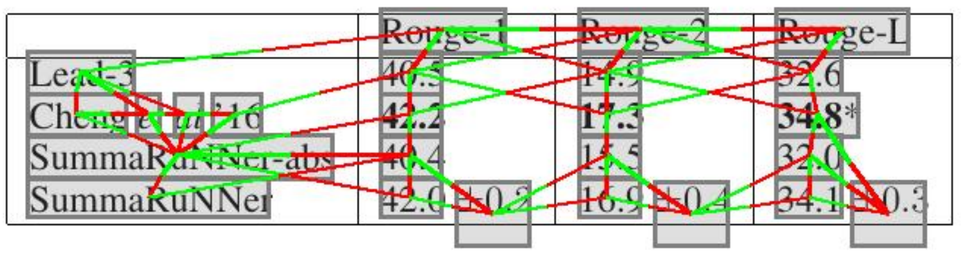

Table 2: Performance of various models on the **entire Daily Mail test set** using the **limited length recall** variants of Rouge at **275 bytes**. SummaRuNNer is statistically indistinguishable from the model of (Cheng and Lapata 2016) at 95% C.I. on Rouge-1 and Rouge-2.

One potential reason SummaRuNNer does not consistently outperform the extractive model of (Cheng and Lapata 2016) is the additional supervised training they used to create sentence-level extractive labels to train their model. Our model instead uses an unsupervised greedy approximation to create extractive labels from abstractive summaries, and as a result, may be more noisy than their ground truth.

We also notice that the abstractively trained SummaRuNNer underperforms its extractive counterpart. Abstractive training is more difficult since the sequence classifier is trained implicitly through the decoder which in turn depends only on the summary representation. In the future, we will investigate better design and training mechanism for the abstractive version.

4.6 Results on CNN/Daily Mail corpus

We also report the performance of SummaRuNNer on the joint CNN/Daily Mail corpus. The only other work that reports performance on this dataset is the abstractive encoder-decoder based model of (Nallapati et al. 2016), in which they use full-length F1 as the metric since neural abstractive approaches can learn when to stop generating words in the summary. In order to do a fair comparison with their work, we use the same metric as them. On this dataset, SummaRuNNer significantly outperforms their model as shown in Table 3. The superior performance of our model is not entirely surprising since abstractive summarization is a much harder problem, but the table serves to quantify the current performance gap between extractive and abstractive approaches to summarization. The results also demonstrate the difficulty of using the F1 metric for extractive summarization since SummaRuNNer, with its top three sentences with highest prediction probability as the summary, errs on the side of high recall at the expense of precision. Dynamically adjusting the summary length based on predicted probability distribution may help balance precision and recall and may further boost F1 performance, but we have not experimented with it in this work.